



Probabilistische Datenwissenschaft für die Psychologie

Dirk Ostwald

Inhaltsverzeichnis

Willkommen	3
Zitation	3
Lizenz	3
I. Mathematische Grundlagen	4
1. Sprache und Logik	6
1.1. Mathematik ist eine Sprache	6
1.2. Grundbausteine	7
1.3. Aussagenlogik	8
1.4. Äquivalenzumformungen	13
1.5. Beweistechniken	15
2. Mengen	16
2.1. Grundlegende Definitionen	16
2.2. Verknüpfungen	19
2.3. Spezielle Mengen	21
3. Summen, Produkte, Potenzen	25
3.1. Summen	25
3.2. Produkte	28
3.3. Potenzen	30
4. Funktionen	32
4.1. Definition und Eigenschaften	32
4.2. Funktionentypen	37
4.3. Elementare Funktionen	41
5. Folgen, Grenzwerte, Stetigkeit	45
5.1. Folgen	45
5.2. Grenzwerte	47
5.3. Stetigkeit	50
6. Differenzialrechnung	52
6.1. Definitionen und Rechenregeln	52
6.2. Analytische Optimierung	57
6.3. Differenzialrechnung multivariater reellwertiger Funktionen	60
7. Integralrechnung	69
7.1. Unbestimmte Integrale	69
7.2. Bestimmte Integrale	71
7.3. Uneigentliche Integrale	79

7.4. Mehrdimensionale Integrale	80
8. Vektoren	82
8.1. Reeller Vektorraum	82
8.2. Euklidischer Vektorraum	87
8.3. Lineare Unabhängigkeit	93
8.4. Vektorraumbasen	95
9. Matrizen	99
9.1. Definition	99
9.2. Grundlegende Matrixoperationen	101
9.3. Matrixmultiplikation	106
9.4. Matrixinversion	110
9.5. Determinanten	113
9.6. Spezielle Matrizen	117
9.7. Eigenanalyse	121
9.8. Orthonormalzerlegung	127
9.9. Singulärwertzerlegung	129
10. Deskriptivstatistiken	131
10.1. Häufigkeitsverteilungen	133
10.2. Histogramme	135
10.3. Empirische Verteilungsfunktionen	138
10.4. Quantile und Boxplots	141
10.5. Maße der Zentralen Tendenz	144
10.6. Maße der Variabilität	153
II. Imperative Programmierung	160
11. Informatische Grundbegriffe	163
11.1. Rechnerarchitektur	164
11.2. Algorithmen und Programme	165
11.3. Programmiersprachen	166
11.4. R	169
11.5. Visual Studio Code	170
12. Arithmetik und Variablen	171
12.1. Arithmetik	171
12.2. Präzedenz	174
12.3. Variablen	177
13. Datenstrukturen	184
13.1. Datenstrukturen in R	185
14. Vektoren	189
14.1. Erzeugung	189
14.2. Charakterisierung	194
14.3. Indizierung	196
14.4. Arithmetik numerischer Vektoren	199
14.5. Attribute	201

15. Matrizen und Arrays	203
15.1. Matrizen	203
15.2. Arrays	209
16. Listen, Dataframes und Tibbles	212
16.1. Listen	212
16.2. Dataframes	218
16.3. Tibbles	222
17. Datenmanagement	226
17.1. Das FAIR Datenideal	226
17.2. Dateiformate	228
17.3. Verzeichnismangement	230
17.4. Datenimport und Datenexport	233
18. Kontrollstrukturen	235
18.1. Konditionale Kontrollstrukturen	235
18.2. Iterative Kontrollstrukturen	239
III. Wahrscheinlichkeitstheorie	242
19. Wahrscheinlichkeitsräume	246
19.1. Definition und erste Eigenschaften	246
19.2. Wahrscheinlichkeitsfunktionen	251
19.3. Beispiele bei endlichem Ergebnisraum	252
19.4. Literaturhinweise	256
20. Wahrscheinlichkeiten	257
20.1. Gemeinsame Wahrscheinlichkeiten	257
20.2. Bedingte Wahrscheinlichkeiten	259
20.3. Unabhängige Ereignisse	263
20.4. Literaturhinweise	264
21. Zufallsvariablen	265
21.1. Definition	265
21.2. Wahrscheinlichkeitsmassefunktionen	270
21.3. Wahrscheinlichkeitsdichtefunktionen	273
21.4. Kumulative Verteilungsfunktionen	277
21.5. Zufallsergebnisse und Zufallsvariablen	283
21.6. Literaturhinweise	284
22. Zufallsvektoren	285
22.1. Definition und multivariate Verteilungen	285
22.2. Marginalverteilungen	289
22.3. Bedingte Verteilungen	291
22.4. Unabhängige Zufallsvariablen	293
23. Erwartungswerte	297
23.1. Definition	297
23.2. Eigenschaften	299

23.3. Bedingter Erwartungswert	300
24. Varianzen	303
24.1. Definition	303
24.2. Eigenschaften	305
24.3. Bedingte Varianz	306
25. Kovarianzen	308
25.1. Definition	308
25.2. Eigenschaften	309
25.3. Bedingte Kovarianz	314
25.4. Kovarianzmatrizen	315
26. Ungleichungen	319
26.1. Wahrscheinlichkeitsungleichungen	319
26.2. Erwartungswertungleichungen	321
27. Grenzwerte	325
27.1. Gesetze der Großen Zahlen	325
27.2. Zentrale Grenzwertsätze	326
27.3. Literaturhinweise	329
28. Transformationen	330
28.1. Univariate Transformationstheoreme	332
28.2. Multivariate WDF Transformationstheoreme	334
28.3. Operationstheoreme	335
29. Normalverteilungen	336
29.1. Konstruktion multivariater Normalverteilungen	336
29.2. Definition multivariater Normalverteilungen	339
29.3. Sphärizität	340
29.4. Marginale, bedingte und gemeinsame Verteilungen	341
29.5. Multivariate Transformationen	345
29.6. Univariate Transformationen	346
29.7. Literaturhinweise	360
IV. Inferenz	361
30. Frequentistische Inferenz	362
30.1. Frequentistische Inferenzmodelle	362
30.2. Statistiken und Schätzer	364
30.3. Standardannahmen und Standardproblemstellungen	366
31. Punktschätzung	370
31.1. Maximum-Likelihood Schätzung	371
31.2. Schätzereigenschaften bei endlichen Stichproben	375
31.3. Asymptotische Schätzereigenschaften	392
31.4. Eigenschaften von Maximum-Likelihood Schätzern	398
31.5. Literaturhinweise	399

32. Konfidenzintervalle	400
32.1. Definition	400
32.2. Beispiele für Konfidenzintervalle	401
32.3. Anwendungsbeispiel	413
32.4. Literaturhinweise	414
33. Hypothesentests	415
33.1. Testhypothesen und Tests	416
33.2. Testgütekriterien und Testkonstruktion	419
33.3. Testbeispiele	425
33.4. Konfidenzintervalle und Hypothesentests	440
33.5. Literaturhinweise	442
V. Allgemeines Lineares Modell	443
34. Regression	444
34.1. Methode der kleinsten Quadrate	445
34.2. Einfache lineare Regression	450
34.3. Literaturhinweise	455
35. Korrelation	456
35.1. Grundlagen	456
35.2. Korrelation und Bestimmtheitsmaß	460
35.3. Korrelation und linear-affine Abhängigkeit	464
35.4. Literaturhinweise	466
36. Modellformulierung	467
36.1. Allgemeine Theorie	467
36.2. Identifizierbarkeit und Schätzbarkeit	470
36.3. Designspektrum	473
36.4. Literaturhinweise	474
37. Parameterschätzung	475
37.1. Betaparameterschätzung	475
37.2. Varianzparameterschätzung	477
37.3. Unabhängig identisch normalverteilte Zufallsvariablen	478
37.4. Einfache lineare Regression	479
37.5. Frequentistische Schätzerverteilungen	482
37.6. Literaturhinweise	485
38. T-Statistiken	486
38.1. Definition und Beispiele	486
38.2. Konfidenzintervalle für Betaparameterkomponenten	490
38.3. Literaturhinweise	492
39. F-Statistiken	494
39.1. Likelihood-Quotienten-Statistiken	494
39.2. Definition und Verteilung	496
39.3. Literaturhinweise	500

40. T-Tests	502
40.1. Einstichproben-T-Tests	502
40.2. Modellevaluation	504
40.3. Zweistichproben-T-Tests	506
40.4. Literaturhinweise	511
41. Einfaktorielle Varianzanalyse	512
41.1. Anwendungsszenario	512
41.2. Anwendungsbeispiel	512
41.3. Modellformulierung	514
41.4. Modellschätzung	518
41.5. Modellevaluation	521
41.6. Literaturhinweise	531
42. Zweifaktorielle Varianzanalyse	532
42.1. Anwendungsszenario	532
42.2. Modellformulierung	536
42.3. Modellschätzung	544
42.4. Modellevaluation	549
42.5. Literaturhinweise	554
43. Partielle Korrelation	555
43.1. Motivation	555
43.2. Bedingte Korrelation	557
43.3. Bedingte Korrelation bei Normalverteilung	559
43.4. Partielle Korrelation	560
43.5. Literaturhinweise	563
44. Multiple Regression	564
44.1. Anwendungsszenario	564
44.2. Modellformulierung	566
44.3. Modellschätzung	566
44.4. Modellevaluation	573
44.5. Literaturhinweise	574
VI. Multivariate Inferenz	575
45. Multivariate Deskriptivstatistiken	576
45.1. Datenanalyseszenarien	576
45.2. Deskriptivstatistiken	577
45.3. Literaturhinweise	582
46. Einstichproben-T^2-Tests	583
46.1. Anwendungsszenario	583
46.2. Modellformulierung und Modellevaluation	584
46.3. Modellevaluation	585
46.4. Anwendungsbeispiel	598
46.5. Literaturhinweise	599

47. Einfaktorielle Varianzanalyse	600
47.1. Anwendungsszenario	600
47.2. Modellformulierung und Modellschätzung	602
47.3. Modellevaluation	604
47.4. Anwendungsbeispiel	613
47.5. Literaturhinweise	614
48. Kanonische Korrelationsanalyse	615
48.1. Anwendungsszenario und Grundzüge	615
48.2. Mathematischer Hintergrund	617
48.3. Modellformulierung	620
48.4. Modellschätzung	624
48.5. Anwendungsbeispiel	626
48.6. Literaturhinweise	628
VII. Psychometrische Verfahren	629
49. Psychologische Tests	630
49.1. Zum Begriff des psychologische Tests	630
49.2. Testgütekriterien	632
49.3. Normwerte	633
49.4. Testtheorien	635
50. Klassische Testtheorie	638
50.1. Modelle multipler Testmessungen	638
50.2. Reliabilität	655
50.3. Interne Konsistenz	663
51. Faktoranalyse	668
51.1. Faktoranalysemodelle	669
51.2. Schätzverfahren	678
51.3. Modellevaluation	686
51.4. Modellinterpretation	691
Referenzen	696

Willkommen

Herzlich willkommen zur Arbeitsversion von *Probabilistische Datenwissenschaft für die Psychologie (PDWP)*, einem Lehrbuch zur datenwissenschaftlichen Methodenlehre am Institut für Psychologie der Otto-von-Guericke-Universität Magdeburg.



Abbildung 1. Probabilistische Datenwissenschaft für die Psychologie

Zitation

Ostwald, D. (2024) Probabilistische Datenwissenschaft für die Psychologie. [10.5281/zenodo.10730199](https://zenodo.org/doi/10.5281/zenodo.10730199)

Lizenz

Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz.

Teil I.

Mathematische Grundlagen

Als Sprache der naturwissenschaftlichen Modellbildung ist die Mathematik auch die Sprache der probabilistischen Datenwissenschaft der Psychologie. Die Funktion der Mathematik ist dabei die einer Vermittlerin zwischen intuitiver Wissenschaftssprache und der Vielfalt von Programmiersprachen, die zur Implementation datenwissenschaftlicher Analysen eingesetzt werden. Mathematische Begriffsbildungen erlauben es, Konzepte sehr genau und allgemein verständlich zu greifen und unabhängig von den Ungenauigkeiten der naturwissenschaftlichen Alltagssprache oder den Idiosynkratien verschiedener Programmiersprachen zu kommunizieren. Natürlich kann die mathematische Sprache diese Rolle nicht losgelöst von wissenschaftssprachlicher Intuition oder praktischer Anwendung einnehmen. Die Mathematik ist in der probabilistischen Datenanalyse also nie Selbstzweck, sondern immer als angewandte Mathematik zu verstehen.

In diesem Teil stellen wir nach einigen Gedanken zu den grundlegenden Konzepten der Mathematik (Kapitel 1) mit den Mengen und den Funktionen (Definition 2.1 und Kapitel 4) die beiden Grundpfeiler der modernen Mathematik in komprimierter Form dar. Wir ergänzen diese Darstellung durch einige im notationelle Aspekte in Kapitel 3. Eng verwoben mit dem Begriff der Funktion sind die Differentialrechnung und die Integralrechnung (Kapitel 6 und Kapitel 7). Auch hier geht es nicht um eine erschöpfende Darstellung, sondern insbesondere um die Erläuterung einiger Grundlagen, die für die datenwissenschaftlich zentralen Konzepte der Optimierung sowie dem Umgang mit Wahrscheinlichkeitsfunktionen zentral sind. Wir ergänzen diese Abschnitte mit einigen oberflächlichen Gedanken zu den analytischen Grundlagen der Differenzialrechnung und Integralrechnung in Kapitel 5. Im Umgang mit den großen Datenmengen der zeitgenössischen Wissenschaft hat sich die Matrixschreibweise als hilfreich erwiesen. Wir widmen uns diesem Teilbereich der linearen Algebra zunächst für einspaltige Matrizen in Kapitel 14 und dann ausführlich in Kapitel 15.1, wiederum mit dem Ziel, notationelle und rechnerische Grundlagen für spätere Abschnitte bereitzustellen.

Insgesamt dient dieser Teil also vor allem der Kurzeinführung von Grundlagen, die in anderen Teilen wieder aufgegriffen werden, und dabei vor allem auch der Einführung einer einheitlichen Notation. Je nach Bedarf und Interesse ist natürlich ein vertieftes Selbststudium der verschiedenen Inhalte angezeigt. In den folgenden Literaturhinweisen führen wir überblicksartig Quellen und Leseempfehlungen zu den hier betrachteten Inhalten.

Literaturhinweise

Bärwolff (2017) und Arens et al. (2018) bilden die primäre Grundlage für viele hier dargestellten Themen und bieten einen exzellenten und umfassenden Einstieg in die behandelte Materie. Auf internationaler Ebene geben Spivak (2008) und Strang (2009) sehr gut lesbare Einführungen in die Differenzialrechnung bzw. die lineare Algebra. Ein tieferes Verständnis insbesondere analytischer Konzepte liefern zum Beispiel Abbott (2015) und Chiossi (2021). S. Searle (1982) bietet eine kompakte Darstellung derjenigen Aspekte der Matrizenrechnung und linearen Algebra, die im Bereich der probabilistischen Datenanalyse hauptsächlich zum Einsatz kommen. Deisenroth et al. (2020) bietet einen Überblick über viele der hier behandelten Themen unter dem Moniker des maschinellen Lernens.

1. Sprache und Logik

1.1. Mathematik ist eine Sprache

Mathematik ist die Sprache der naturwissenschaftlichen Modellbildung. So entspricht zum Beispiel der Ausdruck

$$F = ma \tag{1.1}$$

im Sinne des zweiten Newtonschen Axioms einer Theorie zur Bewegung von Objekten unter der Einwirkung von Kräften (Newton (1687)). Gleichermäßen entspricht der Ausdruck

$$\max_{q(z)} \int q(z) \ln \left(\frac{p(y, z)}{q(z)} \right) dz \tag{1.2}$$

im Sinne der Variational Inference der zeitgenössischen Theorie zur Funktionsweise des Gehirns (Friston (2005)). Mathematische Symbolik dient dabei insbesondere der genauen Kommunikation wissenschaftlicher Erkenntnisse und zielt darauf ab, komplexe Sachverhalte exakt und effizient zu beschreiben. Wie beim reflektierten Umgang mit jeder Form von Sprache steht also die Frage “Was soll das heißen?” als Leitfrage im Umgang mit mathematischen Inhalten und Symbolismen immer im Vordergrund.

Als Sprachgebäude weist die Mathematik einige Besonderheiten auf. Zum einen sind ihre Inhalte oft abstrakt. Dies rührt daher, dass sich die Mathematik um eine möglichst breite Allgemeinverständlichkeit und Anwendbarkeit bemüht. Mathematische Zugänge zu den Phänomenen der Welt sind dabei an einer möglichst einfachen Transferierbarkeit von Erkenntnissen in andere Kontexte interessiert. Um dies zu ermöglichen, versucht die Mathematik möglichst genau und verständlich, also im Sinne präziser Begriffe zu arbeiten. Sie geht dabei streng hierarchisch vor, sodass an späterer Stelle eingeführte Begriffe oft ein gutes Verständnis der ihnen zugrundeliegenden und an früherer Stelle eingeführten Begriffe voraussetzen.

Die Genauigkeit der mathematischen Sprache impliziert eine hohe Informationsdichte. Sie ist daher eher nüchtern und lässt überflüssiges weg, sodass in mathematischen Texten im besten Fall *alles* für die Kommunikation einer Idee relevant ist. Als Rezipient:in mathematischer Texte nimmt man ihre Informationsdichte anhand des hohen Verbrauchs kognitiver Energie beim Lesen eines mathematischen Textes wahr. Dieser hohe Energieverbrauch gebietet insbesondere Ruhe und Langsamkeit bei einem auf ein gutes Verständnis abzielenden Lesen. Als Leitsatz im Umgang mit mathematischen Texten mag dabei folgendes Zitat dienen: “Einen mathematischen Text kann man nicht lesen wie einen Roman, man muss ihn sich erarbeiten” (Unger (2000)). Nach dem Lesen eines kurzen mathematischen Textes sollte man sich immer kritisch fragen, ob man das Gelesene wirklich verstanden hat oder ob man zur Klärung des Sachverhaltes weitere Quellen heranziehen sollte. Auch ist es hilfreich, sich im Sinne des berühmten Zitats “What I cannot create, I do not understand” von Richard Feynman eigene Aufzeichnungen anzufertigen und mathematische Sprachgebäude selbst nachzukonstruieren.

Möchte man sich also die Welt der naturwissenschaftlichen Modellbildung erschließen, so ist es hilfreich, beim Umgang mit ihrer mathematischen Ausdrucksweise und Symbolik die gleichen Strategien wie beim Erlernen einer Fremdsprache anzuwenden. Hierzu gehört neben dem Eintauchen in den entsprechenden Sprachraum, also der ständigen Exposition mit mathematischen Ausdrucksweisen, sicherlich auch zunächst einmal das Auswendiglernen von Begriffen und das Übersetzen von Texten in die Alltagssprache. Ein tiefes und sicheres Verständnis mathematischer Modellbildung ergibt sich dann insbesondere durch die Anwendung mathematischer Herangehensweisen in schriftlicher und mündlicher Form.

1.2. Grundbausteine

Im Folgenden stellen wir mit den Begriffen der *Definition*, des *Theorems* und des *Beweises* drei Grundbausteine mathematischer Kommunikation vor, die uns durchgängig begleiten.

Definition

Eine *Definition* ist ein Grundstein eines mathematischen Systems, der innerhalb dieses Systems weder begründet noch deduktiv abgeleitet wird. Definitionen können nur nach ihrer Nützlichkeit innerhalb eines mathematischen Systems bewertet werden. Eine Definition lernt man am besten erst einmal auswendig und hinterfragt sie erst dann, wenn man ihren Nutzen in der Anwendung verstanden hat oder von diesem nicht überzeugt ist. Etwas Entspannung und Ruhe beim Umgang mit auf den ersten Blick komplexen Definitionen ist generell hilfreich. Um zu kennzeichnen, dass wir ein Symbol als etwas definieren, nutzen wir die Schreibweise “:=”. Zum Beispiel definiert der Ausdruck “ $a := 2$ ” das Symbol a als die Zahl Zwei. Definitionen enden in diesem Text immer mit dem Symbol •.

Theorem

Ein *Theorem* ist eine mathematische Aussage, die mittels eines Beweises als richtig (wahr) erkannt werden kann. Das heißt, ein Theorem wird immer aus Definitionen und/oder anderen Theoremen hergeleitet. Theoreme sind in diesem Sinne die empirischen Ergebnisse der Mathematik. Im Deutschen werden Theoreme oft auch als *Sätze* bezeichnet. In der angewandten, datenanalytischen Mathematik sind Theoreme häufig für Berechnungen hilfreich. Es lohnt sich also, sie auswendig zu lernen, da sie meist die Grundlage für Datenauswertung und Dateninterpretation bilden. Oft tauchen in Theoremen Gleichungen auf. Diese ergeben sich dabei aus den Voraussetzungen des Theorems. Um Gleichungen zu kennzeichnen, nutzen wir das Gleichheitszeichen “=”. So besagt beispielsweise der Ausdruck “ $a = 2$ ” in einem gegebenen Kontext, dass aufgrund bestimmter Voraussetzungen die Variable a den Wert zwei hat. Theoreme enden in diesem Text immer mit dem Symbol ◦.

Beweis

Ein *Beweis* ist eine Argumentationskette, die auf bekannte Definitionen und Theoreme zurückgreift, um die Richtigkeit (Wahrheit) eines Theorems zu belegen. Kurze Beweise tragen dabei oft zum Verständnis eines Theorems bei, lange Beweise eher nicht. Beweise sind also insbesondere die Antwort auf die Frage, warum eine mathematische Aussage gilt (“Warum ist das so?”). Beweise lernt man nicht auswendig. Wenn Beweise kurz sind, ist es sinnvoll, sie durchzuarbeiten, da sie meist als bekannt vorausgesetzte Inhalte wiederholen. Wenn sie lang sind, ist es sinnvoller sie zunächst zu übergehen, um sich nicht in Details

zu verlieren und vom eigentlichen Weg durch das mathematische Gebäude abzukommen. Beweise enden in diesem Text immer mit dem Symbol $\square$.

Neben Definitionen, Theoremen und Beweisen gibt es mit *Axiomen*, *Lemmata*, *Korollaren* und *Vermutungen* noch weitere typische Bausteine mathematischer Texte. Wir werden diese Begriffe nicht verwenden und geben deshalb für sie nur einen kurzen Überblick.

- *Axiome* sind unbeweisbare Theoreme, in dem Sinne, als dass sie als Grundannahmen zum Aufbau mathematischer Systeme dienen. Der Übergang zwischen Definitionen und Axiomen ist dabei oft fließend. Da wir mathematisch nicht besonders tief arbeiten, bevorzugen wir in den allermeisten Fällen den Begriff der Definition.
- Ein *Lemma* ist ein “Hilfstheorem”, also eine mathematische Aussage, die zwar bewiesen wird, aber nicht so bedeutend ist wie ein Theorem. Da wir einerseits auf bedeutende Inhalte fokussieren und andererseits mathematische Aussagen nicht diskriminieren wollen, verzichten wir auf diesen Begriff und nutzen stattdessen durchgängig den Begriff des Theorems.
- Ein *Korollar* ist eine mathematische Aussage, die sich durch einen einfachen Beweis aus einem Theorem ergibt. Da die “Einfachheit” mathematischer Beweise eine relative Eigenschaft ist, verzichten wir auf diesen Begriff und nutzen stattdessen auch hier durchgängig den Begriff des Theorems.
- *Vermutungen* sind mathematische Aussagen, von denen unbekannt ist, ob sie beweisbar oder widerlegbar sind. Da wir im Bereich der angewandten Mathematik arbeiten, treffen wir nicht auf Vermutungen.

1.3. Aussagenlogik

Nachdem wir nun einige Grundbausteine mathematischer Sprachgebäude kennengelernt haben, wollen wir uns mit der *Aussagenlogik* einem einfachen System nähern, das es erlaubt, Beziehungen zwischen mathematischen Aussagen herzustellen und zu formalisieren. Die Aussagenlogik spielt zum Beispiel in der Definition von Mengenoperationen, bei Optimierungsbedingungen von Funktionen und in vielen Beweisen eine tragende Rolle. In der datenanalytischen Anwendung ist die Aussagenlogik die Grundlage der Booleschen Logik der Programmierung. In der mathematischen Psychologie ist die Aussagenlogik die Grundlage der Repräsentationstheorie des Messens.

Wir beginnen mit der Definition des Begriffs der mathematischen *Aussage*.

Definition 1.1 (Aussage). Eine *Aussage* ist ein Satz, dem eindeutig die Eigenschaft *wahr* oder *falsch* zugeordnet werden kann.

•

Das Adjektiv *wahr* kann auch als *richtig* verstanden werden. Wir kürzen wahr mit “w” und falsch mit “f” ab. Im Körper der reellen Zahlen ist zum Beispiel die Aussage $1 + 1 = 2$ wahr und die Aussage $1 + 1 = 3$ falsch. Man beachte, dass die Binärität des Wahrheitsgehalts von Aussagen eine Grundannahme der Aussagenlogik und damit formal wissenschaftlich und nicht empirisch zu verstehen ist. Wahrheitsgehalte beziehen sich nicht auf Definitionen, Definitionen sind immer wahr.

Eine erste Möglichkeit, mit Aussagen zu arbeiten, ist, sie zu negieren. Dies führt auf folgende Definition.

Definition 1.2 (Negation). A sei eine Aussage. Dann ist die *Negation von A* die Aussage, die falsch ist, wenn A wahr ist und die wahr ist, wenn A falsch ist. Die Negation von A wird mit $\neg A$, gesprochen als “nicht A ”, bezeichnet.

•

Beispielsweise ist die Negation der Aussage “Die Sonne scheint” die Aussage “Die Sonne scheint nicht”. Die Negation der Aussage $1 + 1 = 2$ ist die Aussage $1 + 1 \neq 2$ und die Negation der Aussage $x > 1$ ist die Aussage $x \leq 1$. Tabellarisch stellt man die Definition der Negation einer Aussage A wie folgt dar:

Tabelle 1.1. Wahrheitstafel der Negation

A	$\neg A$
w	f
f	w

Tabellen dieser Form nennt man *Wahrheitstafeln*. Sie sind ein beliebtes Hilfsmittel in der Aussagenlogik, das wir im Folgenden oft einsetzen werden.

Möchte man zwei Aussagen logisch verbinden, so bieten sich die Begriffe der *Konjunktion* und der *Disjunktion* an.

Definition 1.3 (Konjunktion). A und B seien Aussagen. Dann ist die *Konjunktion von A und B* die Aussage, die dann und nur dann wahr ist, wenn A und B beide wahr sind. Die Konjunktion von A und B wird mit $A \wedge B$, gesprochen als “ A und B ”, bezeichnet.

•

Die Definition der Konjunktion impliziert folgende Wahrheitstafel:

Tabelle 1.2. Wahrheitstafel der Konjunktion

A	B	$A \wedge B$
w	w	w
w	f	f
f	w	f
f	f	f

Als Beispiel sei A die Aussage $2 \geq 1$ und B die Aussage $2 > 1$. Da sowohl A und B wahr sind, ist auch die Aussage $2 \geq 1 \wedge 2 > 1$ wahr. Als weiteres Beispiel sei A die Aussage $1 \geq 1$ und B die Aussage $1 > 1$. Hier ist nur A wahr und B falsch. Also ist die Aussage $1 \geq 1 \wedge 1 > 1$ falsch.

Definition 1.4 (Disjunktion). A und B seien Aussagen. Dann ist die *Disjunktion von A und B* die Aussage, die dann und nur dann wahr ist, wenn mindestens eine der beiden Aussagen A und B wahr ist. Die Disjunktion von A und B wird mit $A \vee B$, gesprochen als “ A oder B ”, bezeichnet.

•

Die Definition der Disjunktion impliziert folgende Wahrheitstafel:

Tabelle 1.3. Wahrheitstafel der Disjunktion

A	B	$A \vee B$
w	w	w
w	f	w
f	w	w
f	f	f

$A \vee B$ ist also insbesondere auch dann wahr, wenn A und B beide wahr sind. Damit ist das hier betrachtete “oder” genauer ein “und/oder”. Man nennt die Disjunktion daher auch ein “nicht-exklusives oder”. Als Beispiel sei A die Aussage $2 \geq 1$ und B die Aussage $2 > 1$. A ist wahr und B ist wahr. Also ist die Aussage $2 \geq 1 \vee 2 > 1$ wahr. Sei nun wiederum A die Aussage $1 \geq 1$ wahr und B die Aussage $1 > 1$. Dann ist A wahr und B falsch. Also ist die Aussage $1 \geq 1 \vee 1 > 1$ wahr.

Eine Möglichkeit, Aussagen in einen mechanischen logischen Zusammenhang zu stellen, ist die *Implikation*. Diese ist wie folgt definiert.

Definition 1.5 (Implikation). A und B seien Aussagen. Dann ist die *Implikation*, bezeichnet mit $A \Rightarrow B$, die Aussage, die dann und nur dann falsch ist, wenn A wahr und B falsch ist. A heißt dabei die *Voraussetzung* (*Prämisse*) und B der *Schluss* (*Konklusion*) der Implikation. $A \Rightarrow B$ spricht man als “aus A folgt B ”, “ A impliziert B ”, oder “wenn A , dann B ”.

•

Die Definition der Implikation kann durch folgende Wahrheitstafel dargestellt werden:

Tabelle 1.4. Wahrheitstafel der Implikation

A	B	$A \Rightarrow B$
w	w	w
w	f	f
f	w	w
f	f	w

Ein intuitives Verständnis der Definition der Implikation im Sinne obiger Wahrheitstafel ergibt sich am ehesten, indem man sie als Versuch liest, die intuitive Vorstellung einer Folgerung im Kontext der Aussagenlogik abzubilden und zu formalisieren. Um dies nachzuvollziehen, liest man die Wahrheitszustände in Tabelle 1.4 am besten in der Reihenfolge Wahrheitszustand von A , Wahrheitszustand von $A \Rightarrow B$ und betrachtet schließlich den Wahrheitszustand von B . Liest man die Wahrheitstafel auf diese Weise, so sieht man, dass wenn A wahr ist und $A \Rightarrow B$ wahr ist, B wahr ist. Konstruiert man basierend auf einer wahren Aussage also (zum Beispiel durch das Umformen von Gleichungen) eine wahre Implikation, so folgt, dass auch B wahr ist. Ist dies nicht möglich, wenn also gilt, dass, wenn A wahr ist, $A \Rightarrow B$ immer falsch ist, dann ist auch B falsch. So mag man Aussagen widerlegen. Schließlich sieht man, dass wenn A falsch ist und $A \Rightarrow B$ wahr ist, B wahr oder

falsch sein kann. Nur aus einer wahren Voraussetzung folgt also bei wahrer Implikation immer eine wahre Konklusion. Insbesondere genügt die Definition der Implikation damit der Forderung “Aus Falschem folgt beliebiges (ex falso sequitur quodlibet)”. Man kann aus falschen Aussagen also mithilfe der Implikation nichts Sinnvolles folgern.

Im Kontext der Implikation ergeben sich die Begriffe der *hinreichenden* und der *notwendigen* Bedingungsassertionen: Wenn $A \Rightarrow B$ wahr ist, sagt man, dass “ A hinreichend für B ist” und dass “ B notwendig für A ist”. Diese Sprechweise erklärt sich im Kontext der Implikation folgendermaßen: Wenn $A \Rightarrow B$ wahr ist, gilt dass, wenn A wahr ist, auch B wahr ist. Die Wahrheit von A reicht also für die Wahrheit von B aus. A ist also hinreichend (ausreichend) für B . Weiterhin gilt, dass wenn $A \Rightarrow B$ wahr ist, dass wenn B falsch ist, dann auch A falsch ist. Die Wahrheit von B ist also für die Wahrheit von A notwendig.

Schließlich wollen wir zu Illustration von Tabelle 1.4 noch ein lebensnahes Beispiel geben. Dazu sei A die Aussage “Die Brücke ist beschädigt”, B die Aussage “Die Brücke ist gesperrt” und $A \Rightarrow B$ die Implikation “Wenn die Brücke beschädigt ist, dann ist die Brücke gesperrt”. Wir diskutieren die vier Fälle von Tabelle 1.4.

- A wahr, B wahr, $A \Rightarrow B$ wahr. Wenn A wahr ist, dann ist die Brücke beschädigt und wenn dazu auch B wahr ist, dann ist die Brücke außerdem gesperrt. In diesem Fall ist die Implikation “Wenn die Brücke beschädigt ist, dann ist die Brücke gesperrt” also offensichtlich wahr.
- A wahr, B falsch, $A \Rightarrow B$ falsch. Wenn A wahr ist, dann ist die Brücke beschädigt und wenn B falsch ist, dann ist die Brücke also nicht gesperrt. In diesem Fall ist die Implikation “Wenn die Brücke beschädigt ist, dann ist die Brücke gesperrt” also offensichtlich falsch.
- A falsch, B wahr, $A \Rightarrow B$ wahr. Wenn A falsch ist, dann ist die Brücke nicht beschädigt und wenn B wahr ist, dann ist die Brücke gesperrt. Da die Brücke allerdings nicht beschädigt ist, kann man keine Schlussfolgerung hinsichtlich der Implikation “Wenn die Brücke beschädigt ist, dann ist die Brücke gesperrt” ziehen, man kann also insbesondere nicht zeigen, dass diese falsch wäre. Also unterstellt man, dass die Aussage “Wenn die Brücke beschädigt ist, dann ist die Brücke gesperrt” wahr ist.
- A falsch, B falsch, $A \Rightarrow B$ wahr. Wenn A falsch ist, dann ist die Brücke nicht beschädigt und wenn B falsch ist, dann ist die Brücke nicht gesperrt. Da die Brücke allerdings nicht beschädigt ist, kann man keine Schlussfolgerung hinsichtlich der Implikation “Wenn die Brücke beschädigt ist, dann ist die Brücke gesperrt” ziehen, man kann also insbesondere nicht zeigen, dass diese falsch wäre. Also unterstellt man wiederum, dass die Aussage “Wenn die Brücke beschädigt ist, dann ist die Brücke gesperrt” wahr ist.

Eine sehr häufig auftretender Zusammenhang zwischen zwei Aussagen ist ihre *Äquivalenz*.

Definition 1.6 (Äquivalenz). A und B seien Aussagen. Die *Äquivalenz von A und B* ist die Aussage, die dann und nur dann wahr ist, wenn A und B beide wahr sind oder wenn A und B beide falsch sind. Die Äquivalenz von A und B wird mit $A \Leftrightarrow B$ bezeichnet und gesprochen als “ A genau dann, wenn B ” oder “ A ist äquivalent zu B ”.

•

Die Definition der Äquivalenz impliziert folgende Wahrheitstafel:

Tabelle 1.5. Wahrheitstafel der Äquivalenz

A	B	$A \Leftrightarrow B$
w	w	w
w	f	f
f	w	f
f	f	w

Die Definition des Begriffes der *logischen Äquivalenz* erlaubt es unter anderem, die Äquivalenz zweier Aussagen mithilfe von Implikationen nachzuweisen.

Definition 1.7 (Logische Äquivalenz). Zwei Aussagen heißen *logisch äquivalent*, wenn ihre Wahrheitstabeln gleich sind.

•

Als Beispiele für logische Äquivalenzen, die häufig in Beweisargumentationen genutzt werden, zeigen wir die Aussagen folgenden Theorems.

Theorem 1.1 (Logische Äquivalenzen). *A und B seien zwei Aussagen. Dann sind folgende Aussagen logisch äquivalent:*

- (1) $A \Leftrightarrow B$ und $(A \Rightarrow B) \wedge (B \Rightarrow A)$
- (2) $A \Rightarrow B$ und $(\neg B) \Rightarrow (\neg A)$

◦

Beweis. Nach Definition des Begriffes der logischen Äquivalenz müssen wir zeigen, dass die Wahrheitstabeln der betrachteten Aussagen gleich sind. Wir zeigen erst (1), dann (2).

(1) Wir erinnern an die Wahrheitstafel von $A \Leftrightarrow B$:

A	B	$A \Leftrightarrow B$
w	w	w
w	f	f
f	w	f
f	f	w

Wir betrachten weiterhin die Wahrheitstafel von $(A \Rightarrow B) \wedge (B \Rightarrow A)$:

A	B	$A \Rightarrow B$	$B \Rightarrow A$	$(A \Rightarrow B) \wedge (B \Rightarrow A)$
w	w	w	w	w
w	f	f	w	f
f	w	w	f	f
f	f	w	w	w

Der Vergleich der Wahrheitstafel von $A \Leftrightarrow B$ mit den ersten beiden und der letzten Spalte der Wahrheitstafel von $(A \Rightarrow B) \wedge (B \Rightarrow A)$ zeigt ihre Gleichheit.

(2) Wir erinnern an die Wahrheitstafel von $A \Rightarrow B$:

A	B	$A \Rightarrow B$
w	w	w
w	f	f
f	w	w
f	f	w

Wir betrachten weiterhin die Wahrheitstafel von $(\neg B) \Rightarrow (\neg A)$:

A	B	$\neg B$	$\neg A$	$(\neg B) \Rightarrow (\neg A)$
w	w	f	f	w
w	f	w	f	f
f	w	f	w	w
f	f	w	w	w

Der Vergleich der Wahrheitstafel von $A \Rightarrow B$ mit den ersten beiden und der letzten Spalte der Wahrheitstafel von $(\neg B) \Rightarrow (\neg A)$ zeigt ihre Gleichheit. $\square$

Die erste Aussage von Theorem 1.1 besagt, dass die Aussage “ A und B sind äquivalent” logisch äquivalent zur Aussage “Aus A folgt B ” und zur Aussage “Aus B folgt A ” ist. Dies ist die Grundlage für viele sogenannte *direkte Beweise* mithilfe von Äquivalenzumformungen. Die zweite Aussage von Theorem 1.1 besagt, dass die Aussage “Aus A folgt B ” logisch äquivalent zur Aussage “Aus nicht B folgt nicht A ” ist. Dies ist die Grundlage für die Technik des *indirekten Beweises*. Wir betrachten diese Beweistechniken im folgenden Abschnitt genauer. Zunächst fassen wir die Bedeutungen der in diesem Abschnitt eingeführte Symbole noch einmal in untenstehender Tabelle zusammen.

Tabelle 1.10. Symbolübersicht. A und B sind Aussagen, d.h. A und B sind entweder wahr oder falsch.

Symbol	Bedeutung	Bemerkung
$\neg A$	Nicht A	Wahr, wenn A falsch ist und umgekehrt
$A \wedge B$	A und B	Nur wahr, wenn A und B beide wahr sind
$A \vee B$	A und/oder B	Wahr, wenn mindestens eine der Aussagen wahr ist
$A \Rightarrow B$	Aus A folgt B	B ist notwendig für A , A ist hinreichend für B
$A \Leftrightarrow B$	A ist äquivalent zu B	Es gelten $A \Rightarrow B$ und $B \Rightarrow A$

1.4. Äquivalenzumformungen

Mathematische Probleme führen oft auf den Fall, dass Information über eine unbekannte Variable implizit mithilfe einer Gleichung oder einer Ungleichung dargestellt wird. Um die Information über die entsprechende Variable explizit darzustellen, also im Falle von Gleichungen ihren Wert zu bestimmen oder im Falle von Ungleichungen den Zahlenbereich zu ermitteln, in dem der Wert der Variable liegt, nutzt man *Äquivalenzumformungen*. Wir betrachten hier lediglich die aus der Schulmathematik bekannten Äquivalenzumformungen von (Un)Gleichungen bezüglich reeller Variablen. Äquivalenzumformungen von (Un)Gleichungen haben dabei die Eigenschaft, dass sich der Wahrheitsgehalt der durch eine (Un)Gleichung formulierten Aussage bei Anwendung der entsprechenden Umformung nicht ändert. Dabei werden stets beide Seiten der (Un)Gleichung umgeformt. Damit es sich bei der Anwendung einer mathematischen Operation auf eine (Un)gleichungsaussage A in

eine (Un)gleichungsaussage B um eine Äquivalenzumformung der Form $A \Leftrightarrow B$ handelt, müssen bekanntlich sowohl $A \Rightarrow B$ als auch $B \Rightarrow A$ gelten. Dies impliziert, dass es sich bei Äquivalenzumformungen um umkehrbare (invertierbare) Operationen handelt.

Bei Gleichungen sind zulässige Äquivalenzumformungen insbesondere

- die Addition einer Zahl auf beiden Seiten der Gleichung,
- die Subtraktion einer Zahl auf beiden Seiten der Gleichung,
- die Multiplikation mit einer Zahl auf beiden Seiten der Gleichung,
- die Division durch eine von null verschiedene Zahl auf beiden Seiten der Gleichung,
- die Anwendung einer invertierbaren Funktion auf beiden Seiten der Gleichung.

Beispiel

Wir betrachten unter Vorgriff auf Kapitel 4.3 die Aussage

$$2 \exp(x) - 2 = 0. \quad (1.3)$$

Dann gelten

$$\begin{aligned} & 2 \exp(x) - 2 = 0 \\ \Leftrightarrow & 2 \exp(x) - 2 + 2 = +2 \\ \Leftrightarrow & 2 \exp(x) = 2 \\ \Leftrightarrow & \frac{1}{2} \cdot 2 \exp(x) = \frac{1}{2} \cdot 2 \\ \Leftrightarrow & \exp(x) = 1 \\ \Leftrightarrow & \ln(\exp(x)) = \ln(1) \\ \Leftrightarrow & x = 0. \end{aligned} \quad (1.4)$$

Zusammengefasst gilt also

$$2 \exp(x) - 2 \Leftrightarrow x = 0. \quad (1.5)$$

Bei Ungleichungen sind zulässige Äquivalenzumformungen insbesondere

- die Addition einer Zahl auf beiden Seiten der Ungleichung,
- die Subtraktion einer Zahl auf beiden Seiten der Ungleichung,
- die Multiplikation mit einer Zahl auf beiden Seiten der Gleichung, wobei die Multiplikation mit einer negativen Zahl das Ungleichungszeichen umkehrt,
- die Division durch eine von null verschiedene positiven Zahl auf beiden Seiten der Gleichung, wobei die Division mit einer negativen Zahl das Ungleichungszeichen umkehrt,
- die Anwendung einer invertierbaren monotonen Funktion auf beiden Seiten der Ungleichung.

Beispiel

Wir betrachten die Aussage

$$-5x - 2 \geq 8. \quad (1.6)$$

Dann gelten

$$\begin{aligned}
 & -5x - 2 \geq 8 \\
 \Leftrightarrow & -5x - 2 + 2 \geq 8 + 2 \\
 \Leftrightarrow & -5x \geq 10 \\
 \Leftrightarrow & -\frac{1}{5} \cdot 5x \geq \frac{1}{5} \cdot 10 \\
 \Leftrightarrow & -x \geq 2 \\
 \Leftrightarrow & x \leq -2.
 \end{aligned} \tag{1.7}$$

Zusammengefasst gilt also

$$-5x - 2 \geq 8 \Leftrightarrow x \leq -2. \tag{1.8}$$

1.5. Beweistechniken

In diesem Abschnitt wollen wir mit den Begriffen der *direkten* und *indirekten Beweise* sowie des *Beweises durch Widerspruch* drei fundamentale Beweistechniken skizzieren. Dabei wird vor allem die erste im Folgenden immer wieder zur Begründung von Theoremen herangezogen werden.

Es gilt dabei

- *Direkte Beweise* nutzen Äquivalenzumformungen, um $A \Rightarrow B$ zu zeigen.
- *Indirekte Beweise* nutzen die logische Äquivalenz von $A \Rightarrow B$ und $(\neg B) \Rightarrow (\neg A)$.
- *Beweise durch Widerspruch* zeigen, dass $(\neg B) \wedge A$ falsch ist.

Damit ausgestattet wollen wir nun folgendes Theorem mithilfe eines direkten Beweises, eines indirekten Beweises und eines Beweises durch Widerspruch beweisen (vgl. Arens et al. (2018)).

Theorem 1.2 (Quadrate positiver Zahlen). *Es seien a und b zwei positive Zahlen. Dann gilt $a^2 < b^2 \Rightarrow a < b$.*

◦

Beweis. Wir geben zunächst einen *direkten Beweis*. Dazu sei $a^2 < b^2$ die Aussage A und $a < b$ die Aussage B . Dann gilt

$$a^2 < b^2 \Leftrightarrow 0 < b^2 - a^2 \Leftrightarrow 0 < (b + a)(b - a) \Leftrightarrow 0 < (b - a) \Leftrightarrow a < b. \tag{1.9}$$

Wir geben nun einen *indirekten Beweis*. Es sei $a^2 \geq b^2$ die Aussage $\neg A$. Weiterhin sei $a \geq b$ die Aussage $\neg B$. Dann gilt

$$a \geq b \Leftrightarrow a^2 \geq ab \wedge ab \geq b^2 \Leftrightarrow a^2 \geq b^2. \tag{1.10}$$

Schließlich geben wir einen *Beweis durch Widerspruch*. Wir zeigen, dass die Annahme $(\neg B) \wedge A$ auf eine falsche Aussage führt. Es gilt

$$a \geq b \wedge a^2 < b^2 \Leftrightarrow a^2 \geq ab \wedge a^2 < b^2 \Leftrightarrow ab \leq a^2 < b^2. \tag{1.11}$$

Weiterhin gilt

$$a \geq b \wedge a^2 < b^2 \Leftrightarrow ab \geq b^2 \wedge a^2 < b^2 \Leftrightarrow a^2 < b^2 \leq ab. \tag{1.12}$$

Insgesamt gilt dann also die falsche Aussage

$$ab \leq a^2 < b^2 \leq ab \Leftrightarrow ab < ab. \tag{1.13}$$

□

2. Mengen

2.1. Grundlegende Definitionen

Mengen fassen mathematische Objekte zusammen und bilden die Grundlage der modernen Mathematik. Wir beginnen mit folgender Definition.

Definition 2.1 (Mengen). Nach Cantor (1895) ist eine *Menge* M definiert als “eine Zusammenfassung von bestimmten wohlunterschiedenen Objekten m unserer Anschauung oder unseres Denkens (welche die Elemente der Menge genannt werden) zu einem Ganzen”. Wir schreiben

$$m \in M \text{ bzw. } m \notin M \quad (2.1)$$

um auszudrücken, dass m ein Element bzw. kein Element von M ist.

•

Zur Definition von Mengen gibt es mindestens folgende Möglichkeiten:

- Auflisten der Elemente in geschweiften Klammern, z.B. $M := \{1, 2, 3\}$.
- Angabe der Eigenschaften der Elemente, z.B. $M := \{x \in \mathbb{N} | x < 4\}$.
- Gleichsetzen mit einer anderen eindeutig definierten Menge, z.B. $M := \mathbb{N}_3$.

Die Schreibweise $\{x \in \mathbb{N} | x < 4\}$ wird gelesen als “ $x \in \mathbb{N}$, für die gilt, dass $x < 4$ ist”, wobei die Bedeutung von $\mathbb{N}$ im Folgenden noch zu erläutern sein wird. Es ist wichtig zu erkennen, dass Mengen *ungeordnete* mathematische Objekte sind, das heißt, dass die Reihenfolge der Auflistung der Elemente einer Menge keine Rolle spielt. Zum Beispiel bezeichnen $\{1, 2, 3\}$, $\{1, 3, 2\}$ und $\{2, 3, 1\}$ dieselbe Menge, nämlich die Menge der ersten drei natürlichen Zahlen.

Grundlegende Beziehungen zwischen mehreren Mengen werden in der nächsten Definition festgelegt.

Definition 2.2 (Teilmengen und Mengengleichheit). M und N seien zwei Mengen.

- Eine Menge M heißt *Teilmenge* einer Menge N , wenn für jedes Element $m \in M$ gilt, dass auch $m \in N$. Ist M eine Teilmenge von N , so schreibt man

$$M \subseteq N \quad (2.2)$$

und nennt M *Untermenge* von N und N *Obermenge* von M .

- Eine Menge M heißt *echte Teilmenge* einer Menge N , wenn für jedes Element $m \in M$ gilt, dass auch $m \in N$, es aber zumindest ein Element $n \in N$ gibt, für das gilt $n \notin M$. Ist M eine echte Teilmenge von N , so schreibt man

$$M \subset N. \quad (2.3)$$

- Zwei Mengen M und N heißen *gleich*, wenn für jedes Element $m \in M$ gilt, dass auch $m \in N$, und wenn für jedes Element $n \in N$ gilt, dass auch $n \in M$. Sind die Mengen M und N gleich, so schreibt man

$$M = N. \quad (2.4)$$

•

Beispiel

Betrachten wir zum Beispiel die Mengen $M := \{1\}$, $N := \{1, 2\}$, und $O := \{1, 2\}$. Dann gilt mit obigen Definitionen, dass $M \subset N$, weil $1 \in M$ und $1 \in N$, aber $2 \in N$ und $2 \notin M$. Weiterhin gilt, dass $N \subseteq O$, weil $1 \in N$ und $1 \in O$ sowie $2 \in N$ und $2 \in O$ und es kein Element von O gibt, welches nicht in N ist. Ebenso gilt $O \subseteq N$, weil $1 \in O$ und $1 \in N$ sowie $2 \in O$ und $2 \in N$ und es kein Element von N gibt, welches nicht in O ist. Schließlich gilt sogar $N = O$, weil für jedes Element $n \in N$ gilt, dass auch $n \in O$, und gleichzeitig für jedes Element $o \in O$ gilt, dass auch $o \in N$. Wir stellen diese Zusammenhänge schematisch mithilfe von *Venn-Euler-Diagrammen* in Abbildung 2.1 dar.

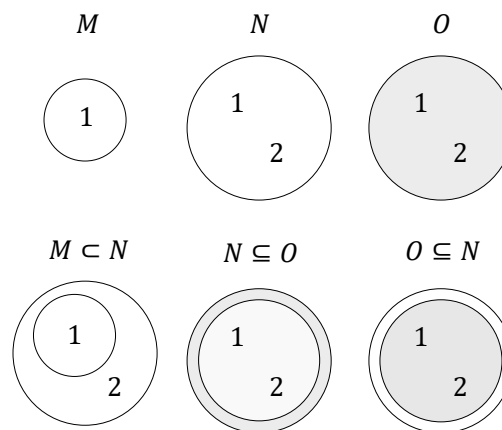


Abbildung 2.1. Venn-Euler Diagramme der Teilmengeneigenschaften der Mengen $M := \{1\}$, $N := \{1, 2\}$ und $O := \{1, 2\}$.

Eine wichtige Eigenschaft einer Menge ist die Anzahl der in ihr enthaltenen Elemente. Diese wird als *Kardinalität* der Menge bezeichnet.

Definition 2.3 (Kardinalität). Die Anzahl der Elemente einer Menge M heißt *Kardinalität* und wird mit $|M|$ bezeichnet.

•

Eine besondere Menge ist die Menge ohne Elemente.

Definition 2.4. Eine Menge mit Kardinalität null heißt *leere Menge* und wird mit $\emptyset$ bezeichnet.

•

Als Beispiele seien $M := \{1, 2, 3\}$, $N = \{a, b, c, d\}$ und $O := \emptyset$. Dann gelten $|M| = 3$, $N = 4$ und $|O| = 0$.

Zu jeder Menge kann man die Menge aller Teilmengen dieser Menge betrachten. Dies führt auf den wichtigen Begriff der *Potenzmenge*.

Definition 2.5 (Potenzmenge). Die Menge aller Teilmengen einer Menge M heißt *Potenzmenge von M* und wird mit $\mathcal{P}(M)$ bezeichnet.

•

Man beachte, dass die leere Untermenge von M und M selbst auch immer Elemente von $\mathcal{P}(M)$ sind. Wir betrachten zunächst vier Beispiele zum Begriff der Potenzmenge.

- $M_0 := \emptyset$ sei die leere Menge. Dann gilt

$$\mathcal{P}(M_0) = \emptyset. \quad (2.5)$$

- M_1 sei die einelementige Menge $M_1 := \{a\}$. Dann gilt

$$\mathcal{P}(M_1) = \{\emptyset, \{a\}\}. \quad (2.6)$$

- Es sei $M_2 := \{a, b\}$. Dann hat M_2 sowohl ein- als auch zweielementige Teilmengen und es gilt

$$\mathcal{P}(M_2) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}. \quad (2.7)$$

- Schließlich sei $M_3 := \{a, b, c\}$. Dann hat M_3 unter anderem ein-, zwei-, und dreielementige Teilmengen und es gilt

$$\mathcal{P}(M_3) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}. \quad (2.8)$$

Theorem 2.1 (Kardinalität der Potenzmenge). Gegeben sei eine Menge M mit Kardinalität $|M| = n$ und $\mathcal{P}(M)$ sei ihre Potenzmenge. Dann gilt $|\mathcal{P}(M)| = 2^n$

◦

Beweis. Um die Aussage des Theorems zu beweisen assoziieren wir jedes Element P der Potenzmenge von M eindeutig mit einer binären Folge der Länge n , wobei der Eintrag an der i ten Stelle repräsentiert, ob das i te Element von M ein Element von P ist oder nicht. Seien beispielsweise $M := \{m_1, m_2, m_3\}$ und $P := \{m_2, m_3\}$. Dann entspricht P die binäre Folge 011. Der leeren Menge $P := \emptyset$ entspricht die binäre Folge 000 und der Ausgangsmenge $P = M$ entspricht die binäre Folge 111. Es ergibt sich also die Frage, wieviele eindeutige binäre Folgen der Länge n es gibt. Da es für jedes Element der Folge zwei mögliche Zustände gibt, ergeben sich n Faktoren $2 \cdot 2 \cdots 2$, also 2^n . $\square$

In den obigen Beispielen haben wir die Fälle

- $|M_0| = 0 \Rightarrow |\mathcal{P}(M_0)| = 2^0 = 1,$
- $|M_1| = 1 \Rightarrow |\mathcal{P}(M_1)| = 2^1 = 2,$
- $|M_2| = 2 \Rightarrow |\mathcal{P}(M_2)| = 2^2 = 4,$
- $|M_3| = 3 \Rightarrow |\mathcal{P}(M_3)| = 2^3 = 8,$

wovon man sich durch Nachzählen der Elemente der entsprechenden Potenzmengen überzeugt.

2.2. Verknüpfungen

Zwei Mengen können auf unterschiedliche Weise miteinander verknüpft werden. Das Ergebnis einer solchen Verknüpfung ist eine weitere Menge. Wir bezeichnen die Verknüpfung zweier Mengen als *Mengenoperation* und geben folgende Definitionen.

Definition 2.6 (Mengenoperationen). M und N seien zwei Mengen.

- Die *Vereinigung von M und N* ist definiert als die Menge

$$M \cup N := \{x | x \in M \vee x \in N\}, \quad (2.9)$$

wobei $\vee$ gemäß Definition 1.4 als *nicht-exklusives oder*, also als und/oder, zu verstehen ist.

- Der *Durchschnitt von M und N* ist definiert als die Menge

$$M \cap N := \{x | x \in M \wedge x \in N\}. \quad (2.10)$$

Wenn für M und N gilt, dass $M \cap N = \emptyset$, dann heißen M und N *disjunkt*.

- Die *Differenz von M und N* ist definiert als die Menge

$$M \setminus N := \{x | x \in M \wedge x \notin N\}. \quad (2.11)$$

Die Differenz von M und N heißt, insbesondere bei $N \subseteq M$, auch das *Komplement von N bezüglich M* und wird mit N^c bezeichnet. In diesem Kontext wird M auch als die *Grundmenge* oder das *Universum* bezeichnet.

- Die *symmetrische Differenz von M und N* ist definiert als die Menge

$$M \Delta N := \{x | (x \in M \vee x \in N) \wedge x \notin M \cap N\}, \quad (2.12)$$

Die symmetrische Differenz kann also als *exklusives oder* verstanden werden.

•

Beispiel

Als Beispiel betrachten wir die Mengen $M := \{1, 2, 3\}$ und $N := \{2, 3, 4, 5\}$. Dann gelten

- $M \cup N = \{1, 2, 3, 4, 5\}$, weil $1 \in M$, $2 \in M$, $3 \in M$, $4 \in N$ und $5 \in N$.
- $M \cap N = \{2, 3\}$, weil nur für 2 und 3 gilt, dass $2 \in M$, $3 \in M$ und auch $2 \in N$, $3 \in N$. Für 1 gilt lediglich, dass $1 \in M$ und für 4 und 5 gelten lediglich, dass $4 \in N$ und $5 \in N$.
- $M \setminus N = \{1\}$, weil $1 \in M$, aber $1 \notin N$ und $2 \in M$, aber auch $2 \in N$.
- $N \setminus M = \{4, 5\}$, weil $2 \in N$ und $3 \in N$, aber auch $2 \in M$ und $3 \in M$. Dies zeigt insbesondere, dass die Differenz von M und N *nicht* symmetrisch ist, also dass *nicht* zwangsläufig gilt, dass $M \setminus N$ gleich $N \setminus M$ ist.
- $M \Delta N = \{1, 4, 5\}$, weil $1 \in M$, aber $1 \notin \{2, 3\}$, $2 \in M$, aber $2 \in \{2, 3\}$, $3 \in M$, aber $3 \in \{2, 3\}$, $4 \in N$, aber $4 \notin \{2, 3\}$ und $5 \in N$, aber $5 \notin \{2, 3\}$.

Abbildung 2.2 visualisiert die in diesem Beispiel betrachteten Mengenoperationen.

Schließlich wollen wir noch den Begriff der Partition einer Menge einführen.

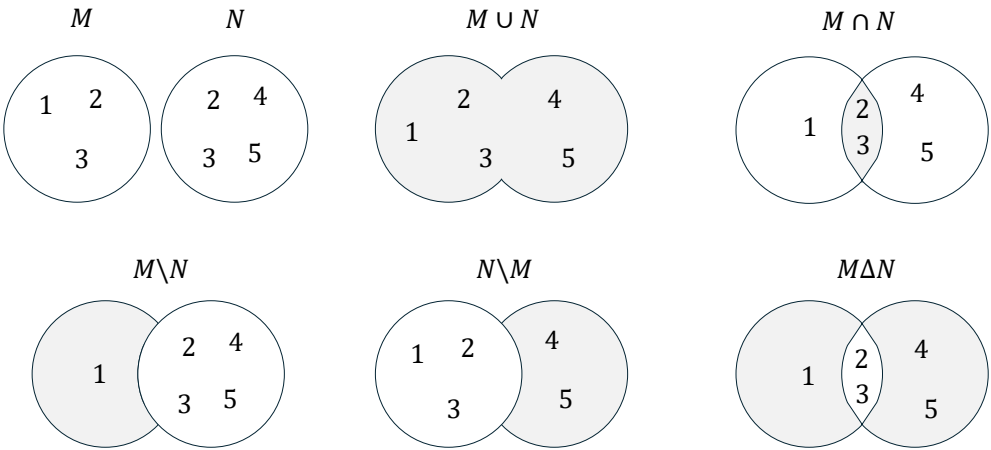


Abbildung 2.2. Venn-Euler Diagramme der im Beispiel für $M := \{1, 2, 3\}$ und $N := \{2, 3, 4, 5\}$ betrachteten Mengenoperationen. Die grau hinterlegten Flächen entsprechen jeweils den sich ergebenden Mengen.

Definition 2.7 (Partition). M sei eine Menge und $P := \{N_i\}$ sei eine Menge von Mengen N_i mit $i = 1, \dots, n$, sodass gilt

$$(M = \cup_{i=1}^n N_i) \wedge (N_i \cap N_j = \emptyset \text{ f\"ur } i, j = 1, \dots, n \text{ und } i \neq j).$$
(2.13)

Dann heit P eine *Partition* von M .

•

Die Partition einer Menge entspricht also dem Aufteilen der Menge in disjunkte Teilmengen. Partitionen sind generell nicht eindeutig, das heit, es gibt meist verschiedene Mglichkeiten eine gegebene Menge zu partitionieren.

Als Beispiel betrachten wir die Menge $M := \{1, 2, 3, 4, 5, 6\}$. Dann sind $P_1 := \{\{1\}, \{2, 3, 4, 5, 6\}\}$, $P_2 := \{\{1, 2, 3\}, \{4, 5, 6\}\}$ und $P_3 := \{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$ drei mgliche Partitionen von M . Abbildung 2.3 visualisiert die in diesem Beispiel betrachteten Partitionen.

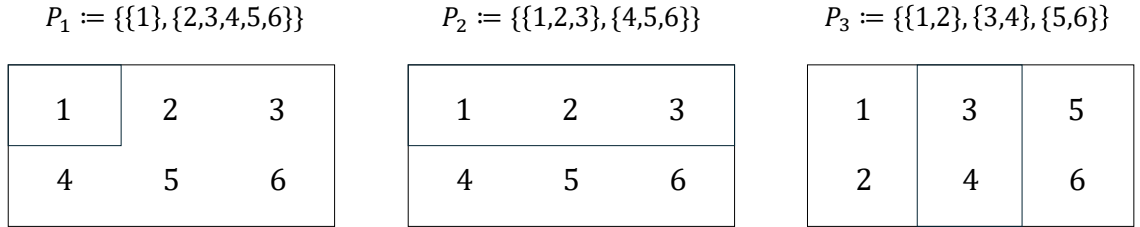


Abbildung 2.3. Diagramme der im Beispiel fr $M := \{1, 2, 3, 4, 5, 6\}$ betrachteten Partitionen.

2.3. Spezielle Mengen

Zahlenmengen

In der Naturwissenschaft versucht man, in der Vorstellung intuitiv als diskret oder kontinuierlich identifizierte Phänomene der Welt mit Zahlen zu beschreiben. Je nach Art des Phänomens bieten sich dazu verschiedene Zahlenmengen an. Die Mathematik stellt unter anderem die in folgender Definition gegebenen Zahlenmengen bereit.

Definition 2.8 (Zahlenmengen). Es bezeichnen

- $\mathbb{N} := \{1, 2, 3, \dots\}$ die *natürlichen Zahlen*,
- $\mathbb{N}_n := \{1, 2, 3, \dots, n\}$ die *natürlichen Zahlen der Ordnung n* ,
- $\mathbb{N}^0 := \mathbb{N} \cup \{0\}$ die *natürlichen Zahlen* und null,
- $\mathbb{Z} := \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ die *ganzen Zahlen*,
- $\mathbb{Q} := \{\frac{p}{q} | p \in \mathbb{Z}, q \in \mathbb{N}\}$ die *rationalen Zahlen*,
- $\mathbb{R}$ die *reellen Zahlen*, und
- $\mathbb{C} := \{a + ib | a, b \in \mathbb{R}, i := \sqrt{-1}\}$ die *komplexen Zahlen*.

•

Die natürlichen und die ganzen Zahlen eignen sich zum Quantifizieren diskreter Phänomene. Die rationalen und insbesondere die reellen Zahlen eignen sich zum Quantifizieren kontinuierlicher Phänomene. $\mathbb{R}$ umfasst dabei die rationalen Zahlen und die sogenannten *irrationalen Zahlen* $\mathbb{R} \setminus \mathbb{Q}$. Rationale Zahlen sind Zahlen, die sich, durch Brüche ganzer und natürlicher Zahlen ausdrücken lassen. Dies sind alle ganzen Zahlen sowie die negativen und positiven Dezimalzahlen wie beispielsweise $-\frac{9}{10} = -0.9$, $\frac{1}{3} = 1.3\bar{3}$ und $\frac{196}{100} = 1.96$. Irrationale Zahlen sind Zahlen, die sich nicht als rationale Zahlen ausdrücken lassen. Beispiele für irrationale Zahlen sind die *Eulersche Zahl* $e \approx 2.71$, die *Kreiszahl* $\pi \approx 3.14$ und die Quadratwurzel von 2, $\sqrt{2} \approx 1.41$.

Die reellen Zahlen enthalten als Teilmengen die natürlichen, ganzen und die rationalen Zahlen. Es gibt also sehr viele reelle Zahlen. Tatsächlich hat Cantor (1874) bewiesen, dass es mehr reelle Zahlen als natürliche Zahlen gibt, obwohl es sowohl unendlich viele reelle Zahlen als auch unendlich viele natürliche Zahlen gibt. Diese Eigenschaft der reellen Zahlen bezeichnet man als die *Überabzählbarkeit* der reellen Zahlen. Wir vertiefen diesen Begriff in Kapitel 4. Insbesondere gilt

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}. \quad (2.14)$$

Zwischen zwei reellen Zahlen gibt es unendlich viele weitere reelle Zahlen. Positiv-Unendlich ∞ und Negativ-Unendlich $-\infty$ sind keine Zahlen, mit denen in der Standardmathematik gerechnet werden könnte. Sie gehören auch nicht zu den in obiger Definition gegebenen Zahlenmengen, es gelten also $\infty \notin \mathbb{R}$ und $-\infty \notin \mathbb{R}$. Komplexe Zahlen eignen sich zur Beschreibung zweidimensionaler kontinuierlicher Phänomene. Dabei werden die Werte der ersten Dimension im reellen Teil a und die Werte der zweiten Dimension im komplexen Teil b einer komplexen Zahl repräsentiert. Komplexe Zahlen kommen beispielsweise bei der Modellierung physikalischer Phänomene und im Bereich der Fourieranalyse zum Einsatz.

Wichtige Teilmengen der reellen Zahlen sind die sogenannten *Intervalle*. Wir geben folgende Definitionen.

Intervalle

Definition 2.9. Zusammenhängende Teilmengen der reellen Zahlen heißen *Intervalle*. Für $a, b \in \mathbb{R}$ unterscheidet man

- das *abgeschlossene Intervall*

$$[a, b] := \{x \in \mathbb{R} | a \leq x \leq b\}, \quad (2.15)$$

- das *offene Intervall*

$$]a, b[:= \{x \in \mathbb{R} | a < x < b\}, \quad (2.16)$$

- und die *halboffenen Intervalle*

$$]a, b] := \{x \in \mathbb{R} | a < x \leq b\} \text{ und } [a, b[:= \{x \in \mathbb{R} | a \leq x < b\}. \quad (2.17)$$

•

Exemplarisch stellen wir in Abbildung 2.4 die Intervalle $[1, 2]$, $]1, 2]$, $[1, 2[$ und $]1, 2[$ grafisch dar. Dabei stellt man sich die reellen Zahlen als kontinuierlicher Zahlenstrahl vor und muss jeweils beachten, ob die linken bzw. rechten Endpunkte Teil des Intervalls sind oder nicht. Wie oben erwähnt sind Positiv-Unendlich (∞) und Negativ-Unendlich $-\infty$ keine Elemente von $\mathbb{R}$. Es gilt also immer $] - \infty, b]$ oder $] - \infty, b[$ bzw. $]a, \infty[$ oder $[a, \infty[$, sowie $\mathbb{R} =] - \infty, \infty[$.

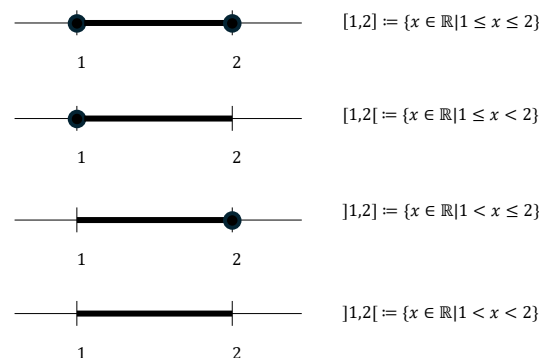


Abbildung 2.4. Darstellung von Intervallen am Zahlenstrahl. Der schwarze Punkt signalisiert dabei, dass die entsprechende Zahl Teil des Intervalls ist.

Kartesische Produkte

Oft möchte man mehrere unabhängige Eigenschaften eines Phänomens gleichzeitig quantitativ beschreiben. Zu diesem Zweck können die oben definierten eindimensionalen Zahlenmenge durch Bildung *Kartesischer Produkte* auf mehrdimensionale Zahlenmengen erweitert werden. Die Elemente Kartesischer Produkte nennt man *geordnete Tupel* oder *Vektoren*.

Definition 2.10 (Kartesische Produkte). M und N seien zwei Mengen. Dann ist das *Kartesische Produkt der Mengen M und N* die Menge aller geordneten Tupel (m, n) mit $m \in M$ und $n \in N$, formal

$$M \times N := \{(m, n) | m \in M, n \in N\}. \quad (2.18)$$

Das Kartesische Produkt einer Menge M mit sich selbst wird bezeichnet mit

$$M^2 := M \times M. \quad (2.19)$$

Seien weiterhin $M_1, M_2, \dots, M_n$ Mengen. Dann ist das *Kartesische Produkt der Mengen* $M_1, \dots, M_n$ die Menge aller geordneten n -Tupel $(m_1, \dots, m_n)$ mit $m_i \in M_i$ für $i = 1, \dots, n$, formal

$$\prod_{i=1}^n M_i := M_1 \times \dots \times M_n := \{(m_1, \dots, m_n) | m_i \in M_i \text{ für } i = 1, \dots, n\}. \quad (2.20)$$

Das n -fache Kartesische Produkt einer Menge M mit sich selbst wird bezeichnet mit

$$M^n := \prod_{i=1}^n M := \{(m_1, \dots, m_n) | m_i \in M\}. \quad (2.21)$$

•

Im Gegensatz zu Mengen sind die in Definition 2.10 eingeführten Tupel *geordnet*. Das heißt, für Mengen gilt beispielsweise $\{1, 2\} = \{2, 1\}$, aber für Tupel gilt $(1, 2) \neq (2, 1)$.

Beispiel

Es seien $M := \{1, 2\}$ und $N := \{1, 2, 3\}$. Dann ist das Kartesische Produkt $M \times N$ gegeben durch

$$M \times N := \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3)\} \quad (2.22)$$

und das Kartesische Produkt $N \times M$ ist gegeben durch

$$N \times M := \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (3, 2)\} \quad (2.23)$$

Das Kartesische Produkt ist also im Allgemeinen nicht kommutativ, es gilt also nicht notwendigerweise, dass $M \times N = N \times M$. Man mag sich die in diesem Beispiel konstruierten Mengen $M \times N \neq N \times M$ von Tabelle 2.1 und Tabelle 2.2 konstruieren.

Tabelle 2.1. Kartesisches Produkt $M \times N$

(m, n)	$n = 1$	$n = 2$	$n = 3$
$m = 1$	(1, 1)	(1, 2)	(1, 3)
$m = 2$	(2, 1)	(2, 2)	(2, 3)

Tabelle 2.2. Kartesisches Produkt $N \times M$

(n, m)	$m = 1$	$m = 2$
$n = 1$	(1, 1)	(1, 2)
$n = 2$	(2, 1)	(2, 2)
$n = 3$	(3, 1)	(3, 2)

$\mathbb{R}$ hoch n

Wie oben beschrieben eignen sich insbesondere die reellen Zahlen zur Beschreibung kontinuierlicher Phänomene. Zur simultanen Beschreibung mehrere Aspekte eines kontinuierlichen Phänomens bietet sich entsprechend die *Menge der reellen Tupel n -ter Ordnung*, kurz $\mathbb{R}$ hoch n , an.

Definition 2.11 (Menge der reellen Tupel n -ter Ordnung). Das n -fache Kartesische Produkt der reellen Zahlen mit sich selbst wird bezeichnet mit

$$\mathbb{R}^n := \prod_{i=1}^n \mathbb{R} := \{x := (x_1, \dots, x_n) \mid x_i \in \mathbb{R}\} \quad (2.24)$$

und wird “ $\mathbb{R}$ hoch n ” gesprochen. Wir schreiben die Elemente von $\mathbb{R}^n$ als Spalten

$$x := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad (2.25)$$

und nennen sie n -dimensionale Vektoren. Zu Abgrenzung nennen wir die Elemente von $\mathbb{R}^1 = \mathbb{R}$ auch *Skalare*.

•

Beispiele

Vertraute Beispiele von $\mathbb{R}^n$ sind $\mathbb{R}^1$ als Menge der reellen Zahlen, $\mathbb{R}^2$ als Menge der reellen Tupel im Modell der zweidimensionalen Ebene und $\mathbb{R}^3$ als Menge der reellen Tripel im Modell des dreidimensionalen Raumes wie in Abbildung 2.5 visualisiert.

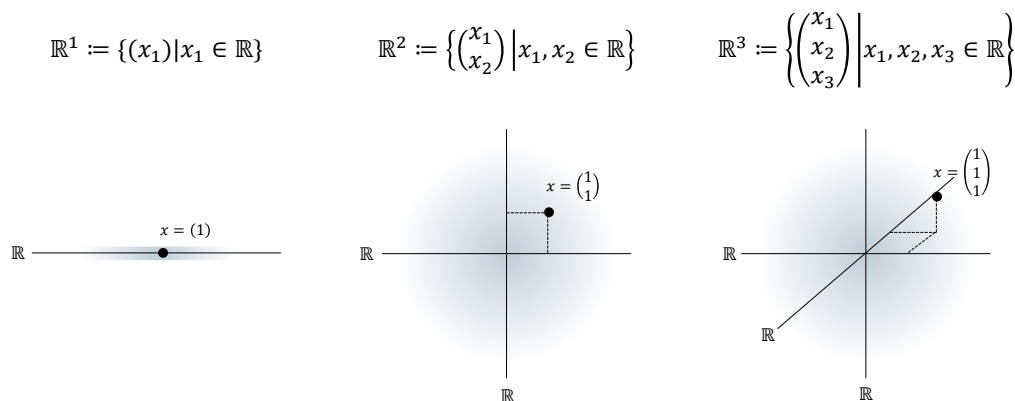


Abbildung 2.5. $\mathbb{R}^n$ für $n = 1, n = 2$ und $n = 3$. Die schwarzen Punkte stellen jeweils ein Element der entsprechenden Menge dar.

Ein Beispiel für ein $x \in \mathbb{R}^4$ ist

$$x = \begin{pmatrix} 0.16 \\ 1.76 \\ 0.23 \\ 7.11 \end{pmatrix}. \quad (2.26)$$

3. Summen, Produkte, Potenzen

Diese Einheit führt Schreibweisen für die Grundrechenarten ein.

3.1. Summen

Definition 3.1 (Summenzeichen). Es bezeichnet

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n. \quad (3.1)$$

Dabei stehen

- Σ für das griechische *Sigma*, mnemonisch für *Summe*,
- das Subskript $i = 1$ für den *Laufindex* und den *Startindex*,
- das Superskript n für den *Endindex* und
- $x_1, x_2, \dots, x_n$ für die *Summanden*.

•

Für eine sinnvolle Benutzung des Summenzeichens ist es essenziell, mithilfe des Subskripts und des Superskripts den Anfang und das Ende der Summation festzulegen. Die genaue Bezeichnung des Laufindex ist dagegen für den Wert der Summe irrelevant, es gilt

$$\sum_{i=1}^n x_i = \sum_{j=1}^n x_j. \quad (3.2)$$

Manchmal wird der Laufindex auch als Element einer *Indexmenge* angegeben. Ist z.B. die Indexmenge $I := \{1, 5, 7\}$ definiert, so ist

$$\sum_{i \in I} x_i := x_1 + x_5 + x_7. \quad (3.3)$$

Im Folgenden wollen wir kurz einige Beispiele für die Benutzung des Summenzeichens betrachten.

- *Summation vordefinierter Summanden*. Es seien $x_1 := 2$, $x_2 := 10$, $x_3 := -4$. Dann gilt

$$\sum_{i=1}^3 x_i = x_1 + x_2 + x_3 = 2 + 10 - 4 = 8. \quad (3.4)$$

- *Summation gewichteter vordefinierter Summanden.* Es seien wiederum $x_1 := 2$, $x_2 := 10$, $x_3 := -4$. Weiterhin seien die *Wichtungskoeffizienten* $a_1 := \frac{1}{2}$, $a_2 := \frac{1}{5}$, $a_3 := 2$ definiert. Dann gilt

$$\sum_{i=1}^3 a_i x_i = a_1 x_1 + a_2 x_2 + a_3 x_3 = \frac{1}{2} \cdot 2 + \frac{1}{5} \cdot 10 + 2 \cdot (-4) = 1 + 2 - 8 = -5. \quad (3.5)$$

Ausdrücke der Form $\sum_{i=1}^n a_i x_i$ werden auch als *Linearkombinationen* der $x_1, \dots, x_n$ mit den *Koeffizienten* oder *Wichtungsparametern* $a_1, \dots, a_n$ bezeichnet.

- *Summation natürlicher Zahlen.* Es gilt

$$\sum_{i=1}^5 i = 1 + 2 + 3 + 4 + 5 = 15. \quad (3.6)$$

- *Summation gerader natürlicher Zahlen.* Es gilt

$$\sum_{i=1}^5 2i = 2 \cdot 1 + 2 \cdot 2 + 2 \cdot 3 + 2 \cdot 4 + 2 \cdot 5 = 2 + 4 + 6 + 8 + 10 = 30. \quad (3.7)$$

- *Summation ungerader natürlicher Zahlen.* Es gilt

$$\sum_{i=1}^5 (2i-1) = 2 \cdot 1 - 1 + 2 \cdot 2 - 1 + 2 \cdot 3 - 1 + 2 \cdot 4 - 1 + 2 \cdot 5 - 1 = 1 + 3 + 5 + 7 + 9 = 25. \quad (3.8)$$

Der Umgang mit dem Summenzeichen kann oft durch die Anwendung folgender Rechenregeln vereinfacht werden.

Theorem 3.1 (Rechenregeln für Summen).

(1) *Summen gleicher Summanden*

$$\sum_{i=1}^n x = nx \quad (3.9)$$

(2) *Assoziativität bei Summen gleicher Länge*

$$\sum_{i=1}^n x_i + \sum_{i=1}^n y_i = \sum_{i=1}^n (x_i + y_i) \quad (3.10)$$

(3) *Distributivität bei Multiplikation mit einer Konstante*

$$\sum_{i=1}^n a x_i = a \sum_{i=1}^n x_i \quad (3.11)$$

(4) *Aufspalten von Summen mit $1 < m < n$*

$$\sum_{i=1}^n x_i = \sum_{i=1}^m x_i + \sum_{i=m+1}^n x_i \quad (3.12)$$

(5) *Umindizierung*

$$\sum_{i=0}^n x_i = \sum_{j=m}^{n+m} x_{j-m} \quad (3.13)$$

◦

Beweis. Man überzeugt sich von diesen Rechenregeln durch Ausschreiben der Summen und Anwenden der Rechenregeln von Addition und Multiplikation. Wir zeigen hier exemplarisch die Assoziativität bei Summen gleicher Länge und die Distributivität bei Multiplikation mit einer Konstante. Hinsichtlich ersterer haben wir

$$\begin{aligned}\sum_{i=1}^n x_i + \sum_{i=1}^n y_i &= x_1 + x_2 + \cdots + x_n + y_1 + y_2 + \cdots + y_n \\ &= x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n \\ &= \sum_{i=1}^n (x_i + y_i).\end{aligned}\tag{3.14}$$

Hinsichtlich letzterer gilt

$$\begin{aligned}\sum_{i=1}^n ax_i &= ax_1 + ax_2 + \cdots + ax_n \\ &= a(x_1 + x_2 + \cdots + x_n) \\ &= a \sum_{i=1}^n x_i.\end{aligned}\tag{3.15}$$

□

Beispiele

Als erstes Beispiel für die Anwendung der in Theorem 3.1 festgehaltenen Rechenregeln betrachten wir die Auswertung eines *Mittelwertes* (manchmal auch *Durchschnitt* genannt). Dazu seien $x_1, x_2, \dots, x_n$ reelle Zahlen. Der Mittelwert dieser Zahlen entspricht der Summe von $x_1, x_2, \dots, x_n$ geteilt durch die Anzahl der Zahlen n . Dabei ist es nach Aussage (3) von Theorem 3.1 irrelevant, ob zunächst die Zahlen aufaddiert werden und dann die resultierende Summe durch n geteilt wird oder die Zahlen jeweils einzeln durch n geteilt werden und die entsprechenden Ergebnisse dann aufaddiert werden. Genauer gilt durch Anwendung Theorem 3.1 (3) mit $a = 1/n$, dass

$$\frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^n \frac{x_i}{n}.\tag{3.16}$$

So ist zum Beispiel der Mittelwert von $x_1 := 1, x_2 := 4, x_3 := 2, x_4 := 1$ gegeben durch

$$\frac{1}{4} \sum_{i=1}^4 x_i = \frac{1}{4}(1 + 4 + 2 + 1) = \frac{8}{4} = 2 = \frac{8}{4} = \frac{1}{4} + \frac{4}{4} + \frac{2}{4} + \frac{1}{4} = \sum_{i=1}^4 \frac{x_i}{4}.\tag{3.17}$$

Als zweites Beispiel betrachten wir die in Theorem 3.1 (5) festgehaltene Umindizierungsregel. Dazu seien $n := 3$ und $m := 2$, sowie $x_0 := 2, x_1 := 3, x_2 := 5$ und $x_3 := 10$. Dann gilt offenbar

$$\begin{aligned}\sum_{i=0}^3 x_i &= x_0 + x_1 + x_2 + x_3 \\ &= 2 + 3 + 5 + 10 \\ &= 20.\end{aligned}\tag{3.18}$$

Ebenso gilt aber auch

$$\begin{aligned}
 \sum_{j=m}^{n+m} x_{j-m} &= \sum_{j=2}^{3+2} x_{j-2} \\
 &= \sum_{j=2}^5 x_{j-2} \\
 &= x_{2-2} + x_{3-2} + x_{4-2} + x_{5-2} \\
 &= x_0 + x_1 + x_2 + x_3 \\
 &= 2 + 3 + 5 + 10 \\
 &= 20.
 \end{aligned} \tag{3.19}$$

Doppelsummen

Nicht selten trifft man in der Anwendung auf Ausdrücke, die mehrere Summationen hintereinander ausführen. Dabei gelten für jede der Summen die oben gelisteten Definitionen und Rechenregeln. Man macht sich die Bedeutung von Doppelsummen am besten dadurch klar, dass man sie von innen nach außen ausschreibt und dabei beachtet, dass für jede Iteration der inneren Summe der Laufindex der äußeren Summe konstant bleibt.

Folgende Beispiele mögen dies verdeutlichen.

(1)

$$\begin{aligned}
 \sum_{i=1}^2 \sum_{j=1}^3 (i+j) &= \sum_{i=1}^2 (i+1 + i+2 + i+3) \\
 &= (1+1 + 1+2 + 1+3) + (2+1 + 2+2 + 2+3) \\
 &= 9 + 12 \\
 &= 21
 \end{aligned} \tag{3.20}$$

(2)

$$\begin{aligned}
 \sum_{i=1}^2 \sum_{j=1}^3 (x_i + y_j) &= \sum_{i=1}^2 (x_i + y_1 + x_i + y_2 + x_i + y_3) \\
 &= (x_1 + y_1 + x_1 + y_2 + x_1 + y_3) + (x_2 + y_1 + x_2 + y_2 + x_2 + y_3)
 \end{aligned} \tag{3.21}$$

(3)

$$\begin{aligned}
 \sum_{i=1}^3 \sum_{j=1}^2 (x_i y_j) &= \sum_{i=1}^3 (x_i y_1 + x_i y_2) \\
 &= (x_1 y_1 + x_1 y_2) + (x_2 y_1 + x_2 y_2) + (x_3 y_1 + x_3 y_2)
 \end{aligned} \tag{3.22}$$

3.2. Produkte

Eine analoge Schreibweise zum Summenzeichen bietet das Produktzeichen für Produkte.

Definition 3.2 (Produktzeichen). Es bezeichnet

$$\prod_{i=1}^n x_i = x_1 \cdot x_2 \cdot \dots \cdot x_n. \quad (3.23)$$

Dabei stehen

- $\prod$ für das griechische *Pi*, mnemonisch für *Produkt*,
- das Subskript $i = 1$ für den *Laufindex* und den *Startindex*,
- das Superskript n für den *Endindex*,
- $x_1, x_2, \dots, x_n$ für die *Produktterme*

•

Analog zum Summenzeichen gilt, dass das Produktzeichen nur mit Subskript und Superskripten zu Lauf- und Endindex Sinn ergibt. Die genaue Bezeichnung des Laufindex ist wiederum irrelevant, es gilt

$$\prod_{i=1}^n x_i = \prod_{j=1}^n x_j. \quad (3.24)$$

Auch hier wird in seltenen Fällen der Laufindex als Element einer Indexmenge angegeben. Ist z.B. die Indexmenge $J := \mathbb{N}_2^0$ definiert, so ist

$$\prod_{j \in J} x_j := x_0 \cdot x_1 \cdot x_2. \quad (3.25)$$

Ein Beispiel für die Benutzung des Produktzeichens ist etwa die Definition der *Fakultät* einer natürlichen Zahl n durch

$$n! := \prod_{i=1}^n i. \quad (3.26)$$

So ist etwa

$$3! := \prod_{i=1}^3 i = 1 \cdot 2 \cdot 3 = 6. \quad (3.27)$$

Auch für Produkte gibt es eine Reihe von Rechenregeln, die den Umgang mit ihnen oft vereinfachen. Wir listen einige in folgendem Theorem. Dabei machen wir vorgehend Gebrauch der in Kapitel 3.3 definierten Schreibweise von *Potenzen*.

Theorem 3.2 (Rechenregeln für Produkte).

(1) *Produkte gleicher Faktoren*

$$\prod_{i=1}^n x = x^n \quad (3.28)$$

(2) *Potenzierung von Konstanten*

$$\prod_{i=1}^n a x_i = a^n \prod_{i=1}^n x_i \quad (3.29)$$

(3) Aufspalten von Produkten mit $1 < m < n$

$$\prod_{i=1}^n x_i = \prod_{i=1}^m x_i \prod_{j=m+1}^n x_j \quad (3.30)$$

(4) Produkt von Produkten

$$\prod_{i=1}^n x_i y_i = \prod_{i=1}^n x_i \prod_{i=1}^n y_i \quad (3.31)$$

◦

3.3. Potenzen

Produkte von Zahlen mit sich selbst können mithilfe der Potenzschreibweise abgekürzt werden.

Definition 3.3 (Potenz). Für $a \in \mathbb{R}$ und $n \in \mathbb{N}^0$ ist die n -te Potenz von a definiert durch

$$a^0 := 1 \text{ und } a^{n+1} := a^n \cdot a. \quad (3.32)$$

Weiterhin ist für $a \in \mathbb{R} \setminus 0$ und $n \in \mathbb{N}^0$ die negative n -te Potenz von a definiert durch

$$a^{-n} := (a^n)^{-1} := \frac{1}{a^n}. \quad (3.33)$$

a wird dabei *Basis* und n wird *Exponent* genannt.

•

Die Art der Definition von a^{n+1} mit Rückbezug auf die Potenz a^n in obiger Definition nennt man *rekursiv*. Die Definition $a^0 := 1$ nennt man dabei den *Rekursionsanfang*; er macht die rekursive Definition von a^{n+1} erst möglich. Die Definition $a^{n+1} := a^n \cdot a$ nennt man auch *Rekursionsschritt*. Folgende Rechenregeln vereinfachen das Rechnen mit Potenzen.

Theorem 3.3 (Rechenregeln für Potenzen). Für $a, b \in \mathbb{R}$ und $n, m \in \mathbb{Z}$ mit $a \neq 0$ bei negativen Exponenten gelten folgende Rechenregeln:

$$a^n a^m = a^{n+m} \quad (3.34)$$

$$(a^n)^m = a^{nm} \quad (3.35)$$

$$(ab)^n = a^n b^n \quad (3.36)$$

◦

Anstelle eines Beweises betrachten wir folgende drei Beispiele.

$$(1) \quad 2^2 \cdot 2^3 = (2 \cdot 2) \cdot (2 \cdot 2 \cdot 2) = 2^5 = 2^{2+3} \quad (3.37)$$

$$(2) \quad (3^2)^3 = (3 \cdot 3)^3 = (3 \cdot 3) \cdot (3 \cdot 3) \cdot (3 \cdot 3) = 3^6 = 3^{2 \cdot 3} \quad (3.38)$$

(3)

$$(2 \cdot 4)^2 = (2 \cdot 4) \cdot (2 \cdot 4) = (2 \cdot 2) \cdot (4 \cdot 4) = 2^2 \cdot 4^2 \quad (3.39)$$

In enger Beziehung zur Potenz steht die Definition der n -ten Wurzel:

Definition 3.4 (n -te Wurzel). Für $a \in \mathbb{R}$ und $n \in \mathbb{N}$ ist die n -te Wurzel von a definiert als die reelle Zahl r mit

$$r^n = a. \quad (3.40)$$

•

Beim Rechnen mit Wurzeln ist die Potenzschreibweise von Wurzeln oft hilfreich, da sie die direkte Anwendung der Rechenregeln für Potenzen ermöglicht.

Theorem 3.4 (Potenzschreibweise der n -ten Wurzel). Es sei $a \in \mathbb{R}$, $n \in \mathbb{N}$ und r die n -te Wurzel von a . Dann gilt

$$r = a^{\frac{1}{n}} \quad (3.41)$$

◦

Beweis. Es gilt

$$\left(a^{\frac{1}{n}}\right)^n = a^{\frac{1}{n}} \cdot a^{\frac{1}{n}} \cdot \dots \cdot a^{\frac{1}{n}} = a^{\sum_{i=1}^n \frac{1}{n}} = a^1 = a. \quad (3.42)$$

Also gilt mit Definition 3.4, dass $r = a^{\frac{1}{n}}$. □

Das Rechnen mit Quadratwurzeln wird durch die Potenzschreibweise

$$\sqrt{x} = x^{\frac{1}{2}} \quad (3.43)$$

sehr erleichtert. So gilt zum Beispiel

$$\frac{2\pi}{\sqrt{2\pi}} = \frac{2\pi}{(2\pi)^{\frac{1}{2}}} = (2\pi)^1 \cdot (2\pi)^{-\frac{1}{2}} = (2\pi)^{1-\frac{1}{2}} = (2\pi)^{\frac{1}{2}} = \sqrt{2\pi}. \quad (3.44)$$

Folgender Zusammenhang zwischen Quadrat, Wurzel und Betrag wird des Öfteren genutzt.

Theorem 3.5 (Wurzel und Betrag). Es sei $x \in \mathbb{R}$ und es bezeichne

$$|x| := \begin{cases} x & \text{für } x \geq 0 \\ -x & \text{für } x < 0 \end{cases} \quad (3.45)$$

den Betrag von x . Dann gilt

$$\sqrt{x^2} = |x|. \quad (3.46)$$

◦

Beweis. Wir betrachten die Fälle $x \geq 0$ und $x < 0$. Sei zunächst $x \geq 0$. Dann gilt

$$x \geq 0 \Rightarrow x^2 \geq 0 \Rightarrow \sqrt{x^2} = x = |x|. \quad (3.47)$$

Sei nun $x < 0$. Dann gilt

$$x < 0 \Rightarrow x^2 > 0 \Rightarrow \sqrt{x^2} = -x = |x|. \quad (3.48)$$

□

4. Funktionen

Funktionen bilden neben den Mengen den zweiten Grundpfeiler der modernen Mathematik. In dieser Einheit definieren wir den Begriff der Funktion, führen erste Eigenschaften von Funktionen ein und geben eine Übersicht über einige elementare Funktionen.

4.1. Definition und Eigenschaften

Definition 4.1 (Funktion). Eine *Funktion* oder *Abbildung* f ist eine Zuordnungsvorschrift, die jedem Element einer Menge D genau ein Element einer Menge Z zuordnet. D wird dabei *Definitionsmenge* von f und Z wird *Zielmenge* von f genannt. Wir schreiben

$$f : D \rightarrow Z, x \mapsto f(x), \quad (4.1)$$

wobei $f : D \rightarrow Z$ gelesen wird als “die Funktion f bildet alle Elemente der Menge D eindeutig auf Elemente in Z ab” und $x \mapsto f(x)$ gelesen wird als “ x , welches ein Element von D ist, wird durch die Funktion f auf $f(x)$ abgebildet, wobei $f(x)$ ein Element von Z ist”. Der Pfeil $\rightarrow$ steht für die Abbildung zwischen den Mengen D und Z , der Pfeil $\mapsto$ steht für die Abbildung zwischen einem Element von D und einem Element von Z .

•

Es ist zentral, zwischen der *Funktion* f als Zuordnungsvorschrift und einem *Wert der Funktion* $f(x)$ als Element von Z zu unterscheiden. x ist das *Argument* der Funktion (der *Input* der Funktion), $f(x)$ der Wert, den die Funktion f für das Argument x annimmt (der *Output* der Funktion). Üblicherweise folgt in der Definition einer Funktion $f(x)$ die Definition der *funktionalen Form von f* , also einer Regel, wie aus x der Wert $f(x)$ zu bilden ist. Zum Beispiel wird in folgender Definition einer Funktion

$$f : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, x \mapsto f(x) := x^2 \quad (4.2)$$

die Definition der Potenz genutzt. Man beachte, dass Funktionen immer *eindeutig* sind, das heißt, dass sie jedem $x \in D$ bei jeder Anwendung der Funktion immer ein und dasselbe $f(x) \in Z$ zuordnen.

Abbildung 4.1 visualisiert einige Aspekte der Funktionsdefinition. Abbildung 4.1 A stellt dabei zunächst die zentralen Begriffe der Funktionsdefinition bildlich dar. Abbildung 4.1 B visualisiert ein erstes Beispiel für eine Funktion, bei der die den Argumenten zugeordneten Funktionswerte nicht durch eine Rechenregel, sondern durch direkte Definition bestimmt sind. Abbildung 4.1 C und Abbildung 4.1 D bedienen sich der Bildsprache, um zwei Aspekte der Definition von Funktionen mithilfe von Gegenbeispielen hervorzuheben: Bei der Zeichnung in Abbildung 4.1 C handelt es sich nicht um die Darstellung einer Funktion, da nach Definition 4.1 eine Funktion *jedem Element* einer Menge D genau ein Element einer Zielmenge Z zuordnet. Die hier angedeutete Zuordnungsvorschrift ordnet

dem Element $2 \in D$ aber kein Element in Z zu und ist deshalb keine Funktion. Gleichsam gilt nach Definition 4.1, dass eine Funktion jedem Element einer Menge D *genau ein Element einer Zielmenge* Z zuordnet. Die in Abbildung 4.1 D angedeutete Zuordnungsvorschrift ordnet $1 \in D$ aber sowohl das Element $a \in Z$ als auch das Element $b \in Z$ zu und ist deshalb mit Definition 4.1 nicht vereinbar.

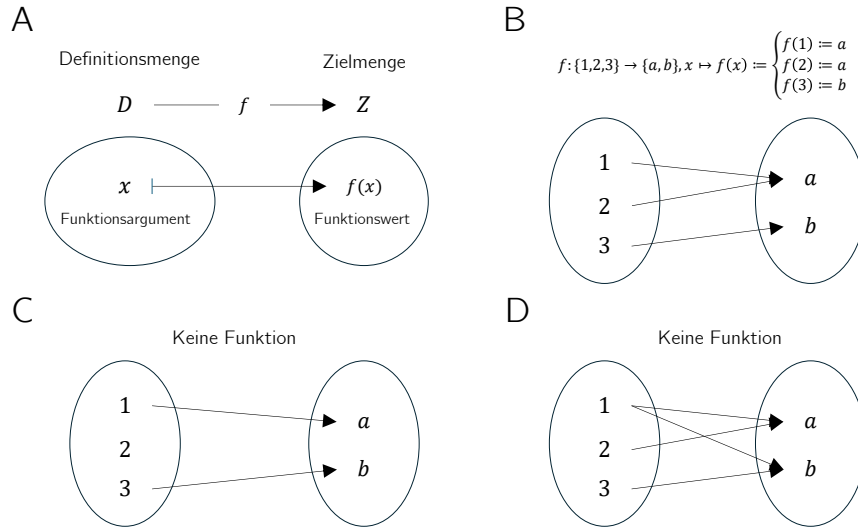


Abbildung 4.1. Aspekte der Funktionsdefinition

Funktionen setzen also Elemente von Mengen miteinander in Beziehung. Die Mengen dieser Elemente erhalten spezielle Bezeichnungen.

Definition 4.2 (Bildmenge, Wertebereich, Urbildmenge, Urbild). Es sei $f : D \rightarrow Z, x \mapsto f(x)$ eine Funktion und es seien $D' \subseteq D$ und $Z' \subseteq Z$. Die Menge

$$f(D') := \{z \in Z \mid \text{Es gibt ein } x \in D' \text{ mit } z = f(x)\} \quad (4.3)$$

heißt die *Bildmenge von D'* und $f(D) \subseteq Z$ heißt der *Wertebereich* von f . Weiterhin heißt die Menge

$$f^{-1}(Z') := \{x \in D \mid f(x) \in Z'\} \quad (4.4)$$

die *Urbildmenge von Z'* . Ein $x \in D$ mit $z = f(x) \in Z$ heißt ein *Urbild von z* .

•

Man beachte, dass der Wertebereich $f(D)$ von f und die Zielmenge Z von f nicht notwendigerweise identisch sein müssen.

Beispiel

Um die in Definition 4.2 eingeführten Begriffe zu verdeutlichen, betrachten wir die in Abbildung 4.2 A dargestellte Funktion

$$f : \{1, 2, 3, 4, 5\} \rightarrow \{a, b, c, d\}, x \mapsto f(x) := \begin{cases} f(1) := b \\ f(2) := d \\ f(3) := c \\ f(4) := c \\ f(5) := d \end{cases} \quad (4.5)$$

Nach Definition 4.2 ist eine Bildmenge immer bezüglich einer Teilmenge D' der Definitionsmenge D definiert. Sei also wie in Abbildung 4.2 B dargestellt $D' := \{2, 3\} \subset D$. Dann ist die Bildmenge von D' die Menge der $z \in Z$, für die ein $x \in D'$ existiert, sodass $z = f(x)$. Diejenigen $z \in Z$ für die ein $x \in D'$ mit $z = f(x)$ existiert sind aber hier gerade $c, d \in Z$, da $f(2) = d$ und $f(3) = c$. Darüberhinaus gibt es keine $z \in Z$ mit $f(x) = z$ und $x \in \{2, 3\}$. Der Wertebereich $f(D)$ ist nach Definition 4.2 die Teilmenge der $z \in Z$, für die gilt, dass ein $x \in D$ existiert, für das $z = f(x)$ ist. Dies ist für alle Elemente von Z der Fall, außer für $a \in Z$, da für dieses kein $x \in D$ existiert mit $a = f(x)$. Der Wertebereich von f ist für die betrachtete Funktion also durch $f(D) := \{b, c, d\}$ gegeben.

Umgekehrt ist nach Definition 4.2 ist eine Urbildmenge immer bezüglich einer Teilmenge Z' der Zielmenge Z definiert. Sei also wie in Abbildung 4.2 C dargestellt $Z' := \{c, d\} \subset Z$. Dann ist die Urbildmenge von Z' die Menge der $x \in D$, für die gilt, dass $f(x) \in Z'$. Diejenigen Elemente von D , deren Funktionswerte unter f durch Elemente von Z' gegeben sind, sind gerade $\{2, 3, 4, 5\}$. $1 \in D$ dagegen ist kein Element von Z' , da $f(1) = a$ auf $a \in Z$ abbildet und $a \notin Z'$. Nichtsdestotrotz ist natürlich $1 \in D$ ein Urbild von b .

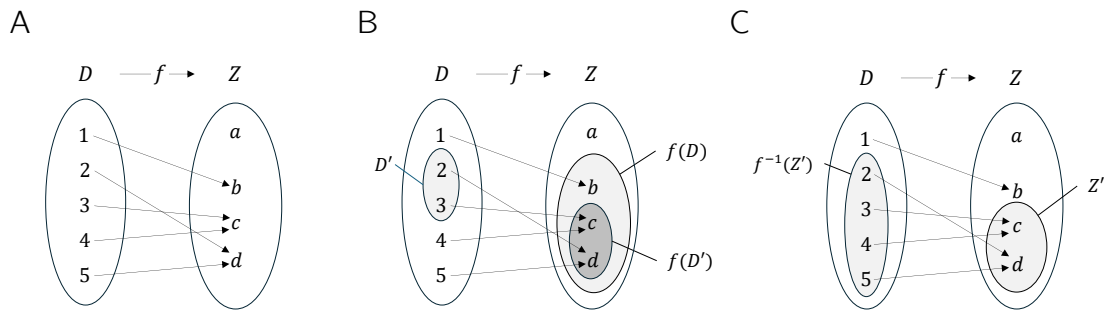


Abbildung 4.2. (A) Visualisierung der Funktion Gleichung 4.5. (B) Wertebereich und eine Bildmenge von f . (C) Urbildmenge einer Menge Z' unter f

Grundlegende Eigenschaften von Funktionen werden in folgender Definition benannt.

Definition 4.3 (Injektivität, Surjektivität, Bijektivität). $f : D \rightarrow Z, x \mapsto f(x)$ sei eine Funktion. f heißt *injektiv*, wenn es zu jedem Bild $z \in f(D)$ genau ein Urbild $x \in D$ gibt. Äquivalent gilt, dass f injektiv ist, wenn aus $x_1, x_2 \in D$ mit $x_1 \neq x_2$ folgt, dass $f(x_1) \neq f(x_2)$ ist. f heißt *surjektiv*, wenn $f(D) = Z$ gilt, wenn also jedes Element der Zielmenge Z ein Urbild in der Definitionsmenge D hat. Schließlich heißt f *bijektiv*, wenn f injektiv und surjektiv ist. Bijektive Funktionen werden auch *eineindeutige Funktionen* (engl. *one-to-one mappings*) genannt.

•

Beispiele

Wir verdeutlichen Definition 4.3 zunächst anhand dreier (Gegen)Beispiele in Abbildung 4.3. Abbildung 4.3 A visualisiert dabei die *nicht-injektive* Funktion

$$f : \{1, 2, 3\} \rightarrow \{a, b\}, x \mapsto f(x) := \begin{cases} f(1) := a \\ f(2) := a \\ f(3) := b \end{cases} . \quad (4.6)$$

Die Funktion ist nicht-injektiv, weil es zum Element a in der Bildmenge von f mehr als ein Urbild in der Definitionsmenge von f gibt, nämlich die Elemente 1 und 2. Abbildung 4.3 B visualisiert die *nicht-surjektive* Funktion

$$g : \{1, 2, 3\} \rightarrow \{a, b, c, d\}, x \mapsto g(x) := \begin{cases} g(1) := a \\ g(2) := b \\ g(3) := d \end{cases} . \quad (4.7)$$

Die Funktion ist nicht surjektiv, weil das Element c in der Zielmenge von f kein Urbild in der Definitionsmenge von f hat. Abbildung 4.3 C schließlich visualisiert die bijektive Funktion

$$h : \{1, 2, 3\} \rightarrow \{a, b, c\}, x \mapsto h(x) := \begin{cases} h(1) := a \\ h(2) := b \\ h(3) := c \end{cases} . \quad (4.8)$$

Zu *jedem* Element in der Zielmenge von h gibt es *genau ein* Urbild, die Funktion ist also injektiv und surjektiv und damit bijektiv.

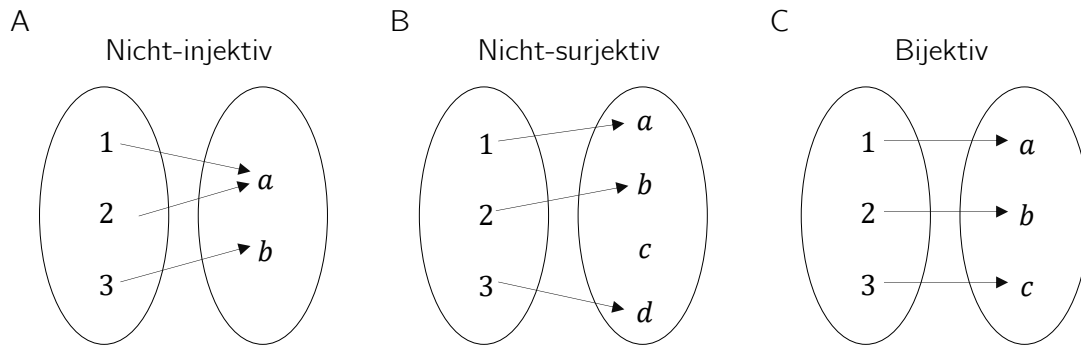


Abbildung 4.3. Injektivität, Surjektivität, Bijektivität.

Als weiteres Beispiel betrachten wir die Funktion

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := x^2 \quad (4.9)$$

Diese Funktion ist nicht injektiv, weil z.B. für $x_1 = 2 \neq -2 = x_2$ gilt, dass $f(x_1) = 2^2 = 4 = (-2)^2 = f(x_2)$. Weiterhin ist f auch nicht surjektiv, weil z.B. $-1 \in \mathbb{R}$ kein Urbild unter f hat. Schränkt man die Definitionsmenge von f allerdings auf die nicht-negativen reellen Zahlen ein, definiert man also die Funktion

$$\tilde{f} : [0, \infty[\rightarrow [0, \infty[, x \mapsto \tilde{f}(x) := x^2, \quad (4.10)$$

so ist $\tilde{f}$ im Gegensatz zu f injektiv und surjektiv, also bijektiv.

Zur Überabzählbarkeit der reellen Zahlen

Wir wollen an dieser Stelle die Begriffe der Surjektivität und Bijektivität von Funktionen nutzen, um den Begriff der *Überabzählbarkeit der reellen Zahlen* noch etwas vertiefen (vgl. Kapitel 2.3). Dieser besagt, dass es verschiedene Arten von Unendlichkeiten gibt, was zunächst sicherlich intuitiv nicht leicht zu fassen ist. Weiterhin sind manche Aspekte der Wahrscheinlichkeitstheorie so gefasst, dass sie der Überabzählbarkeit der reellen Zahlen

Genüge tun, sodass zumindest ein grobes Verständnis dieser Eigenschaft zur Motivation mancher Spitzfindigkeiten der Wahrscheinlichkeitstheorie dienen kann. Speziell wollen wir an dieser Stelle das *Cantorsche Diagonalargument* (Cantor (1891), Gray (1994)) zur Überabzählbarkeit der reellen Zahlen grob skizzieren. Wie beginnen damit, zunächst festzuhalten, was vor dem Hintergrund der bijektiven Abbildungen mit einer *endlichen* und einer *abzählbar unendlichen* Menge gemeint sein soll.

Definition 4.4. Eine Menge M heißt *endlich*, wenn es eine bijektive Abbildung der Form

$$f : \mathbb{N}_n \rightarrow M$$

(4.11)

auf die Elemente von M gibt. Weiterhin heißt eine Menge M *abzählbar unendlich*, wenn es eine bijektive Abbildung der Form

$$f : \mathbb{N} \rightarrow M$$

(4.12)

gibt.

•

Im Falle einer endlichen Menge kann jedem Element der Menge also im Sinne einer bijektiven Abbildung eine natürliche Zahl mit maximalem Wert $n < \infty$ zugeordnet werden. Im Falle einer abzählbar unendlichen Menge kann jedem Element der Menge im Sinne einer bijektiven Abbildung zumindest eine natürliche Zahl zugeordnet werden. Die Elemente dieser Mengen können also abgezählt werden.

Cantor (1891) zeigte nun, dass dies im Allgemeinen für die reellen Zahlen nicht zutrifft, dass also keine bijektive Abbildung zwischen den reellen Zahlen und den natürlichen Zahlen konstruiert werden kann und es somit mehr reelle Zahlen als natürliche Zahlen geben muss. Speziell kann man argumentieren, dass sogar das Intervall $[0, 1]$ schon mehr reelle Zahlen enthält als es natürliche Zahlen gibt, dass also keine surjektive Funktion von $\mathbb{N}$ nach $[0, 1]$ konstruiert werden kann.

Das Argument von Cantor (1891) hat die folgende Form. Man nehme an, $f : \mathbb{N} \rightarrow [0, 1]$ sei eine injektive Funktion, die tabellarisch wie beispielsweise in Tabelle 4.1 dargestellt sei.

Tabelle 4.1. Abbildung $f : \mathbb{N} \rightarrow [0, 1]$. Die Punkte $\dots$ und $:$ zeigen an, dass die Tabelle bis ins Unendliche nach rechts und nach unten fortgesetzt wird.

n	$f(n)$										
1	0.	3	1	4	1	5	9	2	6	5	...
2	0.	8	9	4	5	9	7	8	1	0	...
3	0.	9	6	2	3	2	1	5	8	7	...
4	0.	5	3	6	8	8	8	9	7	3	...
5	0.	7	4	3	8	1	8	0	9	7	...
$\vdots$	$\vdots$										

Man konstruiere nun eine weitere Zahl in $[0, 1]$ dadurch, dass man zu den fett gedruckten Diagonalelementen in Tabelle 4.1 jeweils 1 addiert, wenn der entsprechende Wert kleiner als 9 ist und ansonsten den Wert auf 0 setzt. Dies ist in Tabelle 4.2 gezeigt.

Tabelle 4.2. Konstruktion einer weiteren reellen Zahl in $[0, 1]$.

n	$f(n)$										
1	0.	4	1	4	1	5	9	2	6	5	...
2	0.	8	0	4	5	9	7	8	1	0	...
3	0.	9	6	3	3	2	1	5	8	7	...
4	0.	5	3	6	9	8	8	9	7	3	...
5	0.	7	4	3	8	2	8	0	9	7	...
$\vdots$	$\vdots$										

Die so konstruierte Zahl

0.40392... (4.13)

kann nicht in Tabelle 4.1 auftauchen, da sie sich von $f(1)$ durch ihren Eintrag an der ersten Dezimalstelle, von $f(2)$ durch ihren Eintrag an der zweiten Dezimalstelle, von $f(3)$ durch ihren Eintrag an der dritten Dezimalstelle, ..., von $f(n)$ durch ihren Eintrag an der n ten Dezimalstelle, und so ins Unendliche weiter unterscheidet. Also gibt es zumindest eine Zahl in $[0, 1]$, der keine natürliche Zahl zugeordnet werden kann, die injektive Abbildung f ist nicht surjektiv und damit auch nicht bijektiv. Somit ist aber $[0, 1]$ nicht abzählbar unendlich und wird entsprechend *überabzählbar* genannt.

4.2. Funktionentypen

Durch Verkettung lassen sich aus Funktionen weitere Funktionen bilden.

Definition 4.5 (Verkettung von Funktionen). Es seien $f : D \rightarrow Z$ und $g : Z \rightarrow S$ zwei Funktionen, wobei die Wertemenge von f mit der Definitionsmenge von g übereinstimmen sollen. Dann ist durch

$$g \circ f : D \rightarrow S, x \mapsto (g \circ f)(x) := g(f(x))$$

(4.14)

eine Funktion definiert, die die *Verkettung von f und g* genannt wird.

•

Die Schreibweise für verkettete Funktionen ist etwas gewöhnungsbedürftig. Wichtig ist es zu erkennen, dass $g \circ f$ die verkettete Funktion und $(g \circ f)(x)$ ein Element in der Zielmenge der verketteten Funktion bezeichnen. Intuitiv wird bei der Auswertung von $(g \circ f)(x)$ zunächst die Funktion f auf x angewendet und dann die Funktion g das Element auf $f(x)$ von R angewendet. Dies ist in der funktionalen Form $g(f(x))$ festgehalten. Der Einfachheit halber benennt man die Verkettung zweier Funktionen auch oft mit einem einzelnen Buchstaben und schreibt beispielsweise, $h := g \circ f$ mit $h(x) = g(f(x))$. Leicht zur Verwirrung kann es führen, wenn Elemente in der Zielmenge von f mit y bezeichnet werden, also die Schreibweise $y = f(x)$ und $h(x) = g(y)$ genutzt wird. Allerdings ist diese Schreibweise manchmal zur notationellen Vereinfachung nötig. Wir visualisieren Definition 4.5 in Abbildung 4.4.

Als Beispiel für die Verkettung zweier Funktionen betrachten wir

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := -x^2$$

(4.15)

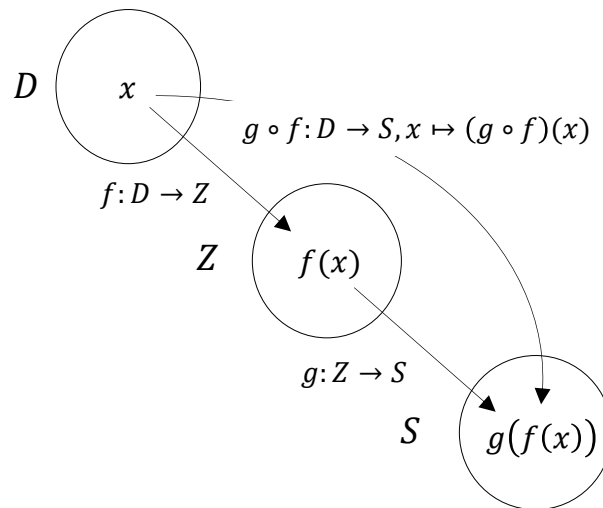


Abbildung 4.4. Verkettung von Funktionen.

und

$$g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto g(x) := \exp(x). \quad (4.16)$$

Die Verkettung von f und g ergibt sich in diesem Fall zu

$$g \circ f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto (g \circ f)(x) := g(f(x)) = \exp(-x^2). \quad (4.17)$$

Eine erste Anwendung der Verkettung von Funktionen findet sich in folgender Definition.

Definition 4.6 (Inverse Funktion). Es sei $f : D \rightarrow Z, x \mapsto f(x)$ eine bijektive Funktion. Dann heißt die Funktion f^{-1} mit

$$f^{-1} \circ f : D \rightarrow D, x \mapsto (f^{-1} \circ f)(x) := f^{-1}(f(x)) = x \quad (4.18)$$

inverse Funktion, Umkehrfunktion oder einfach *Inverse von f* .

•

Inverse Funktionen sind immer bijektiv. Dies folgt, weil f bijektiv ist und damit jedem $x \in D$ genau ein $f(x) = z \in Z$ zugeordnet wird. Damit wird aber auch jedem $z \in Z$ genau ein $x \in D$, nämlich $f^{-1}(f(x)) = x$ zugeordnet. Intuitiv macht die inverse Funktion von f den Effekt von f auf ein Element x rückgängig. Wir visualisieren Definition 4.6 in Abbildung 4.5 A.

Betrachtet man konkret den Graphen einer Funktion in einem Kartesischen Koordinatensystem, so führt die Anwendung von einem Wert auf der x -Achse zu einem Wert auf der y -Achse. Die Anwendung der inversen Funktion führt dementsprechend von einem Wert auf der y -Achse zu einem Wert auf der x -Achse. Wir visualisieren dies in Abbildung 4.5 B. Betrachten wir beispielsweise die Funktion

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := 2x =: y. \quad (4.19)$$

Dann ist die inverse Funktion von f gegeben durch

$$f^{-1} : \mathbb{R} \rightarrow \mathbb{R}, y \mapsto f^{-1}(y) := \frac{1}{2}y, \quad (4.20)$$

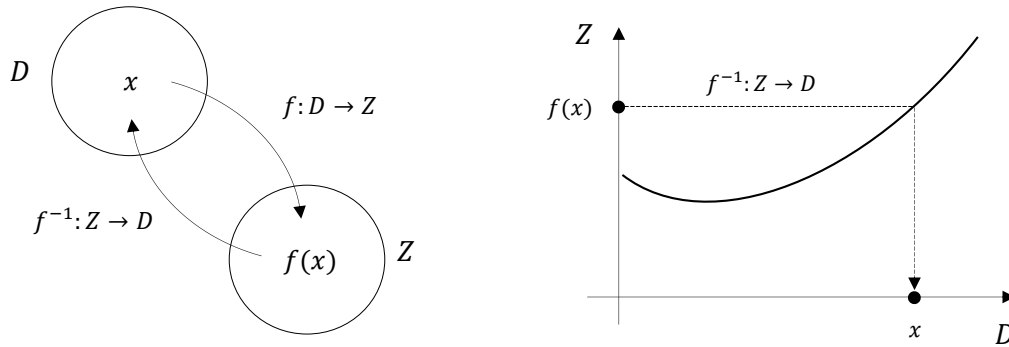


Abbildung 4.5. Inverse Funktion.

weil für jedes $x \in \mathbb{R}$ gilt, dass

$$(f^{-1} \circ f)(x) := f^{-1}(f(x)) = f^{-1}(2x) = \frac{1}{2} \cdot 2x = x. \quad (4.21)$$

Eine wichtige Klasse von Funktionen sind *lineare Abbildungen*.

Definition 4.7 (Lineare Abbildung). Eine Abbildung $f : D \rightarrow Z, x \mapsto f(x)$ heißt *lineare Abbildung*, wenn für $x, y \in D$ und einen Skalar c gelten, dass

$$f(x + y) = f(x) + f(y) \quad (\text{Additivität})$$

und

$$f(cx) = cf(x) \quad (\text{Homogenität})$$

Eine Abbildung, für die obige Eigenschaften nicht gelten, heißt *nicht-lineare Abbildung*.

•

Lineare Abbildungen sind oft als “gerade Linien” bekannt. Die allgemeine Definition linearer Abbildungen ist mit dieser Intuition nicht komplett kongruent. Insbesondere sind lineare Abbildungen nur solche Funktionen, die den Nullpunkt auf den Nullpunkt abbilden. Wir zeigen dazu folgendes Theorem.

Theorem 4.1 (Lineare Abbildung der Null). $f : D \rightarrow Z$ sei eine lineare Abbildung. Dann gilt

$$f(0) = 0. \quad (4.22)$$

◦

Beweis. Wir halten zunächst fest, dass mit der Additivität von f gilt, dass

$$f(0) = f(0 + 0) = f(0) + f(0). \quad (4.23)$$

Addition von $-f(0)$ auf beiden Seiten obiger Gleichung ergibt dann

$$f(0) - f(0) = f(0) + f(0) - f(0) \Leftrightarrow f(0) = 0 \quad (4.24)$$

und damit ist alles gezeigt. $\square$

Wir wollen den Begriff der linearen Abbildung noch an zwei Beispielen verdeutlichen.

- Für $a \in \mathbb{R}$ ist die Abbildung

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := ax \quad (4.25)$$

eine lineare Abbildung, weil gilt, dass

$$f(x+y) = a(x+y) = ax + ay = f(x) + f(y) \text{ und } f(cx) = acx = cax = cf(x). \quad (4.26)$$

- Für $a, b \in \mathbb{R}$ ist dagegen die Abbildung

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := ax + b \quad (4.27)$$

nicht-linear, weil z.B. für $a := b := 1$ gilt, dass

$$f(x+y) = 1(x+y) + 1 = x+y+1 \neq x+1+y+1 = f(x) + f(y). \quad (4.28)$$

Eine Abbildung der Form $f(x) := ax + b$ heißt *linear-affine Abbildung* oder *linear-affine Funktion*. Etwas unsauber werden Funktionen der Form $f(x) := ax + b$ auch manchmal als *lineare Funktionen* bezeichnet.

Neben den bisher diskutierten Funktionentypen gibt es noch viele weitere Klassen von Funktionen. In folgender Definition klassifizieren wir Funktionen anhand der Dimensionalität ihrer Definitions- und Zielmengen. Diese Art der Funktionsklassifikation ist oft hilfreich, um sich einen ersten Überblick über ein mathematisches Modell zu verschaffen.

Definition 4.8 (Funktionenarten). Wir unterscheiden

- *univariate reellwertige Funktionen* der Form

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x), \quad (4.29)$$

- *multivariate reellwertige Funktionen* der Form

$$f : \mathbb{R}^n \rightarrow \mathbb{R}, x \mapsto f(x) = f(x_1, \dots, x_n), \quad (4.30)$$

- und *multivariate vektorwertige Funktionen* der Form

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m, x \mapsto f(x) = \begin{pmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{pmatrix}, \quad (4.31)$$

wobei $f_i, i = 1, \dots, m$ die *Komponenten(funktionen)* von f genannt werden.

•

In der Physik werden multivariate reellwertige Funktionen *Skalarfelder* und multivariate vektorwertige Funktionen *Vektorfelder* genannt. In manchen Anwendungen treten auch *matrixvariante matrixwertige Funktionen* auf.

4.3. Elementare Funktionen

Als *elementare Funktionen* bezeichnen wir eine kleine Schar von univariaten reellwertigen Funktionen, die häufig als Bausteine komplexerer Funktionen auftreten. Dies sind die *Polynomfunktionen*, die *Exponentialfunktion*, die *Logarithmusfunktion* und die *Gammafunktion*. Im Folgenden geben wir die Definitionen dieser Funktionen sowie ihre wesentlichen Eigenschaften als Theoreme an und stellen ihre Graphen an. Für Beweise der Eigenschaften der hier vorgestellten Funktionen verweisen wir auf die weiterführende Literatur.

Definition 4.9 (Polynomfunktionen). Eine Funktion der Form

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := \sum_{i=0}^k a_i x^i = a_0 + a_1 x^1 + a_2 x^2 + \dots + a_k x^k$$

(4.32)

heißt *Polynomfunktion* *k*-ten Grades mit Koeffizienten $a_0, a_1, \dots, a_k \in \mathbb{R}$. Typische Polynomfunktionen sind in Tabelle 4.3 aufgelistet.

Tabelle 4.3. Polynomfunktionen.

Name	Form	Koeffizienten
Konstante Funktion	$f(x) = a$	$a_0 := a, a_i := 0, i > 0$
Identitätsfunktion	$f(x) = x$	$a_0 := 0, a_1 := 1, a_i := 0, i > 1$
Linear-affine Funktion	$f(x) = ax + b$	$a_0 := b, a_1 := a, a_i := 0, i > 1$
Quadratfunktion	$f(x) = x^2$	$a_0 := 0, a_1 := 0, a_2 := 1, a_i := 0, i > 2$

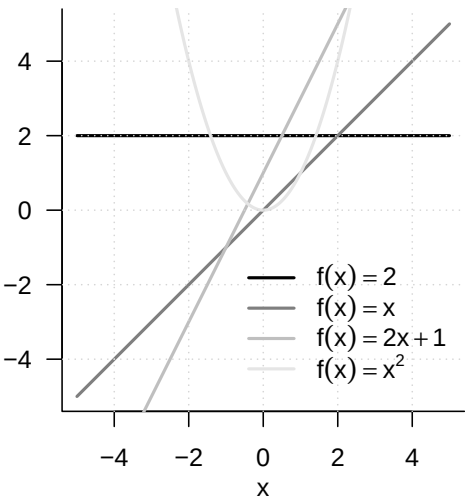


Abbildung 4.6. Ausgewählte Polynomfunktionen

Abbildung 4.6 zeigt die Graphen der in Definition 4.9 aufgelisteten Polynomfunktionen.

Ein wichtiges Funktionenpaar sind die Exponentialfunktion und die Logarithmusfunktion.

Theorem 4.2 (Exponentialfunktion und ihre Eigenschaften). *Die Exponentialfunktion ist definiert als*

$$\exp : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto \exp(x) := e^x := \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \quad (4.33)$$

Die Exponentialfunktion hat unter anderem die in Tabelle 4.4 aufgeführten Eigenschaften.

Tabelle 4.4. Eigenschaften der Exponentialfunktion.

Eigenschaft	Bedeutung
Wertebereich	$x \in]-\infty, 0[\Rightarrow \exp(x) \in]0, 1[$ $x \in]0, \infty[\Rightarrow \exp(x) \in]1, \infty[$
Monotonie	$x < y \Rightarrow \exp(x) < \exp(y)$
Spezielle Werte	$\exp(0) = 1$ und $\exp(1) = e$
Summationseigenschaft	$\exp(x + y) = \exp(x) \exp(y)$
Subtraktionseigenschaft	$\exp(x - y) = \frac{\exp(x)}{\exp(y)}$

◦

Insbesondere nimmt die Exponentialfunktion nur positive Werte an und schneidet die y -Achse bei $x = 0$. Die Zahl $\exp(1) := e \approx 2.71\dots$ heißt *Eulersche Zahl*. Schließlich gilt mit den speziellen Werten der Exponentialfunktion auch

$$\exp(x) \exp(-x) = \exp(x - x) = \exp(0) = 1. \quad (4.34)$$

Theorem 4.3 (Logarithmusfunktion und ihre Eigenschaften). *Die Logarithmusfunktion ist definiert als inverse Funktion der Exponentialfunktion,*

$$\ln :]0, \infty[\rightarrow \mathbb{R}, x \mapsto \ln(x) \text{ mit } \ln(\exp(x)) = x \text{ für alle } x \in \mathbb{R}. \quad (4.35)$$

Die Logarithmusfunktion hat unter anderem die in Tabelle 4.5 aufgeführten Eigenschaften.

Tabelle 4.5. Eigenschaften der Logarithmusfunktion.

Eigenschaft	Bedeutung
Wertebereich	$x \in]0, 1[\Rightarrow \ln(x) \in]-\infty, 0[$ $x \in]1, \infty[\Rightarrow \ln(x) \in]0, \infty[$
Monotonie	$x < y \Rightarrow \ln(x) < \ln(y)$
Spezielle Werte	$\ln(1) = 0$ und $\ln(e) = 1$
Produkteigenschaft	$\ln(xy) = \ln(x) + \ln(y)$
Potenzeigenschaft	$\ln(x^c) = c \ln(x)$
Divisionseigenschaft	$\ln\left(\frac{1}{x}\right) = -\ln(x)$

◦

Im Gegensatz zur Exponentialfunktion nimmt die Logarithmusfunktion sowohl negative als auch positive Werte an . Die Logarithmusfunktion schneidet die x -Achse bei $x = 1$. Die Produkteigenschaft und die Potenzigenschaften sind beim Rechnen mit der Logarithmusfunktion zentral. Man merkt sie sich intuitiv als “Die Logarithmusfunktion wandelt Produkte in Summen und Potenzen in Produkte um.” Die Graphen der Exponential- und Logarithmusfunktion sind in [Abbildung 4.7](#) abgebildet.

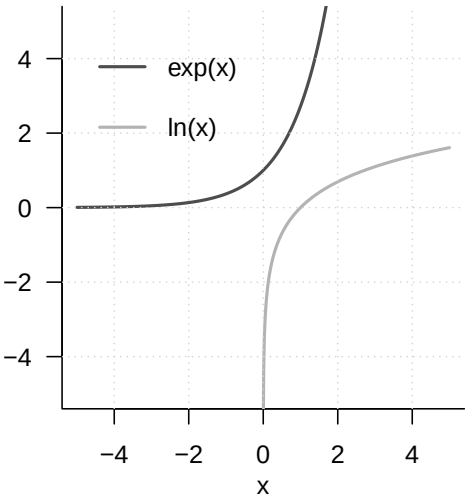


Abbildung 4.7. Exponentialfunktion und Logarithmusfunktion

Ein häufiger Begleiter in der Wahrscheinlichkeitstheorie ist die *Gammafunktion*.

Definition 4.10 (Gammafunktion). Die *Gammafunktion* ist definiert durch

$$\Gamma : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto \Gamma(x) := \int_0^\infty \xi^{x-1} \exp(-\xi) d\xi$$

(4.36)

Die Gammafunktion hat unter anderem die in [Tabelle 4.6](#) aufgeführten Eigenschaften.

Tabelle 4.6. Eigenschaften der Gammafunktion.

Eigenschaft	Bedeutung
Spezielle Werte	$\Gamma(1) = 1$
	$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
	$\Gamma(n) = (n-1)!$ für $n \in \mathbb{N}$
Rekursionseigenschaft	Für $x > 0$ gilt $\Gamma(x+1) = x\Gamma(x)$

•

Ein Ausschnitt des Graphen der Gammafunktion ist in [Abbildung 4.8](#) dargestellt.

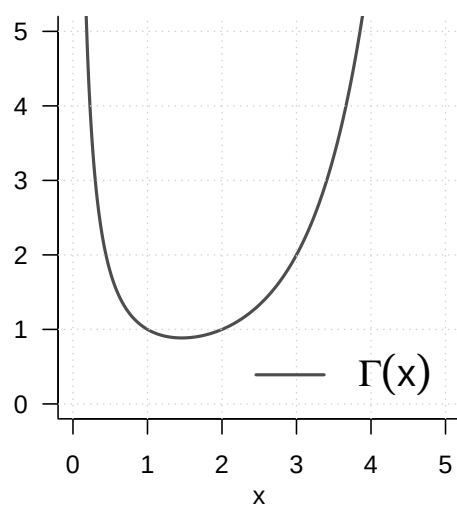


Abbildung 4.8. Gammafunktion

5. Folgen, Grenzwerte, Stetigkeit

Die in diesem Kapitel behandelten Themen sind in der probabilistischen Datenwissenschaft nicht zentral, sondern bilden Grundbausteine der reellen Analysis. Durch die enge Verschränkung der modernen Wahrscheinlichkeitstheorie mit analytischen Ansätzen dienen sie jedoch dem Verständnis bestimmter Resultate der Wahrscheinlichkeitstheorie. Ein Beispiel für ein solches Resultat ist der *Zentrale Grenzwertsatz*. Der Zentrale Grenzwertsatz wiederum bildet die Grundlage für die weit verbreitete Normalverteilungsannahme in der probabilistischen Datenwissenschaft. Die hier betrachteten Grundlagen erhöhen also mittelbar das Verständnis datenwissenschaftlicher Prinzipien. In aller Kürze ist der Zentrale Grenzwertsatz eine Aussage über die *Grenzfunktion* einer *Funktionenfolge*, genauer einer Folge von Zufallsvariablen. Das Wissen um das Wesen von *Folgen*, *Funktionenfolgen* und ihren *Grenzwerten* erlaubt also einen informierten Zugang zum Studium des Zentralen Grenzwertsatzes. Weiterhin ermöglichen die in diesem Kapitel behandelten Themen zumindest einen ersten Einstieg in das Verständnis der Stetigkeit und Glattheit von Funktionen, die beispielsweise in der nichtlinearen Optimierung zur Bestimmung von Parameterschätzern probabilistischer Modelle wichtige Grundkonzepte bilden.

5.1. Folgen

Wir beginnen mit der Definition des Begriffs der *reellen Folge*.

Definition 5.1 (Reelle Folge). Eine *reelle Folge* ist eine Funktion der Form

$$f : \mathbb{N} \rightarrow \mathbb{R}, n \mapsto f(n). \quad (5.1)$$

Die Funktionswerte $f(n)$ einer reellen Folge werden üblicherweise mit x_n bezeichnet und *Folgenglieder* genannt. Übliche Schreibweisen für Folgen sind

$$(x_1, x_2, \dots) \text{ oder } (x_n)_{n=1}^{\infty} \text{ oder } (x_n)_{n \in \mathbb{N}} \text{ oder } (x_n). \quad (5.2)$$

•

Man beachte, dass weil es unendlich viele natürliche Zahlen gibt, eine reelle Folge immer unendlich viele Folgenglieder hat. Dies sollte man sich insbesondere bei der Schreibweise $(x_1, x_2, \dots)$ bewusst machen. Wir wollen zwei Standardbeispiele für reelle Folgen betrachten.

Beispiele für reelle Folgen

(1) Reelle Folgen der Form

$$f : \mathbb{N} \rightarrow \mathbb{R}, n \mapsto f(n) := \left(\frac{1}{n}\right)^{\frac{p}{q}} \text{ mit } p, q \in \mathbb{N} \quad (5.3)$$

nennt man *harmonische Folgen*. Für $p := q := 1$ hat eine harmonische Folge die Folgengliederform

$$\left(\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \dots\right). \quad (5.4)$$

(2) Reelle Folgen der Form

$$f : \mathbb{N} \rightarrow \mathbb{R}, n \mapsto f(n) := q^n \text{ mit } q \in]-1, 1[\quad (5.5)$$

nennt man *geometrische Folgen*. Für $q := \frac{1}{2}$ hat eine geometrische Folge die Folgengliederform

$$\begin{aligned} \left(\left(\frac{1}{2}\right)^1, \left(\frac{1}{2}\right)^2, \left(\frac{1}{2}\right)^3, \dots\right) &= \left(\left(\frac{1}{2}\right)^1, \left(\frac{1}{2}\right)^2, \left(\frac{1}{2}\right)^3, \dots\right) \\ &= \left(\frac{1^1}{2^1}, \frac{1^2}{2^2}, \frac{1^3}{2^3}, \dots\right) \\ &= \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots\right) \end{aligned} \quad (5.6)$$

Neben den reellen Folgen, also Folgen reeller Zahlen, kann man auch Folgen anderer mathematischer Objekte betrachten. Eine wichtige Folgenart sind die *Funktionenfolgen*.

Definition 5.2 (Funktionsfolge). Es sei ϕ eine Menge univariater reellwertiger Funktionen mit Definitionsmenge $D \subseteq \mathbb{R}$. Dann ist eine Funktionenfolge eine Funktion der Form

$$F : \mathbb{N} \rightarrow \phi, n \mapsto F(n). \quad (5.7)$$

Die Funktionswerte $F(n)$ einer Funktionenfolge werden üblicherweise mit f_n bezeichnet und *Folgenglieder* genannt. Übliche Schreibweisen für Funktionenfolgen sind

$$(f_1, f_2, \dots) \text{ oder } (f_n)_{n=1}^{\infty} \text{ oder } (f_n)_{n \in \mathbb{N}} \text{ oder } (f_n). \quad (5.8)$$

•

Die Definition einer Funktionenfolge ist offenbar analog zur Definition einer reellen Folge. Der Unterschied zwischen einer reellen Folge und einer Funktionenfolge ist, dass die Folgenglieder einer reellen Folge reelle Zahlen, die Folgenglieder einer Funktionenfolge dagegen univariate reellwertige Funktionen sind. Auch hier wollen wir zwei Standardbeispiele diskutieren.

Beispiele für Funktionenfolgen

(1) Wir betrachten die Menge ϕ der univariaten reellwertigen Funktionen der Form

$$\phi := \{f_n | f_n : [0, 1] \rightarrow \mathbb{R}, x \mapsto f_n(x) := x^n \text{ für } n \in \mathbb{N}\} \quad (5.9)$$

Dann definiert

$$F : \mathbb{N} \rightarrow \phi, n \mapsto F(n) := f_n \quad (5.10)$$

eine Funktionenfolge. Für die Funktionswerte der Folgenglieder von F gilt

$$f_1(x) := x^1, f_2(x) := x^2, f_3(x) := x^3, \dots \quad (5.11)$$

(2) Wir betrachten die Menge ϕ der univariaten reellwertigen Funktionen der Form

$$\phi := \{f_n | f_n : [-a, a] \rightarrow \mathbb{R}, x \mapsto f_n(x) := \sum_{k=0}^n \frac{x^k}{k!} \text{ für } n \in \mathbb{N}\} \quad (5.12)$$

Dann definiert

$$F : \mathbb{N} \rightarrow \phi, n \mapsto F(n) := f_n \quad (5.13)$$

eine Funktionenfolge. Für die Funktionswerte der Folgenglieder von F gilt

$$f_1(x) := \sum_{k=0}^1 \frac{x^k}{k!}, f_2(x) := \sum_{k=0}^2 \frac{x^k}{k!}, f_3(x) := \sum_{k=0}^3 \frac{x^k}{k!}, \dots \quad (5.14)$$

5.2. Grenzwerte

Wenn man die Folgenglieder einer Folge betrachtet, kann man sich fragen, welche Werte eine Folge wohl annimmt, wenn der Folgenindex n sehr groß wird, also gegen unendlich strebt. Wenn in diesem Fall die Folgenglieder sehr ähnliche Werte annehmen (und nicht etwa auch unendlich groß werden), so ist man auf den Begriff des *Grenzwerts* für reelle Folgen bzw. der *Grenzfunktion* für Funktionenfolgen geführt.

Definition 5.3 (Grenzwert einer reellen Folge). $x \in \mathbb{R}$ heißt Grenzwert einer reellen Folge $(x_n)_{n=1}^\infty$, wenn es zu jedem $\epsilon > 0$ ein $m \in \mathbb{N}$ gibt, so dass

$$|x_n - x| < \epsilon \text{ für alle } n \geq m. \quad (5.15)$$

Eine Folge, die einen Grenzwert besitzt, wird *konvergente Folge* genannt, eine Folge die keinen Grenzwert besitzt, wird *divergente Folge* genannt. Dafür, dass $x \in \mathbb{R}$ Grenzwert der Folge $(x_n)_{n=1}^\infty$ ist, schreibt man auch

$$\lim_{n \rightarrow \infty} x_n = x \text{ oder } x_n \rightarrow x \text{ für } n \rightarrow \infty \text{ oder } x_n \xrightarrow{n \rightarrow \infty} x. \quad (5.16)$$

•

Nach Definition 5.3 kann der Grenzwert einer Folge existieren, muss es aber nicht. So hat zum Beispiel die Folge

$$f : \mathbb{N} \rightarrow \mathbb{R}, n \mapsto f(n) := n \quad (5.17)$$

keinen Grenzwert, da hier sowohl n als auch $f(n)$ unendlich groß werden. Für eine konvergente Folge, also eine Folge deren Grenzwert existiert, gilt nach Definition 5.3, dass sich für ein beliebig kleines, aber positives, ϵ ein $m \in \mathbb{N}$ angeben lässt, sodass alle Folgenglieder x_n mit $n \geq m$ der Abstand in positiver oder negativer Richtung vom Grenzwert kleiner als ϵ ist. Die Folgenglieder mit $n \geq m$ liegen also beliebig nah am Grenzwert x . Oft mag es so sein, dass für ein kleineres ϵ ein größeres m erforderlich ist.

Beispiele

Die oben betrachteten Beispiele für reelle Folgen dagegen haben Grenzwerte.

- (1) Für die verallgemeinerten harmonischen Folgen gilt mit $p, q \in \mathbb{N}$

$$\lim_{n \rightarrow \infty} \left(\frac{1}{n} \right)^{\frac{p}{q}} = 0. \quad (5.18)$$

- (2) Für die geometrischen Folgen gilt mit $q \in]-1, 1[$

$$\lim_{n \rightarrow \infty} q^n = 0. \quad (5.19)$$

Man nennt die harmonischen und geometrischen Folgen entsprechend auch *Nullfolgen*. Für allgemeine Beweise von Gleichung 5.18 und Gleichung 5.19 verweisen wir auf die weiterführende Literatur. Tatsächlich sind diese Beweise nicht trivial und rühren an den Grundannahmen über das Wesen der reellen Zahlen. Wir visualisieren die ersten zehn Folgenglieder sowie die Grenzwerte der harmonischen Folge für $p := q := 1$ und der geometrischen Folge für $q := 1/2$ in Abbildung 5.1. Die durch Punkte dargestellten Folgenglieder liegen für höheres n offenbar zunehmend näher an dem durch eine graue Linie dargestellten Grenzwert.

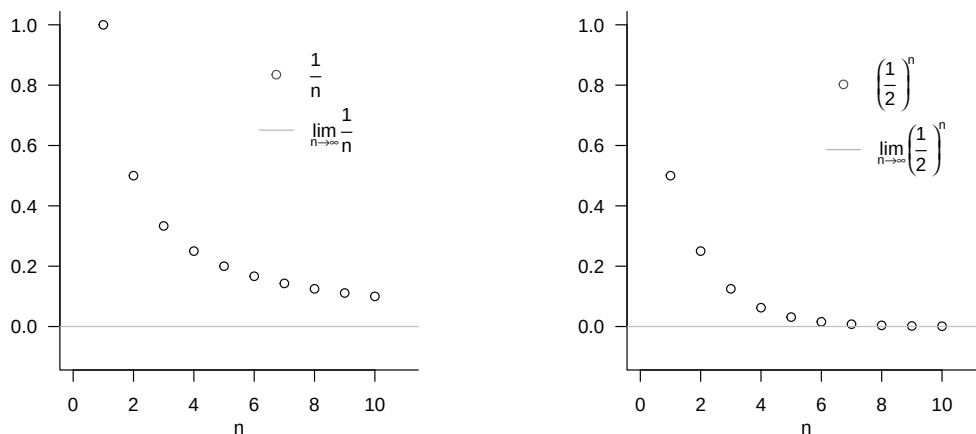


Abbildung 5.1. Beispiele für Grenzwerte reeller Folgen.

Für den Spezialfall der harmonischen Folge mit $p = q = 1$ wollen wir die Argumentation für den Grenzwert

$$\lim_{n \rightarrow \infty} \frac{1}{n} = 0 \quad (5.20)$$

kurz darstellen. Nach Definition 5.3 gilt Gleichung 5.20 genau dann, wenn sich für ein beliebiges $\epsilon > 0$ (also insbesondere ein beliebig kleines) ein $m \in \mathbb{N}$ angeben lässt, so dass für alle $n \geq m$ gilt, dass

$$\left| \frac{1}{n} - 0 \right| < \epsilon \Leftrightarrow \frac{1}{n} < \epsilon.$$

Setzt man also mit der Aufrundungsfunktion $\lceil \cdot \rceil$ für ein beliebiges $\epsilon > 0$

$$m := \left\lceil \frac{1}{\epsilon} \right\rceil + 1,$$

also z.B. bei $\epsilon := 0.01$ entsprechend $m := \lceil \frac{1}{0.01} \rceil + 1 = 101$, so gilt für alle $n \geq m$, dass $\frac{1}{n} < \epsilon$. Da $\epsilon > 0$ beliebig gewählt wurde, ist diese Konstruktion für alle $\epsilon > 0$ möglich.

Für Funktionenfolgen ist eine Möglichkeit der Erweiterung der Begriffe der Konvergenz und des Grenzwertes folgende.

Definition 5.4 (Punktweise Konvergenz und Grenzfunktion einer Funktionenfolge). $F = (f_n)_{n \in \mathbb{N}}$ sei eine Funktionenfolge von univariaten reellwertigen Funktionen mit Definitionsbereich D . F heißt *punktweise konvergent*, wenn die reelle Folge $(f_n(x))_{n \in \mathbb{N}}$ für jedes $x \in D$ eine konvergente Folge ist, also einen Grenzwert besitzt. Die Funktion, die jedem $x \in D$ diesen Grenzwert von $(f_n(x))_{n \in \mathbb{N}}$ zuordnet, heißt dann die *Grenzfunktion der Funktionenfolge* F und hat die Form

$$f : D \rightarrow \mathbb{R}, x \mapsto f(x) := \lim_{n \rightarrow \infty} f_n(x). \quad (5.21)$$

•

Man beachte, dass die Grenzwerte von konvergenten reellen Folgen reelle Zahlen sind, die Grenzfunktionen von punktweise konvergenten Funktionenfolgen dagegen sind Funktionen. Neben der punktweisen Konvergenz von Funktionenfolgen gibt es noch den mächtigeren Begriff der *gleichmäßigen Konvergenz* von Funktionenfolgen, für den wir aber auf die weiterführende Literatur verweisen. Als Beispiel betrachten wir die Grenzfunktionen der oben diskutierten Funktionenfolgen, wobei wir für Beweise auf die weiterführende Literatur verweisen.

Beispiele

- (1) Wir betrachten die Funktionenfolge

$$F : \mathbb{N} \rightarrow \phi, n \mapsto F(n) \quad (5.22)$$

mit

$$\phi := \{f_n | f_n : [0, 1] \rightarrow \mathbb{R}, x \mapsto f_n(x) := x^n \text{ für } n \in \mathbb{N}\} \quad (5.23)$$

Dann ist F punktweise konvergent mit Grenzfunktion

$$f : [0, 1] \rightarrow \mathbb{R}, x \mapsto f(x) := \begin{cases} 0, & \text{für } x \in [0, 1[\\ 1, & \text{für } x = 1 \end{cases} \quad (5.24)$$

da $f_n(x) := x^n$ für $x \in [0, 1[$ eine geometrische Folge und damit eine Nullfolge ist und $f_n(x) := x^n$ für $x = 1$ eine konstante Folge ist, für die alle Folgenglieder den Abstand 0 von 1 haben. Die Funktionenfolge F konvergiert also gegen eine Funktion, die auf dem gesamten Intervall $[0, 1]$ gleich null ist, außer im Punkt 1. Diese Funktion hat offenbar einen Sprung.

- (2) Wir betrachten die Funktionenfolge

$$F : \mathbb{N} \rightarrow \phi, n \mapsto F(n) \quad (5.25)$$

mit

$$\phi := \{f_n | f_n : [-a, a] \rightarrow \mathbb{R}, x \mapsto f_n(x) := \sum_{k=0}^n \frac{x^k}{k!} \text{ für } n \in \mathbb{N}\} \quad (5.26)$$

Dann ist F punktweise konvergent mit Grenzfunktion

$$f : [-a, a] \rightarrow \mathbb{R}, x \mapsto f(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!} =: \exp(x) \quad (5.27)$$

Die Funktionenfolge F konvergiert also gegen die Exponentialfunktion auf $[-a, a]$. Umgekehrt betrachtet ist die Exponentialfunktion gerade durch

$$\exp(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad (5.28)$$

definiert.

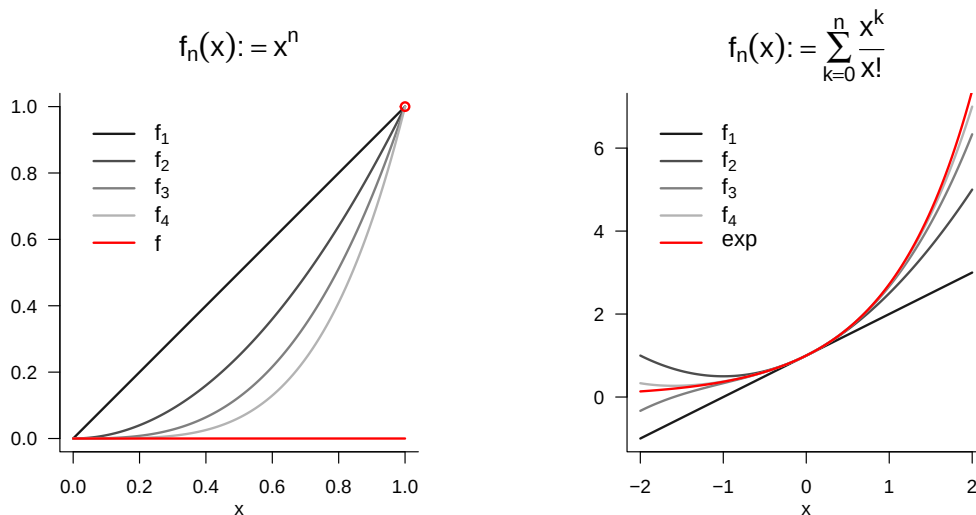


Abbildung 5.2. Beispiele für Grenzwerte von Funktionenfolgen

5.3. Stetigkeit

In diesem Abschnitt versuchen wir uns dem Begriff der *Stetigkeit* einer Funktion zu nähern. Intuitiv ist eine Funktion stetig, wenn sie keine Sprünge hat oder äquivalent, wenn kleine Änderungen in ihren Argumenten stets zu nur kleinen Änderungen in ihren Funktionswerten (und damit eben keinen Sprüngen) führen. Zur Definition der Stetigkeit benötigen wir zunächst den Begriff des *Grenzwertes einer Funktion*.

Definition 5.5 (Grenzwert einer Funktion).

Für $D \subseteq \mathbb{R}$ und $Z \subseteq \mathbb{R}$ sei $f : D \rightarrow Z, x \mapsto f(x)$ eine Funktion und es seien $a, b \in \mathbb{R}$. b heißt *Grenzwert der Funktion f für x gegen a* , wenn

- (1) es eine reelle Folge $(x_n)_{n=1}^{\infty}$ mit Folgengliedern in D mit Grenzwert a gibt, also $\lim_{n \rightarrow \infty} x_n = a$ gilt, und
- (2) für jede solche Folge gilt, dass b der Grenzwert der Folge der Funktionswerte $f(x_n)$ der Folgenglieder von $(x_n)_{n=1}^{\infty}$ ist, also $\lim_{n \rightarrow \infty} f(x_n) = b$ gilt.

Wenn b Grenzwert der Funktion f für x gegen a ist, so schreibt man auch $\lim_{x \rightarrow a} f(x) = b$.

•

In Abbildung 5.3 visualisieren wir den Grenzwert der Exponentialfunktion in $a = 1$ durch Darstellung von Folgengliedern $x_n \rightarrow 1$ als Punkte parallel zur x-Achse und den entsprechenden Folgengliedern $f(x_n)$ auf dem Funktionengraf. Offenbar gilt $\lim_{x \rightarrow 1} \exp(x) = e \approx 2.71$.

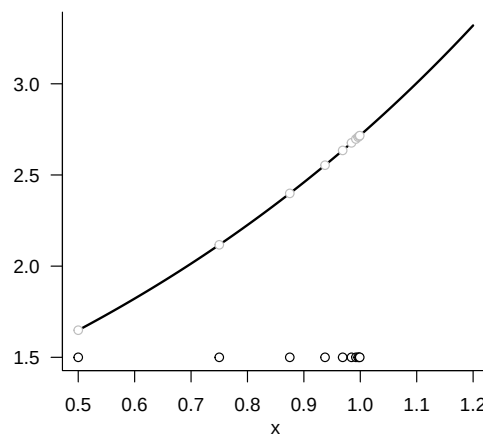


Abbildung 5.3. Beispiel für einen Grenzwert einer Funktion. Die Folge der x -Werte ist hier definiert als $x_n := 1 - (1/2)^n$ für $i = 1, 2, \dots$

Wir können nun den Begriff der Stetigkeit einer Funktion definieren.

Definition 5.6 (Stetigkeit einer Funktion). Eine Funktion $f : D \rightarrow Z$ mit $D \subseteq \mathbb{R}, Z \subseteq \mathbb{R}$ heißt *stetig in* $a \in D$, wenn

$$\lim_{x \rightarrow a} f(x) = f(a). \quad (5.29)$$

Ist f in jedem $x \in D$ stetig, so heißt f *stetig auf* D .

•

Man beachte, dass für eine in a stetige Funktion folgt, dass

$$\lim_{x \rightarrow a} f(x) = f\left(\lim_{x \rightarrow a} x\right) \quad (5.30)$$

Bei stetigen Funktion können also Grenzwertbildung und Auswertung der Funktion vertauscht werden.

6. Differenzialrechnung

Die Differenzialrechnung befasst sich mit der Änderung von Funktionen. Sie bildet einerseits die Grundlage für die mathematische Modellierung dynamischer System mithilfe von *Differenzialgleichungen*, also der Beschreibung von Funktionen anhand ihrer Änderungsraten. Zum anderen bildet die Differenzialrechnung die Grundlage der *Optimierung*, also des Bestimmens von Extremstellen von Funktionen. In Kapitel 6.1 führen wir zunächst den Begriff der Ableitung und mit ihm verbundene elementare Rechenregeln ein. In Kapitel 6.2 widmen wir uns dann der Frage, wie man mithilfe von Ableitungen Extremstellen von Funktionen bestimmen kann.

6.1. Definitionen und Rechenregeln

Wir beginnen mit folgender Definition.

Definition 6.1 (Differenzierbarkeit und Ableitung). Es sei $I \subseteq \mathbb{R}$ ein Intervall und

$$f : I \rightarrow \mathbb{R}, x \mapsto f(x) \quad (6.1)$$

eine univariate reellwertige Funktion. f heißt in $a \in I$ *differenzierbar*, wenn der Grenzwert

$$f'(a) := \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \quad (6.2)$$

existiert. $f'(a)$ heißt dann die *Ableitung von f an der Stelle a* . Ist f differenzierbar für alle $x \in I$, so heißt f *differenzierbar* und die Funktion

$$f' : I \rightarrow \mathbb{R}, x \mapsto f'(x) \quad (6.3)$$

heißt *Ableitung von f* .

•

Für $h > 0$ heißt der Ausdruck

$$\frac{f(a+h) - f(a)}{h} \quad (6.4)$$

Newtonscher Differenzquotient. Wie in Abbildung 6.1 dargestellt, misst der Newtonsche Differenzquotient die Änderung $f(a+h) - f(a)$ von f auf der y -Achse pro Strecke h auf der x -Achse. Wenn also zum Beispiel $f(a)$ und $f(a+h)$ die Position eines Objektes zu einem Zeitpunkt a und zu einem späteren Zeitpunkt $a+h$ repräsentieren, dann ist $f(a+h) - f(a)$ die von diesem Objekt in der Zeit h zurückgelegte Strecke, also seine durchschnittliche Geschwindigkeit über den Zeitraum h . Für $h \rightarrow 0$ misst der Newtonsche Differenzquotient dann die instantane Änderungsrate von f in a , also in diesem Beispiel die Geschwindigkeit des Objektes zu einem Zeitpunkt a .

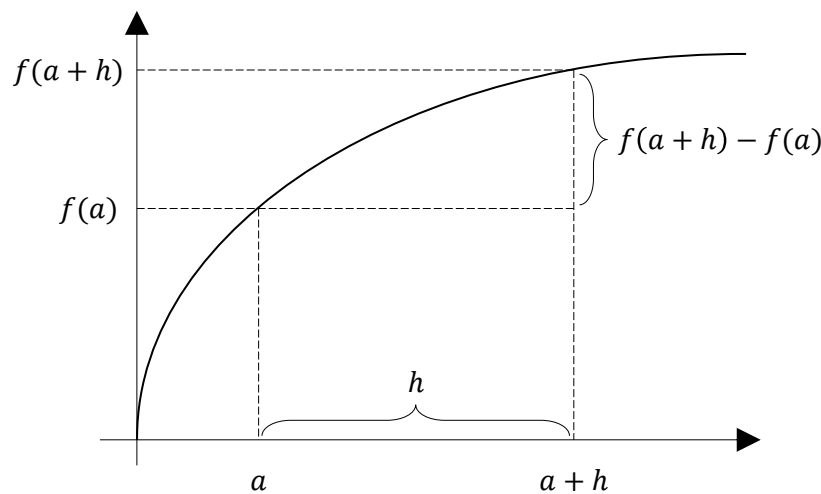


Abbildung 6.1. Elemente des Newtonschen Differenzquotienten.

Aus mathematischer Sicht ist es wichtig, bei der Definition der Ableitung zwischen den Symbolen $f'(a)$ und f' zu unterscheiden. Wie üblich bezeichnet $f'(a)$ den Wert einer Funktion, also eine Zahl. f' dagegen bezeichnet eine Funktion, nämlich die Funktion, deren Werte als $f'(a)$ für alle $a \in \mathbb{R}$ bestimmt sind.

Es existieren in der Literatur verschiedene, historisch gewachsene Notationen für Ableitungen, welche alle das identische Konzept der Ableitung repräsentieren.

Definition 6.2 (Notation für Ableitungen univariater reellwertiger Funktionen). Es sei f eine univariate reellwertige Funktion. Äquivalente Schreibweisen für die Ableitung von f und die Ableitung von f an einer Stelle x sind

- die *Lagrange-Notation* f' und $f'(x)$,
- die *Leibniz-Notation* $\frac{df}{dx}$ und $\frac{df(x)}{dx}$,
- die *Newton-Notation* $\dot{f}$ und $\dot{f}(x)$,
- die *Euler-Notation* Df und $Df(x)$.

•

Wir werden im Folgenden für univariate reellwertige Funktionen vor allem die Lagrange-Notation f' und $f'(x)$ als Bezeichner wählen. In Berechnungen nutzen wir auch eine adaptierte Form der Leibniz-Notation und verstehen dort die Schreibweise $\frac{d}{dx}f(x)$ als den Auftrag, die Ableitung von f zu berechnen. Die Newton-Notation wird vor allem eingesetzt, wenn das Funktionsargument die Zeit repräsentiert und das dann üblicherweise mit t für *time* bezeichnet wird. $\dot{f}(t)$ bezeichnet dann entsprechend die Änderungsrate von f zum Zeitpunkt t . Die Euler-Notation ist vor allem im Kontext multivariater reell- oder vektorwertiger Funktionen nützlich.

Wir setzen eine Reihe von Ableitungen elementarer Funktionen als bekannt voraus, diese sind in Tabelle 6.1 zusammengestellt. Für Beweise verweisen wir auf die weiterführende Literatur.

Tabelle 6.1. Ableitungen elementarer Funktionen

Name	Definition	Ableitung
Polynomfunktion	$f(x) := \sum_{i=0}^k a_i x^i$	$f'(x) = \sum_{i=1}^k i a_i x^{i-1}$
Konstante Funktion	$f(x) := a$	$f'(x) = 0$
Identitätsfunktion	$f(x) := x$	$f'(x) = 1$
Linear-affine Funktion	$f(x) := ax + b$	$f'(x) = a$
Quadratfunktion	$f(x) := x^2$	$f'(x) = 2x$
Exponentialfunktion	$f(x) := \exp(x)$	$f'(x) = \exp(x)$
Logarithmusfunktion	$f(x) := \ln(x)$	$f'(x) = \frac{1}{x}$

In Abbildung 6.2 visualisieren wir die Identitätsfunktion, eine lineare Funktion und die Quadratfunktion zusammen mit ihrer jeweiligen Ableitung. In Abbildung Abbildung 6.3 visualisieren wir die Exponential- und Logarithmusfunktionen zusammen mit ihrer jeweiligen Ableitung.

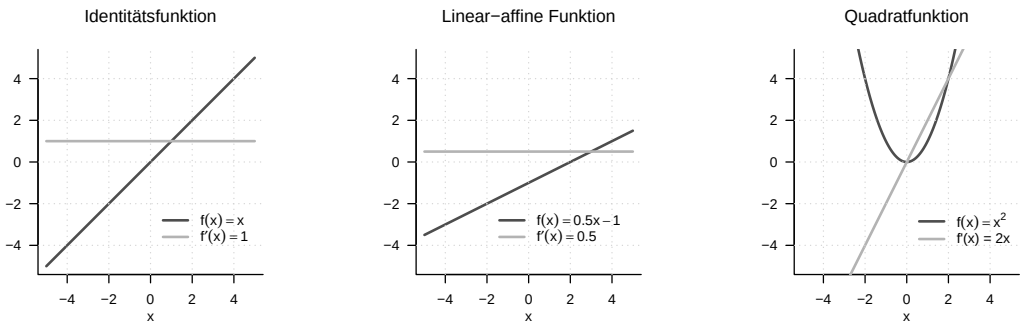


Abbildung 6.2. Ableitungen dreier elementarer Funktionen. Die Ableitung der Identitätsfunktion ist die konstante Funktion mit Wert 1, die Ableitung der betrachteten linear-affinen Funktion ist die konstante Funktion mit Wert 0.5 und die Ableitung der Quadratfunktion ist die linear-affine Funktion mit $a = 2$ und $b = 0$

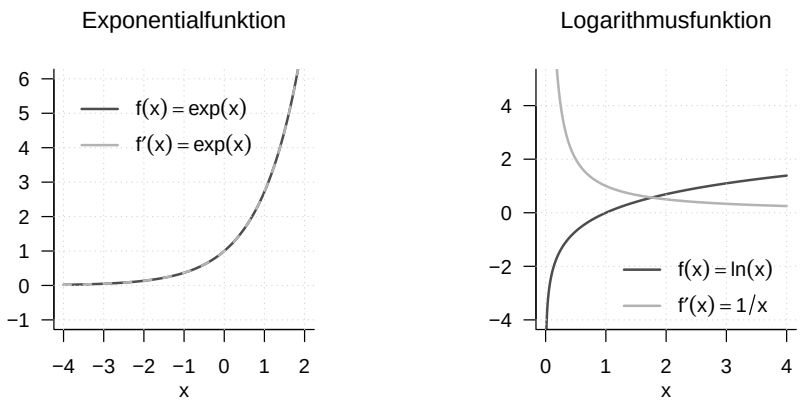


Abbildung 6.3. Ableitungen von Exponentialfunktion und Logarithmusfunktion. Die Ableitung der Exponentialfunktion ist die Exponentialfunktion selbst, die Ableitung der Logarithmusfunktion hat die Funktionswerte $1/x$

Basierend auf der Definition der Ableitung einer univariaten reellwertigen Funktion lassen sich leicht weitere Ableitungen einer solchen Funktion definieren.

Definition 6.3 (Höhere Ableitungen). Es sei f eine univariate reellwertige Funktion und

$$f^{(1)} := f' \quad (6.5)$$

sei die Ableitung von f . Die k -te Ableitung von f ist rekursiv definiert durch

$$f^{(k)} := (f^{(k-1)})' \quad \text{für } k > 1, \quad (6.6)$$

unter der Annahme, dass $f^{(k-1)}$ differenzierbar ist. Insbesondere ist die *zweite Ableitung* von f definiert durch die Ableitung von f' , also

$$f'' := (f')'. \quad (6.7)$$

•

In Analogie zu oben Gesagtem schreiben wir in Berechnungen auch $\frac{d^2}{dx^2} f(x)$ für den Auftrag, die zweite Ableitung einer Funktion f zu bestimmen. Die nullte Ableitung $f^{(0)}$ von f ist f selbst. Der Tradition und Einfachheit halber schreibt man für $k < 4$ gemäß der Lagrange-Notation meist f' , f'' und f''' anstelle von $f^{(1)}$, $f^{(2)}$ und $f^{(3)}$.

Beispiel

Es sei

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := x^2 \quad (6.8)$$

Dann gilt

$$f^{(1)}(x) = f'(x) = \frac{d}{dx}(x^2) = 2x \quad (6.9)$$

Weiterhin gelten

$$\begin{aligned} f^{(2)}(x) &= (f^{(2-1)})'(x) = (f^{(1)})'(x) = \frac{d}{dx}(2x) = 2, \\ f^{(3)}(x) &= (f^{(3-1)})'(x) = (f^{(2)})'(x) = \frac{d}{dx}(2) = 0, \\ f^{(4)}(x) &= (f^{(4-1)})'(x) = (f^{(3)})'(x) = \frac{d}{dx}(0) = 0. \end{aligned} \quad (6.10)$$

Man berechnet also durchgängig lediglich erste Ableitungen.

Zum Bestimmen der Ableitung einer Funktion sind eine Reihe von Rechenregeln hilfreich, die es erlauben, die Ableitung einer Funktion aus den Ableitungen ihrer Unterfunktionen herzuleiten. Für Beweise der in folgendem Theorem eingeführten Rechenregeln verweisen wir auf die weiterführende Literatur.

Theorem 6.1 (Rechenregeln für Ableitungen). Für $i = 1, \dots, n$ seien g_i reellwertige univariate differenzierbare Funktionen. Dann gelten folgende Rechenregeln:

(1) *Summenregel*

$$\text{Für } f(x) := \sum_{i=1}^n g_i(x) \text{ gilt } f'(x) = \sum_{i=1}^n g'_i(x). \quad (6.11)$$

(2) *Produktregel*

$$\text{Für } f(x) := g_1(x)g_2(x) \text{ gilt } f'(x) = g'_1(x)g_2(x) + g_1(x)g'_2(x). \quad (6.12)$$

(3) Quotientenregel

$$\text{Für } f(x) := \frac{g_1(x)}{g_2(x)} \text{ gilt } f'(x) = \frac{g_1'(x)g_2(x) - g_1(x)g_2'(x)}{g_2^2(x)}. \quad (6.13)$$

(4) Kettenregel

$$\text{Für } f(x) := g_1(g_2(x)) \text{ gilt } f'(x) = g_1'(g_2(x))g_2'(x). \quad (6.14)$$

◦

Beispiele

(1) Summenregel

Es sei

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := 4x^3 + 3x^2 \quad (6.15)$$

Dann hat f die Form

$$f(x) = \sum_{i=1}^2 g_i(x) = g_1(x) + g_2(x) \text{ mit } g_1(x) := 4x^3 \text{ und } g_2(x) := 3x^2. \quad (6.16)$$

Es gelten außerdem

$$g_1'(x) = \frac{d}{dx}(4x^3) = 12x^2 \text{ und } g_2'(x) = \frac{d}{dx}(3x^2) = 6x. \quad (6.17)$$

Also gilt mit der Summenregel

$$f'(x) = \sum_{i=1}^2 g_i'(x) = g_1'(x) + g_2'(x) = 12x^2 + 6x. \quad (6.18)$$

(2) Produktregel

Es sei

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := x^2 \sin(x). \quad (6.19)$$

Dann hat f die Form

$$f(x) = g_1(x)g_2(x) \text{ mit } g_1(x) := x^2 \text{ und } g_2(x) := \sin(x). \quad (6.20)$$

Es gelten außerdem

$$g_1'(x) = \frac{d}{dx}(x^2) = 2x \text{ und } g_2'(x) = \frac{d}{dx}(\sin x) = \cos(x). \quad (6.21)$$

Also gilt mit der Produktregel

$$f'(x) = g_1'(x)g_2(x) + g_1(x)g_2'(x) = 2x \sin(x) + x^2 \cos(x). \quad (6.22)$$

(3) Quotientenregel

Es sei

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := \frac{x^2}{x+1}. \quad (6.23)$$

Dann hat f die Form

$$f(x) = \frac{g_1(x)}{g_2(x)} \text{ mit } g_1(x) := x^2 \text{ und } g_2(x) := x + 1. \quad (6.24)$$

s gelten außerdem

$$g_1'(x) = \frac{d}{dx}(x^2) = 2x \text{ und } g_2'(x) = \frac{d}{dx}(1 + x) = 1. \quad (6.25)$$

Also gilt mit der Quotientenregel

$$f'(x) = \frac{g_1'(x)g_2(x) - g_1(x)g_2'(x)}{g_2^2(x)} = \frac{2x(x+1) - x^2}{(x+1)^2} = \frac{2x^2 + 2x - x^2}{(x+1)^2} = \frac{x^2 + 2x}{(x+1)^2}. \quad (6.26)$$

(4) Kettenregel

Es sei

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := \exp(-x^2) \quad (6.27)$$

Dann hat f die Form

$$f(x) = g_1(g_2(x)) \text{ mit } g_1(x) := \exp(x) \text{ und } g_2(x) := -x^2 \quad (6.28)$$

Es gelten außerdem

$$g_1'(x) = \frac{d}{dx}(\exp(x)) = \exp(x) \text{ und } g_2'(x) = \frac{d}{dx}(-x^2) = -2x. \quad (6.29)$$

Also gilt mit der Kettenregel

$$f'(x) = g_1'(g_2(x))g_2'(x) = \exp(-x^2)(-2x) = -2x \exp(-x^2). \quad (6.30)$$

Man kann sich die resultierende Ableitung der Kettenregel also merken als “Die Ableitung der äußeren Funktion an der Stelle der inneren Funktion mal die Ableitung der inneren Funktion.”

6.2. Analytische Optimierung

Eine wichtige Anwendung der Differenzialrechnung ist das Bestimmen von Extremstellen von Funktionen. Dabei geht es im Kern um die Frage, für welche Werte ihrer Definitionsmenge eine Funktion ein Maximum oder ein Minimum annimmt. Bei einfachen Funktionen ist dies analytisch möglich. Die generelle Vorgehensweise dabei ist oft auch unter dem Stichwort “Kurvendiskussion” bekannt. In der Anwendung ist ein analytisches Vorgehen zur Optimierung von Funktionen meist nicht möglich und es werden numerische Algorithmen zur Bestimmung von Extremstellen genutzt. Ein Verständnis dieser Algorithmen setzt allerdings ein Verständnis der Prinzipien der analytischen Optimierung voraus. In diesem Abschnitt geben wir eine Einführung in die analytische Optimierung von univariaten reellwertigen Funktionen. Wir beginnen damit, die Begriffe der erwähnten Maxima und Minima von univariaten reellwertigen Funktionen zu präzisieren.

Definition 6.4 (Extremstellen und Extremwerte). Es seien $U \subseteq \mathbb{R}$ und $f : U \rightarrow \mathbb{R}$ eine univariate reellwertige Funktion. f hat an der Stelle $x_0 \in U$

- ein *lokales Minimum*, wenn es ein Intervall $I :=]a, b[$ gibt mit $x_0 \in]a, b[$ und

$$f(x_0) \leq f(x) \text{ für alle } x \in I \cap U, \quad (6.31)$$

- ein *globales Minimum*, wenn gilt, dass

$$f(x_0) \leq f(x) \text{ für alle } x \in U, \quad (6.32)$$

- ein *lokales Maximum*, wenn es ein Intervall $I :=]a, b[$ gibt mit $x_0 \in]a, b[$ und

$$f(x_0) \geq f(x) \text{ für alle } x \in I \cap U, \quad (6.33)$$

- ein *lokales Maximum*, wenn gilt, dass

$$f(x_0) \geq f(x) \text{ für alle } x \in U. \quad (6.34)$$

Der Wert $x_0 \in U$ der Definitionsmenge von f heißt entsprechend *lokale* oder *globale Minimalstelle* oder *Maximalstelle*, der Funktionswert $f(x_0) \in \mathbb{R}$ heißt entsprechend *lokales* oder *globales Minimum* oder *Maximum*. Generell heißt der Wert $x_0 \in U$ *Extremstelle* und der Funktionswert $f(x_0) \in \mathbb{R}$ *Extremwert*.

•

Extremstellen von Funktionen werden häufig mit

$$\underset{x \in I \cap U}{\operatorname{argmin}} f(x) \text{ oder } \underset{x \in I \cap U}{\operatorname{argmax}} f(x) \quad (6.35)$$

bezeichnet und Extremwerte von Funktionen werden häufig mit

$$\min_{x \in I \cap U} f(x) \text{ oder } \max_{x \in I \cap U} f(x) \quad (6.36)$$

bezeichnet. Abbildung 6.4 visualisiert die verschiedenen Arten von Extremwerten nach Definition 6.4.

Die analytische Optimierung von univariaten reellwertigen Funktionen basiert auf den sogenannten *notwendigen* und *hinreichenden Bedingungen für Extrema*. Erstere macht eine Aussage über das Verhalten der ersten Ableitung einer Funktion an einer Extremstelle, letztere macht eine Aussage über das Verhalten einer Funktion an einer Stelle, die bestimmten Forderungen an ihre erste und zweite Ableitung genügt.

Theorem 6.2 (Notwendige Bedingung für Extrema). *f sei eine univariate reellwertige Funktion. Dann gilt*

$$x_0 \text{ ist Extremstelle von } f \Rightarrow f'(x_0) = 0. \quad (6.37)$$

◦

Wenn x_0 eine Extremstelle von f ist, dann ist also die erste Ableitung von f in x_0 gleich null. Anstelle eines Beweises überlegen wir uns, dass zum Beispiel an einer lokalen Maximalstelle x_0 von f gilt: links von x_0 steigt f an, rechts von x_0 fällt f ab. In x_0 aber steigt f weder an noch fällt f ab, es ist also nachvollziehbar, dass $f'(x_0) = 0$ ist.

Theorem 6.3 (Hinreichende Bedingungen für lokale Extrema). *f sei eine zweimal differenzierbare univariate reellwertige Funktion.*

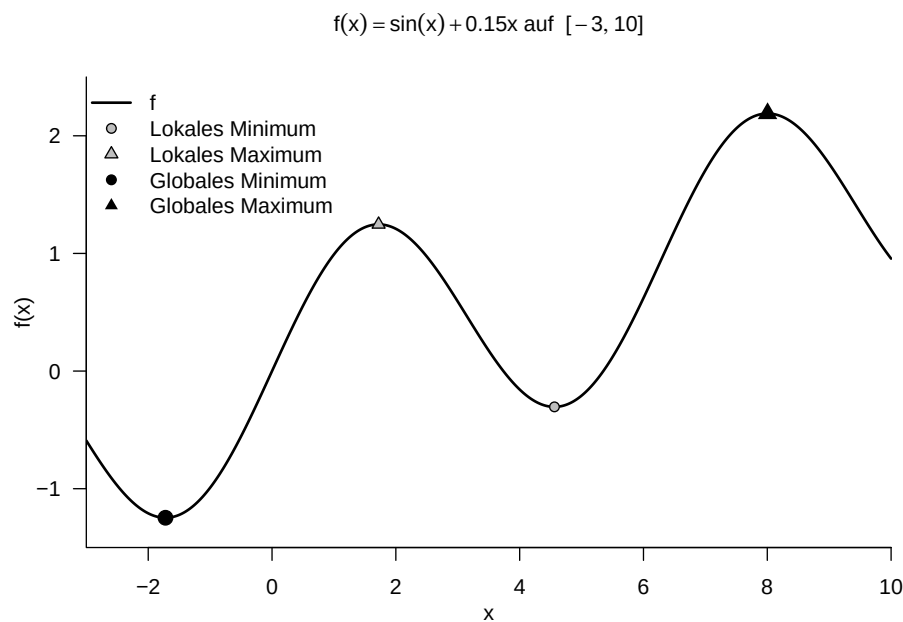


Abbildung 6.4. Extremstellen einer Funktion.

- Wenn für $x_0 \in U \subseteq \mathbb{R}$

$$f'(x_0) = 0 \text{ und } f''(x_0) > 0 \quad (6.38)$$

gilt, dann hat f an der Stelle x_0 ein Minimum.

- Wenn für $x_0 \in U \subseteq \mathbb{R}$

$$f'(x_0) = 0 \text{ und } f''(x_0) < 0 \quad (6.39)$$

gilt, dann hat f an der Stelle x_0 ein Maximum.

◦

Wir verzichten wiederum auf einen Beweis und verdeutlichen uns die Bedingung an dem in Abbildung 6.5 gezeigten Beispiel. Hier ist offenbar $x_0 = 1$ eine lokale Minimalstelle von $f(x) = (x-1)^2$. Man erkennt: links von x_0 fällt f ab, rechts von x_0 steigt f an. In x_0 steigt f weder an, noch fällt f ab, also ist $f'(x_0) = 0$. Weiter gilt, dass links und rechts von x_0 und in x_0 die Änderung f'' von f' positiv ist: links von x_0 schwächt sich die Negativität von f' zu 0 ab und rechts von x_0 verstärkt sich die Positivität von f' .

Insbesondere die hinreichenden Bedingungen für lokale Extrema legen folgendes *Standardverfahren* zur Bestimmung von lokalen Extremstellen nahe.

Theorem 6.4 (Standardverfahren der analytischen Optimierung). *f sei eine univariate reellwertige Funktion. Lokale Extremstellen von f können mit folgendem Standardverfahren der analytischen Optimierung identifiziert werden:*

1. Berechnen der ersten und zweiten Ableitung von f .
2. Bestimmen von Nullstellen x^* von f' durch Auflösen von $f'(x^*) = 0$ nach x^* . Die Nullstellen von f' sind dann Kandidaten für Extremstellen von f .

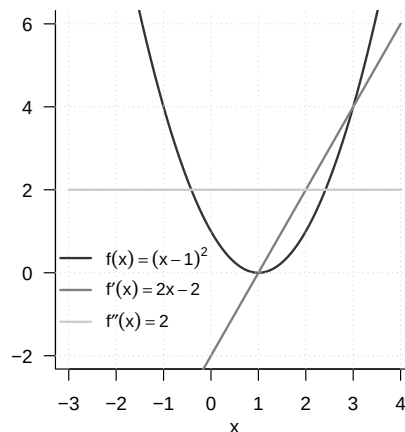


Abbildung 6.5. Analytische Optimierung von $f(x) := (x - 1)^2$

3. *Evaluation von $f''(x^*)$: Wenn $f''(x^*) > 0$ ist, dann ist x^* lokale Minimumstelle von f ; wenn $f''(x^*) < 0$ ist, dann ist x^* lokale Maximumstelle von f ; wenn $f''(x^*) = 0$ ist, dann ist x^* keine Extremstelle von f .*

◦

Beispiel

Anstelle eines Beweises betrachten wir beispielhaft die Funktion

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := (x - 1)^2. \quad (6.40)$$

aus Abbildung 6.5. Die erste Ableitung von f ergibt sich mit der Kettenregel zu

$$f'(x) = \frac{d}{dx} ((x - 1)^2) = 2(x - 1) \cdot \frac{d}{dx} (x - 1) = 2x - 2. \quad (6.41)$$

Die zweite Ableitung von f ergibt sich zu

$$f''(x) = \frac{d}{dx} f'(x) = \frac{d}{dx} (2x - 2) = 2 > 0 \text{ für alle } x \in \mathbb{R}. \quad (6.42)$$

Auflösen von $f'(x^*) = 0$ nach x^* ergibt

$$f'(x^*) = 0 \Leftrightarrow 2x^* - 2 = 0 \Leftrightarrow 2x^* = 2 \Leftrightarrow x^* = 1. \quad (6.43)$$

$x^* = 1$ ist folglich eine Minimalstelle von f mit zugehörigen Minimalwert $f(1) = 0$.

6.3. Differenzialrechnung multivariater reellwertiger Funktionen

Wir erinnern zunächst an den Begriff der multivariaten reellwertigen Funktion.

Definition 6.5 (Multivariate reellwertige Funktion). Eine Funktion der Form

$$f : \mathbb{R}^n \rightarrow \mathbb{R}, x \mapsto f(x) = f(x_1, \dots, x_n) \quad (6.44)$$

heißt *multivariate reellwertige Funktion*.

•

Die Argumente multivariater reellwertiger Funktionen sind also reelle n -Tupel der Form $x := (x_1, \dots, x_n)$, während ihre Funktionswerte reelle Zahlen sind. Ein Beispiel für eine multivariate reellwertige Funktion mit $n := 2$ ist

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \quad (6.45)$$

Wir visualisieren diese Funktion in Abbildung 6.6. Dabei zeigt die rechte Abbildung eine Darstellung mithilfe sogenannter *Isokonturen*, also Linien im Definitionsbereich der Funktion, für die die Funktion identische Werte annimmt. Die entsprechenden Werte sind für ausgewählte Isokonturen in der Abbildung vermerkt.

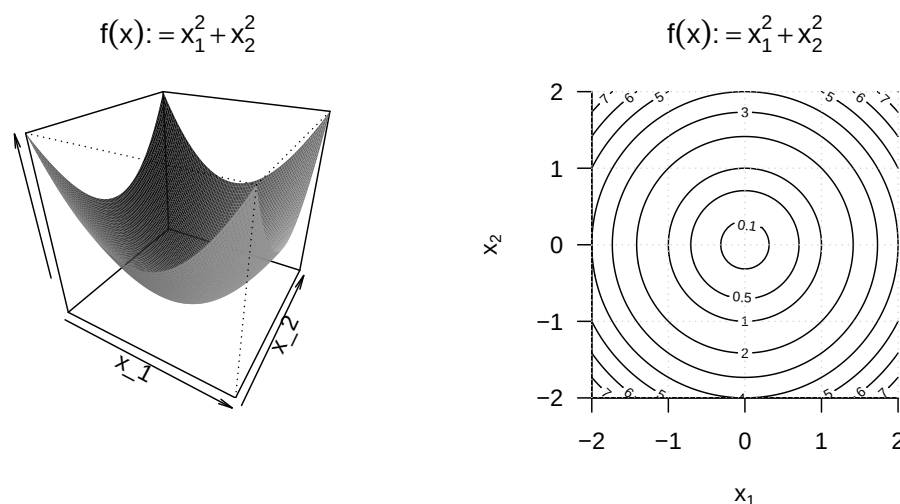


Abbildung 6.6. Visualisierungen einer bivariaten Funktion. Die linke Darstellung visualisiert die Funktionswerte als Fläche über der zweidimensionalen Definitionsmenge der Funktion. Auch wenn diese Darstellung einen gewissen dreidimensionalen Charakter hat, handelt es sich um eine bivariate Funktion mit entsprechend zweidimensionaler Definitionsmenge. Die rechte Abbildung visualisiert Linien mit gleichen Funktionswerten in der zweidimensionalen Definitionsmenge der Funktion.

Wir wollen nun beginnen, die Begriffe der Differenzierbarkeit und der Ableitung univariater reellwertiger Funktionen auf den Fall multivariater reellwertiger Funktionen zu erweitern. Dazu führen wir zunächst die Begriffe der *partiellen Differenzierbarkeit* und der *partiellen Ableitung* ein.

Definition 6.6 (Partielle Differenzierbarkeit und partielle Ableitung). Es sei $D \subseteq \mathbb{R}^n$ eine Menge und

$$f : D \rightarrow \mathbb{R}, x \mapsto f(x) \quad (6.46)$$

eine multivariate reellwertige Funktion. f heißt in $a \in D$ nach x_i *partiell differenzierbar*, wenn der Grenzwert

$$\frac{\partial}{\partial x_i} f(x) := \lim_{h \rightarrow 0} \frac{f(a + he_i) - f(a)}{h} \quad (6.47)$$

existiert. $\frac{\partial}{\partial x_i} f(a)$ heißt dann die *partielle Ableitung von f nach x_i an der Stelle a* . Wenn f für alle $x \in D$, nach x_i partiell differenzierbar ist, dann heißt f nach x_i *partiell differenzierbar* und die Funktion

$$\frac{\partial}{\partial x_i} f : D \rightarrow \mathbb{R}, x \mapsto \frac{\partial}{\partial x_i} f(x) \quad (6.48)$$

heißt *partielle Ableitung von f nach x_i* . f heißt *partiell differenzierbar in $x \in D$* , wenn für alle $i = 1, \dots, n$ in $x \in D$ nach x_i partiell differenzierbar ist, und

f heißt *partiell differenzierbar*, wenn f für alle $i = 1, \dots, n$ in allen $x \in D$ nach x_i partiell differenzierbar ist.

•

In Definition 6.6 bezeichnet $e_i \in \mathbb{R}^n$ den i ten kanonischen Einheitsvektor, für den gilt, dass $e_{ij} = 1$ für $i = j$ und $e_{ij} = 0$ für $i \neq j$ mit $j = 1, \dots, n$ (vgl. Definition 8.14). In Analogie und Verallgemeinerung zum Newtonschen Differenzquotienten misst der hier auftretende Differenzquotient

$$\frac{f(x + he_i) - f(x)}{h} \quad (6.49)$$

damit die Änderung $f(x + he_i) - f(x)$ von f pro Strecke h in Richtung e_i . Wir visualisieren die Bestandteile dieses Quotienten im Falle einer bivariaten Funktion in Abbildung 6.7.

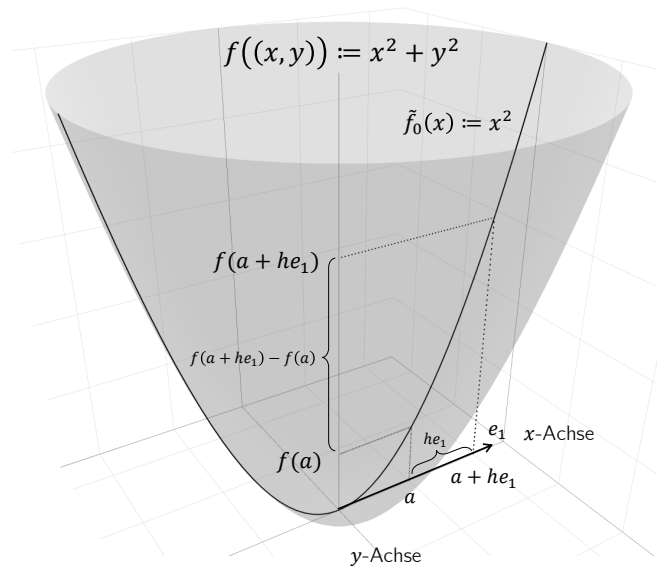


Abbildung 6.7. Komponenten des partiellen Newtonschen Differenzquotienten

Für $h \rightarrow 0$ misst der Differenzquotient entsprechend die *Änderungsrate* von f in x in Richtung e_i . Wie bei der Betrachtung von Ableitungen gilt, dass $\frac{\partial}{\partial x_i} f(x)$ eine Zahl, $\frac{\partial}{\partial x_i} f$ dagegen eine Funktion ist. Praktisch berechnet man $\frac{\partial}{\partial x_i} f$ als die (einfache) Ableitung

$$\frac{d}{dx_i} \tilde{f}_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n}(x_i) \quad (6.50)$$

der univariaten reellwertigen Funktion

$$\tilde{f} : \mathbb{R} \rightarrow \mathbb{R}, x_i \mapsto \tilde{f}_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n}(x_i) := f(x_1, \dots, x_i, \dots, x_n). \quad (6.51)$$

Man betrachtet für die i te partielle Ableitung also alle x_j mit $j \neq i$ als Konstanten und ist auf das gewohnte Berechnen von Ableitungen von univariaten reellwertigen Funktionen geführt. Wir wollen das Vorgehen zum Berechnen von partiellen Ableitungen an einem ersten Beispiel verdeutlichen.

Beispiel (1)

Wir betrachten die Funktion

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \quad (6.52)$$

Weil die Definitionsmenge dieser Funktion zweidimensional ist, kann man zwei partielle Ableitungen berechnen

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \rightarrow \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) \text{ und } \frac{\partial}{\partial x_2} f : \mathbb{R}^2 \rightarrow \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x). \quad (6.53)$$

Um die erste dieser partiellen Ableitungen zu berechnen, betrachtet man die Funktion

$$f_{x_2} : \mathbb{R} \rightarrow \mathbb{R}, x_1 \mapsto f_{x_2}(x_1) := x_1^2 + x_2^2, \quad (6.54)$$

wobei x_2 hier die Rolle einer Konstanten einnimmt. Um explizit zu machen, dass x_2 kein Argument der Funktion ist, die Funktion aber weiterhin von x_2 abhängt haben wir die Subskriptnotation $f_{x_2}(x_1)$ verwendet. Um nun die partielle Ableitung zu berechnen, berechnen wir die (einfache) Ableitung von f_{x_2} ,

$$f'_{x_2}(x) = 2x_1. \quad (6.55)$$

Es ergibt sich also

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \rightarrow \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) = \frac{\partial}{\partial x_1} (x_1^2 + x_2^2) = f'_{x_2}(x) = 2x_1. \quad (6.56)$$

Analog gilt mit der entsprechenden Formulierung von f_{x_1} , dass

$$\frac{\partial}{\partial x_2} f : \mathbb{R}^2 \rightarrow \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x) = \frac{\partial}{\partial x_2} (x_1^2 + x_2^2) = f'_{x_1}(x) = 2x_2. \quad (6.57)$$

Wie bei der Ableitung einer univariaten reellwertigen Funktion ist es auch für eine multivariate reellwertige Funktion möglich, rekursiv eine höhere Ableitung zu definieren.

Definition 6.7 (Zweite partielle Ableitungen). $f : \mathbb{R}^n \rightarrow \mathbb{R}$ sei eine multivariate reellwertige Funktion und $\frac{\partial}{\partial x_i} f$ sei die partielle Ableitung von f nach x_i . Dann ist die zweite partielle Ableitung von f nach x_i und x_j definiert als

$$\frac{\partial^2}{\partial x_j \partial x_i} f(x) := \frac{\partial}{\partial x_j} \left(\frac{\partial}{\partial x_i} f \right). \quad (6.58)$$

•

Man beachte, dass es zu jeder partiellen Ableitung $\frac{\partial}{\partial x_i} f$ für $i = 1, \dots, n$ insgesamt n zweite partielle Ableitungen $\frac{\partial^2}{\partial x_j \partial x_i} f$ für $j = 1, \dots, n$ gibt. Die so resultierenden n^2 zweiten partiellen Ableitungen sind jedoch nicht alle verschieden. Dies ist eine wesentliche Aussage des *Satzes von Schwarz*

Theorem 6.5 (Satz von Schwarz). $f : \mathbb{R}^n \rightarrow \mathbb{R}$ sei eine partiell differenzierbare multivariate reellwertige Funktion. Dann gilt

$$\frac{\partial^2}{\partial x_j \partial x_i} f(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x) \text{ für alle } 1 \leq i, j \leq n. \quad (6.59)$$

◦

Für einen Beweis verweisen wir auf die weiterführende Literatur. Der Satz von Schwarz besagt insbesondere auch, dass bei Bildung der zweiten partiellen Ableitungen die Reihenfolge des partiellen Ableitens irrelevant ist. Der Satz erleichtert auf diese Weise die Berechnung von zweiten partiellen Ableitungen und hilft darüberhinaus, analytische Fehler bei der Berechnung ihrer zweiten partiellen Ableitungen aufzudecken. Wir verdeutlichen dies in Fortführung obigen Beispiels.

Beispiel (1)

Wir wollen die partiellen Ableitungen zweiter Ordnung der Funktion

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2 \quad (6.60)$$

berechnen. Mit den Ergebnissen für die partiellen Ableitungen erster Ordnung dieser Funktion ergibt sich

$$\begin{aligned} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_1} (2x_1) = 2 \\ \frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_1} (2x_2) = 0 \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_2} (2x_1) = 0 \\ \frac{\partial^2}{\partial x_2 \partial x_2} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_2} (2x_2) = 2 \end{aligned} \quad (6.61)$$

Offenbar gilt

$$\frac{\partial^2}{\partial x_1 x_2} f(x) = \frac{\partial^2}{\partial x_2 x_1} f(x). \quad (6.62)$$

Beispiel (2)

Als weiteres Beispiel wollen wir die partiellen Ableitungen erster und zweiter Ordnung der Funktion

$$f : \mathbb{R}^3 \rightarrow \mathbb{R}, x \mapsto f(x) := x_1^2 + x_1 x_2 + x_2 \sqrt{x_3} \quad (6.63)$$

berechnen. Mit den Rechenregeln für Ableitungen ergibt sich für die partiellen Ableitungen erster Ordnung

$$\begin{aligned}\frac{\partial}{\partial x_1} f(x) &= \frac{\partial}{\partial x_1} (x_1^2 + x_1 x_2 + x_2 \sqrt{x_3}) = 2x_1 + x_2, \\ \frac{\partial}{\partial x_2} f(x) &= \frac{\partial}{\partial x_2} (x_1^2 + x_1 x_2 + x_2 \sqrt{x_3}) = x_1 + \sqrt{x_3}, \\ \frac{\partial}{\partial x_3} f(x) &= \frac{\partial}{\partial x_3} (x_1^2 + x_1 x_2 + x_2 \sqrt{x_3}) = \frac{x_2}{2\sqrt{x_3}}.\end{aligned}\tag{6.64}$$

Für die zweiten partiellen Ableitungen hinsichtlich x_1 ergibt sich

$$\begin{aligned}\frac{\partial^2}{\partial x_1 \partial x_1} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_1} (2x_1 + x_2) = 2, \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_2} (2x_1 + x_2) = 1, \\ \frac{\partial^2}{\partial x_3 \partial x_1} f(x) &= \frac{\partial}{\partial x_3} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_3} (2x_1 + x_2) = 0.\end{aligned}\tag{6.65}$$

Für die zweiten partiellen Ableitungen hinsichtlich x_2 ergibt sich

$$\begin{aligned}\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_1} (x_1 + \sqrt{x_3}) = 1, \\ \frac{\partial^2}{\partial x_2 \partial x_2} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_2} (x_1 + \sqrt{x_3}) = 0, \\ \frac{\partial^2}{\partial x_3 \partial x_2} f(x) &= \frac{\partial}{\partial x_3} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_3} (x_1 + \sqrt{x_3}) = \frac{1}{2\sqrt{x_3}}.\end{aligned}\tag{6.66}$$

Für die zweiten partiellen Ableitungen hinsichtlich x_3 ergibt sich

$$\begin{aligned}\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_3} f(x) \right) = \frac{\partial}{\partial x_1} \left(\frac{x_2}{2\sqrt{x_3}} \right) = 0, \\ \frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_3} f(x) \right) = \frac{\partial}{\partial x_2} \left(\frac{x_2}{2\sqrt{x_3}} \right) = \frac{1}{2\sqrt{x_3}}, \\ \frac{\partial^2}{\partial x_3 \partial x_3} f(x) &= \frac{\partial}{\partial x_3} \left(\frac{\partial}{\partial x_3} f(x) \right) = \frac{\partial}{\partial x_3} \left(\frac{x_2}{2\sqrt{x_3}} \right) = -\frac{1}{4} x_2 x_3^{-\frac{3}{2}}.\end{aligned}\tag{6.67}$$

Weiterhin erkennt man, dass die Reihenfolge der partiellen Ableitungen irrelevant ist, denn es gilt

$$\begin{aligned}\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial^2}{\partial x_2 \partial x_1} f(x) = 1, \\ \frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_1} f(x) = 0, \\ \frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_2} f(x) = \frac{1}{2\sqrt{x_3}}.\end{aligned}\tag{6.68}$$

Wie oben gesehen gibt es für eine multivariate reellwertige Funktion $f : \mathbb{R}^n \rightarrow \mathbb{R}$ insgesamt n erste partielle Ableitungen und n^2 zweite partielle Ableitungen. Diese werden im *Gradienten* und der *Hesse-Matrix* einer multivariaten reellwertigen Funktion zusammengefasst.

Definition 6.8 (Gradient). $f : \mathbb{R}^n \rightarrow \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann ist der *Gradient* $\nabla f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix} \in \mathbb{R}^n. \quad (6.69)$$

•

Man beachte, dass Gradienten multivariate vektorwertige Funktionen der

$$\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n, x \mapsto \nabla f(x) \quad (6.70)$$

sind. Für $n = 1$ gilt $\nabla f(x) = f'(x)$. Eine wichtige Eigenschaft des Gradienten ist, dass $-\nabla f(x)$ die Richtung des steilsten Abstiegs von f in $\mathbb{R}^n$ anzeigt. Diese Einsicht ist aber nicht trivial und soll an späterer Stelle vertieft werden. Als Beispiele betrachten wir die Gradienten der oben analysierten Funktionen.

Beispiel (1)

Für die in Gleichung 6.60 betrachtete Funktion $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} \in \mathbb{R}^2. \quad (6.71)$$

In Abbildung 6.8 visualisieren wir farbkodiert ausgewählte Werte dieses Gradienten für $(0.7, 0.7)^T$, $(-0.3, 0.1)^T$, $(-0.5, -0.4)^T$ und $(0.1, -1.0)^T$.

Beispiel (2)

Für die in Gleichung 6.63 betrachtete Funktion $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \frac{\partial}{\partial x_3} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 + x_2 \\ x_1 + \sqrt{x_3} \\ \frac{x_2}{2\sqrt{x_3}} \end{pmatrix} \in \mathbb{R}^3. \quad (6.72)$$

Schließlich widmen wir uns der Zusammenfassung der zweiten partiellen Ableitungen einer multivariaten reellwertigen Funktion in der *Hesse-Matrix*.

Definition 6.9 (Hesse-Matrix). $f : \mathbb{R}^n \rightarrow \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann ist die *Hesse-Matrix* $\nabla^2 f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_n} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(x) & \frac{\partial^2}{\partial x_n \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_n \partial x_n} f(x) \end{pmatrix} \in \mathbb{R}^{n \times n}. \quad (6.73)$$

•

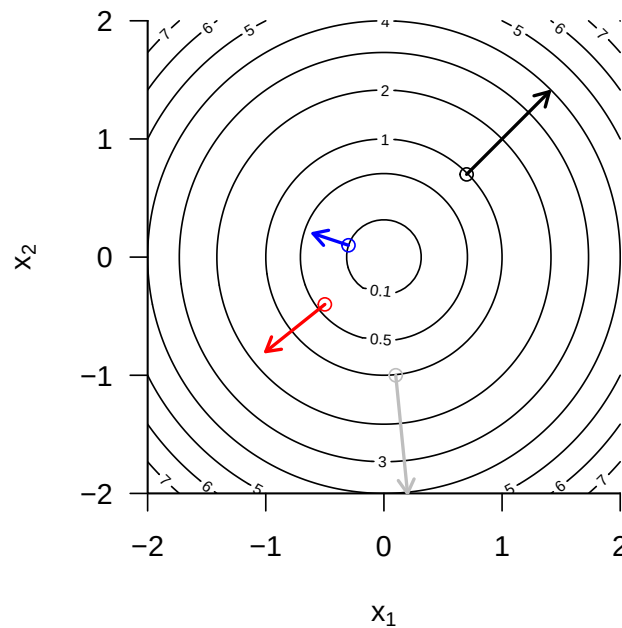


Abbildung 6.8. Exemplarische Gradientenwerte der bivariaten Funktion $f(x) = x_1^2 + x_2^2$.

Man beachte, dass Hesse-Matrizen multivariate matrixwertige Abbildungen der Form

$$\nabla^2 f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}, x \mapsto \nabla^2 f(x) \quad (6.74)$$

sind. Für $n = 1$ gilt $\nabla^2 f(x) = f''(x)$. Weiterhin folgt aus

$$\frac{\partial^2}{\partial x_i \partial x_j} f(x) = \frac{\partial^2}{\partial x_j \partial x_i} f(x) \text{ für } 1 \leq i, j \leq n \quad (6.75)$$

dass die Hesse-Matrix symmetrisch ist, dass also gilt

$$(\nabla^2 f(x))^T = \nabla^2 f(x). \quad (6.76)$$

Beispiel (1)

Für die in Gleichung 6.60 betrachtete Funktion $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 2}. \quad (6.77)$$

Die Hesse-Matrix dieser Funktion ist also eine konstante Funktion, die nicht von x abhängt.

Beispiel (2)

Für die in Gleichung 6.63 betrachtete Funktion $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \frac{\partial^2}{\partial x_1 \partial x_3} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) & \frac{\partial^2}{\partial x_2 \partial x_3} f(x) \\ \frac{\partial^2}{\partial x_3 \partial x_1} f(x) & \frac{\partial^2}{\partial x_3 \partial x_2} f(x) & \frac{\partial^2}{\partial x_3 \partial x_3} f(x) \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & \frac{1}{2\sqrt{3}} \\ 0 & \frac{1}{2\sqrt{3}} & -\frac{1}{4}x_2x_3^{-3/2} \end{pmatrix}. \quad (6.78)$$

Im Gegensatz zu Beispiel (1) ist die Hesse-Matrix der hier betrachteten Funktion keine konstante Funktion und ihr Wert hängt vom Wert des Funktionsarguments $x \in \mathbb{R}^3$ ab.

7. Integralrechnung

Dieses Kapitel gibt einen Überblick über zentrale Begriffe der Integralrechnung. Das Hauptaugenmerk liegt dabei durchgängig auf der Klärung von Begrifflichkeiten, ihrer mathematischen Symbolik und der durch sie vermittelten Intuition und weniger auf der konkreten Berechnung von Integralen.

7.1. Unbestimmte Integrale

Wir beginnen mit der Definition des unbestimmten Integrals und dem Begriff der Stammfunktion.

Definition 7.1 (Unbestimmtes Integral und Stammfunktion). Für ein Intervall $I \subseteq \mathbb{R}$ sei $f : I \rightarrow \mathbb{R}$ eine univariate reellwertige Funktion. Dann heißt eine differenzierbare Funktion $F : I \rightarrow \mathbb{R}$ mit der Eigenschaft

$$F' = f \tag{7.1}$$

Stammfunktion von f . Ist F eine Stammfunktion von f , dann heißt

$$\int f(x) dx := F + c \text{ mit } c \in \mathbb{R} \tag{7.2}$$

unbestimmtes Integral der Funktion f . Das unbestimmte Integral einer Funktion bezeichnet damit die Menge aller Stammfunktionen einer Funktion.

•

Obige Definition besagt, dass die Ableitung der Stammfunktion einer Funktion f eben f ist. Das unbestimmte Integral einer Funktion f ist darüber hinaus die Menge *aller* durch Addition verschiedener Konstanten $c \in \mathbb{R}$ gegebenen Stammfunktionen von f . Eine solche Konstante $c \in \mathbb{R}$ heißt auch *Integrationskonstante*; es gilt natürlich $\frac{d}{dx}c = 0$. Das Symbol $\int f(x) dx$ ist als $F + c$ definiert. $f(x)$ wird in diesem Ausdruck *Integrand* genannt. $\int$ und dx haben keine eigentliche Bedeutung, sondern sind lediglich Symbole.

Für die in vorherigen Abschnitten eingeführten elementaren Funktionen ergeben sich die in Tabelle 7.1 aufgelisteten Stammfunktionen. Man überzeugt sich davon durch Ableiten der jeweiligen Stammfunktion mithilfe der Rechenregeln der Differenzialrechnung. Die uneigentlichen Integrale dieser elementaren Funktionen ergeben sich dann direkt aus diesen Stammfunktionen durch Addition einer Integrationskonstanten.

Tabelle 7.1. Stammfunktionen elementarer Funktionen

Name	Definition	Stammfunktion
Polynomfunktion	$f(x) := \sum_{i=0}^k a_i x^i$	$F(x) = \sum_{i=0}^k \frac{a_i}{i+1} x^{i+1}$
Konstante Funktion	$f(x) := a$	$F(x) = ax$

Tabelle 7.1. Stammfunktionen elementarer Funktionen

Name	Definition	Stammfunktion
Identitätsfunktion	$f(x) := x$	$F(x) = \frac{1}{2}x^2$
Linear-affine Funktion	$f(x) := ax + b$	$F(x) = \frac{1}{2}ax^2 + bx$
Quadratfunktion	$f(x) := x^2$	$F(x) = \frac{1}{3}x^3$
Exponentialfunktion	$f(x) := \exp(x)$	$F(x) = \exp(x)$
Logarithmusfunktion	$f(x) := \ln(x)$	$F(x) = x \ln x - x$

Die in nachfolgendem Theorem zusammengestellten Rechenregeln sind oft hilfreich, um Stammfunktionen von Funktionen zu bestimmen, die sich aus Funktionen mit bekannten Stammfunktionen zusammensetzen.

Theorem 7.1 (Rechenregeln für Stammfunktionen). *f und g seien univariate reellwertige Funktion, die Stammfunktionen besitzen, und g sei invertierbar. Dann gelten folgende Rechenregeln für die Bestimmung von Stammfunktionen*

(1) *Summenregel*

$$\int af(x) + bg(x) dx = a \int f(x) dx + b \int g(x) dx \text{ für } a, b \in \mathbb{R} \quad (7.3)$$

(2) *Partielle Integration*

$$\int f'(x)g(x) dx = f(x)g(x) - \int f(x)g'(x) dx \quad (7.4)$$

(3) *Substitutionsregel*

$$\int f(g(x))g'(x) dx = \int f(t) dt \text{ mit } t = g(x) \quad (7.5)$$

◦

Beweis. Für einen Beweis der Summenregel verweisen wir auf die weiterführende Literatur. Die Regel der partiellen Integration ergibt sich durch Integration der Produktregel der Differentiation. Wir erinnern uns, dass gilt

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x). \quad (7.6)$$

Integration beider Seiten der Gleichung und Berücksichtigung der Summenregel für Stammfunktionen ergibt dann

$$\begin{aligned} \int (f(x)g(x))' dx &= \int f'(x)g(x) + f(x)g'(x) dx \\ &\Leftrightarrow f(x)g(x) = \int f'(x)g(x) dx + \int f(x)g'(x) dx \\ &\Leftrightarrow \int f'(x)g(x) dx = f(x)g(x) - \int f(x)g'(x) dx. \end{aligned} \quad (7.7)$$

Die Substitutionsregel ergibt sich für $F' = f$ durch Anwendung der Kettenregel der Differentiation auf die verkettete Funktion $F(g)$. Speziell gilt zunächst

$$(F(g(x)))' = F'(g(x))g'(x) = f(g(x))g'(x). \quad (7.8)$$

Integration beider Seiten der Gleichung

$$(F(g(x)))' = f(g(x))g'(x) \quad (7.9)$$

ergibt dann

$$\begin{aligned} \int (F(g(x)))' dx &= \int f(g(x))g'(x) dx \\ \Leftrightarrow F(g(x)) + c &= \int f(g(x))g'(x) dx \\ \Leftrightarrow \int f(g(x))g'(x) dx &= \int f(t) dt \text{ mit } t := g(x). \end{aligned} \quad (7.10)$$

Dabei ist die rechte Seite der letzten obigen Gleichung zu verstehen als $F(g(x)) + c$, also als Stammfunktion von f evaluiert an der Stelle $t := g(x)$. Das dt ist nicht durch $dg(x)$ zu ersetzen, sondern rein notationeller Natur. $\square$

Unbestimmte Integrale nehmen in der Lösung von Differenzialgleichungen einen zentralen Platz ein. Naheliegender ist aber zunächst die Anwendung unbestimmter Integrale im Kontext der Auswertung *bestimmter Integrale*, wie im nächsten Abschnitt eingeführt.

7.2. Bestimmte Integrale

Anschaulich entspricht ein bestimmtes Integral der vorzeichenbehafteten und auf ein Intervall $[a, b]$ beschränkten Fläche zwischen dem Graphen einer Funktion f und der x -Achse (vgl. Abbildung 7.1). *Vorzeichenbehaftet* heißt dabei, dass Flächen zwischen der x -Achse und positiven Werten von f positiv zur Fläche beitragen, Flächen zwischen der x und negativen Werten von f dagegen negativ. So ergeben sich zum Beispiel der Wert des in Abbildung 7.1 A gezeigten bestimmten Integral zu 0.68, der Wert des in Abbildung 7.1 B gezeigten bestimmten Integrals zu 0.95 (die eingezeichnete Fläche ist offensichtlich größer als in Abbildung 7.1 A) und der Wert des in Abbildung 7.1 C gezeigten bestimmten Integrals zu 0 (die eingezeichneten positiven und negativen Flächen gleichen sich genau aus). Letzteres Beispiel legt auch die Interpretation des Integrals als Durchschnittswert einer Funktion f über einem Intervall $[a, b]$ nahe.

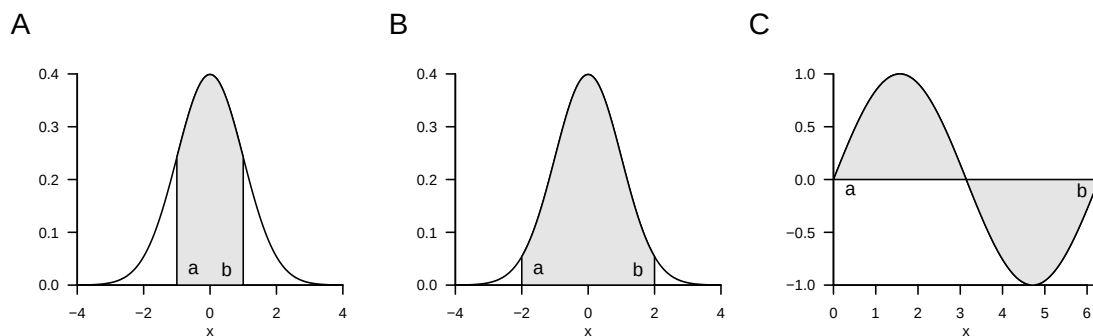


Abbildung 7.1. Beispiele bestimmter Integrale

Um den Begriff des *bestimmten Integrals* im Sinne des *Riemannsches Integrals* einführen zu können, müssen wir zunächst etwas Vorarbeit leisten. Wir beginnen damit, einen Begriff für die Aufteilung eines Intervalls in kleinere Abschnitte einzuführen.

Definition 7.2 (Zerlegung eines Intervalls und Feinheit). Es sei $[a, b] \subset \mathbb{R}$ ein Intervall und $x_0, x_1, x_2, \dots, x_n \in [a, b]$ eine Menge von Punkten mit

$$a =: x_0 < x_1 < x_2 \cdots < x_n := b \quad (7.11)$$

und

$$\Delta x_i := x_i - x_{i-1} \text{ für } i = 1, \dots, n. \quad (7.12)$$

Dann heißt die Menge

$$Z := \{[x_0, x_1], [x_1, x_2], \dots, [x_{n-1}, x_n]\} \quad (7.13)$$

der durch $x_0, x_1, x_2, \dots, x_n$ definierten Teilintervalle von $[a, b]$ eine *Zerlegung von $[a, b]$* . Weiterhin heißt

$$Z_{\max} := \max_{i \in \mathbb{N}_n} \Delta x_i, \quad (7.14)$$

also die größte der Teilintervalllängen Δx_i , die *Feinheit von Z* .

•

Anschaulich ist Δx_i die Breite der Rechtecke in Abbildung 7.2, wie wir in der Folge sehen werden. Mithilfe der Begriffe der Zerlegung eines Intervalls können wir nun den Begriff der *Riemannschen Summen* einführen.

Definition 7.3 (Riemannsche Summen). $f : [a, b] \rightarrow \mathbb{R}$ sei eine beschränkte Funktion auf $[a, b]$, d.h. $|f(x)| < c$ für $0 < c < \infty$ und alle $x \in [a, b]$, Z sei eine Zerlegung von $[a, b]$ mit Teilintervalllängen Δx_i für $i = 1, \dots, n$. Weiterhin sei ξ_i für $i = 1, \dots, n$ ein beliebiger Punkt im Teilintervall $[x_{i-1}, x_i]$ der Zerlegung Z . Dann heißt

$$R(Z) := \sum_{i=1}^n f(\xi_i) \Delta x_i \quad (7.15)$$

Riemannsche Summe von f auf $[a, b]$ bezüglich der Zerlegung Z .

•

Wählt man zum Beispiel in der Riemannschen Summe in jedem Teilintervall das Maximum von f , so ergibt sich die sogenannte *Riemannsche Obersumme*,

$$R_o(Z) := \sum_{i=1}^n \left(\max_{[x_{i-1}, x_i]} f(\xi_i) \right) \cdot \Delta x_i. \quad (7.16)$$

Wählt man in jedem Teilintervall dagegen das Minimum von f , so ergibt sich die sogenannte *Riemannsche Untersumme*,

$$R_u(Z) := \sum_{i=1}^n \left(\min_{[x_{i-1}, x_i]} f(\xi_i) \right) \cdot \Delta x_i. \quad (7.17)$$

Abbildung 7.2 verdeutlicht die Definition dieser Riemannschen Summen: die dunkelgrauen Rechtecke haben jeweils die Fläche $[x_{i-1}, x_i] \cdot \min_{[x_{i-1}, x_i]} f(\xi)$ und bilden damit die Summenterme in der Riemannschen Untersumme

$$R_u(Z) := \sum_{i=1}^4 \left(\min_{[x_{i-1}, x_i]} f(\xi_i) \right) \cdot \Delta x_i. \quad (7.18)$$

Die vertikale Kombination aus dunkelgrauen und hellgrauen Rechtecken hat jeweils die Fläche $[x_{i-1}, x_i] \cdot \max_{[x_{i-1}, x_i]} f(\xi)$ und bilden damit die Summenterme in der Riemannschen Obersumme

$$R_o(Z) := \sum_{i=1}^4 \left(\max_{[x_{i-1}, x_i]} f(\xi_i) \right) \cdot \Delta x_i. \quad (7.19)$$

Stellt man sich nun vor, dass man Δx_i für alle $i = 1, \dots, n$ gegen null gehen lässt, verkleinert man die Feinheit der Zerlegung Z also immer weiter, so werden sich die Werte von $\min_{[x_{i-1}, x_i]} f(\xi_i)$ und $\max_{[x_{i-1}, x_i]} f(\xi_i)$ und damit auch die Werte von $R_u(Z)$ und $R_o(Z)$ immer weiter annähern. Diesen Grenzprozess macht man sich in der Definition des Riemannschen Integrals zunutze.

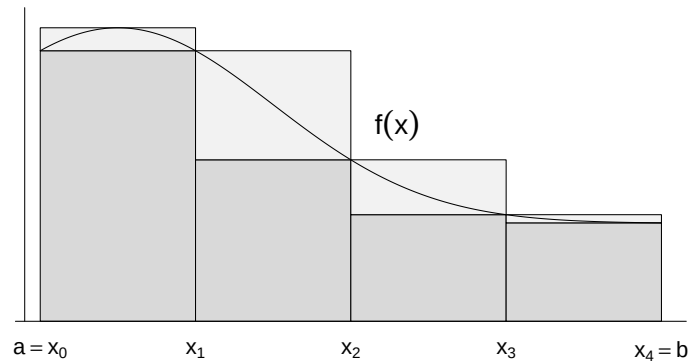


Abbildung 7.2. Riemannsche Summen

Definition 7.4 (Bestimmtes Riemannsches Integral). $f : [a, b] \rightarrow \mathbb{R}$ sei eine beschränkte reellwertige Funktion auf $[a, b]$. Weiterhin sei für Z_k mit $k = 1, 2, 3, \dots$ eine Folge von Zerlegungen von $[a, b]$ mit zugehörigen Feinheit $Z_{\max, k}$. Wenn für jede Folge von Zerlegungen $Z_1, Z_2, \dots$ mit $|Z_{\max, k}| \rightarrow 0$ für $k \rightarrow \infty$ und für beliebig gewählte Punkte ξ_{ki} mit $i = 1, \dots, n$ im Teilintervall $[x_{k,i-1}, x_{k,i}]$ der Zerlegung Z_k gilt, dass die Folge der zugehörigen Riemannschen Summen $R(Z_1), R(Z_2), \dots$ gegen den gleichen Grenzwert strebt, dann heißt f auf $[a, b]$ *integrierbar*. Der entsprechende Grenzwert der Folge von Riemannschen Summen wird *bestimmtes Riemannsches Integral* genannt und mit

$$\int_a^b f(x) dx := \lim_{k \rightarrow \infty} R(Z_k) \text{ für } |Z_{\max, k}| \rightarrow 0 \quad (7.20)$$

bezeichnet. Die Werte a und b bezeichnet man in diesem Kontext als *untere* und *obere* Integrationsgrenzen, respektive, $f(x)$ als *Integrand* und x als *Integrationsvariable*.

•

Die Riemannsche Integrierbarkeit einer Funktion und der Wert eines bestimmten Riemannschen Integrals sind also im Sinne einer Grenzwertbildung definiert. Die Theorie der Riemannschen Integrale lässt sich allerdings um die Hauptsätze der Differenzial- und Integralrechnung erweitern, sodass zur konkreten Berechnung eines bestimmten Integrals die Bildung von Zerlegungen und die Bestimmung eines Grenzwertes nur selten nötig ist. Der Einfachheit halber verzichten wir in der Folge auf die Bezeichnungen *Riemannsche* und sprechen einfach von *bestimmten Integralen*.

Ein erster Schritt zur Vereinfachung der Berechnung von bestimmten Integralen ist das Feststellen folgender Rechenregeln, für deren Beweis wir auf die weiterführende Literatur verweisen.

Theorem 7.2 (Rechenregeln für bestimmte Integrale). *Es seien f und g integrierbare Funktionen auf $[a, b]$. Dann gelten folgende Rechenregeln.*

(1) *Linearität. Für $c_1, c_2 \in \mathbb{R}$ gilt*

$$\int_a^b (c_1 f(x) + c_2 g(x)) dx = c_1 \int_a^b f(x) dx + c_2 \int_a^b g(x) dx. \quad (7.21)$$

(2) *Additivität. Für $a < c < b$ gilt*

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx. \quad (7.22)$$

(3) *Vorzeichenwechsel bei Umkehrung der Integralgrenzen*

$$\int_a^b f(x) dx = - \int_b^a f(x) dx. \quad (7.23)$$

(4) *Unabhängigkeit von der Wahl der Integrationsvariablen*

$$\int_a^b f(x) dx = \int_a^b f(y) dy. \quad (7.24)$$

(5) *Unabhängigkeit des Integrals von Art des Intervalls*

$$\int_a^b f(x) dx = \int_{]a,b[} f(x) dx = \int_{[a,b]} f(x) dx = \int_{]a,b]} f(x) dx = \int_{[a,b[} f(x) dx, \quad (7.25)$$

wobei $\int_I$ das bestimmte Integral von f auf dem Intervall $I \subseteq \mathbb{R}$ bezeichnet.

◦

Man beachte, dass die Linearität des bestimmten Integrals analog zur Assoziativität bei Summen gleicher Länge und der Distributivität bei Multiplikation mit einer Konstante ist (vgl. Theorem 3.1),

$$\sum_{i=1}^n (c_1 x_i + c_2 y_i) = c_1 \sum_{i=1}^n x_i + c_2 \sum_{i=1}^n y_i,$$

die Additivität bestimmter Integrale analog zum Aufspalten von Summen ist, und dass auch die Unabhängigkeit des Werts eines Integrals von der Wahl der Integrationsvariablen in der Unabhängigkeit des Werts einer Summe von ihrem Laufindex ihre Entsprechung findet. Eine grafische Darstellung der Integralrechenregel der Additivität zeigt

Abbildung 7.3. Die Summe der durch die bestimmten Integrale gegebenen Flächen $\int_a^c f(x) dx$ und $\int_c^b f(x) dx$ mit $a < c < b$ ergibt sich dabei zur Fläche von $\int_a^b f(x) dx$.

Mithilfe der Additivität und dem Nullintervallintegral

$$\int_a^a f(x) dx = 0 \quad (7.26)$$

lässt sich der im Folgenden wichtige Vorzeichenwechsel bei Umkehrung der Integralgrenzen begründen. Sei dazu wie in Abbildung 7.3

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx. \text{ und } \int_b^a f(x) dx = \int_b^c f(x) dx + \int_c^a f(x) dx. \quad (7.27)$$

Addiert man beide Gleichungen, so erhält man

$$\begin{aligned} \int_a^b f(x) dx + \int_b^a f(x) dx &= \int_a^c f(x) dx + \int_c^b f(x) dx + \int_b^c f(x) dx + \int_c^a f(x) dx \\ &\Leftrightarrow \int_a^a f(x) dx = \int_a^b f(x) dx + \int_b^a f(x) dx \\ &\Leftrightarrow 0 = \int_a^b f(x) dx + \int_b^a f(x) dx \\ &\Leftrightarrow \int_a^b f(x) dx = - \int_b^a f(x) dx. \end{aligned} \quad (7.28)$$

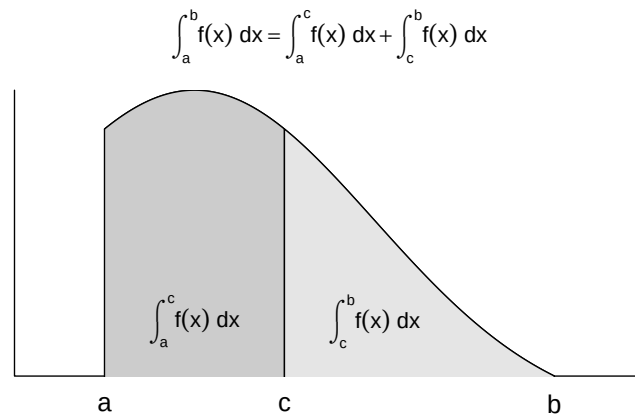


Abbildung 7.3. Additivität bestimmter Integrale

Die *Hauptsätze der Differenzial- und Integralrechnung* schließlich ermöglichen es, bestimmte Integrale einer Funktion f direkt mithilfe der Stammfunktion F von f zu berechnen. Für den Beweis des ersten Hauptsatzes der Differenzial- und Integralrechnung benötigen wir den *Mittelwertsatz der Integralrechnung*, welchen wir hier ohne Beweis wiedergeben und in Abbildung 7.4 veranschaulichen.

Theorem 7.3 (Mittelwertsatz der Integralrechnung). *Für eine stetige Funktion $f : [a, b] \rightarrow \mathbb{R}$ existiert ein $\xi \in]a, b[$ mit*

$$\int_a^b f(x) dx = f(\xi)(b - a) \quad (7.29)$$

◦

Der Mittelwertsatz der Integralrechnung garantiert die Existenz eines $\xi \in [a, b]$, sodass das bestimmte Integral $\int_a^b f(x) dx$ gleich dem Produkt aus der “Rechteckhöhe” $f(\xi)$ und der “Rechteckbreite” $(b - a)$ ist. In Abbildung 7.4 liegt dieses ξ genau mittig zwischen a und b .

Dass die sich so ergebene grau eingefärbte Rechteckfläche gleich $\int_a^b f(x) dx$ ist, ergibt sich aus der visuell zumindest nachvollziehbaren Tatsache, dass die Flächen zwischen $f(x)$ und $f(\xi)$ im Intervall $[a, \xi]$ und zwischen $f(\xi)$ und $f(x)$ im Intervall $[\xi, b]$ den gleichen Betrag haben, erstere aber mit einem negativen Vorzeichen behaftet ist. Der Mittelwertsatz der Integralrechnung garantiert im Allgemeinen aber nur die Existenz eines $\xi \in [a, b]$ mit der diskutierten Eigenschaft, gibt aber keine Formel zu Bestimmung von ξ an.

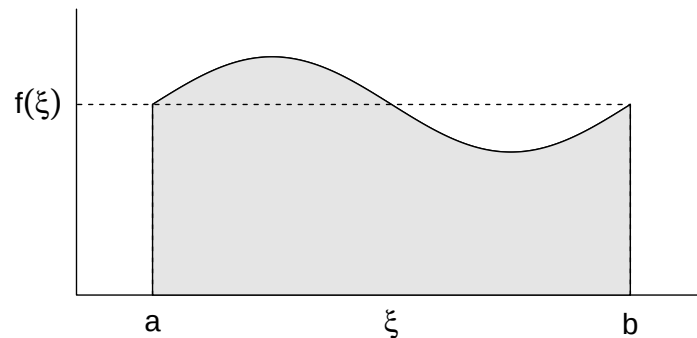


Abbildung 7.4. Zum Mittelwertsatz der Integralrechnung

Mit dieser Vorarbeit können wir nun den ersten Hauptsatz der Differenzial- und Integralrechnung formulieren und beweisen.

Theorem 7.4 (Erster Hauptsatz der Differenzial- und Integralrechnung). *Ist $f : I \rightarrow \mathbb{R}$ eine auf dem Intervall $I \subset \mathbb{R}$ stetige Funktion, dann ist die Funktion*

$$F : I \rightarrow \mathbb{R}, x \mapsto F(x) := \int_a^x f(t) dt \text{ mit } x, a \in I \quad (7.30)$$

eine Stammfunktion von f .

◦

Beweis. Wir betrachten den Differenzquotienten

$$\frac{1}{h}(F(x+h) - F(x)) \quad (7.31)$$

Mit der Definition $F(x) := \int_a^x f(t) dt$ und der Additivität des bestimmten Integrals gilt dann

$$\frac{1}{h}(F(x+h) - F(x)) = \frac{1}{h} \left(\int_a^{x+h} f(t) dt - \int_a^x f(t) dt \right) = \frac{1}{h} \int_x^{x+h} f(t) dt \quad (7.32)$$

Mit dem Mittelwertsatz der Integralrechnung gibt es also ein $\xi \in]x, x+h[$, sodass

$$\frac{1}{h}(F(x+h) - F(x)) = f(\xi) \quad (7.33)$$

Grenzwertbildung ergibt dann

$$\lim_{h \rightarrow 0} \frac{1}{h}(F(x+h) - F(x)) = \lim_{h \rightarrow 0} f(\xi) \text{ für } \xi \in]x, x+h[\Leftrightarrow F'(x) = f(x). \quad (7.34)$$

□

Der zweite Hauptsatz der Differenzial- und Integralrechnung besagt, wie man mithilfe der Stammfunktion ein bestimmtes Integral berechnet.

Theorem 7.5 (Zweiter Hauptsatz der Differenzial- und Integralrechnung). *Ist F eine Stammfunktion einer stetigen Funktion $f : I \rightarrow \mathbb{R}$ auf einem Intervall I , so gilt für $a, b \in I$ mit $a \leq b$*

$$\int_a^b f(x) dx = F(b) - F(a) =: F(x)|_a^b. \quad (7.35)$$

◦

Beweis. Mit den Rechenregeln für bestimmte Integrale und dem ersten Hauptsatz der Differenzial- und Integralrechnung ergibt sich

$$F(b) - F(a) = \int_a^b f(t) dt - \int_a^a f(t) dt = \int_a^a f(t) dt + \int_a^b f(t) dt = \int_a^b f(x) dx. \quad (7.36)$$

□

Wir wollen den Zweiten Hauptsatz der Differenzial- und Integralrechnung in drei Beispielen anwenden (vgl. Abbildung 7.5).

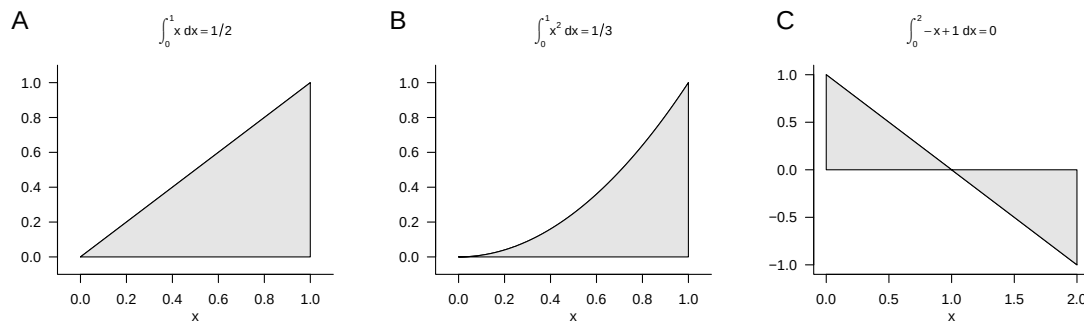


Abbildung 7.5. Beispiele zum Zweiten Hauptsatz der Differenzial- und Integralrechnung

Beispiel (1)

Wir betrachten die Identitätsfunktion

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := x \quad (7.37)$$

und wollen das bestimmte Integral dieser Funktion auf dem Intervall $[0, 1]$, also

$$\int_0^1 f(x) dx = \int_0^1 x dx \quad (7.38)$$

berechnen. Dazu erinnern wir uns, dass eine Stammfunktion von f durch

$$F : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto F(x) := \frac{1}{2}x^2 \quad (7.39)$$

gegeben ist, weil

$$F'(x) = \frac{d}{dx} \left(\frac{1}{2}x^2 \right) = 2 \cdot \frac{1}{2}x^{2-1} = x. \quad (7.40)$$

Einsetzen in den zweiten Hauptsatz der Differenzial- und Integralrechnung ergibt dann

$$\int_0^1 x \, dx = \frac{1}{2}1^2 - \frac{1}{2}0^2 = \frac{1}{2}. \quad (7.41)$$

Dieses Ergebnis ist mit der Intuition, die sich anhand der grauen Fläche in Abbildung 7.5 A ergibt, kongruent.

Beispiel (2)

Als nächstes betrachten wir die Quadratfunktion

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := x^2 \quad (7.42)$$

und wollen das bestimmte Integral dieser Funktion auf dem Intervall $[0, 1]$, also

$$\int_0^1 f(x) \, dx = \int_0^1 x^2 \, dx \quad (7.43)$$

berechnen. Dazu erinnern wir uns, dass eine Stammfunktion von f durch

$$F : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto F(x) := \frac{1}{3}x^3 \quad (7.44)$$

gegeben ist, weil

$$F'(x) = \frac{d}{dx} \left(\frac{1}{3}x^3 \right) = 3 \cdot \frac{1}{3}x^{3-1} = x^2. \quad (7.45)$$

Einsetzen in den zweiten Hauptsatz der Differenzial- und Integralrechnung ergibt dann

$$\int_0^1 x^2 \, dx = \frac{1}{3}1^3 - \frac{1}{3}0^3 = \frac{1}{3}. \quad (7.46)$$

Dieses Ergebnis ist mit der Intuition, die sich aus dem Vergleich der grauen Flächen in Abbildung 7.5 A und Abbildung 7.5 B ergibt, kongruent.

Beispiel (3)

Schließlich betrachten wir die linear-affine Funktion

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := -x + 1 \quad (7.47)$$

und wollen das bestimmte Integral dieser Funktion auf dem Intervall $[0, 2]$, also

$$\int_0^2 f(x) \, dx = \int_0^2 -x + 1 \, dx \quad (7.48)$$

berechnen. Dazu erinnern wir uns, dass eine Stammfunktion der linearen Funktion mit $a = -1$ und $b = 1$ (vgl. Tabelle 7.1) durch

$$F : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto F(x) := -\frac{1}{2}x^2 + x \quad (7.49)$$

gegeben ist, weil

$$F'(x) = \frac{d}{dx} \left(-\frac{1}{2}x^2 + x \right) = -2 \cdot \frac{1}{2}x^{2-1} + 1 \cdot x^{1-1} = -x + 1. \quad (7.50)$$

Einsetzen in den Zweiten Hauptsatz der Differenzial- und Integralrechnung ergibt dann

$$\int_0^2 -x + 1 \, dx = \left(-\frac{1}{2}2^2 + 2 \right) - \left(-\frac{1}{2}0^2 + 0 \right) = -2 + 2 - 0 = 0. \quad (7.51)$$

Dieses Ergebnis ist mit der Intuition kongruent, dass sich die positive und die negative graue Fläche in Abbildung 7.5 C ausgleichen, kongruent.

7.3. Uneigentliche Integrale

Uneigentliche Integrale sind bestimmte Integrale, bei denen mindestens eine Integrationsgrenze keine reelle Zahl, sondern $-\infty$ oder ∞ ist. Wir beleuchten die Natur uneigentlicher Integrale mit folgender Definition und einem Beispiel.

Definition 7.5 (Uneigentliche Integrale). $f : \mathbb{R} \rightarrow \mathbb{R}$ sei eine univariate reellwertige Funktion. Mit den Definitionen

$$\int_{-\infty}^b f(x) dx := \lim_{a \rightarrow -\infty} \int_a^b f(x) dx \text{ und } \int_a^{\infty} f(x) dx := \lim_{b \rightarrow \infty} \int_a^b f(x) dx \quad (7.52)$$

und der Additivität von Integralen

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^b f(x) dx + \int_b^{\infty} f(x) dx \quad (7.53)$$

wird der Begriff des bestimmten Integrals auf die unbeschränkten Integrationsintervalle $]-\infty, b]$, $[a, \infty[$ und $]-\infty, \infty[$ erweitert. Integrale mit unbeschränkten Integrationsintervallen heißen *uneigentliche Integrale*. Wenn die entsprechenden Grenzwerte existieren, sagt man, dass die uneigentlichen Integrale *konvergieren*.

•

Als Beispiel betrachten wir das uneigentliche Integral der Funktion

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := \frac{1}{x^2} \quad (7.54)$$

auf dem Intervall $[1, \infty[$, also

$$\int_1^{\infty} \frac{1}{x^2} dx. \quad (7.55)$$

Nach den Festlegungen in der Definition uneigentlicher Integrale gilt

$$\int_1^{\infty} \frac{1}{x^2} dx = \lim_{b \rightarrow \infty} \int_1^b \frac{1}{x^2} dx. \quad (7.56)$$

Mit der Stammfunktion $F(x) = -x^{-1}$ von $f(x) = x^{-2}$ ergibt sich für das bestimmte Integral in obiger Gleichung

$$\int_1^b \frac{1}{x^2} dx = F(b) - F(1) = -\frac{1}{b} - \left(-\frac{1}{1}\right) = -\frac{1}{b} + 1. \quad (7.57)$$

Es ergibt sich also

$$\int_1^{\infty} \frac{1}{x^2} dx = \lim_{b \rightarrow \infty} \int_1^b \frac{1}{x^2} dx = \lim_{b \rightarrow \infty} \left(-\frac{1}{b} + 1\right) = -\lim_{b \rightarrow \infty} \frac{1}{b} + \lim_{b \rightarrow \infty} 1 = 0 + 1 = 1. \quad (7.58)$$

7.4. Mehrdimensionale Integrale

Bisher haben wir nur Integrale univariater reellwertiger Funktionen betrachtet. Der Integralbegriff lässt sich auch auf multivariate reellwertige Funktionen erweitern. Allerdings ist dann der Integrationsbereich der Funktion nicht notwendigerweise so einfach zu beschreiben wie ein Intervall. Insbesondere sind beispielsweise schon bei bivariaten Funktionen arbiträr geformte Integrationsbereiche möglich. Wir wollen hier nun den einfachsten Fall eines *Hyperrechtecks* betrachten. In diesem Fall können wir mehrdimensionale bestimmte Integrale wie folgt definieren.

Definition 7.6 (Mehrdimensionale bestimmte Integrale auf Hyperrechtecken). $f : \mathbb{R}^n \rightarrow \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann heißen Integrale der Form

$$\int_{[a_1, b_1] \times \dots \times [a_n, b_n]} f(x) dx = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n) dx_1 \dots dx_n \quad (7.59)$$

mehrdimensionale bestimmte Integrale auf Hyperrechtecken. Weiterhin heißen Integrale der Form

$$\int_{\mathbb{R}^n} f(x) dx = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_1 \dots dx_n \quad (7.60)$$

mehrdimensionale uneigentliche Integrale.

•

Als Beispiel für Definition 7.6 betrachten wir das zweidimensionale bestimmte Integral der Funktion

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}, (x_1, x_2) \mapsto f(x_1, x_2) := x_1^2 + 4x_2 \quad (7.61)$$

auf dem Rechteck $[0, 1] \times [0, 1]$. Der *Satz von Fubini* der Theorie mehrdimensionaler Integrale besagt, dass man mehrdimensionale Integrale in beliebiger Koordinatenfolge auswerten kann. Es gilt also zum Beispiel, dass

$$\int_{a_1}^{b_1} \left(\int_{a_2}^{b_2} f(x_1, x_2) dx_2 \right) dx_1 = \int_{a_2}^{b_2} \left(\int_{a_1}^{b_1} f(x_1, x_2) dx_1 \right) dx_2. \quad (7.62)$$

In diesem Sinne betrachten wir für das Beispiel

$$\int_0^1 \int_0^1 x_1^2 + 4x_2 dx_1 dx_2 = \int_0^1 \left(\int_0^1 x_1^2 + 4x_2 dx_1 \right) dx_2 \quad (7.63)$$

also zunächst das innere Integral. x_2 nimmt dabei die Rolle einer Konstanten ein. Eine Stammfunktion von $g(x_1) := x_1^2 + 4x_2$ ist $G(x_1) = \frac{1}{3}x_1^3 + 4x_2x_1$, wovon man sich durch Ableiten von G überzeugt. Es ergibt sich also für das innere Integral

$$\begin{aligned} \int_0^1 x_1^2 + 4x_2 dx_1 &= G(1) - G(0) \\ &= \frac{1}{3} \cdot 1^3 + 4x_2 \cdot 1 - \frac{1}{3} \cdot 0^3 - 4x_2 \cdot 0 \\ &= \frac{1}{3} + 4x_2. \end{aligned} \quad (7.64)$$

Betrachten des äußeren Integrals

$$\int_0^1 4x_2 + \frac{1}{3} dx_2$$

ergibt dann mit der Stammfunktion

$$H(x_2) = 2x_2^2 + \frac{1}{3}x_2 \quad (7.65)$$

von

$$h(x_2) := 4x_2 + \frac{1}{3}, \quad (7.66)$$

dass

$$\begin{aligned} \int_0^1 \int_0^1 x_1^2 + 4x_2 dx_1 dx_2 &= \int_0^1 4x_2 + \frac{1}{3} dx_2 \\ &= H(1) - H(0) \\ &= 2 \cdot 1^2 + \frac{1}{3} \cdot 1 - 2 \cdot 0^2 - \frac{1}{3} \cdot 0 \\ &= \frac{7}{3}. \end{aligned} \quad (7.67)$$

8. Vektoren

In der naturwissenschaftlichen Modellbildung betrachtet man häufig Phänomene, die sich durch das Vorliegen mehrerer quantitativer Merkmale auszeichnen. So ist zum Beispiel die Position eines Objektes im dreidimensionalen Raum durch drei Koordinaten hinsichtlich der drei Achsen eines Kartesischen Koordinatensystems festgelegt. Analog mag der Gesundheitszustand einer Person durch das Vorliegen dreier Messwerte, beispielsweise eines Selbstauskunftscore, eines Biomarkers und einer Expert:inneneinschätzung charakterisiert sein. Zum modellieren und analysieren solcher mehrdimensionalen quantitativen Phänomene stellt die Mathematik mit dem reellen Vektorraum ein vielseitig einsetzbares Hilfsmittel bereit. In diesem Kapitel wollen wir zunächst den Begriff des reellen Vektorraums und das grundlegende Rechnen mit Vektoren einführen (Kapitel 8.1). Eine Vektorraumstruktur, die sich stark an der dreidimensionalen räumlichen Intuition orientiert bietet dann der Euklidische Vektorraum (Kapitel 8.2). Mithilfe der Vektorrechnung können alle Vektoren eines Vektorraums aus einer kleinen Schar ausgezeichneter Vektoren gebildet werden. Die diesem Prinzip zugrundeliegenden Konzepte diskutieren wir in Kapitel 8.3 und Kapitel 8.4.

8.1. Reeller Vektorraum

Wir beginnen mit der allgemeinen Definition eines Vektorraums, die grundlegende Regeln zum Rechnen mit Vektoren festlegt.

Definition 8.1 (Vektorraum). Es seien V eine nichtleere Menge und S eine Menge von Skalaren. Weiterhin sei eine Abbildung

$$+ : V \times V \rightarrow V, (v_1, v_2) \mapsto +(v_1, v_2) =: v_1 + v_2, \quad (8.1)$$

genannt *Vektoraddition*, definiert. Schließlich sei eine Abbildung

$$\cdot : S \times V \rightarrow V, (s, v) \mapsto \cdot(s, v) =: sv, \quad (8.2)$$

genannt *Skalarmultiplikation* definiert. Dann wird das Tupel $(V, S, +, \cdot)$ genau dann *Vektorraum* genannt, wenn für beliebige Elemente $v, w, u \in V$ und $a, b \in S$ folgende Bedingungen gelten:

(1) *Kommutativität der Vektoraddition.*

$$v + w = w + v.$$

(2) *Assoziativität der Vektoraddition.*

$$(v + w) + u = v + (w + u)$$

(3) *Existenz eines neutralen Elements der Vektoraddition.*

$$\text{Es gibt einen Vektor } 0 \in V \text{ mit } v + 0 = 0 + v = v.$$

(4) *Existenz inverser Elemente der Vektoraddition*

Für alle Vektoren $v \in V$ gibt es einen Vektor $-v \in V$ mit $v + (-v) = 0$.

(5) *Existenz eines neutralen Elements der Skalarmultiplikation.*

Es gibt einen Skalar $1 \in S$ mit $1 \cdot v = v$.

(6) *Assoziativität der Skalarmultiplikation.*

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c.$$

(7) *Distributivität hinsichtlich der Vektoraddition.*

$$a \cdot (v + w) = a \cdot v + a \cdot w.$$

(8) *Distributivität hinsichtlich der Skalaraddition.*

$$(a + b) \cdot v = a \cdot v + b \cdot v.$$

•

Es fällt auf, dass Definition 8.1 zwar festlegt, wie mit Vektoren gerechnet werden soll, jedoch keine Aussage darüber macht, was ein Vektor, über ein Element einer Menge hinaus, eigentlich ist. Dies ist der Tatsache geschuldet, dass es verschiedenste mathematische Objekte gibt, für die Vektorraumstrukturen definiert werden können. Beispiele dafür sind die Menge der reellen m -Tupel, die Menge der Matrizen, die Menge der Polynome, die Menge der Lösungen eines linearen Gleichungssystems, die Menge der reellen Folgen, die Menge der stetigen Funktionen u.v.a.m.

Wir sind hier zunächst nur am Vektorraum der Menge reellen m -Tupel interessiert. Wir erinnern dazu daran, dass wir die reellen m -Tupel mit

$$\mathbb{R}^m := \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \mid x_i \in \mathbb{R} \text{ für alle } 1 \leq i \leq m \right\} \quad (8.3)$$

bezeichnen und $\mathbb{R}^m$ als “ $\mathbb{R}$ hoch m ” aussprechen. Die Elemente $x \in \mathbb{R}^m$ nennen wir *reelle Vektoren* oder auch einfach *Vektoren*. Wir wollen nun der Definition eines Vektorraums die Menge $\mathbb{R}^m$ zugrunde legen. Dazu definieren wir zunächst die Vektoraddition für Elemente von $\mathbb{R}^m$ und die Skalarmultiplikation für Elemente von $\mathbb{R}$ und $\mathbb{R}^m$

Definition 8.2 (Vektoraddition und Skalarmultiplikation in $\mathbb{R}^m$). Für alle $x, y \in \mathbb{R}^m$ und $a \in \mathbb{R}$ sei die *Vektoraddition* durch

$$+ : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}^m, (x, y) \mapsto x + y = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} := \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_m + y_m \end{pmatrix} \quad (8.4)$$

und die *Skalarmultiplikation* durch

$$\cdot : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m, (a, x) \mapsto ax = a \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} := \begin{pmatrix} ax_1 \\ \vdots \\ ax_m \end{pmatrix} \quad (8.5)$$

definiert.

•

Es ergibt sich dann folgendes Resultat.

Theorem 8.1 (Reeller Vektorraum). $(\mathbb{R}^m, +, \cdot)$ mit den Rechenregeln der Addition und Multiplikation in $\mathbb{R}$ einen Vektorraum.

◦

Für einen Beweis, auf den wir hier verzichten wollen, muss man die Bedingungen (1) bis (8) aus Definition 8.1 für die hier betrachtete Menge und die hier festgelegten Formen der Vektoraddition und der Skalarmultiplikation nachweisen. Diese ergeben sich aber leicht aus den Rechenregeln von Addition und Multiplikation in $\mathbb{R}$ und der Tatsache, dass Vektoraddition und Skalarmultiplikation für Elemente von $\mathbb{R}^m$ in Definition 8.2 komponentenweise definiert wurden. Wir definieren damit den Begriff des *reellen Vektorraums*.

Definition 8.3 (Reeller Vektorraum). Für $\mathbb{R}^m$ seien $+$ und $\cdot$ die in Definition 8.2 definierte Vektoraddition und Skalarmultiplikation. Dann nennen wir auf Grundlage von Theorem 8.1 den Vektorraum $(\mathbb{R}^m, +, \cdot)$ den *reellen Vektorraum*

•

Auf Grundlage von Definition 8.3 wollen wir uns nun das Rechnen mit reellen Vektoren anhand einiger Beispiele verdeutlichen.

Beispiele

(1) Für

$$x := \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \in \mathbb{R}^4 \text{ und } y := \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^4$$

gilt

$$x + y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1+2 \\ 2+1 \\ 3+0 \\ 4+1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \\ 5 \end{pmatrix} \in \mathbb{R}^4.$$

In **R** implementiert dieses Beispiel wie folgt

```
x = matrix(c(1,2,3,4), nrow = 4) # Vektordefinition
y = matrix(c(2,1,0,1), nrow = 4) # Vektordefinition
x + y                               # Vektoraddition
```

```
      [,1]
[1,]    3
[2,]    3
[3,]    3
[4,]    5
```

(2) Für

$$x := \begin{pmatrix} 2 \\ 3 \end{pmatrix} \in \mathbb{R}^2 \text{ und } y := \begin{pmatrix} 1 \\ 3 \end{pmatrix} \in \mathbb{R}^2$$

gilt

$$x - y = \begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 2-1 \\ 3-3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in \mathbb{R}^2.$$

In **R** implementiert man dieses Beispiel wie folgt

```
x = matrix(c(2,3), nrow = 2) # Vektordefinition
y = matrix(c(1,3), nrow = 2) # Vektordefinition
x - y                         # Vektorsubtraktion
```

```
      [,1]
[1,]    1
[2,]    0
```

(3) Für

$$x := \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} \in \mathbb{R}^3 \text{ und } a := 3 \in \mathbb{R}$$

gilt

$$ax = 3 \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \cdot 2 \\ 3 \cdot 1 \\ 3 \cdot 3 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \\ 9 \end{pmatrix} \in \mathbb{R}^3.$$

In **R** implementiert man dieses Beispiel wie folgt

```
x = matrix(c(2,1,3), nrow = 3) # Vektordefinition
a = 3                             # Skalardefinition
a*x                               # Skalarmultiplikation
```

```
      [,1]
[1,]    6
[2,]    3
[3,]    9
```

Für $m \in \{1, 2, 3\}$ kann man sich reelle Vektoren und das Rechnen mit ihnen visuell veranschaulichen. Für $m > 3$, wenn also zum Beispiel für eine Person mehr als drei quantitative Merkmale zu ihrem Gesundheitszustand vorliegen, was in der Anwendung regelmäßig der Fall ist, ist dies nicht möglich. Trotzdem mag die visuelle Intuition für $m \leq 3$ einen Einstieg in das Verständnis von Vektorräumen erleichtern. Wir fokussieren hier auf den Fall $m := 2$. In diesem Fall liegen die betrachteten reellen Vektoren in der zweidimensionalen Ebene und werden üblicherweise als Punkte oder Pfeile visualisiert (Abbildung 8.1).

Abbildung 8.2 visualisiert die Vektoraddition

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}. \quad (8.6)$$

Der Summenvektor entspricht dabei der Diagonale des von den beiden Summanden aufgespannten Parallelogramms.

Abbildung 8.3 visualisiert die Vektorsubtraktion

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} -3 \\ -1 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \quad (8.7)$$

Der resultierende Vektor entspricht dabei der Diagonale des von dem ersten Vektors und dem entgegengesetzten Vektor des zweiten Vektors aufgespannten Parallelogramms.

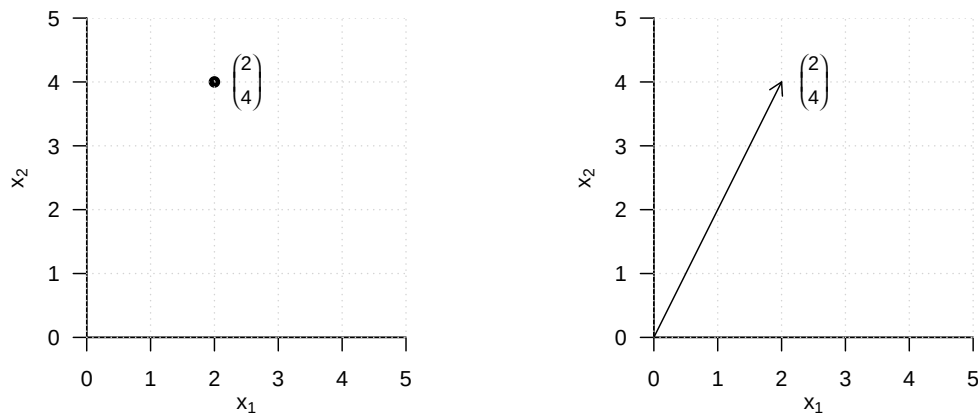


Abbildung 8.1. Visualisierung von Vektoren in $\mathbb{R}^2$

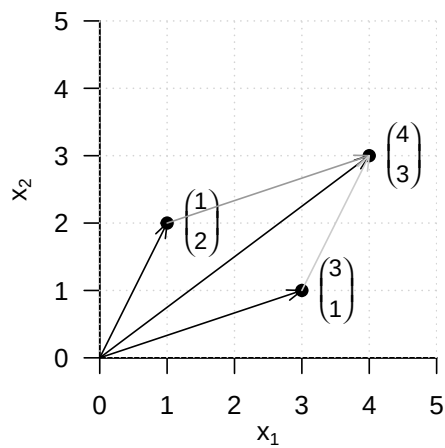


Abbildung 8.2. Vektoraddition in $\mathbb{R}^2$

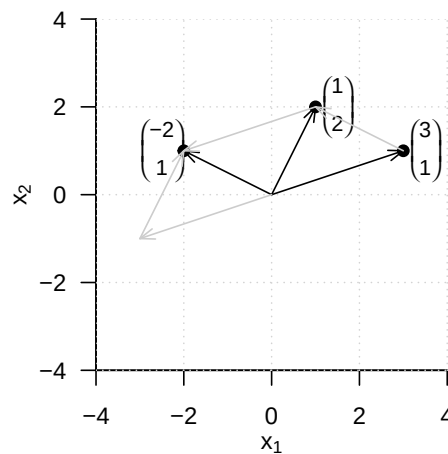


Abbildung 8.3. Vektorsubtraktion in $\mathbb{R}^2$

Abbildung 8.4 schließlich visualisiert die Skalarmultiplikation

$$3 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \quad (8.8)$$

Die Multiplikation eines Vektors mit einem Skalar ändert dabei immer nur seine Länge, nicht jedoch seine Richtung.

8.2. Euklidischer Vektorraum

Der reelle Vektorraum kann durch Definition des *Skalarprodukts* im Sinne eines *Euklidischen Vektorraums* mit räumlich-geometrischer Intuition versehen werden. Diese ermöglicht es insbesondere, Begriffe wie die *Länge eines Vektors*, den *Abstand zwischen zwei Vektoren*, und nicht zuletzt den *Winkel zwischen zwei Vektoren* zu definieren und zu berechnen. Wir führen zunächst das *Skalarprodukt* ein.

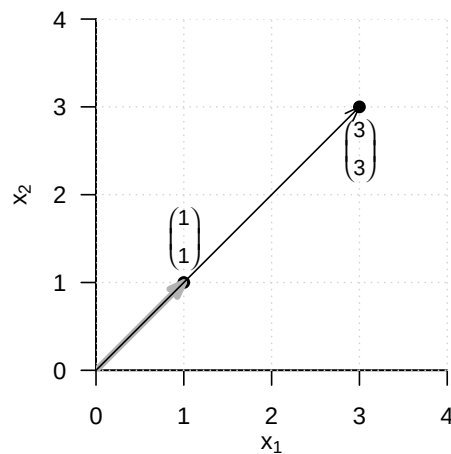
Definition 8.4 (Skalarprodukt auf $\mathbb{R}^m$). Das *Skalarprodukt auf $\mathbb{R}^m$* ist definiert als die Abbildung

$$\langle \rangle : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}, (x, y) \mapsto \langle (x, y) \rangle := \langle x, y \rangle := \sum_{i=1}^m x_i y_i. \quad (8.9)$$

•

Das Skalarprodukt heißt Skalarprodukt, weil es einen Skalar ergibt, nicht etwa, weil mit Skalaren multipliziert wird. Das Skalarprodukt steht in enger Beziehung zum Matrixprodukt, wie wir an späterer Stelle sehen werden. Wir betrachten zunächst ein Beispiel und seine Implementation in **R**.

Beispiel

Abbildung 8.4. Skalarmultiplikation in $\mathbb{R}^2$

Es seien

$$x := \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \text{ und } y := \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} \quad (8.10)$$

Dann ergibt sich

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + x_3 y_3 = 1 \cdot 2 + 2 \cdot 0 + 3 \cdot 1 = 2 + 0 + 3 = 5. \quad (8.11)$$

In $\mathbf{R}$ gibt es verschiedene Möglichkeiten, ein Skalarprodukt auszuwerten. Wir führen zwei von ihnen für das gegebene Beispiel untenstehend auf.

```
# Vektordefinitionen
x = matrix(c(1,2,3), nrow = 3)
y = matrix(c(2,0,1), nrow = 3)

# Skalarprodukt mithilfe von R's komponentenweiser Multiplikation und sum() Funktion
sum(x*y)
```

```
[1] 5
```

```
# Skalarprodukt mithilfe von R's Matrixtransposition und -multiplikation
t(x) %*% y
```

```
      [,1]
[1,]      5
```

Mithilfe des Skalarprodukts kann der Begriff des reellen Vektorraums zum Begriff des *reellen kanonischen Euklidischen Vektorraums* erweitert werden.

Definition 8.5 (Euklidischer Vektorraum). Das Tupel $((\mathbb{R}^m, +, \cdot), \langle \rangle)$ aus dem reellen Vektorraum $(\mathbb{R}^m, +, \cdot)$ und dem Skalarprodukt $\langle \rangle$ auf $\mathbb{R}^m$ heißt *reeller kanonischer Euklidischer Vektorraum*.

•

Generell heißt jedes Tupel aus einem Vektorraum und einem Skalarprodukt “Euklidischer Vektorraum”. Informell sprechen wir aber oft auch einfach von $\mathbb{R}^m$ als “Euklidischer Vektorraum” und insbesondere bei $((\mathbb{R}^m, +, \cdot), \langle \rangle)$ vom “Euklidischen Vektorraum”. Ein Euklidischer Vektorraum ist ein Vektorraum mit geometrischer Struktur, die durch das Skalarprodukt induziert wird. Insbesondere bekommen im Euklidischen Vektorraum nun die geometrischen Begriffe von *Länge*, *Abstand* und *Winkel* eine Bedeutung. Wir definieren sie wie folgt.

Definition 8.6. $((\mathbb{R}^m, +, \cdot), \langle \rangle)$ sei der Euklidische Vektorraum.

(1) Die *Länge* eines Vektors $x \in \mathbb{R}^m$ ist definiert als

$$\|x\| := \sqrt{\langle x, x \rangle}. \quad (8.12)$$

(2) Der *Abstand* zweier Vektoren $x, y \in \mathbb{R}^m$ ist definiert als

$$d(x, y) := \|x - y\|. \quad (8.13)$$

(3) Der *Winkel* α zwischen zwei Vektoren $x, y \in \mathbb{R}^m$ mit $x, y \neq 0$ ist definiert durch

$$0 \leq \alpha \leq \pi \text{ und } \cos \alpha := \frac{\langle x, y \rangle}{\|x\| \|y\|} \quad (8.14)$$

•

Die Länge $\|x\|$ eines Vektors $x \in \mathbb{R}^m$ heißt auch *Euklidische Norm von x* oder ℓ_2 -*Norm von x* oder einfach *Norm von x* . Sie wird häufig auch mit $\|x\|_2$ bezeichnet. Wir betrachten drei Beispiele für die Bestimmung der Länge eines Vektors und ihre entsprechende **R** Implementation. Wir veranschaulichen diese Beispiele in Abbildung 8.5.

Beispiel (1)

$$\left\| \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right\| = \sqrt{\left\langle \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right\rangle} = \sqrt{2^2 + 0^2} = \sqrt{4} = 2.00 \quad (8.15)$$

```
norm(matrix(c(2,0),nrow = 2), type = "2")      # Vektorlänge = l_2 Norm
```

```
[1] 2
```

Beispiel (2)

$$\left\| \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\| = \sqrt{\left\langle \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\rangle} = \sqrt{2^2 + 2^2} = \sqrt{8} \approx 2.83 \quad (8.16)$$

```
norm(matrix(c(2,2),nrow = 2), type = "2")      # Vektorlänge = l_2 Norm
```

```
[1] 2.828427
```

Beispiel (3)

$$\left\| \begin{pmatrix} 2 \\ 4 \end{pmatrix} \right\| = \sqrt{\left\langle \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix} \right\rangle} = \sqrt{2^2 + 4^2} = \sqrt{20} \approx 4.47 \quad (8.17)$$

```
norm(matrix(c(2,4),nrow = 2), type = "2")           # Vektorlänge = l_2 Norm
```

```
[1] 4.472136
```

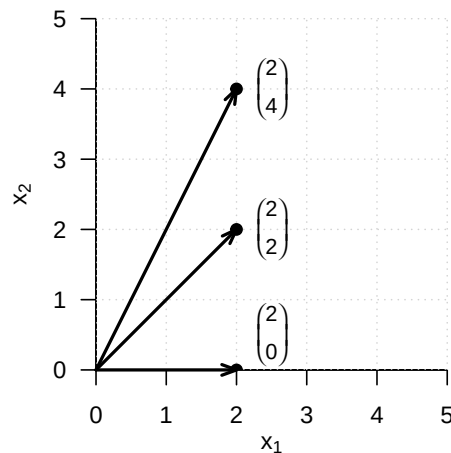


Abbildung 8.5. Vektorlänge in $\mathbb{R}^2$

Für den Abstand $d(x, y)$ zweier Vektoren $x, y \in \mathbb{R}^m$ halten wir ohne Beweis fest, dass er zum einen nicht-negativ und symmetrisch ist, also dass

$$d(x, y) \geq 0, d(x, x) = 0 \text{ und } d(x, y) = d(y, x) \quad (8.18)$$

gelten. Zudem erfüllt $d(x, y)$ die sogenannte *Dreiecksungleichung*, die besagt, dass die direkte Wegstrecke zwischen zwei Punkten im Raum immer kürzer ist als eine indirekte Wegstrecke über einen dritten Punkt,

$$d(x, y) \leq d(x, z) + d(z, y). \quad (8.19)$$

Damit erfüllt $d(x, y)$ wichtige Aspekte der räumlichen Anschauung. Wir geben zwei Beispiele für die Bestimmung von Abständen von Vektoren in $\mathbb{R}^2$, die wir in [Abbildung 8.6](#) visualisieren.

Beispiel (1)

$$d\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right) = \left\| \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\| = \left\| \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right\| = \sqrt{(-1)^2 + (-1)^2} = \sqrt{2} \approx 1.41 \quad (8.20)$$

```
norm(matrix(c(1,1),nrow = 2) - matrix(c(2,2),nrow = 2), type = "2")
```

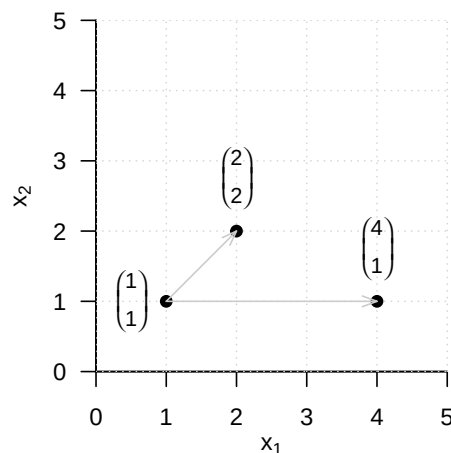
```
[1] 1.414214
```

Beispiel (2)

$$d\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 1 \end{pmatrix}\right) = \left\| \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 4 \\ 1 \end{pmatrix} \right\| = \left\| \begin{pmatrix} -3 \\ 0 \end{pmatrix} \right\| = \sqrt{(-3)^2 + 0^2} = \sqrt{9} = 3 \quad (8.21)$$

```
norm(matrix(c(1,1),nrow = 2) - matrix(c(4,1),nrow = 2), type = "2")
```

[1] 3

**Abbildung 8.6.** Vektorabstände in $\mathbb{R}^2$

Schließlich halten wir fest, dass für die Berechnung des Winkels zwischen zwei Vektoren anhand obiger Definition gilt, dass die Kosinusfunktion $\cos$ auf $[0, \pi]$ bijektiv, also invertierbar mit der Umkehrfunktion $\arccos$, der Arkuskosinusfunktion, ist. Auch für den Begriff des Winkels wollen wir zwei Beispiele betrachten. Man beachte dabei insbesondere, dass die Definition 8.6 den Winkel in Radians angibt. Für eine Angabe in Grad ist eine entsprechende Umrechnung erforderlich.

Beispiel (1)

$$\arccos\left(\frac{\left\langle \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 \\ 3 \end{pmatrix} \right\rangle}{\left\| \begin{pmatrix} 3 \\ 0 \end{pmatrix} \right\| \left\| \begin{pmatrix} 3 \\ 3 \end{pmatrix} \right\|}\right) = \arccos\left(\frac{3 \cdot 3 + 3 \cdot 0}{\sqrt{3^2 + 0^2} \cdot \sqrt{3^2 + 3^2}}\right) = \arccos\left(\frac{9}{3 \cdot \sqrt{18}}\right) = \frac{\pi}{4} \approx 0.785 \quad (8.22)$$

Die Umrechnung in Grad ergibt dann

$$0.785 \cdot \frac{180^\circ}{\pi} = 45^\circ \quad (8.23)$$

In **R** implementiert man dies wie folgt.

```
x = matrix(c(3,0), nrow = 2)           # Vektor 1
y = matrix(c(0,3), nrow = 2)           # Vektor 2
w = acos(sum(x*y)/(sqrt(sum(x*x))*sqrt(sum(y*y)))) * 180/pi # Winkel in Grad
print(w)
```

[1] 45

Beispiel (2)

$$\alpha = \arccos \left(\frac{\left\langle \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \end{pmatrix} \right\rangle}{\left\| \begin{pmatrix} 3 \\ 0 \end{pmatrix} \right\| \left\| \begin{pmatrix} 0 \\ 3 \end{pmatrix} \right\|} \right) = \arccos \left(\frac{3 \cdot 0 + 0 \cdot 3}{\sqrt{3^2 + 0^2} \cdot \sqrt{0^2 + 3^2}} \right) = \arccos \left(\frac{0}{3 \cdot 3} \right) = \frac{\pi}{2} \approx 1.57 \quad (8.24)$$

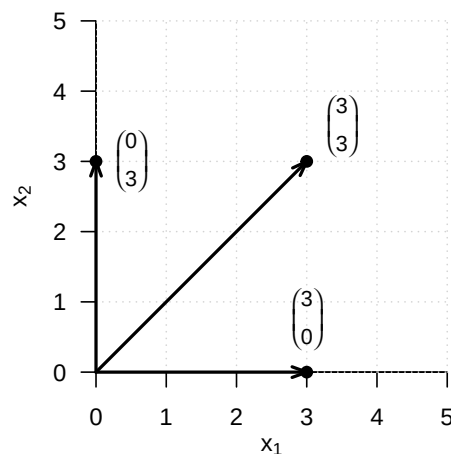
Die Umrechnung in Grad ergibt dann

$$\frac{\pi}{2} \cdot \frac{180^\circ}{\pi} = 90^\circ \quad (8.25)$$

Die entsprechende **R** Implementation lautet wie folgt.

```
x = matrix(c(3,0), nrow = 2)           # Vektor 1
y = matrix(c(0,3), nrow = 2)           # Vektor 2
w = acos(sum(x*y)/(sqrt(sum(x*x))*sqrt(sum(y*y)))) * 180/pi # Winkel in Grad
print(w)
```

[1] 90

Abbildung 8.7. Winkel in $\mathbb{R}^2$

Die Tatsache, dass zwei Vektoren einen rechten Winkel bilden können, also gewissermaßen maximal nicht-parallel sein können, ist ein wichtiges geometrisches Prinzip und wird deshalb mit folgender Definition speziell ausgezeichnet.

Definition 8.7 (Orthogonalität und Orthonormalität von Vektoren). $((\mathbb{R}^m, +, \cdot), \langle \rangle)$ sei der Euklidische Vektorraum.

(1) Zwei Vektoren $x, y \in \mathbb{R}^m$ heißen *orthogonal*, wenn gilt, dass

$$\langle x, y \rangle = 0 \quad (8.26)$$

(2) Zwei Vektoren $x, y \in \mathbb{R}^m$ heißen *orthonormal*, wenn gilt, dass

$$\langle x, y \rangle = 0 \text{ und } \|x\| = \|y\| = 1. \quad (8.27)$$

•

Für orthogonale und orthonormale Vektoren gilt also insbesondere auch

$$\cos \alpha = \frac{\langle x, y \rangle}{\|x\| \|y\|} = \frac{0}{\|x\| \|y\|} = 0, \quad (8.28)$$

also

$$\alpha = \frac{\pi}{2} = 90^\circ. \quad (8.29)$$

8.3. Lineare Unabhängigkeit

In diesem Abschnitt führen wir den Begriff der *linearen Unabhängigkeit* von Vektoren ein. Wir definieren dazu zunächst den Begriff der *Linearkombination* von Vektoren.

Definition 8.8 (Linearkombination). $\{v_1, v_2, \dots, v_k\}$ sei eine Menge von k Vektoren eines Vektorraums V und $a_1, a_2, \dots, a_k$ seien Skalare. Dann ist die *Linearkombination* der Vektoren in $\{v_1, v_2, \dots, v_k\}$ mit den *Koeffizienten* $a_1, a_2, \dots, a_k$ definiert als der Vektor

$$w := \sum_{i=1}^k a_i v_i \in V. \quad (8.30)$$

•

Beispiel

Es seien

$$v_1 := \begin{pmatrix} 2 \\ 1 \end{pmatrix}, v_2 := \begin{pmatrix} 1 \\ 1 \end{pmatrix}, v_3 := \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ und } a_1 := 2, a_2 := 3, a_3 := 0. \quad (8.31)$$

Dann ergibt sich die Linearkombination von v_1, v_2, v_3 mit den Koeffizienten a_1, a_2, a_3 zu

$$\begin{aligned} w &= a_1 v_1 + a_2 v_2 + a_3 v_3 \\ &= 2 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} + 3 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0 \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 7 \\ 5 \end{pmatrix}. \end{aligned} \quad (8.32)$$

Basierend auf dem Begriff der Linearkombination kann man nun den Begriff der *Linearen Unabhängigkeit* von Vektoren definieren.

Definition 8.9 (Lineare Unabhängigkeit). V sei ein Vektorraum. Eine Menge $W := \{w_1, w_2, \dots, w_k\}$ von Vektoren in V heißt *linear unabhängig*, wenn die einzige Repräsentation des Nullelements $0 \in V$ durch eine Linearkombination der $w \in W$ die sogenannte *triviale Repräsentation*

$$0 = a_1 w_1 + a_2 w_2 + \dots + a_k w_k \text{ mit } a_1 = a_2 = \dots = a_k = 0 \quad (8.33)$$

ist. Wenn die Menge W nicht linear unabhängig ist, dann heißt sie *linear abhängig*. •

Um zu prüfen, ob eine gegebene Menge von Vektoren linear abhängig oder unabhängig ist, muss man prinzipiell für jede mögliche Linearkombination der gegebenen Vektoren, ob sie Null ist. Theorem 8.2 und Theorem 8.3 zeigen, wie dies für zwei bzw. endliche viele Vektoren auch mit weniger Aufwand gelingen kann.

Theorem 8.2 (Lineare Abhängigkeit von zwei Vektoren). V sei ein Vektorraum. Zwei Vektoren $v_1, v_2 \in V$ sind linear abhängig, wenn einer der Vektoren ein skalares Vielfaches des anderen Vektors ist. ◦

Beweis. v_1 sei ein skalares Vielfaches von v_2 , also

$$v_1 = \lambda v_2 \text{ mit } \lambda \neq 0. \quad (8.34)$$

Dann gilt

$$v_1 - \lambda v_2 = 0. \quad (8.35)$$

Dies aber entspricht der Linearkombination

$$a_1 v_1 + a_2 v_2 = 0 \quad (8.36)$$

mit $a_1 = 1 \neq 0$ und $a_2 = -\lambda \neq 0$. Es gibt also eine Linearkombination des Nullelementes, die nicht die triviale Repräsentation ist, und damit sind v_1 und v_2 nicht linear unabhängig. □

Theorem 8.3 (Lineare Abhängigkeit einer Menge von Vektoren). V sei ein Vektorraum und $w_1, \dots, w_k \in V$ sei eine Menge von Vektoren in V . Wenn einer der Vektoren w_i mit $i = 1, \dots, k$ eine Linearkombination der anderen Vektoren ist, dann ist die Menge der Vektoren linear abhängig. ◦

Beweis. Die Vektoren $w_1, \dots, w_k$ sind genau dann linear abhängig, wenn gilt, dass $\sum_{i=1}^k a_i w_i = 0$ mit mindestens einem $a_i \neq 0$. Es sei also zum Beispiel $a_j \neq 0$. Dann gilt

$$0 = \sum_{i=1}^k a_i w_i = \sum_{i=1, i \neq j}^k a_i w_i + a_j w_j \quad (8.37)$$

Also folgt

$$a_j w_j = - \sum_{i=1, i \neq j}^k a_i w_i \quad (8.38)$$

und damit

$$w_j = -a_j^{-1} \sum_{i=1, i \neq j}^k a_i w_i = - \sum_{i=1, i \neq j}^k (a_j^{-1} a_i) w_i \quad (8.39)$$

Also ist w_j eine Linearkombination der w_i für $i = 1, \dots, k$ mit $i \neq j$. □

8.4. Vektorraumbasen

In diesem Abschnitt wollen wir den Begriff der *Vektorraumbasis* einführen. Eine Basis eines Vektorraums ist eine Untermenge von Vektoren des Vektorraums, die zur Darstellung aller Vektoren des Vektorraums genutzt werden kann. Im Sinne der linearen Kombination von Vektoren enthält also eine Vektorraumbasis alle nötige Information zur Konstruktion des entsprechenden Vektorraums. Allerdings ist eine Vektorraumbasis in der Regel nicht eindeutig und die viele Vektorräume haben in der Tat unendlich viele Basen. Die folgenden Definition sagt zunächst aus, wie aus einer beschränkten Anzahl von Vektoren mithilfe von Linearkombinationen unendlich viele Vektoren gebildet werden können.

Definition 8.10 (Lineare Hülle und Aufspannen). V sei ein Vektorraum und es sei $W := \{w_1, \dots, w_k\} \subset V$. Dann ist die *lineare Hülle* von W definiert als die Menge aller Linearkombinationen der Elemente von W ,

$$\text{Span}(W) := \left\{ \sum_{i=1}^k a_i w_i \mid a_1, \dots, a_k \text{ sind skalare Koeffizienten} \right\} \quad (8.40)$$

Man sagt, dass eine Menge von Vektoren $W \subseteq V$ *einen Vektorraum V aufspannt*, wenn jedes $v \in V$ als eine Linearkombination von Vektoren in W geschrieben werden kann.

•

Wir definieren nun den Begriff der *Basis* eines Vektorraums.

Definition 8.11 (Basis). V sei ein Vektorraum und es sei $B \subseteq V$. B heißt eine *Basis* von V , wenn

- (1) die Vektoren in B linear unabhängig sind und
- (2) die Vektoren in B den Vektorraum V aufspannen.

•

Basen von Vektorräumen haben folgende wichtige Eigenschaften.

Theorem 8.4 (Eigenschaften von Basen).

- (1) *Alle Basen eines Vektorraums beinhalten die gleiche Anzahl von Vektoren.*
- (2) *Jede Menge von m linear unabhängigen Vektoren ist Basis eines m -dimensionalen Vektorraums.*

◦

Für einen Beweis dieses sehr tiefen Theorems verweisen wir auf die weiterführende Literatur. Die mit obigem Theorem benannte eindeutige Anzahl der Vektoren einer Basis eines Vektorraums heißt die *Dimension des Vektorraums*. Da es in der Regel unendlich viele Mengen von m linear unabhängigen Vektoren in einem Vektorraum gibt, haben Vektorräume in der Regel unendlich viele Basen.

Betrachtet man nun einen einzelnen Vektor in einem Vektorraum, so kann man sich fragen, wie man diesen mithilfe einer Vektorraumbasis darstellen kann. Dies führt auf folgende Begriffsbildungen.

Definition 8.12 (Basisdarstellung und Koordinaten). $B := \{b_1, \dots, b_m\}$ sei eine Basis eines m -dimensionalen Vektorraumes V und es sei $v \in V$. Dann heißt die Linearkombination

$$v = \sum_{i=1}^m c_i b_i \quad (8.41)$$

die *Darstellung von v bezüglich der Basis B* und die Koeffizienten $c_1, \dots, c_m$ heißen die *Koordinaten von v bezüglich der Basis B* .

•

Bei fester Basis sind auch die Koordinaten eines Vektors bezüglich dieser Basis fest und eindeutig. Dies ist die Aussage folgenden Theorems.

Theorem 8.5 (Eindeutigkeit der Basisdarstellung). *Die Basisdarstellung eines $v \in V$ bezüglich einer Basis B ist eindeutig.*

◦

Beweis. Ohne Beschränkung der Allgemeinheit nehmen wir an, dass der Vektorraum von Dimension m ist. Nehmen wir an, dass zwei Darstellungen von v bezüglich der Basis B existieren, also dass

$$\begin{aligned} v &= a_1 b_1 + \dots + a_m b_m \\ v &= c_1 b_1 + \dots + c_m b_m \end{aligned} \quad (8.42)$$

Subtraktion der unteren von der oberen Gleichung ergibt

$$0 = (a_1 - c_1)b_1 + \dots + (a_m - c_m)b_m \quad (8.43)$$

Weil die $b_1, \dots, b_m$ linear unabhängig sind, gilt aber, dass $(a_i - c_i) = 0$ für alle $i = 1, \dots, m$ und somit sind die beiden Darstellungen von v bezüglich der Basis B identisch. $\square$

Zum Abschluss dieses Abschnitts wollen wir eine spezielle Basis des reellen Vektorraums betrachten.

Definition 8.13 (Orthonormalbasis von $\mathbb{R}^m$). Eine Menge von m Vektoren $v_1, \dots, v_m \in \mathbb{R}^m$ heißt *Orthonormalbasis* von $\mathbb{R}^m$, wenn $v_1, \dots, v_m$ jeweils die Länge 1 haben und wechselseitig orthogonal sind, also wenn

$$\langle v_i, v_j \rangle = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases} \quad (8.44)$$

•

Wir wollen zunächst ein Beispiel für eine Orthonormalbasis betrachten.

Beispiel (1)

Es ist

$$B_1 := \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \quad (8.45)$$

eine Orthonormalbasis von $\mathbb{R}^2$, denn B_1 besteht aus zwei Vektoren und es gelten

$$\left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\rangle = 1 \cdot 1 + 0 \cdot 0 = 1 + 0 = 1 \quad (8.46)$$

sowie

$$\left\langle \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\rangle = 0 \cdot 0 + 1 \cdot 1 = 0 + 1 = 1 \quad (8.47)$$

und

$$\left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\rangle = 1 \cdot 0 + 0 \cdot 1 = 0 + 0 = 0 \quad (8.48)$$

Für allgemeine reelle Vektorräume werden Basen der Form von B_1 mit dem Begriff der *kanonischen Basis* speziell ausgezeichnet.

Definition 8.14 (Kanonische Basis und kanonische Einheitsvektoren). Die Orthonormalbasis

$$B := \{e_1, \dots, e_m \mid e_{i_j} = 1 \text{ für } i = j \text{ und } e_{i_j} = 0 \text{ für } i \neq j\} \subset \mathbb{R}^m \quad (8.49)$$

heißt die *kanonische Basis* von $\mathbb{R}^m$ und die e_{i_j} heißen *kanonische Einheitsvektoren*.

•

B_1 aus Beispiel (1) ist also die kanonische Basis von $\mathbb{R}^2$.

Die kanonische Basis von $\mathbb{R}^3$ ist

$$B := \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}. \quad (8.50)$$

Allerdings gibt es auch nicht kanonische Orthonormalbasen. Dazu betrachten wir ein weiteres Beispiel

Beispiel (2)

Es ist auch

$$B_2 := \left\{ \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\} \quad (8.51)$$

eine Orthonormalbasis von $\mathbb{R}^2$, denn B_2 besteht aus zwei Vektoren und es gelten

$$\left\langle \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} + \frac{1}{2} = 1, \quad (8.52)$$

sowie

$$\left\langle \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = \left(-\frac{1}{\sqrt{2}}\right) \cdot \left(-\frac{1}{\sqrt{2}}\right) + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} + \frac{1}{2} = 1 \quad (8.53)$$

und

$$\left\langle \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = -\frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = -\frac{1}{2} + \frac{1}{2} = 0 \quad (8.54)$$

Wir visualisieren die beiden Orthonormalbasen B_1 und B_2 von $\mathbb{R}^2$ in Abbildung 8.8.

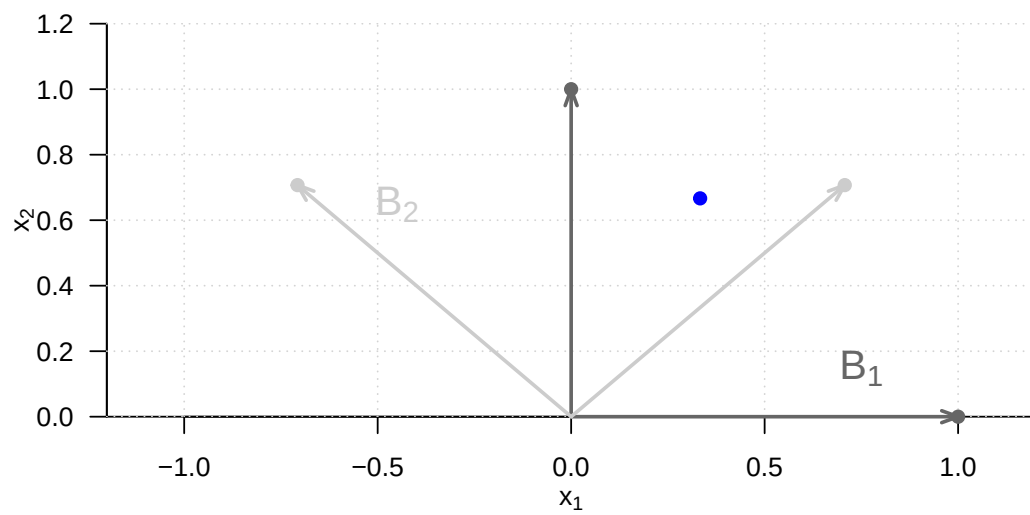


Abbildung 8.8. Zwei Basen von $\mathbb{R}^2$

9. Matrizen

Matrizen sind die Worte der Sprache der modernen Datenanalyse. Ein Verständnis moderner datenanalytischer Verfahren und ihrer Implementation ist ohne ein Grundverständnis des Matrixbegriffs und ein Wissen um die grundlegenden Matrixoperationen nicht möglich. Matrizen können dabei sehr unterschiedliche Rollen spielen. So können Matrizen zum Beispiel Daten, experimentelle Designs und Modellparameter repräsentieren. Im Kontext der Linearen Algebra dienen Matrizen zur Repräsentation linearer Abbildungen und von Vektorräumen, hier werden Vektoren dann als spezielle Matrizen aufgefasst.

In diesem Kapitel geben wir eine Einführung zum Umgang mit Matrizen, wobei wir auf abstrakte Begrifflichkeiten der Linearen Algebra im Wesentlichen verzichten. Wir führen zunächst den Matrixbegriff ein und diskutieren dann mit der Matrixaddition, Matrixsubtraktion, Skalarmultiplikation und der Matrixtransposition erste grundlegende Matrixoperationen (Kapitel 9.1 und Kapitel 9.2). Wir führen dann die zentralen Begriffe der Matrixmultiplikation und der Matrixinversion ein (Kapitel 9.3 und Kapitel 9.4). Mit der Matrixdeterminante diskutieren wir dann in Kapitel 9.5 eine erste Maßzahl zur Beschreibung von Matrizen. Wir schließen in Kapitel 9.6 mit einer Übersicht zu besonders häufig auftretenden Matrizen.

9.1. Definition

Wir beginnen mit der Definition einer Matrix.

Definition 9.1. Eine Matrix ist eine rechteckige Anordnung von Zahlen, die wie folgt bezeichnet wird

$$A := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} := (a_{ij})_{1 \leq i \leq n, 1 \leq j \leq m}. \quad (9.1)$$

•

Matrizen bestehen aus *Zeilen* (*rows*) und *Spalten* (*columns*). Die Matrixeinträge a_{ij} werden mit einem *Zeilenindex* i und einem *Spaltenindex* j indiziert. Zum Beispiel gilt für

$$A := \begin{pmatrix} 2 & 7 & 5 & 2 \\ 8 & 2 & 5 & 6 \\ 6 & 4 & 0 & 9 \\ 9 & 2 & 1 & 2 \end{pmatrix}, \quad (9.2)$$

dass $a_{32} = 4$. Die *Größe* oder *Dimension* einer Matrix ergibt sich aus der Anzahl ihrer Zeilen $n \in \mathbb{N}$ und Spalten $m \in \mathbb{N}$. Matrizen mit $n = m$ heißen *quadratische Matrizen*.

In der Folge benötigen wir nur Matrizen mit reellen Einträgen, also $a_{ij} \in \mathbb{R}$ für alle $i = 1, \dots, n$ und $j = 1, \dots, m$. Wir nennen die Matrizen mit reellen Einträge *reelle Matrizen* und bezeichnen die Menge der reellen Matrizen mit n Zeilen und m Spalten mit $\mathbb{R}^{n \times m}$. An dem Ausdruck

$$A \in \mathbb{R}^{n \times m} \quad (9.3)$$

können wir also ablesen, dass A eine reelle Matrix mit n Zeilen und m Spalten ist. Wir identifizieren dabei die Menge $\mathbb{R}^{1 \times 1}$ mit der Menge $\mathbb{R}$, die Menge $\mathbb{R}^{n \times 1}$ mit der Menge $\mathbb{R}^n$. Reelle Matrizen mit einer Spalte und n Zeilen entsprechen also n -dimensionalen reellen Vektoren und reelle Matrizen mit einer Spalte und einer Zeile entsprechen reellen Zahlen.

Definition von Matrizen in R

In **R** werden Matrizen definiert, indem **R** Vektoren mithilfe der `matrix()` Funktion in die Repräsentation einer mathematischen Matrix transformiert werden. Die Einträge eines **R** Vektors werden dabei anhand der spezifizierten Zeilenanzahl `nrow` anhand ihrer Gesamtanzahl auf die Matrix verteilt. Wollen wir beispielsweise die Matrix

$$A := \begin{pmatrix} 2 & 3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \quad (9.4)$$

in **R** definieren, so ergibt sich

```
# Spaltenweise Definition von A (R default)
A = matrix(c(2,1,3,6,0,5), nrow = 2)
print(A)
```

```
      [,1] [,2] [,3]
[1,]    2    3    0
[2,]    1    6    5
```

R folgt hier per default einer sogenannten *column-major-order*, das heißt, die Elemente des **R** Vektors `c(2,1,3,6,0,5)` werden der Reihe nach von oben nach unten in die Spalten der Matrix von links nach rechts überführt. Einen etwas klareren Zusammenhang zwischen dem visuellen Layout des **R** Codes und der resultierenden Matrix erhält man, indem man den **R** Vektor mithilfe von Zeilenumbrüchen anhand des intendierten Matrixlayouts formatiert und dann die *column-major-order* mithilfe des Arguments `byrow = TRUE` zu einer *row-major-order* umstellt. Es wird dann zunächst die erste Zeile der Matrix von links nach rechts mit den Elementen des **R** Vektors gefüllt wird und dann die zweite Zeile usw. bis alle Elemente des Vektors auf die Matrix verteilt sind.

```
# Reihenweise Definition von A (R default)
A = matrix(c(2,3,0,
             1,6,5),
           nrow = 2,
           byrow = TRUE)
print(A)
```

```
      [,1] [,2] [,3]
[1,]    2    3    0
[2,]    1    6    5
```

```
# Zeilenweise Definition von B
B = matrix(c(4,1,0,
             -4,2,0),
           nrow = 2,
           byrow = TRUE)
print(B)
```

```
      [,1] [,2] [,3]
[1,]    4    1    0
[2,]   -4    2    0
```

9.2. Grundlegende Matrixoperationen

Man kann mit Matrizen rechnen. Dabei sind folgende Matrixoperationen grundlegend:

- Die Addition von Matrizen gleicher Größe, genannt *Matrixaddition*,
- die Subtraktion von Matrizen gleicher Größe, genannt *Matrixsubtraktion*,
- die Multiplikation einer Matrix mit einem Skalar, genannt *Skalarmultiplikation*,
- das Vertauschen der Zeilen- und Spalten einer Matrix, genannt *Matrixtransposition*.

Wir führen diese Operationen in der Folge in Operatorform, also als Funktionen ein. Dies dient insbesondere dazu, bei jeder Operation mit Hilfe ihrer Definitionsmenge zu betonen, von welcher Art die Objekte der jeweiligen Operation sind und mithilfe ihrer Bildmenge zu betonen, von welcher Art das Resultat der jeweiligen Operation ist.

9.2.1. Matrixaddition

Definition 9.2. Es seien $A, B \in \mathbb{R}^{n \times m}$. Dann ist die *Addition* von A und B definiert als die Abbildung

$$+ : \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{n \times m}, (A, B) \mapsto +(A, B) := A + B \quad (9.5)$$

mit

$$\begin{aligned} A + B &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix} \\ &:= \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1m} + b_{1m} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2m} + b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \cdots & a_{nm} + b_{nm} \end{pmatrix}. \end{aligned} \quad (9.6)$$

•

Die Definition der Matrixaddition legt insbesondere fest, dass nur Matrizen gleicher Größe addiert werden können und dass die Operation der Matrixaddition elementweise definiert ist.

Beispiel

Es seien $A, B \in \mathbb{R}^{2 \times 3}$ definiert als

$$A := \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \text{ und } B := \begin{pmatrix} 4 & 1 & 0 \\ -4 & 2 & 0 \end{pmatrix}. \quad (9.7)$$

Da A und B gleich groß sind, können wir sie addieren

$$\begin{aligned} C = A + B &= \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} + \begin{pmatrix} 4 & 1 & 0 \\ -4 & 2 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 2+4 & -3+1 & 0+0 \\ 1-4 & 6+2 & 5+0 \end{pmatrix} \\ &= \begin{pmatrix} 6 & -2 & 0 \\ -3 & 8 & 5 \end{pmatrix}. \end{aligned} \quad (9.8)$$

In **R** führt man obige Rechnung wie folgt aus.

```
# Definition
A = matrix(c(2, -3, 0,
             1, 6, 5),
           nrow = 2,
           byrow = TRUE)
B = matrix(c(4, 1, 0,
             -4, 2, 0),
           nrow = 2,
           byrow = TRUE)

# Addition
C = A + B
print(C)
```

```
      [,1] [,2] [,3]
[1,]    6  -2    0
[2,]   -3    8    5
```

9.2.2. Matrixsubtraktion

Die Subtraktion von Matrizen gleicher Größe ist analog zur Addition definiert.

Definition 9.3 (Matrixsubtraktion). Es seien $A, B \in \mathbb{R}^{n \times m}$. Dann ist die *Subtraktion* von A und B definiert als die Abbildung

$$- : \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{n \times m}, (A, B) \mapsto -(A, B) := A - B \quad (9.9)$$

mit

$$\begin{aligned} A - B &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} - \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix} \\ &:= \begin{pmatrix} a_{11} - b_{11} & a_{12} - b_{12} & \cdots & a_{1m} - b_{1m} \\ a_{21} - b_{21} & a_{22} - b_{22} & \cdots & a_{2m} - b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} - b_{n1} & a_{n2} - b_{n2} & \cdots & a_{nm} - b_{nm} \end{pmatrix}. \end{aligned} \quad (9.10)$$

•

Wie bei der Matrixaddition legt die Definition der Matrixsubtraktion fest, dass nur Matrizen gleicher Größe voneinander subtrahiert werden können und dass die Subtraktion zweier gleich großer Matrizen elementweise definiert ist.

Beispiel

Wir können die im Beispiel zur Matrixaddition definierten Matrizen A und B auch voneinander subtrahieren,

$$\begin{aligned} D = A - B &= \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} - \begin{pmatrix} 4 & 1 & 0 \\ -4 & 2 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 2-4 & -3-1 & 0-0 \\ 1+4 & 6-2 & 5-0 \end{pmatrix} \\ &= \begin{pmatrix} -2 & -4 & 0 \\ 5 & 4 & 5 \end{pmatrix}. \end{aligned} \quad (9.11)$$

In **R** führt man diese Rechnung wie folgt aus.


```
# Subtraktion
D = A - B
print(D)
```

```
      [,1] [,2] [,3]
[1,]   -2   -4    0
[2,]    5    4    5
```

9.2.3. Skalarmultiplikation

Die *Skalarmultiplikation* einer Matrix bezeichnet die Multiplikation eines Skalars mit einer Matrix.

Definition 9.4 (Skalarmultiplikation). Es sei $c \in \mathbb{R}$ ein Skalar und $A \in \mathbb{R}^{n \times m}$. Dann ist die *Skalarmultiplikation* von c und A definiert als die Abbildung

$$\cdot : \mathbb{R} \times \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{n \times m}, (c, A) \mapsto \cdot(c, A) := cA \quad (9.12)$$

mit

$$cA = c \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} := \begin{pmatrix} ca_{11} & ca_{12} & \cdots & ca_{1m} \\ ca_{21} & ca_{22} & \cdots & ca_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{n1} & ca_{n2} & \cdots & ca_{nm} \end{pmatrix}. \quad (9.13)$$

•

Die Skalarmultiplikation ist mit dieser Definition also elementweise definiert.

Beispiel

Es seien $c := -3$ und $A \in \mathbb{R}^{4 \times 3}$ definiert als

$$A := \begin{pmatrix} 3 & 1 & 1 \\ 5 & 2 & 5 \\ 2 & 7 & 1 \\ 3 & 4 & 2 \end{pmatrix}. \quad (9.14)$$

Dann ergibt sich

$$B := cA = -3 \begin{pmatrix} 3 & 1 & 1 \\ 5 & 2 & 5 \\ 2 & 7 & 1 \\ 3 & 4 & 2 \end{pmatrix} = \begin{pmatrix} -3 \cdot 3 & -3 \cdot 1 & -3 \cdot 1 \\ -3 \cdot 5 & -3 \cdot 2 & -3 \cdot 5 \\ -3 \cdot 2 & -3 \cdot 7 & -3 \cdot 1 \\ -3 \cdot 3 & -3 \cdot 4 & -3 \cdot 2 \end{pmatrix} = \begin{pmatrix} -9 & -3 & -3 \\ -15 & -6 & -15 \\ -6 & -21 & -3 \\ -9 & -12 & -6 \end{pmatrix}. \quad (9.15)$$

In **R** führt man diese Skalarmultiplikation aus wie folgt.

```
# Definitionen
A = matrix(c(3,1,1,
             5,2,5,
             2,7,1,
             3,4,2),
           nrow = 4,
           byrow = TRUE)

c = -3

# Skalarmultiplikation
B = c*A
print(B)
```

	[,1]	[,2]	[,3]
[1,]	-9	-3	-3
[2,]	-15	-6	-15
[3,]	-6	-21	-3
[4,]	-9	-12	-6

Mithilfe der Definition von Matrixaddition und Skalarmultiplikation ist es möglich, einen Vektorraum zu definieren, dessen Elemente die reellen Matrizen sind. Insbesondere legt diese Definition auch die Rechenregeln beim Umgang mit Matrixaddition und Skalarmultiplikation fest.

Theorem 9.1 (Vektorraum der reellwertigen Matrizen). *Das Tripel $(\mathbb{R}^{n \times m}, +, \cdot)$ mit der oben definierten Matrixaddition und Skalarmultiplikation ist ein Vektorraum. Insbesondere gelten damit für $A, B, C \in \mathbb{R}^{n \times m}$ und $r, s, t \in \mathbb{R}$ folgende Rechenregeln:*

- (1) *Kommutativität der Addition: $A + B = B + A$.*
- (2) *Assoziativität der Addition: $(A + B) + C = A + (B + C)$.*
- (3) *Existenz eines neutralen Elements der Addition: $\exists 0 \in \mathbb{R}^{n \times m}$ mit $A + 0 = 0 + A = A$.*
- (4) *Existenz inverser Elemente der Addition: $\forall A \exists -A \in \mathbb{R}^{n \times m}$ mit $A + (-A) = 0$.*
- (5) *Existenz eines neutralen Elements der Skalarmultiplikation: $\exists 1 \in \mathbb{R}$ mit $1 \cdot A = A$.*
- (6) *Assoziativität der Skalarmultiplikation: $r \cdot (s \cdot t) = (r \cdot s) \cdot t$.*
- (7) *Distributivität hinsichtlich der Matrixaddition: $r \cdot (A + B) = r \cdot A + r \cdot B$.*
- (8) *Distributivität hinsichtlich der Skalaraddition: $(r + s) \cdot A = r \cdot A + s \cdot A$.*

◦

Wir verzichten auf einen Beweis, der sich mit einigem Notationsaufwand direkt aus dem elementweisen Charakter von Matrixaddition und Skalarmultiplikation sowie den aus dem Umgang mit den reellen Zahlen bekannten Rechenregeln ergibt. Das im Theorem erwähnte neutrale Element der Addition wird *Nullmatrix* genannt, wir werden dazu später eine allgemeine Notation einführen. Die inversen Elemente der Addition sind durch

$$-A := (-a_{ij})_{1 \leq i \leq n, 1 \leq j \leq m} \quad (9.16)$$

gegeben und erlauben es, die Matrixsubtraktion als Spezialfall der Matrixaddition zu betrachten.

9.2.4. Matrixtransposition

Eine weitere häufig auftretende grundlegende Matrixoperation ist das Vertauschen der Zeilen- und Spaltenanordnung einer Matrix, genannt *Matrixtransposition*.

Definition 9.5 (Matrixtransposition). Es sei $A \in \mathbb{R}^{n \times m}$. Dann ist die *Transposition* von A definiert als die Abbildung

$$\cdot^T : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{m \times n}, A \mapsto \cdot^T(A) := A^T \quad (9.17)$$

mit

$$A^T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}^T := \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{pmatrix}. \quad (9.18)$$

•

Für $A \in \mathbb{R}^{n \times m}$ gilt damit also immer $A^T \in \mathbb{R}^{m \times n}$. Weiterhin gelten folgende Rechenregeln der Matrixtransposition, wie man sich an Beispielen klar macht:

(1) Für $A \in \mathbb{R}^{1 \times 1}$ gilt

$$A^T = A. \quad (9.19)$$

(2) Es gilt

$$(A^T)^T = A. \quad (9.20)$$

(3) Es gilt

$$(a_{ii})_{1 \leq i \leq \min(n,m)} = (a_{ii})_{1 \leq i \leq \min(n,m)}^T. \quad (9.21)$$

Letztere Eigenschaft der Transposition besagt, dass die Elemente auf der Hauptdiagonalen einer Matrix bei Transposition unberührt bleiben.

Beispiel

Es sei $A \in \mathbb{R}^{2 \times 3}$ definiert durch

$$A := \begin{pmatrix} 2 & 3 & 0 \\ 1 & 6 & 5 \end{pmatrix}, \quad (9.22)$$

Dann gilt $A^T \in \mathbb{R}^{3 \times 2}$ und speziell

$$A^T := \begin{pmatrix} 2 & 1 \\ 3 & 6 \\ 0 & 5 \end{pmatrix}. \quad (9.23)$$

Weiterhin gilt offenbar $\min(m, n) = 2$ und folglich

$$(a_{11}) = (a_{11})^T \text{ und } (a_{22}) = (a_{22})^T. \quad (9.24)$$

In **R** führt man die Transposition einer Matrix wie folgt durch.

```
# Definition
A = matrix(c(2,3,0,
             1,6,5),
           nrow = 2,
           byrow = TRUE)
print(A)
```

```
      [,1] [,2] [,3]
[1,]    2    3    0
[2,]    1    6    5
```

```
# Transposition
AT = t(A)
print(AT)
```

```
      [,1] [,2]
[1,]    2    1
[2,]    3    6
[3,]    0    5
```

Schließlich gelten in der Verbindung mit der Matrixaddition, Matrixsubtraktion und der Skalarmultiplikation folgende Rechenregeln, wie man sich an Beispielen klar macht:

(1) Für $A, B \in \mathbb{R}^{n \times m}$ gilt

$$(A + B)^T = A^T + B^T. \quad (9.25)$$

(2) Für $A, B \in \mathbb{R}^{n \times m}$ gilt

$$(A - B)^T = A^T - B^T. \quad (9.26)$$

(3) Für $c \in \mathbb{R}$ und $A \in \mathbb{R}^{n \times m}$ gilt

$$(cA)^T = cA^T. \quad (9.27)$$

9.3. Matrixmultiplikation

Die Matrixmultiplikation ist die zentrale Operation beim Rechnen mit Matrizen. Sie ist definiert wie folgt.

Definition 9.6 (Matrixmultiplikation). Es seien $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{m \times k}$. Dann ist die *Matrixmultiplikation* von A und B definiert als die Abbildung

$$\cdot : \mathbb{R}^{n \times m} \times \mathbb{R}^{m \times k} \rightarrow \mathbb{R}^{n \times k}, (A, B) \mapsto \cdot(A, B) := AB \quad (9.28)$$

mit

$$\begin{aligned} AB &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1k} \\ b_{21} & b_{22} & \cdots & b_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mk} \end{pmatrix} \\ &:= \begin{pmatrix} \sum_{i=1}^m a_{1i}b_{i1} & \sum_{i=1}^m a_{1i}b_{i2} & \cdots & \sum_{i=1}^m a_{1i}b_{ik} \\ \sum_{i=1}^m a_{2i}b_{i1} & \sum_{i=1}^m a_{2i}b_{i2} & \cdots & \sum_{i=1}^m a_{2i}b_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^m a_{ni}b_{i1} & \sum_{i=1}^m a_{ni}b_{i2} & \cdots & \sum_{i=1}^m a_{ni}b_{ik} \end{pmatrix} \\ &= \left(\sum_{i=1}^m a_{ji}b_{il} \right)_{1 \leq j \leq n, 1 \leq l \leq k} \end{aligned} \quad (9.29)$$

•

Das Matrixprodukt AB ist also nur dann definiert, wenn A genau so viele Spalten hat wie B Zeilen hat. Informell gilt für die beteiligten Matrixgrößen dabei die Merkregel

$$(n \times m)(m \times k) = (n \times k). \quad (9.30)$$

Der Eintrag $(AB)_{ij}$ in AB entspricht der Summe der multiplizierten i ten Zeile von A und j ten Spalte von B . Zum Berechnen von $(AB)_{ij}$ geht man für $i = 1, \dots, n$ und $j = 1, \dots, k$ also in Gedanken wie folgt vor:

- (1) Man legt die Transposition der i ten Zeile von A über die j te Spalte von B .
- (2) Weil A genau m Spalten hat und B genau m Zeilen hat, gibt es dann zu jedem Element der Zeile aus A ein korrespondierendes Element in der Spalte von B .
- (3) Man multipliziert die korrespondierenden Elemente miteinander.
- (4) Die Summe dieser Produkte ist dann der Eintrag mit Index ij in AB .

Beispiel

$A \in \mathbb{R}^{2 \times 3}$ und $B \in \mathbb{R}^{3 \times 2}$ seien definiert als

$$A := \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \text{ und } B := \begin{pmatrix} 4 & 2 \\ -1 & 0 \\ 1 & 3 \end{pmatrix}. \quad (9.31)$$

Wir wollen $C := AB$ und $D := BA$ berechnen. Mit $n = 2, m = 3$ und $k = 2$ wissen wir schon, dass $C \in \mathbb{R}^{2 \times 2}$ und $D \in \mathbb{R}^{3 \times 3}$, weil

$$(2 \times 3)(3 \times 2) = (2 \times 2) \quad (9.32)$$

und

$$(3 \times 2)(2 \times 3) = (3 \times 3). \quad (9.33)$$

Es gilt hier also sicher $AB \neq BA$. Für C ergibt sich dann

$$\begin{aligned} C &= AB \\ &= \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \begin{pmatrix} 4 & 2 \\ -1 & 0 \\ 1 & 3 \end{pmatrix} \\ &= \begin{pmatrix} 2 \cdot 4 + (-3) \cdot (-1) + 0 \cdot 1 & 2 \cdot 2 + (-3) \cdot 0 + 0 \cdot 3 \\ 1 \cdot 4 + 6 \cdot (-1) + 5 \cdot 1 & 1 \cdot 2 + 6 \cdot 0 + 5 \cdot 3 \end{pmatrix} \\ &= \begin{pmatrix} 8 + 3 + 0 & 4 + 0 + 0 \\ 4 - 6 + 5 & 2 + 0 + 15 \end{pmatrix} \\ &= \begin{pmatrix} 11 & 4 \\ 3 & 17 \end{pmatrix}. \end{aligned} \quad (9.34)$$

In **R** nutzt man für die Matrixmultiplikation den `%%` Operator.

```
# Definitionen
A = matrix(c(2,-3,0,
            1, 6,5),
          nrow = 2,
          byrow = TRUE)
B = matrix(c( 4,2,
            -1,0,
            1,3),
          nrow = 3,
          byrow = TRUE)

# Matrixmultiplikation
C = A %% B
print(C)
```

```
      [,1] [,2]
[1,]   11    4
[2,]    3   17
```

Für D ergibt sich weiterhin

$$\begin{aligned} D &= BA \\ &= \begin{pmatrix} 4 & 2 \\ -1 & 0 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \\ &= \begin{pmatrix} 4 \cdot 2 + 2 \cdot 1 & 4 \cdot (-3) + 2 \cdot 6 & 4 \cdot 0 + 2 \cdot 5 \\ (-1) \cdot 2 + 0 \cdot 1 & (-1) \cdot (-3) + 0 \cdot 6 & (-1) \cdot 0 + 0 \cdot 5 \\ 1 \cdot 2 + 3 \cdot 1 & 1 \cdot (-3) + 3 \cdot 6 & 1 \cdot 0 + 3 \cdot 5 \end{pmatrix} \\ &= \begin{pmatrix} 8 + 2 & -12 + 12 & 0 + 5 \\ -2 + 0 & 3 + 0 & 0 + 0 \\ 2 + 3 & -3 + 18 & 0 + 15 \end{pmatrix} \\ &= \begin{pmatrix} 10 & 0 & 10 \\ -2 & 3 & 0 \\ 5 & 15 & 15 \end{pmatrix} \end{aligned} \quad (9.35)$$

In **R** überprüft man diese Rechnung wie folgt.

```
# Definitionen
A = matrix(c(2,-3,0,
            1, 6,5),
           nrow = 2,
           byrow = TRUE)
B = matrix(c( 4,2,
            -1,0,
            1,3),
           nrow = 3,
           byrow = TRUE)

# Matrixmultiplikation
D = B %*% A
print(D)
```

```
      [,1] [,2] [,3]
[1,]   10    0   10
[2,]   -2    3    0
[3,]    5   15   15
```

Ist allerdings eine Matrixmultiplikation aufgrund nicht-adäquater Matrizengrößen nicht definiert, so lässt sich diese auch nicht numerisch auswerten.

```
# Beispiel für eine undefinierte Matrixmultiplikation
E = t(A) %*% B      # (3 x 2)(3 x 2)
```

```
Error in t(A) %*% B: non-conformable arguments
```

Folgendes Theorem, das wir nicht beweisen wollen, stellt den Bezug zwischen dem Skalarprodukt zweier Vektoren und der Multiplikation zweier Matrizen her. Dieser ergibt sich im Wesentlichen durch die Identifikation von $\mathbb{R}^n$ und $\mathbb{R}^{n \times 1}$ und der Tatsache, dass nach Definition der Eintrag $(AB)_{ij}$ im Produkt von $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{m \times k}$ dem Vektorskalarprodukt der i ten Spalte von A^T und der j ten Spalte von B entspricht.

Theorem 9.2 (Matrixmultiplikation und Vektorskalarprodukt). *Es seien $x, y \in \mathbb{R}^n$. Dann gilt*

$$\langle x, y \rangle = x^T y. \quad (9.36)$$

Weiterhin seien für $A \in \mathbb{R}^{n \times m}$ für $i = 1, \dots, n$

$$\bar{a}_i := (a_{ji})_{1 \leq j \leq m} \in \mathbb{R}^m \quad (9.37)$$

die Spalten von A^T und für $B \in \mathbb{R}^{m \times k}$ für $i = 1, \dots, k$

$$\bar{b}_j := (b_{ij})_{1 \leq i \leq m} \in \mathbb{R}^m \quad (9.38)$$

die Spalten von B , also

$$A^T = (\bar{a}_1 \quad \bar{a}_2 \quad \dots \quad \bar{a}_n) \in \mathbb{R}^{m \times n} \text{ und } B = (\bar{b}_1 \quad \bar{b}_2 \quad \dots \quad \bar{b}_k) \in \mathbb{R}^{m \times k}. \quad (9.39)$$

Dann gilt

$$AB = (\langle \bar{a}_i, \bar{b}_j \rangle)_{1 \leq i \leq n, 1 \leq j \leq k}. \quad (9.40)$$

◦

9.3.1. Rechenregeln der Matrixmultiplikation

Im Folgenden stellen wir einige grundlegende Rechenregeln der Matrixmultiplikation, insbesondere auch in Kombination mit anderen Matrixoperationen zusammen.

Für Beweise der folgenden zwei Theoreme zur Assoziativität und Distributivität, die sich im Wesentlichen mit den entsprechenden Rechenregeln für Summen und Produkte der reellen Zahlen ergeben, verweisen wir auf die weiterführende Literatur.

Theorem 9.3 (Assoziativität). *Es seien $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{m \times k}$, $C \in \mathbb{R}^{k \times p}$ und $c \in \mathbb{R}$. Dann gelten*

(1) *Die Multiplikation von Matrizen ist assoziativ, es gilt*

$$A(BC) = (AB)C. \quad (9.41)$$

(2) *Die Kombination von Matrizenmultiplikation und Skalarmultiplikation ist assoziativ,*

$$c(AB) = (cA)B = A(cB). \quad (9.42)$$

◦

Die Assoziativität von Matrizenmultiplikation und Skalarmultiplikation erkennt man leicht bei Betrachtung des j , l ten Elements von $c(AB)$, $(cA)B$ und $A(cB)$ anhand von

$$c \left(\sum_{i=1}^m a_{ji} b_{il} \right) = \sum_{i=1}^m (c a_{ji}) b_{il} = \sum_{i=1}^m a_{ji} (c b_{il}). \quad (9.43)$$

Theorem 9.4 (Distributivität). *Es seien $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{m \times p}$. Dann gelten*

$$(A + B)C = AC + BC \quad (9.44)$$

und

$$C^T(A + B) = C^T A + C^T B \quad (9.45)$$

◦

Im Gegensatz zur Kommutativität der Multiplikation reeller Zahlen ist die Matrixmultiplikation im Allgemeinen nicht kommutativ.

Theorem 9.5 (Nichtkommutativität). *Es seien $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{m \times p}$. Dann gilt im Allgemeinen*

$$AB \neq BA. \quad (9.46)$$

◦

Beweis. Im Fall $p \neq n$ ist BA nicht definiert, wir betrachten also nur den Fall $p = n$. Wir zeigen durch Angabe eines Gegenbeispiels mit $A, B \in \mathbb{R}^{2 \times 2}$, dass im Allgemeinen $AB = BA$ nicht gilt. Es seien

$$A := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \text{ und } B := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \quad (9.47)$$

Dann gilt

$$AB = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = BA. \quad (9.48)$$

□

Theorem 9.6 (Kombination von Matrixmultiplikation und Transposition). *Es seien $A \in \mathbb{R}^{m \times n}$ und $B \in \mathbb{R}^{n \times k}$. Dann gilt*

$$(AB)^T = B^T A^T. \quad (9.49)$$

◦

Beweis. Ein Beweis ergibt sich wie folgt

$$\begin{aligned} (AB)^T &= \left(\left(\sum_{i=1}^m a_{ji} b_{il} \right)_{1 \leq j \leq n, 1 \leq l \leq k} \right)^T \\ &= \left(\sum_{i=1}^m a_{ij} b_{li} \right)_{1 \leq i \leq k, 1 \leq j \leq n} \\ &= \left(\sum_{i=1}^m b_{li} a_{ij} \right)_{1 \leq j \leq k, 1 \leq l \leq n} \\ &= B^T A^T. \end{aligned} \quad (9.50)$$

□

9.4. Matrixinversion

Um den Begriff der inversen Matrix zu motivieren, betrachten wir zunächst das Problem des *Lösens eines linearen Gleichungssystems*. Dazu seien $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$ und $b \in \mathbb{R}^n$ und es gelte

$$Ax = b. \quad (9.51)$$

A und b seien als bekannt vorausgesetzt, x sei unbekannt. Konkret seien beispielsweise

$$A := \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \text{ und } b := \begin{pmatrix} 5 \\ 11 \end{pmatrix}. \quad (9.52)$$

Dann liegt folgendes lineares Gleichungssystem mit zwei Gleichungen und zwei Unbekannten vor:

$$Ax = b \Leftrightarrow \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 11 \end{pmatrix} \Leftrightarrow \begin{array}{rcl} 1x_1 + 2x_2 & = & 5 \\ 3x_1 + 4x_2 & = & 11 \end{array} \quad (9.53)$$

Ziel des Lösens von linearen Gleichungssystemen ist bekanntlich, herauszufinden, für welche x das Gleichungssystem erfüllt ist. Um in diesem Kontext den Begriff der inversen Matrix von A einzuführen, vereinfachen wir die Situation weiter. Wir nehmen an, dass $A = a$ eine 1×1 Matrix, also ein Skalar, sei und ebenso x und b , dass wir also für $a, x, b \in \mathbb{R}$ die Gleichung

$$ax = b \quad (9.54)$$

haben. Um diese Gleichung nach x aufzulösen würde man natürlich beide Seiten der Gleichung mit dem *multiplikativem Inversen* von a multiplizieren, wobei das *multiplikative Inverse* von a den Wert bezeichnet, der mit a multipliziert 1 ergibt. Dieser ist bekanntlich durch

$$a^{-1} = \frac{1}{a} \quad (9.55)$$

gegeben. Dann würde gelten

$$ax = b \Leftrightarrow a^{-1}ax = a^{-1}b \Leftrightarrow 1 \cdot x = a^{-1}b \Leftrightarrow x = \frac{b}{a}. \quad (9.56)$$

Ganz konkret etwa

$$2x = 6 \Leftrightarrow 2^{-1}2x = 2^{-1}6 \Leftrightarrow \frac{1}{2}2x = \frac{1}{2}6 \Leftrightarrow x = 3. \quad (9.57)$$

Analog zu dem Fall, dass die Matrizen in $Ax = b$ allesamt Skalare sind, möchte man im Fall eines linearen Gleichungssystems beide Seiten der Gleichung mit dem *multiplikativen Inversen* A^{-1} von A multiplizieren können, sodass eine Gleichung der Form

$$A^{-1}A = "1". \quad (9.58)$$

resultiert. Dann hätte man nämlich

$$Ax = b \Leftrightarrow A^{-1}Ax = A^{-1}b \Leftrightarrow x = A^{-1}b. \quad (9.59)$$

Diese intuitive Idee des multiplikativen Inversen einer Matrix A wird im Folgenden unter dem Begriff der *inversen Matrix* formalisiert. Dazu benötigen wir zunächst den Begriff der *Einheitsmatrix*.

Definition 9.7 (Einheitsmatrix). Die Matrix

$$I_n := (a_{ij})_{1 \leq i \leq n, 1 \leq j \leq n} \in \mathbb{R}^{n \times n} := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \quad (9.60)$$

mit $a_{ij} = 1$ für $i = j$ und $a_{ij} = 0$ für $i \neq j$ heißt *n-dimensionale Einheitsmatrix*.

•

In **R** wird I_n mit dem Befehl `diag(n)` erzeugt. Die Einheitsmatrix ist für die Matrixmultiplikation das Analog zur 1 bei der Multiplikation reeller Zahlen. Das ist die Aussage folgenden Theorems.

Theorem 9.7 (Neutrales Element der Matrixmultiplikation). I_n ist das neutrale Element der Matrixmultiplikation, das heißt es gilt für $A \in \mathbb{R}^{n \times m}$, dass

$$I_n A = A \text{ und } A I_m = A. \quad (9.61)$$

◦

Beweis. Es sei $B = (b_{ij}) = I_n A \in \mathbb{R}^{n \times m}$. Dann gilt für alle $1 \leq i \leq n$ und alle $1 \leq j \leq m$

$$d_{ij} = 0 \cdot a_{1j} + 0 \cdot a_{2j} + \cdots + 0 \cdot a_{i-1,j} + 1 \cdot a_{ij} + \cdots + 0 \cdot a_{i+1,j} + 0 \cdot a_{nj} = a_{ij}. \quad (9.62)$$

Analog zeigt man dies für $A I_m$. □

Mit dem Begriff der Einheitsmatrix können wir jetzt die Begriffe der inversen Matrix und der invertierbaren Matrix definieren:

Definition 9.8 (Invertierbare Matrix und inverse Matrix). Eine quadratische Matrix $A \in \mathbb{R}^{n \times n}$ heißt *invertierbar*, wenn es eine quadratische Matrix $A^{-1} \in \mathbb{R}^{n \times n}$ gibt, so dass

$$A^{-1}A = AA^{-1} = I_n \quad (9.63)$$

ist. Die Matrix A^{-1} heißt die *inverse Matrix* von A .

•

Man beachte, dass sich die Begriffe der inversen Matrix und der Invertierbarkeit *nur* auf quadratische Matrizen beziehen. Insbesondere können quadratische Matrizen invertierbar sein, müssen es aber nicht sein (lineare Gleichungssysteme können also Lösungen haben, müssen es aber nicht). Nicht invertierbare Matrizen nennt man auch *singuläre* Matrizen, invertierbare Matrizen manchmal auch *nicht-singuläre* Matrizen. Schließlich beachte man, dass Definition 9.8 lediglich aussagt, was eine inverse Matrix ist, aber nicht wie man sie berechnet.

Beispiel für eine invertierbare Matrix

Die Matrix

$$A := \begin{pmatrix} 2.0 & 1.0 \\ 3.0 & 4.0 \end{pmatrix} \quad (9.64)$$

ist invertierbar und ihre inverse Matrix ist gegeben durch

$$A^{-1} = \begin{pmatrix} 0.8 & -0.2 \\ -0.6 & 0.4 \end{pmatrix}, \quad (9.65)$$

denn

$$\begin{pmatrix} 2.0 & 1.0 \\ 3.0 & 4.0 \end{pmatrix} \begin{pmatrix} 0.8 & -0.2 \\ -0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.8 & -0.2 \\ -0.6 & 0.4 \end{pmatrix} \begin{pmatrix} 2.0 & 1.0 \\ 3.0 & 4.0 \end{pmatrix}, \quad (9.66)$$

wovon man sich durch Nachrechnen überzeugt.

Beispiel für eine nicht-invertierbare Matrix

Die Matrix

$$B := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad (9.67)$$

ist nicht invertierbar, denn wäre B invertierbar, dann gäbe es

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad (9.68)$$

mit

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (9.69)$$

Das würde aber bedeuten, dass $0 = 1$ in $\mathbb{R}$ und das ist ein Widerspruch. Also kann B nicht invertierbar sein.

Zum Berechnen inverser Matrizen

2×2 bis etwa 5×5 Matrizen kann man prinzipiell per Hand invertieren, dazu stellt die Lineare Algebra verschiedene Verfahren bereit. Wir wollen hier auf eine Einführung in die Matrizeninvertierung per Hand verzichten, da in der Anwendung Matrizen standardmäßig

numerisch invertiert werden. Die numerische Matrixinversion ist dann auch ein großes Feld der Forschung zur Numerischen Mathematik, die eine Vielzahl von Algorithmen zu diesem Zweck bereitstellt. In **R** werden Matrizen per default mit der Funktion `solve()`, in Anlehnung an das Lösen linearer Gleichungssysteme, invertiert. Für das obige Beispiel einer invertierbaren Matrix ergibt sich dabei folgender **R** Code.

```
# Definition
A = matrix(c(2,1,
             3,4),
           nrow = 2,
           byrow = TRUE)

# Berechnen von A^{-1}
print(solve(A))
```

```
      [,1] [,2]
[1,]  0.8 -0.2
[2,] -0.6  0.4
```

```
# Überprüfen der Eigenschaften einer inversen Matrix
print(solve(A) %*% A)
```

```
      [,1] [,2]
[1,]    1    0
[2,]    0    1
```

```
# Bei der umgekehrten Berechnung ergeben sich kleine Rundungsfehler
print(A %*% solve(A))
```

```
      [,1]      [,2]
[1,]    1 -5.551115e-17
[2,]    0  1.000000e+00
```

Nicht-invertierbare Matrizen sind dabei natürlich auch numerisch nicht-invertierbar, wie folgende Fehlermeldung in **R** bezüglich obigen Beispiels einer nicht-invertierbaren Matrix demonstriert.

```
B = matrix(c(1,0,
             0,0),
           nrow = 2,
           byrow = 2)
solve(B)
```

```
Error in solve.default(B): Lapack routine dgesv: system is exactly singular: U[2,2] = 0
```

9.5. Determinanten

Die Determinante ist eine vielseitig einsetzbare Maßzahl einer quadratischen Matrix. Für das Verständnis der *Eigenanalyse* und der *Matrixzerlegung* ist der Begriff der Determinante im Kontext des *charakteristischen Polynoms* grundlegend.

Allgemein ist eine Determinante eine nichtlineare Abbildung der Form

$$|\cdot| : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}, A \mapsto |A|, \quad (9.70)$$

das heißt, eine Determinante ordnet einer quadratischen Matrix A die reelle Zahl $|A|$ zu. Die Zahl $|A|$ wird dabei rekursiv anhand folgender Definition bestimmt.

Definition 9.9 (Determinante). Für $A = (a_{ij})_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$ mit $n > 1$ sei $A_{ij} \in \mathbb{R}^{(n-1) \times (n-1)}$ die Matrix, die aus A durch Entfernen der i ten Zeile und der j ten Spalte entsteht. Dann heißt die Zahl

$$|A| := a_{11} \quad \text{für } n = 1 \quad (9.71)$$

$$|A| := \sum_{j=1}^n a_{1j}(-1)^{1+j} \det(A_{1j}) \quad \text{für } n > 1 \quad (9.72)$$

die *Determinante* von A .

•

Die Definition führt die Bestimmung der Determinante einer quadratischen Matrix also sukzessive durch Streichen von Zeilen und Spalten auf die Determinante einer 1×1 Matrix zurück, die durch ihr einziges Element gegeben ist. Für

$$A := \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \in \mathbb{R}^{3 \times 3} \quad (9.73)$$

ergeben sich dabei zum Beispiel folgende Matrizen der Form $A_{ij} \in \mathbb{R}^{(3-1) \times (3-1)}$:

$$A_{11} = \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix}, A_{12} = \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix}, A_{21} = \begin{pmatrix} 2 & 3 \\ 8 & 9 \end{pmatrix}, A_{22} = \begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix}. \quad (9.74)$$

Für die Berechnung der Determinanten von zwei- und dreidimensionalen quadratischen Matrizen gibt es direkte, nicht-rekursive Rechenregeln, die in folgendem Theorem festgehalten sind.

Theorem 9.8 (Determinanten von zwei- und dreidimensionalen Matrizen).

Es sei $A = (a_{ij})_{1 \leq i, j \leq 2} \in \mathbb{R}^{2 \times 2}$. Dann gilt

$$|A| = a_{11}a_{22} - a_{12}a_{21}. \quad (9.75)$$

Es sei $A = (a_{ij})_{1 \leq i, j \leq 3} \in \mathbb{R}^{3 \times 3}$. Dann gilt

$$|A| = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31}. \quad (9.76)$$

◦

Beweis. Für $A \in \mathbb{R}^{2 \times 2}$ gilt nach Definition

$$\begin{aligned} |A| &= \sum_{j=1}^2 a_{1j}(-1)^{1+j} |A_{1j}| \\ &= a_{11}(-1)^{1+1} |A_{11}| + a_{12}(-1)^{1+2} |A_{12}| \\ &= a_{11}|(a_{22})| - a_{12}|(a_{21})| \\ &= a_{11}a_{22} - a_{12}a_{21}. \end{aligned} \quad (9.77)$$

Für $A \in \mathbb{R}^{3 \times 3}$ gilt nach Definition und mit der Formel für Determinanten von 2×2 Matrizen

$$\begin{aligned}
 |A| &= \sum_{j=1}^n a_{1j}(-1)^{1+j}|A_{1j}| \\
 &= a_{11}(-1)^{1+1}|A_{11}| + a_{12}(-1)^{1+2}|A_{12}| + a_{13}(-1)^{1+3}|A_{13}| \\
 &= a_{11}|A_{11}| - a_{12}|A_{12}| + a_{13}|A_{13}| \\
 &= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\
 &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \\
 &= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} \\
 &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31}.
 \end{aligned} \tag{9.78}$$

□

Für die Bestimmung der Determinanten von 2×2 und 3×3 Matrizen gilt somit die sogenannte *Sarrusche Merkregel*:

“Summe der Produkte auf den Diagonalen minus Summe der Produkte auf den Gegendiagonalen.”

Dabei bezieht sich die Merkregeln bei 3×3 Matrizen auf das Schema

$$\left(\begin{array}{ccc|cc} a_{11} & a_{12} & a_{13} & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} & a_{31} & a_{32} \end{array} \right). \tag{9.79}$$

Beispiele für Determinanten von 2×2 und 3×3 Matrizen

Es seien

$$A := \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix}, B := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ und } C := \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{9.80}$$

Dann ergeben sich

$$|A| = 2 \cdot 4 - 1 \cdot 3 = 8 - 3 = 5 \tag{9.81}$$

und

$$|B| = 1 \cdot 0 - 0 \cdot 0 = 0 - 0 = 0 \tag{9.82}$$

und

$$|C| = 2 \cdot 1 \cdot 3 + 0 \cdot 0 \cdot 0 + 0 \cdot 0 \cdot 0 - 0 \cdot 0 \cdot 3 - 0 \cdot 0 \cdot 0 - 0 \cdot 1 \cdot 0 = 2 \cdot 1 \cdot 3 = 6. \tag{9.83}$$

In **R** rechnet man dies mithilfe der `det()` Funktion wie folgt nach.

```
# Matrixdefinition und Determinantenberechnung
A = matrix(c(2,1,
             3,4),
           nrow = 2,
           byrow = TRUE)
det(A)
```

[1] 5

```
# Matrixdefinition und Determinantenberechnung
B = matrix(c(1,0,
             0,0),
           nrow = 2,
           byrow = TRUE)
det(B)
```

[1] 0

```
# Matrixdefinition und Determinantenberechnung
C = matrix(c(2,0,0,
             0,1,0,
             0,0,3),
           nrow = 3,
           byrow = TRUE)
det(C)
```

[1] 6

Für Determinanten bestehen zahlreiche Rechenregeln im Zusammenspiel mit Matrixmultiplikation und Matrixinversion. Ohne Beweis stellen wir diese in folgendem Theorem zusammen.

Theorem 9.9 (Rechenregeln für Determinanten).

(Determinantenmultiplikationssatz). Für $A, B \in \mathbb{R}^{n \times n}$ gilt

$$|AB| = |A||B|. \quad (9.84)$$

(Transposition). Für $A \in \mathbb{R}^{n \times n}$ gilt

$$|A| = |A^T|. \quad (9.85)$$

(Inversion). Für eine invertierbare Matrix $A \in \mathbb{R}^{n \times n}$ gilt

$$|A^{-1}| = \frac{1}{|A|}. \quad (9.86)$$

(Dreiecksmatrizen). Für Matrizen $A = (a_{ij})_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$ mit $a_{ij} = 0$ für $i > j$ oder $a_{ij} = 0$ für $j > i$ gilt

$$|A| = \prod_{i=1}^n a_{ii}. \quad (9.87)$$

◦

Folgendes sehr tiefgehendes Theorem, welches wir nicht vollständig beweisen wollen, gibt eine Möglichkeit an, anhand der Determinante einer quadratischen Matrix zu bestimmen, ob sie invertierbar ist.

Theorem 9.10. $A \in \mathbb{R}^{n \times n}$ ist dann und nur dann invertierbar, wenn gilt, dass $|A| \neq 0$. Es gilt also

$$A \text{ ist invertierbar} \Leftrightarrow |A| \neq 0 \text{ und } A \text{ ist nicht invertierbar} \Leftrightarrow |A| = 0. \quad (9.88)$$

◦

Beweis. Wir deuten einen Beweis lediglich an und zeigen, dass aus der Invertierbarkeit von A folgt, dass $|A|$ nicht gleich Null sein kann. Nehmen wir also an, dass A invertierbar ist. Dann gibt es eine Matrix B mit $AB = I_n$ und mit dem Determinantenmultiplikationssatz folgt

$$|AB| = |A||B| = |I_n| = 1. \quad (9.89)$$

Also kann $|A| = 0$ nicht gelten, denn sonst wäre $0 = 1$. □

Visuelle Intuition

Der abstrakte Begriff der Determinante einer quadratischen Matrix kann mithilfe des Vektorraumbegriffs etwas veranschaulicht werden. Dazu seien $a_1, \dots, a_n \in \mathbb{R}^n$ die Spalten von $A \in \mathbb{R}^{n \times n}$. Dann gilt (wie wir nicht beweisen wollen), dass $|A|$ dem signierten Volumen des von $a_1, \dots, a_n \in \mathbb{R}^n$ aufgespannten Parallelotops entspricht. Um dies visuell zu veranschaulichen betrachten wir die Matrizen

$$A_1 = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}, A_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, A_3 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \quad (9.90)$$

mit den jeweiligen Determinanten

$$|A_1| = 3 \cdot 2 - 1 \cdot 1 = 5, \quad |A_2| = 2 \cdot 2 - 0 \cdot 0 = 4, \quad |A_3| = 2 \cdot 2 - 2 \cdot 2 = 0. \quad (9.91)$$

Abbildung 9.1 visualisiert die entsprechende Intuition.

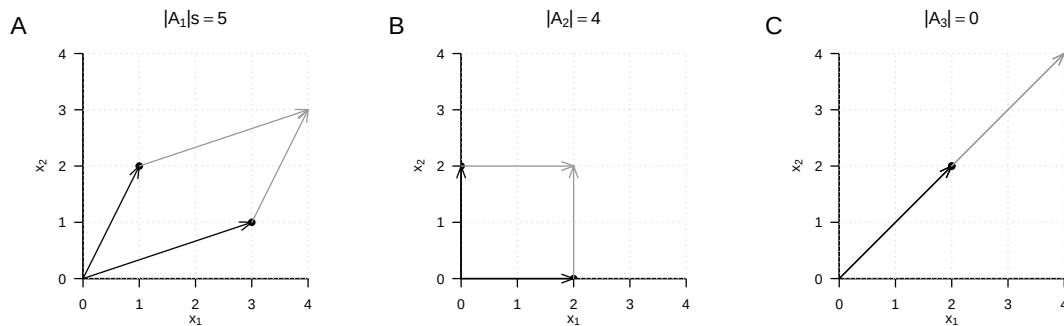


Abbildung 9.1. Determinanten als Parallelotopvolumina.

9.6. Spezielle Matrizen

In dieser Sektion stellen wir einige häufig auftretende Typen von Matrizen und ihre Eigenschaften zusammen. Zum Beweis der allermeisten Eigenschaften verweisen wir dabei auf die weiterführende Literatur.

9.6.1. Einheitsmatrizen

Die Einheitsmatrix und die Einheitsvektoren haben wir bereits kennengelernt. Wir fassen sie hier noch einmal in einer gemeinsamen Definition zusammen.

Definition 9.10 (Einheitsmatrix und Einheitsvektoren). Wir bezeichnen die *Einheitsmatrix* mit

$$I_n := (i_{jk})_{1 \leq j \leq n, 1 \leq k \leq n} \in \mathbb{R}^{n \times n} \text{ mit } i_{jk} = 1 \text{ für } j = k \text{ und } i_{jk} = 0 \text{ für } j \neq k. \quad (9.92)$$

Wir bezeichnen die *Einheitsvektoren* $e_i, i = 1, \dots, n$ mit

$$e_i := (e_{ij})_{1 \leq j \leq n} \in \mathbb{R}^n \text{ mit } e_{ij} = 1 \text{ für } i = j \text{ und } e_{ij} = 0 \text{ für } i \neq j. \quad (9.93)$$

•

Die Einheitsmatrix I_n besteht nur aus Nullen und Diagonalelementen gleich Eins, die Einheitsvektoren bestehen nur aus Nullen und einer Eins in der jeweils indizierten Komponente. Es gilt

$$I_n = (e_1 \quad \cdots \quad e_n) \in \mathbb{R}^{n \times n} \quad (9.94)$$

Für $n = 3$ gilt also zum Beispiel

$$I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ und } e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \quad (9.95)$$

Weiterhin gelten für die Einheitsvektoren bekanntlich für $1 \leq i, j \leq n$

$$e_i^T e_j = 0 \text{ für } i \neq j, e_i^T e_i = 1 \text{ und } e_i^T v = v^T e_i = v_i \text{ für } v \in \mathbb{R}^n. \quad (9.96)$$

9.6.2. Einsmatrizen und Nullmatrizen

Definition 9.11 (Nullmatrizen, Nullvektoren, Einsmatrizen, Einsvektoren). Wir bezeichnen *Nullmatrizen* und *Nullvektoren* mit

$$0_{nm} := (0)_{1 \leq i \leq n, 1 \leq j \leq m} \in \mathbb{R}^{n \times m} \text{ und } 0_n := (0)_{1 \leq i \leq n} \in \mathbb{R}^n. \quad (9.97)$$

Wir bezeichnen *Einsmatrizen* und *Einsvektoren* mit

$$1_{nm} := (1)_{1 \leq i \leq n, 1 \leq j \leq m} \in \mathbb{R}^{n \times m} \text{ und } 1_n := (1)_{1 \leq i \leq n} \in \mathbb{R}^n. \quad (9.98)$$

•

0_{nm} und 0_n bestehen also nur aus Nullen und 1_{nm} und 1_n bestehen nur aus Einsen. Es gilt also beispielsweise

$$0_{32} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, 0_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, 1_{32} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \text{ und } 1_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \quad (9.99)$$

Weiterhin gelten zum Beispiel

$$0_n 0_n^T = 0_{nn} \text{ und } 1_n 1_n^T = 1_{nn}, \quad (9.100)$$

wovon man sich durch Nachrechnen überzeugt.

9.6.3. Diagonalmatrizen

Definition 9.12 (Diagonalmatrix). Eine Matrix $D \in \mathbb{R}^{n \times m}$ heißt *Diagonalmatrix*, wenn $d_{ij} = 0$ für $1 \leq i \leq n, 1 \leq j \leq m$ mit $i \neq j$.

•

Eine quadratische Diagonalmatrix $D \in \mathbb{R}^{n \times n}$ mit den Diagonalelementen $d_1, \dots, d_n \in \mathbb{R}$ schreibt man auch als

$$D = \text{diag}(d_1, \dots, d_n). \quad (9.101)$$

Zum Beispiel gelten

$$D := \text{diag}(1, 2, 3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \quad (9.102)$$

und für $\sigma^2 \in \mathbb{R}$

$$\Sigma = \text{diag}(\sigma^2, \sigma^2, \sigma^2) = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 I_3. \quad (9.103)$$

In folgendem Theorem stellen wir einige wichtige Eigenschaften von quadratischen Diagonalmatrizen zusammen.

Theorem 9.11 (Eigenschaften quadratischer Diagonalmatrizen).

(Determinante.) $D := \text{diag}(d_1, \dots, d_n) \in \mathbb{R}^{n \times n}$ sei eine quadratische Diagonalmatrix. Dann gilt

$$|D| = \prod_{i=1}^n d_i. \quad (9.104)$$

◦

9.6.4. Symmetrische Matrizen

Symmetrische Matrizen sind quadratische Matrizen, die bei Transposition unverändert bleiben:

Definition 9.13. Eine Matrix $S \in \mathbb{R}^{n \times n}$ heißt *symmetrisch*, wenn $S^T = S$.

•

Ein Beispiel für eine symmetrische Matrix ist

$$S := \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix}. \quad (9.105)$$

In folgendem Theorem stellen wir einige wichtige Eigenschaften symmetrischer Matrizen zusammen.

Theorem 9.12 (Eigenschaften symmetrischer Matrizen).

(Summation.) $S_1 \in \mathbb{R}^{n \times n}$ und $S_2 \in \mathbb{R}^{n \times n}$ seien symmetrische Matrizen. Dann gilt

$$S_1 + S_2 = (S_1 + S_2)^T. \quad (9.106)$$

(Inverse.) S sei eine invertierbare symmetrische Matrix und S^{-1} ihre Inverse. Dann ist auch S^{-1} eine symmetrische Matrix, das heißt es gilt

$$(S^{-1})^T = S^{-1}. \quad (9.107)$$

◦

9.6.5. Orthogonale Matrizen

Definition 9.14. Eine Matrix $Q \in \mathbb{R}^{n \times n}$ heißt *orthogonal*, wenn $Q^T Q = I_n$.

•

Die Spalten einer orthogonalen Matrix sind also paarweise orthogonal, es gilt für

$$Q = (q_1 \quad \cdots \quad q_n) \text{ mit } q_i \in \mathbb{R}^n \text{ für } 1 \leq i \leq n, \quad (9.108)$$

dass

$$q_i^T q_j = 0 \text{ für } i \neq j \text{ und } q_i^T q_j = 1 \text{ für } i = j \text{ mit } 1 \leq i, j \leq n. \quad (9.109)$$

Theorem 9.13 (Eigenschaften orthogonaler Matrizen). $Q \in \mathbb{R}^{n \times n}$ sei eine orthogonale Matrix. Dann gelten folgende Eigenschaften von Q .

(Inverse.) Die Inverse von Q ist Q^T , es gilt

$$Q^{-1} = Q^T. \quad (9.110)$$

(Transposition) Die Zeilen von Q sind orthonormal, es gilt

$$Q Q^T = I_n \quad (9.111)$$

◦

Beweis. (Inverse) Unter der Annahme, dass Q^{-1} existiert, gilt

$$Q^T Q = I_n \Leftrightarrow Q^T Q Q^{-1} = I_n Q^{-1} \Leftrightarrow Q^{-1} = Q^T. \quad (9.112)$$

(Transposition) Es gilt

$$Q^T Q = I_n \Leftrightarrow Q Q^T Q = Q I_n \Leftrightarrow Q Q^T Q Q^T = Q Q^T \Leftrightarrow Q Q^T = I_n. \quad (9.113)$$

□

9.6.6. Positiv-definite Matrizen

Positiv-definite Matrizen sind für die probabilistische Modellbildung unter Verwendung multivariater Normalverteilungen zentral.

Definition 9.15. Eine quadratische Matrix $C \in \mathbb{R}^{n \times n}$ heißt positiv-definit (p.d.), wenn

- C eine symmetrische Matrix ist und
- für alle $x \in \mathbb{R}^n, x \neq 0_n$ gilt, dass $x^T C x > 0$ ist.

•

In folgendem Theorem stellen wir einige wichtige Eigenschaften positiv-definiten Matrizen zusammen.

Theorem 9.14 (Eigenschaften positiv-definiten Matrizen).

(Inverse.) $C \in \mathbb{R}^{n \times n}$ sei eine positiv-definite Matrix. Dann gilt, dass C^{-1} existiert und ebenfalls positiv-definit ist.

◦

9.7. Eigenanalyse

Mit der *Eigenanalyse einer quadratischen Matrix*, der *Orthonormalzerlegung einer symmetrischen Matrix* und der *Singulärwertzerlegung einer beliebigen Matrix* behandeln wir in diesem Abschnitt drei eng zusammenhängende Konzepte der Matrixtheorie, die in vielen Gebieten der datenanalytischen Anwendung zentrale Rollen spielen. Allerdings erschließt sich die Bedeutung dieser Konzepte dann vor allem im jeweiligen Anwendungskontext, so dass dieser Abschnitt notwendigerweise etwas abstrakt anmuten mag.

Eigenvektoren und Eigenwerte

Unter der *Eigenanalyse* einer quadratischen Matrix versteht man das bestimmen ihrer *Eigenvektoren* und *Eigenwerte*. Diese sind für eine quadratische Matrix wie folgt definiert.

Definition 9.16 (Eigenvektor und Eigenwert). $A \in \mathbb{R}^{m \times m}$ sei eine quadratische Matrix. Dann heißt jeder vom Nullvektor 0_m verschiedene Vektor $v \in \mathbb{R}^m$, für den mit einem Skalar $\lambda \in \mathbb{R}$ gilt, dass

$$Av = \lambda v \quad (9.114)$$

ist, ein *Eigenvektor* von A und λ heißt dann ein *Eigenwert* von A .

•

Nach Definition hat also jeder Eigenvektor einen zugehörigen Eigenwert, allerdings können die Eigenwerte verschiedener Eigenvektoren durchaus identisch sein. Intuitiv bedeutet die Definition von Eigenvektor und Eigenwert, dass ein Eigenvektor einer Matrix durch Multiplikation mit eben dieser Matrix in seiner Länge, nicht aber in seiner Richtung, verändert wird. Der zugehörige Eigenwert des Eigenvektors entspricht dem Faktor der Längenänderung. Allerdings ist die Zuordnung von Eigenvektoren und Eigenwerten nicht eindeutig, wie folgendes Theorem zeigt.

Theorem 9.15 (Multiplikativität von Eigenvektoren). $A \in \mathbb{R}^{m \times m}$ sei eine quadratische Matrix. Wenn $v \in \mathbb{R}^m$ Eigenvektor von A mit Eigenwert $\lambda \in \mathbb{R}$ ist, dann ist für $c \in \mathbb{R}$ auch $cv \in \mathbb{R}^m$ Eigenvektor von A und zwar wiederum mit Eigenwert $\lambda \in \mathbb{R}$.

◦

Beweis. Es gilt

$$Av = \lambda v \Leftrightarrow cAv = c\lambda v \Leftrightarrow A(cv) = \lambda(cv). \quad (9.115)$$

Also ist cv ein Eigenvektor von A mit Eigenwert λ . □

Um nun die Uneindeutigkeit in der Definition des zu einem Eigenwert zugeordneten Eigenvektors aufzulösen, nutzen wir die Konvention, nur diejenigen Vektoren also Eigenvektoren zu einem Eigenwert λ zu betrachten, die die Länge 1 haben, für die also gilt, dass

$$\|v\| = 1. \quad (9.116)$$

Sollten wir also einen Eigenvektor v zu einem Eigenwert λ einer Matrix A finden, der nicht von der Länge 1 ist, so können wir ihn immer mit $\|v\|^{-1}$ multiplizieren. Der resultierende Vektor $v' = v/\|v\|$ hat dann die Länge 1 und ist nach Theorem 9.15 ebenso ein Eigenvektor

von A zum Eigenwert λ . Bevor wir uns der Bestimmung von Eigenwerten und Eigenvektoren widmen, wollen wir die Konzepte von Eigenwert und Eigenvektor für den Fall einer 2×2 Matrix an einem Beispiel veranschaulichen

Beispiel

Es sei

$$A := \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad (9.117)$$

Dann ist der Vektor der Länge 1

$$v := \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (9.118)$$

ein Eigenvektor von A zum Eigenwert $\lambda = 3$, da gilt, dass

$$\begin{aligned} Av &= \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} 3 \\ 3 \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} 3 \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= 3 \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) \\ &= \lambda v. \end{aligned} \quad (9.119)$$

Inspektion von Abbildung 9.2 zeigt dementsprechend, dass für die hier definierte Matrix A die Vektoren v und Av in die gleiche Richtung zeigen, dass aber Av um den Faktor λ länger ist als v .

Der Vektor

$$w := \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (9.120)$$

dagegen hat zwar die Länge 1, ist aber im Gegensatz zu v kein Eigenvektor von A , da es im Falle von

$$Aw = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad (9.121)$$

keinen Skalar λ geben kann, der mit Null, dem zweiten Eintrag von w , multipliziert einen Wert ungleich Null ergeben kann. Inspektion von Abbildung 9.2 zeigt dementsprechend, dass der aus der Multiplikation von w mit A resultierende Vektor in eine andere Richtung zeigt als w .

Bestimmung von Eigenwerten und Eigenvektoren

Folgendes Theorem besagt, wie die Eigenwerte und Eigenvektoren einer quadratischen Matrix berechnet werden können.

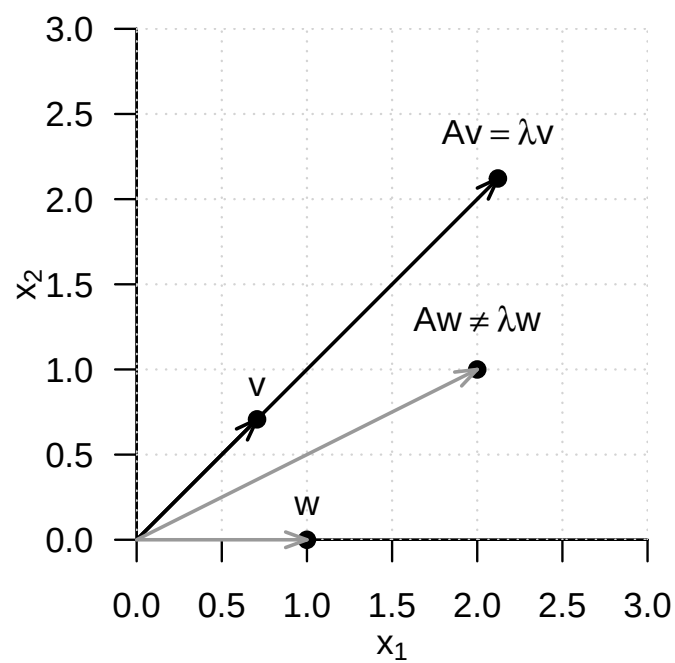


Abbildung 9.2. Eigenvektor einer 2×2 Matrix. Für die Matrix

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad (9.122)$$

ist v ein Eigenvektor, w jedoch nicht

Theorem 9.16 (Bestimmung von Eigenwerten und Eigenvektoren). $A \in \mathbb{R}^{m \times m}$ sei eine quadratische Matrix. Dann ergeben sich die Eigenwerte von A als die Nullstellen des charakteristischen Polynoms

$$\chi_A(\lambda) := |A - \lambda I_m| \quad (9.123)$$

von A . Weiterhin seien $\lambda_i^*, i = 1, 2, \dots$ die auf diese Weise bestimmten Eigenwerte von A . Die entsprechenden Eigenvektoren $v_i, i = 1, 2, \dots$ von A können dann durch Lösen der linearen Gleichungssysteme

$$(A - \lambda_i^* I_m) v_i = 0_m \text{ für } i = 1, 2, \dots \quad (9.124)$$

bestimmt werden.

◦

Beweis. (1) Bestimmen von Eigenwerten

Wir halten zunächst fest, dass mit der Definition von Eigenvektoren und Eigenwerten gilt, dass

$$Av = \lambda v \Leftrightarrow Av - \lambda v = 0_m \Leftrightarrow (A - \lambda I_m)v = 0_m. \quad (9.125)$$

Für den Eigenwert λ wird der Eigenvektor v also durch Multiplikation mit $(A - \lambda I_m)$ auf den Nullvektor 0_m abgebildet. Weil aber per Definition $v \neq 0_m$ gilt, ist die Matrix $(A - \lambda I_m)$ somit nicht invertierbar: sowohl der Nullvektor als auch v werden durch A auf 0_m abgebildet, die Abbildung

$$f: \mathbb{R}^m \rightarrow \mathbb{R}^m, x \mapsto (A - \lambda I_m)x \quad (9.126)$$

ist also nicht bijektiv, und $(A - \lambda I_m)^{-1}$ kann nicht existieren. Die Tatsache, dass $(A - \lambda I_m)$ nicht invertierbar ist, ist aber äquivalent dazu, dass die Determinante von $(A - \lambda I_m)$ gleich Null ist. Also ist

$$\chi_A(\lambda) = |A - \lambda I_m| = 0 \quad (9.127)$$

eine notwendige und hinreichende Bedingung dafür, dass λ ein Eigenwert von A ist.

(2) Bestimmen von Eigenvektoren

Es sei λ_i^* ein Eigenwert von A . Dann gilt mit den obigen Überlegungen, dass Auflösen von

$$(A - \lambda_i^* I_m) v_i^* = 0_m \quad (9.128)$$

nach v_i^* einen Eigenvektor zum Eigenwert λ_i^* ergibt. □

Allgemein müssen zur Bestimmung von Eigenwerten und Eigenvektoren also Polynomnullstellen bestimmt und lineare Gleichungssysteme gelöst werden. Dies kann für kleine Matrizen mit $m \leq 4$ durchaus manuell geschehen. Die in der Anwendung auftretenden Matrizen sind jedoch meist weitaus größer, so dass zur Eigenanalyse numerische Verfahren der Nullstellenbestimmung und des Lösen linearer Gleichungssysteme eingesetzt werden, die zum Beispiel in Funktionen wie **R's** `eigen()`, **SciPy's** `linalg.eig()` oder **Julia's** `eigvals()` und `eigvecs()` genutzt werden. Für Details zu diesen Verfahren verweisen wir auf die weiterführende Literatur, zum Beispiel Burden et al. (2016) und Richter & Wick (2017). Wir wollen Theorem 9.16 hier lediglich anhand eines Beispiels illustrieren.

Beispiel

Dazu sei wiederum

$$A := \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad (9.129)$$

Wir wollen zunächst die Eigenwerte von A berechnen. Nach Theorem 9.16 sind dies die Nullstellen des charakteristischen Polynoms von A . Wir berechnen also zunächst das charakteristische Polynom von A durch

$$\chi_A(\lambda) = \left| \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| = \left| \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} \right| = (2-\lambda)^2 - 1. \quad (9.130)$$

Mithilfe der pq-Formel zur Lösung quadratischer Gleichungen findet man dann

$$(2 - \lambda_{1/2}^*)^2 - 1 = 0 \Leftrightarrow \lambda_1^* = 3 \text{ oder } \lambda_2^* = 1. \quad (9.131)$$

Die Eigenwerte von A sind also $\lambda_1 = 3$ und $\lambda_2 = 1$. Die zugehörigen Eigenvektoren ergeben sich dann für $i = 1, 2$ durch Lösen des linearen Gleichungssystems

$$(A - \lambda_i I_2)v_i = 0_2. \quad (9.132)$$

Speziell ergibt sich hier, dass für $\lambda_1 = 3$ aus

$$(A - 3I_2)v_1 = 0_2 \Leftrightarrow \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} v_{1_1} \\ v_{1_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (9.133)$$

folgt, dass

$$v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (9.134)$$

ein Eigenvektor zum Eigenwert λ_1 ist und dass für $\lambda_2 = 1$ aus

$$(A - 1I_2)v_2 = 0_2 \Leftrightarrow \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_{2_1} \\ v_{2_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (9.135)$$

folgt, dass

$$v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad (9.136)$$

ein Eigenvektor zum Eigenwert $\lambda_2 = 1$ ist. Weiterhin gelten hier offenbar

$$v_1^T v_2 = 0 \text{ und } \|v_1\| = \|v_2\| = 1. \quad (9.137)$$

Folgender **R** Code demonstriert die Bestimmung der Eigenwerte und Eigenvektoren der hier betrachteten Matrix mithilfe der `eigen()` Funktion.

```
# Matrixdefinition
A = matrix(c(2,1,
             1,2),
           nrow = 2,
           byrow = TRUE)

# Eigenanalyse
eigen(A)

eigen() decomposition
$values
[1] 3 1

$vectors
      [,1]      [,2]
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068
```

Zum Abschluss dieses Abschnittes betrachten wir zwei technische Theoreme, die Aussagen zum Zusammenhang spezieller Matrixprodukte und ihrer Eigenwerte und Eigenvektoren machen. Wir benötigen dieses Theoreme im Kontext der Kanonischen Korrelationsanalyse (Kapitel 48).

Theorem 9.17 (Eigenwerte und Eigenvektoren von Matrixprodukten). *Für $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{m \times n}$ sind die Eigenwerte von $AB \in \mathbb{R}^{n \times n}$ und $BA \in \mathbb{R}^{m \times m}$ gleich. Weiterhin gilt, dass für einen Eigenvektor v zu einem von Null verschiedenen Eigenwert λ von AB $w := Bv$ ein Eigenvektor von BA zum Eigenwert λ ist.*

◦

Für einen Beweis verweisen wir auf Mardia et al. (1979), S. 468. Wir demonstrieren die Aussage dieses Theorems anhand untenstehenden **R** Codes.

```
A = matrix(1:6, nrow = 2, byrow = T)      # Matrix A \in \mathbb{R}^{2 \times 3}
B = matrix(1:6, ncol = 2, byrow = T)      # Matrix B \in \mathbb{R}^{3 \times 2}
EAB = eigen(A %*% B)                      # Eigenanalyse von AB \in \mathbb{R}^{2 \times 2}
EBA = eigen(B %*% A)                      # Eigenanalyse von BA \in \mathbb{R}^{3 \times 3}
w = B %*% EAB$eigenvectors[,1]           # Eigenvektor von BA
cat("Eigenwerte von AB :", EAB$values[1:2],
    "\nEigenwerte von BA :", EBA$values[1:2],
    "\nBAw mit w = Bv      :", B %*% A %*% w,
    "\nlw mit w = Bv      :", EBA$values[1] * w)
```

```
Eigenwerte von AB : 85.57934 0.4206623
Eigenwerte von BA : 85.57934 0.4206623
BAw mit w = Bv    : -191.1333 -416.7586 -642.3839
lw mit w = Bv     : -191.1333 -416.7586 -642.3839
```

Theorem 9.18. *Für $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{p \times n}$, $a \in \mathbb{R}^m$ und $b \in \mathbb{R}^p$ gilt, dass der einzige von Null verschiedene Eigenwert von $Aab^T B \in \mathbb{R}^{n \times n}$ gleich $b^T B A a$ mit zugehörigem Eigenvektor Aa ist.*

◦

Für einen Beweis verweisen wir auf Mardia et al. (1979), S. 468. Wir demonstrieren die Aussage dieses Theorems anhand untenstehenden **R** Codes.

```
A = matrix(1:6, nrow = 2, byrow = T)      # Matrix A \in \mathbb{R}^{2 \times 3}
B = matrix(1:8, ncol = 2, byrow = T)      # Matrix B \in \mathbb{R}^{4 \times 2}
a = matrix(1:3, nrow = 3, byrow = T)      # Vektor a \in \mathbb{R}^{3 \times 1}
b = matrix(1:4, nrow = 4, byrow = T)      # Vektor b \in \mathbb{R}^{4 \times 1}
EAabTB = eigen(A %*% a %*% t(b) %*% B)    # Eigenanalyse von Aab^T B \in \mathbb{R}^{4 \times 4}
cat("Eigenwerte von AabTB :", EAabTB$values,
    "\nbTBaAa      :", t(b) %*% B %*% A %*% a,
    "\nAa         :", A %*% a,
    "\n(AabTB)Aa   :", (A %*% a %*% t(b) %*% B) %*% A %*% a, # Mv
    "\n(bTBaAa)Aa  :", as.vector((t(b) %*% B %*% A %*% a)) * (A %*% a)) # = \lambda v
```

```
Eigenwerte von AabTB : 2620 0
bTBaAa              : 2620
Aa                  : 14 32
(AabTB)Aa           : 36680 83840
(bTBaAa)Aa          : 36680 83840
```


9.8. Orthonormalzerlegung

Mit dem Begriff der *Zerlegung* einer Matrix wird das Aufspalten einer gegebenen Matrix in das Matrixprodukt mehrerer Matrizen bezeichnet. Verschiedenste Matrixzerlegungen spielen in vielen mathematischen Anwendungen eine wichtige Rolle, für einen Überblick siehe beispielsweise Golub & Van Loan (2013). In diesem Abschnitt führen wir mit der *Orthonormalzerlegung einer symmetrischen Matrix* eine spezielle Matrixzerlegung ein, die direkt auf der Eigenanalyse aufbaut. Wir halten zunächst folgendes grundlegendes Theorem zu den Eigenwerten und Eigenvektoren symmetrischer Matrizen fest.

Theorem 9.19 (Eigenwerte und Eigenvektoren symmetrischer Matrizen).

$S \in \mathbb{R}^{m \times m}$ sei eine symmetrische Matrix. Dann gelten

- (1) Die Eigenwerte von S sind reell.
- (2) Die Eigenvektoren zu je zwei verschiedenen Eigenwerten von S sind orthogonal.

◦

Beweis. Wir setzen die Tatsache, dass eine symmetrische Matrix m reelle Eigenwerte hat, als gegeben voraus und zeigen lediglich, dass die Eigenvektoren zu je zwei verschiedenen Eigenwerten einer symmetrischen Matrix orthogonal sind. Ohne Beschränkung der Allgemeinheit seien also $\lambda_i, \lambda_j \in \mathbb{R}$ mit $1 \leq i, j \leq m$ und $\lambda_i \neq \lambda_j$ zwei verschiedenen Eigenwerte von S mit zugehörigen Eigenvektoren q_i und q_j , respektive. Dann ergibt sich wie unten gezeigt, dass

$$\lambda_i q_i^T q_j = \lambda_j q_i^T q_j. \quad (9.138)$$

Mit $q_i \neq 0_m, q_j \neq 0_m$ und $\lambda_i \neq \lambda_j$ folgt damit $q_i^T q_j = 0$, weil es keine andere Zahl c als die Null gibt, für die bei $a, b \in \mathbb{R}$ und $a \neq b$ gilt, dass

$$ac = bc. \quad (9.139)$$

Um abschließend

$$\lambda_i q_i^T q_j = \lambda_j q_i^T q_j. \quad (9.140)$$

zu zeigen, halten wir zunächst fest, dass

$$S q_i = \lambda_i q_i \Leftrightarrow (S q_i)^T = (\lambda_i q_i)^T \Leftrightarrow q_i^T S^T = q_i^T \lambda_i^T \Leftrightarrow q_i^T S = q_i^T \lambda_i \Leftrightarrow q_i^T S q_j = \lambda_i q_i^T q_j \quad (9.141)$$

und

$$S q_j = \lambda_j q_j \Leftrightarrow q_j^T S = q_j^T \lambda_j \Leftrightarrow q_j^T S q_i = \lambda_j q_j^T q_i \Leftrightarrow (q_j^T S q_i)^T = (\lambda_j q_j^T q_i)^T \Leftrightarrow q_i^T S q_j = \lambda_j q_i^T q_j \quad (9.142)$$

gelten. Sowohl $\lambda_i q_i^T q_j$ als auch $\lambda_j q_i^T q_j$ sind also mit $q_i^T S q_j$ und damit auch miteinander identisch. $\square$

Offenbar haben wir nur Aussage (2) von Theorem 9.19 bewiesen. Ein vollständiger Beweis des Theorems findet sich zum Beispiel bei Strang (2009). Wir merken außerdem an, dass, weil wir nach Konvention Eigenvektoren der Länge 1 betrachten, die in Theorem 9.19 angesprochenen orthogonalen Eigenvektoren insbesondere auch orthonormal sind. Mithilfe von Theorem 9.19 können wir nun die Orthonormalzerlegung einer symmetrischen Matrix formulieren und ihre Existenz beweisen.

Theorem 9.20 (Orthonormalzerlegung einer symmetrischen Matrix). $S \in \mathbb{R}^{m \times m}$ sei eine symmetrische Matrix mit m verschiedenen Eigenwerten. Dann kann S geschrieben werden als

$$S = Q \Lambda Q^T, \quad (9.143)$$

wobei $Q \in \mathbb{R}^{m \times m}$ eine orthogonale Matrix ist und $\Lambda \in \mathbb{R}^{m \times m}$ eine Diagonalmatrix ist.

◦

Beweis. Es seien $\lambda_1 > \lambda_2 > \dots > \lambda_m$ die der Größe nach geordneten Eigenwerte von S und $q_1, \dots, q_m$ die zugehörigen orthonormalen Eigenvektoren. Mit

$$Q := (q_1 \quad q_2 \quad \dots \quad q_m) \in \mathbb{R}^{m \times m} \text{ und } \Lambda := \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m) \in \mathbb{R}^{m \times m}, \quad (9.144)$$

folgt dann mit den Definitionen von Eigenwerten und Eigenvektoren zunächst, dass

$$Sq_i = \lambda_i q_i \text{ für } i = 1, \dots, m \Leftrightarrow SQ = Q\Lambda. \quad (9.145)$$

Rechtseitige Multiplikation mit Q^T ergibt dann mit $QQ^T = I_m$, dass

$$SQQ^T = Q\Lambda Q^T \Leftrightarrow SI_m = Q\Lambda Q^T \Leftrightarrow S = Q\Lambda Q^T. \quad (9.146)$$

□

Man nennt das Aufspalten von S in das Matrixprodukt $Q\Lambda Q^T$ aufgrund der Diagonalität von Λ auch eine *Diagonalisierung von S* . Wie im Beweis gezeigt, wählt man zur Darstellung von S in Diagonaldarstellung für die Diagonalelemente von Λ die der Größe nach geordneten Eigenwerte von S und für die Spalten von Q die jeweils zugehörigen Eigenvektoren von S . Wir verdeutlichen dies an einem Beispiel.

Beispiel

Für die symmetrische Matrix

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad (9.147)$$

mit den oben bestimmten Eigenwerten $\lambda_1 = 3$ und $\lambda_2 = 1$ sowie den zugehörigen orthonormalen Eigenvektoren

$$v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad (9.148)$$

seien

$$Q := (v_1 \quad v_2) \text{ und } \Lambda = \text{diag}(\lambda_1, \lambda_2). \quad (9.149)$$

Dann ergibt sich offenbar

$$\begin{aligned} Q\Lambda Q^T &= (v_1 \quad v_2) \text{diag}(\lambda_1, \lambda_2) (v_1 \quad v_2)^T \\ &= \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \right) \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \right) \\ &= \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 3 & -1 \\ 3 & 1 \end{pmatrix} \right) \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \right) \\ &= \frac{1}{2} \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\ &= A \end{aligned}$$

und wir haben Theorem 9.20 für dieses Beispiel verifiziert.

Symmetrische Quadratwurzel einer Matrix

Die Definition der Orthonormalzerlegung einer symmetrischen Matrix erlaubt es, den Begriff der *symmetrischen Quadratwurzel einer Matrix* einzuführen.

Definition 9.17 (Symmetrische Quadratwurzel einer Matrix). $S \in \mathbb{R}^{m \times m}$ sei eine invertierbare symmetrische Matrix mit positiven Eigenwerten. Dann sind für $r \in \mathbb{N}^0$ und $s \in \mathbb{N}$ die rationalen Potenzen von S mit der orthonormalen Matrix $Q \in \mathbb{R}^{m \times m}$ der Eigenvektoren von S und der Diagonalmatrix $\Lambda = \text{diag}(\lambda_i) \in \mathbb{R}^{m \times m}$ der zugehörigen Eigenwerte $\lambda_1, \dots, \lambda_m$ von S definiert als

$$S^{r/s} = Q\Lambda^{r/s}Q^T \text{ mit } \Lambda^{r/s} = \text{diag}(\lambda_i^{r/s}). \quad (9.150)$$

Der Spezialfall $r := 1, s := 2$ wird als *symmetrische Quadratwurzel von S* bezeichnet und hat die Form

$$S^{1/2} = Q\Lambda^{1/2}Q^T \text{ mit } \Lambda^{1/2} = \text{diag}(\lambda_i^{1/2}). \quad (9.151)$$

•

Wir halten fest, dass mit Definition 9.17 offenbar gilt, dass

$$(S^{1/2})^2 = Q\Lambda^{1/2}Q^TQ\Lambda^{1/2}Q^T = Q\Lambda^{1/2}\Lambda^{1/2}Q^T = Q\Lambda Q^T = S. \quad (9.152)$$

Weiterhin gilt, dass

$$(S^{-1/2})^2 = Q\Lambda^{-1/2}Q^TQ\Lambda^{-1/2}Q^T = Q\Lambda^{-1/2}\Lambda^{-1/2}Q^T = Q\Lambda^{-1}Q^T = S^{-1}. \quad (9.153)$$

Schließlich gilt, dass

$$\begin{aligned} S^{-1/2}SS^{-1/2} &= Q\Lambda^{-1/2}Q^TQ\Lambda Q^TQ\Lambda^{-1/2}Q^T \\ &= Q\Lambda^{-1/2}\Lambda\Lambda^{-1/2}Q^T \\ &= Q\Lambda\Lambda^{-1}Q^T \\ &= I_m \end{aligned} \quad (9.154)$$

9.9. Singulärwertzerlegung

Eine vielseitig einsetzbare Matrixzerlegung einer beliebigen Matrix ist die *Singulärwertzerlegung*. Wir sind an dieser Stelle lediglich an dem Zusammenhang von Singulärwertzerlegung und Eigenanalyse interessiert und verweisen für eine ausführliche Diskussion der Singulärwertzerlegung auf die weiterführende Literatur, beispielsweise Strang (2009). Der Begriff der Singulärwertzerlegung ist wie folgt definiert.

Definition 9.18 (Singulärwertzerlegung). $Y \in \mathbb{R}^{m \times n}$ sei eine Matrix. Dann heißt die Zerlegung

$$Y = USV^T, \quad (9.155)$$

wobei $U \in \mathbb{R}^{m \times m}$ eine orthogonale Matrix ist, $S \in \mathbb{R}^{m \times n}$ eine Diagonalmatrix ist und $V \in \mathbb{R}^{n \times n}$ eine orthogonale Matrix ist, *Singulärwertzerlegung* von Y . Die Diagonalelemente von S heißen die *Singulärwerte* von Y .

•

Singulärwertzerlegungen werden auf Englisch *singular value decompositions* genannt und entsprechend mit *SVD* abgekürzt. Wir verzichten auf eine Diskussion der Berechnung einer Singulärwertzerlegung und weisen lediglich daraufhin, dass Singulärwertzerlegungen zum Beispiel in **R** mit der Funktion `svd()`, in **SciPy** mit `scipy.linalg.svd()` und in **Julia** mit `svd()` berechnet werden können. Folgendes Theorem beschreibt den Zusammenhang zwischen Singulärwertzerlegung und Eigenanalyse und wird an vielen Stellen eingesetzt.

Theorem 9.21 (Singulärwertzerlegung und Eigenanalyse).

$Y \in \mathbb{R}^{m \times n}$ sei eine Matrix und

$$Y = USV^T \quad (9.156)$$

sei ihre Singulärwertzerlegung. Dann gilt:

- Die Spalten von U sind die Eigenvektoren von YY^T ,
- die Spalten von V sind die Eigenvektoren von Y^TY und
- die Singulärwerte sind die Quadratwurzeln der zugehörigen Eigenwerte.

◦

Beweis. Wir halten zunächst fest, dass mit

$$(YY^T)^T = YY^T \text{ und } (Y^TY)^T = Y^TY, \quad (9.157)$$

YY^T und Y^TY symmetrische Matrizen sind und somit Orthonormalzerlegungen haben. Wir halten weiterhin fest, dass mit $V^TV = I_n$, $U^TU = I_m$ und $S^T = S$ gilt, dass

$$YY^T = USV^T(USV^T)^T = USV^TVS^TU^T = USSU^T =: U\Lambda U^T \quad (9.158)$$

und

$$Y^TY = (USV^T)^T USV^T = V S^T U^T U S^T V^T =: V\Lambda V^T \quad (9.159)$$

ist, wobei wir $\Lambda := SS$ definiert haben. Weil das Produkt von Diagonalmatrizen wieder eine Diagonalmatrix ist, ist Λ eine Diagonalmatrix und per Definition sind U und V orthogonale Matrizen. Wir haben also YY^T und Y^TY in Form der Orthonormalzerlegungen

$$YY^T = U\Lambda U^T \text{ und } Y^TY = V\Lambda V^T \quad (9.160)$$

geschrieben, wobei für die Diagonalelemente von Λ gilt, dass sie die quadrierten Werte der Diagonalelemente von S sind. $\square$

10. Deskriptivstatistiken

Ziel der Auswertung von Deskriptivstatistiken ist es, Datensätze möglichst effizient der menschlichen Betrachtung zugänglich zu machen. Sobald die Anzahl der Datenpunkte eines Datensatzes eine gewisse Anzahl überschreitet, sind reine Zahlendarstellungen für die menschliche Kognition meist eher schwer zu erfassen. Hier setzt die Deskriptivstatistik an und bietet neben graphischen Darstellungen größerer Datensätze auch Maßzahlen an, die im Sinne der Datenreduktion bestimmte charakteristische Aspekte eines Datensatzes, wie seinen “durchschnittlichen” Wert oder seine Variabilität beschreiben. In diesem Abschnitt wollen wir mit den univariaten Deskriptivstatistiken erste Verfahren und Maße kennenlernen, die es erlauben, Datensätze, in denen pro Studieneinheit eine Zahl vorliegt, zusammenzufassen. Wir betrachten dabei insbesondere *Häufigkeitsverteilungen* und *Histogramme*, *Verteilungsfunktionen* und *Quantile*, sowie *Maße der zentralen Tendenz*, die Aussagen über “durchschnittliche” Datenwerte erlauben und *Maße der Datenvariabilität*, die Aussagen über die Streuung der Datenwerte erlauben.

In Abgrenzung zur probabilistischen Modellierung im Rahmen der Frequentistischen und Bayesianischen Inferenz basiert die deskriptive Statistik nicht auf der Wahrscheinlichkeitstheorie, kann also losgelöst von dieser betrachtet werden. Allerdings ergeben viele Aspekte der Deskriptivstatistik letztlich nur vor dem Hintergrund wahrscheinlichkeitstheoretischer Betrachtungen Sinn und umgekehrt sind viele Begriffe der Wahrscheinlichkeitstheorie durch deskriptivstatistische Konzepte motiviert. Wenn also der vorliegende Abschnitt ohne einen wahrscheinlichkeitstheoretischen Hintergrund auskommt, so erschließt sich seine volle Bedeutung sicherlich erst im Kontext der Begriffe der Wahrscheinlichkeitstheorie.

Wir fokussieren in diesem Abschnitt auf Deskriptivstatistiken zur Beschreibung eines Datensatzes der Form

$$y := (y_1, \dots, y_n) \text{ mit } y_i \in \mathbb{R} \text{ für } i = 1, \dots, n. \quad (10.1)$$

Dabei ist es zunächst unerheblich, ob ein solcher Datensatz in Form eines Spaltenvektors $y \in \mathbb{R}^{n \times 1}$ oder eines Zeilenvektors $y \in \mathbb{R}^{1 \times n}$ vorliegt. Weiterhin fokussieren wir auf die praktische Auswertung der betrachteten Deskriptivstatistiken, geben also viele Beispiele für ihre Berechnung mit **R**.

Anwendungsbeispiel

Als Anwendungsbeispiel betrachten wir durchgängig einen Datensatz zur evidenzbasierten Evaluation von Psychotherapie bei Depression. Dazu nehmen wir an, dass $n = 100$ Patient:innen zufällig auf eine Treatment- und eine Kontrollstudienbedingung aufgeteilt wurden und zu Beginn der jeweiligen Intervention mithilfe des BDI-II Fragebogens zu ihrer Depressionsschwere befragt wurden. Die nach den Interventionen erhobenen BDI-II Werte wollen wir in diesem Kapitel nicht betrachten. Wir nehmen also an, dass wir einen Datensatz wie durch folgenden **R** Code in Tabelle 10.1 dargestellt vorliegen haben.

```
D = read.csv("./_data/110-deskriptive-statistik.csv")
knitr::kable(D[,1:2], digits = 2, align = "c")
```

```
# Laden des Datensatzes
# RMarkdowntabellenoutput
```

Tabelle 10.1. BDI-II Werte (PRE) von $n = 100$ Patient:innen in einer Treatment (T) und einer Kontrollstudiengruppe (C) vor Beginn einer psychologischen Intervention

GROUP	PRE
T	17
T	20
T	16
T	18
T	21
T	17
T	17
T	17
T	18
T	18
T	20
T	17
T	16
T	18
T	16
T	18
T	17
T	14
T	18
T	18
T	20
T	20
T	21
T	19
T	19
T	17
T	20
T	21
T	17
T	17
T	19
T	20
T	22
T	20
T	21
T	20
T	16
T	15
T	15
T	18
T	21
T	17
T	18
T	21
T	17
T	19
T	16
T	17
T	17
T	18
C	22
C	19
C	21
C	18
C	19
C	17
C	20
C	19
C	19
C	19
C	17
C	20
C	18
C	20
C	18
C	18
C	15
C	18
C	17
C	19
C	19
C	19
C	20
C	19
C	20
C	19
C	21
C	19
C	19
C	18
C	20
C	16
C	21
C	19

Tabelle 10.1. BDI-II Werte (PRE) von $n = 100$ Patient:innen in einer Treatment (T) und einer Kontrollstudiengruppe (C) vor Beginn einer psychologischen Intervention

GROUP	PRE
C	20
C	18
C	23
C	18
C	19
C	18
C	21
C	17
C	20
C	21
C	21
C	20
C	18
C	18
C	19
C	19

10.1. Häufigkeitsverteilungen

Wir beginnen mit den Begriffen der *absoluten* und der *relativen Häufigkeitsverteilung*. Für Datensätze mit Werten, in denen der gleiche Wert mehrfach auftritt, können Häufigkeitsverteilungen einen ersten Eindruck von der Beschaffenheit des Datensatzes liefern.

Definition 10.1 (Absolute und relative Häufigkeitsverteilungen). $y := (y_1, \dots, y_n)$ sei ein Datensatz und $W := \{w_1, \dots, w_k\}$ mit $k \leq n$ seien die im Datensatz vorkommenden Werte. Dann heißt die Funktion

$$h : W \rightarrow \mathbb{N}, w \mapsto h(w) := \text{Anzahl der } y_i \text{ aus } y \text{ mit } y_i = w \quad (10.2)$$

die *absolute Häufigkeitsverteilung* der Werte von y und die Funktion

$$r : W \rightarrow [0, 1], w \mapsto r(w) := \frac{h(w)}{n} \quad (10.3)$$

heißt die *relative Häufigkeitsverteilung* der Werte von y .

•

Häufigkeitsverteilungen können in **R** durch Kombination der `table()` und `barplot()` Funktion ausgewertet und dargestellt werden. Folgender **R** Code erzeugt zunächst die absolute und relative Häufigkeitsverteilungen der Pre-Interventions-BDI-II Daten des Beispieldatensatzes.

```
D = read.csv("./_data/110-deskriptive-statistik.csv") # Laden des Datensatzes
y = D$PRE # Double vector der PRE Werte
n = length(y) # Anzahl der Datenwerte (100)
H = as.data.frame(table(y)) # absolute Häufigkeitsverteilung (dataframe)
names(H) = c("w", "ah") # Spaltenbenennung
H$rh = H$ah/n # relative Häufigkeitsverteilung
print(H, digits = 1) # Ausgabe
```

```
  w ah  rh
1 14  1 0.01
2 15  3 0.03
3 16  6 0.06
4 17 17 0.17
5 18 21 0.21
6 19 20 0.20
7 20 17 0.17
8 21 12 0.12
9 22  2 0.02
10 23  1 0.01
```

An der Ausgabe des Dataframes H können wir ablesen, dass im betrachteten Datensatz 10 verschiedene Werte w zwischen 14 und 23 vorkommen, wobei zum Beispiel die Werte 18 und 22 jeweils 21 bzw. 2 mal auftreten. Sowohl die absolute als auch die relative Häufigkeitsverteilung zeigen, dass Werte um 18 im vorliegenden Datensatz gehäuft auftreten, Extremwerte wie 13 oder 23 dagegen eher selten. Man beachte, dass es sich bei der relativen Häufigkeitsverteilung lediglich um eine skalierte Form der absoluten Häufigkeitsverteilung handelt. Qualitativ, also hinsichtlich des gehäuft oder selteneren Vorkommens bestimmter Werte, unterscheiden sich beide Häufigkeitsverteilungen nicht. Dies wird insbesondere auch deutlich, wenn man wie in Abbildung 10.1 beide Häufigkeitsverteilungen visualisiert. Folgender **R** nutzt zu diesem Zweck die Funktion `barplot()`. Die Höhe des über einem Wert w aufgetragenen Balkens entspricht dabei den Funktionswerten $h(w)$ bzw. $r(w)$. Die visuelle Einsicht entspricht natürlich der Inspektion des Dataframes H . Beide Häufigkeitsverteilungen zeigen, dass Werte um 18 im vorliegenden Datensatz gehäuft auftreten, wohingegen niedrigere oder höhere Werte wie 13 bzw. 23 dagegen eher selten sind.

```
pdf(
  file      = "_figures/110-haeufigkeitsverteilungen.pdf",
  width     = 7,
  height    = 4)
par(
  mfcol     = c(1,2),
  las       = 1,
  cex       = 0.7,
  cex.main  = 1.3)
h           = H$h
r           = H$r
names(h)    = H$w
names(r)    = H$w
barplot(h,
  col      = "gray90",
  xlab     = "w",
  ylab     = "h(w)",
  ylim     = c(0,25),
  main     = "Absolute Häufigkeitsverteilung")
barplot(r,
  col      = "gray90",
  xlab     = "w",
  ylab     = "r(w)",
  ylim     = c(0,.25),
  main     = "Relative Häufigkeitsverteilung")
dev.off()
```

PDF Kopiefunktion
Dateiname
Breite (inch)
Höhe (inch)
Abbildungsparameter
Zeilen und Spaltenlayout
x-Tick Orientierung
Annotationsvergrößerung
Titelvergrößerung
h(w) Werte
r(w) Werte
barplot braucht w Werte als names
barplot braucht w Werte als names
Balkendiagramm
absolute Haeufigkeiten
Balkenfarbe
x Achsenbeschriftung
y Achsenbeschriftung
y Achsengrenzen
Titel
Balkendiagramm
absolute Haeufigkeiten
Balkenfarbe
x Achsenbeschriftung
y Achsenbeschriftung
y Achsengrenzen
Titel
Beenden der Abbildungsbearbeitung

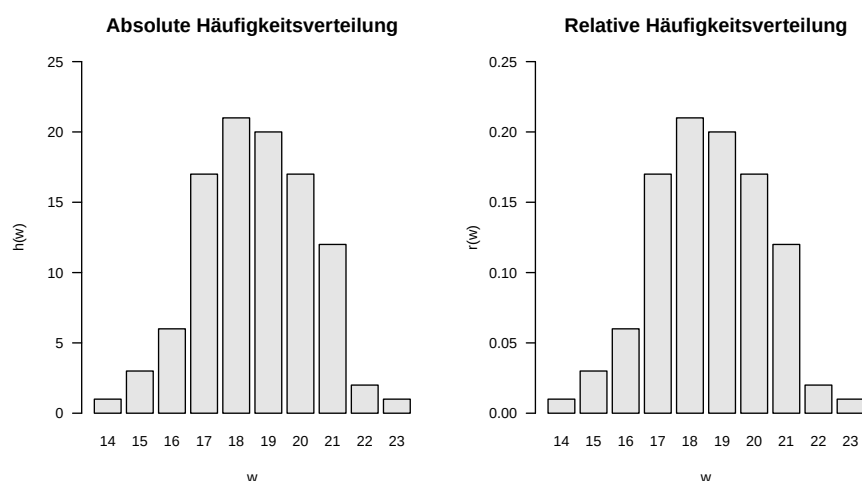


Abbildung 10.1. Absolute und relative Häufigkeitsverteilungen der Pre-Interventions-BDI-II-Werte .

10.2. Histogramme

Wenn in einem Datensatz keine Werte mehrfach vorkommen, nehmen die absoluten und relativen Häufigkeitsverteilungen für jeden dieser Werte den Wert 1 bzw. $1/n$ an und bieten dann keinen Mehrwert über die visuelle Inspektion des Datensatzes hinaus. Die in diesem Abschnitt vorgestellten Histogramme bieten jedoch eine Möglichkeit, auch in diesem Fall visuelle Darstellungen von Häufigkeitsverteilungen zu generieren. Dabei werden verschiedene Werte, die in einem wohldefinierten Sinne “nah beieinander” liegen, zu Klassen zusammengefasst und die absoluten oder relativen Häufigkeiten dieser Klassen bestimmt. Da bei Datensätzen, in denen die gleichen Werte mehrfach vorkommen, diese Werte der Natur nach in der gleichen Klasse liegen, können Histogramme auch in dem in Kapitel 10.1 betrachteten Fall sinnvoll angewendet werden. Histogramme verallgemeinern also die im vorherigen Abschnitt betrachteten Häufigkeitsverteilungen bei Datensätzen mit mehrfach auftretenden Werten.

Wir geben zunächst eine Definition des Histogrammbegriffs. Man beachte, dass in dieser Definition der Vorgang der Klassenbildung, die Bestimmung ihrer absoluten und relativen Häufigkeiten, sowie die Visualisierung dieser Häufigkeiten eng verzahnt sind. Dies ist der Konvention geschuldet, dass der Begriff des Histogramms sowohl eine bestimmte Form der Datenanalyse als auch ihre Visualisierung beschreibt.

Definition 10.2 (Histogramm). Ein *Histogramm* ist ein Diagramm, in dem zu einem Datensatz $y = (y_1, \dots, y_n)$ mit Werten $W := \{w_1, \dots, w_m\}$ mit $m \leq n$ für $j = 1, \dots, k$ über benachbarten Intervallen $[b_{j-1}, b_j[$ mit $b_0 := \min W$ und $b_k := \max W$ Rechtecke mit der Breite $d_j := b_j - b_{j-1}$ und der Höhe $h(w)$ oder $r(w)$ für $w \in [b_{j-1}, b_j[$ abgebildet sind. Dabei werden die Intervalle $[b_{j-1}, b_j[$ Klassen (engl. *bins*) genannt.

•

Natürlich ist das visuelle Erscheinungsbild eines Histogramms stark davon abhängig, wie genau die Werte eines Datensatzes zu Klassen zusammengefasst werden. Der entscheidende Parameter dafür ist nach Definition 10.2 die Anzahl der Klassen k , die zwischen dem Minimalwert b_0 und dem Maximalwert b_k des Datensatzes gewählt werden. Im Folgenden stellen wir einige konventionelle Werte für die Anzahl der Klassen k vor und berechnen und visualisieren sie am Beispiel des Beispieldatensatzes. Dabei gehen wir durchgängig davon aus, dass die benachbarten Intervalle $[b_{j-1}, b_j[$ mit $j = 1, \dots, k$ für die Anzahl von Klassen k mit konstanter Breite $d = b_j - b_{j-1}$ gewählt werden.

Zur Berechnung und Visualisierung von Histogrammen nutzen wir die **R** Funktion `hist()`. Bei dieser werden die Klassen $[b_{j-1}, b_j[$ für $j = 1, \dots, k$ als das Argument `breaks` festgelegt. Speziell ist `breaks` der **R** Vector `c(b_0, b_1, \dots, b_k)` mit Länge $k + 1$. Um den Effekt verschiedener Wahlen der Anzahl der Klassen zu veranschaulichen, lesen wir zunächst den Datensatz ein und bestimmen seine minimalen und maximalen Werte b_0 und b_k .

```
# Datensatz
D = read.csv("../data/110-deskriptive-statistik.csv")
y = D$PRE
b_0 = min(y)
b_k = max(y)

# Laden des Datensatzes
# Pre-Interventions-BDI-II-Daten
# b_0
# b_k
```

```
Minimum b_0      : 14
Maximum b_k      : 23
```

Eine erste Möglichkeit der Wahl der Klassen k besteht darin, möglichst genau eine gewünschte konstante Breite d zu erzielen. Hierzu setzt man

$$k := \lceil (b_k - b_0)d \rceil. \quad (10.4)$$

Man beachte, dass die Anwendung der Aufrundungsfunktion impliziert, dass die so resultierende Klassenbreite höchstens dem gewünschten Wert entspricht, aber auch kleiner sein kann. Folgender **R** Code verdeutlicht diesen Effekt.

```
# Histogramm mit gewünschter Klassenbreite d = 2
d      = 2
b_0    = min(y)
b_k    = max(y)
k      = ceiling((b_k - b_0)/d)
b      = seq(b_0, b_k, length.out = k + 1)

# gewünschte Klassenbreite
# b_0
# b_k
# Anzahl der Klassen
# Klassen [b_{j-1}, b_j[ (breaks)
```

```
Gewünschte Klassenbreite d : 2
Klassenanzahl k          : 5
Intervalle [b_{j-1}, b_j] : 14 15.8 17.6 19.4 21.2 23
Erzielte Klassenbreite d : 1.8
```

Die Wahl einer gewünschten Klassenbreite von $d = 1.5$ dagegen führt für den vorliegenden Datensatz auf eine resultierende Klassenbreite, die dem Wunsch entspricht.

```
# Histogramm mit gewünschter Klassenbreite d = 1.5
d      = 1.5
k      = ceiling((b_k - b_0)/d)
b      = seq(b_0, b_k, length.out = k + 1)

# gewünschte Klassenbreite
# Anzahl der Klassen
# Klassen [b_{j-1}, b_j[ (breaks)
```

```
Gewünschte Klassenbreite d : 1.5
Klassenanzahl k          : 6
Intervalle [b_{j-1}, b_j] : 14 15.5 17 18.5 20 21.5 23
Erzielte Klassenbreite d : 1.5
```

Eine weitere Möglichkeit ist es, die Anzahl der Klassen abhängig von der Anzahl der Datenpunkte im Datensatz zu machen. Die Excelstandardwahl ist dabei

$$k := \lceil \sqrt{n} \rceil. \quad (10.5)$$

Folgender **R** Code implementiert diese Wahl für den den vorliegenden Datensatz.

```
# Histogramm mit Excelstandard
n      = length(y)
k      = ceiling(sqrt(n))
b      = seq(b_0, b_k, length.out = k + 1)

# Anzahl Datenpunkte
# Anzahl der Klassen
# Klassen [b_{j-1}, b_j[ (breaks)
```

```
Anzahl Datenpunkte n      : 100
Klassenanzahl k          : 10
Intervalle [b_{j-1}, b_j] : 14 14.9 15.8 16.7 17.6 18.5 19.4 20.3 21.2 22.1 23
Erzielte Klassenbreite d : 0.9
```

Ähnlich gelagert wie der Excelstandard ist die von Sturges (1926) vorgeschlagene Klassenanzahl, welche die Darstellung unter einer impliziten Normalverteilungsannahme optimiert und durch

$$k := \lceil \log_2 n + 1 \rceil \quad (10.6)$$

gegeben ist. Im vorliegenden Fall resultiert hier:

```
# Klassenanzahl nach Sturges (1926)
n      = length(y)
k      = ceiling(log2(n)+1)
b      = seq(b_0, b_k, length.out = k + 1)

# Anzahl Datenpunkte
# Anzahl der Klassen
# Klassen [b_{j-1}, b_j[ (breaks)
```

```
Anzahl Datenpunkte n      : 100
Klassenanzahl k          : 8
Intervalle [bj-1, bj]    : 14 15.125 16.25 17.375 18.5 19.625 20.75 21.875 23
Erzielte Klassenbreite d : 1.125
```

Eine weitere Möglichkeit ist es, Deskriptivstatistiken des Datensatzes in die Wahl der Klassenanzahl einfließen zu lassen. Als Maß für die Streuung der Daten im Datensatz nutzt die Klassenanzahl nach Scott (1979) die empirische Standardabweichung (cf. Kapitel 10.6), um durch das Histogramm eine Dichteschätzung bei Normalverteilung zu realisieren. Wir wollen den nichtparametrischen Hintergrund dieses Vorgehens hier nicht vertiefen, sondern halten lediglich fest, dass die Wahl der Klassenanzahl nach Scott (1979) durch Festlegung der Klassenbreite zu

$$d := 3.49S/\sqrt[3]{n} \quad (10.7)$$

und die anschließende Wahl der Klassenanzahl durch

$$k := \lceil (b_k - b_0)d \rceil \quad (10.8)$$

gegeben ist, wobei S die Standardabweichung des Datensatzes bezeichnet. Im vorliegenden Fall resultieren hier:

```
# Klassenanzahl nach Scott (1979)
n      = length(y)
S      = sd(y)
d      = ceiling(3.49*S/(n^(1/3)))
k      = ceiling((b_k - b_0)/d)
b      = seq(b_0, b_k, length.out = k + 1)

# Anzahl Datenwerte
# Stichprobenstandardabweichung
# Klassenbreite
# Anzahl der Klassen
# Klassen [b_{j-1}, b_j] (breaks)
```

```
Anzahl Datenpunkte n      : 100
Standardabweichung        : 1.740167
Klassenbreite nach Scott d : 2
Klassenanzahl k          : 5
Intervalle [bj-1, bj]    : 14 15.8 17.6 19.4 21.2 23
Erzielte Klassenbreite d : 1.8
```

Per Default nutzt die in **R** implementierte `hist()` Funktion eine Modifikation der von Sturges (1926) vorgeschlagenen Klassenanzahl und bietet eine Reihe von weiteren Möglichkeiten zur Wahl der Klassen. Einen Überblick dazu liefert `?hist`. Folgender **R** Code visualisiert die absoluten Häufigkeiten der Werte des Beispieldatensatzes anhand der oben spezifizierten Klassen.

```
# Abbildungsparameter
pdf(
  file      = "../figures/t10-histogramme.pdf",
  width     = 6,
  height    = 9)
par(
  mfcol     = c(3,2),
  las       = 1,
  cex       = .7,
  cex.main  = .9)
x_min      = 12
x_max      = 25
y_min      = 0
y_max      = 45
label      = c("", "", "Excel", "Sturges", "Scott")

# Histogramme mit vordefinierter Klassenanzahl und Klassenbreite
for(i in 1:5){
  hist(
    y,
    breaks = bs[[i]],
    xlim   = c(x_min, x_max),
    ylim   = c(y_min, y_max),
    ylab    = "Häufigkeit",
    xlab    = "BDI",
    main    = sprintf(paste(label[i], "k = %.0f, d = %.2f"), ks[i], ds[i]))
}

# R Default Histogramm

# PDF Kopiefunktion
# Dateiname
# Breite (inch)
# Höhe (inch)
# Abbildungsparameter
# Zeilen und Spaltenlayout
# x-Tick Orientierung
# Annotationsvergrößerung
# Titelvergrößerung
# x-Achsenlimit
# x-Achsenlimit
# y-Achsenlimit
# y-Achsenlimit
# Titel

# Histogramm
# Datensatz
# breaks Argument
# x-Achsenlimits
# y-Achsenlimits
# y-Achsenbezeichnung
# x-Achsenbezeichnung
# Titel
```

```

hist(                                     # Histogramm
y,                                       # Datensatz
xlim = c(x_min, x_max),               # x-Achsenlimits
ylim = c(y_min, y_max),               # y-Achsenlimits
ylab = "Häufigkeit",                  # y-Achsenbezeichnung
xlab = "",                             # x-Achsenbezeichnung
main = "R Default")                  # Titel
mtext(LETTERS[6], adj=0, line=2, cex = 1.2, at = 8) # Subplot Buchs
dev.off()                             # Beenden der Abbildungsbearbeitung

```

10.3. Empirische Verteilungsfunktionen

Neben der Frage, wie häufig in einem Datensatz bestimmte Werte oder Klassen von Werten vorkommen, kann es interessant sein, zu untersuchen, wie häufig Werte vorkommen, die kleiner als ein bestimmter Wert sind. Obwohl die Antwort auf diese Frage natürlich in den absoluten und relativen Häufigkeiten bzw. den entsprechenden Histogrammen implizit ist, ist es manchmal von Vorteil, diese Information direkt verfügbar zu haben. Die in diesem Abschnitt eingeführten Begriffe der *kumulativen absoluten und relativen Häufigkeitsverteilung* sowie der *empirischen Verteilungsfunktion* bezeichnen die entsprechenden Analoga zu den in Kapitel 10.1 und Kapitel 10.2 diskutierten Begriffen. Wir beginnen mit der Definition der *kumulativen absoluten* und *kumulativen relativen Häufigkeitsverteilungen*.

Definition 10.3 (Kumulative absolute und relative Häufigkeitsverteilungen). $y = (y_1, \dots, y_n)$ sei ein Datensatz, $W := \{w_1, \dots, w_k\}$ mit $k \leq n$ die im Datensatz vorkommenden Zahlenwerte und h und r die absoluten und relativen Häufigkeitsverteilungen der Werte von y , respektive. Dann heißt die Funktion

$$H : W \rightarrow \mathbb{N}, w \mapsto H(w) := \sum_{w' \leq w} h(w') \quad (10.9)$$

die *kumulative absolute Häufigkeitsverteilung* von y und die Funktion

$$R : W \rightarrow [0, 1], w \mapsto R(w) := \sum_{w' \leq w} r(w') \quad (10.10)$$

die *kumulative relative Häufigkeitsverteilung* der Zahlwerte von y .

•

Im Sinne der Definition Definition 10.1 gilt für die kumulativen Häufigkeitsverteilungen also

$$H(w) = \text{Anzahl der } y_i \text{ aus } y \text{ mit } y_i \leq w \quad (10.11)$$

und

$$R(w) = \text{Anzahl der } y_i \text{ aus } y \text{ mit } y_i \leq w \text{ geteilt durch } n. \quad (10.12)$$

In **R** erlaubt die Funktion `cumsum()` die Berechnung kumulativer Summen und ermöglicht damit die Auswertung kumulativer Häufigkeitsverteilungen, wie folgender **R** Code anhand des Beispieldatensatzes demonstriert.

```

D = read.csv("./_data/110-deskriptive-statistik.csv") # Laden des Datensatzes
y = D$PRE                                             # Pre-Interventions-BDI-II-Daten
n = length(y)                                         # Anzahl der Datenwerte
H = as.data.frame(table(y))                           # absolute Häufigkeitsverteilung als Dataframe
names(H) = c("w", "h")                               # Spaltenbenennung
H$h = cumsum(H$h)                                     # kumulative absolute Häufigkeitsverteilung
H$r = H$h/n                                           # relative Häufigkeitsverteilung
H$R = cumsum(H$r)                                     # kumulative relative Häufigkeitsverteilung
print(H, digits = 1)

```

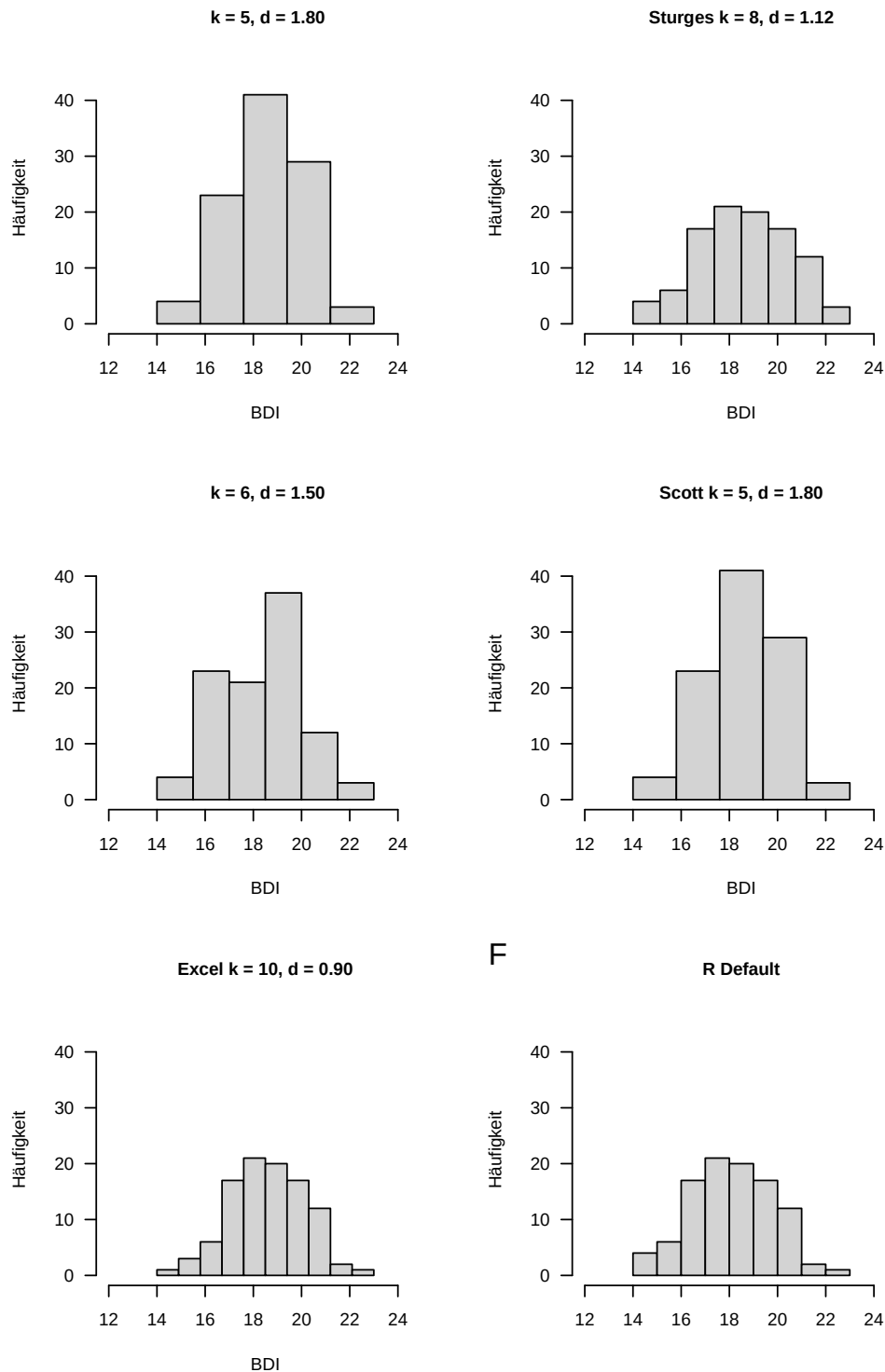


Abbildung 10.2. Histogramme der absoluten Häufigkeiten der Pre-Interventions-BDI-II-Werte. Man beachte, dass der qualitative Eindruck, dass im Datensatz Werte um 18 gehäuft auftreten und niedrigere und höhere Werte seltener sind, über alle Wahlen der Klassenanzahl konstant ist. Das genaue quantitative Erscheinungsbild ist dagegen von der exakten Wahl der Klassenanzahl und Klassenbreite abhängig.

	w	h	H	r	R
1	14	1	1	0.01	0.01
2	15	3	4	0.03	0.04
3	16	6	10	0.06	0.10
4	17	17	27	0.17	0.27
5	18	21	48	0.21	0.48
6	19	20	68	0.20	0.68
7	20	17	85	0.17	0.85
8	21	12	97	0.12	0.97
9	22	2	99	0.02	0.99
10	23	1	100	0.01	1.00

Man beachte dabei insbesondere die Gegenüberstellung von h und H sowie von r und R , die die kumulativen Summen in Definition 10.3 verdeutlichen. Folgender **R** Code ermöglicht die Darstellung der so bestimmten kumulativen Häufigkeitsverteilungen wie in Abbildung 10.3 gezeigt. Offensichtlich ist die absolute und die relative Häufigkeit von Werten kleiner als der kleinste Wert im Datensatz Null. Die absolute und relative Häufigkeit von Werten kleiner als der größte Wert im Datensatz ist dagegen n bzw. $n/n = 1$.

```
pdf(
  file      = "../figures/110-kumulative-haeufigkeitsverteilungen.pdf",
  width     = 10,
  height    = 5,
  Hw        = H$H,
  Rw        = H$R,
  names(Hw) = H$w,
  names(Rw)  = H$w,
  par(
    mfcol = c(1,2),
    las   = 1,
    cex   = 1)
  barplot(
    Hw,
    col = "gray90",
    xlab = "w",
    ylab = "H(w)",
    ylim = c(0,110),
    las = 1,
    main = "Absolute kumulative Häufigkeit PRE")
  barplot(
    Rw,
    col = "gray90",
    xlab = "w",
    ylab = "R(w)",
    ylim = c(0,1),
    las = 1,
    main = "Relative kumulative Häufigkeit PRE")
  dev.off()
```

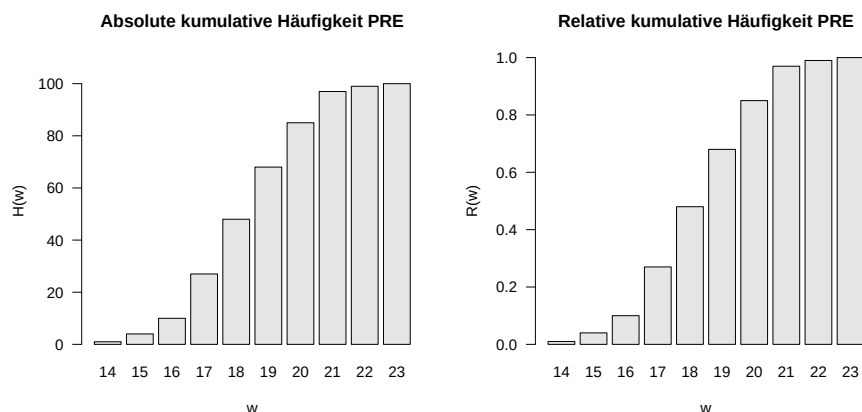


Abbildung 10.3. Absolute und relative kumulative Häufigkeitsverteilungen der Pre-Interventions-BDI-II-Werte .

Das kumulative Analogon zu einem Histogramm ist die *empirische Verteilungsfunktion*. Im Unterschied zu einem Histogramm ist für die Bestimmung einer empirischen Verteilungsfunktion allerdings keine Definition von Klassen erforderlich. An die Stelle der Intervalleinteilung der reellen Zahlen zwischen dem Minimum und dem Maximum eines Datensatz

treten bei der empirischen Verteilungsfunktion einfach die reellen Zahlen, wie folgende Definition verdeutlicht.

Definition 10.4 (Empirische Verteilungsfunktion). $y = (y_1, \dots, y_n)$ sei ein Datensatz. Dann heißt die Funktion

$$F : \mathbb{R} \rightarrow [0, 1], x \mapsto F(x) := \frac{\text{Anzahl der } y_i \text{ aus } y \text{ mit } y_i \leq x}{n} \quad (10.13)$$

die empirische Verteilungsfunktion (EVF) von y .

•

Die empirische Verteilungsfunktion wird manchmal auch die *empirische kumulative Verteilungsfunktion* genannt. Allgemein beschränkt man sich bei der empirischen Verteilungsfunktion wie in Definition 10.4 auf die Angabe der relativen Häufigkeiten. Natürlichweise kommen in Datensätzen nicht alle reellen Zahlen vor. Dies impliziert, dass die relative Häufigkeit, die reellen Zahlen zugeordnet wird, die zwischen zwei benachbarten Werten des Datensatzes liegen, derjenigen des kleineren der beiden Werte entspricht. Typischerweise sind empirische Verteilungsfunktionen also Treppenfunktionen. In **R** kann die Funktion `ecdf()` zur Evaluation und Visualisierung der empirischen Verteilungsfunktion eingesetzt werden, wie folgender **R** Code anhand des Beispieldatensatzes demonstriert und in Abbildung 10.4 visualisiert ist. Man liest aus Abbildung 10.4 zum Beispiel ab, dass die relative Häufigkeit von Werten kleiner als 17.9 etwa 0.3 ist. Die gleiche relative Häufigkeit haben Werte kleiner als 17.8.

```
D = read.csv("../data/110-deskriptive-statistik.csv")
y = D$PRE
evf = ecdf(y)
pdf(
  file = "../figures/110-ecdf.pdf",
  width = 8,
  height = 5)
plot(
  evf,
  xlab = "BDI",
  ylab = "Relative Häufigkeit",
  main = "Empirische Verteilungsfunktion")
dev.off()
```

```
# Laden des Datensatzes
# Pre-Interventions-BDI-II-Daten
# Evaluation der EVF
# PDF Kopiefunktion
# Dateiname
# Breite (inch)
# Höhe (inch)
# plot() weiß mit ecdf object umzugehen
# ecdf Objekt
# x-Achsenbeschriftung
# y-Achsenbeschriftung
# Titel
# Beenden der Abbildungsbearbeitung
```

10.4. Quantile und Boxplots

Die empirische Verteilungsfunktion gibt eine zumindest graphisch dargestellte Antwort auf die Frage, wie groß der relative Anteil der Werte eines Datensatzes ist, die kleiner als ein gegebener Wert sind. Die in diesem Abschnitt eingeführten Quantile können als Inversion dieser Datensatzinterrogation verstanden werden. Speziell gibt man bei Quantilen einen relativen Anteil $0 \leq p \leq 1$ der Werte eines Datensatzes vor und fragt, welcher Wert des Datensatzes so beschaffen ist, dass eben $p \cdot 100\%$ der Werte des Datensatzes kleiner als oder gleich diesem Wert sind. Wir nutzen folgende Definition des Begriffs des *p-Quantils*.

Definition 10.5 (*p*-Quantil). $y = (y_1, \dots, y_n)$ sei ein Datensatz und

$$y_s = (y_{(1)}, y_{(2)}, \dots, y_{(n)}) \text{ mit } \min_{1 \leq i \leq n} y_i = y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)} = \max_{1 \leq i \leq n} y_i \quad (10.14)$$

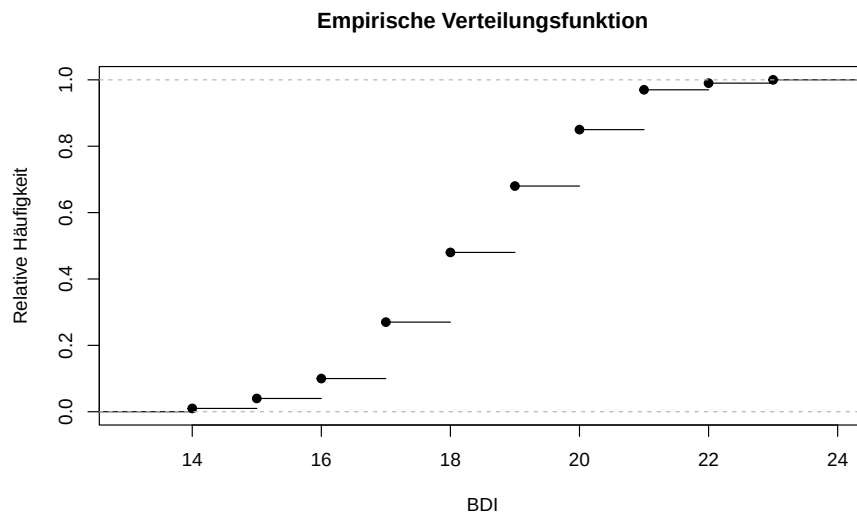


Abbildung 10.4. Empirische Verteilungsfunktion der PRE Werte.

sei der zugehörige aufsteigend sortierte Datensatz. Weiterhin bezeichne $\lfloor \cdot \rfloor$ die Abrundungsfunktion. Dann heißt für ein $p \in [0, 1]$ die Zahl

$$y_p := \begin{cases} y_{(\lfloor np+1 \rfloor)} & \text{falls } np \neq \mathbb{N} \\ \frac{1}{2} (y_{(np)} + y_{(np+1)}) & \text{falls } np \in \mathbb{N} \end{cases} \quad (10.15)$$

das p -Quantil des Datensatzes y .

•

Man beachte, dass das p -Quantil y_p nach Definition 10.5 entweder ein Wert des Datensatzes oder ein Wert zwischen zwei Werten des Datensatzes ist. Nach Konstruktion von y_p gilt dabei insbesondere, dass mindestens $p \cdot 100\%$ aller Werte des Datensatzes kleiner oder gleich y_p und mindestens $(1-p) \cdot 100\%$ aller Werte des Datensatzes größer als y_p sind. Das p -Quantil teilt den geordneten Datensatz y_s also im Verhältnis p zu $(1-p)$ auf. Je nach Wahl von p erhalten die p -Quantile spezielle Bezeichnungen:

- $y_{0.25}, y_{0.50}, y_{0.75}$ heißen *unteres Quartil*, *Median* und *oberes Quartil*, respektive,
- $y_{j \cdot 0.10}$ für $j = 1, \dots, 9$ heißen *Dezile* und
- $y_{j \cdot 0.01}$ für $j = 1, \dots, 99$ heißen *Percentile*.

Um den Begriff des p -Quantils zu verdeutlichen, betrachten wir ein an Henze (2018), Kapitel 5 orientiertes Beispiel. Dazu seien zunächst der in untenstehender Tabelle dargestellte Datensatz und seine aufsteigend sortierte Version gegeben.

i	1	2	3	4	5	6	7	8	9	10
y_i	8.5	1.5	75	4.5	6.0	3.0	3.0	2.5	6.0	9.0
$y_{(i)}$	1.5	2.5	3.0	3.0	4.5	6.0	6.0	8.5	9.0	75

Wir wollen als erstes das untere Quartil, also das 0.25-Quantil, dieses Datensatzes bestimmen. Es ist offenbar $n = 10$ und wir haben $p := 0.25$ gewählt. Basierend auf Definition 10.5 finden wir, dass

$$np = 10 \cdot 0.25 = 2.5 \notin \mathbb{N}. \quad (10.16)$$

Also gilt nach Definition 10.5, dass

$$y_{0.25} = x_{(\lfloor 2.5+1 \rfloor)} = x_{(3)} = 3.0. \quad (10.17)$$

Das untere Quartil des betrachteten Datensatzes ist also der Wert 3.0 des Datensatzes.

Als zweites wollen wir das 0.80-Quantil des Datensatzes bestimmen. Es ist offenbar weiterhin $n = 10$ und wir haben nun $p := 0.80$ gewählt. Wir finden in diesem Fall, dass

$$np = 10 \cdot 0.80 = 8 \in \mathbb{N}. \quad (10.18)$$

Also folgt nach Definition 10.5

$$y_{0.80} = \frac{1}{2} (y_{(8)} + y_{(8+1)}) = \frac{1}{2} (y_{(8)} + y_{(9)}) = \frac{8.5 + 9.0}{2} = 8.75. \quad (10.19)$$

Das $y_{0.80}$ -Quantil des betrachteten Datensatzes ist mit 8.75 also der Wert zwischen den Werten 8.5 und 9 des Datensatzes. In **R** wertet man die so bestimmten Quantile wie folgt aus.

```
y = c(8.5,1.5,12,4.5,6.0,3.0,3.0,2.5,6.0,9.0)
n = length(y)
y_s = sort(y)
p = 0.25
y_p = y_s[floor(np + 1)]
cat("0.25 Quantil: ", round(y_p,2))
```

```
# Beispieldaten
# Anzahl Datenwerte
# sortierter Datensatz
# np \notin \mathbb{N}
# 0.25 Quantil
# Ausgabe
```

```
0.25 Quantil: 3
```

```
p = 0.80
y_p = (1/2)*(y_s[np] + y_s[np + 1])
cat("0.80 Quantil: ", round(y_p,2))
```

```
# np \in \mathbb{N}
# 0.80 Quantil
# Ausgabe
```

```
0.80 Quantil: 8.75
```

Mit `quantile()` stellt **R** auch eine Funktion bereit, die die Quantilbestimmung eines Datensatzes automatisiert. Dabei ist allerdings zu beachten, dass die in Definition 10.5 gegebene Definition des Quantilsbegriffs nur eine von mindestens neun verschiedenen Quantildefinitionen ist und die genauen Werte sich von Definition zu Definition unterscheiden können. Die spezielle Quantildefinition, die von `quantile()` genutzt wird, wird über das Funktionsargument `type` ausgewählt. Auf theoretischer Ebene geben Hyndman & Fan (1996) dazu einen Überblick, auf Implementationsebene `?quantile`. Wir betrachten exemplarisch die Bestimmung des 0.25-Quantils Pre-Interventions BDI Beispieldatensatzes anhand von `type = 8`.

```
D = read.csv("./_data/i10-deskriptive-statistik.csv")
y = D$PRE
y_p = quantile(y, 0.25, type = 8)
```

```
# Laden des Datensatzes
# Pre-Interventions-BDI-II-Daten
# 0.25 Quantil, Definition 1
```

```
0.25 Quantil (Type = 8): 17
```

Ein *Boxplot* schließlich visualisiert eine Quantil-basierte Zusammenfassung eines Datensatzes, wobei typischerweise das Minimum und das Maximum eines Datensatzes sowie das untere Quartil, der Median und das obere Quartil visualisiert werden. In **R** erstellt man einen Boxplot mithilfe der `boxplot()` Funktion, wie folgender **R** Code anhand des Pre-Interventions-BDI-II-Datensatzes demonstriert.

```

D      = read.csv("./_data/110-deskriptive-statistik.csv")      # Laden des Datensatzes
pdf(    # PDF Kopiefunktion
  file  = "./_figures/110-boxplot.pdf",                        # Dateiname
  width = 8,                                                    # Breite (inch)
  height = 5)                                                  # Höhe (inch)
boxplot(D$PRE,                                                # Boxplot
  horizontal = T,                                             # Datensatz
  range = 0,                                                 # horizontale Darstellung
  xlab = "BDI",                                              # Whiskers bis zu min y und max y
  main = "Boxplot",                                         # x Achsenbeschriftung
  dev.off())                                                 # Titel
                                                         # Beenden der Abbildungsbearbeitung

```

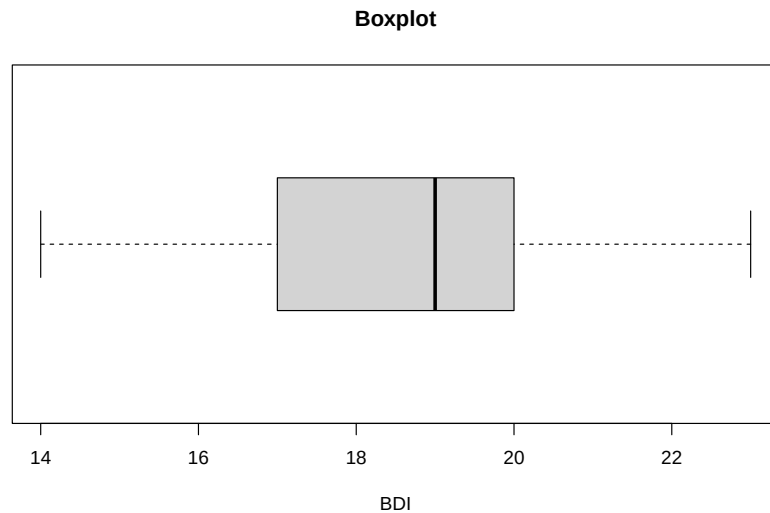


Abbildung 10.5. Boxplot der Pre-Interventions-BDI-Werte des Beispieldatensatzes.

In [Abbildung 10.5](#) werden $\min y$ und $\max y$ als sogenannte “Whiskerendpunkte” dargestellt, $y_{0.25}$ und $y_{0.75}$ bilden die unteren und oberen Grenze der zentralen grauen Box und der Median $y_{0.50}$ wird als Strich in der zentralen grauen Box abgebildet. Auch hier gilt, dass es sehr viele Boxplotvariationen gibt, vgl. McGill et al. (1978). Eine genaue Erläuterung der graphisch dargestellten Datensatzeigenschaften ist also folglich immer nötig. Schließlich sei angemerkt, dass die Darstellung von Datensätzen mithilfe von Boxplots in den letzten zwanzig Jahren etwas aus der Mode gekommen ist und man sie nur noch sehr selten in aktuellen wissenschaftlichen Publikationen findet.

10.5. Maße der Zentralen Tendenz

In diesem Abschnitt betrachten wir einige grundlegende Kennzahlen von Datensätzen, die für die deskriptive Statistik zentral sind und Aussagen zu dem “durchschnittlichen” Wert eines Datensatzes machen. Wir beginnen mit der Definition des Mittelwerts.

10.5.1. Mittelwert

Definition 10.6 (Mittelwert). $y = (y_1, \dots, y_n)$ sei ein Datensatz. Dann heißt

$$\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i \quad (10.20)$$

der *Mittelwert* von y .

•

Der Mittelwert eines Datensatzes ist also die Summe der Datenwerte geteilt durch die Anzahl der Datenwerte. Der hier definierte Mittelwert wird auch als *arithmetisches Mittel* oder *Durchschnitt* bezeichnet. Im Kontext der Frequentistischen Inferenz werden wir den Mittelwert vor dem Hintergrund des Modells der Zufallsstichprobe als *Stichprobenmittel* bezeichnen. Insbesondere erlauben die Modelle der Frequentistischen Inferenz auch, die funktionale Form des Mittelwertes tiefer zu begründen, wie wir an gegebener Stelle sehen werden. Mithilfe der Summenfunktion berechnet man den Mittelwert eines Datensatzes in **R** wie folgt.

```
D = read.csv("../data/110-deskriptive-statistik.csv") # Laden des Datensatzes
y = D$PRE # Pre-Interventions-BDI-II-Daten
n = length(y) # Anzahl der Werte
y_bar = (1/n)*sum(y) # Mittelwertberechnung
```

Mittelwert der Pre-BDI-Interventions Daten 18.61

Mit der `mean()` Funktion stellt **R** natürlich auch eine Funktion zur automatisierten Berechnung des Mittelwerts bereit.

```
cat("Mittelwert der PRE-BDI-II-Daten ", mean(y)) # Ausgabe
```

Mittelwert der PRE-BDI-II-Daten 18.61

In folgendem Theorem halten wir zunächst zwei Eigenschaften des Mittelwerts eines Datensatzes fest, die seine Einordnung als Maß der zentralen Tendenz eines Datensatzes plausibel machen.

Theorem 10.1 (Zentralitätseigenschaften des Mittelwerts). $y = (y_1, \dots, y_n)$ sei ein Datensatz und $\bar{y}$ sei der Mittelwert von y . Dann gelten

(1) Die Summe der Abweichungen vom Mittelwert ist Null,

$$\sum_{i=1}^n (y_i - \bar{y}) = 0. \quad (10.21)$$

(2) Die absoluten Summen negativer und positiver Abweichungen vom Mittelwert sind gleich, d.h. wenn $j = 1, \dots, n_j$ die Datenpunktindizes mit $(y_j - \bar{y}) < 0$ und $k = 1, \dots, n_k$ die Datenpunktindizes mit $(y_k - \bar{y}) \geq 0$ bezeichnen, dann gilt mit $n_j + n_k$, dass

$$\left| \sum_{j=1}^{n_j} (y_j - \bar{y}) \right| = \left| \sum_{k=1}^{n_k} (y_k - \bar{y}) \right|. \quad (10.22)$$

◦

Beweis. (1) Es gilt

$$\begin{aligned}
 \sum_{i=1}^n (y_i - \bar{y}) &= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} \\
 &= \sum_{i=1}^n y_i - n\bar{y} \\
 &= \sum_{i=1}^n y_i - \frac{n}{n} \sum_{i=1}^n y_i \\
 &= \sum_{i=1}^n y_i - \sum_{i=1}^n y_i \\
 &= 0.
 \end{aligned} \tag{10.23}$$

(2) Seien $j = 1, \dots, n_j$ die Indizes mit $(y_j - \bar{y}) < 0$ und $k = 1, \dots, n_k$ die Indizes mit $(y_k - \bar{y}) \geq 0$, so dass $n = n_j + n_k$. Dann gilt

$$\begin{aligned}
 \sum_{i=1}^n (y_i - \bar{y}) &= 0 \\
 \Leftrightarrow \sum_{j=1}^{n_j} (y_j - \bar{y}) + \sum_{k=1}^{n_k} (y_k - \bar{y}) &= 0 \\
 \Leftrightarrow \sum_{k=1}^{n_k} (y_k - \bar{y}) &= -\sum_{j=1}^{n_j} (y_j - \bar{y}) \\
 \Leftrightarrow \left| \sum_{j=1}^{n_j} (y_j - \bar{y}) \right| &= \left| \sum_{k=1}^{n_k} (y_k - \bar{y}) \right|.
 \end{aligned} \tag{10.24}$$

□

Nach Theorem 10.1 liegt der Mittelwert eines Datensatzes also gerade so, dass sich in Summe die Abweichungen der Datenpunkte von diesem Wert zu Null ausgleichen und dass sich positive und negative Abweichungen vom Mittelwert die Waage halten. In **R** kann man Theorem 10.1 am Beispieldatensatz wie folgt verifizieren.

```

# Datensatz
D = read.csv("../data/110-deskriptive-statistik.csv")
y = D$PRE
s = sum(y - mean(y))
s_1 = sum(y[y <= mean(y)] - mean(y))
s_2 = sum(y[y > mean(y)] - mean(y))

# Laden des Datensatzes
# Pre-Interventions-BDI-II-Daten
# Summe der Abweichungen vom Mittelwert
# Summe aller negativen Abweichungen
# Summe aller positiven Abweichungen

```

```

Summe der Abweichungen vom Mittelwert      : 0
Summen der positiven und negativen Abweichungen : -71.28 71.28

```

Theorem 10.2 (Summationsseigenschaften des Mittelwerts). *Gegeben seien zwei Datensätze $y = (y_1, \dots, y_n)$ und $z = (z_1, \dots, z_n)$ gleichen Umfangs und es sei durch*

$$y + z := (y_1 + z_1, \dots, y_n + z_n) \tag{10.25}$$

die Summe dieser Datensätze definiert. Weiterhin seien $\bar{y}$ und $\bar{z}$ die Mittelwerte der Datensätze y und z , respektive. Dann gilt

$$\overline{y + z} = \bar{y} + \bar{z} \tag{10.26}$$

◦

Beweis. Es gilt

$$\begin{aligned}
 \overline{y+z} &:= \frac{1}{n} \sum_{i=1}^n (y_i + z_i) \\
 &= \frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{n} \sum_{i=1}^n z_i \\
 &=: \bar{y} + \bar{z}.
 \end{aligned} \tag{10.27}$$

□

Betrachtet man die Summe zweier Datensätze, so ist es für die Bestimmung ihres Mittelwerts also unerheblich, ob man zunächst die Datensätze aufaddiert und dann den Mittelwert des summierten Datensatzes bestimmt oder ob man für jeden der beiden Datensätze den Mittelwert bestimmt und diese aufaddiert. Wir verifizieren dieses Ergebnis exemplarisch mithilfe der Pre- und Post-Interventions BDI-II-Daten des Beispieldatensatzes in **R** wie folgt.

```

D      = read.csv("../data/110-deskriptive-statistik.csv")
y      = D$PRE
y_bar  = mean(y)
z      = D$POS
z_bar  = mean(z)
yz_bar_1 = mean(y + z)
yz_bar_2 = y_bar + z_bar
# Laden des Datensatzes
# Pre-Interventions-BDI-II-Daten
# Mittelwert der Pre-Interventions-BDI-II-Daten
# Post-Interventions BDI-II-Daten
# Mittelwert der Post-Interventions BDI-II-Daten
# Mittelwert der summierten Pre- und Post-Daten
# Summe der Mittelwerte der Pre- und Post-Date

```

```

Mittelwert von z      : 31.68
Summe der Mittelwerte von x und y : 31.68

```

Theorem 10.3 (Mittelwert bei linear-affiner Transformation). $y = (y_1, \dots, y_n)$ sei ein Datensatz und $\bar{y}$ sei der Mittelwert von y . Weiterhin sei für $a, b \in \mathbb{R}$ mit

$$ay + b := (ay_1 + b, \dots, ay_n + b) \tag{10.28}$$

ein linear-affin transformierter Datensatz von y definiert. Dann gilt, dass

$$\overline{ay + b} = a\bar{y} + b \tag{10.29}$$

◦

Beweis. Es gilt

$$\begin{aligned}
 \overline{ay + b} &:= \frac{1}{n} \sum_{i=1}^n (ay_i + b) \\
 &= \sum_{i=1}^n \left(\frac{1}{n} ay_i + \frac{1}{n} b \right) \\
 &= \left(\sum_{i=1}^n \frac{1}{n} ay_i \right) + \left(\sum_{i=1}^n \frac{1}{n} b \right) \\
 &= a \left(\frac{1}{n} \sum_{i=1}^n y_i \right) + \frac{1}{n} \sum_{i=1}^n b \\
 &= a\bar{y} + \frac{1}{n} nb \\
 &= a\bar{y} + b.
 \end{aligned} \tag{10.30}$$

□

Eine linear-affine Transformation eines Datensatz transformiert den Mittelwert des Datensatzes also linear-affin. Dies kann man sich zunutze machen, wenn man sich überlegt, wie der Mittelwert lauten sollte, wenn man einen Datensatz neu skaliert, also z.B. von Gramm in Kilogramm umrechnet. Natürlich kann man zur Sicherheit aber auch einfach den Mittelwert des neu skalierten Datensatzes bestimmen. Wir demonstrieren diese Eigenschaft exemplarisch durch Reskalierung des Pre-Interventions-BDI-II-Datensatzes mit dem Wert $a = 2$ unter gleichzeitiger Addition der Konstante $b := 5$.

```
D = read.csv("../data/110-deskriptive-statistik.csv") # Laden des Datensatzes
y = D$PRE # Pre-Interventions-BDI-II-Daten
y_bar = mean(y) # Mittelwert der Pre-Interventions-BDI-II-Daten
a = 2 # Multiplikationskonstante
b = 5 # Additionskonstante
z = a*y + b # linear-affine Transformation der PRE Daten
z_bar = mean(z) # Mittelwert der transformierten PRE Daten
ay_bar_b = a*y_bar + b # Transformation des PRE Daten Mittelwerts
```

```
Mittelwert des transformierten Datensatzes : 42.22
Transformation des Mittelwertes : 42.22
```

10.5.2. Median

In den oben definierten Mittelwert geht jeder Wert eines Datensatzes summativ mit gleichem Gewicht $1/n$ ein. Dies betrifft insbesondere auch solche Werte, die im Vergleich zu den übrigen Werten des Datensatzes sehr untypisch sind. Eine Möglichkeit, die "Mitte" eines Datensatzes unabhängig von solchen untypischen Werten zu bestimmen, besteht in der Bestimmung seines Medians, den wir nachfolgend definieren. Wir haben den Median bereits in Abschnitt Kapitel 10.4 als das 0.5-Quantil kennengelernt, mit dem er identisch ist.

Definition 10.7 (Median). $y = (y_1, \dots, y_n)$ sei ein Datensatz und $y_s = (y_{(1)}, \dots, y_{(n)})$ der zugehörige aufsteigend sortierte Datensatz. Dann ist der Median von y definiert als

$$\tilde{y} := \begin{cases} y_{((n+1)/2)} & \text{falls } n \text{ ungerade,} \\ \frac{1}{2} (y_{(n/2)} + y_{(n/2+1)}) & \text{falls } n \text{ gerade.} \end{cases} \quad (10.31)$$

•

Definition 10.7 impliziert, dass mindestens 50% aller Datenwerte y_i des Datensatzes kleiner oder gleich $\tilde{y}$ sind und dass gleichzeitig mindestens 50% aller Datenwerte y_i des Datensatzes größer oder gleich $\tilde{y}$ sind. Anstelle eines formalen Beweises dieser Aussagen verweisen wir auf Abbildung 10.6. Wie in der Definition des Medians unterscheiden wir dort die Fälle, dass die Anzahl der Datenpunkte ungerade ist (hier $n = 5$, Abbildung 10.6 A) oder gerade ist ($n = 6$, Abbildung 10.6 B). Im Fall der ungeraden Anzahl an Datenpunkten ist der Median der Datenwert mit der Indexzahl, die in der Mitte der Indexzahlen des sortierten Datensatzes liegt, im Beispiel Abbildung 10.6 A also der Wert $y_{(3)}$. Damit sind im vorliegenden Fall 3/5, also 60% der Datenwerte (speziell $y_{(1)}$, $y_{(2)}$ und $y_{(3)}$) kleiner oder gleich dem Median, gleichzeitig aber auch 3/5, also 60% der Datenwerte (speziell $y_{(3)}$, $y_{(4)}$ und $y_{(5)}$) größer oder gleich dem Median. Im Fall einer geraden Anzahl von Datenpunkten ist die Lage des Medians etwas intuitiver: Der Median liegt in der Mitte der beiden mittleren Werte des sortierten Datensatzes (in Abbildung 10.6 B speziell in der Mitte der Werte $y_{(3)}$ und $y_{(4)}$). Damit sind dann 3/6, also 50% der Datenwerte (speziell

A

Beispiel für n ungerade

$$n := 5 \Rightarrow \left(\frac{5+1}{2}\right) = (3) \Rightarrow \tilde{y} := y_{(3)}$$

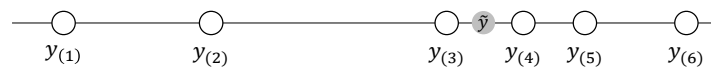


$$y_{(1)}, y_{(2)}, y_{(3)} \leq \tilde{y} \leq y_{(3)}, y_{(4)}, y_{(5)}$$

B

Beispiel für n gerade

$$n := 6 \Rightarrow \left(\frac{6}{2}\right) = (3), \left(\frac{6}{2} + 1\right) = (4) \Rightarrow \tilde{y} := \frac{1}{2}(y_{(3)} + y_{(4)})$$



$$y_{(1)}, y_{(2)}, y_{(3)} < \tilde{y} < y_{(4)}, y_{(5)}, y_{(6)}$$

Abbildung 10.6. Bestimmung und Eigenschaften des Medians.

$y_{(1)}$, $y_{(2)}$ und $y_{(3)}$ kleiner als der Median und $3/6$, also 50% der Datenwerte (speziell $y_{(4)}$, $y_{(5)}$ und $y_{(6)}$) größer als der Median.

Folgender **R** Code bestimmt den Median des Beispieldatensatzes anhand von Definition 10.7

```
D = read.csv("./_data/110-deskriptive-statistik.csv") # Laden des Datensatzes
y = D$PRE # Pre-Interventions-BDI-II-Daten
n = length(y) # Anzahl der Werte
y_s = sort(y) # aufsteigend sortierter Vektor
if(n %% 2 == 1){ # n ungerade, n mod 2 == 1
  y_tilde = y_s[(n+1)/2] # Medianformel, Fall n ungerade
} else { # n gerade, n mod 2 == 0
  y_tilde = (y_s[n/2] + y_s[n/2 + 1])/2 # Medianformel, Fall n gerade
```

Median der PRE-BDI-II-Daten: 19

Natürlich stellt **R** mit `median()` auch eine Funktion zur direkten Bestimmung, wie folgender **R** Code demonstriert.

```
cat("Median der PRE-BDI-II-Daten: ", median(y)) # Ausgabe
```

Median der PRE-BDI-II-Daten: 19

Folgende Simulation zeigt beispielhaft, dass der Median als Maß der Mitte eines Datensatzes weniger anfällig für Ausreißer ist als der Mittelwert.

```
D = read.csv("./_data/110-deskriptive-statistik.csv") # Laden des Datensatzes
y = D$PRE # Pre-Interventions-BDI-II-Daten
```

Mittelwert und Median für Datensatz ohne Ausreißer: 18.61 19

```

z      = y
z[1]   = 10000                                # neuer simulierter Datensatz
                                              # ... mit einem Extremwert

```

Mittelwert und Median für Datensatz mit Ausreißer: 118.44 19

Obwohl der Median also als robusteres Maß für die zentrale Tendenz eines Datensatzes als der Mittelwert erscheint, ist die Bestimmung von Mittelwerten sicherlich das üblichere Vorgehen. Dies ist einerseits dadurch begründet, dass sich der Mittelwert aufgrund seiner mathematisch weniger komplexen Definition einfacher im analytischen Kontext zur Gestaltung von probabilistischen Modellen eignet. Andererseits ist das Auftreten von Extremwerten üblicherweise ein Anzeichen von Datenerhebungsfehlern (z.B. sollte eine BDI-II Wert von 10.000 schlicht nicht vorkommen) oder großen Unterschieden zwischen den betrachteten experimentellen Einheiten, die durch visuelle Inspektion des Datensatzes meist vor der Bestimmung von Maßen der zentralen Tendenz auffallen. Es gibt allerdings ein ganzes Unterfeld der Statistik, dass sich mit - im weitesten Sinne - Ausreißer unabhängigen Maßen beschäftigt, vgl. z.B. Huber (1981).

10.5.3. Modalwert

Wenn in einem Datensatz manche Werte wiederholt auftreten, ist mit dem *Modalwert* als dem in einem Datensatz am häufigsten vorkommenden Wert ein drittes Maß der zentralen Tendenz gegeben. Wir nutzen folgende Definition.

Definition 10.8 (Modalwert). $y := (y_1, \dots, y_n)$ mit $y_i \in \mathbb{R}$ sei ein Datensatz, $W := \{w_1, \dots, w_k\}$ mit $k \leq n$ seien die im Datensatz vorkommenden verschiedenen Zahlenwerte und $h : W \rightarrow \mathbb{N}$ sei die absolute Häufigkeitsverteilung der Zahlwerte von y . Dann ist der *Modalwert* (oder *Modus*) von y definiert als

$$\operatorname{argmax}_{w \in W} h(w), \quad (10.32)$$

also als der im Datensatz am häufigsten vorkommende Wert.

•

In **R** kann man den Modalwert eines entsprechenden Datensatzes wie folgt bestimmen.

```

D      = read.csv("./_data/110-deskriptive-statistik.csv")      # Laden des Datensatzes
y      = D$PRE                                                    # Pre-Interventions-BDI-II-Daten
n      = length(y)                                                # Anzahl der Datenwerte (100)
H      = as.data.frame(table(y))                                  # absolute Häufigkeitsverteilung (dataframe)
names(H) = c("a", "w")                                           # konsistente Benennung
mode   = H$w[which.max(H$w)]                                     # Modalwert

cat("Modalwert der PRE-BDI-II-Daten: ", as.numeric(as.vector(mode))) # Ausgabe als numeric vector, nicht factor

```

Modalwert der PRE-BDI-II-Daten: 21

10.5.4. Datenverteilungsformen und Maße zentraler Tendenz

Wie sinnvoll es ist, mit dem Mittelwert, dem Median oder dem Modalwert eine Angabe über die zentrale Tendenz eines Datensatzes zu geben, hängt stark von der Beschaffenheit des Datensatzes ab. Hat man beispielsweise einen Datensatz hochaufgelöster Daten vorliegen, in dem kein Datenwert mehrfach auftaucht, eignet sich der Modalwert sicherlich nicht, um eine quantitative Aussage über die zentrale Tendenz des Datensatzes zu machen. Man sollte die Angabe eines Maßes der zentralen Tendenz also immer im Kontext der Gesamtbeschaffenheit des Datensatzes sowie, wenn zutreffend, des inferenzstatistischen Modells, das man zur Interrogation des Datensatzes nutzt, betrachten. In Abbildung 10.7 geben wir vier Beispiele, wie sich die oben betrachteten Maße der zentralen Tendenz in verschiedenen Szenarien der Datenwertverteilung verhalten können. Diese vier Szenarien sind natürlich nicht erschöpfend und jeder Datensatz ist in Hinblick auf ein sinnvolles Maß der zentralen Tendenz kritisch zu betrachten.

Abbildung 10.7 A visualisiert die Lage von Mittelwert, Median und Modalwert in einem Datensatz, der sich symmetrisch um einen Wert von etwa $y = 50$ verteilt und für den Werte nah bei diesem zentralen Wert häufig vorkommen und Werte fern von diesem zentralen Wert eher selten. Wie wir später sehen, entspricht dies den typischen Charakteristika normalverteilter Datenwerte. In diesem Fall liegen Mittelwert, Median und Modalwert recht nahe beieinander und jedes dieser Maße beschreibt in der Tat die zentrale Tendenz des Datensatzes.

Abbildung 10.7 B visualisiert die Lage von Mittelwert, Median und Modalwert in einem Datensatz, in dem die Datenwerte um zwei Werte von etwa $y = 25$ und $y = 75$ gehäuft auftreten und jeweils in positive und negative Richtung von diesen Werten entfernt eher selten auftreten. So liegen zum Beispiel um $y = 50$ recht wenige Datenpunkte dieses Datensatzes. Mittelwert und Median nehmen nun aber gerade Werte um $y = 50$ an. Als solche tragen sie bei diesem Datensatz also keine Bedeutung darüber, welche Werte besonders häufig vorkommen, sondern lediglich darüber, wo sich der mittlere “physikalische Schwerpunkt” des Datensatzes befindet. Der Modalwert hingegen liegt, aufgrund des etwas häufigeren Auftretens von Werten um $y = 75$, bei dem größeren der beiden Datenhäufungspunkte. Insgesamt betrachtet beschreibt keines der drei Maße bei einem solchen Datensatz, den man auch *bimodal* nennt, die zentrale Tendenz der Datenwerte wirklich zufriedenstellend.

Abbildung 10.7 C visualisiert die Lage von Mittelwert, Median und Modalwert in einem Datensatz, in dem die Datenwerte in der betrachteten Breite etwa durchgängig gleichhäufig auftreten. Hier liegt der Modalwert relativ arbiträr bei etwa $y = 15$, da sich dort eine minimale Häufung von Datenwerten zeigt. Mittelwert und Median liegen wiederum im Sinne des “physikalischen Schwerpunkts” des Datensatzes bei etwa $y = 50$, auch wenn die Werte dieses Datensatzes dort keinerlei Tendenz haben, häufiger als andersorts aufzutreten. Wirklich zufriedenstellend ist auch hier die Beschreibung der zentralen Tendenz des Datensatzes durch die betrachteten Maße nicht. Zusammen betrachtet kann man anhand von Abbildung 10.7 B und Abbildung 10.7 C vielleicht festhalten, dass wenn ein Datensatz keine klare zentrale Tendenz hat, die Maße der zentralen Tendenz diese auch nicht abbilden können.

Abbildung 10.7 D schließlich zeigt die Lage von Mittelwert, Median, und Modalwert für einen Datensatz, der schiefsymmetrisch um Werte von etwa $y = 0.5$ verteilt ist. Die Daten dieses Datensatzes nehmen nur positive Werte an, viele Datenwerte liegen in der Tat um

$y = 0.5$, aber es kommen auch höhere Datenwerte vor. Wir werden an späterer Stelle sehen, dass dieses Erscheinungsbild für quadrierte normalverteilte Datenwerte typisch ist. Im psychologischen Kontext haben insbesondere Reaktionszeiten oft eine ähnliche Verteilung, da sie nicht negativ sein können, in den meisten Fällen etwa im Bereich von 0.3 bis 0.6 Sekunden liegen, aber in manchen Fällen durch Unaufmerksamkeit oder Stimulusvariabilität auch sehr viel länger sein können. In diesem Fall beschreibt der Median die tatsächliche zentrale Tendenz der Datenwerte etwas besser als der Mittelwert, da dieser durch die selten vorkommenden Ausreißerwerte etwas zu höheren Werten verschoben ist.

Insgesamt ist die Angabe eines oder mehrerer Maße der zentralen Tendenz also immer kritisch vor dem Hintergrund der Verteilung der Werte eines Datensatzes zu betrachten und qualitativ einzuordnen.

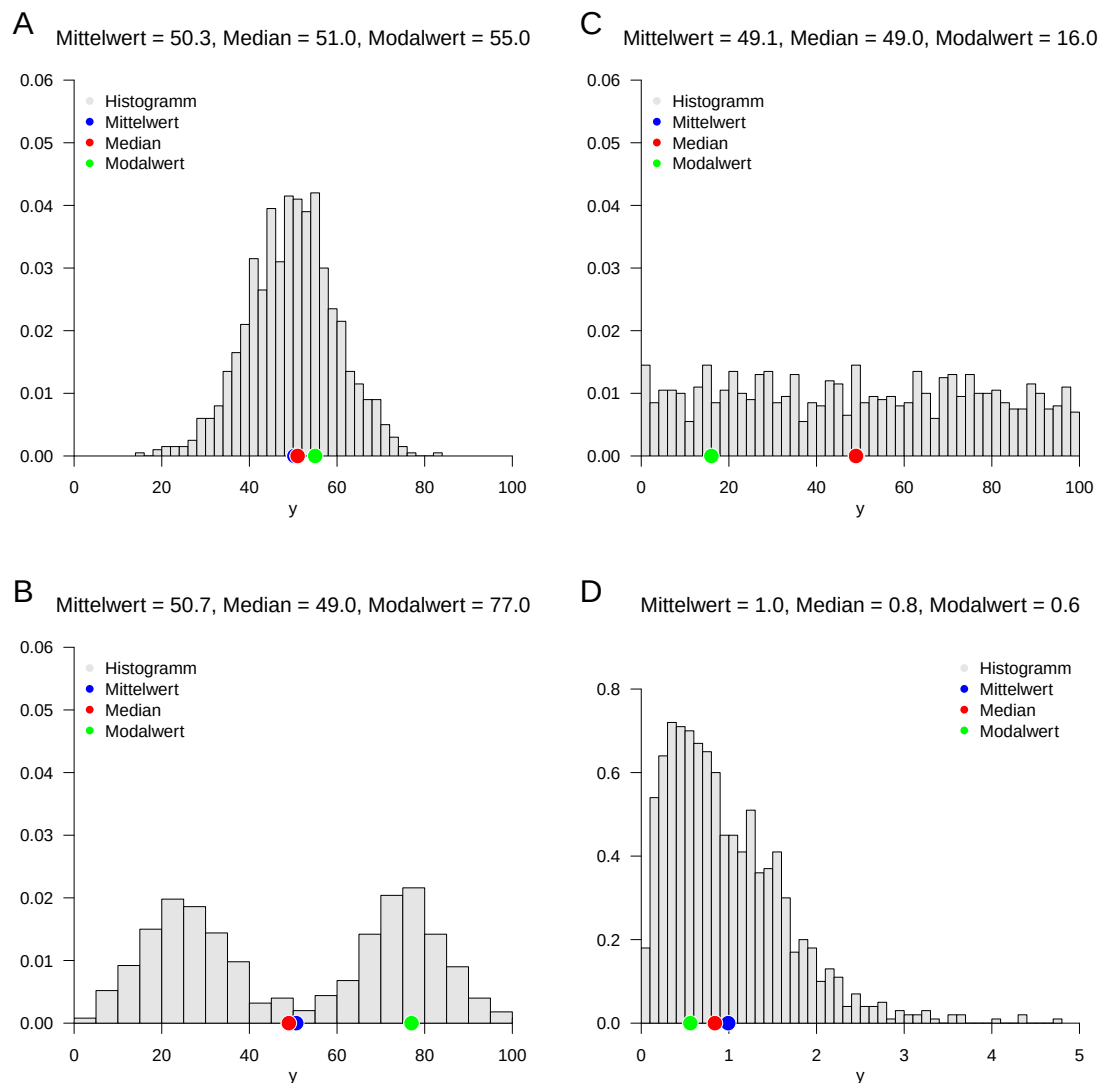


Abbildung 10.7. Visuelle Intuition zu Maßen zentraler Tendenz bei verschiedenen Datenverteilungsformen

10.6. Maße der Variabilität

Neben der Frage, welchen “durchschnittlichen Wert” ein Datensatz annimmt, ist es oft von Interesse zu quantifizieren, wie sehr die Werte im Datensatz streuen oder anders ausgedrückt, wie “variabel” sie sind. Als quantitative Maße für die Variabilität eines Datensatzes betrachten wir hier die *Spannbreite*, die *empirische Varianz* und die *empirische Standardabweichung*. Wie auch für den Mittelwert erschließt sich die volle Bedeutung insbesondere letztere Maße vor allem im Kontext inferenzstatistischer Verfahren.

10.6.1. Spannbreite

Die Spannbreite eines Datensatzes ist die Differenz zwischen seinem größten und seinem kleinsten Wert.

Definition 10.9 (Spannbreite). $y = (y_1, \dots, y_n)$ sei ein Datensatz. Dann ist die *Spannbreite* von $y_1, \dots, y_n$ definiert als

$$r := \max(y_1, \dots, y_n) - \min(y_1, \dots, y_n). \quad (10.33)$$

•

In **R** bestimmt man die Spannbreite entweder durch Evaluation des Maximums und des Minimums des Datensatzes mithilfe der `max()` und `min()` Funktionen oder direkt mithilfe der `range()` Funktion. Folgender **R** Code demonstriert die Anwendung der `max()` und `min()` Funktionen.

```
D      = read.csv("../data/110-deskriptive-statistik.csv")      # Laden des Datensatzes
y      = D$PRE                                                  # Pre-Interventions-BDI-II-Daten
y_max  = max(y)                                                  # Maximum der PRE Werte
y_min  = min(y)                                                  # Minimum der PRE Werte
r      = y_max - y_min                                           # Spannbreite
```

Spannbreite der PRE-BDI-II-Daten: 9

Folgender **R** Code demonstriert die Anwendung der `range()` Funktion.

```
MinMax = range(y)                                               # "automatische" Berechnung von min(x), max(x)
r      = MinMax[2] - MinMax[1]                                   # Spannbreite
```

Spannbreite der PRE-BDI-II-Daten: 9

10.6.2. Empirische Varianz

Ein typisches statistisches Maß für die Variabilität von Datensätzen ist die standardisierte Summe der quadrierten Differenzen zwischen den einzelnen Datensatzwerten und ihrem Mittelwert. Die quadrierten Differenzen zwischen den einzelnen Datensatzwerten und ihrem Mittelwert werden auch als *Abweichungsquadrate* bezeichnet. Je nach Form der Standardisierung, also der Bildung eines Durchschnitts der Abweichungsquadrate, unterscheidet man eine *unkorrigierte empirische Varianz* und eine *empirische Varianz*, wie folgende Definition festsetzt.

Definition 10.10 (Unkorrigierte empirische Varianz und empirische Varianz). $y = (y_1, \dots, y_n)$ sei ein Datensatz mit $n > 1$ und $\bar{y}$ sei sein Mittelwert. Dann ist die *unkorrigierte empirische Varianz* von y definiert als

$$\tilde{s}^2 := \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (10.34)$$

und die *empirische Varianz* von y ist definiert als

$$s^2 := \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \quad (10.35)$$

•

Wir haben in Theorem 10.1 gesehen, dass die Summe der Differenzen zwischen den Datenwerten und ihrem Mittelwert immer gleich Null ist. Die Quadrierung der Differenzen

$$y_i - \bar{y} \text{ für } i = 1, \dots, n \quad (10.36)$$

in Definition 10.10 führt nun gerade dazu, dass sich negative und positive Differenzen von Datenpunkten nicht ausgleichen und, je nach Höhe der positiven und negativen Differenzen, mit

$$\sum_{i=1}^n (y_i - \bar{y})^2 \quad (10.37)$$

ein höheres oder geringeres Maß für die Gesamtabweichung der Datenwerte von ihrem Mittelwert bestimmt werden kann. Dabei ist die Summe der Abweichungsquadrate gerade dann Null, wenn alle Datenwerte identisch und damit auch identisch mit ihrem Mittelwert sind. Alternativ wäre zum Beispiel auch die Bestimmung der Absolutbeträge $|y_i - \bar{y}|$ denkbar, allerdings erfüllt das so entstehende Maß der Variabilität andere Eigenschaften als das Maß der quadrierten Abweichungen in Definition 10.10, weshalb dieses, auch in Hinblick auf seine Nähe zu probabilistischen Normalverteilungsmodellen in der Regel bevorzugt wird. Die unkorrigierte empirische Varianz $\tilde{s}^2$ und die (korrigierte) empirische Varianz s^2 unterscheiden sich dann hinsichtlich der Form der Durchschnittsbildung der quadrierten Abweichungsquadrate. Für $\tilde{s}^2$ wird die Summe der Abweichungsquadrate durch n , für s^2 durch $n-1$ geteilt. $\tilde{s}^2$ ist also immer etwas kleiner als s^2 . Allerdings wird dieser Unterschied für großes n eher gering ausfallen, man denke beispielsweise an den numerischen Unterschied zwischen $\frac{1}{3}$ und $\frac{1}{2}$ bei kleinem n sowie zwischen $\frac{1}{1001}$ und $\frac{1}{1000}$ bei großem n . Man beachte, dass die empirische Varianz für $n = 1$ nicht definiert ist, die unkorrigierte empirische Varianz jedoch auch in diesem Spezialfall prinzipiell bestimmt werden kann. Der Begriff *unkorrigiert* (und implizit *korrigiert*) kann sich an dieser Stelle noch nicht erschließen, sondern ergibt nur vor dem Hintergrund eines Frequentistischen Inferenzmodells Sinn und wird in Kapitel 37 erläutert werden.

Wir fassen die obige Diskussion in folgendem Theorem zusammen

Theorem 10.4 (Verhältnis von unkorrigierter empirischer Varianz und empirischer Varianz). $y = (y_1, \dots, y_n)$ sei ein Datensatz und $\tilde{s}^2$ und s^2 seien seine unkorrigierte empirische Varianz und empirische Varianz, respektive. Dann gelten

(1) (Umrechnung)

$$\tilde{s}^2 = \frac{n-1}{n} s^2 \text{ und } s^2 = \frac{n}{n-1} \tilde{s}^2. \quad (10.38)$$

(2) (Ungleichung)

$$0 \leq \tilde{s}^2 \leq s^2. \quad (10.39)$$

◦

Beweis. (1) Nach Definition 10.10 gilt zum einen

$$\begin{aligned} \tilde{s}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ \Leftrightarrow \frac{n-1}{n-1} \tilde{s}^2 &= \frac{n-1}{n-1} \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ \Leftrightarrow \tilde{s}^2 &= \frac{n-1}{n} \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ \Leftrightarrow \tilde{s}^2 &= \frac{n-1}{n} s^2 \end{aligned} \quad (10.40)$$

und zum anderen

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ \Leftrightarrow \frac{n}{n} s^2 &= \frac{n}{n} \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ \Leftrightarrow s^2 &= \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ \Leftrightarrow s^2 &= \frac{n}{n-1} \tilde{s}^2. \end{aligned} \quad (10.41)$$

(2) Mit der Nichtnegativität der Quadratfunktion sowie $\frac{1}{n} > 0$ und $\frac{1}{n-1} > 0$ für $n > 1$ ergeben sich $\tilde{s}^2 \geq 0$ und $\tilde{s}^2 > 0$ direkt. Schließlich gilt für $\sum_{i=1}^n (y_i - \bar{y})^2 \neq 0$

$$\frac{1}{n} < \frac{1}{n-1} \Leftrightarrow \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 < \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \Leftrightarrow \tilde{s}^2 < s^2. \quad (10.42)$$

und für $\sum_{i=1}^n (y_i - \bar{y})^2 = 0$ offenbar $\tilde{s}^2 = s^2$. Insgesamt gilt also

$$0 \leq \tilde{s}^2 \leq s^2. \quad (10.43)$$

□

Folgender **R** Code demonstriert die Bestimmung der unkorrigierten empirischen Varianz und der empirischen Varianz anhand der in Definition 10.10 festgelegten Formeln.

```
D = read.csv("../data/110-deskriptive-statistik.csv") # Laden des Datensatzes
y = D$PRE # Pre-Interventions-BDI-II-Daten
n = length(y) # Anzahl der Werte
S2_tilde = (1/n)*sum((y - mean(y))^2) # unkorrigierte empirische Varianz
S2 = (1/(n-1))*sum((y - mean(y))^2) # empirische Varianz
```

```
Unkorrigierte empirische Varianz der PRE-BDI-II-Daten : 2.998
Empirische Varianz der PRE-BDI-II-Daten : 3.028
```

Die **R** Funktion `var()` bestimmt per default die empirische Varianz. Mithilfe von Aussage (1) in Theorem 10.4 kann basierend auf dieser Funktion auch die unkorrigierte empirische Varianz bestimmt werden.

```
s2 = var(y) # empirische Varianz
s2_tilde = ((n-1)/n)*var(y) # unkorrigierte empirische Varianz
```

Unkorrigierte empirische Varianz der PRE-BDI-II-Daten: 2.998
 Empirische Varianz der PRE-BDI-II-Daten : 3.028

Folgendes Theorem besagt, wie sich die empirische Varianz bei linear-affiner Transformation eines Datensatzes verhält.

Theorem 10.5 (Varianz bei linear-affinen Transformationen). $y = (y_1, \dots, y_n)$ sei ein Datensatz mit empirischer Varianz S_y^2 und $z = (ay_1 + b, \dots, ay_n + b)$ sei der mit $a, b \in \mathbb{R}$ linear-affin transformierte Datensatz mit empirischer Varianz S_z^2 . Dann gilt

$$s_z^2 = a^2 s_y^2. \quad (10.44)$$

◦

Beweis.

$$\begin{aligned} s_z^2 &= \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (ay_i + b - (a\bar{y} + b))^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (ay_i + b - a\bar{y} - b)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (a(y_i - \bar{y}))^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n a^2 (y_i - \bar{y})^2 \\ &= a^2 \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= a^2 S_y^2. \end{aligned} \quad (10.45)$$

□

Beispielsweise ändert also das Umrechnen eines Datensatzes von Meter in Zentimeter die Varianz des Ausgangsdatsatzes um den Faktor 100^2 . Wir reproduzieren Theorem 10.5 beispielhaft mithilfe folgenden **R** Codes.

```
y      = D$PRE                                # Pre-Interventions-BDI-II-Daten
s2y     = var(y)                               # Empirische Varianz von y_1,...,y_n
a       = 2                                   # Multiplikationskonstante
b       = 5                                   # Additionskonstante
z       = a*y + b                             # z_i = az_i + b
s2z     = var(z)                               # Empirische Varianz von z_1,...,z_n
cat("Empirische Varianz von z_1,...,z_n durch direkte Berechnung : ", round(s2z,3)) # Ausgabe
```

Empirische Varianz von z_1,...,z_n durch direkte Berechnung : 12.113

```
s2z     = a^2*s2y                             # Empirische Varianz von z_1,...,z_n
cat("Empirische Varianz von z_1,...,z_n nach Theorem : ", round(s2z,3)) # Ausgabe
```

Empirische Varianz von z_1,...,z_n nach Theorem : 12.113

Folgendes Theorem zeigt eine Möglichkeit auf, die unkorrigierte empirische Varianz eines Datensatzes allein durch die Bestimmung von Mittelwerten zu bestimmen. Das Theorem findet seine analytische Entsprechung im Verschiebungssatz der Varianz, Theorem 24.5.

Theorem 10.6 (Verschiebungssatz zur unkorrigierten empirischen Varianz). $y = (y_1, \dots, y_n)$ sei ein Datensatz, $y^2 := (y_1^2, \dots, y_n^2)$ sei sein elementweises Quadrat und $\bar{y}$ und $\bar{y}^2$ seien die respektiven Mittelwerte. Dann gilt

$$\tilde{s}^2 = \overline{y^2} - \bar{y}^2 \quad (10.46)$$

◦

Beweis.

$$\begin{aligned} \tilde{s}^2 &:= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (y_i^2 - 2y_i\bar{y} + \bar{y}^2) \\ &= \frac{1}{n} \sum_{i=1}^n y_i^2 - 2\bar{y} \frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{n} \sum_{i=1}^n \bar{y}^2 \\ &= \overline{y^2} - 2\bar{y}\bar{y} + \frac{1}{n} n\bar{y}^2 \\ &= \overline{y^2} - 2\bar{y}^2 + \bar{y}^2 \\ &= \overline{y^2} - \bar{y}^2. \end{aligned} \quad (10.47)$$

□

Wir reproduzieren **thm-verschiebungssatz-der-empirischen-varianz** beispielhaft mithilfe folgenden **R** Codes.

```
y           = D$PRE                               # Pre-Interventions-BDI-II-Daten
s2_tilde    = mean(y^2) - (mean(y))^2              # \bar{y^2} - \bar{y}^2
cat("Korrigierte empirische Varianz nach Transformation : ", round(s2z,3)) # Ausgabe
```

```
Korrigierte empirische Varianz nach Transformation : 12.113
```

Man beachte, dass der Verschiebungssatz der unkorrigierten empirischen Varianz für die empirische Varianz nicht zutrifft, da in diesem Fall der Multiplikationsfactor $\frac{1}{n-1}$ nicht auf die entsprechenden Mittelwerte führt. Insbesondere haben wir oben schon gesehen, dass für den betreffenden Beispieldatensatz gilt, dass $s^2 = 3.028$.

10.6.3. Empirische Standardabweichung

Die empirische Varianz ergibt sich wie oben gesehen als durchschnittliches Abweichungsquadrat eines Datensatzwerts von seinem Mittelwert. Ist die Einheit des Datensatzes dabei zum Beispiel in einer Reaktionsszeitaufgabe die Millisekunde, so ergibt sich durch Differenzbildung und anschließende Quadrierung die Einheit Millisekunde², eine nicht sehr intuitive Einheit. Um ein Streuungsmaß zu erhalten, dessen Einheit der Einheit des Ausgangsdatsatzes entspricht, wird auf die empirische Varianz häufig noch die Quadratwurzelfunktion angewendet. Dies führt auf den Begriff der *empirischen Standardabweichung*.

Definition 10.11 (Empirische Standardabweichung). $y = (y_1, \dots, y_n)$ sei ein Datensatz. Die *empirische Standardabweichung* von y ist definiert als

$$s := \sqrt{s^2} \quad (10.48)$$

und die *unkorrigierte empirische Stichprobenstandardabweichung* von y ist definiert als

$$\tilde{s} := \sqrt{\tilde{s}^2}. \quad (10.49)$$

•

In Analogie zu den Eigenschaften der empirischen Varianz und ihrer unkorrigierten Version kann auch die unkorrigierte empirischen Standardabweichung leicht in die empirischen Standardabweichung umgerechnet werden und umgekehrt, es gelten hier offenbar

$$\tilde{s} = \sqrt{\frac{n-1}{n}} s \text{ und } s = \sqrt{\frac{n}{n-1}} \tilde{s}. \quad (10.50)$$

Folgender **R** Code demonstriert die Bestimmung der unkorrigierten empirischen Varianz und der empirischen Varianz anhand der in Definition 10.11 festgelegten Formeln.

```
y = D$PRE # Pre-Interventions-BDI-II-Daten
n = length(y) # Anzahl der Werte
s = sqrt((1/(n-1))*sum((y - mean(y))^2)) # Standardabweichung
```

Empirische Standardabweichung : 1.74

Die **R** Funktion `sd()` bestimmt per default die empirische Varianz.

```
s = sd(y)
```

Empirische Standardabweichung : 1.74

Transformiert man einen Datensatz linear-affin, so geht die Multiplikationskonstante der Transformation lediglich im Sinne ihres Betrages in die Transformation der empirischen Standardabweichung ein. Dies ist die Aussage des folgenden Theorems.

Theorem 10.7 (Empirische Standardabweichung bei linear-affinen Transformationen). $y = (y_1, \dots, y_n)$ sei ein Datensatz mit empirischer Standardabweichung s_y und $z = (az_1 + b, \dots, az_n + b)$ sei der mit $a, b \in \mathbb{R}$ linear-affin transformierte Datensatz mit empirischer Standardabweichung s_z . Dann gilt

$$s_z = |a| s_y. \quad (10.51)$$

◦

Beweis.

$$\begin{aligned}
 S_y &:= \left(\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right)^{1/2} \\
 &= \left(\frac{1}{n-1} \sum_{i=1}^n (ay_i + b - (a\bar{x} + b))^2 \right)^{1/2} \\
 &= \left(\frac{1}{n-1} \sum_{i=1}^n (a(y_i - \bar{x}))^2 \right)^{1/2} \\
 &= \left(\frac{1}{n-1} \sum_{i=1}^n a^2 (y_i - \bar{x})^2 \right)^{1/2} \\
 &= (a^2)^{1/2} \left(\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{x})^2 \right)^{1/2}.
 \end{aligned} \tag{10.52}$$

Also gilt $S_z = aS_y$, wenn $a \geq 0$ und $S_z = -aS_y$, wenn $a < 0$. Dies aber entspricht $S_z = |a|S_y$. $\square$

Wir reproduzieren Theorem 10.7 beispielhaft mithilfe folgenden **R** Codes einmal für den Fall $a \geq 0$ und einmal für den Fall $a < 0$

```
# a >= 0
D = read.csv("../data/110-deskriptive-statistik.csv")
y = D$PRE
a = 2
b = 5
z = a*y + b
sz = sd(z)

# Laden des Datensatzes
# Pre-Interventions-BDI-II-Daten
# Multiplikationskonstante
# Additionskonstante
# y_i = ax_i + b
# Korrigierte empirische Standardabweichung von y
```

Empirische Standardabweichung von z_1,...z_n durch direkte Berechnung : 3.48

```
sz = a*sd(y)

# Korrigierte empirische Standardabweichung von y
```

Empirische Standardabweichung von z_1,...z_n nach Theorem : 3.48

```
# a < 0
D = read.csv("../data/110-deskriptive-statistik.csv")
y = D$PRE
a = -3
b = 10
z = a*y + b
sz = sd(z)

# Laden des Datensatzes
# Pre-Interventions-BDI-II-Daten
# Multiplikationskonstante
# Additionskonstante
# y_i = ax_i + b
# Korrigierte empirische Standardabweichung von y
```

Empirische Standardabweichung nach Transformation : 5.221

```
sz = -a*sd(y)

# Korrigierte empirische Standardabweichung von y
```

Empirische Standardabweichung nach Transformation : 5.221

Teil II.

Imperative Programmierung

Zur Bedeutung der imperativen Programmierung in der Psychologie

Heutzutage liegen wissenschaftliche Daten in der Regel in digitaler Form vor. Um diese digitalen Daten auszuwerten, werden sie mithilfe eines Computers analysiert. Dazu schreibt man spezielle Computerprogramme, die genau auf die jeweilige Fragestellung zugeschnitten sind. Solche Programme bezeichnet man als Datenanalyseskripte.

Die computergestützte Datenanalyse folgt dabei einer typischen Struktur. Zunächst wird ein digitaler Datensatz eingelesen und bereinigt, um fehlerhafte oder unvollständige Daten zu korrigieren und für die Analyse vorzubereiten. Anschließend werden meist deskriptive Statistiken berechnet und visualisiert, um einen ersten Überblick über die Daten zu gewinnen. Darauf folgt häufig ein Schritt der probabilistischen Modellierung und Inferenz, bei der Wahrscheinlichkeitstheorie-basierte Modelle verwendet werden, um Hypothesen zu testen oder Vorhersagen zu treffen. Abschließend werden die Ergebnisse, wiederum mithilfe des Computers, dokumentiert und anschaulich präsentiert.

Zentrales Element der computergestützten Datenanalyse ist das *Datenanalyseskript*, eine Textdatei, die alle Schritte vom Rohdatum bis hin zur finalen Datenvisualisierung dokumentiert und so Datenanalyseprozesse transparent und nachvollziehbar macht. So ermöglicht ein Datenanalyseskript Dritten die Reproduktion wissenschaftlicher Ergebnisse, was einen Grundpfeiler guter wissenschaftlicher Praxis darstellt. Datenanalyseskripte sind damit unverzichtbare Werkzeuge der täglichen wissenschaftlichen Arbeit und essenzielle Bestandteile wissenschaftlicher Publikationen. In den folgenden Abschnitten wollen wir ausgehend von der Diskussion einiger informatischer Grundbegriffe und strukturbildender Elemente der imperativen Programmierung die Entwicklung von Datenanalyseskripten in der Programmierumgebung [R](#) kennenlernen.

Zur Analyse psychologischer Daten kommen verschiedene digitale Werkzeuge zum Einsatz. Besonders verbreitet sind frei verfügbare Programmiersprachen und Plattformen wie [R](#), das häufig in der Datenwissenschaft, Statistik und Psychologie genutzt wird, und [Python](#) mit seinem breiten Einsatzspektrum in der Datenwissenschaft, dem maschinellen Lernen und der künstlichen Intelligenz. Eher noch ein Nischendasein führt die eigentlich datenanalytisch optimierte freie Programmiersprache [Julia](#), die momentan insbesondere in industriellen Anwendungen noch keine weite Verbreitung gefunden hat. Freie Programmiersprachen haben den Vorteil, dass sie häufig von einer breiten Community von Anwendern weiter entwickelt werden und auf diese Weise eine versatile Infrastruktur entwickeln können.

Daneben kommen aber auch im Bereich der Psychologie weiterhin kommerzielle Programmierumgebungen, deren Quellcode nicht einsehbar ist, zum Einsatz. Im Bereich der Neurobiologie sowie in ingenieurstechnischen Anwendungen ist zum Beispiel auch heute noch die kommerzielle Programmierumgebung [Matlab](#) beliebt. Weiterhin werden auch weiterhin einige aus heutiger Sicht eher altmodische Programme verwendet, darunter [SPSS](#), das zur Jahrtausendwende vor allem in den Sozialwissenschaften und in der Psychologie sehr verbreitet war, [JMP](#) mit Einsatzschwerpunkten in Biologie und Psychologie und [STATA](#), das unter anderem in der Medizin und den Wirtschaftswissenschaften populär war.

Die Popularität digitaler Werkzeuge unterliegt einem stetigen Wandel. Eine Möglichkeit, die aktuelle Popularität verschiedener Programmiersprachen und -umgebungen zu messen, ist die Quantifizierung von GoogleSuchanfragen zu Programmiersprachentutorials. Einen Überblick über aktuelle Trends bietet der [PYPL Index](#), der auf der Analyse solcher Suchanfragen basiert.

Zusammenfassend lässt sich festhalten, dass die Digitalisierung insbesondere auch die Wissenschaft nachhaltig prägt. Ein effektives Forschungsdatenmanagement stellt dabei eine akute Herausforderung dar, der sich Forschende stellen müssen. Programmierung entwickelt sich zunehmend zu einem zentralen Handwerkszeug wissenschaftlicher Arbeit, und grundlegende Informatikkenntnisse sind in der modernen Arbeitswelt unverzichtbar geworden. Dies gilt nicht zuletzt auch für Psychotherapeut:innen, etwa im Kontext von Online-Interventionen, digital gestützten Behandlungsangeboten oder der Psychotherapieforschung.

11. Informatische Grundbegriffe

Nach allgemeinem Sprachgebrauch (vgl. [Wikipedia](#)) handelt es sich bei der Informatik um die Wissenschaft der systematischen Darstellung, Speicherung, Verarbeitung und Übertragung von Informationen, wobei besonders die automatische Informationsverarbeitung mit Computern betrachtet wird. Die Informatik ist dabei zugleich Grundlagen- und Formalwissenschaft als auch Ingenieurdisziplin.

Zentrale Komponenten der Informatik sind zum einen die *Computer*, zum anderen *Algorithmen und Programme*. *Computer* sind Maschinen, die Daten speichern und einfache Datenoperationen ausführen können. Diese einfachen Operationen führen sie mit extrem hoher Geschwindigkeit aus. Die Universalität von Computern beruht darauf, dass sie sowohl Daten als auch Programme speichern können.

Programme sind in einer Programmiersprache formulierte *Algorithmen*. *Algorithmen* wiederum sind Folgen von *Anweisungen*, die bestimmte *Operationen* vorschreiben. Bei Algorithmen unterscheidet man verschiedene Ebenen: die *Beschreibungsebene*, wie man sie beispielsweise in Form eines Kochrezepts, einer Bauanleitung oder eines Datenanalyse-Skripts findet; die *Anweisungen* eines Algorithmus, wie “Mehl und Wasser vermengen”, die bildhafte Darstellung des Zusammenschraubens zweier Teile oder die Anweisung $x = c(1,2,3)$ in $\mathbf{R}$, und schließlich die *Durchführungsebene*, also der eigentliche Kochvorgang, der Zusammenbau eines Möbelstücks oder das Ausführen des Datenanalyseskripts.

Die Informatik gliedert sich in mehrere Teilgebiete, die auch für die Psychologie von Bedeutung sein können. Die *Theoretische Informatik* beschäftigt sich mit den Grundlagen der Berechnung, wie der Automatentheorie, der Berechenbarkeitstheorie und der Komplexitätstheorie. Die *Angewandte Informatik* befasst sich mit der Entwicklung und Nutzung von Anwendungssoftware, mit Fragen der Mensch-Computer-Interaktion sowie mit den gesellschaftlichen Auswirkungen der Informatik. Die *Technische Informatik* untersucht die Hardwareseite der Informatik, darunter die Mikroprozessortechnik, die Rechnerarchitektur und die Netzwerktechnik. Schließlich behandelt die *Praktische Informatik* die konkrete Umsetzung in Form von Programmierung, der Entwicklung und Analyse von Algorithmen sowie den Aufbau und die Nutzung von Datenbanken. Anwendung in der psychologischen Grundlagenforschung findet vor allem die praktische Informatik. Allerdings sind auch Fragen der Komplexitätstheorie und damit der theoretischen Informatik in der Entwicklung quantitativer psychologischer Theorien von Interesse (vgl. Van Rooij & Wareham (2012)) und die Mensch-Computer-Interaktion bildet eines der Themengebiete der modernen Arbeitspsychologie (vgl. Dahm (2005)).

Darüberhinaus gibt es zahlreiche Spezialgebiete der Informatik, die eng mit der Psychologie verbunden sind. Dazu gehört insbesondere natürlich der Bereich *Maschinelles Lernen* und *Künstliche Intelligenz*, der seine Ursprünge in der quantitativen Kognitionspsychologie hat und es heutzutage ermöglicht, große Datenmengen automatisiert zu analysieren und Vorhersagen zu treffen. Ein weiteres wichtiges Feld ist die *Computervisualistik*, die sich mit Bilderkennung und Bildsynthese sowie mit der Gestaltung von Virtueller Realität

(VR) und Augmented Reality (AR) beschäftigt. Diese Technologien nutzen einerseits Erkenntnisse der Wahrnehmungspsychologie und eröffnen andererseits neue Möglichkeiten für Experimente, Therapien und Trainings in der Psychologie. Die *Computerlinguistik* trägt mit Verfahren der Spracherkennung und Sprachsynthese dazu bei, menschliche Kommunikation zu analysieren, zu modellieren und in Anwendungen wie Chatbots oder Assistenzsystemen nutzbar zu machen. Schließlich spielt für die Psychologie auch die *Bioinformatik* eine große Rolle, indem sie etwa Methoden zur Auswertung bildgebender Verfahren in der Medizin, die auch in der neuropsychologischen Forschung und Diagnostik genutzt werden, weiter entwickelt.

11.1. Rechnerarchitektur

Auch wenn wir die technischen Aspekte der Informatik hier nicht vertiefen wollen, ist es für die imperative Programmierung hilfreich, zumindest ein rudimentäres Wissen um den typischen physischen Aufbau, also die Hardware, eines Computers zu besitzen. Ein Computer besteht dabei aus mehreren wesentlichen Hardwarekomponenten, die zusammenarbeiten, um alle Rechen- und Speicheraufgaben zu bewältigen. Im Zentrum steht die Zentraleinheit, auch Hauptplatine, Motherboard oder Mainboard genannt. Sie ist die zentrale Leiterplatte, auf der alle wichtigen Bauteile wie Prozessor, Speicher und Anschlüsse miteinander verbunden sind.

Das Herzstück eines Computers ist die CPU (Central Processing Unit, auch Mikroprozessor), die das Rechenwerk, das Steuerwerk und das Leitwerk des Systems vereint. Sie übernimmt alle grundlegenden Berechnungen, steuert den Datenfluss und koordiniert die Abläufe. Zusätzlich verfügt sie über einen kleinen, aber sehr schnellen *Cache*, einen flüchtigen Speicher, der häufig benötigte Daten zwischenspeichert, um die Verarbeitung zu beschleunigen. Typische Namen von aktuellen CPUs sind Intels' Core iX Prozessoren, AMD's Ryzen Prozessoren oder Apple Silicon MX Prozessoren.

Ergänzt wird die CPU durch den RAM (Random Access Memory), den temporären, flüchtigen Arbeitsspeicher des Systems. Hier werden während der Laufzeit des Computers die aktuell benötigten Programme und Daten abgelegt, um schnellen Zugriff zu ermöglichen. Der RAM ist in seiner Kapazität begrenzt, in vielen modernen Systemen zum Beispiel auf 32 - 64 GB.

Für die dauerhafte Speicherung von Daten und Programmen dient der Massenspeicher, der stationäre Speicher des Computers. Heute werden dafür in der Regel SSDs (Solid State Drives) verwendet, die wesentlich schneller und robuster als herkömmliche magnetbasierte Festplatten sind. Auch Cloudspeicher wird zunehmend als Ergänzung oder Ersatz für lokale Massenspeicher genutzt.

Eine weitere wichtige Komponente ist die GPU (Graphics Processing Unit), ein leistungstarker, speziell auf Visualisierungsaufgaben optimierter Prozessor. Die GPU entlastet die CPU bei grafikintensiven Anwendungen wie 3D-Rendering oder Videobearbeitung und spielt zudem eine wachsende Rolle bei rechenintensiven Aufgaben wie dem Training neuronaler Netze, bei denen sie die CPU effektiv unterstützen kann.

Die grundlegende funktionale Architektur moderner Computern geht zurück auf Von Neumann (1945) und wird entsprechend als *Von Neumann-Architektur* bezeichnet. Das entscheidende Charakteristikum der von Neumann-Architektur besteht darin, dass sowohl die

Befehle (Programmanweisungen) als auch die Daten im selben Speicher abgelegt werden, und zwar in Form von Binärzahlen. Die Zentrale Verarbeitungseinheit (CPU) liest sowohl die Befehle als auch die Daten aus diesem gemeinsamen Speicher. Dadurch können Programme während der Ausführung dynamisch verändert oder sogar neu erzeugt werden, da sie im Speicher lediglich als Daten vorliegen.

Speziell beschreibt Von Neumann (1945) folgenden Rechneraufbau: Der Rechner setzt sich aus den grundlegenden Funktionseinheiten Steuerwerk, Rechenwerk, Speicher, Eingabewerk und Ausgabewerk zusammen. Programme und Daten werden gemeinsam in den Speicher eingegeben, sodass Daten, Programme sowie Zwischen- und Endergebnisse im gleichen Speicherbereich abgelegt sind. Der Speicher selbst ist in gleich große, nummerierte (adressierte) Zellen unterteilt, deren Inhalt über die jeweilige Adresse gezielt abgerufen oder verändert werden kann. Die Befehle eines Programms liegen typischerweise in aufeinanderfolgenden Speicherzellen, sodass das Steuerwerk den jeweils nächsten Befehl durch einfaches Erhöhen der Befehlsadresse um eins aufrufen kann. Dieses Prinzip der sequentiellen Abarbeitung der Befehle stellt dabei das Grundprinzip der imperativen Programmierung dar und ist somit auf engste mit der von Neumann-Architektur verknüpft. Durch spezielle Sprungbefehle ist es weiterhin möglich, von dieser festen Reihenfolge abzuweichen und an eine andere Stelle im Programm zu springen. Zu den grundlegenden Befehlen gehören dabei arithmetische Operationen wie Addition und Multiplikation, logische Vergleiche wie das logische UND oder ODER sowie Transportbefehle, mit denen Daten beispielsweise vom Eingabewerk in den Speicher oder vom Speicher ins Rechenwerk übertragen werden. Sämtliche Daten, also sowohl Befehle als auch Adressen, werden dabei binär codiert. Die binäre Encodierung und Dekodierung der Informationen übernimmt dabei eine geeignete Schaltungstechnik im Rechner.

11.2. Algorithmen und Programme

Ein *Realweltproblem* bezeichnet das Problem, das mithilfe eines Computers gelöst werden soll. Ein typisches Beispiel hierfür ist die Auswertung von Fragebogendaten einer psychologischen Studie. Zur Bearbeitung eines solchen Problems erfolgt zunächst eine *Problemspezifikation*, also die genaue sprachliche Fassung des Realweltproblems, wie man sie etwa im Methodenteil einer wissenschaftlichen Publikation findet. Auf dieser Grundlage wird ein *Algorithmus* entworfen, also eine Folge von Anweisungen zur Lösung des Problems, beispielsweise das Einlesen der Daten, die Berechnung von deskriptiven Statistiken und die Durchführung eines statistischen Tests. Schließlich wird der Algorithmus in Form eines *Programms* implementiert, das heißt als in einer Programmiersprache verfasste Textdatei, die von einem Computer ausgeführt werden kann.

Definition 11.1 (Algorithmus). Ein Algorithmus ist eine Folge von Anweisungen, um aus gewissen Eingabedaten bestimmte Ausgabedaten herzuleiten, wobei folgende Bedingungen erfüllt sein müssen

- **Finitheit.** Die Anweisungsfolge ist in einem endlichen Text vollständig beschrieben.
- **Effektivität.** Jede Anweisung muss tatsächlich ausführbar sein.
- **Terminierung.** Der Algorithmus endet nach endlich vielen Anweisungen.
- **Determiniertheit.** Der Ablauf des Algorithmus ist zu jeder Zeit vorgeschrieben.

Wenn E die Menge der zulässigen Eingabedaten und A die Menge der zulässigen Ausgabedaten bezeichnet, dann ist ein Algorithmus also eine Funktion

$$f : E \rightarrow A, e \mapsto f(e) \quad (11.1)$$

Umgekehrt heißen Funktionen, die durch einen Algorithmus beschrieben werden können, *berechenbare Funktionen*.

•

11.3. Programmiersprachen

Programmiersprachen legen die Regeln fest, denen ein Programm gehorchen muss. Sie definieren eine *Syntax*, also das Vokabular und den Aufbau eines Programms, sowie die *Semantik*, also die Bedeutung der erlaubten Anweisungen. Im Folgenden wollen wir einige Begriffsklärungen zur Kategorisierung verschiedener Formen von Programmiersprachen angeben. Dabei unterscheidet man *Maschinensprache* und sogenannte *Höhere Programmiersprachen*, verschiedene *Generationen von Programmiersprachen*, die hier im Vordergrund stehende *Imperative Programmierung* von der *Deklarativen Programmierung*, sowie *kompilierte* von *interpretierten* Programmiersprachen.

Maschinensprache und höhere Programmiersprachen

Maschinensprache besteht aus elementaren Operationsbefehlen wie Speichern, Vergleichen oder Addieren, die als Binärzahlen kodiert werden. Tabelle 11.1 zeigt einige Beispiele solcher Befehle:

Tabelle 11.1. Beispiele für Maschinensprache

Befehl	Binärcode
Addiere Inhalt R1 zu Inhalt R2	1001 0010
Erhöhe Inhalt R um 1	1001 0110
Übertrage Inhalt R1 nach R3	0010 0011

Programme, die in Maschinensprache geschrieben sind, nennt man Maschinenprogramme. De facto führt ein Computer ausschließlich Maschinenprogramme aus, da diese direkt von der Hardware verstanden werden. Für Menschen ist die Programmierung in Maschinensprache jedoch sehr mühselig, weshalb in der Praxis überwiegend höhere Programmiersprachen verwendet werden.

Höhere Programmiersprachen verwenden an die menschliche Sprache angelehnte Wörter und Sätze, um die Programmierung verständlicher und effizienter zu machen. Programme in höheren Programmiersprachen werden anschließend von einem *Compiler* oder einem *Interpreter* in Maschinensprache übersetzt, damit sie vom Computer ausgeführt werden können. Bekannte Beispiele für höhere Programmiersprachen sind FORTRAN, COBOL, C++, Python oder auch **R**.

Generationen von Programmiersprachen

Die *Programmiersprachen der 1. Generation (1GL)* sind die Maschinensprachen. Programme werden dabei direkt in Form von Binärzahlen geschrieben, etwa 10110000 01100001,

was in hexadezimaler Darstellung B0 61 entspricht. Solche Maschinensprachen wurden vor allem auf den ersten Röhren- und Relaiscomputern wie der ENIAC oder frühen IBM-Großrechnern der 1940er- und frühen 1950er-Jahre eingesetzt.

Mit der *2. Generation (2GL)* entstanden ab den 1950er Jahren die Assemblersprachen als erste Form der symbolischen Programmierung. Hierbei werden menschenlesbare Kürzel verwendet, die jedoch noch prozessorspezifisch sind. Ein Beispiel für eine solche Intelprozessor-spezifische Anweisung ist `MOV AL, 61H`. Diese Generation von Programmiersprachen wurde vor allem auf Großrechnern wie der IBM 704 oder frühen Mainframes der 1950er und 1960er Jahre genutzt, die bereits Transistoren statt Röhren verwendeten.

Ab den 1970er Jahren setzten sich die *Programmiersprachen der 3. Generation (3GL)* durch. Dazu zählen höhere Programmiersprachen wie FORTRAN, C, C++ und Java. Sie sind wesentlich programmierfreundlicher und vor allem prozessorunabhängig, was die Portabilität von Programmen deutlich verbessert. Solche Sprachen wurden zunächst auf Minicomputern wie DEC PDP-11 und später auf Workstations, PCs und Servern weit verbreitet verwendet.

Schließlich entwickelten sich ab den 1980er Jahren die *Programmiersprachen der 4. Generation (4GL)*, zu denen Sprachen wie Python, Matlab und **R** gehören. Diese zeichnen sich durch die Minimierung von Code-Overhead, hohe Flexibilität, Unterstützung multiparadigmatischer Ansätze und einen hohen Grad an Automatisierung aus. Sie finden vor allem auf modernen PCs, High-Performance-Computing (HPC) Clustern und Cloud-basierten Systemen Anwendung, die seit den 1990er Jahren zunehmend verfügbar wurden.

Imperative und deklarative Programmierung

Bei der *Imperativen Programmierung* (von *imperare*, lat. “befehlen”) wird der Lösungsweg als Folge von Anweisungen beziehungsweise Befehlen vorgegeben. Diese Befehle verarbeiten Daten, die über Variablen adressiert werden. Innerhalb der imperativen Programmierung unterscheidet man mit der *prozeduralen* und der *objektorientierten* imperativen Programmierung zwei verschiedene Ansätze.

In der *prozeduralen imperativen Programmierung* werden Daten und die sie manipulierenden Befehle getrennt behandelt. Das zentrale Strukturkonzept bilden hier Prozeduren oder Funktionen, die wiederverwendbare Einheiten von Anweisungen kapseln. Demgegenüber werden in der *objektorientierten imperativen Programmierung* Daten und die auf sie wirkenden Befehle als Objekte zusammengefasst. Diese Objekte sind zugleich Datenträger und Funktionsanbieter und bilden das zentrale Strukturkonzept dieses Ansatzes.

In der Praxis kommen häufig Mischformen aus beiden Paradigmen zum Einsatz, da sich die Ansätze je nach Problemstellung gut kombinieren lassen. Ein Beispiel für eine solche Mischform ist die neuere Programmiersprache **Julia**, die keinen klassischen objektorientierten Ansatz, sondern stattdessen einen *multiple dispatch* Ansatz bietet: Funktionen können je nach Typ der Argumente unterschiedliche Implementierungen haben und so Aspekte der objektorientierten und der prozeduralen Programmierung verbinden.

Bei der *Deklarativen Programmierung* steht nicht der Lösungsweg, sondern die Beschreibung des Problems selbst im Vordergrund. Der Programmierende formuliert, was erreicht werden soll, und nicht im Detail, wie der Computer dieses Ziel schrittweise erreichen soll. Der Lösungsweg wird dabei anhand vorgegebener Regeln und Mechanismen automatisch vom System ermittelt.

Ein typisches Merkmal deklarativer Ansätze ist, dass der Fokus auf der Definition von Beziehungen, Bedingungen oder Zielen liegt, anstatt auf einer expliziten Abfolge von Befehlen. Der Programmierende beschreibt also das gewünschte Ergebnis in einer formalen Sprache, und der Computer sucht eigenständig nach einer Lösung, die diesen Bedingungen genügt.

Bekannte Ansätze der deklarativen Programmierung sind zum Beispiel die logische Programmierung, wie sie in der Sprache [Prolog](#) umgesetzt ist, oder die funktionale Programmierung, wie in [Haskell](#). In der logischen Programmierung werden Fakten und Regeln formuliert, aus denen das System per Inferenz Schlüsse zieht. In der funktionalen Programmierung werden Berechnungen durch die Komposition von Funktionen beschrieben, wobei der Zustand der Daten nicht explizit verändert wird. Deklarative Programmierung eignet sich besonders für Probleme, bei denen der Lösungsweg komplex und vielfältig sein kann, aber die gewünschten Eigenschaften des Ergebnisses klar spezifiziert werden können. Typische Anwendungsgebiete sind zum Beispiel Datenbankabfragen (z. B. SQL), Wissensbasen, Constraint-Lösungen und formale Beweise.

Ein modernes Beispiel für den Einsatz deklarativer Programmierung ist das System [Lean](#), ein sogenannter *Theorem Prover*. Lean wird eingesetzt, um mathematische Sätze formal zu beweisen. Hierbei formuliert der Benutzer in einer Mischung aus deklarativen und interaktiven Befehlen die Axiome, Definitionen und zu beweisenden Aussagen. Das System prüft automatisch, ob die angegebenen Schritte gültig sind, und ergänzt nach festgelegten Regeln die fehlenden Teile des Beweises. Damit verbindet Lean deklarative Programmierung mit automatischer Beweisführung und erlaubt es, große mathematische Theorien formal korrekt zu entwickeln. Für eine Einführung in Lean im Rahmen mathematischer Beweisführung, siehe zum Beispiel [The Mechanics of Proof](#).

Kompilierte und interpretierte Programmiersprachen

Kompilierte Programmiersprachen zeichnen sich dadurch aus, dass der gesamte Quellcode vor der Ausführung in Maschinensprache übersetzt wird. Für diese Übersetzung wird ein spezielles Programm, der sogenannte *Compiler*, verwendet. Der vom Compiler erzeugte Maschinencode wird direkt vom Prozessor ausgeführt, was eine sehr schnelle Ausführung des Programms ermöglicht. Da das ausführbare Programm bereits in Maschinensprache vorliegt, muss es zur Laufzeit nicht mehr übersetzt werden. Allerdings ist es notwendig, bei jeder Änderung des Quellcodes das Programm erneut zu kompilieren. Typische Beispiele für kompilierte Programmiersprachen sind Java, C und C++.

Im Gegensatz dazu wird bei *interpretierten Programmiersprachen* der Quellcode während der Ausführung in eine maschinennahe Sprache übersetzt. Dies übernimmt ein spezielles Programm, der sogenannte *Interpreter*. Da die Übersetzung und Ausführung gleichzeitig stattfinden, ist die Ausführung im Vergleich zu kompilierten Programmen in der Regel langsamer. Ein Vorteil besteht jedoch darin, dass der Quellcode bei Änderungen nicht neu interpretiert werden muss – das Programm kann sofort ausgeführt werden. Typische Beispiele für interpretierte Programmiersprachen sind Python und **R**.

Im Sinne der hier eingeführten Begriffsbildungen ist **R** eine interpretierte imperative Programmiersprache der 4. Generation, die perse objektorientiert ist, aber prozedural genutzt werden kann und auf die statistische Datenanalyse zugeschnitten ist.

11.4. R

R ist eine Programmiersprache und ein Softwarepaket, das ursprünglich von Ross Ihaka (Ihaka & Gentleman (1996)) entwickelt wurde. Es handelt sich um einen freien Dialekt der proprietären Software S, die von Becker et al. (Becker et al. (1988)) eingeführt wurde. Heute wird **R** vom [R Core Team](#) und der [R Foundation](#) weiterentwickelt und gepflegt. **R** verfügt über eine große und aktive Community, die bislang rund 20.000 **R**-Pakete beigetragen hat, die das System um zahlreiche Erweiterungen ergänzen. Dabei bleibt der Kern von **R** bewusst evolviert und konservativ, während die Vielzahl an [R-Paketen](#) eine konsistente und progressive Weiterentwicklung ermöglichen. **R** lässt sich einfach von [CRAN](#) herunterladen und installieren.

Mit **R** kann man Datensätze laden, manipulieren und in verschiedenen Formaten speichern. Die Sprache ermöglicht es, eine Vielzahl von Berechnungen auf unterschiedlichen Datenstrukturen durchzuführen. Darüber hinaus bietet **R** ein breites Spektrum statistischer Analysemethoden, die sich direkt auf die Daten anwenden lassen. Typisch ist auch das Schreiben von reproduzierbaren Datenanalyseskripten, mit denen sich Analysen automatisieren und Ergebnisse nachvollziehbar dokumentieren lassen. Schließlich eignet sich **R** hervorragend zur Erstellung von Abbildungen und Visualisierungen, um Daten anschaulich zu präsentieren.

Allerdings kann mit **R** bisher auch einige Dinge, die in der psychologischen Datenwissenschaft von Bedeutung sind, nicht so gut umsetzen. So fehlt es **R** per se an einer modernen, ansprechenden Entwicklungsumgebung, allerdings schafft hier die Einbettung von **R** in VSCode Abhilfe. Im Bereich des Scientific Computing stößt **R** an seine Grenzen. Hier bieten Sprachen wie Python, Matlab oder Julia oft leistungsfähigere und flexiblere Werkzeuge. Schließlich eignet sich **R** kaum zum Programmieren psychologischer Experimente, wofür ebenfalls eher Python oder Matlab zum Einsatz kommen.

Um Hilfe zu **R** zu bekommen, gibt es viele Möglichkeiten. Ganz klassisch hilft oft schon eine schnelle Suche bei [Google](#), eine Anfrage bei [ChatGPT](#) oder eine Interaktion mit [CoPilot](#) in VSCode. Während der Programmierung und wenn man den Namen einer Funktion bereits kennt, kann man direkt in **R** Hilfe aufrufen:

```
?mean  
help(mean)
```

Für umfassendere, längere Tutorials und Hintergrundinformationen lohnt sich ein Blick in die sogenannten *Vignetten* der installierten Pakete:

```
browseVignettes()
```

Darüber hinaus gibt es zahlreiche hilfreiche Online-Ressourcen, zum Beispiel die spezialisierte Suchmaschine [rseek](#) oder die vielen Artikel und Anleitungen auf [R-Bloggers](#), wobei die meisten Informationen auf diesen Plattformen sicherlich auch Teil der über ChatBots zugänglichen Datenbanken wie [ChatGPT](#), [Gemini](#), [Claude](#) oder [DeepSeek](#) sind.

11.5. Visual Studio Code

[Visual Studio Code \(VSCode\)](#) ist ein kostenloser Quelltext-Editor von Microsoft, der seit 2015 für Windows, macOS und Linux verfügbar ist. Grob gesagt ist VSCode so etwas wie die (vielleicht eher bekannte) RStudio Umgebung, nur eben für alle Programmiersprachen. VSCode lässt sich über Extensions flexibel erweitern und als integrierte Entwicklungsumgebung (IDE) für nahezu jede Sprache nutzen, zum Beispiel für **R**, Python, Julia, Shell, Quarto und viele mehr. Seit 2018 gilt VSCode als der beliebteste Editor überhaupt, wie jährliche Umfragen zeigen. Bemerkenswerterweise ist damit ausgerechnet ein Microsoft-Produkt der meistgenutzte Editor auch in der Linux-Welt. VSCode ist sehr konfigurierbar, stark von der Community geprägt und ermöglicht über Microsofts GitHub die Synchronisation der Einstellungen und Erweiterungen über mehrere Endgeräte hinweg. Eine ausführliche Dokumentation mit Anleitungen zur Nutzung man unter www.code.visualstudio.com/docs. Einen Einstieg zum Arbeiten mit **R** in VSCode liefert die entsprechende Dokumentation [R in Visual Studio Code](#).

12. Arithmetik und Variablen

12.1. Arithmetik

Öffnet man ein **R** Terminal, so kann man dieses zunächst einmal als Taschenrechner nutzen, wie folgendes Beispiele zeigt

```
1+1 # Addition
```

```
[1] 2
```

Im Terminal erscheint dabei einerseits das Ergebnis 2, andererseits mit [1] ein Hinweis darauf, dass das Ergebnis der erste Eintrag in dem von **R** erzeugten Ergebnisvektor ist. Vektoren werden wir in Kapitel 14 im Detail behandeln. Weitere Beispiele für den Gebrauch eines **R** Terminals als Taschenrechner sind folgende.

```
2*3 # Multiplikation
```

```
[1] 6
```

```
sqrt(2) # Quadratwurzel (square root)
```

```
[1] 1.414214
```

```
exp(0) # Exponentialfunktion
```

```
[1] 1
```

```
log(1) # Logarithmusfunktion
```

```
[1] 0
```

Tabelle 12.1 gibt einen ersten Überblick über die in **Base R** verfügbaren arithmetischen Operatoren. Neben den vertrauten Operationen der Addition, Subtraktion, Multiplikation, Division und Potenz hält **Base R** zum Beispiel mit der Matrixmultiplikation, der ganzzahligen Teilung ($5 \% \% 2 = 2$) und der Modulooperation, die den ganzzahligen Rest bei ganzzahliger Teilung ausgibt ($5 \% 2 = 2$), auch Operatoren für spezielle arithmetische Operationen bereit.

Tabelle 12.1. Beispiele arithmetischer Operatoren in **Base R**

Operation	Operator
Addition	+
Subtraktion	-
Multiplikation	*
Division	\
Potenz	^
Matrimultiplikation	%*%
Ganzzahlige Teilung	%/%
Modulo	%%

Neben den arithmetischen Operatoren zur Verknüpfung zweier Zahlen bietet **Base R** natürlich auch die Möglichkeit mathematische Standardfunktionen auszuwerten, einen ersten Überblick gibt [Tabelle 12.2](#).

Tabelle 12.2. Beispiele mathematischer Funktionen in **Base R**

Funktion	Aufruf
Exponentialfunktion	<code>exp(x)</code>
Logarithmusfunktion	<code>log(x)</code>
Betrag	<code>abs(x)</code>
Wurzel	<code>sqrt(x)</code>
Aufrunden	<code>ceiling(x)</code>
Abrunden	<code>floor(x)</code>
Mathematisches Runden	<code>round(x)</code>

Wenn auch die in [Tabelle 12.1](#) aufgeführten Operatoren und die in [Tabelle 12.2](#) aufgeführten Funktionen auf den ersten Blick einen etwas anderen Charakter haben, so unterscheidet **Base R** formal nicht zwischen Operatoren und Funktionen. Insbesondere können Operatoren mithilfe der sogenannten *Infixnotation* auch also Funktionen mehrerer Argumente genutzt und verstanden werden. Untenstehender **R** Code zeigt, wie die arithmetische Verknüpfung $2 + 3$ als Ausführung einer Funktion der Form $+ : \mathbb{R}^2 \rightarrow \mathbb{R}$ mit dem Namen “+” verstanden werden kann

```
+^(2,3) # Infixnotation für 2 + 3
```

[1] 5

Schließlich bietet **Base R** wie jede Programmiersprache auch die Möglichkeit Ausdrücke auf ihren logischen Gehalt hin auszuwerten. Dabei ist Logik hier im Sinne der aus [Kapitel 1](#) Aussagenlogik zu verstehen, in der Aussagen einen von zwei Werten annehmen können, wahr (TRUE) oder falsch (FALSE). Die in [Tabelle 12.3](#) aufgeführten Relationsoperatoren `<`, `<=`, `>`, und `>=` werden zumeist auf numerische Werte angewendet, die Gleichheitsoperatoren `==` und `!=` können auf beliebige Datenstrukturen angewendet werden und die Operatoren zur Verknüpfung logischer Werte `&` und `|` entsprechen den aus [Kapitel 1](#) bekannten Begriffen der logischen Konjunktion (“und”) und Disjunktion (“nicht-exklusive oder”)

Tabelle 12.3. Beispiele logischer Operatoren in **R**

Logische Verknüpfung	Operator
Gleich	==
Ungleich	!=
Konjunktion	&
Disjunktion	
Kleiner	<
Kleiner-gleich	<=

Die in Tabelle 12.1, Tabelle 12.2 und Tabelle 12.3 aufgelisteten Operatoren und Funktionen stellen nur einen kleinen Auszug der von **Base R** bereitgestellten Operatoren und Funktionen dar. Einen vollständigen Überblick bietet folgender **R** Aufruf, der die Namen aller von **Base R** bereitgestellten Funktionen auflistet. Wir werden viele dieser Funktionen im weiteren Verlauf kennenlernen.

```
names(methods:::BasicFuncsList)
```

```
[1] "$"                "$<-"              "["
[4] "[<-"              "[[<-"             "[[<-"
[7] "%*%"              "crossprod"        "tcrossprod"
[10] "xtfrm"            "c"                 "all"
[13] "any"              "sum"               "prod"
[16] "max"              "min"               "range"
[19] "cummax"           "rep"               "~"
[22] "cos"              "levels<-"         "cumsum"
[25] "asin"             "anyNA"             "<="
[28] "Conj"             "exp"               "is.matrix"
[31] "log10"            "as.environment"    "ceiling"
[34] "asinh"            "abs"               "as.raw"
[37] "is.infinite"      "is.array"          "floor"
[40] "=="              "sign"              "cummin"
[43] "|"               "round"             "Re"
[46] "!"               "is.numeric"        "acosh"
[49] "!="              "names<-"           "&"
[52] "atanh"           "sinh"              ">="
[55] "sin"             "*"                  "atan"
[58] "+"               "length"            "length<-"
[61] "sinpi"           "-"                  "is.nan"
[64] "/"              "sqrt"              "%/%"
[67] "as.numeric"      "seq.int"           "trunc"
[70] "digamma"         "acos"              "<"
[73] "as.logical"      "cosh"              "Mod"
[76] "tanpi"           "dimnames<-"        "log2"
[79] ">"              "tanh"              "is.na"
[82] "dim"             "signif"            "gamma"
[85] "is.finite"       "as.integer"        "dim<-"
[88] "as.double"       "lgamma"            "Arg"
[91] "loglp"           "trigamma"          "%/%"
[94] "tan"             "cumprod"           "Im"
[97] "expm1"           "as.call"           "log"
[100] "as.complex"      "cospi"             "as.character"
[103] "dimnames"        "names"             "("
[106] ":@"              "="                 "@&"
[109] "{"               "~"                 "&&"
[112] ".C"              "baseenv"           "quote"
[115] ":@"              "<-"               "is.name"
[118] "if"              "||"               "attr<-"
[121] "untracemem"      ".cache_class"      "substitute"
[124] "interactive"     "is.call"           "switch"
[127] "function"        "is.single"         "is.null"
[130] "is.language"     "is.pairlist"       ".External.graphics"
[133] "declare"         "globalenv"         "class<-"
[136] ".Primitive"      "is.logical"        "enc2utf8"
[139] "UseMethod"       ".subset"           "proc.time"
[142] "enc2native"      "repeat"            "::::"
[145] "<<-"            "@<-"              "missing"
[148] "nargs"           "isS4"              ".isMethodsDispatchOn"
[151] "forceAndCall"    "Exec"              ".primTrace"
[154] "storage.mode<-" ".Call"             "unclass"
[157] "gc.time"         ".subset2"          "environment<-"
[160] "emptyenv"        "seq_len"          ".External2"
[163] "is.symbol"       "class"             "on.exit"
[166] "is.raw"          "for"               "is.complex"
[169] "list"            "invisible"         "is.character"
[172] "oldClass<-"     "is.environment"    "attributes"
[175] "break"           "return"            "attr"
[178] "tracemem"        "next"              ".Call.graphics"
```

[181] "standardGeneric"	"is.atomic"	"retracemem"
[184] "expression"	"is.expression"	"call"
[187] "is.object"	"pos.to.env"	"attributes<-"
[190] ".primUntrace"	"...length"	"Tailcall"
[193] ".External"	"oldClass"	".Internal"
[196] ".Fortran"	"browser"	"is.double"
[199] ".class2"	"while"	"nzchar"
[202] "is.list"	"lazyLoadDBfetch"	"unCfillPOSIXlt"
[205] "...elt"	"...names"	"is.integer"
[208] "is.function"	"is.recursive"	"seq_along"
[211] "unlist"	"as.vector"	"lengths"

12.2. Präzedenz

Eine wichtige Eigenschaft bei der Benutzung von Operatoren in der Programmierung ist ihre durch die jeweilige Programmiersprache festgelegte *Präzedenz*. Dabei handelt es sich um von den Entwicklern der Programmiersprache festgelegten Regeln, in welcher Rangfolge Operatoren in Ausdrücken, in denen mehrere von ihnen vorkommen, ausgeführt werden. Aus der Mathematik ist man insbesondere die Präzedenzregel “Punktrechnung geht vor Strichrechnung” gewöhnt. Diese besagt, dass in Ausdrücken, in denen sowohl Produkte oder Divisionen als auch Summen oder Differenzen vorkommen, die Produkte oder Divisionen zunächst ausgeführt werden und ihre Ergebnisse dann in die Summen- oder Differenzbildung eingehen. So ergibt sich zum Beispiel

$$2 \cdot 5 - 3 = 10 - 3 = 7 \quad (12.1)$$

und nicht etwa

$$2 \cdot 5 - 3 \neq 2 \cdot 2 = 4. \quad (12.2)$$

Wir setzen als bekannt voraus, dass die Präzedenzregeln durch Klammerbildung betont bzw. überschrieben werden können. So ist beispielsweise der Ausdruck in Gleichung 12.1 äquivalent zu

$$(2 \cdot 5) - 3 = 10 - 3 = 7 \quad (12.3)$$

und die in Gleichung 12.2 beabsichtigte Rechnung kann durch Klammersetzung als

$$2 \cdot (5 - 3) = 2 \cdot 2 = 4 \quad (12.4)$$

richtig gestellt werden. Diese vertrauten Rechenregeln und ihre Überschreibung durch Klammern finden sich in **R** entsprechend implementiert:

```
2*5 - 3
```

```
[1] 7
```

```
2*(5 - 3)
```

```
[1] 4
```

Eine weitere als bekannt vorausgesetzte Präzedenzregel ist, dass Potenzen eine höhere Präzedenz als Produkte oder Divisionen und Summen oder Differenzen haben. Es ergibt sich also beispielsweise

$$2^3 + 3 = (2 \cdot 2 \cdot 2) + 3 = 8 + 3 = 11 \quad (12.5)$$

und nicht etwa

$$2^3 + 3 \neq 2^{3+3} = 2^6 = 64 \quad (12.6)$$

Entsprechend gelten in **R**

```
2^3 + 3
```

```
[1] 11
```

```
2^(3 + 3)
```

```
[1] 64
```

Eine wichtige Besonderheit hinsichtlich der Präzedenzregeln in **R** ist, dass auch unitäre Vorzeichen, also Vorzeichen, die eine Zahl als eine negative Zahl identifizieren in den Bereich der Operatorpräzedenz fallen. Wenn man geneigt sein sollte, einen Ausdruck der Form -1^2 als $(-1)^2 = 1$ zu interpretieren, so folgt **R** der Regel, dass in -1^2 zunächst die Potenz und dann das Vorzeichen ausgewertet werden, dass also $-1^2 = -(1^2) = -1$ gilt, wie folgender **R** Code demonstriert.

```
-1^2 # -(1^2) = -1
```

```
[1] -1
```

```
(-1)^2 # (-1)^2 = 1
```

```
[1] 1
```

Weiterhin werden in **R** Potenzen von rechts nach links abgearbeitet, Produkte, Divisionen, Summen und Differenzen dagegen von links nach rechts. So gelten beispielsweise

```
2^2^3 # 2^(2^3) = 2^8 = 256
```

```
[1] 256
```

```
(2^2)^3 # (2^2)^3 = 4^3 = 64
```

```
[1] 64
```

und

```
2+3/4*5 # 2+(3/4)*5 = 2+(0.75*5) = 2+3.75 = 5.75
```

```
[1] 5.75
```

```
2+3/(4*5) # 2+3/(4*5) = 2+3/20 = 2+0.15 = 2.15
```

```
[1] 2.15
```

Generell gilt, dass man in der Programmierung Präzedenzregeln nicht raten sollte oder versuchen sollte, sie logisch herzuleiten, und dann darauf vertrauen sollte, dass die Entwickler der Programmiersprache die Präzedenzregeln schon gemäß des eigenen Verständnisses implementiert haben werden. Stattdessen sollte man Berechnungen immer kritisch prüfen und zur Sicherheit die beabsichtigte Rechnung lieber einmal zuviel mit Klammern betonen als zu wenig.

Neben den hier betrachteten Präzedenzregeln für die Grundrechenarten gibt es in **R** eine ganze Reihe weiterer Regeln zum Umgang mit vielen weiteren Operatoren, von denen wir bisher nur sehr wenige kennengelernt haben. Wenn man sich beim Erlernen einer Programmiersprache mit einem neuen Operator vertraut macht, sollte man also für sich auch unbedingt die Präzedenz dieses Operators klären. Der in Abbildung 12.1 gezeigte und mit dem Befehl `?Syntax` aufzurufende Text der **R** Dokumentation diskutiert die in **R** geltenden Präzedenzregeln vollumfänglich und sollte bei der Arbeit mit **R** häufig konsultiert werden.

Operator Syntax and Precedence

Description

Outlines **R** syntax and gives the precedence of operators.

Details

The following unary and binary operators are defined. They are listed in precedence groups, from highest to lowest.

<code>:: :::</code>	access variables in a namespace
<code>\$ @</code>	component / slot extraction
<code>[[[</code>	indexing
<code>^</code>	exponentiation (right to left)
<code>- +</code>	unary minus and plus
<code>:</code>	sequence operator
<code>%any% ></code>	special operators (including <code>%</code> and <code>%/%</code>)
<code>* /</code>	multiply, divide
<code>+ -</code>	(binary) add, subtract
<code>< > <= >= == !=</code>	ordering and comparison
<code>!</code>	negation
<code>& &&</code>	and
<code> </code>	or
<code>~</code>	as in formulae
<code>-> ->></code>	rightwards assignment
<code><- <<-</code>	assignment (right to left)
<code>=</code>	assignment (right to left)
<code>?</code>	help (unary and binary)

Within an expression operators of equal precedence are evaluated from left to right except where indicated. (Note that `=` is not necessarily an operator.)

Abbildung 12.1. Präzedenzregeln in **R**.

12.3. Variablen

In der Programmierung sind *Variablen* abstrakte Behälter für Daten, deren Inhalt im Verlauf einer Programmausführung geändert werden kann. Variablen werden im Programmcode mit Namen bezeichnet und besitzen bei Programmausführung eine Adresse im Arbeitsspeicher. Variablen können verschiedenen *Typs* sein, d.h. unterschiedliche Arten von Daten, und nur diese, repräsentieren. In traditionellen Programmiersprachen wird der Typ einer Variable explizit deklariert, d.h. es wird zu Beginn der Programmausführung festgelegt, welche Art von Daten eine bestimmte Variable repräsentiert. Die Deklaration einer Variable A, die ganze Zahlen (und nur diese) repräsentiert hat etwa die Form

```
VAR A : INTEGER      # A ist eine Variable vom Typ Integer (ganze Zahl)
```

In der Folge kann der Variable A dann ein Wert vom Typ ganze Zahl zugewiesen werden, beispielsweise in der Form

```
A := 1               # Der Variable A wird der numerische Wert 1 zugewiesen
```

In 4GL Sprachen wie **R** wird der Variablentyp üblicherweise direkt durch eine Initialisierungsanweisung festgelegt. So deklariert folgender **R** Code die Variable a gleichzeitig als vom Typ `double` (Dezimalzahl) mit dem Wert 1,

```
a = 1               # a ist eine Variable vom Typ double, ihr Wert ist 1
```

In den allermeisten Programmiersprachen ist der Zuweisungsbefehl `=`. **R** ist dahingehend besonders, als es mit `->` und `<-` Zuweisungsbefehle für Zuweisungen von Werten von links nach rechts oder von links nach rechts erlaubt, wobei letzteres auch durch `=` erreicht wird. Offiziell empfohlener Zuweisungsbefehl für **R** ist `<-`, da dies aber sehr idiosynkratisch ist, benutzen wir `=` wie in jeder anderen Programmiersprache. Ein erstes Beispiel soll den Umgang mit Variablen verdeutlichen.

Beispiel

Wir betrachten folgende Textaufgabe:

“Lina geht ins Schreibwarengeschäft und kauft vier Hefte und zwei Stifte. Ein Heft kostet 1 Euro und ein Stift kostet 3 Euro. Wieviele Dinge kauft Lina insgesamt und wieviel Euro muss Lina insgesamt bezahlen?”

Um die Aufgabe zu lösen, definieren wir zunächst die Variablen `anzahl_hefte` und `anzahl_stifte` und weisen ihnen dabei ihre jeweiligen Werte aus der Aufgabe zu.

```
anzahl_hefte = 4  
anzahl_stifte = 2
```

Nach Zuweisung existieren die Variablen mit den ihnen zugewiesenen Werten nun im Arbeitsspeicher, dem sogenannten *Workspace*. Der Befehl `ls()` zeigt die existierenden benutzbaren Variablen im Arbeitsspeicher an.

```
ls()
```

```
[1] "anzahl_hefte" "anzahl_stifte"
```

Entscheidend ist, dass die Variablennamen jetzt wie Zahlen in Berechnungen genutzt werden können `print()` gibt dabei Variablenwerte in einem **R** Terminal aus.

```
anzahl_dinge = anzahl_hefte + anzahl_stifte  
print(anzahl_dinge)
```

```
[1] 6
```

Um den Gesamtpreis zu berechnen, definieren wir als nächstes die Variablen `preis_heft` und `preis_stift` anhand der Aufgabenspezifikation.

```
preis_heft      = 1  
preis_stift     = 2
```

Der Gesamtpreis berechnet sich dann wie folgt.

```
gesamtpreis = anzahl_hefte*preis_heft + anzahl_stifte*preis_stift  
print(gesamtpreis)
```

```
[1] 8
```

Variablen im Workspace können auch wieder gelöscht werden. `rm()` (remove) erlaubt das Löschen einzelner Variablen

```
rm(gesamtpreis)  
ls()
```

```
[1] "anzahl_dinge" "anzahl_hefte" "anzahl_stifte" "preis_heft"  
[5] "preis_stift"
```

`rm(list=ls())` löscht alle Variablen

```
rm(list = ls())  
ls()
```

```
character(0)
```

Variablennamen

Prinzipiell ist man in der Wahl der Namen für Variablen mit kleineren Einschränkungen frei. Zulässige Variablennamen in **R**

- bestehen aus Buchstaben, Zahlen, Punkten (.) und Unterstrichen (_),
- beginnen mit einem Buchstaben oder einem Punkt, dann allerdings nicht gefolgt von einer Zahl und
- dürfen keine in **R** schon belegten Ausdrücke wie `for`, `if`, `NaN`, ... sein

Hilfestellung zu den Einschränkungen von Variablennamen geben `?reserved`. Eine Funktion, um aus beliebigen Vorschlägen einen syntaktisch zulässigen Variablennamen zu generieren, ist `make.names()`. Ihre Hilfe beschreibt auch die hier skizzierten Regeln zur Variablenbenennung in weiteren Details. Generell sind sinnvolle Variablen kurz, um im Code Platz zu sparen und aussagekräftig, um das menschliche Verständnis des Codes zu erhöhen. Meist bestehen Variablen nur aus Kleinbuchstaben und Unterstrichen.

Variablenrepräsentation

Wir wollen uns noch etwas genauer mit dem Verhältnis von Variablennamen (Bezeichnen) und Variableninhalten (Objekten im Arbeitsspeicher) befassen. Die folgenden Begriffe und Zusammenhänge sind insbesondere dann wichtig, wenn Berechnungen an die Grenzen des verfügbaren Arbeitsspeichers gelangen und optimiert werden müssen.

Wir beginnen mit dem Begriff des *Binding*. Dazu betrachten wir zunächst die folgende Variableninitialisierung:

```
x = 1
```

Intuitiv wird durch dieses Statement eine Variable genannt `x` mit dem Wert 1 erzeugt. Auf technischer Ebene geschehen dabei de-facto geschehen zwei Dinge (vgl. Abbildung 12.2):

1. **R** erzeugt ein Objekt (Vektor mit Wert 1) mit Arbeitsspeicheradresse `'lobstr::obj_addr(x)'`.
2. **R** verbindet dieses Objekt mit dem Namen `x` (*binding*), der das Objekt im Arbeitsspeicher referenziert.

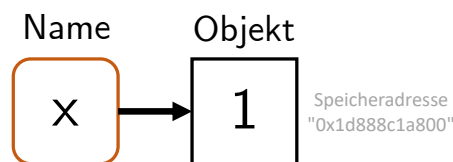


Abbildung 12.2. Binding eines Namens mit einem Objekt im Arbeitsspeicher.

Betrachten wir nun folgendes Statement :

```
y = x
```

Intuitiv wird dabei eine Variable, genannt `y`, mit Wert gleich dem Wert von `x` erzeugt. Auf technischer Ebene wird dabei neuer Name `y` erzeugt, der dasselbe Objekt im Arbeitsspeicher referenziert wie `x` (vgl. .@fig-binding-2).

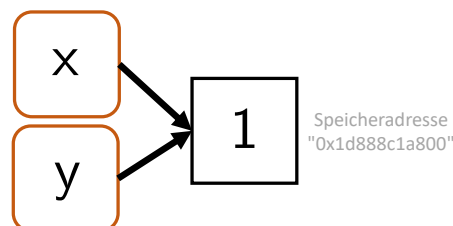


Abbildung 12.3. Binding zweier Namen mit einem Objekt im Arbeitsspeicher.

Man kann sich mithilfe von `lobstr::obj_addr(x)` und `lobstr::obj_addr(y)` davon überzeugen, dass die Adressen des Objekts im Arbeitsspeicher identisch sind

```
print(lobstr::obj_addr(x))
```

```
[1] "0x26bf6520388"
```

```
print(lobstr::obj_addr(y))
```

```
[1] "0x26bf6520388"
```

Das betreffende Objekt wird also bei Allokation zu `y` nicht kopiert und dadurch Arbeitsspeicher eingespart.

Weiterhin nutzt **R** das sogenannte *Copy-on-modify* Prinzip. Dieses Prinzip besagt, dass ein Objekt im Arbeitsspeicher kopiert wird und eine neue Adresse erhält, sobald es modifiziert wird. Das ursprüngliche Objekt im Arbeitsspeicher bleibt dabei unverändert. Abbildung 12.4 und folgender **R** Code verdeutlichen zunächst das *Copy-on-Modify* Prinzip

```
x = 1      # Objekt (z.B. 0x74b) erzeugt, x referenziert Speicheradresse des Objektes
y = x      # y referenziert dieselbe Speicheradresse wie x (0x74b)
y = 3      # y modifiziert, eine modifizierte Kopie (z.B. 0xcd2) wird gespeichert
y          # y referenziert jetzt (0xcd2)
```

```
[1] 3
```

```
x          # x referenziert weiterhin (0x74b)
```

```
[1] 1
```

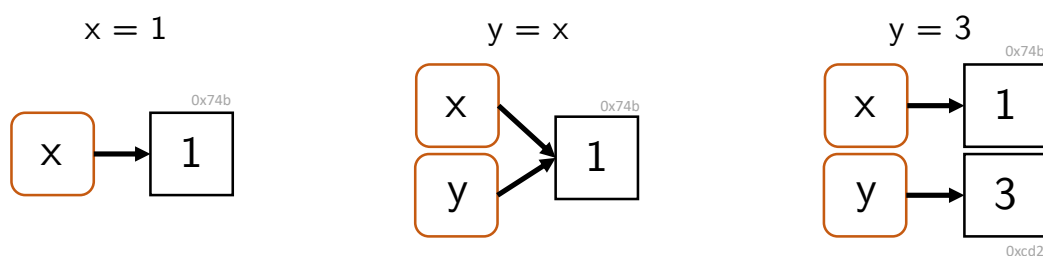


Abbildung 12.4. Copy-on-modify.

Die Tatsache, dass Objekte bei Veränderung im Arbeitsspeicher kopiert werden, die ursprünglichen Objekte aber im Prinzip unverändert erhalten bleiben, wird auch als die *Immutability* von Objekten in **R** bezeichnet. Allerdings greift die Immutability von Objekten in **R** nicht besonders weit, wie folgendes Beispiel zeigt. Dabei wird zunächst ein Objekt einer `1` im Arbeitsspeicher erzeugt, das mit dem Variablennamen `x` referenziert werden kann. Per *Indexing* (vgl. [?@sec-Datenstrukturen-und-Datenmanagement](#)) wird das Objekt dann bei einer beabsichtigten Modifikation im Arbeitsspeicher neu adressiert.

```
x = 1 # Objekt erzeugt, x referenziert Speicheradresse des Objektes
print(lobstr::obj_addr(x)) # Speicheradresse des Objekts
```

```
[1] "0x26bf952f038"
```

```
x[1] = 2 # Objekt verändert
print(lobstr::obj_addr(x)) # neue Speicheradresse
```

```
[1] "0x26bf97c1c40"
```

Das Copy-on-modify-Prinzip gilt auch in umgekehrter Reihenfolge, d.h. wird ein Objekt, auf das ein weiterer Variablenname zeigt, verändert, so ändert sich das ursprünglich referenzierte Objekt nicht. Folgender **R** Code mag dies verdeutlichen.

```
x = 1 # Objekt (z.B. 0x74b) erzeugt, x referenziert Speicheradresse des Objektes
y = x # y referenziert dieselbe Speicheradresse wie x (0x74b)
x = 3 # Ein neues Objekt (z.B. 0x2a08) wird erzeugt, x referenziert nun 0x2a08
y # y referenziert weiterhin das ursprüngliche Objekt (0x74b)
```

```
[1] 1
```

Schließlich ist auch ein *Unbinding* von Objekten im Arbeitsspeicher möglich, z.B. als Folge folgenden **R** Codes (vgl. Abbildung 12.5):

```
x = 1 # x referenziert Objekt (0x74b)
x = "a" # x referenziert Objekt (0x2a08), Objekt (0x74b) jetzt ohne Referenz
```

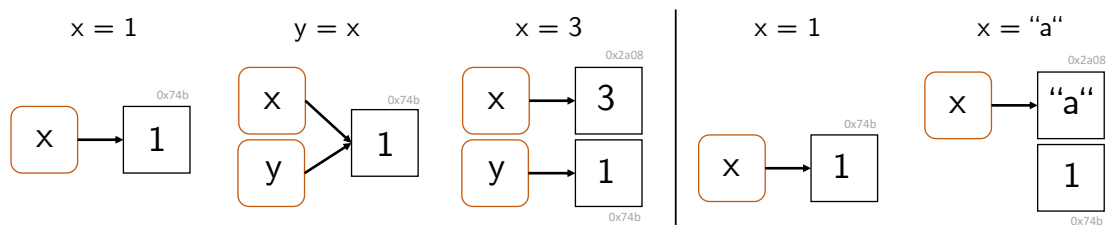


Abbildung 12.5. Unbinding.

Das Löschen von nicht-referenzierten Objekten im Arbeitsspeicher wird als *Garbage collection* bezeichnet. **R** löscht nicht referenzierte Objekte im Arbeitsspeicher automatisch, allerdings meist erst dann, wenn der Speicher benötigt wird, es also wirklich nötig ist. Generell funktioniert dies aber gut und es ist nicht nötig, aktiv die von **R** bereitgestellte Garbage Collection Funktion `gc()` zu benutzen.

Repräsentation fehlender Werte

In der datenanalytischen Programmierung hat man es desöfteren mit fehlenden Werten zu tun, z.B. wenn eine experimentelle Messung bei einer Versuchsperson abgebrochen wurde. Zur Repräsentation fehlender Werte nutzt **R** `NA` (*not applicable*). Versucht man mit `NA` zu rechnen, so ergibt dies meist wieder `NA`:

```
3 * NA # Multiplikation eines NA Wertes ergibt NA
```

```
[1] NA
```

```
NA < 2 # Relationaler Vergleich eines NA Wertes ergibt NA
```

```
[1] NA
```

```
NA^0 # NA hoch 0 ergibt 1, weil jeder Wert hoch 0 eins ergibt (?)
```

```
[1] 1
```

```
NA & FALSE # NA UND FALSE ergibt FALSE, weil jeder Wert UND FALSE FALSE ergibt
```

```
[1] FALSE
```

Auf NA testet man mit `is.na()`:

```
x = c(NA, 5, NA, 10) # Vektor mit NAs  
x == NA # Kein Testen auf NAs : 5 == NA ist NA, nicht FALSE
```

```
[1] NA NA NA NA
```

```
is.na(x) # Logisches Testen auf NA
```

```
[1] TRUE FALSE TRUE FALSE
```

Neben NA besitzt **R** mit NaN auch eine Möglichkeit zur Darstellung mathematisch nicht definierter Werte. Im Gegensatz zu NA ist NaN immer vom numerischen Typ. Die Ergebnisse beim Rechnen mit NaN ähneln denen mit NA.

```
3 * NaN # Multiplikation eines NaN Wertes ergibt NaN
```

```
[1] NaN
```

```
NaN < 2 # Relationaler Vergleich eines NaN Wertes ergibt NA
```

```
[1] NA
```

```
NaN^0 # NaN hoch 0 ergibt 1, weil jeder Wert hoch 0 eins ergibt (?)
```

```
[1] 1
```

```
NaN & FALSE # NaN UND FALSE ergibt FALSE, weil jeder Wert UND FALSE FALSE ergibt
```

```
[1] FALSE
```

Auf NaN testet man mit `is.nan()`


```
x = c(NaN, 5, NaN, 10) # Vektor mit NaNs  
is.nan(x)              # Logisches Testen auf NaN
```

13. Datenstrukturen

In jeder Programmiersprache gibt es abstrakte Behälter, um verschiedene Arten von Daten zu repräsentieren und mit ihnen zu arbeiten. In klassischen 3GL-Programmiersprachen unterscheidet man *fundamentale*, *zusammengesetzte* und *selbstdefinierte* Datenstrukturen.

Fundamentale Datenstrukturen sind in der Programmiersprache vordefiniert. Typische Beispiele sind Behälter für logische Werte (`TRUE`, `FALSE`), ganze Zahlen (`int8`), die z.B. die 256 Werte von -128 bis 127 repräsentieren, Gleitkommazahlen (`single`, `double`) mit unterschiedlicher Genauigkeit, Schriftzeichen (`a`, `b`), die oft in Anführungszeichen gesetzt werden. Mit diesen grundlegenden Datentypen sind typischerweise passende Operationen verknüpft, wie `AND/OR` für logische Werte, `+`, `-` für Ganzzahlen, `+`, `-`, `*` und `\` für Gleitkommazahlen und Konkatenation für Zeichen.

Zusammengesetzte Datenstrukturen sind ebenfalls vordefiniert und dienen dazu, mehrere Variablen gleichen Typs zusammenzufassen. Dazu zählen in **R** zum Beispiel Vektoren, Listen, Arrays, Matrizen und Dataframes, die in diesem Abschnitt behandelt werden. Auch für zusammengesetzte Strukturen existieren spezifische Operationen, wie die Vektorindizierung oder die Matrixmultiplikation.

Schließlich lassen sich in 3GL-Programmiersprachen auch *selbstdefinierte Datenstrukturen* erstellen. Sie ermöglichen die Anpassung der Programmumgebung an spezifische Anwendungsfälle und sind daher besonders praxisrelevant. In der Regel werden sie aus vordefinierten Strukturen zusammengesetzt und können im Rahmen objektorientierter Programmierung auch mit eigenen Operationen assoziiert sein.

Wenn man eine Programmiersprache erlernt ist ein erster fundamentaler Schritt, sich mit ihren Datenstrukturen zumindest rudimentär vertraut zu machen. Feinheiten in der Datenstrukturrepräsentation wird man dann in der täglichen praktischen Arbeit mit der Programmiersprache kennenlernen. Um sich mit den Datenstrukturen einer Programmiersprache vertraut zu machen bieten sich folgende Leitfragen an, denen wir in diesem Abschnitt folgen wollen:

Fundamentale Datenstrukturen

- Welche fundamentalen Datenstrukturen bietet die Sprache an?
- Welche Operationen darauf sind bereits definiert?
- Wie lautet die Syntax zur Definition einer Variable eines fundamentalen Datentyps?
- Wie lautet die Syntax, um vordefinierte Operationen aufzurufen?

Zusammengesetzte Datenstrukturen

- Welche Container und zugehörige Operationen bietet die Programmiersprache?
- Wie lautet die Syntax zum Umgang mit einem Containers?

Selbstdefinierte Datenstrukturen

- Wie erzeugt man selbstdefinierte Datenstrukturen und zugehörige Operationen?
- Wie lautet die Syntax zum Umgang mit einer selbstdefinierten Datenstruktur?

13.1. Datenstrukturen in **R**

Als 4GL Sprache geht **R** in seiner Datenstruktur über die klassischen Datenstrukturen von 3GL Sprachen hinaus. In **R** ist zunächst erstmal alles ein *Objekt*. Egal, ob es sich um einen einzelnen Wert, einen Vektor, eine Funktion oder eine komplexe Datenstruktur handelt – alles wird von **R** als Objekt repräsentiert. Allgemein bezeichnet der Begriff des *Objekts* die Möglichkeit, in einer Sinneinheit sowohl Daten als auch Funktionalität zu repräsentieren. Wir sind hier allerdings primär an der Datenrepräsentation interessiert. Jedes **R** Objekt lässt sich eindeutig beschreiben durch drei zentrale Eigenschaften: seinen *Modus*, seine *Länge* und optional durch weitere *Attribute*.

Der *Modus* eines Objekts gibt an, wie seine Komponenten strukturiert sind. Man unterscheidet dabei zwischen *atomaren* Objekten und *rekursiven* Objekten. *Atomare Objekte* bestehen aus Komponenten gleichen Datentyps. Typische Beispiele sind numerische Vektoren, logische Vektoren oder Zeichenketten. Rekursive Objekte hingegen können Komponenten unterschiedlichen Typs enthalten. Dazu gehören insbesondere Listen und Dataframes, die es erlauben, heterogene Daten zusammenzufassen.

Neben dem Modus besitzt jedes Objekt eine *Länge*, die angibt, wie viele Elemente es enthält. Bei einem Vektor ist die Länge zum Beispiel die Anzahl der enthaltenen Werte, bei einer Liste die Zahl ihrer Komponenten.

Darüber hinaus können Objekte in **R** mit zusätzlichen *Attributen* versehen sein. Attribute sind Metadaten, die zusätzliche Informationen über das Objekt bereitstellen, etwa Namen der Elemente (`names`), Dimensionen (`dim`) für Matrizen und Arrays oder Klasseninformationen (`class`), die für das objektorientierte Programmieren in **R** relevant sind.

Modus, Typ und Klasse

Die Klassifizierung von Datenstrukturen ist etwas komplex, da nebeneinander die Begriffe des *Modus*, des *Typs* und der *Klasse* eines Objekts existieren. In der praktischen Anwendung sind die Feinheiten dieser Klassifizierung meist nicht von Belang, außerde sind der Modus, Typ, und die Klasse eines Objekts auch oft identisch,

Grob gesagt entspricht der *Modus* der grundlegenden Natur eines Objekts und ist insbesondere so gewählt, dass sich eine hohe Kompatibilität zum Datenstrukturklassifikationssystem von **S** (Becker et al. (1988)) besteht. Beispiele für Modi sind `numeric`, `character`, `logical`, oder `list`. Der Modus eines Objektes kann mit der Funktion `mode()` abgerufen werden, wie folgendes Beispiel zeigt.

```
mode(42) # numeric
```

```
[1] "numeric"
```

```
mode(1L) # numeric
```

```
[1] "numeric"
```

```
mode("hallo") # character
```

```
[1] "character"
```

```
mode(TRUE) # logical
```

```
[1] "logical"
```

```
mode(list(1, 2, 3)) # list
```

```
[1] "list"
```

```
mode(factor(c("a", "b"))) # R Faktor => numeric
```

```
[1] "numeric"
```

```
mode(Sys.Date()) # R Datum => numeric
```

```
[1] "numeric"
```

```
mode(lm(mpg ~ wt, data=mtcars)) # R Lineares Model => list
```

```
[1] "list"
```

Etwas näher an der internen Repräsentation eines Objekts im Speicher und damit etwas technischerer Natur ist der *Typ* eines **R** Objekts. Dies wird insbesondere hinsichtlich Objekten von Modus **numeric** deutlich. Hier gibt es die klassischen 3GL Typenunterscheidungen **integer** oder **double**. Ein anderes Beispiel ist der Modus *Faktor*. Wie wir an späterer Stelle sehen werden, sehen *Faktoren* in **R** für den Nutzer aus wie kategorische Daten, oft von Schriftzeichenart, ihr Modus ist allerdings **numeric** und ihr Typ **integer**. Dies hängt damit zusammen, dass Faktoren in **R** oft dazu genutzt werden auf der Benutzerebene experimentelle Bedingungen zu spezifizieren, die meist aussagekräftige Namen wie “Treamtent” oder “Kontrolle” haben, in der Datenanalyse aber durch mathematische Matrizen mit ganzzahligen Einträgen repräsentiert werden. Der Typ eines Objektes kann mit der Funktion `typeof()` abgerufen werden, wie folgendes Beispiel zeigt. Man beachte die Gemeinsamkeiten und Unterschiede zu den oben aufgerufenen Modi der gleichen Objekte.

```
typeof(42) # double
```

```
[1] "double"
```

```
typeof(1L) # integer
```

```
[1] "integer"
```

```
typeof("hallo")           # character
```

```
[1] "character"
```

```
typeof(TRUE)             # logical
```

```
[1] "logical"
```

```
typeof(list(1,2,3))       # list
```

```
[1] "list"
```

```
typeof(factor(c("a", "b"))) # R Faktor => integer
```

```
[1] "integer"
```

```
typeof(Sys.Date())        # R Datum  => double
```

```
[1] "double"
```

```
typeof(lm(mpg ~ wt, data=mtcars)) # R Lineares Model = list
```

```
[1] "list"
```

Die Klasse eines Objektes schließlich spezifiziert die Art eines Objektes aus der Perspektive der objekt-orientierten Programmierung in **R**. Dabei ist es für ein Objekt insbesondere relevant, wie generische Funktionen wie z.B. `print()` oder `summary()` mit dem Objekt umgehen. Im Gegensatz zu Modi und Typen können Programmierende die Klasse von Objekten selbst festlegen und so generische Funktionen entwickeln, die auf viele Objektarten anwendbar sind und trotz ihres gleichen Namens mit unterschiedlichen Objekten unterschiedliche Operationen ausführen. Die Klasse eines Objektes kann mit der Funktion `class()` abgerufen werden, wie folgendes Beispiel zeigt. Man beachte die Gemeinsamkeiten und Unterschiede zu den oben aufgerufenen Modi und Typen der gleichen Objekte.

```
class(42)                 # numeric
```

```
[1] "numeric"
```

```
class(1L)                 # integer
```

```
[1] "integer"
```

```
class("hallo")           # character
```

```
[1] "character"
```

```
class(TRUE) # logical
```

```
[1] "logical"
```

```
class(list(1,2,3)) # list
```

```
[1] "list"
```

```
class(factor(c("a", "b"))) # R Faktor => "factor"
```

```
[1] "factor"
```

```
class(Sys.Date()) # R Datum => "Date"
```

```
[1] "Date"
```

```
class(lm(mpg ~ wt, data=mtcars)) # R Lineares Model => "lm"
```

```
[1] "lm"
```

Zusammengefasst lässt sich festhalten: Der Modus beschreibt die allgemeine Art eines Objekts, der Typ beschreibt seine interne, technische Speicherung und die Klasse definiert die Sichtweise von R für die funktionale Verarbeitung des Objekts. Für viele Objekte stimmen diese drei Begriffe weitgehend überein, aber bei speziellen Objekten wie Faktoren, Datumswerten oder benutzerdefinierten Objekten unterscheiden sie sich.

Atomare und rekursive Objekte

{=html}

14. Vektoren

In diesem Abschnitt wollen wir nun mit den **R** *Vektoren* eine erste grundlegende Datenstruktur genauer kennenlernen. Viele Prinzipien, die am Beispiel von Vektoren erläutert werden können, übertragen sich sinngemäß auf komplexere Datenstrukturen in **R**, sodass ein kompetenter Umgang auch mit der hier betrachteten basalen Datenstruktur im Anwendungskontext unverzichtbar ist. Es sei angemerkt, dass sich der Begriff des *Vektors* in diesem Abschnitt streng auf die entsprechenden **R** Datenstruktur, nicht aber auf das gleichnamige mathematische Objekt beziehen. Auch wenn **R** Vektoren in manchen Eigenschaften ihrem mathematischen Pendant ähneln und bestimmte Operationen zwischen **R** Vektoren Operationen zwischen den entsprechenden mathematischen Objekten abbilden können, so haben Vektoren in der imperativen Programmierung oft viel mehr und andere Eigenschaften als die entsprechenden mathematischen Objekte.

Allgemein betrachten für die Präsentation von Datenstruktur immer zunächst, wie eine bestimmte Datenstruktur mit **R** Code *erzeugt* wird, durch welche Eigenschaften die Datenstruktur *charakterisiert* ist, wie einzelne Elemente der Struktur *indiziert* und bearbeitet werden können und welche *arithmetischen Operationen* für die Datenstruktur zulässig sind. Schließlich betrachten wir auch immer eventuelle weitere Besonderheiten der Datenstruktur. Im Kontext von **R** Vektoren sind dies beispielsweise die Begriffe der *Datentypangleichung* (*Coercion*), dem *Recycling* oder die Repräsentation fehlender Werte in Datensätzen.

R Vektoren sind geordnete Folgen von Datenwerten. Die einzelnen Datenwerte eines Vektors heißen *Elemente* des Vektors. Vektoren, deren Elemente alle vom gleichen Datentyp sind, heißen *atomar*. Mit dem Begriff *Vektor* soll hier immer ein atomarer Vektor gemeint sein. Die zentralen Datentypen für Vektoren sind die oben eingeführten *numeric*, (*double*, *integer*), *logical* und *character*. Abbildung 14.1 zeigt graphische Darstellungen entstprechender Vektoren.



Abbildung 14.1. Vektoren in **R**.

14.1. Erzeugung

Wir betrachten zunächst die Erzeugung einelementiger Vektoren, die auch als *Elementarwerte* bezeichnet werden. Wie alle Datenstrukturen werden auch Elementarwerte durch Zuweisungen der Form

```
Elementarwert <- Wert
```

oder

```
Elementarwert = Wert
```

Dabei bestimmt die Spezifikation von **Wert**, welcher Art der erzeugte Elementarwert ist.

Elementarwerte vom Typ **numeric** können durch Zuweisung einer Zahl erzeugt werden. Per Default wird dabei ein Elementarwert vom Typ **double** erzeugt. Die Zahlen können dabei ohne Dezimalstellen, in Dezimalnotation oder wissenschaftlicher Notation spezifiziert werden. Weitere mögliche Werte zur Erzeugung eines Elementarwerts vom Typ **double** sind **Inf** und **-Inf** für positiv und negativ Unendlich und **NaN** (Not-a-Number) als Platzhalter für numerische Werte.

```
x = 1                      # Einelementiger Vektor
typeof(x)                  # vom Typ double
```

```
[1] "double"
```

```
y = 2.1e2                  # Einelementiger Vektor
typeof(y)                  # vom Typ double
```

```
[1] "double"
```

```
z = Inf                    # Einelementiger Vektor
typeof(z)                  # vom Typ double
```

```
[1] "double"
```

Numerische Elementarwerte vom Typ **integer** wird wie **double** ohne Dezimalstellen erzeugt, gefolgt von einem **L** für **long integer**.

```
x = 1L                     # Einelementiger Vektor
typeof(x)                  # vom Typ integer
```

```
[1] "integer"
```

```
y = 200L                   # Einelementiger Vektor
typeof(y)                  # vom Typ integer
```

```
[1] "integer"
```

Elementarwerte vom Typ **logical** können durch Zuweisung der logischen Werte **TRUE** oder **FALSE** bzw. abgekürzt **T** oder **F** erzeugt werden.

```
x = TRUE                   # Einelementiger Vektor
typeof(x)                  # vom Typ logical
```

```
[1] "logical"
```



```
y = F                                # Einelementiger Vektor
typeof(y)                            # vom Typ logical
```

```
[1] "logical"
```

Elementarwerte vom Typ `character` schließlich werden durch die Zuweisung von Werten in Hochkommata (`'a'`) oder Anführungszeichen (`"a"`) erzeugt

```
x = "a"                              # Einelementiger Vektor
typeof(x)                            # vom Typ character
```

```
[1] "character"
```

```
y = 'test'                          # Einelementiger Vektor
typeof(y)                            # vom Typ character
```

```
[1] "character"
```

Um nun mehrere Elementarwerte in einem Vektor zusammenzufügen, konkateniert (von lat. *catena* (Kette), engl. *concatenate* (verketten)) man sie mit der Funktion `c()`.

```
x = c(1,2,3)                        # numeric vector [1,2,3]
s = c("a", "b", "c")                # character vector ["a", "b", "c"]
l = c(TRUE, FALSE)                  # logical vector [TRUE, FALSE]
```

```
x
```

```
[1] 1 2 3
```

```
s
```

```
[1] "a" "b" "c"
```

```
l
```

```
[1] TRUE FALSE
```

Es können nicht nur Elementarwerte, sondern auch Vektoren mit Elementarwerten oder anderen Vektoren konkateniert werden. Mit obigem Vektor `x` erzeugt man beispielsweise folgende Vektoren `y` und `z`:

```
y = c(0,x,4)                        # numeric vector [0,1,2,3,4] mit
z = c(x,y)                          # numeric vector [1,2,3,0,1,2,3,4]
```

```
y
```

```
[1] 0 1 2 3 4
```

z

[1] 1 2 3 0 1 2 3 4

Vektoren müssen nicht mit bereits bestehenden Werten erzeugt werden, sondern es können auch “leere” Vektoren von einem gewünschten Datentyp erzeugt werden. **R** stellt dafür die Funktionen `vector()` sowie `double()`, `integer()`, `logical()` und `character()` bereit. Dies ist potenziell hilfreich, um zu Beginn einer längeren Berechnung bereits Arbeitsspeicher des Computers belegt zu haben, sodass während der Programmausführung nicht nach Arbeitsspeicher gesucht werden muss, was unter Umständen die Programmlaufzeit erhöht.

```
# Erzeugen "leerer" Vektoren mit vector()
v = vector("double",3)      # double vector [0,0,0]
w = vector("integer",3)     # integer vector [0,0,0]
l = vector("logical",2)     # logical vector [FALSE, FALSE]
s = vector("character",4)   # character vector ["", "", "", ""]

# Erzeugen leere Vektoren mit vector()
v = double(3)               # double vector [0,0,0]
w = integer(3)              # integer vector [0,0,0]
l = logical(2)              # logical vector [FALSE, FALSE]
s = character(4)            # character vector ["", "", "", ""]
```

Allerdings sind die so erzeugten “leeren” Vektoren nicht wirklich leer: `v` und `w` bestehen aus Nullen, `l` aus den logischen Werten `FALSE` und `s` aus den Zeichen `""`. Zur Initialisierung von Vektoren als Container und anschließender Befüllung im Rahmen eines Programms sind `vector()`, `double()`, `integer()`, `logical()` und `character()` deshalb nur bedingt geeignet: Übersieht der Code die Befüllung einzelner Einträge, so wird mit den von diesen Funktionen erzeugten Werten gerechnet, was nicht unbedingt beabsichtigt ist. In Kapitel 18 lernen wir eine Methode kennen, um Datenstrukturen zu initialisieren und dabei einigermaßen sicheren Code zu generieren.

Sequenzen

Ein häufiges Hilfsmittel in der Programmierung sind Vektoren, die Folgen von numerischen Werten, z.B. als Repräsentation der Definitionsmengen einer univariaten Funktion, darstellen. **R** stellt zu ihrer Erzeugung den sogenannten Colonoperator `:` und die Funktion `seq()` sowie ihre Varianten zur Verfügung.

Der Colonoperators in der Form `a:b` erzeugt Sequenzen ganzzahliger Schritte zwischen *a* und *b* anhand von

$$(s_i)_{i=0,1,2,\dots} \text{ mit } s_i = a + i \text{ für } i = 0, 1, 2, \dots \text{ und } s_i \leq b \quad (14.1)$$

Dabei meint “zwischen *a* und *b*” hier also speziell, dass *a* die erste Zahl der generierten Sequenz ist und *b* die maximal mögliche. Folgende Beispiele verdeutlichen dies

```
x = 0:5      # 0 ist Sequenzbeginn, 5 kann inkludiert werden
y = 1.5:6.1  # 1.5 ist Sequenzbeginn, Sequenz endet bei 5.5 < 6.1
```

x

```
[1] 0 1 2 3 4 5
```

```
y
```

```
[1] 1.5 2.5 3.5 4.5 5.5
```

Die Funktion `seq()` mit der Syntax `seq(from, to, by)`, wobei `from` den Startwert, `to` den maximalen Endwert und `by` die Schrittweite bezeichnen bietet weitere Möglichkeiten Sequenzen nicht zur ganzzahliger Schrittweiten zu erzeugen. Die folgenden Beispiele sollen dies verdeutlichen:

```
x_1 = seq(0,5)           # wie 0:5
x_2 = seq(0,1,len = 5)   # 5 Zahlen zwischen 0 (inkl.) und 1 (inkl.)
x_3 = seq(0,2,by = .15)   # 0.15 Schritte zwischen 0 (inkl.) und 2 (max.)
x_4 = seq(1,0,by = -.1)   # -0.1 Schritte zwischen 1 (inkl.) und 0 (min.)
```

```
x_1
```

```
[1] 0 1 2 3 4 5
```

```
x_2
```

```
[1] 0.00 0.25 0.50 0.75 1.00
```

```
x_3
```

```
[1] 0.00 0.15 0.30 0.45 0.60 0.75 0.90 1.05 1.20 1.35 1.50 1.65 1.80 1.95
```

```
x_4
```

```
[1] 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
```

Die `seq()` Variante `seq.int()` kann zur Erzeugung von ganzzahligen Folgen genutzt werden und ist unter Umständen schneller als `seq()`. `seq_len()` erzeugt ganzzahlige Folgen einer vorgegebenen Länge, `seq_along()` generiert ganzzahlige Folgen anhand der Einträge eines Vektors als weitere Varianten

```
x_1 = seq.int(0,5)       # wie 0:5, [0,1,2,3,4,5]
x_2 = seq_len(5)         # Natuerliche Zahlen bis 5, [1,2,3,4,5]
x_3 = seq_along(c("a","b")) # wie seq_len(length(c("a", "b")))
```

```
x_1
```

```
[1] 0 1 2 3 4 5
```

```
x_2
```

```
[1] 1 2 3 4 5
```

```
x_3
```

```
[1] 1 2
```

14.2. Charakterisierung

Zur Charakterisierung von Vektoren stehen mit `length()` und `typeof()` Funktionen bereit, die die Anzahl der Einträge eines Vektors und den Datentyp der Einträge eines Vektors ausgeben, wie folgende Beispiele verdeutlichen.

```
x = 0:10          # Vektor
length(x)         # Anzahl der Elemente des Vektors
```

```
[1] 11
```

```
x = 1:3L          # Vektor
typeof(x)         # Datentyp des atomic vectors
```

```
[1] "integer"
```

```
y = c(T,F,T)      # Vektor
typeof(y)         # Der Datentyp des atomic vectors
```

```
[1] "logical"
```

`is.logical()`, `is.double()`, `is.integer()`, `is.character()` testen den Datentyp eines Vektors und geben einen logischen Binärwert aus.

```
is.double(x)      # Testen obigen Vektors
```

```
[1] FALSE
```

```
is.logical(y)     # Testen obigen Vektors
```

```
[1] TRUE
```

Datentypangleichung (Coercion)

Versucht man in **R** verschiedene Datentypen in einem Vektor zu kontaktieren, zum Beispiel Elementarwerte vom Typ `numeric`, `character` und `logical`, so erzwingt **R** für den resultierenden Vektor einen einheitlichen Datentyp. Es gilt dabei folgende Rangfolge

$$\text{character} > \text{double} > \text{integer} > \text{logical}$$

Ist also beispielsweise einer der Elementarwerte vom Typ `character`, so ist der resultierende Vektor vom Typ `character` und numerische oder logische Werte werden in diesen umgewandelt.

```
x = c(1, "a", TRUE)      # Konkatination verschiedener Datentypen
typeof(x)                # erzwungener Datentyp
```

```
[1] "character"
```

```
x                      # character vector
```

```
[1] "1"    "a"    "TRUE"
```

Ist dagegen beispielsweise einer der Elementarwerte vom Typ `integer` und ein anderer vom Typ `logical`, so ist der resultierende Vektor vom Typ `integer` und die logischen Werte werden in ganze Zahlen umgewandelt

```
x = c(1L, TRUE, FALSE)  # Kombination gemischter Datentypen (integer schlaegt logical)
typeof(x)                # Erzeugter Vektor ist vom Typ integer
```

```
[1] "integer"
```

```
x
```

```
[1] 1 1 0
```

Da **R** diese Datentypangleichung implizit vornimmt, sollte man darauf achten nur Werte gleichen Datentyps mit der Funktion `c()` zu konkatinieren. Um möglichst gewandt zu sein, nimmt **R** auch noch subtilere Datenangleichungen vor. So ist es zum Beispiel möglich, die Summe logischer Werte zu berechnen, wobei die logischen Werten `TRUE` und `FALSE` implizit als die numerischen Werte 0 und 1 behandelt werden.

```
x = c(TRUE, FALSE, TRUE, TRUE) # logical Vektor
s = sum(x)                     # Summation in integer gewandelter logical Elemente
s
```

```
[1] 3
```

Explizite Datentypangleichung erlauben die Funktionen `as.logical()`, `as.integer()`, `as.double()`, und `as.character()`. Wendet man diese auf einen Vektor an, so ist der resultierende Vektor vom gewünschten Typ.

```
x = c(0,1,1,0)          # double Vektor
typeof(x)
```

```
[1] "double"
```

```
x
```

```
[1] 0 1 1 0
```

```
y = as.logical(x)          # ... umgewandelt in logical
typeof(y)
```

```
[1] "logical"
```

```
y
```

```
[1] FALSE  TRUE  TRUE FALSE
```

```
z = as.integer(x)
typeof(z)
```

```
[1] "integer"
```

```
z
```

```
[1] 0 1 1 0
```

14.3. Indizierung

In **R** und vielen anderen Programmiersprachen werden einzelne oder mehrere Vektoreinträge durch *Indizierung* adressiert. Indizierung wird dabei auch *indexing*, *subsetting* oder *slicing* genannt. In **R** werden zur Indizierung eckige Klammern der Form `[ ]` benutzt. **R** nutzt *one-based indexing*, d.h. der Index des ersten Elements ist 1 und nicht etwa 0 wie in anderen Programmiersprachen.

Die Indizierung von Vektoreinträgen kann genutzt werden, um einen bestimmten Eintrag eines Vektors in eine andere Datenstruktur zu kopieren:

```
x      = c("a", "b", "c")      # character vector ["a", "b", "c"]
y      = x[2]                  # Kopie von "b" in y
y      # Ausgabe
```

```
[1] "b"
```

Ebenso kann die Indizierung von Vektoreinträgen genutzt werden, um einen bestimmten Eintrag des indizierten Vektors zu manipulieren:

```
x      = c("a", "b", "c")      # character vector ["a", "b", "c"]
x[3]   = "d"                   # Änderung von x zu x = ["a", "b", "d"]
x      # Ausgabe
```

```
[1] "a" "b" "d"
```

Prinzipiell gilt in **R**, dass die Indizierung

- (1) mit einem Vektor positiver Zahlen die entsprechenden Einträge,
- (2) mit einem Vektor negativer Zahlen die entsprechenden komplementären Einträge,

- (3) mit einem logischen Vektor die Einträge mit dem logischen Wert `TRUE`, und
- (4) mit einem `character` die entsprechend benannten Einträge adressiert.

Wir wollen diese Prinzipien anhand folgender Beispiele verdeutlichen

(1) Indizierung mit einem Vektor positiver Zahlen

```
x = c(1,4,9,16,25)      # [1,4,9,16,25] = [1^2, 2^2, 3^2, 4^2, 5^2]
y = x[1:3]              # 1:3 erzeugt Vektor [1,2,3], x[1:3] = [1,4,9]
z = x[c(1,3,5)]         # c(1,3,5) erzeugt Vektor [1,3,5], x[c(1,3,5)] = [1,9,25]
x
```

```
[1] 1 4 9 16 25
```

```
y
```

```
[1] 1 4 9
```

```
z
```

```
[1] 1 9 25
```

(2) Indizierung mit einem Vektor negativer Zahlen

```
x = c(1,4,9,16,25)      # [1,4,9,16,25] = [1^2, 2^2, 3^2, 4^2, 5^2]
y = x[c(-2,-4)]         # Alle Komponenten ausser 2 und 4, x[c(-2,-4)] = [1,9,25]
x
```

```
[1] 1 4 9 16 25
```

```
y
```

```
[1] 1 9 25
```

```
z = x[c(-1,2)]          # Gemischte Indizierung nicht erlaubt (Fehlermeldung)
```

Error in `x[c(-1, 2)]`: only 0's may be mixed with negative subscripts

(3) Indizierung mit einem logischen Vektor

```
x = c(1,4,9,16,25)      # [1,4,9,16,25] = [1^2, 2^2, 3^2, 4^2, 5^2]
y = x[c(T,T,F,F,T)]     # TRUE Komponenten, x[c(T,T,F,F,T)] = [1,4,25]
z = x[x > 5]            # x > 5 = [F,F,T,T,T], x[x > 5] = [9,16,25]
x
```

```
[1] 1 4 9 16 25
```

```
y
```

```
[1] 1 4 25
```

```
z
```

```
[1]  9 16 25
```

(4) Indizierung mit einem character Vektor

```
x = c(1,4,9,16,25)      # [1,4,9,16,25] = [1^2, 2^2, 3^2, 4^2, 5^2]
names(x) = c("a","b")   # Benennung der Komponenten als [a b <NA> <NA> <NA>]
y = x["a"]              # x["a"] = 1
```

Es bleibt anzumerken, dass **R** hat eine hohe Flexibilität bei Indizierung aufweist, d.h. oft keine Fehler ausgibt, obwohl eine Indizierung per se keinen Sinn ergibt. So verursachen z.B. out-of-range Indizes, also Indizes die die Anzahl der Einträge eines Vektors überschreiten keine Fehler, sondern geben **NA** (not applicable) aus

```
x = c(1,4,9,16,25)      # [1,4,9,16,25] = [1^2, 2^2, 3^2, 4^2, 5^2]
y = x[10]               # x[10] = NA (Not Applicable)
y
```

```
[1] NA
```

Weiterhin verursachen nichtganzzahlige Indizes keine Fehler, sondern werden auf die nächste ganze Zahl abgerundet.

```
y = x[4.9]              # x[4.9] = x[4] = 16
y
```

```
[1] 16
```

```
z = x[-4.9]              # x[-4.9] = x[-4] = [1,4,9,25]
z
```

```
[1]  1  4  9 25
```

Leere Indizes schließlich indizieren den gesamten Vektor

```
y = x[]                  # y = x
```

Die Flexibilität von **R** bei nicht sinnhafter Indizierung hat einerseits den Vorteil, dass Programme seltener mit einem Fehler unterbrochen werden. Auf der anderen Seite hat die Indizierungsflexibilität den Nachteil, dass unbeabsichtigte Fehlindizierungen unbemerkt bleiben und Datenanalyseprogramme so zu nicht intendierten Ergebnissen kommen können. Im Umgang mit der Indizierung in **R** ist also ein hohes Maß an Aufmerksamkeit und im Idealfall die Validierung von erstellten Datenanalyseprogramme anhand simulierter Daten mit bekannten Ergebnissen geboten. Ein ähnliches Maß an Vorsicht ist aufgrund von **R**'s *Recycling* Funktionsweise geboten, die wir im nächsten Abschnitt kennenlernen werden.

14.4. Arithmetik numerischer Vektoren

Mit **R** Vektoren kann man arithmetische Operationen ausführen, d.h. mit ihnen rechnen. Wir unterscheiden dabei *unitäre arithmetische Operationen*, d.h. arithmetische Operationen, die auf einen Vektor angewendet werden und *binäre arithmetische Operationen*, bei denen zwei Vektoren miteinander verknüpft werden. Allgemein gilt, dass arithmetische Vektoroperatoren in **R** elementweise ausgewertet werden, allerdings sind bei binären Operatoren einige Besonderheiten zu beachten.

Betrachten wir zunächst einige unitäre arithmetische Operatoren und die Anwendung von Funktionen auf einen Vektor. In folgendem Beispiel werden die unitären Operatoren `-` und `^2`, sowie die Funktion `log()` auf einen Vektor `a` angewendet. Die Auswertung der Operatoren bzw. der Funktion geschieht dabei elementweise, d.h. der Operator oder die Funktion werden auf jedes Element des Vektors unabhängig von seinen anderen Elementen angewendet.

```
a = seq(0,1,len=11) # a = [ 0.0 , 0.1 , ..., 0.9 , 1.0 ]
a
```

```
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

```
b = -a # b = [-0.0 , -0.1 , ..., -0.9 , -1.0 ]
b
```

```
[1] 0.0 -0.1 -0.2 -0.3 -0.4 -0.5 -0.6 -0.7 -0.8 -0.9 -1.0
```

```
v = a^2 # v = [ 0.0^2 , 0.1^2 , ..., 0.9^2 , 1.0^2 ]
v
```

```
[1] 0.00 0.01 0.04 0.09 0.16 0.25 0.36 0.49 0.64 0.81 1.00
```

```
w = log(a) # w = [ln(0.0), ln(0.1), ..., ln(0.9), ln(1.0)]
w
```

```
[1] -Inf -2.3025851 -1.6094379 -1.2039728 -0.9162907 -0.6931472
[7] -0.5108256 -0.3566749 -0.2231436 -0.1053605 0.0000000
```

Binäre arithmetische Operatoren, also die Verknüpfung von zwei Vektoren, geschieht ebenfalls elementweise. Bei der Verknüpfung zweier Vektoren gleicher Länge werden dabei die Vektoreinträge mit den gleichen Indizes anhand des binären arithmetischen Operators verknüpft, als Resultat ergibt sich ein Vektor mit der gleichen Länge wie die beiden Ausgangsvektoren

```
a = c(1,2,3) # a = [1,2,3]
b = c(2,1,4) # b = [2,1,4]
v = a + b # v = [1,2,3] + [2,1,4] = [1+2,2+1,3+4] = [ 3, 3, 7]
v
```

```
[1] 3 3 7
```

```
w = a - b          # w = [1,2,3] - [2,1,4] = [1-2,2-1,3-4] = [ -1,  1,  -1]
w
```

```
[1] -1  1 -1
```

```
x = a * b          # x = [1,2,3] * [2,1,4] = [1*2,2*1,3*4] = [  2,  2,  12]
x
```

```
[1]  2  2 12
```

```
y = a / b          # y = [1,2,3] / [2,1,4] = [1/2,2/1,3/4] = [0.50,  2,  0.75]
y
```

```
[1] 0.50 2.00 0.75
```

Verküpft man in **R** einen Vektor und einen Skalar, also einen Vektor der Länge 1, so erweitert **R** den Skalar auf einen Vektor der gleichen Länge und führt die elementweise Verknüpfung dieser Vektoren aus, wie folgende Beispiele verdeutlichen.

```
a = c(1,2,3)      # a = [1,2,3]
b = 2              # b = [2]
v = a + b          # v = [1,2,3] + [2,2,2] = [1+2,2+2,3+2] = [  3,  4,  5]
v
```

```
[1] 3 4 5
```

```
w = a - b          # w = [1,2,3] - [2,2,2] = [1-2,2-2,3-2] = [ -1,  2,  1]
w
```

```
[1] -1  0  1
```

```
x = a * b          # x = [1,2,3] * [2,2,2] = [1*2,2*2,3*2] = [  2,  4,  6]
x
```

```
[1] 2 4 6
```

```
y = a / b          # y = [1,2,3] / [2,2,2] = [1/2,2/2,3/2] = [0.5,  1,  1.5]
y
```

```
[1] 0.5 1.0 1.5
```

Unter dem Begriff des *Recycling* erlaubt **R** neben der elementweisen Arithmetik für Vektoren gleicher Länge bzw. für Skalare und Vektoren außerdem auch eine Arithmetik mit Vektoren unterschiedlicher Länge. Dazu werden Elemente des kürzeren von zwei Vektoren zum Aufstocken auf die Länge des längeren Vektors genutzt, also in diesem Sinn *recycled*. Wiederum führt diese Eigenschaft der Vektorarithmetik in **R** dazu, dass Programme seltener mit einem Fehler abbrechen und dass das Recycling unter Umständen in informiert geschriebenem Code elegante Lösungen für Berechnungen anbieten kann. Allerdings hat es den klaren Nachteil, dass auch unbeabsichtigte Fehlallokationen von Vektoren, z.B. durch

das Nichtvorhandensein von Datenpunkten unentdeckt bleiben können. Auch hier ist also bei der Erstellung von Programmen hohe Aufmerksamkeit und insbesondere ausführlichen Testen angezeigt. Speziell kann man bei Recycling der Vektorarithmetik in **R** zwei Fälle unterscheiden: Ist die Länge des längeren der beiden Vektoren ein ganzzahliges Vielfaches der Länge des kürzeren Vektors, so wird der kürzere Vektor durch die entsprechend häufige Wiederholung seiner Elemente in ihrer originären Reihenfolge expandiert. Folgendes Beispiel soll dies verdeutlichen:

```
x = 1:2          # x = [1,2], length(x) = 2
x
```

```
[1] 1 2
```

```
y = 3:6          # y = [3,4,5,6], length(y) = 4
y
```

```
[1] 3 4 5 6
```

```
v = x + y        # v = [1,2,1,2] + [3,4,5,6] = [4,6,6,8]
v
```

```
[1] 4 6 6 8
```

In diesem Sinn ist dann die Arithmetik von Vektoren und Skalaren, also Vektoren der Länge 1, in Spezialfall dieses Prinzips. Ist die Länge des längeren der beiden Vektoren kein ganzzahliges Vielfaches der Länge des kürzeren Vektors, so werden Elemente des kürzeren Vektor in ihrer originären Reihenfolge so lange zum Auffüllen des kürzeren Vektors genutzt, bis beide Vektoren die gleiche Länge haben. Folgendes Beispiel soll dies verdeutlichen:

```
x = c(1,3,5)      # x = [1,3,5], length(x) = 3
x
y = c(2,4,6,8,10)  # y = [2,4,6,8,10], length(y) = 5
y
v = x + y          # v = [1,3,5,1,3] + [2,4,6,8,10] = [3,7,11,9,13]
v
```

14.5. Attribute

Attribute in **R** sind Metadaten von Objekten in Form von Schlüssel-Wert-Paaren. Generell ruft die Funktion `attributes()` alle Attribute eines Objektes auf. Vektoren haben per se keine Attribute

```
a = 1:3          # ein numerischer Vektor
attributes(a)     # Aufrufen aller Attribute von
```

```
NULL
```

Die Funktion `attr()` kann zum Aufrufen und Definieren von Attributen genutzt werden. Folgender Code definiert für den Vektor `a` als Attribut mit dem Schlüssel (Namen) `"S"` und dem Werte `"W"`.

```
attr(a, "S") = "W" # a bekommt Attribut mit Schluessel S und Wert W
attr(a, "S")      # Das Attribut mit Schluessel S hat den Wert W
```

```
[1] "W"
```

Allerdings werden Attribute werden bei Operationen oft entfernt.

```
b = a[1]          # Kopie des ersten Elements von a in Vektor b
attributes(b)     # Aufrufen aller Attribute von b
```

```
NULL
```

Die Spezifikation des Attributs **names** gibt den Elementen eines Vektors Namen (Bezeichnungen), die zur Indizierung genutzt werden können und die bei Operationen nicht entfernt werden.

```
v = c(x=1,y=2,z=3) # Elementnamengeneration bei Vektorerzeugung
v                  # Vektorausgabe
```

```
x y z
1 2 3
```

Die Indizierung mithilfe von Namen hat folgende Form

```
v["y"]           # Indizierung per Namen
```

```
y
2
```

Im Nachgang der Definition eines Vektors kann die Funktion **names()** genutzt werden, um für die Vektoreinträge Namen zu Definieren und diese nachzuschlagen.

```
y = 4:6           # Erzeugung eines Vektors
names(y) = c("a","b","c") # Definition von Namen
names(y)          # Elementnamenaufruf
```

```
[1] "a" "b" "c"
```

Benannte Vektoreinträge können hilfreich sein, wenn der Vektor eine Sinneinheit bildet, z.B. die Eigenschaften einer Versuchsperson wie Alter, Größe und Gewicht repräsentiert

```
p = c(age      = 31,      # Alter (Jahre), Groesse (cm), Gewicht (kg) einer Person
      height = 198,
      weight  = 75)
p                  # Vektorausgabe
```

Allerdings ist die Arbeit mit benannten Vektoreinträgen eher selten. Das Prinzip, Einträge einer Datenstruktur anstelle eines numerischen Indizes mit einer sinntragenden Bezeichnung zu versehen ist in **R** allerdings eines der Leitmotive, wie wir im Umgang mit der zentralen Datenstruktur der Dataframes in Kapitel 16 sehen werden.

15. Matrizen und Arrays

R Matrizen und Arrays können als höherdimensionale Generalisierungen von **R** Vektoren verstanden werden. Wie bei den Vektoren ist es wichtig, Matrizen in der Programmierung von ihren gleichnamigen mathematischen Objekten abzugrenzen. So können Matrizen in der Programmierung einfach der tabellarischen Speicherung numerischer Daten dienen, ohne dass dabei das primäre Ziel die Anwendung der mathematischen Matrizenrechnung sein muss. Andersherum bilden die elementweisen arithmetischen Operationen mit Matrizen in der Programmierung in der Mathematik eher sehr spezielle mathematische Operationen, wie das Hadamard-Produkt ab. In diesem Abschnitt wollen wir Matrizen und Arrays primär als Datencontainer in der imperativen Programmierung mit **R** untersuchen. Eine Einführung in die Matrizenrechnung mit den gleichen Datenstrukturen gibt Kapitel [15.1](#).

15.1. Matrizen

Abstrakt gesprochen sind **R** Matrizen zweidimensionale, rechteckige Datenstrukturen der Form

$$M = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1n_c} \\ m_{21} & m_{22} & \cdots & m_{2n_c} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n_r1} & m_{n_r2} & \cdots & m_{n_rn_c} \end{pmatrix} \quad (15.1)$$

Die Elemente m_{ij} , $i = 1, \dots, n_r$, $j = 1, \dots, n_c$ der Matrix sind dabei notwendigerweise vom gleichen Datentyp. In obigem Schema bezeichnet n_r die Anzahl der Zeilen (rows) und n_c die Anzahl der Spalten (columns) der Matrix M . Jedes Element einer Matrix hat damit einen Zeilenindex i und einen Spaltenindex j . Intuitiv sind **R** Matrizen also numerisch indizierte Tabellen, formal sind Matrizen in **R** zweidimensional interpretierte atomare Vektoren.

15.1.1. Erzeugung

Zur Erzeugung einer Matrix wird die **R** Funktion `matrix()` genutzt, die Matrizen mit den Elementen eines Vektors befüllt. Die allgemeine Syntax dieser Funktion ist `M = matrix(v, nrow, ncol, byrow)`, wobei

- `v` einen Vektor der Länge `nrow · ncol`,
- `nrow` die gewünschte Zeilenanzahl von `M`,
- `ncol` die gewünschte Spaltenanzahl von `M`,
- die logische Variable `byrow` die Befüllungsordnung von `M`

bezeichnen. Da sich bei gegebener Länge von `v` und gegebener Zeilen- oder Spaltenanzahl die entsprechende Spalten- oder Zeilenanzahl der Matrix ergibt, ist es lediglich nötig entweder `nrow` oder `ncol` zu spezifizieren. Folgende Beispiele verdeutlichen dies

```
matrix(c(1:12), nrow = 3)           # 3 x 4 Matrix der Zahlen 1,...,12, byrow = F
```

```
      [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

```
matrix(c(1:12), ncol = 4)           # 3 x 4 Matrix der Zahlen 1,...,12, byrow = F
```

```
      [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

```
matrix(c(1:12), nrow = 3, byrow = T) # 3 x 4 Matrix der Zahlen 1,...,12, byrow = T
```

```
      [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
```

Bestehende Matrizen gleicher Zeilenanzahl können mit der Funktion `cbind()` spaltenweise konkateniert werden:

```
A = matrix(c(1:4), nrow = 2)        # 2 x 2 Matrix der Zahlen 1,...,4
A                                     # Ausgabe
```

```
      [,1] [,2]
[1,]    1    3
[2,]    2    4
```

```
B = matrix(c(5:10), nrow = 2)       # 2 x 3 Matrix der Zahlen 5,...,10
B                                     # Ausgabe
```

```
      [,1] [,2] [,3]
[1,]    5    7    9
[2,]    6    8   10
```

```
C = cbind(A,B)                       # spaltenweise Konkatenierung von A und B
C                                     # Ausgabe
```

```
      [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    5    7    9
[2,]    2    4    6    8   10
```

Äquivalent können bestehende Matrizen gleicher Spaltenanzahl mit der Funktion `rbind()` zeilenweise konkateniert werden

```
A = matrix(c(1:6) , nrow = 2, byrow = T)      # 2 x 3 Matrix der Zahlen 1,...,6
print(A)
```

```
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

```
B = matrix(c(7:9), nrow = 1)                  # 1 x 3 Matrix der Zahlen 5,...,10
print(B)
```

```
      [,1] [,2] [,3]
[1,]    7    8    9
```

```
C = rbind(A,B)                                # reihenweise Konkatenierung von A und B
print(C)
```

```
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

15.1.2. Charakterisierung

Mit der vertrauten Funktion `typeof()` kann der elementare Datentyp einer Matrix ausgegeben werden:

```
typeof(matrix(c(1,0,0,1) , nrow = 2))      # 2 x 2 double Matrix
```

```
[1] "double"
```

```
typeof(matrix(c(1L,0L,0L,1L), nrow = 2))   # 2 x 2 integer Matrix
```

```
[1] "integer"
```

```
typeof(matrix(c(T,T,F,F) , nrow = 2))      # 2 x 2 logical Matrix
```

```
[1] "logical"
```

```
typeof(matrix(c("a","b","c"), nrow = 1))   # 1 x 3 character Matrix
```

```
[1] "character"
```

Die Zeilen- bzw. Spaltenanzahl einer Matrix können mit den Funktionen `nrow()` (number of rows) und `ncol()` (number of columns) ausgegeben werden.

```
C = matrix(1:12, nrow = 3)                  # 3 x 4 Matrix
nrow(C)                                     # Anzahl Zeilen
```

```
[1] 3
```

```
ncol(C) # Anzahl Spalten
```

```
[1] 4
```

15.1.3. Indizierung

Generell gilt, dass Matrixelemente mit einem Zeilenindex und einem Spaltenindex indiziert werden. Der erste zweier Matrixindices indiziert dabei immer die Zeile, der zweite immer die Spalte. Die übrigen Prinzipien der Indizierung von Matrizen entsprechen im Wesentlichen den Prinzipien der Indizierung von Vektoren. Eine Besonderheit ist, dass ein fehlender Zeilen- oder Spaltenindex alle Elemente der entsprechenden Dimension indiziert.

```
A = matrix(c(2:7)^2, nrow = 2) # 2 x 3 Matrix der Zahlen 2^2,...,7^2
A # Ausgabe
```

```
      [,1] [,2] [,3]
[1,]    4   16   36
[2,]    9   25   49
```

```
A[1,3] # Element in 1. Zeile, 3. Spalte von A [36]
```

```
[1] 36
```

```
A[2,2] # Element in 2. Zeile, 2. Spalte von A [35]
```

```
[1] 25
```

```
A[2,] # Alle Elemente der 2. Zeile [9,25,49]
```

```
[1] 9 25 49
```

```
A[,3] # Alle Elemente der 3. Spalte [36,49]
```

```
[1] 36 49
```

```
A[1:2,1:2] # Submatrix der ersten zwei Zeilen und Spalten
```

```
      [,1] [,2]
[1,]    4   16
[2,]    9   25
```

```
A[A>10] # Elemente von A groesser 10 [16,25,36,49]
```

```
[1] 16 25 36 49
```



```
A[1,c(F,F,T)] # Element in 1. Zeile, 3. Spalte von A [36]
```

```
[1] 36
```

15.1.4. Arithmetik

Wendet man unitäre arithmetische Operatoren und Funktionen auf Matrizen an, so werden diese wie bei den Vektoren elementweise ausgewertet:

```
A = matrix(c(1:4), nrow = 2) # 2 x 2 Matrix der Zahlen 1,2,3,4
      [,1] [,2]
[1,]    1    3
[2,]    2    4

B = A^2 # B[i,j] = A[i,j]^2, 1 <= i,j <= 2
      [,1] [,2]
[1,]    1    9 # 1^2, 3^2
[2,]    4   16 # 2^2, 4^2

C = sqrt(B) # C[i,j] = sqrt(A[i,j]^2), 1 <= i,j <= 2
      [,1] [,2]
[1,]    1    3 # sqrt(1^2), sqrt(3^2)
[2,]    2    4 # sqrt(2^2), sqrt(4^2)

D = exp(A) # D[i,j] = exp(A[i,j]), 1 <= i,j <= 2
      [,1] [,2]
[1,]  2.7 20.0 # exp(1), exp(3)
[2,]  7.4 54.6 # exp(2), exp(4)
```

Wie Vektoren können auch Matrizen mit identischen Zeilen- und Spaltenanzahlen mit den binären arithmetischen Operatoren $+$, $-$, $*$ und $/$ elementweise verknüpft werden. Folgende Beispiele verdeutlichen dies:

```
A = matrix(c(1:4), nrow = 2) # 2 x 2 Matrix der Zahlen 1,2,3,4
      [,1] [,2]
[1,]    1    3
[2,]    2    4

B = matrix(c(5:8), nrow = 2) # 2 x 2 Matrix der Zahlen 5,6,7,8
      [,1] [,2]
[1,]    5    7
[2,]    6    8

C = A + B # C[i,j] = A[i,j] + B[i,j], 1 <= i,j <= 2
      [,1] [,2]
[1,]    6   10 # 1 + 5, 3 + 7
[2,]    8   12 # 2 + 6, 4 + 8

D = A * B
      [,1] [,2]
[1,]    5   21 # 1 * 5, 3 * 7
[2,]   12   32 # 2 * 6, 4 * 8
```

Darüber hinaus kann mit **R** Matrizen im Sinne d der Matrizenrechnung der linearen Algebra gerechnet werden. Wir zeigen hier nur einige wenige Beispiele, eine ausführliche

Darstellung der mathematischen Prinzipien und ihrer **R** Implementation findet sich in Kapitel 15.1.

```
C = A % * % B          # 2 x 2 Matrixprodukt
      [,1] [,2]
[1,]    23    31      # 1*5 + 3*6, 1*7+3*8
[2,]    34    46      # 2*5 + 4*6, 2*7+4*8

A_T = t(A)             # Transposition von A
      [,1] [,2]
[1,]     1     2      # A[1,1], A[2,1]
[2,]     3     4      # A[1,2], A[2,2]

A_inv = solve(A)        # Inverse von A
      [,1] [,2]
[1,]    -2  1.5
[2,]     1 -0.5

A_det = det(A)          # Determinante von A
[1] -2                  # 1*4 - 2*3
```

15.1.5. Attribute

Formal sind Matrizen atomare Vektoren mit einem `dim` Attribut

```
A = matrix(1:12, nrow = 4 )      # 4 x 3 Matrix
attributes(A)                    # Aufrufen der Attribute von A
```

```
$dim
[1] 4 3
```

Wie bei den Vektoren kann man auch den Matrizenelementen Namen geben, allerdings ist dies auch bei Matrizen eher ungewöhnlich. Speziell definieren die Funktionen `rownames()` und `colnames()` das Matrixattribut `dimnames`.

```
rownames(A) = c("P1","P2","P3","P4") # Benennung der Zeilen von A
colnames(A) = c("Age", "Hgt", "Wgt")  # Benennung der Spalten von A
A                                         # A mit Attribut dimnames
```

```
      Age Hgt Wgt
P1     1   5   9
P2     2   6  10
P3     3   7  11
P4     4   8  12
```

```
attr(A, "dimnames")                  # Aufrufen des Attributs dimnames
```

```
[[1]]
[1] "P1" "P2" "P3" "P4"

[[2]]
[1] "Age" "Hgt" "Wgt"
```

15.2. Arrays

R Arrays sind d -dimensionale, hyperrechteckige Datenstrukturen. Intuitiv können Arrays als die Generalisierung von Matrizen auf mehr als zwei Dimensionen verstanden werden. Für $d = 1$ entspricht ein Array einem Vektor, für $d = 2$ entspricht ein Array einer Matrix. Wie bei Vektoren und Matrizen müssen die Elemente eines Arrays vom gleichen Datentyp sein. Arrays sind dadurch charakterisiert, wieviele Elemente $d_i, i = 1, \dots, d$ in der i ten Dimension repräsentieren können, jedes Element eines Arrays wird dabei durch d -dimensionalen Index $i_1, i_2, \dots, i_d$ indiziert.

15.2.1. Erzeugung

In Analogie zur Funktion `matrix()` wird die Funktion `array()` genutzt, um einen Array mit Vektorelementen zu befüllen. Die allgemeine Syntax dieser Funktion ist `A = array(v, dim)`, wobei `v` ein Vektor von Arraydaten und `dim` ein Vektor ganzzahliger Werte ist, der die maximalen Indizes in jeder der `length(dim)` Dimensionen des Arrays repräsentiert. So erzeugt beispielsweise

```
A = array(1:12, dim = c(2,2,3)) # 2 x 2 x 3 Array der Zahlen 1,...,12
A
```

```
, , 1
      [,1] [,2]
[1,]    1    3
[2,]    2    4

, , 2
      [,1] [,2]
[1,]    5    7
[2,]    6    8

, , 3
      [,1] [,2]
[1,]    9   11
[2,]   10   12
```

einen $2 \times 2 \times 3$ Array. Dreidimensionale Arrays kann man sich intuitiv als im Raum hintereinander aufgestellte Matrizen vorstellen, die dritte Dimension repräsentiert dann dabei die “Seiten” des Arrays. Allerdings ist es im Umgang mit höherdimensionalen Arrays sinnvoll, zu versuchen, sich von räumlichen Vorstellungen von Arrays zu lösen und Arrays stattdessen als Datencontainer zu verstehen, deren Elemente eineindeutig durch ihre Indizes adressiert werden können.

15.2.2. Charakterisierung

Die `length()` Funktion gibt die Anzahl der Elemente eines Arrays aus:

```
A = array(1:12, dim = c(2,2,3))    # 2 x 2 x 3 Array der Zahlen 1,...,12
length(A)                        # Die Anzahl der Elemente des Arrays
```

```
[1] 12
```

Die Funktion `typeof()` gibt den elementaren Datentyp eines Arrays aus:

```
typeof(A)                        # Der Typ des Arrays
```

```
[1] "integer"
```

Die Funktion `dim()` gibt die Dimensionen eines Arrays aus:

```
dim(A)                          # Dimension des Arrays
```

```
[1] 2 2 3
```

15.2.3. Indizierung

Die Indizierung von Arrays erfolgt analog zu Vektor- und Matrixindizierung, mit dem Unterschied, dass zur Indizierung eines Arrayelements drei oder mehr Indices benötigt werden.

```
A      = array(1:12, dim = c(2,2,3))    # 2 x 2 x 3 Array der Zahlen 1,...,12
a_223  = A[2,2,3]                      # Arrayelement mit Indexadresse [2,2,3] (12)
```

15.2.4. Arithmetik

Wie bei Vektoren und Matrizen werden unitäre arithmetische Operatoren bei Arrays elementweise ausgewertet. Ebenso analog werden binäre arithmetische Operatoren bei Arrays identischer Dimensionalität elementweise ausgewertet

```
A      = array(1:4, dim = c(1,2,2))    # 1 x 2 x 2 Array der Zahlen 1,...,4
A
B      = array(5:8, dim = c(1,2,2))    # 1 x 2 x 2 Array der Zahlen 5,...,8
B
C      = A^2                            # elementweise Potenzierung
C
D      = A + B                          # elementweise Addition
D
```

15.2.5. Attribute

Formal sind Arrays atomic vectors mit einem `dim` Attribut

```
A = array(1:12, dim = c(2,2,3))      # 2 x 2 x 3 Array der Zahlen 1,...,12
attributes(A)                        # Aufrufen der Attribute von A
```

```
$dim
[1] 2 2 3
```

Wie bei den Matrizen kann die Funktion `dimnames()` zur Benennung der Arraydimensionen benutzt werden, aber auch hier die Benennung von Arraydimensionen eher ungewöhnlich.

16. Listen, Dataframes und Tibbles

16.1. Listen

Listen sind geordnete Folgen von R Objekten. Listen werden als *rekursiver* Datentyp bezeichnet, da sie Objekte verschiedener Datentypen, insbesondere auch wiederum Listen bündeln können. Obwohl sich die intuitive Sichtweise, dass Listen Objekte “enthalten” anbietet, “enthalten” Listen defacto keine Objekte, sondern lediglich die Referenzen zu den entsprechenden Objekten im Arbeitsspeicher (Abbildung 16.1). Listen sind ein wesentlicher Baustein der zentralen **dataframe** Datenstruktur in **R**.

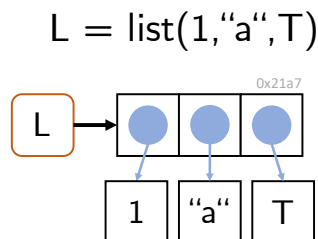


Abbildung 16.1. Listenrepräsentation. Eine Liste in **R** ist eine indizierbare Folge von Referenzen zu **R** Objekten im Arbeitsspeicher.

Erzeugung

Listen können durch die direkte Konkatenation der Listenelemente mithilfe der Funktion `list()` erzeugt werden.

```
# Erzeugung einer Liste mit einem Vektor-, Matrix-, und Funktionselement
L = list(c(1,4,5), matrix(1:8, nrow = 2), exp)
print(L)
```

```
[[1]]
[1] 1 4 5

[[2]]
  [,1] [,2] [,3] [,4]
[1,]  1   3   5   7
[2,]  2   4   6   8

[[3]]
function (x) .Primitive("exp")
```

Insbesondere können auch Listen selbst Elemente von Listen sein, was ihre Bezeichnung als *rekursiver* Datentyp motiviert.

```
L = list(list(1))          # Liste mit Element 1 in einer Liste
print(L)                  # Ausgabe
```

Die von den **R** Vektoren vertraute Funktion `c()` kann auch zum Konkatenieren von mehreren Listen genutzt werden.

```
L = c(list(pi), list("a")) # Konkatenation zweier Listen
print(L)                  # Ausgabe
```

Charakterisierung

Der Datentyp von Listen ist `list`:

```
L = list(1:2, "a", log)    # Erzeugung einer Liste
typeof(L)                  # Typenbestimmung
```

```
[1] "list"
```

`length()` gibt die Anzahl der Listenelemente auf dem obersten Listenlevel aus, die Elementinhalte werden also ignoriert.

```
L = list(1:2, list("a", pi), exp) # Liste mit drei Toplevelelementen
length(L)                         # length() ignoriert Elementinhalte, length() von L ist also 3
```

```
[1] 3
```

Die Dimension, Zeilen- und Spaltenanzahl von Listen ist `NULL`:

```
L = list(1:2, "a", sin)     # eine Liste
dim(L)                      # Die Dimension von Listen ist NULL
```

```
NULL
```

```
nrow(L)                     # Die Zeilenanzahl von Listen ist NULL
```

```
NULL
```

```
ncol(L)                     # Die Spaltenanzahl von Listen ist NULL
```

```
NULL
```

Indizierung

Bei der Indizierung von Listen ist zu beachten, dass Listen auf zwei unterschiedliche Arten indiziert werden können. Einfache eckige Klammern `[ ]` indizieren Listenelemente als Listen:

```
L = list(1:3, "a", exp)      # eine Liste
l1 = L[1]                   # Indizierung eines Listenelements
print(l1)                   # Ausgabe
```

```
[[1]]
[1] 1 2 3
```

```
typeof(l1)                  # Typbestimmung von l1
```

```
[1] "list"
```

Doppelte eckige Klammern `[[ ]]` indizieren den Inhalt von Listenelementen:

```
L = list(1:3, "a", exp)      # eine Liste
i2 = L[[2]]                  # Indizierung des Listenelementinhalts
print(i2)                    # Ausgabe des Inhalts von Listenelement L[2]
typeof(i2)                   # Typbestimmung von i2
```

Auch beim Ersetzen von Listeninhalten sind diese Konventionen zu beachten. So ist es zum Beispiel nicht möglich durch den Ausdruck `L[1] = c(1,2,3)` das erste Listenelement zu ändern, da der Vektor `c(1,2,3)` keine Liste ist:

```
L      = list(1:3, "a", exp)  # eine Liste
L[1]   = 4:6                  # keine Typkonversion, Fehlermeldung
```

```
Warning in L[1] = 4:6: number of items to replace is not a multiple of
replacement length
```

```
L[1]    = list(4:6)           # Ersetzung des 1. Listenelementes
L[[3]]  = "c"                 # Ersetzung des 3. Listenelementinhaltes
```

Indizierung

Die Prinzipien der Listenindizierung sind analog zu denen der Vektorindizierung. Vektoren positiver Zahlen adressieren die entsprechenden Elemente:

```
L = list(1:3, "a", pi)      # eine Liste
l = L[c(1,3)]               # 1. und 3. Listenelement
l
```

```
[[1]]
[1] 1 2 3
```

```
[[2]]
[1] 3.141593
```

Vektoren negativer Zahlen adressieren die zu ihren Werten komplementären Elemente: __


```
L = list(1:3, "a", pi)          # eine Liste
l = L[-c(1,3)]                 # 2. Listenelement
l
```

```
[[1]]
[1] "a"
```

Logische Vektoren adressieren Elemente mit Index **TRUE**:

```
L = list(1:3, "a", pi)          # eine Liste
l = L[c(T,T,F)]                 # 1. und 2. Listenelement
l
```

```
[[1]]
[1] 1 2 3
```

```
[[2]]
[1] "a"
```

Zur nicht-numerischen Addressierung können auch Listenelementen Namen gegeben werden. Dies ist nach der Erzeugung einer Liste mithilfe der **names()** Funktion möglich:

```
K      = list(1:2, TRUE)          # eine unbenannte Liste
names(K) = c("Frodo", "Sam")      # Namensgebung mit names()
K
```

```
$Frodo
[1] 1 2
```

```
$Sam
[1] TRUE
```

Ebenso ist es möglich, die Namen der Listenelemente direkt bei Erzeugung der Liste durch Zuweisung zu vergeben:

```
L = list(greta = 1:3, luisa = "a", carla = "b") # Listenerzeugung mit Namenvergabe
```

Sowohl Listenelemente als auch Listenelementinhalte können dann mit ihren Namen adressiert werden:

```
L = list(greta = 1:3, luisa = "a", carla = "b") # Listenerzeugung mit Namenvergabe
L["greta"]                                     # Listenelementausgabe
```

```
$greta
[1] 1 2 3
```

```
typeof(L["greta"])                          # Listenelementtyp
```

```
[1] "list"
```

```
L[["greta"]] # Listenelementinhaltsausgabe
```

```
[1] 1 2 3
```

```
typeof(L[["luisa"]]) # Listenelementinhaltstyp
```

```
[1] "character"
```

Insbesondere können Listenelementinhalte auch mit dem `$` Operator und ihrem Namen indiziert werden. Dies stellt im Rahmen der Dataframes die gebräuchlichste Art der Listenindizierung dar:

```
L = list(greta = 1:3, luisa = "a", carla = "b") # Listenerzeugung mit Namenvergabe  
L$greta # Listenelementinhaltsausgabe
```

```
[1] 1 2 3
```

```
typeof(L$greta) # Listenelementinhaltstyp
```

```
[1] "integer"
```

```
L$luisa # Listenelementinhaltsausgabe
```

```
[1] "a"
```

```
typeof(L$luisa) # Listenelementinhaltstyp
```

```
[1] "character"
```

```
L$carla # Listenelementinhaltsausgabe
```

```
[1] "b"
```

```
typeof(L$carla) # Listenelementinhaltstyp
```

```
[1] "character"
```

Arithmetik

Eine allgemeine Listenarithmetik ist nicht definiert, da Listenelementinhalte ja unterschiedlichen Datentyps sein können, für die unitäre oder binäre Operationen nicht definiert sind:

```
L1 = list(1:3, "a" )      # eine Liste
L2 = list(T, exp )       # eine Liste
L1+L2                    # Versuch der Listenaddition
```

Error in L1 + L2: non-numeric argument to binary operator

Sind Inhalte von Listenelementen allerdings von einem gemeinsamen Datentyp, für den bestimmte binäre Operationen definiert sind, so können sie auch z.B. arithmetisch verknüpft werden:

```
L1 = list(1:3, pi )      # eine Liste
L2 = list(4:6, exp )     # eine Liste
L1[[1]] + L2[[1]]        # Addition der 1. Listenelementinhalte, [1+4, 2+5, 3+6]
```

```
[1] 5 7 9
```

```
L2[[2]](1)              # Anwendung des 2. Listenelementinhalts, exp(1)
```

```
[1] 2.718282
```

Copy-on-modify

Wie bei Vektoren gilt bei Listen das *Copy-on-Modify* Prinzip. Insbesondere gilt bei Listen, dass sogenannte *Shallow copies* bei Modifikation einer Liste erzeugt werden. Dabei wird nur das Listenobjekt selbst kopiert und erhält eine neue Adresse im Arbeitsspeicher, nicht aber die durch die Liste nicht-modifizierend gebundenen Objekte, die an den gleichen Adressen im Arbeitsspeicher verbleiben. `lobstr::ref()` erlaubt es, dieses Verhalten zu studieren.

```
L1 = list(1,2,3)         # Erzeugen einer Liste als Objekt
lobstr::ref(L1)           # Ausgabe der Referenzen
```

```
[1:0x1996ab07ae8] <list>
[2:0x19967cb40e8] <dbl>
[3:0x19967cb40b0] <dbl>
[4:0x19967cb4078] <dbl>
```

```
L2 = L1                  # L1 und L2 referenzieren dasselbe Objekt
lobstr::ref(L2)           # Ausgabe der Referenzen
```

```
[1:0x1996ab07ae8] <list>
[2:0x19967cb40e8] <dbl>
[3:0x19967cb40b0] <dbl>
[4:0x19967cb4078] <dbl>
```

```
L2[[3]] = 4          # Copy-on-Modify mit shallow Objekt Kopie
lobstr::ref(L2)      # Ausgabe der Referenzen
```

```
[1:0x1996ac74688] <list>
[2:0x19967cb40e8] <dbl>
[3:0x19967cb40b0] <dbl>
[4:0x19967d1be00] <dbl>
```

Offenbar ändert die Zuweisung `L2 = L1` zunächst nichts an den Arbeitsspeicheradressen auf die L1 und L2 sowie ihre Elemente verweisen. Die Modifikation des Inhalts des dritten Listenelements von L2 durch `L2[[3]] = 4` allerdings ändert dann die Adresse des Listenobjekts im Arbeitsspeicher, da eine Kopie des Listenobjekts erzeugt wird. Gleichzeitig ändern sich gemäß dem shallow copy-Prinzip die Adressen der ersten beiden Inhalte von L2 aber nicht. Abbildung 16.2 visualisiert diese Prinzipien an einem Beispiel.

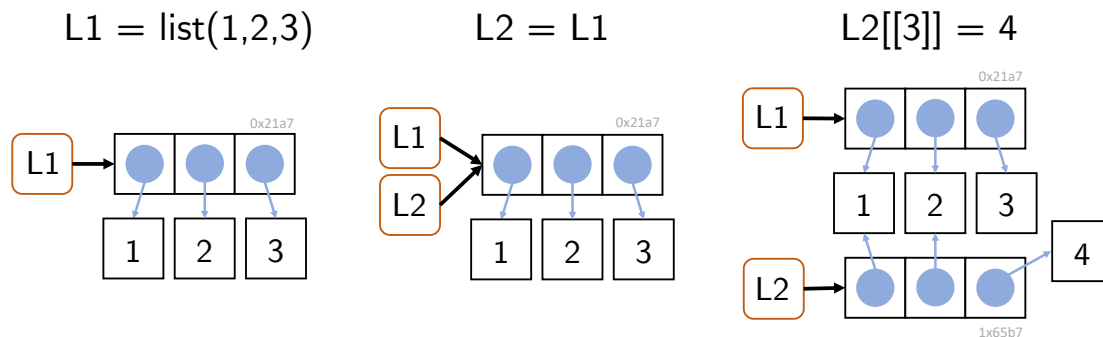


Abbildung 16.2. Effekte der Zuweisung `L2 = L1` für eine Liste L1 sowie der Modifikation eines Elements von L2.

16.2. Dataframes

Dataframes sind die zentrale Datenstruktur in **R**. Dataframes stellt man sich am besten als Tabelle vor, deren Spalten Namen haben. Technisch ist ein Dataframe eine **R** Liste, deren Elemente **R** Vektoren gleicher Länge sind. Die Listenelemente entsprechen dabei den Spalten einer Tabelle und die Vektorelemente an gleicher Position entsprechen den Zeilen einer Tabelle.

Erzeugung

Dataframes werden mit der Funktion `data.frame()` erzeugt. Dabei werden als Argument der Funktion die Spaltennamen und ihre vektorwertigen Inhalte festgelegt:

```
D = data.frame(x = letters[1:4], # 1. Spalte mit Name x
               y = 1:4,          # 2. Spalte mit Name y
               z = c(T,T,F,T))   # 3. Spalte mit Name z
print(D)                        # Ausgabe des Dataframes als Tabelle
```

```
D = data.frame(c(1,2,3),
               c("a", "b", "c"),
               c(T,F,F))
```

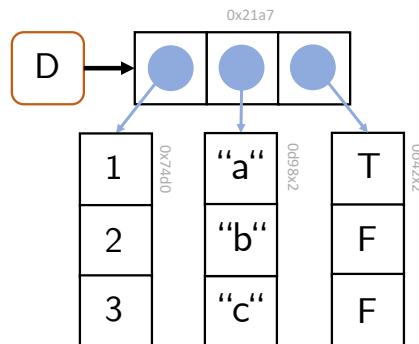


Abbildung 16.3. Dataframerepräsentation. Ein **R** Dataframe ist einer **R** Liste deren Elemente **R** Vektoren gleicher Länge sind.

```
  x y    z
1 a 1 TRUE
2 b 2 TRUE
3 c 3 FALSE
4 d 4 TRUE
```

Die Spalten des Dataframes müssen die gleiche Länge haben, ansonsten ist ein Dataframe nicht definiert:

```
D = data.frame(x = letters[1:4], # 1. Spalte mit Name x
               y = 1:4,          # 2. Spalte mit Name y
               z = c(T,T,F))     # 3. Spalte mit Name z
```

```
Error in data.frame(x = letters[1:4], y = 1:4, z = c(T, T, F)): arguments imply differing number of rows: 4,
```

Die Spalten eines Dataframes können offenbar unterschiedlichen Datentyps sein, dies ist natürlich mit dem Charakter von **R** Listen kongruent.

Charakterisierung

Mit den Funktionen `names()` und `colnames()` können die Namen der Spalten eines Dataframes ausgegeben werden, mit der Funktion `rownames()` seine Zeilennamen. Per Default sind die Zeilennamen eines Dataframes die ganzen Zahlen 1,2,...:

```
D = data.frame(age = c(30,35,40,45), # 1. Spalte
               height = c(178,189,165,171), # 2. Spalte
               weight = c(67, 76, 81, 92)) # 3. Spalte
names(D) # names gibt die Spaltennamen aus
```

```
[1] "age"    "height" "weight"
```

```
colnames(D) # colnames entspricht names
```

```
[1] "age" "height" "weight"
```

```
rownames(D) # default rownames sind 1,2,...
```

```
[1] "1" "2" "3" "4"
```

Des Weiteren ist ein Dataframe durch die Anzahl seiner Zeilen und Spalten charakterisiert. Diese können mit den Funktionen `nrow()` bzw. `ncol()` ausgegeben werden. Die Funktion `length()` gibt für Dataframes wie für Listen die Anzahl der Listeneinträge, bei Dataframes also entsprechend die Spaltenanzahl, aus.

```
nrow(D) # Zeilenanzahl
```

```
[1] 4
```

```
ncol(D) # Spaltenanzahl
```

```
[1] 3
```

```
length(D) # Länge ist die Spaltenanzahl
```

```
[1] 3
```

Eine hilfreiche Funktion zur Charakterisierung eines Dataframes ist die Funktion `str()`, die in kompakter Form wesentliche Aspekte eines Dataframes anzeigt. `str()` steht dabei für *structure* und bezeichnet die Spalten des Dataframes als *variables* und Zeileneinträge als *observations*:

```
str(D)
```

```
'data.frame':  4 obs. of  3 variables:
 $ age   : num  30 35 40 45
 $ height: num  178 189 165 171
 $ weight: num   67  76  81  92
```

Attribute

Technisch sind Dataframes Listen mit Attributen für (column) `names` und `row.names`. Dataframes sind von derclass “data.frame”:

```
typeof(D)
```

```
[1] "list"
```

```
attributes(D)
```

```
$names
[1] "age"      "height" "weight"

$class
[1] "data.frame"

$row.names
[1] 1 2 3 4
```

Indizierung

Dataframes können sowohl wie Matrizen als auch wie Listen indiziert werden. Die Prinzipien der Indizierung entsprechen dabei denen von Vektoren. Folgendes Beispiel zeigt zunächst, wie ein Dataframe im Sinne einer Liste mit einfachen Klammern indiziert wird. Das entsprechende Listenelement ist bei Dataframes dann selbst ein Dataframe:

```
D = data.frame(x = letters[1:4], # 1. Spalte mit Name x
               y = 1:4,          # 2. Spalte mit Name y
               z = c(T,T,F,T))   # 3. Spalte mit Name z
class(D)                        # D ist ein Dataframe
```

```
[1] "data.frame"
```

```
v = D[1]                        # 1. Listenelement als Dataframe
v
```

```
  x
1 a
2 b
3 c
4 d
```

```
class(v)                        # v ist ein Dataframe
```

```
[1] "data.frame"
```

Des Weiteren können Dataframes wie Listen mit doppelten Klammern indiziert werden, es werden dadurch entsprechend die Inhalte eines Listenelements adressiert:

```
w = D[[1]]                     # Inhalt des 1. Listenelements
w
```

```
[1] "a" "b" "c" "d"
```

```
class(w)                        # w ist ein character vector
```

```
[1] "character"
```

Die typische Indizierung eines Dataframes erfolgt mit der von den Listen bekannten \$ Indizierung, die den Inhalt der entsprechenden Dataframe Spalte adressiert:

```
y = D$y # $ zur Indizierung der y Spalte
y
```

```
[1] 1 2 3 4
```

```
class(y) # v ist ein Vektor vom Typ "integer" (!)
```

```
[1] "integer"
```

Ist man an speziellen Inhalten des Dataframes in spezifischen Zeilen und Spalten interessiert, so bieten sich die Prinzipien der Matrixindizierung an, wie folgendes Beispiel zeigt:

```
D[2:3,-2] # 1. Index fuer Zeilen, 2. Index fuer Spalten
```

```
  x      z
2 b  TRUE
3 c FALSE
```

```
D[c(T,F,T,F),] # 1. Index fuer Zeilen, 2. Index fuer Spalten
```

```
  x y      z
1 a 1  TRUE
3 c 3 FALSE
```

```
D[,c("x", "z")] # 1. Index fuer Zeilen, 2. Index fuer Spalten
```

```
  x      z
1 a  TRUE
2 b  TRUE
3 c FALSE
4 d  TRUE
```

Copy-on-modify

Zur optimalen Nutzung des Arbeitsspeichers gelten die Copy-on-Modify Prinzipien für Listen auch für Dataframes. Weiterhin gelten dabei, dass die Modifikation einer Spalte zur Kopie der entsprechenden Spalte führt, wohingegen die Modifikation einer Zeile in der Kopie des gesamten Dataframes resultiert.

16.3. Tibbles

In Ergänzung zu den **R** Dataframes haben in den letzten zehn Jahren die *Tibbles* (diminutiv für *tables*) als Teil des **tidyverse** Pakets Verbreitung gefunden. Tibbles sind programmiertechnisch optimierte Dataframes und stellen ein Beispiel für die datenwissenschaftliche Verallgemeinerung der **R** Anwendung über den Bereich der linearen statistischen Modellbildung hinaus dar. Zur Nutzung von Tibbles muss das **tidyverse** Paket installiert und der Tibble Code geladen sein:

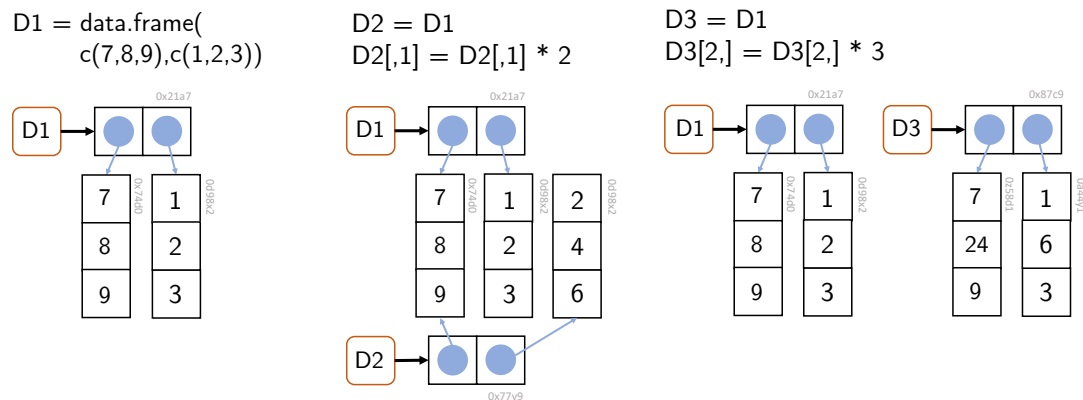


Abbildung 16.4. Effekte von Spalten und Zeilen Modifikationen bei Dataframes. Die Modifikation einer Spalte führt zur Kopie der entsprechenden Spalte, die Modifikation einer Zeile zur Kopie des gesamten Dataframes.

```
install.packages("tidyverse")           # Einmaliger Paket Download und Installation
library(tibble)                         # Sessionspezifisches Laden des tibble Codes
```

Im Großen und Ganzen entsprechen Tibbles den Dataframes. Wie bei Dataframes handelt es sich bei Tibbles um Listen von Vektorreferenzen, wobei auch hier die Vektoren die gleiche Länge haben und die benannten Spalten einer Tabelle darstellen. Tibbles können mit der Funktion `tibble()` erzeugt werden. Die Anzeige im **R** Terminal ist im Vergleich zu Dataframes optimiert, wie folgendes Beispiel zeigt:

```
library(tibble)                         # Paket
T = tibble(x = c("a", "c", "d"), y = c(1,2,3), z = c(TRUE,FALSE,FALSE)) # Erzeugung
print(T)                               # Ausgabe
```

```
# A tibble: 3 x 3
  x         y z
<chr> <dbl> <lgl>
1 a         1 TRUE
2 c         2 FALSE
3 d         3 FALSE
```

```
attributes(T)                           # Attribute
```

```
$class
[1] "tbl_df"      "tbl"        "data.frame"

$row.names
[1] 1 2 3

$names
[1] "x" "y" "z"
```

16.3.1. Tibbles vs. Dataframes

Tibbles wurden 2016 im Rahmen des `tidyverse` Pakets eingeführt, um wahrgenommene Schwächen der **R** Dataframes auszugleichen. Dazu gehörte unter anderem, dass die

Funktion `data.frame()` character Vektoren automatisiert in den **R** Datentyp `factor` umwandelte. Dies ist seit einer **R** Revision von 2019 allerdings nicht mehr der Fall, wie folgendes Beispiel zeigt:

```
D = data.frame(x = 1:3, y = c("a", "b", "c"))      # character vector y
str(D)                                             # ... bleibt character vector
```

```
'data.frame':  3 obs. of  2 variables:
 $ x: int  1 2 3
 $ y: chr  "a" "b" "c"
```

Per Default wandelt `data.frame()` weiterhin nicht legale Spaltennamen wie eine Zahl (1) in legale Spaltennamen (X1) um. Möchte man dies nicht, so kann man dieses Verhalten allerdings durch Setzen des Arguments `check.names` abstellen

```
names(data.frame(`1` = 1))                        # Per Default Umwandlung von nicht legalen Variablennamen
```

```
[1] "X1"
```

```
names(data.frame(`1` = 1, check.names = FALSE)) # Dieser Default kann jedoch abgestellt werden.
```

```
[1] "1"
```

Wie in Kapitel 14 gesehen, hat **R** die Tendenz, bei von ihrer Länge unpassenden Datentypen selbst Daten durch Recycling zu erzeugen, was leicht zu Fehlern führen kann. So recyclet die Funktion `data.frame()` bei der Erzeugung eines Dataframes zum Beispiel Vektoren, wenn einer davon ein ganzzahliges Vielfaches des anderen ist:

```
D = data.frame(x = 1:2, y = 3:6)      # length(y) = 2 * length(x)
print(D)                               # Ausgabe
```

```
  x y
1 1 3
2 2 4
3 1 5
4 2 6
```

`tibble()` dagegen ist hier genauer und gibt im obigen Fall einen Fehler aus:

```
T = tibble(x = 1:2, y = 3:6)          # tibble gibt hier eine Fehler aus
```

```
Error in `tibble()`:
! Tibble columns must have compatible sizes.
* Size 2: Existing data.
* Size 4: Column `y`.
i Only values of size one are recycled.
```

Vektoren der Länge 1 werden allerdings auch von `tibble()` recyclet:

```
T = tibble(x = 1:3, y = 3)      # Vektoren mit length 1
print(T)                       # Ausgabe
```

```
# A tibble: 3 x 2
      x     y
  <int> <dbl>
1     1     3
2     2     3
3     3     3
```

Dataframes sind manchmal inkonsistent, was die von ihnen ausgegeben Datentypen angeht. So ergibt, wie unteres Beispiel zeigt, die Indizierung einer Dataframespalte mit ihrem Namen etwa einen Dataframe, die Indizierung der gleichen Spalte mit ihrem Namen in Matrixindizierung allerdings einen Vektor:

```
D = data.frame(x = 1:3, y = 4:6)  # Dataframe
str(D["x"])                       # Indizierung mit Namen gibt einen Dataframe aus
```

```
'data.frame':  3 obs. of  1 variable:
 $ x: int  1 2 3
```

```
str(D[, "x"])                     # Matrixindizierung mit Namen gibt einen Vektor aus
```

```
int [1:3] 1 2 3
```

Tibbles sind in dieser Hinsicht konsistenter:

```
T = tibble(x = 1:3, y = 4:6)     # Tibble
str(T["x"])                      # Indizierung mit Namen ergibt tibble
```

```
tibble [3 x 1] (S3: tbl_df/tbl/data.frame)
 $ x: int [1:3] 1 2 3
```

```
str(T[, "x"])                    # Matrixindizierung mit Namen ergibt tibble
```

```
tibble [3 x 1] (S3: tbl_df/tbl/data.frame)
 $ x: int [1:3] 1 2 3
```

Ein zusätzliches Feature von Tibbles gegenüber Dataframes ist, dass die Funktion `tibble()` es erlaubt, Spaltennamen schon bei Erzeugung eines Tibble zu referenzieren, was bei Dataframes nicht möglich ist. Zusätzliches Feature von Tibbles

```
T = tibble(x = 1, y = 2*x)       # Die Spalte x wird definiert und benutzt
D = data.frame(x = 1, y = 2*x)   # data.frame erlaubt dies nicht
```

```
Error in eval(expr, envir, enclos): object 'x' not found
```

Zusammenfassend lässt sich festhalten, dass Tibbles in mancher Hinsicht weniger flexibel und damit präziser als Dataframes funktionieren. Generell verhalten sich Tibbles in `tidyverse` Kontexten konsistenter als Dataframes. Allerdings sind die Unterschiede zwischen Dataframes und Tibbles wie oben gesehen eher subtil, sodass ein effizienter Umgang mit Dataframes auch einen effizienten Umgang mit Tibbles ermöglicht. Da nicht alle **R** Projekte die Pakete des `tidyverse` nutzen, werden wohl auch in Zukunft Dataframes die Standarddatenstruktur für tabellarische Daten in **R** bleiben.

17. Datenmanagement

In diesem Abschnitt wollen wir einige allgemeine Begrifflichkeiten zum Umgang mit Forschungsdaten einführen, gängige Formate zur längerfristigen Speicherung von Daten diskutieren und uns schließlich mit dem praktischen Einlesen und Speichern von Daten im Rahmen des **R** Verzeichnismanagements beschäftigen.

17.1. Das FAIR Datenideal

Das FAIR Datenideal ist ein Prinzip zum Management von Forschungsdaten, dass der weiten Verbreitung, Nützlichkeit und Transparenz von Forschungsdaten im Digitalzeitalterrechnung trägt. Unter *Forschungsdaten* werden dabei im allgemeinen “elektronisch repräsentierte analoge oder digitale Daten, die im Zuge wissenschaftlicher Vorhaben entstehen oder genutzt werden, z.B. durch Beobachtungen, Experimente, Simulationsrechnungen, Erhebungen, Befragungen, Quellenforschungen, Aufzeichnungen von Audio- und Videosequenzen, Digitalisierung von Objekten, und Auswertungen.” (Rat für Informationsinfrastrukturen (2017)) verstanden. Konkret handelt sich dabei etwa um Zahlenarrays oder Characterarrays, aber auch Computersoftware, wie **R** Skripte, digitale Werkzeuge, Workflows und Analysepipelines.

Neben diesen, manchmal auch als *Primärdaten* bezeichneten Datenformen sind für das FAIR Datenideal sogenannte *Metadaten* von Bedeutung. Dabei handelt es sich um Daten, die Informationen über andere Daten repräsentieren. Man unterscheidet dabei beispielsweise *deskriptive*, *strukturelle* und *administrative Metadaten* (Riley (2017), Ulrich et al. (2022)).

- *Deskriptive Metadaten* dienen dem Auffinden und der Identifikation einer Datenquelle. Beispiele für deskriptive Metadaten sind etwa Titel der einer wissenschaftlichen Publikation oder ihr digital object identifier (DOI, Rosenblatt (1997)). DOIs sind alphanumerische Kennungen, die digitale Objekte eindeutig identifizieren und als persistente Links zu Objekten im Internet funktionieren.
- *Strukturelle Metadaten* sind Metadaten über Datencontainer und repräsentieren den strukturellen Aufbau einer Datenquelle. Beispiele sind die Ordnung der Seiten eines Buches oder die for-Schleifenkodierung dreidimensionaler Datenobjekte.
- *Administrative Metadaten* schließlich sind Daten, die das Management einer Datenquelle erleichtern. Beispiele für administrative Metadaten sind Provenienz, das Dateiformat, die Zugangsrechte oder andere technische Informationen zu einer Datenquelle.

Das FAIR Ideal zum Forschungsdatenmanagement hat seinen Ursprung in Diskussionen im Rahmen der FORCE11 Konferenzserie und wurde 2016 durch Wilkinson et al. (2016) akademisch kommuniziert. Es besagt im Kern, dass Forschungsdaten *Findable*, *Accessible*, *Interoperable*, und *Reusable* für Menschen und Maschinen aufbereitet und vorgehalten

werden sollen. Speziell formuliert das FAIR Datenideal folgende Ansprüche an das Management von Forschungs(meta)daten:

Findability (Auffindbarkeit)

- F1. (Meta)Daten haben einen persistenten global einzigartigen Identifikator.
- F2. Daten werden mit Metadaten angereichert.
- F3. Metadaten sind zweifelsfrei einem Datensatz zuzuordnen.
- F4. (Meta)Daten sind in einer durchsuchbaren Ressource indexiert.

Accessibility (Zugänglichkeit)

- A1. (Meta)Daten sind mit standardisierten Protokollen abrufbar.
- A1.1. Das genutzte Protokoll ist offen, kostenlos und nutzbar.
- A1.2. Das Protokoll ermöglicht Authentifizierung und Rechtevergabe.
- A2. Metadaten bleiben zugänglich, auch wenn Daten nicht mehr vorliegen.

Interoperability (Interoperabilität)

- I1. (Meta)Daten nutzen eine formale, zugängliche, gemeinsam genutzte und breit anwendbare Sprache zur Wissensrepräsentation.
- I2. (Meta)Daten nutzen Vokabularen, die den FAIR-Prinzipien folgen.
- I3. (Meta)Daten enthalten qualifizierte Referenzen auf andere (Meta)Daten.

Reusability (Wiederverwendbarkeit)

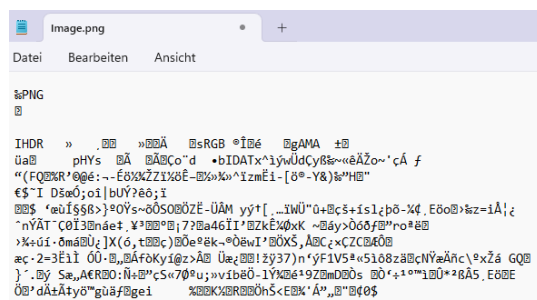
- R1. (Meta)Daten haben eine Vielzahl genauer und relevanter Attribute.
- R1.1. (Meta)Daten enthalten eine eindeutige Nutzungslizenz.
- R1.2. (Meta)Daten enthalten detaillierte Provenienz-Informationen.
- R1.3. (Meta)Daten genügen den Standards der jeweiligen Fachcommunity.

Der Anspruch an ein FAIRes Forschungsdatenmanagement ist mittlerweile recht verbreitet, wobei die FAIR Prinzipien mit mehr oder weniger bürokratischem Aufwand umgesetzt werden können. Es bleibt aber festzuhalten, dass zurzeit die FAIR Prinzipien im Wesentlichen ein anzustrebendes Datenmanagementideal sind, und längst nicht alle Forschungsdaten FAIR verfügbar sind. Tatsächlich ist der Umgang mit digitalen Forschungsdaten auch weiterhin noch sehr unstrukturiert und gerade das deutsche Universitätssystem begreift das digitale Datenmanagement nur sehr langsam als eine Kernaufgabe. Im größeren, wenn auch Projekt-basiertem, und damit nicht nachhaltigem Rahmen, versucht die Initiative zur [Nationalen Forschungsdateninfrastruktur \(NFDI\)](#), das deutsche digitale Forschungsdatenmanagement zu. Leider bleibt weiterhin festzuhalten, dass viele durch öffentliche Mittel finanzierte Wissenschaftler:innen von ihnen erhobene Forschungsdaten weder systematisch aufbereiten noch der Öffentlichkeit zur Verfügung stellen wollen. Hier ist also weiterhin auf einen Kulturwandel zu mehr wissenschaftlicher Transparenz zu hoffen und offene öffentlich-finanziert Forschung (Open Research, vgl. z.B. Toelch & Ostwald (2018), Miedema (2022), Bertram et al. (2023)) bleibt ein anzustrebendes Ideal.

17.2. Dateiformate

Allgemeine definiert Dateiformat Syntax und Semantik von Daten innerhalb einer Datei. Dateiformate sind dabei bijektive Abbildungen von Information auf den binären Langzeitspeicher. Allgemein unterscheidet man *Datendateiformate* gegenüber *Softwaredateiformaten*, *textuelle* gegenüber *binären Dateiformaten* und *offene* gegenüber *proprietären*, d.h. urheberrechtlich geschützten, Dateiformaten. Bei binären Datendateiformaten ist das Einlesen von Daten, ihre Inspektion, und ihre Manipulation generell nur mit spezieller Software möglich. Beispiele für binäre Dateiformate sind *.docx*, *.pdf*, *.xlsx*, *.jpg*, und *.mp4* Dateien. Binäre Dateiformate sind dabei oft proprietär. Früher wurden binäre Dateiformate aufgrund ihres geringeren Speicherbedarfs auch für Forschungsdaten bevorzugt eingesetzt. Heutzutage werden dagegen textuelle Datendateiformate wie *.txt*, *.csv*, *.tsv*, und *.json* favorisiert. Bei diesen Formaten ist das Einlesen der gespeicherten Daten, ihre Inspektion und ihre Manipulation mit einfachen allgemeinen Editoren, wie beispielsweise VSCode problemlos möglich und die Dateiformate sind generell nicht proprietär. Abbildung 17.1 zeigt die Darstellung eines binären Datenformats und eines textuellen Datenformats im Windows Editor. Generell bieten sich im Forschungskontext und im Sinne der FAIR Prinzipien also offene, textuelle Datendateiformate an. Exemplarisch wollen wir hier als textuelle Dateiformate das *.csv* und *.tsv*, sowie das *.json* Dateiformat genauer betrachten.

Binäres Datenformat



Textuelles Datenformat

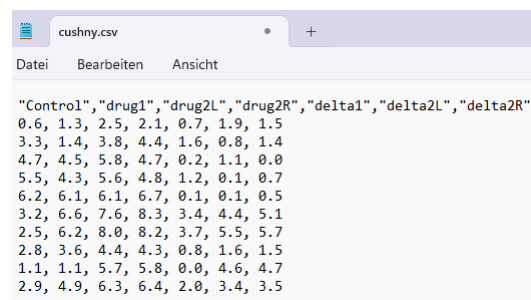


Abbildung 17.1. Binäre und textuelle Dateiformate in Editordarstellung.

17.2.1. .csv und .tsv. Dateien

.csv (comma-separated values) und *.tsv* (tab-separated values) Dateien sind wichtige Dateiformate zur Speicherung einfacher, tabellarisch strukturierter Daten. Allgemein repräsentieren *.csv* und *.tsv* zeilenweise miteinander verknüpfte Datensätze. Dabei werden Datenfelder innerhalb einer Zeile durch Kommata (*.csv*) oder Tabulatorzeichen (*.tsv*) und einzelne Datensätze durch Zeilenumbrüche voneinander abgegrenzt. Der Datensatz in der ersten Zeile einer *.csv* oder *.tsv* Datei enthält typischerweise die Definition der Spaltennamen und wird als Kopfdatensatz (Header) bezeichnet. Ein typisches Beispiel einer *.csv* ist in Abbildung 17.1 (rechts) dargestellt. Dabei repräsentieren die Zeilen experimentelle Einheiten (hier Versuchspersonen) und die Spalten verschiedene Variablen, für die jeweils ein Wert bei der Versuchsperson bestimmt wurde.

Im Kontext der Datenanalyse unterscheidet man bei `.csv` und `.tsv` Dateien eine *wide format* und eine *long format* Datenrepräsentation. Bei der *wide format* Datenrepräsentation werden alle Variablen einer experimentellen Einheit wie in obigem Beispiel in einer Zeile repräsentiert, vgl. Tabelle 17.1. Allerdings ist es für manche Datenanalysen hilfreicher, die Daten einer experimentellen Einheit über Zeilen zu verteilen. Dies führt auf das sogenannte *long format*, wie in Tabelle 17.2 gezeigt. Generell ist das *wide format* für ein erstes intuitives Verständnis eines Datensatzes oft hilfreicher, wohingegen das *long format* Daten im Rahmen modell-basierter Analysen meist konsistenter mit der Modellarchitektur repräsentiert.

Tabelle 17.1. Wide Format (EE: Experimentelle Einheit, V1,V2,V3: Variablen)

EE	V1	V2	V3
1	10.1	67.5	4
2	12.9	51.2	2
3	20.4	70.8	5
4	17.5	60.3	7

Tabelle 17.2. Long Format (EE: Experimentelle Einheit, V1,V2,V3: Variablen)

Einheit	Variable	Messwert
1	V1	10.1
1	V2	67.5
1	V3	4
2	V1	12.9
2	V2	51.2
2	V3	2
3	V1	20.4
3	V2	70.8
3	V3	5
4	V1	17.5
4	V2	60.3
4	V3	7

17.2.2. `.json` Dateien

`.json` (JavaScript Object Notation), ist ein textuelles Datenformat, das zum Speichern strukturierter Daten in Key-Value-Form verwendet wird. Es weist Ähnlichkeiten mit R-Listen auf, insbesondere mit benannten Listenelementen, und eignet sich daher gut als Format für die Speicherung von Metadaten.

Eine `.json`-Datei besteht aus Objekten, die in geschweiften Klammern `{ }` notiert werden und eine durch Kommata getrennte Liste von Eigenschaften enthalten. Jede Eigenschaft ist ein Key-Value-Paar, wobei der Key immer als String in Hochkommata (`" "`) geschrieben wird. Der Value kann wiederum ein Objekt, ein Array, ein String, ein Boolean oder eine Zahl sein. Abbildung 17.2 zeigt den Inhalt einer `.json` Datei zur Speicherung von Studierendeninformationen.

```
{
  "Vorname" : "Maxi",
  "Nachname" : "Musterfrau",
  "Matrikelnummer" : 12345,
  "Fachsemester" : 2,
  "Studiengang" : "BSc Psychologie",
  "Module" :
  {
    "Deskriptive Statistik" : { "Abgeschlossen" : TRUE, "Note" : 1.0 },
    "Inferenzstatistik" : { "Abgeschlossen" : FALSE, "Note" : NA }
  }
}
```

Abbildung 17.2. .json Dateiformat.

17.3. Verzeichnismangement

Da Verzeichnissysteme von Computern üblicherweise in Worten (z.B. `C:\Users`) statt numerisch dargestellt sind, ist eine Grundlage für das automatisierte Einlesen von langfristig gespeicherten Daten in den Arbeitsspeicher und umgekehrt des Speicherns von Daten aus dem Arbeitsspeicher im langfristigen Speicher das Arbeiten mit sogenannten *strings* (Zeichenketten). Strings werden in **R** mit Anführungszeichen oder Hochkommata erzeugt:

```
# Anführungszeichen sind der String Standard
c("Dies ist ein character vector")
```

```
[1] "Dies ist ein character vector"
```

```
# Hochkommata koennen bei Anführungszeichen im String helfen
c('Dies ist ein "string"')
```

```
[1] "Dies ist ein \"string\""
```

Die Funktion `paste()` konvertiert Vektoren in character und fügt sie elementweise zusammen.

```
# Konvertierung und Konkatenation einelementiger double vectors
paste(1,2)
```

```
[1] "1 2"
```

```
# Konkatenation einelementiger character strings
paste("Dies ist", "ein String")
```

```
[1] "Dies ist ein String"
```

Die Funktion `paste()` hat darüber hinaus eine Reihe weiterer Funktionalitäten, die manchmal hilfreich sein können:

```
# Vector recylcing, elementweise Veknuepfungen
paste(c("Rote", "Gelbe"), "Blume")
```

```
[1] "Rote Blume" "Gelbe Blume"
```



```
# Separatorspezifikation
paste(c("Rote", "Gelbe"), "Blume", sep = "-")
```

```
[1] "Rote-Blume" "Gelbe-Blume"
```

```
# Zusammenfuegen mit spezifiziertem Separator
paste(c("Rote", "Gelbe"), "Blume", collapse = ", ")
```

```
[1] "Rote Blume, Gelbe Blume"
```

Die Funktion `toString()` ist eine `paste()` Variante für numerische Vektoren:

```
# Konversion eines double Vektors in formatierten String
toString(1:10)
```

```
[1] "1, 2, 3, 4, 5, 6, 7, 8, 9, 10"
```

```
# mit Moeglichkeit der Beschraenkung auf width Zeichen
toString(1:10, width = 10)
```

```
[1] "1, 2, ...."
```

Um nun die Daten einer Datei im Langzeitspeicher in den Arbeitspeicher zu laden, benötigt man ihre Adresse in der Verzeichnisstruktur des Computers. Abbildung 17.3 zeigt einen Ausschnitt der Verzeichnisstruktur eines typischen Windows-Rechners.

Die Adressen von Dateien in der Verzeichnisstruktur heißen *Dateipfade* (*paths*). Ein Pfad besteht aus einer durch Schrägstriche getrennten Liste von Verzeichnisnamen. Dabei werden unter Windows traditionell linksgeneigte Schrägstriche (`\`, *backslash*) und unter Mac und Linux rechtsgeneigte Schrägstriche (`/`, *slash*) genutzt. Bei Windows findet sich zusätzlich in der Regel an oberster Stelle der Verzeichnisstruktur der Festplattenname (z.B. `C:\` oder `D:\`), unter Mac der Benutzername (`/Users/`) und Linux das Heimverzeichnis (`/home/`). Die folgenden Beispiele sind Windows-basiert.

```
# Pfad der auf einem Verzeichnisnamen endet
C:\Users\dirko\Nextcloud\Home\pdwp-online\dirk-ostwald.github.io\_data

# Pfad der auf einem Dateinamen endet
C:\Users\dirko\Nextcloud\Home\pdwp-online\dirk-ostwald.github.io\_data\cushny.csv
```

Absolute Dateipfade geben die Adresse eines Verzeichnisses oder einer Datei in der Gesamtverzeichnisstruktur der Festplatte und sind wenig anfällig für Dateiverwechselungen. *Relative Dateipfade* beziehen sich dagegen auf einen Speicherort in Relation zum aktuellen Verzeichnis, wobei `.` und `..` das aktuelle und das übergeordnete Verzeichnis bezeichnen. Bei relativen Dateipfaden kann es durchaus zu Dateiverwechselungen kommen. Generell wird also die Verwendung adaptiv generierter absoluter Pfade empfohlen.

```
# Dateiname in absoluter Pfadform
fname = "D:\Lehre\Daten\cushny.csv"
```

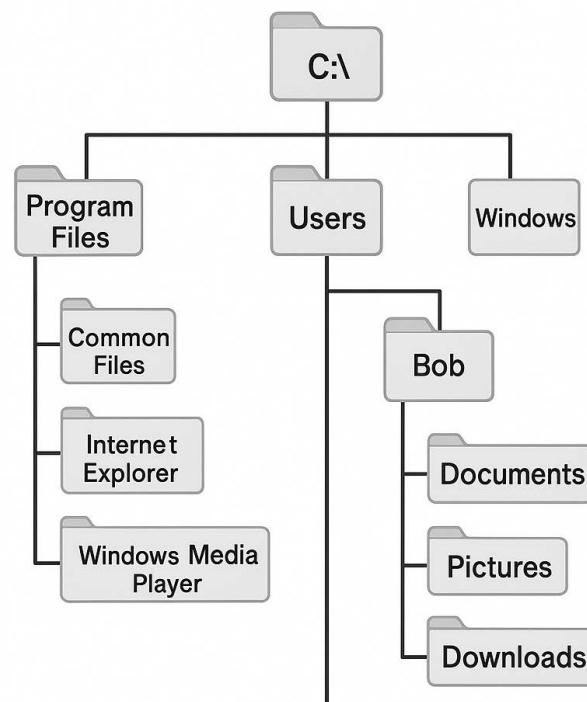


Abbildung 17.3. Windows Verzeichnisstruktur

R besitzt ein sogenanntes *working directory* aus dem per Default Dateien gelesen werden. Bei geöffnetem **R** Terminal gibt `getwd()` das working directory an.

```
getwd()
```

```
[1] "C:/Users/dirko/Nextcloud/Home/pdwp-online/dirk-ostwald.github.io/_part_2"
```

`setwd()` ändert das working directory. Dabei ist zu beachten, dass **R** nur mit forward slashes / arbeitet. Die manuelle Spezifikation von Windowspfaden benötigt deshalb doppelte backward slashes \\

```
setwd("C:\\Users\\dirko\\Nextcloud\\Home\\pdwp-online\\dirk-ostwald.github.io\\_data")
getwd()
```

Mithilfe der `file.path()` Funktion können Betriebssystem-konforme Verzeichnis und Dateipfade erzeugt werden:

```
file.path("D:", "Nextcloud", "Home", "Teaching")
```

Die Funktion `dirname()` gibt das Verzeichnis an, das ein Verzeichnis oder eine Datei enthält:

```
getwd()
```

```
[1] "C:/Users/dirko/Nextcloud/Home/pdwp-online/dirk-ostwald.github.io/_part_2"
```

```
dirname(getwd())
```

```
[1] "C:/Users/dirko/Nextcloud/Home/pdwp-online/dirk-ostwald.github.io"
```

Die Funktion `basename()` gibt die unterste Ebene eines Datei- oder Verzeichnispfades an:

```
getwd()
```

```
[1] "C:/Users/dirko/Nextcloud/Home/pdwp-online/dirk-ostwald.github.io/_part_2"
```

```
basename(getwd())
```

```
[1] "_part_2"
```

17.4. Datenimport und Datenexport

Zum Einlesen von einfach strukturierten .csv Dateien bietet sich die **R** Funktion `read.csv()` an. Diese Funktion liest eine Datei ein und speichert ihre Inhalte in einem Dataframe.

```
D = read.csv("C:\\Users\\dirko\\Nextcloud\\Home\\pdwp-online\\dirk-ostwald.github.io\\_part_2\\data.csv")
print(D)
```

	Control	drug1	drug2L	drug2R	delta1	delta2L	delta2R
1	0.6	1.3	2.5	2.1	0.7	1.9	1.5
2	3.3	1.4	3.8	4.4	1.6	0.8	1.4
3	4.7	4.5	5.8	4.7	0.2	1.1	0.0
4	5.5	4.3	5.6	4.8	1.2	0.1	0.7
5	6.2	6.1	6.1	6.7	0.1	0.1	0.5
6	3.2	6.6	7.6	8.3	3.4	4.4	5.1
7	2.5	6.2	8.0	8.2	3.7	5.5	5.7
8	2.8	3.6	4.4	4.3	0.8	1.6	1.5
9	1.1	1.1	5.7	5.8	0.0	4.6	4.7
10	2.9	4.9	6.3	6.4	2.0	3.4	3.5

Folgender **R** Code definiert den absoluten Pfad der Datei adaptiv in der Verzeichnisstruktur des Computers mithilfe des **R** `working directories`.

```
wdir    = getwd()                # Working directory
print(wdir)                      # Ausgabe
ddir    = file.path(wdir, "_part_2/_data") # Datenverzeichnispfad
print(ddir)                      # Ausgabe
fname   = "cushny.csv"          # (base) filename
print(fname)                    # Ausgabe
fpath   = file.path(ddir, fname) # filepath
print(fpath)                   # Ausgabe
D       = read.csv(fpath)       # Einlesen der Datei
print(D)                       # Ausgabe des Dataframes
```

Darüber hinaus bietet **R** eine Vielzahl von Möglichkeiten des Datenimports mithilfe spezialisierte Funktionen:

Für .csv und andere Textdateien:

- `read.csv()`, `read.csv2()`, `read.delim()`, `read.delim2()` als `read.table()` Varianten.
- `readlines` für low-level Textdateiimport.
- `fromJSON()` aus dem Paket `rjson` für .json Dateien.

Für binäre Dateien:

- `read.xlsx()` und `read.xlsx2()` aus dem Paket `xlsx` für Excel .xlsx Dateien.
- `read.spss()` aus dem Paket `foreign` für SPSS .sav Dateien.
- `readMat` aus dem Paket `R.matlab` für Matlab .mat Dateien.

Für Datenbanken:

- SQL Datenbanken können mithilfe der Pakete `DBI` und `RSQLite` abgefragt werden.

18. Kontrollstrukturen

Perse wird imperativer Programmiercode streng sequentiell Befehl für Befehl ausgeführt. Manchmal möchte man von dieser rein sequentiellen Befehlsreihenfolge allerdings abweichen und die Ausführungsreihenfolge von Befehlen flexibler kontrollieren. Prinzipielle Werkzeuge zu diesem Zweck sind *konditionale Kontrollstrukturen* und *iterative Kontrollstrukturen*. *Konditionale Kontrollstrukturen* steuern die Programmausführung, indem sie festlegen, dass bestimmte Befehle nur dann ausgeführt werden, wenn definierte Bedingungen erfüllt sind. **R** bietet zu diesem Zweck `if()` und `switch()` Statement. Oft möchte man bestimmte Codeabschnitte auch in Schleifen häufig wiederholen (*iterieren*). Zu diesem Zweck bietet **R** `for()`, `while()`, und `repeat()` Schleifen. In diesem Abschnitt stellen wir die grundlegenden konditionalen und iterativen Kontrollstrukturen in **R** vor und folgen dabei der Darstellung in Cotton (2013), Kapitel 8 und Kapitel 9, sowie Wickham (2019), Kapitel 5.

18.1. Konditionale Kontrollstrukturen

Konditionale Kontrollstrukturen erlauben die bedingungsabhängige Ausführung von Befehlen.

if-statements

Die grundlegende konditionale Kontrollstruktur ist das if-statement. Die allgemeine Syntax für ein if-statement in **R** hat folgende Form:

```
if (Bedingung){  
  TrueActions  
}
```

Dabei gilt, dass wenn die Variable `Bedingung` den Wert `TRUE` hat, also wahr ist, die mit `TrueActions` bezeichneten Befehle ausgeführt werden. Ist der Wert der Variable `Bedingung` dagegen `FALSE`, so werden die Befehle `TrueActions` nicht ausgeführt und die Programmausführung wird in der nächsten Zeile nach dem if-statement fortgesetzt. Beispielsweise wird in folgendem Code der Befehl `print("x ist größer als 0")` ausgeführt, da `x` ja nach vorheriger Zuordnung den Wert `1` hat.

```
x = 1  
if(x > 0){  
  print("x ist größer als 0")  
}
```

```
[1] "x ist größer als 0"
```

Das if-statement kann durch eine **else** Bedingung erweitert werden, um zu spezifizieren, welche Befehle in dem Fall ausgeführt werden sollen, dass die Variable **Bedingung** den Wert **FALSE** hat.

```
if (Bedingung){
  TrueActions
} else {
  FalseActions
}
```

Hier werden die Befehle **TrueActions** ausgeführt, wenn die Variable **Bedingung** den Wert **TRUE** hat und es werden die Befehle **FalseActions** ausgeführt, wenn die Variable **Bedingung** den Wert **FALSE** hat. Beispielsweise wird in folgendem Code der Befehl `print("y ist nicht größer als 0")` ausgeführt, da die Bedingung `y > 0` nicht zutrifft.

```
y = -1
if(y > 0){
  print("y ist größer als 0")
} else{
  print("y ist nicht größer als 0")
}
```

```
[1] "y ist nicht größer als 0"
```

R umfasst die in Tabelle 18.1 aufgeführten Relationsoperatoren zur Prüfung binärer logischer Bedingungen.

Tabelle 18.1. Relationsoperatoren in **R**

Relationsoperator	Bedeutung
<code>==</code>	Gleich
<code>!=</code>	Ungleich
<code><, ></code>	Kleiner, Größer
<code><=, >=</code>	Kleiner gleich, Größer gleich
<code>&</code>	UND
<code> </code>	ODER

Dabei werden die Relationsoperatoren `<`, `<=`, `>`, `>=` zumeist nur auf numerische Werte angewendet. Die Gleichheitsoperatoren `==`, `!=` können auf beliebige Datenstrukturen angewendet werden, um deren Gleichheit zu überprüfen. Die logischen Verknüpfungen `&` und `|` werden zumeist auf logische Werte angewendet. Die Funktion `xor()` schließlich implementiert das exklusive ODER. Es können beispielsweise folgende Tests auf Gleichheit und Relation durchgeführt werden

```
# Variablendefinition
x = 1
y = 2

# Test auf Gleichheit
if(x == y){
```

```
print("x ist gleich y")
} else {
  print("x ist ungleich y")
}
```

```
[1] "x ist ungleich y"
```

```
# Test auf Ungleichheit
if(x < y){
  print("x ist kleiner als y")
} else {
  print("x ist größer oder gleich y")
}
```

```
[1] "x ist kleiner als y"
```

Die if-statement Tests können durch logische Bedingungen erweitert werden, wie folgendes Beispiel demonstrieren soll:

```
# Variablendefinition
x = 2
y = 2

# logisches UND/ODER
if(x == y | x < y){
  print("x ist kleiner oder gleich y")
} else {
  print("x ist größer als y")
}
```

```
[1] "x ist kleiner oder gleich y"
```

```
# logisches UND
if(x > y & x != y){
  print("x ist größer als y")
} else {
  print("x ist kleiner oder gleich y")
}
```

```
[1] "x ist kleiner oder gleich y"
```

Bei if-Statements ist es entscheidend, dass die Bedingung einem einzigen logischen Wert entspricht. Folgende Bedingungen sind beispielsweise fehlerhaft:

```
if("x"){
  print(1)
}
```

```
Error in if ("x") {: argument is not interpretable as logical
```

```
if(NA){
  print(1)
}
```

Error in if (NA) {: missing value where TRUE/FALSE needed

```
if(c(T,F)){
  print(1)
}
```

Error in if (c(T, F)) {: the condition has length > 1

switch-statements

Eine weitere konditionale Kontrollstruktur ist das switch-statement. Switch-statements sind dadurch motiviert, dass kombinierte if-else-statements werden leicht unübersichtlich werden können, wie folgendes Beispiel zeigt:

```
x = 2
if (x == 1){
  print("Aktion 1")
} else if(x == 2){
  print("Aktion 2")
} else if(x == 3){
  print("Aktion 3")
} else if(x == 4){
  print("Aktion 4")
}
```

```
[1] "Aktion 2"
```

Hat man also einen Fall vorliegen, in dem in Abhängigkeit des Wertes einer Variable eine von vielen Aktionen ausgeführt werden soll, so kann sich ein switch-statement zur Verbesserung der Codelesbarkeit anbieten. **R** implementiert switch-statements in der `switch()` Funktion, die in Abhängigkeit vom Datentyp ihrer Variable leicht unterschiedlich funktioniert. Ist die switch Variable vom Typ Integer, so wird die *ite* Aktion ausgeführt, wie folgendes Beispiel demonstriert:

```
x = 2
switch(x,
  print("Aktion 1"),
  print("Aktion 2"),
  print("Aktion 3"),
  print("Aktion 4"))
```

```
# switch Variable
# x = 2
# 1. Aktion
# 2. Aktion
# 3. Aktion
# 4. Aktion
```

```
[1] "Aktion 2"
```

Ist die switch-Variable dagegen vom Typ character, so wird die Aktion mit dem entsprechenden Namen ausgeführt:

```
x = "a"
switch(x,
  a = print("Aktion a"),
  b = print("Aktion b"),
  c = print("Aktion c"),
  d = print("Aktion d"))
```

```
# switch Variable
# x = 2
# a Aktion
# b Aktion
# c Aktion
# d Aktion
```



```
[1] "Aktion a"
```

Generell lässt sich zu if- und switch-statements festhalten, dass sich jedes if-statement äquivalent als switch-statement formulieren lässt und andersherum. If-statements sind dabei in der Anwendung sicherlich häufiger anzutreffen als switch-statements. Generell macht es für das menschliche Codeverständnis Sinn, if-else-statements nicht zu komplex zu formulieren. Ist dies allerdings nötig, so können komplexe if-else-statements durch switch-statements vereinfacht werden.

18.2. Iterative Kontrollstrukturen

Iterative Kontrollstrukturen dienen dazu, bestimmte Befehle wiederholt auszuführen. Die grundlegende iterative Kontrollstruktur in **R** ist die **for**-Schleife. Diese hat die generelle Form

```
for (item in vector){  
  perform_action  
}
```

Dabei wird der Befehl `perform_action` für jedes `item` in `vector` einmal ausgeführt und der Wert von `item` jeweils aktualisiert. Folgendes Beispiel soll dies verdeutlichen:

```
for (i in 1:3){  
  print(i)  
}
```

```
[1] 1  
[1] 2  
[1] 3
```

Der Vektor besteht hier aus `1:3 = [1,2,3]`, der ausgeführte Befehl ist `print(i)` und die `items` sind die Einträge des Vektors. Allerdings muss der Befehl innerhalb der **for**-Schleife nicht zwangsläufig auf die `items` Bezug nehmen, wie folgendes Beispiel zeigt:

```
for (i in 1:3){  
  print('a')  
}
```

```
[1] "a"  
[1] "a"  
[1] "a"
```

Hier wird dreimal der konstante Wert `a` ausgegeben. Typischerweise wird innerhalb einer **for**-Schleife etwas erzeugt und gespeichert. Dabei sollte generell die entsprechende Speicherstruktur vorallokiert werden, da dann der entsprechende Platz im Arbeitsspeicher schon bereitgestellt ist und nicht während der Ausführung der **for**-Schleife bereitgestellt werden muss. Der Effizienzgewinn dieser Strategie mag in den meisten Fällen vernachlässigbar sein, allerdings erlaubt dieses Vorgehen auch insbesondere, Fehler leichter zu detektieren, wenn es beispielsweise Dimensionserwartungsunterschiede oder fehlende Einträge

gibt, die dann schon als NaN alloziert sind. Folgendes Beispiel demonstriert das iterative Auffüllen eines Vektors mit Mittelwerten.

```
ns      = 3                                # Anzahl an Simulationen
X_bar = rep(NaN, ns)                       # Speicherstruktur

# Simulationsiterationen
for(i in 1:ns){                            # Iterationsindices sind typischerweise i,j,k
  X      = rnorm(12)                       # Realisierung von 12 Z-Variablen
  X_bar[i] = mean(X)                       # Mittelwert der i-ten Realisierung
  print(X_bar)                             # Anzeige zur Demonstration
}
```

```
[1] -0.496586      NaN      NaN
[1] -0.49658598 -0.01176742      NaN
[1] -0.49658598 -0.01176742  0.49624963
```

Um linear entlang eines Vektors mit variablen Einträgen zu iterieren, bietet sich die Funktion `seq_along()` an. Diese erzeugt die linearen Indizes 1,2, 3, ... bis zur Länge des betreffenden Vektors, wie folgendes Beispiel demonstriert:

```
mu      = c(0,5,50)                       # Drei Erwartungswertparameter
X_bar = rep(NaN, ns)                       # Speicherstruktur

# Simulationsiterationen
for(i in seq_along(mu)){                   # Iterationsindices sind typischerweise i,j,k
  X      = rnorm(12, mu[i], 1)             # Realisierung von 12 N(\mu,1) Variablen
  X_bar[i] = mean(X)                       # Mittelwert der i-ten Realisierung
  print(X_bar)                             # Anzeige zur Demonstration
}
```

```
[1] 0.07846975      NaN      NaN
[1] 0.07846975 4.79444022      NaN
[1] 0.07846975 4.79444022 50.35093790
```

`for`-Schleifen können auch geschachtelt werden, also innerhalb von `for`-Schleifen weitere `for`-Schleifen ausgeführt werden. Dies mag etwas gewöhnungsbedürftig sein. Entscheidend ist dabei, dass man sich klarmacht, dass für *jede* Iteration der äußeren Schleife *alle* Iterationen der inneren Schleife durchlaufen werden. Es ist oft hilfreich, sich die generelle Logik einer geschachtelten Schleifenstruktur zunächst mithilfe von `print()`-statements klarzumachen, wie folgendes Beispiel demonstriert:

```
n_i = 2                                # Anzahl der Iterationen äußere Schleife
n_j = 3                                # Anzahl der Iterationen innere Schleife
for(i in 1:n_i){                       # Äußere Schleife
  for (j in 1:n_j){                     # Innere Schleife
    print(c(i,j))                       # Anzeigen der Iterationsindices [i,j]
  }
}
```

```
[1] 1 1
[1] 1 2
[1] 1 3
[1] 2 1
[1] 2 2
[1] 2 3
```

Folgendes Beispiel etwa erzeugt eine 4 x 5 Matrix, der Elemente den jeweiligen Spaltenindizes entsprechen. Dabei bedient die innere **for**-Schleife mit der Iteratorvariable **j** die Spaltenindizes und die äußere **for**-Schleife mit der Iteratorvariable **i** bedient die Zeilenindizes:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	2	3	4	5
[2,]	1	2	3	4	5
[3,]	1	2	3	4	5
[4,]	1	2	3	4	5

Analog erzeugt folgendes Beispiel eine 4 x 5 Matrix, der Elemente den jeweiligen Spaltenindizes entsprechen. Dabei bedient die innere **for**-Schleife mit der Iteratorvariable **j** wiederum die Spaltenindizes und die äußere **for**-Schleife mit der Iteratorvariable **i** bedient die Zeilenindizes:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	1	1	1	1
[2,]	2	2	2	2	2
[3,]	3	3	3	3	3
[4,]	4	4	4	4	4

In folgendem Beispiel schließlich wird mithilfe zweier geschachtelter **for**-Schleifen und eines **if**-statements eine 4 x 5 Matrix, deren Elemente gleich den Spaltenindizes sind, wenn der jeweilige Spaltenindex größer oder gleich dem Zeilenindex ist, und deren Elemente ansonsten Null sind erzeugt:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	2	3	4	5
[2,]	0	2	3	4	5
[3,]	0	0	3	4	5
[4,]	0	0	0	4	5

Teil III.

Wahrscheinlichkeitstheorie

Vorbemerkungen

Die Wahrscheinlichkeitstheorie ist ein mathematisches Modell zur Beschreibung von und zum quantitativen Schlussfolgern über *Zufallsvorgänge* der Wirklichkeit (Abbildung 18.1). Unter Zufallsvorgängen verstehen wir dabei alle Phänomene, die von uns nicht mit absoluter Sicherheit vorhergesagt werden können, deren Ergebnis also mit Unsicherheit behaftet ist. Offensichtliche und vertraute Beispiele für Zufallsvorgänge sind das Werfen eines Würfels oder einer Münze. Allerdings ist der Begriff des Zufallsvorgangs und damit der Anwendungsbereich der Wahrscheinlichkeitstheorie als sehr viel weiter gefasst zu verstehen. Nicht mit vollständiger Sicherheit vorhersagbar und damit mit Unsicherheit behaftet sind zum Beispiel auch der Ausgang einer Wahl, das morgige Wetter, der Messwert einer EEG-Elektrode zu einem bestimmten Zeitpunkt nach Applikation eines Reizes oder der Effekt einer Psychotherapieintervention auf den Gesundheitszustand einer Patient:in. Beginnt man darüber nachzudenken, welche Phänomene der Wirklichkeit mit Unsicherheit behaftet sind, so fällt es schwer, nichttriviale Phänomene anzugeben hinsichtlich deren Ergebnis man vollständige Sicherheit besitzt.

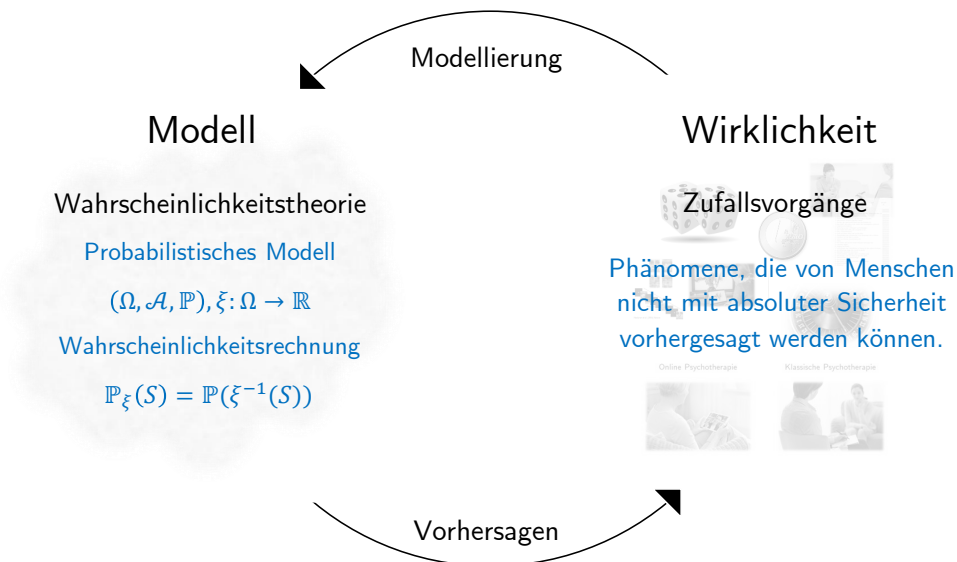


Abbildung 18.1. *Wahrscheinlichkeitstheorie als Modell von Zufallsvorgängen.* Ausgangspunkt der Wahrscheinlichkeitstheorie ist die Absicht, über einen Zufallsvorgang, also ein mit Unsicherheit behaftetes Phänomen der Wirklichkeit, logisch-quantitative Schlüsse zu ziehen. Die Repräsentation zentraler Aspekte des Zufallsvorgangs mithilfe wahrscheinlichkeitstheoretischer Begrifflichkeiten bezeichnet man als Modellierung. Das wahrscheinlichkeitstheoretische Modell selbst garantiert dann im Sinne der Wahrscheinlichkeitsrechnung die Korrektheit logisch-quantitativer Schlussfolgerungen, welche zur Vorhersage von Aspekten des Zufallsvorgangs genutzt werden können.

Als mathematisches Modell von Zufallsvorgängen erlaubt die Wahrscheinlichkeitstheorie insbesondere das vernunftbasierte, quantitative Schlussfolgern über Zufallsvorgänge. Dies schlägt sich primär in der sogenannten *Wahrscheinlichkeitsrechnung* nieder. Quantitative Schlussfolgerungen der Wahrscheinlichkeitsrechnung haben beispielsweise folgende Form: Wenn ich annehme, dass das Ereignis A mit Wahrscheinlichkeit x und das Ereignis B mit Wahrscheinlichkeit y eintritt, dann ergibt sich für die Wahrscheinlichkeit von Ereignis C eine Wahrscheinlichkeit von z . Dabei ist der Schluss auf die Wahrscheinlichkeit von C logisch-mathematisch abgesichert, in dem Sinne, wie zum Beispiel logisch-mathematisch

abgesichert ist, dass $1 + 1 = 2$ ist. Ob die Annahmen hinsichtlich der Wahrscheinlichkeiten von A und B aber den Gegebenheiten des Zufallsvorgangs in der Wirklichkeit entsprechen, darüber macht die Wahrscheinlichkeitstheorie keine Aussagen.

Die Wahrscheinlichkeitstheorie selbst bedient sich dabei der mathematischen Theorie der Mengen und Funktionen. Spätestens seit Kolmogoroff (1933) herrscht dabei ein axiomatischer Zugang vor: Man fragt in der Wahrscheinlichkeitstheorie selbst nicht, was denn eine Wahrscheinlichkeit sei oder inwieweit die Vorhersagen der Wahrscheinlichkeitstheorie mit der Wirklichkeit übereinstimmen, sondern versucht, ein in sich schlüssiges formal-mathematisches System von unbegründeten, aber intuitiv plausiblen, Grundannahmen und ihren Folgerungen zu entwickeln. Ausgangspunkt dieser Entwicklung ist das *Wahrscheinlichkeitsraummodell* eines Zufallsvorgangs, das wir in Kapitel 19 einführen werden. In der Tat gibt es neben dem formal-mathematischen System der Wahrscheinlichkeitstheorie bis heute mathematisch-philosophische Diskussionen darüber, was genau denn unter dem Begriff der “Wahrscheinlichkeit eines Ereignisses” zu verstehen ist (vgl. Hájek (2019)). Dabei sind grob gesagt zwei etwas gegensätzliche Interpretationen vorherrschend, die sogenannte *Frequentistische Interpretation* und die sogenannte *Bayesianische Interpretation*.

Nach der *Frequentistischen Interpretation* ist die Wahrscheinlichkeit eines Ereignisses die idealisierte relative Häufigkeit, mit der ein Ereignis unter den gleichen äußeren Bedingungen eintreten pflegt. Zum Beispiel ist die Frequentistische Interpretation der Aussage “Mit einer Wahrscheinlichkeit von $1/6$ zeigt der Würfel im nächsten Wurf eine 2” die folgende: “Wenn man einen Würfel unendlich oft werfen würde und dabei die relative Häufigkeit des Ereignisses, dass der Würfel eine 2 zeigt, bestimmen würde, dann wäre diese relative Häufigkeit gleich $1/6$ ”. Man beachte bei dieser Interpretation, dass man de-facto die Wahrscheinlichkeit eines Ereignisses nicht empirisch bestimmen kann, da man einen Würfel nicht unendlich oft werfen kann. Natürlich kann man die Wahrscheinlichkeit in dieser Interpretation aber empirisch schätzen. Schätzvorgänge selbst wiederum sind allerdings kein Teil der Wahrscheinlichkeitstheorie, sondern der *Frequentistischen* oder *Bayesianischen Inferenz*.

Nach der *Bayesianischen Interpretation* ist die Wahrscheinlichkeit eines Ereignisses der Grad der Sicherheit, den eine Beobachter:in aufgrund ihrer subjektiven Einschätzung der Lage dem Eintreten des Ereignisses zuschreibt. Zum Beispiel ist die Bayesianische Interpretation der Aussage “Mit einer Wahrscheinlichkeit von $1/6$ zeigt der Würfel im nächsten Wurf eine Zwei” dann etwa die folgende: “Basierend auf meiner eigenen und der tradierten Erfahrung mit dem Werfen eines Würfels bin ich mir zu 16.6% sicher, dass der Würfel beim nächsten Wurf eine Zwei zeigt.”

In Modellen von tatsächlich zumindest unter ähnlichen Umständen wiederholbaren Zufallsvorgängen wie dem Werfen eines Würfels ist der Unterschied zwischen Frequentistischer und Bayesianischer Interpretation oft eher subtil. Es gibt aber wie oben angedeutet viele Zufallsvorgänge, die mit Wahrscheinlichkeiten beschrieben werden können, bei denen aufgrund ihrer Einmaligkeit eine Frequentistische Interpretation nicht angemessen ist. Zum Beispiel machen Aussagen der Form “Die Wahrscheinlichkeit dafür, dass die weltweiten Hitzerekorde im Jahr 2023 nicht auf den Klimawandel zurückzuführen sind, ist kleiner als 0.01” (vgl. Philip et al. (2020)) nur unter der Bayesianischen Interpretation Sinn, da es sich bei den Wetteraufzeichnungen des Jahres 2023 um ein einmaliges, nicht wiederholbares Ereignis handelt.

Obwohl also die Interpretation des Begriffes der Wahrscheinlichkeit durchaus nicht eindeutig ist, unterscheiden sich die formalen Definitionen und Rechenregeln für Wahrscheinlich-

keiten nicht. Sowohl die Frequentistische als auch die Bayesianische Inferenz, auf die wir an späterer Stelle eingehen, haben mit der Wahrscheinlichkeitstheorie also ein identisches mathematisches Bezugssystem und gemeinsames Fundament. Im Wesentlichen unterscheidet sich dies nicht so sehr von vielen anderen Formen der mathematischen Modellierung. Beispielsweise kann die Ableitung einer Funktion als die Geschwindigkeit eines Objekts oder als die Wachstumsrate einer Bakterienpopulation interpretiert werden. Obwohl sich diese Phänomene intuitiv sehr stark in Bezug auf die Realität unterscheiden, ist das logische Schlussfolgern über beide Phänomene im Bereich der Mathematik identisch.

19. Wahrscheinlichkeitsräume

Ein *Wahrscheinlichkeitsraum* ist ein sehr allgemein gehaltenes formal-mathematisches Modell eines Zufallsvorgangs. Die zentrale Bedeutung dieses Modells für die Wahrscheinlichkeitstheorie und probabilistische Inferenz ergibt sich daraus, dass das Wahrscheinlichkeitsraummodell eine Anleitung dafür gibt, wie man beliebige Zufallsvorgänge, über die man quantitativ schlussfolgern möchte, in das formal-mathematische System der Wahrscheinlichkeitstheorie übersetzen kann. Gleichzeitig garantiert das Wahrscheinlichkeitsraummodell und die auf ihm aufgebauten Konzepte, dass die Mechanik der Wahrscheinlichkeitsrechnung zu logisch sinnvollen quantitativen Schlüssen über Zufallsvorgänge der Wirklichkeit führen. In diesem Kapitel führen den Begriff des Wahrscheinlichkeitsraums ein (Kapitel 19.1) und geben dann mithilfe von Wahrscheinlichkeitsfunktionen (Kapitel 19.2) erste Beispiele für die Modellierung von Zufallsvorgängen durch Wahrscheinlichkeitsräume (Kapitel 19.3).

19.1. Definition und erste Eigenschaften

Wir beginnen mit der Definition des Wahrscheinlichkeitsraummodells nach Kolmogoroff (1933), das wir dann nachfolgend in seinen Einzelteilen aus Frequentistischer Perspektive erläutern wollen.

Definition 19.1 (Wahrscheinlichkeitsraum). Ein *Wahrscheinlichkeitsraum* ist ein Triple $(\Omega, \mathcal{A}, \mathbb{P})$, wobei

- Ω eine beliebige nichtleere Menge von *Ergebnissen* ω ist und *Ergebnismenge* heißt,
- $\mathcal{A}$ eine Menge von Teilmengen von Ω mit den Eigenschaften

- $\Omega \in \mathcal{A}$,
- für alle $A \in \mathcal{A}$ gilt, dass auch $A^c := \Omega \setminus A \in \mathcal{A}$,
- aus $A_1, A_2, \dots \in \mathcal{A}$ folgt, dass auch $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

ist, σ -Algebra auf Ω genannt wird und *Ereignissystem* heißt,

- $\mathbb{P}$ eine Abbildung der Form $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$ mit den Eigenschaften

- $\mathbb{P}(A) \geq 0$ für alle $A \in \mathcal{A}$ (*Nicht-Negativität*),
- $\mathbb{P}(\Omega) = 1$ (*Normiertheit*),
- $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ für paarweise disjunkte $A_i \in \mathcal{A}$ (σ -Additivität)

ist und *Wahrscheinlichkeitsmaß* heißt.

•

Ergebnismenge und Mechanik

Wir beginnen mit Erläuterungen zum Begriff der *Ergebnismenge* Ω und der impliziten Mechanik des Wahrscheinlichkeitsraummodells. Um den Einstieg zu erleichtern betrachten wir im Folgenden zunächst vor allem *endliche Wahrscheinlichkeitsräume*, bei denen die Kardinalität von Ω nicht unendlich groß ist. Es sei also $|\Omega| < \infty$, Ω habe also nur endlich viele Elemente. Zum Modellieren des Werfen eines Würfels könnte man zum Beispiel $\Omega := \{1, 2, 3, 4, 5, 6\}$ definieren.

Hinter der formalen Definition des Wahrscheinlichkeitsraummodells stehen folgende Frequentistisch-geprägten Annahmen über seine Mechanik als Modell eines Zufallsvorgangs. Wir stellen uns zunächst sequentielle Durchgänge eines Zufallsvorgangs vor, also zum Beispiel das wiederholte Werfen eines Würfels. Nach Annahme des Wahrscheinlichkeitsraummodells wird in jedem dieser Durchgänge genau ein ω aus Ω *realisiert*, also als tatsächlich vorliegend ausgewählt. Wirft man zum Beispiel einen Würfel und fällt eine Zwei, so sagt man, dass eine Zwei realisiert wurde. Die Wahrscheinlichkeit, mit der ein ω aus Ω in einem Durchgang realisiert wird, wird durch den Wert $\mathbb{P}(\{\omega\}) \in [0, 1]$ beschrieben. Ist zum Beispiel $\mathbb{P}(\{\omega\}) = 1$, so wird dieses ω in jedem Durchgang des Zufallsvorgangs realisiert; ist $\mathbb{P}(\{\omega\}) = 0$, so wird dieses ω in keinem Durchgang des Zufallsvorgangs realisiert; und ist $\mathbb{P}(\{\omega\}) = 1/2$, so wird ω in etwa der Hälfte der Durchgänge des Zufallsvorgangs realisiert. Beim Modell des Werfens eines fairen Würfels nimmt man üblicherweise $\mathbb{P}(\{\omega\}) = 1/6$ für alle $\omega \in \Omega$ an. Hier könnte zum Beispiel im ersten Durchgang eine Vier realisiert werden, im zweiten Durchgang eine Eins, im dritten Durchgang eine Fünf, dann vielleicht wieder eine Vier und so weiter.

Ereignisse und Ereignissystem

Den Begriff des *Ereignisses* $A \in \mathcal{A}$ stellt man sich am besten als konzeptionelle Zusammenfassung ein oder mehrerer Ergebnisse vor. Beim Werfen eines Würfels sind mögliche Ereignisse zum Beispiel “Es fällt eine gerade Augenzahl”, das heißt $\omega \in \{2, 4, 6\}$; “Es fällt eine Augenzahl größer als Zwei”, das heißt $\omega \in \{3, 4, 5, 6\}$; oder etwa “Es fällt eine Eins oder eine Fünf”, das heißt $\omega \in \{1, 5\}$. Man beachte, dass zum Beispiel das Ereignis “Es fällt eine gerade Augenzahl” vor dem Hintergrund der Mechanik des Wahrscheinlichkeitsraums genau dann eintritt, wenn in einem Durchgang des Zufallsvorgangs *Werfen eines Würfels* das realisierte ω ein Element der Menge $\{2, 4, 6\}$ ist, wenn also zum Beispiel eine Vier fällt. Man mag das Eintreten des Ereignisses “Es fällt eine Augenzahl größer als Zwei” also auch lesen als “In einem Durchgang des Zufallsvorgangs *Werfen eines Würfels* wird ein Element von $\{3, 4, 5, 6\}$ realisiert”, d.h. konkret fällt entweder eine Drei, eine Vier, eine Fünf oder eine Sechs. Natürlich sind auch die Ergebnisse $\omega \in \Omega$ selbst mögliche Ereignisse, so dass zum Beispiel folgende Interpretationen gelten: Das Ereignis “Es fällt eine Eins” entspricht der Realisation $\omega \in \{1\}$ und das Ereignis “Es fällt eine Sechs” entspricht der Realisation $\omega \in \{6\}$. Betrachtet man in diesem Zusammenhang ein Ergebnis $\omega \in \Omega$ als Ereignis, so nennt man es *Elementarereignis* und schreibt es als einelementige Menge $\{\omega\}$.

Insgesamt entspricht dieses Vorgehen zur Beschreibung zufälliger Ereignisse dem inhärenten Ziel der Definition des Wahrscheinlichkeitsraums. Kolmogoroff (1933) schreibt dazu “Wir haben die eigentlichen Objekte unserer weiteren Betrachtungen - die zufälligen Ereignisse - als Mengen definiert.” Dies hat den Vorteil, dass Mengen mathematische Objekte sind, mit denen mathematisch gearbeitet werden kann und damit ein Aspekt der Wirklichkeit, ein “zufälliges Ereignis”, in den Modellbereich der Mathematik übersetzt wurde.

Alleiniger Sinn des *Ereignissystems* $\mathcal{A}$ ist es nun, alle Ereignisse, die sich basierend auf einer gegebenen Ergebnismenge bei Auswahl eines $\omega \in \Omega$ ergeben können, mathematisch zu repräsentieren. Es soll also keine Ereignisse in der Wirklichkeit geben, die nicht im Vorhinein im Wahrscheinlichkeitsraummodells des Zufallsvorgangs mitgedacht wurden. Wäre dies der Fall, so wäre das Modell defizitär, da es für bestimmte, in der Wirklichkeit eintretende Ereignisse keine Wahrscheinlichkeiten angeben könnte. Das Ereignissystem $\mathcal{A}$ soll also die vollständige Menge aller möglichen Ereignisse bei vorgegebenem Ω sein. Die Forderung, dass $\mathcal{A}$ zu diesem Zweck die sogenannten σ -Algebra Kriterien erfüllen soll, begründet sich dabei intuitiv wie folgt.

- Es soll zunächst einmal sichergestellt sein, dass $\omega \in \Omega$ für ein beliebiges ω , dass also irgendein Ergebnis realisiert wird, eines der möglichen Ereignisse ist. Dies entspricht der Eigenschaft $\Omega \in \mathcal{A}$.
- Zu jedem Ereignis soll es weiterhin auch möglich sein, dass dieses Ereignis gerade nicht eintritt. Dies entspricht der Eigenschaft, dass aus $A \in \mathcal{A}$ folgen soll, dass $A^c := \Omega \setminus A$ auch in $\mathcal{A}$ ist. Dies impliziert insbesondere auch, dass $\emptyset = \Omega \setminus \Omega \in \mathcal{A}$. Ein Ereignis ist also, dass kein Elementarereignis eintritt, allerdings passiert dies nur mit Wahrscheinlichkeit Null, $\mathbb{P}(\emptyset) = 0$. Es tritt also sicher immer zumindest ein Elementarereignis ein.
- Schließlich soll die Kombination von Ereignissen auch immer ein Ereignis sein. Bei der Modellierung des Werfen eines Würfels soll also zum Beispiel neben den Ereignissen “Es fällt eine gerade Zahl” und “Es fällt eine Zahl größer Zwei” auch das Ereignis “Es fällt eine gerade Zahl und/oder diese Zahl ist größer als Zwei” ein Ereignis sein. Dies entspricht, in allgemeiner Form, dass aus $A_1, A_2, \dots \in \mathcal{A}$ folgen soll, dass auch $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ ist.

Wenn auch die Begriffe des Ereignissystems und der σ -Algebra etwas abstrakt anmuten mögen, so stellt ihre Definition in der praktischen Modellierung von Zufallsvorgängen meist keine großen Herausforderung dar, da sowohl für endliche Ergebnismengen als auch für unendliche (abzählbare und überabzählbare) Ergebnismengen passende Ereignissysteme schon lange bekannt sind. So erfüllt bei Ergebnismengen mit endlicher Kardinalität die Potenzmenge der Ergebnismenge immer die Anforderungen eines Ereignissystems und kann immer zur Modellformulierung eines Zufallsvorgangs mit endlicher Ergebnismenge genutzt werden. Dies ist die Aussage folgenden Theorems.

Theorem 19.1 (Ereignissystem bei endlicher Ergebnismenge). $\Omega := \{\omega_1, \omega_2, \dots, \omega_n\}$ mit $n \in \mathbb{N}$ sei eine endliche Menge. Dann ist die Potenzmenge $\mathcal{P}(\Omega)$ von Ω eine σ -Algebra auf Ω und damit ein geeignetes Ereignissystem im Wahrscheinlichkeitsraummodell.

◦

Beweis. Die Potenzmenge von Ω ist die Menge aller Teilmengen von Ω . Wir überprüfen die σ -Algebra Eigenschaften. Zunächst gilt, dass Ω selbst eine der Teilmengen von Ω ist, also ist die erste σ -Algebra Eigenschaft erfüllt. Sei nun A eine Teilmenge von Ω . Dann ist auch $A^c = \Omega \setminus A$ eine Teilmenge von Ω und somit ist auch die zweite σ -Algebra Eigenschaft erfüllt. Schließlich betrachten wir die Vereinigung von k Teilmengen $A_1, A_2, \dots, A_k \subseteq \Omega$. Dann ist $\bigcup_{i=1}^k A_i$ die Menge der $\omega \in \Omega$ für die gilt, dass $\omega \in A_1$ und/oder $\omega \in A_2 \dots$ und/oder $\omega \in A_k$. Da für alle diese ω gilt, dass $\omega \in \Omega$ ist also auch $\bigcup_{i=1}^k A_i$ eine Teilmenge von Ω und damit auch die dritte σ -Algebra Eigenschaft erfüllt. Die Potenzmenge erfüllt also die geforderten Eigenschaften an ein Ereignissystem. $\square$

Bei überabzählbaren Ergebnismengen wie den reellen Zahlen $\mathbb{R}$ oder dem n -dimensionalen reellen Raum $\mathbb{R}^n$ ist die Konstruktion eines geeigneten Ereignissystems komplexer, so dass

wir in dieser Hinsicht für formale Entwicklungen auf die weiterführende Literatur verweisen wollen (z.B. Meintrup & Schäffler (2005), Schmidt (2009)). Mit der auf Borel (1898) zurückgehenden sogenannten *Borelschen σ -Algebra* ist jedoch ein Mengensystem bekannt, das den Anforderungen einer σ -Algebra auf überabzählbaren Ergebnismengen genügt. Wir bezeichnen die Borelsche σ -Algebra auf $\mathbb{R}$ mit $\mathcal{B}(\mathbb{R})$ und die Borelsche σ -Algebra auf $\mathbb{R}^n$ mit $\mathcal{B}(\mathbb{R}^n)$. Als Menge von Teilmengen von $\mathbb{R}$ bzw. $\mathbb{R}^n$ enthalten $\mathcal{B}(\mathbb{R})$ bzw. $\mathcal{B}(\mathbb{R}^n)$ alle Mengen, an denen man hinsichtlich ihrer durch $\mathbb{P}$ zugeordneten Wahrscheinlichkeit interessiert sein mag. Intuitiv mag man sich die Borelschen σ -Algebren $\mathcal{B}(\mathbb{R})$ und $\mathcal{B}(\mathbb{R}^n)$ also als die Potenzmengen von $\mathbb{R}$ bzw. $\mathbb{R}^n$ denken, auch wenn dies formal falsch ist. Tatsächlich enthält die Borelsche σ -Algebra nur Teilmengen, die durch abzählbare Mengenoperationen generiert werden, nicht aber durch überabzählbare.

Insgesamt ergibt sich also folgendes Vorgehen zur Auswahl von Ereignissystemen in Abhängigkeit von der Ergebnismenge Ω . Ist Ω endlich, so wählt man als Ereignissystem $\mathcal{A}$ die Potenzmenge $\mathcal{P}(\Omega)$ von Ω . Ist Ω gegeben durch $\mathbb{R}$, so wählt man als Ereignissystem $\mathcal{A}$ die Borelsche σ -Algebra $\mathcal{B}(\mathbb{R})$. Ist Ω schließlich gegeben durch $\mathbb{R}^n$, so wählt man für $\mathcal{A}$ die Borelsche σ -Algebra $\mathcal{B}(\mathbb{R}^n)$. In Spezialfällen und für sehr spezielle Ergebnismengen Ω mag man von diesem Vorgehen abweichen wollen, allgemein decken die drei betrachteten Fälle jedoch die meisten Anwendungen ab.

Wahrscheinlichkeitsmaß $\mathbb{P}$

Mit Ω und $\mathcal{A}$, die in Tupleform $(\Omega, \mathcal{A})$ auch als *Messraum* bezeichnet werden, haben wir bisher die *strukturelle Basis* eines Wahrscheinlichkeitsraummodells genauer betrachtet. Viele Wahrscheinlichkeitsräume, zum Beispiel für Zufallsvorgänge die reellen Zahlen betreffend, sind hinsichtlich ihres Messraums identisch. Das *Wahrscheinlichkeitsmaß* $\mathbb{P}$ nun repräsentiert die probabilistischen Charakteristika eines Wahrscheinlichkeitsraummodells und formt damit die *funktionelle Basis* eines Wahrscheinlichkeitsraummodells. Wir werden im Folgenden, insbesondere nach Einführung der Begriffe der Zufallsvariablen und Zufallsvektoren, sehr viele verschiedene Wahrscheinlichkeitsmaße kennenlernen. An dieser Stelle wollen wir zunächst nur allgemeine Eigenschaften von Wahrscheinlichkeitsmaßen betrachten.

Mit der Definition

$$\mathbb{P} : \mathcal{A} \rightarrow [0, 1], A \mapsto \mathbb{P}(A) \quad (19.1)$$

gilt zunächst einmal, dass ein Wahrscheinlichkeitsmaß auf einer Menge von Mengen definiert ist und den Elementen dieser Menge, also den Mengen $A \in \mathcal{A}$, Wahrscheinlichkeiten, also Werte im Intervall $[0, 1]$, zuordnet. Natürlich gilt mit $\{\omega\} \in \mathcal{A}$ für alle $\omega \in \Omega$, dass $\mathbb{P}$ auch den Elementareignissen Wahrscheinlichkeiten zuordnet, aber eben nicht nur. Wir betonen auch, dass nach Definition die Wahrscheinlichkeit $\mathbb{P}(A)$ eines Ereignisses $A \in \mathcal{A}$ eine Zahl im Intervall $[0, 1]$ ist und nicht etwa eine Prozentzahl oder ein Verhältnis. Wir wollen nachfolgend die definierenden Eigenschaften der *Nicht-Negativität*, der *Normiertheit* und der σ -*Additivität* von $\mathbb{P}$ näher beleuchten.

Die *Nicht-Negativität* $\mathbb{P}(A) \geq 0$ für alle $A \in \mathcal{A}$ ist natürlich in der Definition $[0, 1]$ der Zielmenge von $\mathbb{P}$ implizit. Tatsächlich ist die Abbildungsform von $\mathbb{P}$ eine von uns vorgenommene Ergänzung der Formulierung von Kolmogoroff (1933), die der Klarheit dienen soll. Formal folgt die Form der Zielmenge von $\mathbb{P}$ eigentlich aus den definierenden Eigenschaften der Nicht-Negativität, Normiertheit und σ -Additivität von $\mathbb{P}$.

Die *Normiertheit* $\mathbb{P}(\Omega) = 1$ entspricht der Tatsache, dass in jedem Durchgang eines Zufallsvorgangs sicher gilt, dass ein realisiertes ω ein Element von Ω ist. In jedem Durchgang eines Zufallsvorgangs tritt also ein Elementarereignis ein und, je nach Beschaffenheit des Messraums, noch viele weitere. Beim Modell des Werfen eines Würfels gilt also, dass das in einem Durchgang des Zufallsvorgangs realisierte Ergebnis/Elementarereignis mit Wahrscheinlichkeit 1 ein Element von $\Omega := \{1, 2, 3, 4, 5, 6\}$ ist. Ist das realisierte Ergebnis zum Beispiel eine Eins, so treten neben dem Ereignis “Es fällt eine Eins” auch noch die Ereignisse “Eine ungerade Zahl fällt”, “Eine Zahl kleiner als Drei fällt”, “Eine ungerade Zahl kleiner als Drei fällt” und viele weitere ein.

Die σ -Additivität des Wahrscheinlichkeitsmaßes $\mathbb{P}$ schließlich bildet das Fundament der *Wahrscheinlichkeitsrechnung*, also die Grundlage für das Rechnen mit Wahrscheinlichkeiten. Die σ -Additivität von $\mathbb{P}$ erlaubt es nämlich, aus bereits bekannten Ereigniswahrscheinlichkeiten die Wahrscheinlichkeiten anderer Ereignisse zu berechnen. Man kann damit basierend auf einer Definition von Ω , $\mathcal{A}$ und $\mathbb{P}$ also Wahrscheinlichkeiten für alle möglichen Ereignisse eines Wahrscheinlichkeitsraummodells berechnen. Ob diese Wahrscheinlichkeiten nun aber tatsächlich etwas mit den realen Ereignissen bezüglich eines Zufallsvorgangs der Wirklichkeit zu tun haben, kommt darauf an, ob die Modellierung einigermaßen gelungen ist oder nicht. Dabei werden berechnete Wahrscheinlichkeiten aber zumindest rational, also nach den Regeln der Vernunft, d.h. der Logik und der Mathematik, bestimmt. Insgesamt erlaubt das Wahrscheinlichkeitsmodell damit das schlussfolgernde Nachdenken über mit Unsicherheit behaftete Phänomene.

Wir wollen abschließend das auf der σ -Additivität von $\mathbb{P}$ beruhende Rechnen mit Wahrscheinlichkeiten noch an zwei grundlegenden Beispielen verdeutlichen. Das erste Beispiel besagt, dass die Wahrscheinlichkeit dafür, dass das in einem Durchgang eines Zufallsvorgangs realisierte ω kein Element der Ergebnismenge ist, gleich Null ist.

Theorem 19.2 (Wahrscheinlichkeit des unmöglichen Ereignisses). *$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum. Dann gilt*

$$\mathbb{P}(\emptyset) = 0. \quad (19.2)$$

◦

Beweis. Für $i = 1, 2, \dots$ sei $A_i := \emptyset$. Dann ist $A_1, A_2, \dots$ eine Folge disjunkter Ereignisse, weil gilt, dass $\emptyset \cap \emptyset = \emptyset$ und es ist $\cup_{i=1}^{\infty} A_i = \emptyset$. Mit der σ -Additivität von $\mathbb{P}$ folgt dann, dass

$$\mathbb{P}(\emptyset) = \mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} \mathbb{P}(\emptyset). \quad (19.3)$$

Das unendliche Aufaddieren der Zahl $\mathbb{P}(\emptyset) \in [0, 1]$ soll also wieder $\mathbb{P}(\emptyset)$ ergeben. Dies ist aber nur möglich, wenn $\mathbb{P}(\emptyset) = 0$. $\square$

Man beachte, dass hier intuitiv natürlich eine mögliche Unzulänglichkeit des Wahrscheinlichkeitsraums als Modell für Zufallsvorgänge in der Wirklichkeit auftritt: Fällt beim Würfelspiel der Würfel zum Beispiel unerreichbar unter das Sofa, so ist ein Elementarereignis $\omega \notin \Omega$ eingetreten, obwohl seine modellierte Wahrscheinlichkeit gleich Null ist.

Als zweites Beispiel wollen wir zeigen, dass die σ -Additivität, die in der Definition des Wahrscheinlichkeitsraums (nur) für die Vereinigung unendlich vieler disjunkter Ereignisse definiert ist, die σ -Additivität endlich vieler disjunkter Ereignisse, wie sie in in der Anwendung oft vorkommen, impliziert.

Theorem 19.3 (σ -Additivität bei endlichen Folgen disjunkter Ereignisse). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $A_1, \dots, A_n$ sei eine endliche Folge paarweise disjunkter Ereignisse. Dann gilt

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i). \quad (19.4)$$

◦

Beweis. Wir betrachten eine unendliche Folge von paarweise disjunkten Ereignissen $A_1, A_2, \dots$ wobei für ein $n \in \mathbb{N}$ gelten soll, dass $A_i := \emptyset$ für $i > n$. Dann gilt mit der σ -Additivität von $\mathbb{P}$ zunächst, dass

$$\mathbb{P}(\cup_{i=1}^n A_i) = \mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^n \mathbb{P}(A_i) + \sum_{i=n+1}^{\infty} \mathbb{P}(A_i). \quad (19.5)$$

Mit $\mathbb{P}(A_i) = \mathbb{P}(\emptyset) = 0$ für $i = n+1, n+2, \dots$ folgt dann direkt

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i) + 0 = \sum_{i=1}^n \mathbb{P}(A_i). \quad (19.6)$$

□

19.2. Wahrscheinlichkeitsfunktionen

In diesem Abschnitt wollen wir mit den *Wahrscheinlichkeitsfunktionen* eine erste Möglichkeit kennenlernen, für Wahrscheinlichkeitsräume mit endlicher Ergebnismenge Wahrscheinlichkeitsmaße festzulegen. In Kapitel 19.3 nutzen wir dieses Hilfsmittel intensiv, um erste Beispiele für die Modellierung von Zufallsvorgängen mithilfe von Wahrscheinlichkeitsräumen geben zu können. In Kapitel 21 und Kapitel 22 werden wir Wahrscheinlichkeitsfunktionen unter der Bezeichnung *Wahrscheinlichkeitsmassefunktionen* erneut begegnen. Wir definieren hier den Begriff der Wahrscheinlichkeitsfunktion zunächst wie folgt.

Definition 19.2 (Wahrscheinlichkeitsfunktion). Ω sei eine endliche Menge. Dann heißt eine Funktion $\pi : \Omega \rightarrow [0, 1]$ *Wahrscheinlichkeitsfunktion*, wenn gilt, dass

$$\sum_{\omega \in \Omega} \pi(\omega) = 1. \quad (19.7)$$

Sei weiterhin $\mathbb{P}$ ein Wahrscheinlichkeitsmaß. Dann heißt die durch

$$\pi : \Omega \rightarrow [0, 1], \omega \mapsto \pi(\omega) := \mathbb{P}(\{\omega\}) \quad (19.8)$$

definierte Funktion *Wahrscheinlichkeitsfunktion* von $\mathbb{P}$ auf Ω .

•

Wir merken an, dass weil $\mathbb{P}$ per Definition σ -additiv ist, insbesondere auch gilt, dass

$$\mathbb{P}(\Omega) = \mathbb{P}(\cup_{\omega \in \Omega} \{\omega\}) = \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = \sum_{\omega \in \Omega} \pi(\omega) = 1. \quad (19.9)$$

Zur Konstruktion von Wahrscheinlichkeitsmaßen durch Wahrscheinlichkeitsfunktionen stellt folgendes Theorem die formale Basis bereit. Es besagt insbesondere, dass bei endlichem Ω die Wahrscheinlichkeiten *aller* möglichen Ereignisse aus den Wahrscheinlichkeiten der Elementarereignisse $\pi(\omega)$ berechnet werden können.

Theorem 19.4. $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum mit endlicher Ergebnismenge und $\pi : \Omega \rightarrow [0, 1]$ sei eine Wahrscheinlichkeitsfunktion. Dann existiert ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf Ω mit π als Wahrscheinlichkeitsfunktion von $\mathbb{P}$. Dieses Wahrscheinlichkeitsmaß ist definiert als

$$\mathbb{P} : \mathcal{A} \rightarrow [0, 1], A \mapsto \mathbb{P}(A) := \sum_{\omega \in A} \pi(\omega). \quad (19.10)$$

◦

Beweis. Wir überprüfen zunächst die Wahrscheinlichkeitsmaßeigenschaften von $\mathbb{P}$. Weil $\pi(\omega) \in [0, 1]$ für alle $\omega \in \Omega$, gilt auch immer $\sum_{\omega \in A} \pi(\omega) \geq 0$ und damit die Nicht-Negativität von $\mathbb{P}$. Ferner folgt wie oben gesehen mit der Normiertheit von π direkt die Normiertheit von $\mathbb{P}$. Seien nun $A_1, A_2, \dots \in \mathcal{A}$. Dann gilt

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{\omega \in \cup_{i=1}^{\infty} A_i} \pi(\omega) = \sum_{i=1}^{\infty} \sum_{\omega \in A_i} \pi(\omega) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \quad (19.11)$$

und damit die σ -Additivität von $\mathbb{P}$. □

Definiert man also für eine gegebene Ergebnismenge Ω eine Funktion $\pi : \Omega \rightarrow [0, 1]$ und stellt sicher, dass sich die Funktionswerte $\pi(\omega)$ über alle $\omega \in \Omega$ hinweg zu 1 summieren und interpretiert den einzelnen Funktionswert $\pi(\omega)$ dann als die Wahrscheinlichkeit $\mathbb{P}(\{\omega\})$ des Elementarereignisses $\{\omega\} \in \mathcal{A}$, so hat man ein Wahrscheinlichkeitsmaß konstruiert.

19.3. Beispiele bei endlichem Ergebnisraum

Aus dem bis hierin Gesagtem lässt sich nun zusammenfassend folgendes Vorgehen zur Modellierung eines Zufallsvorganges mithilfe eines Wahrscheinlichkeitsraums $(\Omega, \mathcal{A}, \mathbb{P})$ festhalten:

- (1) In einem ersten Schritt überlegt man sich eine sinnvolle Definition der Ergebnismenge Ω , also der Ergebnisse bzw. Elementarereignisse, die in jedem Durchgang des Zufallsvorgangs realisiert werden sollen.
- (2) In einem zweiten Schritt wählt man dann ein geeignetes Ereignissystem; im Falle einer endlichen Ergebnismenge bietet sich die Potenzmenge $\mathcal{P}(\Omega)$, im Falle der überabzählbaren Ergebnismenge $\Omega := \mathbb{R}$ bietet sich die Borelsche σ -Algebra $\mathcal{B}(\mathbb{R})$ an.
- (3) Schließlich definiert man ein Wahrscheinlichkeitsmaß $\mathbb{P}$, das die Wahrscheinlichkeiten für das Auftreten aller möglichen Ereignisse repräsentiert. Im Falle einer endlichen Ergebnismenge gelingt dies insbesondere durch Definition der Wahrscheinlichkeiten der Elementarereignisse. In der Folge verdeutlichen wir dieses Vorgehen anhand von Beispielen. Für eine überabzählbare Ergebnismenge $\Omega := \mathbb{R}$ bietet sich die Definition von $\mathbb{P}$ mithilfe von *Wahrscheinlichkeitsdichtefunktionen* an, wie wir später sehen werden.

Würfeln mit einem Würfel

Wir modellieren das Würfeln mit einem Würfel. Es ist sicherlich sinnvoll, die Ergebnismenge als $\Omega := \{1, 2, 3, 4, 5, 6\}$ zu definieren. Allerdings wäre auch die Definition von $\Omega := \{\cdot, \cdot\cdot, \cdot\cdot\cdot, \cdot\cdot\cdot\cdot, \cdot\cdot\cdot\cdot\cdot, \cdot\cdot\cdot\cdot\cdot\cdot\}$ in äquivalenter Weise möglich.

Da es sich um eine endliche Ergebnismenge handelt, wählen wir als σ -Algebra $\mathcal{A}$ die Potenzmenge $\mathcal{P}(\Omega)$. $\mathcal{A}$ enthält dann automatisch alle möglichen Ereignisse. Die Kardinalität von $\mathcal{A} := \mathcal{P}(\Omega)$ ist $|\mathcal{P}(\Omega)| = 2^{|\Omega|} = 2^6 = 64$. In Tabelle 19.1 listen wir sechs dieser 64 Ereignisse in ihrer verbalen Beschreibung und als Teilmenge A von Ω auf.

Tabelle 19.1. Ausgewählte Ereignisse beim Modell des Werfen eines Würfels

Beschreibung	Mengenform
Es fällt eine beliebige Augenzahl	$\omega \in A = \Omega$
Keine Augenzahl fällt	$\omega \in A = \emptyset$
Es fällt eine Augenzahl größer als 4	$\omega \in A = \{5, 6\}$
Es fällt eine gerade Augenzahl	$\omega \in A = \{2, 4, 6\}$
Es fällt eine Sechs	$\omega \in A = \{6\}$
Eine Eins, eine Drei oder eine Sechs fällt	$\omega \in A = \{1, 3, 6\}$

Damit ist die Definition des Messraum $(\Omega, \mathcal{A})$ in der Modellierung des Werfens eines Würfels abgeschlossen.

Wie in Kapitel 19.2 beschrieben, kann das Wahrscheinlichkeitsmaß $\mathbb{P}$ durch Festlegung von $\mathbb{P}(\{\omega\})$ für alle $\omega \in \Omega$ festgelegt werden. Für das Modell eines unverfälschten Würfels würde man

$$\mathbb{P}(\{\omega\}) := \frac{1}{|\Omega|} := 1/6 \text{ für alle } \omega \in \Omega \quad (19.12)$$

wählen. Für ein Modell eines verfälschten Würfels, der das Werfen einer Sechs bevorzugt, könnte man zum Beispiel definieren, dass

$$\mathbb{P}(\{\omega\}) := \frac{1}{8} \text{ für } \omega \in \{1, 2, 3, 4, 5\} \text{ und } \mathbb{P}(\{6\}) := \frac{3}{8}. \quad (19.13)$$

Im Fall des unverfälschten Würfel ergibt sich beispielsweise die Wahrscheinlichkeit für das Ereignis “Es fällt eine gerade Augenzahl” mit der σ -Additivät von $\mathbb{P}$ dann zu

$$\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(\{2\} \cup \{4\} \cup \{6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}. \quad (19.14)$$

Im Fall des obigen Modells eines verfälschten Würfels ergibt sich für das gleiche Ereignis die Wahrscheinlichkeit zu

$$\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(\{2\} \cup \{4\} \cup \{6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{8} + \frac{1}{8} + \frac{3}{8} = \frac{5}{8}. \quad (19.15)$$

Das betrachtete Ereignis hat im Modell des verfälschten Würfels eine höhere Wahrscheinlichkeit als im Modell des unverfälschten Würfels, was intuitiv sinnvoll ist, da die Sechs eine gerade Zahl ist.

Gleichzeitiges Würfeln mit einem blauen und einem roten Würfel

Wir wollen nun das gleichzeitige Werfen eines blauen und eines roten Würfels modellieren. Dazu ist es sinnvoll, die Ergebnismenge als

$$\Omega := \{(r, b) | r \in \{1, 2, 3, 4, 5, 6\}, b \in \{1, 2, 3, 4, 5, 6\}\} \quad (19.16)$$

mit Kardinalität $|\Omega| = 36$ zu definieren, wobei r die Augenzahl des blauen Würfels und b die Augenzahl des roten Würfels repräsentieren soll.

Wiederum bietet sich die Wahl der Potenzmenge von Ω als σ -Algebra an, wir definieren also wieder $\mathcal{A} := \mathcal{P}(\Omega)$. Die Anzahl der in diesem Modell möglichen Ereignisse ergibt sich zu $|\mathcal{A}| = 2^{|\Omega|} = 2^{36} = 68.719.476.736$. In Tabelle 19.2 listen wir sechs dieser Ereignisse in ihrer verbalen Beschreibung und als Teilmenge A von Ω auf.

Tabelle 19.2. Ausgewählte Ereignisse beim Modell des Werfens eines roten und eines blauen Würfels

Beschreibung	Mengenform
Auf dem roten Würfel fällt eine Drei	$\omega \in A = \{(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)\}$
Auf dem blauen Würfel fällt eine Drei	$\omega \in A = \{(1, 3), (2, 3), (3, 3), (4, 3), (5, 3), (6, 3)\}$
Auf beiden Würfeln fällt eine Drei	$\omega \in A = \{(3, 3)\}$
Es fällt eine Pasch	$\omega \in A = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$
Die Summe der gefallen Zahlen ist Vier	$\omega \in A = \{(1, 3), (3, 1), (2, 2)\}$

Die Definition des Messraum $(\Omega, \mathcal{A})$ ist damit abgeschlossen. Ein Wahrscheinlichkeitsmaß $\mathbb{P}$ kann wiederum durch Definition von $\mathbb{P}(\{\omega\})$ für alle $\omega \in \Omega$ festgelegt werden. Für das Modell zweier unverfälschter Würfel würde man

$$\mathbb{P}(\{\omega\}) := \frac{1}{|\Omega|} = \frac{1}{36} \text{ für alle } \omega \in \Omega \quad (19.17)$$

wählen. Unter diesem Wahrscheinlichkeitsmaß ergibt sich dann beispielsweise die Wahrscheinlichkeit für das Ereignis “Die Summe der gefallen Zahlen ist Vier” mit der σ -Additivät von $\mathbb{P}$ zu

$$\begin{aligned} \mathbb{P}(\{(1, 3), (3, 1), (2, 2)\}) &= \mathbb{P}(\{(1, 3)\} \cup \{(3, 1)\} \cup \{(2, 2)\}) \\ &= \mathbb{P}(\{(1, 3)\}) + \mathbb{P}(\{(3, 1)\}) + \mathbb{P}(\{(2, 2)\}) \\ &= 1/36 + 1/36 + 1/36 \\ &= 1/12. \end{aligned}$$

Werfen einer Münze

Wir modellieren das Werfen einer Münze, deren eine Seite Kopf (Heads) und deren andere Seite Zahl (Tails) zeigt. Es ist sinnvoll, die Ergebnismenge als $\Omega := \{H, T\}$ zu definieren, wobei H “Heads” und T “Tails” repräsentiert. Allerdings wäre auch jede andere binäre Definition von Ω möglich, z.B. $\Omega := \{0, 1\}$, $\Omega := \{-1, 1\}$ oder $\Omega := \{1, 2\}$.

Die Potenzmenge $\mathcal{A} := \mathcal{P}(\Omega)$ enthält alle möglichen Ereignisse. In diesem Fall können wir das gesamte Mengensystem $\mathcal{A}$ leicht, wie in Tabelle 19.3 gezeigt, komplett auflisten.

Tabelle 19.3. Ereignissystem $\mathcal{A}$ beim Modell des Werfens einer Münze

Beschreibung	Mengenform
Weder Kopf noch Zahl fällt	$\omega \in A = \emptyset$
Kopf fällt	$\omega \in A = \{H\}$
Zahl fällt	$\omega \in A = \{T\}$
Kopf oder Zahl fällt	$\omega \in A = \{H, T\}$

Die Definition des Messraums $(\Omega, \mathcal{A})$ ist damit abgeschlossen.

Ein Wahrscheinlichkeitsmaß $\mathbb{P}$ kann wiederum durch Definition von $\mathbb{P}(\{\omega\})$ für alle $\omega \in \Omega$ festgelegt werden. Die Normiertheit von Ω bedingt hier insbesondere, dass

$$\mathbb{P}(\Omega) = 1 \Leftrightarrow \mathbb{P}(\{H\}) + \mathbb{P}(\{T\}) = 1 \Leftrightarrow \mathbb{P}(\{T\}) = 1 - \mathbb{P}(\{H\}). \quad (19.18)$$

Bei Festlegung der Wahrscheinlichkeit des Elementarereignisses $\{H\}$ wird also die Wahrscheinlichkeit des Elementarereignis $\{T\}$ sofort mit festgelegt, andersherum gilt dies natürlich ebenso. Für das Modell einer fairen Münze würde man $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) := 1/2$ wählen. Die Wahrscheinlichkeiten aller möglichen Ereignisse ergeben sich in diesem Fall zu

$$\mathbb{P}(\emptyset) = 0, \mathbb{P}(\{H\}) = 1/2, \mathbb{P}(\{T\}) = 1/2 \text{ und } \mathbb{P}(\{H, T\}) = 1. \quad (19.19)$$

Zweifaches Werfen einer Münze

Wir modellieren das zweifache Werfen einer Münzen. Basierend auf dem Modell des einfachen Münzwurfs ist es sinnvoll, die Ergebnismenge als $\Omega := \{HH, HT, TH, TT\}$ zu definieren. Die Potenzmenge $\mathcal{A} := \mathcal{P}(\Omega)$ enthält wiederum alle $2^{|\Omega|} = 2^4 = 16$ möglichen Ereignisse. In untenstehender Tabelle listen wir vier davon.

Tabelle 19.4. Ausgewählte Ereignisse beim Modell des zweifachen Werfens einer Münze

Beschreibung	Mengenform
Kopf fällt im ersten Wurf	$\omega \in A = \{HH, HT\}$
Kopf fällt im zweiten Wurf	$\omega \in A = \{HH, TH\}$
Kopf fällt nicht	$\omega \in A = \{TT\}$
Zahl fällt mindestens einmal	$\omega \in A = \{HT, TH, TT\}$

Die Definition des Messraum $(\Omega, \mathcal{A})$ ist damit abgeschlossen. Ein Wahrscheinlichkeitsmaß $\mathbb{P}$ kann wiederum durch Definition von $\mathbb{P}(\{\omega\})$ für alle $\omega \in \Omega$ festgelegt werden. Für das Modell des zweifachen Werfens einer fairen Münze würde man

$$\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) := \frac{1}{4} \quad (19.20)$$

definieren.

19.4. Literaturhinweise

Die Monographie “Grundbegriffe der Wahrscheinlichkeitsrechnung” von Andrey Kolmogorov (Kolmogoroff (1933)) symbolisiert die Grundlage der modernen Wahrscheinlichkeitsrechnung. Neben der hier diskutierten axiomatischen Einführung des Wahrscheinlichkeitsraummodells betrachtet Kolmogoroff (1933) noch viele weitere Aspekte der Wahrscheinlichkeitrechnung und bietet so einen gut lesbaren Einstieg in das gesamte Gebiet der Wahrscheinlichkeitstheorie. Natürlich ist der von Kolmogoroff (1933) formulierte Zugang nur ein vorläufiges Endprodukt der langen geschichtlichen Entwicklung der Wahrscheinlichkeitstheorie. Schließlich ist die Entwicklung der mathematischen Modellierung von Zufallsvorgängen auch mit Kolmogoroff (1933) keinesfalls an einem Ende angelangt. Spätere Arbeiten im 20. Jahrhundert betrafen insbesondere die Interpretation des Wahrscheinlichkeitsbegriffs (vgl. De Finetti (1975)) oder führen verallgemeinerte quantitative Maße subjektiver Unsicherheit ein (vgl. Walley (1991)). Einen aktuellen Überblick zur Interpretation des Wahrscheinlichkeitsbegriffs und seiner formalen Grundlagen gibt Hájek (2019).

20. Wahrscheinlichkeiten

In diesem Abschnitt führen wir mit den Begriffen der *gemeinsamen Wahrscheinlichkeit* zweier Ereignisse und der *bedingten Wahrscheinlichkeit* eines Ereignisses zwei elementare Formen von Wahrscheinlichkeiten ein. Intuitiv bezieht sich der Begriff der gemeinsamen Wahrscheinlichkeit auf die Wahrscheinlichkeit des “gleichzeitigen” Eintretens zweier Ereignisse A und B und der Begriff der bedingten Wahrscheinlichkeit auf die Wahrscheinlichkeit des Eintretens eines Ereignisses A , “wenn man um das Eintreten eines anderen Ereignisses B weiß”. Ist es für die Wahrscheinlichkeit eines Ereignisses A unerheblich, ob ein Ereignis B eingetreten ist oder nicht, so nennt man A und B *unabhängige Ereignisse*. Intuitiv modellieren unabhängige Ereignisse die Abwesenheit gegenseitiger Einflüsse.

20.1. Gemeinsame Wahrscheinlichkeiten

Der Begriff der gemeinsamen Wahrscheinlichkeit zweier Ereignisse ist wie folgt definiert.

Definition 20.1 (Gemeinsame Wahrscheinlichkeit). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und es seien $A, B \in \mathcal{A}$. Dann heißt

$$\mathbb{P}(A \cap B) \tag{20.1}$$

die *gemeinsame Wahrscheinlichkeit von A und B* .

•

Wie oben angemerkt entspricht $\mathbb{P}(A \cap B)$ der Wahrscheinlichkeit dafür, dass die Ereignisse A und B “gleichzeitig” eintreten. Dies verdeutlicht man sich am besten vor dem Hintergrund der Mechanik des Wahrscheinlichkeitsraummodells. Danach ist $\mathbb{P}(A \cap B)$ eben die Wahrscheinlichkeit, dass in einem Durchgang eines Zufallsvorgangs ein ω realisiert wird, für das sowohl $\omega \in A$ als auch $\omega \in B$ gelten.

Beispiel

Als erstes Beispiel für eine gemeinsame Wahrscheinlichkeit zweier Ereignisse wollen wir wieder das Wahrscheinlichkeitsraummodells des Werfens eines fairen Würfels betrachten. Seien dazu A das Ereignis “Es fällt eine gerade Augenzahl”, also $A := \{2, 4, 6\}$ und B das Ereignis “Es fällt eine Augenzahl größer als Drei”, also $B := \{4, 5, 6\}$. Mengentheoretisch gilt dann

$$A \cap B = \{2, 4, 6\} \cap \{4, 5, 6\} = \{4, 6\}. \tag{20.2}$$

Die Interpretation von $A \cap B = \{4, 6\}$ ist dabei gerade “Es fällt eine gerade Augenzahl und diese Augenzahl ist größer als Drei”. Bei Annahme der Fairness des Würfels, also für

$\mathbb{P}(\{4\}) = \mathbb{P}(\{6\}) := 1/6$ können wir mithilfe der σ -Additivität von $\mathbb{P}$ die Wahrscheinlichkeit dieses Ereignisses leicht berechnen. Es ergibt sich

$$\begin{aligned}
 \mathbb{P}(A \cap B) &= \mathbb{P}(\{2, 4, 6\} \cap \{4, 5, 6\}) \\
 &= \mathbb{P}(\{4, 6\}) \\
 &= \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) \\
 &= \frac{1}{6} + \frac{1}{6} \\
 &= \frac{1}{3}.
 \end{aligned} \tag{20.3}$$

Beim Nachdenken über gemeinsame Wahrscheinlichkeiten ist es natürlich wichtig, die gemeinsame Wahrscheinlichkeit $\mathbb{P}(A \cap B)$ nicht mit der Wahrscheinlichkeit $\mathbb{P}(A \cup B)$ des Ereignisses $A \cup B$ zur verwechseln. Es sei daran erinnert, dass die Vereinigung zweier Mengen $\cup$ dem *inklusive* oder, also einem *und/oder* entspricht (vgl. Kapitel 1.3 und Kapitel 2.2). Das Ereignis $A \cup B$ entspricht also dem Ereignis, dass das Ereignis A und/oder das Ereignis B eintritt. Insbesondere ist $\omega \in A \cup B$ also auch schon dann erfüllt, wenn für das Ergebnis eines Durchgang eines Zufallsvorgangs *nur* $\omega \in A$ oder *nur* $\omega \in B$ gilt. Konkret ergibt sich etwa für die Ereignisse $A := \{2, 4, 6\}$ und $B := \{4, 5, 6\}$ aus obigem Würfelbeispiel

$$A \cup B = \{2, 4, 6\} \cup \{4, 5, 6\} = \{2, 4, 5, 6\} \tag{20.4}$$

mit der Interpretation “Es fällt eine gerade Augenzahl und/oder es fällt eine Augenzahl größer als Drei”. Für die entsprechende Wahrscheinlichkeit ergibt sich

$$\mathbb{P}(\{2, 4, 5, 6\}) = \frac{2}{3}, \tag{20.5}$$

so dass in diesem Fall offenbar $\mathbb{P}(A \cap B) \neq \mathbb{P}(A \cup B)$ gilt.

Mithilfe folgendem Theorems wollen wir in diesem Abschnitt schließlich noch einige nützliche Eigenschaften zum Rechnen mit Wahrscheinlichkeiten festhalten, die sich direkt aus der Verbindung von Mengenverknüpfungen und der σ -Additivität von Wahrscheinlichkeitsmaßen ergeben.

Theorem 20.1 (Weitere Eigenschaften von Wahrscheinlichkeiten). *($\Omega, \mathcal{A}, \mathbb{P}$) sei ein Wahrscheinlichkeitsraum und es seien $A, B \in \mathcal{A}$ Ereignisse. Dann gelten*

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.
3. $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
5. $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

◦

Beweis. Die zweite, dritte, und vierte Aussage dieses Theorems basieren auf elementaren mengentheoretischen Aussagen und der σ -Additivität von $\mathbb{P}$. Wir wollen diese elementaren mengentheoretischen Aussagen hier nicht beweisen, sondern verweisen jeweils auf die Intuition, die durch die Venn-Diagramme in Abbildung 20.1 vermittelt wird.

Zu 1.: Wir halten zunächst fest, dass aus $A^c := \Omega \setminus A$ folgt, dass $A^c \cup A = \Omega$ und dass $A^c \cap A = \emptyset$. Mit der Normiertheit und der σ -Additivität von $\mathbb{P}$ folgt dann

$$\mathbb{P}(\Omega) = 1 \Leftrightarrow \mathbb{P}(A^c \cup A) = 1 \Leftrightarrow \mathbb{P}(A^c) + \mathbb{P}(A) = 1 \Leftrightarrow \mathbb{P}(A^c) = 1 - \mathbb{P}(A). \tag{20.6}$$

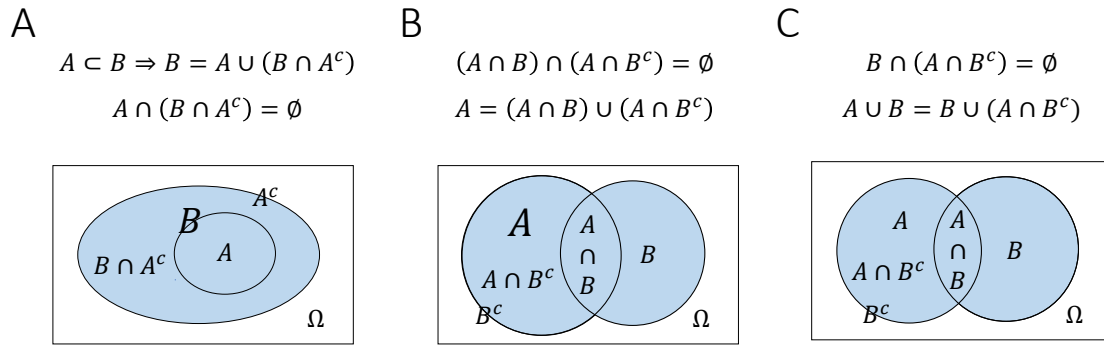


Abbildung 20.1. Venn-Diagramme zum Beweis von Theorem 20.1

Zu 2.: Zunächst gilt (vgl. Abbildung A)

$$A \subset B \Rightarrow B = A \cup (B \cap A^c) \text{ mit } A \cap (B \cap A^c) = \emptyset. \quad (20.7)$$

Mit der σ -Additivität von $\mathbb{P}$ folgt dann aber

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c). \quad (20.8)$$

Mit $\mathbb{P}(B \cap A^c) \geq 0$ folgt dann $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Zu 3.: Zunächst gilt (vgl. Abbildung B)

$$(A \cap B) \cap (A \cap B^c) = \emptyset \text{ und } A = (A \cap B) \cup (A \cap B^c). \quad (20.9)$$

Mit der σ -Additivität von $\mathbb{P}$ folgt dann

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B^c) \quad (20.10)$$

Zu 4.: Zunächst gilt (vgl. Abbildung C)

$$B \cap (A \cap B^c) = \emptyset \text{ und } A \cup B = B \cup (A \cap B^c). \quad (20.11)$$

Mit der σ -Additivität von $\mathbb{P}$ folgt dann

$$\mathbb{P}(A \cup B) = \mathbb{P}(B) + \mathbb{P}(A \cap B^c). \quad (20.12)$$

Mit 3. folgt dann

$$\mathbb{P}(A \cup B) = \mathbb{P}(B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) \quad (20.13)$$

Zu 5.: Die Aussage folgt direkt aus 4. mit $\mathbb{P}(A \cap B) = \emptyset$ und $\mathbb{P}(\emptyset) = 0$. □

20.2. Bedingte Wahrscheinlichkeiten

Wir wenden uns nun dem Begriff der bedingten Wahrscheinlichkeit zu.

Definition 20.2 (Bedingte Wahrscheinlichkeit). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ seien Ereignisse mit $\mathbb{P}(B) > 0$. Die *bedingte Wahrscheinlichkeit des Ereignisses A gegeben das Ereignis B* ist definiert als

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \quad (20.14)$$

Weiterhin heißt das für ein festes $B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$ definierte Wahrscheinlichkeitsmaß

$$\mathbb{P}(\cdot | B) : \mathcal{A} \rightarrow [0, 1], A \mapsto \mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (20.15)$$

die *bedingte Wahrscheinlichkeit* gegeben Ereignis B .

•

Wir halten fest: Die bedingte Wahrscheinlichkeit $\mathbb{P}(A|B)$ eines Ereignisses A gegeben ein Ereignis B ist die mit $1/\mathbb{P}(B)$ skalierte gemeinsame Wahrscheinlichkeit $\mathbb{P}(A \cap B)$ der Ereignisse A und B . Legt man in der Formulierung eines probabilistischen Modells also die gemeinsame Wahrscheinlichkeit $\mathbb{P}(A \cap B)$ sowie die Wahrscheinlichkeit $\mathbb{P}(B) > 0$ des Ereignisses B fest, so legt man insbesondere auch die bedingte Wahrscheinlichkeit $\mathbb{P}(A|B)$ des Ereignisses A gegeben das Ereignis B fest.

Im Unterschied zur gemeinsamen Wahrscheinlichkeit definiert $\mathbb{P}(\cdot|B)$ für ein fest gewähltes B ein Wahrscheinlichkeitsmaß für alle $A \in \mathcal{A}$. Es gelten also insbesondere

- $\mathbb{P}(A|B) \geq 0$ für alle $A \in \mathcal{A}$,
- $\mathbb{P}(\Omega|B) = 1$ und
- $\mathbb{P}(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$ für paarweise disjunkte $A_1, A_2, \dots \in \mathcal{A}$.

Man sollte sich in dieser Hinsicht intuitiv merken, dass die Regeln zum Rechnen mit Wahrscheinlichkeiten für die Ereignisse links des vertikalen Strichs gelten. Wir weisen ferner daraufhin, dass es keinen Grund gibt, die bedingten Wahrscheinlichkeiten

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ und } \mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \quad (20.16)$$

zu verwechseln (vgl. Herzog & Ostwald (2013)). Insbesondere folgt aus $\mathbb{P}(A) \neq \mathbb{P}(B)$ immer direkt $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$. Schließlich sei angemerkt, dass eine Verallgemeinerung der bedingten Wahrscheinlichkeit in Definition 20.2 auf den Fall $\mathbb{P}(B) = 0$ möglich, aber technisch aufwändig ist. Wir verweisen dafür auf die weiterführende Literatur, z.B. Meintrup & Schäffler (2005) und Schmidt (2009).

Beispiel

Als erstes Beispiel für eine bedingte Wahrscheinlichkeit betrachten wir erneut das Modell $(\Omega, \mathcal{A}, \mathbb{P})$ des fairen Würfels. Wir wollen die bedingte Wahrscheinlichkeit für das Ereignis “Es fällt eine gerade Augenzahl” gegeben das Ereignis “Es fällt eine Zahl größer als Drei” berechnen. Wir haben oben bereits gesehen, dass das Ereignis “Es fällt eine gerade Augenzahl” der Teilmenge $A := \{2, 4, 6\}$ von Ω entspricht, und dass das Ereignis “Es fällt eine Zahl größer als Drei” der Teilmenge $B := \{4, 5, 6\}$ von Ω entspricht. Weiterhin haben wir gesehen, dass unter der Annahme, dass der modellierte Würfel fair ist, gilt, dass

$$\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} \quad (20.17)$$

und dass

$$\mathbb{P}(\{4, 5, 6\}) = \mathbb{P}(\{4\}) + \mathbb{P}(\{5\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}. \quad (20.18)$$

Schließlich hatten wir auch gesehen, dass das Ereignis $A \cap B$, also das Ereignis “Es fällt eine gerade Augenzahl, die größer als Drei ist”, die Wahrscheinlichkeit

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{2, 4, 6\} \cap \{4, 5, 6\}) = \mathbb{P}(\{4, 6\}) = \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} \quad (20.19)$$

hat. Nach Definition der bedingten Wahrscheinlichkeit ergibt sich dann direkt

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{4, 6\})}{\mathbb{P}(\{4, 5, 6\})} = \frac{2}{6} \cdot \frac{6}{3} = \frac{2}{3}. \quad (20.20)$$

In diesem Zusammenhang bietet sich eine Interpretation der bedingten Wahrscheinlichkeit $\mathbb{P}(A|B)$ als eine Abnahme subjektiver Unsicherheit bzw. als Zugewinn an subjektiver Information gegenüber der unbedingten Wahrscheinlichkeit $\mathbb{P}(A)$ an: Wenn man weiß, dass eine Augenzahl größer als Drei gefallen ist, dass also das Ereignis $\omega \in B$ vorliegt ist, ist die Wahrscheinlichkeit, dass es sich bei ω um eine gerade Augenzahl handelt $2/3$. Wenn man dagegen nicht weiß, dass das Ereignis $\omega \in B$ vorliegt (und auch sonst keine Information über ω hat) ist die Wahrscheinlichkeit für das Fallen einer geraden Augenzahl nur $1/2$. Bedingen auf dem Vorliegen eines Ereignisses entspricht also der Inkorporation von Information und damit der Abnahme von Unsicherheit in wahrscheinlichkeitstheoretische Modellen. Dies ist die Grundlage der Bayesianischen Statistik.

Zum Schluss dieses Abschnittes wollen wir noch drei technische Konsequenzen der Definition der bedingten Wahrscheinlichkeit betrachten, die wir als Theoreme formulieren.

Das erste Theorem betrifft den Zusammenhang von gemeinsamen und bedingten Wahrscheinlichkeiten und reiteriert, wie gemeinsame Wahrscheinlichkeiten aus bedingten und totalen Wahrscheinlichkeiten berechnet werden können.

Theorem 20.2 (Gemeinsame und bedingte Wahrscheinlichkeiten). *Es seien $(\Omega, \mathcal{A}, \mathbb{P})$ ein W-Raum und $A, B \in \mathcal{A}$ mit $\mathbb{P}(A) > 0$ und $\mathbb{P}(B) > 0$. Dann gilt*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A). \quad (20.21)$$

◦

Beweis. Mit der Definition der jeweiligen bedingten Wahrscheinlichkeit folgen direkt

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \quad (20.22)$$

und

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A). \quad (20.23)$$

□

Ebenso wie das Festlegen von $\mathbb{P}(A \cap B)$ und $\mathbb{P}(A)$ die bedingte Wahrscheinlichkeit $\mathbb{P}(B|A)$ festlegt, legt das Festlegen von $\mathbb{P}(A)$ und $\mathbb{P}(B|A)$ also die gemeinsame Wahrscheinlichkeit $\mathbb{P}(A \cap B)$ fest.

Das nachfolgende sogenannte *Gesetz der totalen Wahrscheinlichkeit* besagt, wie basierend auf gemeinsamen Wahrscheinlichkeiten unbedingte, sogenannte *totale Wahrscheinlichkeiten* berechnet werden können.

Theorem 20.3 (Gesetz der totalen Wahrscheinlichkeit). *$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $A_1, \dots, A_k$ sei eine Partition von Ω . Dann gilt für jedes $B \in \mathcal{A}$, dass*

$$\mathbb{P}(B) = \sum_{i=1}^k \mathbb{P}(B \cap A_i) = \sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i). \quad (20.24)$$

◦

Beweis. Für $i = 1, \dots, k$ sei $C_i := B \cap A_i$, so dass $\cup_{i=1}^k C_i = B$ und $C_i \cap C_j = \emptyset$ für $1 \leq i, j \leq k, i \neq j$. Wir verdeutlichen diese Festlegungen in Abbildung 20.2 mithilfe eines Venn-Diagramms.

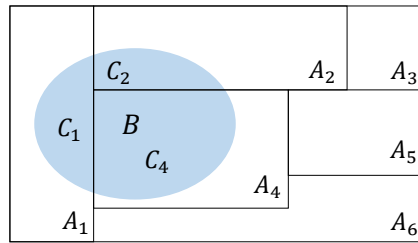


Abbildung 20.2. Venn-Diagramm zum Beweis von Theorem 20.3.

Also gilt

$$\mathbb{P}(B) = \sum_{i=1}^k \mathbb{P}(C_i) = \sum_{i=1}^k \mathbb{P}(B \cap A_i) = \sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i). \quad (20.25)$$

□

Intuitiv entspricht $\mathbb{P}(B)$ also der gewichteten Summe der bedingten Wahrscheinlichkeiten $\mathbb{P}(B|A_i)$ wobei die Wichtungsfaktoren gerade die unbedingten Wahrscheinlichkeiten $\mathbb{P}(A_i)$ für $i = 1, \dots, k$ sind.

Schließlich betrachten wir mit dem *Bayesschen Theorem* eine Formel zur alternativen Berechnung von bedingten Wahrscheinlichkeiten.

Theorem 20.4 (Bayessches Theorem). *($\Omega, \mathcal{A}, \mathbb{P}$) sei ein Wahrscheinlichkeitsraum und $A_1, \dots, A_k$ sei eine Partition von Ω mit $\mathbb{P}(A_i) > 0$ für alle $i = 1, \dots, k$. Wenn $\mathbb{P}(B) > 0$ gilt, dann gilt für jedes $i = 1, \dots, k$, dass*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i)}. \quad (20.26)$$

◦

Beweis. Mit der Definition der bedingten Wahrscheinlichkeit und dem Gesetz der totalen Wahrscheinlichkeit gilt

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i)}. \quad (20.27)$$

□

Man beachte, dass das Theorem von Bayes unabhängig von der Frequentistischen oder Bayesianischen Interpretation von Wahrscheinlichkeiten ist und lediglich eine Aussage zum Rechnen mit bedingten Wahrscheinlichkeiten macht. Im Rahmen der Frequentistischen Inferenz wird das Theorem von Bayes allerdings recht selten benutzt. Im Rahmen der Bayesianischen Inferenz dagegen ist das Theorem von Bayes zentral. In diesem Kontext wird $\mathbb{P}(A_i)$ dann oft die *Prior Wahrscheinlichkeit des Ereignisses* A_i und $\mathbb{P}(A_i|B)$ die *Posterior Wahrscheinlichkeit des Ereignisses* A_i genannt. Wie oben erläutert entspricht $\mathbb{P}(A_i|B)$ der Wahrscheinlichkeit von A_i , wenn man um das Eintreten von B weiß.

20.3. Unabhängige Ereignisse

Die Unabhängigkeit von Ereignissen dient der Modellierung der Abwesenheit von gegenseitigen Einflüssen von Ereignissen. Ihre Definition besagt, dass sich die gemeinsame Wahrscheinlichkeit zweier Ereignisse aus dem Produkt der Wahrscheinlichkeiten der einzelnen Ereignisse ergeben soll. Man spricht in diesem Kontext auch von der *Faktorisierung der gemeinsamen Wahrscheinlichkeit* der Ereignisse. Der Sinn dieser Definition erschließt sich dann im Lichte des Begriffs der bedingten Wahrscheinlichkeit in Theorem 20.5. Wir betrachten zunächst die Definition.

Definition 20.3 (Unabhängige Ereignisse). Zwei Ereignisse $A \in \mathcal{A}$ und $B \in \mathcal{A}$ heißen *unabhängig*, wenn

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \quad (20.28)$$

Eine Menge von Ereignissen $\{A_i | i \in I\} \subset \mathcal{A}$ mit beliebiger Indexmenge I heißt *unabhängig*, wenn für jede endliche Untermenge $J \subseteq I$ gilt, dass

$$\mathbb{P}(\cap_{j \in J} A_j) = \prod_{j \in J} \mathbb{P}(A_j). \quad (20.29)$$

•

Man beachte, dass die Unabhängigkeit bestimmter Ereignissen in der Definition eines wahrscheinlichkeitstheoretischen Modells vorausgesetzt werden kann oder auch aus der Definition eines wahrscheinlichkeitstheoretischen Modells folgen kann. Sind zwei Ereignisse nicht unabhängig, so sagt man auch, dass diese Ereignisse abhängig sind. Ohne Beweis merken wir an, dass die Bedingung der beliebigen Untermengen von I in Definition 20.3 die paarweise Unabhängigkeit der $A_i, i \in I$ sichert (vgl. DeGroot & Schervish (2012)). Schließlich weisen wir daraufhin, dass unabhängige Ereignisse nicht mit disjunkten Ereignissen, also Ereignissen A und B für die $A \cap B = \emptyset$ gilt, verwechselt werden sollten. Insbesondere sind disjunkte Ereignisse mit von Null verschiedenen Wahrscheinlichkeiten $\mathbb{P}(A) > 0$ und $\mathbb{P}(B) > 0$ nie unabhängig, da in diesem Fall $\mathbb{P}(A)\mathbb{P}(B) > 0$ und $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$ gelten und damit offenbar gilt, dass $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$.

Der Sinn der Faktorisierung der gemeinsamen Wahrscheinlichkeit erschließt sich nun anhand folgenden Theorems.

Theorem 20.5 (Bedingte Wahrscheinlichkeit unter Unabhängigkeit). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ seien unabhängige Ereignisse mit $\mathbb{P}(B) \geq 0$. Dann gilt

$$\mathbb{P}(A|B) = \mathbb{P}(A). \quad (20.30)$$

◦

Beweis. Unter den Annahmen des Theorems gilt

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A). \quad (20.31)$$

□

Bei gegebener Unabhängigkeit zweier Ereignisse A und B ist es für die Wahrscheinlichkeit des Ereignisses A also unerheblich, ob auch B eintritt oder nicht, die Wahrscheinlichkeit $\mathbb{P}(A)$ bleibt gleich. Damit wird die Unabhängigkeit von Ereignissen also gerade als Faktorisierung der gemeinsamen Wahrscheinlichkeit von A und B modelliert, damit $\mathbb{P}(A|B) = \mathbb{P}(A)$ folgt. Aus Sicht der Modellierung subjektiver Unsicherheit durch Wahrscheinlichkeiten bedeutet die Unabhängigkeit zweier Ereignisse also, dass das Wissen um das Vorliegen eines der beiden Ereignisse die Wahrscheinlichkeit für das Vorliegen des anderen Ereignisses nicht ändert. Andersherum bedeutet die Abhängigkeit zweier Ereignisse, dass das Wissen um das Vorliegen eines der beiden Ereignisse die Wahrscheinlichkeit für das Vorliegen des anderen Ereignisses verändert, also entweder erhöht oder verringert.

20.4. Literaturhinweise

Viele der in diesem Abschnitt eingeführten Begrifflichkeiten sind auf engste mit der geschichtlichen Genese der Wahrscheinlichkeitstheorie verwoben, so dass keine einzelnen Referenzen angegeben werden sollen. Einen Einstieg in die Geschichte der Wahrscheinlichkeitstheorie der letzten zwei Jahrhunderte bietet Hald (1990), einen Überblick über modernere Entwicklungen gibt Von Plato (1994). Das Theorem von Bayes wird allgemein auf Bayes (1763) zurückgeführt, auch wenn es nicht das eigentliche Hauptthema dieser Arbeit ist.

21. Zufallsvariablen

Mit dem Begriff der *Zufallsvariable* führen wir in diesem Kapitel das probabilistische Standardmodell für einen univariaten Datenpunkt ein. Zentral ist dabei die Möglichkeit, die Verteilungen von Zufallsvariablen mithilfe von Wahrscheinlichkeitsmassefunktion und Wahrscheinlichkeitsdichtefunktionen im Rahmen der probabilistischen Modellformulierung festzulegen. Weiterhin impliziert die Konstruktion von Zufallsvariablen als Abbildungen zufälliger Ergebnisse mit der Untersuchung der Verteilungen der so transformierten zufälligen Ergebnisse das Kernthema statistischer Inferenz.

21.1. Definition

Wir skizzieren zunächst die Konstruktion einer Zufallsvariable und ihrer Verteilung anhand von Abbildung 21.1. Dazu sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und

$$\xi : \Omega \rightarrow \mathcal{X}, \omega \mapsto \xi(\omega) \quad (21.1)$$

eine Abbildung. Weiterhin sei $\mathcal{S}$ eine σ -Algebra auf der Zielmenge $\mathcal{X}$ dieser Abbildung. Für jedes $S \in \mathcal{S}$ sei die Urbildmenge von S gegeben durch (vgl. Definition 4.2)

$$\xi^{-1}(S) := \{\omega \in \Omega \mid \xi(\omega) \in S\}. \quad (21.2)$$

Wenn nun $\xi^{-1}(S) \in \mathcal{A}$ für alle $S \in \mathcal{S}$ gilt, dann nennt man die Abbildung ξ *messbar*. Nehmen wir also an ξ sei messbar. Dann kann allen $S \in \mathcal{S}$ die Wahrscheinlichkeit

$$\mathbb{P}_\xi : \mathcal{S} \rightarrow [0, 1], S \mapsto \mathbb{P}_\xi(S) := \mathbb{P}(\xi^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) \in S\}) \quad (21.3)$$

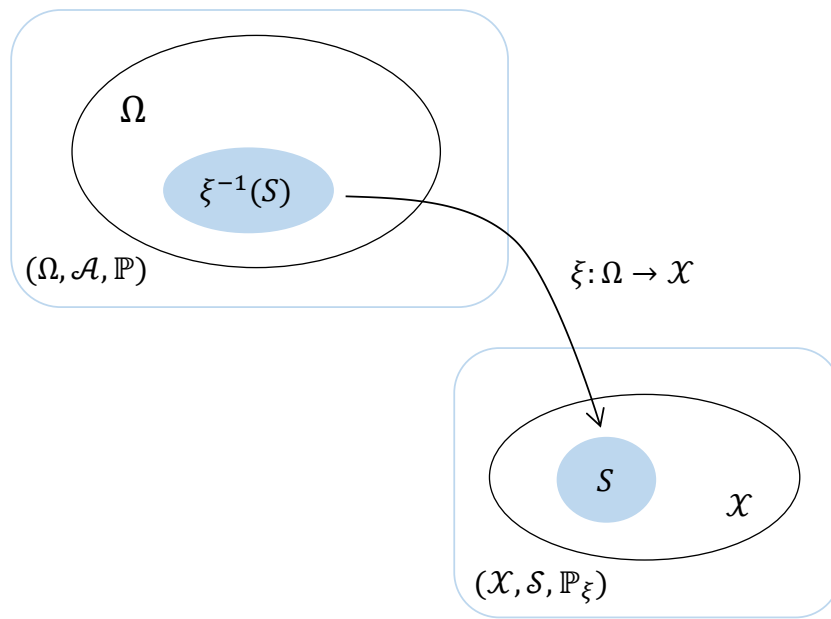
zugeordnet werden. In diesem Kontext nennt man ξ nun eine *Zufallsvariable* und $\mathbb{P}_\xi$ heißt das *Bildmaß* oder die *Verteilung* von ξ . Insgesamt wurde damit ausgehend von dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mithilfe der Zufallsvariable ξ der Wahrscheinlichkeitsraum $(\mathcal{X}, \mathcal{S}, \mathbb{P}_\xi)$ konstruiert. Formal ist eine Zufallsvariable damit wie folgt definiert.

Definition 21.1 (Zufallsvariable). Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $(\mathcal{X}, \mathcal{S})$ ein *Messraum*. Dann ist eine *Zufallsvariable* definiert als eine Abbildung $\xi : \Omega \rightarrow \mathcal{X}$ mit der *Messbarkeitseigenschaft*

$$\{\omega \in \Omega \mid \xi(\omega) \in S\} \in \mathcal{A} \text{ für alle } S \in \mathcal{S}. \quad (21.4)$$

•

Nach Definition 21.1 sind Zufallsvariablen weder “zufällig” noch “Variablen”, sondern messbare Abbildungen. Fragt man nach der Bedeutung des Zufalls für die Werte $\xi(\omega)$ von Zufallsvariablen, so vermittelt weiterhin die implizite Frequentistische Mechanik des Wahrscheinlichkeitsraums $(\Omega, \mathcal{A}, \mathbb{P})$ eine entsprechende Intuition: In jedem Durchgang des



$$\mathbb{P}(\xi^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) \in S\}) =: \mathbb{P}_\xi(S)$$

Abbildung 21.1. Konstruktion von Zufallsvariable und Verteilung.

modellierten Zufallsvorgangs wird dabei ein ω anhand von $\mathbb{P}$ realisiert und dann (deterministisch) auf $\xi(\omega)$ abgebildet. Wir definieren dementsprechend die Begriffe des *Ergebnisraums* und der *Realisierung* einer Zufallsvariable.

Definition 21.2 (Realisierung einer Zufallsvariable). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{S})$ sei ein Messraum und $\xi : \Omega \rightarrow \mathcal{X}$ sei eine Zufallsvariable. Dann heißt $\mathcal{X}$ der *Ergebnisraum* der Zufallsvariable ξ und ein $\xi(\omega) \in \mathcal{X}$ heißt eine *Realisierung* der Zufallsvariable ξ .

•

Beispiel

Aufbauend auf dem in Kapitel 19.3 betrachteten Beispiel eines Wahrscheinlichkeitsraums zur Modellierung des gleichzeitigen Würfels mit einem blauen und einem roten Würfel wollen wir mit der Summe der Würfelaugen Zahlen ein erstes Beispiel für eine Zufallsvariable und ihre Verteilung betrachten. Wir haben in Kapitel 19.3 gesehen, dass ein sinnvolles Wahrscheinlichkeitsraummodell für das gleichzeitige Würfeln mit einem blauen und einem roten Würfel durch $(\Omega, \mathcal{A}, \mathbb{P})$ mit

- $\Omega := \{(r, b) \mid r \in \mathbb{N}_6, b \in \mathbb{N}_6\}$,
- $\mathcal{A} := \mathcal{P}(\Omega)$ und
- $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$ mit $\mathbb{P}(\{(r, b)\}) = \frac{1}{36}$ für alle $(r, b) \in \Omega$.

gegeben ist. Die Auswertung der Summe der beiden Würfelaugen Zahlen wird dann sinnvoller Weise durch die Abbildung

$$\xi : \Omega \rightarrow \mathcal{X}, (r, b) \mapsto \xi((r, b)) := r + b \quad (21.5)$$

beschrieben, wobei offenbar $\mathcal{X} := \{2, 3, \dots, 12\}$ gelten muss. Der Ergebnisraum der Zufallsvariable ist also wiederum endlich und $\mathcal{S} := \mathcal{P}(\mathcal{X})$ ist eine sinnvolle σ -Algebra auf $\mathcal{X}$. Mithilfe der σ -Additivität von $\mathbb{P}$ können wir nun die Verteilung $\mathbb{P}_\xi$ von ξ für alle Elementarereignisse $\{x\} \in \mathcal{S}$ berechnen und damit insbesondere auch die Messbarkeit von ξ nachweisen, wie in Tabelle 21.1 gezeigt.

Tabelle 21.1. Wahrscheinlichkeitsverteilung der Augensumme beim Wurf zweier Würfel

$\mathbb{P}_\xi(\{x\})$	$\mathbb{P}(\xi^{-1}(\{x\}))$	$\mathbb{P}(\{\omega \xi(\omega) \in \{x\}\})$	
$\mathbb{P}_\xi(\{2\})$	$\mathbb{P}(\xi^{-1}(\{2\}))$	$\mathbb{P}(\{(1, 1)\})$	$\frac{1}{36}$
$\mathbb{P}_\xi(\{3\})$	$\mathbb{P}(\xi^{-1}(\{3\}))$	$\mathbb{P}(\{(1, 2), (2, 1)\})$	$\frac{2}{36}$
$\mathbb{P}_\xi(\{4\})$	$\mathbb{P}(\xi^{-1}(\{4\}))$	$\mathbb{P}(\{(1, 3), (3, 1), (2, 2)\})$	$\frac{3}{36}$
$\mathbb{P}_\xi(\{5\})$	$\mathbb{P}(\xi^{-1}(\{5\}))$	$\mathbb{P}(\{(1, 4), (4, 1), (2, 3), (3, 2)\})$	$\frac{4}{36}$
$\mathbb{P}_\xi(\{6\})$	$\mathbb{P}(\xi^{-1}(\{6\}))$	$\mathbb{P}(\{(1, 5), (5, 1), (2, 4), (4, 2), (3, 3)\})$	$\frac{5}{36}$
$\mathbb{P}_\xi(\{7\})$	$\mathbb{P}(\xi^{-1}(\{7\}))$	$\mathbb{P}(\{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\})$	$\frac{6}{36}$
$\mathbb{P}_\xi(\{8\})$	$\mathbb{P}(\xi^{-1}(\{8\}))$	$\mathbb{P}(\{(2, 6), (6, 2), (3, 5), (5, 3), (4, 4)\})$	$\frac{5}{36}$
$\mathbb{P}_\xi(\{9\})$	$\mathbb{P}(\xi^{-1}(\{9\}))$	$\mathbb{P}(\{(3, 6), (6, 3), (4, 5), (5, 4)\})$	$\frac{4}{36}$
$\mathbb{P}_\xi(\{10\})$	$\mathbb{P}(\xi^{-1}(\{10\}))$	$\mathbb{P}(\{(4, 6), (6, 4), (5, 5)\})$	$\frac{3}{36}$
$\mathbb{P}_\xi(\{11\})$	$\mathbb{P}(\xi^{-1}(\{11\}))$	$\mathbb{P}(\{(5, 6), (6, 5)\})$	$\frac{2}{36}$
$\mathbb{P}_\xi(\{12\})$	$\mathbb{P}(\xi^{-1}(\{12\}))$	$\mathbb{P}(\{(6, 6)\})$	$\frac{1}{36}$

Die Wahrscheinlichkeiten der Elementarereignisse in $\mathcal{S}$ wiederum erlauben mithilfe des Begriffs der Wahrscheinlichkeitsmassefunktion (vgl. Kapitel 19.2) das Berechnen beliebiger Ereigniswahrscheinlichkeiten hinsichtlich der Würfelaugen Zahlsumme. Insgesamt haben wir basierend auf $(\Omega, \mathcal{A}, \mathbb{P})$ und ξ also ein weiteres Wahrscheinlichkeitsraummodell $(\mathcal{X}, \mathcal{S}, \mathbb{P}_\xi)$ konstruiert.

Folgender **R** Code demonstriert, wie mithilfe der computerbasierten Erzeugung zufälliger Ergebnisse die Konstruktion des hier betrachteten Beispiels für einen Durchgang eines Zufallsvorgangs simuliert werden kann.

```
# Wahrscheinlichkeitsraummodellformulierung
Omega = list()
idx = 1
for(r in 1:6){
  for(b in 1:6){
    Omega[[idx]] = c(r,b)
    idx = idx + 1 }}
K = length(Omega)
pi = rep(1/K, 1, K)

# Ergebnisrauminitialisierung
# Ergebnisindexinitialisierung
# Ergebnisse roter Würfel
# Ergebnisse blauer Würfel
# \omega \in \Omega
# Ergebnisindexupdate
# Kardinalität von \Omega
# Wahrscheinlichkeitsfunktion \pi

# Durchgang des Zufallsvorgangs
omega = Omega[[which(rmultinom(1, 1, pi) == 1)]] # Auswahl von \omega anhand \mathbb{P}(\{\omega\})

# Realisierung der Zufallsvariable
xi_omega = sum(omega) # \xi(\omega)

omega : 6 5
xi(omega) : 11
```

Im Kontext des Vermessens zufällig ausgewählter experimenteller Einheiten dienen Zufallsvariablen oft als Modelle für Messvorgänge. Betrachten wie beispielsweise die Bestimmung des Wertes eines psychologischen Fragebogens (BDI-II) an zufällig ausgewählten Proband:innen (Abbildung 21.2), so ergibt sich folgende Interpretation: Der Ergebnisraum des

zugrundeliegenden Wahrscheinlichkeitsraums (Ω) soll die Gesamtheit aller in Frage kommender Proband:innen darstellen und die Auswahl einer Proband:in aus diesem Raum die Auswahl eines Ergebnisses ω , welche mit Wahrscheinlichkeit $\mathbb{P}(\{\omega\})$ geschehen soll. Wird nun eine bestimmte Eigenschaft dieser Proband:in in idealisierter Weise gemessen, so handelt es sich dabei um eine deterministische Abbildung auf einen dieser Proband:in zugeordneten Messwert $\xi(\omega)$ im Raum der Messwerte $\mathcal{X}$. Die Messwerte selbst unterliegen dann einer Wahrscheinlichkeitsverteilung, die durch die zugrundeliegende Verteilung und die Art des Messvorgangs bestimmt wird.

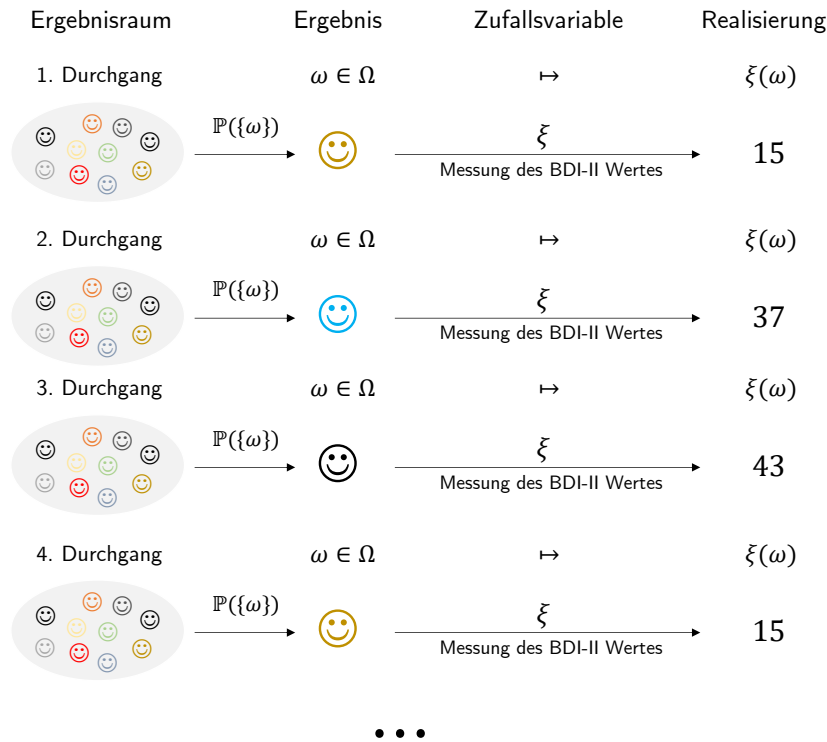


Abbildung 21.2. Zufallsvariable als Modell eines Messvorgangs.

Notation

Die Konventionen zur Notation der mit Zufallsvariablen assoziierten Wahrscheinlichkeiten und Verteilungen sind etwas gewöhnungsbedürftig, so dass wir sie in folgender Definition festhalten wollen.

Definition 21.3 (Notation für Zufallsvariablen). Es seien $(\Omega, \mathcal{A}, \mathbb{P})$ und $(\mathcal{X}, \mathcal{S}, \mathbb{P}_\xi)$ Wahrscheinlichkeitsräume und $\xi : \Omega \rightarrow \mathcal{X}$ sei eine Zufallsvariable. Dann gelten mit $S \in \mathcal{S}$ und

$x \in \mathcal{X}$ folgende Notationskonventionen für Ereignisse in $\mathcal{A}$:

$$\begin{aligned}\{\xi \in S\} &:= \{\omega \in \Omega \mid \xi(\omega) \in S\} \\ \{\xi = x\} &:= \{\omega \in \Omega \mid \xi(\omega) = x\} \\ \{\xi \leq x\} &:= \{\omega \in \Omega \mid \xi(\omega) \leq x\} \\ \{\xi < x\} &:= \{\omega \in \Omega \mid \xi(\omega) < x\} \\ \{\xi \geq x\} &:= \{\omega \in \Omega \mid \xi(\omega) \geq x\} \\ \{\xi > x\} &:= \{\omega \in \Omega \mid \xi(\omega) > x\}\end{aligned}$$

Aus diesen Konventionen ergeben sich folgende Konventionen für Wahrscheinlichkeiten von Verteilungen

$$\begin{aligned}\mathbb{P}_\xi(\xi \in S) &= \mathbb{P}(\{\xi \in S\}) = \mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) \in S\}) \\ \mathbb{P}_\xi(\xi = x) &= \mathbb{P}(\{\xi = x\}) = \mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) = x\}) \\ \mathbb{P}_\xi(\xi \leq x) &= \mathbb{P}(\{\xi \leq x\}) = \mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) \leq x\}) \\ \mathbb{P}_\xi(\xi < x) &= \mathbb{P}(\{\xi < x\}) = \mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) < x\}) \\ \mathbb{P}_\xi(\xi \geq x) &= \mathbb{P}(\{\xi \geq x\}) = \mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) \geq x\}) \\ \mathbb{P}_\xi(\xi > x) &= \mathbb{P}(\{\xi < x\}) = \mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) > x\}).\end{aligned}$$

Oft wird zudem auf das Subskript bei Verteilungssymbolen verzichtet und zum Beispiel $\mathbb{P}_\xi(\xi \in S)$ nur als $\mathbb{P}(\xi \in S)$ geschrieben, solange sich aus dem Kontext keine Verwechslungsmöglichkeit der beiden Wahrscheinlichkeitsmaße ergeben kann.

•

Der Verzicht auf die explizite Bezeichnung von Funktionen eindeutig bestimmter Zufallsvariablen, wie er in Definition 21.3 bezüglich $\mathbb{P}_\xi$ und $\mathbb{P}$ angerissen wird, ist ein grundlegendes Phänomen moderner, anwendungsorientierter Wahrscheinlichkeitstheorie. Speziell werden in den meisten Fällen Funktionen von Zufallsvariablen mithilfe ihrer Argumente identifiziert. So sollte der verständigen Leser:in anhand von “ $\mathbb{P}(\xi \in S)$ ” klar sein, dass es bei $\mathbb{P}$ um das Bildmaß der Zufallsvariable ξ geht, bei “ $\mathbb{P}(\{\omega \in \Omega \mid \xi(\omega) \in S\})$ ” dagegen um das Wahrscheinlichkeitsmaß auf dem Messraum $(\Omega, \mathcal{A})$. Die Identifikation von Funktionen anhand ihrer Argumente, die insbesondere auch bei den in folgenden Abschnitten diskutierten *Wahrscheinlichkeitsmassefunktionen*, *Wahrscheinlichkeitsdichtefunktionen* und *Kumulativen Verteilungsfunktionen* zum Tragen kommt, bezeichnet man in der informatischen Anwendung auch als *Multimethode* (*multiple dispatch*).

Theorem 21.1 (Arithmetik reeller Zufallsvariablen). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ sei der reelle Messraum, $\xi_1 : \Omega \rightarrow \mathbb{R}$ und $\xi_2 : \Omega \rightarrow \mathbb{R}$ seien reellwertige Zufallsvariablen und $c \in \mathbb{R}$ sei eine Konstante. Weiterhin seien

$$\begin{aligned}\xi_1 + c : \Omega &\rightarrow \mathbb{R}, \omega \mapsto (\xi_1 + c)(\omega) := \xi_1(\omega) + c \text{ für } c \in \mathbb{R} \\ c\xi_1 : \Omega &\rightarrow \mathbb{R}, \omega \mapsto (c\xi_1)(\omega) := c\xi_1(\omega) \text{ für } c \in \mathbb{R} \\ \xi_1 + \xi_2 : \Omega &\rightarrow \mathbb{R}, \omega \mapsto (\xi_1 + \xi_2)(\omega) := \xi_1(\omega) + \xi_2(\omega) \\ \xi_1\xi_2 : \Omega &\rightarrow \mathbb{R}, \omega \mapsto (\xi_1\xi_2)(\omega) := \xi_1(\omega)\xi_2(\omega)\end{aligned}\tag{21.6}$$

die Addition einer Konstante zu einer reellwertigen Zufallsvariable, die Multiplikation einer reellwertigen Zufallsvariable mit einer Konstante, die Addition zweier reellwertiger

Zufallsvariablen und die Multiplikation zweier reellwertigen Zufallsvariablen, respektive. Dann sind auch $\xi_1 + c$, $c\xi_1$, $\xi_1 + \xi_2$ und $\xi_1\xi_2$ reellwertige Zufallsvariablen.

◦

Für einen Beweis dieses Theorems verweisen wir auf die weiterführende Literatur, beispielsweise Hesse (2009). Intuitiv besagt Theorem 21.1, dass sowohl Addition einer zufälligen Größe zu einer konstanten Größe, als auch die Multiplikation einer zufälligen Größe mit einer Konstante, als auch die Addition zweier zufälliger Größen und schließlich auch die Multiplikation zweier zufälliger Größen immer wieder zufällige Größen mit ihren dann eigenen Verteilungen ergeben.

21.2. Wahrscheinlichkeitsmassefunktionen

In diesem Abschnitt führen wir mit den *Wahrscheinlichkeitsmassefunktionen* (WMFen) ein Hilfsmittel ein, um Verteilungen von Zufallsvariablen mit *diskretem* (genauer *endlichem* oder *abzählbarem*) Ergebnisraum zu definieren. Wir illustrieren den Begriff anhand dreier wichtiger Beispiele, den Bernoulli- und Binomialzufallsvariablen sowie den diskret-gleichverteilten Zufallsvariablen. Wir definieren zunächst den Begriff der WMF wie folgt.

Definition 21.4 (Diskrete Zufallsvariable und Wahrscheinlichkeitsmassefunktion). Eine Zufallsvariable ξ heißt *diskret*, wenn ihr Ergebnisraum $\mathcal{X}$ endlich oder abzählbar ist und eine Funktion der Form

$$p : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}, x \mapsto p(x) \quad (21.7)$$

existiert, für die gilt

- (1) $\sum_{x \in \mathcal{X}} p(x) = 1$
- (2) $\mathbb{P}_{\xi}(\xi = x) = p(x)$ für alle $x \in \mathcal{X}$.

Eine entsprechende Funktion p heißt *Wahrscheinlichkeitsmassefunktion* (WMF) von ξ .

•

Wir erinnern daran, dass eine Menge *abzählbar* heißt, wenn sie bijektiv auf $\mathbb{N}$ abgebildet werden kann. Im Deutschen nennt man WMFen auch *Zähldichten*. Im Englischen nennt man WMFen *probability mass functions* (PMFs), an diesem Begriff orientieren wir uns hier. Ohne Beweis halten wir fest, dass jede Funktion $p : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$, die die Normiertheitseigenschaft $\sum_{x \in \mathcal{X}} p(x) = 1$ besitzt als WMF einer Zufallsvariable interpretiert werden kann.

Beispiele

Wir wollen mit den *Bernoulli-Zufallsvariablen*, den *Binomial-Zufallsvariablen* und den *Diskrete-gleichverteilten Zufallsvariablen* drei erste Beispiel für die Definition von Verteilungen mithilfe von WMFen betrachten.

Definition 21.5 (Bernoulli Zufallsvariable). Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} = \{0, 1\}$ und WMF

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \mu^x(1 - \mu)^{1-x} \text{ mit } \mu \in [0, 1]. \quad (21.8)$$

Dann sagen wir, dass ξ einer *Bernoulli-Verteilung mit Parameter $\mu \in [0, 1]$* unterliegt und nennen ξ eine *Bernoulli-Zufallsvariable*. Wir kürzen dies mit $\xi \sim \text{Bern}(\mu)$ ab. Die WMF einer Bernoulli-Zufallsvariable bezeichnen wir mit

$$\text{Bern}(x; \mu) := \mu^x(1 - \mu)^{1-x}. \quad (21.9)$$

•

Bernoulli-Zufallsvariablen können immer dann zur probabilistischen Modellierung genutzt werden, wenn das betrachtete Phänomen binär ist und die möglichen Werte der Zufallsvariable bijektiv auf $\{0, 1\}$ abgebildet werden können. Man beachte, dass die funktionale Form der Bernoulli-Zufallsvariable $\text{Bern}(x; \mu)$ nur für $x \in \{0, 1\}$ Sinn ergibt und nicht etwa für $x \in \{\text{Heads}, \text{Tails}\}$. Bildet man aber die möglichen Ergebnisse eines Münzwurfs auf $\{0, 1\}$ ab, also definiert etwa $0 := \text{Heads}$ und $1 := \text{Tails}$, so kann eine Bernoulli-Zufallsvariable durchaus als Modell eines Münzwurfs dienen.

Der Parameter $\mu \in [0, 1]$ einer Bernoulli-Zufallsvariable ist die Wahrscheinlichkeit dafür, dass die Zufallsvariable ξ den Wert 1 annimmt, dies erkennt man anhand von

$$\mathbb{P}(\xi = 1) = \mu^1(1 - \mu)^{1-1} = \mu. \quad (21.10)$$

Wir visualisieren die WMFen von Bernoulli-Zufallsvariablen für $\mu := 0.1, \mu := 0.5$ und $\mu := 0.7$ in Abbildung 21.3.

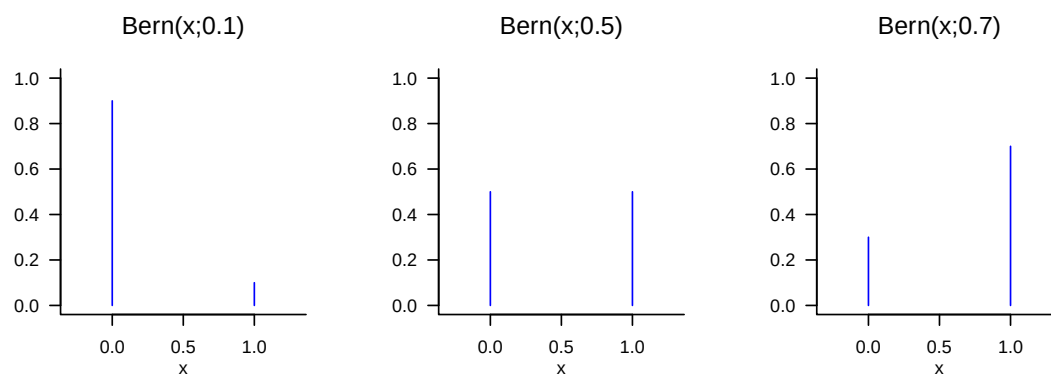


Abbildung 21.3. WMFen von Bernoulli-Zufallsvariablen.

Definition 21.6 (Binomialzufallsvariable). Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := \mathbb{N}_n^0$ und WMF

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \binom{n}{x} \mu^x(1 - \mu)^{n-x} \text{ für } \mu \in [0, 1]. \quad (21.11)$$

Dann sagen wir, dass ξ einer *Binomialverteilung mit Parametern* $\mu \in [0, 1]$ und $n \in \mathbb{N}$ unterliegt und nennen ξ eine *Binomial-Zufallsvariable*. Wir kürzen dies mit $\xi \sim \text{Bin}(\mu, n)$ ab. Die WMF einer Binomial-Zufallsvariable bezeichnen wir mit

$$\text{Bin}(x; \mu, n) := \binom{n}{x} \mu^x (1 - \mu)^{n-x}. \quad (21.12)$$

•

Ohne Beweis halten wir fest, dass eine Binomial-Zufallsvariable als Modell der Summe von n unabhängig und identisch verteilten Bernoulli-Zufallsvariablen genutzt werden kann. Insbesondere gilt also $\text{Bin}(x; \mu, 1) = \text{Bern}(x; \mu)$. Binomial-Zufallsvariablen haben die Eigenschaft, dass mit n einer ihrer Parameter nicht nur die funktionale Form ihrer WMF, sondern auch ihren Ergebnisraum $\mathcal{X}$ festlegt. Wir visualisieren die WMFen von Binomial-Zufallsvariablen für $(\mu, n) := (0.1, 5)$, $(\mu, n) := (0.5, 10)$ und $(\mu, n) := (0.7, 15)$ in Abbildung 21.4.

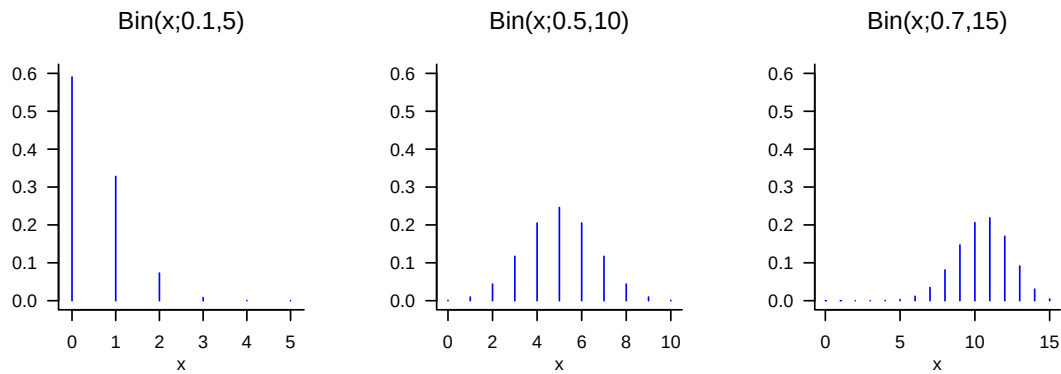


Abbildung 21.4. WMFen von Binomial-Zufallsvariablen.

Definition 21.7 (Diskret-gleichverteilte Zufallsvariable). Es sei ξ eine diskrete Zufallsvariable mit endlichem Ergebnisraum $\mathcal{X}$ und WMF

$$p : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}, x \mapsto p(x) := \frac{1}{|\mathcal{X}|}. \quad (21.13)$$

Dann sagen wir, dass ξ einer *diskreten Gleichverteilung* unterliegt und nennen ξ eine *diskret-gleichverteilte Zufallsvariable*. Wir kürzen dies mit $\xi \sim U(|\mathcal{X}|)$ ab. Die WMF einer diskret-gleichverteilten Zufallsvariable bezeichnen wir mit

$$U(x; |\mathcal{X}|) := \frac{1}{|\mathcal{X}|}. \quad (21.14)$$

•

Diskrete-gleichverteilte Zufallsvariable können offenbar immer dann zur probabilistischen Modellierung genutzt werden, wenn die möglichen diskreten Ergebnisse des modellierten Phänomens die gleiche Wahrscheinlichkeit haben. Im Fall der diskret-gleichverteilten

Zufallsvariablen braucht es zur Definition ihrer funktionalen Form nach Festlegung des Ergebnisraums keinen weiteren Parameter. Offenbar gilt für $\mathcal{X} := \{0, 1\}$

$$U(x; |\mathcal{X}|) = \text{Bern}(x; 0.5) = \text{Bin}(x; 1, 0.5) \quad (21.15)$$

Wir visualisieren die WMFen von diskret-gleichverteilten Zufallsvariablen für $\mathcal{X} := \{0, 1\}$, $\mathcal{X} := \{-3, -2, -1, 0, 1\}$ und $\mathcal{X} := \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ in Abbildung 21.5. Man beachte, dass Definition 21.7 formal die Sequentialität der Elemente von $\mathcal{X}$ nicht erfordert, man kann also genauso eine diskret-gleichverteilte Zufallsvariable mit Ergebnisraum $\mathcal{X} := \{1, 5, 7\}$ oder auch nicht numerischen Ergebnisraum $\mathcal{X} := \{a, b, x, y\}$ definieren.

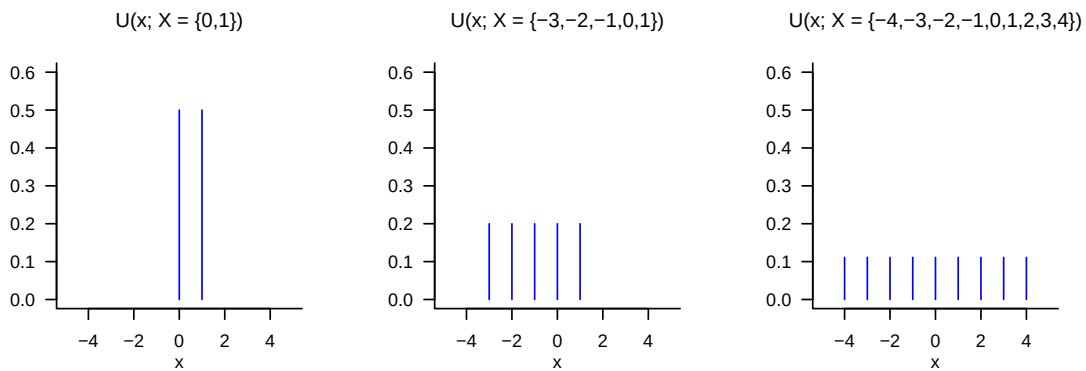


Abbildung 21.5. WMFen von diskret-gleichverteilten Zufallsvariablen.

21.3. Wahrscheinlichkeitsdichtefunktionen

In diesem Abschnitt führen wir mit den *Wahrscheinlichkeitsdichtefunktionen* (WDFen) ein Hilfsmittel ein, um Verteilungen von Zufallsvariablen mit *kontinuierlichem* (genauer *überabzählbarem*) Ergebnisraum zu definieren. Wir illustrieren den Begriff zunächst an der grundlegendsten aller Zufallsvariablen, der normalverteilten Zufallsvariable. Mit der Gamma-Zufallsvariable, der Beta-Zufallsvariable und der gleichverteilten Zufallsvariable wollen wir dann noch drei Beispiele von Zufallsvariablen betrachten, die sowohl in der Modellformulierung der Frequentistischen als auch der Bayesianischen Inferenz an vielen Stellen eingesetzt werden. Wir definieren den Begriff der WDF wie folgt.

Definition 21.8 (Kontinuierliche Zufallsvariable und Wahrscheinlichkeitsdichtefunktion). Eine Zufallsvariable ξ heißt *kontinuierlich*, wenn $\mathbb{R}$ der Ergebnisraum von ξ ist und eine Funktion

$$p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, x \mapsto p(x) \quad (21.16)$$

existiert, für die gilt

- (1) $\int_{-\infty}^{\infty} p(x) dx = 1$ und
- (2) $\mathbb{P}_{\xi}(\xi \in [a, b]) = \int_a^b p(x) dx$ für alle $a, b \in \mathbb{R}$ mit $a \leq b$.

Eine entsprechende Funktion p heißt *Wahrscheinlichkeitsdichtefunktion (WDF)* von ξ .

•

Im Englischen nennt man WDFen *probability density functions (PDFs)*. Ohne Beweis halten wir fest, dass jede Funktion $p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$, für deren uneigentliches Integral $\int_{-\infty}^{\infty} p(x) dx = 1$ gilt, die also normiert ist, als WDF einer Zufallsvariable interpretiert werden kann.

Im Umgang mit WDFen und in Abgrenzung zu WMFen sollte man sich die *Dichteigenschaft* einer WDF immer bewusst machen: Die Werte einer WDF stellen keine Wahrscheinlichkeiten, sondern Wahrscheinlichkeitsdichten dar, Wahrscheinlichkeiten werden aus WDFen anhand von Definition 21.8 (2) durch Integration berechnet. Wie im physikalischen Sinn ergibt sich die einem reellen Intervall zugeordnete Wahrscheinlichkeit(smasse) also erst durch “Multiplikation” mit dem entsprechenden “Intervallvolumen”. Man denke hierzu auch an die Approximation des bestimmten Integrals $\int_a^b p(x) dx$ durch einen Riemannschen Summenterm (vgl. Definition 7.3). Intuitiv gilt also

$$(\text{Wahrscheinlichkeits})\text{Masse} = (\text{Wahrscheinlichkeits})\text{Dichte} \cdot (\text{Mengen})\text{Volumen}, \quad (21.17)$$

wobei sich das Volumen im Sinne des Lebesgue-Maßes auf die Breite des Intervalls $[a, b]$ bezieht. Wie in der physikalischen Analogie ist die Wahrscheinlichkeitsmasse eines Intervalls ohne Volumen gleich Null,

$$\mathbb{P}(\xi = a) = \int_a^a p(x) dx = 0. \quad (21.18)$$

Ferner gilt, dass bei entsprechend kleinen Intervallen WDFen auch Werte größer als 1 annehmen können, auch wenn dies für die Wahrscheinlichkeit, die sich dann durch entsprechende Integration ergibt, nicht der Fall sein kann. Schließlich sei trotz dieser technischen Feinheiten folgende Intuition betont: Betrachtet man die graphische Darstellung der WDF einer Zufallsvariable und stellt sich eine Zerlegung von $\mathbb{R}$ in gleich große Intervalle vor, so besitzt die Zufallsvariable natürlich eine höhere Wahrscheinlichkeit dafür, Werte in einem Intervall mit assoziierter höherer Wahrscheinlichkeitsdichte anzunehmen als Werte in einem Intervall mit assoziierter relativ niedrigerer Wahrscheinlichkeitsdichte.

Normalverteilte Zufallsvariablen

Mit der normalverteilten Zufallsvariable wollen wir als erstes Beispiel für die Definition einer kontinuierlichen Zufallsvariable mithilfe einer WDF nun die wichtigste Zufallsvariable der probabilistischen Modellbildung einführen.

Definition 21.9 (Normalverteilte Zufallsvariable). Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (21.19)$$

Dann sagen wir, dass ξ einer *Normalverteilung mit Parametern* $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ unterliegt und nennen ξ eine *normalverteilte Zufallsvariable*. Wir kürzen dies mit $\xi \sim N(\mu, \sigma^2)$ ab. Die WDF einer normalverteilten Zufallsvariable bezeichnen wir mit

$$N(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (21.20)$$

Eine normalverteilte Zufallsvariable mit $\mu = 0$ und $\sigma^2 = 1$ nennt man eine *standardnormalverteilte Zufallsvariable*

•

Wir visualisieren die WDFen von normalverteilten Zufallsvariablen für $(\mu, \sigma^2) := (0, 1)$, $(\mu, \sigma^2) := (-2.5, 10)$ und $(\mu, \sigma^2) := (3, 0.5)$ in Abbildung 21.6. Man macht sich an dieser Abbildung graphisch klar, dass die WDFen von normalverteilten Zufallsvariablen immer genau einen Werte höchster Wahrscheinlichkeitsdichte haben und zwar an der Stelle des Parameters $\mu \in \mathbb{R}$. Dies ergibt sich durch die Tatsache, dass das Argument der Exponentialfunktion in der funktionalen Form von $N(x; \mu, \sigma^2)$ aufgrund des negativen Vorzeichens des Quadrates von $x - \mu$ und der Positivität von σ^2 immer nicht-positiv ist und die Exponentialfunktion auf den nicht-positiven reellen Zahlen ihr Maximum bei $x = \mu$, also $x - \mu = 0$ annimmt. Weiterhin macht man sich graphisch klar, dass der Parameter $\sigma^2 > 0$ die Breite der WDF einer normalverteilten Zufallsvariable enkodiert.

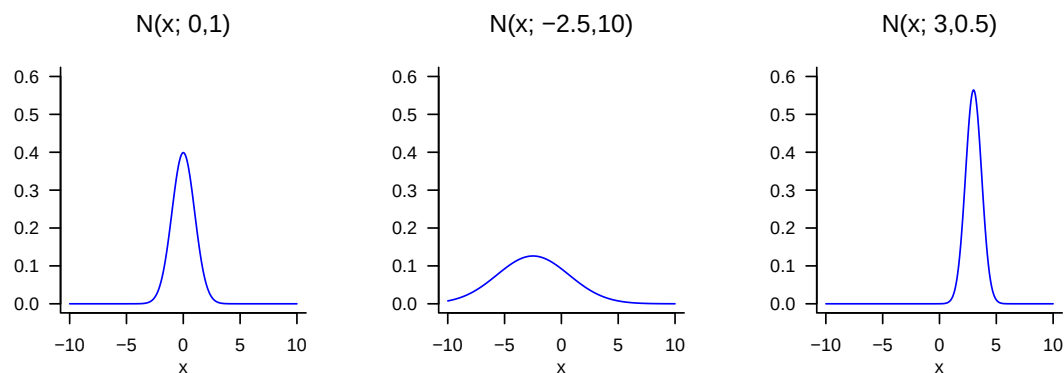


Abbildung 21.6. WDFen von normalverteilten Zufallsvariablen.

Weitere Beispiele

Wir wollen mit den *Gamma-Zufallsvariablen*, den *Beta-Zufallsvariablen* und den *gleichverteilten Zufallsvariablen* drei weitere Beispiele für die Definition von Verteilungen mithilfe von WDFen betrachten.

Definition 21.10 (Gamma-Zufallsvariable). Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := \mathbb{R}_{>0}$ und WDF

$$p : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right), \quad (21.21)$$

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass ξ einer *Gammaverteilung mit Formparameter $\alpha > 0$ und Skalenparameter $\beta > 0$* unterliegt und nennen ξ eine *gammaverteilte Zufallsvariable*. Wir kürzen dies mit $\xi \sim G(\alpha, \beta)$ ab. Die WDF einer gammaverteilten Zufallsvariable bezeichnen wir mit

$$G(x; \alpha, \beta) := \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right). \quad (21.22)$$

•

Die spezielle Gamma-Zufallsvariable mit WDF $G(x; \frac{n}{2}, 2)$ wird *Chi-Quadrat* (χ^2) *Verteilung mit n Freiheitsgraden* genannt. Wir visualisieren die WDFen von Gamma-Zufallsvariablen für $(\alpha, \beta) := (1, 1)$, $(\alpha, \beta) := (2, 2)$ und $(\alpha, \beta) := (5, 1)$ in Abbildung 21.7.

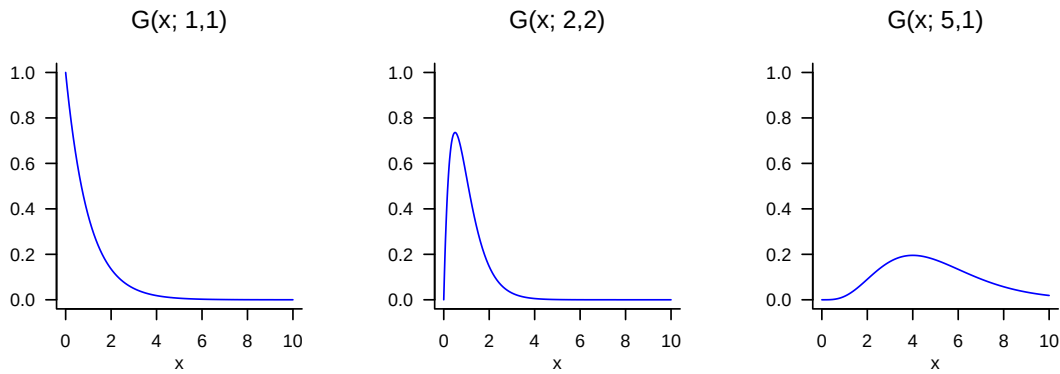


Abbildung 21.7. WDFen von Gamma-Zufallsvariablen.

Definition 21.11 (Beta-Zufallsvariable). Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := [0, 1]$ und WDF

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \text{ mit } \alpha, \beta \in \mathbb{R}_{>0}, \quad (21.23)$$

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass ξ einer *Beta-Verteilung* mit Parametern $\alpha > 0$ und $\beta > 0$ unterliegt, und nennen ξ eine *Beta-Zufallsvariable*. Wir kürzen dies mit $\xi \sim \text{Beta}(\alpha, \beta)$ ab. Die WDF einer Beta-Zufallsvariable bezeichnen wir mit

$$\text{Beta}(x; \alpha, \beta) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}. \quad (21.24)$$

•

Dadurch, dass der Ergebnisraum einer Beta-Zufallsvariable auf das Intervall $\mathcal{X} := [0, 1]$ (für $\alpha < 1, \beta < 1$ genauer $\mathcal{X} :=]0, 1[$) beschränkt ist, bietet sich eine Beta-Zufallsvariable unter anderem dafür an, Wahrscheinlichkeiten von Wahrscheinlichkeiten (also Werten zwischen 0 und 1) zu beschreiben. Wir visualisieren die WDFen von Beta-Zufallsvariablen für $(\alpha, \beta) := (1, 1)$, $(\alpha, \beta) := (3, 2)$ und $(\alpha, \beta) := (10, 5)$ in Abbildung 21.8.

Mit den gleichverteilten Zufallsvariablen betrachten wir abschließend noch das Analogon zu diskret-gleichverteilten Zufallsvariablen für den Fall kontinuierlicher Zufallsvariablen.

Definition 21.12 (Gleichverteilte Zufallsvariable). Es sei ξ eine kontinuierliche Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, x \mapsto p(x) := \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}. \quad (21.25)$$

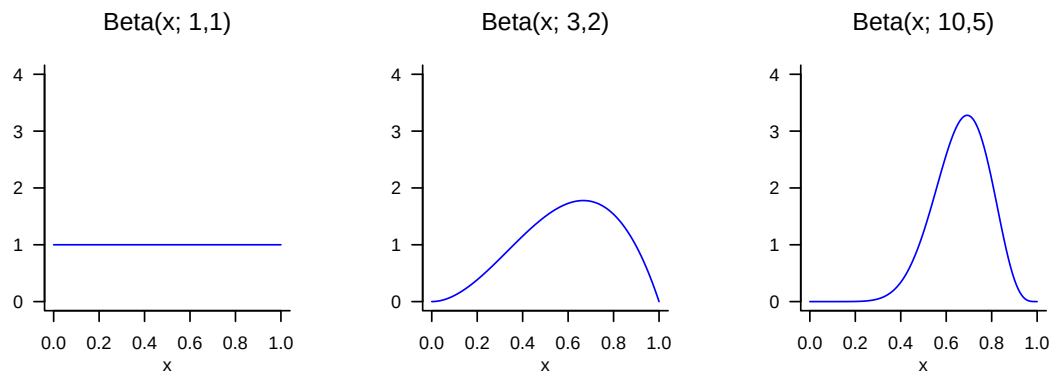


Abbildung 21.8. WDFen von Beta-Zufallsvariablen.

Dann sagen wir, dass ξ einer *Gleichverteilung mit Parametern a und b* unterliegt und nennen ξ eine *gleichverteilte Zufallsvariable*. Wir kürzen dies mit $\xi \sim U(a, b)$ ab. Die WDF einer gleichverteilten Zufallsvariable bezeichnen wir mit

$$U(x; a, b) := \frac{1}{b - a}. \quad (21.26)$$

•

Wir visualisieren die WDFen von gleichverteilten Zufallsvariablen für $(a, b) := (0, 1)$, $(a, b) := (-3, 1)$ und $(a, b) := (-4, 4)$ in Abbildung 21.9.

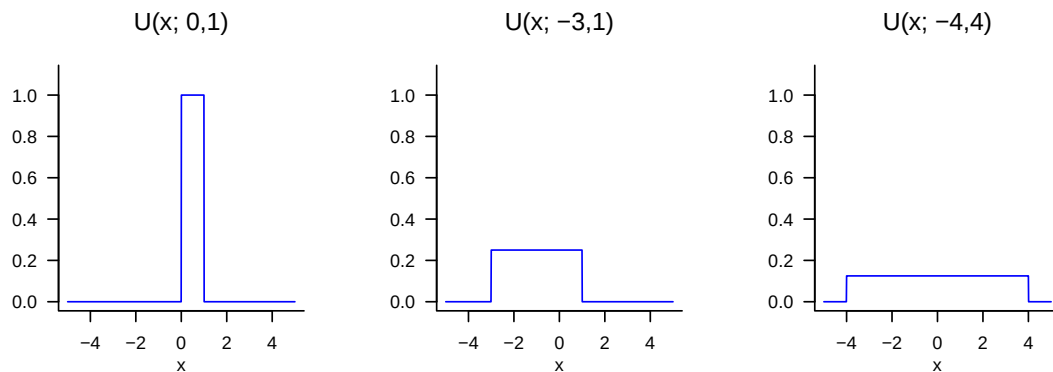


Abbildung 21.9. WDFen von gleichverteilten-Zufallsvariablen.

21.4. Kumulative Verteilungsfunktionen

Im letzten Abschnitt dieses Kapitels wollen wir mit den *kumulativen Verteilungsfunktionen (KVFen)* eine weitere Möglichkeit einführen, die Verteilungen von diskreten oder kontinuierlichen Zufallsvariablen festzulegen. Es ist allgemein allerdings eher üblich, dies mithilfe

von WMFen oder WDFen zu tun. Trotzdem sind KVFen an vielen Stellen nützlich, da sie sowohl für diskrete als auch kontinuierliche Zufallsvariablen einen direkten Zusammenhang zwischen Werten der Zufallsvariable und bestimmten Wahrscheinlichkeiten herstellen, der in der Anwendung ohne Summation oder Integration auskommt. Wir betrachten zunächst die allgemeine Definition von KVFen für sowohl diskrete als auch kontinuierliche Zufallsvariablen und wenden uns dann den KVFen von diskreten und den KVFen von kontinuierlichen Zufallsvariablen im einzelnen zu. Für eine beliebige Zufallsvariable definieren wir den Begriff der KVF wie folgt.

Definition 21.13 (Kumulative Verteilungsfunktion). Die *kumulative Verteilungsfunktion* (KVF) einer Zufallsvariable ξ ist definiert als

$$P : \mathbb{R} \rightarrow [0, 1], x \mapsto P(x) := \mathbb{P}(\xi \leq x). \quad (21.27)$$

•

Man beachte, dass $P(x)$ für jedes $x \in \mathbb{R}$ definiert ist, auch wenn für ein gegebenes $x \in \mathbb{R}$ gilt, dass $x \notin \mathcal{X}$. Mithilfe von KVFen können zum Beispiel *Überschreitungswahrscheinlichkeiten* und *Intervallwahrscheinlichkeiten* von Zufallsvariablen direkt ausgewertet werden. Dies ist der Inhalt folgender Theoreme.

Theorem 21.2 (Überschreitungswahrscheinlichkeit). *Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X}$ und P ihre kumulative Verteilungsfunktion. Dann gilt für die Überschreitungswahrscheinlichkeit $\mathbb{P}(\xi > x)$, dass*

$$\mathbb{P}(\xi > x) = 1 - P(x) \text{ für alle } x \in \mathcal{X}. \quad (21.28)$$

◦

Beweis. Die Ereignisse $\{\xi > x\}$ und $\{\xi \leq x\}$ sind disjunkt und

$$\Omega = \{\omega \in \Omega | \xi(\omega) > x\} \cup \{\omega \in \Omega | \xi(\omega) \leq x\} = \{\xi > x\} \cup \{\xi \leq x\}. \quad (21.29)$$

Mit der σ -Additivität von $\mathbb{P}$ folgt dann

$$\begin{aligned} \mathbb{P}(\Omega) &= 1 \\ \Leftrightarrow \mathbb{P}(\{\xi > x\} \cup \{\xi \leq x\}) &= 1 \\ \Leftrightarrow \mathbb{P}(\{\xi > x\}) + \mathbb{P}(\{\xi \leq x\}) &= 1 \\ \Leftrightarrow \mathbb{P}(\{\xi > x\}) &= 1 - \mathbb{P}(\{\xi \leq x\}) \\ \Leftrightarrow \mathbb{P}(\{\xi > x\}) &= 1 - P(x). \end{aligned} \quad (21.30)$$

□

Theorem 21.3 (Intervallwahrscheinlichkeiten). *Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X}$ und P ihre KVF. Dann gilt für die Intervallwahrscheinlichkeit $\mathbb{P}(\xi \in]x_1, x_2])$, dass*

$$\mathbb{P}(\xi \in]x_1, x_2]) = P(x_2) - P(x_1) \text{ für alle } x_1, x_2 \in \mathcal{X} \text{ mit } x_1 < x_2. \quad (21.31)$$

◦

Beweis. Wir betrachten die Ereignisse $\{\xi \leq x_1\}, \{x_1 < \xi \leq x_2\}$ und $\{\xi \leq x_2\}$, wobei

$$\{\xi \leq x_1\} \cap \{x_1 < \xi \leq x_2\} = \emptyset \text{ und } \{\xi \leq x_1\} \cup \{x_1 < \xi \leq x_2\} = \{\xi \leq x_2\}. \quad (21.32)$$

gelten. Mit der σ -Additivität von $\mathbb{P}$ gilt dann

$$\begin{aligned} \mathbb{P}(\{\xi \leq x_1\} \cup \{x_1 < \xi \leq x_2\}) &= \mathbb{P}(\{\xi \leq x_2\}) \\ \Leftrightarrow \mathbb{P}(\{\xi \leq x_1\}) + \mathbb{P}(\{x_1 < \xi \leq x_2\}) &= \mathbb{P}(\{\xi \leq x_2\}) \\ \Leftrightarrow \mathbb{P}(\{x_1 < \xi \leq x_2\}) &= \mathbb{P}(\{\xi \leq x_2\}) - \mathbb{P}(\{\xi \leq x_1\}) \\ \Leftrightarrow \mathbb{P}(\{x_1 < \xi \leq x_2\}) &= P(x_2) - P(x_1) \\ \Leftrightarrow \mathbb{P}(\xi \in]x_1, x_2]) &= P(x_2) - P(x_1). \end{aligned} \quad (21.33)$$

□

Folgendes Theorem gibt drei zentrale Eigenschaften von KVFn an. Dabei besagt die dritte Eigenschaft, dass eine KVF keine Sprünge hat, wenn man sich Grenzpunkten von rechts nähert. Tatsächlich sind die diskutierten Eigenschaften auch gerade die definierenden Eigenschaften von KVFn, das heißt, jede Funktion P , die die Eigenschaften von Theorem 21.4 erfüllt, kann als eine KVF einer Zufallsvariable interpretiert werden. Für einen Beweis dieser Tatsache verweisen wir auf die weiterführende Literatur.

Theorem 21.4 (Eigenschaften von kumulative Verteilungsfunktionen). *Es sei ξ eine Zufallsvariable und P ihre kumulative Verteilungsfunktion. Dann hat P die folgenden Eigenschaften*

- (1) P ist monoton steigend, i.e., wenn $x_1 < x_2$, dann gilt $P(x_1) \leq P(x_2)$.
- (2) $\lim_{x \rightarrow -\infty} P(x) = 0$ und $\lim_{x \rightarrow \infty} P(x) = 1$.
- (3) P ist rechtsseitig stetig, d.h., $P(x) = \lim_{y \rightarrow x, y > x} P(y)$ für alle $x \in \mathbb{R}$

◦

Beweis. Wir betrachten die Eigenschaften nacheinander.

- (1) Wir halten zunächst fest, dass für Ereignisse $A \subset B$ gilt, dass $\mathbb{P}(A) \leq \mathbb{P}(B)$. Wir halten dann fest, dass für $x_1 < x_2$,

$$\{\xi \leq x_1\} = \{\omega \in \Omega | \xi(\omega) \leq x_1\} \subset \{\omega \in \Omega | \xi(\omega) \leq x_2\} = \{\xi \leq x_2\}. \quad (21.34)$$

Also gilt

$$\mathbb{P}(\{\xi \leq x_1\}) \leq \mathbb{P}(\{\xi \leq x_2\}) \Rightarrow P(x_1) \leq P(x_2). \quad (21.35)$$

- (2) Für einen Beweis verweisen wir auf die weiterführende Literatur.
- (3) Wir definieren

$$P(x^+) = \lim_{y \rightarrow x, y > x} P(y). \quad (21.36)$$

Seien nun $y_1 > y_2 > \dots$ so, dass $\lim_{n \rightarrow \infty} y_n = x$. Dann gilt

$$\{\xi \leq x\} = \cap_{n=1}^{\infty} \{\xi \leq y_n\}. \quad (21.37)$$

Es gilt also

$$P(x) = \mathbb{P}(\{\xi \leq x\}) = \mathbb{P}(\cap_{n=1}^{\infty} \{\xi \leq y_n\}) = \lim_{n \rightarrow \infty} \mathbb{P}(\{\xi \leq y_n\}) = P(x^+), \quad (21.38)$$

wobei wir die dritte Gleichung unbegründet stehen lassen.

□

KVFen von diskreten Zufallsvariablen

Anhand von Abbildung 21.10 und Abbildung 21.11 wollen wir uns obige Eigenschaften von KVFen diskreter Zufallsvariablen visuell verdeutlichen. Dabei sollte man immer vor Augen haben, dass der Wert $P(x)$ einer KVF für den Ergebniswert x der Zufallsvariable x der Wahrscheinlichkeit $\mathbb{P}_\xi(\xi \leq x)$ entspricht, also die Wahrscheinlichkeit dafür ist, dass die Zufallsvariable ξ Werte kleiner oder gleich x annimmt. Liest man diese Wahrscheinlichkeiten entsprechend aus den Darstellungen der korrespondierenden WMF ab, so ergeben sich die funktionale Form der KVFen im Vergleich mit den entsprechenden WMFen intuitiv. Weiterhin erschließen sich auch folgende Eigenschaften der KVFen von diskreten Zufallsvariablen intuitiv: Wenn $a < b$ und $\mathbb{P}(a < \xi < b) = 0$ ist, dann ist die KVF von ξ konstant horizontal auf $]a, b[$. Weiterhin gilt, dass an jedem Punkt x mit $\mathbb{P}(\xi = x) > 0$ die KVF um den Betrag $\mathbb{P}(\xi = x)$ springt, an dieser Stelle also linksseitig nicht stetig ist. Allgemein ist die KVF einer diskreten Zufallsvariable mit Ergebnisraum $\mathbb{N}_0$ durch

$$P : \mathbb{R} \rightarrow [0, 1], x \mapsto P(x) := \sum_{k=0}^{\lfloor x \rfloor} \mathbb{P}(\xi = k) \quad (21.39)$$

gegeben, wobei $\lfloor x \rfloor$ die Abrundungsfunktion bezeichnet.

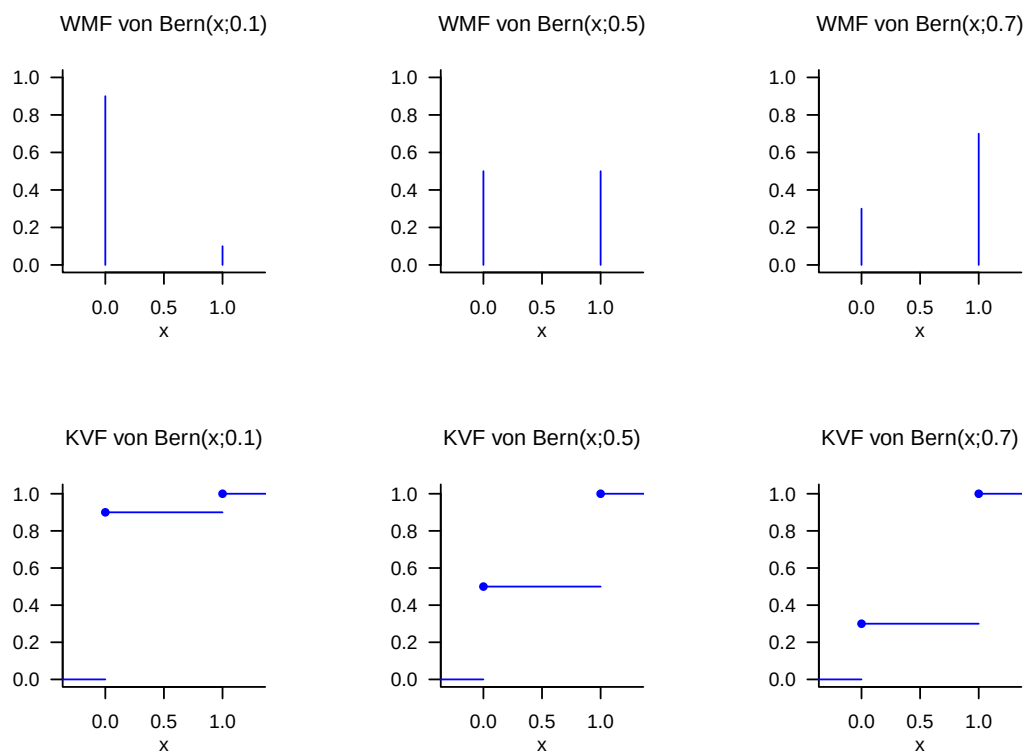


Abbildung 21.10. WMFen und KVFen von Bernoulli-Zufallsvariablen.

KVFen von kontinuierlichen Zufallsvariablen

Die KVFen von kontinuierlichen Zufallsvariablen sind analytisch etwas zugänglicher als die KVFen von diskreten Zufallsvariablen, da sie keine Unstetigkeitsstellen aufweisen. Wir

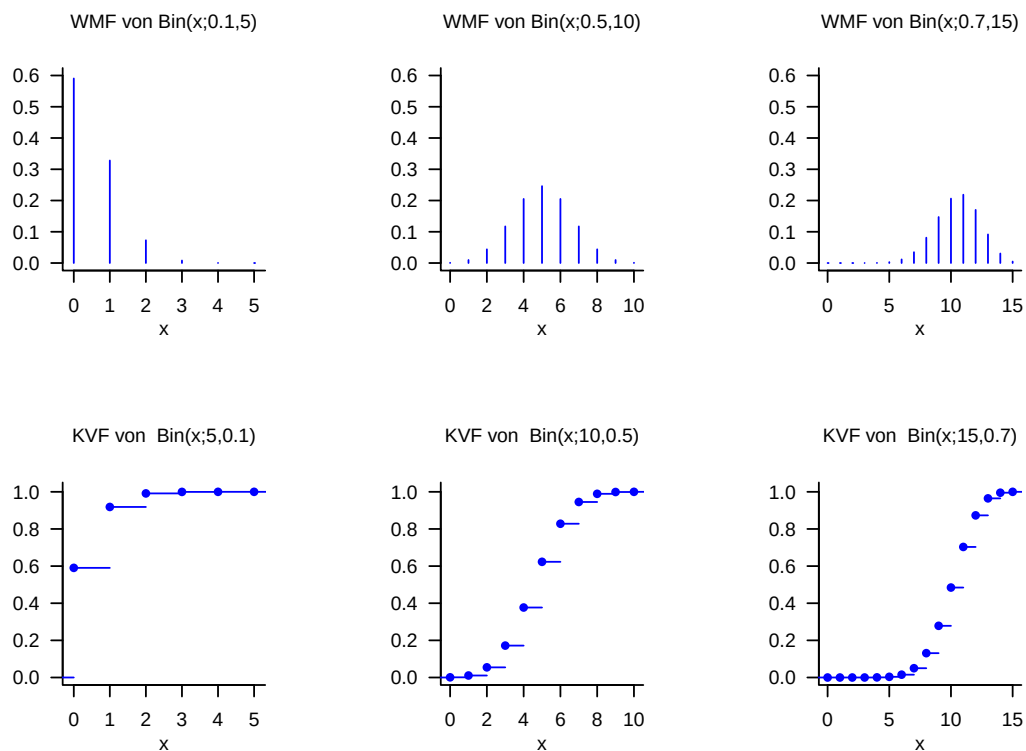


Abbildung 21.11. WMFen und KVFen von Binomial-Zufallsvariablen.

haben zunächst folgendes, vielleicht etwas überraschendes Theorem.

Theorem 21.5 (Kumulative Verteilungsfunktionen von kontinuierlichen Zufallsvariablen). ξ sei eine kontinuierliche Zufallsvariable mit WDF p und KVF P . Dann gelten

$$P(x) = \int_{-\infty}^x p(s) ds \text{ und } p(x) = P'(x). \quad (21.40)$$

◦

Beweis. Wir halten zunächst fest, dass weil $\mathbb{P}(\xi = x) = 0$ für alle $x \in \mathbb{R}$ gilt, die KVF von ξ keine Sprünge hat, d.h. P ist stetig. Mit den Definitionen von WDF und KVF, folgt, dass P die Form einer Stammfunktion von p hat. Dass p die Ableitung von P ist folgt dann direkt aus dem ersten Hauptsatz der Differential- und Integralrechnung, **1. Hauptsatz der Differential- und Integralrechnung**. $\square$

Für kontinuierliche Zufallsvariablen gilt also, dass die KVF der Zufallsvariable eine Stammfunktion der entsprechenden WDF ist, und umgekehrt, dass die WDF die Ableitung der KVF ist. Im Kontext des *Theorem von Radon-Nikodym* wird diese Einsicht auf generelle Maße generalisiert (vgl. Schmidt (2009)). KVFen kontinuierlicher Zufallsvariablen werden auch oft als kumulative Dichtefunktionen (KDFen) bezeichnet.

Als Beispiel betrachten wir die KVF einer normalverteilten Zufallsvariable. Es sei $\xi \sim N(\mu, \sigma^2)$. Dann ist die WDF von ξ bekanntlich durch

$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (21.41)$$

gegeben. Für die KVF von ξ folgt entsprechend, dass

$$P : \mathbb{R} \rightarrow]0, 1[, x \mapsto P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x \exp\left(-\frac{1}{2\sigma^2}(s - \mu)^2\right) ds. \quad (21.42)$$

Interessanterweise kann das definierende Integral der KVF einer normalverteilten Zufallsvariable nur numerisch, nicht aber analytisch berechnet werden. Wir visualisieren ausgewählte WDFen und KVFen von normalverteilten Zufallsvariablen in Abbildung 21.12.

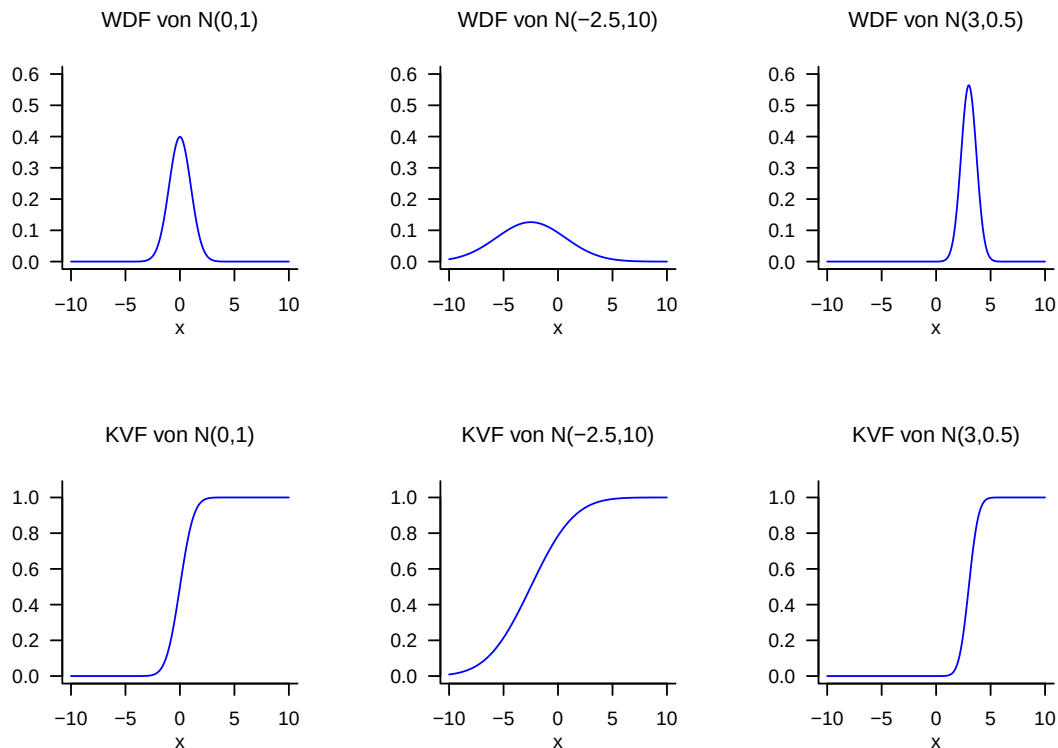


Abbildung 21.12. WDFen und KVFen von normalverteilten Zufallsvariablen.

Inverse Kumulative Verteilungsfunktion

Wir beschließen diesen Abschnitt mit dem Begriff der *inversen kumulativen Verteilungsfunktion*, einem technischen Hilfsmittel, dass insbesondere bei Konfidenzintervallen und Hypothesentests der Frequentistischen Inferenz zur Bestimmung kritischer Werte benutzt wird. Wir definieren die inverse KVF wie folgt.

Definition 21.14 (Inverse Kumulative Verteilungsfunktion). ξ sei eine kontinuierliche Zufallsvariable mit KVF P . Dann heißt die Funktion

$$P^{-1} :]0, 1[\rightarrow \mathbb{R}, q \mapsto P^{-1}(q) := \{x \in \mathbb{R} | P(x) = q\} \quad (21.43)$$

die *inverse kumulative Verteilungsfunktion* von ξ .

•

Nach Definition 21.14 gilt offenbar, dass die Funktion P^{-1} die Umkehrfunktion von P ist, also

$$P^{-1}(P(x)) = x. \quad (21.44)$$

Da bekanntlich gilt, dass

$$P(x) = q \Leftrightarrow \mathbb{P}(\xi \leq x) = q \text{ für } q \in]0, 1[, \quad (21.45)$$

ist $P^{-1}(q)$ also der Wert x von ξ , so dass $\mathbb{P}(\xi \leq x) = q$. Wir visualisieren die KVFen und inversen KVFen normalverteilter Zufallsvariablen in Abbildung 21.13. Im Falle einer normalverteilten Zufallsvariable $\xi \sim N(0, 1)$ gelten beispielsweise, dass

$$P(1.645) = 0.950 \Leftrightarrow P^{-1}(0.950) = 1.645, \quad (21.46)$$

und

$$P(1.906) = 0.975 \Leftrightarrow P^{-1}(0.975) = 1.960. \quad (21.47)$$

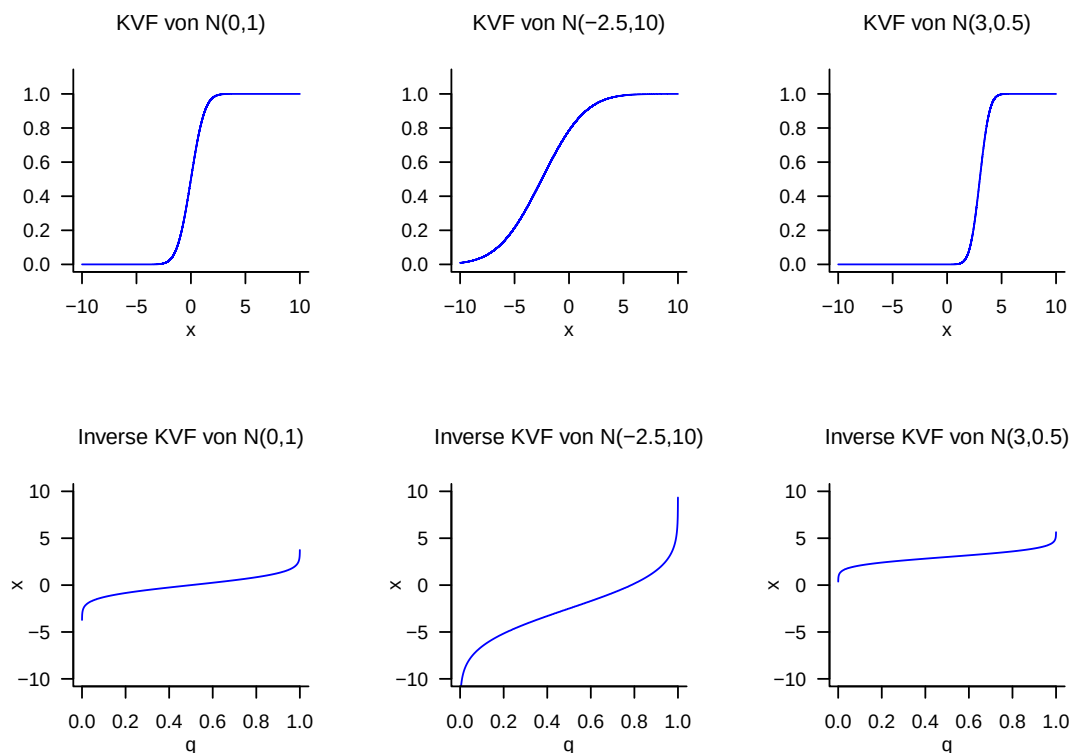


Abbildung 21.13. KVFen und inverse KVFen von normalverteilten Zufallsvariablen.

21.5. Zufallsergebnisse und Zufallsvariablen

Mit den in Kapitel 19 diskutierten *Ergebnissen* und den in diesem Kapitel diskutierten *Werten von Zufallsvariablen* haben wir nun zwei Konzepte kennengelernt, die die Unsicherheit über einen oft numerischen Wert eines Zufallsvorgangs beschreiben können. So kann man sich zum Beispiel die Augenzahl, die ein Würfel beim einmaligen Werfen annimmt,

als Realisierung eines Ergebnisses oder einer Zufallsvariable vorstellen. Tatsächlich gibt es auch keine standardisierte Antwort auf die Frage, ob man einen Zufallsvorgang nun lediglich mit einem Wahrscheinlichkeitsraum oder aber mit einem Wahrscheinlichkeitsraum, einer Zufallsvariable, und dem durch beide induzierten Wahrscheinlichkeitsraum modellieren sollte. In der Anwendung wird meist der Begriff der Zufallsvariable bevorzugt und entsprechende WMFen oder WDFen angegeben, ohne dass ein zugrundeliegender Wahrscheinlichkeitsraum oder die Abbildungsform der Zufallsvariable spezifiziert würde. Folgende Formulierung ist beispielsweise typisch:

ξ sei eine normalverteilte Zufallsvariable mit Erwartungswertparameter μ und Varianzparameter σ^2 .

Implizit werden in dieser Aussage basierend auf der Definition der normalverteilten Zufallsvariable (vgl. Definition 21.9) für ξ der Ergebnisraum $\mathcal{X} := \mathbb{R}$, die σ -Algebra $\mathcal{S} := \mathcal{B}(\mathbb{R})$, und die Verteilung $\mathbb{P}_\xi := N(\mu, \sigma^2)$ festgelegt, also der “induzierte” Wahrscheinlichkeitsraum $(\mathbb{R}, \mathcal{B}(\mathbb{R}), N(\mu, \sigma^2))$ betrachtet. Allerdings bleibt unklar, durch welchen Wahrscheinlichkeitsraum und welche Abbildungsform genau $(\mathbb{R}, \mathcal{B}(\mathbb{R}), N(\mu, \sigma^2))$ nun induziert wurde. Dies kann allerdings immer durch die Annahme, dass ξ die Identitätsfunktion ergänzt werden. Konkret könnte man für den zugrundeliegenden Wahrscheinlichkeitsraum hier $(\mathbb{R}, \mathcal{B}(\mathbb{R}), N(\mu, \sigma^2))$ und für ξ dann

$$\xi : \mathbb{R} \rightarrow \mathbb{R}, \omega \mapsto \xi(\omega) := \omega := x \quad (21.48)$$

wählen. Da ξ dann an dem im Rahmen des zugrundeliegenden Wahrscheinlichkeitsraum zufällig realisiertem Ergebnis ω nichts ändert und dieses lediglich in x umbenannt wird, entspricht die Verteilung von ξ dann direkt dem Wahrscheinlichkeitsmaß des zugrundeliegenden Wahrscheinlichkeitsraums.

Allgemein mag man festhalten, dass das Modellieren von Zufallsvorgängen mithilfe von Zufallsvariablen das elementare Wahrscheinlichkeitsraummodell also als Spezialfall der Zufallsvariable als Identitätsabbildung impliziert, über dies hinausgehend aber die Möglichkeit eröffnet, durch von der Identitätsabbildung verschiedene Zufallsvariablen die Transformation von Wahrscheinlichkeitsmaßen auf unterschiedlichen Messräumen zu formalisieren.

21.6. Literaturhinweise

Die Genese des Begriffs der Zufallsvariablen ist eng mit der Entwicklung der Wahrscheinlichkeitstheorie in den letzten drei Jahrhunderten verflochten, so dass keine für den Begriff entscheidene Publikation angegeben werden kann. Die mathematischen Entwicklung des Begriffs der Normalverteilung durch Abraham De Moivre (1667-1754), Pierre Simon Laplace (1749-1827), Johann Carl Friedrich Gauss (1777-1855) und viele andere, ihre deskriptiv-statistischen Entsprechungen in der empirischen Forschung des 19. Jahrhunderts, sowie ihre multivariate Generalisierung im ausgehenden 19. Jahrhundert werden ausführlich in Stigler (1986) dargestellt.

22. Zufallsvektoren

Zufallsvektoren sind Tupel von Zufallsvariablen. Da jede Zufallsvariable einer Wahrscheinlichkeitsverteilung unterliegt, unterliegt auch ein Zufallsvektor einer Wahrscheinlichkeitsverteilung. Da ein Zufallsvektor aus zwei oder mehr Zufallsvariablen besteht, beschreibt die Verteilung eines Zufallsvektors die *gemeinsame Verteilung* von zwei oder mehr Zufallsvariablen. In diesem Kapitel wollen wir zunächst den Begriff des Zufallsvektors und seiner assoziierten Wahrscheinlichkeitsverteilung, die oft einfach als eine *multivariate Verteilung* bezeichnet wird, einführen. Wir wollen dann diskutieren, wie die Begriffe der WMF, WDF und KVF von Zufallsvariablen auf Zufallsvektoren übertragen werden können (Kapitel 22.1). Mit den *marginalen Verteilungen* und den *bedingten Verteilungen* führen wir dann nachfolgend Begriffe ein, die in der Betrachtung von Zufallsvariablen nicht auftreten. Schließlich führen wir mit dem Begriff der *unabhängigen Zufallsvariablen* das probabilistische Standardmodell für univariate Datensätze ein. Dies stellt einen Spezialfall der gemeinsamen Verteilung der Zufallsvariablen eines Zufallsvektors dar.

22.1. Definition und multivariate Verteilungen

Die Konstruktion und Definition eines Zufallsvektors ist analog zu der einer Zufallsvariable, mit dem Unterschied, dass es sich bei einer Zufallsvariable um eine skalarwertige, bei einem Zufallsvektor dagegen um eine vektorwertige Abbildung auf dem Ergebnisraum eines Wahrscheinlichkeitsraums handelt.

Definition 22.1 (Zufallsvektor). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $(\mathcal{X}, \mathcal{S})$ sei ein n -dimensionaler Messraum. Ein n -dimensionaler *Zufallsvektor* ist definiert als eine Abbildung

$$\xi : \Omega \rightarrow \mathcal{X}, \omega \mapsto \xi(\omega) := \begin{pmatrix} \xi_1(\omega) \\ \vdots \\ \xi_n(\omega) \end{pmatrix} \quad (22.1)$$

mit der *Messbarkeitseigenschaft*

$$\{\omega \in \Omega \mid \xi(\omega) \in S\} \in \mathcal{A} \text{ für alle } S \in \mathcal{S}. \quad (22.2)$$

•

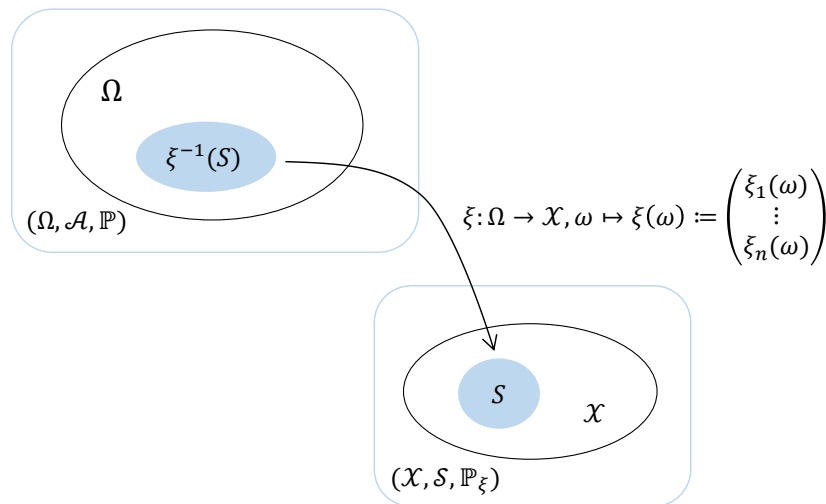
Das Standardbeispiel für den Ergebnisraum eines Zufallsvektors ist $\mathbb{R}^n$, das Standardbeispiel für die auf ihm definierte σ -Algebra ist die n -dimensionale Borelsche σ -Algebra $\mathcal{B}(\mathbb{R}^n)$. Für eine explizite und formale Einführung der n -dimensionalen Borelschen σ -Algebra verweisen wir auf die weiterführende Literatur (z.B. Schmidt (2009)). Wir begnügen uns hier wieder mit der (weiterhin formal falschen) Intuition der n -dimensionalen Borelschen σ -Algebra als Menge aller Teilmengen des $\mathbb{R}^n$. Das Standardbeispiel für einen n -dimensionalen Messraum ist damit $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

Wie bei allen vektorwertigen Funktionen nennen wir die den Zufallsvektor konstituierenden Funktionen ξ_i die *Komponentenfunktionen* von ξ . Legen wir den n -dimensionalen Messraum $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ dem Zufallsvektor zugrunde, so haben diese die Form

$$\xi_i : \Omega \rightarrow \mathbb{R}, \omega \mapsto \xi_i(\omega). \quad (22.3)$$

Ohne Beweis halten wir fest, dass der Zufallsvektor ξ messbar ist, wenn für alle $i = 1, \dots, n$ die Funktionen ξ_i messbar sind und umgekehrt. Damit sind die Komponentenfunktionen eines Zufallsvektors (letztlich nach Definition) Zufallsvariablen. Ein n -dimensionaler Zufallsvektor wird also als eine Konkatenation von n Zufallsvariablen betrachtet, für $n := 1$ entspricht ein Zufallsvektor einer Zufallsvariable.

Zufallsvektoren werden manchmal auch als “multivariate Zufallsvariablen” bezeichnet. Tatsächlich stehen bei der Betrachtung von Zufallsvektoren auch zunächst primär wahrscheinlichkeitstheoretische Aspekte und nicht etwa Aspekte der geometrischen Vektorraumtheorie im Vordergrund. Die Betrachtung von Vektorraumstrukturen ist im Kontext probabilistischer Standardmodelle wie dem Allgemeinen Linearen Modell aber durchaus üblich, so dass wir hier den Begriff des Zufallsvektors bevorzugen (vgl. Christensen (2011)). Trotzdem werden wir Zufallsvektoren, wie in vielen Texten der Probabilistik üblich, auch oft in Zeilenform, also etwa als $\xi := (\xi_1, \dots, \xi_n)$, notieren.



$$\mathbb{P}(\xi^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \in S\}) =: \mathbb{P}_\xi(S)$$

Abbildung 22.1. Konstruktion von Zufallsvektor und multivariater Verteilung.

Das durch die Konstruktion eines Zufallsvektors definierte Bildmaß heißt die *multivariate Verteilung des Zufallsvektors*, wie in folgender Definition ausgeführt (vgl. Abbildung 22.1).

Definition 22.2 (Multivariate Verteilung). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{S})$ sei ein n -dimensionaler Messraum und

$$\xi : \Omega \rightarrow \mathcal{X}, \omega \mapsto \xi(\omega) \quad (22.4)$$

sei ein Zufallsvektor. Dann heißt das Wahrscheinlichkeitsmaß $\mathbb{P}_\xi$, definiert durch

$$\mathbb{P}_\xi : \mathcal{S} \rightarrow [0, 1], S \mapsto \mathbb{P}_\xi(S) := \mathbb{P}(\xi^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \in S\}) \quad (22.5)$$

die *multivariate Verteilung des Zufallsvektor* ξ .

•

Der Einfachheit halber spricht man oft auch nur von *der Verteilung des Zufallsvektors* ξ oder einer *multivariaten Verteilung*. Die Notationskonventionen für Zufallsvariablen Definition 21.3 gelten für Zufallsvektoren analog. Zum Beispiel gelten

$$\begin{aligned}\mathbb{P}_\xi(\xi \in S) &:= \mathbb{P}(\{\xi \in S\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \in S\}) \\ \mathbb{P}_\xi(\xi = x) &:= \mathbb{P}(\{\xi = x\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) = x\}) \\ \mathbb{P}_\xi(\xi \leq x) &:= \mathbb{P}(\{\xi \leq x\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \leq x\}) \\ \mathbb{P}_\xi(x_1 \leq \xi \leq x_2) &:= \mathbb{P}(\{x_1 \leq \xi \leq x_2\}) = \mathbb{P}(\{\omega \in \Omega | x_1 \leq \xi(\omega) \leq x_2\})\end{aligned}\tag{22.6}$$

wobei die Relationsoperatoren $<, \leq, >, \geq$ hier *komponentenweise* verstanden werden. So heißt beispielsweise $x \leq y$ für $x, y \in \mathbb{R}^n$, dass für alle Komponenten $x_i, y_i, i = 1, \dots, n$ gilt, dass $x_i \leq y_i$. Eben dieser Konvention folgt auch die Definition der *multivariaten kumulativen Verteilungsfunktion* in Generalisierung von Definition 21.13. Wie bei den Zufallsvariablen werden die qualifizierenden Subskripte bezüglich eines speziellen Zufallsvektors im Sinne der Multimethode meist weggelassen. Wir nutzen diese Konvention schon in folgender Definition der multivariaten kumulativen Verteilungsfunktion eines Zufallsvektors ξ .

Definition 22.3 (Multivariate kumulative Verteilungsfunktionen). ξ sei ein Zufallsvektor mit Ergebnisraum $\mathcal{X}$. Dann heißt eine Funktion der Form

$$P : \mathcal{X} \rightarrow [0, 1], x \mapsto P(x) := \mathbb{P}(\xi \leq x)\tag{22.7}$$

multivariate kumulative Verteilungsfunktion von ξ .

•

Wie kumulative Verteilungsfunktionen können auch multivariate kumulative Verteilungsfunktionen zur Definition von multivariaten Verteilungen genutzt werden. Häufiger ist allerdings, wie im univariaten Fall, die Definition multivariater Verteilungen durch *multivariate Wahrscheinlichkeitsmasse* - oder *Wahrscheinlichkeitsdichtefunktionen*. Wir generalisieren die Definitionen diskreter und kontinuierlicher Zufallsvariablen und ihren assoziierten Wahrscheinlichkeitsmasse- und Wahrscheinlichkeitsdichtefunktionen wie folgt (vgl. Definition 21.4 und Definition 21.8).

Definition 22.4 (Diskreter Zufallsvektor und multivariate Wahrscheinlichkeitsmassefunktion). Ein Zufallsvektor ξ heißt *diskret*, wenn sein Ergebnisraum $\mathcal{X}$ endlich oder abzählbar ist und eine Funktion

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x)\tag{22.8}$$

existiert, für die gilt

- (1) $\sum_{x \in \mathcal{X}} p(x) = 1$
- (2) $\mathbb{P}(\xi = x) = p(x)$ für alle $x \in \mathcal{X}$.

Ein entsprechende Funktion p heißt *multivariate Wahrscheinlichkeitsmassefunktion* (WMF) von ξ .

•

Der Begriff der multivariaten WMF ist offenbar direkt analog zum Begriff der WMF. Wie univariate WMFen sind multivariate WMFen nicht-negativ und normiert. Der Einfachheit halber spricht man oft einfach von *der WMF eines Zufallsvektors* und verzichtet bei ihrer Bezeichnung, wenn der betreffende Zufallsvektor aus dem Kontext klar ist, auf das ξ Subskript, schreibt also oft einfach p anstelle von p_ξ .

Beispiel

Zur Illustration des Begriffs des diskreten Zufallsvektors und seiner WMF wollen wir ein Beispiel betrachten. Dazu sei $\xi := (\xi_1, \xi_2)$ ein Zufallsvektor, der Werte in $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ annimmt, wobei $\mathcal{X}_1 := \{1, 2, 3\}$ und $\mathcal{X}_2 = \{1, 2, 3, 4\}$ seien. Dann entspricht der Ergebnisraum von ξ der in Tabelle 22.1 spezifizierten Menge an Tupeln (x_1, x_2) .

Tabelle 22.1. Stichprobenrealisation

(x_1, x_2)	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$x_1 = 1$	(1, 1)	(1, 2)	(1, 3)	(1, 4)
$x_1 = 2$	(2, 1)	(2, 2)	(2, 3)	(2, 4)
$x_1 = 3$	(3, 1)	(3, 2)	(3, 3)	(3, 4)

Eine exemplarische bivariate WMF der Form

$$p : \{1, 2, 3\} \times \{1, 2, 3, 4\} \rightarrow [0, 1], (x_1, x_2) \mapsto p(x_1, x_2) \quad (22.9)$$

ist dann durch Tabelle 22.2 definiert.

Tabelle 22.2. Gemeinsame Wahrscheinlichkeitsverteilung $p(x_1, x_2)$

$p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$x_1 = 1$	0.1	0.0	0.2	0.1
$x_1 = 2$	0.1	0.2	0.0	0.0
$x_1 = 3$	0.0	0.1	0.1	0.1

Man beachte, dass die so spezifizierte Funktion p den Normiertheits- und Nichtnegativitätsansprüchen an eine WMF genügt. Insbesondere gilt hier

$$\sum_{x \in \mathcal{X}} p(x) = \sum_{x_1=1}^3 \sum_{x_2=1}^4 p(x_1, x_2) = 1. \quad (22.10)$$

Den Begriff des kontinuierlichen Zufallsvektors und der multivariaten Wahrscheinlichkeitsdichtefunktion definieren wir wie folgt.

Definition 22.5 (Kontinuierlicher Zufallsvektor und multivariate Wahrscheinlichkeitsdichtefunktion). Ein Zufallsvektor ξ heißt *kontinuierlich*, wenn sein Ergebnisraum durch $\mathbb{R}^n$ gegeben ist und eine Funktion

$$p : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, x \mapsto p(x), \quad (22.11)$$

existiert, für die gilt, dass

- (1) $\int_{\mathbb{R}^n} p(x) dx = 1$ und
- (2) $\mathbb{P}(x_1 \leq \xi \leq x_2) = \int_{x_{1_1}}^{x_{2_1}} \cdots \int_{x_{1_n}}^{x_{2_n}} p(s_1, \dots, s_n) ds_1 \cdots ds_n$.

Eine entsprechende Funktion p heißt *multivariate Wahrscheinlichkeitsdichtefunktion (WDF) von ξ* .

•

Offenbar ist der Begriff der multivariaten WDF eines kontinuierlichen Zufallsvektors analog zum Begriff der WDF einer kontinuierlichen Zufallsvariable und wie univariate WDFen sind multivariate WDFen nicht-negativ und normiert. Der Einfachheit halber spricht man auch hier oft einfach von *multivariaten WDFen* und verzichtet auf die den Zufallsvektor identifizierenden Subskripte. Wie für kontinuierliche Zufallsvariablen gilt für kontinuierliche Zufallsvektoren

$$\mathbb{P}(\xi = x) = \mathbb{P}(x \leq \xi \leq x) = \int_{x_1}^{x_1} \cdots \int_{x_n}^{x_n} p(s_1, \dots, s_n) ds_1 \cdots ds_n = 0. \quad (22.12)$$

Das Standardbeispiel für eine multivariate WDF ist die *multivariate Normalverteilung*, welcher wir ein eigenes Kapitel widmen.

22.2. Marginalverteilungen

Hat man die Verteilung eines Zufallsvektors spezifiziert, so kann man sich fragen, welche Verteilungen daraus für die einzelnen Komponenten des Zufallsvektors, also die Zufallsvariablen, die zusammen den Zufallsvektor bilden, folgen. Im Kontext eines Zufallsvektors nennt man diese die *univariaten Marginalverteilungen* des Zufallsvektors. Folgende Definition ist grundlegend.

Definition 22.6 (Univariate Marginalverteilung). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{S})$ sei ein n -dimensionaler Messraum, $\xi : \Omega \rightarrow \mathcal{X}$ sei ein Zufallsvektor, $\mathbb{P}$ sei die Verteilung von ξ , $\mathcal{X}_i \subset \mathcal{X}$ sei der Ergebnisraum der i ten Komponente ξ_i von ξ , und $\mathcal{S}_i$ sei eine σ -Algebra auf ξ_i . Dann heißt die durch

$$\mathbb{P}_{\xi_i} : \mathcal{S}_i \rightarrow [0, 1], S \mapsto \mathbb{P}(\mathcal{X}_1 \times \cdots \times \mathcal{X}_{i-1} \times S \times \mathcal{X}_{i+1} \times \cdots \times \mathcal{X}_n) \text{ für } S \in \mathcal{S}_i \quad (22.13)$$

definierte Verteilung die *ite univariate Marginalverteilung von ξ* .

•

Konkret kann man sowohl für diskrete als auch für kontinuierliche Zufallsvektoren die WMFen bzw. WDFen ihrer Komponenten direkt aus der entsprechenden multivariaten WMF bzw. WDF bestimmen. Dies ist die Aussage folgenden Theorems.

Theorem 22.1 (Marginale Wahrscheinlichkeitsmasse und Wahrscheinlichkeitsdichtefunktionen). (1) $\xi = (\xi_1, \dots, \xi_n)$ sei ein n -dimensionaler diskreter Zufallsvektor mit WMF p und Komponentenergebnisräumen $\mathcal{X}_1, \dots, \mathcal{X}_n$. Dann ergibt sich die WMF der i ten Komponente ξ_i von ξ als

$$p : \mathcal{X}_i \rightarrow [0, 1], x_i \mapsto p(x_i) := \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n). \quad (22.14)$$

(2) $\xi = (\xi_1, \dots, \xi_n)$ sei ein n -dimensionaler kontinuierlicher Zufallsvektor mit WDF p und Komponentenergebnisraum $\mathbb{R}$. Dann ergibt sich die WDF der i ten Komponente ξ_i von ξ als

$$p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, x_i \mapsto p(x_i) := \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_n} p(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n. \quad (22.15)$$

◦

Die WMFen der univariaten Marginalverteilungen diskreter Zufallsvektoren ergeben sich also durch Summation über alle Werte der zu der jeweils betrachteten Zufallsvariable komplementären Zufallsvariablen und die WDFen der univariaten Marginalverteilungen kontinuierlicher Zufallsvektoren ergeben sich analog durch Integration über alle Werte der zu der jeweils betrachteten Zufallsvariable komplementären Zufallsvariablen. Für einen Beweis von Theorem 22.1 verweisen wir auf die weiterführende Literatur.

Beispiel

In Fortführung des in Kapitel 22.1 betrachteten Beispiels eines zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ ergeben sich für die dort definierte WMF für die marginalen WMFen die an den Rändern der von Tabelle 22.3 aufgelisteten WMFen anhand von

$$p(x_1) = \sum_{x_2=1}^4 p(x_1, x_2) \text{ und } p(x_2) = \sum_{x_1=1}^3 p(x_1, x_2). \quad (22.16)$$

Tabelle 22.3. Gemeinsame Wahrscheinlichkeitsverteilung $p(x_1, x_2)$ mit Randverteilungen

$p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$p(x_1)$
$x_1 = 1$	0.1	0.0	0.2	0.1	0.4
$x_1 = 2$	0.1	0.2	0.0	0.0	0.3
$x_1 = 3$	0.0	0.1	0.1	0.1	0.3
$p(x_2)$	0.2	0.3	0.3	0.2	

Für die Werte $p(x_1)$ werden die entsprechenden Werte von $p(x_1, x_2)$ also zeilenweise und für die Werte von $p(x_2)$ spaltenweise addiert. Man beachte, dass aus der Normiertheit von $p(x_1, x_2)$ die Normiertheit von $p(x_1)$ und $p(x_2)$ direkt folgt, da sich die Gesamtsumme an Wahrscheinlichkeitsmasse nicht ändert:

$$1 = \sum_{x_1=1}^3 \sum_{x_2=1}^4 p(x_1, x_2) = \sum_{x_1=1}^3 p(x_1) = \sum_{x_2=1}^4 p(x_2). \quad (22.17)$$

Ein Realisierungsbeispiel mithilfe relativer Häufigkeiten mag den Begriff der marginalen WMF intuitiv verdeutlichen. Nehmen wir an, wir hätten $n = 100$ unabhängige Realisierungen von ξ vorliegen. Um die Wahrscheinlichkeiten $p(x_1, x_2)$ zu schätzen, würden wir die Anzahl der Realisierungen von (x_1, x_2) zählen und durch n teilen. Hätten wir beispielsweise 12 Realisierungen von $(3, 2)$ vorliegen, so würden wir $p(3, 2) \approx 12/100 = 0.12$ schätzen.

Die Frage nach der marginalen Wahrscheinlichkeit von $x_2 = 2$ entspräche dann der Frage, wie oft unter den Realisierungen solche zu finden sind, bei denen $x_2 = 2$ ist, irrespektive des Wertes von x_1 . Dies wäre gerade die Anzahl der Realisierungen der Form $(1, 2)$, $(2, 2)$ und $(3, 2)$. Gäbe es von diesen beispielsweise 0, 22 und 12 respektive, so würde man die Wahrscheinlichkeit $p(2)$ natürlicherweise durch

$$\frac{0 + 22 + 12}{100} = \frac{0}{100} + \frac{22}{100} + \frac{12}{100} = 0.00 + 0.22 + 0.12 = 0.34 \quad (22.18)$$

schätzen. Anstelle der Wahrscheinlichkeiten $p(1, 2)$, $p(2, 2)$, $p(3, 2)$ addiert man hier also die entsprechenden relativen Häufigkeiten.

Marginale Verteilungen im Fall von kontinuierlichen Zufallsvektoren behandeln wir am Standardbeispiel der multivariaten Normalverteilung in **sec-multivariate-normalverteilungen**.

22.3. Bedingte Verteilungen

Hat man die Verteilung eines Zufallsvektors spezifiziert, so kann man sich fragen, welche Verteilung daraus für eine einzelne Komponenten des Zufallsvektors folgt, wenn man den Wert einer anderen Komponente als bekannt annimmt. Dies führt auf den Begriff der *bedingten Verteilung*, welcher sich natürlicherweise aus dem Begriff der bedingten Wahrscheinlichkeit (vgl. Kapitel 20.2) ergibt. Wir erinnern uns zunächst, dass für einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ und zwei Ereignisse $A, B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$ die bedingte Wahrscheinlichkeit von A gegeben B definiert ist als

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \quad (22.19)$$

Analog wird für zwei Zufallsvariablen ξ_1, ξ_2 mit Ereignisräumen $\mathcal{X}_1, \mathcal{X}_2$ und (messbaren) Mengen $S_1 \in \mathcal{X}_1, S_2 \in \mathcal{X}_2$ die bedingte Verteilung von ξ_1 gegeben ξ_2 mithilfe der Ereignisse

$$A := \{\xi_1 \in S_1\} \text{ und } B := \{\xi_2 \in S_2\} \quad (22.20)$$

definiert. So ergibt sich zum Beispiel die bedingte Wahrscheinlichkeit, dass $\xi_1 \in S_1$ gegeben, dass $\xi_2 \in S_2$ und unter der Annahme, dass $\mathbb{P}(\{\xi_2 \in S_2\}) > 0$, zu

$$\mathbb{P}(\{\xi_1 \in S_1\}|\{\xi_2 \in S_2\}) = \frac{\mathbb{P}(\{\xi_1 \in S_1\} \cap \{\xi_2 \in S_2\})}{\mathbb{P}(\{\xi_2 \in S_2\})}. \quad (22.21)$$

Wir betrachten zunächst die Definition der bedingten Verteilungen von diskreten Zufallsvektoren, die lediglich aus zwei Zufallsvariablen bestehen.

Definition 22.7 (Bedingte Wahrscheinlichkeitsmassefunktion und diskrete bedingte Verteilung). $\xi := (\xi_1, \xi_2)$ sei ein diskreter Zufallsvektor mit Ergebnisraum $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$, WMF $p(x_1, x_2)$ und marginalen WMFen $p(x_1)$ und $p(x_2)$. Die bedingte WMF von ξ_1 gegeben $\xi_2 = x_2$ ist dann für $p(x_2) > 0$ definiert als

$$p : \mathcal{X}_1 \rightarrow [0, 1], x_1 \mapsto p(x_1|x_2) := \frac{p(x_1, x_2)}{p(x_2)} \quad (22.22)$$

Analog ist für $p(x_1) > 0$ die bedingte WMF von ξ_2 gegeben $\xi_1 = x_1$ definiert als

$$p : \mathcal{X}_2 \rightarrow [0, 1], x_2 \mapsto p(x_2|x_1) := \frac{p(x_1, x_2)}{p(x_1)} \quad (22.23)$$

Die bedingten Verteilungen mit WMFen $p(x_1|x_2)$ und $p(x_2|x_1)$ heißen dann die *diskreten bedingten Verteilungen von ξ_1 gegeben $\xi_2 = x_2$ bzw. ξ_2 gegeben $\xi_1 = x_1$* .

•

In Analogie zur Definition der bedingten Wahrscheinlichkeit von Ereignissen gilt also

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{\mathbb{P}(\{\xi_1 = x_1\} \cap \{\xi_2 = x_2\})}{\mathbb{P}(\{\xi_2 = x_2\})}. \quad (22.24)$$

Es ist dabei entscheidend zu erkennen, dass bedingte Verteilungen lediglich normalisierte gemeinsame Verteilungen sind.

Beispiel

In Fortführung des in Kapitel 22.1 betrachteten Beispiels eines zweidimensionalen Zufallsvektors $\xi := (\xi_1, \xi_2)$ und seiner in Kapitel 22.2 bestimmten Marginalverteilungen ergeben sich die in Tabelle 22.4 dargestellten bedingte WMFen für $p(x_1|x_2)$ und die in Tabelle 22.5 dargestellten bedingte WMFen für $p(x_2|x_1)$.

Tabelle 22.4. Bedingte Wahrscheinlichkeitsverteilungen $p(x_1|x_2)$

$p(x_1 x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$p(x_1 = 1 x_2)$	$\frac{0.1}{0.2} = 0.5$	$\frac{0.0}{0.3} = 0.0$	$\frac{0.2}{0.3} = 0.6\bar{6}$	$\frac{0.1}{0.2} = 0.5$
$p(x_1 = 2 x_2)$	$\frac{0.1}{0.2} = 0.5$	$\frac{0.2}{0.3} = 0.6\bar{6}$	$\frac{0.0}{0.3} = 0.0$	$\frac{0.0}{0.2} = 0.0$
$p(x_1 = 3 x_2)$	$\frac{0.0}{0.2} = 0.0$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.1}{0.2} = 0.5$

Tabelle 22.5. Bedingte Wahrscheinlichkeitsverteilungen $p(x_2|x_1)$

$p(x_2 x_1)$	$p(x_2 = 1 x_1)$	$p(x_2 = 2 x_1)$	$p(x_2 = 3 x_1)$	$p(x_2 = 4 x_1)$
$x_1 = 1$	$\frac{0.1}{0.4} = 0.25$	$\frac{0.0}{0.4} = 0.00$	$\frac{0.2}{0.4} = 0.50$	$\frac{0.1}{0.4} = 0.25$
$x_1 = 2$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.2}{0.3} = 0.6\bar{6}$	$\frac{0.0}{0.3} = 0.00$	$\frac{0.0}{0.3} = 0.00$
$x_1 = 3$	$\frac{0.0}{0.3} = 0.00$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.1}{0.3} = 0.3\bar{3}$

Man beachte, dass zum einen gilt, dass

$$\sum_{x_1=1}^3 p(x_1|x_2) = 1 \text{ für alle } x_2 \in \mathcal{X}_2 \text{ und } \sum_{x_2=1}^4 p(x_2|x_1) = 1 \text{ für alle } x_1 \in \mathcal{X}_1, \quad (22.25)$$

die bedingten WMFen sind also normiert. Zum anderen beachte man die qualitative Ähnlichkeit der WMFen $p(x_1, x_2)$ und $p(x_1|x_2)$ bzw. $p(x_2|x_1)$, die sich einfach daraus ergibt, dass einerseits $p(x_1, x_2)$ und $p(x_1|x_2)$ für alle $x_2 \in \mathcal{X}_2$ bis auf den gemeinsamen Skalierungsfaktor $1/p(x_2)$ und andererseits $p(x_1, x_2)$ und $p(x_2|x_1)$ für alle $x_1 \in \mathcal{X}_1$ bis auf den gemeinsamen Skalierungsfaktor $1/p(x_1)$ identisch sind. Im Fall eines kontinuierlichen Zufallsvektors sind die analogen bedingten WDFen definiert wie folgt.

Definition 22.8 (Bedingte Wahrscheinlichkeitsdichtefunktion und kontinuierliche bedingte Verteilung). $\xi := (\xi_1, \xi_2)$ sei ein kontinuierlicher Zufallsvektor mit Ergebnisraum $\mathbb{R}^2$, WDF $p(x_1, x_2)$ und marginalen WDFen $p(x_1)$ und $p(x_2)$. Die bedingte WDF von ξ_1 gegeben $\xi_2 = x_2$ ist dann für $p(x_2) > 0$ definiert als

$$p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, x_1 \mapsto p(x_1|x_2) := \frac{p(x_1, x_2)}{p(x_2)} \quad (22.26)$$

Analog ist für $p(x_1) > 0$ die bedingte WMF von ξ_2 gegeben $\xi_1 = x_1$ definiert als

$$p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, x_2 \mapsto p(x_2|x_1) := \frac{p(x_1, x_2)}{p(x_1)} \quad (22.27)$$

Die Verteilungen mit WDFen p und p heißen dann die *kontinuierlichen bedingten Verteilungen von ξ_1 gegeben $\xi_2 = x_2$ und ξ_2 gegeben $\xi_1 = x_1$* , respektive.

•

Man beachte, dass im kontinuierlichen Fall zwar $\mathbb{P}(\xi = x) = 0$, aber nicht notwendig auch $p(x) = 0$ gilt. Die bedingten Verteilungen multivariater Normalverteilungen diskutieren wir in einem eigenen Kapitel.

22.4. Unabhängige Zufallsvariablen

Ähnlich wie die bedingte Wahrscheinlichkeiten von Ereignissen lässt sich auch das Konzept der unabhängigen Ereignisse auf Zufallsvektoren übertragen. Wir definieren zunächst den Begriff der unabhängigen Zufallsvariablen.

Definition 22.9 (Unabhängige Zufallsvariablen). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $\xi := (\xi_1, \xi_2)$ ein zweidimensionaler Zufallsvektor. Die Zufallsvariablen ξ_1, ξ_2 mit Ergebnisräumen $\mathcal{X}_1, \mathcal{X}_2$ heißen *unabhängig*, wenn für alle $S_1 \subseteq \mathcal{X}_1$ und $S_2 \subseteq \mathcal{X}_2$ gilt, dass

$$\mathbb{P}(\xi_1 \in S_1, \xi_2 \in S_2) = \mathbb{P}(\xi_1 \in S_1)\mathbb{P}(\xi_2 \in S_2). \quad (22.28)$$

•

Definition 22.9 besagt, dass die Ereignisse $\{\xi_1 \in S_1\}$ und $\{\xi_2 \in S_2\}$ unabhängig sind. Es gilt also auch, dass

$$\mathbb{P}(\{\xi_1 \in S_1\} | \{\xi_2 \in S_2\}) = \mathbb{P}(\{\xi_1 \in S_1\}) \quad (22.29)$$

und das Wissen um das Eintreten des Ereignisses $\{\xi_2 \in S_2\}$ verändert die Wahrscheinlichkeit des Ereignisses $\{\xi_1 \in S_1\}$ nicht. Das Faktorisierungsprinzip zur Modellierung probabilistischer Unabhängigkeit überträgt sich auf WMFen und WDFen von Zufallsvektoren. Dies ist die Aussagen folgenden Theorems.

Theorem 22.2 (Unabhängigkeit und Faktorisierung).

(1) $\xi := (\xi_1, \xi_2)$ sei ein diskreter Zufallsvektor mit Ergebnisraum $\mathcal{X}_1 \times \mathcal{X}_2$, WMF $p(x_1, x_2)$ und marginalen WMFen $p(x_1), p(x_2)$. Dann gilt

$$\begin{aligned} \xi_1 \text{ und } \xi_2 \text{ sind unabhängige Zufallsvariablen} &\Leftrightarrow \\ p(x_1, x_2) &= p(x_1)p(x_2) \text{ für alle } (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2. \end{aligned} \quad (22.30)$$

(2) $\xi := (\xi_1, \xi_2)$ sei ein kontinuierlicher Zufallsvektor mit Ergebnisraum $\mathbb{R}^2$, WDF p und marginalen WDFen p, p . Dann gilt

$$\begin{aligned} \xi_1 \text{ und } \xi_2 \text{ sind unabhängige Zufallsvariablen} &\Leftrightarrow \\ p(x_1, x_2) &= p(x_1)p(x_2) \text{ für alle } (x_1, x_2) \in \mathbb{R}^2. \end{aligned} \quad (22.31)$$

◦

Generell ist die Unabhängigkeit zweier Zufallsvariablen also äquivalent zur Faktorisierung ihrer gemeinsamen WMF oder WDF. Für einen Beweis von Theorem 22.2 verweisen wir auf die weiterführende Literatur. Theorem 22.2 für weite Aspekte der probabilistischen Modellierung grundlegend.

Beispiel

Wir betrachten erneut den zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ aus Kapitel 22.1, dessen gemeinsame und marginale WMFen die in Tabelle 22.3 spezifizierten Formen haben. Wir fragen zunächst, ob ξ_1 und ξ_2 unabhängig sind. Dies ist nicht der Fall, da hier gilt, dass

$$p(1, 1) = 0.10 \neq 0.08 = 0.40 \cdot 0.20 = p(1)p(1). \quad (22.32)$$

Möchten wir basierend auf den Marginalverteilungen von ξ eine gemeinsame Verteilung erzeugen, in der ξ_1 und ξ_2 unabhängig sind, so muss sich jeder Eintrag der gemeinsamen Verteilung $p(\xi_1, \xi_2)$ aus dem jeweiligen Produkt der Marginalwahrscheinlichkeiten ergeben. Die gemeinsame Verteilung von ξ_1 und ξ_2 unter dieser Annahme der Unabhängigkeit von ξ_1 und ξ_2 bei gleichen Marginalverteilungen wie in Tabelle 22.3 ergibt sich zu der in Tabelle 22.6 gemeinsamen WMF.

Tabelle 22.6. Gemeinsame WMF bei Annahme der Unabhängigkeit von ξ_1 und ξ_2

$p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$p(x_1)$
$x_1 = 1$	0.08	0.12	0.12	0.08	0.40
$x_1 = 2$	0.06	0.09	0.09	0.06	0.30
$x_1 = 3$	0.06	0.09	0.09	0.06	0.30
$p(x_2)$	0.20	0.30	0.30	0.20	

Weiterhin ergeben sich im Falle der Unabhängigkeit von ξ_1 und ξ_2 beispielsweise die bedingten WMFen $p(x_2|x_1)$ wie in Tabelle 22.7 dargestellt. Im Falle der Unabhängigkeit von ξ_1 und ξ_2 ändert sich die Verteilung von ξ_2 gegeben (oder im Wissen um) den Wert von ξ_1 also nicht und entspricht jeweils der Marginalverteilung von ξ_2 . Dies entspricht natürlich der Intuition der Unabhängigkeit von Ereignissen im Kontext elementarer Wahrscheinlichkeiten.

Tabelle 22.7. Bedingte WMFen bei Annahme der Unabhängigkeit von ξ_1 und ξ_2

$p(x_2 x_1)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$p(x_2 x_1 = 1)$	$\frac{0.08}{0.40} = 0.2$	$\frac{0.12}{0.40} = 0.3$	$\frac{0.12}{0.40} = 0.3$	$\frac{0.08}{0.40} = 0.2$
$p(x_2 x_1 = 2)$	$\frac{0.06}{0.30} = 0.2$	$\frac{0.09}{0.30} = 0.3$	$\frac{0.09}{0.30} = 0.3$	$\frac{0.06}{0.30} = 0.2$
$p(x_2 x_1 = 3)$	$\frac{0.06}{0.30} = 0.2$	$\frac{0.09}{0.30} = 0.3$	$\frac{0.09}{0.30} = 0.3$	$\frac{0.06}{0.30} = 0.2$

Wir wollen abschließend den Begriff der unabhängigen Zufallsvariablen für mehr als zwei Zufallsvariablen definieren.

Definition 22.10 (n unabhängige Zufallsvariablen). $\xi := (\xi_1, \dots, \xi_n)$ sei ein n -dimensionaler Zufallsvektor mit Ergebnisraum $\mathcal{X} = \times_{i=1}^n \mathcal{X}_i$. Die n Zufallsvariablen $\xi_1, \dots, \xi_n$ heißen *unabhängig*, wenn für alle $S_i \subseteq \mathcal{X}_i, i = 1, \dots, n$ gilt, dass

$$\mathbb{P}(\xi_1 \in S_1, \dots, \xi_n \in S_n) = \prod_{i=1}^n \mathbb{P}(\xi_i \in S_i). \quad (22.33)$$

Wenn der Zufallsvektor eine n -dimensionale WMF oder WDF p mit marginalen WMFen oder WDFen $p_{\xi_i}, i = 1, \dots, n$ besitzt, dann ist die Unabhängigkeit von $\xi_1, \dots, \xi_n$ gleichbedeutend mit der Faktorisierung der gemeinsamen WMF oder WDF, also mit

$$p(\xi_1, \dots, \xi_n) = \prod_{i=1}^n p(x_i). \quad (22.34)$$

•

Es handelt bei Definition 22.10 also um eine direkte Generalisierung des zweidimensionalen Falls.

Sind n Zufallsvariablen nicht nur unabhängig, sondern haben sie auch alle die gleiche Verteilung, so nennt man sie *unabhängig und identisch verteilt (u.i.v.)*.

Definition 22.11 (Unabhängig und identisch verteilte Zufallsvariablen). n Zufallsvariablen $\xi_1, \dots, \xi_n$ heißen *unabhängig und identisch verteilt (u.i.v.)*, wenn

- (1) $\xi_1, \dots, \xi_n$ unabhängige Zufallsvariablen sind, und
- (2) die Marginalverteilungen der ξ_i übereinstimmen, also gilt, dass

$$\mathbb{P}(\xi_i) = \mathbb{P}(\xi_j) \text{ für alle } 1 \leq i, j \leq n. \quad (22.35)$$

Wenn die Zufallsvariablen $\xi_1, \dots, \xi_n$ unabhängig und identisch verteilt sind und die i te Marginalverteilung $\mathbb{P}$ ist, so schreibt man auch

$$\xi_1, \dots, \xi_n \sim \mathbb{P}. \quad (22.36)$$

•

Man sagt kurz, dass $\xi_1, \dots, \xi_n$ *u.i.v.* sind. Im Englischen spricht man von *independent and identically distributed (i.i.d)* Zufallsvariablen. U.i.v. Zufallsvariablen spielen an vielen Stellen der probabilistischen Modellierung eine wichtige Rolle. So werden, wie wir an späterer Stelle sehen werden, additive Fehlerterme in probabilistischen Modellen meist durch u.i.v. Zufallsvariablen modelliert.

Schließlich halten wir fest, dass n u.i.v. normalverteilte Zufallsvariablen als

$$\xi_1, \dots, \xi_n \sim N(\mu, \sigma^2) \quad (22.37)$$

geschrieben werden. In einem eigenen Kapitel zeigen wir, wie genau die gemeinsame Verteilung von n u.i.v. normalverteilten Zufallsvariablen im Sinne eines Zufallsvektors beschaffen ist.

23. Erwartungswerte

23.1. Definition

Definition 23.1 (Erwartungswert einer Zufallsvariable). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und ξ sei eine Zufallsvariable. Dann ist der *Erwartungswert von ξ* definiert als

- $\mathbb{E}(\xi) := \sum_{x \in \mathcal{X}} x p(x)$, wenn $\xi : \Omega \rightarrow \mathcal{X}$ diskret mit WMF p ist.
- $\mathbb{E}(\xi) := \int_{-\infty}^{\infty} x p(x) dx$, wenn $\xi : \Omega \rightarrow \mathbb{R}$ kontinuierlich mit WDF p ist.

•

Der Erwartungswert ist also eine skalare Zusammenfassung der Verteilung einer Zufallsvariable. Eine Definition des Erwartungswertes, die ohne eine Fallunterscheidung in kontinuierliche und diskrete Zufallsvariablen auskommt, ist möglich, erfordert aber mit der Einführung des Lebesgue-Integrals einigen technischen Aufwand. Wir verweisen dafür auf die weiterführende Literatur (vgl. Schmidt (2009), Meintrup & Schäffler (2005)). Intuitiv entspricht der Erwartungswert einer Zufallsvariable dem im langfristigen Mittel zu erwartenden Wert der Zufallsvariable. Wir betrachten drei Beispiele für Definition 23.1.

Beispiele

Theorem 23.1 (Erwartungswert einer diskreten Zufallsvariable). ξ sei eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := \{-1, 0, 1\}$ und WMF

$$p(-1) = \frac{1}{4}, \quad p(0) = \frac{1}{2}, \quad p(1) = \frac{1}{4}. \quad (23.1)$$

Dann gilt

$$\mathbb{E}(\xi) = 0. \quad (23.2)$$

◦

Beweis. Nach Definition 23.1 ergibt sich

$$\begin{aligned} \mathbb{E}(\xi) &= \sum_{x \in \mathcal{X}} x p(x) \\ &= -1 \cdot p(-1) + 0 \cdot p(0) + 1 \cdot p(1) \\ &= -1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} \\ &= 0. \end{aligned} \quad (23.3)$$

□

Theorem 23.2 (Erwartungswert einer Bernoulli Zufallsvariable). *Es sei $\xi \sim \text{Bern}(\mu)$. Dann gilt $\mathbb{E}(\xi) = \mu$.*

◦

Beweis. ξ ist diskret mit $\mathcal{X} = \{0, 1\}$. Also gilt nach Definition 23.1

$$\begin{aligned}
 &= \sum_{x \in \{0, 1\}} x \text{Bern}(x; \mu) \\
 &= 0 \cdot \mu^0(1 - \mu)^{1-0} + 1 \cdot \mu^1(1 - \mu)^{1-1} \\
 &= 1 \cdot \mu^1(1 - \mu)^0 \\
 &= \mu.
 \end{aligned} \tag{23.4}$$

□

Theorem 23.3 (Erwartungswert einer normalverteilten Zufallsvariable). *Es sei $\xi \sim N(\mu, \sigma^2)$. Dann gilt $\mathbb{E}(\xi) = \mu$.*

◦

Beweis. Wir verzichten auf einen Beweis.

□

In Verallgemeinerung von Definition 23.1 gilt folgende Definition für den Erwartungswert einer Funktion einer Zufallsvariable.

Definition 23.2 (Erwartungswert einer Funktion einer Zufallsvariable). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, ξ sei eine Zufallsvariable mit Ergebnisraum $\mathcal{X}$ und $f : \mathcal{X} \rightarrow \mathcal{Z}$ sei eine Funktion mit Zielmenge $\mathcal{Z}$. Dann ist der *Erwartungswert der Funktion f der Zufallsvariable ξ* definiert als

- $\mathbb{E}(f(\xi)) := \sum_{x \in \mathcal{X}} f(x) p(x)$, wenn $\xi : \Omega \rightarrow \mathcal{X}$ diskret mit WMF p ist,
- $\mathbb{E}(f(\xi)) := \int_{-\infty}^{\infty} f(x) p(x) dx$, wenn $\xi : \Omega \rightarrow \mathbb{R}$ kontinuierlich mit WDF p ist.

•

Der Erwartungswert einer Zufallsvariable ergibt sich anhand von Definition 23.2 als der Spezialfall, in dem gilt, dass

$$f : \mathcal{X} \rightarrow \mathcal{Z}, x \mapsto f(x) := x. \tag{23.5}$$

In der englischsprachigen Literatur ist Definition 23.2 auch als *law of the unconscious statistician* bekannt und wird oft auch direkt zur Definition des Erwartungswertes herangezogen. Folgendes Theorem gibt ein erstes Beispiel für den Erwartungswert einer Funktion einer Zufallsvariable.

Theorem 23.4 (Erwartungswert bei linear-affiner Transformation einer Zufallsvariable). *ξ sei eine Zufallsvariable mit Ergebnisraum $\mathcal{X}$ und es sei*

$$f : \mathcal{X} \rightarrow \mathcal{Z}, x \mapsto f(x) := ax + b \text{ für } a, b \in \mathbb{R} \tag{23.6}$$

eine linear-affine Funktion. Dann gilt

$$\mathbb{E}(f(\xi)) = \mathbb{E}(a\xi + b) = a\mathbb{E}(\xi) + b. \tag{23.7}$$

◦

Beweis. Die Aussage des Theorems folgt mit den Linearitätseigenschaften von Summen und Integralen. Wir betrachten den Fall einer diskreten Zufallsvariable ξ mit endlichem Ergebnisraum $\mathcal{X}$ und WMF p . Es gilt

$$\begin{aligned}
 \mathbb{E}(f(\xi)) &= \mathbb{E}(a\xi + b) \\
 &= \sum_{x \in \mathcal{X}} (ax + b)p(x) \\
 &= \sum_{x \in \mathcal{X}} axp(x) + bp(x) \\
 &= a \sum_{x \in \mathcal{X}} xp(x) + b \sum_{x \in \mathcal{X}} p(x) \\
 &= a\mathbb{E}(\xi) + b.
 \end{aligned} \tag{23.8}$$

□

In Analogie zu Definition 23.2 definiert man für die Funktion eines Zufallsvektors den Erwartungswert dieser Transformation wie folgt.

Definition 23.3 (Erwartungswert einer Funktion eines Zufallsvektors). $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, ξ sei ein Zufallsvektor mit Ergebnisraum $\mathcal{X}$ und $f : \mathcal{X} \rightarrow \mathcal{Z}$ sei eine Funktion mit Zielmenge $\mathcal{Z}$. Dann ist der *Erwartungswert der Funktion f des Zufallsvektors ξ* definiert als

- $\mathbb{E}(f(\xi)) := \sum_{x \in \mathcal{X}} f(x)p(x)$, wenn $\xi : \Omega \rightarrow \mathcal{X}$ diskret mit WMF p ist,
- $\mathbb{E}(f(\xi)) := \int_{-\infty}^{\infty} f(x)p(x)dx$, wenn $\xi : \Omega \rightarrow \mathbb{R}$ kontinuierlich mit WDF p ist.

•

23.2. Eigenschaften

Theorem 23.5 (Erwartungswert bei linear-affiner Kombination). ξ sei ein n -dimensionaler Zufallsvektor mit Komponenten $\xi_1, \dots, \xi_n$ und Ergebnisraum $\mathcal{X} := \mathcal{X}_1 \times \dots \times \mathcal{X}_n$. Weiterhin sei

$$f : \mathcal{X} \rightarrow \mathcal{Z}, x \mapsto f(x) := a_0 + \sum_{i=1}^n a_i x_i \text{ für } a_0, \dots, a_n \in \mathbb{R}. \tag{23.9}$$

eine linear-affine Kombination der Komponenten von ξ . Dann gilt

$$\mathbb{E}(f(\xi)) = \mathbb{E} \left(a_0 + \sum_{i=1}^n a_i \xi_i \right) = a_0 + \sum_{i=1}^n a_i \mathbb{E}(\xi_i). \tag{23.10}$$

◦

Beweis. Wie unten gezeigt, folgt das Theorem mit den Linearitätseigenschaften von Summen und Integralen, sowie Fubini's Theorem zur Vertauschung von Summations- bzw. Integrationsreihenfolge. Zur

Vereinfachung der Notation schreiben wir untenstehend $\sum_{x_i \in \mathcal{X}_i}$ als $\sum_{x_i}$. Es gilt dann

$$\begin{aligned}
 \mathbb{E}(f(\xi)) &= \mathbb{E}\left(a_0 + \sum_{i=1}^n a_i \xi_i\right) \\
 &= \mathbb{E}(a_0 + a_1 \xi_1 + \dots + a_n \xi_n) \\
 &= \sum_{x_1} \dots \sum_{x_n} (a_0 + a_1 x_1 + \dots + a_n x_n) p(x_1, \dots, x_n) \\
 &= \sum_{x_1} \dots \sum_{x_n} a_0 p(x_1, \dots, x_n) + a_1 x_1 p(x_1, \dots, x_n) + \dots + a_n x_n p(x_1, \dots, x_n) \\
 &= \sum_{x_1} \dots \sum_{x_n} a_0 p(x_1, \dots, x_n) + \sum_{x_1} \dots \sum_{x_n} a_1 x_1 p(x_1, \dots, x_n) + \dots + \sum_{x_1} \dots \sum_{x_n} a_n x_n p(x_1, \dots, x_n) \\
 &= a_0 \sum_{x_1} \dots \sum_{x_n} p(x_1, \dots, x_n) + a_1 \sum_{x_1} \dots \sum_{x_n} x_1 p(x_1, \dots, x_n) + \dots + a_n \sum_{x_1} \dots \sum_{x_n} x_n p(x_1, \dots, x_n) \\
 &= a_0 + a_1 \sum_{x_1} x_1 \sum_{x_2} \dots \sum_{x_n} p(x_1, \dots, x_n) + \dots + a_n \sum_{x_n} x_n \sum_{x_{n-1}} \dots \sum_{x_1} p(x_1, \dots, x_n) \\
 &= a_0 + a_1 \sum_{x_1} x_1 p(x_1) + \dots + a_n \sum_{x_n} x_n p(x_n) \\
 &= a_0 + a_1 \mathbb{E}(\xi_1) + \dots + a_n \mathbb{E}(\xi_n) \\
 &= a_0 + \sum_{i=1}^n a_i \mathbb{E}(\xi_i)
 \end{aligned}$$

(23.11)

□

23.3. Bedingter Erwartungswert

Betrachtet man bei der Bildung eines Erwartungswertes nun anstelle der Verteilung einer Zufallsvariable die bedingte Verteilung einer Zufallsvariable, so ist man entsprechend auf den Begriff des bedingten Erwartungswerts geführt.

Definition 23.4 (Bedingter Erwartungswert). Gegeben sei ein Zufallsvektor $\xi := (\xi_1, \xi_2)$ mit Ergebnisraum $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$, WMF oder WDF $p(x_1, x_2)$ und bedingter WMF oder WDF $p(x_1|x_2)$ für alle $x_2 \in \mathcal{X}_2$. Dann ist der *bedingte Erwartungswert* von ξ_1 gegeben $\xi_2 = x_2$ definiert als

- $\mathbb{E}(\xi_1|\xi_2 = x_2) := \sum_{x_1 \in \mathcal{X}_1} x_1 p(x_1|x_2)$, wenn ξ ein diskreter Zufallsvektor ist.
- $\mathbb{E}(\xi_1|\xi_2 = x_2) := \int_{\mathcal{X}_1} x_1 p(x_1|x_2) dx_1$, wenn ξ ein kontinuierlicher Zufallsvektor ist.

•

Der bedingte Erwartungswert einer Zufallsvariable ist also der Erwartungswert einer Zufallsvariable in einer bedingten Verteilung. Man beachte, dass wir in Definition 23.4 den bedingten Erwartungswert für einen festen Wert x_2 von ξ_2 definiert habe. Bei einem festen Wert x_2 von ξ_2 ist $\mathbb{E}(\xi_1|\xi_2 = x_2)$ damit ein fester Wert und durch Austauschen der Subskripte erhält man entsprechend $\mathbb{E}(\xi_2|\xi_1 = x_1)$.

Allgemein ist $\mathbb{E}(\xi_1|\xi_2)$ allerdings eine Zufallsvariable, da ξ_2 eine Zufallsvariable ist und nur mit einer gewissen Wahrscheinlichkeit einen Wert $\xi_2 = x_2$ annimmt. Wir werden später sehen, dass analog zu Definition 23.4 bedingte Varianzen, bedingte Kovarianzen und bedingte Korrelationen definiert werden. In der Psychologie ist der Begriff des bedingten Erwartungswerts zentral, denn in der klassischen Testtheorie sind die wahren Werte von

Personen bei Testmessungen

als bedingte Erwartungswerte definiert. Zur Verdeutlichung von Definition 23.4 betrachten wir ein Beispiel für einen diskreten Zufallsvektor.

Beispiel

Für einen zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$, der Werte in $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ mit $\mathcal{X}_1 := \{1, 2, 3\}$ und $\mathcal{X}_2 = \{1, 2, 3, 4\}$ annehme seien bedingte WMFen für $p(x_2|x_1)$ für alle $x_1 \in \mathcal{X}_1$ wie in Tabelle 23.1 spezifiziert

Tabelle 23.1. Bedingte Wahrscheinlichkeitsverteilung von x_2 gegeben x_1

$p(x_2 x_1)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$p(x_2 x_1 = 1)$	$\frac{1}{4}$	0	$\frac{1}{2}$	$\frac{1}{4}$
$p(x_2 x_1 = 2)$	$\frac{1}{3}$	$\frac{2}{3}$	0	0
$p(x_2 x_1 = 3)$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Dann ergeben sich die bedingten Erwartungswerte

$$\begin{aligned}
 \mathbb{E}(\xi_2|\xi_1 = 1) &= \sum_{x_2 \in \mathcal{X}_2} x_2 p(x_2|x_1 = 1) = 1 \cdot \frac{1}{4} + 2 \cdot 0 + 3 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = \frac{11}{4} \\
 \mathbb{E}(\xi_2|\xi_1 = 2) &= \sum_{x_2 \in \mathcal{X}_2} x_2 p(x_2|x_1 = 2) = 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} + 3 \cdot 0 + 4 \cdot 0 = \frac{5}{3} \\
 \mathbb{E}(\xi_2|\xi_1 = 3) &= \sum_{x_2 \in \mathcal{X}_2} x_2 p(x_2|x_1 = 3) = 1 \cdot 0 + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} = \frac{8}{3}.
 \end{aligned} \tag{23.12}$$

Das folgende Theorem, der sogenannte *Satz vom iterierten Erwartungswert*, stellt einen Zusammenhang zwischen dem (marginalen) Erwartungswert einer Zufallsvariable und dem bedingten Erwartungswert dieser Zufallsvariable her. Das Theorem wird manchmal auch als *Satz vom totalen Erwartungswert* bezeichnet.

Theorem 23.6 (Satz vom iterierten Erwartungswert). *Gegeben sei ein Zufallsvektor $\xi := (\xi_1, \xi_2)$. Dann gilt*

$$\mathbb{E}(\xi_1) = \mathbb{E}(\mathbb{E}(\xi_1|\xi_2)). \tag{23.13}$$

◦

Beweis. Wir beschränken uns auf den Beweis für einen diskreten Zufallsvektor, der Beweis für einen kon-

tinuierlichen Zufallsvektor folgt analog. Es gilt

$$\begin{aligned}
 \mathbb{E}(\xi_1) &= \sum_{x_1 \in \mathcal{X}_1} x_1 p(x_1) \\
 &= \sum_{x_1 \in \mathcal{X}_1} x_1 \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2) \\
 &= \sum_{x_1 \in \mathcal{X}_1} x_1 \sum_{x_2 \in \mathcal{X}_2} p(x_1|x_2)p(x_2) \\
 &= \sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} x_1 p(x_1|x_2)p(x_2) \\
 &= \sum_{x_2 \in \mathcal{X}_2} \sum_{x_1 \in \mathcal{X}_1} x_1 p(x_1|x_2)p(x_2) \\
 &= \sum_{x_2 \in \mathcal{X}_2} \mathbb{E}(\xi_1|\xi_2 = x_2)p(x_2) \\
 &= \mathbb{E}(\mathbb{E}(\xi_1|\xi_2)).
 \end{aligned} \tag{23.14}$$

□

Offenbar bezeichnen die verschiedenen Erwartungswerte in Theorem 23.6 Erwartungswert bezüglich unterschiedlicher Verteilungen: der Erwartungswert auf der linken Seite der Gleichung bezeichnet den Erwartungswert bezüglich der marginalen Verteilung von ξ_1 , der äußere Erwartungswert auf der rechten Seite der Gleichung bezeichnet den Erwartungswert bezüglich der marginalen Verteilung von ξ_2 und der innere Erwartungswert auf der rechten Seite der Gleichung bezeichnet den bedingten Erwartungswert von ξ_1 gegeben ξ_2

24. Varianzen

24.1. Definition

Definition 24.1 (Varianz). ξ sei eine Zufallsvariable mit endlichem Erwartungswert $\mathbb{E}(\xi)$. Dann ist die *Varianz von ξ* definiert als

$$\mathbb{V}(\xi) := \mathbb{E}((\xi - \mathbb{E}(\xi))^2). \quad (24.1)$$

•

Die Varianz einer Zufallsvariable ξ mit Ergebnisraum $\mathcal{X}$ ist nach Definition 23.2 also der Erwartungswert der Funktion

$$f : \mathcal{X} \rightarrow \mathbb{R}, x \mapsto f(x) := (x - \mathbb{E}(\xi))^2 \quad (24.2)$$

und damit die erwartete quadrierte Abweichung einer Zufallsvariable von ihrem Erwartungswert. Die Quadrierung der Abweichung der Zufallsvariable von ihrem Erwartungswert ist dabei nötig, da andernfalls mit Theorem 23.4 immer gelten würde, dass

$$\mathbb{E}(\xi - \mathbb{E}(\xi)) = \mathbb{E}(\xi) - \mathbb{E}(\xi) = 0. \quad (24.3)$$

Intuitiv misst die Varianz also die *Streuung* oder *Variabilität* einer Zufallsvariable. Insbesondere besagt die *Chebyshev Ungleichung*

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq x) \leq \frac{\mathbb{V}(\xi)}{x^2}. \quad (24.4)$$

Beispielweise gilt für eine Zufallsvariable immer, dass die Wahrscheinlichkeit für eine absolute Abweichung vom doppelten ihrer Standardabweichung höchstens $1/4$ ist, also Frequentistisch betrachtet nur für etwa ein Viertel ihrer Realisierungen zutrifft, und die Wahrscheinlichkeit für eine absolute Abweichung vom dreifachen ihrer Standardabweichung höchstens $1/9$ ist, also Frequentistisch betrachtet nur für etwa ein Zehntel ihrer Realisierungen zutrifft, jeweils unabhängig davon, von welcher genauen Form die Verteilung der Zufallsvariable ist. Dies folgt mit der Chebyshev Ungleichung aus

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq 2\sqrt{\mathbb{V}(\xi)}) \leq \frac{\mathbb{V}(\xi)}{(2\sqrt{\mathbb{V}(\xi)})^2} = \frac{1}{4} \quad (24.5)$$

und

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq 3\sqrt{\mathbb{V}(\xi)}) \leq \frac{\mathbb{V}(\xi)}{(3\sqrt{\mathbb{V}(\xi)})^2} = \frac{1}{9}. \quad (24.6)$$

Neben der Varianz gibt es viele weitere Maße für die Variabilität von Zufallsvariablen. Hier seien beispielsweise die erwartete absolute Abweichung einer Zufallsvariable von ihrem Erwartungswert $\mathbb{E}(|\xi - \mathbb{E}(\xi)|)$ und die sogenannte *Entropie* $-\mathbb{E}(\ln p(x))$ genannt. Wir verdeutlichen Definition 24.1 zunächst an einigen Beispielen.

Beispiele

Theorem 24.1. ξ sei eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := \{-1, 0, 1\}$ und WMF

$$p(-1) = \frac{1}{4} \quad p(0) = \frac{1}{2} \quad p(1) = \frac{1}{4}. \quad (24.7)$$

Dann gilt

$$\mathbb{V}(\xi) = \frac{1}{2}. \quad (24.8)$$

◦

Beweis. Nach Definition 24.1 ergibt sich mit dem in Theorem 23.1 bestimmten Erwartungswert derselben Zufallsvariable

$$\begin{aligned} \mathbb{V}(\xi) &= \mathbb{E}((\xi - \mathbb{E}(\xi))^2) \\ &= \mathbb{E}((\xi - 0)^2) \\ &= \mathbb{E}(\xi^2) \\ &= \sum_{x \in \mathcal{X}} x^2 p_{\xi}(x) \\ &= (-1)^2 \cdot p_{\xi}(-1) + 0^2 \cdot p_{\xi}(0) + 1^2 \cdot p_{\xi}(1) \\ &= 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} \\ &= \frac{1}{2}. \end{aligned} \quad (24.9)$$

□

Theorem 24.2 (Varianz einer Bernoulli-Zufallsvariable). *Es sei $\xi \sim \text{Bern}(\mu)$. Dann ist die Varianz von ξ gegeben durch*

$$\mathbb{V}(\xi) = \mu(1 - \mu). \quad (24.10)$$

◦

Beweis. ξ ist eine diskrete Zufallsvariable und es gilt $\mathbb{E}(\xi) = \mu$. Also gilt

$$\begin{aligned} \mathbb{V}(\xi) &= \mathbb{E}((\xi - \mu)^2) \\ &= \sum_{x \in \{0,1\}} (x - \mu)^2 \text{Bern}(x; \mu) \\ &= (0 - \mu)^2 \mu^0 (1 - \mu)^{1-0} + (1 - \mu)^2 \mu^1 (1 - \mu)^{1-1} \\ &= \mu^2 (1 - \mu) + (1 - \mu)^2 \mu \\ &= (\mu^2 + (1 - \mu)\mu) (1 - \mu) \\ &= (\mu^2 + \mu - \mu^2) (1 - \mu) \\ &= \mu(1 - \mu). \end{aligned} \quad (24.11)$$

□

Theorem 24.3 (Varianz einer normalverteilten Zufallsvariable). *Es sei $\xi \sim N(\mu, \sigma^2)$. Dann ist die Varianz von ξ gegeben durch*

$$\mathbb{V}(\xi) = \sigma^2. \quad (24.12)$$

◦

Beweis. Wir verzichten auf einen Beweis. □

24.2. Eigenschaften

Die Varianz hat die Eigenschaft, keine negativen Werte anzunehmen.

Theorem 24.4 (Nichtnegativität der Varianz). ξ sei eine Zufallsvariable. Dann gilt

$$\mathbb{V}(\xi) \geq 0. \quad (24.13)$$

◦

Beweis. Wir betrachten den Fall einer diskreten Zufallsvariable. Dann gilt zunächst

$$(x - \mathbb{E}(\xi))^2 \geq 0 \text{ für alle } x \in \mathcal{X}. \quad (24.14)$$

Weiterhin gilt für die WMF p von ξ , dass

$$p(x) \geq 0 \text{ für alle } x \in \mathcal{X}. \quad (24.15)$$

Also folgt

$$p(x)(x - \mathbb{E}(\xi))^2 \geq 0 \text{ für alle } x \in \mathcal{X}. \quad (24.16)$$

Damit gilt dann aber

$$\mathbb{V}(\xi) = \sum_{x \in \mathcal{X}} p(x)(x - \mathbb{E}(\xi))^2 dx \geq 0, \quad (24.17)$$

Analog zeigt man die Nichtnegativität der Varianz bei kontinuierlichen Zufallsvariablen. $\square$

Das Berechnen von Varianzen wird durch folgendes Theorem, den sogenannten *Varianzverschiebungssatz* oft erleichtert, insbesondere, wenn der Erwartungswert der quadrierten Zufallsvariable leicht zu bestimmen oder bekannt ist.

Theorem 24.5 (Varianzverschiebungssatz). ξ sei eine Zufallsvariable. Dann gilt

$$\mathbb{V}(\xi) = \mathbb{E}(\xi^2) - \mathbb{E}(\xi)^2. \quad (24.18)$$

◦

Beweis. Mit der Definition der Varianz und der Linearität des Erwartungswerts gilt

$$\begin{aligned} \mathbb{V}(\xi) &= \mathbb{E}((\xi - \mathbb{E}(\xi))^2) \\ &= \mathbb{E}(\xi^2 - 2\xi\mathbb{E}(\xi) + \mathbb{E}(\xi)^2) \\ &= \mathbb{E}(\xi^2) - 2\mathbb{E}(\xi)\mathbb{E}(\xi) + \mathbb{E}(\mathbb{E}(\xi)^2) \\ &= \mathbb{E}(\xi^2) - 2\mathbb{E}(\xi)^2 + \mathbb{E}(\xi)^2 \\ &= \mathbb{E}(\xi^2) - \mathbb{E}(\xi)^2. \end{aligned} \quad (24.19)$$

 $\square$

In Analogie zu Theorem 23.4 gilt für die Varianz folgendes Resultat.

Theorem 24.6 (Varianz bei linear-affiner Transformation einer Zufallsvariable). ξ sei eine Zufallsvariable und es sei

$$f : \mathcal{X} \rightarrow \mathcal{Z}, x \mapsto f(x) := ax + b \text{ für } a, b \in \mathbb{R} \quad (24.20)$$

eine linear-affine Funktion. Dann gilt

$$\mathbb{V}(f(\xi)) = \mathbb{V}(a\xi + b) = a^2\mathbb{V}(\xi). \quad (24.21)$$

◦

Beweis. Es ergibt sich

$$\begin{aligned}
 \mathbb{V}(f(\xi)) &= \mathbb{V}(a\xi + b) \\
 &= \mathbb{E}((a\xi + b - \mathbb{E}(a\xi + b))^2) \\
 &= \mathbb{E}((a\xi + b - a\mathbb{E}(\xi) - b)^2) \\
 &= \mathbb{E}((a\xi - a\mathbb{E}(\xi))^2) \\
 &= \mathbb{E}(a(\xi - \mathbb{E}(\xi))^2) \\
 &= \mathbb{E}(a^2(\xi - \mathbb{E}(\xi))^2) \\
 &= a^2\mathbb{E}((\xi - \mathbb{E}(\xi))^2) \\
 &= a^2\mathbb{V}(\xi).
 \end{aligned} \tag{24.22}$$

□

24.3. Bedingte Varianz

Definition 24.2 (Bedingte Varianz). Gegeben sei ein Zufallsvektor $\xi := (\xi_1, \xi_2)$ mit Ergebnisraum $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$. Dann ist die *bedingte Varianz von ξ_1 gegeben $\xi_2 = x_2$* definiert als

$$\mathbb{V}(\xi_1|\xi_2 = x_2) = \mathbb{E}((\xi_1 - \mathbb{E}(\xi_1|\xi_2 = x_2))^2 | \xi_2 = x_2) \tag{24.23}$$

und die *bedingte Standardabweichung von ξ_1 gegeben $\xi_2 = x_2$* ist definiert als

$$\mathbb{S}(\xi_1|\xi_2 = x_2) = \sqrt{\mathbb{V}(\xi_1|\xi_2 = x_2)}. \tag{24.24}$$

•

Die bedingte Varianz ist im Sinne des bedingten Erwartungswerts definiert, ebenso wie ein bedingter Erwartungswert ist eine bedingte Varianz also eine Zufallsvariable. Auch für die bedingte Varianz gilt der Verschiebungssatz, wie folgendes Theorem besagt.

Theorem 24.7 (Kovarianzverschiebungssatz der bedingten Varianz). Gegeben sei ein Zufallsvektor $\xi := (\xi_1, \xi_2)$. Dann gilt

$$\mathbb{V}(\xi_1|\xi_2 = x_2) = \mathbb{E}(\xi_1^2|\xi_2 = x_2) - \mathbb{E}(\xi_1|\xi_2 = x_2)^2 \tag{24.25}$$

◦

Beweis. Wir beschränken uns auf den Beweis für einen diskreten Zufallsvektor. Der Beweis für einen kontinuierlichen Zufallsvektor folgt analog. Mit der Definition des bedingten Erwartungswerts (Definition 23.4) gilt

$$\begin{aligned}
 \mathbb{V}(\xi_1|\xi_2 = x_2) &= \mathbb{E}((\xi_1 - \mathbb{E}(\xi_1|\xi_2 = x_2))^2 | \xi_2 = x_2) \\
 &= \mathbb{E}(\xi_1^2 - 2\xi_1\mathbb{E}(\xi_1|\xi_2 = x_2) + \mathbb{E}(\xi_1|\xi_2 = x_2)^2 | \xi_2 = x_2) \\
 &= \sum_{x_1 \in \mathcal{X}_1} p(x_1|x_2) (x_1^2 - 2x_1\mathbb{E}(\xi_1|\xi_2 = x_2) + \mathbb{E}(\xi_1|\xi_2 = x_2)^2) \\
 &= \sum_{x_1 \in \mathcal{X}_1} x_1^2 p(x_1|x_2) - \sum_{x_1 \in \mathcal{X}_1} 2x_1\mathbb{E}(\xi_1|\xi_2 = x_2)p(x_1|x_2) + \sum_{x_1 \in \mathcal{X}_1} \mathbb{E}(\xi_1|\xi_2 = x_2)^2 p(x_1|x_2) \\
 &= \mathbb{E}(\xi_1^2|\xi_2 = x_2) - 2\mathbb{E}(\xi_1|\xi_2 = x_2) \sum_{x_1 \in \mathcal{X}_1} x_1 p(x_1|x_2) + \mathbb{E}(\xi_1|\xi_2 = x_2)^2 \sum_{x_1 \in \mathcal{X}_1} p(x_1|x_2) \\
 &= \mathbb{E}(\xi_1^2|\xi_2 = x_2) - 2\mathbb{E}(\xi_1|\xi_2 = x_2)\mathbb{E}(\xi_1|\xi_2 = x_2) + \mathbb{E}(\xi_1|\xi_2 = x_2)^2 \cdot 1 \\
 &= \mathbb{E}(\xi_1^2|\xi_2 = x_2) - 2\mathbb{E}(\xi_1|\xi_2 = x_2)^2 + \mathbb{E}(\xi_1|\xi_2 = x_2)^2 \\
 &= \mathbb{E}(\xi_1^2|\xi_2 = x_2) - \mathbb{E}(\xi_1|\xi_2 = x_2)^2
 \end{aligned} \tag{24.26}$$

□

Folgendes Theorem, der sogenannte *Satz von der iterierten Varianz* stellt einen Zusammenhang zwischen der Varianz einer Zufallsvariable, ihrer bedingten Varianz und der Varianz ihres bedingten Erwartungswerts her.

Theorem 24.8 (Satz von der iterierten Varianz). *Gegeben sei ein Zufallsvektor $\xi := (\xi_1, \xi_2)$. Dann gilt*

$$\mathbb{V}(\xi_1) = \mathbb{E}(\mathbb{V}(\xi_1|\xi_2)) + \mathbb{V}(\mathbb{E}(\xi_1|\xi_2)) \quad (24.27)$$

◦

Beweis. Wir beschränken uns auf den Beweis für einen diskreten Zufallsvektor. Der Beweis für einen kontinuierlichen Zufallsvektor folgt analog. Mit dem Kovarianzverschiebungssatz (Theorem 25.2), dem Satz vom iterierten Erwartungswert (Theorem 23.6) und dem Kovarianzverschiebungssatz der bedingten Varianz (Theorem 24.7) gilt

$$\begin{aligned} \mathbb{V}(\xi_1) &= \mathbb{E}(\xi_1^2) - \mathbb{E}(\xi_1)^2 \\ &= \mathbb{E}(\mathbb{E}(\xi_1^2|\xi_2)) - \mathbb{E}(\mathbb{E}(\xi_1|\xi_2))^2 \\ &= \sum_{x_2 \in \mathcal{X}_2} p(x_2) \mathbb{E}(\xi_1^2|\xi_2 = x_2) - \mathbb{E}(\mathbb{E}(\xi_1|\xi_2))^2 \\ &= \sum_{x_2 \in \mathcal{X}_2} p(x_2) (\mathbb{V}(\xi_1|\xi_2 = x_2) + \mathbb{E}(\xi_1|\xi_2 = x_2)^2) - \mathbb{E}(\mathbb{E}(\xi_1|\xi_2))^2 \\ &= \sum_{x_2 \in \mathcal{X}_2} p(x_2) \mathbb{V}(\xi_1|\xi_2 = x_2) + \sum_{x_2 \in \mathcal{X}_2} p(x_2) \mathbb{E}(\xi_1|\xi_2 = x_2)^2 - \mathbb{E}(\mathbb{E}(\xi_1|\xi_2))^2 \\ &= \mathbb{E}(\mathbb{V}(\xi_1|\xi_2)) + (\mathbb{E}(\mathbb{E}(\xi_1|\xi_2)^2) - \mathbb{E}(\mathbb{E}(\xi_1|\xi_2))^2) \\ &= \mathbb{E}(\mathbb{V}(\xi_1|\xi_2)) + \mathbb{V}(\mathbb{E}(\xi_1|\xi_2)) \end{aligned} \quad (24.28)$$

□

25. Kovarianzen

25.1. Definition

Definition 25.1 (Kovarianz). Die *Kovarianz* zweier Zufallsvariablen ξ_1 und ξ_2 ist definiert als

$$\mathbb{C}(\xi_1, \xi_2) := \mathbb{E}((\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2))). \quad (25.1)$$

•

Die Definition der Kovarianz nach Definition 25.1 lässt implizit, dass die in ihr auftauchenden Erwartungswerte unterschiedlicher Natur sind. Dabei geht die Definition der Kovarianz zunächst einmal davon aus, dass ξ_1 und ξ_2 die Komponenten eines Zufallsvektors $\xi := (\xi_1, \xi_2)$ mit Ergebnisraum $\mathcal{X}_1 \times \mathcal{X}_2$, gemeinsamer Verteilung $\mathbb{P}(\xi_1, \xi_2)$ und marginalen Verteilungen $\mathbb{P}(\xi_1)$ und $\mathbb{P}(\xi_2)$ sind. Bei den Erwartungswerten $\mathbb{E}(\xi_1)$ und $\mathbb{E}(\xi_2)$ handelt es sich um die Erwartungswerte der Komponenten ξ_1 und ξ_2 bezüglich dieser Marginalverteilungen. Die Kovarianz selbst ist dann der Erwartungswert der Funktion

$$f : \mathcal{X}_1 \times \mathcal{X}_2 \rightarrow \mathbb{R}, (x_1, x_2) \mapsto f(x_1, x_2) := (x_1 - \mathbb{E}(\xi_1))(x_2 - \mathbb{E}(\xi_2)) \quad (25.2)$$

bezüglich der gemeinsamen Verteilung $\mathbb{P}(\xi_1, \xi_2)$. Wir wollen dies am Beispiel eines diskreten Zufallsvektors verdeutlichen.

Beispiel

Es $\xi := (\xi_1, \xi_2)$ ein diskreter Zufallsvektor mit Ergebnisraum $\mathcal{X} := \{1, 2\} \times \{1, 2, 3\}$ und der in Tabelle 25.1 dargestellten gemeinsamen WMF $p(x_1, x_2)$ und marginalen WMFen $p(x_1)$ und $p(x_2)$.

Tabelle 25.1. Gemeinsame und marginale WMFen des Zufallsvektors ξ

$p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$p(x_1)$
$x_1 = 1$	0.10	0.05	0.15	0.30
$x_1 = 2$	0.60	0.05	0.05	0.70
$p(x_2)$	0.70	0.10	0.20	

Mit der Definition der Kovarianz von ξ_1 und ξ_2 gilt dann unter Beachtung von Gleichung 25.2

$$\begin{aligned}\mathbb{C}(\xi_1, \xi_2) &= \mathbb{E}(f(\xi_1, \xi_2)) \\ &= \sum_{x_1=1}^2 \sum_{x_2=1}^3 f(x_1, x_2) p(x_1, x_2) \\ &= \sum_{x_1=1}^2 \sum_{x_2=1}^3 (x_1 - \mathbb{E}(\xi_1))(x_2 - \mathbb{E}(\xi_2)) p(x_1, x_2)\end{aligned}\quad (25.3)$$

Einsetzen der Werte aus Tabelle 25.1 ergibt dann zunächst

$$\mathbb{E}(\xi_1) = \sum_{x_1=1}^2 x_1 p(x_1) = 1 \cdot 0.3 + 2 \cdot 0.7 = 1.7 \quad (25.4)$$

und

$$\mathbb{E}(\xi_2) = \sum_{x_2=1}^3 x_2 p(x_2) = 1 \cdot 0.7 + 2 \cdot 0.1 + 3 \cdot 0.2 = 1.5. \quad (25.5)$$

und damit dann schließlich

$$\begin{aligned}\mathbb{C}(\xi_1, \xi_2) &= \sum_{x_1=1}^2 \sum_{x_2=1}^3 (x_1 - 1.7)(x_2 - 1.5) p(x_1, x_2) \\ &= \sum_{x_1=1}^2 (x_1 - 1.7)(1 - 1.5)p(x_1, 1) + (x_1 - 1.7)(2 - 1.5)p(x_1, 2) + (x_1 - 1.7)(3 - 1.5)p(x_1, 3) \\ &= (1 - 1.7)(1 - 1.5)p(1, 1) + (1 - 1.7)(2 - 1.5)p(1, 2) + (1 - 1.7)(3 - 1.5)p(1, 3) \\ &\quad + (2 - 1.7)(1 - 1.5)p(2, 1) + (2 - 1.7)(2 - 1.5)p(2, 2) + (2 - 1.7)(3 - 1.5)p(2, 3) \\ &= (-0.7) \cdot (-0.5) \cdot 0.10 + (-0.7) \cdot 0.5 \cdot 0.05 + (-0.7) \cdot 1.5 \cdot 0.15 \\ &\quad + 0.3 \cdot (-0.5) \cdot 0.60 + 0.3 \cdot 0.5 \cdot 0.05 + 0.3 \cdot 1.5 \cdot 0.05 \\ &= 0.035 - 0.0175 - 0.1575 - 0.09 + 0.0075 + 0.0225 \\ &= -0.2.\end{aligned}\quad (25.6)$$

Die Kovarianz der Zufallsvariablen ξ_1 und ξ_2 mit der in obiger Tabelle festgelegter Verteilung ist also $\mathbb{C}(\xi_1, \xi_2) = -0.2$. Im Gegensatz zur Varianz kann die Kovarianz also offenbar auch negative Werte annehmen.

25.2. Eigenschaften

Wir betrachten im Folgenden einige grundlegende Eigenschaften der Kovarianz.

Theorem 25.1 (Symmetrie der Kovarianz). ξ_1 und ξ_2 seien zwei Zufallsvariablen. Dann gilt

$$\mathbb{C}(\xi_1, \xi_2) = \mathbb{C}(\xi_2, \xi_1) \quad (25.7)$$

◦

Beweis. Mit der Kommutativität der Multiplikation gilt

$$\mathbb{C}(\xi_1, \xi_2) = \mathbb{E}((\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2))) = \mathbb{E}((\xi_2 - \mathbb{E}(\xi_2))(\xi_1 - \mathbb{E}(\xi_1))) = \mathbb{C}(\xi_2, \xi_1) \quad (25.8)$$

□

Wie das Berechnen von Varianzen wird auch das Berechnen von Kovarianzen manchmal durch folgendes Theorem erleichtert.

Theorem 25.2 (Kovarianzverschiebungssatz). ξ_1 und ξ_2 seien Zufallsvariablen. Dann gilt

$$\mathbb{C}(\xi_1, \xi_2) = \mathbb{E}(\xi_1 \xi_2) - \mathbb{E}(\xi_1)\mathbb{E}(\xi_2). \quad (25.9)$$

◦

Beweis. Mit Definition 25.1 gilt

$$\begin{aligned} \mathbb{C}(\xi_1, \xi_2) &= \mathbb{E}((\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2))) \\ &= \mathbb{E}(\xi_1 \xi_2 - \xi_1 \mathbb{E}(\xi_2) - \mathbb{E}(\xi_1) \xi_2 + \mathbb{E}(\xi_1)\mathbb{E}(\xi_2)) \\ &= \mathbb{E}(\xi_1 \xi_2) - \mathbb{E}(\xi_1)\mathbb{E}(\xi_2) - \mathbb{E}(\xi_1)\mathbb{E}(\xi_2) + \mathbb{E}(\xi_1)\mathbb{E}(\xi_2) \\ &= \mathbb{E}(\xi_1 \xi_2) - \mathbb{E}(\xi_1)\mathbb{E}(\xi_2). \end{aligned} \quad (25.10)$$

□

Natürlich ist Theorem 25.2 nur dann wirklich nützlich, wenn $\mathbb{E}(\xi_1 \xi_2)$ leicht zu berechnen sind. Der Varianzverschiebungssatz nach Theorem 24.5 ergibt sich aus Theorem 25.2 für den Fall $\xi_1 = \xi_2 := \xi$ anhand von

$$\mathbb{V}(\xi) = \mathbb{C}(\xi, \xi) = \mathbb{E}(\xi \xi) - \mathbb{E}(\xi)\mathbb{E}(\xi) = \mathbb{E}(\xi^2) - \mathbb{E}(\xi)\mathbb{E}(\xi). \quad (25.11)$$

Auch in Hinblick auf die Kovarianz $\mathbb{C}(\xi_1, \xi_2)$ ist man daran interessiert, wie sich diese unter Anwendung linear-affiner Transformationen auf ξ_2 und ξ_2 verhält. Folgendes Resultat zeigt, dass die Kovarianz zweier Zufallsvariablen von ihrem Maßstab abhängt.

Theorem 25.3 (Kovarianz bei linear-affinen Transformationen). ξ_1 und ξ_2 seien Zufallsvariablen mit gemeinsamen Ergebnisraum $\mathcal{X}_1 \times \mathcal{X}_2$ und es seien

$$\begin{aligned} f_1 : \mathcal{X}_1 &\rightarrow \mathcal{Z}_1, x_1 \mapsto f(x_1) := ax_1 + b \\ f_2 : \mathcal{X}_2 &\rightarrow \mathcal{Z}_2, x_2 \mapsto f(x_2) := cx_2 + d \end{aligned} \quad (25.12)$$

für $a, b, c, d \in \mathbb{R}$ zwei linear-affine Funktionen. Dann gilt

$$\mathbb{C}(f_1(\xi_1), f_2(\xi_2)) = \mathbb{C}(a\xi_1 + b, c\xi_2 + d) = ac\mathbb{C}(\xi_1, \xi_2). \quad (25.13)$$

◦

Beweis. Es gilt

$$\begin{aligned} \mathbb{C}(a\xi_1 + b, c\xi_2 + d) &= \mathbb{E}((a\xi_1 + b - \mathbb{E}(a\xi_1 + b))(c\xi_2 + d - \mathbb{E}(c\xi_2 + d))) \\ &= \mathbb{E}((a\xi_1 + b - a\mathbb{E}(\xi_1) - b)(c\xi_2 + d - c\mathbb{E}(\xi_2) - d)) \\ &= \mathbb{E}(a(\xi_1 - \mathbb{E}(\xi_1))(c(\xi_2 - \mathbb{E}(\xi_2)))) \\ &= \mathbb{E}(ac((\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2)))) \\ &= ac\mathbb{C}(\xi_1, \xi_2). \end{aligned} \quad (25.14)$$

□

Mithilfe des Begriffes des Kovarianz ist es möglich, weitgreifenderer Aussagen über die Varianzen von Summen und Differenzen von Zufallsvariablen zu treffen als es in Kapitel 24 der Fall war, wo lediglich *unabhängige* Zufallsvariablen betrachtet wurden. Mit dem folgenden Theorem betrachten wir zunächst den sehr allgemeinen Fall der Kovarianz zweier Zufallsvariablen, die sich als Linearkombinationen von n Zufallsvariablen $\xi_1, \dots, \xi_n$ und m Zufallsvariablen $\zeta_1, \dots, \zeta_m$ unter Addition reeller Konstanten ergeben. Hieraus ergeben sich dann eine Reihe von in der Anwendung, insbesondere in der Klassischen Testtheorie, wichtiger Spezialfälle.

Theorem 25.4 (Kovarianz von Linearkombinationen von Zufallsvariablen). *Gegeben seien n Zufallsvariablen $\xi_1, \dots, \xi_n$ und $n + 1$ reelle Konstanten $a_0, a_1, \dots, a_n$ sowie m Zufallsvariablen $\zeta_1, \dots, \zeta_m$ und $m + 1$ reelle Konstanten $b_0, b_1, \dots, b_m$. Dann gilt*

$$\mathbb{C} \left(a_0 + \sum_{i=1}^n a_i \xi_i, b_0 + \sum_{j=1}^m b_j \zeta_j \right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathbb{C}(\xi_i, \zeta_j). \quad (25.15)$$

◦

Beweis.

$$\begin{aligned} & \mathbb{C} \left(a_0 + \sum_{i=1}^n a_i \xi_i, b_0 + \sum_{j=1}^m b_j \zeta_j \right) \\ &= \mathbb{E} \left(\left(a_0 + \sum_{i=1}^n a_i \xi_i - \mathbb{E} \left(a_0 + \sum_{i=1}^n a_i \xi_i \right) \right) \left(b_0 + \sum_{j=1}^m b_j \zeta_j - \mathbb{E} \left(b_0 + \sum_{j=1}^m b_j \zeta_j \right) \right) \right) \\ &= \mathbb{E} \left(\left(a_0 + \sum_{i=1}^n a_i \xi_i - a_0 - \mathbb{E} \left(\sum_{i=1}^n a_i \xi_i \right) \right) \left(b_0 + \sum_{j=1}^m b_j \zeta_j - b_0 - \mathbb{E} \left(\sum_{j=1}^m b_j \zeta_j \right) \right) \right) \\ &= \mathbb{E} \left(\left(\sum_{i=1}^n a_i \xi_i - \sum_{i=1}^n a_i \mathbb{E}(\xi_i) \right) \left(\sum_{j=1}^m b_j \zeta_j - \sum_{j=1}^m b_j \mathbb{E}(\zeta_j) \right) \right) \\ &= \mathbb{E} \left(\sum_{i=1}^n a_i \xi_i \sum_{j=1}^m b_j \zeta_j - \sum_{i=1}^n a_i \xi_i \sum_{j=1}^m b_j \mathbb{E}(\zeta_j) - \sum_{i=1}^n a_i \mathbb{E}(\xi_i) \sum_{j=1}^m b_j \zeta_j + \sum_{i=1}^n a_i \mathbb{E}(\xi_i) \sum_{j=1}^m b_j \mathbb{E}(\zeta_j) \right) \\ &= \mathbb{E} \left(\sum_{i=1}^n a_i \xi_i \sum_{j=1}^m b_j \zeta_j \right) - \mathbb{E} \left(\sum_{i=1}^n a_i \xi_i \sum_{j=1}^m b_j \mathbb{E}(\zeta_j) \right) - \mathbb{E} \left(\sum_{i=1}^n a_i \mathbb{E}(\xi_i) \sum_{j=1}^m b_j \zeta_j \right) + \mathbb{E} \left(\sum_{i=1}^n a_i \mathbb{E}(\xi_i) \sum_{j=1}^m b_j \mathbb{E}(\zeta_j) \right) \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathbb{E}(\xi_i \zeta_j) - 2 \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathbb{E}(\xi_i) \mathbb{E}(\zeta_j) + \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathbb{E}(\xi_i) \mathbb{E}(\zeta_j) \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j (\mathbb{E}(\xi_i \zeta_j) - 2\mathbb{E}(\xi_i) \mathbb{E}(\zeta_j) + \mathbb{E}(\xi_i) \mathbb{E}(\zeta_j)) \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathbb{E}((\xi_i - \mathbb{E}(\xi_i))(\zeta_j - \mathbb{E}(\zeta_j))) \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathbb{C}(\xi_i, \zeta_j) \end{aligned}$$

□

Mit folgenden Theorem betrachten wir nun einen ersten Spezialfall von Theorem 25.4.

Theorem 25.5 (Kovarianz bei paarweiser Addition von Zufallsvariablen). *$\xi_1, \xi_2, \zeta_1, \zeta_2$ seien vier Zufallsvariablen. Dann gilt*

$$\mathbb{C}(\xi_1 + \xi_2, \zeta_1 + \zeta_2) = \mathbb{C}(\xi_1, \zeta_1) + \mathbb{C}(\xi_1, \zeta_2) + \mathbb{C}(\xi_2, \zeta_1) + \mathbb{C}(\xi_2, \zeta_2). \quad (25.16)$$

◦

Beweis. Es seien $n := 2, m := 2, a_0 = b_0 := 0$ und $a_i = b_i := 1$ für $i = 1, 2$. Dann gilt mit Theorem 25.4

$$\begin{aligned}
 \mathbb{C}(\xi_1 + \xi_2, \zeta_1 + \zeta_2) &= \mathbb{C}\left(a_0 + \sum_{i=1}^2 a_i \xi_i, b_0 + \sum_{j=1}^2 b_j \zeta_j\right) \\
 &= \sum_{i=1}^2 \sum_{j=1}^2 a_i b_j \mathbb{C}(\xi_i, \zeta_j) \\
 &= \sum_{i=1}^2 \sum_{j=1}^2 \mathbb{C}(\xi_i, \zeta_j) \\
 &= \mathbb{C}(\xi_1, \zeta_1) + \mathbb{C}(\xi_1, \zeta_2) + \mathbb{C}(\xi_2, \zeta_1) + \mathbb{C}(\xi_2, \zeta_2).
 \end{aligned} \tag{25.17}$$

□

Folgendes Theorem besagt nun, wie man im Allgemeinen die Varianz einer Linearkombinationen von Zufallsvariablen unter Addition einer Konstante bestimmt werden kann.

Theorem 25.6 (Varianz einer Linearkombination von Zufallsvariablen). *Gegeben seien n Zufallsvariablen $\xi_1, \dots, \xi_n$ und $n + 1$ reelle Konstanten $a_0, a_1, \dots, a_n$. Dann gilt*

$$\mathbb{V}\left(a_0 + \sum_{i=1}^n a_i \xi_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \mathbb{C}(\xi_i, \xi_j). \tag{25.18}$$

◦

Beweis. Mit $\mathbb{V}(\xi) = \mathbb{C}(\xi, \xi)$ und Theorem 25.4 gilt zunächst

$$\begin{aligned}
 \mathbb{V}\left(a_0 + \sum_{i=1}^n a_i \xi_i\right) &= \mathbb{C}\left(a_0 + \sum_{i=1}^n a_i \xi_i, a_0 + \sum_{i=1}^n a_i \xi_i\right) \\
 &= \mathbb{C}\left(a_0 + \sum_{i=1}^n \xi_i, a_0 + \sum_{j=1}^n a_j \xi_j\right) \\
 &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathbb{C}(\xi_i, \xi_j) \\
 &= \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n a_i a_j \mathbb{C}(\xi_i, \xi_j) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j=i}}^n a_i a_j \mathbb{C}(\xi_i, \xi_j) \\
 &= \sum_{i=1}^n a_i a_i \mathbb{C}(\xi_i, \xi_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n a_i a_j \mathbb{C}(\xi_i, \xi_j) \\
 &= \sum_{i=1}^n a_i^2 \mathbb{V}(\xi_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n a_i a_j \mathbb{C}(\xi_i, \xi_j) \\
 &= \sum_{i=1}^n a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \mathbb{C}(\xi_i, \xi_j),
 \end{aligned} \tag{25.19}$$

Dabei wurde in der vierten Gleichung die Doppelsumme in solche Terme aufgespalten für die $i = j$ und für die $i \neq j$ und in siebten Gleichung ausgenutzt, dass

$$a_i a_j \mathbb{C}(\xi_i, \xi_j) = a_j a_i \mathbb{C}(\xi_j, \xi_i). \tag{25.20}$$

Wir verdeutlichen die darauffolgende Identität

$$\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n a_i a_j \mathbb{C}(\xi_i, \xi_j) = 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \mathbb{C}(\xi_i, \xi_j), \quad (25.21)$$

am Beispiel $n = 3$ untenstehend. Analog mag man sich vorstellen, über alle Elemente außer der Diagonalelemente der $i = 1, \dots, n$ Zeilen und $j = 1, \dots, n$ Spalten einer symmetrischen Matrix mit Einträgen $a_i a_j \mathbb{C}(\xi_i, \xi_j)$ zu summieren. Die linke Seite oberer Gleichung entspricht dann dem zeilenweisen Vorgehen der Summationsbildung für alle Spalteneinträge außer dem, der in der Spalte der jeweils betrachteten Zeile steht, d.h. dem Diagonaleintrag. Die rechte Seite der obigen Gleichung entspricht dann dem Vorgehen, die Symmetrie der Matrix auszunutzen, also die Tatsache zu berücksichtigen, dass die Einträge der Matrix rechts oberhalb und links unterhalb der Diagonalen identisch sind, so dass es genügt, die Einträge der oberen linken Hälfte aufzuaddieren und zu verdoppeln. Dabei werden für jede Zeile $i = 1, \dots, n$ nur gerade die Spalten $j = i + 1, \dots, n$ betrachtet, die rechts von der Diagonale stehen und ihre Einträge aufsummiert. In der letzten Zeile steht dabei nur ein Diagonalelement, das nicht zur Summe gehört, weshalb der Index der äußeren Summe nur bis $n - 1$ läuft. Konkret ergibt sich für $n := 3$

$$\begin{aligned} & \sum_{i=1}^3 \sum_{\substack{j=1 \\ j \neq i}}^3 a_i a_j \mathbb{C}(\xi_i, \xi_j) \\ &= \sum_{\substack{j=1 \\ j \neq 1}}^3 a_1 a_j \mathbb{C}(\xi_1, \xi_j) + \sum_{\substack{j=1 \\ j \neq 2}}^3 a_2 a_j \mathbb{C}(\xi_2, \xi_j) + \sum_{\substack{j=1 \\ j \neq 3}}^3 a_3 a_j \mathbb{C}(\xi_3, \xi_j) \\ &= a_1 a_2 \mathbb{C}(\xi_1, \xi_2) + a_1 a_3 \mathbb{C}(\xi_1, \xi_3) + a_2 a_1 \mathbb{C}(\xi_2, \xi_1) + a_2 a_3 \mathbb{C}(\xi_2, \xi_3) + a_3 a_1 \mathbb{C}(\xi_3, \xi_1) + a_3 a_2 \mathbb{C}(\xi_3, \xi_2) \\ &= a_1 a_2 \mathbb{C}(\xi_1, \xi_2) + a_2 a_1 \mathbb{C}(\xi_2, \xi_1) + a_1 a_3 \mathbb{C}(\xi_1, \xi_3) + a_3 a_1 \mathbb{C}(\xi_3, \xi_1) + a_2 a_3 \mathbb{C}(\xi_2, \xi_3) + a_3 a_2 \mathbb{C}(\xi_3, \xi_2) \\ &= 2\mathbb{C}(\xi_1, \xi_2) + 2\mathbb{C}(\xi_1, \xi_3) + 2\mathbb{C}(\xi_2, \xi_3) \\ &= 2a_1 a_2 (\mathbb{C}(\xi_1, \xi_2) + a_1 a_3 \mathbb{C}(\xi_1, \xi_3) + a_2 a_3 \mathbb{C}(\xi_2, \xi_3)) \\ &= 2 \left(\sum_{j=2}^3 a_1 a_j \mathbb{C}(\xi_1, \xi_j) + \sum_{j=3}^3 a_2 a_j \mathbb{C}(\xi_2, \xi_j) \right) \\ &= 2 \left(\sum_{j=1+1}^3 a_1 a_j \mathbb{C}(\xi_1, \xi_j) + \sum_{j=2+1}^3 a_2 a_j \mathbb{C}(\xi_2, \xi_j) \right) \\ &= 2 \sum_{i=1}^2 \sum_{j=i+1}^3 a_i a_j \mathbb{C}(\xi_i, \xi_j). \end{aligned} \quad (25.22)$$

□

Theorem 25.7 (Varianzen spezieller Linearkombinationen von Zufallsvariablen).

(1) (Varianz bei Addition zweier Zufallsvariablen). Gegeben seien zwei Zufallsvariablen ξ und ζ . Dann gilt

$$\mathbb{V}(\xi + \zeta) = \mathbb{V}(\xi) + \mathbb{V}(\zeta) + 2\mathbb{C}(\xi, \zeta) \quad (25.23)$$

(2) (Varianz bei Subtraktion zweier Zufallsvariablen). Gegeben seien zwei Zufallsvariablen ξ und ζ . Dann gilt

$$\mathbb{V}(\xi - \zeta) = \mathbb{V}(\xi) + \mathbb{V}(\zeta) - 2\mathbb{C}(\xi, \zeta) \quad (25.24)$$

◦

Beweis. (1) Es seien $n := 2$, $\xi_1 := \xi$, $\xi_2 := \zeta$, $a_0 := 0$, $a_1 := 1$ und $a_2 := 1$. Dann gilt mit dem Theorem

zur Varianz einer Linearkombination von Zufallsvariablen

$$\begin{aligned}
 \mathbb{V}(\xi + \zeta) &= \mathbb{V}(a_0 + a_1\xi_1 + a_2\xi_2) \\
 &= \mathbb{V}\left(a_0 + \sum_{i=1}^2 a_i\xi_i\right) \\
 &= \sum_{i=1}^2 a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{i=1}^{2-1} \sum_{j=i+1}^2 a_i a_j \mathbb{C}(\xi_i, \xi_j) \\
 &= \sum_{i=1}^2 a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{i=1}^1 \sum_{j=i+1}^2 a_i a_j \mathbb{C}(\xi_i, \xi_j) \\
 &= \sum_{i=1}^2 a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{j=1+1}^2 a_1 a_j \mathbb{C}(\xi_1, \xi_j) \\
 &= \sum_{i=1}^2 a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{j=2}^2 a_1 a_j \mathbb{C}(\xi_1, \xi_j) \\
 &= a_1^2 \mathbb{V}(\xi_1) + a_2^2 \mathbb{V}(\xi_2) + 2a_1 a_2 \mathbb{C}(\xi_1, \xi_2) \\
 &= 1^2 \cdot \mathbb{V}(\xi) + 1^2 \cdot \mathbb{V}(\zeta) + 2 \cdot 1 \cdot 1 \cdot \mathbb{C}(\xi, \zeta) \\
 &= \mathbb{V}(\xi) + \mathbb{V}(\zeta) + 2\mathbb{C}(\xi, \zeta).
 \end{aligned} \tag{25.25}$$

(2) Es seien $n := 2$, $\xi_1 := \xi$, $\xi_2 := \zeta$, $a_0 := 0$, $a_1 := 1$ und $a_2 := -1$. Dann gilt mit dem Theorem zur Varianz einer Linearkombination von Zufallsvariablen

$$\begin{aligned}
 \mathbb{V}(\xi - \zeta) &= \mathbb{V}(a_0 + a_1\xi_1 + a_2\xi_2) \\
 &= \mathbb{V}\left(a_0 + \sum_{i=1}^2 a_i\xi_i\right) \\
 &= \sum_{i=1}^2 a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{i=1}^{2-1} \sum_{j=i+1}^2 a_i a_j \mathbb{C}(\xi_i, \xi_j) \\
 &= \sum_{i=1}^2 a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{i=1}^1 \sum_{j=i+1}^2 a_i a_j \mathbb{C}(\xi_i, \xi_j) \\
 &= \sum_{i=1}^2 a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{j=1+1}^2 a_1 a_j \mathbb{C}(\xi_1, \xi_j) \\
 &= \sum_{i=1}^2 a_i^2 \mathbb{V}(\xi_i) + 2 \sum_{j=2}^2 a_1 a_j \mathbb{C}(\xi_1, \xi_j) \\
 &= a_1^2 \mathbb{V}(\xi_1) + a_2^2 \mathbb{V}(\xi_2) + 2a_1 a_2 \mathbb{C}(\xi_1, \xi_2) \\
 &= 1^2 \cdot \mathbb{V}(\xi) + (-1)^2 \cdot \mathbb{V}(\zeta) + 2 \cdot 1 \cdot (-1) \cdot \mathbb{C}(\xi, \zeta) \\
 &= \mathbb{V}(\xi) + \mathbb{V}(\zeta) - 2\mathbb{C}(\xi, \zeta).
 \end{aligned} \tag{25.26}$$

□

25.3. Bedingte Kovarianz

Wie den Erwartungswert und die Varianz kann man auch die Kovarianz in einer bedingten gemeinsamen Verteilung zweier Zufallsvariablen betrachten. Dies führt auf den Begriff der *bedingten Kovarianz*.

Definition 25.2 (Bedingte Kovarianz). Gegeben sei ein Zufallsvektor $\xi := (\xi_1, \xi_2, \xi_3)$ mit Ergebnisraum $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ und WMF oder WDF $p(x_1, x_2, x_3)$ und bedingter WMF oder WDF $p(x_1, x_2 | x_3)$ für alle $x_3 \in \mathcal{X}_3$. Dann ist die bedingte Kovarianz von ξ_1 und ξ_2 gegeben $\xi_3 = x_3$ definiert als

$$\mathbb{C}(\xi_1, \xi_2 | \xi_3 = x_3) = \mathbb{E}((\xi_1 - \mathbb{E}(\xi_1 | \xi_3 = x_3))(\xi_2 - \mathbb{E}(\xi_2 | \xi_3 = x_3)) | \xi_3 = x_3) \tag{25.27}$$

•

Die bedingte Kovarianz ist also im Sinne des bedingten Erwartungswerts des Zufallsvektors (ξ_1, ξ_2) definiert und ist, wie der bedingte Erwartungswert und die bedingte Varianz im Allgemeinen eine Zufallsvariable dar. Schließlich halten wir fest, dass auch für die bedingte Kovarianz der Kovarianzverschiebungssatz gilt.

Theorem 25.8 (Verschiebungssatz der bedingten Kovarianz). *Gegeben sei ein Zufallsvektor $\xi := (\xi_1, \xi_2, \xi_3)$. Dann gilt*

$$\mathbb{C}(\xi_1, \xi_2 | \xi_3) = \mathbb{E}(\xi_1 \xi_2 | \xi_3) - \mathbb{E}(\xi_1 | \xi_3) \mathbb{E}(\xi_2 | \xi_3). \quad (25.28)$$

◦

Beweis. Ein Beweis ergibt sich durch die Ersetzung der entsprechenden Erwartungswerte im Beweis von Theorem 25.2 durch die entsprechenden bedingten Erwartungswerte. $\square$

25.4. Kovarianzmatrizen

Das multivariate Analogon der Varianz einer Zufallsvariable ist die *Kovarianzmatrix eines Zufallsvektors*. Diese enkodiert neben den Varianzen der Komponenten des Zufallsvektors auch ihre paarweisen Kovarianzen und ist wie folgt definiert.

Definition 25.3 (Kovarianzmatrix eines Zufallsvektors). ξ sei ein n -dimensionaler Zufallsvektor. Dann ist die *Kovarianzmatrix* von ξ definiert als die $n \times n$ Matrix

$$\mathbb{C}(\xi) := \mathbb{E}((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T). \quad (25.29)$$

•

Die Kovarianzmatrix ist in Definition 25.3 formal analog zur Kovarianz zweier Zufallsvariablen definiert. Eine direkte Rückführung des Begriffs der Kovarianzmatrix eines Zufallsvektors auf den Begriff aus dem univariaten Kontext bekannten Begriff der Kovarianz zweier Zufallsvariablen erlaubt folgendes Theorem.

Theorem 25.9 (Eigenschaften der Kovarianzmatrix). ξ sei ein m -dimensionaler Zufallsvektor und $\mathbb{C}(\xi)$ sei seine Kovarianzmatrix. Dann gelten

(1) (Elemente) Die Elemente von $\mathbb{C}(\xi)$ sind die Kovarianzen der Komponenten von ξ ,

$$\mathbb{C}(\xi) = (\mathbb{C}(\xi_i, \xi_j))_{1 \leq i, j \leq m}. \quad (25.30)$$

(2) (Kovarianzmatrixverschiebungssatz) Es gilt

$$\mathbb{C}(\xi) = \mathbb{E}(\xi \xi^T) - \mathbb{E}(\xi) \mathbb{E}(\xi)^T. \quad (25.31)$$

(3) (Linear-affine Transformation) Für $A \in \mathbb{R}^{n \times m}$ und $b \in \mathbb{R}^n$ gilt

$$\mathbb{C}(A\xi + b) = A\mathbb{C}(\xi)A^T. \quad (25.32)$$

(4) (Matrizeigenschaften) $\mathbb{C}(\xi)$ ist symmetrisch und positiv-semidefinit.

◦

Beweis. (1) Es gilt

$$\begin{aligned}
 \mathbb{C}(\xi) &:= \mathbb{E}((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T) \\
 &= \mathbb{E} \left(\left(\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} - \begin{pmatrix} \mathbb{E}(\xi_1) \\ \vdots \\ \mathbb{E}(\xi_n) \end{pmatrix} \right) \left(\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} - \begin{pmatrix} \mathbb{E}(\xi_1) \\ \vdots \\ \mathbb{E}(\xi_n) \end{pmatrix} \right)^T \right) \\
 &= \mathbb{E} \left(\begin{pmatrix} \xi_1 - \mathbb{E}(\xi_1) \\ \vdots \\ \xi_n - \mathbb{E}(\xi_n) \end{pmatrix} \begin{pmatrix} \xi_1 - \mathbb{E}(\xi_1) \\ \vdots \\ \xi_n - \mathbb{E}(\xi_n) \end{pmatrix}^T \right) \\
 &= \mathbb{E} \left(\begin{pmatrix} \xi_1 - \mathbb{E}(\xi_1) \\ \vdots \\ \xi_n - \mathbb{E}(\xi_n) \end{pmatrix} (\xi_1 - \mathbb{E}(\xi_1) \quad \dots \quad \xi_n - \mathbb{E}(\xi_n)) \right) \\
 &= \mathbb{E} \begin{pmatrix} (\xi_1 - \mathbb{E}(\xi_1))(\xi_1 - \mathbb{E}(\xi_1)) & \dots & (\xi_1 - \mathbb{E}(\xi_1))(\xi_n - \mathbb{E}(\xi_n)) \\ \vdots & \ddots & \vdots \\ (\xi_n - \mathbb{E}(\xi_n))(\xi_1 - \mathbb{E}(\xi_1)) & \dots & (\xi_n - \mathbb{E}(\xi_n))(\xi_n - \mathbb{E}(\xi_n)) \end{pmatrix} \\
 &= (\mathbb{E}((\xi_i - \mathbb{E}(\xi_i))(\xi_j - \mathbb{E}(\xi_j))))_{1 \leq i, j \leq n} \\
 &= (\mathbb{C}(\xi_i, \xi_j))_{1 \leq i, j \leq n}.
 \end{aligned} \tag{25.33}$$

(2) Mit den Eigenschaften von Erwartungswerten gilt

$$\begin{aligned}
 \mathbb{C}(\xi) &= \mathbb{E}((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T) \\
 &= \mathbb{E}(\xi \xi^T - \xi \mathbb{E}(\xi)^T - \mathbb{E}(\xi) \xi^T + \mathbb{E}(\xi) \mathbb{E}(\xi)^T) \\
 &= \mathbb{E}(\xi \xi^T) - \mathbb{E}(\xi) \mathbb{E}(\xi)^T - \mathbb{E}(\xi) \mathbb{E}(\xi)^T + \mathbb{E}(\xi) \mathbb{E}(\xi)^T \\
 &= \mathbb{E}(\xi \xi^T) - \mathbb{E}(\xi) \mathbb{E}(\xi)^T.
 \end{aligned} \tag{25.34}$$

(3) Mit den Eigenschaften von Erwartungswerten gilt

$$\begin{aligned}
 \mathbb{C}(A\xi + b) &= \mathbb{E}((A\xi + b - \mathbb{E}(A\xi + b))(A\xi + b - \mathbb{E}(A\xi + b))^T) \\
 &= \mathbb{E}((A\xi + b - A\mathbb{E}(\xi) - b)(A\xi + b - A\mathbb{E}(\xi) - b)^T) \\
 &= \mathbb{E}((A(\xi - \mathbb{E}(\xi)))(A(\xi - \mathbb{E}(\xi)))^T) \\
 &= \mathbb{E}(A(\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T A^T) \\
 &= A\mathbb{E}((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T) A^T \\
 &= A\mathbb{C}(\xi)A^T.
 \end{aligned} \tag{25.35}$$

(4) Die Symmetrie von $\mathbb{C}(\xi)$ folgt aus der Symmetrie der Kovarianz einer Zufallsvariable mit

$$\mathbb{C}(\xi_i, \xi_j) = \mathbb{C}(\xi_j, \xi_i) \text{ für alle } i = 1, \dots, m, j = 1, \dots, m. \tag{25.36}$$

Um die positive Semidefinitheit von $\mathbb{C}(\xi)$ nachzuweisen, ist zu zeigen, dass $a^T \mathbb{C}(\xi) a \geq 0$ für alle $a \in \mathbb{R}^m$ mit $a \neq 0_m$. Sei also $a \in \mathbb{R}^m$ mit $a \neq 0$. Dann gilt mit Aussage (3) für $A := a^T \in \mathbb{R}^{1 \times m}$, dass

$$a^T \mathbb{C}(\xi) a = \mathbb{C}(a^T \xi). \tag{25.37}$$

Weiterhin gilt mit der Definition der Kovarianzmatrix aber, dass

$$\mathbb{C}(a^T \xi) = \mathbb{E}((a^T \xi - \mathbb{E}(a^T \xi))^2) = \mathbb{V}(a^T \xi). \tag{25.38}$$

Da mit den Eigenschaften der Varianz die Varianz der Zufallsvariable $a^T \xi$ aber immer nichtnegativ ist, folgt

$$a^T \mathbb{C}(\xi) a = \mathbb{V}(a^T \xi) \geq 0 \tag{25.39}$$

und damit die positive Semidefinitheit von $\mathbb{C}(\xi)$. $\square$

Die Diagonalelemente von $\mathbb{C}(\xi)$ sind die Varianzen der Komponenten von ξ , da

$$\mathbb{V}(\xi_i) = \mathbb{C}(\xi_i, \xi_i) \text{ für } i = 1, \dots, m. \quad (25.40)$$

Eigenschaften (2) und (3) sind im Wesentlichen analog zu den Eigenschaften der Varianz.

Die Kovarianzmatrix eines Zufallsvektors ξ ist also die Matrix der Kovarianzen der Komponenten von ξ . Damit ist auch die Kovarianzmatrix direkt im Sinne des Begriffs der Kovarianz von Zufallsvektoren gegeben. Da die Kovarianz einer Zufallsvariable mit sich selbst bekanntlich ihre Varianz ist, enthält die Kovarianzmatrix auf ihrer Diagonalen die Varianzen der Komponenten von ξ .

Folgendes Theorem dokumentiert eine Schreibweise für die Kovarianzmatrix eines partitionierten Zufallsvektors im Sinne von Erwartungswerten von Zufallsvektorprodukten an, die zum Beispiel im Rahmen der Kanonischen Korrelationsanalyse hilfreich ist.

Theorem 25.10 (Kovarianzmatrizen von Zufallsvektoren). *Es seien*

$$\zeta = \begin{pmatrix} \xi \\ v \end{pmatrix} \text{ mit } \mathbb{E}(\zeta) := \mathbf{0}_m \quad (25.41)$$

ein $m_\xi + m_v$ -dimensionaler Zufallsvektor und sein Erwartungswertvektor, respektive. Dann kann die $m \times m$ Kovarianzmatrix von ζ geschrieben werden als

$$\mathbb{C}(\zeta) = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi_1\xi_2} \\ \Sigma_{v\xi} & \Sigma_{vv} \end{pmatrix} \in \mathbb{R}^{m \times m} \quad (25.42)$$

wobei

$$\begin{aligned} \Sigma_{\xi\xi} &:= \mathbb{E}(\xi\xi^T) \in \mathbb{R}^{m_\xi \times m_\xi} \\ \Sigma_{\xi_1\xi_2} &:= \mathbb{E}(\xi_1\xi_2^T) \in \mathbb{R}^{m_\xi \times m_v} \\ \Sigma_{v\xi} &:= \mathbb{E}(v\xi^T) \in \mathbb{R}^{m_v \times m_\xi} \\ \Sigma_{vv} &:= \mathbb{E}(vv^T) \in \mathbb{R}^{m_v \times m_v} \end{aligned} \quad (25.43)$$

◦

Beweis. Nach Definition der Kovarianzmatrix eines Zufallsvektors gilt

$$\begin{aligned} \mathbb{C}(z) &= \mathbb{E}((\zeta - \mathbb{E}(\zeta))(\zeta - \mathbb{E}(\zeta))^T) \\ &= \mathbb{E}((\zeta - \mathbf{0}_m)(\zeta - \mathbf{0}_m)^T) \\ &= \mathbb{E}(\zeta\zeta^T) \\ &= \mathbb{E}\left(\begin{pmatrix} \xi \\ v \end{pmatrix} (\xi^T \quad v^T)\right) \\ &= \mathbb{E}\left(\begin{pmatrix} \xi\xi^T & \xi_1\xi_2^T \\ v\xi^T & vv^T \end{pmatrix}\right) \\ &= \begin{pmatrix} \mathbb{E}(\xi\xi^T) & \mathbb{E}(\xi_1\xi_2^T) \\ \mathbb{E}(v\xi^T) & \mathbb{E}(vv^T) \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi_1\xi_2} \\ \Sigma_{v\xi} & \Sigma_{vv} \end{pmatrix} \end{aligned} \quad (25.44)$$

□

Schließlich ist man in manchen Anwendungen an einer normalisierten, maßstabsunabhängigen Repräsentation der Kovarianzen eines Zufallsvektors interessiert. Wie im univariaten Fall bietet sich hierfür die Normalisierung der Kovarianz zweier Zufallsvariablen mithilfe ihrer jeweiligen Varianzen im Sinne einer Korrelation an. Diese Überlegung führt auf den Begriff der *Korrelationsmatrix* eines Zufallsvektors.

Definition 25.4 (Korrelationsmatrix). ξ sei ein n -dimensionaler Zufallsvektor. Dann ist die *Korrelationsmatrix* von ξ definiert als die $n \times n$ Matrix

$$\mathbb{R}(\xi) := (\rho_{ij})_{1 \leq i, j \leq n} = \left(\frac{\mathbb{C}(\xi_i, \xi_j)}{\sqrt{\mathbb{V}(\xi_i)} \sqrt{\mathbb{V}(\xi_j)}} \right)_{1 \leq i, j \leq n}. \quad (25.45)$$

•

Da es sich bei den Varianzen der Komponenten von ξ um die Diagonalelemente der Kovarianzmatrix von ξ handelt, ist die Korrelationsmatrix natürlich in der Kovarianzmatrix implizit. Weiterhin gelten, wie immer für Korrelationen, für die Einträge ρ_{ij} , $1 \leq i, j \leq n$ der Korrelationsmatrix, dass

$$\rho_{ij} \in [-1, 1] \text{ für } 1 \leq i, j \leq n \text{ und } \rho_{ii} = 1 \text{ für } 1 \leq i \leq n. \quad (25.46)$$

26. Ungleichungen

Das Thema dieses Kapitels sind Ungleichungen, die in der Wahrscheinlichkeitstheorie häufig zur Abschätzung von Wahrscheinlichkeiten und Erwartungswerten genutzt werden. Wir gliedern die Ungleichungen entsprechend in *Wahrscheinlichkeitsungleichungen* (*Markov Ungleichung* und *Chebyshev Ungleichung*, Kapitel 26.1) und *Erwartungswertungleichungen* (*Cauchy-Schwarz Ungleichung*, *Korrelationsungleichung* und *Jensensche Ungleichung*, Kapitel 26.2).

26.1. Wahrscheinlichkeitsungleichungen

Die *Markov Ungleichung* stellt einen Bezug zwischen den Überschreitungswahrscheinlichkeiten (vgl. Theorem 21.2) und dem Erwartungswert einer *nicht-negativen* Zufallsvariablen, also einer Zufallsvariable, für die $\mathbb{P}(\xi \geq 0) = 1$ ist, her. Im Beweis dieser Ungleichung wollen wir nur den Fall einer kontinuierlichen Zufallsvariable betrachten.

Theorem 26.1 (Markov Ungleichung). ξ sei eine Zufallsvariable mit $\mathbb{P}(\xi \geq 0) = 1$. Dann gilt für alle $x \in \mathbb{R}$, dass

$$\mathbb{P}(\xi \geq x) \leq \frac{\mathbb{E}(\xi)}{x}. \quad (26.1)$$

◦

Beweis. Wir betrachten den Fall einer kontinuierlichen Zufallsvariable ξ mit WDF p . Wir halten zunächst fest, dass

$$\mathbb{E}(\xi) = \int_{-\infty}^{\infty} s p(s) ds = \int_0^{\infty} s p(s) ds = \int_0^x s p(s) ds + \int_x^{\infty} s p(s) ds, \quad (26.2)$$

weil ξ nicht-negativ ist. Es folgt dann

$$\mathbb{E}(\xi) \geq \int_x^{\infty} s p(s) ds \geq \int_x^{\infty} x p(s) ds = x \int_x^{\infty} p(s) ds = x \mathbb{P}(\xi \geq x). \quad (26.3)$$

Dabei gilt die erste Ungleichung, weil

$$\int_0^x s p(s) ds \geq 0, \quad (26.4)$$

und die zweite Ungleichung gilt, weil $x \leq \xi$ für $\xi \in [x, \infty[$. Es folgt also, dass

$$\mathbb{E}(\xi) \geq x \mathbb{P}(\xi \geq x) \Leftrightarrow \mathbb{P}(\xi \geq x) \leq \frac{\mathbb{E}(\xi)}{x}. \quad (26.5)$$

□

Gilt beispielweise für eine nichtnegative Zufallsvariable ξ , dass $\mathbb{E}(\xi) = 1$ ist, dann folgt aus der Markov Ungleichung, dass

$$\mathbb{P}(\xi \geq 100) \leq 0.01. \quad (26.6)$$

Beispiel

Als Beispiel für die Markov Ungleichung betrachten wir den Fall einer Gamma-Zufallsvariable $\xi \sim G(\alpha, \beta)$. Gamma-Zufallsvariablen sind bekanntlich per Definition nicht-negativ (vgl. Definition 21.10) und wir haben gesehen, dass für den Erwartungswert einer Gamma-Zufallsvariable $\mathbb{E}(\xi) = \alpha\beta$ gilt. Wir betrachten konkret den Fall $\alpha := 5$ und $\beta := 2$, so dass ξ auch einer χ^2 -Zufallsvariable mit Freiheitsgradparameter $n = 10$ entspricht. In Abbildung 26.1 A stellen wir die KVF P dieser Zufallsvariable dar, in Abbildung 26.1 B visualisieren wir die in der Markov Ungleichung betrachteten Größen $\mathbb{E}(\xi)/x$ und die Überschreitungswahrscheinlichkeit $\mathbb{P}(\xi \geq x) = 1 - P(x)$. Offensichtlich trifft die Markov Ungleichung zu.

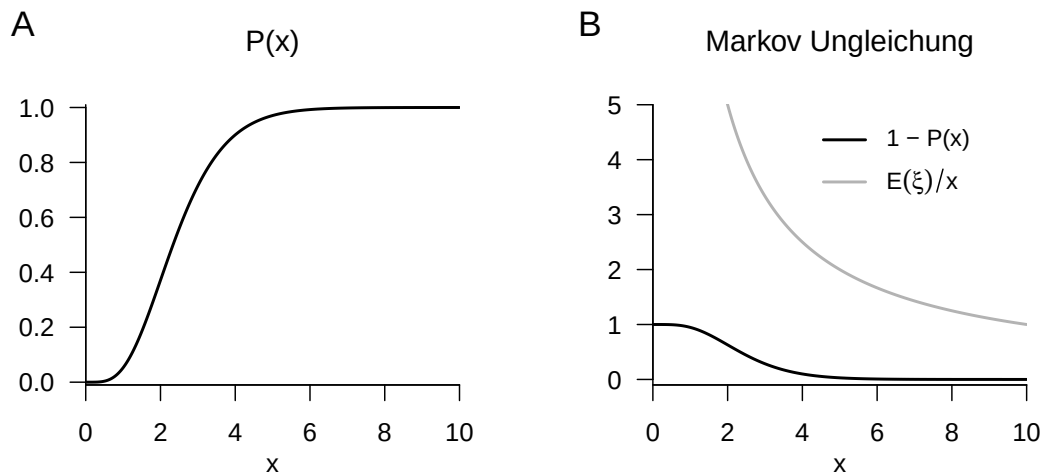


Abbildung 26.1. Markov Ungleichung am Beispiel einer Gamma-Zufallsvariable.

Die Chebychev Ungleichung setzt die Wahrscheinlichkeit dafür, dass eine Zufallsvariable Werte weit von ihrem Erwartungswert entfernt annimmt, in Bezug zur ihrer Varianz. Die Chebychev Ungleichung liefert damit eine Begründung dafür, warum die in Definition 24.1 formulierte Größe als ein Maß für die Streuung einer Zufallsvariable verstanden werden kann. Im Beweis der Chebyshev Ungleichung wird an entscheidender Stelle auf die Markov Ungleichung zurück gegriffen.

Theorem 26.2 (Chebyshev Ungleichung). *Es sei ξ eine Zufallsvariable mit Varianz $\mathbb{V}(\xi)$. Dann gilt für alle $x \in \mathbb{R}$*

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq x) \leq \frac{\mathbb{V}(\xi)}{x^2}. \quad (26.7)$$

◦

Beweis. Wir halten zunächst fest, dass für $a, b \in \mathbb{R}$ gilt, dass aus $a^2 \geq b^2$ folgt, dass $|a| \geq b$. Dazu betrachten wir die folgenden vier möglichen Fälle.

- (1) $a^2 \geq b^2$ für $a \geq 0$ und $b \geq 0$. Dann gilt

$$a^2 \geq b^2 \Rightarrow \sqrt{a^2} \geq \sqrt{b^2} \Rightarrow a \geq b \Rightarrow |a| \geq b. \quad (26.8)$$

- (2) $a^2 \geq b^2$ für $a \leq 0$ und $b \geq 0$. Dann gilt

$$a^2 \geq b^2 \Rightarrow \sqrt{a^2} \geq \sqrt{b^2} \Rightarrow -a \geq b \Rightarrow |a| \geq b. \quad (26.9)$$

(3) $a^2 \geq b^2$ für $a \geq 0$ und $b \leq 0$. Dann gilt

$$a^2 \geq b^2 \Rightarrow \sqrt{a^2} \geq \sqrt{b^2} \Rightarrow a \geq -b \geq b \Rightarrow |a| \geq b. \quad (26.10)$$

(4) $a^2 \geq b^2$ für $a \leq 0$ und $b \leq 0$. Dann gilt

$$a^2 \geq b^2 \Rightarrow \sqrt{a^2} \geq \sqrt{b^2} \Rightarrow -a \geq -b \geq b \Rightarrow |a| \geq b. \quad (26.11)$$

Als nächstes definieren wir $v := (\xi - \mathbb{E}(\xi))^2$. Dann gilt offenbar $v \geq 0$ und es folgt aus der Markov Ungleichung

$$\begin{aligned} \mathbb{P}(v \geq x^2) &\leq \frac{\mathbb{E}(v)}{x^2} \\ \Leftrightarrow \mathbb{P}((\xi - \mathbb{E}(\xi))^2 \geq x^2) &\leq \frac{\mathbb{E}((\xi - \mathbb{E}(\xi))^2)}{x^2} \\ \Leftrightarrow \mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq x) &\leq \frac{\mathbb{V}(\xi)}{x^2}. \end{aligned} \quad (26.12)$$

□

Beispielweise gilt für eine Zufallsvariable immer, dass die Wahrscheinlichkeit für eine absolute Abweichung vom doppelten ihrer Standardabweichung höchstens $1/4$ ist, also Frequentistisch betrachtet nur für etwa ein Viertel ihrer Realisierungen zutrifft, und die Wahrscheinlichkeit für eine absolute Abweichung vom dreifachen ihrer Standardabweichung höchstens $1/9$ ist, also Frequentistisch betrachtet nur für etwa ein Zehntel ihrer Realisierungen zutrifft, jeweils unabhängig davon, von welcher genauen Form die Verteilung der Zufallsvariable ist. Dies folgt mit der Chebyshev Ungleichung aus

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq 2\sqrt{\mathbb{V}(\xi)}) \leq \frac{\mathbb{V}(\xi)}{(2\sqrt{\mathbb{V}(\xi)})^2} = \frac{1}{4} \quad (26.13)$$

und

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq 3\sqrt{\mathbb{V}(\xi)}) \leq \frac{\mathbb{V}(\xi)}{(3\sqrt{\mathbb{V}(\xi)})^2} = \frac{1}{9}. \quad (26.14)$$

26.2. Erwartungswertungleichungen

Die Cauchy-Schwarz Ungleichung ist eine zentrale Ungleichung der modernen Mathematik, die in verschiedenen mathematischen Bereichen wie der Analysis, der Vektorraumtheorie und eben auch der Wahrscheinlichkeitstheorie zur Anwendung kommt (vgl. Steele (2006)). In Bezug auf Erwartungswerte von Zufallsvariablen hat sie die folgende Form.

Theorem 26.3 (Cauchy-Schwarz Ungleichung). ξ und v seien zwei Zufallsvariablen und $\mathbb{E}(\xi v)$ sei endlich. Dann gilt

$$\mathbb{E}(\xi v)^2 \leq \mathbb{E}(\xi^2) \mathbb{E}(v^2). \quad (26.15)$$

◦

Für einen Beweis verweisen wir auf den Beweis von Theorem 4.6.2 in DeGroot & Schervish (2012). Analog zu Theorem 26.3 gilt zum Beispiel für Vektoren $x, y \in \mathbb{R}^n$, dass

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \langle x, y \rangle. \quad (26.16)$$

Im Kontext der probabilistischen Datenanalyse ist die Anwendung der Cauchy-Schwarz Ungleichung vor allem im Beweis der sogenannten *Korrelationsungleichung* von Relevanz.

Theorem 26.4 (Korrelationsungleichung). ξ und v seien Zufallsvariablen mit $\mathbb{V}(\xi), \mathbb{V}(v) > 0$. Dann gelten

$$\frac{\mathbb{C}(\xi, v)^2}{\mathbb{V}(\xi)\mathbb{V}(v)} \leq 1 \text{ und } -1 \leq \rho(\xi, v) \leq 1. \quad (26.17)$$

◦

Beweis. Mit der Cauchy-Schwarz-Ungleichung für zwei Zufallsvariablen α und β gilt, dass

$$\mathbb{E}(\alpha\beta)^2 \leq \mathbb{E}(\alpha^2) \mathbb{E}(\beta^2). \quad (26.18)$$

Wir definieren nun $\alpha := \xi - \mathbb{E}(\xi)$ und $\beta := v - \mathbb{E}(v)$. Dann besagt die Cauchy-Schwarz Ungleichung gerade, dass

$$\mathbb{E}((\xi - \mathbb{E}(\xi))(v - \mathbb{E}(v)))^2 \leq \mathbb{E}((\xi - \mathbb{E}(\xi))^2) \mathbb{E}((v - \mathbb{E}(v))^2). \quad (26.19)$$

Also gilt

$$\mathbb{C}(\xi, v)^2 \leq \mathbb{V}(\xi)\mathbb{V}(v) \Leftrightarrow \frac{\mathbb{C}(\xi, v)^2}{\mathbb{V}(\xi)\mathbb{V}(v)} \leq 1. \quad (26.20)$$

Weiterhin folgt aus der Definition der Korrelation dann sofort, dass auch

$$\rho(\xi, v)^2 \leq 1. \quad (26.21)$$

Dann gilt aber auch

$$|\rho(\xi, v)^2| \leq 1 \Leftrightarrow -1 \leq \rho(\xi, v) \leq 1, \quad (26.22)$$

denn

$$\rho(\xi, v)^2 \leq 1 \Rightarrow \sqrt{\rho(\xi, v)^2} \leq \sqrt{1} \Rightarrow \rho(\xi, v) \leq 1 \Rightarrow |\rho(\xi, v)| \leq 1 \text{ für } \rho(\xi, v) \geq 0 \quad (26.23)$$

und

$$\rho(\xi, v)^2 \leq 1 \Rightarrow \sqrt{\rho(\xi, v)^2} \leq \sqrt{1} \Rightarrow -\rho(\xi, v) \leq 1 \Rightarrow |\rho(\xi, v)| \leq 1 \text{ für } \rho(\xi, v) \leq 0 \quad (26.24)$$

□

Die Korrelationsungleichung wird manchmal auch als *Kovarianzungleichung* bezeichnet. Insbesondere besagt sie, dass die Korrelation von Zufallsvariablen normalisiert ist, also immer Werte zwischen -1 und 1 inklusive annimmt (vgl. Kapitel 25).

Die *Jensensche Ungleichung* schließlich liefert Abschätzungen für den Erwartungswert einer durch eine konvexe oder konkave Funktion transformierte Zufallsvariable. Sie kommt in der Betrachtung von Parameterschätzereigenschaften (vgl. Kapitel 37) und insbesondere als Grundlage der Variationalen Bayesianischen Inferenz zum Einsatz. Wir erinnern daran, dass sich eine konvexe Funktion g dadurch auszeichnet, dass der Funktionsgraph von g über einem Intervall $[x_1, x_2]$ immer unter der verbindenden Geraden zwischen den Funktionswerten $g(x_1)$ und $g(x_2)$ liegt, wohingegen bei einer konkaven Funktion g dieser immer über der verbindenden Geraden zwischen den Funktionswerten $g(x_1)$ zu $g(x_2)$ liegt. Wir visualisieren dies für eine konvexe Funktion in Abbildung 26.2.

Theorem 26.5 (Jensensche Ungleichung). ξ sei eine Zufallsvariable und $g : \mathbb{R} \rightarrow \mathbb{R}$ eine konvexe Funktion, d.h.

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2) \quad (26.25)$$

für alle $x_1, x_2 \in \mathbb{R}, \lambda \in [0, 1]$. Dann gilt

$$\mathbb{E}(g(\xi)) \geq g(\mathbb{E}(\xi)). \quad (26.26)$$

Analog sei $g : \mathbb{R} \rightarrow \mathbb{R}$ eine konkave Funktion, d.h.

$$g(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda g(x_1) + (1 - \lambda)g(x_2) \quad (26.27)$$

für alle $x_1, x_2 \in \mathbb{R}, \lambda \in [0, 1]$. Dann gilt

$$\mathbb{E}(g(\xi)) \leq g(\mathbb{E}(\xi)). \quad (26.28)$$

◦

Beweis. Es sei g eine konvexe Funktion. Dann gilt für die Tangente t von g in $x_0 \in \mathbb{R}$ für alle $x \in \mathbb{R}$, dass

$$g(x) \geq t(x) := g(x_0) + g'(x_0)(x - x_0) \quad (26.29)$$

Wir setzen nun $x := \xi$ und $x_0 := \mathbb{E}(\xi)$. Dann gilt mit obiger Ungleichung, dass

$$g(\xi) \geq g(\mathbb{E}(\xi)) + g'(\mathbb{E}(\xi))(\xi - \mathbb{E}(\xi)) \quad (26.30)$$

Erwartungswertbildung ergibt dann

$$\begin{aligned} \mathbb{E}(g(\xi)) &\geq \mathbb{E}(g(\mathbb{E}(\xi)) + g'(\mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))) \\ &\Leftrightarrow \mathbb{E}(g(\xi)) \geq g(\mathbb{E}(\xi)) + g'(\mathbb{E}(\xi))\mathbb{E}(\xi - \mathbb{E}(\xi)) \\ &\Leftrightarrow \mathbb{E}(g(\xi)) \geq g(\mathbb{E}(\xi)) + g'(\mathbb{E}(\xi))(\mathbb{E}(\xi) - \mathbb{E}(\xi)) \\ &\Leftrightarrow \mathbb{E}(g(\xi)) \geq g(\mathbb{E}(\xi)). \end{aligned} \quad (26.31)$$

Sei nun g eine konkave Funktion. Dann ist $-g$ eine konvexe Funktion. Mit der Jensenschen Ungleichung für konvexe Funktionen folgt dann die Jensensche Ungleichung für konkave Funktionen aus

$$\begin{aligned} \mathbb{E}(-g(\xi)) &\geq -g(\mathbb{E}(\xi)) \\ &\Leftrightarrow -\mathbb{E}(g(\xi)) \geq -g(\mathbb{E}(\xi)) \\ &\Leftrightarrow \mathbb{E}(g(\xi)) \leq g(\mathbb{E}(\xi)). \end{aligned} \quad (26.32)$$

□

Im Kontext der Variationalen Bayesianischen Inferenz ist grundlegend, dass der Logarithmus eine konkave Funktion ist und damit für eine beliebige Zufallsvariable ξ gilt, dass

$$\mathbb{E}(\ln \xi) \leq \ln \mathbb{E}(\xi). \quad (26.33)$$

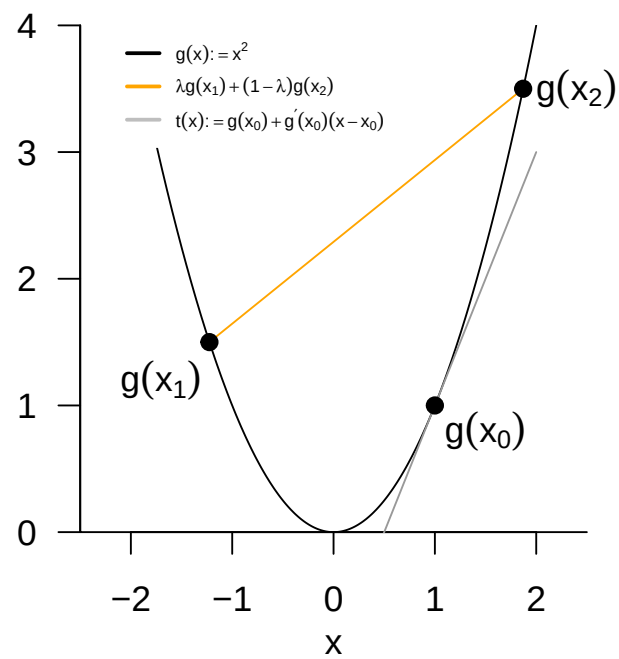


Abbildung 26.2. Darstellung der konvexen Funktion $g(x) := x^2$ mit $x_1 := -\sqrt{1.5}$, $x_2 := \sqrt{3.5}$, $\lambda \in [0, 1]$ und $x_0 := 1$.

27. Grenzwerte

In diesem Kapitel beschäftigen wir uns mit für die probabilistische Modellbildung und Datenanalyse grundlegenden Grenzwertaussagen zu Folgen von Zufallsvariablen. Dabei liefern die *Gesetze der Großen Zahlen* (Kapitel 27.1) zunächst eine grundlegende Begründung für die Mittelwertbildung im Rahmen der probabilistischen Inferenz. Die *Zentralen Grenzwertsätze* liefern dann die Begründung für die weite Verbreitung von Normalverteilungsannahmen zu Störvariablen im Rahmen der probabilistischen Modellformulierung (Kapitel 27.2). Die mathematische Tiefe dieser Grenzwertaussagen kann in dieser einführenden Betrachtung nicht ausgeschöpft werden, so dass wir uns mit zahlreichen Vereinfachungen begnügen müssen. Ein minimales Vorwissen zu Funktionenfolgen und ihren Grenzfunktionen liefert Kapitel 5.

27.1. Gesetze der Großen Zahlen

Es gibt ein *Schwaches Gesetz der Großen Zahlen* und ein *Starkes Gesetz der Großen Zahlen*. Intuitiv besagen beide Gesetze, dass sich das Stichprobenmittel von unabhängigen und identisch verteilten Zufallsvariablen für eine große Anzahl an Zufallsvariablen dem Erwartungswert der zugrundeliegenden Verteilung nähert. Das Schwache und das Starke Gesetz der Großen Zahlen unterscheiden sich in Hinblick auf die zu ihrer Formulierung benutzen Formen der *Konvergenz von Zufallsvariablen*. Das Schwache Gesetz basiert auf der *Konvergenz in Wahrscheinlichkeit*. Das Starke Gesetz basiert auf der *fast sicheren Konvergenz*. Wir begnügen uns hier mit dem Begriff der Konvergenz in Wahrscheinlichkeit und damit dem Schwachen Gesetz der Großen Zahlen.

Definition 27.1 (Konvergenz in Wahrscheinlichkeit). Eine Folge von Zufallsvariablen $\xi_1, \xi_2, \dots$ *konvergiert gegen eine Zufallsvariable ξ in Wahrscheinlichkeit*, wenn für jedes noch so kleine $\epsilon > 0$ gilt, dass

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\xi_n - \xi| < \epsilon) = 1 \Leftrightarrow \lim_{n \rightarrow \infty} \mathbb{P}(|\xi_n - \xi| \geq \epsilon) = 0. \quad (27.1)$$

Die Konvergenz von $\xi_1, \xi_2, \dots$ gegen ξ in Wahrscheinlichkeit wird geschrieben als

$$\xi_n \xrightarrow[n \rightarrow \infty]{P} \xi. \quad (27.2)$$

•

$\xi_n \xrightarrow[n \rightarrow \infty]{P} \xi$ heißt also, dass sich die Wahrscheinlichkeit dafür, dass ξ_n in dem zufälligen Intervall

$$]\xi - \epsilon, \xi + \epsilon[\quad (27.3)$$

liegt, unabhängig davon, wie klein dieses Intervall sein mag, 1 nähert, wenn n gegen Unendlich geht. Intuitiv heißt das, dass sich für $n \rightarrow \infty$ und eine konstante Zufallsvariable $\xi := a$ die Verteilung von ξ_n mehr und mehr um a konzentriert, wenn n gegen Unendlich strebt. Mithilfe der Konvergenz in Wahrscheinlichkeit formuliert man das Schwache Gesetz der Großen Zahlen wie folgt.

Theorem 27.1 (Schwaches Gesetz der Großen Zahlen). $\xi_1, \dots, \xi_n$ seien unabhängig und gleichverteilte Zufallsvariablen mit $\mathbb{E}(\xi_i) = \mu$ für alle $i = 1, \dots, n$. Weiterhin bezeichne

$$\bar{\xi}_n := \frac{1}{n} \sum_{i=1}^n \xi_i \quad (27.4)$$

das Stichprobenmittel der $\xi_i, i = 1, \dots, n$. Dann konvergiert $\bar{\xi}_n$ in Wahrscheinlichkeit gegen μ ,

$$\bar{\xi}_n \xrightarrow[n \rightarrow \infty]{P} \mu. \quad (27.5)$$

◦

Für einen Beweis dieses Theorems verweisen wir auf die weiterführende Literatur, so zum Beispiel auf Abschnitt 5.1 in Georgii (2009). Intuitiv heißt

$$\bar{\xi}_n \xrightarrow[n \rightarrow \infty]{P} \mu \quad (27.6)$$

also, dass die Wahrscheinlichkeit, dass das Stichprobenmittel nahe dem Erwartungswert der zugrundeliegenden Verteilung liegt, sich 1 nähert, wenn n gegen Unendlich strebt.

Simulation

Wir betrachten den Fall von u.i.v. normalverteilten Zufallsvariablen $\xi_1, \dots, \xi_n \sim N(0, 1)$. Abbildung 27.1 A zeigt Realisationen der von Stichprobenmitteln $\bar{\xi}_n$ als Funktion von n . Man erkennt, dass für größeres n mehr Realisierungen von $\bar{\xi}_n$ in der Nähe des Erwartungswerts der $\xi_i, i = 1, \dots, n$ liegen. Basierend auf diesen Stichprobenmittelrealisationen zeigt Abbildung 27.1 B Schätzungen der Wahrscheinlichkeit $\mathbb{P}(|\bar{\xi}_n - \mu| \geq \epsilon)$ als Funktionen von n und ϵ . Für ein großes ϵ reicht ein geringes n aus um die Wahrscheinlichkeit für eine absolute Abweichung des Stichprobenmittels vom Erwartungswert klein werden zu lassen, für ein kleineres ϵ ist dafür ein größeres n nötig. In jedem Fall sinken die Wahrscheinlichkeiten jedoch mit größerem n .

27.2. Zentrale Grenzwertsätze

Die Zentralen Grenzwertsätze besagen intuitiv, dass die Summe von unabhängigen Zufallsvariablen mit Erwartungswert Null *asymptotisch*, also für unendlich viele Zufallsvariablen, normalverteilt mit Erwartungswertparameter Null ist. Modelliert man eine beliebige Messgröße v also als Summe eines deterministischen Einflusses μ und der Summe

$$\varepsilon := \sum_{i=1}^n \xi_i \quad (27.7)$$

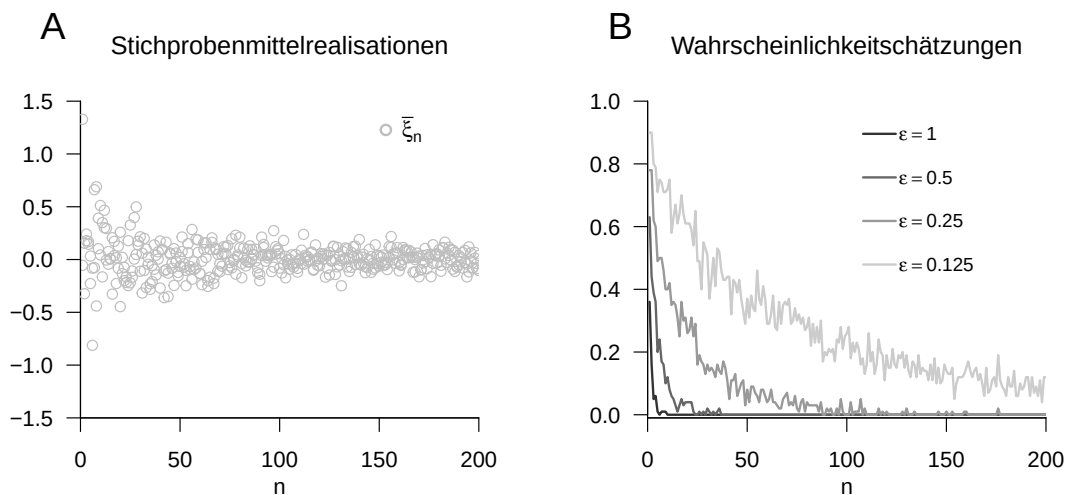


Abbildung 27.1. Simulation des schwachen Gesetz der Großen Zahlen.

einer Vielzahl von unabhängigen Zufallsvariablen $\xi_i, i = 1, \dots, n$, die unbekannte Störeinflüsse modellieren sollen, so ist für großes n die Annahme

$$v = \mu + \varepsilon \text{ mit } \varepsilon \sim N(0, \sigma^2) \quad (27.8)$$

mathematisch gerechtfertigt. Wie wir später sehen werden, liegt die Annahme in Gleichung Gleichung 27.8 einer großen Vielzahl von probabilistischen Modellen zugrunde.

Formal werden verschiedene Formen von Zentralen Grenzwertsätzen, je nach Ausgestaltung der zugrundeliegenden Annahmen und ihrer Beweisführung unterschieden. In der sogenannten *Lindenberg und Lévy Form* des Zentralen Grenzwertsatzes werden unabhängig und identische Zufallsvariablen vorausgesetzt. In der *Liapunov Form* dagegen werden nur unabhängige Zufallsvariablen vorausgesetzt. In beiden Formulierungen des Zentralen Grenzwertsatzes ist die betrachtete Konvergenz von Zufallsvariablen die *Konvergenz in Verteilung*, welche wir zunächst einführen.

Definition 27.2 (Konvergenz in Verteilung). Eine Folge $\xi_1, \xi_2, \dots$ von Zufallsvariablen *konvergiert in Verteilung gegen eine Zufallsvariable ξ* , wenn

$$\lim_{n \rightarrow \infty} P_{\xi_n}(x) = P_{\xi}(x) \quad (27.9)$$

für alle ξ an denen P_{ξ} stetig ist. Die Konvergenz in Verteilung von $\xi_1, \xi_2, \dots$ gegen ξ wird geschrieben als

$$\xi_n \xrightarrow[n \rightarrow \infty]{D} \xi. \quad (27.10)$$

Gilt $\xi_n \xrightarrow[n \rightarrow \infty]{D} \xi$, dann heißt die Verteilung von ξ die *asymptotische Verteilung der Folge $\xi_1, \xi_2, \dots$*

•

Die Konvergenz in Verteilung ist also eine Aussage zur Konvergenz von Funktionenfolgen, speziell von KVFn. Ohne Begründung merken wir an, dass die oben betrachtete

Konvergenz in Wahrscheinlichkeit die Konvergenz in Verteilung impliziert. Wir geben nun zunächst den Zentralen Grenzwertsatz nach Lindenberg und Lévy an.

Theorem 27.2 (Zentraler Grenzwertsatz nach Lindenberg und Lévy). $\xi_1, \dots, \xi_n$ seien unabhängig und identisch verteilte Zufallsvariablen mit

$$\mathbb{E}(\xi_i) := \mu \text{ und } \mathbb{V}(\xi_i) := \sigma^2 > 0 \text{ für alle } i = 1, \dots, n. \quad (27.11)$$

Weiterhin sei ζ_n die Zufallsvariable definiert als

$$\zeta_n := \sqrt{n} \left(\frac{\bar{\xi}_n - \mu}{\sigma} \right). \quad (27.12)$$

Dann gilt für alle $z \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} P_{\zeta_n}(z) = \Phi(z), \quad (27.13)$$

wobei Φ die kumulative Verteilungsfunktion der Standardnormalverteilung bezeichnet.

◦

Wir zeigen an späterer Stelle, dass damit für $n \rightarrow \infty$ auch gilt, dass

$$\sum_{i=1}^n \xi_i \sim N(\mu, n\sigma^2) \text{ und } \bar{\xi}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right). \quad (27.14)$$

Simulation

Wir betrachten den Fall von u.i.v. χ^2 -Zufallsvariablen $\xi_1, \dots, \xi_n \sim \chi^2(3)$. Offenbar ist die funktionale Form der $\chi^2(3)$ -Verteilung von der Standardnormalverteilung recht verschieden, insbesondere nehmen χ^2 -Zufallsvariablen mit von Null verschiedener Wahrscheinlichkeit nur nicht-negative Werte an (vgl. Kapitel 21.3). Nichtsdestotrotz resultiert ihre standardisierte Summe asymptotisch in einer Normalverteilung, wie in Abbildung 27.2 visualisiert. Dazu nutzen wir auf Ebene der Implementation die Tatsache, für die χ^2 -Zufallsvariablen $\xi_i, i = 1, \dots, n$ mit Freiheitsgradparameter 3 bekanntlich gilt (vgl. Kapitel 23 und Kapitel 24)

$$\mathbb{E}(\xi_i) = 3 \text{ und } \mathbb{V}(\xi_i) = 6 \quad (27.15)$$

Die Abbildungen in Abbildung 27.2 A zeigen Histogrammschätzer der Wahrscheinlichkeitsdichte von

$$\zeta_n := \sqrt{n} \left(\frac{\bar{\xi}_n - \mu}{\sigma} \right) \quad (27.16)$$

basierend auf 1000 Realisationen von ζ_n für $n = 2$ und $n = 200$, sowie die WDF von $N(0, 1)$. Offenbar ist die Verteilung der Realisationen von ζ_2 der Standardnormalverteilung noch sehr unähnlich, wohingegen sich die Verteilung der Realisationen von ζ_{200} der Standardnormalverteilung schon annähert. Abbildung 27.2 B zeigt die entsprechenden geschätzten KVFen über die Theorem 27.2 formal eine Aussage trifft.

Theorem 27.3 (Zentraler Grenzwertsatz nach Liapounov). $\xi_1, \dots, \xi_n$ seien unabhängige aber nicht notwendigerweise identisch verteilten Zufallsvariablen mit

$$\mathbb{E}(\xi_i) := \mu_i \text{ und } \mathbb{V}(\xi_i) := \sigma_i^2 > 0 \text{ für alle } i = 1, \dots, n. \quad (27.17)$$

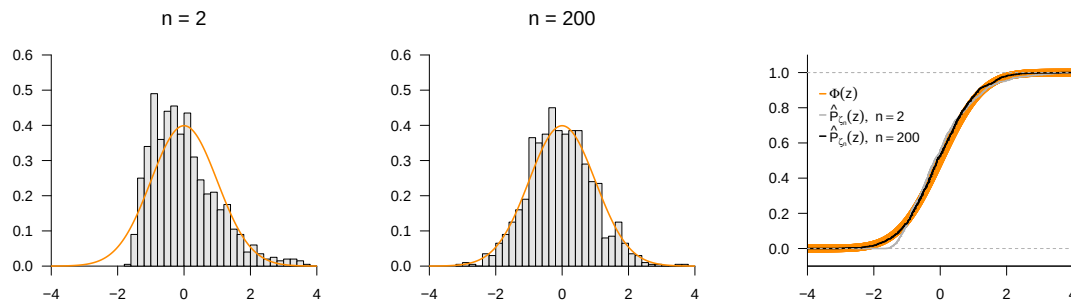


Abbildung 27.2. Simulation des Zentralen Grenzwertsatzes nach Lindenberg und Lévy.

Weiterhin sollen für $\xi_1, \dots, \xi_n$ folgende Eigenschaften gelten:

$$\mathbb{E}(|\xi_i - \mu_i|^3) < \infty \text{ und } \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \mathbb{E}(|\xi_i - \mu_i|^3)}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0. \quad (27.18)$$

Dann gilt für die Zufallsvariable ζ_n definiert als

$$\zeta_n := \frac{\sum_{i=1}^n \xi_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}, \quad (27.19)$$

für alle $z \in \mathbb{R}$, dass

$$\lim_{n \rightarrow \infty} P_{\zeta_n}(z) = \Phi(z), \quad (27.20)$$

wobei Φ KVF der Standardnormalverteilung bezeichnet.

◦

Wir zeigen an späterer Stelle, dass damit für $n \rightarrow \infty$ auch gilt, dass

$$\sum_{i=1}^n \xi_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right) \quad (27.21)$$

27.3. Literaturhinweise

Zur mathematik-geschichtlichen Genese der Zentralen Grenzwertsätze siehe z.B. Fischer (2011).

28. Transformationen

Die Transformation von Zufallsvariablen und ihren Verteilungen ist ein zentrales Thema der probabilistischen Modellierung und in besonderem Maße der Frequentistischen Inferenz. Mit einer *Transformation* sollen hier die Anwendung von Funktionen auf Zufallsvariablen sowie die arithmetische Verknüpfung mehrerer Zufallsvariablen gemeint sein. Die zentrale Fragestellung dabei ist folgende: “Wenn eine Zufallsvariable ξ eine durch ihre WDF fest vorgegebene Verteilung hat, wie ist dann eine Zufallsvariable v , die sich durch Transformation von ξ ergibt, verteilt?” Für die in diesem Kapitel behandelten Fälle gilt dabei, dass man explizit WDFen für die Verteilung der transformierten Zufallsvariable angeben kann. Diese gehören zu den klassischen Resultaten der Frequentistischen Inferenz und sind für das Verständnis von klassischer Frequentistischer Inferenzverfahren wie Konfidenzintervallen und Hypothesentests essentiell.

Intuitiv kann man sich die Transformation einer Zufallsvariable anhand der Transformation ihrer unabhängig und identisch verteilten Realisierungen klar machen. Betrachtet man beispielsweise die Zufallsvariable $\xi \sim N(0, 1)$ und ihre Transformation $v := \xi^2$ und sind $x_1 = 0.10, x_2 = -0.20, x_3 = 0.80$ drei Realisierungen von drei unabhängigen Kopien von ξ , so entspricht dies den Realisierungen $y_1 = x_1^2 = 0.01, y_2 = x_2^2 = 0.04, y_3 = x_3^2 = 0.64$ von v . In diesem Beispiel fällt auf, dass v keine negativen Werte annimmt, die Verteilung von v ordnet negativen Werten daher Wahrscheinlichkeitsdichten von 0 zu. Untenstehender **R** Code simuliert diese Überlegungen. Abbildung 28.1 zeigt das Histogramm der gewonnenen Realisierungen der hier betrachteten Zufallsvariable ξ und Abbildung 28.2 zeigt das Histogramm der quadrierten Realisierungen von ξ , also unabhängig und identisch verteilte Realisierungen von v .

```
# Simulationsspezifikation
n      = 1e4                # Anzahl von u.i.v Realisierungen (ZVen)
mu     = 1                  # Erwartungswertparameter von \xi
sigsqr = 2                  # Varianzparameter von \xi

# Quadrieren einer Zufallsvariable
x      = rnorm(n, mu, sqrt(sigsqr)) # Realisierungen x_i, i = 1,...,n von \xi
y      = x^2                 # Realisierungen y_i = x_i^2 von \upsilon

# Ausgabe der ersten acht Werte der Realisierungen von \xi und von \upsilon
print(x[1:8], digits = 2)
```

```
[1] 2.15 -2.18 0.22 1.54 1.32 1.38 0.85 -0.16
```

```
print(y[1:8], digits = 2)
```

```
[1] 4.605 4.757 0.050 2.382 1.734 1.912 0.724 0.026
```

Grundlegend für die nachfolgenden Betrachtungen ist folgendes Theorem, das wir nicht beweisen wollen.

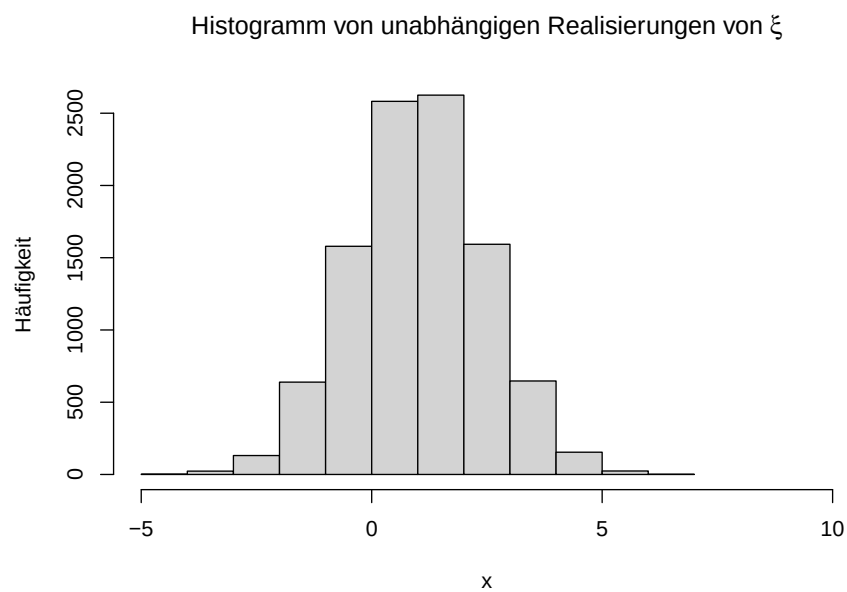


Abbildung 28.1. Histogramm von 10.000 Realisierungen unabhängig und identisch normalverteilter Zufallsvariablen.

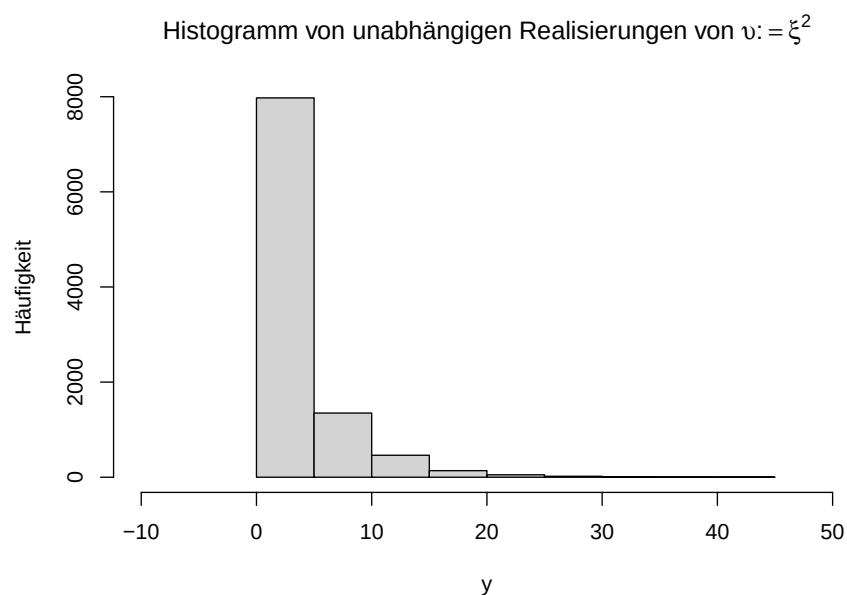


Abbildung 28.2. Histogramm von 10.000 quadrierten Realisierungen unabhängig und identisch normalverteilter Zufallsvariablen.

Theorem 28.1 (Transformation eines Zufallsvektors). $\xi : \Omega \rightarrow \mathcal{X}$ sei ein Zufallsvektor und $f : \mathcal{X} \rightarrow \mathbb{R}^m$ sei eine multivariate vektorwertige Funktion. Dann ist

$$v : \Omega \rightarrow \mathbb{R}, \omega \mapsto v(\omega) := (f \circ \xi)(\omega) := f(\xi(\omega)) \quad (28.1)$$

ein Zufallsvektor.

◦

Das Theorem formalisiert die oben etablierte Intuition, dass die Anwendung einer (deterministischen) Funktion auf eine zufällige Größe im Allgemeinen wieder eine zufällige Größe ergibt. In einem Beweis von Theorem 28.1 müsste die Messbarkeit von v als Folge der Messbarkeit von ξ nachgewiesen werden. Im Folgenden ist oft $\mathcal{X} := \mathbb{R}$ und $f : \mathbb{R} \rightarrow \mathbb{R}$. Wir schreiben in diesem Fall in der Regel einfach $v := f(\xi)$ und nennen v die *transformierte Zufallsvariable*. In Kapitel 28.1 betrachten wir drei Theoreme zur Transformation von Zufallsvariablen in Abhängigkeit von der Art der betrachteten Funktion f . In Kapitel 28.2 betrachten wir ein Theorem zur Transformation von Zufallsvektoren bei bijektiver Transformationsfunktion. In Kapitel 28.3 schließlich betrachten wir, wie der Verteilungen von Verknüpfungen von Zufallsvariablen analytisch bestimmt werden können.

28.1. Univariate Transformationstheoreme

Dabei liefert Theorem 28.2 eine Formel zur Berechnung der WDF p_v von $v := f(\xi)$, wenn ξ eine Zufallsvariable mit WDF p_ξ ist und f eine bijektive Funktion ist. Theorem 28.3 gibt weiterhin eine vereinfachte Formel zur Berechnung der WDF p_v von $v := f(\xi)$ an, wenn f speziell eine linear-affine Funktion ist. Theorem 28.4 schließlich gibt eine Formel zur Berechnung der WDF p_v von $v := f(\xi)$ an, wenn f zumindest in Teilen bijektiv ist.

Theorem 28.2 (Univariate WDF Transformation bei bijektiven Abbildungen). ξ sei eine Zufallsvariable mit WDF p_ξ für die $\mathbb{P}(]a, b[) = 1$ gilt, wobei a und/oder b entweder endlich oder unendlich seien. Weiterhin sei

$$v := f(\xi), \quad (28.2)$$

wobei die univariate reellwertige Funktion $f :]a, b[\rightarrow \mathbb{R}$ differenzierbar und bijektiv auf $]a, b[$ sei. $f(]a, b[)$ sei das Bild von $]a, b[$ unter f . Schließlich sei $f^{-1}(y)$ der Wert der Umkehrfunktion von $f(x)$ für $y \in f(]a, b[)$ und $f'(x)$ sei die Ableitung von f an der Stelle x . Dann ist die WDF von v gegeben durch

$$p_v : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_v(y) := \begin{cases} \frac{1}{|f'(f^{-1}(y))|} p_\xi(f^{-1}(y)) & \text{für } y \in f(]a, b[) \\ 0 & \text{für } y \in \mathbb{R} \setminus f(]a, b[). \end{cases} \quad (28.3)$$

◦

Beweis. Wir halten zunächst fest, dass weil f eine differenzierbare bijektive Funktion auf $]a, b[$ ist, f entweder strikt wachsend oder strikt fallend ist. Nehmen wir zunächst an, dass f auf $]a, b[$ strikt wachsend ist. Dann ist auch f^{-1} für alle $y \in f(]a, b[)$ wachsend, und es gilt

$$P_v(y) = \mathbb{P}(v \leq y) = \mathbb{P}(f(\xi) \leq y) = \mathbb{P}(f^{-1}(f(\xi)) \leq f^{-1}(y)) = \mathbb{P}(\xi \leq f^{-1}(y)) = P_\xi(f^{-1}(y)).$$

P_v ist also differenzierbar an allen Stellen y , an denen sowohl f^{-1} als auch P_ξ differenzierbar sind. Mit der Kettenregel und dem Satz von der Umkehrabbildung $(f^{-1}(x))' = 1/f'(f^{-1}(x))$, folgt dann, dass die WDF p_v sich ergibt wie folgt:

$$p_v(y) = \frac{d}{dy} P_v(y) = \frac{d}{dy} P_\xi(f^{-1}(y)) = p_\xi(f^{-1}(y)) \frac{d}{dy} f^{-1}(y) = \frac{1}{f'(f^{-1}(y))} p_\xi(f^{-1}(y)),$$

Weil f^{-1} strikt wachsend ist, ist $d/dy(f^{-1}(y))$ positiv und das Theorem trifft zu. Analog gilt, dass wenn f auf $]a, b[$ strikt fallend ist, dann ist auch f^{-1} für alle $y \in f(]a, b[)$ fallend und es gilt

$$P_v(y) = \mathbb{P}(f(\xi) \leq y) = \mathbb{P}(f^{-1}(f(\xi)) \geq f^{-1}(y)) = \mathbb{P}(\xi \geq f^{-1}(y)) = 1 - P_\xi(f^{-1}(y)),$$

Mit der Kettenregel und dem Satz von der Umkehrabbildung folgt dann

$$p_v(y) = \frac{d}{dy}(1 - P_v(y)) = -\frac{d}{dy} P_\xi(f^{-1}(y)) = -p_\xi(f^{-1}(y)) \frac{d}{dy} f^{-1}(y) = -\frac{1}{f'(f^{-1}(y))} p_\xi(f^{-1}(y)).$$

Weil f^{-1} strikt fallend ist, ist $d/dy(f^{-1}(y))$ negativ, so dass $-d/dy(f^{-1}(y))$ gleich $|d/dy(f^{-1}(y))|$ ist und das Theorem trifft zu. $\square$

Ein wichtiger Anwendungsfall von Theorem 28.2 ist Theorem 28.3.

Theorem 28.3 (Univariates WDF Transformationstheorem bei linear-affinen Abbildungen). ξ sei eine Zufallsvariable mit WDF p_ξ und es sei

$$v = f(\xi) \text{ mit } f(\xi) := a\xi + b \text{ für } a \neq 0. \quad (28.4)$$

Dann ist die WDF von v gegeben durch

$$p_v : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_v(y) := \frac{1}{|a|} p_\xi\left(\frac{y-b}{a}\right). \quad (28.5)$$

◦

Beweis. Wir halten zunächst fest, dass

$$f^{-1} : \mathbb{R} \rightarrow \mathbb{R}, y \mapsto f^{-1}(y) = \frac{y-b}{a} \quad (28.6)$$

ist, weil dann $f \circ f^{-1} = \text{id}_{\mathbb{R}}$ gilt, wie man anhand von

$$f(f^{-1}(x)) = a\left(\frac{x-b}{a}\right) + b = x - b + b = x \text{ für alle } x \in \mathbb{R} \quad (28.7)$$

einsieht. Wir halten weiterhin fest, dass

$$f' : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f'(x) = \frac{d}{dx}(ax + b) = a. \quad (28.8)$$

Also folgt mit Theorem 28.2, dass

$$\begin{aligned} p_v : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_v(y) &= \frac{1}{|f'(f^{-1}(y))|} p_\xi(f^{-1}(y)) \\ &= \frac{1}{|a|} p_\xi\left(\frac{y-b}{a}\right). \end{aligned} \quad (28.9)$$

$\square$

Ein wichtiger Anwendungsfall dieses Theorems ist die in **sec-normalverteilungstransformationen** betrachtete *Z-Transformation*. Das folgende Theorem, dass wir nicht beweisen wollen, verallgemeinert Theorem 28.2 auf den Fall nur stückweise bijektiver Abbildungen.

Theorem 28.4 (Univariate WDF Transformation bei stückweise bijektiven Abbildungen). ξ sei eine Zufallsvariable mit Ergebnisraum $\mathcal{X}$ und WDF p_ξ . Weiterhin sei

$$v = f(\xi), \quad (28.10)$$

wobei f so beschaffen sei, dass der Ergebnisraum von ξ in eine endliche Anzahl von Mengen $\mathcal{X}_1, \dots, \mathcal{X}_k$ mit einer entsprechenden Anzahl von Mengen $\mathcal{Y}_1 := f(\mathcal{X}_1), \dots, \mathcal{Y}_k := f(\mathcal{X}_k)$ im Ergebnisraum $\mathcal{Y}$ von v partitioniert werden kann (wobei nicht notwendigerweise $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset, 1 \leq i, j \leq k$ gelten muss), so dass die Abbildung f für alle $\mathcal{X}_1, \dots, \mathcal{X}_k$ bijektiv ist (d.h. f ist eine stückweise bijektive Abbildung). Für $i = 1, \dots, k$ bezeichne f_i^{-1} die Umkehrfunktion von f auf $\mathcal{Y}_i$. Schließlich nehmen wir an, dass die Ableitungen f'_i für alle $i = 1, \dots, k$ existieren und stetig sind. Dann ist eine WDF von v durch

$$p_v : \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_v(y) := \sum_{i=1}^k 1_{\mathcal{Y}_i}(y) \frac{1}{|f'_i(f_i^{-1}(y))|} p_\xi(f_i^{-1}(y)). \quad (28.11)$$

gegeben.

◦

Ein wichtiger Anwendungsfall ist die in [?@sec-normalverteilungstransformationen](#) betrachtete χ^2 -Transformation.

28.2. Multivariate WDF Transformationstheoreme

Theorem 28.5 liefert eine Formel zur Berechnung der WDF p_v von $v := f(\xi)$, wenn ξ ein Zufallsvektor mit WDF p_ξ ist und f eine bijektive multivariate vektorwertige Funktion ist. Es handelt sich dabei um eine direkte Generalisierung von Theorem 28.2 und wir verzichten auf einen Beweis.

Theorem 28.5 (Multivariate WDF Transformation bei bijektiven Abbildungen). ξ sei ein n -dimensionaler Zufallsvektor mit Ergebnisraum $\mathbb{R}^n$ und WDF p_ξ . Weiterhin sei

$$v := f(\xi), \quad (28.12)$$

wobei die multivariate vektorwertige Funktion $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ differenzierbar und bijektiv auf $[a, b]$ sei. Schließlich seien

$$J^f(x) = \left(\frac{\partial}{\partial x_j} f_i(x) \right)_{1 \leq i \leq n, 1 \leq j \leq n} \in \mathbb{R}^{n \times n} \quad (28.13)$$

die Jacobi-Matrix von f an der Stelle $x \in \mathbb{R}^n$, $|J^f(x)|$ die Determinante von $J^f(x)$, und es sei $|J^f(x)| \neq 0$ für alle $x \in \mathbb{R}^n$. Dann ist eine WDF von v durch

$$p_v : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_v(y) := \begin{cases} \frac{1}{|J^f(f^{-1}(y))|} p_\xi(f^{-1}(y)) & \text{für } y \in f(\mathbb{R}^n) \\ 0 & \text{für } y \in \mathbb{R}^n \setminus f(\mathbb{R}^n) \end{cases} \quad (28.14)$$

gegeben.

◦

Wichtige Anwendungsfälle sind die in [?@sec-normalverteilungstransformationen](#) betrachteten T - und F -Transformationen.

28.3. Operationstheoreme

Das folgende sogenannte *Konvolutionstheorem* liefert eine Formel zur Berechnung der WDF p_v von $v := \xi_1 + \xi_2$, wenn ξ_1 und ξ_2 zwei Zufallsvariablen mit WDFen p_{ξ_1} und p_{ξ_2} sind.

Theorem 28.6 (Summe unabhängiger Zufallsvariablen (Konvolution)). ξ_1 und ξ_2 seien zwei kontinuierliche unabhängige Zufallsvariablen mit WDF p_{ξ_1} und p_{ξ_2} , respektive. $v := \xi_1 + \xi_2$ sei die Summe von ξ_1 und ξ_2 . Dann ergibt sich eine WDF der Verteilung von v als

$$p_v(y) = \int_{-\infty}^{\infty} p_{\xi_1}(y - x_2) p_{\xi_2}(x_2) dx_2 = \int_{-\infty}^{\infty} p_{\xi_1}(x_1) p_{\xi_2}(y - x_1) dx_1 \quad (28.15)$$

Die Formel für die WDF p_v heißt *Konvolution* oder *Faltung* von p_{ξ_1} und p_{ξ_2} .

◦

Beweis. Wir nutzen das multivariate WDF Transformationstheorem für bijektive Abbildungen. Dazu definieren wir zunächst

$$f: \mathbb{R}^2 \rightarrow \mathbb{R}^2, x \mapsto f(x) := \begin{pmatrix} x_1 + x_2 \\ x_2 \end{pmatrix} := \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad (28.16)$$

Die inverse Funktion von f ist dann gegeben durch

$$f: \mathbb{R}^2 \rightarrow \mathbb{R}^2, z \mapsto f(z) := \begin{pmatrix} z_1 - x_2 \\ z_2 \end{pmatrix} \quad (28.17)$$

weil dann $f \circ f^{-1} = \text{id}_{\mathbb{R}^2}$ gilt, wie man anhand von

$$f^{-1}(f(x)) = f^{-1} \begin{pmatrix} x_1 + x_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 - x_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad (28.18)$$

einsieht. Die Jacobimatrix von f ergibt sich zu

$$Jf(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} f_1(x) & \frac{\partial}{\partial x_2} f_1(x) \\ \frac{\partial}{\partial x_1} f_2(x) & \frac{\partial}{\partial x_2} f_2(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial x_1} (x_1 + x_2) & \frac{\partial}{\partial x_2} (x_1 + x_2) \\ \frac{\partial}{\partial x_1} x_2 & \frac{\partial}{\partial x_2} x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad (28.19)$$

und die Jacobideterminante damit zu $|Jf(x)| = 1$. Wir halten weiterhin fest, dass die Unabhängigkeit von ξ_1 und ξ_2 impliziert, dass

$$p_{\xi_1, \xi_2}(x_1, x_2) = p_{\xi_1}(x_1) p_{\xi_2}(x_2) \quad (28.20)$$

impliziert. Einsetzen und Integration hinsichtlich x_2 ergibt dann ergibt dann für $z \in f(\mathbb{R}^2)$

$$\begin{aligned} p_{\zeta}(z) &= \frac{1}{|Jf(f^{-1}(z))|} p_{\xi}(f^{-1}(z)) \\ &= \frac{1}{1} p_{\xi_1, \xi_2}(z_1 - x_2, x_2) \\ &= p_{\xi_1}(z_1 - x_2) p_{\xi_2}(x_2) \end{aligned} \quad (28.21)$$

Integration über x_2 ergibt dann eine WDF für die marginale Verteilung von ζ_1

$$p_{\zeta_1}(z_1) = \int_{-\infty}^{\infty} p_{\xi_1}(z_1 - x_2) p_{\xi_2}(x_2) dx_2 \quad (28.22)$$

Mit $\zeta_1 = \xi_1 + \xi_2 = v$ ergibt sich dann die erste Form des Konvolutionstheorems zu

$$p_v(y) = \int_{-\infty}^{\infty} p_{\xi_1}(y - x_2) p_{\xi_2}(x_2) dx_2. \quad (28.23)$$

□

Wichtige in **sec-normalverteilungstransformationen** betrachtete Anwendungsfälle sind die *Summentransformation* und die *Mittelwerttransformation*.

29. Normalverteilungen

Die multivariate Normalverteilung ist die multivariate Generalisierung der univariaten Normalverteilung. Die Motivation für die verbreiteten Normalverteilungsannahmen in der probabilistischen Modellierung liegt bekanntermaßen im Zentralen Grenzwertsatz: In probabilistischen Modellen repräsentieren probabilistische Terme die Summation sehr vieler Zufallsvorgänge, die durch die deterministischen Bestandteile des jeweiligen Modells, also eine formalisierte wissenschaftliche Theorie, nicht erklärt werden. Nach dem Zentralen Grenzwertsatz ist die Summe dieser nicht erklärten Zufallsvorgänge dann aber gerade normalverteilt.

Über diesen grundlegenden Aspekt hinaus hat die Normalverteilung viele günstige mathematische Eigenschaften, die ihren Einsatz in vielen Bereichen der probabilistischen Modellierung ermöglichen. Anwendungen multivariater Normalverteilungen finden sich damit im Kontext des Allgemeinen Linearen Modells, den Generalisierungen des Allgemeinen Modells zu Hierarchischen Linearen Modellen oder Multivariaten Allgemeinen Linearen Modellen, der Bayesianischen Inferenz bei Normalverteilungsannahmen und nicht zuletzt der Theorie probabilistischer Filter, wie zum Beispiel dem Kalman-Bucy Filter.

29.1. Konstruktion multivariater Normalverteilungen

In diesem Abschnitt wollen wir zeigen, wie ein bivariat normalverteilter Zufallsvektor durch Transformation und Konkatination zweier univariat normalverteilter Zufallsvariablen konstruiert werden kann. Dazu erinnern wir zunächst an den Begriff der normalverteilten Zufallsvariable

Definition 29.1 (Normalverteilte Zufallsvariable). ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (29.1)$$

Dann sagen wir, dass ξ einer *Normalverteilung* (oder *Gauß-Verteilung*) mit *Erwartungswertparameter* $\mu \in \mathbb{R}$ und *Varianzparameter* $\sigma^2 > 0$ unterliegt und nennen ξ eine *normalverteilte Zufallsvariable*. Wir kürzen dies mit $\xi \sim N(\mu, \sigma^2)$ ab. Die WDF einer normalverteilten Zufallsvariable bezeichnen wir mit

$$N(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (29.2)$$

•

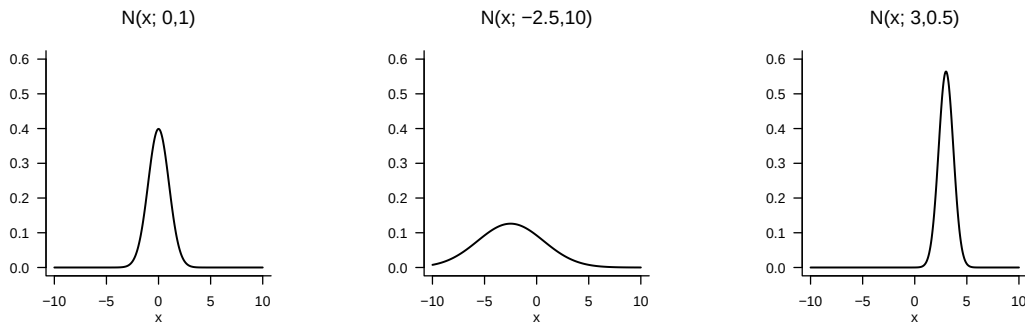


Abbildung 29.1. Wahrscheinlichkeitsdichtefunktionen univariater Normalverteilungen.

Visuell entspricht der Parameter μ einer normalverteilten Zufallsvariable dem Wert höchster Wahrscheinlichkeitsdichte und der Parameter σ^2 spezifiziert die Breite der WDF (Abbildung 29.1). Weiterhin gelten für den Erwartungswert und die Varianz einer normalverteilten Zufallsvariable bekanntlich

$$\mathbb{E}(\xi) = \mu \text{ und } \mathbb{V}(\xi) = \sigma^2. \quad (29.3)$$

Eine normalverteilte Zufallsvariable der Form $\xi \sim N(0, 1)$ schließlich heißt auch *standardnormalverteilt*.

Folgendes Theorem zeigt, wie zwei unabhängige, univariat standardnormalverteilte Zufallsvariablen kombiniert werden können, um einen bivariat verteilten Zufallsvektor zu konstruieren. Die Verteilung eines ebensolchen Zufallsvektors wird dann als *bivariate Normalverteilung* bezeichnet.

Theorem 29.1 (Konstruktion bivariater Normalverteilungen). $\zeta_1 \sim N(0, 1)$ und $\zeta_2 \sim N(0, 1)$ seien zwei unabhängige standardnormalverteilte Zufallsvariablen. Weiterhin seien $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$ und $\rho \in]-1, 1[$. Schließlich seien

$$\begin{aligned} \xi_1 &:= \sigma_1 \zeta_1 + \mu_1 \\ \xi_2 &:= \sigma_2 (\rho \zeta_1 + (1 - \rho^2)^{1/2} \zeta_2) + \mu_2. \end{aligned} \quad (29.4)$$

Dann hat die WDF des Zufallsvektors $\xi := (\xi_1, \xi_2)^T$, also der gemeinsamen Verteilung von ξ_1 und ξ_2 , die Form

$$p : \mathbb{R}^2 \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right), \quad (29.5)$$

wobei $n := 2$ und $\mu \in \mathbb{R}^2$ und $\Sigma \in \mathbb{R}^{2 \times 2}$ durch

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ und } \Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_2 \sigma_1 & \sigma_2^2 \end{pmatrix} \quad (29.6)$$

gegeben sind.

◦

Für einen Beweis des Theorems verweisen wir auf DeGroot & Schervish (2012).

Beispiel

Folgender **R** Code zeichnet das obige Theorem anhand konkreter Beispielwerte für $\mu_1, \mu_2, \sigma_1, \sigma_2$ und ρ nach und gibt die Parameter μ und Σ der resultierenden bivariaten WDF aus.

```
# Parameterdefinitionen
mu_1 = 5.0          # \mu_1
mu_2 = 4.0          # \mu_2
sig_1 = 1.5         # \sigma_1
sig_2 = 1.0         # \sigma_2
rho = 0.9           # \rho

# Realisierungen der standardnormalverteilten ZVen
n = 100             # Anzahl Realisierungen
zeta_1 = rnorm(n)    # \zeta_1 \sim N(0,1)
zeta_2 = rnorm(n)    # \zeta_2 \sim N(0,1)

# Evaluation von Realisierungen von \xi_1 und \xi_2
xi_1 = sig_1*zeta_1 + mu_1 # Realisierungen von zeta_1
xi_2 = sig_2*(rho*zeta_1 + sqrt(1-rho^2)*zeta_2) + mu_2 # Realisierungen von zeta_2

# Parameter der gemeinsamen Verteilung von \xi_1 und \xi_2
mu = matrix(c(mu_1, # \mu \in \mathbb{R}^2
              mu_2),
            nrow = 2, byrow = TRUE)
Sigma = matrix(c(sig_1^2, rho*sig_1*sig_2, # \Sigma \in \mathbb{R}^{2 \times 2}
                 rho*sig_1*sig_2, sig_2^2),
              nrow = 2, byrow = TRUE)

print(mu)
```

```
      [,1]
[1,] 5
[2,] 4
```

```
print(Sigma)
```

```
      [,1] [,2]
[1,] 2.25 1.35
[2,] 1.35 1.00
```

Die durch obigen **R** Code generierten Realisierungen von $\xi = (\xi_1, \xi_2)^T$, sowie die Isokon-
turen der durch Theorem 29.1 postulierten WDF sind in Abbildung 29.2 dargestellt.

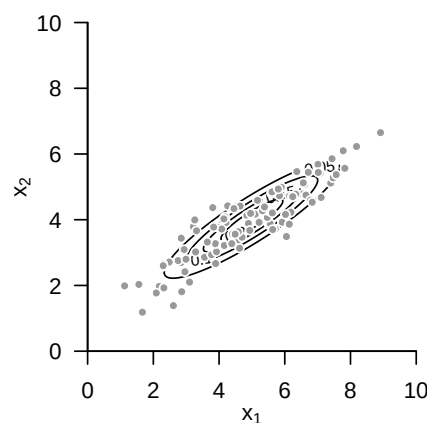


Abbildung 29.2. Konstruktion bivariater Normalverteilungen.

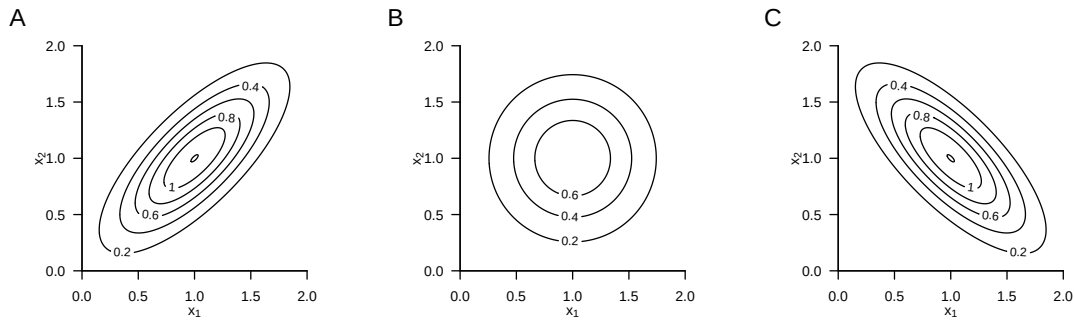


Abbildung 29.3. Wahrscheinlichkeitsdichtefunktionen bivariater Normalverteilungen.

29.2. Definition multivariater Normalverteilungen

Wir wollen die multivariate Normalverteilung nun formal einführen und erste Eigenschaften angeben. Wir nutzen dazu folgende Definition.

Definition 29.2. ξ sei ein n -dimensionaler Zufallsvektor mit Ergebnisraum $\mathbb{R}^n$ und WDF

$$p : \mathbb{R}^n \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right). \quad (29.7)$$

Dann sagen wir, dass ξ einer *multivariaten (oder n -dimensionalen) Normalverteilung* mit *Erwartungswertparameter* $\mu \in \mathbb{R}^n$ und positiv-definitem *Kovarianzmatrixparameter* $\Sigma \in \mathbb{R}^{n \times n}$ unterliegt und nennen ξ einen *(multivariat) normalverteilten Zufallsvektor*. Wir kürzen dies mit $\xi \sim N(\mu, \Sigma)$ ab. Die WDF eines multivariat normalverteilten Zufallsvektors bezeichnen wir mit

$$N(x; \mu, \Sigma) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right). \quad (29.8)$$

•

Abbildung 29.3 A,B,C zeigen die Isokonturen der WDFen bivariat normalverteilter Zufallsvektoren für $\mu = (1, 1)^T$ und

$$\Sigma_A := \begin{pmatrix} 0.20 & 0.15 \\ 0.15 & 0.20 \end{pmatrix}, \quad \Sigma_B := \begin{pmatrix} 0.20 & 0.00 \\ 0.00 & 0.20 \end{pmatrix}, \quad \Sigma_C := \begin{pmatrix} 0.20 & -0.15 \\ -0.15 & 0.20 \end{pmatrix} \quad (29.9)$$

respektive. Abbildung 29.4 A,B und C zeigen jeweils 200 Realisierungen der entsprechenden Zufallsvektoren.

Ohne Beweis halten wir fest, dass wie im Fall einer univariat normalverteilten Zufallsvariable, der Erwartungswert und die Kovarianzmatrix eines normalverteilten Zufallsvektors durch die entsprechenden Parameter gegeben sind.

Theorem 29.2 (Erwartungswert und Kovarianzmatrix normalverteilter Zufallsvektoren). $\xi \sim N(\mu, \Sigma)$ sei ein multivariat normalverteilter Zufallsvektor mit Erwartungswertparameter $\mu \in \mathbb{R}^n$ und Kovarianzmatrixparameter $\Sigma \in \mathbb{R}^{n \times n}$ pd. Dann gelten

$$\mathbb{E}(\xi) = \mu \text{ und } \text{C}(\xi) = \Sigma. \quad (29.10)$$

◦

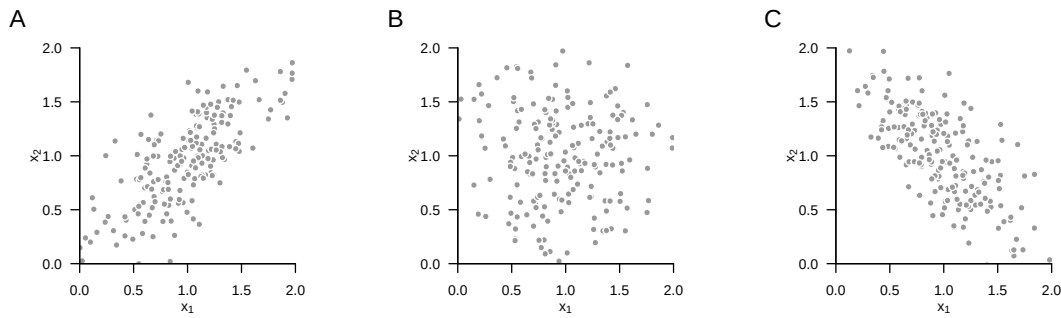


Abbildung 29.4. Realisierungen bivariat normalverteilter Zufallsvektoren.

Wie im Falle der univariat normalverteilten Zufallsvariable entspricht der Parameter $\mu \in \mathbb{R}^n$ dem Wert höchster Wahrscheinlichkeitsdichte der multivariaten Normalverteilung. Analog zum Varianzparameter der univariat normalverteilten Zufallsvariable spezifizieren die Diagonalelemente von $\Sigma \in \mathbb{R}^{n \times n}$ pd die Breite der WDF bezüglich der Zufallsvektorkomponenten $\xi_1, \dots, \xi_n$. Allgemein spezifiziert im Falle des multivariat normalverteilten Zufallsvektors das i, j te Element von $\Sigma \in \mathbb{R}^{n \times n}$ pd hier nun die Kovarianz der Zufallsvektorkomponenten ξ_i und ξ_j .

29.3. Sphärizität

Folgendes Theorem ist für die grundlegende Theorie des Allgemeinen Linearen Modells zentral.

Theorem 29.3 (Sphärische multivariate Normalverteilung). *Für $i = 1, \dots, n$ seien $N(x_i; \mu_i, \sigma^2)$ die WDFen von n unabhängigen univariaten normalverteilten Zufallsvariablen $\xi_1, \dots, \xi_n$ mit $\mu_1, \dots, \mu_n \in \mathbb{R}$ und $\sigma^2 > 0$. Weiterhin sei $N(x; \mu, \sigma^2 I_n)$ die WDF eines n -variaten normalverteilten Zufallsvektors ξ mit Erwartungswertparameter $\mu := (\mu_1, \dots, \mu_n)^T \in \mathbb{R}^n$. Dann gilt*

$$p_\xi(x) = p_{\xi_1, \dots, \xi_n}(x_1, \dots, x_n) = \prod_{i=1}^n p_{\xi_i}(x_i) \quad (29.11)$$

und insbesondere

$$N(x; \mu, \sigma^2 I_n) = \prod_{i=1}^n N(x_i; \mu_i, \sigma^2) \quad (29.12)$$

für alle $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$.

◦

Beweis. Wir zeigen die Identität der multivariaten WDF $N(x; \mu, \sigma^2 I_n)$ mit dem Produkt von n univaria-

ten WDFen $N(x_i; \mu_i, \sigma^2 I_n)$, wobei μ_i der i te Eintrag von $\mu \in \mathbb{R}^n$ ist. Es ergibt sich

$$\begin{aligned}
 N(x; \mu, \sigma^2 I_n) &= (2\pi)^{-\frac{n}{2}} |\sigma^2 I_n|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T (\sigma^2 I_n)^{-1} (x - \mu)\right) \\
 &= \left(\prod_{i=1}^n 2\pi^{-\frac{1}{2}}\right) (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} (x - \mu)^T (x - \mu)\right) \\
 &= \left(\prod_{i=1}^n (2\pi\sigma^2)^{-\frac{1}{2}}\right) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_i)^2\right) \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2} (x_i - \mu_i)^2\right) \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x_i - \mu_i)^2\right) \\
 &= \prod_{i=1}^n N(x_i; \mu_i, \sigma^2).
 \end{aligned} \tag{29.13}$$

□

Einen Kovarianzmatrixparameter der Form $\Sigma = \sigma^2 I_n$ nennt man auch *sphärisch*, da die Isokonturen der WDF eines normalverteilten Zufallsvektors mit einem solchen Kovarianzmatrixparameter *Sphären* bilden (zum Beispiel Kreise bei $n = 2$ und Kugeln bei $n = 3$). Eine multivariate Normalverteilung mit sphärischem Kovarianzmatrixparameter nennt man entsprechend eine *sphärische Normalverteilung*. Theorem 29.3 besagt, dass die WDF eines n -dimensionalen normalverteilten Zufallsvektors mit sphärischem Kovarianzparameter der gemeinsamen WDF von n unabhängigen univariat normalverteilten Zufallsvariablen entspricht und umgekehrt. Eine Realisierung eines n -dimensionalen normalverteilten Zufallsvektors entspricht also den Realisierungen von n unabhängigen univariat normalverteilten Zufallsvariablen und umgekehrt. Man beachte, dass die Identität der Verteilungen der $\xi_i, i = 1, \dots, n$ hier nicht vorausgesetzt ist, insbesondere können sich ihre Erwartungswertparameter $\mu_i, i = 1, \dots, n$ explizit unterscheiden.

29.4. Marginale, bedingte und gemeinsame Verteilungen

Multivariate Normalverteilungen haben die Eigenschaft, dass auch alle anderen assoziierten Verteilungen Normalverteilungen sind und deren Erwartungswert- und Kovarianzmatrixparameter aus den Parametern der jeweils komplementären Verteilung errechnet werden können. Insbesondere gilt zum einen, dass die uni- und multivariaten Marginalverteilungen multivariater Normalverteilungen wiederum Normalverteilungen sind. Zum anderen lassen sich wie alle multivariaten Verteilungen multivariate Normalverteilungen multiplikativ in eine marginale und eine bedingte Verteilung zerlegen. Insbesondere sind nun aber bei multivariaten Normalverteilungen diese Verteilungen wiederum (multivariate) Normalverteilungen, deren Parameter aus den Parametern der gemeinsamen Verteilung errechnet werden können und umgekehrt. Wir fassen obige Erkenntnisse formal in den folgenden drei Theoremen zusammen.

Theorem 29.4 (Marginale Normalverteilungen). *Es sei $n := k + l$ und $\xi = (\xi_1, \dots, \xi_n)^T$ sei ein n -dimensionaler normalverteilter Zufallsvektor mit Erwartungswertparameter*

$$\mu = \begin{pmatrix} \mu_v \\ \mu_\zeta \end{pmatrix} \in \mathbb{R}^n, \tag{29.14}$$

mit $\mu_v \in \mathbb{R}^k$ and $\mu_\zeta \in \mathbb{R}^l$ und Kovarianzmatrixparameter

$$\Sigma = \begin{pmatrix} \Sigma_{vv} & \Sigma_{v\zeta} \\ \Sigma_{\zeta v} & \Sigma_{\zeta\zeta} \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad (29.15)$$

mit $\Sigma_{vv} \in \mathbb{R}^{k \times k}$, $\Sigma_{v\zeta} \in \mathbb{R}^{k \times l}$, $\Sigma_{\zeta v} \in \mathbb{R}^{l \times k}$, und $\Sigma_{\zeta\zeta} \in \mathbb{R}^{l \times l}$. Dann sind $v := (\xi_1, \dots, \xi_k)^T$ und $\zeta := (\xi_{k+1}, \dots, \xi_n)^T$ k - und l -dimensionale normalverteilte Zufallsvektoren, respektive, und es gilt

$$v \sim N(\mu_v, \Sigma_{vv}) \text{ und } \zeta \sim N(\mu_\zeta, \Sigma_{\zeta\zeta}). \quad (29.16)$$

◦

Die Marginalverteilungen einer multivariaten Normalverteilung sind also auch Normalverteilungen und die Parameter der Marginalverteilungen ergeben sich aus den Parametern der gemeinsamen Verteilung. Für Beweise dieses Theorems verweisen wir auf Mardia et al. (1979) und Anderson (2003). Abbildung 29.5 visualisiert Theorem 29.4 für den Fall $n := 2, k := 1, l := 1$,

$$\mu := \begin{pmatrix} 1 \\ 2 \end{pmatrix} \in \mathbb{R}^2 \text{ und } \Sigma := \begin{pmatrix} 0.10 & 0.08 \\ 0.08 & 0.15 \end{pmatrix} \in \mathbb{R}^{2 \times 2}. \quad (29.17)$$

Abbildung 29.5 A zeigt dabei die WDF des bivariaten Zufallsvektors ξ und Abbildung 29.5 B und C die WDFen der entsprechenden marginalen Zufallsvariablen v und ζ .

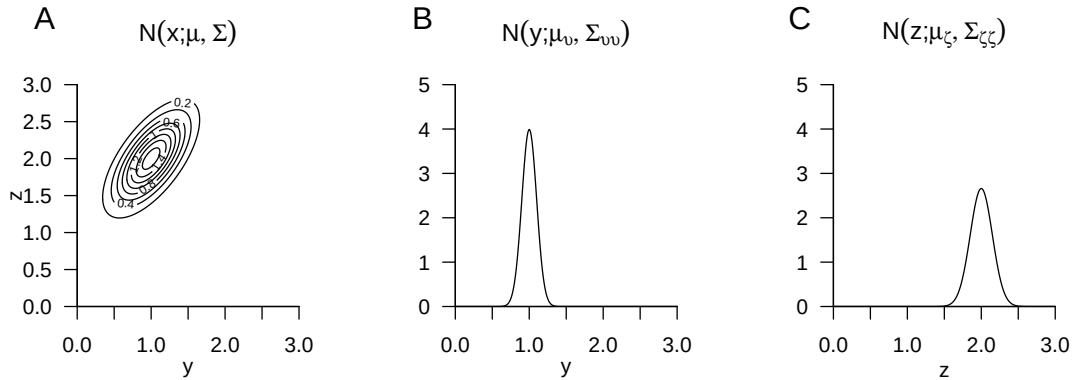


Abbildung 29.5. Marginale Verteilungen eines bivariaten normalverteilten Zufallsvektor.

Mithilfe einer marginalen und einer bedingten multivariaten Normalverteilung lässt sich eine gemeinsame multivariate Normalverteilung konstruieren, deren Parameter sich aus den Parametern der marginalen und bedingten Verteilung ergeben. Dies ist die zentrale Aussage folgenden Theorems.

Theorem 29.5 (Gemeinsame Normalverteilungen). ξ sei ein m -dimensionaler normalverteilter Zufallsvektor mit WDF

$$p_\xi : \mathbb{R}^m \rightarrow \mathbb{R}_{>0}, x \mapsto p_\xi(x) := N(x; \mu_\xi, \Sigma_{\xi\xi}) \text{ mit } \mu_\xi \in \mathbb{R}^m, \Sigma_{\xi\xi} \in \mathbb{R}^{m \times m}, \quad (29.18)$$

$A \in \mathbb{R}^{n \times m}$ sei eine Matrix, $b \in \mathbb{R}^n$ sei ein Vektor und v sei ein n -dimensionaler bedingt normalverteilter Zufallsvektor mit bedingter WDF

$$p_{v|\xi}(\cdot|x) : \mathbb{R}^n \rightarrow \mathbb{R}_{>0}, y \mapsto p_{v|\xi}(y|x) := N(y; A\xi + b, \Sigma_{vv}) \text{ mit } \Sigma_{vv} \in \mathbb{R}^{n \times n}. \quad (29.19)$$

Dann ist der $m + n$ -dimensionale Zufallsvektor $(\xi, v)^T$ normalverteilt mit (gemeinsamer) WDF

$$p_{\xi, v} : \mathbb{R}^{m+n} \rightarrow \mathbb{R}_{>0}, \begin{pmatrix} x \\ y \end{pmatrix} \mapsto p_{\xi, v} \left(\begin{pmatrix} x \\ y \end{pmatrix} \right) = N \left(\begin{pmatrix} x \\ y \end{pmatrix}; \mu_{\xi, v}, \Sigma_{\xi, v} \right), \quad (29.20)$$

mit $\mu_{\xi, v} \in \mathbb{R}^{m+n}$ und $\Sigma_{\xi, v} \in \mathbb{R}^{(m+n) \times (m+n)}$ und insbesondere

$$\mu_{\xi, v} = \begin{pmatrix} \mu_\xi \\ A\mu_\xi + b \end{pmatrix} \text{ und } \Sigma_{\xi, v} = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\xi}A^T \\ A\Sigma_{\xi\xi} & \Sigma_{vv} + A\Sigma_{\xi\xi}A^T \end{pmatrix}. \quad (29.21)$$

◦

Insbesondere ergeben sich die Parameter der gemeinsamen Verteilung also als linear-affine Transformation der Parameter der induzierenden Verteilungen. Abbildung 29.6 visualisiert Theorem 29.5 für den Fall $m := 1, n := 1, \mu_\xi := 1, \Sigma_{\xi\xi} := 0.2, A := 1, b := 1$ und $\Sigma_{vv} := 0.1$. Abbildung 29.6 A zeigt dabei die WDF der Zufallsvariable ξ , Abbildung 29.6 B zeigt die WDF der bedingten Verteilung der Zufallsvariable v gegeben ξ und Abbildung 29.6 C schließlich zeigt die WDFen des induzierten bivariaten Zufallsvektors (ξ, v) .

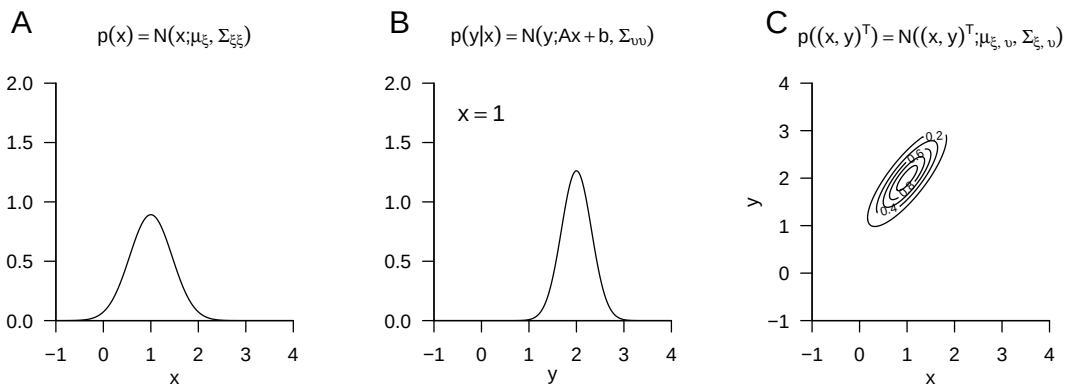


Abbildung 29.6. Gemeinsame Verteilungen einer marginalen und einer auf dieser bedingten normalverteilten Zufallsvariable.

Die Definition einer multivariaten Normalverteilung erlaubt es weiterhin, die bedingten Verteilungen aller Komponenten des entsprechenden Zufallsvektors direkt mithilfe der Parameter der multivariaten Normalverteilung zu bestimmen. Dies ist die zentrale Aussage folgenden Theorems.

Theorem 29.6 (Bedingte Normalverteilungen). (ξ, v) sei ein $m + n$ -dimensionaler normalverteilter Zufallsvektor mit WDF

$$p_{\xi, v} : \mathbb{R}^{m+n} \rightarrow \mathbb{R}_{>0}, \begin{pmatrix} x \\ y \end{pmatrix} \mapsto p_{\xi, v} \left(\begin{pmatrix} x \\ y \end{pmatrix} \right) := N \left(\begin{pmatrix} x \\ y \end{pmatrix}; \mu_{\xi, v}, \Sigma_{\xi, v} \right), \quad (29.22)$$

mit

$$\mu_{\xi, v} = \begin{pmatrix} \mu_{\xi} \\ \mu_v \end{pmatrix}, \Sigma_{\xi, v} = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi v} \\ \Sigma_{v\xi} & \Sigma_{vv} \end{pmatrix}, \quad (29.23)$$

mit $x, \mu_{\xi} \in \mathbb{R}^m, y, \mu_v \in \mathbb{R}^n$ und $\Sigma_{\xi\xi} \in \mathbb{R}^{m \times m}, \Sigma_{\xi v} \in \mathbb{R}^{m \times n}, \Sigma_{vv} \in \mathbb{R}^{n \times n}$. Dann ist die bedingte Verteilung von ξ gegeben v eine m -dimensionale Normalverteilung mit bedingter WDF

$$p_{\xi|v}(\cdot|y) : \mathbb{R}^m \rightarrow \mathbb{R}_{>0}, x \mapsto p_{\xi|v}(x|y) := N(x; \mu_{\xi|v}, \Sigma_{\xi|v}) \quad (29.24)$$

mit

$$\mu_{\xi|v} = \mu_{\xi} + \Sigma_{\xi v} \Sigma_{vv}^{-1} (y - \mu_v) \in \mathbb{R}^m \quad (29.25)$$

und

$$\Sigma_{\xi|v} = \Sigma_{\xi\xi} - \Sigma_{\xi v} \Sigma_{vv}^{-1} \Sigma_{v\xi} \in \mathbb{R}^{m \times m}. \quad (29.26)$$

◦

Im Zusammenspiel mit Theorem 29.5 und Theorem 29.4 können die Parameter bedingter und marginaler Normalverteilungen also aus den Parametern der komplementären bedingten und marginalen Normalverteilungen bestimmt werden. Abbildung 29.7 visualisiert Theorem 29.6 für den Fall $m := 2, n := 1$,

$$\mu := \begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ und } \Sigma := \begin{pmatrix} 0.12 & 0.09 \\ 0.09 & 0.12 \end{pmatrix} \quad (29.27)$$

Dabei zeigt Abbildung 29.7 A die WDF des bivariaten Zufallsvektors $(\xi, v)^T$ und Abbildung 29.7 B und C zeigen die WDF der bedingten Verteilung der Zufallsvariable ξ gegeben $v = 1.5$ und $v = 2.8$, respektive.

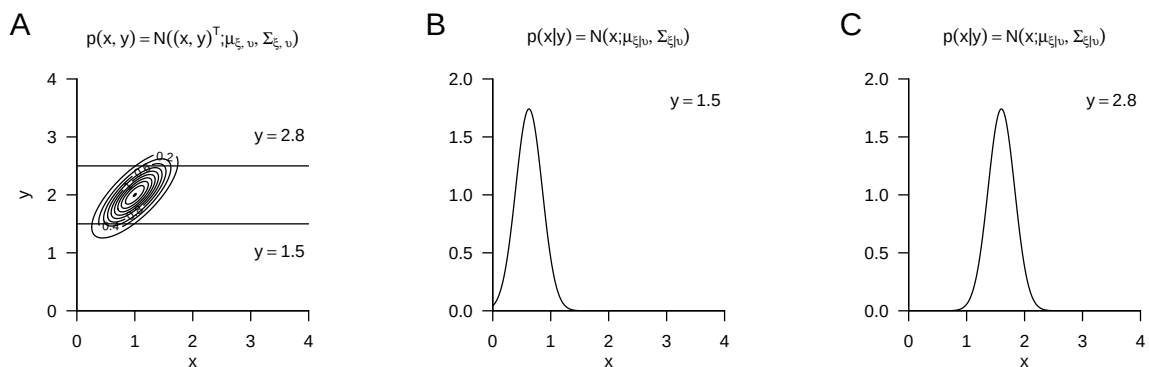


Abbildung 29.7. Bedingte Normalverteilungen

29.5. Multivariate Transformationen

In diesem Abschnitt stellen wir einige Resultate zu den Verteilungen transformierter normalverteilter Zufallsvektoren zusammen. Wir verzichten dabei auf Beweise.

Theorem 29.7 (Invertierbare lineare Transformation eines normalverteilten Zufallsvektors). $\xi \sim N(\mu_\xi, \Sigma_\xi)$ sei ein normalverteilter n -dimensionaler Zufallsvektor und es sei $\zeta := A\xi$ mit einer invertierbaren Matrix $A \in \mathbb{R}^{n \times n}$. Dann gilt

$$\zeta \sim N(\mu_\zeta, \Sigma_\zeta) \text{ mit } \mu_\zeta = A\mu_\xi \text{ und } \Sigma_\zeta = A\Sigma_\xi A^T. \quad (29.28)$$

◦

Nach Theorem 29.7 ergibt die invertierbare lineare Transformation eines multivariat normalverteilten Zufallsvektors also wiederum einen multivariat normalverteilten Zufallsvektor und die Parameter der Verteilung dieses normalverteilten Zufallsvektors ergeben sich aus den Parametern der Verteilung des ursprünglichen Zufallsvektors und der Transformationsmatrix.

Beispiel

Als Beispiel betrachten wir die invertierbare lineare Transformation eines bivariaten normalverteilten Zufallsvektors ξ . Es seien

$$\mu_\xi := \begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ und } \Sigma_\xi := \begin{pmatrix} 0.20 & 0.15 \\ 0.15 & 0.20 \end{pmatrix} \quad (29.29)$$

der Erwartungswert- und Kovarianzmatrixparameter von ξ , respektive, und es sei

$$A := \begin{pmatrix} -2 & 1 \\ -1 & 2 \end{pmatrix} \quad (29.30)$$

die Transformationsmatrix. Da $|A| = -3 \neq 0$ ist A invertierbar und es gilt nach Theorem 29.7, dass

$$\zeta \sim N(\mu_\zeta, \Sigma_\zeta) \text{ mit } \mu_\zeta = A\mu_\xi = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \text{ und } A\Sigma_\xi A^T = \begin{pmatrix} 0.40 & 0.05 \\ 0.05 & 0.40 \end{pmatrix}. \quad (29.31)$$

Abbildung 29.8 A zeigt Isokonturen der WDF von ξ und Realisierungen $x^{(i)} \in \mathbb{R}^2$ von ξ für $i = 1, \dots, 50$. Abbildung 29.8 B zeigt die transformierten Realisierungen $z^{(i)} = Ax^{(i)} \in \mathbb{R}^2$ von ζ sowie die Isokonturen der WDF von ζ nach Theorem 29.7.

Die Tatsache, dass ein linear transformierter normalverteilter Zufallsvektor wiederum normalverteilt ist und dass sich die Parameter der Verteilung des transformierten Zufallsvektors aus den Parametern der Verteilung des ursprünglichen Zufallsvektors sowie den Transformationsparametern bestimmen lassen, bleibt auch im Falle einer nicht notwendigerweise invertierbaren linearen Transformation und auch im Falle einer nicht notwendigerweise invertierbaren linear-affinen Transformation wahr. Dies ist die Aussage folgenden zentralen Theorems. Für einen Beweis verweisen wir auf Anderson (2003).

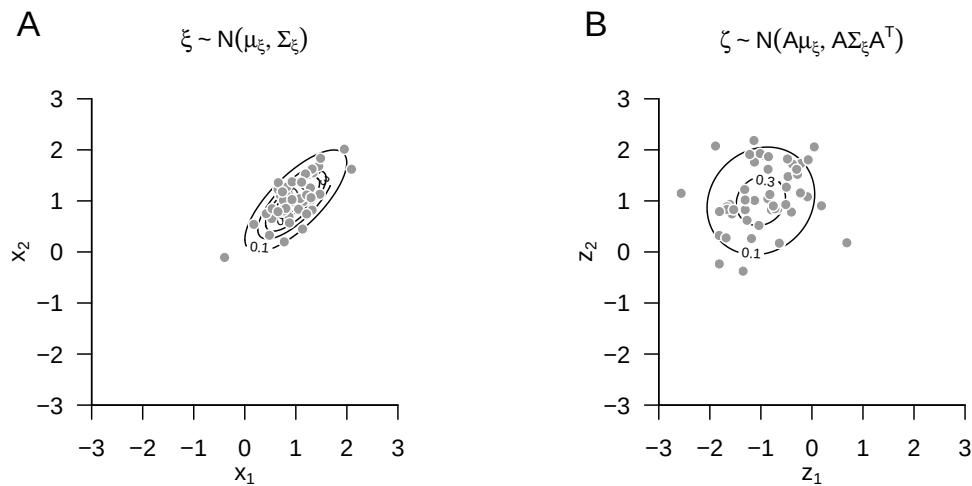


Abbildung 29.8. Invertierbare lineare Transformation eines normalverteilten Zufallsvektors

Theorem 29.8 (Linear-affine Transformation eines normalverteilten Zufallsvektors). $\xi \sim N(\mu_\xi, \Sigma_\xi)$ sei ein normalverteilter n -dimensionaler Zufallsvektor und es sei

$$\zeta := Ax + b \text{ mit } A \in \mathbb{R}^{m \times n} \text{ und } b \in \mathbb{R}^m. \quad (29.32)$$

Dann gilt

$$\zeta \sim N(\mu_\zeta, \Sigma_\zeta) \text{ mit } \mu_\zeta = A\mu + b \in \mathbb{R}^m \text{ und } \Sigma_\zeta = A\Sigma A^T \in \mathbb{R}^{m \times m}. \quad (29.33)$$

◦

29.6. Univariate Transformationen

In diesem Abschnitt diskutieren wir sechs Transformationen normalverteilter Zufallsvariablen, die in der Frequentistischen Inferenz zentrale Rollen spielen und bei denen es sich um Anwendungen von den in **§29.5** eingeführten Transformationstheoremen handelt. Unsere Aussagen sind dabei von der allgemeinen Form “Wenn $\xi_i, i = 1, \dots, n$ unabhängig und identisch normalverteilte Zufallsvariablen sind und $v := f(\xi_1, \dots, \xi_n)$ eine Transformation dieser Zufallsvariablen ist, dann ist die WDF von v durch die Formel $p_v := \{\text{Formel}\}$ gegeben und man nennt die Verteilung von v *Verteilungsname*”. Aussagen dieser Form sind für die Frequentistische Inferenz zentral, weil

- (1) die Zentralen Grenzwertsätze die Annahme additiv unabhängig normalverteilter Störvariablen, und damit normalverteilter Daten, begründet,
- (2) wie wir in **§29.6** sehen werden, es sich bei Schätzern und Statistiken um Transformationen von Zufallsvariablen handelt, und
- (3) Parameterschätzergütekriterien, Konfidenzintervalle und Hypothesentests der Frequentistischen Inferenz durch die Verteilungen der jeweiligen Schätzer und Statistiken charakterisiert und begründet sind.

Summationstransformation

In diesem Abschnitt betrachten wir die resultierende Verteilung bei Summation und Mittelwertbildung von unabhängig und identisch normalverteilten Zufallsvariablen. Speziell besagt das Theorem 29.9, dass die Summe unabhängig normalverteilter Zufallsvariablen wiederum normalverteilt ist und gibt die Parameter dieser Verteilung an, während Theorem 29.10 besagt, dass das Stichprobenmittel unabhängig normalverteilter Zufallsvariablen wiederum normalverteilt ist und gibt die Parameter dieser Verteilung an.

Theorem 29.9 (Summationstransformation). *Für $i = 1, \dots, n$ seien $\xi_i \sim N(\mu_i, \sigma_i^2)$ unabhängige normalverteilte Zufallsvariablen. Dann gilt für die Summe $v := \sum_{i=1}^n \xi_i$, dass*

$$v \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right) \quad (29.34)$$

Für unabhängige und identisch normalverteilte Zufallsvariablen $\xi_i \sim N(\mu, \sigma^2)$ gilt folglich

$$v \sim N(n\mu, n\sigma^2). \quad (29.35)$$

◦

Beweis. Wir skizzieren mithilfe von Theorem 28.6, dass für $\xi_1 \sim N(\mu_1, \sigma_1^2)$, $\xi_2 \sim N(\mu_2, \sigma_2^2)$, und $v := \xi_1 + \xi_2$ gilt, dass $v \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Für $n > 2$ folgt das Theorem dann durch Iteration. Mit der Definition der WDF der Normalverteilung erhalten wir zunächst

$$\begin{aligned} p_v(y) &= \int_{-\infty}^{\infty} p_{\xi_1}(x_1) p_{\xi_2}(y - x_1) dx_1 \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{y - x_1 - \mu_2}{\sigma_2}\right)^2\right) dx_1 \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - \frac{1}{2}\left(\frac{y - x_1 - \mu_2}{\sigma_2}\right)^2\right) dx_1. \end{aligned} \quad (29.36)$$

Mit einigem algebraischen Aufwand erhält man die Identität

$$-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - \frac{1}{2}\left(\frac{y - x_1 - \mu_2}{\sigma_2}\right)^2 = -\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)} - \frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2 \sigma_1^2 - \mu_1 \sigma_2^2)^2}{2\sigma_1^2 \sigma_2^2 (\sigma_1^2 + \sigma_2^2)}, \quad (29.37)$$

so dass weiterhin gilt, dass

$$\begin{aligned} p_v(y) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)} - \frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2 \sigma_1^2 - \mu_1 \sigma_2^2)^2}{2\sigma_1^2 \sigma_2^2 (\sigma_1^2 + \sigma_2^2)}\right) dx_1 \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \exp\left(-\frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2 \sigma_1^2 - \mu_1 \sigma_2^2)^2}{2\sigma_1^2 \sigma_2^2 (\sigma_1^2 + \sigma_2^2)}\right) dx_1 \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2 \sigma_1^2 - \mu_1 \sigma_2^2)^2}{2\sigma_1^2 \sigma_2^2 (\sigma_1^2 + \sigma_2^2)}\right) dx_1. \end{aligned} \quad (29.38)$$

Für das verbleibende Integral zeigt man mithilfe der Integration durch Substitution, dass

$$\int_{-\infty}^{\infty} \exp\left(-\frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2 \sigma_1^2 - \mu_1 \sigma_2^2)^2}{2\sigma_1^2 \sigma_2^2 (\sigma_1^2 + \sigma_2^2)}\right) dx_1 = \frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \quad (29.39)$$

Es ergibt sich also

$$\begin{aligned}
 p_v(y) &= \frac{1}{2\pi\sigma_1\sigma_2} \frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \\
 &= \frac{(2\pi)^{-1}(2\pi)^2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \\
 &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right).
 \end{aligned} \tag{29.40}$$

Schließlich folgt, dass

$$p_v(y) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left(-\frac{1}{2(\sigma_1^2 + \sigma_2^2)} (y - (\mu_1 + \mu_2))^2\right) = N(y; \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \tag{29.41}$$

Ein einfacheres Vorgehen ergibt sich vermutlich nach Fouriertransformation der WDF im Sinne der sogenannten charakteristischen Funktion einer Zufallsvariable. In diesem Fall würde die Faltung der WDFen der Multiplikation der charakteristischen Funktionen entsprechen. $\square$

Ein wichtiger Anwendungsfall von Theorem 29.9 ist das nachfolgende Theorem 29.10 sowie die in Gleichung 27.14 und Gleichung 27.21 erwähnten Generalisierungen der Zentralen Grenzwertsätze. Wir visualisieren Theorem 29.9 exemplarisch in Abbildung 29.9.

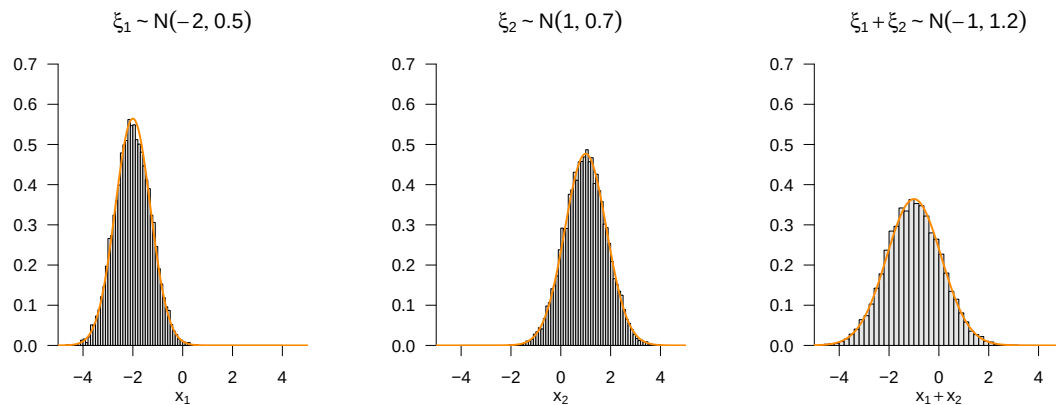


Abbildung 29.9. Summation normalverteilter Zufallsvariablen.

Mittelwertstransformation

Theorem 29.10 (Mittelwertstransformation). Für $i = 1, \dots, n$ seien $\xi_i \sim N(\mu, \sigma^2)$ unabhängig und identisch normalverteilte Zufallsvariablen. Dann gilt für das Stichprobenmittel $\bar{\xi}_n := \frac{1}{n} \sum_{i=1}^n \xi_i$, dass

$$\bar{\xi}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{29.42}$$

◦

Beweis. Wir halten zunächst fest, dass mit dem Theorem zur Summe von unabhängig normalverteilten Zufallsvariablen gilt, dass $\bar{\xi}_n = \frac{1}{n}v$ mit $v := \sum_{i=1}^n \xi_i \sim N(n\mu, n\sigma^2)$. Einsetzen in Theorem 28.3 ergibt

dann

$$\begin{aligned}
 p_{\bar{\xi}_n}(\bar{x}_n) &= \frac{1}{|1/n|} N(n\bar{x}_n; n\mu, n\sigma^2) \\
 &= \frac{n}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{1}{2n\sigma^2} (n\bar{x}_n - n\mu)^2\right) \\
 &= \frac{n}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{1}{2n\sigma^2} (n\bar{x}_n - n\mu)^2\right) \\
 &= nn^{-\frac{1}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(n\bar{x}_n)^2}{2n\sigma^2} + \frac{2(n\bar{x}_n)(n\mu)}{2n\sigma^2} - \frac{(n\mu)^2}{2n\sigma^2}\right) \\
 &= \sqrt{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{n\bar{x}_n^2}{2\sigma^2} + \frac{2n\bar{x}_n\mu}{2\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right) \\
 &= \frac{1}{1/\sqrt{n}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\bar{x}_n^2}{2(\sigma^2/n)} + \frac{2\bar{x}_n\mu}{2(\sigma^2/n)} - \frac{\mu^2}{2(\sigma^2/n)}\right) \\
 &= \frac{1}{\sqrt{2\pi(\sigma^2/n)}} \exp\left(-\frac{1}{2(\sigma^2/n)} (\bar{x}_n - \mu)^2\right) \\
 &= N(\bar{x}_n; \mu, \sigma^2/n)
 \end{aligned} \tag{29.43}$$

□

Wichtige Anwendungsfälle von Theorem 29.10 sind die Analyse von Erwartungswertschätzern in Kapitel 31 sowie die in Gleichung 27.14 erwähnte Generalisierung des Zentralen Grenzwertsatzes nach Lindenberg-Lévy. Wir visualisieren Theorem 29.10 exemplarisch in Abbildung 29.10.

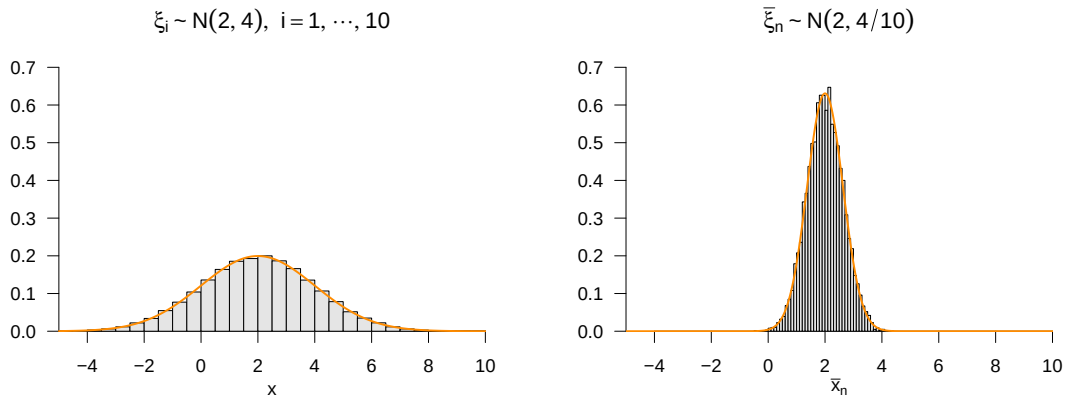


Abbildung 29.10. Mittelwertbildung bei normalverteilten Zufallsvariablen.

Z-Transformation

Das Theorem 29.11 besagt, dass Subtraktion des Erwartungswertparameters und gleichzeitige Division mit der Wurzel des Varianzparameters die Verteilung einer normalverteilten Zufallsvariable in eine Standardnormalverteilung transformiert.

Theorem 29.11 (Z-Transformation). *Es sei $v \sim N(\mu, \sigma^2)$ eine normalverteilte Zufallsvariable. Dann ist die Zufallsvariable*

$$Z := \frac{v - \mu}{\sigma} \tag{29.44}$$

eine standardnormalverteilte Zufallsvariable, es gilt also $Z \sim N(0, 1)$.

◦

Beweis. Wir nutzen Theorem 28.3. Dazu halten wir zunächst fest, dass die Z -Transformation einer Funktion der Form

$$f(v) := \frac{v - \mu}{\sigma} =: Z \quad (29.45)$$

entspricht. Wir stellen weiterhin fest, dass die Umkehrfunktion von f durch

$$f^{-1}(Z) := \sigma Z + \mu \quad (29.46)$$

gegeben ist, da für alle $z \in \mathbb{R}$ mit $z = y - \mu\sigma$ gilt, dass

$$\zeta^{-1}(z) = \zeta^{-1}\left(\frac{y - \mu}{\sigma}\right) = \frac{\sigma(y - \mu)}{\sigma} + \mu = y - \mu + \mu = y. \quad (29.47)$$

Schließlich stellen wir fest, dass für die Ableitung f' von f gilt, dass

$$f'(y) = \frac{d}{dy}\left(\frac{y - \mu}{\sigma}\right) = \frac{d}{dy}\left(\frac{y}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}. \quad (29.48)$$

Einsetzen in das univariate WDF Transformationstheorem für lineare Funktionen ergibt dann

$$\begin{aligned} p_Z(z) &= \frac{1}{|1/\sigma|} N(\sigma z + \mu; \mu, \sigma^2) \\ &= \frac{1}{1/\sqrt{\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\sigma z + \mu - \mu)^2\right) \\ &= \frac{\sqrt{\sigma^2}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\sigma^2 z^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \\ &= N(z; 0, 1) \end{aligned} \quad (29.49)$$

also, dass $Z \sim N(0, 1)$. □

Wichtige Anwendungsfälle von Theorem 29.11 sind neben der häufig angewandten Standardisierung von normalverteilten Zufallsvariablen im Sinne der sogenannten Z -Werte (Z -Scores) die Z -Konfidenzintervallstatistik und die Z -Teststatistik. Wir visualisieren Theorem 29.11 exemplarisch in Abbildung 29.11.

χ^2 -Transformation

Mit der χ^2 -Transformation führen wir nun eine erste Transformation unabhängig und identisch normalverteilter Zufallsvariablen ein, die *nicht* wiederum auf eine Normalverteilung führt. Speziell besagt Theorem 29.12, dass die Summe quadrierter unabhängiger standardnormalverteilter Zufallsvariablen eine χ^2 -verteilte Zufallsvariable ist. Dazu erinnern wir zunächst an den Begriff der χ^2 -Zufallsvariable als Spezialfall der in Kapitel 21 betrachteten Gammazufallsvariablen (vgl. Definition 21.10).

Definition 29.3 (χ^2 -Zufallsvariable). U sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}_{>0}$ und WDF

$$p : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, u \mapsto p(u) := \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} u^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}u\right), \quad (29.50)$$

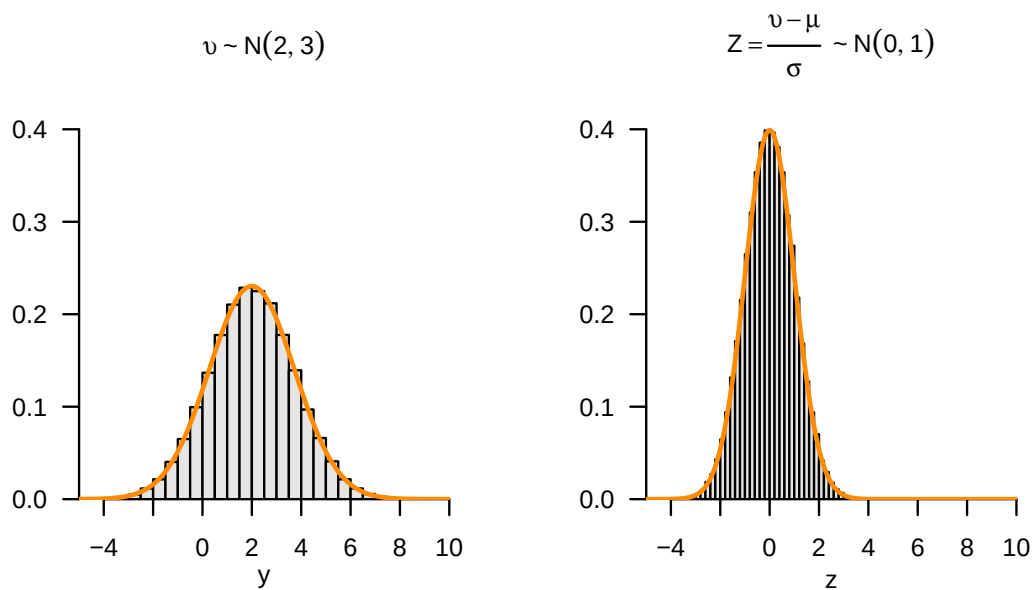


Abbildung 29.11. Z -Transformation normalverteilter Zufallsvariablen.

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass U einer χ^2 -Verteilung mit Freiheitsgradparameter n unterliegt und nennen U eine χ^2 -Zufallsvariable mit Freiheitsgradparameter n . Wir kürzen dies mit $U \sim \chi^2(n)$ ab. Die WDF einer χ^2 -Zufallsvariable bezeichnen wir mit

$$\chi^2(u; n) := \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} u^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}u\right). \quad (29.51)$$

•

Wir erinnern daran, dass die WDF der χ^2 -Verteilung der WDF $G(u; \frac{n}{2}, 2)$ einer Gammaverteilung entspricht. In Abbildung 29.12 visualisieren wir exemplarisch einige WDFen von χ^2 -Zufallsvariablen. Wir beobachten, dass mit ansteigendem n sich $\chi^2(u; n)$ verbreitert und Wahrscheinlichkeitsmasse zur größeren Werten von u verschoben wird.

Theorem 29.12 (χ^2 -Transformation). $Z_1, \dots, Z_n \sim N(0, 1)$ seien unabhängig und identisch standardnormalverteilte Zufallsvariablen. Dann ist die Zufallsvariable

$$U := \sum_{i=1}^n Z_i^2 \quad (29.52)$$

eine χ^2 -verteilte Zufallsvariable mit Freiheitsgradparameter n , es gilt also $U \sim \chi^2(n)$. Insbesondere gilt für $Z \sim N(0, 1)$ und $U := Z^2$, dass $U \sim \chi^2(1)$.

◦

Beweis. Wir zeigen das Theorem nur für den Fall $n := 1$ mithilfe von Theorem 28.4. Danach ist die WDF einer Zufallsvariable $U := f(Z)$, welche aus der Transformation einer Zufallsvariable Z mit WDF p_Z durch eine stückweise bijektive Abbildung hervorgeht, gegeben durch

$$p_U(u) = \sum_{i=1}^k 1_{u_i} \frac{1}{|f'_i(f_i^{-1}(u))|} p_Z(f_i^{-1}(u)). \quad (29.53)$$

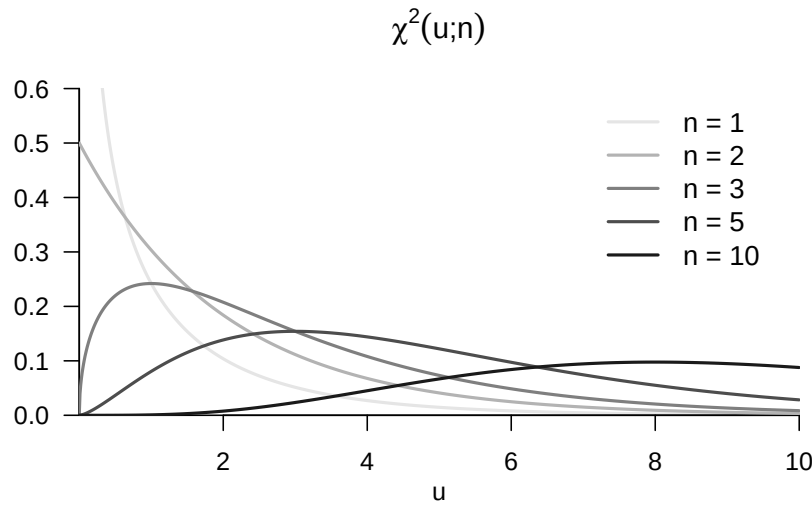


Abbildung 29.12. WDFen von χ^2 Zufallsvariablen.

Wir definieren

$$\mathcal{U}_1 :=]-\infty, 0[, \mathcal{U}_2 :=]0, \infty[, \text{ und } \mathcal{U}_i := \mathbb{R}_{>0} \text{ für } i = 1, 2, \quad (29.54)$$

sowie

$$f_i : \mathcal{Z}_i \rightarrow \mathcal{U}_i, x \mapsto f_i(z) := z^2 =: u \text{ für } i = 1, 2. \quad (29.55)$$

Die Ableitung und die Umkehrfunktion der f_i ergeben sich zu

$$f'_i : \mathcal{Z}_i \rightarrow \mathcal{Z}_i, x \mapsto f'_i(z) = 2z \text{ für } i = 1, 2, \quad (29.56)$$

und

$$f_1^{-1} : \mathcal{U}_1 \rightarrow \mathcal{U}_1, u \mapsto f_1^{-1}(u) = -\sqrt{u} \text{ und } f_2^{-1} : \mathcal{U}_2 \rightarrow \mathcal{U}_2, u \mapsto f_2^{-1}(u) = \sqrt{u}, \quad (29.57)$$

respektive. Einsetzen in Gleichung (29.53) ergibt dann

$$\begin{aligned} p_U(u) &= 1_{\mathcal{U}_1}(u) \frac{1}{|f'_1(f_1^{-1}(u))|} p_\zeta(f_1^{-1}(u)) + 1_{\mathcal{U}_2}(u) \frac{1}{|f'_2(f_2^{-1}(u))|} p_\zeta(f_2^{-1}(u)) \\ &= \frac{1}{|2(-\sqrt{u})|} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(-\sqrt{u})^2\right) + \frac{1}{|2(\sqrt{u})|} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{u})^2\right) \\ &= \frac{1}{2\sqrt{u}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u\right) + \frac{1}{2\sqrt{u}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u\right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{u}} \exp\left(-\frac{1}{2}u\right). \end{aligned} \quad (29.58)$$

Andererseits gilt, dass mit $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, die PDF einer χ^2 -Zufallsvariable U mit $n = 1$ durch

$$\frac{1}{\Gamma(\frac{1}{2})} u^{\frac{1}{2}-1} \exp\left(-\frac{1}{2}u\right) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{u}} \exp\left(-\frac{1}{2}u\right) \quad (29.59)$$

gegeben ist. Also gilt, dass wenn $Z \sim N(0, 1)$ ist, dann ist $U := Z^2 \sim \chi^2(1)$. $\square$

Wichtige Anwendungsfälle sind die U -Konfidenzintervallstatistik sowie die im folgenden eingeführten t - und f -Zufallsvariablen. Wir visualisieren Theorem 29.12 exemplarisch in Abbildung 29.13.

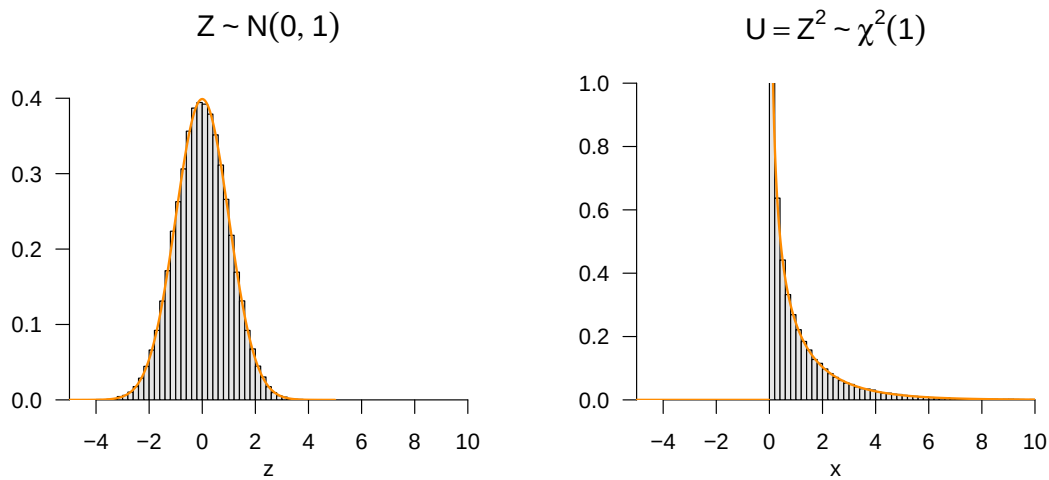


Abbildung 29.13. χ^2 -Transformation normalverteilter Zufallsvariablen.

T -Transformation

Das in diesem Abschnitt betrachtete Theorem geht auf Student (1908) zurück und ist das zentrale und stilprägende Resultat der Entwicklung der Frequentistischen Inferenz in der ersten Hälfte der 20. Jahrhunderts. Hald (2007) und Zabell (2008) und geben hierzu einen historischen Überblick. Das zentrale Theorem 29.13 besagt dabei, dass die Zufallsvariable, die sich durch Division einer standardnormalverteilten Zufallsvariable durch die Quadratwurzel einer χ^2 -verteilten Zufallsvariable geteilt durch ein n , ergibt, eine t -verteilte Zufallsvariable ist. Dabei ist eine t -verteilte Zufallsvariable wie folgt definiert.

Definition 29.4 (t -Zufallsvariable). T sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, t \mapsto p(t) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad (29.60)$$

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass T einer t -Verteilung mit Freiheitsgradparameter n unterliegt und nennen T eine t -Zufallsvariable mit Freiheitsgradparameter n . Wir kürzen dies mit $T \sim t(n)$ ab. Die WDF einer t -Zufallsvariable bezeichnen wir mit

$$T(t; n) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}. \quad (29.61)$$

•

In Abbildung 29.14 visualisieren wir exemplarisch einige WDFen von t -Zufallsvariablen. Wir beobachten, dass die t -Verteilung immer um 0 symmetrisch ist und ein steigendes n Wahrscheinlichkeitsmasse aus den Ausläufen zum Zentrum verschiebt. Wir merken an, dass ab etwa $n = 30$ gilt, dass $T(t; n) \approx N(0, 1)$.

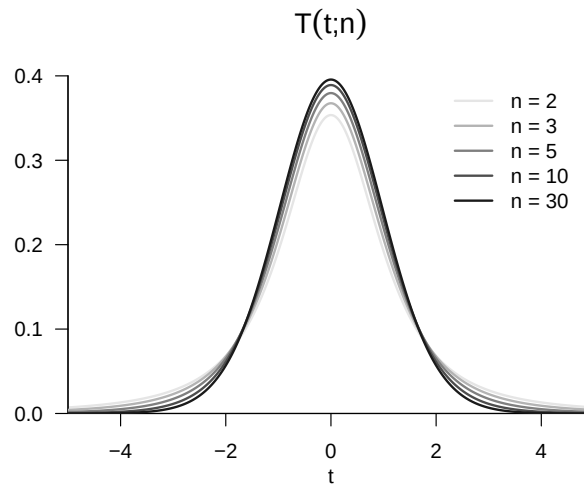


Abbildung 29.14. WDFen von T -Zufallsvariablen.

Theorem 29.13 (T -Transformation). $Z \sim N(0, 1)$ sei eine standardnormalverteilte Zufallsvariable, $U \sim \chi^2(n)$ sei eine χ^2 -Zufallsvariable mit Freiheitsgradparameter n , und Z und U seien unabhängig. Dann ist die Zufallsvariable

$$T := \frac{Z}{\sqrt{U/n}} \quad (29.62)$$

eine t -verteilte Zufallsvariable mit Freiheitsgradparameter n , es gilt also $T \sim t(n)$.

◦

Beweis. Wir halten zunächst fest, dass die zweidimensionale WDF der gemeinsamen (unabhängigen) Verteilung von Z und U durch

$$p_{Z,U}(z, u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} u^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}u\right). \quad (29.63)$$

gegeben ist. Wir betrachten dann die multivariate vektorwertige Abbildung

$$f: \mathbb{R}^2 \rightarrow \mathbb{R}^2, (z, u) \mapsto f(z, u) := \left(\frac{z}{\sqrt{u/n}}, u \right) =: (t, w) \quad (29.64)$$

und benutzen das multivariate WDF Transformationstheorem für bijektive Abbildungen um die WDF von (t, w) herzuleiten. Dazu erinnern wir uns, dass wenn ξ ein n -dimensionaler Zufallsvektor mit WDF p_ξ und $v := f(\xi)$ für eine differenzierbare und bijektive Abbildung $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ ist, die WDF des Zufallsvektors v durch

$$p_v: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_v(y) := \frac{1}{|Jf(f^{-1}(y))|} p_\xi(f^{-1}(y)) \quad (29.65)$$

gegeben ist. Für die im vorliegenden Fall betrachtete Abbildung halten wir zunächst fest, dass

$$f^{-1}: \mathbb{R}^2 \rightarrow \mathbb{R}^2, (t, w) \mapsto f^{-1}(t, w) := (\sqrt{w/nt}, w). \quad (29.66)$$

Dies ergibt sich direkt aus

$$f^{-1}(f(z, u)) = f^{-1}\left(\frac{z}{\sqrt{u/n}}, u\right) = \left(\frac{\sqrt{u/n}z}{\sqrt{u/n}}, u\right) = (z, u) \text{ für alle } (z, u) \in \mathbb{R}^2. \quad (29.67)$$

Wir halten dann fest, dass die Determinante der Jacobi-Matrix von f an der Stelle (z, u) durch

$$|Jf(z, u)| = \begin{vmatrix} \frac{\partial}{\partial z} \left(\frac{z}{\sqrt{u/n}} \right) & \frac{\partial}{\partial u} \left(\frac{z}{\sqrt{u/n}} \right) \\ \frac{\partial}{\partial z} u & \frac{\partial}{\partial u} u \end{vmatrix} = \left(\frac{v}{n} \right)^{-1/2}, \quad (29.68)$$

gegeben ist, sodass folgt, dass

$$\frac{1}{|Jf(f^{-1}(z, u))|} = \left(\frac{w}{n} \right)^{1/2}. \quad (29.69)$$

Einsetzen in Gleichung (29.65) ergibt dann

$$p_{T,W}(t, w) = \left(\frac{w}{n} \right)^{1/2} p_{Z,V}(\sqrt{w/nt}, w), \quad (29.70)$$

Es folgt also

$$\begin{aligned} p_T(t) &= \int_0^\infty p_{T,W}(t, w) dw \\ &= \int_0^\infty \left(\frac{w}{n} \right)^{1/2} p_{Z,V}(\sqrt{w/nt}, w) dw \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{w/nt})^2\right) \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} w^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}w\right) \left(\frac{w}{n}\right)^{1/2} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\frac{w}{n}t^2\right) w^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}w\right) w^{1/2} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\frac{w}{n}t^2 - \frac{1}{2}w\right) w^{\frac{n}{2}-1} w^{1/2} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{w}{n}t^2 + w\right)\right) w^{\frac{n+1}{2}-1} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\left(1 + \frac{t^2}{n}\right)w\right) w^{\frac{n+1}{2}-1} dw \end{aligned} \quad (29.71)$$

Wir stellen dann fest, dass der Integrand auf der linken Seite der obigen Gleichung dem Kern einer Gamma WDF mit Parametern $\alpha = \frac{n+1}{2}$ und $\beta = \frac{2}{1+\frac{t^2}{n}}$ entspricht, wie man leicht einsieht:

$$\begin{aligned} \Gamma(w; \alpha, \beta) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} w^{\alpha-1} \exp\left(-\frac{w}{\beta}\right) \\ \Rightarrow \Gamma\left(w; \frac{n+1}{2}, \frac{2}{1+\frac{t^2}{n}}\right) &= \frac{1}{\Gamma(\frac{n+1}{2})\left(\frac{2}{1+\frac{t^2}{n}}\right)^{\frac{n+1}{2}}} w^{\frac{n+1}{2}-1} \exp\left(-\frac{w}{\frac{2}{1+\frac{t^2}{n}}}\right) \\ &= \frac{1}{\Gamma(\frac{n+1}{2})\left(\frac{2}{1+\frac{t^2}{n}}\right)^{\frac{n+1}{2}}} \exp\left(-\frac{1}{2}\left(1 + \frac{t^2}{n}\right)w\right) w^{\frac{n+1}{2}-1}. \end{aligned}$$

Es ergibt sich also

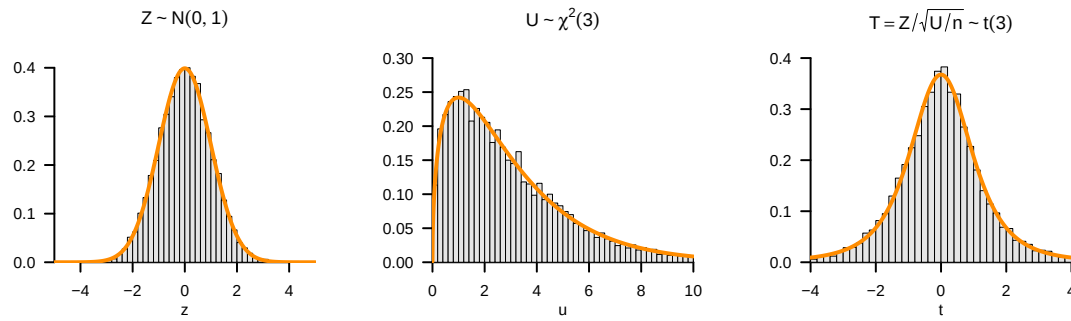
$$p_T(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \Gamma\left(w; \frac{n+1}{2}, \frac{2}{1+\frac{t^2}{n}}\right) dw. \quad (29.72)$$

Schließlich stellen wir fest, dass der Integralterm in obiger Gleichung dem Normalisierungsterm einer Gamma WDF entspricht. Abschließend ergibt sich also

$$p_T(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \Gamma\left(\frac{n+1}{2}\right) \left(\frac{2}{1+\frac{t^2}{n}}\right)^{\frac{n+1}{2}}. \quad (29.73)$$

Die Verteilung von $Z/\sqrt{U/n}$ hat also die WDF einer t -Zufallsvariable. $\square$

Wichtige Anwendungsfälle sind die T -Konfidenzintervallstatistik sowie die T -Teststatistiken der Theorie von Hypothesentests im Kontext des Allgemeinen Linearen Modells. Wir visualisieren Theorem 29.13 exemplarisch in Abbildung 29.15.

Abbildung 29.15. T -Transformation normalverteilter Zufallsvariablen.

Nichtzentrale T -Transformation

In diesem Abschnitt betrachten wir den Fall, dass der Erwartungswertparameter der Zählervariable der in Theorem 29.13 betrachteten Zufallsvariable T von Null verschieden ist, dass es sich bei der Zählervariable also nicht um eine nach $N(0, 1)$, sondern eine nach $N(\mu, 1)$ verteilte Zufallsvariable für ein beliebiges $\mu \in \mathbb{R}$ handelt. Die so entstehende Zufallsvariable T folgt dann einer sogenannten *nichtzentralen- t -Verteilung*. Eine frühe ausführliche Diskussion dieser Verteilung findet sich zum Beispiel in Johnson & Welch (1940). Eine entsprechende nichtzentralen- t -verteilte Zufallsvariable ist wie folgt definiert (vgl. Lehmann (1986)).

Definition 29.5 (Nichtzentrale t -Zufallsvariable). T sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, t \mapsto p(t) := \frac{1}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n}{2}\right) (n\pi)^{\frac{1}{2}}} \times \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left(-\frac{\tau}{2}\right) \exp\left(-\frac{1}{2} \left(t \left(\frac{\tau}{n}\right)^{\frac{1}{2}} - \delta\right)^2\right) d\tau. \quad (29.74)$$

Dann sagen wir, dass T einer nichtzentralen t -Verteilung mit Nichtzentralitätsparameter δ und Freiheitsgradparameter n unterliegt und nennen T eine *nichtzentrale t -Zufallsvariable mit Nichtzentralitätsparameter δ und Freiheitsgradparameter n* . Wir kürzen dies mit $t(\delta, n)$ ab. Die WDF einer nichtzentralen t -Zufallsvariable bezeichnen wir mit $t(T; \delta, n)$. Die KVF und inverse KVF einer nichtzentralen t -Zufallsvariable bezeichnen wir mit $\Psi(\cdot; \delta, n)$ und $\Psi^{-1}(\cdot; \delta, n)$, respektive.

•

Ohne Beweis merken wir an, dass eine nichtzentrale t -Zufallsvariable mit $\delta = 0$ einer t -Zufallsvariable entspricht, es gelten also

$$t(T; 0, n) = t(T; n) \quad (29.75)$$

sowie

$$\Psi(T; 0, n) = \Psi(T; n) \text{ und } \Psi^{-1}(T; 0, n) = \Psi^{-1}(T; n). \quad (29.76)$$

In Abbildung 29.16 visualisieren wir exemplarisch einige WDFen von nichtzentralen t -Zufallsvariablen. Wir beobachten, dass ein positiver Nichtzentralitätsparameter δ die Verteilung nach rechts verschiebt und die Verteilungen mit steigendem Freiheitsgradparameter δ sich entsprechend lokalisierten Normalverteilungen mit Varianzparameter 1 annähern. Man beachte auch die Nichtsymmetrie der WDFen für kleine Freiheitsgradparameter bei von Null verschiedenem positivem Nichtzentralitätsparameter.

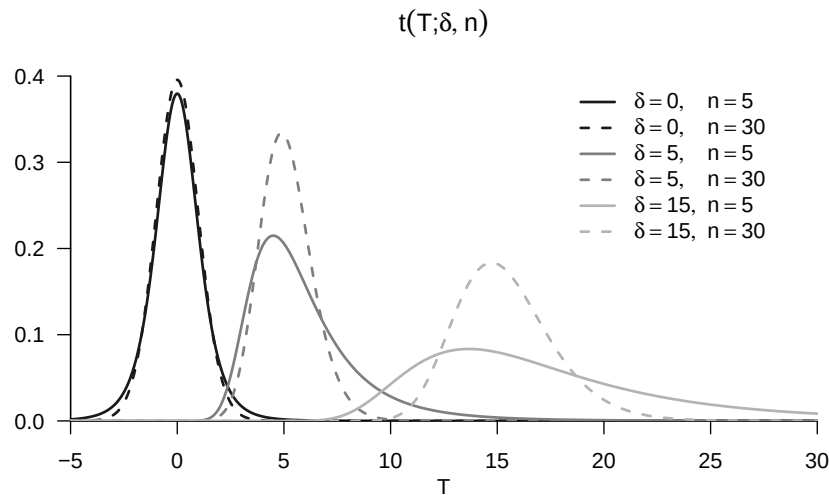


Abbildung 29.16. WDFen von Nichtzentralen- T -Zufallsvariablen.

Eine nichtzentrale t -Zufallsvariable ist das Resultat einer nichtzentralen T -Transformation, wie folgendes Theorem besagt.

Theorem 29.14 (Nichtzentrale T -Transformation). $v \sim N(\mu, 1)$ sei eine normalverteilte Zufallsvariable, $U \sim \chi^2(n)$ sei eine χ^2 Zufallsvariable mit Freiheitsgradparameter n , und v und U seien unabhängige Zufallsvariablen. Dann ist die Zufallsvariable

$$T := \frac{v}{\sqrt{U/n}} \quad (29.77)$$

eine nichtzentrale t -Zufallsvariable mit Nichtzentralitätsparameter μ und Freiheitsgradparameter n , also $T \sim t(\mu, n)$.

◦

Wir verzichten auf einen Beweis. Wichtige Anwendungsfälle sind die Testgütefunktionen der T -Test Varianten im Kontext des Allgemeinen Linearen Modells. Wir visualisieren Theorem 29.14 exemplarisch in Abbildung 29.17.

F -Transformation

Das in diesem Abschnitt zentrale Theorem 29.15 besagt, dass die Zufallsvariable, die sich durch Division zweier χ^2 verteilter Zufallsvariablen, jeweils geteilt durch ihre jeweiligen Freiheitsgradparameter, eine F -verteilte Zufallsvariable ist. Dabei ist eine F -verteilte Zufallsvariable wie folgt definiert.

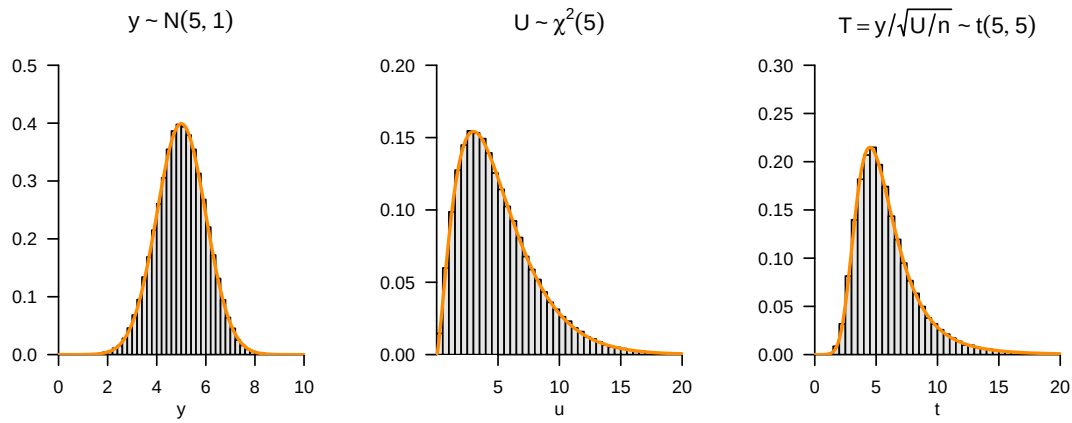


Abbildung 29.17. Nichtzentrale T -Transformation normalverteilter Zufallsvariablen.

Definition 29.6 (f -Zufallsvariable). F sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}_{>0}$ und WDF

$$p_F : \mathbb{R} \rightarrow \mathbb{R}_{>0}, f \mapsto p_F(f) := m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} \frac{f^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n} f\right)^{\frac{m+n}{2}}}, \quad (29.78)$$

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass F einer f -Verteilung mit Freiheitsgradparametern n, m unterliegt und nennen F eine f -Zufallsvariable mit Freiheitsgradparametern n, m . Wir kürzen dies mit $F \sim f(n, m)$ ab. Die WDF einer f -Zufallsvariable bezeichnen wir mit

$$F(f; n, m) := m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} \frac{f^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n} f\right)^{\frac{m+n}{2}}}. \quad (29.79)$$

•

In Abbildung 29.18 visualisieren wir exemplarisch einige WDFen von f -Zufallsvariablen. Wir beobachten, dass die Form der WDFen zunächst primär durch den Freiheitsgradparameter n und dann sekundär durch den Freiheitsgradparameter m bestimmt werden.

Theorem 29.15 (F -Transformation). $V \sim \chi^2(n)$ und $W \sim \chi^2(m)$ seien zwei unabhängige χ^2 -Zufallsvariablen mit Freiheitsgradparametern n und m , respektive. Dann ist die Zufallsvariable

$$F := \frac{V/n}{W/m} \quad (29.80)$$

eine f -verteilte Zufallsvariable mit Freiheitsgradparametern n, m , es gilt also $F \sim f(n, m)$.

◦

Das Theorem kann bewiesen werden, in dem man zunächst ein Transformationstheorem für Quotienten von Zufallsvariablen mithilfe von Theorem 28.1 und Marginalisierung herleitet und dieses Theorem dann auf die WDF von χ^2 -verteilten Zufallsvariablen anwendet. Wir visualisieren Theorem 29.15 exemplarisch in Abbildung 29.19. Wichtige Anwendungsfälle von Theorem 29.15 sind die im Rahmen der Theorie des Allgemeinen Linearen Modells betrachteten F -Statistiken.

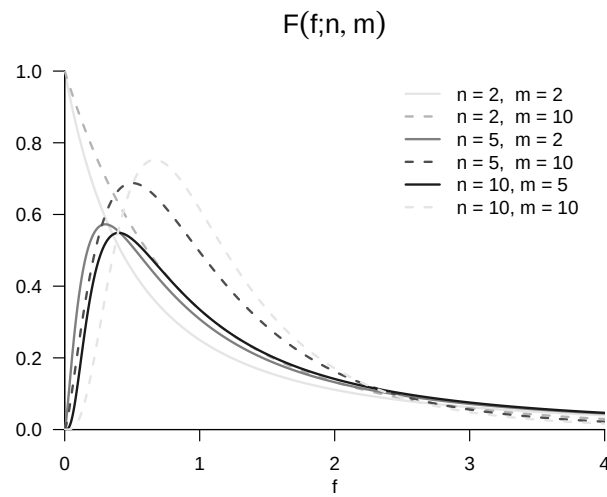


Abbildung 29.18. WDFen von f -verteilten Zufallsvariablen.

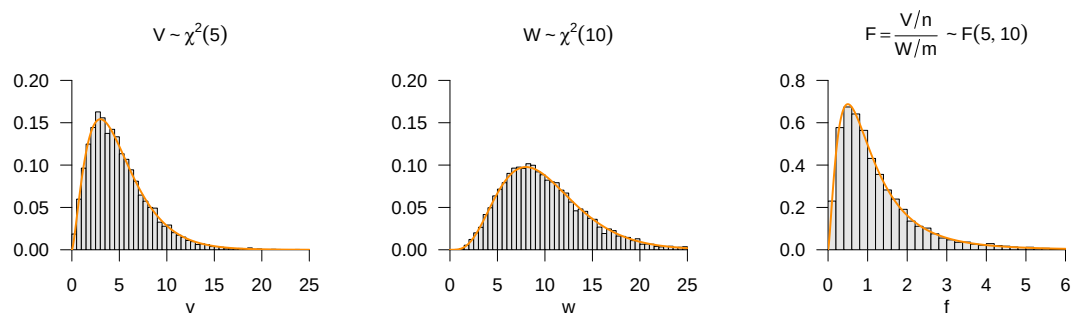


Abbildung 29.19. F -Transformation normalverteilter Zufallsvariablen.

29.7. Literaturhinweise

Die Entwicklung der bivariaten Normalverteilung hat ihre Ursprünge in der statistischen Literatur zur Mitte des 19. Jahrhunderts, insbesondere in den Arbeiten von Francis Galton (1822-1911). Die mathematische Formalisierung der bivariaten Normalverteilung geht dabei wohl insbesondere auf Pearson (1896) zurück (Seal (1967)). Die ursprüngliche Formulierung der multivariaten Normalverteilung wird bei Edgeworth (1892) verortet. Tong (1990) gibt einen umfassenden Überblick zur Theorie und Anwendung der multivariaten Normalverteilung.

Teil IV.

Inferenz

30. Frequentistische Inferenz

30.1. Frequentistische Inferenzmodelle

Mit folgender Definition wollen wir zunächst einige grundlegende Begrifflichkeiten bei der Betrachtung Frequentistischer Inferenzmodelle einführen.

Definition 30.1 (Frequentistische Inferenzmodelle). Ein *Frequentistisches Inferenzmodell* ist ein Tupel

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \quad (30.1)$$

bestehend aus einem *Datenraum* $\mathcal{Y}$, einer σ -Algebra $\mathcal{A}$ auf $\mathcal{Y}$ und einer mindestens zweielementigen Menge $\{\mathbb{P}_\theta | \theta \in \Theta\}$ von Wahrscheinlichkeitsmaßen auf $(\mathcal{Y}, \mathcal{A})$, die durch $\theta \in \Theta$ indiziert sind. Wenn $\Theta \subset \mathbb{R}^k$ ist, heißt ein Frequentistisches Inferenzmodell auch *parametrisches Frequentistisches Inferenzmodell* und Θ heißt *Parameterraum* des Frequentistischen Inferenzmodells. Ein Frequentistisches Inferenzmodell $\mathcal{M}$ heißt ein *diskretes Modell*, wenn $\mathcal{Y}$ endlich oder abzählbar ist und jedes $\mathbb{P}_\theta$ eine WMF p_θ besitzt. Ein Frequentistisches Inferenzmodell $\mathcal{M}$ heißt ein *stetiges Modell*, wenn $\mathcal{Y} \subset \mathbb{R}^n$ ist und jedes $\mathbb{P}_\theta$ eine WDF p_θ besitzt. Wenn der Datenraum $\mathcal{Y}$ eines Frequentistischen Inferenzmodells $\mathcal{M}$ eindimensional ist, also zum Beispiel $\mathcal{Y} := \mathbb{R}$, spricht man von einem *univariaten Frequentistischen Inferenzmodell*. Wenn der Datenraum $\mathcal{Y}$ eines Frequentistischen Inferenzmodells $\mathcal{M}$ mehrdimensional ist, also zum Beispiel $\mathcal{Y} := \mathbb{R}^m$ für $m > 1$, spricht man von einem *multivariaten Frequentistischen Inferenzmodell*. Für ein Frequentistisches Inferenzmodell $\mathcal{M}_0 := (\mathcal{Y}_0, \mathcal{A}_0, \{\mathbb{P}_\theta^0 | \theta \in \Theta\})$ wird das Frequentistische Inferenzmodell $\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\})$, für das $\mathcal{Y}$ das n -fache kartesische Produkt von $\mathcal{Y}_0$ mit sich selbst, $\mathcal{A}$ die entsprechende Produkt- σ -Algebra und $\{\mathbb{P}_\theta | \theta \in \Theta\}$ die entsprechende Menge an Produktmaßen ist, ein (zu $\mathcal{M}_0$ gehöriges) *Frequentistisches Produktmodell* genannt.

•

Vor dem Hintergrund eines Frequentistischen Inferenzmodells wird der Vorgang der Datenbeobachtung durch einen Zufallsvektor y , der Werte in $\mathcal{Y}$ annimmt und dessen Verteilung einer der prinzipiell möglichen Verteilungen $\mathbb{P}_\theta$ entspricht, beschrieben. Man nennt diesen Zufallsvektor *Daten*, *Beobachtung*, *Messung* oder *Stichprobe*. Im Gegensatz zum Wahrscheinlichkeitsraummodell betrachtet man bei Frequentistischen Inferenzmodellen also explizit zwei oder mehr Wahrscheinlichkeitsmaße, die die Verteilung von y mutmaßlich bestimmen. Eine Realisierung von y , also konkret vorliegende Datenwerte $y \in \mathcal{Y}$, nennt man *Datensatz*, *Beobachtungswert*, *Messwert* oder *Stichprobenwert*. Erwartungswerte und (Ko)Varianzen von y bezüglich $\mathbb{P}_\theta$ schreibt man meist als $\mathbb{E}_\theta(y)$, $\mathbb{V}_\theta(y)$ und $\mathbb{C}_\theta(y)$. Frequentistische Produktmodelle modellieren die n -fache unabhängige Wiederholung eines Zufallsvorgangs. Die entsprechende Menge von Zufallsvektoren $y_1, \dots, y_n$ entspricht dann einer Menge von n unabhängigen Zufallsvektoren.

In einem konkreten Datenanalyseproblem auf Grundlage eines parameterischen Frequentistischen Produktmodells nimmt man an, dass die beobachteten Werte $y_1, \dots, y_n$ von $y_1, \dots, y_n$

durch genau ein Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ mit Parameter $\theta \in \Theta$ generiert wurde. In der Anwendung wird dieses $\theta \in \Theta$ dann als *wahrer, aber unbekannter, Parameterwert* bezeichnet. Der wahre, aber unbekannte Parameterwert θ bleibt dabei auch nach jeglicher Form von Inferenz unbekannt. Allgemeines Ziel von parameterischen Inferenzverfahren ist es damit, basierend auf einem vorliegenden Datensatz eine möglichst valide Aussage hinsichtlich des wahren, aber unbekannten Parameters θ zu treffen. In diesem Sinne ist der wahre, aber unbekannte Parameterwert, nur indirekt beobachtbar. Dies wird manchmal auch durch die Sprechweisen ausgedrückt, dass der wahre, aber unbekannte Parameterwert *unbeobachtbar* oder *latent*, d.h. nicht unmittelbar sichtbar oder zu erfassen, ist. In der mathematischen Analyse von Inferenzverfahren betrachtet man alle möglichen wahren, aber unbekannten, Parameterwerte, verzichtet deshalb also meist auf eine explizite notationelle Auszeichnung des in einem Anwendungskontext unterstellten wahren, aber unbekannten, Parameterwerts.

Beispiele

Mit dem univariaten Normalverteilungsmodell und dem Bernoullimodell wollen wir zwei erste Beispiele für frequentistische Inferenzmodelle geben.

Definition 30.2 (Normalverteilungsmodell). Das univariate parametrische Produktmodell

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \quad (30.2)$$

mit

$$\mathcal{Y} := \mathbb{R}^n, \mathcal{A} := \mathcal{B}(\mathbb{R}^n), \mathbb{P}_\theta := N(\mu, \sigma^2), \theta := (\mu, \sigma^2), \Theta := \mathbb{R} \times \mathbb{R}_{>0}, \quad (30.3)$$

also

$$\{\mathbb{P}_\theta | \theta \in \Theta\} := \left\{ \prod_{i=1}^n N(\mu, \sigma^2) | (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \right\}, \quad (30.4)$$

und damit

$$y_1, \dots, y_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \quad (30.5)$$

heißt *Normalverteilungsmodell*.

•

Das Normalverteilungsmodell ist Grundlage vieler populärer statistischen Verfahren die im Rahmen des Allgemeinen Linearen Modells integrativ betrachtet werden. Man beachte, dass die Annahme normalverteilter Daten dabei durch additive normalverteilte Fehlerterme motiviert ist, wie wir in Kapitel 27 schon kurz angerissen haben und an späterer Stelle vertiefen werden.

Definition 30.3 (Bernoullimodell). Das univariate parametrische Produktmodell

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \quad (30.6)$$

mit

$$\mathcal{Y} := \{0, 1\}^n, \mathcal{A} := \mathcal{P}(\{0, 1\}^n), \mathbb{P}_\theta := \text{Bern}(\mu), \theta := \mu, \Theta :=]0, 1[, \quad (30.7)$$

also

$$\{\mathbb{P}_\theta | \theta \in \Theta\} := \left\{ \prod_{i=1}^n \text{Bern}(\mu) | \mu \in]0, 1[\right\}, \quad (30.8)$$

und damit

$$y_1, \dots, y_n \sim \text{Bern}(\mu) \text{ mit } \mu \in]0, 1[, \quad (30.9)$$

heißt *Bernoullimodell*.

•

30.2. Statistiken und Schätzer

Vor dem Hintergrund Frequentistischer Inferenzmodelle wollen wir nun formalisieren, was unter den Begriffen einer *Statistik* und eines *Schätzers* zu verstehen ist.

Definition 30.4 (Statistik). $\mathcal{M}$ sei ein Frequentistisches Inferenzmodell und $(\Sigma, \mathcal{S})$ sei ein Messraum. Dann ist eine *Statistik* ein Zufallsvektor der Form

$$S : \mathcal{Y} \rightarrow \Sigma. \quad (30.10)$$

•

Sowohl Daten als auch Statistiken werden in der Frequentistischen Inferenz also durch Zufallsvektoren (im univariaten Fall entsprechend durch Zufallsvariablen) modelliert. Allerdings unterscheiden sich diese Zufallsvektoren hinsichtlich ihrer intuitiven Bedeutung fundamental: Daten repräsentieren den Ausgang von Messvorgängen unter Unsicherheit, Statistiken dagegen modellieren von Datenwissenschaftler:innen konstruierte Funktionen von Daten. Diese liefern im besten Fall datenbasierte Informationen, aus denen sich Schlüsse über die latenten datengenerierenden Zufallsvorgänge ziehen lassen. Die Tatsache, dass Statistiken zufällig sind ergibt sich dabei daraus, dass sie als Funktionen auf zufällige Daten angewendet werden. (vgl. etwa Theorem 21.1).

Beispiele

$\mathcal{M}$ sei das Normalverteilungsmodell. Dann sind zum Beispiel folgende Zufallsvariablen Statistiken:

- Das *Stichprobenmittel*

$$\bar{y} : \mathbb{R}^n \rightarrow \mathbb{R}, y \mapsto \bar{y}(y) := \frac{1}{n} \sum_{i=1}^n y_i, \quad (30.11)$$

- Die *Stichprobenvarianz*

$$s^2 : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto s^2(y) := \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}(y))^2, \quad (30.12)$$

- Die *Stichprobenstandardabweichung*

$$s : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto s(y) := \sqrt{s^2(y)}, \quad (30.13)$$

Oft bleibt wie hier das Wesen von Statistiken als Zufallsvariablen oder Zufallsvektoren notationell eher implizit. Dies ändert allerdings nichts an der fundamental zu beachtenden Tatsache, dass Statistiken als Funktionen von vom Zufall abhängigen Werten selbst wiederum Zufallsvariablen oder Zufallsvektoren sind.

Definition 30.5 (Schätzer). $\mathcal{M}$ sei ein Frequentistisches Inferenzmodell, $(\Sigma, \mathcal{S})$ sei ein Messraum und $\tau : \Theta \rightarrow \Sigma$ sei eine Abbildung, die jedem $\theta \in \Theta$ eine Kenngröße $\tau(\theta) \in \Sigma$ zuordnet. Dann heißt eine Statistik

$$\hat{\tau} : \mathcal{Y} \rightarrow \Sigma \quad (30.14)$$

ein *Schätzer* für τ .

•

Schätzer schätzen also Funktionen der Parameter eines parametrischen Frequentistischen Inferenzmodells. Typische Beispiele für solche Funktionen sind

- $\tau(\theta) := \theta$ für die Schätzung des Parameters θ ,
- $\tau(\theta) := \theta_i$ mit $\theta \in \mathbb{R}^d, d > 1$ für die Schätzung einer Komponente des Parameters θ ,
- $\tau(\theta) := \mathbb{E}_\theta(y_1)$ für die Schätzung des Erwartungswerts,
- $\tau(\theta) := \mathbb{V}_\theta(y_1)$ für die Schätzung der Varianz.

Im Falle $\tau(\theta) := \theta$, also der Schätzung von Parametern, schreibt man üblicherweise $\hat{\theta}$. Man beachte, dass Schätzer Zahlwerte in Σ annehmen, bei der Schätzung von Parametern etwa in Θ . Sie heißen deshalb auch *Punktschätzer*. Dies ist ein Charakteristikum Frequentistischer Inferenzverfahren. Im Rahmen der Bayesianischen Inferenz können Schätzer auch generalisierte Formen annehmen, zum Beispiel werden dort auch Wahrscheinlichkeitsverteilungen als Schätzer betrachtet. Schließlich ist festzuhalten, dass die Definition eines Schätzers keinerlei Aussage über die Validität von Schätzern macht. Nicht jeder Schätzer ist damit *per se* ein guter Schätzer. In der Frequentistischen Inferenz definiert man deshalb zusätzlich *Schätzgütekriterien*, wie in Kapitel 31 ausführlich dargestellt.

Beispiel

$\mathcal{M}$ sei das Normalverteilungsmodell. Dann ist zum Beispiel das Stichprobenmittel $\bar{y} : \mathbb{R}^n \rightarrow \mathbb{R}$ ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}, (\mu, \sigma^2) \mapsto \tau(\mu, \sigma^2) := \mu. \quad (30.15)$$

Ebenso ist $\bar{y}$ ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}, (\mu, \sigma^2) \mapsto \tau(\mu, \sigma^2) := \mathbb{E}_{\mu, \sigma^2}(y_1). \quad (30.16)$$

Weiterhin ist die konstante Funktion

$$\hat{\tau} : \mathbb{R}^n \rightarrow \mathbb{R}, y \mapsto \hat{\tau}(y) := 42 \quad (30.17)$$

ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, (\mu, \sigma^2) \mapsto \tau(\mu, \sigma^2) := \sigma^2. \quad (30.18)$$

Dass eine Funktion $\hat{\tau} : \mathcal{Y} \rightarrow \Sigma$ ein Schätzer ist impliziert also keinesfalls, dass sie ein guter Schätzer ist.

30.3. Standardannahmen und Standardproblemstellungen

Wir wollen die in diesem Kapitel bisher betrachteten Konzepte zunächst noch einmal unter dem Begriff der *datenanalytischen Standardannahmen der Frequentistischen Inferenz* zusammenfassen (vgl. auch Abbildung 30.1). Dazu sei $\mathcal{M}$ ein univariates parametrisches Frequentistisches Produktmodell und es seien $y_1, \dots, y_n \sim p_\theta$ die Zufallsvariablen der Stichprobe, die wir etwa in einem Zufallsvektor $y := (y_1, \dots, y_n)$ zusammenfassen können. Von einem konkret vorliegenden Datensatz $y_1, \dots, y_n$ mit $y_i \in \mathbb{R}$, $i = 1, \dots, n$, den wir etwa in einem n -dimensionalen Vektor $y := (y_1, \dots, y_n)^T \in \mathbb{R}^n$ zusammenfassen können, wird dann angenommen, dass er eine der möglichen Realisierungen von y auf Grundlage einer Verteilung $\mathbb{P}_\theta$ mit wahren, aber unbekannten, Parameter θ ist. Aus Frequentistischer Sicht kann man dabei die Beobachtung eines Datensatzes unendlich oft wiederholen und zu jeder Datenrealisierung Schätzer oder Statistiken auswerten, so zum Beispiel das Stichprobenmittel:

Datenrealisierung $y^{(1)} := (y_1^{(1)}, \dots, y_n^{(1)})$ mit $\bar{y}^{(1)} = \frac{1}{n} \sum_{i=1}^n y_i^{(1)}$

Datenrealisierung $y^{(2)} := (y_1^{(2)}, \dots, y_n^{(2)})$ mit $\bar{y}^{(2)} = \frac{1}{n} \sum_{i=1}^n y_i^{(2)}$

Datenrealisierung $y^{(3)} := (y_1^{(3)}, \dots, y_n^{(3)})$ mit $\bar{y}^{(3)} = \frac{1}{n} \sum_{i=1}^n y_i^{(3)}$

Datenrealisierung $y^{(4)} := (y_1^{(4)}, \dots, y_n^{(4)})$ mit $\bar{y}^{(4)} = \frac{1}{n} \sum_{i=1}^n y_i^{(4)}$

Datenrealisierung $y^{(5)} := (y_1^{(5)}, \dots, y_n^{(5)})$ mit $\bar{y}^{(5)} = \frac{1}{n} \sum_{i=1}^n y_i^{(5)}$

...

Vor diesem Hintergrund behandelt die Frequentistische Inferenz dann üblicherweise folgende Standardproblemstellungen:

- (1) *Punktschätzung.* Ziel der Punktschätzung ist es, auf Grundlage beobachteter Daten einen präzisen und im Frequentistischen Sinn möglichst guten Tipp für den wahren, aber unbekannten, Parameterwert abzugeben.
- (2) *Konfidenzintervallbestimmung.* Ziel der Konfidenzintervallbestimmung ist es, basierend auf der angenommenen Datenverteilung und den beobachteten Daten durch eine Intervallschätzung einen möglichst sicheren, wenn auch oft unpräzisen, Tipp für den wahren, aber unbekannten, Parameterwert abzugeben.
- (3) *Hypothesentests.* Ziel des Frequentistischen Hypothesentestens ist es, basierend auf der angenommenen Verteilung der Daten in einer möglichst zuverlässigen Form zu entscheiden, ob ein wahrer, aber unbekannter Parameterwert in einer von zwei sich gegenseitig ausschließenden Untermengen des Parameterraumes

Verfahren zur Lösung dieser Problemstellungen bezeichnen wir als *Frequentistische Inferenzverfahren*. Um die *Qualität* von Frequentistischen Inferenzverfahren zu beurteilen, betrachtet man in der Frequentistischen Inferenz üblicherweise die Verteilungen von Schätzern und Statistiken unter der Annahme von $y = (y_1, \dots, y_n) \sim p_\theta$. Man fragt zum Beispiel nach der Verteilung der oben skizzierten $\bar{y}^{(1)}$, $\bar{y}^{(2)}$, $\bar{y}^{(3)}$, $\bar{y}^{(4)}$, ... also der Verteilung der Zufallsvariable $\bar{y}_n$. Wenn ein Inferenzverfahren auf Grundlage dieser Annahmen für "gut" befunden wird, so heißt das also insbesondere nur, dass das Verfahren bei häufiger Anwendung "im Mittel gut" ist. Im Einzelfall, also im Normalfall nur eines vorliegenden Datensatzes, kann sie auch "schlecht" sein. Wir werden diese Denkweise insbesondere im Kontext

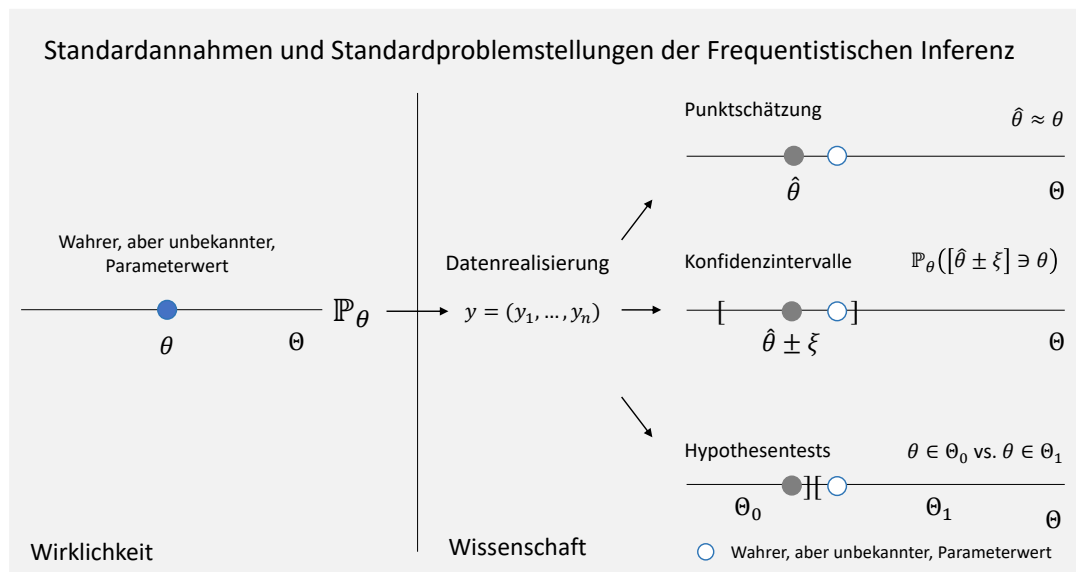


Abbildung 30.1. Standardannahmen und Standardproblemstellungen Frequentistischer Inferenz. Die Frequentistische Inferenz unterstellt, dass es in der Wirklichkeit einen wahren, aber unbekannten, Parameterwert des Wahrscheinlichkeitsmaßes $\mathbb{P}_\theta$ gibt, dass als Modell für die Erhebung eines Datensatzes dient. Ein konkret vorliegender Datensatz $y = (y_1, \dots, y_n)$ ist dann eine (und insbesondere *nur* eine) der möglichen Realisierungen des anhand $\mathbb{P}_\theta$ verteilten Zufallsvektors $y := (y_1, \dots, y_n)$. Auf Grundlage dieser Realisierung beabsichtigen die Verfahren zur Behandlung der Frequentistischen Standardproblemstellungen von Punktschätzung, Konfidenzintervallbestimmung und Hypothesentestauswertung möglichst valide Aussagen hinsichtlich des wahren, aber unbekannten Parameterwertes zu machen, in den Kapiteln Kapitel 31, Kapitel 32 und Kapitel 33 diskutiert werden soll. Der tatsächliche wahre, aber unbekannte, Parameterwert aber bleibt auch nach Abschluss eines Inferenzverfahrens immer unbekannt.

der Punktschätzung (Kapitel 31) vertiefen. Ebenso beurteilt die Frequentistische Inferenz die *Stärke empirischer Evidenz* vor dem Hintergrund der angenommenen Verteilung von Schätzern und Statistiken in Szenarien, in denen angenommen wird, das interessierende Effekt *nicht* existieren (sogenannte “Nullhypothesen”). Diese Denkweise verdeutlichen wir insbesondere im Rahmen der Betrachtung von Konfidenzintervallen (Kapitel 32) und Hypothesentests (Kapitel 33).

30.3.1. Anwendungsbeispiel

Für ein erstes Anwendungsbeispiel Frequentistischer Inferenzverfahren betrachten wir die evidenzbasierte Evaluation einer Psychotherapie bei Depression. Dazu sei der in Tabelle 41.1 dargestellte Datensatz von an $n = 12$ Patient:innen Differenzen von Prä- und Post-Therapie erhobenen BDI-II Scores gegeben (dBDI; Beck (1961), Beck et al. (1996)). Die dBDI Werte sollen dabei die Reduktion des BDI-II Scores der Patient:innen über den Zeitraum der Therapie widerspiegeln. Hohe positive Werte von dBDI entsprechen also einer starken Abnahme der durch den BDI-II Score quantifizierten Depressionssymptomatik, Werte um Null entsprechen keiner wesentlichen Änderung und negative Werte entsprechen einer Zunahme der durch den BDI-II Score quantifizierten Depressionssymptomatik.

Tabelle 30.1. Prä-Post-Therapie BDI-II Reduktionsscores von $n = 12$ Patient:innen

dBDI
-1
3
-2
9
3
-2
4
5
5
1
9
4

Für jeden der $n := 12$ dBDI Werte legen wir nun das Modell

$$y_i := \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \dots, n \quad (30.19)$$

zugrunde. Damit wird der dBDI der i ten Patient:in also mithilfe einer über die Gruppe von Patient:innen identischen BDI-II Score Reduktion $\mu \in \mathbb{R}$ und einer Patient:innen-spezifischen normalverteilten BDI-II Score Reduktionsabweichung ε_i erklärt und es wird angenommen, dass sich diese Reduktionsabweichungen zwischen Patient:innen nicht gegenseitig beeinflussen. Intuitiv wird also davon ausgegangen, dass die Therapie einen Effekt

hat, der bei allen Patient:innen zur gleichen BDI-II Score Reduktion μ führt und sich die Unterschiede in den beobachteten dBDI Werten durch eine Vielzahl weiterer Zufallsvorgänge, die in der Summe normalverteilt und zentriert sind erklären lässt. Alternativ mag man diese Abweichungen als Realisierungen der Unsicherheit verstehen mit der das Modell in Gleichung 30.19 behaftet ist.

Aus Gleichung 30.19 folgt dann direkt

$$y_1, \dots, y_n \sim N(\mu, \sigma^2), \quad (30.20)$$

denn für $i = 1, \dots, n$ und mit

$$y_i = f(\varepsilon_i) \text{ mit } f: \mathbb{R} \rightarrow \mathbb{R}, e_i \mapsto f(e_i) := e_i + \mu. \quad (30.21)$$

gilt für die WDFen der y_i , dass

$$\begin{aligned} p_{y_i}(y_i) &= \frac{1}{|1|} p_{\varepsilon_i} \left(\frac{y_i - \mu}{1} \right) \\ &= N(y_i - \mu; 0, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (y_i - \mu - 0)^2 \right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (y_i - \mu)^2 \right) \\ &= N(y_i; \mu, \sigma^2) \end{aligned} \quad (30.22)$$

Die Standardproblemstellungen der Frequentistischen Inferenz führen in diesem Anwendungsszenario dann auf folgende Fragen, die wir jeweils in den Kapiteln zur Punktschätzung, Konfidenzintervallbestimmung, und Hypothesentestevaluation wieder aufgreifen wollen:

- (1) Was sind sinnvolle Tipps für die wahren, aber unbekannten, Parameterwerte μ und σ^2 , also den wahren, aber unbekannten, Erwartungswert der BDI-II Score Reduktion und ihre wahre, aber unbekannte, Varianz? Wie gut ist die Therapie also in diesem quantitativen Sinn, wenn wir versuchen, die Patient:innen-abhängigen Abweichungen zu berücksichtigen und wie groß ist die in der Datengeneration inhärente Unsicherheit?
- (2) Wie kann im Sinne einer Intervallschätzung eine möglichst sichere Schätzung des wahren, aber unbekannten, Erwartungswert der BDI-II Score Reduktion gelingen? Wie unpräzise muss eine solche Schätzung sein, um möglichst verlässlich zu sein?
- (3) Entscheiden wir uns sinnvollerweise für eine der Hypothesen, dass die Therapie nicht wirksam ist ($\mu = 0$) oder dass sie etwa im positiven ($\mu > 0$) oder auch im negativen Sinne wirksam ist ($\mu < 0$)? Und wenn wir uns für eine dieser Hypothesen entscheiden sollten, mit welcher Fehlerwahrscheinlichkeit täten wir dies? Wie hoch ist also die einer solchen Entscheidung innewohnende Unsicherheit?

31. Punktschätzung

In diesem Kapitel gehen wir immer im Sinne der in Kapitel 30 eingeführten Begrifflichkeiten immer von einem parametrischem Produktmodell

$$\mathcal{M} := \{\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}\} \quad (31.1)$$

mit n -dimensionalen Stichprobenraum (z.B. $\mathcal{Y} := \mathbb{R}^n$), d -dimensionalen Parameteraum $\Theta \subset \mathbb{R}^d$ und gegebener WMF oder WDF p_θ für alle $\theta \in \Theta$ aus. $y := (y_1, \dots, y_n)$ bezeichnet die zu $\mathcal{M}$ gehörende Stichprobe unabhängig und identisch verteilter Zufallsvariablen, es gilt also durchgängig

$$y_1, \dots, y_n \sim \mathbb{P}_\theta. \quad (31.2)$$

Wesen und Ziel der hier behandelten *Punktschätzung* ist es, basierend auf der Stichprobe einen möglichst guten Tipp für eine interessierende Kennzahl der Verteilung $\mathbb{P}_\theta$ einer Stichprobenvariable anzugeben. Dabei ist der Tipp von der gleichen mathematischen Wesensart wie die entsprechende Kennzahl, also zum Beispiel ein skalarer Wert für einen skalaren Parameter. Dies ist nicht die einzige Möglichkeit der Schätzung, mit den Konfidenzintervallen werden wir in Kapitel 32 eine Möglichkeit der Schätzung von skalaren Werten durch Intervalle kennenlernen und die Bayesianische Inferenz nutzt zur Schätzung von skalaren Werten in aller Regel Wahrscheinlichkeitsverteilungen. Die zu schätzenden Kennzahlen von $\mathbb{P}_\theta$ sind oft schlicht die wahren, aber unbekannten, Parameter selbst. Wir widmen uns diesem Fall ausführlich in Kapitel ?? . Allerdings sind viele grundlegende Resultate der Frequentistischen Punktschätzung auch dann valide, wenn es sich bei zu schätzenden Kennzahlen nicht um die Parameter selbst, sondern, bei parameterischen Produktmodellen, Funktionen von ihnen handelt, wie zum Beispiel die Schätzung des Erwartungswerts, der Varianz, oder der Standardabweichung von $\mathbb{P}_\theta$. Beginnen wollen wir allerdings mit der *Parameterschätzung*. Um den wahren, aber unbekannten, Parameter eines parametrischen Produktmodells oder auch allgemein eines Frequentistischen Inferenzmodells zu schätzen, nutzt man in der Frequentistischen Inferenz sogenannte *Parameterpunktschätzer*.

Definition 31.1. $\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\})$ sei ein Frequentistisches Inferenzmodell, $(\Theta, \mathcal{S})$ sei ein Messraum und $\hat{\theta} : \mathcal{Y} \rightarrow \Theta$ sei eine Abbildung. Dann nennt man $\hat{\theta}$ einen *Parameterpunktschätzer* für θ .

•

Parameterpunktschätzer werden meist auch einfach als *Parameterschätzer* bezeichnet. Im Sinne von Definition 30.5 sind Parameterpunktschätzer Schätzer mit $\tau := \text{id}_\Theta$. Parameterpunktschätzer sind also Funktionen von Daten und nehmen Zahlwerte im Parameterraum an. Als als Funktionen von Zufallsvariablen sind Parameterschätzer natürlich auch Zufallsvariablen. Oft wird dabei notationell nicht zwischen $\hat{\theta}$ als Zufallsvariable und $\hat{\theta}(y)$ als Wert dieser Zufallsvariable unterschieden.

Definition 31.1 macht offenbar keine Angabe darüber, wie ein Parameterpunktschätzer zu konstruieren ist oder inwieweit er dann ein sinnvoller Schätzer sein mag. Im Folgenden

werden wir mit der *Maximum-Likelihood Schätzung* zunächst ein allgemeines Prinzip diskutieren, das es erlaubt, für ein gegebenes Frequentistisches Inferenzmodell Parameterschätzer zu bestimmen, die, wie wir an späterer Stelle sehen werden, garantiert bestimmte wünschenswerte Eigenschaften haben (Kapitel 31.4). Dabei beziehen sich diese Eigenschaften allgemein auf sein qualitatives Verteilungsverhalten bei festen Stichprobenumfang bzw. im Grenzübergang zu einem unendlich großen Stichprobenumfang. Wir führen diese Eigenschaften allgemein und insbesondere auch in der Schätzung auf andere Kennzahlen von $\mathbb{P}_\theta$ in Kapitel 31.2 und Kapitel 31.3 ein.

31.1. Maximum-Likelihood Schätzung

Die Grundidee der Maximum-Likelihood Schätzung ist es, als Tipp für einen wahren, aber unbekannten, Parameterwert denjenigen Parameterwert zu wählen, für den die Wahrscheinlichkeit der beobachteten Daten maximal ist. Dafür ist es zunächst nötig, die Wahrscheinlichkeit beobachteter Daten eines Frequentistischen Inferenzmodells als Funktion des betreffenden Parameters zu betrachten. Dies ermöglichen und formalisieren die *Likelihood-Funktion* und ihr Logarithmus, die *Log-Likelihood-Funktion*. Wir definieren diese Begriffe hier für parametrische Produktmodelle.

Definition 31.2 (Likelihood-Funktion und Log-Likelihood-Funktion). $\mathcal{M}$ sei ein parametrisches Produktmodell mit WMF oder WDF p_θ . Dann ist die *Likelihood-Funktion* definiert als

$$L : \Theta \rightarrow [0, \infty[, \theta \mapsto L(\theta) := \prod_{i=1}^n p_\theta(y_i) \quad (31.3)$$

und die *Log-Likelihood-Funktion* ist definiert als

$$\ell_n : \Theta \rightarrow \mathbb{R}, \theta \mapsto \ell(\theta) := \ln L(\theta). \quad (31.4)$$

•

Die Likelihood-Funktion ist also eine Funktion des Parameters und ihre Funktionswerte sind die Werte der gemeinsamen WMF bzw. WDF beobachteter Datenwerte $y_1, \dots, y_n$. Generell gibt es keinen Grund anzunehmen, dass eine Likelihood-Funktion über dem Parameterraum zu 1 integriert, die Likelihood-Funktion ist also im Allgemeinen keine WMF oder WDF. Die Log-Likelihood Funktion ist schlicht die logarithmierte Likelihood-Funktion. Ein nach dem Prinzip der Maximum-Likelihood Schätzung gewonnener Parameterschätzer soll nun die Likelihood-Funktion bzw. die Log-Likelihood-Funktion maximieren. Dies führt auf folgende Definition des Begriffs des *Maximum-Likelihood Schätzers*.

Definition 31.3 (Maximum-Likelihood Schätzer). $\mathcal{M}$ sei ein parametrisches Produktmodell mit Parameter $\theta \in \Theta$. Ein *Maximum-Likelihood Schätzer* von θ ist definiert als

$$\hat{\theta}^{\text{ML}} : \mathcal{Y} \rightarrow \Theta, y \mapsto \hat{\theta}^{\text{ML}}(y) := \operatorname{argmax}_{\theta \in \Theta} L(\theta) = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta) \quad (31.5)$$

•

Man beachte bei Definition 31.3, dass eine Maximumstelle der Log-Likelihood-Funktion der Maximumstelle der Likelihood-Funktion entspricht, weil die Logarithmusfunktion eine monoton steigende Funktion ist. Das Arbeiten mit der Log-Likelihood-Funktion ist allerdings oft einfacher als das direkte Arbeiten mit der Likelihood-Funktion, zum Beispiel, wenn in der WMF oder WDF des Modells Exponentialfunktionen auftauchen. Weiterhin beachte man bei Definition 31.3, dass Definition 31.2 impliziert, dass

$$\hat{\theta}^{\text{ML}}(y) = \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_{\theta}(y_i) = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \ln p_{\theta}(y_i) \quad (31.6)$$

was die Abhängigkeit eines Maximum-Likelihood Schätzers von den Daten verdeutlicht.

Mit Definition 31.3 handelt es sich bei der Maximum-Likelihood Schätzung also um das Problem, Extremalstellen einer Funktion zu bestimmen. Für diese Extremalstellen stellt die Differentialrechnung bekanntlich notwendige und hinreichende Bedingungen bereit (vgl. Kapitel 6.2). In ihrer Anwendung auf die Gewinnung von Maximum-Likelihood Schätzern begnügt man sich zumeist aufgrund der funktionellen Form der betrachteten Funktionen mit dem Erfülltsein der notwendigen Bedingung. Je nach Beschaffenheit der Log-likelihood Funktion bieten sich dann Methoden entweder der analytischen Optimierung oder der numerischen Optimierung an. In den folgenden klassischen Beispielen nutzen wir einen analytischen Zugang anhand folgendem standardisierten Vorgehen:

- (1) Formulierung der Log-Likelihood-Funktion.
- (2) Bestimmung der ersten Ableitung der Log-Likelihood-Funktion und Nullsetzen.
- (3) Auflösen nach potentiellen Maximumstellen.

In Theorem 31.1 zeigen wir, dass der Maximum-Likelihood Schätzer für den Parameter des Bernoullimodells aus Definition 30.3 durch das entsprechende Stichprobenmittel gegeben ist und in Theorem 31.2 zeigen wir, dass die Maximum-Likelihood Schätzer für den Erwartungswert- und Varianzparameter des Normalverteilungsmodells aus Definition 30.2 durch das Stichprobenmittel und eine modifizierte Stichprobenvarianz, respektive, gegeben sind.

Beispiele

Theorem 31.1 (Maximum-Likelihood Schätzer des Bernoullimodells). *$\mathcal{M}$ sei das Bernoullimodell, es gelte also $y_1, \dots, y_n \sim \text{Bern}(\mu)$. Dann ist*

$$\hat{\mu}^{\text{ML}} : \{0, 1\}^n \rightarrow [0, 1], y \mapsto \hat{\mu}^{\text{ML}}(y) := \frac{1}{n} \sum_{i=1}^n y_i \quad (31.7)$$

ein Maximum-Likelihood Schätzer von μ

◦

Beweis. Wir formulieren zunächst die Log-Likelihood-Funktion. Für die Likelihood-Funktion gilt

$$L :]0, 1[\rightarrow]0, 1[, \mu \mapsto L(\mu) := \prod_{i=1}^n \mu^{y_i} (1 - \mu)^{1-y_i} = \mu^{\sum_{i=1}^n y_i} (1 - \mu)^{n - \sum_{i=1}^n y_i}. \quad (31.8)$$

Logarithmieren ergibt

$$\ell :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto \ell(\mu) = \ln \mu \sum_{i=1}^n y_i + \ln(1 - \mu) \left(n - \sum_{i=1}^n y_i \right). \quad (31.9)$$

Wir werten dann die Ableitung der Log-Likelihood-Funktion aus. Es gilt

$$\begin{aligned}\frac{d}{d\mu}\ell(\mu) &= \frac{d}{d\mu} \left(\ln \mu \sum_{i=1}^n y_i + \ln(1-\mu) \left(n - \sum_{i=1}^n y_i \right) \right) \\ &= \frac{d}{d\mu} \ln \mu \sum_{i=1}^n y_i + \frac{d}{d\mu} \ln(1-\mu) \left(n - \sum_{i=1}^n y_i \right) \\ &= \frac{1}{\mu} \sum_{i=1}^n y_i - \frac{1}{1-\mu} \left(n - \sum_{i=1}^n y_i \right).\end{aligned}\quad (31.10)$$

Nullsetzen ergibt dann folgende *Maximum-Likelihood-Gleichung* als notwendige Bedingung für einen Maximum-Likelihood Schätzer im Bernoullimodell:

$$\frac{1}{\hat{\mu}^{\text{ML}}} \sum_{i=1}^n y_i - \frac{1}{1-\hat{\mu}^{\text{ML}}} \left(n - \sum_{i=1}^n y_i \right) = 0. \quad (31.11)$$

Auflösen der Maximum-Likelihood-Gleichung nach $\hat{\mu}^{\text{ML}}$ ergibt dann

$$\begin{aligned}\frac{1}{\hat{\mu}^{\text{ML}}} \sum_{i=1}^n y_i - \frac{1}{1-\hat{\mu}^{\text{ML}}} \left(n - \sum_{i=1}^n y_i \right) &= 0 \\ \Leftrightarrow \hat{\mu}^{\text{ML}}(1-\hat{\mu}^{\text{ML}}) \left(\frac{1}{\hat{\mu}^{\text{ML}}} \sum_{i=1}^n y_i - \frac{1}{1-\hat{\mu}^{\text{ML}}} \left(n - \sum_{i=1}^n y_i \right) \right) &= 0 \\ \Leftrightarrow \sum_{i=1}^n y_i - \hat{\mu}^{\text{ML}} \sum_{i=1}^n y_i - n\hat{\mu}^{\text{ML}} + \hat{\mu}^{\text{ML}} \sum_{i=1}^n y_i &= 0 \\ \Leftrightarrow n\hat{\mu}^{\text{ML}} &= \sum_{i=1}^n y_i \\ \Leftrightarrow \hat{\mu}^{\text{ML}} &= \frac{1}{n} \sum_{i=1}^n y_i.\end{aligned}\quad (31.12)$$

$\hat{\mu}^{\text{ML}} = \frac{1}{n} \sum_{i=1}^n y_i$ ist also ein Kandidat für einen Maximum-Likelihood Schätzer von μ . Dies könnte durch Betrachten der zweiten Ableitung von ℓ verifiziert werden, worauf wir hier aber verzichten wollen. $\square$

Theorem 31.2 (Maximum-Likelihood Schätzer des Normalverteilungsmodells). *$\mathcal{M}$ sei das Normalverteilungsmodell, es gilt also $y_1, \dots, y_n \sim N(\mu, \sigma^2)$. Dann sind*

$$\hat{\mu}^{\text{ML}} : \mathbb{R}^n \rightarrow \mathbb{R}, y \mapsto \hat{\mu}^{\text{ML}}(y) := \frac{1}{n} \sum_{i=1}^n y_i \quad (31.13)$$

und

$$\hat{\sigma}^{2\text{ML}} : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto \hat{\sigma}^{2\text{ML}}(y) := \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}})^2. \quad (31.14)$$

Maximum-Likelihood Schätzer für μ und σ^2 , respektive.

◦

Beweis. Wir formulieren zunächst die Log-Likelihood-Funktion. Für die Likelihood-Funktion ergibt sich

$$\begin{aligned}L : \mathbb{R} \times \mathbb{R}_{>0} &\rightarrow \mathbb{R}_{>0}, (\mu, \sigma^2) \mapsto L(\mu, \sigma^2) := \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right).\end{aligned}\quad (31.15)$$

Logarithmieren ergibt dann

$$\ell : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}, (\mu, \sigma^2) \mapsto \ell_n(\mu, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2. \quad (31.16)$$

Die Auswertung der partiellen Ableitungen der Log-Likelihood-Funktion ergeben dann

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma^2) = -\frac{\partial}{\partial \mu} \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{\partial}{\partial \mu} (y_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) \quad (31.17)$$

und

$$\frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2) = -\frac{n}{2} \frac{\partial}{\partial \sigma^2} \ln \sigma^2 - \frac{\partial}{\partial \sigma^2} \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2. \quad (31.18)$$

Das System der Maximum-Likelihood Gleichungen als Ausdruck der notwendigen Bedingungen für Extremstellen der Log-Likelihood-Funktion hat in diesem Fall also die Form

$$\sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}}) = 0 \text{ und } -\frac{n}{2\hat{\sigma}^{2\text{ML}}} + \frac{1}{2\hat{\sigma}^{4\text{ML}}} \sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}})^2 = 0. \quad (31.19)$$

Lösen des Systems der Maximum-Likelihood Gleichungen ergibt dann zunächst

$$\sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}}) = 0 \Leftrightarrow \sum_{i=1}^n y_i = n\hat{\mu}^{\text{ML}} \Leftrightarrow \hat{\mu}^{\text{ML}} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (31.20)$$

Damit ist

$$\hat{\mu}^{\text{ML}} = \frac{1}{n} \sum_{i=1}^n y_i \quad (31.21)$$

ein potentieller Maximum-Likelihood Schätzer von μ . Einsetzen dieses Schätzers in die zweite Maximum-Likelihood Gleichung ergibt dann

$$\begin{aligned} -\frac{n}{2\hat{\sigma}^{2\text{ML}}} + \frac{1}{2\hat{\sigma}^{4\text{ML}}} \sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}})^2 &= 0 \\ \Leftrightarrow -n\hat{\sigma}^{2\text{ML}} + \sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}})^2 &= 0 \\ \Leftrightarrow \hat{\sigma}^{2\text{ML}} &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}})^2. \end{aligned} \quad (31.22)$$

Also ist

$$\hat{\sigma}^{2\text{ML}} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}})^2 \quad (31.23)$$

ein potentieller Maximum-Likelihood Schätzer von σ^2 . Beide potentiellen Maximum-Likelihood Schätzer können durch Betrachten der zweiten Ableitung von ℓ verifiziert werden, worauf wir hier verzichten wollen. $\square$

Man beachte bei Theorem 31.2, dass $\hat{\mu}^{\text{ML}}$ mit dem Stichprobenmittel $\bar{y}$ identisch ist, aber $\hat{\sigma}^{2\text{ML}}$ nicht mit der Stichprobenvarianz S^2 übereinstimmt. Im Gegensatz zur Stichprobenvarianz findet sich im Maximum-Likelihood Schätzer von σ^2 der multiplikative Faktor $\frac{1}{n}$, nicht, wie in der Stichprobenvarianz, der multiplikative Faktor $\frac{1}{n-1}$. Wir werden auf diesen Unterschied im Kontext der Schätzereigenschaften zurückkommen.

Anwendungsbeispiel

Zum Abschluss dieses Abschnitts wollen wir Theorem 31.2 im Kontext des Anwendungsbeispiels aus Kapitel 30.3.1 betrachten. Wir hatten dort den beobachteten dBDI Werten das Normalverteilungsmodell

$$y_1, \dots, y_n \sim N(\mu, \sigma^2) \quad (31.24)$$

zugrundegelegt. Die Maximum-Likelihood Schätzer für die Parameter dieses Modells lassen sich dann anhand von Theorem 31.2 mithilfe der **R** Stichprobenmittel- und Stichprobenvarianzfunktionen `mean()` und `var()` und unter Beachtung der Identität

$$\frac{n-1}{n}s^2 = \frac{n-1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}^{\text{ML}})^2 = \hat{\sigma}^{2\text{ML}} \quad (31.25)$$

wie in folgendem **R** Code auswerten.

```
D      = read.csv("../data/bdi-ii-datensatz.csv") # Datensatzeinlesen
y      = D$BDI # Datenauswahl
mu_hat = mean(y) # Maximum-Likelihood Schätzung des Erwartungswertparameters
n      = length(y) # Anzahl der Datenpunkte
sigsqr_hat = ((n-1)/n)*var(y) # Maximum-Likelihood Schätzung des Varianzparameters
cat("mu_hat      :", mu_hat, "\nsigsqr_hat :", sigsqr_hat) # Ausgabe
```

```
mu_hat      : 3.166667
sigsqr_hat : 12.63889
```

Basierend auf dem Prinzip der Maximum-Likelihood Schätzung und den vorliegenden $n = 12$ Datenpunkten sind also

$$\hat{\mu}^{\text{ML}} = 3.17 \text{ und } \hat{\sigma}^{2\text{ML}} = 12.6 \quad (31.26)$$

Tipps für die wahren, aber unbekannten, Parameter des Modells.

31.2. Schätzereigenschaften bei endlichen Stichproben

Allgemein betreffen Frequentistische Schätzereigenschaften die Verteilung von Schätzern in Abhängigkeit der Verteilung der ihn zugrundeliegenden Daten. Weil Daten in der Frequentistischen Inferenz zufällig sind, sind auch Schätzer zufällig. Speziell werden beobachtete Datenwerte als Realisierungen von Zufallsvariablen interpretiert. Schätzer als Funktionen von Zufallsvariablen sind damit auch Zufallsvariablen, auch wenn sie natürlich bei Vorliegen eines konkreten Datensatzes nur einen konkreten Wert annehmen. Wir unterscheiden zwischen *Schätzereigenschaften bei endlichen Stichproben* und *Asymptotischen Schätzereigenschaften*. Erstere sind Inhalt dieses Abschnittes und betreffen die Eigenschaften eines Schätzer für einen festen Stichprobenumfang n , letztere sind Inhalt von Kapitel 31.3 und betreffen die Eigenschaften eines Schätzers im Grenzfall $n \rightarrow \infty$ von großen Stichprobenumfängen.

Es sei zunächst (Σ, S) ein Messraum und $\hat{\tau} : \mathcal{Y} \rightarrow \Sigma$ ein Schätzer von $\tau : \Theta \rightarrow \Sigma$ (vgl. Definition 30.5). In der Folge betrachten wir neben Parameterschätzern der Form

$$\tau : \Theta \rightarrow \Sigma, \tau(\theta) := \theta \quad (31.27)$$

auch wiederholt zunächst solche Schätzer, die bei parametrischen Produktmodellen nur Funktionen der Parameter wie den Erwartungswert, die Varianz und die Standardabweichung der Stichprobenvariablen schätzen. Da nach Annahme die Verteilungen der Stichprobenvariablen $y_1, \dots, y_n$ identisch sind, handelt es sich dabei um Schätzer der Form

$$\tau : \Theta \rightarrow \Sigma, \theta \mapsto \tau(\theta) \text{ mit } \tau(\theta) := \mathbb{E}_\theta(y_1), \tau(\theta) := \mathbb{V}_\theta(y_1), \text{ und } \tau(\theta) := \mathbb{S}_\theta(y_1). \quad (31.28)$$

Speziell wollen wir in diesem Abschnitt vier Aspekte von Schätzereigenschaften bei endlichen Stichproben beleuchten. In Kapitel 31.2.1 beschäftigen wir uns zunächst mit der *Erwartungstreue* eines Schätzers. Dabei heißt ein Schätzer *erwartungstreu*, wenn sein Erwartungswert mit dem wahren, aber unbekannten, Wert $\tau(\theta)$ für alle $\theta \in \Theta$ identisch ist. In Kapitel 31.2.2 führen wir mit den Begriffen der *Varianz* und des *Standardsfehlers* eines Schätzers als Bezeichnungen für die Varianz der Zufallsvariable $\hat{\tau}(y)$ und die Standardabweichung der Zufallsvariable $\hat{\tau}(y)$ zwei Maße für die Frequentistische Variabilität von Schätzern ein. Mit dem *mittlere quadratischen Fehler* eines Schätzers $\hat{\tau}$ als Erwartungswert der quadrierten Abweichung von $\hat{\tau}(y)$ von $\tau(\theta)$ führen wir dann in Kapitel 31.2.3 eine Schätzereigenschaft ein, die es erlaubt die Genauigkeit und die Variabilität eines Schätzers im Sinne eines sogenannten *Bias-Variance-Tradeoffs* miteinander in Beziehung zu setzen. Die in Kapitel 31.2.4 diskutierte *Cramér-Rao-Ungleichung* schließlich gibt eine untere Schranke für die Varianz erwartungstreuer Schätzer an. Ein erwartungstreuer Schätzer mit Varianz gleich der in der Cramér-Rao-Ungleichung gegebenen unteren Schranke hat die kleinstmögliche Varianz aller erwartungstreuen Schätzer und ist in diesem Sinne ein optimaler Schätzer.

31.2.1. Erwartungstreue

Der Begriff der Erwartungstreue eines Schätzers ergibt sich im Kontext des *Fehlers* und des *systematischen Fehlers* eines Schätzers wie folgt.

Definition 31.4 (Fehler, Systematischer Fehler und Erwartungstreue). y sei eine Stichprobe eines Frequentischen Inferenzmodells und $\hat{\tau}$ sei ein Schätzer für τ .

- Der *Fehler* von $\hat{\tau}$ ist definiert als

$$\hat{\tau}(y) - \tau(\theta). \quad (31.29)$$

- Der *systematische Fehler* (engl. *Bias*) von $\hat{\tau}$ ist definiert als

$$B(\hat{\tau}) := \mathbb{E}_{\theta}(\hat{\tau}(y)) - \tau(\theta). \quad (31.30)$$

- Der Schätzer $\hat{\tau}$ heißt *erwartungstreu* (engl. *unbiased*), wenn

$$B(\hat{\tau}) = 0 \Leftrightarrow \mathbb{E}_{\theta}(\hat{\tau}(y)) = \tau(\theta) \text{ für alle } \theta \in \Theta \text{ und alle } n \in \mathbb{N}. \quad (31.31)$$

Andernfalls heißt $\hat{\tau}$ *verzerrt* (engl. *biased*).

•

Man beachte, dass in Definition 31.4 der Fehler eines Schätzers von der spezifischen Realisation der Stichprobe y abhängt. Der systematische Fehler dagegen ist der erwartete Fehler über Stichprobenrealisationen und damit im Sinne eines Erwartungswerts von einer spezifischen Realisation unabhängig. Für den speziellen Fall eines Parameterpunktschätzers gilt nach Definition 31.4, dass er erwartungstreu ist, wenn gilt, dass

$$\mathbb{E}_{\theta}(\hat{\theta}(y)) = \theta. \quad (31.32)$$

Als erste Beispiele für erwartungstreue Schätzer betrachten wir in folgendem Theorem das Stichprobenmittel und die Stichprobenvarianz als Schätzer für den Erwartungswert und die Varianz einer Stichprobenvariable.

Theorem 31.3 (Erwartungstreue von Stichprobenmittel und Stichprobenvarianz). $y := (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells. Dann gelten

(1) Das Stichprobenmittel

$$\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i \quad (31.33)$$

ist ein erwartungstreuer Schätzer des Erwartungswerts $\mathbb{E}_\theta(y_1)$.

(2) Die Stichprobenvarianz

$$S^2 := \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (31.34)$$

ist ein erwartungstreuer Schätzer der Varianz $\mathbb{V}_\theta(y_1)$.

◦

Beweis. (1) Die Erwartungstreue des Stichprobenmittels ergibt mit den Eigenschaften des Erwartungswerts (vgl. [?@thm-eigenschaften-des-erwartungswerts](#)) aus

$$\mathbb{E}_\theta(\bar{y}) = \mathbb{E}_\theta\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta(y_i) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta(y_1) = \frac{1}{n} n \mathbb{E}_\theta(y_1) = \mathbb{E}_\theta(y_1). \quad (31.35)$$

(2) Um die Erwartungstreue der Stichprobenvarianz zu zeigen, halten wir zunächst fest, dass mit den Eigenschaften der Varianz gilt, dass (vgl. [?@thm-eigenschaften-der-varianz](#))

$$\mathbb{V}_\theta(\bar{y}) = \mathbb{V}_\theta\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}_\theta(y_i) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}_\theta(y_1) = \frac{1}{n^2} n \mathbb{V}_\theta(y_1) = \frac{\mathbb{V}_\theta(y_1)}{n}. \quad (31.36)$$

Weiterhin gilt für den Term der summierten quadratischen Abweichungen in der Stichprobenvarianz, dass

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1))^2 - n(\bar{y} - \mathbb{E}_\theta(y_1))^2, \quad (31.37)$$

weil

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1) - \bar{y} + \mathbb{E}_\theta(y_1))^2 \\ &= \sum_{i=1}^n ((y_i - \mathbb{E}_\theta(y_1)) - (\bar{y} - \mathbb{E}_\theta(y_1)))^2 \\ &= \sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1))^2 - 2(\bar{y} - \mathbb{E}_\theta(y_1)) \left(\sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1)) \right) + \sum_{i=1}^n (\bar{y} - \mathbb{E}_\theta(y_1))^2 \\ &= \sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1))^2 - 2(\bar{y} - \mathbb{E}_\theta(y_1)) \left(\sum_{i=1}^n y_i - n \mathbb{E}_\theta(y_1) \right) + n(\bar{y} - \mathbb{E}_\theta(y_1))^2 \\ &= \sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1))^2 - 2(\bar{y} - \mathbb{E}_\theta(y_1)) \left(n \left(\frac{1}{n} \sum_{i=1}^n y_i \right) - n \mathbb{E}_\theta(y_1) \right) + n(\bar{y} - \mathbb{E}_\theta(y_1))^2 \\ &= \sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1))^2 - 2n(\bar{y} - \mathbb{E}_\theta(y_1))^2 + n(\bar{y} - \mathbb{E}_\theta(y_1))^2 \\ &= \sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1))^2 - n(\bar{y} - \mathbb{E}_\theta(y_1))^2. \end{aligned} \quad (31.38)$$

Zusammen ergibt sich also

$$\mathbb{E}_\theta((n-1)S^2) = \mathbb{E}_\theta\left(\sum_{i=1}^n (y_i - \bar{y})^2\right) \quad (31.39)$$

$$= \mathbb{E}_\theta\left(\sum_{i=1}^n (y_i - \mathbb{E}_\theta(y_1))^2 - n(\bar{y} - \mathbb{E}_\theta(y_1))^2\right) \quad (31.40)$$

$$= \sum_{i=1}^n \mathbb{E}_\theta((y_i - \mathbb{E}_\theta(y_1))^2) - n\mathbb{E}_\theta((\bar{y} - \mathbb{E}_\theta(y_1))^2) \quad (31.41)$$

$$= n\mathbb{V}_\theta(y_1) - n\mathbb{V}_\theta(\bar{y}) \quad (31.42)$$

$$= n\mathbb{V}_\theta(y_1) - n\frac{\mathbb{V}_\theta(y_1)}{n} \quad (31.43)$$

$$= n\mathbb{V}_\theta(y_1) - \mathbb{V}_\theta(y_1) \quad (31.44)$$

$$= (n-1)\mathbb{V}_\theta(y_1). \quad (31.45)$$

Schließlich ergibt sich dann

$$\mathbb{E}_\theta(S^2) = \mathbb{E}_\theta\left(\frac{1}{n-1}(n-1)S^2\right) = \frac{1}{n-1}\mathbb{E}_\theta((n-1)S^2) = \frac{1}{n-1}(n-1)\mathbb{V}_\theta(y_1) = \mathbb{V}_\theta(y_1) \quad (31.46)$$

und damit die Erwartungstreue der Stichprobenvarianz als Schätzer der Varianz. $\square$

Natürlich sind in Theorem 31.3 aufgrund der identischen Verteilung der Stichprobenvariablen eines parametrischen Produktmodells das Stichprobenmittel und die Stichprobenvarianz auch erwartungstreue Schätzer des Erwartungswertes und der Varianz einer beliebigen Stichprobenvariablen y_i mit $1 \leq i \leq n$. Man beachte, dass im Beweis der Erwartungstreue der Stichprobenvarianz der Nenner $n-1$ in der Definition der Stichprobenvarianz eine entscheidende Rolle spielt.

Obwohl die Stichprobenvarianz ein unverzerrter Schätzer der Varianz einer Stichprobenvariable eines parametrischen Produktmodells ist, trifft dies auf die Stichprobenstandardabweichung als Schätzer der Standardabweichung nicht zu. Dies ist Inhalt des folgenden Theorems.

Theorem 31.4 (Verzerrtheit der Stichprobenstandardabweichung). *$y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells. Dann ist die Stichprobenstandardabweichung*

$$S := \sqrt{S^2} \quad (31.47)$$

ein verzerrter Schätzer der Standardabweichung $\mathbb{S}_\theta(y_1)$.

◦

Beweis. Wir halten zunächst fest, dass $\sqrt{\cdot}$ eine strikt konkave Funktion und $\sigma^2 > 0$ ist. Dann aber gilt mit der Jensenschen Ungleichung $\mathbb{E}(f(\xi)) < f(\mathbb{E}(\xi))$ für strikt konkave Funktionen (vgl. Theorem 26.5), dass

$$\mathbb{E}_\theta(S) = \mathbb{E}_\theta(\sqrt{S^2}) < \sqrt{\mathbb{E}_\theta(S^2)} = \sqrt{\mathbb{V}_\theta(y_1)} = \mathbb{S}_\theta(y_1). \quad (31.48)$$

$\square$

Allgemein führen nichtlineare Transformationen von erwartungstreuen Schätzern oft auf verzerrte Schätzer, was wir hier aber nicht weiter vertiefen wollen. Folgender **R** Code demonstriert exemplarisch die Begriffe der Unverzerrtheit und Verzerrtheit von Stichprobenmittel, Stichprobenvarianz und Stichprobenstandardabweichung am Beispiel

eines parametrischen Produktmodells mit Stichprobenverteilung

$$y_1, \dots, y_{12} \sim N(1.7, 2) \quad (31.49)$$

Dabei werden die Erwartungswerte der Schätzer anhand ihrer Stichprobenmittel über viele Realisierungen von $y_1, \dots, y_{12}$ als Funktion der Anzahl an Realisierungen (Simulationen) geschätzt.

```
# Modellformulierung
set.seed(0)                                # Zufallszahlengenerator
mu     = 1.7                               # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr = 2                                 # wahrer, aber unbekannter, Varianzparameter
n      = 12                                # Stichprobenumfang n
nsim   = 5e4                               # Anzahl der Simulationen
y_bar  = rep(NA, nsim)                     # Stichprobenmittellarray
s_sqr  = rep(NA, nsim)                     # Stichprobenvarianzarray
s      = rep(NA, nsim)                     # Stichprobenstandardabweichungarray

# Simulationsiterationen
for(sim in 1:nsim){

  # Stichprobenrealisation von y_1, ..., y_{12}
  y = rnorm(n, mu, sqrt(sigsqr))

  # Erwartungswert-, Varianz-, Standardabweichungsschätzer
  y_bar[sim] = mean(y)                     # Stichprobenmittel
  s_sqr[sim] = var(y)                       # Stichprobenvarianz
  s[sim]     = sd(y)                       # Stichprobenstandardabweichung
}

# Erwartungswertschätzung
E_hat_y_bar = cumsum(y_bar)/(1:nsim) # \mathbb{E}(\bar{y}) Schätzungen
E_hat_s_sqr = cumsum(s_sqr)/(1:nsim) # \mathbb{E}(S^2) Schätzungen
E_hat_s     = cumsum(s)/(1:nsim)     # \mathbb{E}(S) Schätzungen
```

Abbildung 31.1 visualisiert die Ergebnisse obiger Simulation. Gezeigt sind Schätzungen der Erwartungswerte von Stichprobenmittel, Stichprobenvarianz und Stichprobenstandardabweichung als Funktion der Anzahl an Realisierungen der Stichprobenvariablen $y_1, \dots, y_{12}$ sowie die wahren, aber unbekannten, Werte des Erwartungswerts, der Varianz und der Standardabweichung der y_i mit $1 \leq i \leq 12$. Es fällt auf, dass diese Schätzungen bei geringer Realisierungsanzahl variabler ausfallen. Ab einer Schätzung basierend auf etwa 10000 Realisierungen von $y_1, \dots, y_{12}$ entsprechen die Stichprobenmittel von $\bar{y}$ und S^2 gemäß ihrer Erwartungstreue ihren wahren, aber unbekannten, Werten. Die Stichprobenstandardabweichung dagegen zeigt gemäß ihrer Verzerrtheit auch bei weiter ansteigenden Anzahlen von der Realisierungen von $y_1, \dots, y_{12}$ konstant eine zu niedrige Schätzung der wahren, aber unbekannten, Standardabweichung.

31.2.2. Varianz und Standardfehler

Im vorherigen Abschnitt haben wir den Erwartungswert eines Schätzers betrachtet. In diesem Abschnitt betrachten wir seine Varianz und seine Standardabweichung und führen die mit diesen assoziierten Begriffe ein. Wir nutzen folgende Definition.

Definition 31.5 (Varianz und Standardfehler). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines Frequentistischen Inferenzmodells und $\hat{\tau}$ sei ein Schätzer von τ .

- Die *Varianz* von $\hat{\tau}$ ist definiert als

$$\mathbb{V}_\theta(\hat{\tau}) := \mathbb{E}_\theta((\hat{\tau}(y) - \mathbb{E}_\theta(\hat{\tau}(y)))^2). \quad (31.50)$$

- Der *Standardfehler* von $\hat{\tau}$ ist definiert als

$$\text{SE}(\hat{\tau}) := \sqrt{\mathbb{V}_\theta(\hat{\tau})}. \quad (31.51)$$

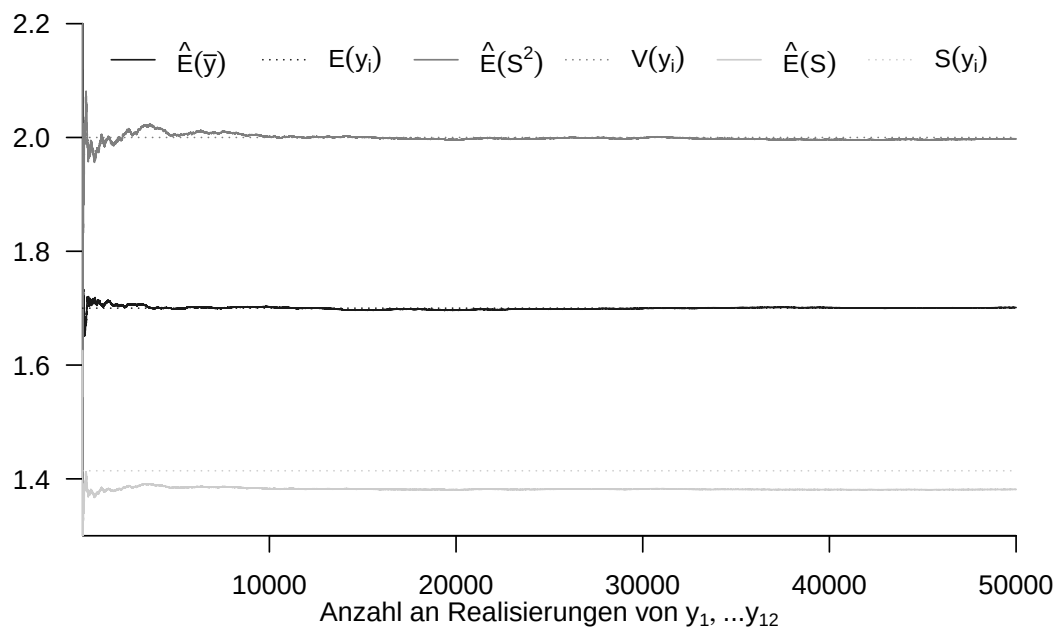


Abbildung 31.1. Simulation der Erwartungstreue von Stichprobenmittel und Stichprobenvarianz als Schätzer des Erwartungswerts und der Varianz bei normalverteilten Stichprobenvariablen und Simulation der Verzerrtheit der Stichprobenstandardabweichung als Schätzer der Standardabweichung bei normalverteilten Stichprobenvariablen

•

Die Varianz eines Schätzers $\hat{\tau}$ ist also als die Varianz der Zufallsvariable $\hat{\tau}(y)$ definiert. Der Standardfehler eines Schätzers $\hat{\tau}$ ist als die Standardabweichung von $\hat{\tau}(y)$ definiert. Als erstes Beispiel für einen Standardfehler betrachten wir den *Standardfehler des Stichprobenmittels*.

Theorem 31.5 (Standardfehler des Stichprobenmittels). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells. Dann ist der Standardfehler des Stichprobenmittels gegeben durch

$$SE(\bar{y}) = \frac{\mathbb{S}_\theta(y_1)}{\sqrt{n}}. \quad (31.52)$$

.

◦

Beweis. Mit der Varianz des Stichprobenmittels ergibt sich

$$SE(\bar{y}) = \sqrt{\mathbb{V}_\theta(\bar{y})} = \sqrt{\frac{\mathbb{V}_\theta(y_1)}{n}} = \frac{\mathbb{S}_\theta(y_1)}{\sqrt{n}}. \quad (31.53)$$

□

Der Standardfehler des Mittelwerts beschreibt die Variabilität des Stichprobenmittels. Da die Standardabweichung $\mathbb{S}_\theta(y_1)$ unbekannt ist, ist auch der Standardfehler $SE(\bar{y})$ unbekannt, kann also nur geschätzt werden. Mit der Stichprobenstandardabweichung als verzerrter Schätzer der Standardabweichung $\mathbb{S}_\theta(y_1)$ ergibt sich ein ebenfalls verzerrter Schätzer für den Standardfehler des Stichprobenmittels zu

$$\hat{SE}(\bar{y}) = \frac{S}{\sqrt{n}}. \quad (31.54)$$

Als zweites Beispiel wollen wir den Standardfehler des Maximum-Likelihood Schätzers für den Parameter eines Bernoulli-Modells betrachten.

Theorem 31.6 (Standardfehler des Maximum-Likelihood Schätzers des Bernoullimodellparameters). *Es sei $y = (y_1, \dots, y_n)$ die Stichproben eines Bernoullimodells und $\hat{\mu}^{ML}$ sei der Maximum-Likelihood Schätzer für den Bernoullimodellparameter μ . Dann ist der Standardfehler von $\hat{\mu}^{ML}$ gegeben durch*

$$SE(\hat{\mu}^{ML}) = \sqrt{\frac{\mu(1-\mu)}{n}}. \quad (31.55)$$

◦

Beweis. Es gilt

$$SE(\hat{\mu}^{ML}) = \sqrt{\mathbb{V}_\mu(\hat{\mu}^{ML})} = \sqrt{\mathbb{V}_\mu\left(\frac{1}{n} \sum_{i=1}^n y_i\right)} = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \mathbb{V}_\mu(y_i)} = \sqrt{\frac{n\mu(1-\mu)}{n^2}} = \sqrt{\frac{\mu(1-\mu)}{n}}, \quad (31.56)$$

wobei die dritte Gleichung mit der Unabhängigkeit der y_i und die vierte Gleichung mit der Varianz $\mathbb{V}_\mu(y_1) = \mathbb{V}_\mu(y_i) = \mu(1-\mu)$ der Stichprobenvariablen folgt. □

Wie im Falle des Standardfehlers des Stichprobenmittels ist auch der Standardfehler des Maximum-Likelihood Schätzers des Bernoullimodellparameters ein wahrer, aber unbekannter, Wert. Ein Schätzer für $SE(\hat{\mu}^{\text{ML}})$ ergibt sich mit dem Maximum-Likelihood Schätzer für den Bernoullimodellparameter durch

$$\hat{SE}(\hat{\mu}^{\text{ML}}) = \sqrt{\frac{\hat{\mu}^{\text{ML}}(1 - \hat{\mu}^{\text{ML}})}{n}}. \quad (31.57)$$

Folgender **R** Code simuliert die Verteilung des Maximum-Likelihood Schätzers für den Parameter eines Bernoullimodells mit wahrem, aber unbekanntem, Parameterwert $\mu := 0.4$ für die Stichprobenumfänge $n = 20, n = 100$ und $n = 200$. Abbildung 31.2 visualisiert die resultierenden Verteilungen mithilfe von Histogrammen. Die Variabilität der Schätzwerte, also die Breite der Histogrammverteilungen, hängt dabei offenbar vom Stichprobenumfang ab und höhere Stichprobenumfänge resultieren in einer geringeren Variabilität des Schätzers. Diesen Gedanken werden wir im Abschnitt Kapitel 31.3 vertiefen.

```
# Modellformulierung
mu      = 0.4                                # wahrer, aber unbekannter, Parameterwert
n_all   = c(20,100,200)                     # Stichprobenumfänge n
ns      = 1e4                                # Anzahl der Simulationen
mu_hat  = matrix(rep(NA, length(n_all)*ns), nrow = length(n_all)) # Maximum-Likelihood Schätzearray

# Stichprobenumfängeiterationen
for(i in seq_along(n_all)){
  # Simulationsiterationen
  for(s in 1:ns){
    y      = rbinom(n_all[i],1,mu)           # Stichprobenrealisation von y_1,...,y_n
    mu_hat[i,s] = mean(y)                   # Stichprobenmittel
  }
}
```

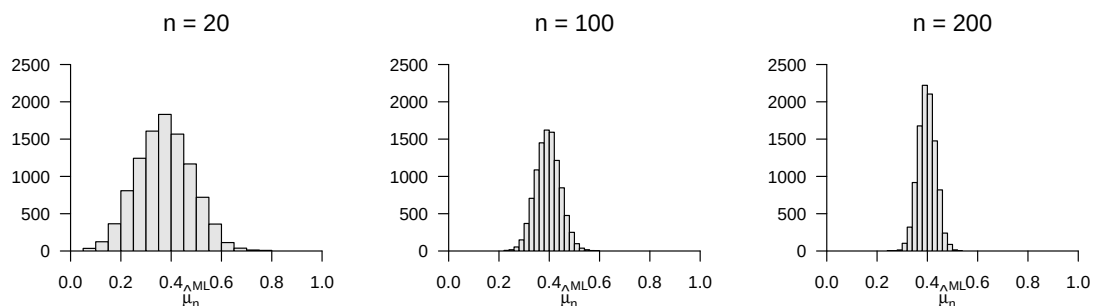


Abbildung 31.2. Simulation der Verteilung des Maximum-Likelihood Schätzers eines Bernoullimodells. Die Variabilität des Schätzers hängt dabei offenbar vom Stichprobenumfang n ab.

31.2.3. Mittlerer quadratischer Fehler

Mit der Erwartungstreue und der Varianz eines Schätzers haben wir in den beiden vorherigen Abschnitten zwei unabhängige Kriterien für die Güte von Schätzern kennengelernt. Der in diesem Abschnitt eingeführte *Mittlere quadratische Fehler* eines Schätzers ermöglicht eine integrierte Betrachtung der Genauigkeit (Erwartungstreue) und Variabilität (Varianz) eines Schätzers im Sinne seiner sogenannten *Bias-Varianz-Zerlegung*. Wir definieren den mittleren quadratischen Fehler eines Schätzers zunächst wie folgt.

Definition 31.6 (Mittlerer quadratischer Fehler). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}$ ein Schätzer für τ . Dann ist der *mittlere quadratische Fehler* (engl. *mean squared error*) von $\hat{\tau}$ definiert als

$$MQF(\hat{\tau}) := \mathbb{E}_\theta((\hat{\tau}(y) - \tau(\theta))^2). \quad (31.58)$$

•

Der mittlere quadratische Fehler von $\hat{\tau}$ ist also die erwartete quadrierte Abweichung von $\hat{\tau}(y)$ von $\tau(\theta)$. Man beachte, dass in Abgrenzung dazu die Varianz von $\hat{\tau}$ die erwartete quadrierte Abweichung von $\hat{\tau}$ von $\mathbb{E}_\theta(\hat{\tau}(y))$ ist. Dabei kann, wie in Kapitel 31.2.1 gesehen $\mathbb{E}_\theta(\hat{\tau}(y))$ mit $\tau(\theta)$ übereinstimmen, ein Schätzer also erwartungstreu sein, er muss es aber nicht. Nutzt man den mittleren quadratischen Fehler als Gütekriterium für einen Schätzer, zum Beispiel indem man versucht, einen Schätzer mit möglichst geringem mittleren quadratischen Fehler zu konstruieren, so kann man dabei eventuelle leichte Abweichungen von der Erwartungstreue zugunsten einer geringen Schätzervarianz in Kauf nehmen. Für den mittleren quadratischen Fehler gilt nämlich folgendes Theorem.

Theorem 31.7 (Zerlegung des mittleren quadratischen Fehlers). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells, $\hat{\tau}$ sei ein Schätzer für τ , und $MQF(\hat{\tau})$ sei der mittlere quadratische Fehler von $\hat{\tau}$. Dann gilt

$$MQF(\hat{\tau}) = B(\hat{\tau})^2 + \mathbb{V}_\theta(\hat{\tau}). \quad (31.59)$$

◦

Beweis. Zur Vereinfachung der Notation seien $\tau := \tau(\theta)$, $\hat{\tau} := \hat{\tau}(y)$ und $\bar{\tau}_n := \mathbb{E}_\theta(\hat{\tau}(y))$. Dann gilt:

$$\begin{aligned} \mathbb{E}_\theta((\hat{\tau} - \tau)^2) &= \mathbb{E}_\theta((\hat{\tau} - \bar{\tau}_n + \bar{\tau}_n - \tau)^2) \\ &= \mathbb{E}_\theta((\hat{\tau} - \bar{\tau}_n)^2 + 2(\hat{\tau} - \bar{\tau}_n)(\bar{\tau}_n - \tau) + (\bar{\tau}_n - \tau)^2) \\ &= \mathbb{E}_\theta((\hat{\tau} - \bar{\tau}_n)^2) + 2\mathbb{E}_\theta((\hat{\tau} - \bar{\tau}_n)(\bar{\tau}_n - \tau)) + \mathbb{E}_\theta((\bar{\tau}_n - \tau)^2) \\ &= \mathbb{E}_\theta((\hat{\tau} - \bar{\tau}_n)^2) + 2\mathbb{E}_\theta(\hat{\tau}\bar{\tau}_n - \hat{\tau}\tau - \bar{\tau}_n\bar{\tau}_n + \bar{\tau}_n\tau) + \mathbb{E}_\theta((\bar{\tau}_n - \tau)^2) \\ &= \mathbb{E}_\theta((\hat{\tau} - \bar{\tau}_n)^2) + 2(\bar{\tau}_n\bar{\tau}_n - \bar{\tau}_n\tau) + \mathbb{E}_\theta((\bar{\tau}_n - \tau)^2) \\ &= \mathbb{E}_\theta((\hat{\tau} - \bar{\tau}_n)^2) + 0 + \mathbb{E}_\theta((\bar{\tau}_n - \tau)^2) \\ &= \mathbb{E}_\theta((\bar{\tau}_n - \tau)^2) + \mathbb{E}_\theta((\hat{\tau} - \bar{\tau}_n)^2) \\ &= \mathbb{E}_\theta((\mathbb{E}_\theta(\hat{\tau}) - \tau)^2) + \mathbb{E}_\theta((\hat{\tau} - \mathbb{E}_\theta(\hat{\tau}))^2) \\ &= (\mathbb{E}_\theta(\hat{\tau}) - \tau)^2 + \mathbb{V}_\theta(\hat{\tau}) \\ &= B(\hat{\tau})^2 + \mathbb{V}_\theta(\hat{\tau}). \end{aligned} \quad (31.60)$$

□

31.2.4. Cramér-Rao-Ungleichung

Hat man mehrere erwartungstreue Schätzer vorliegen, so gilt, dass derjenige Schätzer mit der kleinsten Varianz am verlässlichsten seinen Zweck erfüllt. Weil aber die Stichprobenrealisierungen Frequentistischer Inferenzmodelle in aller Regel variabel sind, kann auch die Variabilität erwartungstreuer Schätzer nicht beliebig klein sein. Die *Cramér-Rao-Ungleichung* gibt eine untere Schranke für die Varianz erwartungstreuer Schätzer an. Ein erwartungstreu Schätzer mit Varianz gleich dieser unteren Schranke hat damit die

kleinstmögliche Varianz aller erwartungstreuer Schätzer und ist - in diesem Sinne - ein optimaler Schätzer.

Die Cramér-Rao-Ungleichung basiert auf dem Begriff der sogenannten *Fisher-Information*, welche wiederum auf dem Begriff der *Scorefunktion* eines Frequentistischen Inferenzmodells beruht. Wir führen im Folgenden also zunächst diese beiden Begrifflichkeiten ein, bevor die Cramér-Rao-Ungleichung formuliert und bewiesen werden soll.

Dabei gelten die vorgestellten Resultate allgemein nur unter einer Reihe mathematischer Annahmen, den sogenannten *Fisher-Regularitätsbedingungen*. Diese bestehen für ein Frequentistisches Inferenzmodell mit WMF oder WDF p_θ und Parameterraum Θ darin, dass angenommen wird, dass (1) Θ eine offene Menge ist, der wahre, aber unbekannte, Parameterwert damit nicht an einer Parameterraumgrenze liegen kann, (2) die Teilmenge von Θ , auf der p_θ von Null verschiedene Werte annimmt, nicht von θ abhängt, (3) das Modell selbst identifizierbar ist, dass also WMFen oder WDFen mit unterschiedliche Parameterwerten unterschiedliche Funktionen sind und damit unterschiedliche Stichprobenverteilungen implizieren, (4) die Likelihood-Funktion des Modells zweimal stetig differenzierbar und (5) dass für die Likelihood-Funktion Integration und Differentiation vertauscht werden dürfen. Wir setzen die Fisher-Regularitätsbedingungen also als erfüllt voraus und wollen nur Modelle mit eindimensionalen Parameterräumen $\Theta \subseteq \mathbb{R}$ betrachten. Wir definieren zunächst die Begriffe der *Scorefunktion* und der *Fisher-Information* wie folgt.

Definition 31.7 (Scorefunktion und Fisher-Information). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells mit eindimensionalem Parameter $\theta \in \Theta \subseteq \mathbb{R}$ und ℓ sei die zugehörige Log-Likelihood-Funktion. Dann gelten:

- Die erste Ableitung von ℓ wird *Scorefunktion der Stichprobe* genannt und wird mit

$$S(\theta) := \frac{d}{d\theta} \ell(\theta) \quad (31.61)$$

bezeichnet. Für $n = 1$ schreiben wir $S(\theta) := S_1(\theta)$ und nennen $S(\theta)$ *Scorefunktion einer Zufallsvariable*.

- Die negative zweite Ableitung von ℓ wird *Fisher-Information der Stichprobe* genannt und mit

$$I(\theta) := -\frac{d^2}{d\theta^2} \ell(\theta) \quad (31.62)$$

bezeichnet. Für $n = 1$ schreiben wir $I(\theta) := I_1(\theta)$ und nennen $I(\theta)$ die *Fisher-Information einer Zufallsvariable*.

•

Da Likelihood- und Log-Likelihood-Funktionen von der Realisierung einer Stichprobe abhängen, sind sie vor dem Hintergrund eines Frequentistischen Inferenzmodells zufällige Funktionen. Da die Fisher-Information als Funktion der Log-Likelihood-Funktion damit auch eine Zufallsvariable ist, muss man zwischen den *beobachteten* und den *erwarteten* Werten der Fisher-Information unterscheiden.

Definition 31.8 (Beobachtete und erwartete Fisher-Information). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells mit eindimensionalem Parameter $\theta \in \Theta \subseteq \mathbb{R}$, ℓ sei die zugehörige Log-Likelihood-Funktion und $\hat{\theta}^{\text{ML}}$ sei ein Maximum-Likelihood-Schätzer von θ . Dann gelten:

- Die *beobachtete Fisher-Information der Stichprobe* ist definiert als

$$I(\hat{\theta}^{\text{ML}}) := -\frac{d^2}{d\theta^2} \ell(\hat{\theta}^{\text{ML}}), \quad (31.63)$$

die beobachtete Fisher-Information der Stichprobe ist also die Fisher-Information an der Stelle des Maximum-Likelihood-Schätzers $\hat{\theta}^{\text{ML}}$.

- Die *erwartete Fisher-Information der Stichprobe* ist definiert als

$$J(\theta) := \mathbb{E}_\theta(I(\theta)). \quad (31.64)$$

Für $n = 1$ schreiben wir $J(\theta) := J_1(\theta)$ und nennen $J(\theta)$ die *erwartete Fisher-Information einer Zufallsvariable*.

•

Bevor wir diese Begrifflichkeiten anhand des Bernoullimodells (Theorem 31.10) und des Normalverteilungsmodells (Theorem 31.12 und Theorem 31.11) verdeutlichen wollen, führen wir mit der *Additivität der Fisher-Information* bei parametrischen Produktmodellen (Theorem 31.8) und dem Erwartungswert und der Varianz der Scorefunktion (Theorem 31.9) noch wichtige Eigenschaften der genannten Begriffe ein, die die folgende Diskussion vereinfachen.

Theorem 31.8 (Additivität der Fisher-Information). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells mit Parameter $\theta \in \Theta \subseteq \mathbb{R}$, ℓ sei die zugehörige Log-Likelihood-Funktion und $I(\theta)$ und $J(\theta)$ seien die Fisher-Information und die erwartete Fisher-Information der Stichprobe, respektive. Dann gilt

$$I(\theta) = nI_1(\theta) \text{ und } J(\theta) = nJ_1(\theta). \quad (31.65)$$

◦

Beweis. Wir zeigen das Resultat für die erwartete Fisher-Information, das Resultat für die beobachtete Fisher-Information gilt dann implizit. Mit der Linearität von Ableitungen und Erwartungswerten gilt

$$\begin{aligned} J(\theta) &= \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \ell(\theta) \right) \\ &= \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \ln \left(\prod_{i=1}^n p_\theta(y_i) \right) \right) \\ &= \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \sum_{i=1}^n \ln p_\theta(y_i) \right) \\ &= \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \sum_{i=1}^n \ln p_\theta(y_1) \right) \\ &= \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \ell_1(\theta) n \right) \\ &= n \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \ell_1(\theta) \right) \\ &= nJ(\theta). \end{aligned} \quad (31.66)$$

□

Nach Theorem 31.8 genügt es zur Berechnung der beobachteten oder erwarteten Fisher-Information einer Stichprobe bei parametrischen Produktmodellen also, die beobachtete oder erwartete Fisher-Information einer der Zufallsvariablen der Stichprobe zu berechnen. Weitere Vereinfachungen in der Bestimmung von Fisher-Informationen und der Begründung der Cramér-Rao-Ungleichung ergeben sich durch die im folgenden Theorem formulierten Identitäten.

Theorem 31.9 (Erwartungswert und Varianz der Scorefunktion). *Der Erwartungswert der Scorefunktion einer Zufallsvariable ist*

$$\mathbb{E}_\theta(S(\theta)) = 0 \quad (31.67)$$

und die Varianz der Scorefunktion einer Zufallsvariable ist

$$\mathbb{V}_\theta(S(\theta)) = J(\theta). \quad (31.68)$$

◦

Beweis. Wir betrachten nur den Fall, dass p_θ eine WDF ist und zeigen zunächst, dass $\mathbb{E}_\theta(S(\theta)) = 0$ ist.

$$\begin{aligned} \mathbb{E}_\theta(S(\theta)) &= \int S(\theta) p_\theta(x) dx \\ &= \int \frac{d}{d\theta} \ell(\theta) p_\theta(x) dx \\ &= \int \frac{d}{d\theta} \ln L(\theta) p_\theta(x) dx \\ &= \int \frac{1}{L(\theta)} \frac{d}{d\theta} L(\theta) p_\theta(x) dx \\ &= \int \frac{1}{p_\theta(x)} \frac{d}{d\theta} L(\theta) p_\theta(x) dx \\ &= \int \frac{d}{d\theta} L(\theta) dx \\ &= \frac{d}{d\theta} \int p_\theta(x) dx \\ &= \frac{d}{d\theta} 1 \\ &= 0. \end{aligned} \quad (31.69)$$

Mit der Definition der Varianz folgt dann sofort, dass $\mathbb{V}_\theta(S(\theta)) = \mathbb{E}_\theta(S(\theta)^2)$ ist. Als nächstes zeigen wir,

dass $J(\theta) = \mathbb{E}_\theta(S(\theta)^2)$ und deshalb $\mathbb{V}_\theta(S(\theta)) = J(\theta)$ ist.

$$\begin{aligned}
 J(\theta) &= \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \ln L(\theta) \right) \\
 &= \mathbb{E}_\theta \left(-\frac{d}{d\theta} \frac{\frac{d}{d\theta} L(\theta)}{L(\theta)} \right) \\
 &= \mathbb{E}_\theta \left(-\frac{\frac{d^2}{d\theta^2} L(\theta) L(\theta) - \frac{d}{d\theta} L(\theta) \frac{d}{d\theta} L(\theta)}{L(\theta) L(\theta)} \right) \\
 &= -\mathbb{E}_\theta \left(\frac{\frac{d^2}{d\theta^2} L(\theta)}{L(\theta)} \right) + \mathbb{E}_\theta \left(\frac{\left(\frac{d}{d\theta} L(\theta) \right)^2}{(L(\theta))^2} \right) \\
 &= -\int \frac{\frac{d^2}{d\theta^2} L(\theta)}{L(\theta)} p_\theta(x) dx + \int \frac{\left(\frac{d}{d\theta} L(\theta) \right)^2}{(L(\theta))^2} p_\theta(x) dx \\
 &= -\frac{d^2}{d\theta^2} \int p_\theta(x) dx + \int \left(\frac{1}{L(\theta)} \frac{d}{d\theta} L(\theta) \right)^2 p_\theta(x) dx \\
 &= -\frac{d^2}{d\theta^2} 1 + \int \left(\frac{d}{d\theta} \ln L(\theta) \right)^2 p_\theta(x) dx \\
 &= \mathbb{E}_\theta(S(\theta)^2).
 \end{aligned} \tag{31.70}$$

□

Der Erwartungswert der Ableitung der Log-Likelihood-Funktion ist also immer Null und die erwartete Fisher-Information ist immer gleich der Varianz der Scorefunktion. Wir wollen die Scorefunktion und die verschiedenen Formen der Fisher-Information nun für die uns vertrauten Frequentistischen Inferenzmodelle konkret berechnen. Nachfolgendes Theorem fasst zunächst die Ergebnisse für das Bernoullimodell zusammen.

Theorem 31.10 (Scorefunktion und Fisher-Informationen des Bernoullimodells). *Es sei $y = (y_1, \dots, y_n)$ die Stichprobe eines Bernoullimodells mit Parameter $\mu \in]0, 1[$. Dann gelten:*

- Die Scorefunktion der Stichprobe ist

$$S :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto S(\mu) := \frac{1}{\mu} \sum_{i=1}^n y_i - \frac{1}{1-\mu} \left(n - \sum_{i=1}^n y_i \right). \tag{31.71}$$

- Die Fisher-Information der Stichprobe ist

$$I :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto I(\mu) := I(\mu) = \frac{ny}{\mu^2} + \frac{n(1-y)^2}{1-\mu}. \tag{31.72}$$

- Die beobachtete Fisher-Information der Stichprobe ist

$$I :]0, 1[\rightarrow \mathbb{R}, \hat{\mu}^{ML} \mapsto I(\hat{\mu}^{ML}) := \frac{ny}{\hat{\mu}_n^{ML^2}} + \frac{n(1-y)}{1-\hat{\mu}^{ML}}. \tag{31.73}$$

- Die erwartete Fisher-Information der Stichprobe ist

$$J :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto J(\mu) := \frac{n}{\mu(1-\mu)}. \tag{31.74}$$

◦

Beweis. Die Scorefunktion wurde bereits im Kontext der Maximum-Likelihood-Schätzung von μ hergeleitet. Wir betrachten die Fisher-Information einer einzelnen Bernoulli-Zufallsvariable y .

$$\begin{aligned}
 I(\mu) &:= -\frac{d^2}{d\mu^2} \ell_1(\mu) \\
 &= -\frac{d^2}{d\mu^2} \ln p_\mu(y) \\
 &= -\frac{d^2}{d\mu^2} (y \ln \mu + (1-y) \ln(1-\mu)) \\
 &= -\frac{\partial}{\partial \mu} \left(\frac{\partial}{\partial \mu} (y \ln \mu + (1-y) \ln(1-\mu)) \right) \\
 &= -\frac{\partial}{\partial \mu} \left(\frac{y}{\mu} + \frac{(1-y)}{1-\mu} \right) \\
 &= -\left(-\frac{y}{\mu^2} - \frac{(1-y)^2}{(1-\mu)^2} \right) \\
 &= \frac{y}{\mu^2} + \frac{(1-y)^2}{(1-\mu)^2}.
 \end{aligned} \tag{31.75}$$

Damit ergibt sich die erwartete Fisher-Information der Zufallsvariable y als

$$\begin{aligned}
 J(\mu) &= \mathbb{E}_\mu(I(\mu)) \\
 &= \mathbb{E}_\mu \left(\frac{y}{\mu^2} + \frac{(1-y)^2}{(1-\mu)^2} \right) \\
 &= \frac{\mathbb{E}_\mu(y)}{\mu^2} + \frac{(1-\mathbb{E}_\mu(y))^2}{(1-\mu)^2} \\
 &= \frac{\mu}{\mu^2} + \frac{(1-\mu)^2}{(1-\mu)^2} \\
 &= \frac{1}{\mu(1-\mu)}.
 \end{aligned} \tag{31.76}$$

Mit der Additivitätseigenschaft der Fisher-Information und der Definition der beobachteten Fisher-Information ergibt sich dann sofort

$$I(\mu) = \frac{ny}{\mu^2} + \frac{n(1-y)^2}{(1-\mu)^2} \quad \text{und} \quad J(\mu) = \frac{n}{\mu(1-\mu)}. \tag{31.77}$$

□

Die Scorefunktion und die Fisher-Informationen des Normalverteilungsmodells betrachten wir lediglich unter der zusätzlichen Annahme eines bekannten Varianzparameters (Theorem 31.11) bzw. eines bekannten Erwartungswertparameters (Theorem 31.12)

Theorem 31.11 (Scorefunktion und Fisher-Informationen des Normalverteilungsmodells bei bekanntem Varianzparameter). *$y = (y_1, \dots, y_n)$ sei die Stichprobe eines Normalverteilungsmodells und der Varianzparameter σ^2 sei als bekannt vorausgesetzt. Dann gelten:*

- Die Scorefunktion der Stichprobe ist

$$S : \mathbb{R} \rightarrow \mathbb{R}, \mu \mapsto S(\mu) := \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu). \tag{31.78}$$

- Die Fisher-Information der Stichprobe ist

$$I : \mathbb{R} \rightarrow \mathbb{R}, \mu \mapsto I(\mu) := \frac{n}{\sigma^2}. \tag{31.79}$$

- Die beobachtete Fisher-Information der Stichprobe ist

$$I(\hat{\mu}_n^{ML}) = \frac{n}{\sigma^2}. \quad (31.80)$$

- Die erwartete Fisher-Information der Stichprobe ist

$$J : \mathbb{R} \rightarrow \mathbb{R}, \mu \mapsto J(\mu) := \frac{n}{\sigma^2}. \quad (31.81)$$

◦

Beweis. Wir erinnern uns, dass die Log-Likelihood-Funktion eines Normalverteilungsmodells bei bekanntem Varianzparameter σ^2 durch

$$\ell : \mathbb{R} \rightarrow \mathbb{R}, \mu \mapsto \ell(\mu) := -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \quad (31.82)$$

gegeben ist. Damit ergibt sich die Scorefunktion als

$$S(\mu) = \frac{\partial}{\partial \mu} \ell(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) \quad (31.83)$$

Die Fisher-Information der Stichprobe ergibt sich als

$$I(\mu) = -\frac{d^2}{d\mu^2} \ell(\mu) = -\frac{\partial}{\partial \mu} S(\mu) = -\frac{1}{\sigma^2} \frac{\partial}{\partial \mu} \left(\sum_{i=1}^n y_i - n\mu \right) = \frac{n}{\sigma^2}. \quad (31.84)$$

Die beobachtete Fisher-Information ist die Fisher-Information an der Stelle des Maximum-Likelihood Schätzes $\hat{\mu}_n^{ML}$. Die erwartete Fisher-Information schließlich ergibt sich als

$$J(\mu) = \mathbb{E}_\mu(I(\mu)) = \mathbb{E}_\mu\left(\frac{n}{\sigma^2}\right) = \frac{n}{\sigma^2}. \quad (31.85)$$

□

Theorem 31.12 (Scorefunktion und Fisher-Informationen des Normalverteilungsmodells bei bekanntem Erwartungswertparameter). *$y = (y_1, \dots, y_n)$ sei die Stichprobe eines Normalverteilungsmodells und der Varianzparameter σ^2 sei als bekannt vorausgesetzt. Dann gelten:*

- Die Scorefunktion der Stichprobe ist gegeben durch

$$S : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto S(\sigma^2) := -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2 \quad (31.86)$$

- Die Fisher-Information der Stichprobe ist gegeben durch

$$I : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto I(\sigma^2) := \frac{1}{\sigma^6} \sum_{i=1}^n (y_i - \mu)^2 - \frac{n}{2\sigma^4} \quad (31.87)$$

- Die beobachtete Fisher-Information der Stichprobe ist gegeben durch

$$I(\hat{\sigma}_n^{2ML}) = \frac{n}{2\hat{\sigma}_{ML}^4} \quad (31.88)$$

- Die erwartete Fisher-Information der Stichprobe ist gegeben durch

$$J : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto J(\sigma^2) := \frac{n}{2\sigma^4}. \quad (31.89)$$

◦

Beweis. Wir erinnern uns, dass die Log-Likelihood-Funktion der Stichprobe eines Normalverteilungsmodells bei bekanntem Erwartungswert-Parameter μ durch

$$\ell : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto \ell(\sigma^2) := -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2. \quad (31.90)$$

gegeben ist. Die Scorefunktion ergibt sich also als

$$S(\sigma^2) = \frac{\partial}{\partial \sigma^2} \ell(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2. \quad (31.91)$$

Die Fisher-Information der Stichprobe ergibt sich als

$$\begin{aligned} I(\sigma^2) &= -\frac{\partial}{\partial \sigma^2} S(\sigma^2) \\ &= -\frac{\partial}{\partial \sigma^2} \left(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2 \right) \\ &= \frac{1}{\sigma^6} \sum_{i=1}^n (y_i - \mu)^2 - \frac{n}{2\sigma^4}. \end{aligned} \quad (31.92)$$

Die beobachtete Fisher-Information ist die Fisher-Information an der Stelle des Maximum-Likelihood Schätzes $\hat{\sigma}_n^{2\text{ML}}$,

$$\begin{aligned} I(\hat{\sigma}_n^{2\text{ML}}) &= \frac{\sum_{i=1}^n (y_i - \mu)^2}{(\hat{\sigma}_n^{2\text{ML}})^3} - \frac{n}{2(\hat{\sigma}_n^{2\text{ML}})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \mu)^2}{\frac{1}{n^3} \left(\sum_{i=1}^n (y_i - \mu)^2 \right)^3} - \frac{n}{2(\hat{\sigma}_n^{2\text{ML}})^2} \\ &= \frac{1}{\frac{1}{n^3} \left(\sum_{i=1}^n (y_i - \mu)^2 \right)^2} - \frac{n}{2(\hat{\sigma}_n^{2\text{ML}})^2} \\ &= \frac{n}{(\hat{\sigma}_n^{2\text{ML}})^2} - \frac{n}{2(\hat{\sigma}_n^{2\text{ML}})^2} \\ &= \frac{n}{2(\hat{\sigma}_n^{2\text{ML}})^2} \\ &= \frac{n}{2\hat{\sigma}_n^{4\text{ML}}}. \end{aligned} \quad (31.93)$$

Die erwartete Fisher-Information ergibt sich schließlich als

$$\begin{aligned} J(\sigma^2) &= \mathbb{E}_{\sigma^2} (I(\sigma^2)) \\ &= \mathbb{E}_{\sigma^2} \left(\frac{1}{\sigma^6} \sum_{i=1}^n (y_i - \mu)^2 - \frac{n}{2\sigma^4} \right) \\ &= \frac{1}{\sigma^6} \sum_{i=1}^n \mathbb{E}_{\sigma^2} ((y_i - \mu)^2) - \frac{n}{2\sigma^4} \\ &= \frac{1}{\sigma^6} \sum_{i=1}^n \sigma^2 - \frac{n}{2\sigma^4} \\ &= \frac{n\sigma^2}{\sigma^6} - \frac{n}{2\sigma^4} \\ &= \frac{n}{\sigma^4} - \frac{n}{2\sigma^4} \\ &= \frac{n}{2\sigma^4}. \end{aligned} \quad (31.94)$$

□

Mit den oben diskutierten Eigenschaften der Scorefunktion können wir nun die Cramér-Rao-Ungleichung formulieren und beweisen.

Theorem 31.13 (Cramér-Rao-Ungleichung). *Gegeben sei ein Frequentistisches Inferenzmodell mit eindimensionalen Parameter $\theta \in \Theta \subseteq \mathbb{R}$, WMF oder WDF p_θ und $\hat{\tau}$ sei ein erwartungstreuer Schätzer von $\tau(\theta)$. Dann gilt*

$$\mathbb{V}_\theta(\hat{\tau}) \geq \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{J(\theta)}. \quad (31.95)$$

Speziell gilt für einen Parameterschätzer mit $\tau(\theta) := \theta$ und somit

$$\hat{\tau} = \hat{\theta} \text{ und } \left(\frac{d}{d\theta}\tau(\theta)\right)^2 = 1, \quad (31.96)$$

dass

$$\mathbb{V}_\theta(\hat{\theta}) \geq \frac{1}{J(\theta)}. \quad (31.97)$$

Die rechte Seite obiger Ungleichungen wird dabei Cramér-Rao-Schranke genannt.

◦

Beweis. Wir halten zunächst fest, dass für die Zufallsvariablen $S(\theta)$ und $\hat{\tau}$ mit der Korrelationsungleichung (Theorem 26.4) und der Identität von $\mathbb{V}_\theta(S(\theta))$ und $J(\theta)$ (Theorem 31.9) gilt, dass

$$\frac{\mathbb{C}_\theta(S(\theta), \hat{\tau})^2}{\mathbb{V}_\theta(S(\theta))\mathbb{V}_\theta(\hat{\tau})} \leq 1 \Leftrightarrow \mathbb{V}_\theta(\hat{\tau}) \geq \frac{\mathbb{C}_\theta(S(\theta), \hat{\tau})^2}{J(\theta)}. \quad (31.98)$$

Mit dem Kovarianzverschiebungssatz (Theorem 25.2), der Tatsache, dass der Erwartungswert der Scorefunktion immer Null ist (Theorem 31.9) und der vorausgesetzten Erwartungstreue von $\hat{\tau}$ ergibt sich dann zunächst

$$\begin{aligned} \mathbb{C}_\theta(S(\theta), \hat{\tau}) &= \mathbb{E}_\theta(S(\theta)\hat{\tau}) - \mathbb{E}_\theta(S(\theta))\mathbb{E}_\theta(\hat{\tau}) \\ &= \mathbb{E}_\theta(S(\theta)\hat{\tau}) \\ &= \int S(\theta) \hat{\tau} p_\theta(x) dx \\ &= \int \frac{d}{d\theta} \ln L(\theta) \hat{\tau} p_\theta(x) dx \\ &= \int \frac{\frac{d}{d\theta} L(\theta)}{L(\theta)} \hat{\tau} p_\theta(x) dx \\ &= \int \frac{\frac{d}{d\theta} L(\theta)}{p_\theta(x)} \hat{\tau} p_\theta(x) dx \\ &= \int \frac{d}{d\theta} L(\theta) \hat{\tau} dx \\ &= \frac{d}{d\theta} \int L(\theta) \hat{\tau} dx \\ &= \frac{d}{d\theta} \int \hat{\tau} p_\theta(x) dx \\ &= \frac{d}{d\theta} \mathbb{E}_\theta(\hat{\tau}) \\ &= \frac{d}{d\theta} \tau(\theta). \end{aligned} \quad (31.99)$$

Damit folgt dann aber direkt

$$\mathbb{V}_\theta(\hat{\tau}) \geq \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{J(\theta)}. \quad (31.100)$$

□

Für Parameterschätzer gilt also insbesondere, dass die Varianz eines erwartungstreuen Parameterschätzers $\hat{\theta}$ immer größer oder gleich der reziproken erwarteten Fisher-Information $J(\theta)$ ist. Im Fall, dass sogar

$$\mathbb{V}_{\theta}(\hat{\theta}) = \frac{1}{J(\theta)} \quad (31.101)$$

ist, ist die Varianz des Parameterschätzers minimal und der Schätzer somit als optimaler Schätzer im Sinne der Cramér-Rao-Ungleichung nachgewiesen. Wir kommen auf diesen Gedanken in Kapitel 31.4 zurück.

31.3. Asymptotische Schätzereigenschaften

In diesem Abschnitt geben wir eine Kurzeinführung in die *Asymptotische Statistik* (Vaart (1998)). Die Asymptotische Statistik ist der Bereich der Frequentistischen Inferenz, der sich mit dem Verhalten von Statistiken und Schätzern bei großen Stichprobenumfängen n beschäftigt. Dabei werden Methoden der Asymptotischen Statistik zum einen benutzt um, wie hier, qualitative Schätzereigenschaften zu studieren und andererseits um Schätzereigenschaften bei großen Stichprobenumfänge approximieren zu können. Da Stichprobenumfänge heutzutage durchaus groß sein können („Big Data“), sind die Methoden der Asymptotischen Statistik also für die Anwendung gut motiviert und dort vielseitig einsetzbar.

In Fortführung von Kapitel 31.2 wollen wir in diesem Abschnitt vier asymptotische Schätzereigenschaften beleuchten. Um zu betonen, dass in diesem Abschnitt die Eigenschaften eines Schätzers $\hat{\tau}$ vom Stichprobenumfang n abhängen, schreiben wir in diesem Abschnitt für einen Schätzer $\hat{\tau}_n$. In Kapitel 31.3.1 betrachten wir die *Asymptotische Erwartungstreue* eines Schätzers. Dabei heißt ein Schätzer $\hat{\tau}_n$ für τ *asymptotisch erwartungstreu*, wenn der Erwartungswert von $\hat{\tau}_n$ für große Stichprobenumfänge $n \rightarrow \infty$ mit dem wahren, aber unbekannten, Wert $\tau(\theta)$ identisch ist. In Kapitel 31.3.2 führen wir den Begriff der *Konsistenz* eines Schätzers ein. Intuitiv heißt ein Schätzer $\hat{\tau}_n$ für τ *konsistent*, wenn für große Stichprobenumfänge $n \rightarrow \infty$ die Wahrscheinlichkeit dafür, dass $\hat{\tau}_n(y)$ vom wahren, aber unbekannten, Wert $\tau(\theta)$ abweicht, beliebig klein wird. Für große Stichprobenumfänge resultieren die Verteilungen von Schätzern oft in Normalverteilungen. In Kapitel 31.3.3 führen wir mit dem Begriff der *Asymptotischen Normalverteilung* eine entsprechende Formalisierung ein. Ein Schätzer $\hat{\tau}_n$ für τ heißt dann *asymptotisch normalverteilt*, wenn für große Stichprobenumfänge $n \rightarrow \infty$, die Verteilung von $\hat{\tau}_n$ durch eine Normalverteilung gegeben ist. In Kapitel 31.3.4 schließlich betrachten wir mit der *Asymptotischen Effizienz* ein Optimalitätskriterium für Schätzer bei gegen unendlich strebenden Stichprobenumfängen mit folgender Bedeutung: ein Schätzer $\hat{\tau}_n$ für τ heißt *asymptotisch effizient*, wenn für große Stichprobenumfänge $n \rightarrow \infty$ die Verteilung von $\hat{\tau}_n$ durch eine Normalverteilung mit Erwartungswertparameter $\tau(\theta)$ und Varianzparameter gleich der Cramér-Rao-Schranke gegeben ist.

31.3.1. Asymptotische Erwartungstreue

Die Asymptotische Erwartungstreue eines Schätzers verfeinert den Begriff des erwartungstreuen Schätzers wie folgt.

Definition 31.9 (Asymptotische Erwartungstreue). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}_n$ sei ein Schätzer für τ . $\hat{\tau}_n$ heißt *asymptotisch erwartungstreu*, wenn gilt, dass

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\theta}(\hat{\tau}_n(y)) = \tau(\theta) \text{ für alle } \theta \in \Theta. \quad (31.102)$$

•

Asymptotisch erwartungstreue Schätzer sind also nur erwartungstreu, wenn der Stichprobenumfang gegen Unendlich geht. Erwartungstreue Schätzer sind immer auch asymptotisch erwartungstreu, da ihre Erwartungstreue vom Stichprobenumfang unabhängig sind. Als Beispiel für einen nur asymptotisch erwartungstreuen Schätzer betrachten wir den Maximum-Likelihood Schätzer des Varianzparameters des Normalverteilungsmodells.

Theorem 31.14 (Asymptotische Erwartungstreue des Varianzparameterschätzers). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines Normalverteilungsmodells mit Varianzparameter σ^2 . Dann ist der Maximum-Likelihood Schätzer von σ^2

$$\hat{\sigma}_n^{2ML} := \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2 \quad (31.103)$$

nicht erwartungstreu, aber asymptotisch erwartungstreu.

◦

Beweis. Mit der Erwartungstreue der Stichprobenvarianz ergibt sich zunächst (vgl. Theorem 31.3)

$$\mathbb{E}_{\mu, \sigma^2}(\hat{\sigma}_n^{2ML}) = \mathbb{E}_{\mu, \sigma^2} \left(\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2 \right) = \frac{1}{n} \mathbb{E}_{\mu, \sigma^2} \left(\sum_{i=1}^n (y_i - \bar{y}_n)^2 \right) = \frac{n-1}{n} \sigma^2. \quad (31.104)$$

Also gilt

$$\mathbb{E}_{\mu, \sigma^2}(\hat{\sigma}_n^{2ML}) \neq \sigma^2 \quad (31.105)$$

und $\hat{\sigma}_n^{2ML}$ kein erwartungstreuer Schätzer von σ^2 . Allerdings gilt auch

$$\frac{n-1}{n} \rightarrow 1 \text{ für } n \rightarrow \infty, \quad (31.106)$$

so dass

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\mu, \sigma^2}(\hat{\sigma}_n^{2ML}) = \lim_{n \rightarrow \infty} \frac{n-1}{n} \sigma^2 = \sigma^2 \lim_{n \rightarrow \infty} \frac{n-1}{n} = \sigma^2 \quad (31.107)$$

gilt und der Maximum-Likelihood Schätzer von σ^2 damit asymptotisch erwartungstreu ist. $\square$

Folgender **R** Code simuliert das Verhalten des Maximum-Likelihood Schätzers für den Varianzparameter eines Normalverteilungsmodells mit wahren, aber unbekannten, Parametern $\mu = 1$ und $\sigma^2 = 2$ in Abhängigkeit des Stichprobenumfangs.

```
# Modellformulierung
mu      = 1                                # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr  = 2                                # wahrer, aber unbekannter, Varianzparameter
n       = seq(1, 100, by = 2)              # Stichprobenumfänge
ns      = 1e3                              # Anzahl Simulation pro Stichprobenumfang
sigsqr_ml = matrix(rep(NA, length(n)*ns), ncol = length(n)) # \hat{\sigma^2}^{ML} Array

# Simulation
for(i in seq_along(n)){
  for(s in 1:ns){
    y      = rnorm(n[i], mu, sqrt(sigsqr))  # Stichprobenrealisation
    sigsq_r_ml[s,i] = ((n[i]-1)/n[i])*var(y) # \hat{\sigma^2}^{ML}
  }
}
E_sigsqr_ml = colMeans(sigsqr_ml)          # Schätzererwartungswertschaetzung
```

Wir visualisieren den basierend auf obigen Simulationen geschätzten Erwartungswert des Schätzers in Abhängigkeit des Stichprobenumfangs in Abbildung 31.3. Für kleine Stichprobenumfänge unterschätzt der Schätzer den wahren, aber unbekannten, Varianzparameter deutlich, für größere Stichprobenumfänge dagegen nicht.

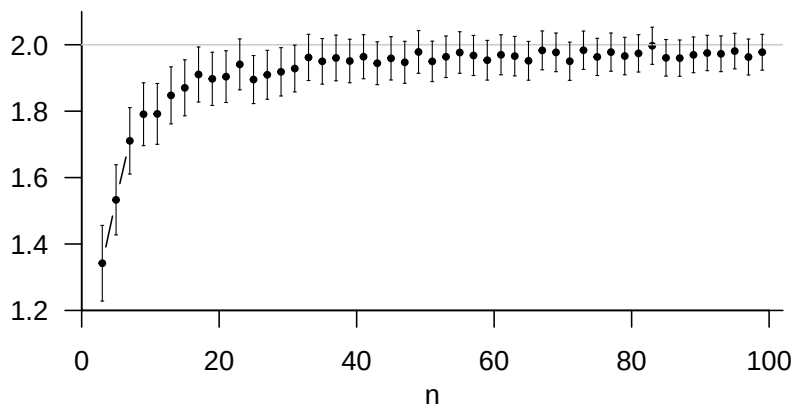


Abbildung 31.3. Simulation des Erwartungswerts des Maximum-Likelihood Schätzers für den Varianzparameter eines Normalverteilungsmodells. Bei kleinen Stichprobenumfängen unterschätzt der Maximum-Likelihood Schätzer den wahren, aber unbekannten, Varianzparameter systematisch, bei größeren Stichprobenumfängen ist der Schätzer in etwa erwartungstreu.

31.3.2. Konsistenz

Der Begriff der Konsistenz eines Schätzers verallgemeinert das Schwache Gesetz der Großen Zahlen von Stichprobenmitteln und Erwartungswerten (vgl. Theorem 27.1) auf beliebige Schätzer und wahre, aber unbekannte, Parameterwerte.

Definition 31.10 (Konsistenz). $y := (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}_n$ sei ein Schätzer von τ . Dann heißt eine Folge von Schätzern $\hat{\tau}_1, \hat{\tau}_2, \dots$ eine *konsistente Folge von Schätzern*, wenn für jedes noch so kleine $\epsilon > 0$ und jedes $\theta \in \Theta$ gilt, dass

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta} (|\hat{\tau}_n(y) - \tau(\theta)| \geq \epsilon) = 0. \quad (31.108)$$

Wenn $\hat{\tau}_1, \hat{\tau}_2, \dots$ eine konsistente Folge von Schätzern ist, dann heißt $\hat{\tau}_n$ ein *konsistenter Schätzer*.

•

Die Intuition zu Definition 31.10 entspricht der dem Gesetz der Großen Zahlen. Für Stichprobenumfänge mit $n \rightarrow \infty$ wird die Wahrscheinlichkeit, dass $\hat{\tau}_n(y)$ beliebig nah bei $\tau(\theta)$ liegt bzw. die Wahrscheinlichkeit, dass $\hat{\tau}_n(y)$ weit von $\tau(\theta)$ abweicht, klein. Allerdings müssen diese Eigenschaften für alle möglichen wahren, aber unbekannten, Parameterwerte gelten. Wie unten gezeigt werden soll ist das Stichprobenmittel ein konsistenter Schätzer für

den Erwartungswertparameter eines Normalverteilungsmodells. Analog zu Kapitel 27.1 demonstrieren wir die Bedeutung der Konsistenz zunächst mithilfe der Simulation für $\mu = 1$ und $\sigma^2 = 2$ anhand folgenden **R** Codes.

```
# Modellformulierung
mu      = 1                                # wahrer, aber unbekannter, Wert von \mu
sigsqr  = 2                                # wahrer, aber unbekannter, Wert von \sigma^2
n       = seq(1,1e3,by = 10)               # Stichprobenumfang n
eps     = c(0.15, 0.10, 0.05)             # \epsilon Werte
ne      = length(eps)                     # Anzahl \epsilon Werte
nn      = length(n)                       # Anzahl Stichprobenumfänge
ns      = 1000                             # Anzahl Simulationen
E       = array(rep(NaN,nn*ne*ns),dim = c(nn,ne,ns)) # Ereignisindikatorarray

# Simulation
for(e in seq_along(eps)){                  # \epsilon Iterationen
  for(i in seq_along(n)){                  # n Iterationen
    for(s in 1:ns){                       # Simulationsiterationen

      # Stichprobenrealisationen
      y = rnorm(n[i], mu, sqrt(sigsqr))
      if(abs(mean(y) - mu) >= eps[e]){      # |y_bar - \mu| \ge \epsilon
        E[i,e,s] = 1
      } else {                             # |y_bar - \mu| < \epsilon
        E[i,e,s] = 0
      }
    }
  }
}

# Schätzung von \mathbb{P}(|\hat{\tau}_n(y) - \tau(\theta)| \ge \epsilon)
P_hat   = apply(E, c(1,2), mean)
```

Wir visualisieren die Schätzung der Wahrscheinlichkeit $\mathbb{P}(|\hat{\tau}_n(y) - \tau(\theta)| \geq \epsilon)$ für dieses Beispiel in Abbildung 31.4 in Abhängigkeit des Stichprobenumfangs und des Kriteriumswerts ϵ . Bei größeren Werten von ϵ genügen geringe Stichprobenumfänge für eine geringe Wahrscheinlichkeit der Abweichung des Schätzers vom wahren, aber unbekannten, Parameterwert, bei kleineren Werten von ϵ sind dazu größere Stichprobenumfänge nötig.

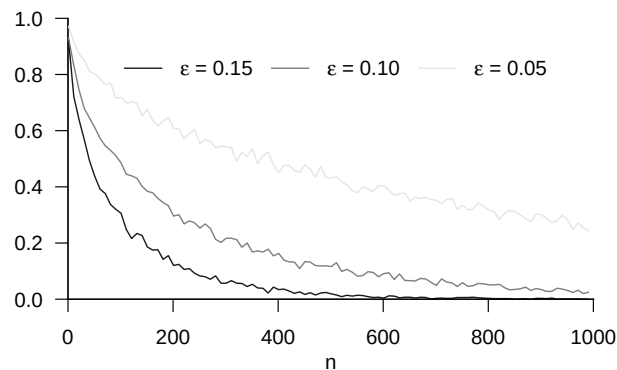


Abbildung 31.4. Simulation der Konsistenz des Stichprobenmittels als Schätzer für den Erwartungswertparameter des Normalverteilungsmodells.

Die in Theorem 31.15 und Theorem 31.16 angegebenen Kriterien für die Konsistenz von Schätzern vereinfachen den Nachweis der Konsistenz eines Schätzers mithilfe des Mittleren Quadratischen Fehlers (Definition 31.6).

Theorem 31.15 (Mittlerer Quadratischer Fehler Kriterium für Konsistenz). $y := (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}_n$ sei

ein Schätzer von τ . Wenn gilt, dass

$$\lim_{n \rightarrow \infty} MQF(\hat{\tau}_n) = 0, \quad (31.109)$$

dann ist $\hat{\tau}_n$ ein konsistenter Schätzer.

◦

Beweis. Mit der Chebychev-Ungleichung gilt, dass

$$\mathbb{P}_\theta(|\hat{\tau}_n(y) - \tau(\theta)| \geq \epsilon) \leq \frac{\mathbb{E}_\theta((\hat{\tau}_n(y) - \tau(\theta))^2)}{\epsilon^2} \quad (31.110)$$

Grenzwertbildung ergibt dann

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(|\hat{\tau}_n(y) - \tau(\theta)| \geq \epsilon) \leq \frac{1}{\epsilon^2} \lim_{n \rightarrow \infty} \mathbb{E}_\theta((\hat{\tau}_n(y) - \tau(\theta))^2). \quad (31.111)$$

Wenn also $\lim_{n \rightarrow \infty} \mathbb{E}_\theta((\hat{\tau}_n(y) - \tau(\theta))^2) = 0$ gilt, dann gilt mit $\mathbb{P}_\theta(|\hat{\tau}_n(y) - \tau(\theta)| \geq \epsilon) \geq 0$, dass

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta(|\hat{\tau}_n(y) - \tau(\theta)| \geq \epsilon) = 0. \quad (31.112)$$

Also ist $\hat{\tau}_n$ ein konsistenter Schätzer.

□

Theorem 31.16 (Bias-Varianz-Kriterium für Konsistenz). $y := (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}_n$ sei ein Schätzer von τ . Wenn

$$\lim_{n \rightarrow \infty} B(\hat{\tau}_n) = 0 \text{ und } \lim_{n \rightarrow \infty} \mathbb{V}_\theta(\hat{\tau}_n) = 0 \quad (31.113)$$

gelten, dann ist $\hat{\tau}_n$ ein konsistenter Schätzer

◦

Beweis. Wenn $n \rightarrow \infty$, dann gilt $B(\hat{\tau}_n) \rightarrow 0$, also auch $B(\hat{\tau}_n)^2 \rightarrow 0$. Wenn für $n \rightarrow \infty$ sowohl $B(\hat{\tau}_n)^2 \rightarrow 0$ als auch $\mathbb{V}_\theta(\hat{\tau}_n) \rightarrow 0$, dann gilt auch $\lim_{n \rightarrow \infty} MQF(\hat{\tau}_n) = 0$. Also gilt mit Theorem 31.15, dass $\hat{\tau}_n$ konsistent ist.

□

Als Anwendung von Theorem 31.16 wollen wir mit folgendem Theorem die Konsistenz des Stichprobenmittels als Schätzer für den Erwartungswert bei Normalverteilung nachweisen.

Theorem 31.17 (Konsistenz des Erwartungswertschätzers bei Normalverteilung). *Es sei y die Stichprobe eines Normalverteilungsmodells. Dann ist $\bar{y}_n$ ein konsistenter Schätzer von $\mathbb{E}(y_1)$.*

◦

Beweis. Mit der Erwartungstreue des Stichprobenmittels als Schätzer für den Erwartungswert gilt zunächst

$$\lim_{n \rightarrow \infty} B(\bar{y}_n) = 0. \quad (31.114)$$

Weiterhin gilt mit der Varianz des Stichprobenmittels

$$\lim_{n \rightarrow \infty} \mathbb{V}_\theta(\bar{y}_n) = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{V}(y_1) = 0. \quad (31.115)$$

Mit dem Bias-Varianz-Kriterium folgt dann die Konsistenz von $\bar{y}_n$ als Schätzer von $\mathbb{E}(y_1)$

□

Man beachte, dass Theorem 31.17 natürlich auch schon im Schwachen Gesetz der Großen Zahlen impliziert ist (vgl. Theorem 27.1).

31.3.3. Asymptotische Normalität

In manchen Fällen nähert sich die Frequentistische Verteilung eines Schätzers bei großen Stichprobenumfängen einer Normalverteilung. Man nennt einen solchen Schätzer dann *asymptotisch normalverteilt*

Definition 31.11 (Asymptotisch normalverteilter Schätzer). $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\theta}_n$ sei ein Parameterschätzer für θ . Weiterhin sei

$$\tilde{\theta} \sim N(\mu, \sigma^2) \quad (31.116)$$

eine normalverteilte Zufallsvariable mit Erwartungswertparameter μ und Varianzparameter σ^2 . Wenn $\hat{\theta}_n$ in Verteilung gegen $\tilde{\theta}$ konvergiert, wenn also für die KVFen P_n und P von $\hat{\theta}_n$ und $\tilde{\theta}$, respektive, gilt, dass

$$\lim_{n \rightarrow \infty} P_n(\hat{\theta}_n) = P(\tilde{\theta}), \quad (31.117)$$

dann heißt $\hat{\theta}_n$ *asymptotisch normalverteilt* und wir schreiben

$$\hat{\theta}_n \overset{a}{\sim} N(\mu, \sigma^2). \quad (31.118)$$

•

Als Beispiel für einen asymptotisch normalverteilten Schätzer betrachten wir in Kapitel 31.3.4 den Maximum-Likelihood-Schätzer des Bernoullimodellparameters. Es ist bemerkenswert, dass dieser Schätzer asymptotisch normalverteilt ist, da die Stichprobenvariablen des Bernoullimodells lediglich die Werte Null und Eins annehmen.

31.3.4. Asymptotische Effizienz

In manchen Fällen lassen sich neben der asymptotischen Normalität eines Schätzers auch in der Form des Frequentistischen Modells begründete Aussagen zum Erwartungswertparameter und Varianzparameter dieser asymptotischen Normalverteilung machen. Der Begriff des *effizienten Schätzers* formuliert einen solchen Spezialfall.

Definition 31.12. $y = (y_1, \dots, y_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\theta}_n$ sei ein Parameterschätzer für θ . Weiterhin sei $J(\theta)$ die erwartete Fisher-Information der Stichprobe y . Wenn gilt, dass

$$\hat{\theta}_n \overset{a}{\sim} N(\theta, J(\theta)^{-1}), \quad (31.119)$$

dann heißt $\hat{\theta}_n$ *asymptotisch effizient*.

•

Ein asymptotisch effizienter Schätzer ist offenbar auch immer asymptotisch normalverteilt und erwartungstreu. Die Varianz der asymptotischen Verteilung eines Schätzers nennt man auch die *asymptotische Varianz*. Für einen asymptotisch effizienten Schätzers ist die asymptotische Varianz mit der Cramér-Rao-Schranke identisch und damit in der Menge der erwartungstreuen asymptotisch normalverteilten Schätzer minimal. Als Beispiel für einen asymptotisch effizienten Schätzer betrachten wir den Maximum-Likelihood-Schätzer des Bernoullimodellparameters. Folgender **R** simuliert die Frequentistische Verteilung dieses Schätzers und bestimmt die WDF seiner asymptotischen Verteilung

```
# Modellformulierung
mu      = 0.4
n_all   = c(1e1, 5e1, 1e2)
ns      = 1e4
mu_hat  = matrix(rep(NaN, length(n_all)*ns), nrow = length(n_all))
mu_hat_r = 1e3
mu_hat_y = seq(0, 1, len = mu_hat_r)
mu_hat_p = matrix(rep(NaN, length(n_all)*mu_hat_r), nrow = length(n_all))

# Stichprobenumfangiterationen
for(i in seq_along(n_all)){
  # Simulationsiterationen
  for(s in 1:ns){
    y = rbinom(n_all[i], 1, mu)
    mu_hat[i, s] = mean(y)
  }
  mu_hat_p[i, ] = dnorm(mu_hat_y, mu, sqrt(mu*(1-mu)/n_all[i]))
}
```

wahrer, aber unbekannter, Parameterwert
Stichprobenumfang
Anzahl Stichprobenrealisierungen
Maximum-Likelihood-Schätzerarray
Maximum-Likelihood Schätzerraumaufloesung
Maximum-LikelihoodSchätzerraum
Maximum-Likelihood WDF Array

Stichprobenrealisation
Maximum-Likelihood-Schätzerrealisation

WDF der asymptotischen Verteilung

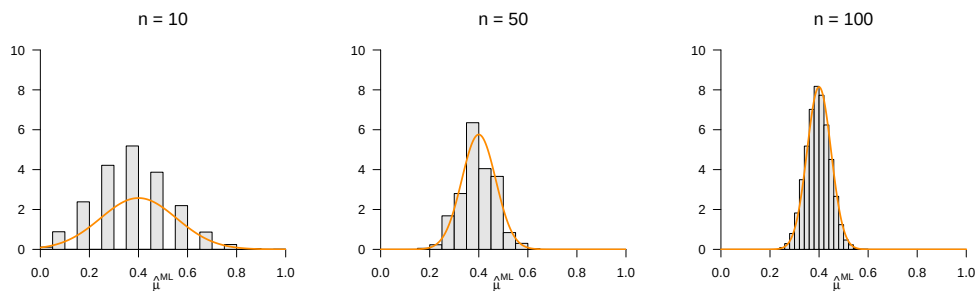


Abbildung 31.5. Simulation der asymptotischen Effizienz des Maximum-Likelihood-Parameterschätzers für den Parameter eines Bernoullimodells. Mit steigendem Stichprobenumfang gleicht sich die durch ein Histogramm dargestellte Verteilung der simulierten Schätzerwerte der in Definition 31.12 formulierten Normalverteilung

31.4. Eigenschaften von Maximum-Likelihood Schätzern

Das Maximum-Likelihood-Prinzip zur Gewinnung von Schätzern für Parameter Frequentistischer Inferenzmodelle ist durch folgendes Theorem zu den Eigenschaften von Maximum-Likelihood-Schätzern begründet.

Theorem 31.18 (Eigenschaften von Maximum-Likelihood Schätzern). *y sei die Stichprobe eines parametrischen Produktmodells und $\hat{\theta}_n^{ML}$ sei ein Maximum-Likelihood Schätzer für θ . Dann gilt, dass $\hat{\theta}_n^{ML}$*

- (1) nicht notwendigerweise erwartungstreu, aber
- (2) asymptotisch erwartungstreu,
- (3) konsistent,

- (4) *asymptotisch normalverteilt und*
- (5) *asymptotisch effizient*

ist.

◦

Für einen nicht unaufwändigen Beweis von Theorem 31.18 verweisen wir auf Held & Sabanés Bové (2014), Abschnitt 3.4. Nutzt man also das Maximum-Likelihood Prinzip zur Gewinnung eines Schätzers, so erfüllt der gewonnene Schätzer also garantiert die in Theorem 31.18 Schätzer Gütekriterien.

31.5. Literaturhinweise

Die in diesem Kapitel vorgestellten Resultate gehen in ganz wesentlicher Weise auf Fisher (1922) zurück. Aldrich (1997) gibt dazu eine historische Einordnung.

32. Konfidenzintervalle

Konfidenzintervalle stellen eine Intervallschätzung wahrer, aber unbekannter, Parameter dar, die so konstruiert ist, dass sie in den meisten Fällen zutrifft. Dabei wird die gegenüber der Punktschätzung gewonnene Sicherheit hinsichtlich der Akkuratheit der Schätzung durch einen Verlust der Genauigkeit der Schätzung erkauft. Wo im Bereich der Punktschätzung zum Beispiel die Schätzung eines wahren, aber unbekannten, Parameterwertes von $\theta = 2$ durch einen Punktschätzer der Form $\hat{\theta}$ zwar sehr genau erfolgt, z.B. durch $\hat{\theta} = 2.14$. Jedoch ist diese Schätzung vor dem Hintergrund der verschwundenen Wahrscheinlichkeit einer kontinuierlichen Zufallsvariable genau *einen* reellen Wert anzunehmen mit hoher Sicherheit falsch. Durch eine Intervallschätzung des wahren, aber unbekannten Parameters durch z.B. $[1.94, 2.34]$ ist eine gröbere Schätzung gegeben, die jedoch so konstruiert werden kann, dass sie mit einer hohen gewünschten Wahrscheinlichkeit zutrifft. Um diese Intuition formal darzustellen, fokussieren wir in diesem Kapitel zunächst auf eindimensionale Parameterräume, also $\Theta \subseteq \mathbb{R}$ und damit in der Tat nur Konfidenzintervalle, d.h. Teilmengen von $\mathbb{R}$. Eine Generalisierung der hier vorgestellten Konzepte auf höher dimensionale Parameterräume im Sinne von Konfidenzmengen ist jedoch relativ unproblematisch möglich.

32.1. Definition

Definition 32.1 (δ -Konfidenzintervall). Es sei y die Stichprobe eines Frequentistischen Inferenzmodells mit wahrem, aber unbekannten Parameter, $\theta \in \Theta$, es sei $\delta \in]0, 1[$ und es seien $G_u(y)$ und $G_o(y)$. Dann heißt ein Intervall der Form

$$\kappa(y) := [G_u(y), G_o(y)], \quad (32.1)$$

so dass

$$\mathbb{P}_\theta(\kappa(y) \ni \theta) = \mathbb{P}_\theta(G_u(y) \leq \theta \leq G_o(y)) = \delta \text{ für alle } \theta \in \Theta \text{ gilt} \quad (32.2)$$

ein δ -Konfidenzintervall für θ . δ ist die Überdeckungswahrscheinlichkeit von $\kappa(y)$ für θ und wird meist *Konfidenzlevel* genannt. Die Statistiken $G_u(y)$ und $G_o(y)$ heißen die unteren und oberen Grenzen des Konfidenzintervalls, respektive.

•

Man beachte in Definition 32.1 dass, wie in allen Frequentistischen Inferenzmodellen, der Parameter θ ein wahrer, aber unbekannter, Wert und damit insbesondere fest, nicht zufällig, ist. Weil die oberen und unteren Grenzen eines Konfidenzintervalls als Funktionen der zufälligen Stichprobe Zufallsvariablen sind, ist das durch sie definierte Konfidenzintervall ein zufälliges Intervall. Die etwas ungewöhnliche Schreibweise $\kappa(y) \ni \theta$ bedeutet schlicht $\theta \in \kappa(y)$. Da $\kappa(y)$ in dem Ausdruck $\mathbb{P}_\theta(\kappa(y) \ni \theta)$ wie beschrieben die zufällige Entität ist, steht $\kappa(y)$ konventionellerweise links, man denke zum Beispiel an einen Ausdruck wie $\mathbb{P}(\xi = x)$. Ein δ -Konfidenzintervall überdeckt den wahren, aber unbekannten,

Parameter θ nach Definition mit Wahrscheinlichkeit δ . Dabei wird oft eine hohe Überdeckungswahrscheinlichkeit von $\delta := 0.95$ gewählt, in diesem Fall spricht man von einem *95%-Konfidenzintervall*.

Intuitiv mag man δ -Konfidenzintervalle auf zwei Arten interpretieren. Im ersten Fall geht man von der Wiederholung der unabhängigen Realisierung von Stichproben bei immer identischen wahren, aber unbekannten, Parameter θ aus. Wiederholt man die Realisierung von Daten also “unter immer den gleichen Umständen” und bei identischen wahren, aber, unbekannten, Parameter θ , so überdeckt ein δ -Konfidenzintervall diesen wahren, aber unbekannten, Parameter im langfristigen Mittel in $\delta \cdot 100\%$ der realisierten Fälle. Alternativ gilt diese Frequentistische Wahrscheinlichkeit für die Überdeckung des wahren, aber unbekannten, Parameters nach Definition 32.1 jedoch auch für jeden beliebigen wahren, aber unbekannten, Parameterwert $\theta_i, i = 1, 2, \dots$. Auch wenn man also unterschiedliche, wahre, aber unbekannte, Parameterwerte $\theta_1, \theta_2, \dots$ betrachtet und in jedem Fall eine, von den anderen Realisierungen unabhängige, Realisierung der Stichprobe erfasst, so überdecken die entsprechenden δ -Konfidenzintervalle diese wahren, aber unbekannten, Parameter im langfristigen Mittel in $\delta \cdot 100\%$ der Fälle. Intuitiv braucht man also “eine Studie”, also die Untersuchung eines wahren, aber unbekannten, Parameterwerts, nicht unter den gleichen Umständen “unendlich oft wiederholen”, um von der Überdeckungswahrscheinlichkeit eines Konfidenzintervalls zu profitieren, sondern es genügt in “unterschiedlichen Studien”, also den Untersuchungen unterschiedlicher wahrer, aber unbekannten, Parameter, Konfidenzintervalle gemäß Definition 32.1 zu bestimmen, auch in diesen Fall ist ihre Überdeckungswahrscheinlichkeit für die wahren, aber unbekannten Parameter, gesichert. Wir demonstrieren diese beiden Interpretationen in der Folge mithilfe einer Simulation.

Um nun für gegeben Frequentistische Inferenzmodelle δ -Konfidenzintervalle durch eine konkrete Angabe der Statistiken $G_u(y)$ und $G_o(y)$ zu konstruieren, geht man vor wie folgt. Zunächst definiert man das Frequentistische Inferenzmodell und legt damit die Verteilung der Stichprobe y fest. In einem zweiten Schritt definiert man eine Statistik, also eine Funktion der Stichprobe, die als Grundlage für $G_u(y)$ und $G_o(y)$ dienen mag und analysiert ihre, auf der Stichprobenverteilung basierende, Verteilung. Hat man die entsprechende Verteilung gefunden, so kann man diese dazu nutzen, die Überdeckungswahrscheinlichkeit des wahren, aber unbekannten, Parameters durch ein entsprechend definiertes Konfidenzintervall zu sichern. Wir zeichnen dieses Verfahren in der Entwicklung und den konstruktiven Beweisen der folgenden Beispiele nach.

32.2. Beispiele für Konfidenzintervalle

Konfidenzintervall für den Erwartungswertparameter des Normalverteilungsmodells

Wir betrachten die Konstruktion eines δ -Konfidenzintervalls für den Erwartungswertparameter des Normalverteilungsmodells. Zu diesem Zweck definieren wir zunächst folgende Konfidenzintervallstatistik.

Definition 32.2 (*T-Konfidenzintervallstatistik*). Gegeben sei das Normalverteilungsmodell

$$y_1, \dots, y_n \sim N(\mu, \sigma^2) \quad (32.3)$$

Dann heißt die mit dem Stichprobenmittel und der Stichprobenstandardabweichung

$$\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i \text{ und } S := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad (32.4)$$

definierte Statistik

$$T := \sqrt{n} \frac{\bar{y} - \mu}{S} \quad (32.5)$$

T-Konfidenzintervallstatistik.

•

Für die Verteilung der T -Konfidenzintervallstatistik gilt folgendes Theorem.

Theorem 32.1 (Verteilung der T -Konfidenzintervallstatistik). *Die T -Konfidenzintervallstatistik ist eine t -verteilte Zufallsvariable mit Parameter $n - 1$, es gilt also*

$$T \sim t(n - 1) \quad (32.6)$$

◦

Man beachte, dass die T -Konfidenzintervallstatistik nach Definition 32.2 eine Funktion der Stichprobe ist, während ihre Verteilung nach Theorem 32.1 unabhängig von den wahren, aber unbekannten, Parametern der Stichprobenverteilung ist. Man nennt dies auch die *Pivoteigenschaft* der T -Konfidenzeigenschaft. Für die folgenden Entwicklungen erinnern wir daran, dass wir die WDF einer t -verteilten Zufallsvariable mit t , die KVF einer t -verteilten Zufallsvariable mit Ψ und die inverse KVF einer t -verteilten Zufallsvariable mit Ψ^{-1} bezeichnen. Folgender **R** Code simuliert zunächst die Verteilung der T -Konfidenzintervallstatistik.

```
# Modellformulierung
mu      = 10                                # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr  = 4                                # wahrer, aber unbekannter, Varianzparameter
n       = 12                                # Stichprobenumfang
ns      = 1e4                               # Anzahl Stichprobenrealisierungen
res     = 1e3                               # Ausgangsraumauflösung

# analytische Definitionen und Resultate
yx      = seq(3,17,len = res)              # y_i Raum
ssqr    = seq(0,20,len = res)              # S^2 Raum
tx      = seq(-4,4,len = res)              # T Raum
p_y_i   = dnorm(yx,mu,sqrt(sigsqr))         # y_i WDF
p_y_bar = dnorm(yx,mu,sqrt(sigsqr/n))       # y_bar WDF
p_sqr   = dchisq(ssqr,n-1)                 # S^2 WDF
p_t     = dt(tx,n-1)                       # T WDF

# Simulation
y_i     = rep(NA,ns)                       # y_i Array
y_bar   = rep(NA,ns)                       # \bar{y} Array
S       = rep(NA,ns)                       # S Array
TKS     = rep(NA,ns)                       # T-Konfidenzintervallstatistik Array
for(s in 1:ns){
  y      = rnorm(n,mu,sqrt(sigsqr))         # Stichprobenrealisierung
  y_i[s] = y[1]                            # Stichprobenrealisierung y_i mit i = 1
  y_bar[s] = mean(y)                       # Stichprobenmittelrealisierung
  S[s]   = sd(y)                           # Stichprobenstandardabweichungrealisierung
  TKS[s] = sqrt(n)*((y_bar[s]-mu)/S[s])     # T-Konfidenzintervallstatistikrealisierung
}
```

In Abbildung 32.1 visualisieren wir die Verteilung der T -Konfidenzintervallstatistik als Resultat der ihr zugrundeliegenden Verteilungen der Stichprobenvariablen, des Stichprobenmittels und der Stichprobenvarianz (vgl. ?@sec-t-transformation). Mithilfe der Verteilung der T -Konfidenzintervallstatistik können wir jetzt folgendes Theorem zum Konfidenzintervall für den Erwartungswertparameter des Normalverteilungsmodells beweisen.

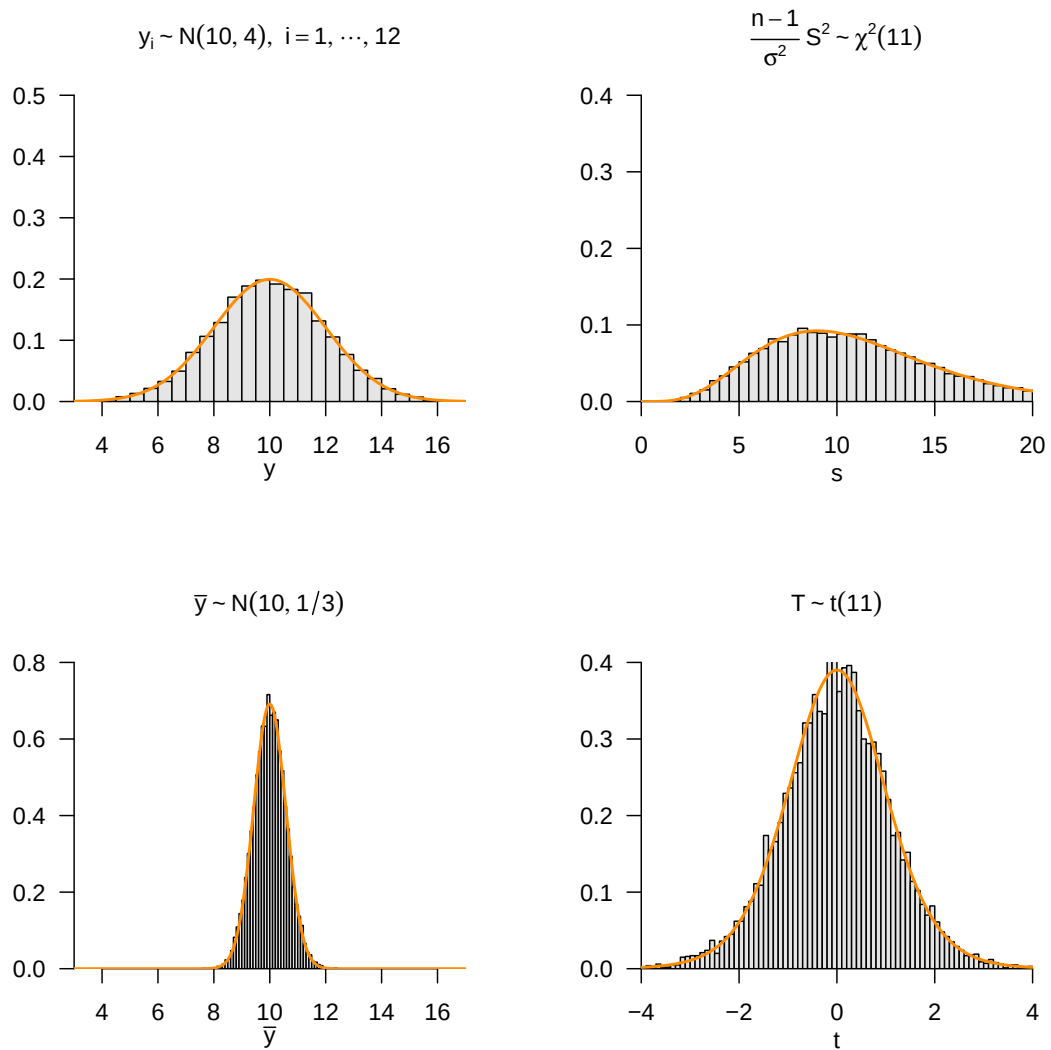


Abbildung 32.1. Simulation der Verteilung der T -Konfidenzintervallstatistik und der ihr zugrundeliegenden Verteilungen der Stichprobenvariable, des Stichprobenmittels und der Stichprobenvarianz.

Theorem 32.2 (Konfidenzintervall für den Erwartungswertparameter des Normalverteilungsmodells). *Gegeben sei das Normalverteilungsmodell*

$$y_1, \dots, y_n \sim N(\mu, \sigma^2) \quad (32.7)$$

mit wahren, aber unbekannten, Parametern μ und σ^2 , es sei $\delta \in]0, 1[$ und es sei

$$t_\delta := \Psi^{-1} \left(\frac{1+\delta}{2}; n-1 \right). \quad (32.8)$$

mit der inversen KVF Ψ^{-1} einer t -verteilten Zufallsvariable. Dann gilt für das Intervall

$$\kappa(y) := \left[\bar{y} - \frac{S}{\sqrt{n}} t_\delta, \bar{y} + \frac{S}{\sqrt{n}} t_\delta \right], \quad (32.9)$$

mit dem Stichprobenmittel und der Stichprobenstandardabweichung

$$\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i \text{ und } S := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad (32.10)$$

respektive, dass

$$\mathbb{P}_\mu(\kappa(y) \ni \mu) = \delta. \quad (32.11)$$

◦

Beweis. Für $\delta \in]0, 1[$ seien zunächst

$$t_1 := \Psi^{-1} \left(\frac{1-\delta}{2}; n-1 \right) \text{ und } t_2 := \Psi^{-1} \left(\frac{1+\delta}{2}; n-1 \right) \quad (32.12)$$

definiert. Dann gilt

$$\frac{1+\delta}{2} - \frac{1-\delta}{2} = \delta \quad (32.13)$$

und weiterhin gilt mit der Symmetrie der WDF der t -Verteilung, dass

$$t_1 = -t_2. \quad (32.14)$$

Per Definition gilt dann aber mit Definition 32.2 und Theorem 32.1, dass

$$\mathbb{P}_\mu(-t_\delta \leq T \leq t_\delta) = \delta. \quad (32.15)$$

Damit folgt dann aber direkt

$$\begin{aligned} \delta &= \mathbb{P}_\mu(-t_\delta \leq T \leq t_\delta) \\ &= \mathbb{P}_\mu \left(-t_\delta \leq \frac{\sqrt{n}}{S} (\bar{y} - \mu) \leq t_\delta \right) \\ &= \mathbb{P}_\mu \left(-\frac{S}{\sqrt{n}} t_\delta \leq \bar{y} - \mu \leq \frac{S}{\sqrt{n}} t_\delta \right) \\ &= \mathbb{P}_\mu \left(-\bar{y} - \frac{S}{\sqrt{n}} t_\delta \leq -\mu \leq -\bar{y} + \frac{S}{\sqrt{n}} t_\delta \right) \\ &= \mathbb{P}_\mu \left(\bar{y} + \frac{S}{\sqrt{n}} t_\delta \geq \mu \geq \bar{y} - \frac{S}{\sqrt{n}} t_\delta \right) \\ &= \mathbb{P}_\mu \left(\bar{y} - \frac{S}{\sqrt{n}} t_\delta \leq \mu \leq \bar{y} + \frac{S}{\sqrt{n}} t_\delta \right) \\ &= \mathbb{P}_\mu \left(\left[\bar{y} - \frac{S}{\sqrt{n}} t_\delta, \bar{y} + \frac{S}{\sqrt{n}} t_\delta \right] \ni \mu \right) \\ &= \mathbb{P}_\mu(\kappa(y) \ni \mu). \end{aligned} \quad (32.16)$$

□

Der entscheidende Schritt zur Sicherung der Überdeckungswahrscheinlichkeit δ des wahren, aber unbekannten, Erwartungswertparameters durch das in Theorem 32.2 definierte Konfidenzintervall ist die Definition von

$$t_\delta := \Psi^{-1}\left(\frac{1+\delta}{2}; n-1\right). \quad (32.17)$$

Wie im Beweis von Theorem 32.2 nachgezeichnet ist die Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den wahren, aber unbekannten, Erwartungswertparameter äquivalent zu der Tatsache, dass bei Wahl eben dieses t_δ die T -Konfidenzintervallstatistik eine Wahrscheinlichkeit von δ dafür hat, einen Wert im Intervall $[-t_\delta, t_\delta]$ anzunehmen. Wir visualisieren die Wahl von t_δ für Fall $\delta := 0.95$ und $n := 5$ in Abbildung 32.2. In diesem Fall ergibt sich

$$-t_\delta = \Psi^{-1}(0.025; 4) = -2.57 \text{ und } t_\delta = \Psi^{-1}(0.975; 4) = 2.57. \quad (32.18)$$

Abbildung 32.2 A zeigt diese Wahl aus Perspektive der WDF der T -Konfidenzintervallstatistik. Die von $-t_\delta$ und t_δ eingeschlossene Wahrscheinlichkeitsmasse beträgt nach Konstruktion δ , T nimmt mit einer Wahrscheinlichkeit von δ also einen Wert zwischen $-t_\delta$ und t_δ an. Abbildung 32.2 B zeigt die entsprechende Perspektive der KVF der T -Konfidenzintervallstatistik. Basierend auf der Vorgabe von $\frac{1-\delta}{2}$ und $\frac{1+\delta}{2}$ werden anhand der inversen KVF Ψ^{-1} die entsprechenden Werte für $-t_\delta$ und t_δ bestimmt. Man beachte, dass die hier gegebene Zentralität der Wahrscheinlichkeitsmasse in Definition 32.1 nicht implizit ist, sondern sich aus den Gegebenheiten der Verteilung der T -Konfidenzintervallstatistik, insbesondere ihrer Symmetrie um 0, ergibt.

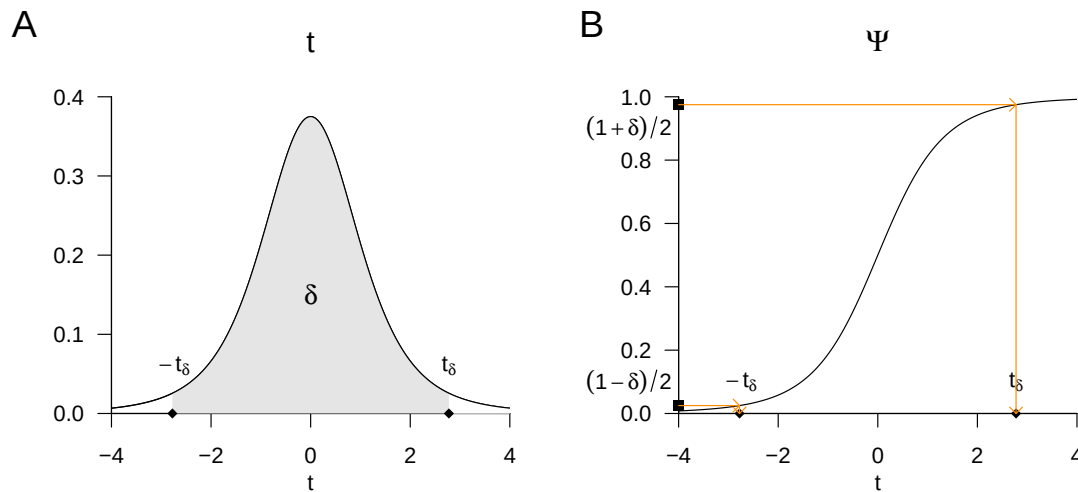


Abbildung 32.2. Sicherung der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Erwartungswertparameter des Normalverteilungsmodells für $\delta := 0.95$ und $n := 5$ aus Perspektive der WDF (A) und der KVF (B) der Verteilung der T -Konfidenzintervallstatistik.

Abschließend wollen wir die Überdeckungswahrscheinlichkeit des durch Theorem 32.2 gegebenen Konfidenzintervalls mithilfe einer Simulation demonstrieren. Wir betrachten dabei lediglich die erste Interpretation eines Konfidenzintervalls bei konstantem, wahren, aber unbekannten, Parameter. Folgender **R** Code bestimmt in diesem Sinne zu jeder Stichprobenrealisierung das entsprechende Konfidenzintervall.

```
# Modellformulierung
set.seed(1)                                     # Random number generator seed
mu      = 2                                     # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr  = 1                                     # wahrer, aber unbekannter, Varianzparameter
sigma   = sqrt(sigsqr)                         # wahrer, aber unbekannter, Standardabweichungsparameter
n       = 12                                    # Stichprobenumfang
delta   = 0.95                                 # Konfidenzbedingung
t_delta = qt((1+delta)/2,n-1)                 # \Psi^{-1}((\delta + 1)/2, n-1)

# Stichprobenrealisierungen
ns      = 1e2                                  # Anzahl Stichprobenrealisierungen
y_bar   = rep(NA,ns)                          # Stichprobenmittellarray
S       = rep(NA,ns)                          # Standardabweichungsarray
kappa   = matrix(rep(NA,2*ns), ncol = 2)      # Konfidenzintervallarray
for(i in 1:ns){
  y      = rnorm(n,mu,sigma)                  # Stichprobenrealisierung
  y_bar[i] = mean(y)                         # Stichprobenmittel
  S[i]   = sd(y)                             # Stichprobenstandardabweichung
  kappa[i,1] = y_bar[i] - (S[i]/sqrt(n))*t_delta # untere Konfidenzintervallgrenze
  kappa[i,2] = y_bar[i] + (S[i]/sqrt(n))*t_delta # obere Konfidenzintervallgrenze
}
```

Wir visualisieren die Ergebnisse dieser Simulation in Abbildung 32.3.

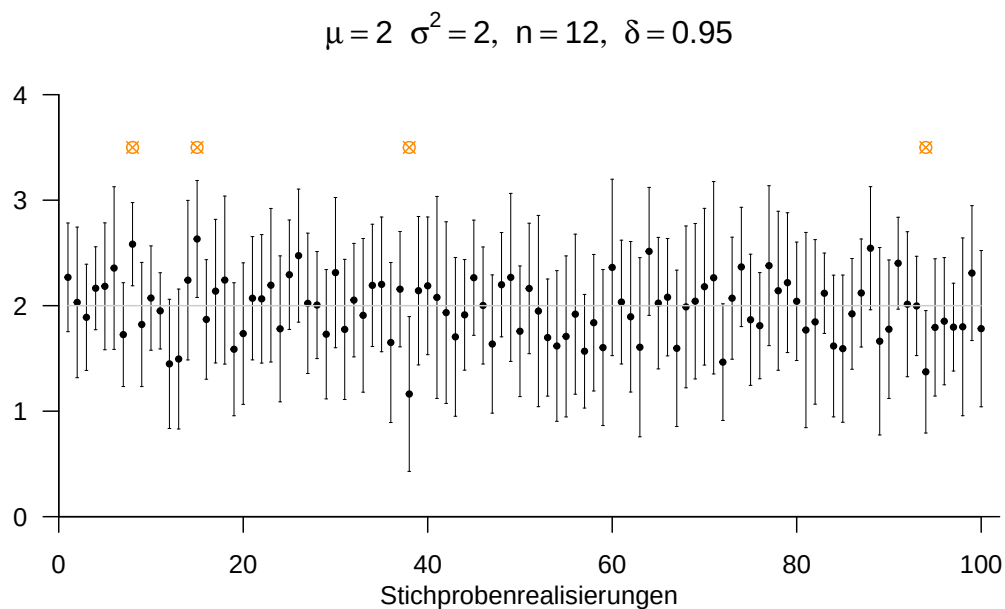


Abbildung 32.3. Simulation der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Erwartungswertparameter des Normalverteilungsmodells bei konstanten, wahren, aber unbekannten, Erwartungswertparameter $\mu := 2$ für $\sigma^2 := 2, n := 12$ und einer gewünschten Überdeckungswahrscheinlichkeit von $\delta := 0.95$. Die Abbildung zeigt für jede Stichprobenrealisierung das Konfidenzintervall und den entsprechenden Erwartungswertparameterschätzer. In der vorliegenden Simulation überdecken die Konfidenzintervalle den durch eine graue Linie eingezeichneten immer gleichen wahren, aber unbekannten, Erwartungswertparameter $\mu := 2$ in 96 von 100 Fällen. Die Stichprobenrealisierungen, für die dies nicht der Fall sind, sind mit einem orangen Kreis markiert

```
# Anzahl Simulationen mit \theta_1, \theta_2,...
set.seed(1)                                     # random number generator seed
ns      = 1e2                                  # Anzahl Simulationen
mu      = 2*seq(0,1,len = ns)                 # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr  = 1                                     # wahrer, aber unbekannter, Varianzparameter
sigma   = sqrt(sigsqr)                         # wahrer, aber unbekannter, Standardabweichungsparameter
n       = 12                                    # Stichprobenumfang
delta   = 0.95                                 # Konfidenzbedingung
t_delta = qt((1+delta)/2,n-1)                 # \Psi^{-1}((\delta + 1)/2, n-1)

# Simulation
y_bar   = rep(NA,ns)                          # Stichprobenmittellarray
S       = rep(NA,ns)                          # Standardabweichungsarray
kappa   = matrix(rep(NA,2*ns), ncol = 2)      # Konfidenzintervallarray
```



```

for(i in 1:ns){
  y      = rnorm(n,mu[i],sigma)      # Stichprobenrealisierung
  y_bar[i] = mean(y)                 # Stichprobenmittel
  S[i]    = sd(y)                    # Stichprobenstandardabweichung
  kappa[i,1] = y_bar[i] - (S[i]/sqrt(n))*t_delta # untere Konfidenzintervallgrenze
  kappa[i,2] = y_bar[i] + (S[i]/sqrt(n))*t_delta # obere Konfidenzintervallgrenze
}

```

Wir visualisieren die Ergebnisse dieser Simulation in Abbildung 32.4.

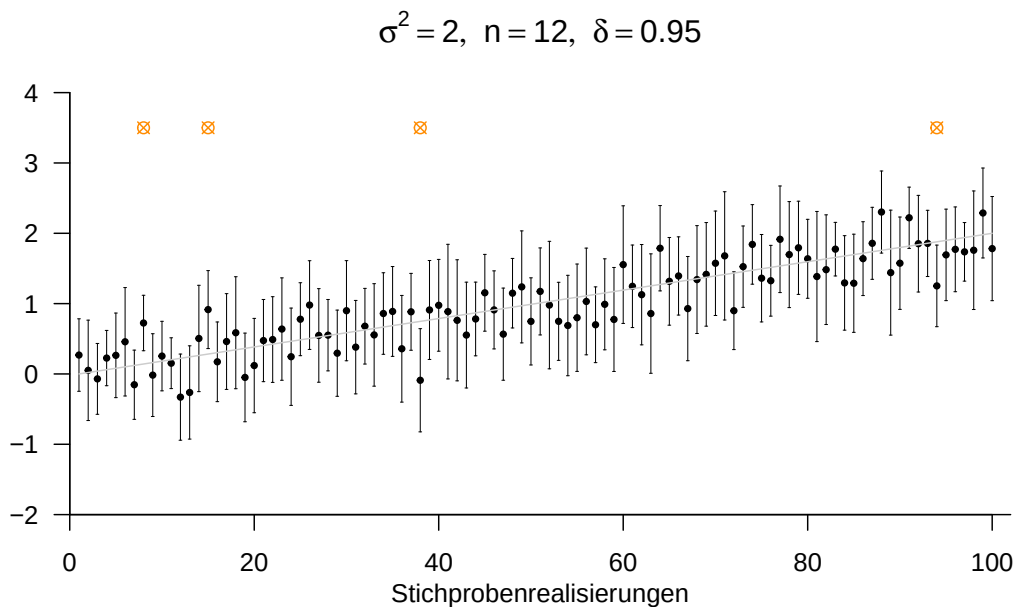


Abbildung 32.4. Simulation der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Erwartungswertparameter des Normalverteilungsmodells bei variablem, wahren, aber unbekannten, Erwartungswertparameter μ für $\sigma^2 := 2, n := 12$ und einer gewünschten Überdeckungswahrscheinlichkeit von $\delta := 0.95$. Die Abbildung zeigt für jede Stichprobenrealisierung das Konfidenzintervall und den entsprechenden Erwartungswertparameterschätzer. In der vorliegenden Simulation überdecken die Konfidenzintervalle den durch eine graue Linie eingezeichneten variablen wahren, aber unbekannten, Erwartungswertparameter μ in 95 von 100 Fällen. Die Stichprobenrealisierungen, für die dies nicht der Fall ist, sind mit einem orangen Kreis markiert

Konfidenzintervall für den Varianzparameter des Normalverteilungsmodells

Wir betrachten die Konstruktion eines δ -Konfidenzintervalls für den Varianzparameter des Normalverteilungsmodells. Zu diesem Zweck definieren zunächst folgende Konfidenzintervallstatistik.

Definition 32.3 (*U-Konfidenzintervallstatistik*). Gegeben sei das Normalverteilungsmodell

$$y_1, \dots, y_n \sim N(\mu, \sigma^2) \quad (32.19)$$

Dann heißt die mit der Stichprobenvarianz

$$\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (32.20)$$

definierte Statistik

$$U := \frac{n-1}{\sigma^2} S^2 \quad (32.21)$$

U-Konfidenzintervallstatistik.

•

Für die Verteilung der *U*-Konfidenzintervallstatistik gilt folgendes Theorem.

Theorem 32.3 (Verteilung der *U*-Konfidenzintervallstatistik). *Die U-Konfidenzintervallstatistik ist eine χ^2 -verteilte Zufallsvariable mit Parameter $n - 1$, es gilt also*

$$U \sim \chi^2(n - 1) \quad (32.22)$$

◦

Für einen Beweis von Theorem 32.3 verweisen wir auf Casella & Berger (2002). Wie die *T*-Konfidenzintervallstatistik besitzt auch die *U*-Konfidenzintervallstatistik die Pivoteigenschaft, da sie eine Funktion der Stichprobe ist, aber ihre Verteilung nach Theorem 32.3 von den wahren, aber unbekannten, Verteilungsparametern der Stichprobe nicht abhängt. Für die folgenden Entwicklungen erinnern wir daran, dass wir die WDF einer χ^2 -verteilten Zufallsvariable mit χ^2 , die KVF einer χ^2 -verteilten Zufallsvariable mit Ξ und die inverse KVF einer χ^2 -verteilten Zufallsvariable mit Ξ^{-1} bezeichnen. Folgender **R** Code simuliert zunächst die Verteilung der *U*-Konfidenzintervallstatistik.

```
# Modellformulierung
mu      = 10                # wahrer Erwartungswertparameter
sigsqr  = 4                 # wahrer bekannter Varianzparameter
n        = 12              # Stichprobenumfang
ns       = 1e4             # Anzahl Stichprobenrealisierungen
res      = 1e3             # Ausgangsraumauflösung

# analytische Definitionen und Resultate
yx       = seq(3,17,len = res) # y_i Raum
ux       = seq(0,30,len = res)  # U Raum
p_y_i    = dnorm(yx,mu,sqrt(sigsqr)) # y_i WDF
p_y_bar  = dnorm(yx,mu,sqrt(sigsqr/n)) # y_bar WDF
p_u      = dchisq(ux,n-1)       # U WDF

# Simulation
y_i      = rep(NaN,ns)         # y_i Array
y_bar    = rep(NaN,ns)         # \bar{y} Array
S_sqr    = rep(NaN,ns)         # S^2 Array
UKS      = rep(NaN,ns)         # U-Konfidenzintervallstatistik Array
for(s in 1:ns){
  y       = rnorm(n,mu,sqrt(sigsqr)) # Stichprobenrealisierung
  y_i[s]  = y[1]                 # Stichprobenrealisierung y_i mit i = 1
  y_bar[s] = mean(y)             # Stichprobenmittelrealisierung
  S_sqr[s] = var(y)              # Stichprobenvarianzrealisierung
  UKS[s]  = ((n-1)/sigsqr)*S_sqr[s] # U-Konfidenzintervallstatistikrealisierung
}
```

Mithilfe der Verteilung der *U*-Konfidenzintervallstatistik können wir jetzt folgendes Theorem zum Konfidenzintervall für den Varianzparameter des Normalverteilungsmodells beweisen.

Theorem 32.4 (Konfidenzintervall für den Varianzparameter des Normalverteilungsmodells). *Gegeben sei das Normalverteilungsmodell*

$$y_1, \dots, y_n \sim N(\mu, \sigma^2) \quad (32.23)$$

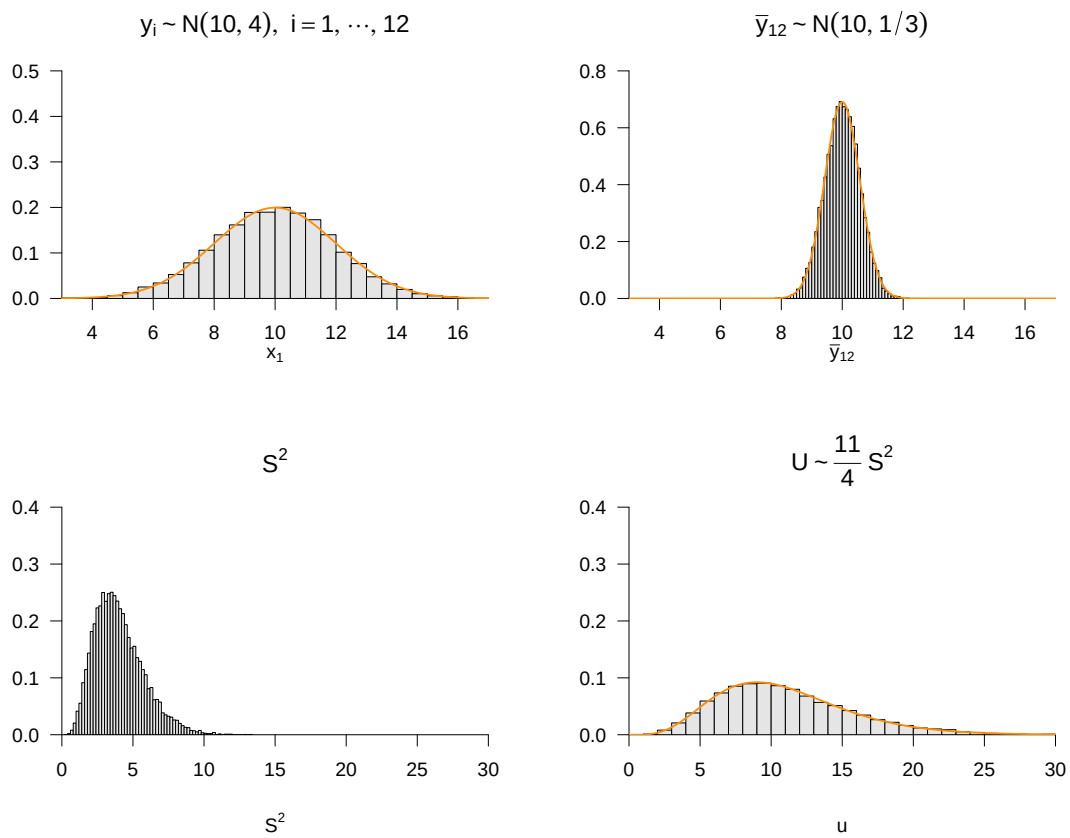


Abbildung 32.5. Simulation der Verteilung der U -Konfidenzintervallstatistik und der ihr zugrundeliegenden Verteilungen der Stichprobenvariable, des Stichprobenmittels und der Stichprobenvarianz.

mit wahren, aber unbekannten, Parametern μ und σ^2 , es sei $\delta \in]0, 1[$ und es seien

$$u_\delta := \Xi^{-1} \left(\frac{1-\delta}{2}; n-1 \right) \text{ und } u'_\delta := \Xi^{-1} \left(\frac{1+\delta}{2}; n-1 \right) \quad (32.24)$$

mit der inversen KVF Ξ^{-1} einer χ^2 -verteilten Zufallsvariable. Dann gilt für das Intervall

$$\kappa(y) := \left[\frac{(n-1)S^2}{u'_\delta}, \frac{(n-1)S^2}{u_\delta} \right]. \quad (32.25)$$

mit der Stichprobenvarianz

$$S^2 := \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2, \quad (32.26)$$

dass

$$\mathbb{P}_{\sigma^2}(\kappa(y) \ni \sigma^2) = \delta. \quad (32.27)$$

◦

Beweis. Per Definition gilt mit Definition 32.3 und Theorem 32.3, dass

$$\mathbb{P}_{\sigma^2}(u_\delta \leq U \leq u'_\delta) = \delta. \quad (32.28)$$

Damit folgt dann aber direkt

$$\begin{aligned} \delta &= \mathbb{P}_{\sigma^2}(u_\delta \leq U \leq u'_\delta) \\ &= \mathbb{P}_{\sigma^2} \left(u_\delta \leq \frac{n-1}{\sigma^2} S^2 \leq u'_\delta \right) \\ &= \mathbb{P}_{\sigma^2} \left(u_\delta^{-1} \geq \frac{\sigma^2}{(n-1)S^2} \geq u'_\delta{}^{-1} \right) \\ &= \mathbb{P}_{\sigma^2} \left(\frac{(n-1)S^2}{u_\delta} \geq \sigma^2 \geq \frac{(n-1)S^2}{u'_\delta} \right) \\ &= \mathbb{P}_{\sigma^2} \left(\frac{(n-1)S^2}{u'_\delta} \leq \sigma^2 \leq \frac{(n-1)S^2}{u_\delta} \right) \\ &= \mathbb{P}_{\sigma^2} \left(\left[\frac{(n-1)S^2}{u'_\delta}, \frac{(n-1)S^2}{u_\delta} \right] \ni \sigma^2 \right). \end{aligned} \quad (32.29)$$

□

Wie im Falle von Theorem 32.2 ist der entscheidene Schritt zur Sicherung der Überdeckungswahrscheinlichkeit δ des wahren, aber unbekannten, Varianzparameters durch das in Theorem 32.4 definierte Konfidenzintervall die Definition von

$$u_\delta := \Xi^{-1} \left(\frac{1-\delta}{2}; n-1 \right) \text{ und } u'_\delta := \Xi^{-1} \left(\frac{1+\delta}{2}; n-1 \right) \quad (32.30)$$

Wie im Beweis von Theorem 32.4 nachgezeichnet ist die Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den wahren, aber unbekannten, Varianzparameter äquivalent zu der Tatsache, dass bei Wahl eben dieser Werte von u_δ und u'_δ die U -Konfidenzintervallstatistik eine Wahrscheinlichkeit von δ dafür hat, einen Wert im Intervall $[u_\delta, u'_\delta]$ anzunehmen. Wir visualisieren die Wahl von u_δ und u'_δ für Fall $\delta := 0.95$ und $n := 10$ in Abbildung 32.6. In diesem Fall ergibt sich

$$u_\delta := \Xi^{-1}(0.025; 9) = 2.70 \text{ und } u'_\delta := \Xi^{-1}(0.975; 9) = 19.0. \quad (32.31)$$

Abbildung 32.6 A zeigt diese Wahl aus Perspektive der WDF der U -Konfidenzintervallstatistik. Die von u_δ und u'_δ eingeschlossene Wahrscheinlichkeitsmasse beträgt nach Konstruktion δ , U nimmt mit einer Wahrscheinlichkeit von δ also einen Wert zwischen u_δ und u'_δ an. Abbildung 32.6 B zeigt die entsprechende Perspektive der KVF der U -Konfidenzintervallstatistik. Basierend auf der Vorgabe von $\frac{1-\delta}{2}$ und $\frac{1+\delta}{2}$ werden anhand der inversen KVF Ψ^{-1} die entsprechenden Werte für u_δ und u'_δ bestimmt. Man beachte, dass in diesem Fall die Wahrscheinlichkeitsmasse recht arbiträr hinsichtlich des Modalwerts der Verteilung der U -Konfidenzintervallstatistik lokalisiert ist. Dementsprechend gibt es weitergehende Verfahren, die Überdeckungswahrscheinlichkeit einer Konfidenzintervallstatistik so zu lokalisieren, dass sie beispielsweise ein maximales Intervall in ihrem Ergebnisraum einnimmt oder eine Symmetrieeigenschaft um den Erwartungswert erfüllt, die wir hier aber nicht vertiefen wollen.

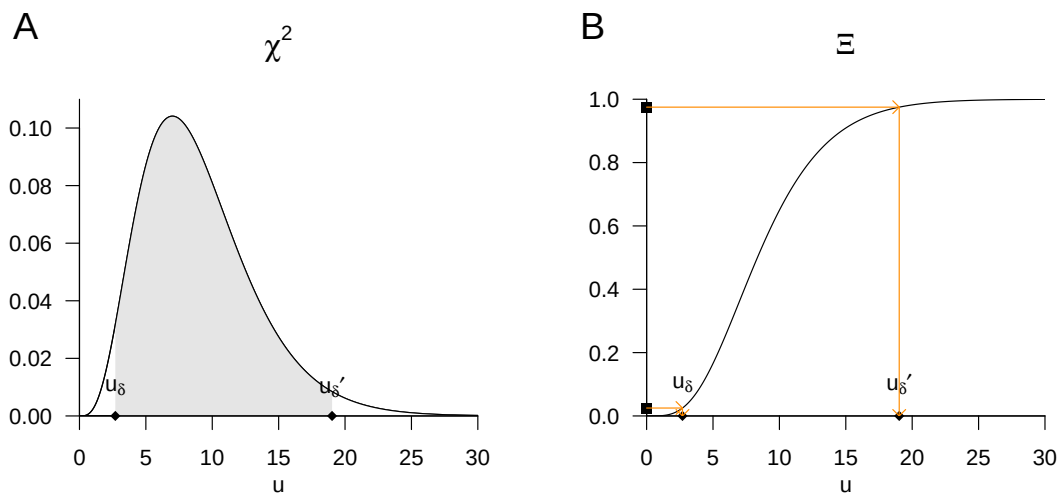


Abbildung 32.6. Sicherung der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Varianzparameter des Normalverteilungsmodells für $\delta := 0.95$ und $n := 10$ aus Perspektive der WDF (A) und der KVF (B) der Verteilung der U -Konfidenzintervallstatistik.

Abschließend wollen wir die Überdeckungswahrscheinlichkeit des durch Theorem 32.4 gegebenen Konfidenzintervalls mithilfe einer Simulation demonstrieren. Dazu betrachten wir zunächst die in Kapitel 32.1 gegebene erste Interpretation eines Konfidenzintervalls bei immer gleichem wahren, aber unbekannten, Parameter. Folgender **R** Code bestimmt in diesem Sinne zu jeder Stichprobenrealisierung das entsprechende Konfidenzintervall.

```
# Modellformulierung
set.seed(1)                                # random number generator seed
mu      = 2                                # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr  = 2                                # wahrer, aber unbekannter, Varianzparameter
n        = 12                               # Stichprobenumfang
delta    = 0.95                             # Konfidenzbedingung
u_delta_u = qchisq((1-delta)/2, n - 1)      # \chi^2((1-\delta)/2; n - 1)
u_delta_o = qchisq((1+delta)/2, n - 1)      # \chi^2((1+\delta)/2; n - 1)

# Stichprobenrealisierungen
ns      = 1e2                               # Anzahl Simulationen
y_bar   = rep(NA, ns)                       # Stichprobenmittellarray
S2      = rep(NA, ns)                       # Stichprobenvarianzarray
kappa   = matrix(rep(NA, 2*ns), ncol = 2)  # Konfidenzintervallarray
for(i in 1:ns){
  y      = rnorm(n, mu, sqrt(sigsqr))        # Stichprobenrealisierung
  S2[i]  = var(y)                           # Stichprobenvarianz
  kappa[i,1] = (n-1)*S2[i]/u_delta_o         # untere Konfidenzintervallgrenze
  kappa[i,2] = (n-1)*S2[i]/u_delta_u         # obere Konfidenzintervallgrenze
}
```

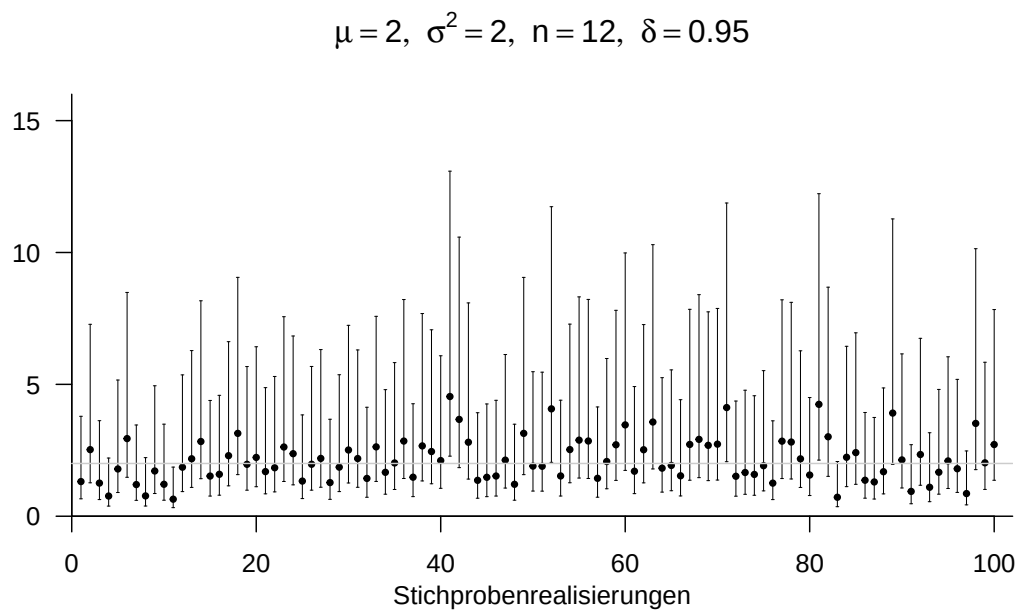


Abbildung 32.7. Simulation der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Varianzparameter des Normalverteilungsmodells bei konstanten, wahren, aber unbekannten, Varianzparameter $\sigma^2 := 2$ für $\mu := 2, n := 12$ und einer gewünschten Überdeckungswahrscheinlichkeit von $\delta := 0.95$. Die Abbildung zeigt für jede Stichprobenrealisierung das Konfidenzintervall und den entsprechenden Varianzparameterschätzer. In der vorliegenden Simulation überdecken die Konfidenzintervalle den durch eine graue Linie eingezeichneten immer gleichen wahren, aber unbekannten, Varianzparameter $\sigma^2 := 2$ in 95 von 100 Fällen. Die Stichprobenrealisierungen, für die dies nicht der Fall ist, sind mit einem orangen Kreis markiert. Man beachte, dass die Konfidenzintervalle nicht symmetrisch um den Varianzparameterschätzer angeordnet sind

32.3. Anwendungsbeispiel

Zum Abschluss dieses Abschnitts wollen wir die Evaluation von Konfidenzintervallen für den Erwartungswert und den Varianzparameter bei Normalverteilung nun im Kontext des Anwendungsbeispiels von Kapitel 30.3.1. Dazu werten wir zunächst einmal die unverzerrten Punktschätzer von μ und σ^2 , also das Stichprobenmittel und die Stichprobenvarianz des Datensatzes mithilfe folgenden **R** Codes aus.

```
D = read.csv("../data/bdi-ii-datensatz.csv") # Datensatzeinlesen
y = D$dBDI # Datenauswahl
mu_hat = mean(y) # Stichprobenmittel
sigsqr_hat = var(y) # Stichprobenvarianz
cat("mu_hat : ", mu_hat, "\nsigsqr_hat : ", sigsqr_hat) # Ausgabe
```

```
mu_hat : 3.166667
sigsqr_hat : 13.78788
```

Basierend auf diesen Schätzern und den vorliegenden $n = 12$ Datenpunkten sind also

$$\hat{\mu} = 3.17 \text{ und } \hat{\sigma}^2 = 13.8 \quad (32.32)$$

sinnvolle Tipps für μ und σ^2 . Um neben diesen Punktschätzern, die zwar sehr genau sind, mit einer Wahrscheinlichkeit von 0 aber den wahren, aber unbekannten Parametern, exakt entsprechen, werten wir zusätzlich die 95%-Konfidenzintervallschätzungen für μ und σ^2 aus. Folgender **R** Code bestimmt das 95%-Konfidenzintervall für den Erwartungswertparameter.

```
# Konfidenzintervall für den Erwartungswertparameter
delta = 0.95 # Konfidenzlevel
n = length(y) # Anzahl Datenpunkte
t_delta = qt((1+delta)/2, n-1) # \psi^{-1}((\delta+1)/2, n-1)
y_bar = mean(y) # Stichprobenmittel
s = sd(y) # Stichprobenstandardabweichung
mu_hat = y_bar # Erwartungswertparameterschätzer
kappa_u = y_bar - (s/sqrt(n))*t_delta # untere Konfidenzintervallgrenze
kappa_o = y_bar + (s/sqrt(n))*t_delta # obere Konfidenzintervallgrenze
cat("kappa_u:", kappa_u, "\nkappa_o:", kappa_o) # Ausgabe
```

```
kappa_u: 0.8074098
kappa_o: 5.525923
```

Das 0.95-Konfidenzintervall für den Erwartungswertparameter ist also

$$\kappa(y) = [0.80, 5.52]. \quad (32.33)$$

Im langfristigen Mittel überdeckt ein auf diese Weise berechnetes Konfidenzintervall den wahren, aber unbekannten, Erwartungswertparameter in 95 von 100 Fällen. In diesem Sinne liegt der wahre, aber unbekannte, Therapieeffekt also sehr sicher in einem Intervall zwischen 0.80 und 5.52 BDI-II Score Pre-Post-Differenzen.

Folgender **R** Code bestimmt das 95%-Konfidenzintervall für den Varianzparameter.

```
# Konfidenzintervall für den Varianzparameter
delta = 0.95 # Konfidenzlevel
n = length(y) # Anzahl Datenpunkte
u_delta_u = qchisq((1-delta)/2, n-1) # \chi^2((1-\delta)/2; n-1)
u_delta_o = qchisq((1+delta)/2, n-1) # \chi^2((1+\delta)/2; n-1)
s2 = var(y) # Stichprobenstandardabweichung
sigsqr_hat = s2 # Varianzparameterschätzer
kappa_u = (n-1)*s2/u_delta_o # untere Konfidenzintervallgrenze
kappa_o = (n-1)*s2/u_delta_u # obere Konfidenzintervallgrenze
cat("kappa_u:", kappa_u, "\nkappa_o:", kappa_o) # Ausgabe
```

kappa_u: 6.919084
kappa_o: 39.74756

Das 0.95-Konfidenzintervall für den Varianzparameter ist also

$$\kappa(y) = [6.91, 39.74]. \quad (32.34)$$

Im langfristigen Mittel überdeckt ein auf diese Weise berechnetes Konfidenzintervall den wahren, aber unbekannten, Varianzparameter in 95 von 100 Fällen. In diesem Sinne liegt die wahre, aber unbekannte, Therapieeffektstreuung also sehr sicher in einem Intervall zwischen 6.91 und 39.74 quadriertern BDI-II Score Pre-Post-Differenzen.

32.4. Literaturhinweise

Die in diesem Kapitel vorgestellten Ergebnisse gehen in ganz wesentlicher Weise auf Neyman (1937) zurück.

33. Hypothesentests

Die grundlegende Logik Frequentistischer Hypothesentests kann am Beispiel eines Normalverteilungsmodells für einen beobachteten univariaten Datensatz $y_1, \dots, y_n$ grob wie folgt umrissen werden. Man unterstellt zunächst, dass der beobachtete Datensatz eine Realisierung der Stichprobe $y_1, \dots, y_n \sim N(\mu, \sigma^2)$ ist und berechnet dann basierend auf dem Datensatz eine *Teststatistik*, zum Beispiel das anhand der Stichprobenstandardabweichung und der Stichprobengröße normalisierte Stichprobenmittel $\sqrt{n} \frac{\bar{y}}{s}$.

Man fragt sich dann, wie wahrscheinlich es wohl wäre, den beobachteten oder einen extremeren Wert der Teststatistik unter der Annahme eines *Nullmodels* zu observieren. Dabei versteht man unter einem *Nullmodell* intuitiv ein Wahrscheinlichkeitsverteilungsmodell bei dem kein “interessanter Effekt” vorliegt, also im Sinne des Normalverteilungsmodells zum Beispiel $\mu = 0$ gilt. Dabei ist der Begriff der Wahrscheinlichkeit natürlich Frequentistisch zu verstehen, also als idealisierte relative Häufigkeit, wenn man viele Stichprobenrealisationen des Nullmodells generieren würde. Je nach Beschaffenheit des zugrundeliegenden Frequentistischen Inferenzmodells und der betrachteten Teststatistik kann es dabei durchaus möglich sein, diese Wahrscheinlichkeit exakt zu bestimmen.

Ist nun die betrachtete Wahrscheinlichkeit dafür, den beobachteten oder einen extremeren Wert der Teststatistik unter Annahme des Nullmodells zu observieren groß, so schließt man intuitiv, dass “es wohl ganz plausibel ist, dass das Nullmodell die Daten generiert hat”. Im Wissenschaftsjargon spricht dann manchmal von einem “statistisch nicht-signifikanten Ergebnis”. Ist die betrachtete Wahrscheinlichkeit dafür, den beobachteten oder einen extremeren Wert der Teststatistik unter Annahme des Nullmodells zu observieren dagegen klein, so schließt man intuitiv, dass “es wohl nicht so plausibel, dass das Nullmodell die Daten generiert hat”. Im Wissenschaftsjargon spricht man in diesem Fall manchmal von einem “statistisch signifikanten Ergebnis”.

Wie immer in der Frequentistischen Statistik weiß man nach Durchführung einer solchen Prozedur natürlich nicht, ob im vorliegenden Fall nun wirklich das Nullmodell oder ein anderes Modell die Daten generiert hat, sondern man weiß nur, wie oft man bei dieser Prozedur im Mittel richtig oder falsch liegen würde, wenn alle Annahmen zuträfen und man diese Prozedur sehr oft wiederholen würde.

In den folgenden Abschnitten wollen wir diese intuitiven Gedanken formalisieren. Dabei ist es wichtig, immer zwischen “Hypothesen” im Sinne der Frequentistischen Inferenz und dem generellen Begriff der wissenschaftlichen Hypothese zu unterscheiden. Das Aufstellen einer wissenschaftlichen Hypothese bedingt keinesfalls, dass ein Frequentistischer Hypothesentest anzuwenden ist, sondern lediglich, so man denn quantitativ arbeiten möchte, dass es Sinn macht seine Unsicherheit im Lichte beobachteter Daten, die potentiell über die (wissenschaftliche) Hypothese aussagekräftig sind, zu quantifizieren und zu kommunizieren. Frequentistische Hypothesentests sind nur eine der vielen Möglichkeiten, dies zu tun, wenn auch eine sehr populäre. Es sei trotzdem schon an dieser Stelle erwähnt, dass

das “Nullhypothesen-Signifikanz-Testen”, wie im folgenden dargelegt, im wissenschaftlichen Kontext durchaus nicht unumstritten ist (vgl. zum Beispiel Amrhein & Greenland (2018) und McShane et al. (2019)).

33.1. Testhypothesen und Tests

Im Kontext von Frequentistischen Hypothesentests wird der Begriff des Frequentistischen Inferenzmodells (vgl. Definition 30.1) zunächst durch die sogenannten *Testhypothesen* zu einem *Testszenario* erweitert. Wir nutzen folgende Definition.

Definition 33.1 (Testhypothesen und Testszenario). Gegeben sei ein Frequentistisches Inferenzmodell mit Stichprobe y , Ergebnisraum $\mathcal{Y}$ und Parameterraum Θ . Weiterhin sei $\{\Theta_0, \Theta_1\}$ eine Partition des Parameterraums, so dass

$$\Theta = \Theta_0 \cup \Theta_1 \text{ und } \Theta_0 \cap \Theta_1 = \emptyset. \quad (33.1)$$

Dann ist eine *Testhypothese* eine Aussage über den wahren, aber unbekannten, Parameterwert θ in Hinblick auf die Untermengen Θ_0 und Θ_1 des Parameterraums. Speziell werden die Aussagen

- $\theta \in \Theta_0$ als *Nullhypothese* und
- $\theta \in \Theta_1$ als *Alternativhypothese*

bezeichnet. Der Einfachheit halber bezeichnet man auch Θ_0 und Θ_1 direkt als Nullhypothese und Alternativhypothese, respektive. Die Einheit aus Frequentistischem Inferenzmodell und Testhypothesen wird als *Testszenario* bezeichnet.

•

Je nach Beschaffenheit von Θ_0 und Θ_1 unterscheidet man einerseits *einfache* und *zusammengesetzte* und andererseits *einseitige* und *zweiseitige* Testhypothesen.

Definition 33.2 (Einfache und zusammengesetzte Testhypothesen). Für die Testhypothesen Θ_i mit $i = 0, 1$ gilt:

- Enthält Θ_i nur ein einziges Element, so heißt Θ_i *einfach*.
- Enthält Θ_i mehr als ein Element, so heißt Θ_i *zusammengesetzt*.

•

Man beachte, dass da nach Annahme der wahren, aber unbekannte, Parameter θ die Verteilung $\mathbb{P}_\theta$ der Stichprobe festlegt, eine einfache Testhypothese der Festlegung der Verteilung der Stichprobe auf genau eine Verteilung entspricht. Eine zusammengesetzte entspricht dagegen einer Menge möglicher Verteilungen der Stichprobe. Ein Beispiel für eine einfache Testhypothese in einem Testszenario mit Parameterraum $\Theta := \mathbb{R}$ ist

$$\Theta_0 := \{0\}, \quad (33.2)$$

die entsprechend zusammengesetzte Alternativhypothese ist dann gegeben durch

$$\Theta_1 = \mathbb{R} \setminus \{0\}. \quad (33.3)$$

Die Nullhypothese, also die Aussage “ $\theta \in \Theta_0$ ” entspricht dann der Aussage “ $\theta = 0$ ”, da Θ_0 nur eben dieses eine Element enthält.

Ist wie in diesem Beispiel der Parameterraum eindimensional, so unterscheidet man weiterhin einseitige und zweiseitige Null- und Alternativhypothesen.

Definition 33.3 (Einseitige und zweiseitige Testhypothesen). Gegeben sei ein Testszenario mit eindimensionalem Parameterraum $\Theta := \mathbb{R}$ und es sei $\theta_0 \in \Theta$. Dann werden zusammengesetzte Nullhypothesen der Form $\Theta_0 :=]-\infty, \theta_0]$ oder $\Theta_0 := [\theta_0, \infty[$ *einseitige Nullhypothesen* genannt und auch in der Form $H_0 : \theta \leq \theta_0$ bzw. $H_0 : \theta \geq \theta_0$ geschrieben. Die entsprechenden Alternativhypothesen haben dabei die Form $\Theta_1 :=]\theta_0, \infty[$ bzw. $\Theta_1 :=]-\infty, \theta_0[$, auch geschrieben als $H_1 : \theta > \theta_0$ bzw. $H_1 : \theta < \theta_0$. Bei einer einfachen Nullhypothese der Form $\Theta_0 := \{\theta_0\}$, auch geschrieben als $H_0 : \theta = \theta_0$, wird die Alternativhypothese $\Theta_1 := \Theta \setminus \{\theta_0\}$, auch geschrieben als $H_1 : \theta \neq \theta_0$, *zweiseitige Alternativhypothese* genannt.

•

Vor dem Hintergrund eines Testszenarios definieren wir nun den Begriff des *Hypothesentests*, den wir kurz einfach als *Test* bezeichnen wollen.

Definition 33.4 (Test). Gegeben sei ein Testszenario. Dann ist ein *Test* eine Abbildung ϕ aus dem Ergebnisraum der Stichprobe $\mathcal{Y}$ in die Menge $\{0, 1\}$, also

$$\phi : \mathcal{Y} \rightarrow \{0, 1\}, y \mapsto \phi(y), \quad (33.4)$$

wobei

- $\phi(y) = 0$ den Vorgang des Nichtablehnens der Nullhypothese und
- $\phi(y) = 1$ den Vorgang des Ablehnens der Nullhypothese

repräsentieren.

•

Die Formalisierung des Testbegriffs ist nicht trivial, da Tests, wie Schätzer und Konfidenzintervalle, Funktionen von Zufallsvariablen, nämlich gerade den Stichprobenvariablen sind. Eigentlich sind Tests damit auf Zufallsvektorräumen definiert. Der Einfachheit halber betrachten wir in Definition 33.4 eine konkrete Realisierung $y \in \mathcal{Y}$ der Stichprobe y , die durch ϕ in die Menge $\{0, 1\}$ abgebildet wird. Der Funktionswert $\phi(y)$ von ϕ ist vor diesem Hintergrund also eine Realisierung der Zufallsvariable $\phi(y)$.

In der Anwendung ist man oft an Tests interessiert, die eine bestimmte Struktur haben, wir formalisieren diese unter dem Begriff der *Standardtests*.

Definition 33.5 (Standardtest). Gegeben sei ein Testszenario. Dann ist ein *Standardtest* ϕ definiert als die Verkettung einer *Teststatistik*

$$\gamma : \mathcal{Y} \rightarrow \Gamma \quad (33.5)$$

und einer *Entscheidungsregel*

$$\delta : \Gamma \rightarrow \{0, 1\} \quad (33.6)$$

kann also geschrieben werden als

$$\phi := \delta \circ \gamma : \mathcal{Y} \rightarrow \{0, 1\}. \quad (33.7)$$

•

Wie oben angemerkt gibt es auch bei Definition 33.5 zu beachten, dass die Teststatistik eigentlich eine Funktion der Stichprobenvariablen, also von Zufallsvariablen ist, die wir hier als Funktion der Werte dieser Zufallsvariablen in $\mathcal{Y}$ definiert haben. Ebenso gibt es zu beachten, dass die Entscheidungsregel eine Funktion der somit zufälligen Teststatistik ist, die wir hier gleichfalls als Funktion der Werte dieser Zufallsvariable mit Ergebnisraum Γ geschrieben haben. Sowohl Teststatistik und Entscheidungsregel sind in einem Testszenario also Zufallsvariablen. Entsprechend ist, wenn y eine Realisierung der Stichprobe y ist, $\gamma(y) \in \Gamma$ eine Realisierung von $\gamma(y)$ und $(\delta \circ \gamma)(y)$ eine Realisierung von $(\delta \circ \gamma)(y)$.

Die verteilungstheoretischen Eigenschaften eines Tests ergeben sich aus den ihnen zugrundeliegenden verteilungstheoretischen Eigenschaften des entsprechenden Frequentistischen Inferenzmodells und damit natürlich insbesondere der Verteilung der Stichprobenvariablen. Eine wichtige Brücke zwischen diesen beiden Ebenen der Verteilung der Stichprobenvariablen auf der einen Seite und der Verteilung der Testergebnisse auf der anderen Seite bilden die Begriffe des *kritischen Bereichs* und des *Ablehnungsbereichs* eines Tests.

Definition 33.6 (Kritischer Bereich eines Tests). Gegeben sei ein Testszenario und ein Test ϕ . Dann heißt die Untermenge K des Ergebnisraums $\mathcal{Y}$ der Stichprobe y , für die der Test den Wert 1 annimmt, *kritischer Bereich* des Tests, formal

$$K := \{y \in \mathcal{Y} | \phi(y) = 1\} \subset \mathcal{Y}. \quad (33.8)$$

•

Man beachte, dass vor dem Hintergrund von Definition 33.6 die zufälligen Ereignisse $\{y \in K\}$ und $\{\phi(y) = 1\}$, also dass die Stichprobe einen Wert im kritischen Bereichs des Tests annimmt bzw. dass der Test den Wert 1 annimmt, äquivalent sind und damit insbesondere auch die gleiche Wahrscheinlichkeit haben. Fragt man also nach der Wahrscheinlichkeit, dass ein Test den Wert 1 annimmt, also die Nullhypothese abgelehnt wird, so entspricht diese Wahrscheinlichkeit genau der Wahrscheinlichkeit, dass die Stichprobe einen Wert im kritischen Bereichs des Tests annimmt. Da die Verteilung der Stichprobe aber als bekannt vorausgesetzt ist, kann die Wahrscheinlichkeit für das Ablehnen der Nullhypothese darauf basierend bestimmt werden. Hat man insbesondere einen Standardtest vorliegen, so überträgt sich das Gesagte unmittelbar auch auf die zwischen Stichprobe und Test geschaltete Teststatistik. Dies führt auf die folgende Definition.

Definition 33.7 (Ablehnungsbereich eines Standardtests). Gegeben sei ein Testszenario und ein Standardtest ϕ mit Teststatistik γ . Die Untermenge A des Ergebnisraums Γ der Teststatistik, für die der Test den Wert 1 annimmt, *Ablehnungsbereich des Tests*, formal

$$A := \{\gamma(y) \in \Gamma | \phi(y) = 1\} \subset \Gamma. \quad (33.9)$$

•

Wie zum Begriff des kritischen Bereichs angemerkt gilt auch hier, dass die Ereignisse $\{\phi(y) = 1\}$ und $\{\gamma(y) \in A\}$ äquivalent sind und damit insbesondere auch die gleiche Wahrscheinlichkeit besitzen. Insgesamt gelten mit Definition 33.6 und Definition 33.7 für einen Standardtest also

$$\{y \in K\} \Leftrightarrow \{\gamma(y) \in A\} \Leftrightarrow \{\phi(y) = 1\} \quad (33.10)$$

und

$$\mathbb{P}_\theta(\{y \in K\}) = \mathbb{P}_\theta(\{\gamma(y) \in A\}) = \mathbb{P}_\theta(\{\phi(y) = 1\}), \quad (33.11)$$

wobei das Subskript θ bei der Verteilung der Teststatistik und des Tests andeuten soll, dass diese Verteilungen durch den Parameter der Stichprobenverteilung festgelegt sind.

In der Anwendung basiert die in Definition 33.5 allgemein angegebene Form der Entscheidungsregel eines Standardtest meist darauf, dass eine beobachtete Teststatistik mit Ergebnisraum $\Gamma := \mathbb{R}$ einen bestimmten sogenannten *kritischen Wert* $k \in \mathbb{R}$ überschreitet oder unterschreitet. Dies führt auf die Konzepte der *einseitigen* und *zweiseitigen kritischen Wert-basierte Tests*.

Definition 33.8 (Kritischer Wert-basierte Tests). Ein *kritischer Wert-basierter Test* ist ein Standardtest, bei dem die Entscheidungsregel δ von einem kritischen Wert k der Teststatistik mit Ergebnisraum $\mathbb{R}$ abhängt. Speziell ist

- ein *einseitiger kritischer Wert-basierter Test* von der Form

$$\phi : \mathcal{Y} \rightarrow \{0, 1\}, y \mapsto \phi(y) := 1_{\{\gamma(y) \geq k\}} = \begin{cases} 1 & \gamma(y) \geq k \\ 0 & \gamma(y) < k \end{cases}, \quad (33.12)$$

- ein *zweiseitiger kritischer Wert-basierter Test* von der Form

$$\phi : \mathcal{Y} \rightarrow \{0, 1\}, y \mapsto \phi(y) := 1_{\{|\gamma(y)| \geq k\}} = \begin{cases} 1 & |\gamma(y)| \geq k \\ 0 & |\gamma(y)| < k \end{cases}. \quad (33.13)$$

•

Mit der Definition kritischer Wert-basierter Tests ist die praktische Durchführung eines Hypothesentests nun vorgezeichnet. Wie immer in der Frequentistischen Inferenz legt man vorliegenden Daten zunächst ein Frequentistisches Inferenzmodell zugrunde, nimmt also an, dass die vorliegenden Daten eine Realisierung einer Stichprobe sind. Basierend auf dieser Realisierung berechnet man eine Teststatistik und vergleicht diese abschließend mit einem kritischen Wert, um dann entweder die Nullhypothese nicht abzulehnen oder die Nullhypothese abzulehnen. Im folgenden Abschnitt wollen wir nun der Frage nachgehen, wie vor dem Hintergrund von Null- und Alternativhypothese dabei der kritische Wert eines kritischen Wert-basierten Tests so bestimmt werden kann, dass man im Sinne der Frequentistischen Wahrscheinlichkeit möglichst gute Testentscheidungen trifft.

33.2. Testgütekriterien und Testkonstruktion

Die Tatsache, dass in einem Testszenario der wahre, aber unbekannte, Parameter im Bereich der Nullhypothese oder der Alternativhypothese liegen kann und man gleichzeitig basierend auf dem Wert des Tests die Nullhypothese entweder ablehnen oder nicht ablehnen kann, impliziert, dass eine Testentscheidung richtig oder falsch sein kann. Untenstehende Definition soll dahingehend zunächst begriffliche Klarheit schaffen.

Definition 33.9 (Richtige Testentscheidungen und Testfehler). Gegeben seien ein Test-szenario und ein Test. Dann gibt es mit dem Nichtablehnen der Nullhypothese $\phi(y) = 0$, wenn die Nullhypothese $\theta \in \Theta_0$ zutrifft und dem Ablehnen der Nullhypothese $\phi(y) = 1$, wenn die Alternativhypothese $\theta \in \Theta_1$ zutrifft zwei Formen der *richtigen Testentscheidung*. Ebenso gibt es zwei Arten von *Testfehlern*: Das Ablehnen der Nullhypothese $\phi(y) = 1$, wenn die Nullhypothese $\theta \in \Theta_0$ zutrifft, heißt *Typ I Fehler* und das Nichtablehnen der Nullhypothese, wenn die Alternativhypothese $\theta \in \Theta_1$ zutrifft, heißt *Typ II Fehler*.

		Testentscheidung	
		$\phi(y) = 0$	$\phi(y) = 1$
Wahrer, aber unbekannter, Parameter	$\theta \in \Theta_0$	Richtige Entscheidung	Typ I Fehler
	$\theta \in \Theta_1$	Typ II Fehler	Richtige Entscheidung

Abbildung 33.1. Richtige Testentscheidungen und Typ I und Typ II Fehler

Abbildung 33.1 gibt eine Übersicht zu den möglichen richtigen Testentscheidungen und Testfehlern bei Durchführung eines Tests. Natürlich möchte man generell meist eine richtige Testentscheidung treffen. Das entscheidene Werkzeug, um vor dem Frequentistischen Hintergrund des Testszenarios gute Tests zu konstruieren, ist die sogenannte *Testgütefunktion*.

Definition 33.10 (Testgütefunktion). Gegeben sei ein Testszenario und ein Test ϕ . Dann ist die *Testgütefunktion* von ϕ definiert als

$$q_\phi : \Theta \rightarrow [0, 1], \theta \mapsto q_\phi(\theta) := \mathbb{P}_\theta(\phi(y) = 1). \quad (33.14)$$

Für $\theta \in \Theta_1$ heißt q_ϕ auch *Trennschärfefunktion* oder *Powerfunktion*.

Man beachte, dass $\mathbb{P}_\theta$ in Definition 33.10 die Verteilung der Zufallsvariable $\phi(y)$ unter der Annahme, dass die Verteilung von y durch θ festgelegt ist, bezeichnen soll. Für jedes $\theta \in \Theta$ liefert q_ϕ also die Wahrscheinlichkeit dafür, dass die Nullhypothese durch den Test ϕ abgelehnt wird. Für diese Wahrscheinlichkeiten gelten insbesondere mit den Begriffen des kritischen Bereichs (vgl. Definition 33.6) und des Ablehnungsbereichs (vgl. Definition 33.7) wie bereits gesehen

$$\mathbb{P}_\theta(\phi(y) = 1) = \mathbb{P}_\theta(\gamma \in A) = \mathbb{P}_\theta(y \in K). \quad (33.15)$$

Die Testgütefunktion ist spezifisch für einen gegebenen Test. Ändert sich der Test, zum Beispiel, weil bei einem kritischen Wert-basierten Test ein anderer kritischer Wert gewählt wird, ändern sich obige Wahrscheinlichkeiten und damit die Testgütefunktion.

Mithilfe der Testgütefunktion folgt die Testkonstruktion dann folgenden Überlegungen. Im Idealfall hätte man einen Test ϕ mit

$$q_\phi(\theta) = \mathbb{P}_\theta(\phi(y) = 1) = 0 \text{ für } \theta \in \Theta_0 \text{ und } q_\phi(\theta) = \mathbb{P}_\theta(\phi(y) = 1) = 1 \text{ für } \theta \in \Theta_1. \quad (33.16)$$

Die Testentscheidung eines solchen Tests wäre mit Wahrscheinlichkeit 1 richtig, da ein solcher Test die Nullhypothese mit Wahrscheinlichkeit 0 ablehnt, wenn sie zutrifft, und die Nullhypothese mit Wahrscheinlichkeit 1 ablehnt, wenn sie nicht zutrifft. Allgemeiner sind natürlich kleine Werte von q_ϕ für $\theta \in \Theta_0$, also kleine Wahrscheinlichkeiten dafür, die Nullhypothese abzulehnen, wenn sie zutrifft, und große Werte von q_ϕ für $\theta \in \Theta_1$, also große Wahrscheinlichkeiten dafür, die Nullhypothese abzulehnen, wenn sie nicht zutrifft, zur Testfehlerminimierung günstig. Allerdings bestehen im Allgemeinen Abhängigkeiten zwischen den Werten der Testgütefunktion für $\theta \in \Theta_0$ und $\theta \in \Theta_1$, wie folgende Beispiele illustrieren sollen.

Beispiel (A) Es sei ϕ_a ein Test definiert durch

$$\phi_a : \mathcal{Y} \rightarrow \{0, 1\}, y \mapsto \phi_a(y) := 0. \quad (33.17)$$

ϕ_a sei also ein Test, der die Nullhypothese unabhängig von den beobachteten Daten nie ablehnt. Für ϕ_a gilt dann

$$q_{\phi_a}(\theta) = \mathbb{P}_\theta(\phi(y) = 1) = 0 \text{ für } \theta \in \Theta_0. \quad (33.18)$$

Allerdings gilt für ϕ_a dann auch automatisch

$$q_{\phi_a}(\theta) = \mathbb{P}_\theta(\phi(y) = 1) = 0 \text{ für } \theta \in \Theta_1. \quad (33.19)$$

ϕ_a hat also eine minimale Sensitivität dafür, die Tatsache, dass die Alternativhypothese zutrifft, zu detektieren.

Beispiel (B) Andersherum sei ϕ_b ein Test definiert durch

$$\phi_b : \mathcal{Y} \rightarrow \{0, 1\}, y \mapsto \phi_b(y) := 1. \quad (33.20)$$

ϕ_b sei also ein Test, der die Nullhypothese, unabhängig von den beobachteten Daten immer ablehnt. Für ϕ_b gilt dann

$$q_{\phi_b}(\theta) = \mathbb{P}_\theta(\phi(y) = 1) = 1 \text{ für } \theta \in \Theta_1. \quad (33.21)$$

ϕ_b ist also maximal sensitiv für das Zutreffen der Alternativhypothese. Allerdings gilt für ϕ_b dann auch automatisch

$$q_{\phi_b}(\theta) = \mathbb{P}_\theta(\phi(y) = 1) = 0 \text{ für } \theta \in \Theta_0, \quad (33.22)$$

und ϕ_b resultiert auch immer in der Ablehnung der Nullhypothese, wenn diese zutrifft und generiert in diesem Sinne viele falsch positive Resultate.

Vor dem Hintergrund dieser beiden Extremszenarien muss es also das Ziel der Testkonstruktion sein, eine angemessene Balance zwischen kleinen Werten der Testgütefunktion bei Zutreffen der Nullhypothese und großen Werten der Testgütefunktion bei Zutreffen der Alternativhypothese zu finden. Die populärste Methode, dies zu erreichen ist es, in einem ersten Schritt einen kleinen Wert $\alpha_0 \in [0, 1]$ zu wählen und sicherzustellen, dass

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0, \quad (33.23)$$

dass also die Wahrscheinlichkeit für das Ablehnen der Nullhypothese, wenn diese zutrifft, also die Wahrscheinlichkeit für einen Typ I Fehler, höchstens α_0 beträgt. Konventionelle Werte für ein solches α_0 sind zum Beispiel $\alpha_0 := 0.001$ und $\alpha_0 := 0.05$. Unter allen Tests (und, bei Optimierung von Stichprobengrößen, Frequentistischen Inferenzmodellen), die die Ungleichung (33.23) erfüllen, sucht man dann in einem zweiten Schritt einen Test, für den $q_\phi(\theta)$ für $\theta \in \Theta_1$ so groß wie möglich ist. Dieses zweischrittige Vorgehen ist nicht alternativlos, man könnte ja beispielsweise auch eine lineare Kombinationen von Typ I und Typ II Fehlern simultan minimieren. Allerdings ist das skizzierte zweischrittige Vorgehen das in der Anwendung populärste, so dass wir uns in der Folge darauf beschränken wollen. Ungleichung (33.23) motiviert dann zunächst die Definition der Begriffe des *Level- α_0 -Tests*, des *Signifikanzlevels* α_0 und des *Testumfangs* α .

Definition 33.11 (Level- α_0 -Test, Signifikanzlevel α_0 und Testumfang α). Gegeben seien ein Testszenario, ein Test ϕ , seine Testgütefunktion q_ϕ und ein $\alpha_0 \in [0, 1]$. ϕ heißt ein *Level- α_0 -Test*, wenn gilt, dass

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0. \quad (33.24)$$

Wenn ϕ ein Level- α_0 -Test ist, nennt man den Wert α_0 auch das *Signifikanzlevel* des Tests. Weiterhin heißt die Zahl

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) \in [0, 1] \quad (33.25)$$

der *Testumfang* von ϕ .

•

Nach Definition 33.11 ist der Testumfang α die maximale Wahrscheinlichkeit für einen Typ I Fehler und ein Test ist dann, und nur dann, ein Level- α_0 -Test, wenn diese maximale Wahrscheinlichkeit kleiner oder gleich dem Signifikanzlevel α_0 ist. Es ist dabei für die Anwendung wichtig, sich die feinen begrifflichen Unterschiede zwischen der Wahrscheinlichkeit eines Typ I Fehlers, dem Testumfang und dem Signifikanzlevels eines Tests zu verdeutlichen. Vor dem Hintergrund des Unterschiedes von einfachen und zusammengesetzten Nullhypothesen (vgl. Definition 33.2) muss man zunächst die Begriffe der Typ I Fehler Wahrscheinlichkeit und des Testumfangs differenzieren. Bei einer einfachen Nullhypothese Θ_0 ist der Testumfang immer gleich der Wahrscheinlichkeit eines Typ I Fehlers, da gilt dass

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) = \max_{\theta \in \{\theta_0\}} q_\phi(\theta) = q_\phi(\theta_0) = \mathbb{P}_{\theta_0}(\phi = 1). \quad (33.26)$$

Bei einer zusammengesetzten Nullhypothese Θ_0 gibt es je nach Wert von $\theta \in \Theta_0$ verschiedene Wahrscheinlichkeiten für einen Typ I Fehler. Die größte dieser Wahrscheinlichkeiten ist der Testumfang

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) = \max_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi = 1). \quad (33.27)$$

Ebenso klar sollte man die Begriffe des Signifikanzlevels α_0 und des Testumfangs α voneinander abgrenzen. Ein Signifikanzlevel ist eine frei gewählte obere Grenze für die maximale Wahrscheinlichkeit eines Typ I Fehlers. Die tatsächliche maximale Wahrscheinlichkeit für einen Typ I Fehler, kann mit dieser identisch sein, wie in den meisten Fällen der Kapitel 33.3 diskutierten Beispiele, muss es aber nicht, wie zum Beispiel in multiplen Testszenarien mit nicht unabhängigen Stichprobenvariablen. Man nennt dementsprechend einen Test *exakt*, wenn sein Testumfang mit seinem Signifikanzlevel identisch ist, wenn also

$$\alpha = \alpha_0. \quad (33.28)$$

Ein Test, für den der Testumfang kleiner als sein Signifikanzlevel ist, für den also gilt

$$\alpha < \alpha_0 \quad (33.29)$$

wird *konservativ* genannt. Ein Test schließlich, dessen Testumfang größer als sein Signifikanzlevel ist,

$$\alpha > \alpha_0 \quad (33.30)$$

und der damit natürlich kein Level- α_0 -Test sein kann, wird *liberal* genannt.

p-Wert

Ein definierendes Charakteristikum eines Tests ist es wie gesehen, dass die Wertemenge eines Tests binär ist, Resultat eines Tests ist entweder $\phi = 0$, die Nullhypothese wird nicht abgelehnt, oder $\phi = 1$, die Nullhypothese wird abgelehnt. Als finales Resultat einer Datenanalyse wird dabei die einem Datensatz inhärente Information sehr stark komprimiert. Insbesondere supprimiert das alleinige Berichten des Testergebnisses interessante Information über das Signal-zu-Rauschen-Verhältnis des betrachteten Datensatzes. So ist es ja beispielsweise möglich, dass die Nullhypothese im Kontext eines kritischen Wert-basierten abgelehnt wird, weil die Teststatistik den kritischen Wert nur um wenige Nachkommastellen übertroffen hat oder aber, dass die Teststatistik ein Vielfaches des kritischen Werts angenommen hat. In beiden Fällen wäre das Testergebnis mit $\phi = 1$ identisch. Neben der reinen Testumfangkontrolle eines Tests und des Berichtens des binären Testergebnisses hat es sich deshalb für kritische Wert-basierte Tests eingebürgert, basierend auf dem beobachteten Wert der Teststatistik auch alle Werte des Signifikanzlevels α_0 , für die ein Level- α_0 -Test das Ergebnis $\phi = 1$ hätte, für die die Nullhypothese also abgelehnt werden würden, zu betrachten. Diese Überlegung führt auf folgende allgemeine Definition des sogenannten p-Werts, wobei p für *probability* steht.

Definition 33.12 (p-Wert). ϕ sei ein Test. Dann ist der *p-Wert* das kleinste Signifikanzlevel α_0 , bei dem die Nullhypothese basierend auf einem vorliegendem Wert der Teststatistik abgelehnt werden würde.

•

Insbesondere in einfachen Anwendungsbeispielen, wie dem in Kapitel 33.3.1 betrachteten Einstichproben-T-Test-Szenario spiegeln p-Werte dann die Antwort auf die intuitive Frage, wie wahrscheinlich es im Frequentistischen Sinne wäre, den beobachteten oder einen extremeren Wert der Teststatistik unter der Annahme eines Nullmodells zu observieren. Dabei ist in vielen Bereichen der Grundlagenwissenschaft das Berichten von p-Werten sehr populär, aber auch umstritten (vgl. Wasserstein et al. (2019)). Dabei gilt es insbesondere, p-Werte nicht zu überinterpretieren. Basierend auf dem Gesagten gibt es keinen Grund dies anzunehmen, trotzdem weisen wir vorsorglich daraufhin, dass p-Werte *nicht* die Wahrscheinlichkeit dafür quantifizieren, dass die Nullhypothese wahr ist, man aufgrund eines p-Wertes kleiner als 0.05 *nicht* darauf schließen kann, dass die Alternativhypothese zutrifft, und man aufgrund eines p-Wertes von größer als 0.05 *nicht* darauf schließen kann, dass die Nullhypothese zutrifft. Ebenso wie der Wert einer Teststatistik und eines Tests quantifizieren p-Werte lediglich das in einem vorliegenden Datensatz beobachtete Signal-zu-Rauschen-Verhältnis - nicht weniger, aber auch nicht mehr.

Anmerkungen zur Wahl von Null- und Alternativhypothese

Wir wollen diesen Abschnitt mit einigen Anmerkungen zur Durchführung von Hypothesentests in der Wissenschaft beschließen. Vor dem Hintergrund der skizzierten Theorie der Hypothesentests stellt sich zunächst die Frage, wie man in einem gegebenen Anwendungskontext die Zuordnung von Null- und Alternativhypothese zu den Gegenständen des wissenschaftlichen Interesses, also zweier wissenschaftlichen Hypothesen vornimmt. Möchte man zum Beispiel einen Test durchführen, um im Sinne der Frequentistischen Inferenz zu entscheiden, ob ein bestimmtes Psychotherapieverfahren in einer klinischen Studie wirksam war oder nicht, so stellt sich die Frage, ob man dabei die Abwesenheit eines Therapieeffekts dabei als die Null- oder als die Alternativhypothese verstehen sollte. Dazu sei angemerkt, dass das oben beschriebene zweischrittige Vorgehen zur Testkonstruktion, in dem zunächst durch die Wahl eines Signifikanzlevels der Testumfang begrenzt wird und erst in einem zweiten Schritt dafür gesorgt wird, dass die Wahrscheinlichkeit, die Nullhypothese abzulehnen, wenn die Alternativhypothese zutrifft, möglichst groß ist, eine deutliche Asymmetrie in der Behandlung von Null- und Alternativhypothese impliziert: Man wichtet mit diesem Vorgehen Typ I Fehler als schwerwiegender als Typ II Fehler. Dies wiederum impliziert eine mögliche Strategie zur Festlegung von Null- und Alternativhypothese: Die Nullhypothese ist die wissenschaftliche Hypothese, hinsichtlich deren assoziierter Testentscheidung man eher keinen Fehler machen möchte bzw. deren Fehlerwahrscheinlichkeit man primär kontrollieren möchte. In der Wissenschaft ist es ein gebräuchlicher Standard, die falsche Konfirmation der von einem selbst favorisierten Theorie (also zum Beispiel die falsche Konfirmation, dass ein selbstentwickeltes Psychotherapieverfahren besser wirkt als ein anderes) als einen schwerwiegenderen Fehler als die falsche Ablehnung der eigenen Theorie zu werten. Damit sollte die falsche Konfirmation der eigenen Theorie ein Typ I Fehler, das falsche Ablehnen der eigenen Theorie ein Typ II Fehler sein. Damit nun die falsche Konfirmation der eigenen Theorie einen Typ I Fehler, also das Ablehnen der Nullhypothese bei Zutreffen der Nullhypothese, darstellt, muss die eigene Theorie als Alternativhypothese aufgestellt werden, die Alternativhypothese fälschlicherweise abzulehnen wird damit ein Typ II Fehler. Intuitiv ergibt sich also folgende Zuordnung:

Nicht-favorisierte	wissenschaftliche Hypothese	→	Nullhypothese
Favorisierte	wissenschaftliche Hypothese	→	Alternativhypothese

Anmerkungen zu Hypothesentests in Entscheidungskontexten und Grundlagenwissenschaft

Zum zweiten stellt sich die Frage, ob man zur Evaluation wissenschaftlicher Hypothesen überhaupt einen Hypothesentest durchführen sollte. Oberflächliche betrachtet liefern Hypothesentests zunächst einmal einfache binäre Aussagen der Form *“Die Hypothese ist gegeben die Evidenz abzulehnen oder zu akzeptieren”*. Solche Aussagen sind in einem konkreten Entscheidungskontext hilfreich, wenn tatsächlich eine Entscheidung getroffen werden muss. Allerdings sei dazu angemerkt, dass wie gesehen, Frequentistische Hypothesentests ohne explizite Entscheidungsnutzenfunktion formuliert sind und potentielle Entscheidungskosten damit nicht explizit in die Entscheidungswahl einbezogen werden. Speziell für diesen Zweck gibt es eine Reihe sehr zugänglicher Theorien, die es erlauben, im langfristigen Mittel gute Entscheidungen unter Unsicherheit zu treffen, vgl. zum Beispiel Pratt et al. (1995), Puterman (1994), oder Kochenderfer et al. (2022).

Orientiert man sich von praktisch relevanten Entscheidungskontexten in den Bereich der Grundlagenwissenschaften, deren Wesen es ja gerade ist, keine finalen Wahrheiten zu kennen, sondern nur das Maß an Unsicherheit über den gerade vorherrschenden Theoriestand zu quantifizieren und zu kommunizieren, erscheint die Binarität der Hypothesentestentscheidung im besten Fall überflüssig, im schlimmsten Fall grob irreführend. Prinzipiell sollten Fragestellungen der Grundlagenwissenschaften deshalb gerade nicht als Entscheidungsprobleme formuliert werden. Trotz der weit verbreiteten Meinung, dass Bayesianische Herangehensweisen wie *Positive Predictive Values* oder *Bayes Factors* hier Vorteile bieten würden, ist dem nicht so, so lange die mit einer gewissen Modellpräferenz assoziierte Unsicherheit nicht klar mitkommuniziert wird. Nichtsdestotrotz bleibt das das Frequentistische Hypothesentesten auch in der grundlagenorientierten Wissenschaftsgemeinschaft weiterhin sehr populär, manchmal allerdings nur unter dem Deckmantel der Rufe nach Grundlagenstudien mit "höherer Power". Um einen Zugang zur psychologisch-naturwissenschaftlichen Literatur zu haben, ist es daher bisher unumgänglich, sich auch mit dem grundlagenwissenschaftlich betrachtet eigentlich wenig sinnvollen Hypothesentesten zu beschäftigen.

33.3. Testbeispiele

33.3.1. Einstichproben-T-Test

Das Anwendungsszenario eines Einstichproben-T-Test ist dadurch gekennzeichnet, dass n univariate Datenpunkte einer Stichprobe (Gruppe) randomisierter experimenteller Einheiten betrachtet werden, von denen angenommen wird, dass sie Realisierungen von n unabhängigen und identisch normalverteilten Zufallsvariablen sind. Hinsichtlich der identischen univariaten Normalverteilungen $N(\mu, \sigma^2)$ dieser Zufallsvariablen wird angenommen, dass sowohl der Erwartungswertparameter μ als auch der Varianzparameter σ^2 unbekannt sind. Schließlich wird vorausgesetzt, dass ein Interesse an einem inferentiellen Vergleich des unbekannten Erwartungswertparameters μ mit einem vorgegebenen Wert μ_0 im Sinne eines Hypothesentests besteht.

Dabei gibt es allerdings mindestens vier Szenarien, die potentiell von Interesse sein können. Ein erster Fall wäre das Szenario einer einfachen Nullhypothese und einer einfachen Alternativhypothese,

$$H_0 : \mu = \mu_0 \text{ und } H_1 : \mu = \mu_1 \quad (33.31)$$

Dieser Fall ist in der Theorie sehr gut verstanden und Grundlage des sogenannten *Neymann-Pearson-Lemmas* (Neyman & Pearson (1933)). Seine praktische Relevanz ist aber eher gering, da die Alternativhypothese von einer genauen Spezifikation des Erwartungswertparameters ausgeht. Ein zweiter Fall ist das Szenario einer einfachen Nullhypothese und einer zusammengesetzten Alternativhypothese

$$H_0 : \mu = \mu_0 \text{ und } H_1 : \mu \neq \mu_0 \quad (33.32)$$

In diesem Fall spricht man auch von einer *ungerichteten Hypothese* und nutzt in der Regel einen zweiseitigen Test. Intuitiv entspricht dies der ungerichteten Frage nach inferentieller Evidenz für einen Unterschied. Es ist dieser Fall, den wir im Folgenden detailliert betrachten werden. Schließlich gibt es noch mindestens Szenarien mit zusammengesetzten Null- und Alternativhypothesen, etwa der Form

$$H_0 : \mu \leq \mu_0 \text{ und } H_1 : \mu > \mu_0 \text{ oder } H_0 : \mu \geq \mu_0 \text{ und } H_1 : \mu < \mu_0 \quad (33.33)$$

Man spricht in diesem Fall auch von *gerichteten* Hypothesen und nutzt in der Regel einseitige Tests. Diese Fall betrachten wir im Folgenden jedoch nicht.

Frequentistisches Inferenzmodell

Definition 33.13 (Frequentistisches Inferenzmodell des Einstichproben-T-Tests). Das Frequentistische Inferenzmodell des Einstichproben-T-Tests ist gegeben durch das Normalverteilungsmodell (vgl. Definition 30.2)

$$y_1, \dots, y_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \quad (33.34)$$

•

Wir erinnern daran, dass aus generativer Sicht das Normalverteilungsmodell dem Modell

$$y_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, \dots, n \quad (33.35)$$

entspricht (vgl. Kapitel 30.3.1). Die Annahme unabhängig und identisch normalverteilter Zufallsvariablen als Grundlage der Modellierung der Beobachtung von n Datenpunkten ist wie in Kapitel 30.3.1 gesehen äquivalent zu der Annahme, dass sich jede einen Datenpunkt modellierende Zufallsvariable y_i als Summe aus einem festen, wahren, aber unbekannten, über Zufallsvariablen konstanten Wert μ und aus einem Zufallsvariablen- bzw. Datenpunkt-spezifischen Abweichungsterm ε_i ergibt. Dabei modelliert, wie gesehen, μ den tatsächlichen im wissenschaftlichen Anwendungskontext angenommen Effekt von Interesse und ε_i den Aspekt der Datenvariabilität, der nicht durch diesen Effekt erklärt werden kann und im Sinne des Zentralen Grenzwertsatzes aus der Summation unendlich vieler Störeinflüsse hervorgeht und damit als Unsicherheit über die Erklärung der Datenvariabilität durch den festen Wert μ verbleibt.

Testhypothesen

Wie oben diskutiert betrachten wir hier den Fall des Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese.

Definition 33.14 (Einfache Nullhypothese und zusammengesetzte Alternativhypothese des Einstichproben-T-Tests). Gegeben sei das Frequentistische Inferenzmodell des Einstichproben-T-Tests

$$y_1, \dots, y_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \quad (33.36)$$

und es sei $\Theta := \mathbb{R}$ der Parameterunterraum des Parameters von Interesse μ . Dann sind für den *Nullhypothesenparameterwert* $\mu_0 \in \mathbb{R}$ die *einfache Nullhypothese* und die *zusammengesetzte Alternativhypothese* des Einstichproben-T-Tests gegeben durch

$$\Theta_0 := \{\mu_0\} \Leftrightarrow H_0 : \mu = \mu_0 \text{ und } \Theta_1 := \mathbb{R} \setminus \{\mu_0\} \Leftrightarrow H_1 : \mu \neq \mu_0. \quad (33.37)$$

•

Man beachte, dass die einfache Nullhypothese und die zusammengesetzte Alternativhypothese durch den Wert $\mu_0 \in \mathbb{R}$ parameterisiert sind. Je nach Wahl von μ_0 ergeben sich also verschiedene Hypothesenszenarien. Wird beispielsweise $\mu_0 := 0$ gewählt, so entspricht die Nullhypothese $\Theta_0 := \{0\}$ der Aussage, dass der wahre, aber unbekannte, Parameter μ gleich 0 ist und die Alternativhypothese $\Theta_1 := \mathbb{R} \setminus \{0\}$ der Aussage, dass der wahre, aber unbekannte, Parameter μ ungleich 0 ist. Wird dagegen beispielsweise $\mu_0 := 2$ gewählt, so entspricht die Nullhypothese $\Theta_0 := \{2\}$ der Aussage, dass der wahre, aber unbekannte, Parameter μ gleich 2 ist und die Alternativhypothese $\Theta_1 := \mathbb{R} \setminus \{2\}$ der Aussage, dass der wahre, aber unbekannte, Parameter μ ungleich 2 ist. Im Anwendungskontext ist μ_0 dementsprechend ein frei gewählter und damit natürlich auch bekannter Parameter des Einstichproben-T-Tests ist, wohingegen μ bekanntlich wahr, aber unbekannt ist und bleibt.

Definition der Teststatistik

Mit der *Einstichproben-T-Test-Statistik* definieren wir nun eine Teststatistik, die als Grundlage eines kritischen Wert-basierten Tests dienen kann und deren Betrag eine Abweichung von der Nullhypothese indiziert.

Definition 33.15 (Einstichproben-T-Test-Statistik). Gegeben sei das Testszenario eines Einstichproben-T-Tests mit Stichprobe $y_1, \dots, y_n$, Stichprobenmittel $\bar{y}$, Stichprobenstandardabweichung S und Nullhypothese parameter μ_0 . Dann ist die *Einstichproben-T-Test-Statistik* definiert als

$$T := \sqrt{n} \frac{\bar{y} - \mu_0}{S}. \quad (33.38)$$

•

Offenbar hat die Einstichproben-T-Test-Statistik eine hohe Ähnlichkeit mit der T-Konfidenzintervallstatistik (vgl. Definition 32.2). Man beachte allerdings, dass im Fall der Einstichproben-T-Test-Statistik der Nullhypothese parameter μ_0 nicht identisch mit dem in der T-Konfidenzintervallstatistik auftauchendem wahren, aber unbekannten, Parameterwert μ sein muss.

Da die Einstichproben-T-Test-Statistik im Kontext des Einstichproben-T-Tests zentral ist, macht es Sinn, sich ihrer intuitiven Mechanik bewusst zu sein. Im Zähler des Bruches der Einstichproben-T-Test-Statistik tritt zunächst die Differenz des Stichprobenmittels $\bar{y}$ zum angenommenen Nullhypothese parameter μ_0 auf. Wie gesehen ist das Stichprobenmittel ein unverzerrter Schätzer des Erwartungswertparameters μ der Stichprobe. Die Differenz $\bar{y} - \mu_0$ entspricht also einer Schätzung der Abweichung des wahren, aber unbekannten, Erwartungswertparameters vom Nullhypothese parameter und damit dem Betrage nach der Evidenz für eine Abweichung des wahren, aber unbekannten, Erwartungswertparameters von der Nullhypothese. Grob betrachtet hat man mit dem Zähler $\bar{y} - \mu_0$ also ein Maß für das der Stichprobe innewohnende “Signal” im Sinne der Abweichung von der Nullhypothese oder “systematischer Variabilität”. Der Nenner S erlaubt es dann, dieses Signal in Einheiten der Stichprobenstandardabweichung auszudrücken. Gilt zum Beispiel $\bar{y} - \mu_0 = 2$ und ist $S = 1$, so beträgt die Abweichung des Stichprobenmittels vom Nullhypothese parameter gerade zwei Standardabweichungen, ist dagegen $S = 2$ so beträgt die entsprechende Abweichung gerade eine Standardabweichung. Weiterhin entspricht der Nenner S ja

einem Maß für die beobachtete Datenvariabilität und einem Schätzer für die Standardabweichung σ der Fehlerterme in der generativen Form des Einstichproben-T-Test Modells. Grob betrachtet hat man also im Nenner der Einstichproben-T-Test-Statistik ein Maß für das den Daten innewohnende “Rauschen” oder ihrer “unsystematischen Variabilität”. Insgesamt kann man den Bruch $\frac{\bar{y}-\mu_0}{S}$ also als eine Schätzung des “Signal-zu-Rauschen-Verhältnis” der Daten verstehen. Schließlich wird in der Einstichproben-T-Test-Statistik dieses Verhältnis mit der Wurzel der Stichprobengröße $\sqrt{n}$ gewichtet. Intuitiv entspricht diese Wichtung der Tatsache, dass man einem gegebenen Signal-zu-Rauschen-Verhältnis mehr Validität zumessen kann, wenn es auf einer höheren Anzahl von Datenpunkten basiert, als wenn es auf einer geringen Anzahl von Datenpunkten basiert. Insgesamt hat man mit der Einstichproben-T-Test-Statistik eine skalare Zusammenfassung der den Daten innewohnenden Evidenz gegen die Nullhypothese, bei der sowohl die Datenvariabilität als auch der Datenumfang betrachtet werden.

Verteilung der Teststatistik

Für die Verteilung der Einstichproben-T-Test-Statistik gilt nun folgendes Theorem.

Theorem 33.1 (Verteilung der Einstichproben-T-Test-Statistik). *Gegeben sei das Testszenario eines Einstichproben-T-Tests mit Stichprobe $y_1, \dots, y_n$, Stichprobenmittel $\bar{y}$, Stichprobenstandardabweichung S , Nullhypothese parameter μ_0 und Einstichproben-T-Test-Statistik definiert als*

$$T := \sqrt{n} \frac{\bar{y} - \mu_0}{S}. \quad (33.39)$$

Dann ist T eine nichtzentrale t -Zufallsvariable mit Nichtzentralitätsparameter

$$d = \sqrt{n} \frac{\mu - \mu_0}{\sigma} \quad (33.40)$$

und Freiheitsgradparameter $n - 1$, es gilt also $T \sim t(d, n - 1)$

◦

Man beachte, dass im Falle des Zutreffens der Nullhypothese der Nullhypothese parameter μ_0 mit dem wahren, aber unbekannten, Erwartungswertparameter μ identisch ist und der Nichtzentralitätsparameter der Verteilung der Einstichproben-T-Test-Statistik den Wert $d = 0$ annimmt. Im Falle des Zutreffens der Nullhypothese des Einstichproben-T-Testszenarios ist die Einstichproben-T-Test-Statistik also eine t -verteilte Zufallsvariable mit Freiheitsgradparameter $n - 1$. Wir visualisieren die Verteilung der Einstichproben-T-Test-Statistik exemplarisch für ein Einstichproben-T-Testszenario mit $n = 12$, wahren, aber unbekannten, Parametern $\mu = 3$ und $\sigma^2 = 2$ und Nullhypothese parameter $\mu_0 = 0$ in Abbildung 33.2 (B), Die Parameter dieser Verteilung ergeben sich dabei zu

$$d = \sqrt{n} \frac{\mu - \mu_0}{\sigma} = \sqrt{12} \frac{3 - 0}{\sqrt{2}} \approx 7.34 \text{ und } n - 1 = 11 \quad (33.41)$$

ergeben.

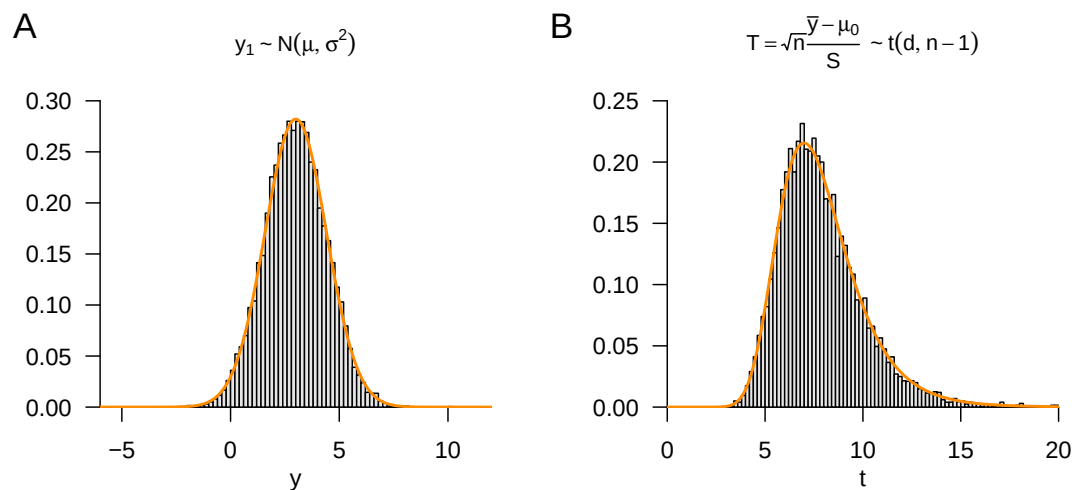


Abbildung 33.2. Verteilung der Einstichproben-T-Test-Statistik für $n = 12$, $\mu = 3$, σ^2 und $\mu_0 = 0$. (A) Verteilung der Stichprobenvariablen. (B) Verteilung der Einstichproben-T-Test-Statistik

Testdefinition

Wir können nun den zweiseitigen Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese definieren und seine Testgütefunktion analysieren.

Definition 33.16 (Zweiseitiger Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese). Gegeben seien das Frequentistische Inferenzmodell des Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese und T bezeichne die Einstichproben-T-Test-Statistik mit Werten $t \in \mathbb{R}$. Dann ist der *zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese* definiert als der zweiseitige kritische Wert-basierte Test

$$\phi : \mathcal{Y} \rightarrow \{0, 1\}, y \mapsto \phi(y) := 1_{\{|t| \geq k\}} = \begin{cases} 1 & |t| \geq k \\ 0 & |t| < k \end{cases}. \quad (33.42)$$

•

Der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese nimmt also den Wert 0 an, wenn der Betrag der Einstichproben-T-Test-Statistik kleiner als der kritische Wert ist und er nimmt den Wert 1 an, wenn der Betrag der Einstichproben-T-Test-Statistik gleich oder größer als der kritische Wert ist.

Testgütefunktion

Für die Kontrolle des Testumfangs durch Wahl eines kritischen Werts und zur Bestimmung der Powerfunktion dieses Tests ist nun folgendes Theorem maßgeblich.

Theorem 33.2 (Testgütefunktion des zweiseitigen Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese). ϕ sei der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. Dann ist die Testgütefunktion von ϕ gegeben durch

$$q_\phi : \mathbb{R} \rightarrow [0, 1], \mu \mapsto q_\phi(\mu) := 1 - \Psi(k; d_\mu, n-1) + \Psi(-k; d_\mu, n-1), \quad (33.43)$$

wobei $\Psi(\cdot; d_\mu, n-1)$ die KVF der nichtzentralen t -Verteilung mit Nichtzentralitätsparameter

$$d_\mu := \sqrt{n} \frac{\mu - \mu_0}{\sigma} \quad (33.44)$$

und Freiheitsgradparameter $n-1$ bezeichnet.

◦

Beweis. Die Testgütefunktion des betrachteten Test im vorliegenden Testszenario ist definiert als

$$q_\phi : \mathbb{R} \rightarrow [0, 1], \mu \mapsto q_\phi(\mu) := \mathbb{P}_\mu(\phi = 1). \quad (33.45)$$

Da die Wahrscheinlichkeiten für $\phi = 1$ und dafür, dass die zugehörige Teststatistik im Ablehnungsbereich des Tests liegt gleich sind, benötigen wir also zunächst die Verteilung der Teststatistik. Wir haben oben bereits gesehen, dass die Einstichproben-T-Test-Statistik

$$T := \sqrt{n} \frac{\bar{y} - \mu_0}{S} \quad (33.46)$$

unter der Annahme $y_1, \dots, y_n \sim N(\mu, \sigma^2)$ anhand einer nichtzentralen t -Verteilung $t(d_\mu, n-1)$ mit Nichtzentralitätsparameter

$$d_\mu := \sqrt{n} \frac{\mu - \mu_0}{\sigma} \quad (33.47)$$

verteilt ist. Der Ablehnungsbereich des zweiseitigen Einstichproben-T-Tests ist

$$A =]-\infty, -k] \cup]k, \infty[. \quad (33.48)$$

Mit diesem Ablehnungsbereich ergibt sich dann

$$\begin{aligned} q_\phi(\mu) &= \mathbb{P}_\mu(\phi = 1) \\ &= \mathbb{P}_\mu(T \in]-\infty, -k] \cup]k, \infty[) \\ &= \mathbb{P}_\mu(T \in]-\infty, -k]) + \mathbb{P}_\mu(T \in [k, \infty[) \\ &= \mathbb{P}_\mu(T \leq -k) + \mathbb{P}_\mu(T \geq k) \\ &= \mathbb{P}_\mu(T \leq -k) + (1 - \mathbb{P}_\mu(T \leq k)) \\ &= 1 - \mathbb{P}_\mu(T \leq k) + \mathbb{P}_\mu(T \leq -k) \\ &= 1 - \Psi(k; d_\mu, n-1) + \Psi(-k; d_\mu, n-1), \end{aligned} \quad (33.49)$$

wobei $\Psi(\cdot; d_\mu, n-1)$ die KVF der nichtzentralen T-Verteilung mit Nichtzentralitätsparameter d_μ und Freiheitsgradparameter $n-1$ bezeichnet. $\square$

In Abbildung 33.3 visualisieren wir die Testgütefunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese aus Theorem 33.2 für $\sigma^2 = 9$ und $\mu_0 = 4$ in Abhängigkeit vom kritischen Wert k . Man beachte dabei zunächst, dass die Testgütefunktion als Funktion von μ sowohl das Szenario des Zutreffens der Nullhypothese $\mu = \mu_0$ als auch das Szenario des Zutreffens der Alternativhypothese $\mu \neq \mu_0$ abdeckt. Man beachte weiterhin, dass der Wert der Testgütefunktion, also die Wahrscheinlichkeit dafür, dass der Test den Wert 1 annimmt, sowohl bei positiven als auch bei negativen Abweichungen des wahren, aber unbekannten, Erwartungswertparameters μ vom Nullhypothesenparameter μ_0 ansteigt. Dies ist natürlich der Tatsache geschuldet ist, dass die Testentscheidung auf dem Betrag der Teststatistik

beruht. Schließlich ist die genaue Form und Lage der Testgütefunktion von der Wahl des kritischen Werts k abhängig. Wird dieser größer gewählt, ist also ein größerer absoluter Wert der Teststatistik für dafür notwendig, dass der Test den Wert 1 annimmt, so ist die Wahrscheinlichkeit dafür, bei ansonsten konstanten Parametern, kleiner als bei kleineren Werten des kritischen Werts.

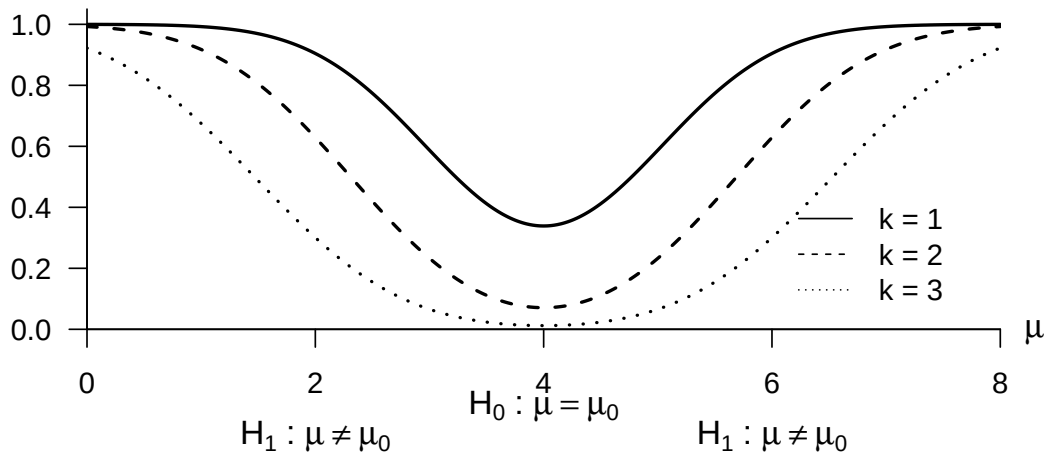


Abbildung 33.3. Testgütefunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese für $\sigma^2 = 9$, $\mu_0 = 4$, $n = 12$ und $k = 1, 2, 3$.

Testumfangkontrolle

Die Werte der Testgütefunktion bei $\mu = \mu_0$ in Abbildung 33.3 geben einen visuellen Eindruck davon, wie der kritische Wert den Testumfang kontrolliert. Die exakte Bestimmung des kritischen Werts bei einem gewünschten Testumfang ist Inhalt folgenden Theorems.

Theorem 33.3 (Testumfangkontrolle für den zweiseitigen Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese). *ϕ sei der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. Dann ist ϕ ein Level- α_0 -Test mit Testumfang α_0 , wenn der kritische Wert definiert ist durch*

$$k_{\alpha_0} := \Psi^{-1} \left(1 - \frac{\alpha_0}{2}; n - 1 \right), \quad (33.50)$$

wobei $\Psi^{-1}(\cdot; n - 1)$ die inverse KVF der t -Verteilung mit Freiheitsgradparameter $n - 1$ bezeichnet.

◦

Beweis. Damit der betrachtete Test ein Level- α_0 -Test ist, muss bekanntlich $q_\phi(\mu) \leq \alpha_0$ für alle $\mu \in \{\mu_0\}$, also hier $q_\phi(\mu_0) \leq \alpha_0$, gelten. Weiterhin ist der Testumfang des betrachteten Tests durch $\alpha = \max_{\mu \in \{\mu_0\}} q_\phi(\mu)$, also hier durch $\alpha = q_\phi(\mu_0)$ gegeben. Wir müssen also zeigen, dass die Wahl von k_{α_0}

garantiert, dass ϕ ein Level- α_0 -Test mit Testumfang α_0 ist. Dazu merken wir zunächst an, dass für $\mu = \mu_0$ gilt, dass

$$\begin{aligned} q_\phi(\mu_0) &= 1 - \Psi(k; d_{\mu_0}, n-1) + \Psi(-k; d_{\mu_0}, n-1) \\ &= 1 - \Psi(k; 0, n-1) + \Psi(-k; 0, n-1) \\ &= 1 - \Psi(k; n-1) + \Psi(-k; n-1), \end{aligned} \quad (33.51)$$

wobei $\Psi(\cdot; d, n-1)$ und $\Psi(\cdot; n-1)$ die KVF der nichtzentralen t -Verteilung mit Nichtzentralitätsparameter d und Freiheitsgradparameter $n-1$ sowie der t -Verteilung mit Freiheitsgradparameter $n-1$, respektive, bezeichnen. Sei nun also $k := k_{\alpha_0}$. Dann gilt

$$\begin{aligned} q_\phi(\mu_0) &= 1 - \Psi(k_{\alpha_0}; n-1) + \Psi(-k_{\alpha_0}; n-1) \\ &= 1 - \Psi(k_{\alpha_0}; n-1) + (1 - \Psi(k_{\alpha_0}; n-1)) \\ &= 2(1 - \Psi(k_{\alpha_0}; n-1)) \\ &= 2 \left(1 - \Psi \left(\Psi^{-1} \left(1 - \frac{\alpha_0}{2}, n-1 \right), n-1 \right) \right) \\ &= 2 \left(1 - 1 + \frac{\alpha_0}{2} \right) \\ &= \alpha_0, \end{aligned} \quad (33.52)$$

wobei die zweite Gleichung mit der Symmetrie der t -Verteilung folgt. Es folgt also direkt, dass bei der Wahl von $k = k_{\alpha_0}$, $q_\phi(\mu_0) \leq \alpha_0$ ist und der betrachtete Test somit ein Level- α_0 -Test ist. Weiterhin folgt direkt, dass der Testumfang des betrachteten Tests bei der Wahl von $k = k_{\alpha_0}$ gleich α_0 ist. $\square$

Man beachte, dass nach Theorem 33.3 der hier betrachtete Tests insbesondere exakt ist, der Testumfang also mit dem Signifikanzlevel identisch ist. In Abbildung 33.4 visualisieren wir die Wahl des kritischen Werts k_{α_0} in einem zweiseitigen Einstichproben-T-Test-Szenario mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese für $\alpha_0 := 0.05$ und $n = 12$.

Folgender **R** Code demonstriert die Bestimmung des kritischen Werts des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese mithilfe der inversen KVF der t -Verteilung, die in **R** als die Funktion `qt()` implementiert ist. Darüberhinaus simuliert der Code 10^6 Stichprobenrealisationen für das hier betrachteten Testszenario bei Zutreffen der Nullhypothese und wertet den betrachteten Test aus. Es zeigt sich, dass die geschätzte Wahrscheinlichkeit dafür, dass der Test bei Zutreffen der Nullhypothese den Wert 1 annimmt mit dem gewünschten Wert von $\alpha_0 = 0.05$ sehr gut übereinstimmt.

```
# Modellparameter
n      = 12                                # Anzahl der Datenpunkte
mu     = 0                                # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr = 2                                # wahrer, aber unbekannter, Varianzparameter

# Testparameter
mu_0    = 0                                # Nullhypothese parameter, hier \mu = \mu_0
alpha_0 = 0.05                             # Signifikanzlevel
k_alpha_0 = qt(1-alpha_0/2, n-1)           # Kritischer Wert

# Simulation der Testumfangkontrolle
set.seed(1)                                # Random number generator seed
nsim    = 1e6                              # Anzahl Simulationen
phi      = rep(NA, nsim)                   # Testentscheidungsarray
for(j in 1:nsim){
  y      = rnorm(n, mu, sigsqr)             # Simulationsiterationen
  y_bar  = mean(y)                         # y_i \sim N(\mu, \Sigma), i = 1, ..., n
  s      = sd(y)                           # Stichprobenmittel
  Tee    = sqrt(n)*(y_bar - mu_0)/s         # Stichprobenstandardabweichung
  if(abs(Tee) > k_alpha_0){                 # Einstichproben-T-Test-Statistik
    phi[j] = 1                             # Test 1_{|t| \ge k_alpha_0}
  } else {                                 # Ablehnen der Nullhypothese
    phi[j] = 0                             # Nichtablehnen der Nullhypothese
  }
}

# Ausgabe
cat("Kritischer Wert      =", k_alpha_0,
    "\nGeschätzter Testumfang alpha =", mean(phi))
```

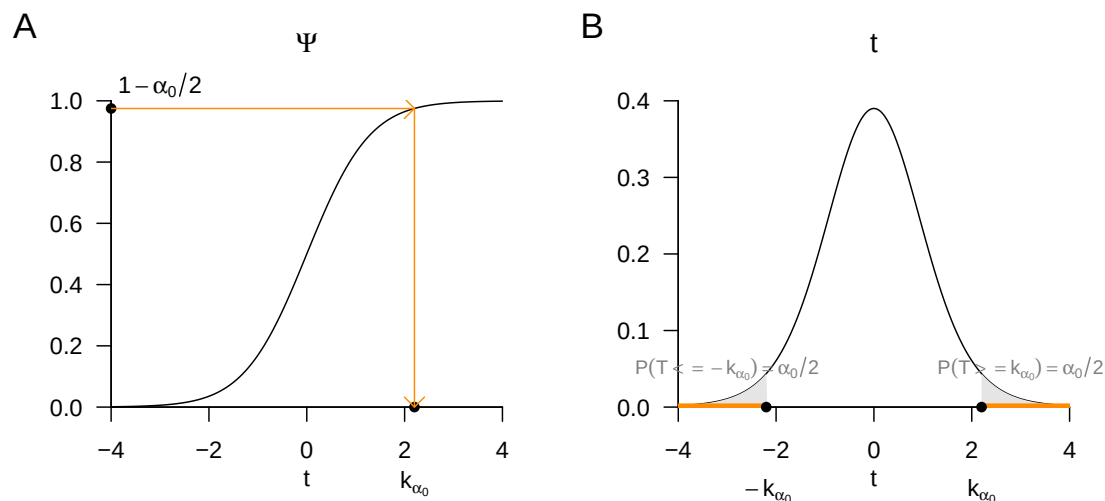


Abbildung 33.4. Bestimmung des kritischen Werts k_{α_0} für den zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese für $n = 12$ und $\alpha_0 := 0.05$ zu Kontrolle des Testumfangs (A) Für ein gewähltes α_0 ist der kritische Wert des betrachteten Tests nach Theorem 33.3 durch den Wert der inversen KVF der t -Verteilung mit Freiheitsgradparameter $n - 1$ an der Stelle $1 - \frac{\alpha_0}{2}$ gegeben. Die Abbildung zeigt die Bestimmung dieses Werts zu $k_{0.05} = 2.2$ anhand der KVF der t -Verteilung mit Freiheitsgradparameter $n - 1$. (B) Die Abbildung zeigt den aus der Wahl von k_{α_0} resultierenden Ablehnungsbereich des betrachteten Tests als grau hinterlegte Flächen unter der WDF der t -Verteilung mit Freiheitsgradparameter $n - 1$. Die symmetrische Verteilung der Teilmengen des Ablehnungsbereichs in den Ausläufern der WDF ergibt sich dabei aus der Definition des Tests als Funktion des Betrages der Einstichproben-T-Test-Statistik, also insbesondere des zweiseitigen Charakters des hier betrachteten Tests.

Kritischer Wert = 2.200985
Geschätzter Testumfang alpha = 0.049755

p-Wert

Der mit einem vorliegenden Wert der Teststatistik des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese assoziierte p-Wert ergibt aus folgendem Theorem wie folgt.

Theorem 33.4 (p-Wert des zweiseitigen Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese). *Gegeben sei der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese und t sei ein Wert der Einstichproben-T-Test-Statistik T . Dann gilt*

$$p\text{-Wert} = 2(1 - \Psi(|t|; n - 1)) \quad (33.53)$$

wobei $\Psi(\cdot; n - 1)$ die KVF der t -Verteilung mit Freiheitsgradparameter $n - 1$ bezeichnet.

◦

Beweis. Nach Definition 33.12 ist der p-Wert das kleinste Signifikanzlevel α_0 bei für den betrachteten Test die Nullhypothese basierend auf dem Wert von t abgelehnt werden würde. Im vorliegenden Fall würde die Nullhypothese für jedes α_0 mit

$$|t| \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n - 1\right) \quad (33.54)$$

abgelehnt werden, vgl. Theorem 33.3. Für diese α_0 gilt, dass

$$\alpha_0 \geq 2(1 - \Psi(|t|; n - 1)), \quad (33.55)$$

denn

$$\begin{aligned} |t| &\geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n - 1\right) \\ \Leftrightarrow \Psi(|t|; n - 1) &\geq \Psi\left(\Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n - 1\right); n - 1\right) \\ \Leftrightarrow \Psi(|t|; n - 1) &\geq 1 - \frac{\alpha_0}{2} \\ \Leftrightarrow \mathbb{P}(T \leq |t|) &\geq 1 - \frac{\alpha_0}{2} \\ \Leftrightarrow \frac{\alpha_0}{2} &\geq 1 - \mathbb{P}(T \leq |t|) \\ \Leftrightarrow \frac{\alpha_0}{2} &\geq \mathbb{P}(T \geq |t|) \\ \Leftrightarrow \alpha_0 &\geq 2\mathbb{P}(T \geq |t|) \\ \Leftrightarrow \alpha_0 &\geq 2(1 - \Psi(|t|; n - 1)). \end{aligned} \quad (33.56)$$

Das kleinste $\alpha_0 \in [0, 1]$ mit

$$\alpha_0 \geq 2\mathbb{P}(T \geq |t|) \quad (33.57)$$

ist dann entsprechend

$$\alpha_0 = 2(1 - \Psi(|t|; n - 1)). \quad (33.58)$$

□

In Abbildung 33.5 visualisieren wir die Bestimmung von p-Werten für $t = 2.26$ und für $t = 3.81$, welche sich zu $p = 0.045$ und $p = 0.003$, respektive, ergeben. Man beachte, dass zum Beispiel der p-Wert zu $t = -2.26$ auch $p = 0.045$ beträgt.

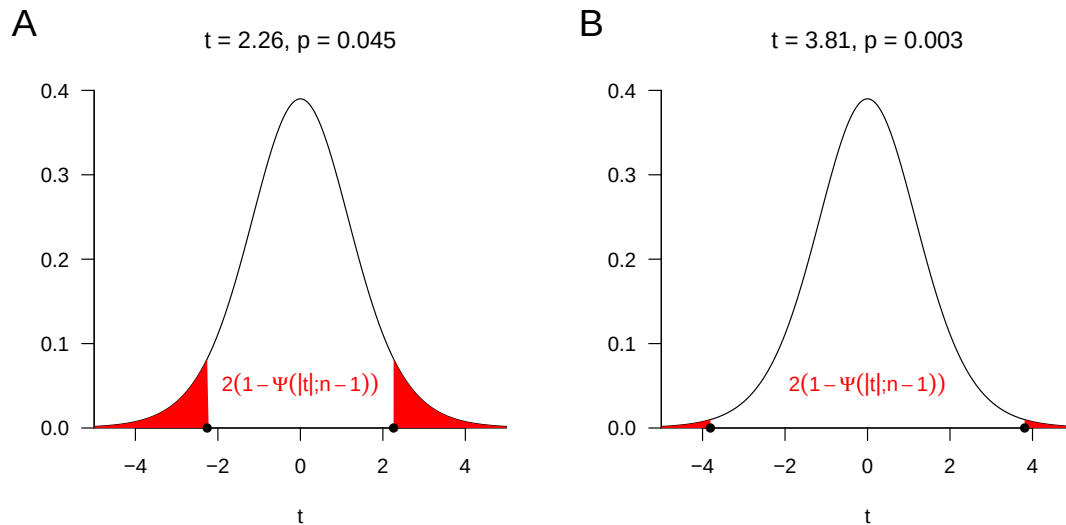


Abbildung 33.5. p-Werte für zwei mögliche Werte der Einstichproben-T-Test-Statistik bei $n = 12$. Die Bereiche im Ergebnisraum der Einstichproben-T-Test-Statistik über denen roten Flächen eingezeichnet sind, markieren die Bereiche mit $T \geq |t|$ gilt. Die summierten ihnen zugeordneten Wahrscheinlichkeiten, also die entsprechenden roten Flächen unter der WDF der t -Verteilungen entsprechen dem p-Wert

Powerfunktion

Die Powerfunktion eines Tests entspricht der Testgütefunktion eines Tests für den Bereich des Parameterraums, der der Alternativhypothese entspricht. Änderungen im Wert der

Powerfunktion eines Tests, oft einfach als *Power des Tests* bezeichnet, ergeben sich also zunächst einmal durch Änderungen des Wertes des wahren, aber unbekannten, Parameters im Bereich der Alternativhypothese. Allerdings hat es sich eingebürgert, die Wahrscheinlichkeit dafür, dass der Test den Wert 1 annimmt, also die Nullhypothese abgelehnt wird, nicht ausschließlich als Funktion des wahren, aber unbekannten, Parameters, sondern auch weiterer und in der praktischen Anwendung relevanter Parameter eines Testszenarios zu betrachten. An erster Stelle ist hier der Stichprobenumfangs n von Interesse. Im Kontext der praktischen Durchführung von *Poweranalysen* fragt man dann meist danach, welcher Stichprobenumfang bei Annahme eines Wertes für den wahren, aber unbekannten, Parameter im Bereich der Alternativhypothese mit einer bestimmten Wahrscheinlichkeit dafür, die Nullhypothese abzulehnen, assoziiert ist. Basierend auf der asymmetrischen Behandlung von Typ I und Typ II Fehlerwahrscheinlichkeiten (vgl. Kapitel 33.2) setzt man dahingehend zunächst ein Signifikanzlevel α_0 zur Kontrolle des Testumfangs fest. Für den hier diskutierten zweiseitigen Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese betrachten wir also die Testgütefunktion

$$q_\phi : \mathbb{R} \rightarrow [0, 1], \mu \mapsto q_\phi(\mu) := 1 - \Psi(k_{\alpha_0}; d_\mu, n - 1) + \Psi(-k_{\alpha_0}; d_\mu, n - 1) \quad (33.59)$$

bei kontrolliertem Testumfang, also für

$$k_{\alpha_0} := \Psi^{-1} \left(1 - \frac{\alpha_0}{2}; n - 1 \right) \quad (33.60)$$

mit festem α_0 als Funktion des Nichtzentralitätsparameters d und des Stichprobenumfangs n . Insbesondere hängt dabei der Nichtzentralitätsparameter d vom Verhältnis der wahren, aber unbekannten, Parameter μ und σ^2 , also dem wahren, aber unbekannten, Signal-zu-Rauschen-Verhältnis des Testszenarios und k_{α_0} von n ab. Diese Überlegungen führen auf folgende Definition.

Definition 33.17 (Powerfunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese). Gegeben sei der zweiseitige Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. Dann ist die *Powerfunktion* des Tests gegeben durch

$$\pi : \mathbb{R} \times \mathbb{N} \rightarrow [0, 1], (d, n) \mapsto \pi(d, n) := 1 - \Psi(k_{\alpha_0}; d, n - 1) + \Psi(-k_{\alpha_0}; d, n - 1) \quad (33.61)$$

•

Folgender **R** Code demonstriert exemplarisch die Auswertung der Powerfunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese in einem Szenario mit $\alpha_0 := 0.05$ mithilfe der **R** Implementationen der KVF und der inversen KVF der nichtzentralen t -Verteilung `pt()` und `qt()`, respektive.

```
alpha_0 = 0.05 # Signifikanzlevel
d_min = -5 # minimaler Nichtzentralitätsparameter
d_max = 5 # maximaler Nichtzentralitätsparameter
d_res = 50 # Auflösung Nichtzentralitätsparameter
d = seq(d_min, d_max, len = d_res) # Nichtzentralitätsparametertraum
n_min = 1 # minimaler Stichprobenumfang
n_max = 30 # maximaler Stichprobenumfang
n_res = 50 # Auflösung Stichprobenumfang
n = seq(n_min, n_max, len = n_res) # maximaler Stichprobenumfang
pi = matrix(rep(NA, d_res*n_res), nrow = d_res) # Powerfunktionsarray
for(i in 1:d_res){ # Nichtzentralitätsparameteriterationen
  for(j in 1:n_res){ # Stichprobenumfangiterationen
    k_alpha_0 = qt(1 - alpha_0/2, n[j]-1) # kritischer Wert
    pi[i,j] = 1-pt(k_alpha_0, n[j]-1, d[i])+pt(-k_alpha_0, n[j]-1, d[i]) # Auswertung der Powerfunktion
  }
}
```

Wir visualisieren die Abhängigkeit der Powerfunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese vom Nichtzentralitätsparameter und Stichprobenumfang in Abbildung 33.6 und Abbildung 33.7. Generell steigt die Powerfunktion des betrachteten Tests mit positiver oder negativer Abweichung des Nichtzentralitätsparameters vom Nullhypothesenszenario $d = 0$ und steigendem Stichprobenumfang n monoton. Je nach Wahl des Signifikanzlevels erfolgt dieser Anstieg steiler oder weniger steil.

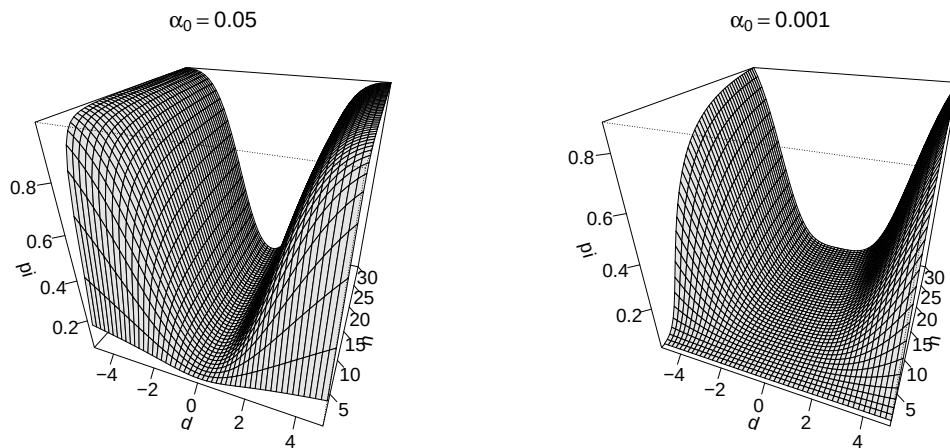


Abbildung 33.6. Powerfunktionen des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. (A) Abhängigkeit der Powerfunktion π vom Nichtzentralitätsparameter d und Stichprobenumfang n bei Wahl von $\alpha_0 := 0.05$. (B) Abhängigkeit der Powerfunktion π vom Nichtzentralitätsparameter d und Stichprobenumfang n bei Wahl von $\alpha_0 := 0.001$.

Praktische Durchführung

Vor dem Hintergrund der in den bisherigen Abschnitten diskutierten Theorie des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese ergibt sich dann zunächst folgendes routiniertes Durchführen des Tests im Rahmen einer Datenanalyse.

Man nimmt an, dass ein vorliegender univariater Datensatz $y_1, \dots, y_n$ eine Realisierung des Frequentistischen Inferenzmodells $y_1, \dots, y_n \sim N(\mu, \sigma^2)$ des Einstichproben-T-Tests mit wahren, aber unbekannten, Parameter μ und $\sigma^2 > 0$ ist. Man nimmt ferner an, dass man entscheiden muss ob für einen gewählten Nullhypothesenparameter μ_0 eher die Nullhypothese $H_0 : \mu = \mu_0$ oder die Alternativhypothese $H_1 : \mu \neq \mu_0$ zutrifft. Um den Testumfang über viele Wiederholungen dieser Testprozedur zu kontrollieren, wählt ein Signifikanzlevel α_0 und bestimmt den zugehörigen kritischen Wert k_{α_0} , so dass zum Beispiel bei einem Stichprobenumfang von $n = 12$ und der Wahl von $\alpha_0 := 0.05$ ein kritischer Wert von $k_{0.05} = 2.20$ gewählt wird. Anhand des Stichprobenumfangs n , des Nullhypothesenparameters μ_0 , des Stichprobenmittels $\bar{y}$ und der Stichprobenstandardabweichung s berechnet man sodann den Wert der Einstichproben-T-Test-Statistik durch

$$t := \sqrt{n} \left(\frac{\bar{y} - \mu_0}{s} \right). \quad (33.62)$$

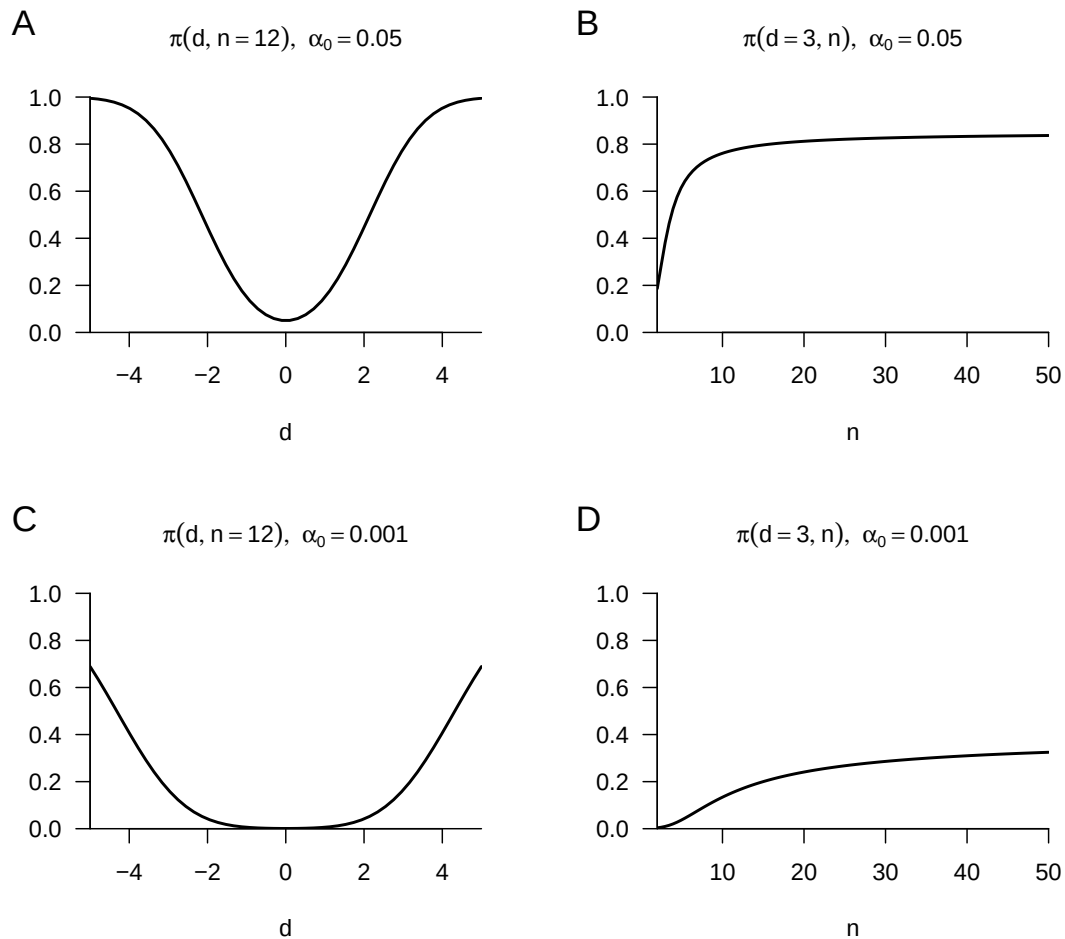


Abbildung 33.7. Schnitte der Powerfunktionen des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. (A) Abhängigkeit der Powerfunktion π vom Nichtzentralitätsparameter d bei konstantem Stichprobenumfang $n = 12$ und Wahl von $\alpha_0 := 0.05$. (B) Abhängigkeit der Powerfunktion vom Stichprobenumfang n π bei Nichtzentralitätsparameter $d = 3$ und Wahl von $\alpha_0 := 0.05$. (C) Abhängigkeit der Powerfunktion π vom Nichtzentralitätsparameter d bei konstantem Stichprobenumfang $n = 12$ und Wahl von $\alpha_0 := 0.001$. (D) Abhängigkeit der Powerfunktion vom Stichprobenumfang n π bei Nichtzentralitätsparameter $d = 3$ und Wahl von $\alpha_0 := 0.001$.

Wenn dieses für den vorliegenden Datensatz so bestimmte t größer als k_{α_0} ist oder wenn t kleiner als $-k_{\alpha_0}$ ist, lehnt man die Nullhypothese ab, andernfalls lehnt man sie nicht ab. Die oben entwickelte Theorie garantiert dann, dass man im langfristigen Mittel in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt. Weiterhin bestimmt man basierend auf dem vorliegenden Wert der Einstichproben-T-Test-Statistik den zugehörigen p-Wert durch

$$\text{p-Wert} = 2(1 - \Psi(|t|; n - 1)) \quad (33.63)$$

Folgender **R** Code demonstriert dieses Vorgehen bei Annahme eines vorliegenden Datenvektors y der Länge n .

```
n      = length(y)           # Stichprobenumfang
mu_0   = 0                  # Nullhypothese parameter
alpha_0 = 0.05              # Signifikanzlevel
k_alpha_0 = qt(1-alpha_0/2, n-1) # kritischer Wert
Tee     = sqrt(n)*((mean(y) - mu_0)/sd(y)) # Einstichproben-T-Test-Statistik
if(abs(Tee) > k_alpha_0){phi = 1} else {phi = 0} # Testauswertung
p = 2*(1 - pt(Tee, n-1))    # p-Wert Evaluation
```

Will man im Rahmen einer Studienplanung eine Poweranalyse zur Optimierung des Stichprobenumfangs im vorliegenden Testszenario durchführen, so gilt natürlich zunächst grundsätzlich, dass mit steigendem Stichprobenumfang die Powerfunktion des Tests ansteigt. Vor dem Gesichtspunkt der Power des Tests ist ein größerer Stichprobenumfang also immer besser als ein kleinerer Stichprobenumfang. Allerdings bleiben dabei mögliche Kosten für die Erhöhung des Stichprobenumfangs, wie zum Beispiel mögliche Risiken für die Studienteilnehmer:innen, unberücksichtigt. Weiterhin ist der Wert, den die Powerfunktion bei einem gewählten Stichprobenumfang immer von den wahren, aber unbekannten, Parameterwerten μ und σ , die in den Wert des Nichtzentralitätsparameters d einfließen, abhängig. Würde man diese Werte in einem gegebenen Anwendungskontext schon sehr genau kennen, so würde man vermutlich keine Studie durchführen wollen. Generell wird im Rahmen der Studienplanung deshalb folgendes Vorgehen favorisiert. Zunächst entscheidet man sich für ein Signifikanzlevel α_0 zur Kontrolle des Testumfangs und evaluiert die Powerfunktion. Man überlegt sich dann einen Nichtzentralitätswert d^* , den man mit einer Power von mindestens β detektieren möchte, wobei ein typischer konventioneller $\beta = 0.8$ ist. Man wertet dann die für einen Powerfunktionswert

$$\pi(d = d^*, n) = \beta \quad (33.64)$$

nötige Stichprobengröße aus. Aufgrund der Monotonie der Powerfunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese im Bereich nicht-negativer Nichtzentralitätsparameter ist dann gewährleistet, dass die Power des Tests für Nichtzentralitätsparameter, die größer als d^* sind, größer oder gleich β sind. Folgender **R** Code implementiert dieses Vorgehen zur Optimierung des Stichprobenumfangs und Abbildung 33.8 visualisiert es.

```
# Powerfunktionsbasierte Stichprobenumfangsoptimierung
alpha_0 = 0.05          # Signifikanzlevel
beta     = 0.8           # gewünschter Powerfunktionswert
d_stern  = 3             # fester Nichtzentralitätsparameter
n_min    = 2             # minimal betrachteter Stichprobenumfang
n_max    = 20            # maximal betrachteter Stichprobenumfang
n_res    = 1e2           # Auflösung des Stichprobenumfangsraums
n        = seq(n_min, n_max, len = n_res) # Stichprobenumfangraum
k_alpha_0 = qt(1-alpha_0/2, n-1) # kritische Werte in Abhängigkeit vom Stichprobenumfang
pi_n     = 1-pt(k_alpha_0, n-1, d_stern)+pt(-k_alpha_0, n-1, d_stern) # Powerfunktion bei festem Nichtzentralitätsparameter
i        = 1             # Indexinitialisierung
n_min    = NaN           # minimales n Initialisierung
while(pi_n[i] < beta){    # Solange \pi(d*, n) < \beta
  n_min = n[i]           # Aufnahme des minimal nötigen ns
  i     = i + 1          # und Erhöhung des Indexes
}
cat("Minimal nötiges n =", ceiling(n_min)) # Ausgabe
```

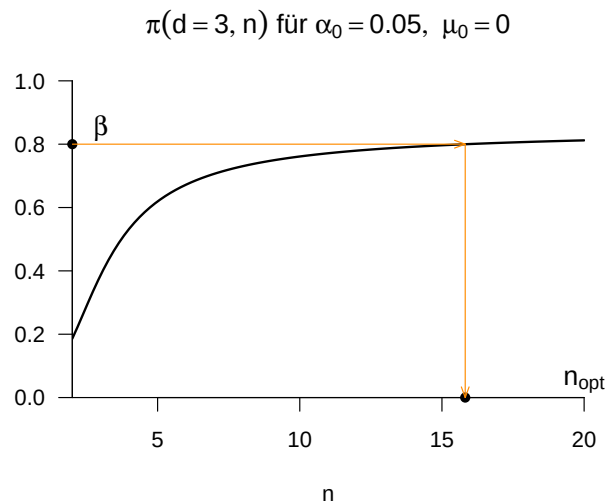

Minimal nötiges $n = 16$ 

Abbildung 33.8. Praktische Durchführung einer Powerfunktionsbasierten Stichprobenumfangoptimierung. Bei gewählten Signifikanzlevel α_0 und fest angenommenen Nichtzentralitätsparameter d^* evaluiert man den Stichprobenumfang für den die Testpowerfunktion einen Wert von β oder größer hat.

Anwendungsbeispiel

Abschließen wollen wir oben skizziertes Vorgehen zur Durchführung eines zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese noch an dem in Kapitel 30.3.1 eingeführten Anwendungsbeispiel demonstrieren. Inhaltlich entspricht in diesem Fall die einfache Nullhypothese $H_0 : \mu = 0$ der Hypothese, dass die Therapie keinen Effekt auf BDI-II Reduktionsscores hat und die zusammengesetzte Alternativhypothese $H_1 : \mu \neq 0$, dass die Therapie einen systematischen von Null verschiedenen Effekt auf BDI-II Reduktionsscores hat. Untenstehender **R** Code wendet das oben demonstrierte Verfahren zur Evaluation der Hypothesen auf den Prä-Post-Therapie BDI-II Reduktionsscore Datensatz von $n = 12$ Patient:innen an und evaluiert darüberhinaus zusätzlich das 95%-Konfidenzintervall für den Erwartungswertparameter.

```
D = read.csv("../data/bdi-ii-datensatz.csv") # Datensatzeinlesen
y = D$dBDI # Datenauswahl
n = length(y) # Stichprobenumfang
mu_hat = mean(y) # Erwartungswertparameterschätzer
delta = 0.95 # Konfidenzlevel
t_delta = qt((1+delta)/2, n-1) # \Psi^{-1}((\delta + 1)/2, n-1)
G_u = mean(y) - (sd(y)/sqrt(n))*t_delta # untere Konfidenzintervallgrenze
G_o = mean(y) + (sd(y)/sqrt(n))*t_delta # obere Konfidenzintervallgrenze
mu_0 = 0 # Nullhypothese parameter, hier \mu = \mu_0
alpha_0 = 0.05 # Signifikanzlevel
k_alpha_0 = qt(1-alpha_0/2, n-1) # kritischer Wert
Tee = sqrt(n)*((mean(y) - mu_0)/sd(y)) # T-Teststatistik
if(abs(Tee) > k_alpha_0){phi = 1} else {phi = 0} # Test 1_{|vert t |vert >= k_alpha_0}
p = 2*(1 - pt(Tee, n-1)) # p-Wert
cat("Parameterschätzwert =", mu_hat, # Ausgabe
    "\n95%-Konfidenzintervall =", G_u, G_o,
    "\nSignifikanzlevel =", alpha_0,
    "\nKritischer Wert =", k_alpha_0,
    "\nTeststatistik =", Tee,
    "\nTestwert =", phi,
    "\np-Wert =", p)
```

Parameterschätzwert = 3.166667

```

95%-Konfidenzintervall = 0.8074098 5.525923
Signifikanzlevel      = 0.05
Kritischer Wert       = 2.200985
Teststatistik         = 2.95423
Testwert              = 1
p-Wert                = 0.01310986

```

Die gleiche Analyse kann auch mit der in **R** implementierten Funktion `t.test()` durchgeführt werden, die Syntax zu ihrer Benutzung und die Formatierung der durch sie bestimmten Ergebnisse finden sich untenstehend

```
t.test(y)
```

```

One Sample t-test

data: y
t = 2.9542, df = 11, p-value = 0.01311
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.8074098 5.5259235
sample estimates:
mean of x
 3.166667

```

Im vorliegenden Fall würde man die Nullhypothese also bei einem Signifikanzlevel von $\alpha_0 = 0.05$ ablehnen. Ob die Nullhypothese allerdings im vorliegenden Fall zutrifft oder nicht bleibt, wie der wahre, aber unbekannte, Erwartungswertparameter unbekannt. Im langfristigen Mittel jedoch lehnt man basierend auf den oben beschriebenen Annahmen die Nullhypothese in nur 5 von 100 Fällen fälschlicherweise ab.

33.4. Konfidenzintervalle und Hypothesentests

In diesem Abschnitt untersuchen wir, inwieweit Konfidenzintervalle und Hypothesentests als äquivalent angesehen werden können. Wir wollen dabei von dem Szenario eines Konfidenzintervalls ausgehen.

Theorem 33.5 (Dualität von Konfidenzintervallen und Hypothesentests). *Es sei y die Stichprobe eines Frequentistischen Inferenzmodells mit Ergebnisraum $\mathcal{Y}$ und Parameterraum Θ . Weiterhin sei für ein $\delta \in]0, 1[$ mit $[G_u(y), G_o(y)]$ ein δ -Konfidenzintervall für den wahren, aber unbekannten, Parameter $\theta \in \Theta$ definiert. Dann gilt, dass der Hypothesentest definiert durch*

$$\phi_\theta : \mathcal{Y} \rightarrow \{0, 1\}, y \mapsto \phi(y) := \begin{cases} 0, & [G_u(y), G_o(y)] \ni \theta_0 \\ 1, & [G_u(y), G_o(y)] \not\ni \theta_0 \end{cases} \quad (33.65)$$

ein Hypothesentest vom Signifikanzlevel $\alpha_0 = 1 - \delta$ für die Hypothesen

$$\Theta_0 := \{\theta_0\} \text{ und } \Theta_1 := \Theta \setminus \{\theta_0\}. \quad (33.66)$$

◦

Beweis. Aufgrund der einfachen Nullhypothese und somit $\alpha_0 = \alpha$ folgt

$$\alpha_0 = \alpha = \mathbb{P}_{\theta_0}(\phi(y) = 1) = \mathbb{P}_{\theta_0}([G_u(y), G_o(y)] \not\ni \theta) = 1 - \mathbb{P}_{\theta_0}([G_u(y), G_o(y)] \ni \theta) = 1 - \delta. \quad (33.67)$$

□

Theorem 33.5 besagt also, dass man mithilfe eines δ -Konfidenzintervalls einen Hypothesentest mit Signifikanzlevel $\alpha_0 = 1 - \delta$ mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese konstruieren kann. Dazu ist die bei diesem Test die Nullhypothese $\theta = \theta_0$ jeweils abzulehnen, wenn das Konfidenzintervall den Nullhypotheseparameter θ_0 nicht überdeckt und ansonsten nicht. Anhand folgenden Theorems wollen wir Theorem 33.5 für das in Kapitel 32.2 betrachtete Konfidenzintervall für den Erwartungswertparameter des Normalverteilungsmodells und den in Kapitel 33.3.1 betrachteten Einstichproben-T-Test konkretisieren.

Theorem 33.6 (Dualität von Erwartungswertkonfidenzintervall und Einstichproben-T-Test). *Gegeben sei das Normalverteilungsmodell und es sei*

$$\kappa := \left[\bar{y} - \frac{S}{\sqrt{n}} t_\delta, \bar{y} + \frac{S}{\sqrt{n}} t_\delta \right]. \quad (33.68)$$

das mithilfe von

$$t_\delta := \Psi^{-1} \left(\frac{1 + \delta}{2}; n - 1 \right) \quad (33.69)$$

in Theorem 32.2 definierte δ -Konfidenzintervall für den Erwartungswertparameter. Dann ist mit Theorem 33.5 der Test

$$\phi : \mathcal{Y} \rightarrow \{0, 1\}, y \mapsto \phi(y) := \begin{cases} 0, & \left[\bar{y} - \frac{S}{\sqrt{n}} t_\delta, \bar{y} + \frac{S}{\sqrt{n}} t_\delta \right] \ni \mu_0 \\ 1, & \left[\bar{y} - \frac{S}{\sqrt{n}} t_\delta, \bar{y} + \frac{S}{\sqrt{n}} t_\delta \right] \not\ni \mu_0 \end{cases} \quad (33.70)$$

ein Test der einfachen Nullhypothese $H_0 : \mu = \mu_0$ und der zusammengesetzten Alternativhypothese $H_1 : \mu \neq \mu_0$ mit Signifikanzlevel $\alpha_0 = 1 - \delta$.

◦

Beweis. Es gilt

$$\mathbb{P}_{\mu_0}(\phi(y) = 1) = 1 - \mathbb{P}_{\mu_0}(\phi(y) = 0) = 1 - \mathbb{P}_{\mu_0} \left(\left[\bar{y} - \frac{S}{\sqrt{n}} t_\delta, \bar{y} + \frac{S}{\sqrt{n}} t_\delta \right] \ni \mu_0 \right) = 1 - \delta. \quad (33.71)$$

□

Folgender **R** Code simuliert diesen Konfidenzintervall-basierten Hypothesentest bei Zutreffen der Nullhypothese und gibt Schätzungen für das Konfidenzlevel und das Signifikanzlevel über 100 Realisierungen einer Stichproben vom Stichprobenumfang $n = 12$ mit wahren, aber unbekannten, Parametern $\mu = 2$ und $\sigma^2 = 1$ an.

```
n = 12 # Stichprobenumfang
mu = 2 # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr = 1 # wahrer, aber unbekannter, Varianzparameter
delta = 0.95 # Konfidenzlevel
t_delta = qt((1+delta)/2, n-1) # \Psi^{-1}((\delta + 1)/2, n-1)
mu_0 = mu # Nullhypotheseparameter bei Zutreffen von H_0
set.seed(1) # random number generator seed
ns = 1e2 # Anzahl Simulationen
y_bar = rep(NA, ns) # Stichprobenmittellarray
s = rep(NA, ns) # Stichprobenstandardabweichungarray
kappa = matrix(rep(NA, 2*ns), ncol = 2) # Konfidenzintervallarray
kfn = rep(NA, ns) # Überdeckungsindikatorarray
phi = rep(NA, ns) # Testarray
for(i in 1:ns){
  y = rnorm(n, mu_0, sqrt(sigsqr)) # Simulationsiterationen
  y_bar[i] = mean(y) # Stichprobenrealisierung
  s[i] = sd(y) # Stichprobenmittel
  kappa[i,1] = y_bar[i] - (s[i]/sqrt(n))*t_delta # Stichprobenstandardabweichung
  kappa[i,2] = y_bar[i] + (s[i]/sqrt(n))*t_delta # untere Konfidenzintervallgrenze
  if(kappa[i,1] <= mu_0 & mu_0 <= kappa[i,2]){ # obere Konfidenzintervallgrenze
    kfn[i] = 1} else{kfn[i] = 0}
  if(kappa[i,1] <= mu_0 & mu_0 <= kappa[i,2]){
    phi[i] = 0} else{phi[i] = 1}
}
cat("Geschätztes Konfidenzniveau =", mean(kfn), # Testevaluation
    "\nGeschätzter Testumfang =", mean(phi)) # Ausgabe
```

Geschätztes Konfidenzniveau = 0.96
 Geschätzter Testumfang = 0.04

Wir visualisieren die Ergebnisse dieser Simulation in Abbildung 33.9.

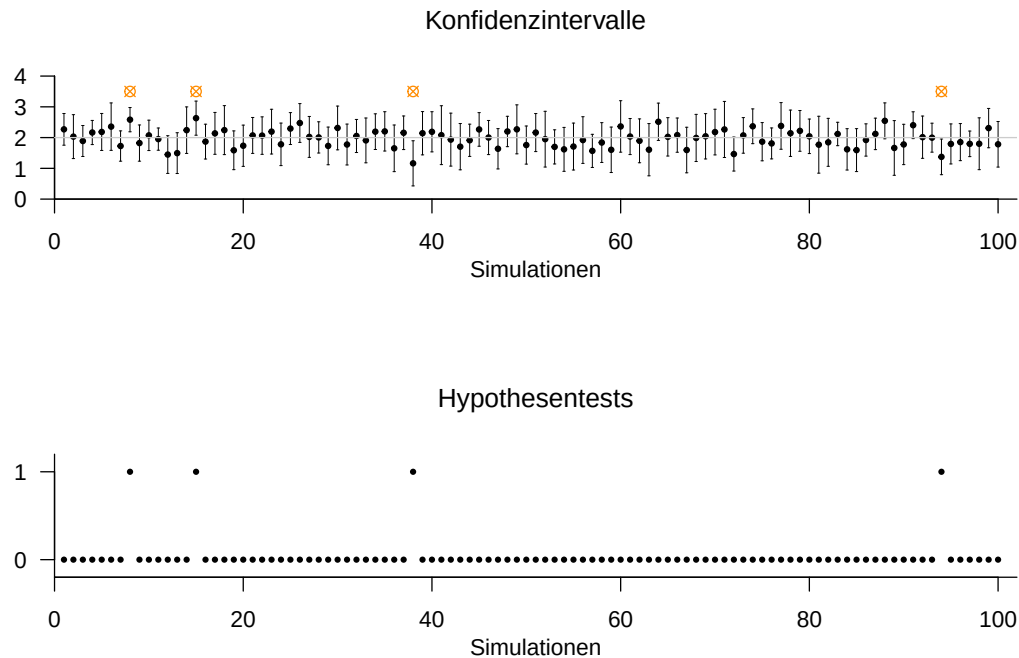


Abbildung 33.9. Dualität von Konfidenzintervall und Hypothesentest am Beispiel des Normalverteilungsmodells. Nutzt man das δ -Konfidenzintervall für den Erwartungswertparameter um bei Nichtüberdeckung des Nullhypotheseparameters die Nullhypothese abzulehnen, so ergibt sich in diesem Fall ein Hypothesentest mit Signifikanzlevel $\alpha_0 = 1 - \delta$.

33.5. Literaturhinweise

Die hier präsentierte Theorie der Hypothesentests geht im Wesentlichen auf Neyman & Pearson (1928) und Neyman & Pearson (1933) zurück. Gigerenzer (2004) und Lehmann (2011) geben historische Einordnungen der Genese des Hypothesentestbegriffs.

Teil V.

Allgemeines Lineares Modell

34. Regression

Fundamentales Ziel von Regressionsanalysen ist es, Beziehungen zwischen unabhängigen und abhängigen Variablen zu modellieren. Ein zentrales Thema dabei ist die Anpassung von Funktionen an beobachtete Datensätze. Mit dem Begriff der Ausgleichsgerade im Rahmen der Methode der kleinsten Quadrate und dem Begriff der einfachen linearen Regression wollen wir uns in diesem Abschnitt diesen zentralen Themen der probabilistischen Datenmodellierung schrittweise nähern. Dabei unterscheiden sich die Konzepte von Ausgleichsgerade und einfacher linearer Regression in einem zentralen Aspekt: bei der Ausgleichsgerade werden unabhängige und abhängige Variable nicht als Zufallsvariablen modelliert, im Rahmen der einfachen linearen Regression nimmt die abhängige Variable dann die Form einer Zufallsvariablen an. Im Kontext der Korrelation schließlich werden sowohl abhängige als auch unabhängige Variable als Zufallsvariablen modelliert.

Um die Konzepte dieses Abschnittes zu verdeutlichen, betrachten wir einen Beispieldatensatz in dem die Anzahl an Psychotherapiestunden als unabhängige Variable x der Symptomreduktion einer Gruppe von $n = 20$ Patient:innen als abhängige Variable y gegenüber gestellt wird (Abbildung 42.3). Die visuelle Inspektion dieses Datensatzes legt nahe, dass ein Mehr an Therapiestunden ein Mehr an Symptomreduktion impliziert. Ziel der Methode der kleinsten Quadrate und der einfachen linearen Regression ist es, diesen intuitiven funktionalen Zusammenhang zwischen unabhängiger und abhängiger Variable auf eine quantitative Basis zu stellen.

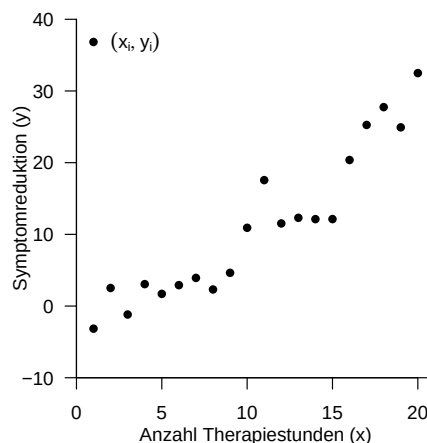


Abbildung 34.1. Beispieldatensatz.

34.1. Methode der kleinsten Quadrate

Wir definieren zunächst den Begriff der *Ausgleichsgerade*.

Definition 34.1 (Ausgleichsgerade). Für $\beta := (\beta_0, \beta_1)^T \in \mathbb{R}^2$ heißt die linear-affine Funktion

$$f_\beta : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f_\beta(x) := \beta_0 + \beta_1 x \quad (34.1)$$

für die für einen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ die Funktion

$$q : \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}, \beta \mapsto q(\beta) := \sum_{i=1}^n (y_i - f_\beta(x_i))^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (34.2)$$

der quadrierten vertikalen Abweichungen der y_i von den Funktionswerten $f_\beta(x_i)$ ihr Minimum annimmt, Ausgleichsgerade für den Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\}$. •

Bei der Ausgleichsgerade handelt es sich also um eine linear-affine Funktion der Form

$$f_\beta : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f_\beta(x) := \beta_0 + \beta_1 x \quad (34.3)$$

Abbildung 34.2 zeigt drei durch jeweils andere Werte von β_0 und β_1 parameterisierte linearaffine Funktionen zusammen mit der Wertemenge des Beispieldatensatzes.

Wie bei allen linear-affinen Funktionen entspricht bei f_β der Wert von β_0 dem Wert, den f_β für $x = 0$ annimmt,

$$f_\beta(0) = \beta_0 + \beta_1 \cdot 0 = \beta_0 \quad (34.4)$$

und damit graphisch dem Schnittpunkt des Funktionsgraphen mit der y -Achse. Da β_0 damit dem Versatz (engl. *offset*) des Funktionsgraphen von $y = 0$ an der Stelle $x = 0$ entspricht, nennt man β_0 auch häufig den *Offsetparameter*. Analog entspricht wie bei allen linear-affinen Funktionen der Wert von β_1 dem Wert der Funktionswertdifferenz pro Argumenteinheitsdifferenz. Beispielsweise gilt etwa für $\beta_0 = 5$ und $\beta_1 = 0.5$, dass

$$\begin{aligned} f_\beta(2) - f_\beta(1) &= (5 + 0.5 \cdot 2) - (5 + 0.5 \cdot 1) = 1 - 0.5 = 0.5 \\ f_\beta(9) - f_\beta(8) &= (5 + 0.5 \cdot 9) - (5 + 0.5 \cdot 8) = 9.5 - 8 = 0.5 \end{aligned} \quad (34.5)$$

Für eine Argumentdifferenz von 1 ergibt sich also eine Funktionswertdifferenz von 0.5. β_1 enkodiert also die Stärke der Änderung der Funktionswerte pro Argumenteinheitsdifferenz und damit die Steigung (engl. *slope*) des Graphen der linear-affinen Funktion. Entsprechend wird β_1 *Steigungparameter* oder *Slopeparameter* genannt.

Nach Definition ist die Ausgleichsgerade nun allerdings nicht eine beliebige linearaffine Funktion der Form f_β , sondern eben jene, die für einen gegebenen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\}$ die Summe der quadrierten vertikalen Abweichungen

$$q(\beta) := \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (34.6)$$

minimiert. Für eine fest vorgegebenen Datensatz von (x_i, y_i) Paaren ist der Wert dieser Summe abhängig von den Werten von β_0 und β_1 und kann deshalb durch Wahl geeigneter

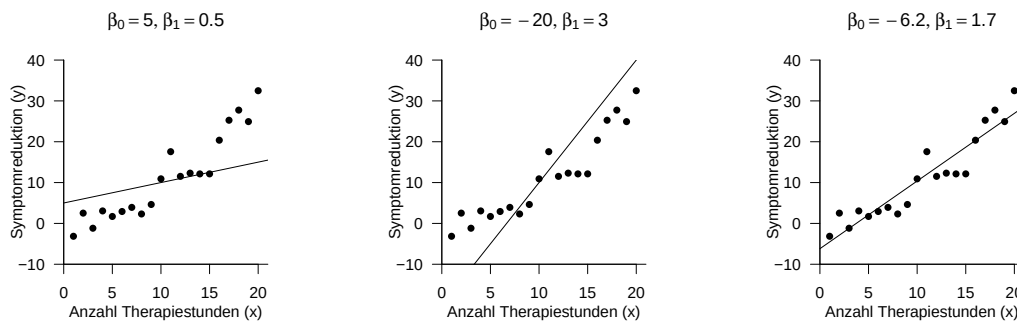


Abbildung 34.2. Linear-affine Funktionen mit unterschiedlichen Parameterwerten vor dem Hintergrund des Beispieldatensatzes.

Werte von β_0 und β_1 minimiert werden. Da hierbei eine Summe von quadrierten Abweichungen zwischen Datenpunkten und Werten der Ausgleichsgerade minimiert wird, spricht man auch oft etwas ungenau von der *Methode der kleinsten Quadrate* (engl. *method of least squares*). Abbildung 1.3 zeigt die vertikalen Abweichungen zwischen y_i und $\beta_0 + \beta_1 x_i$ für $i = 1, \dots, n$ des Beispieldatensatzes als orange Linien sowie die Summe ihrer Quadrate $q(\beta)$ im Titel. Für die Parameterwerte $\beta_0 = -6.2$ und $\beta_1 = 1.7$ (vgl. Abbildung 34.2) nimmt diese Summe ihren kleinsten Wert an.

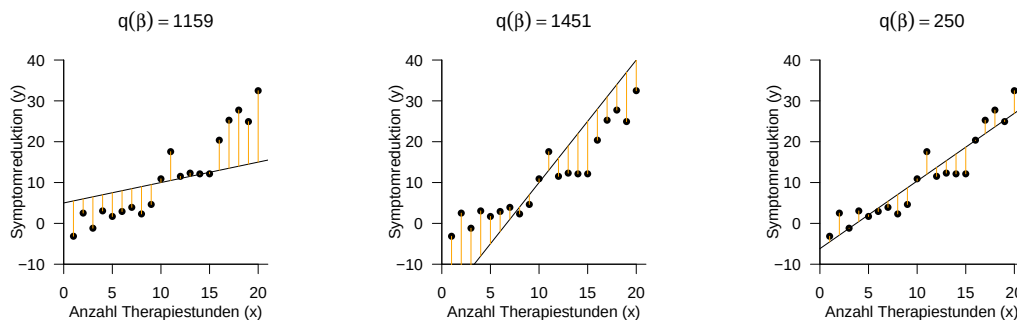


Abbildung 34.3. Vertikale Abweichungen und Quadratsummen bei unterschiedlichen Parameterwerten.

Konkrete Formeln zur Bestimmung der Parameterwerte der Ausgleichsgerade stellt Theorem 34.1 bereit.

Theorem 34.1 (Ausgleichsgerade). *Für einen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ hat die Ausgleichsgerade die Form*

$$f_\beta : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f_\beta(x) := \hat{\beta}_0 + \hat{\beta}_1 x \quad (34.7)$$

wobei mit der Stichprobenkovarianz c_{xy} der (x_i, y_i) -Werte, der Stichprobenvarianz s_x^2 der x_i -Werte und den Stichprobenmitteln $\bar{x}$ und $\bar{y}$ der x_i - und y_i -Werte, respektive, gilt, dass

$$\hat{\beta}_1 = \frac{c_{xy}}{s_x^2} \text{ und } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (34.8)$$

◦

Beweis. Wir betrachten die Summe der quadrierten vertikalen Abweichungen der y_i von den Funktionswerten $f(x_i)$ als Funktion von β_0 und β_1 und bestimmen Werte $\hat{\beta}_0$ und $\hat{\beta}_1$, für die diese Funktion ihr Minimum annimmt, die Summe der quadrierten vertikalen Abweichungen der y_i von den Funktionswerten $f(x_i)$ also minimal wird. Wir betrachten die Funktion

$$q : \mathbb{R}^2 \rightarrow \mathbb{R}, (\beta_0, \beta_1) \mapsto q(\beta_0, \beta_1) := \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (34.9)$$

Um das Minimum dieser Funktion zu bestimmen, berechnen wir zunächst die partiellen Ableitungen hinsichtlich β_0 und β_1 und setzen diese gleich 0. Es ergibt sich zunächst

$$\begin{aligned} \frac{\partial}{\partial \beta_0} q(\beta_0, \beta_1) &= \frac{\partial}{\partial \beta_0} \left(\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \right) \\ &= \sum_{i=1}^n \frac{\partial}{\partial \beta_0} (y_i - (\beta_0 + \beta_1 x_i))^2 \\ &= \sum_{i=1}^n 2(y_i - (\beta_0 + \beta_1 x_i)) \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_1 x_i) \\ &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i). \end{aligned} \quad (34.10)$$

Weiterhin ergibt sich

$$\begin{aligned} \frac{\partial}{\partial \beta_1} q(\beta_0, \beta_1) &= \frac{\partial}{\partial \beta_1} \left(\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \right) \\ &= \sum_{i=1}^n \frac{\partial}{\partial \beta_1} (y_i - (\beta_0 + \beta_1 x_i))^2 \\ &= \sum_{i=1}^n 2(y_i - (\beta_0 + \beta_1 x_i)) \frac{\partial}{\partial \beta_1} (y_i - \beta_0 - \beta_1 x_i) \\ &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i. \end{aligned} \quad (34.11)$$

Nullsetzen beider partieller Ableitungen ergibt dann

$$\begin{aligned} \frac{\partial}{\partial \beta_0} q(\beta_0, \beta_1) &= 0 \text{ und } \frac{\partial}{\partial \beta_1} q(\beta_0, \beta_1) = 0 \\ \Leftrightarrow -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) &= 0 \text{ und } -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \\ \Leftrightarrow \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) &= 0 \text{ und } \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{aligned} \quad (34.12)$$

und weiter

$$\begin{aligned} \sum_{i=1}^n y_i - \sum_{i=1}^n \beta_0 - \beta_1 \sum_{i=1}^n x_i &= 0 \text{ und } \sum_{i=1}^n y_i x_i - \sum_{i=1}^n \beta_0 x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0 \\ \Leftrightarrow \beta_0 n + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \text{ und } \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i \end{aligned} \quad (34.13)$$

Das sich hier ergebende Gleichungssystem

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i \end{aligned} \quad (34.14)$$

wird System der Normalgleichungen genannt und beschreibt die notwendige Bedingung für ein Minimum von q . Auflösen dieses Gleichungssystems nach β_0 und β_1 liefert dann die Werte $\hat{\beta}_0$ und $\hat{\beta}_1$ des Theorems. Um dies zu sehen, halten wir zunächst fest, dass mit der ersten Gleichung des Systems der Normalgleichungen gilt

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \Leftrightarrow \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \Leftrightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (34.15)$$

Einsetzen der Form von $\hat{\beta}_0$ in die zweite Gleichung des Systems der Normalgleichungen ergibt dann zunächst

$$\begin{aligned}
 & \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i \\
 & \Leftrightarrow (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i \\
 & \Leftrightarrow \bar{y} \sum_{i=1}^n x_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i \\
 & \Leftrightarrow -\hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i - \bar{y} \sum_{i=1}^n x_i \\
 & \Leftrightarrow \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) = \sum_{i=1}^n y_i x_i - \bar{y} \sum_{i=1}^n x_i.
 \end{aligned} \tag{34.16}$$

Wir halten nun zunächst fest, dass gilt

$$\begin{aligned}
 \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \bar{x} \sum_{i=1}^n x_i \\
 &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \bar{x} \\
 &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \\
 &= \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
 &= \sum_{i=1}^n (x_i - \bar{x})^2
 \end{aligned} \tag{34.17}$$

Weiterhin halten wir zunächst fest, dass gilt

$$\begin{aligned}
 \sum_{i=1}^n y_i x_i - \bar{y} \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i x_i - \bar{y} \sum_{i=1}^n x_i - n\bar{y}\bar{x} + n\bar{y}\bar{x} \\
 &= \sum_{i=1}^n y_i x_i - \bar{y} \sum_{i=1}^n x_i - \sum_{i=1}^n y_i \bar{x} + \sum_{i=1}^n \bar{y} \bar{x} \\
 &= \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \bar{x} - \sum_{i=1}^n \bar{y} x_i + \sum_{i=1}^n \bar{y} \bar{x} \\
 &= \sum_{i=1}^n (y_i x_i - y_i \bar{x} - \bar{y} x_i + \bar{y} \bar{x}) \\
 &= \sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x}).
 \end{aligned} \tag{34.18}$$

In der Fortsetzung von (1.16) ergibt sich dann

$$\begin{aligned}
 \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) &= \sum_{i=1}^n y_i x_i - \bar{y} \sum_{i=1}^n x_i \\
 \Leftrightarrow \hat{\beta}_1 \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) &= \sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x}) \\
 \Leftrightarrow \hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
 \Leftrightarrow \hat{\beta}_1 &= \frac{c_{xy}}{s_x^2}.
 \end{aligned} \tag{34.19}$$

□

Theorem 34.1 besagt, dass die Parameterwerte, die für einen gegebenen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\}$ die Summe der quadrierten vertikalen Abweichungen für eine

linear-affine Funktion minimieren mithilfe der Stichprobenmittel der x_i - und y_i -Werte, der Stichprobenvarianz der x_i -Werte und der Stichprobenkovarianz der x_i - und y_i -Werte berechnet werden können. Die Terminologie orientiert sich hier an den Begrifflichkeiten der deskriptiven Statistik, insbesondere werden die x_i -Werte häufig nicht als Realisationen von Zufallsvariablen verstanden, der Begriff der Stichprobe wird jedoch trotzdem verwendet. Aus der Anwendungsperspektive können nach Theorem 34.1 die Parameter der Ausgleichsgerade also mithilfe der bekannten Funktionen für die Auswertung deskriptiver Statistiken bestimmt werden. Folgender **R** Code demonstriert dies.

```
# Einlesen des Beispieldatensatzes
D = read.csv("../data/501-regression.csv")

# Stichprobenstatistiken
x_bar = mean(D$x_i)
y_bar = mean(D$y_i)
s2x = var(D$x_i)
cxy = cov(D$x_i, D$y_i)

# Stichprobenmittel der x_i-Werte
# Stichprobenmittel der y_i-Werte
# Stichprobenvarianz der x_i-Werte
# Stichprobenkovarianz der (x_i, y_i)-Werte

# Ausgleichsgeradenparameter
beta_1_hat = cxy/s2x
beta_0_hat = y_bar - beta_1_hat*x_bar

# \hat{\beta}_1, Steigungsparameter
# \hat{\beta}_0, Offset Parameter

# Ausgabe
cat("beta_0_hat:", beta_0_hat,
    "\nbeta_1_hat:", beta_1_hat)
```

```
beta_0_hat: -6.194704
beta_1_hat: 1.657055
```

Eine typische Visualisierung der Ausgleichsgerade eines Datensatzes wie in Abbildung 34.4 implementiert folgender **R** Code.

```
# Datenwerte
plot(
  D$x_i,
  D$y_i,
  pch = 16,
  xlab = "Anzahl Therapiestunden (x)",
  ylab = "Symptomreduktion (y)",
  xlim = c(0, 21),
  ylim = c(-10, 40),
  main = TeX("$\\hat{\\beta}_0 = -6.19, \\hat{\\beta}_1 = 1.66$"))

# Ausgleichsgerade
abline(
  coef = c(beta_0_hat, beta_1_hat),
  lty = 1,
  col = "black")

# Legende
legend(
  "topleft",
  c(TeX("$x_i, y_i$"), TeX("$f(x) = \\hat{\\beta}_0 + \\hat{\\beta}_1 x$")),
  lty = c(0, 1),
  pch = c(16, NA),
  bty = "n")
```

Die Idee, bei einem gegebenen Datensatz von (x_i, y_i) Paaren die Summe der quadrierten vertikalen Abweichungen zwischen einer Funktion der x_i -Werte und den y_i -Werten zu minimieren und so eine Funktion möglichst gut an eine Wertemenge anzupassen, ist nicht auf linear-affine Funktionen beschränkt. Folgende Definition verallgemeinert die Definition der Ausgleichsgerade auf Polynomfunktionen beliebigen Grades.

Definition 34.2 (Ausgleichspolynom). Für $\beta := (\beta_0, \dots, \beta_k)^T \in \mathbb{R}^{k+1}$ heißt die Polynomfunktion k ten Grades

$$f_\beta : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f_\beta(x) := \sum_{i=0}^k \beta_i x^i \quad (34.20)$$

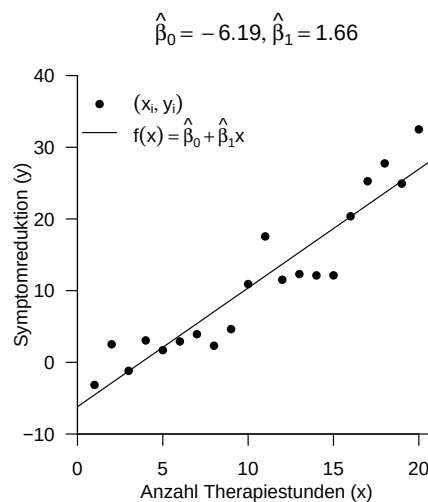


Abbildung 34.4. Ausgleichsgerade für den Beispieldatensatz.

für die für einen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ die Funktion

$$q : \mathbb{R}^{k+1} \rightarrow \mathbb{R}_{\geq 0}, \beta \mapsto q(\beta) := \sum_{i=1}^n (y_i - f_\beta(x_i))^2 = \sum_{i=1}^n \left(y_i - \sum_{i=0}^k \beta_i x^i \right)^2 \quad (34.21)$$

der quadrierten vertikalen Abweichungen der y_i von den Funktionswerten $f_\beta(x_i)$ ihr Minimum annimmt, das *Ausgleichspolynom kten Grades für den Datensatz* $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

•

Die Ausgleichsgerade ist damit das Ausgleichspolynom ersten Grades. Wir wollen den Begriff des Ausgleichspolynoms hier nicht weiter vertiefen und werden insbesondere die Parameterwerte $\hat{\beta}_0, \dots, \hat{\beta}_k$, für die die Funktion q ihr Minimum annimmt an späterer Stelle im Rahmen der Theorie des Allgemeinen Linearen Modells allgemein bestimmen. In [Abbildung 34.5](#) visualisieren beispielhaft die Ausgleichspolynome ersten bis vierten Grades für den Beispieldatensatz wobei der Wert der Funktion q an der Minimumsstelle jeweils im Titel vermerkt ist.

34.2. Einfache lineare Regression

Eine Ausgleichsgerade erlaubt Aussagen über unbeobachtete der abhängigen Variable. Allerdings erlaubt eine Ausgleichsgerade nur implizite Aussagen über die mit der Anpassung einer linear-affinen Funktion an einen Datensatz verbundene Unsicherheit. In der einfachen linearen Regression wird die Idee einer Ausgleichsgerade um eine probabilistische Komponente erweitert. Die einfache lineare Regression erlaubt damit insbesondere im Sinne der Frequentistischen Inferenz Konfidenzintervalle für die Ausgleichsgeradenparameter anzugeben und Hypothesentests bezüglich der Ausgleichsgeradenparameter durchzuführen. Wir wollen hier zunächst nur das Modell der einfachen linearen Regression und die

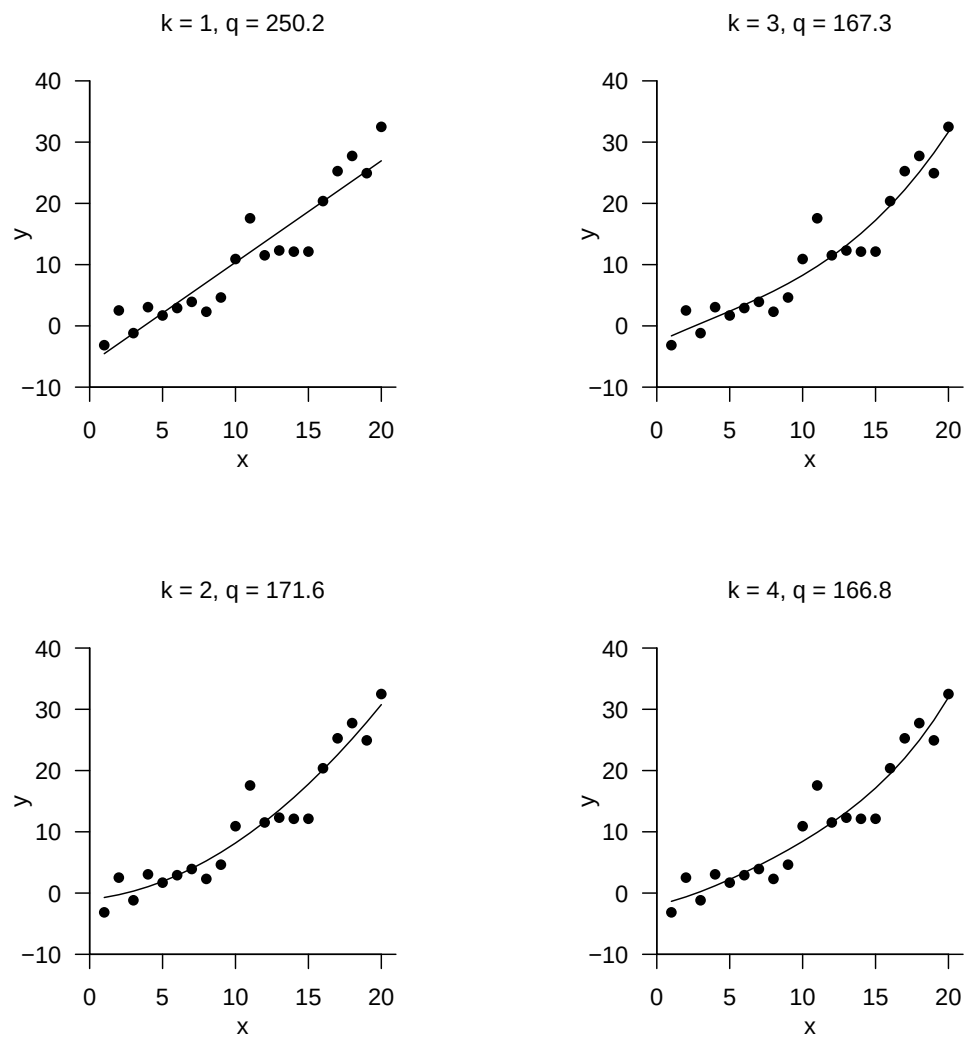


Abbildung 34.5. Ausgleichspolynome ersten bis vierten Grades für den Beispieldatensatz.

auf ihm basierende Maximum-Likelihood Schätzung der Ausgleichsgeradenparameter betrachten.

Die Bewertung der mit dieser Schätzung verbundenen Unsicherheit sowie parameterzentrierte Hypothesentests behandeln wir dann an späterer Stelle zunächst im Kontext des Allgemeinen Linearen Modells. Wir beginnen mit folgender Definition.

Definition 34.3 (Modell der einfachen linearen Regression). Es sei

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \dots, n \quad (34.22)$$

wobei

- y_i beobachtbare Zufallsvariablen sind, die Werte einer abhängigen Variable modellieren,
- $x_i \in \mathbb{R}$ fest vorgegebene Prädiktorwerte oder Regressorwerte sind, die Werte einer unabhängigen Variable modellieren,
- $\beta_0, \beta_1 \in \mathbb{R}$ wahre, aber unbekannte, Offset- und Steigungsparameterwerte sind und
- ε_i unabhängig und identisch normalverteilte nicht-beobachtbare Zufallsvariablen mit wahren, aber unbekanntem, Varianzparameter $\sigma^2 > 0$ sind, die Fehler- oder Störvariablen modellieren.

Dann heißt Gleichung 34.22 *Modell der einfachen linearen Regression*.

•

Im Gegensatz zur Ausgleichsgerade treten im Modell der einfachen linearen Regression also explizit Zufallsvariablen auf. Speziell definiert das Modell der einfachen linearen Regression wie n beobachtbare (abhängige) Zufallsvariablen y_i anhand der Werte x_i einer unabhängigen Variable, der Parameterwerte β_0 und β_1 sowie durch Addition der normalverteilten Fehlervariablen ε_i generiert wird. Das Modell hat dabei drei Parameter, den Offsetparameter β_0 , den Steigungsparameter β_1 und den Varianzparameter σ^2 der normalverteilten Fehlervariablen. Addition der festen Werte β_0 und $\beta_1 x_i$ zu der normalverteilten Zufallsvariable ε_i impliziert dabei eine Normalverteilung von y_i . Dies ist die Aussage folgenden Theorems.

Theorem 34.2 (Datenverteilung der einfachen linearen Regression). *Das Modell der einfachen linearen Regression*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \dots, n \quad (34.23)$$

lässt sich mit $\mu_i := \beta_0 + \beta_1 x_i$ äquivalent in der Form

$$y_i \sim N(\mu_i, \sigma^2) \text{ u.v. für } i = 1, \dots, n \quad (34.24)$$

schreiben.

◦

Beweis. Wir zeigen die Äquivalenz für ein i , die Unabhängigkeit der y_i zeigen wir an späterer Stelle im Rahmen des Allgemeinen Linearen Modells. Die Äquivalenz beider Modellformen für ein i folgt direkt aus der Transformation normalverteilter Zufallsvariablen durch linear-affine Funktionen. Speziell gilt im vorliegenden Fall für $\varepsilon_i \sim N(0, \sigma^2)$, dass

$$y_i = f(\varepsilon_i) \text{ mit } f: \mathbb{R} \rightarrow \mathbb{R}, \varepsilon_i \mapsto f(\varepsilon_i) := \varepsilon_i + (\beta_0 + \beta_1 x_i) \quad (34.25)$$

Mit dem WDF Transformationstheorem bei linear-affinen Abbildungen folgt dann

$$\begin{aligned} p_{y_i}(y_i) &= \frac{1}{|1|} p_{\varepsilon_i} \left(\frac{y_i - \beta_0 - \beta_1 x_i}{1} \right) \\ &= N(y_i - \beta_0 - \beta_1 x_i; 0, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_i - 0)^2 \right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (y_i - (\beta_0 + \beta_1 x_i))^2 \right) \\ &= N(y_i; \beta_0 + \beta_1 x_i, \sigma^2). \end{aligned} \quad (34.26)$$

Definition von $\mu_i := \beta_0 + \beta_1 x_i$ ergibt dann die Aussage des Theorems. $\square$

Theorem 34.2 besagt insbesondere, dass die Datenvariablen y_i univariat normalverteilte Zufallsvariablen sind, deren Erwartungswertparameter jeweils vom Wert der unabhängigen Variable x_i abhängen. Abbildung 1.6 visualisiert das Modell und eine Realisation der einfachen linearen Regression für wahre, aber unbekannte, Parameterwerte $\beta_0 := 0, \beta_1 := 1$ und $\sigma^2 := 1$.

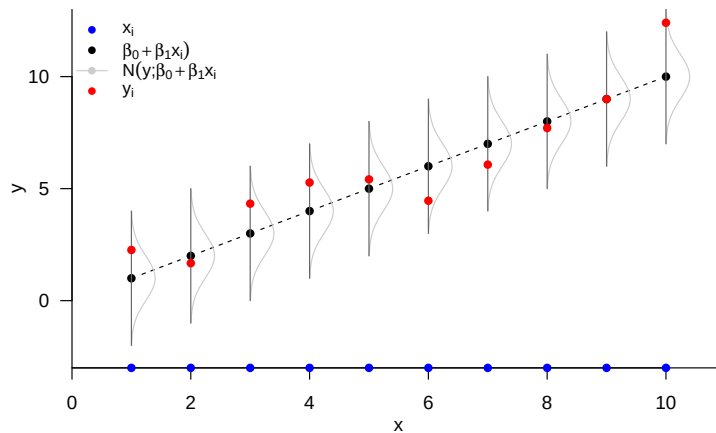


Abbildung 34.6. Modell der einfachen linearen Regression für $\beta_0 := 0, \beta_1 := 1$ und $\sigma^2 := 1$.

Da es sich bei dem Modell der einfachen linearen Regression um ein parametrisches Frequentistisches Modell handelt, können Schätzer für die Modellparameter mithilfe des MaximumLikelihood Prinzips gewonnen werden. Insbesondere stellt sich dabei heraus, dass die Maximum-Likelihood Schätzer des Offset- und des Steigungsparameter mit den Werten der Ausgleichsgeradenparameter identisch sind. Dies ist eine der Aussagen folgenden Theorems. Wir verzichten hier bei den Schätzern aus Gründen der notationstechnischen Übersichtlichkeit auf ^{ML} Superskripte.

Theorem 34.3 (Maximum-Likelihood Schätzung der einfachen linearen Regression).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \dots, n \quad (34.27)$$

das Modell der einfachen linearen Regression. Dann sind Maximum Likelihood Schätzer der Modellparameter β_0, β_1 und σ^2 gegeben durch

$$\hat{\beta}_1 := \frac{c_{xy}}{s_x^2}, \quad \hat{\beta}_0 := \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{und} \quad \hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \quad (34.28)$$

◦

Beweis. Wir zeigen zunächst, dass die Ausgleichsgeradenparameter $\hat{\beta}_0$ und $\hat{\beta}_1$ den entsprechenden Maximum-Likelihood Schätzern gleichen. Dazu halten wir zunächst fest, dass aufgrund der Unabhängigkeit der $y_1, \dots, y_n$ die Likelihood-Funktion des Modells der einfachen linearen Regression bezüglich β_0 und β_1 die Form

$$\begin{aligned} L : \mathbb{R}^2 \rightarrow \mathbb{R}_{>0}, (\beta_0, \beta_1) &\mapsto L(\beta_0, \beta_1) := \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - (\beta_0 + \beta_1 x_i))^2\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2\right) \end{aligned} \quad (34.29)$$

hat. Weil für die Exponentialfunktion für $a < b \leq 0$ gilt, dass $\exp(a) < \exp(b)$, wird der Exponentialterm dieser Likelihood-Funktion maximal, wenn der nicht-negative Term

$$q := \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (34.30)$$

minimal und damit $-q$ maximal wird. Im Rahmen des Beweises der Ausgleichsgeradenform haben wir aber schon gezeigt, dass der Term (1.30) für

$$\hat{\beta}_1 := \frac{c_{xy}}{s_x^2} \text{ und } \hat{\beta}_0 := \bar{y} - \hat{\beta}_1 \bar{x} \quad (34.31)$$

minimal wird und damit $\hat{\beta}_1$ und $\hat{\beta}_0$ die Likelihood-Funktion maximieren.

In einem zweiten Schritt betrachten wir nun die Likelihood-Funktion des Modells der einfachen linearen Regression bezüglich σ^2 an der Stelle von $\hat{\beta}_0$ und $\hat{\beta}_1$. Wir erhalten die Likelihood-Funktion

$$L : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, \sigma^2 \mapsto L(\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2\right) \quad (34.32)$$

und die entsprechende Log-Likelihood-Funktion

$$\ell : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto \ell(\sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \quad (34.33)$$

In Analogie zu der Herleitung des Maximum-Likelihood Schätzers für σ^2 im Normalverteilungsmodell ergibt sich unter Beachtung von

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (34.34)$$

dann hier

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \quad (34.35)$$

□

In der Anwendung ist die Maximum-Likelihood Schätzung der Parameter der einfachen linearen Regression also im Wesentlichen mit der Bestimmung der Ausgleichsgeradenparameter identisch, wie folgender **R** Code zur Schätzung der Parameter für den Beispieldatensatz demonstriert.


```

# Einlesen des Beispieldatensatzes
fname = "../data/501-regression.csv"
D = read.table(fname, sep = ",", header = TRUE)

# Stichprobenstatistiken
n = length(D$y_i)
x_bar = mean(D$x_i)
y_bar = mean(D$y_i)
s2x = var(D$x_i)
cxy = cov(D$x_i, D$y_i)

# Parameteterschätzer
beta_1_hat = cxy/s2x
beta_0_hat = y_bar - beta_1_hat*x_bar
sigsqr_hat = (1/n)*sum((D$y_i-(beta_0_hat+beta_1_hat*D$x_i))^2)

# Ausgabe
cat("beta_0_hat:", beta_0_hat,
    "\nbeta_1_hat:", beta_1_hat,
    "\nsigsqr_hat:", sqrt(sigsqr_hat))

```

```

beta_0_hat: -6.194704
beta_1_hat: 1.657055
sigsqr_hat: 3.536795

```

34.3. Literaturhinweise

Die Idee der Minimierung einer Summe von quadrierten Abweichungen bei der Anpassung einer Polynomfunktion an beobachtete Werte geht auf die Arbeiten von Legendre (1805) und Gauss (1809) im Kontext der Bestimmung von Planetenbahnen zurück. Eine historische Einordnung dazu gibt Stigler (1981). Der Begriff der Regression geht zurück auf Galton (1886). Stigler (1986) gibt dazu einen ausführlichen historischen Überblick.

35. Korrelation

Die Konzepte der Regression und der Korrelation sind eng miteinander verwoben und dienen letztlich der Quantifizierung linear-affiner Abhängigkeiten zwischen unabhängigen und abhängigen Variablen. Ein zentrales Thema dieses Abschnittes sind damit die Äquivalenzen zwischen beiden Konzepten. Im Gegensatz zur Regression werden bei der Korrelation sowohl die abhängige als auch die unabhängige Variable als Zufallsvariablen modelliert, die probabilistische Komplexität der Korrelation ist also etwas höher als die der Regression und des Allgemeinen Linearen Modells. Um uns den Zusammenhängen von Regression und Korrelation zu nähern, erinnern wir Kapitel 35.1 zunächst an den Begriff und die Mechanik der Stichprobenkorrelation. In Kapitel 35.2 widmen wir uns anhand der R^2 Statistik dem Zusammenhang zwischen Stichprobenkorrelation und Ausgleichsgerade und geben ein erstes Beispiel für die Idee einer Varianzzerlegung. In Kapitel 35.3 schließlich diskutieren wir erste Aspekte zum Verständnis der Stichprobenkorrelation im probabilistischen Kontext. Als Anwendungsbeispiel betrachten wir durchgängig den in Kapitel 34 eingeführten Beispieldatensatz zum Zusammenhang von Psychotherapiedauer und Symptomreduktion.

35.1. Grundlagen

Wir erinnern zunächst an den Begriff der Korrelation zweier Zufallsvariablen.

Definition 35.1 (Korrelation). Die Korrelation zweier Zufallsvariablen ξ und v ist definiert als

$$\rho(\xi, v) := \frac{\mathbb{C}(\xi, v)}{\mathbb{S}(\xi)\mathbb{S}(v)} \quad (35.1)$$

wobei $\mathbb{C}(\xi, v)$ die Kovarianz von ξ und v und $\mathbb{S}(\xi)$ und $\mathbb{S}(v)$ die Standardabweichungen von ξ und v , respektive, bezeichnen.

•

Die Zahl $\rho(\xi, v)$ wird auch Korrelationskoeffizient von ξ und v genannt. Wir haben bereits gesehen, dass $-1 \leq \rho(\xi, v) \leq 1$ gilt und dass ξ und v unkorreliert genannt werden, wenn $\rho(\xi, v) = 0$ ist. Weiterhin haben wir bereits gesehen, dass aus der Unabhängigkeit der Zufallsvariablen ξ und v immer die Unkorreliertheit von ξ und v folgt, dass aber im Allgemeinen aus der Unkorreliertheit von ξ und v nicht die Unabhängigkeit von ξ und v folgt.

Liegen von zwei Zufallsvariablen Realisationen als ein bivariater Datensatz vor, so kann man anhand folgender Definition die realisationsspezifische Stichprobenkorrelation bestimmen.

Definition 35.2 (Stichprobenkorrelation). $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ sei ein Datensatz. Weiterhin seien:

- Die Stichprobenmittel der x_i und y_i definiert als

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i \text{ und } \bar{y} := \frac{1}{n} \sum_{i=1}^n y_i \quad (35.2)$$

- Die Stichprobenstandardabweichungen x_i und y_i definiert als

$$s_x := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \text{ und } s_y := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}. \quad (35.3)$$

- Die Stichprobenkovarianz der $(x_1, y_1), \dots, (x_n, y_n)$ definiert als

$$c_{xy} := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (35.4)$$

Dann ist die *Stichprobenkorrelation* der $(x_1, y_1), \dots, (x_n, y_n)$ definiert als

$$r_{xy} := \frac{c_{xy}}{s_x s_y} \quad (35.5)$$

und wird auch *Pearson's Stichprobenkorrelationskoeffizient* genannt.

•

Folgender **R** Code wertet die Stichprobenkorrelation des Beispieldatensatzes aus.

```
# Laden des Beispieldatensatzes
fname = "../data/502-korrelation.csv"
D = read.table(fname, sep = ",", header = TRUE)
x_i = D$x_i
y_i = D$y_i
n = length(x_i)

# "Manuelle" Berechnung der Stichprobenkorrelation
x_bar = (1/n)*sum(x_i)
y_bar = (1/n)*sum(y_i)
s_x = sqrt(1/(n-1)*sum((x_i - x_bar)^2))
s_y = sqrt(1/(n-1)*sum((y_i - y_bar)^2))
c_xy = 1/(n-1) * sum((x_i - x_bar) * (y_i - y_bar))
r_xy = c_xy/(s_x * s_y)
print(r_xy)
```

```
[1] 0.9378162
```

```
# Automatische Berechnung mit cor()
r_xy = cor(x_i, y_i)
print(r_xy)
```

```
[1] 0.9378162
```

Im Beispieldatensatz sind die Anzahl der Therapiestunden und die Symptomreduktion also mit $r_{xy} = 0.93$ hoch korreliert. Allgemein spricht man bei absoluten Werten von r_{xy} größer als etwa 0.70 von hoher Korrelation, bei absoluten Werten von r_{xy} zwischen etwa 0.30 und 0.70 von mittlerer Korrelation und bei absoluten Werten von r_{xy} zwischen 0.00 und 0.30 von niedriger Korrelation. Eine niedrige Korrelation zweier Variablen bedeutet

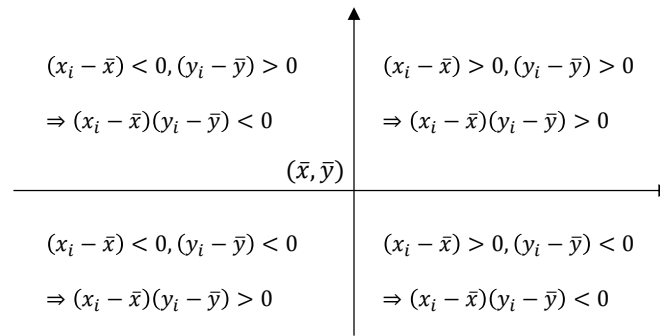


Abbildung 35.1. Mechanik der Stichprobenkorrelationssummenterme.

aber nicht zwangsläufig, dass diese Korrelation irrelevant ist (man denke an Gesundheitsrisikofaktoren), genauso wenig wie eine hohe Korrelation zweier Variablen trivial sein kann (man denke an die Korrelation von Körpergröße und Schuhgröße).

Da die Stichprobenkorrelation lediglich die auf das Intervall $[-1, 1]$ normalisierte Stichprobenkovarianz c_{xy} ist, wird die Höhe der Stichprobenkorrelation und insbesondere ihr Vorzeichen entscheidend durch die Werte Stichprobenkovarianzsummenterme $(x_i - \bar{x})(y_i - \bar{y})$ bestimmt. Dabei ist es entscheidend, wie häufig über die Datenpaare (x_i, y_i) hinweg die x_i und y_i gleichartig oder entgegengesetzt von ihren jeweiligen Stichprobenmitteln abweichen. Dies ist schematisch in Abbildung 35.1 dargestellt. Bei häufiger richtungsgleicher Abweichung von ihren jeweiligen Mittelwerten, sowohl in positiver als auch in negativer Richtung, ergibt das Produkt $(x_i - \bar{x})(y_i - \bar{y})$ eine positive Zahl, trägt also zu einer positiven Stichprobenkovarianz bei. Bei häufiger entgegengesetzter Abweichung von ihren jeweiligen Mittelwerten ergibt das Produkt $(x_i - \bar{x})(y_i - \bar{y})$ häufig eine negative Zahl, trägt also zu einer negativen Stichprobenkovarianz bei. Kommen sowohl richtungsgleiche als auch entgegengesetzte Abweichungen der x_i und y_i häufig vor, so gleichen sich positive und negative Beiträge zur Stichprobenkovarianzsumme eher aus und es resultiert eine geringe Stichprobenkovarianz bzw. ein Stichprobenkorrelationskoeffizient nahe Null. Abbildung 2.2 zeigt bivariate Datensätze von jeweils $n = 30$ Datenpunkten zusammen mit ihren jeweiligen Stichprobenkorrelationskoeffizienten.

Der Vorteil des Stichprobenkorrelationskoeffizienten gegenüber der Stichprobenkovarianz als Zusammenhangsmaß ist es, dass der absolute Wert des Stichprobenkorrelationskoeffizienten bei linear-affiner Transformation der zugrundeliegende Wertemenge gleich bleibt, wohingegen die Stichprobenkovarianz ihren Wert je nach gewähltem Maßstab ändert. Man sagt deshalb auch, dass der Stichprobenkorrelationkoeffizient maßstabsunabhängig ist. Dies ist die zentrale Aussage folgenden Theorems.

Theorem 35.1 (Stichprobenkorrelation bei linear affiner Transformation). *Für einen Datensatz $\{(x_i, y_i)\}_{i=1, \dots, n} \subset \mathbb{R}^2$ sei $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1, \dots, n} \subset \mathbb{R}^2$ der linear-affin transformierte Datensatz mit*

$$(\tilde{x}_i, \tilde{y}_i) = (a_x x_i + b_x, a_y y_i + b_y), a_x, a_y \neq 0. \quad (35.6)$$

gegeben. Dann gilt

$$|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \quad (35.7)$$

o

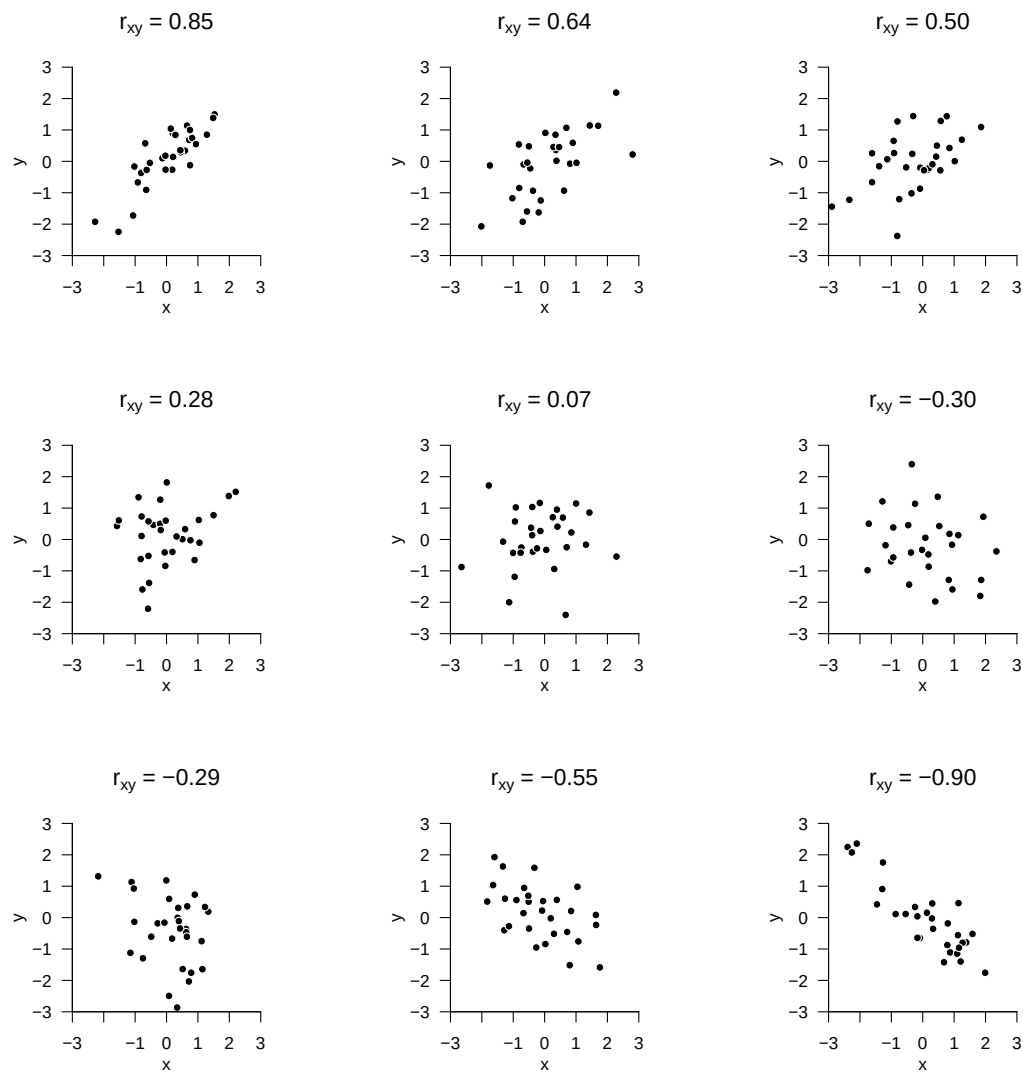


Abbildung 35.2. Beispiele bivariater Datensätze und ihrer Stichprobenkorrelationen.

Beweis.

$$\begin{aligned}
 r_{\tilde{x}\tilde{y}} &:= \frac{\frac{1}{n-1} \sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}}) (\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n \tilde{x}_i - \bar{\tilde{x}} \right)^2} \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n \tilde{y}_i - \bar{\tilde{y}} \right)^2}} \\
 &= \frac{\sum_{i=1}^n (a_x x_i + b_x - (a_x \bar{x} + b_x)) (a_y y_i + b_y - (a_y \bar{y} + b_y))}{\sqrt{\sum_{i=1}^n (a_x x_i + b_x - (a_x \bar{x} + b_x))^2} \sqrt{\sum_{i=1}^n (a_y y_i + b_y - (a_y \bar{y} + b_y))^2}} \\
 &= \frac{a_x a_y \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{a_x^2 \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{a_y^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (35.8) \\
 &= \frac{a_x a_y}{|a_x| |a_y|} \frac{\sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\
 &= \frac{a_x a_y}{|a_x| |a_y|} \frac{c_{xy}}{s_x s_y} \\
 &= \frac{a_x a_y}{|a_x| |a_y|} r_{xy}.
 \end{aligned}$$

Also folgt, durch Durchspielen aller möglichen Vorzeichenfälle, dass

$$|r_{\tilde{x}\tilde{y}}| = |r_{xy}|. \quad (35.9)$$

□

35.2. Korrelation und Bestimmtheitsmaß

Das sogenannte Bestimmtheitsmaß R^2 ist eine beliebte, häufig reportierte Statistik zur Beschreibung der Zusammenhangsstärke der Werte einer unabhängigen und einer abhängigen Variable. Numerisch handelt es sich bei R^2 lediglich um den quadrierten Stichprobenkorrelationskoeffizienten. Ist die Stichprobenkorrelation zum Beispiel $r_{xy} = 0.5$, dann ist $R^2 = 0.5^2 = 0.25$, ist die Stichprobenkorrelation dagegen $r_{xy} = -0.5$, dann gilt analog $R^2 = (-0.5)^2 = 0.25$. An diesen Beispielen erkennt man, dass R^2 weniger Information über die Rohdaten enthält als r_{xy} , da das Vorzeichen und damit die Richtung des Zusammenhangs wegfällt. Per se ist die Angabe von R^2 anstelle von r_{xy} als Deskriptivstatistik zur Beschreibung der Zusammenhangsstärke der Werte von unabhängiger und abhängiger Variable ohne Vorteil. Wir wollen hier trotzdem etwas genauer auf R^2 eingehen, da ein tieferes Verständnis von R^2 einerseits den Einstieg in das Konzept der Varianzzerlegungen erlaubt und andererseits die Zusammenhänge zwischen den Konzepten der Ausgleichsgerade und der Stichprobenkorrelation weiter verdeutlicht. Wir erweitern dazu zunächst die Beschreibung der Ausgleichsgerade aus Kapitel 34.1 durch die Begriffe der erklärten Werte und der Residuen.

Definition 35.3 (Erklärte Werte und Residuen einer Ausgleichsgerade.). Gegeben sei ein Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ und die zu diesem Datensatz gehörende Ausgleichsgerade

$$f_{\hat{\beta}} : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f_{\hat{\beta}}(x) := \hat{\beta}_0 + \hat{\beta}_1 x \quad (35.10)$$

Dann werden für $i = 1, \dots, n$

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (35.11)$$

die durch die Ausgleichsgerade *erklärten Werte* genannt und

$$\hat{\varepsilon}_i := y_i - \hat{y}_i \quad (35.12)$$

werden die *Residuen* der Ausgleichsgerade genannt.

•

Etwas allgemeiner formuliert sind die *erklärten Werte* damit die Datenvorhersage des Modells basierend auf den geschätzten Parameterwerten, während die Residuen die Differenzen zwischen den geschätzten Datenvorhersagen und den beobachteten Datenwerten bezeichnen. Abbildung 35.3 verdeutlicht diese Begriffe am Beispiel der Ausgleichsgerade des Beispieldatensatzes.

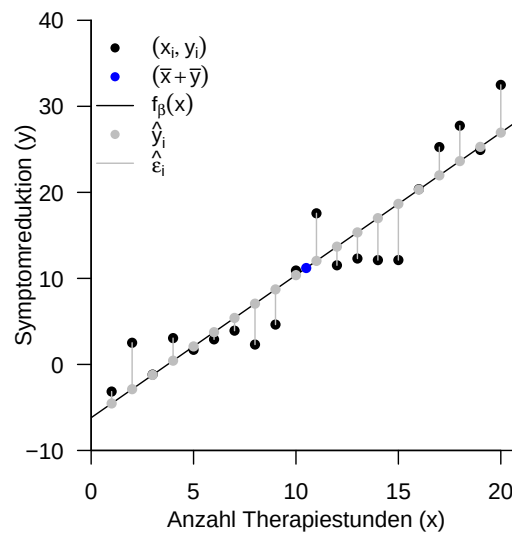


Abbildung 35.3. Ausgleichsgerade, erklärte Werte und Residuen für den Beispieldatensatz..

Mithilfe der Begriffe der erklärten Werte und Residuen lässt sich nun folgende Quadratsummenzerlegung beim Vorliegen einer Ausgleichsgerade eines Datensatzes angeben.

Theorem 35.2 (Quadratsummenzerlegung bei Ausgleichsgerade). *Für einen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ und seine zugehörige Ausgleichsgerade $f_{\hat{\beta}}$ seien für*

$$\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i \text{ und } \hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i, \text{ für } i = 1, \dots, n \quad (35.13)$$

das Stichprobenmittel der y_i -Werte und die durch die Ausgleichsgerade erklärten Werte, respektive. Weiterhin seien

- *die Total Sum of Squares definiert als*

$$SQT := \sum_{i=1}^n (y_i - \bar{y})^2 \quad (35.14)$$

- *die Explained Sum of Squares definiert als*

$$SQE := \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (35.15)$$

- die *Residual Sum of Squares* definiert als

$$SQR := \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (35.16)$$

Dann gilt

$$SQT = SQE + SQR \quad (35.17)$$

◦

Beweis.

$$\begin{aligned} SQT &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^n ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \\ &= \sum_{i=1}^n ((y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2) \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= SQE + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + SQR \\ &= SQE + SQR \end{aligned} \quad (35.18)$$

Dabei ergibt sich die letzte Gleichung mit

$$\bar{\hat{y}} := \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y} \quad (35.19)$$

und damit auch

$$\bar{\hat{y}} = \bar{y} \Leftrightarrow \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n y_i \Leftrightarrow \sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i \Leftrightarrow \bar{\hat{y}} \sum_{i=1}^n \hat{y}_i = \bar{y} \sum_{i=1}^n y_i \quad (35.20)$$

sowie

$$\bar{\hat{y}} = \bar{y} \Leftrightarrow \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n y_i \Leftrightarrow \sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i \Leftrightarrow \sum_{i=1}^n y_i \hat{y}_i = \sum_{i=1}^n \hat{y}_i \hat{y}_i \quad (35.21)$$

aus

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n (y_i \hat{y}_i - y_i \bar{y} - \hat{y}_i \hat{y}_i + \hat{y}_i \bar{y}) \\ &= \sum_{i=1}^n y_i \hat{y}_i - \sum_{i=1}^n y_i \bar{y} - \sum_{i=1}^n \hat{y}_i \hat{y}_i + \sum_{i=1}^n \hat{y}_i \bar{y} \\ &= \sum_{i=1}^n y_i \hat{y}_i - \sum_{i=1}^n \hat{y}_i \hat{y}_i + \bar{y} \sum_{i=1}^n \hat{y}_i - \bar{y} \sum_{i=1}^n y_i \\ &= 0 + 0 \\ &= 0 \end{aligned} \quad (35.22)$$

□

Die Begriffsbildungen von Theorem 35.2 erklären sich intuitiv wie folgt:

- SQT repräsentiert die Gesamtvariabilität der y_i -Werte um ihren Mittelwert $\bar{y}$.

- SQE repräsentiert die Variabilität der erklärten Werte $\hat{y}_i$ um ihren Mittelwert. Große Werte von SQE repräsentieren damit eine große absolute Steigung der y_i mit den x_i und kleine Werte von SQE repräsentieren eine kleine absolute Steigung der y_i mit den x_i . SQE ist somit ein Maß für die Stärke des linearen Zusammenhangs der x_i - und y_i -Werte
- SQR ist die Summe der quadrierten Residuen, denn es gilt

$$\text{SQR} := \sum_{i=1}^n (y_i - \hat{y}_i)^2 := \sum_{i=1}^n \hat{\varepsilon}_i^2 \quad (35.23)$$

Große Werte von SQR repräsentieren damit große Abweichungen der erklärten von den beobachteten y_i -Werten und kleine Werte von SQR repräsentieren geringe Abweichungen der erklärten von den beobachteten y_i -Werten. SQR ist also ein Maß für die Güte der Beschreibung der Datenmenge durch die Ausgleichsgerade.

Die zentrale Aussage des Theorem 2.1 ist nun, dass sich die Gesamtstreuung der y_i -Werte um ihren Mittelwert $\bar{y}$ gerade aus der Summe der Stärke des linearen Zusammenhangs der x_i - und y_i -Werte (also des “deterministischen Einflusses” der x_i auf die y_i) sowie den den Abweichungen von diesem linearen Zusammenhang (also dem “Rauschen”) zusammensetzt. Obwohl es sich bei SQT formal nicht um ein Varianzmaß handelt, spricht man in diesem Zusammenhang auch oft von einer Varianzzerlegung in erklärte Varianz und Residualvarianz. Dieses Motiv ist ein zentraler Aspekt des Allgemeinen Linearen Modells und wird in späteren Kapiteln erneut aufgegriffen werden. Für den Moment erlaubt Theorem 2.1 nun folgende Definition des Bestimmtheitsmaßes R^2 .

Definition 35.4 (Bestimmtheitsmaß R^2). Für einen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ und seine zugehörige Ausgleichsgerade $f_{\hat{\beta}}$ sowie die zugehörigen Explained Sum of Squares SQE und Total Sum of Squares SQT heißt

$$R^2 := \frac{\text{SQE}}{\text{SQT}} \quad (35.24)$$

das *Bestimmtheitsmaß* oder der *Determinationskoeffizient*.

•

Folgendes Theorem liefert nun den oben erwähnten Zusammenhang zwischen dem Bestimmtheitsmaß und der Stichprobenkorrelation.

Theorem 35.3 (Stichprobenkorrelation und Bestimmtheitsmaß). Für einen Datensatz $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^2$ sei R^2 das Bestimmtheitsmaß und r_{xy} sei die Stichprobenkorrelation. Dann gilt

$$R^2 = r_{xy}^2 \quad (35.25)$$

◦

Beweis. Wir halten zunächst fest, dass mit

$$\bar{\hat{y}} := \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y} \quad (35.26)$$

folgt, dass

$$\begin{aligned}
 \text{SQE} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
 &= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x})^2 \\
 &= \sum_{i=1}^n (\hat{\beta}_1 (x_i - \bar{x}))^2 \\
 &= \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2
 \end{aligned} \tag{35.27}$$

Damit ergibt sich dann

$$\begin{aligned}
 R^2 &= \frac{\text{SQE}}{\text{SQT}} \\
 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
 &= \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
 &= \frac{c_{xy}^2 \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}{s_x^4 \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\
 &= \frac{c_{xy}^2 s_x^2}{s_x^4 s_y^2} \\
 &= \frac{c_{xy}^2}{s_x^2 s_y^2} \\
 &= \left(\frac{c_{xy}}{s_x s_y} \right)^2 \\
 &= r_{xy}^2.
 \end{aligned} \tag{35.28}$$

□

Man beachte, dass mit $-1 \leq r_{xy} \leq 1$ aus Theorem 35.3 direkt folgt, dass $0 \leq R^2 \leq 1$. Nach Definition 35.4 gilt $R^2 = 0$ genau dann, wenn $\text{SQE} = 0$ ist. $R^2 = 0$ bedeutet also, dass die erklärte Datenvariabilität durch die Ausgleichsgerade gleich Null ist und beschreibt damit den Fall einer denkbar schlechten Datenerklärung durch die Ausgleichsgerade. Andererseits gilt $R^2 = 1$ genau dann, wenn $\text{SQE} = \text{SQT}$ ist. $R^2 = 1$ bedeutet also, dass die Gesamtstreuung gleich der durch die Ausgleichsgerade erklärten Streuung ist und beschreibt den Fall, dass sämtliche Datenvariabilität durch die Ausgleichsgerade erklärt werden kann. Man sagt deshalb auch oft etwas ungenau, dass R^2 die durch die Ausgleichsgerade erklärte Varianz an der Gesamtvarianz der Daten repräsentiert. Neben der jeweiligen Stichprobenkorrelation ist in Abbildung 35.2 auch jeweils das Bestimmtheitsmaß für die bivariaten Beispieldatensätze aufgeführt.

35.3. Korrelation und linear-affine Abhängigkeit

Die Tatsache, dass stochastische Unabhängigkeit zwar Unkorreliertheit impliziert, dass umgekehrt die Unkorreliertheit zweier Zufallsvariablen aber nicht ihre stochastische Unabhängigkeit impliziert, deutet daraufhin, dass die Korrelation zweier Zufallsvariablen nur

bestimmte Formen der Abhängigkeit zwischen Variablen misst. Abbildung 2.4 verdeutlicht dies anhand dreier Simulationsbeispiele.

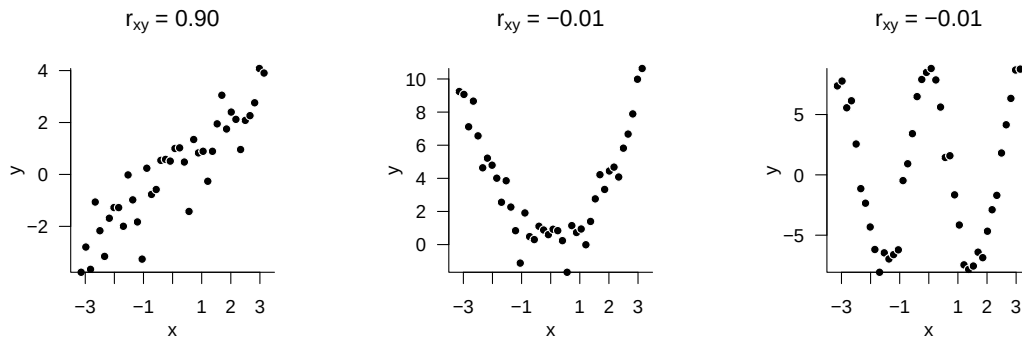


Abbildung 35.4. Korrelation und Abhängigkeit.

Abbildung 35.4 A zeigt eine Realisation des Modells

$$y_i = x_i + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, 1) \text{ für } i = 1, \dots, n. \quad (35.29)$$

Die Stichprobenkorrelationen des Datensatzes $\{(x_i, y_i)\}_{i=1}^n$ ergibt sich hier zu $r_{xy} = 0.90$. Es besteht eine klare Abhängigkeit des Wertes der y_i Realisation y_i vom Wert x_i : je höher der Wert von x_i , desto höher der Erwartungswert für den Wert von y_i . Abbildung 35.4 B zeigt eine Realisation des Modells

$$y_i = x_i^2 + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, 1) \text{ für } i = 1, \dots, n. \quad (35.30)$$

Die Stichprobenkorrelationen des Datensatzes $\{(x_i, y_i)\}_{i=1}^n$ ergibt sich hier zu $r_{xy} = -0.01$. Die Stichprobenkorrelation ist also minimal. Es besteht aber auch hier eine klare Abhängigkeit des Wertes der y_i Realisation y_i vom Wert x_i : je höher oder je niedriger der Wert von x_i , desto höher der Erwartungswert für den Wert von y_i , es besteht ein quadratischer Zusammenhang. Ein ähnliches Bild ergibt sich beim Betrachten von Abbildung 35.4 C einer Realisation des Modells

$$y_i = 8 \cos(2x_i) + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, 1) \text{ für } i = 1, \dots, n. \quad (35.31)$$

Auch hier ergibt sich das Bild einer Abhängigkeit des Wertes der y_i Realisation y_i vom Wert x_i , in diesem Fall im Sinne einer zyklischen Abhängigkeit, die Stichprobenkorrelation ist aber mit $r_{xy} = -0.01$ wiederum minimal. Diese Simulationsbeispiele belegen also intuitiv, dass die Stichprobenkorrelation verschwindend gering sein kann, auch wenn klare Abhängigkeiten zwischen zwei Variablen bestehen. Als allgemeines Maß für die Abhängigkeit zweier Variablen ist die Korrelation also ungeeignet. Diesen Umstand zugrunde liegt die Tatsache, dass die Korrelation lediglich ein Maß für den linear-affinen Zusammenhang zweier Zufallsvariablen, nicht aber für ihre stochastische Abhängigkeit ist. Wir formalisieren und präzisieren diese Aussage in folgenden Theorem.

Theorem 35.4 (Korrelation und linear-affine Abhängigkeit). *x und y seien zwei Zufallsvariablen mit positiver Varianz. Dann besteht genau dann eine lineare-affine Abhängigkeit der Form*

$$y = \beta_0 + \beta_1 x \text{ mit } \beta_0, \beta_1 \in \mathbb{R} \quad (35.32)$$

zwischen x und y , wenn

$$\rho(x, y) = 1 \text{ oder } \rho(x, y) = -1 \quad (35.33)$$

gilt.

◦

Beweis. Wir beschränken uns auf den Beweis der Aussage, dass aus $y = \beta_0 + \beta_1 x$ folgt, dass $\rho(x, y) = \pm 1$ ist. Dazu halten wir zunächst fest, dass mit den Theoremen zu den Eigenschaften von Erwartungswert und Varianz gilt, dass

$$\mathbb{E}(y) = \beta_0 + \beta_1 \mathbb{E}(x) \text{ und } \mathbb{V}(y) = \beta_1^2 \mathbb{V}(x) \quad (35.34)$$

Wegen $\mathbb{V}(x) > 0$ und $\mathbb{V}(y) > 0$ gilt damit $\beta_1 \neq 0$. Es folgt dann

$$\beta_1 > 0 \Rightarrow \mathbb{S}(y) = \beta_1 \mathbb{S}(x) > 0 \text{ und } \beta_1 < 0 \Rightarrow \mathbb{S}(y) = -\beta_1 \mathbb{S}(x) > 0. \quad (35.35)$$

Weiterhin gilt

$$\begin{aligned} y - \mathbb{E}(y) &= \beta_0 + \beta_1 x - \mathbb{E}(y) \\ &= \beta_0 + \beta_1 x - \beta_0 - \beta_1 \mathbb{E}(x) \\ &= \beta_1 x - \beta_1 \mathbb{E}(x) \\ &= \beta_1 (x - \mathbb{E}(x)). \end{aligned} \quad (35.36)$$

Für die Kovarianz von x und y ergibt sich also

$$\begin{aligned} \mathbb{C}(x, y) &= \mathbb{E}((y - \mathbb{E}(y))(x - \mathbb{E}(x))) \\ &= \mathbb{E}(\beta_1 (x - \mathbb{E}(x))(x - \mathbb{E}(x))) \\ &= \beta_1 \mathbb{E}((x - \mathbb{E}(x))^2) \\ &= \beta_1 \mathbb{V}(x). \end{aligned} \quad (35.37)$$

Damit ergibt für die Korrelation von x und y

$$\rho(x, y) = \frac{\mathbb{C}(x, y)}{\mathbb{S}(x)\mathbb{S}(y)} = \pm \frac{\beta_1 \mathbb{V}(x)}{\mathbb{S}(x)\beta_1 \mathbb{S}(x)} = \pm \frac{\beta_1 \mathbb{V}(x)}{\beta_1 \mathbb{V}(x)} = \pm 1 \quad (35.38)$$

□

Die Korrelation zweier Zufallsvariablen wird also genau dann maximal, wenn zwischen den beiden Zufallsvariablen ein linear-affiner Zusammenhang besteht. Dabei impliziert die linear-affine Abhängigkeit von y von x auch immer die linear-affine Abhängigkeit von x von y , denn

$$y = \beta_0 + \beta_1 x \Leftrightarrow -\beta_0 + y = \beta_1 x \Leftrightarrow x = -\frac{\beta_0}{\beta_1} + \frac{1}{\beta_1} y \Leftrightarrow x = \tilde{\beta}_0 + \tilde{\beta}_1 y \quad (35.39)$$

mit

$$\tilde{\beta}_0 = -\frac{\beta_0}{\beta_1} \text{ und } \tilde{\beta}_1 = \frac{1}{\beta_1}. \quad (35.40)$$

35.4. Literaturhinweise

Der Begriff der Korrelation erscheint, allerdings basierend auf früheren Arbeiten zum Beispiel von Bravais (1844), zunächst bei Galton (1890) (vgl. Stigler (1986)) und wird unter anderem durch die Arbeiten von Pearson (1895), Pearson (1896), Pearson (1900), Pearson (1901) im Kontext multivariater Normalverteilungen weiter ausgearbeitet. Eine frühe Studie zum Verhältnis von Korrelation und Kausalität ist Wright (1921).

36. Modellformulierung

36.1. Allgemeine Theorie

Wir definieren das Allgemeine Lineare Modell (ALM) wie folgt.

Definition 36.1 (Allgemeines Lineares Modell). Es sei

$$y = X\beta + \varepsilon \quad (36.1)$$

wobei

- y ein n -dimensionaler beobachtbarer Zufallsvektor ist, der *Daten* genannt wird,
- $X \in \mathbb{R}^{n \times p}$ für $n > p$ und $\text{rg}(X) = p$ eine Matrix ist, die *Designmatrix* genannt wird,
- $\beta \in \mathbb{R}^p$ ein unbekannter Parametervektor ist, der *Betaparametervektor* genannt wird,
- ε ein n -dimensionaler nicht-beobachtbarer Zufallsvektor ist, der *Zufallsfehler* genannt wird und für den angenommen wird, dass mit einem unbekannten Varianzparameter $\sigma^2 > 0$ gilt, dass

$$\varepsilon \sim N(0_n, \sigma^2 I_n) \quad (36.2)$$

Dann heißt Gleichung 36.1 *Allgemeines Lineares Modell (ALM)*.

•

In Gleichung 36.1 bezeichnen wir $X\beta \in \mathbb{R}^n$ als den *deterministischen Aspekt* des ALMs und ε als den *probabilistischen Aspekt* des ALMs. Das ALM postuliert also, dass Daten aus der Addition eines deterministischen Aspektes $X\beta \in \mathbb{R}^n$ unter der Addition eines multivariat normalverteilten probabilistischen Aspektes ε zustande kommen. Man beachte, dass $X\beta \in \mathbb{R}^n$ ein n -dimensionaler Vektor und ε ein n -dimensionaler Zufallsvektor ist. Der resultierende Vektor y ist ein Zufallsvektor, weil er aus der Addition des Zufallsvektors ε zu dem Vektor $X\beta \in \mathbb{R}^n$ resultiert. Das ALM ist also ein probabilistisches Modell bei dem durch y vorliegende Datensätze modelliert werden. Generativ betrachtet entsteht im ALM ein Datensatz $y \in \mathbb{R}^n$ als Realisierung von y dann durch Addition des deterministischen Modellaspekts und einer nicht direkt beobachtbaren Realisierung $e \in \mathbb{R}^n$ von ε ,

$$y = X\beta + e \quad (36.3)$$

mit $y \in \mathbb{R}^n$, $X\beta \in \mathbb{R}^n$ und $e \in \mathbb{R}^n$.

Die Gesamtzahl an Parametern des ALMs beträgt $p+1$, bestehend aus p skalaren Betaparametern und einem Varianzparameter σ^2 . Der Betaparametervektor $\beta \in \mathbb{R}^p$ wird dabei auch Gewichtsvektor oder Effektvektor genannt. Seine Einträge wichten die Einträge der Spalten der Designmatrix $X \in \mathbb{R}^{n \times p}$ in der Erzeugung des deterministischen Modellaspekts

$X\beta \in \mathbb{R}^n$. Die Spalten der Designmatrix werden in unterschiedlichen Kontexten unterschiedlich bezeichnet, gebräuchliche Bezeichnungen sind zum Beispiel Prädiktoren, Regressoren oder Kovariaten. Allgemein betrachtet modellieren die Spalten der Designmatrix unabhängige Variablen und der Datenvektor abhängige Variablen.

Man beachte, dass der Kovarianzmatrixparameter von ε als sphärisch angenommen wird. Damit folgt direkt, dass die $\varepsilon_1, \dots, \varepsilon_n$ unabhängige normalverteilte Zufallsvariablen mit identischem Varianzparameter sind. Weil für ε zusätzlich der Erwartungswertparameter als $0_n \in \mathbb{R}^n$ angenommen wird, sind die $\varepsilon_1, \dots, \varepsilon_n$ auch identisch normalverteilte Zufallsvariablen. Wenn $x_{ij} \in \mathbb{R}$ das ij te Element der Designmatrix $X \in \mathbb{R}^{n \times p}$ bezeichnet, dann gilt damit für jede Komponente $y_i, i = 1, \dots, n$ von y nach Gleichung 36.1, dass

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{ip}\beta_p + \varepsilon_i \text{ mit } \varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2). \quad (36.4)$$

Die in Gleichung 36.1 implizite Verteilung des Datenvektors y halten wir in folgendem Theorem fest.

Theorem 36.1 (Datenverteilung des Allgemeinen Linearen Modells). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (36.5)$$

das ALM. Dann gilt

$$y \sim N(\mu, \sigma^2 I_n) \text{ mit } \mu := X\beta \in \mathbb{R}^n \quad (36.6)$$

◦

Beweis. Mit dem Theorem zur linear-affinen Transformation von multivariaten Normalverteilungen gilt für

$$\varepsilon \sim N(0_n, \sigma^2 I_n) \text{ und } y := I_n \varepsilon + X\beta, \quad (36.7)$$

dass

$$y \sim N(I_n 0_n + X\beta, I_n (\sigma^2 I_n) I_n^T) = N(X\beta, \sigma^2 I_n). \quad (36.8)$$

Mit der Definition $\mu := X\beta \in \mathbb{R}^n$ folgt die Aussage des Theorems dann direkt. $\square$

Im ALM sind die Daten y also ein n -dimensionaler normalverteilter Zufallsvektor mit Erwartungswertparameter $\mu = X\beta \in \mathbb{R}^n$ und Kovarianzmatrixparameter $\sigma^2 I_n \in \mathbb{R}^{n \times n}$. Das ALM ist also eine multivariate Normalverteilung deren Erwartungswertparameter mithilfe einer Designmatrix und eines Betaparametervektors parameterisiert ist. Weiterhin sind die Komponenten $y_1, \dots, y_n$ von y , also die Zufallsvariablen, die skalare Datenpunkte modellieren, damit unabhängige normalverteilte Zufallsvariablen der Form

$$y_i \sim N((X\beta)_i, \sigma^2) \text{ für } i = 1, \dots, n. \quad (36.9)$$

Da im Allgemeinen aber $(X\beta)_i \neq (X\beta)_j$ für $i \neq j$ gilt, sind die $y_i, i = 1, \dots, n$ im Allgemeinen aber nicht identisch verteilt. Das Szenario unabhängig und identisch normalverteilter Zufallsvariablen kann aber natürlich als Spezialfall des ALMs formuliert werden.

Beispiel (1) Unabhängige und identisch normalverteilte Zufallsvariablen

Wir betrachten das Szenario von n unabhängigen und identisch normalverteilten Zufallsvariablen mit Erwartungswertparameter $\mu \in \mathbb{R}$ und Varianzparameter σ^2 ,

$$y_1, \dots, y_n \sim N(\mu, \sigma^2). \quad (36.10)$$

Dann gilt, dass Gleichung 36.10 äquivalent ist zu

$$y_i = \mu + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, \dots, n \text{ mit unabhängigen } \varepsilon_i. \quad (36.11)$$

In Matrixschreibweise ist dies wiederum äquivalent zu dem ALM Spezialfall

$$y \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^{n \times 1}, \beta := \mu \in \mathbb{R}^1, \sigma^2 > 0. \quad (36.12)$$

In **R** können Realisierungen des ALMs leicht mithilfe eines Zufallszahlengenerators für multivariate Normalverteilungen durch Spezifikation der entsprechenden Erwartungs- und Kovarianzparameter gewonnen werden. Folgender **R** Code zeigt, wie n unabhängig und identisch normalverteilte skalare Datenpunkte im Sinne des ALMs realisiert werden können. Man beachte, dass n skalare Datenpunkte dabei einer Realisierung des ALMs entsprechen.

```
# Modellformulierung
library(MASS)
set.seed(0)
n = 12
p = 1
X = matrix(rep(1,n), nrow = n)
I_n = diag(n)
beta = 2
sigsqr = 1

# Multivariate Normalverteilung
# Zufallszahlengeneratorseed
# Anzahl von Datenpunkten
# Anzahl von Betaparametern
# Designmatrix
# n x n Einheitsmatrix
# wahrer, aber unbekannter, Betaparameter
# wahrer, aber unbekannter, Varianzparameter

# Datenrealisierung
y = mvrnorm(1, X %*% beta, sigsqr*I_n)

# eine Realisierung eines n-dimensionalen ZVs
```

Realisierungen: 1.2 2.76 4.4 1.99 1.71 1.07 0.46 2.41 3.27 3.33 1.67 3.26

Beispiel (2) Einfache lineare Regression

Wir betrachten das generative Modell der einfachen linearen Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, \dots, n, \quad (36.13)$$

Wir haben bereits gesehen, dass dieses Modell äquivalent ist zu dem Normalverteilungsmodell der Regression

$$y_i \sim N(\mu_i, \sigma^2) \text{ mit } \mu_i := \beta_0 + \beta_1 x_i \text{ für } i = 1, \dots, n. \quad (36.14)$$

In Matrixschreibweise ist dies wiederum äquivalent zu dem ALM Spezialfall

$$y \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n \times 2}, \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \in \mathbb{R}^2, \sigma^2 > 0. \quad (36.15)$$

R Code zur Simulation von Realisierungen einer einfachen linearen Regression hat dementsprechend eine sehr ähnliche Struktur wie obiger **R** Code zur Simulation von Realisierungen von n unabhängig und identisch normalverteilten Zufallsvariablen.

```

# Modellformulierung
library(MASS)
set.seed(0)
n = 10
p = 2
x = 1:n
X = matrix(c(rep(1,n),x), nrow = n)
I_n = diag(n)
beta = matrix(c(0,1), nrow = p)
sigsqr = 1

# Multivariate Normalverteilung
# Zufallzahlengeneratorseed
# Anzahl von Datenpunkten
# Anzahl von Betaparametern
# Prädiktorwerte
# Designmatrix
# n x n Einheitsmatrix
# wahrer, aber unbekannter, Betaparameter
# wahrer, aber unbekannter, Varianzparameter

# Datenrealisierung
y = mvrnorm(1, X %*% beta, sigsqr*I_n)

# eine Realisierung eines n-dimensionalen ZVs

```

Realisierungen: 3.4 1.99 2.71 3.07 3.46 6.41 8.27 9.33 8.67 11.26

Wir visualisieren obige Realisierung in Abbildung 36.1.

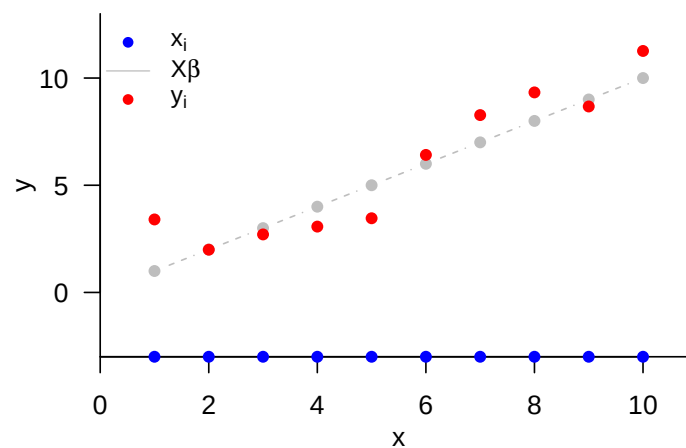


Abbildung 36.1. Realisierung einer einfachen linearen Regression.

36.2. Identifizierbarkeit und Schätzbarkeit

Wir haben oben gesehen, dass die Datenverteilung des ALM durch

$$y \sim N(\mu, \sigma^2 I_n) \text{ mit } \mu := X\beta \in \mathbb{R}^n \text{ für } X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p \quad (36.16)$$

gegeben ist. Das ALM ist also eine multivariate Normalverteilung mit einer speziellen Erwartungswertparameterisierung. Um die Begriffe der Identifizierbarkeit und Schätzbarkeit im Kontext von ALMs einzuführen ist es hilfreich den Begriff der *Parametrisierung einer multivariaten Normalverteilung* zunächst etwas zu verallgemeinern.

Definition 36.2 (Betaparameterisierung). $N(\mu, \sigma^2 I_n)$ sei eine multivariate Normalverteilung mit sphärischem Kovarianzmatrixparameter. Dann bezeichnen wir eine multivariate, vektorwertige Funktion der Form

$$f : \mathbb{R}^p \rightarrow \mathbb{R}^n, \beta \mapsto f(\beta) =: \mu \quad (36.17)$$

als eine Betaparametrisierung von μ .

•

Das ALM beruht offenbar auf der Designmatrix-abhängigen Betaparametrisierung

$$f : \mathbb{R}^p \rightarrow \mathbb{R}^n, \beta \mapsto f(\beta) := X\beta. \quad (36.18)$$

Mit Hilfe des Begriffs der Betaparametrisierung können wir nun den Begriff der *Betaparameteridentifizierbarkeit* formulieren:

Definition 36.3 (Betaparameteridentifizierbarkeit). $N(\mu, \sigma^2 I_n)$ sei eine multivariate Normalverteilung mit sphärischem Kovarianzmatrixparameter und f sei eine Betaparametrisierung von μ . β heißt dann und nur dann identifizierbar, wenn für beliebige $\beta_1, \beta_2 \in \mathbb{R}^p$ gilt, dass aus $f(\beta_1) = f(\beta_2)$ folgt, dass $\beta_1 = \beta_2$ gilt.

•

Die Betaparameter allgemeiner linearer Modelle, deren Designmatrix vollen Rang hat, sind identifizierbar. Dies ist die Aussage folgenden Theorems:

Theorem 36.2 (Betaparameteridentifizierbarkeit bei vollem Designmatrixrang). $y \sim N(X\beta, \sigma^2 I_n)$ sei die Datenverteilung eines ALMs mit $\text{rg}(X) = p$. Dann ist β identifizierbar.

◦

Beweis. Für $X \in \mathbb{R}^{n \times p}$ impliziert $\text{rg}(X) = p$, dass $(X^T X) \in \mathbb{R}^{p \times p}$ eine invertierbare Matrix ist. Dann aber gilt für beliebige $\beta_1, \beta_2 \in \mathbb{R}^p$

$$\begin{aligned} f(\beta_1) = f(\beta_2) &\Leftrightarrow X\beta_1 = X\beta_2 \\ &\Leftrightarrow X^T X\beta_1 = X^T X\beta_2 \\ &\Leftrightarrow (X^T X)^{-1} X^T X\beta_1 = (X^T X)^{-1} X^T X\beta_2 \\ &\Leftrightarrow \beta_1 = \beta_2. \end{aligned} \quad (36.19)$$

□

Im Rahmen der Analyse schätzbarer Funktionen benötigen wir weiterhin den Begriff der Identifizierbarkeit von vektorwertigen Funktionen der Betaparameter. Wir definieren:

Definition 36.4 (Identifizierbarkeit von Funktionen der Betaparameter). $N(\mu, \sigma^2 I_n)$ sei eine multivariate Normalverteilung mit sphärischem Kovarianzmatrixparameter und f sei eine Betaparametrisierung von μ . Weiterhin sei

$$g : \mathbb{R}^p \rightarrow \mathbb{R}^k, \beta \mapsto g(\beta) \quad (36.20)$$

eine Funktion des Betaparametervektors. g heißt dann und nur dann identifizierbar, wenn für beliebige $f(\beta_1), f(\beta_2) \in \mathbb{R}^n$ gilt, dass aus $f(\beta_1) = f(\beta_2)$ folgt, dass $g(\beta_1) = g(\beta_2)$.

•

Schätzbare Funktionen sind lineare Funktionen von β , die identifizierbar sind. Allgemein gilt folgendes Theorem:

Theorem 36.3 (Identifizierbare Betaparameterfunktionen). $N(\mu, \sigma^2 I_n)$ sei eine multivariate Normalverteilung mit sphärischem Kovarianzmatrixparameter und f sei eine Betaparametrisierung von μ . Weiterhin sei

$$g : \mathbb{R}^p \rightarrow \mathbb{R}^k, \beta \mapsto g(\beta) \quad (36.21)$$

eine Funktion des Betaparametervektors. Die Funktion g ist dann und nur dann identifizierbar, wenn g eine Funktion von f ist, wenn also eine Funktion ϕ existiert, so dass

$$g = \phi \circ f \quad (36.22)$$

◦

Beweis. Wir zeigen die Aussage lediglich in eine Richtung. $\Rightarrow$ Wir nehmen an, es existiert eine Funktion ϕ , so dass

$$g = \phi \circ f. \quad (36.23)$$

Dann impliziert die Tatsache, dass eine Funktion einem Argument genau einen Funktionswert zuordnet, dass gilt

$$f(\beta_1) = f(\beta_2) \Leftrightarrow \phi(f(\beta_1)) = \phi(f(\beta_2)) \Leftrightarrow g(\beta_1) = g(\beta_2) \quad (36.24)$$

Also ist g identifizierbar, denn aus $f(\beta_1) = f(\beta_2)$ folgt, dass $g(\beta_1) = g(\beta_2)$. $\square$

Wie oben bereits erwähnt sind schätzbare Funktionen lineare Funktionen von β , die identifizierbar sind. Die klassische Definition einer schätzbaren Funktion ist folgende.

Definition 36.5 (Schätzbare Funktion). $N(X\beta, \sigma^2 I_n)$ sei ein ALM. Dann heißt eine lineare Funktion

$$g : \mathbb{R}^p \rightarrow \mathbb{R}^k, \beta \mapsto g(\beta) := C^T \beta \quad (36.25)$$

mit $C \in \mathbb{R}^{p \times k}$ schätzbar, wenn eine Matrix $P \in \mathbb{R}^{n \times k}$ existiert, so dass

$$C^T \beta = P^T X \beta. \quad (36.26)$$

•

Diese Definition erschließt sich wie folgt: Nach dem Theorem zu identifizierbaren Funktionen muss eine identifizierbare lineare Funktion von β eine Funktion der Form

$$g(\beta) = (\phi \circ f)(\beta) \quad (36.27)$$

sein. Da weiterhin gilt, dass für ein ALM $f(\beta) = X\beta$ und dass g eine lineare Funktion ist, also mithilfe einer Matrix C^T geschrieben werden kann, muss auch ϕ eine lineare Funktion sein. Damit kann aber auch ϕ geschrieben werden als

$$\phi(f(\beta)) = \phi(X\beta) = P^T X \beta \quad (36.28)$$

mit einer geeigneten Matrix $P \in \mathbb{R}^{n \times k}$.

36.3. Designspektrum

Die Wahl von Designmatrix und Betaparameter öffnet eine große Freiheit zur Implementation verschiedenster Erwartungswertparameterszenarien der ALM Datenverteilung. Prinzipiell liegen dabei alle speziellen Designs in einem Kontinuum zwischen folgenden beiden Extrema:

- (1) Die Erwartungswerte aller Datenvariablen sind identisch, d.h.

$$y_i \sim N(\mu, \sigma^2) \text{ u.i.v. für } i = 1, \dots, n \quad (36.29)$$

also

$$y = X\beta + \varepsilon \text{ mit } X := \mathbf{1}_n \in \mathbb{R}^{n \times 1}, \beta := \mu \in \mathbb{R}, \varepsilon \sim N(0_n, \sigma^2 I_n). \quad (36.30)$$

- (2) Die Erwartungswerte aller Datenvariablen sind paarweise verschieden, d.h.

$$y_i \sim N(\mu_i, \sigma^2) \text{ u.v. für } i = 1, \dots, n, \quad (36.31)$$

also

$$y = X\beta + \varepsilon \text{ mit } X := I_n \in \mathbb{R}^{n \times n}, \beta := (\mu_1, \dots, \mu_n)^T \in \mathbb{R}^n, \varepsilon \sim N(0_n, \sigma^2 I_n). \quad (36.32)$$

In Szenario (1) wird jegliche Datenvariabilität dem Zufallsfehlerterm zugeschrieben, in Szenario (2) wird dagegen jegliche Datenvariabilität dem Erwartungswertparameter zugeschrieben. Beide Extremszenarien sind wissenschaftlich nicht ergiebig, da sie keine theoriegeleitete systematische Abhängigkeit zwischen unabhängigen und abhängigen Variablen repräsentieren. Alle im weiteren Verlauf betrachteten ALM Designs liegen damit zwischen den beiden Extremszenarien und repräsentieren verschiedene Formen der systematischen Abhängigkeit zwischen unabhängigen und abhängigen Variablen. Insbesondere unterscheidet man

- (1) *Faktorielle Designs*, bei denen die Designmatrix im Wesentlichen nur 1en und 0en und manchmal -1en enthält. In diesem Fall repräsentieren die Betaparameter Gruppenerwartungswerte und wir werden sehen, dass die Betaparameterschätzer in der Repräsentationen von Gruppenstichprobenmitteln resultieren. Man sagt manchmal, dass faktorielle Designs der Untersuchung von *Unterschiedshypothesen* dienen. Beispiele für faktorielle Designs sind verschiedene Designs zur Implementation von *T-Tests* und *Varianzanalysen*.
- (2) *Parameterische Designs*, bei denen die Designmatrix aus Spalten mit kontinuierlichen reellen Werten besteht. Vor allem in diesem Kontext werden die Spalten der Designmatrix oft als *Regressoren* oder *Prädiktoren* bezeichnet. In diesem Fall repräsentieren die Betaparameter partielle Steigungsparameter und wir werden sehen, dass sich die entsprechenden Betaparameterschätzer als normalisierte Regressor-Daten Kovarianzen ergeben. Man sagt manchmal, dass parametrische Designs zur Untersuchung von *Zusammenhangshypothesen* dienen. Beispiele für parametrische Designs sind verschiedene Designs zur Implementation von *einfacher linearer Regression* und insbesondere *multipler linearer Regression*.
- (3) *Faktoriell-parametrische Designs*, bei denen die Spalten der Designmatrix sowohl faktorielle als auch parametrische Prädiktoren repräsentieren. In diesem Kontext werden die parametrischen Regressoren oft als *Kovariaten* betrachtet. Gemischt

faktoriell-parametrische Designs sind das zentrale Charakteristikum der *Kovarianzanalyse* die auf eine kontrollierte Untersuchung von Unterschiedshypothesen bei Vorliegen weiterer möglicher Abhängigkeiten zwischen unabhängigen und abhängigen Variablen bzw. auf die kontrollierte Untersuchung von Zusammenhangshypothesen bei Vorliegen weiterer möglicher Gruppenunterschiede abzielt.

36.4. Literaturhinweise

Das Allgemeine Lineare Modell hat eine lange Geschichte, deren moderne Inkarnation üblicherweise auf die Arbeiten von Legendre (1805) und Gauss (1809) zurückgeführt wird. Matrixbasierte Formulierungen der multiplen Regression finden sich spätestens bei Aitken (1936) und Scheffé (1959). Eingang in den psychologischen Methodenkanon findet das Allgemeine Lineare Modell spätestens mit Cohen (1968). Seal (1967) gibt einen ausführlichen Überblick zur Geschichte des Allgemeinen Linearen Modells im 19. und der ersten Hälfte des 20. Jahrhunderts. Die in diesem Abschnitt gegebene Diskussion von Identifizierbarkeit und Schätzbarkeit beruht auf der Darstellung in Christensen (2011).

37. Parameterschätzung

In diesem Abschnitt betrachten wir die Frequentistische Punktschätzung von Betaparametervektor und Varianzparameter im ALM. Als Beispielanwendungen betrachten wir das Szenario n unabhängig und identisch normalverteilter Zufallsvariablen und das Szenario der einfachen linearen Regression. Wir schließen mit der Dokumentation der Frequentistischen Parameterschätzerverteilungen des ALMs.

37.1. Betaparameterschätzung

Wir fassen die Frequentistische Punktschätzung des Betaparametervektors in folgendem Theorem zusammen.

Theorem 37.1 (Betaparameterschätzer). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (37.1)$$

das Allgemeine Lineare Modell und es sei

$$\hat{\beta} := (X^T X)^{-1} X^T y. \quad (37.2)$$

Dann gelten:

- (1) $\hat{\beta}$ ist der KQ-Schätzer von $\beta \in \mathbb{R}^p$, für einen beliebigen festen Wert $y \in \mathbb{R}^n$ von y gilt also

$$\hat{\beta} = \operatorname{argmin}_{\tilde{\beta}} (y - X\tilde{\beta})^T (y - X\tilde{\beta}). \quad (37.3)$$

- (2) $\hat{\beta}$ ist ein unverzerrter Maximum-Likelihood-Schätzer von $\beta \in \mathbb{R}^p$.

◦

Beweis. (1) Wir zeigen in einem ersten Schritt, dass $\hat{\beta}$ ein KQ-Schätzer ist, dass also $\hat{\beta}$ für einen beliebigen festen Wert $y \in \mathbb{R}^n$ von y die Summe der Abweichungsquadrate

$$(y - X\tilde{\beta})^T (y - X\tilde{\beta}) = \sum_{i=1}^n (y_i - (X\tilde{\beta})_i)^2 \quad (37.4)$$

minimiert (die Notation $\tilde{\beta}$ für das Minimierungsargument dient hier lediglich dazu, es vom wahren, aber unbekannten, Parameterwert $\beta \in \mathbb{R}^p$ abzugrenzen und ist ansonsten ohne Bedeutung). Dazu halten wir zunächst fest, dass

$$\hat{\beta} = (X^T X)^{-1} X^T y \Leftrightarrow X^T X \hat{\beta} = X^T y \Leftrightarrow X^T y - X^T X \hat{\beta} = 0_p \Leftrightarrow X^T (y - X \hat{\beta}) = 0_p. \quad (37.5)$$

Weiterhin gilt dann auch, dass

$$X^T (y - X \hat{\beta}) = 0_p \Leftrightarrow (X^T (y - X \hat{\beta}))^T = 0_p^T \Leftrightarrow (y - X \hat{\beta})^T X = 0_p^T \quad (37.6)$$

Weiterhin halten wir ohne Beweis fest, dass für jede Matrix $X \in \mathbb{R}^{n \times p}$ gilt, dass

$$z^T X^T X z \geq 0 \text{ für alle } z \in \mathbb{R}^p. \quad (37.7)$$

Wir betrachten nun für festes y und ein beliebiges $\tilde{\beta}$ die Summe der Abweichungsquadrate

$$(y - X\tilde{\beta})^T (y - X\tilde{\beta}) \quad (37.8)$$

Es ergibt sich

$$\begin{aligned} & (y - X\tilde{\beta})^T (y - X\tilde{\beta}) \\ &= (y - X\hat{\beta} + X\hat{\beta} - X\tilde{\beta})^T (y - X\hat{\beta} + X\hat{\beta} - X\tilde{\beta}) \\ &= ((y - X\hat{\beta}) + X(\hat{\beta} - \tilde{\beta}))^T ((y - X\hat{\beta}) + X(\hat{\beta} - \tilde{\beta})) \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) + (y - X\hat{\beta})^T X(\hat{\beta} - \tilde{\beta}) + (\hat{\beta} - \tilde{\beta})^T X^T (y - X\hat{\beta}) + (\hat{\beta} - \tilde{\beta})^T X^T X(\hat{\beta} - \tilde{\beta}) \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) 0_p^T (\hat{\beta} - \tilde{\beta}) + (\hat{\beta} - \tilde{\beta})^T 0_p + (\hat{\beta} - \tilde{\beta})^T X^T X(\hat{\beta} - \tilde{\beta}) \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) + (\hat{\beta} - \tilde{\beta})^T X^T X(\hat{\beta} - \tilde{\beta}) \end{aligned} \quad (37.9)$$

Auf der rechten Seite obiger Gleichung ist nur der zweite Term von $\tilde{\beta}$ abhängig. Da für diesen Term gilt, dass

$$(\hat{\beta} - \tilde{\beta})^T X^T X(\hat{\beta} - \tilde{\beta}) \geq 0 \quad (37.10)$$

nimmt dieser Term genau dann seinen Minimalwert 0 an, wenn

$$(\hat{\beta} - \tilde{\beta}) = 0_p \Leftrightarrow \tilde{\beta} = \hat{\beta} \quad (37.11)$$

Also gilt

$$\hat{\beta} = \operatorname{argmin}_{\tilde{\beta}} (y - X\tilde{\beta})^T (y - X\tilde{\beta}). \quad (37.12)$$

(2) Um zu zeigen, dass $\hat{\beta}$ ein Maximum Likelihood Schätzer ist, betrachten wir für einen beliebigen Wert $y \in \mathbb{R}^n$ von y und festes $\sigma^2 > 0$ die Log-Likelihood Funktion

$$\ell : \mathbb{R}^p \rightarrow \mathbb{R}, \tilde{\beta} \mapsto \ln p_{\tilde{\beta}}(y) = \ln N(y; X\tilde{\beta}, \sigma^2 I_n) \quad (37.13)$$

wobei gilt, dass

$$\begin{aligned} \ln N(y; X\tilde{\beta}, \sigma^2 I_n) &= \ln \left((2\pi)^{-\frac{n}{2}} |\sigma^2 I_n|^{-\frac{1}{2}} \exp \left(-\frac{1}{2\sigma^2} (y - X\tilde{\beta})^T (y - X\tilde{\beta}) \right) \right) \\ &= -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\sigma^2 I_n| - \frac{1}{2\sigma^2} (y - X\tilde{\beta})^T (y - X\tilde{\beta}) \end{aligned} \quad (37.14)$$

Dabei hängt allein der Term $-\frac{1}{2\sigma^2} (y - X\tilde{\beta})^T (y - X\tilde{\beta})$ von $\tilde{\beta}$ ab. Weil aber $(y - X\tilde{\beta})^T (y - X\tilde{\beta}) \geq 0$ gilt, wird dieser Term aufgrund des negativen Vorzeichens maximal, wenn $(y - X\tilde{\beta})^T (y - X\tilde{\beta})$ minimal wird. Dies ist aber wie oben gezeigt genau für $\tilde{\beta} = \hat{\beta}$ der Fall. Die Unverzerrtheit von $\hat{\beta}$ schließlich ergibt sich aus

$$\mathbb{E}(\hat{\beta}) = \mathbb{E} \left((X^T X)^{-1} X^T y \right) = (X^T X)^{-1} X^T \mathbb{E}(y) = (X^T X)^{-1} X^T X \beta = \beta. \quad (37.15)$$

□

Theorem 37.1 gibt mit

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (37.16)$$

eine Formel an, um β anhand der Designmatrix und einer Realisierung $y \in \mathbb{R}^n$ von y konkret zu schätzen. Als Zufallsvektor ist $\hat{\beta}$ ein unverzerrter Schätzer von β und als Maximum-Likelihood-Schätzer insbesondere auch konsistent, asymptotisch normalverteilt und asymptotisch effizient. Wir sehen an späterer Stelle dass $\hat{\beta}$ sogar normalverteilt ist. Neben den genannten Eigenschaften hat $\hat{\beta}$ noch weitere gute Eigenschaften. Zum Beispiel besitzt $\hat{\beta}$ die kleinste Varianz in der Klasse der linearen unverzerrten Schätzer von β . Diese Eigenschaft ist Kernaussage des *Gauss-Markov Theorems*, auf das wir hier aber nicht näher eingehen wollen.

Mithilfe des Betaparameterschätzers können wir die Begriffe der erklärten Daten, des Residuenvektors und der Residuen definieren, die wir an vielen Stellen benötigen werden.

Definition 37.1 (Erklärte Daten, Residuenvektor und Residuen). Es sei

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (37.17)$$

das Allgemeine Lineare Modell und es sei

$$\hat{\beta} := (X^T X)^{-1} X^T y. \quad (37.18)$$

der Betaparameterschätzer. Dann heißt der Zufallsvektor

$$\hat{y} := X\hat{\beta} = X(X^T X)^{-1} X^T y \quad (37.19)$$

die *erklärten Daten*, der Zufallsvektor

$$\hat{\varepsilon} := y - \hat{y} = y - X\hat{\beta} \quad (37.20)$$

heißt *Residuenvektor* und für $i = 1, \dots, n$ heißen die Komponenten dieses Zufallsvektors

$$\hat{\varepsilon}_i := y_i - \hat{y}_i = y_i - (X\hat{\beta})_i \quad (37.21)$$

die *Residuen*.

•

37.2. Varianzparameterschätzung

Wir fassen die Frequentistische Punktschätzung des Varianzparameters in folgendem Theorem zusammen, das wir an dieser Stelle nicht beweisen wollen.

Theorem 37.2 (Varianzparameterschätzer). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (37.22)$$

das Allgemeine Lineare Modell. Dann ist

$$\hat{\sigma}^2 := \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - p} \quad (37.23)$$

ein unverzerrter Schätzer von $\sigma^2 > 0$.

◦

Theorem 37.2 gibt mit

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} \quad (37.24)$$

eine Formel an, um σ^2 anhand der Designmatrix, des Betaparameterschätzers und einer Realisierung $y \in \mathbb{R}^n$ von y zu schätzen. Offenbar gilt mit Theorem 37.2, dass

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{i=1}^n (y_i - (X\hat{\beta})_i)^2 \quad (37.25)$$

$\hat{\sigma}^2$ wird also durch die Summe der quadrierten Residuen, also als eine Summe von Abweichungsquadraten geschätzt. Für einen Beweis von Theorem 37.2 verweisen wir zum Beispiel auf S. R. Searle (1971), S. R. Searle & Gruber (2017) oder Rencher & Schaalje (2008). Aus probabilistischer Perspektive handelt es sich bei $\hat{\sigma}^2$ nicht um einen Maximum-Likelihood-Schätzer, sondern um einen Restricted Maximum-Likelihood-Schätzer von σ^2 (vgl. Harville (1977), Foulley (1993), Starke & Ostwald (2017)). Aus geometrischer Perspektive handelt es sich bei $\hat{\sigma}^2$ um einen KQ-Schätzer (vgl. Christensen (2011)).

37.3. Unabhängig identisch normalverteilte Zufallsvariablen

Als erste Anwendung von Theorem 37.1 und Theorem 37.2 analysieren wir das Szenario von n unabhängigen und identisch normalverteilten Zufallsvariablen mit Erwartungswertparameter $\mu \in \mathbb{R}$ und Varianzparameter σ^2 ,

$$y_i \sim N(\mu, \sigma^2) \text{ für } i = 1, \dots, n. \quad (37.26)$$

Schreibt man dieses Modell in seiner Designmatrixform (vgl. Gleichung 36.10) dann gilt, wie unten gezeigt,

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n y_i =: \bar{y} \text{ und } \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 =: s_y^2 \quad (37.27)$$

In diesem Fall ist also der Betaparameterschätzer mit dem Stichprobenmittel $\bar{y}$ der $y_1, \dots, y_n$ und der Varianzparameterschätzer mit der Stichprobenvarianz s_y^2 der $y_1, \dots, y_n$ identisch.

Gleichung 37.27 ergibt sich wie folgt. Zum einen gilt

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ &= (1_n^T 1_n)^{-1} 1_n^T v \\ &= (1 \quad \dots \quad 1) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}^{-1} (1 \quad \dots \quad 1) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= n^{-1} \sum_{i=1}^n y_i \\ &= \frac{1}{n} \sum_{i=1}^n y_i \\ &=: \bar{y}. \end{aligned} \quad (37.28)$$

Zum anderen gilt

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-1} (y - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= \frac{1}{n-1} (v - 1_n \bar{y})^T (v - 1_n \bar{y}) \\ &= \frac{1}{n-1} \left(\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \bar{y}^T \right) \left(\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \bar{y} \right) \\ &= \frac{1}{n-1} (y_1 - \bar{y} \quad \dots \quad y_n - \bar{y}) \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \\ &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= s_y^2. \end{aligned} \quad (37.29)$$

Wir demonstrieren die Parameterschätzung in diesem Szenario in folgendem **R** Code.

```
# Modellformulierung
library(MASS)
n      = 12
p      = 1
X      = matrix(rep(1,n), nrow = n)
I_n    = diag(n)
beta   = 2

# Multivariate Normalverteilung
# Anzahl Datenpunkte
# Anzahl Betaparameter
# Designmatrix
# n x n Einheitsmatrix
# wahrer, aber unbekannter, Betaparameter
```



```

sigsqr      = 1                                # wahrer, aber unbekannter, Varianzparameter

# Datenrealisierung
y           = mvrnorm(1, X %>% beta, sigsq*I_n)  # eine Realisierung eines n-dimensionalen ZVs

# Parameterschätzung
beta_hat    = solve(t(X) %>% X) %>% t(X) %>% y   # Betaparameterschätzer
eps_hat     = y - X %>% beta_hat                 # Residuenvektor
sigsqr_hat  = (t(eps_hat) %>% eps_hat) / (n-p)    # Varianzparameterschätzer

beta       : 2
hat{beta}  : 2.212857
sigsqr     : 1
hat{sigsqr}: 1.252161

```

Die Frequentistische Bedeutung der Schätzerunverzerrtheit in diesem Szenario simuliert folgender **R** Code.

```

# Modellformulierung
library(MASS)                                # Multivariate Normalverteilung
n           = 12                             # Anzahl Datenpunkte
p           = 1                             # Anzahl Betaparameter
X           = matrix(rep(1,n), nrow = n)     # Designmatrix
I_n         = diag(n)                       # n x n Einheitsmatrix
beta        = 2                             # wahrer, aber unbekannter, Betaparameter
sigsqr      = 1                             # wahrer, aber unbekannter, Varianzparameter

# Frequentistische Simulation
nsim        = 1e4                           # Anzahl Datenrealisierungen
beta_hat    = rep(NaN,nsim)                 # \hat{\beta} Realisierungsarray
sigsqr_hat  = rep(NaN,nsim)                 # \hat{\sigma}^2 Realisierungsarray
for(i in 1:nsim){
  y         = mvrnorm(1, X %>% beta, sigsq*I_n) # Datenrealisierung
  beta_hat[i] = solve(t(X) %>% X) %>% t(X) %>% y # Betaparameterschätzer
  eps_hat    = y - X %>% beta_hat[i]          # Residuenvektor
  sigsq_hat[i] = (t(eps_hat) %>% eps_hat) / (n-p) # Varianzparameterschätzer
}

Wahrer, aber unbekannter, Betaparameter      : 2
Geschätzter Erwartungswert des Betaparameterschätzers : 1.996142
Wahrer, aber unbekannter, Varianzparameter    : 1
Geschätzter Erwartungswert des Varianzparameterschätzers : 1.002743

```

37.4. Einfache lineare Regression

Als zweite Anwendung von Theorem 37.1 und Theorem 37.2 analysieren wir das Szenario der einfachen linearen Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, \dots, n. \quad (37.30)$$

Basierend auf der Designmatrixform Gleichung 36.15 dieses Modells ergibt sich, wie unten gezeigt,

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} - \frac{c_{xy}}{s_x^2} \bar{x} \\ \frac{c_{xy}}{s_x^2} \end{pmatrix} \text{ und } \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2 \quad (37.31)$$

wobei $\bar{x}$ und $\bar{y}$ die Stichprobenmittel der $x_1, \dots, x_n$ und $y_1, \dots, y_n$, c_{xy} die Stichprobenkovarianz der $x_1, \dots, x_n$ und $y_1, \dots, y_n$ und s_x^2 die Stichprobenvarianz der $x_1, \dots, x_n$ bezeichnen. Wie in Kapitel 1 sind die Bezeichnungen Stichprobenkovarianz und Stichprobenvarianz bezüglich der $x_1, \dots, x_n$ hier lediglich formal gemeint, da keine Annahme zugrundeliegt, dass die $x_1, \dots, x_n$ Realisierungen von Zufallsvariablen sind.

Wir halten also fest, dass für eine parametrische Designmatrixspalte sich der entsprechende Betaparameterschätzer aus der Stichprobenkovarianz der respektiven Spalte mit den Daten geteilt durch die Stichprobenvarianz der entsprechenden Spalte ergibt und somit einer standardisierten Stichprobenkovarianz entspricht. Ein Vergleich mit den Parametern der Ausgleichsgerade in Kapitel 34 zeigt weiterhin die Identität der Betaparameterschätzerkomponenten $\hat{\beta}_0$ und $\hat{\beta}_1$ mit den dort unter dem Kriterium der Minimierung der quadrierten vertikalen Abweichungen hergeleiteten Parametern. Dies überrascht nicht, da sowohl $\hat{\beta}$ als auch die Parameter der Ausgleichsgerade bei festem Wert $y \in \mathbb{R}^n$ von y den Wert

$$q(\tilde{\beta}) = \sum_{i=1}^n \left(y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_i) \right)^2 = (y - X\tilde{\beta})^T (y - X\tilde{\beta}) \quad (37.32)$$

hinsichtlich $\tilde{\beta}$ minimieren.

Um die Form des Betaparameterschätzers in Gleichung 37.31 herzuleiten, halten wir zunächst fest, dass

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) \\ &= \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n \bar{x} y_i + \sum_{i=1}^n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - \bar{y} n \bar{x} - \bar{x} n \bar{y} + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \end{aligned} \quad (37.33)$$

Weiterhin halten wir fest, dass

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^n x_i^2 - \sum_{i=1}^n 2x_i \bar{x} + \sum_{i=1}^n \bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n \bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x} n \bar{x} + n \bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - 2n \bar{x}^2 + n \bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - n \bar{x}^2. \end{aligned} \quad (37.34)$$

Aus der Definition von $\hat{\beta}$ ergibt sich

$$\begin{aligned}
 \hat{\beta} &= (X^T X)^{-1} X^T y \\
 &= \left(\begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
 &= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\
 &= \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix}
 \end{aligned} \tag{37.35}$$

Die Inverse von $X^T X$ ist gegeben durch

$$\frac{1}{s_x^2} \begin{pmatrix} \frac{s_x^2}{n} + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}, \tag{37.36}$$

weil

$$\begin{aligned}
 &\frac{1}{s_x^2} \begin{pmatrix} \frac{s_x^2}{n} + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix} \\
 &= \frac{1}{s_x^2} \begin{pmatrix} \frac{n s_x^2}{n} + n\bar{x}^2 - n\bar{x}^2 & \frac{s_x^2 n \bar{x}}{n} + n\bar{x}^2 \bar{x} - \bar{x} \sum_{i=1}^n x_i^2 \\ -\bar{x} n + n\bar{x} & -n\bar{x}^2 + \sum_{i=1}^n x_i^2 \end{pmatrix} \\
 &= \frac{1}{s_x^2} \begin{pmatrix} s_x^2 & s_x^2 \bar{x} - \bar{x} (\sum_{i=1}^n x_i^2 - n\bar{x}^2) \\ 0 & \sum_{i=1}^n x_i^2 - n\bar{x}^2 \end{pmatrix} \\
 &= \frac{1}{s_x^2} \begin{pmatrix} s_x^2 & s_x^2 \bar{x} - \bar{x} s_x^2 \\ 0 & s_x^2 \end{pmatrix} \\
 &= \frac{1}{s_x^2} \begin{pmatrix} s_x^2 & 0 \\ 0 & s_x^2 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
 \end{aligned} \tag{37.37}$$

Es ergibt sich also

$$\begin{aligned}
 \hat{\beta} &= \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_x^2} & -\frac{\bar{x}}{s_x^2} \\ -\frac{\bar{x}}{s_x^2} & \frac{1}{s_x^2} \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix} = \begin{pmatrix} \left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} \right) n\bar{y} - \frac{\bar{x} \sum_{i=1}^n x_i y_i}{s_x^2} \\ \frac{\sum_{i=1}^n x_i y_i}{s_x^2} - \frac{n\bar{x}\bar{y}}{s_x^2} \end{pmatrix} \\
 &= \begin{pmatrix} \frac{n\bar{y}}{n} + \frac{\bar{x}^2 n\bar{y}}{s_x^2} - \frac{\bar{x} \sum_{i=1}^n x_i y_i}{s_x^2} \\ \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x^2} \end{pmatrix} \\
 &= \begin{pmatrix} \bar{y} + \frac{\bar{x} n\bar{x}\bar{y} - \bar{x} \sum_{i=1}^n x_i y_i}{s_x^2} \\ \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x^2} \end{pmatrix} \\
 &= \begin{pmatrix} \bar{y} - \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x^2} \bar{x} \\ \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x^2} \end{pmatrix} \\
 &= \begin{pmatrix} \bar{y} - \frac{c_{xy}}{s_x^2} \bar{x} \\ \frac{c_{xy}}{s_x^2} \end{pmatrix}.
 \end{aligned} \tag{37.38}$$

Wir demonstrieren die Parameterschätzung in diesem Szenario in folgendem **R** Code. Man beachte die weitgehende Übereinstimmung mit der Implementation der Parameterschätzung im Szenario der unabhängig und identisch normalverteilten Zufallsvariablen - lediglich die Designmatrix und die Dimension des Betaparameters ändern sich.

```
# Modellformulierung
library(MASS)
n      = 10
p      = 2
x      = 1:n
X      = matrix(c(rep(1,n),x), nrow = n)
I_n    = diag(n)
beta   = matrix(c(0,1), nrow = p)
sigsqr = 1

# Datenrealisierung
y      = mvrnorm(1, X %*% beta, sigsqr*I_n)

# Parameterschätzung
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
eps_hat  = y - X %*% beta_hat
sigsqr_hat = (t(eps_hat) %*% eps_hat) / (n-p)

# Multivariate Normalverteilung
# Anzahl Datenpunkte
# Anzahl Betaparameter
# Prädiktorwerte
# Designmatrix
# n x n Einheitsmatrix
# wahrer, aber unbekannter, Betaparameter
# wahrer, aber unbekannter, Varianzparameter

# eine Realisierung eines n-dimensionalen ZVs

# Betaparameterschätzer
# Residuenvektor
# Varianzparameterschätzer

beta      : 0 1
hat{beta} : 1.189061 0.7644628
sigsqr    : 1
hat{sigsqr}: 1.242027
```

Analog zum Szenario der unabhängig und identisch normalverteilten Zufallsvariablen kann auch hier die Frequentistische Bedeutung der Schätzerunverzerrtheit simuliert werden.

```
# Modellformulierung
library(MASS)
n      = 10
p      = 2
x      = 1:n
X      = matrix(c(rep(1,n),x), nrow = n)
I_n    = diag(n)
beta   = matrix(c(0,1), nrow = p)
sigsqr = 1

# Frequentistische Simulation
nsim   = 1e4
beta_hat = matrix(rep(NaN,p*nsim), nrow = p)
sigsqr_hat = rep(NaN,nsim)
for(i in 1:nsim){
  y      = mvrnorm(1, X %*% beta, sigsqr*I_n)
  beta_hat[,i] = solve(t(X) %*% X) %*% t(X) %*% y
  eps_hat  = y - X %*% beta_hat[,i]
  sigsqr_hat[i] = (t(eps_hat) %*% eps_hat) / (n-p)
}

# Multivariate Normalverteilung
# Anzahl Datenpunkte
# Anzahl Betaparameter
# Prädiktorwerte
# Designmatrix
# n x n Einheitsmatrix
# wahrer, aber unbekannter, Betaparameter
# wahrer, aber unbekannter, Varianzparameter

# Anzahl Realisierungen des n-dimensionalen ZVs
# \hat{beta} Realisierungsarray
# \hat{sigsqr} Realisierungsarray
# Simulationsiterationen
# Datenrealisierung
# Betaparameterschätzer
# Residuenvektor
# Varianzparameterschätzer

Wahrer, aber unbekannter, Betaparameter      : 0 1
Geschätzter Erwartungswert des Betaparameterschätzers : -0.004855008 1.001228
Wahrer, aber unbekannter, Varianzparameter    : 1
Geschätzter Erwartungswert des Varianzparameterschätzers : 1.00269
```

37.5. Frequentistische Schätzerverteilungen

Wir dokumentieren die Frequentistische Verteilung des Betaparameterschätzers in folgendem Theorem.

Theorem 37.3 (Frequentistische Verteilung des Betaparameterschätzers.). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (37.39)$$

das ALM. Weiterhin sei

$$\hat{\beta} := (X^T X)^{-1} X^T y \quad (37.40)$$

der Betaparameterschätzer. Dann gilt

$$\hat{\beta} \sim N\left(\beta, \sigma^2 (X^T X)^{-1}\right). \quad (37.41)$$

◦

Beweis. Das Theorem folgt direkt mit dem Theorem zur linearen Transformation von multivariaten Normalverteilungen. Speziell gilt hier:

$$\hat{\beta} \sim N \left((X^T X)^{-1} X^T X \beta, (X^T X)^{-1} X^T (\sigma^2 I_n) \left((X^T X)^{-1} X^T \right)^T \right). \quad (37.42)$$

Der Erwartungswertparameter vereinfacht sich dann zu

$$(X^T X)^{-1} X^T X \beta = \beta. \quad (37.43)$$

Der Kovarianzmatrixparameter vereinfacht sich wie folgt:

$$\begin{aligned} (X^T X)^{-1} X^T (\sigma^2 I_n) \left((X^T X)^{-1} X^T \right)^T &= (X^T X)^{-1} X^T (\sigma^2 I_n) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned} \quad (37.44)$$

Dabei hier die erste Gleichung aus der Tatsache, dass sowohl $X^T X$ als auch ihre Inverse $(X^T X)^{-1}$ symmetrische Matrizen sind. Damit folgt dann aber sofort

$$\hat{\beta} \sim N \left(\beta, \sigma^2 (X^T X)^{-1} \right). \quad (37.45)$$

□

Mit Theorem 6.3 folgt also insbesondere auch für den Erwartungswert und die Kovarianzmatrix des Betaparameterschätzers, dass

$$\mathbb{E}(\hat{\beta}) = \beta \text{ und } \mathbb{C}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \quad (37.46)$$

Als Diagonalelemente von $\mathbb{C}(\hat{\beta})$ hängen die Varianzen der Betaparameterschätzerkomponenten also sowohl vom Varianzparameter der Fehlervariablen als auch von der Designmatrix ab. Insbesondere bei festem, wahren, aber unbekannten $\sigma^2 > 0$ kann also die Designmatrix so gewählt werden, dass die Varianz der Betaparameterschätzerkomponenten minimiert wird.

Die Frequentistische Verteilung des Varianzparameterschätzers dokumentieren wir in folgendem Theorem, welches wir an dieser Stelle nicht beweisen wollen.

Theorem 37.4 (Frequentistische Verteilung des Varianzparameterschätzers). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (37.47)$$

das ALM. Weiterhin sei

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - p} \quad (37.48)$$

der Varianzparameterschätzer. Dann gilt

$$\frac{n - p}{\sigma^2} \hat{\sigma}^2 \sim \chi^2(n - p). \quad (37.49)$$

◦

Da es sich bei $(n - p)\hat{\sigma}^2$ um eine Summe normalverteilter Zufallsvariablen handelt, liegt die χ^2 -Verteilung im Lichte der χ^2 -Transformation bei normalverteilten Zufallsvariablen zumindest nahe. Allerdings ist $\hat{\sigma}^2$ selbst nicht χ^2 verteilt, sondern lediglich seine durch Multiplikation mit $\frac{n-p}{\sigma^2}$ skalierte Version. Wir wollen die Frequentistischen Schätzerverteilungen aus Theorem 37.3 und Theorem 37.4 noch an den beiden Standardbeispielen verdeutlichen.

Beispiel (1) Unabhängige und identisch normalverteilte Zufallsvariablen

Es sei

$$y \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^{n \times 1}, \beta := \mu \in \mathbb{R} \text{ und } \sigma^2 > 0 \quad (37.50)$$

das ALM Szenario unabhängiger und identisch normalverteilter Zufallsvariablen bei bekannter Varianz. Wir haben bereits gesehen, dass in diesem Fall $\hat{\beta}$ mit dem Stichprobenmittel $\bar{y}$ identisch ist. Theorem 37.3 impliziert dann mit

$$(X^T X)^{-1} = (1_n^T 1_n)^{-1} = \frac{1}{n}, \quad (37.51)$$

dass

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \quad (37.52)$$

Das Stichprobenmittel von n unabhängigen und identisch normalverteilten Zufallsvariablen mit Erwartungswertparameter μ und Varianzparameter σ^2 ist also normalverteilt mit Erwartungswertparameter μ und Varianzparameter σ^2/n . Wir haben diese Tatsache bereits im Kontext der Transformationen der Normalverteilungen unter dem Begriff der Mittelwertstransformation gesehen.

Beispiel (2) Einfache lineare Regression

Es sei

$$y \sim N(X\beta, \sigma^2 I_n) \text{ mit } \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n \times 2}, \beta \in \mathbb{R}^2 \text{ und } \sigma^2 > 0 \quad (37.53)$$

das Szenario der einfachen linearen Regression. Wir haben bereits gesehen, dass

$$\sigma^2 (X^T X)^{-1} = \frac{\sigma^2}{s_{xx}} \begin{pmatrix} \frac{s_{xx}}{n} + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \text{ mit } s_{xx} := \sum_{i=1}^n (x_i - \bar{x})^2 \quad (37.54)$$

Die Varianz des Offsetparameterschätzers hängt damit sowohl von der Summe der quadrierten Differenzen der Werte der unabhängigen Variable von ihrem Stichprobenmittel und dem Stichprobenmittel der Werte der unabhängigen Variable selbst ab. Die Varianz des Steigungsparameterschätzers hängt dagegen nur von der Summe der quadrierten Differenzen der unabhängigen Variable von ihrem Stichprobenmittel ab. Die Kovarianz von Offset- und Steigungsparameterschätzern schließlich hängt vom Mittelwert der Werte der unabhängigen Variable ab. Folgender **R** Code simuliert die frequentistischen Verteilungen von Beta- und Varianzparameterschätzern im Szenario der einfachen linearen Regression.

```
# Modellformulierung
library(MASS)
n      = 10
p      = 2
x      = 1:n
X      = matrix(c(rep(1,n),x), nrow = n)
I_n    = diag(n)
beta   = matrix(c(0,1), nrow = p)
sigsqr = .5

# Multivariate Normalverteilung
# Anzahl von Datenpunkten
# Anzahl von Betparametern
# Prädiktorwerte
# Designmatrix
# n x n Einheitsmatrix
# wahrer, aber unbekannter, Betaparameter
# wahrer, aber unbekannter, Varianzparameter
```

```

# Frequentistische Simulation
nsim = 10
y = matrix(rep(NaN,n*nsim), nrow = n)
beta_hat = matrix(rep(NaN,p*nsim), nrow = p)
for(i in 1:nsim){
  y[,i] = mvrnorm(1, X %*% beta, sigsqr*I_n)
  beta_hat[,i] = solve(t(X) %*% X) %*% t(X) %*% y[,i]
}

# Anzahl Realisierungen n-dimensionaler ZV
# y Realisierungsarray
# \hat{\beta} Realisierungsarray

# eine Realisierung n-dimensionaler ZV
# \hat{\beta} = (X^T X)^{-1} X^T y

```

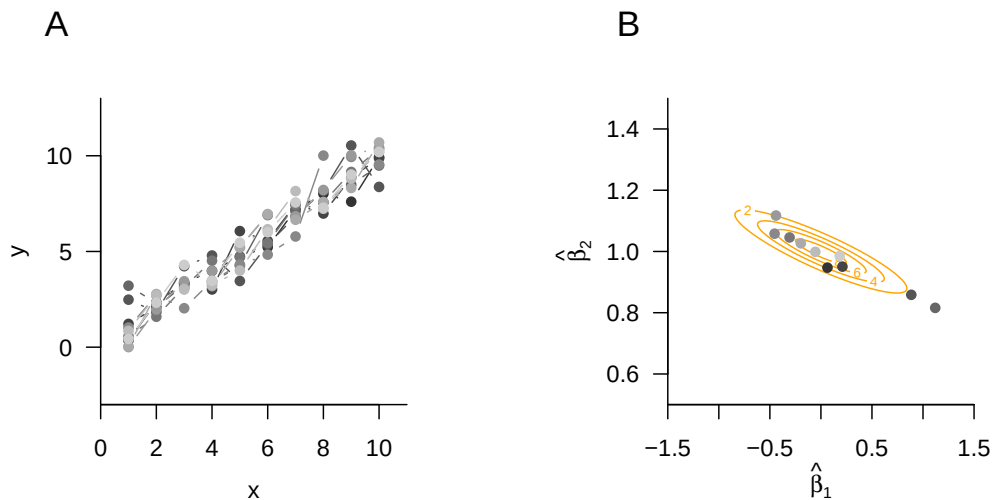


Abbildung 37.1. Frequentistische Betaparameterschätzerverteilung bei einfacher linearer Regression.

Abbildung 37.1 zeigt 10 Realisationen des Modells einer einfachen linearen Regression und Abbildung 37.1 B zeigt die entsprechenden Betaparameterschätzerrealisationen sowie die analytische Verteilung des Betaparameterschätzers.

37.6. Literaturhinweise

Plackett (1949) gibt einen historischen Überblick zur Entwicklung der Betaparameterschätzung und insbesondere des Gauss-Markov Theorems. Das Problem der Varianzparameterschätzung ALM im Sinne der Restricted Maximum Likelihood Methode erscheint zunächst in Patterson & Thompson (1971) (vgl. Harville (1977)), Verbyla (1990)) und bleibt, in verallgemeinerten ALMs, Gegenstand aktueller Forschung (vgl. Lindholm & Wahl (2020)).

38. T-Statistiken

In diesem Abschnitt führen wir T-Statistiken als Maße zur Evaluation von Betaparameterschätzern im ALM ein. T-Statistiken quantifizieren dabei die geschätzten Effekte des Betaparameterschätzers in Bezug zur durch den Varianzparameterschätzer geschätzten Residualvariabilität. Der Wert einer T-Statistik ist also zunächst einmal einfach als Signal-zu-Rauschen Verhältnis (Signal-to-Noise Ratio) zu verstehen.

T-Statistiken erlauben weiterhin die Evaluation von Linearkombinationen der Komponenten des Betaparameterschätzers im Sinne Frequentistischer Konfidenzintervalle und Hypothesentests. Wir betrachten hier zunächst nur die funktionale Form von T-Statistiken und ihre Frequentistische Verteilung zum Zwecke der Konfidenzintervallbestimmung. Der Einsatz von T-Teststatistiken im Rahmen von Einstich- und Zweistichproben T-Tests ist das Thema von Kapitel 40.

38.1. Definition und Beispiele

Vor dem Hintergrund des ALMs und seiner Parameterpunktschätzer definieren wir die T-Statistik wie folgt.

Definition 38.1 (T-Statistik). Es sei

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (38.1)$$

das ALM. Weiterhin seien

$$\hat{\beta} := (X^T X)^{-1} X^T y \text{ und } \hat{\sigma}^2 := \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} \quad (38.2)$$

die Betaparameter- und Varianzparameterschätzer, respektive. Dann ist für einen Kontrastgewichtsvektor $c \in \mathbb{R}^p$ und einen Parameter $\beta_0 \in \mathbb{R}^p$ die T-Statistik definiert als

$$T := \frac{c^T \hat{\beta} - c^T \beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}}. \quad (38.3)$$

•

Geeignete Wahlen des Kontrastgewichtvektors $c \in \mathbb{R}^p$ und des Parameters $\beta_0 \in \mathbb{R}^p$ erlauben eine Vielzahl von Einsatzmöglichkeiten der T-Statistik. Betrachten wir zunächst den Kontrastgewichtvektor. Offenbar dient der Kontrastgewichtvektor dazu, den Zufallsvektor $\hat{\beta} \in \mathbb{R}^p$ in die Zufallsvariable $c^T \hat{\beta}$ zu transformieren und sichert damit die Skalarität der T-Statistik. Weiterhin erlaubt die Wahl von p -dimensionalen Einheitsvektoren für den Kontrastgewichtvektor die Auswahl einzelner Komponenten des Betaparameters zur

Evaluation mithilfe der T-Statistik. Schließlich erlaubt eine generelle Wahl des Kontrastgewichtsvektors die Evaluation beliebiger Linearkombination der Betaparameterkomponenten, wie zum Beispiel Differenzen einzelner Komponenten. Beispielfhaft seien für $\hat{\beta} \in \mathbb{R}^2$ hier folgende Möglichkeiten für die Wahl von $c \in \mathbb{R}^2$ hinsichtlich des Skalarproduktes $c^T \hat{\beta}$ aufgeführt:

$$c := \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Rightarrow c^T \hat{\beta} = \hat{\beta}_1, \quad c := \begin{pmatrix} 0 \\ 1 \end{pmatrix} \Rightarrow c^T \hat{\beta} = \hat{\beta}_2, \quad c := \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow c^T \hat{\beta} = \hat{\beta}_1 - \hat{\beta}_2. \quad (38.4)$$

Die Wahl des Parameters $\beta_0 \in \mathbb{R}^p$ eröffnet die Möglichkeit, die T-Statistik unterschiedlich einzusetzen. Wählt man zum Beispiel $\beta_0 := 0_p$, so erhält man mit der T-Statistik eine Deskriptivstatistik, die es erlaubt, geschätzte Regressoreffekte, also Komponenten oder Linearkombinationen von $\hat{\beta}$, im Sinne eines Signal-zu-Rauschen Verhältnisses in Bezug zu der durch $\hat{\sigma}^2$ quantifizierten Residualdatenvariabilität zu setzen. Der Nenner der T-Statistik stellt dabei sicher, dass insbesondere die adäquate (Ko)Standardabweichung der entsprechenden Betaparameterkomponentenkombination als Bezugsgröße dient, da es sich bei $\hat{\sigma}^2 (X^T X)^{-1}$ bekanntlich um die Kovarianz des Betaparameterschätzers handelt (vgl. Theorem 37.3).

Wählt man für β_0 dagegen β , also den wahren, aber unbekannten, Betaparameterwert, so eröffnet die T-Statistik die Möglichkeit, für einzelne Komponenten des Betaparametervektors Konfidenzintervalle zu bestimmen. Wir vertiefen diesen Aspekt der T-Statistik in Kapitel 38.2. Deklariert man schließlich β_0 im Kontext eines Testszenarios als das Element einer Nullhypothese, so eröffnet die T-Statistik die hypothesentestbasierte Inferenz über Betaparameterkomponenten und ihrer Linearkombinationen. Anwendungsfälle dieser Art diskutieren wir ausführlich in Kapitel 40.

Die Anwendung der T-Statistik zum Zwecke der Frequentistischen Inferenz im Sinne von Konfidenzintervallen und Hypothesentests basiert dabei natürlich auf der Frequentistischen Verteilung der T-Statistik vor dem Hintergrund des ALMs. Diese ist der zentrale Inhalt folgenden Theorems, auf dessen Beweis wir verzichten.

Theorem 38.1 (Frequentistische Verteilung der T-Statistik). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (38.5)$$

das ALM. Weiterhin seien

$$\hat{\beta} := (X^T X)^{-1} X^T y \text{ und } \hat{\sigma}^2 := \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} \quad (38.6)$$

die Betaparameter- und Varianzparameterschätzer, respektive. Schließlich sei für einen Kontrastgewichtsvektor $c \in \mathbb{R}^p$ und einen Parameter $\beta_0 \in \mathbb{R}^p$

$$T := \frac{c^T \hat{\beta} - c^T \beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} \quad (38.7)$$

die T-Statistik. Dann gilt

$$T \sim t(\delta, n - p) \text{ mit } \delta := \frac{c^T \beta - c^T \beta_0}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}}. \quad (38.8)$$

◦

Im Allgemeinen ist die T-Statistik also nichtzentral t-verteilt. Gilt dabei, wie bei der Bestimmung von Konfidenzintervallen (vgl. Kapitel 38.2) $\beta_0 := \beta$ oder gilt in einem Test-szenario bei Zutreffen der Nullhypothese $\beta := \beta_0$ (vgl. Kapitel 40), so ist die T-Statistik sogar t -verteilt, jeweils mit Freiheitsgradparameter $n - p$. Gilt in einem Testszenario dagegen, dass die Nullhypothese nicht zutrifft, so kann die Verteilung der T-Statistik aus Theorem 38.1 zur Herleitung der Testgütefunktion und damit zur Bestimmung der Power des Tests genutzt werden (vgl. Kapitel Kapitel 40). Wir werden diese Aspekte an gegebener Stelle vertiefen. An dieser Stelle wollen wir die T-Statistik und ihre Verteilung zunächst nur an den Beispielen der unabhängigen identisch normalverteilten Zufallsvariablen und der einfachen linearen Regression illustrieren.

Beispiel (1) Unabhängige und identisch normalverteilte Zufallsvariablen

Es sei

$$y \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^{n \times 1}, \beta := \mu \in \mathbb{R} \text{ und } \sigma^2 > 0 \quad (38.9)$$

das ALM Szenario unabhängiger und identisch normalverteilter Zufallsvariablen und es seien $c := 1$ und $\beta_0 := \mu_0 \in \mathbb{R}$. Dann gilt für die T-Statistik

$$T = \frac{c^T \hat{\beta} - c^T \mu_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} = \frac{1^T \bar{v} - 1^T \mu_0}{\sqrt{s_v^2 1^T (1_n^T 1_n)^{-1} 1}} = \sqrt{n} \left(\frac{\bar{v} - \mu_0}{s_v} \right) \quad (38.10)$$

Dies entspricht offenbar der bekannten Einstichproben-T-Teststatistik (vgl. Kapitel 33). Wie diese nimmt die hier betrachtete T-Statistik, bei Konstanz der jeweils komplementären Terme, große absolute Werte für eine große absolute Differenz von $\bar{v} - \mu_0$ (oft als Effekt bezeichnet), sowie für kleine Werte von s_v^2 (also eine geringe Datenvariabilität) und einen großen Wert von n (also einen großen Stichprobenumfang) an. Folgender **R** Code simuliert die Frequentistische Verteilung dieser T-Statistik für die Fälle $\beta = \beta_0$ und $\beta \neq \beta_0$.

```
# Libraries
library(MASS)                                # multivariate Normalverteilung

# Modellformulierung
n      = 12                                  # Anzahl von Datenpunkten
p      = 1                                  # Anzahl von Betaparametern
X      = matrix(c(rep(1,n)), nrow = n)      # Designmatrix
I_n    = diag(n)                            # Einheitsmatrix
beta   = c(0,1)                             # wahre , aber unbekannte , Betaparameter
nscn   = length(beta)                      # Anzahl wahrer, aber unbekannter, Hypothesenszenarien
sigsqr = 1                                  # wahrer, aber unbekannter, Varianzparameter
c      = 1                                  # Kontrastvektor von Interesse
beta_0 = 0                                  # Nullhypothesenbetaparameter

# Frequentistische Simulation
nsim   = 1e4                                # Anzahl Simulationen
delta  = rep(NA, nscn)                     # Anzahl Nichtzentralitätsparameter
Tee    = matrix(rep(NA, nscn*nsim), ncol = nscn) # T-Teststatistik Realisierungsarray
for(s in 1:nscn){                          # Hypothesenszenarien
  delta[s] = ((t(c) %*% beta[s] - t(c) %*% beta_0)/ # Nichtzentralitätsparameter
             sqrt(sigsqr*t(c)%*%solve(t(X)%*%X)%*%c))
  for(i in 1:nsim){
    y      = mvrnorm(1, X %*% beta[s], sigsqr*I_n) # Simulationsiterationen
    beta_hat = solve(t(X) %*% X) %*% t(X) %*% y    # y
    eps_hat  = y - X %*% beta_hat                  # \hat{\beta}
    sigsqr_hat = (t(eps_hat) %*% eps_hat)/(n-p)     # \hat{\epsilon}
    Tee[i,s] = ((t(c) %*% beta_hat - t(c) %*% beta_0)/ # \hat{\sigma}^2
               sqrt(sigsqr_hat*t(c)%*%solve(t(X)%*%X)%*%c)) # T
  }
}
```

Abbildung 38.1 A und B zeigen die resultierenden simulierten und analytischen Verteilungen der T-Statistik.

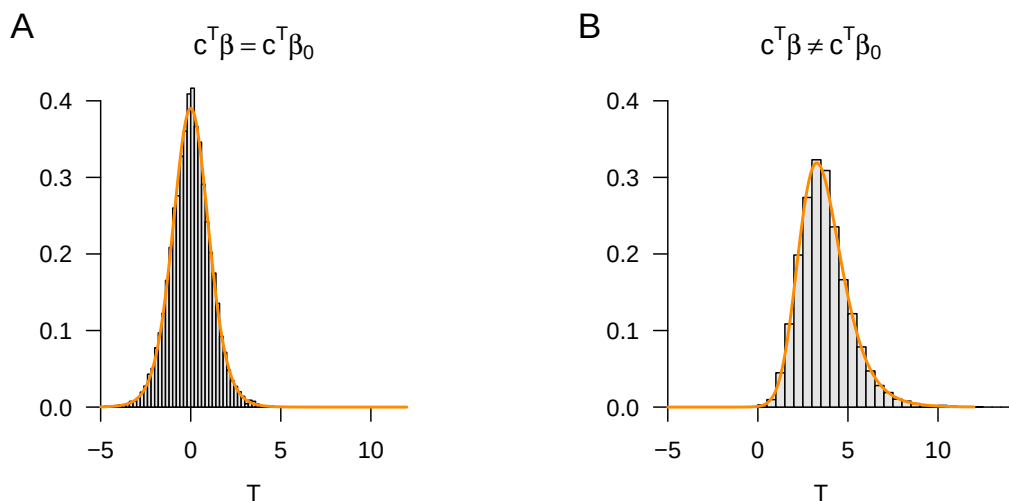


Abbildung 38.1. Verteilungen der T-Statistik bei unabhängig und identisch normalverteilten Zufallsvariablen.

Beispiel (2) Einfache lineare Regression

In diesem Beispiel wollen wir nicht auf die spezifische Form der T-Statistik eingehen, aber anhand einer Simulation demonstrieren, wie sich das Prinzip der T-Statistik im Kontext der einfachen linearen Regression darstellt. Dazu betrachten wir das bekannte Beispielmmodell der einfachen linearen Regression (vgl. Kapitel 36), in diesem Fall mit den wahren, aber unbekannten, Parameterwerten $\beta_A := (1, 0)$ und $\beta_B := (1, 1)$. Weiterhin betrachten wir den Kontrastgewichtsvektor $c := (0, 1)$, so dass die T-Statistik zur Evaluation des Steigungsparameters der einfachen linearen Regression genutzt werden kann. Schließlich betrachten wir in beiden Fällen den Parameter $\beta_0 := (0, 0)^T$, so dass im Fall von β_A gilt, dass $c^T \beta = c^T \beta_0$ und im Fall von β_B gilt, dass $c^T \beta \neq c^T \beta_0$. Folgender **R** Code implementiert die skizzierten Szenarien, Abbildung 38.2 A und B zeigen die resultierenden simulierten und analytischen Verteilungen der T-Statistik.

```
# Modellformulierung
library(MASS)
n = 10
p = 2
x = 1:n
X = matrix(c(rep(1,n),x), ncol = p)
I_n = diag(n)
beta = matrix(c(1,0,1,1), nrow = 2)
nscn = ncol(beta)
sigsqr = 1
c = matrix(c(0,1), nrow = 2)
beta_0 = matrix(c(0,0), nrow = 2)

# multivariate Normalverteilung
# Anzahl von Datenpunkten
# Anzahl von Betaparametern
# Prädiktorwerte
# Designmatrix
# Einheitsmatrix
# wahre, aber unbekannte, Betaparameter
# Anzahl wahrer, aber unbekannter, Hypothesenszenarien
# wahrer, aber unbekannter, Varianzparameter
# Kontrastvektor von Interesse
# Nullhypothesenbetaparameter

# Frequentistische Simulation
nsim = 1e4
delta = rep(NA, nscn)
Tee = matrix(rep(NA, nscn*nsim), ncol = nscn)
for(s in 1:nscn){
  delta[s] = ((t(c) %*% beta[,s] - t(c) %*% beta_0) /
    sqrt(sigsqr*t(c) %*% solve(t(X) %*% X) %*% c))
  for(i in 1:nsim){
    y = mvrnorm(1, X %*% beta[,s], sigsqr*I_n)
    beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
    eps_hat = y - X %*% beta_hat
    sigsqr_hat = (t(eps_hat) %*% eps_hat) / (n-p)
    Tee[i,s] = ((t(c) %*% beta_hat - t(c) %*% beta_0) /
      sqrt(sigsqr_hat*t(c) %*% solve(t(X) %*% X) %*% c))
  }
}

# Anzahl Simulationen
# Anzahl Nichtzentralitätsparameter
# T-Teststatistik Realisierungsarray
# Hypothesenszenarien
# Nichtzentralitätsparameter

# Simulationsiterationen
# y
# \hat{\beta}
# \hat{\epsilon}
# \hat{\sigma}^2
# T
```

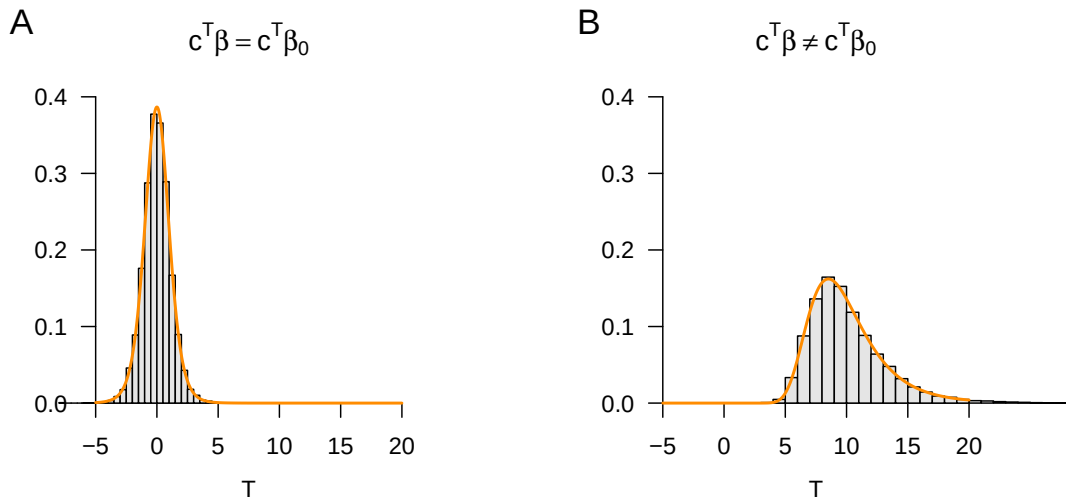


Abbildung 38.2. Verteilungen der T-Statistik bei einfacher linearer Regression.

38.2. Konfidenzintervalle für Betaparameterkomponenten

Mithilfe der T-Statistik können Konfidenzintervalle für die Komponenten des Betaparametervektors bestimmt werden. Das folgende Theorem ist die zentrale Aussage dieses Abschnitts.

Theorem 38.2 (Konfidenzintervalle für Betaparameterkomponenten). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (38.11)$$

das ALM,

$$\hat{\beta} := (X^T X)^{-1} X^T y \text{ und } \hat{\sigma}^2 := \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} \quad (38.12)$$

seien die Betaparameter- und Varianzparameterschätzer, respektive und für ein $\delta \in]0, 1[$ sei

$$t_\delta := \Psi^{-1}\left(\frac{1 + \delta}{2}; n - p\right). \quad (38.13)$$

Schließlich sei für $j = 1, \dots, p$

$$\lambda_j := \left((X^T X)^{-1}\right)_{jj} \text{ das } j\text{te Diagonalelement von } (X^T X)^{-1} \quad (38.14)$$

Dann ist für $j = 1, \dots, p$

$$\kappa_j := \left[\hat{\beta}_j - \hat{\sigma} \sqrt{\lambda_j} t_\delta, \hat{\beta}_j + \hat{\sigma} \sqrt{\lambda_j} t_\delta\right] \quad (38.15)$$

ein δ -Konfidenzintervall für die j te Komponente β_j des Betaparameters $\beta = (\beta_1, \dots, \beta_p)^T$.

◦

Beweis. Wir müssen zeigen, dass

$$\mathbb{P}(\kappa_j \ni \beta_j) = \delta. \quad (38.16)$$

Dazu halten wir zunächst fest, dass für alle $j = 1, \dots, p$ bei Wahl von $\beta_0 = \beta$ und $c := e_j$ nach Theorem 38.1 für $T \sim t(\delta, n - p)$ gilt, dass

$$T = \frac{e_j^T \hat{\beta} - e_j^T \beta}{\sqrt{\hat{\sigma}^2 e_j^T (X^T X)^{-1} e_j}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 ((X^T X)^{-1})_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{\lambda_j}} =: T_j \quad (38.17)$$

und

$$\delta = \frac{e_j^T \beta - e_j^T \beta}{\sqrt{\hat{\sigma}^2 e_j^T (X^T X)^{-1} e_j}} = 0 \quad (38.18)$$

Damit gilt dann auch sofort, dass $T_j \sim t(n - p)$. Weiterhin erinnern wir daran (vgl. Kapitel 32), dass per Definition von t_δ gilt, dass

$$\mathbb{P}(-t_\delta \leq T_j \leq t_\delta). \quad (38.19)$$

Aus der Definition eines δ -Konfidenzintervalls folgt dann

$$\begin{aligned} \delta &= \mathbb{P}(-t_\delta \leq T_j \leq t_\delta) \\ &= \mathbb{P}\left(-t_\delta \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{\lambda_j}} \leq t_\delta\right) \\ &= \mathbb{P}\left(-t_\delta \hat{\sigma} \sqrt{\lambda_j} \leq \hat{\beta}_j - \beta_j \leq t_\delta \hat{\sigma} \sqrt{\lambda_j}\right) \\ &= \mathbb{P}\left(-\hat{\beta}_j - t_\delta \hat{\sigma} \sqrt{\lambda_j} \leq -\beta_j \leq -\hat{\beta}_j + t_\delta \hat{\sigma} \sqrt{\lambda_j}\right) \\ &= \mathbb{P}\left(\hat{\beta}_j + t_\delta \hat{\sigma} \sqrt{\lambda_j} \geq \beta_j \geq \hat{\beta}_j - t_\delta \hat{\sigma} \sqrt{\lambda_j}\right) \\ &= \mathbb{P}\left(\hat{\beta}_j - t_\delta \hat{\sigma} \sqrt{\lambda_j} \leq \beta_j \leq \hat{\beta}_j + t_\delta \hat{\sigma} \sqrt{\lambda_j}\right) \\ &= \mathbb{P}\left([\hat{\beta}_j - \hat{\sigma} \sqrt{\lambda_j} t_\delta, \hat{\beta}_j + \hat{\sigma} \sqrt{\lambda_j} t_\delta]\right) \\ &= \mathbb{P}(\kappa_j \ni \beta_j) \end{aligned} \quad (38.20)$$

und damit ist alles gezeigt. $\square$

Beispiel (1) Unabhängige und identisch normalverteilte Zufallsvariablen

Wie gewohnt betrachten wir als erstes Beispiel die ALM Form des Szenarios unabhängiger und identisch normalverteilter Zufallsvariablen

$$y \sim N(X\beta, \sigma^2 I_n) \text{ mit } X := 1_n \in \mathbb{R}^n, \beta := \mu \in \mathbb{R}, \sigma^2 > 0 \quad (38.21)$$

Dann gelten, wie bereits gesehen

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n v_i =: \bar{v}, \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (v_i - \bar{v})^2 =: s_v^2 \text{ und } \lambda_1 = (1_n^T 1_n)^{-1} = \frac{1}{n}. \quad (38.22)$$

Nach Theorem 38.2 gilt dann, dass

$$\kappa := \left[\bar{v} - \frac{s}{\sqrt{n}} t_\delta, \bar{v} + \frac{s}{\sqrt{n}} t_\delta \right] \quad (38.23)$$

ein δ -Konfidenzintervall für β ist und dieses ist offenbar identisch mit dem bekannten δ -Konfidenzintervall für den Erwartungsparameter der Normalverteilung.

Beispiel (2) Einfache lineare Regression

In diesem Beispiel wollen wir nicht auf die spezifische Form der Konfidenzintervalle für den Offset- und Steigungsparameter eingehen, sondern lediglich anhand folgender Simulation an die Frequentistische Bedeutung eines δ -Konfidenzintervalls erinnern: Realisierungen von δ -Konfidenzintervallen überdecken den wahren, aber unbekannten, Parameterwert mit einer frequentistischen Wahrscheinlichkeit von δ . Abbildung 38.3 zeigt, basierend auf folgendem **R** Code, dass dies in der konkreten Simulation mit $\delta = 0.95$ für den Offsetparameter in 94 von 100 und für den Steigungsparameter in 93 von 100 Fällen der Fall ist.

```
# Modellformulierung
library(MASS)
set.seed(0)
ns      = 1e2
n       = 10
p       = 2
x       = 1:n
X       = matrix(c(rep(1,n),x), ncol = p)
I_n     = diag(n)
beta    = matrix(c(1,2), nrow = 2)
sigsqr  = 1
delta   = 0.95
t_delta = qt((1+delta)/2,n-1)
lambda  = diag(solve(t(X) %*% X))

# multivariate Normalverteilung
# Random number generator seed
# Anzahl Simulationen
# Anzahl von Datenpunkten
# Anzahl von Betaparametern
# Prädiktorwerte
# Designmatrix
# Einheitsmatrix
# wahre, aber unbekannte, Betaparameter
# wahrer, aber unbekannter, Varianzparameter
# Konfidenzbedingung
# \Psi^{-1}((1+\delta)/2,n-1)
# \lambda_j Werte

# Simulation
kappa   = array(rep(NaN, ns*p*p), dim=c(ns,2,2))
beta_hat = matrix(rep(NaN,p*ns), nrow = p)
for(i in 1:ns){
  y      = mvrnorm(1, X %*% beta, sigsqr*I_n)
  beta_hat[,i] = solve(t(X) %*% X) %*% t(X) %*% y
  eps_hat = y - X %*% beta_hat[,i]
  sigsqr_hat = (t(eps_hat) %*% eps_hat)/(n-p)
  for(j in 1:p){
    kappa[i,1,j] = beta_hat[j,i]-sqrt(sigsqr_hat*lambda[j])*t_delta
    kappa[i,2,j] = beta_hat[j,i]+sqrt(sigsqr_hat*lambda[j])*t_delta
  }
}
```

38.3. Literaturhinweise

Box (1981) und Zabell (2008) geben einen historischen Überblick zur Entwicklung der T-Statistik und ihrer Verteilung im Kontext der Arbeiten von Student (1908) und Fisher (1925c), Fisher (1925b) und Fisher (1925a). Die Theorie der Konfidenzintervalle geht auf Neyman (1935) und Neyman (1937) zurück.

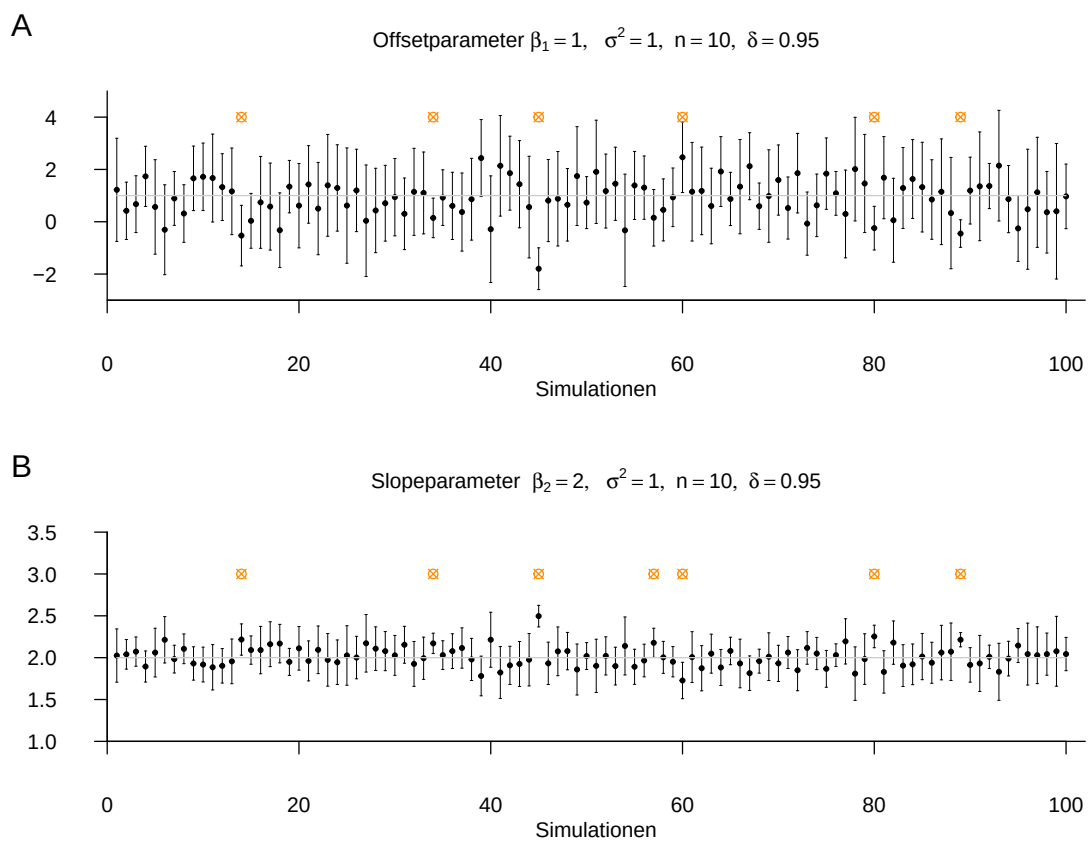


Abbildung 38.3. Konfidenzintervalle für Betaparameterkomponenten bei einfacher linearer Regression.

39. F-Statistiken

In diesem Abschnitt führen wir F-Statistiken vor dem Hintergrund Likelihood-Quotienten-Statistik-basierter Modellvergleiche ein. Die (maximierte oder marginale) Likelihood eines Datensatzes unter einem gegebenen probabilistischen Modell als Modellvergleichskriterium heranzuziehen ist dabei ein weit verbreitetes Verfahren in der probabilistischen Datenanalyse. Im Gegensatz zu T-Statistiken kann das Ziel der Berechnung von F-Statistiken damit insbesondere sein, nicht nur Linearkombinationen von Betaparameterschätzwerten probabilistisch zu evaluieren, sondern die Modellanpassung an einen Datensatz insgesamt zu evaluieren. Die Modellvergleichskapazität von F-Statistiken ist allerdings etwas beschränkt, da sich die F-Statistik nur auf ALMs und insbesondere geschachtelte ALMs bezieht, in denen ein Modell Bestandteil eines anderen Modells ist. F-Statistiken bilden üblicherweise die Grundlage für Hypothesentests im Rahmen varianzanalytischer Verfahren. Der Einsatz von F-Statistiken ist aber per se nicht auf Varianzanalysen beschränkt, sondern kann auch bei parametrischen ALM Designs angebracht sein. Im Folgenden führen wir zunächst den Begriff der Likelihood-Quotienten-Statistik ein und betrachten dann die Definition der F-Statistik vor diesem Hintergrund. Wir schließen mit der Frequentistischen Verteilung der F-Statistik.

39.1. Likelihood-Quotienten-Statistiken

Wir definieren den Begriff der Likelihood-Quotienten-Statistik wie folgt.

Definition 39.1 (Likelihood-Quotienten-Statistik). Gegeben seien zwei Frequentistische Inferenzmodelle

$$\mathcal{M}_0 := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_{\theta_0}^0 \mid \theta_0 \in \Theta_0\}) \text{ und } \mathcal{M}_1 := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_{\theta_1}^1 \mid \theta_1 \in \Theta_1\}) \quad (39.1)$$

mit identischem Datenraum, identischer σ -Algebra und potentiell distinkten Mengen von Wahrscheinlichkeitsmaßen und Parameterräumen. Sei weiterhin v ein Zufallsvektor mit Datenraum $\mathcal{Y}$. Seien schließlich L_0^y und L_1^y die Likelihood-Funktionen von $\mathcal{M}_0$ und $\mathcal{M}_1$, respektive, wobei das Superskript y jeweils an die Datenabhängigkeit der Likelihood Funktion erinnern soll. Dann wird

$$\Lambda := \frac{\max_{\theta_0 \in \Theta_0} L_0^y(\theta_0)}{\max_{\theta_1 \in \Theta_1} L_1^y(\theta_1)} \quad (39.2)$$

Likelihood-Quotienten-Statistik genannt.

•

Eine Likelihood-Quotienten-Statistik setzt die Wahrscheinlichkeitsmasse/dichte eines beobachteten Datensatzes $y \in \mathcal{Y}$ unter zwei Frequentistischen Inferenzmodellen nach Optimierung der jeweiligen Modellparameter ins Verhältnis. Ein hoher Wert der Likelihood-Quotienten-Statistik entspricht einer höheren Wahrscheinlichkeitsmasse/dichte des beobachteten Datensatzes $y \in \mathcal{Y}$ unter $\mathcal{M}_0$ als unter $\mathcal{M}_1$ und vice versa.

Die Wahrscheinlichkeitsmassen/dichten beobachteter Daten nach Modellschätzung unter verschiedenen Modellen zu betrachten ist ein allgemeines Vorgehen zum Vergleich von Modellen. Letztlich erlaubt dieses Vorgehen, verschiedene wissenschaftliche Theorien über die Genese beobachtbarer Daten quantitativ zu vergleichen und die damit verbundene Unsicherheit zu quantifizieren. Modellvergleiche sind ein zentrales Thema in der Bayesianischen Inferenz die die Logik von Likelihood-Quotienten-Statistiken zum Beispiel unter den Begriffen der Bayes Factors oder der des Bayesian Information Criteria auf allgemeine probabilistische Modelle generalisiert. Allerdings sind, wie hier gesehen, Modellvergleiche auch im Rahmen der Frequentistischen Inferenz möglich und sinnvoll, Modellvergleiche sind also kein Alleinstellungsmerkmal der Bayesianischen gegenüber der Frequentistischen Inferenz.

Mit dem *reduzierten Modell* und dem *vollständigen Modell* betrachten wir im Folgenden zwei spezielle Formen von $\mathcal{M}_0$ und $\mathcal{M}_1$, respektive, im Kontext des ALMs.

Definition 39.2 (Vollständiges und reduziertes Modell.). Für $p > 1$ mit $p = p_0 + p_1$ seien

$$X := \begin{pmatrix} X_0 & X_1 \end{pmatrix} \in \mathbb{R}^{n \times p} \text{ mit } X_0 \in \mathbb{R}^{n \times p_0} \text{ und } X_1 \in \mathbb{R}^{n \times p_1} \quad (39.3)$$

sowie

$$\beta := \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \in \mathbb{R}^p \text{ mit } \beta_0 \in \mathbb{R}^{p_0} \text{ und } \beta_1 \in \mathbb{R}^{p_1} \quad (39.4)$$

Partitionierungen einer $n \times p$ Designmatrix und eines p -dimensionalen Betaparametervektors. Dann nennen wir

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (39.5)$$

das *vollständige Modell* und

$$y = X_0\beta_0 + \varepsilon_0 \text{ mit } \varepsilon_0 \sim N(0_n, \sigma^2 I_n) \quad (39.6)$$

das *reduzierte Modell* und sprechen von einer *Partitionierung eines (vollständigen) Modells*.

•

Man sagt auch, dass das reduzierte Modell im vollständigen Modell *geschachtelt* (engl. *nested*) ist. Die Likelihood-Quotienten-Statistik beim Vergleich eines vollständigen und eines reduzierten Modells hat eine einfache Form. Diese ist der zentrale Aspekt folgenden Theorems.

Theorem 39.1 (Likelihood-Quotienten-Statistik von vollständigem und reduziertem Modell.). Für $p = p_0 + p_1, p > 1$ sei eine Partitionierung eines vollständigen ALMs gegeben und es seien $\hat{\sigma}^2$ und $\hat{\sigma}_0^2$ die Maximum-Likelihood-Schätzer des Varianzparameters unter vollständigem und reduziertem Modell, respektive. Weiterhin seien die zwei parametrischen

statistischen Modelle $\mathcal{M}_0$ und $\mathcal{M}_1$ in der Definition der Likelihood-Quotienten-Statistik durch das reduzierte Modell und das vollständige Modell gegeben. Dann gilt

$$\Lambda = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right)^{\frac{n}{2}} \quad (39.7)$$

◦

Beweis. Wir erinnern zunächst daran, dass die Maximum-Likelihood-Schätzer des Varianzparameters durch

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) \text{ und } \hat{\sigma}_0^2 = \frac{1}{n} (y - X_0\hat{\beta}_0)^T (y - X_0\hat{\beta}_0) \quad (39.8)$$

respektive, gegeben sind, wobei $\hat{\beta}$ und $\hat{\beta}_0$ die Maximum-Likelihood-Schätzer der Betaparameter unter vollständigem und reduziertem Modell, respektive, bezeichnen. Weiterhin halten wir fest, dass für die Likelihood-Funktion des vollständigen Modells an der Stelle der Maximum-Likelihood-Schätzer gilt, dass

$$\begin{aligned} L_1^y(\hat{\beta}, \hat{\sigma}^2) &= (2\pi)^{-\frac{n}{2}} (\hat{\sigma}^2)^{-\frac{n}{2}} \exp \left(-\frac{1}{2\hat{\sigma}^2} (y - X\hat{\beta})^T (y - X\hat{\beta}) \right) \\ &= (2\pi)^{-\frac{n}{2}} (\hat{\sigma}^2)^{-\frac{n}{2}} \exp \left(-\frac{n}{2} \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{(y - X\hat{\beta})^T (y - X\hat{\beta})} \right) \\ &= (2\pi)^{-\frac{n}{2}} (\hat{\sigma}^2)^{-\frac{n}{2}} e^{-\frac{n}{2}} \end{aligned} \quad (39.9)$$

und analog, dass für die Likelihood-Funktion des reduzierten Modells an der Stelle der Maximum-Likelihood-Schätzer gilt, dass

$$L_0^y(\hat{\beta}_0, \hat{\sigma}_0^2) = (2\pi)^{-\frac{n}{2}} (\hat{\sigma}_0^2)^{-\frac{n}{2}} e^{-\frac{n}{2}} \quad (39.10)$$

Damit ergibt sich dann aber

$$\Lambda = \frac{\max_{\theta_0 \in \Theta_0} L_0^y(\theta_0)}{\max_{\theta_1 \in \Theta_1} L_1^y(\theta_1)} = \frac{L_0^y(\hat{\beta}_0, \hat{\sigma}_0^2)}{L_1^y(\hat{\beta}, \hat{\sigma}^2)} = \frac{(2\pi)^{-\frac{n}{2}} (\hat{\sigma}_0^2)^{-\frac{n}{2}} e^{-\frac{n}{2}}}{(2\pi)^{-\frac{n}{2}} (\hat{\sigma}^2)^{-\frac{n}{2}} e^{-\frac{n}{2}}} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{-\frac{n}{2}} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right)^{\frac{n}{2}}. \quad (39.11)$$

□

39.2. Definition und Verteilung

Wir definieren nun die F-Statistik vor dem Hintergrund eines vollständigen und eines reduzierten Modells.

Definition 39.3 (F-Statistik). Für $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ und $\sigma^2 > 0$ sei ein ALM der Form

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (39.12)$$

mit der Partitionierung

$$X = \begin{pmatrix} X_0 & X_1 \end{pmatrix}, X_0 \in \mathbb{R}^{n \times p_0}, X_1 \in \mathbb{R}^{n \times p_1}, \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \beta_0 \in \mathbb{R}^{p_0}, \beta_1 \in \mathbb{R}^{p_1} \quad (39.13)$$

mit $p = p_0 + p_1$ gegeben. Weiterhin seien mit

$$\hat{\beta}_0 := (X_0^T X_0)^{-1} X_0^T y \text{ und } \hat{\beta} := (X^T X)^{-1} X^T y \quad (39.14)$$

die Residuenvektoren

$$\hat{\varepsilon}_0 := y - X_0 \hat{\beta}_0 \text{ und } \hat{\varepsilon} := y - X \hat{\beta} \quad (39.15)$$

definiert. Dann ist die *F-Statistik* definiert als

$$F := \frac{(\hat{\varepsilon}_0^T \hat{\varepsilon}_0 - \hat{\varepsilon}^T \hat{\varepsilon}) / p_1}{\hat{\varepsilon}^T \hat{\varepsilon} / (n - p)} \quad (39.16)$$

•

Der Zähler der F-Statistik

$$\frac{\hat{\varepsilon}_0^T \hat{\varepsilon}_0 - \hat{\varepsilon}^T \hat{\varepsilon}}{p_1} \quad (39.17)$$

misst, inwieweit die p_1 Regressoren in X_1 die Residualquadratsumme reduzieren und zwar im Verhältnis zur Anzahl dieser Regressoren. Das heißt, dass bei gleicher Größe der Residualquadratsummenreduktion (und gleichem Nenner) ein größerer F Wert resultiert, wenn diese durch weniger zusätzliche Regressoren resultiert, also p_1 klein ist (und vice versa). Im Sinne der Anzahl der Spalten von X und der entsprechenden Komponenten von β favorisiert die F -Statistik also weniger "komplexe" Modelle.

Für den Nenner der F-Statistik gilt

$$\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - p} = \hat{\sigma}^2 \quad (39.18)$$

wobei $\hat{\sigma}^2$ hier der aufgrund des vollständigen Modells geschätzte Schätzer von σ^2 ist. Werden die Daten tatsächlich unter dem reduzierten Modell generiert, so kann das vollständige Modell dies durch $\hat{\beta}_2 \approx 0_{p_1}$ abbilden und erreicht eine ähnliche σ^2 Schätzung wie das reduzierte Modell. Werden die Daten de-facto unter dem vollständigen Modell generiert, so ist $\hat{\varepsilon}^T \hat{\varepsilon} / (n - p)$ ein besserer Schätzer von σ^2 als $\hat{\varepsilon}_0^T \hat{\varepsilon}_0 / (n - p)$, da sich für diesen die Datenvariabilität, die nicht durch die p_0 Regressoren in X_0 erklärt wird, in der Schätzung von σ^2 widerspiegeln würde. Der Nenner der F-Statistik ist also in beiden Fällen der sinnvollere Schätzer von σ^2 .

Zusammengenommen misst die F-Statistik also die Residualquadratsummenreduktion durch die p_1 Regressoren in X_1 gegenüber den p_0 Regressoren in X_0 pro Datenvariabilitäts (σ^2) - und Regressor (p_1)-Einheit.

Beispiel (1) Einfache lineare Regression

Exemplarisch evaluiert untenstehender **R** Code die F-Statistik im Kontext folgender Partitionierung des Modells der einfachen linearen Regression

$$X = \begin{pmatrix} X_0 & X_1 \end{pmatrix}, X_0 := 1_n, X_1 := (x_1, \dots, x_n)^T, \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad (39.19)$$

und zwar einmal für den Fall, dass $\beta = (1, 0)^T$, also dass das reduzierte Modell das wahre, aber unbekannte, datenerzeugende Modell ist und einmal für den Fall, dass $\beta = (1, 0.5)^T$, also dass das vollständige Modell das wahre, aber unbekannte, datenerzeugende Modell ist. Im Sinne obiger Diskussion ergibt sich im ersten Fall eine F-Statistik nahe Null, im zweiten Fall dagegen eine hohe F-Statistik.

```

# Modellformulierung
library(MASS)
nmod = 2
n = 10
p = 2
p_0 = 1
p_1 = 1
p = p_0 + p_1
x = 1:n
X = matrix(c(rep(1,n),x), nrow = n)
X_0 = X[,1]
I_n = diag(n)
beta = matrix(c(1,0,1,.5), nrow = 2)
nscn = ncol(beta)
sigsqr = 1

# Multivariate Normalverteilung
# Anzahl Modelle
# Anzahl Datenpunkte
# Anzahl Betaparameter
# Anzahl Betaparameter reduziertes Modell
# Anzahl zusätzlicher Betaparameter vollständiges Modell
# Anzahl Betaparameter im vollständigen Modell
# Prädiktorwerte
# Designmatrix des vollständigen Modells
# Designmatrix des reduzierten Modells
# n x n Einheitsmatrix
# wahre , aber unbekannte , Betaparameter
# Anzahl wahrer, aber unbekannter, Hypothesenszenarien
# wahrer, aber unbekannter, Varianzparameter

# Modellsimulation und Evaluierung
Eff = matrix(rep(NA,nscn), nrow = nscn)
for(s in 1:nscn){
  y = mvrnorm(1, X %*%beta[,s], sigsqr*I_n)
  beta_hat_0 = solve(t(X_0)%*%X_0)%*%t(X_0)%*%y
  beta_hat = solve(t(X) %*%X )%*%t(X) %*%y
  eps_0_hat = y-X_0%*%beta_hat_0
  eps_hat = y-X%*%beta_hat
  eps_0_eps_0_hat = t(eps_0_hat) %*% eps_0_hat
  eps_eps_hat = t(eps_hat) %*% eps_hat
  Eff[s] = (((eps_0_eps_0_hat-eps_eps_hat)/p_1)/
            (eps_eps_hat/(n-p))))
}

# F-Statistik Realisierungsarray
# Szenarieniterationen
# Datenrealisierung
# Betaparameterschätzer reduziertes Modell
# Betaparameterschätzer vollständiges Modell
# Residuenvektor reduziertes Modell
# Residuenvektor vollständiges Modell
# RQS reduziertes Modell
# RQS vollständiges Modell
# F-Statistik

```

F-Statistik für beta_1 = 0_{p_1}: 0.4362354
 F-Statistik für beta_1 != 0_{p_1}: 18.91575

Die Likelihood-Quotienten-Statistik und die F-Statistik von vollständigem und reduziertem Modell sind ineinander überführbar. Dies ist die zentrale Aussage folgenden Theorems.

Theorem 39.2 (F-Statistik und Likelihood-Quotienten-Statistik.). *Es sei die Partitionierung eines ALMs in ein vollständiges und ein reduziertes Modell gegeben und F und Λ seien die entsprechenden F- und Likelihood-Quotienten-Statistiken. Dann gilt*

$$F = \frac{n-p}{p_1} \left(\Lambda^{-\frac{2}{n}} - 1 \right). \quad (39.20)$$

◦

Beweis. Wir erinnern zunächst daran, dass die Maximum-Likelihood Schätzer des Varianzparameters durch

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta}) = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n} \quad \text{und} \quad \hat{\sigma}_0^2 = \frac{1}{n} (y - X_0\hat{\beta}_0)^T (y - X_0\hat{\beta}_0) = \frac{\hat{\varepsilon}_0^T \hat{\varepsilon}_0}{n} \quad (39.21)$$

gegeben sind. Mit der Definition der F-Statistik und der Form der Likelihood-Quotienten-Statistik für den Vergleich von reduziertem und vollständigem Modell ergibt sich dann

$$\begin{aligned}
 F &= \frac{(\hat{\varepsilon}_0^T \hat{\varepsilon}_0 - \hat{\varepsilon}^T \hat{\varepsilon}) / p_1}{\hat{\varepsilon}^T \hat{\varepsilon} / (n-p)} \\
 &= \frac{n(\hat{\sigma}_0^2 - \hat{\sigma}^2) / p_1}{n\hat{\sigma}^2 / (n-p)} \\
 &= \frac{n-p}{p_1} \frac{\hat{\sigma}_0^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \\
 &= \frac{n-p}{p_1} \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} - 1 \right) \\
 &= \frac{n-p}{p_1} \left(\Lambda^{-\frac{2}{n}} - 1 \right).
 \end{aligned} \quad (39.22)$$

□

Zwischen der F-Statistik und der Likelihood-Quotienten-Statistik besteht also ein nichtlinearer, reziproker Zusammenhang, den wir für $n = 12, p = 2$ und $p_1 = 1$ in Abbildung 39.1 visualisieren. Man beachte, dass für $\Lambda = 1$ gilt, dass $F = 0$. Ein Wert von $F = 0$ impliziert also, dass das reduzierte Modell gegenüber dem vollständigen Modell im Lichte eines beobachteten Datensatzes die gleiche Plausibilität besitzt.

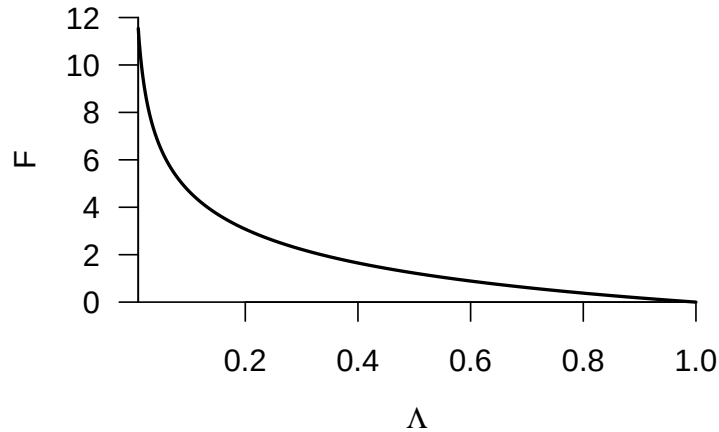


Abbildung 39.1. Zusammenhang von F- und Likelihood-Quotienten-Statistik.

Wir dokumentieren die Frequentistische Verteilung der F-Statistik in folgendem Theorem, auf dessen Beweis wir verzichten.

Theorem 39.3 (F-Statistik). Für $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$ und $\sigma^2 > 0$ sei ein ALM der Form

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (39.23)$$

mit der Partitionierung

$$X = \begin{pmatrix} X_0 & X_1 \end{pmatrix}, X_0 \in \mathbb{R}^{n \times p_0}, X_1 \in \mathbb{R}^{n \times p_1}, \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \beta_0 \in \mathbb{R}^{p_0}, \beta_1 \in \mathbb{R}^{p_1} \quad (39.24)$$

mit $p = p_0 + p_1$ gegeben. Schließlich sei

$$c := \begin{pmatrix} 0_{p_0} \\ 1_{p_1} \end{pmatrix} \in \mathbb{R}^p \quad (39.25)$$

ein Vektor. Dann gilt

$$F \sim f(\delta, p_1, n - p) \text{ mit } \delta := \frac{c^T \beta \left(c^T (X^T X)^{-1} c \right)^{-1} c^T \beta}{\sigma^2} \quad (39.26)$$

◦

Man beachte, dass die F -Statistik eine Funktion des Parameterschätzers, δ dagegen eine Funktion der wahren, aber unbekannten, Parameter ist. Wie die Verteilung der T -Statistik kann die Verteilung der F -Statistik für die Evaluation von Frequentistischen Konfidenzintervallen und Hypothesentests genutzt werden. Insbesondere letzteren Aspekt verdeutlichen wir in Kapitel 47 und Kapitel 42.

Beispiel (1) Einfache lineare Regression

Exemplarisch evaluieren wir mithilfe untenstehenden **R** Codes die Verteilung der F -Statistik im Kontext der Partitionierung des Modells der einfachen linearen Regression und zwar erneut einmal für den Fall, dass $\beta = (1, 0)^T$, also dass das reduzierte Modell das wahre, aber unbekannte, datenerzeugende Modell ist und einmal für den Fall, dass $\beta = (1, 0.5)^T$, also dass das vollständige Modell das wahre, aber unbekannte, datenerzeugende Modell ist. Abbildung 8.2 A und B visualisieren die resultierenden Verteilungen, respektive.

```
# Modellformulierung
library(MASS)
nmod = 2
n = 10
p_0 = 1
p_1 = 1
p = p_0 + p_1
x = 1:n
X = matrix(c(rep(1,n),x), nrow = n)
X_0 = X[,1]
I_n = diag(n)
beta = matrix(c(1,0,1,.5), nrow = 2)
nscn = ncol(beta)
sigsqr = 1
c = matrix(c(0,1), nrow = 2)

# Multivariate Normalverteilung
# Anzahl Modelle
# Anzahl Datenpunkte
# Anzahl Betaparameter im reduzierten Modell
# Anzahl additiver Betaparameter im vollständigen Modell
# Anzahl Betaparameter im vollständigen Modell
# Prädiktorwerte
# Designmatrix des vollständigen Modells
# Designmatrix des reduzierten Modells
# n x n Einheitsmatrix
# wahre, aber unbekannte, Betaparameter
# Anzahl wahrer, aber unbekannter, Hypothesenszenarien
# wahrer, aber unbekannter, Varianzparameter
# Vektor

# Frequentistische Simulation
nsim = 1e4
delta = rep(NA,nscn)
Eff = matrix(rep(NA, nscn*nsim), nrow = nscn)
for(s in 1:nscn){
  delta[s] = (t(t(c)%*%beta[,s])%*%
    solve(t(c)%*%solve(t(X)%*%X)%*%c) %*%
    (t(c)%*%beta[,s])/sigsqr)
  for(i in 1:nsim){
    y = mvrnorm(1, X %*%beta[,s], sigsqr*I_n)
    beta_hat_0 = solve(t(X_0)%*%X_0)%*%t(X_0)%*%y
    beta_hat = solve(t(X) %*%X) %*%t(X) %*%y
    eps_0_hat = y-X_0%*%beta_hat_0
    eps_hat = y-X%*%beta_hat
    eps_0_eps_0_hat = t(eps_0_hat) %*% eps_0_hat
    eps_eps_hat = t(eps_hat) %*% eps_hat
    Eff[s,i] = (((eps_0_eps_0_hat-eps_eps_hat)/p_1)/
      (eps_eps_hat/(n-p))))}

# Anzahl Realisierungen des n-dimensionalen ZVs
# Nichtzentralitätsparameterarray
# F-Statistik Realisierungsarray
# Szenarieniterationen
# Nichtzentralitätsparameter

# Simulationsiterationen
# Datenrealisierung
# Betaparameterschätzer reduziertes Modell
# Betaparameterschätzer vollständiges Modell
# Residuenvektor reduziertes Modell
# Residuenvektor vollständiges Modell
# RQS reduziertes Modell
# RQS vollständiges Modell
# F-Statistik
```

39.3. Literaturhinweise

Die Popularität von F -Statistiken, insbesondere im Kontext der Varianzanalyse, wird allgemein auf Fisher (1925a) zurückgeführt. Seal (1967) gibt einen historischen Überblick. Likelihood-Quotienten-Statistiken werden insbesondere von Neyman & Pearson (1928) und Wilks (1938) betrachtet. Lehmann (2011) gibt einen integrierten historischen Überblick zu beiden Ansätzen. Die hier diskutierte Äquivalenz von Likelihood-Quotienten-Statistik und F -Statistik basiert auf der Darstellung in Seber & Lee (2003).

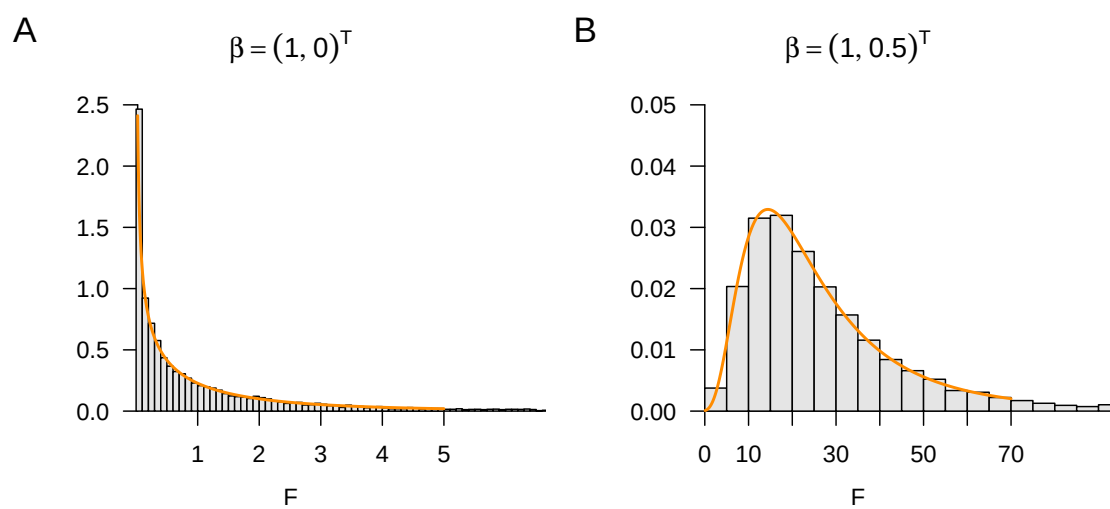


Abbildung 39.2. Exemplarische F-Statistik Verteilungen unter reduziertem und vollständigem wahren, aber unbekannten, datenerzeugenden Modell bei einfacher linearer Regression.

40. T-Tests

40.1. Einstichproben-T-Tests

Anwendungsszenario

Das Anwendungsszenario eines Einstichproben-T-Test ist bekanntlich dadurch gekennzeichnet, dass n univariate Datenpunkte einer Stichprobe (Gruppe) randomisierter experimenteller Einheiten betrachtet werden, von denen angenommen wird, dass sie Realisierungen von n unabhängigen und identisch normalverteilten Zufallsvariablen sind. Hinsichtlich der identischen univariaten Normalverteilungen $N(\mu, \sigma^2)$ dieser Zufallsvariablen wird angenommen, dass sowohl der Erwartungswertparameter μ als auch der Varianzparameter σ^2 unbekannt sind. Schließlich wird vorausgesetzt, dass ein Interesse an einem inferentiellen Vergleich des unbekannten Erwartungswertparameters μ mit einem vorgegebenen Wert μ_0 (z.B. $\mu_0 := 0$) besteht.

Anwendungsbeispiel

Für ein konkretes Anwendungsbeispiel betrachten wir die Analyse von Pre-PostInterventions-BDI-Differenzwerten einer Gruppe von $n = 12$ Patient:innen wie in Tabelle 40.1 dargestellt. Die ersten beiden Spalten dieser Tabelle listen die patientenspezifische BDI Werte vor (PreBDI) und nach (PosBDI) der psychotherapeutischen Intervention, die dritte Spalte dBDI zeigt die entsprechenden PreBDI-PosBDI Differenzwerte. Ein positiver Wert entspricht hier einer Verbesserung der Depressionssymptomatik und ein negativer Wert einer Verschlechterung der Depressionssymptomatik

Tabelle 40.1. Pre- und Post-Intervention BDI Werte

PreBDI	PosBDI	dBDI
29	28	1
32	26	6
33	29	4
28	26	2
31	30	1
32	27	5
33	27	6
30	27	3
29	27	2
30	28	2
33	29	4
31	25	6

Bei der Anwendung eines Einstichproben-T-Tests auf die dBDI Daten dieses Datensatzes nehmen wir also an, dass die dBDI Daten Realisierungen von $n = 12$ unabhängig normalverteilten Zufallsvariablen $y_i \sim N(\mu, \sigma^2)$ sind. Wir nehmen weiterhin an, dass wir daran interessiert sind, unsere Unsicherheit beim inferentiellen Vergleich des wahren, aber unbekannten, Erwartungswertparameters μ mit einem Vergleichswert μ_0 im Sinne eines Hypothesentests zu quantifizieren.

Unabhängig von diesem inferenzstatistischen Vorgehen betrachten wir zunächst die deskriptiven Statistiken der dBDI Daten, wie in Tabelle 40.2 dargestellt. Es fällt insbesondere auf, dass das Stichprobenmittel im Vergleich zur Standardabweichung relativ klein ist. Im Gruppenmittel unterscheiden sich die PreBDI und PosBDI also zwar in positiver Richtung, was eine Verringerung der Depressionssymptomatik impliziert, allerdings streuen die Daten auch über Patient:innen deutlich, wie auch bereits aus Tabelle 40.1 ersichtlich.

Tabelle 40.2. Deskriptivstatistiken der Pre-Post BDI Differenzwerte

	n	Max	Min	Median	Mean	Var	Std
F2F	12	6	1	3.5	3.5	3.73	1.93

Modellformulierung

Wir definieren nun das Einstichproben-T-Test-Modell wie folgt.

Definition 40.1 (Einstichproben-T-Test-Modell). Für $i = 1, \dots, n$ seien y_i Zufallsvariablen, die die n Datenpunkte eines Einstichproben-T-Test-Szenarios modellieren. Dann hat das Einstichproben-T-Test-Modell die strukturelle Form

$$y_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v für } i = 1, \dots, n \text{ mit } \mu \in \mathbb{R} \text{ und } \sigma^2 > 0, \quad (40.1)$$

die Datenverteilungsform

$$y_i \sim N(\mu, \sigma^2) \text{ u.i.v für } i = 1, \dots, n \text{ mit } \mu \in \mathbb{R} \text{ und } \sigma^2 > 0, \quad (40.2)$$

und für den Datenvektor $y = (y_1, \dots, y_n)^T$ die Designmatrixform

$$y = X\beta + \varepsilon \text{ mit } X := \mathbf{1}_n \in \mathbb{R}^{n \times 1}, \beta := \mu \in \mathbb{R}, \varepsilon \sim N(0_n, \sigma^2 I_n) \text{ und } \sigma^2 > 0. \quad (40.3)$$

•

Das Modell des Einstichproben-T-Tests ist offenbar mit dem dem Modell unabhängiger und identisch normalverteilter Zufallsvariablen identisch (vgl. Kapitel 36). Die Äquivalenz von struktureller, Datenverteilungs- und Designmatrixform des Einstichproben-T-TestModells wurde in Kapitel 36 bereits ausführlich diskutiert. Die Simulation von Daten basierend auf dem Einstichproben-T-Test-Modell hat dementsprechend die gleiche Form wie die Simulation unabhängig und identisch normalverteilter Zufallsvariablen. Unterer **R** Code demonstriert dies.

```

# Modellformulierung
library(MASS)
n = 12
p = 1
X = matrix(rep(1,n), nrow = n)
I_n = diag(n)
beta = 5
sigsqr = 14

# Multivariate Normalverteilung
# Anzahl von Datenpunkten
# Anzahl von Betaparameter
# Designmatrix
# n x n Einheitsmatrix
# wahrer, aber unbekannter, Betaparameter
# wahrer, aber unbekannter, Varianzparameter

# Datenrealisierung
y = mvrnorm(1, X %*% beta, sigsqr*I_n)
# eine Realisierung eines n-dimensionalen ZVs

```

Modellschätzung

Da die Form des Einstichproben-T-Test-Modells mit dem Szenario unabhängig und identisch normalverteilter Zufallsvariablen identisch ist, trifft dies auch auf die entsprechenden Beta- und Varianzparameterschätzer zu. Es ergibt sich also folgendes Theorem, das bereits in Kapitel 37 bewiesen wurde.

Theorem 40.1 (Parameterschätzer im Einstichproben-T-Test-Modell). *Gegeben sei die Designmatrixform des Einstichproben-T-Test-Modells. Dann ergeben sich für den Betaparameterschätzer*

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n y_i =: \bar{y} \quad (40.4)$$

und für den Varianzparameterschätzer

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 =: s_y^2 \quad (40.5)$$

$\bar{y}$ und s_y^2 bezeichnen hier also wiederum das Stichprobenmittel und die Stichprobenvarianz der Zufallsvariablen $y_1, \dots, y_n$.

◦

40.2. Modellevaluation

Basierend auf Definition 38.1 formulieren wir nun die T-Teststatistik für das Einstichproben-T-Test Szenario und geben ihre frequentistische Verteilung an.

Theorem 40.2 (T-Teststatistik des Einstichproben-T-Tests). *Gegeben sei die Designmatrixform des Einstichproben-T-Test-Modells. Dann ergibt sich für die T-Teststatistik mit*

$$c := 1 \text{ und } c^T \beta_0 =: \mu_0, \quad (40.6)$$

dass

$$T = \sqrt{n} \left(\frac{\bar{y} - \mu_0}{s_y} \right) \quad (40.7)$$

und es gilt, dass

$$T \sim t(\delta, n-1) \text{ mit } \delta = \sqrt{n} \left(\frac{\mu - \mu_0}{\sigma} \right). \quad (40.8)$$

◦

Beweis. Mit Theorem 38.1 gilt

$$T = \frac{c^T \hat{\beta} - c^T \beta_0}{\sqrt{\hat{\sigma}^2 c^T (X^T X)^{-1} c}} = \frac{1^T \bar{y} - 1^T \mu_0}{\sqrt{s_y^2 1^T (1_n^T 1_n)^{-1} 1}} = \sqrt{n} \left(\frac{\bar{y} - \mu_0}{s_y} \right). \quad (40.9)$$

Weiterhin gilt mit demselben Theorem

$$\delta = \frac{c^T \beta - c^T \beta_0}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}} = \frac{1^T \mu - 1^T \mu_0}{\sqrt{\sigma^2 1^T (1_n^T 1_n)^{-1} 1}} = \sqrt{n} \left(\frac{\mu - \mu_0}{\sigma} \right). \quad (40.10)$$

□

Die Formen der T-Teststatistik und ihre Verteilung im Einstichproben-T-Test-Spezialfall des ALMs sind also natürlicherweise mit den entsprechenden Formen im ALM-freien Kontext identisch. Die entsprechende Theorie zu Konfidenzintervallen und der Kontrolle des Testumfangs bei Einstichproben-T-Tests sowie der Gebrauch der Testgütefunktion zur Evaluation der Testtrennschärfe (Power) folgt also analog.

Anwendungsbeispiel

Folgender **R** Code demonstriert die Evaluation eines 95%-Konfidenzintervalls für den Erwartungswertparameter μ sowie Durchführung eines zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese $\Theta_0 := \{0\}$ und Signifikanzlevel $\alpha_0 := 0.05$ für das oben skizzierte Anwendungsbeispiel.

```
# Datenanalyse
D = read.csv("../data/507-t-tests.csv")
y = D$dBDI[D$COND == "F2F"]
n = length(y)
p = 1
c = 1
mu_0 = 0
delta = 0.95
alpha_0 = 0.05
X = matrix(rep(1,n), nrow = n)
beta_hat = solve(t(X)%*%X)%*%t(X)%*%y
eps_hat = y - X %*% beta_hat
sigsqr_hat = (t(eps_hat) %*% eps_hat)/(n-p)
t_delta = qt((1+delta)/2, n-1)
lambda = diag(solve(t(X) %*% X))
kappa_u = beta_hat - sqrt(sigsqr_hat*lambda)*t_delta
kappa_o = beta_hat + sqrt(sigsqr_hat*lambda)*t_delta
t_num = t(c) %*% beta_hat - mu_0
t_den = sqrt(sigsqr_hat %*% t(c) %*% solve(t(X) %*% X) %*% c)
t = t_num/t_den
pval = 2*(1 - pt(abs(t), n-1))
k_alpha_0 = qt(1-alpha_0/2, n-1)
if(abs(t) > k_alpha_0){phi = 1} else {phi = 0}

# Laden des Datensatzes
# Post-Pre Differenzwerte
# Anzahl Datenpunkte
# Anzahl Betaparameter
# Kontrastgewichtsvektor
# Nullhypothesenparameter
# Konfidenzlevel
# Signifikanzlevel
# Einstichproben-T-Test Designmatrix
# Betaparameterschätzer
# Residuenvektor
# Varianzparameterschätzer
# \Psi^{-1}(1+\delta)/2, n-1)
# \lambda_j Werte
# untere Konfidenzintervallgrenze
# obere Konfidenzintervallgrenze
# Zähler der Einstichproben-T-Teststatistik
# Nenner der Einstichproben-T-Teststatistik
# Wert der Einstichproben-T-Teststatistik
# p-Wert bei zweiseitigem Einstichproben-T-Test
# kritischer Wert
# Einstichproben-T-Test
```

```
Betaparameterschätzer      : 3.5
95%-Konfidenzintervall    : 2.27 4.73
Varianzparameterschätzer  : 3.73
alpha_0                    : 0.05
Kritischer Wert           : 2.2
Einstichproben-T-Teststatistik : 6.28
phi                         : 1
p-Wert                     : 0
```

Die Nullhypothese wird in diesem Fall bei einem einem kritischen Wert von $k_{0.05} = 2.20$ und einem Wert der T-Statistik von $T = 3.91$ abgelehnt. Das 95%-Konfidenzintervall für den wahren, aber unbekannten, Erwartungswertparameter ist $[2.26, 8.07]$, überdeckt also den Nullhypothesenparameterwert $\mu_0 = 0$ nicht.

40.3. Zweistichproben-T-Tests

Anwendungsszenario

Das Anwendungsszenario eines Zweistichproben-T-Tests für unabhängige Stichproben ist bekanntlich dadurch gekennzeichnet, dass insgesamt n univariate Datenpunkte zweier Stichproben (Gruppen) randomisierter experimenteller Einheiten betrachtet werden. Es wird dabei insbesondere angenommen, dass die n_1 univariaten Datenpunkte der ersten Gruppe Realisierungen von n_1 unabhängigen und identisch normalverteilten Zufallsvariablen mit Erwartungswertparameter μ_1 und Varianzparameter σ^2 sind, während weiterhin angenommen wird, dass die n_2 univariaten Datenpunkte der zweiten Gruppe Realisierungen von n_2 unabhängigen und identisch normalverteilten Zufallsvariablen mit Erwartungswertparameter μ_2 und Varianzparameter σ^2 sind. Es wird also insbesondere angenommen, dass sich die wahren, aber unbekannten, Erwartungswertparameter beider Gruppen von Zufallsvariablen unterscheiden können, die Varianzparameter beider Gruppen dagegen werden als identisch angenommen. Schließlich wird vorausgesetzt, dass ein Interesse am inferentiellen Vergleich der unbekannten Erwartungswertparameter μ_1 und μ_2 besteht, so zum Beispiel ihrer Gleichheit $\mu_1 = \mu_2$ oder Verschiedenheit $\mu_1 \neq \mu_2$.

Anwendungsbeispiel

Für ein konkretes Anwendungsbeispiel betrachten wird die Analyse von Pre-Post-Interventions-BDI-Differenzwerten zweier Gruppen von je 12 Patient:innen in unterschiedlichen Therapiesettings, wie in Tabelle 40.3 dargestellt. Die erste Spalte der Tabelle (COND) listet das patientenspezifische Therapiesetting (F2F: face-to-face, ONL: online) auf. Die zweite Spalte der Tabelle (dBDI) listet die entsprechenden patientenspezifischen Pre-Post-Interventions-BDI-Differenzwerte auf. Positive Werte entsprechen hier erneut einer Abnahme der Depressionssymptomatik, negative Werte einer Zunahme der Depressionssymptomatik.

Tabelle 40.3. Pre-Post-BDI Differenzwerte für zwei Stichproben

COND	dBDI
F2F	1
F2F	6
F2F	4
F2F	2
F2F	1
F2F	5
F2F	6
F2F	3
F2F	2
F2F	2
F2F	4
F2F	6
ONL	9
ONL	2

Tabelle 40.3. Pre-Post-BDI Differenzwerte für zwei Stichproben

COND	dBDI
ONL	10
ONL	-1
ONL	4
ONL	11
ONL	10
ONL	15
ONL	4
ONL	5
ONL	10
ONL	10

Zu Anwendung eines Zweistichproben-T-Tests auf die dBDI Daten nehmen wir an, dass die 12 Datenpunkte der F2F Therapiegruppe Realisierungen von $n_1 = 12$ unabhängig und identisch normalverteilten Zufallsvariablen $y_{1j} \sim N(\mu_1, \sigma^2)$ mit $j = 1, \dots, n_1$ sind und dass die 12 Datenpunkte der ONL Therapiegruppe Realisierungen von $n_2 = 12$ unabhängig und identisch normalverteilten Zufallsvariablen $y_{2j} \sim N(\mu_2, \sigma^2)$ mit $j = 1, \dots, n_2$ sind.

Unabhängig von dem unten beschriebenen inferenzstatistischen Vorgehen betrachten wir auch hier zunächst die deskriptiven Statistiken der Therapiesetting-spezifischen dBDI Werte. Diese sind in Tabelle 40.4 aufgeführt.

Tabelle 40.4. Deskriptivstatistiken der Pre-Post BDI Differenzwerte bei unterschiedlichen Therapiesettings

	n	Max	Min	Median	Mean	Var	Std
F2F	12	6	1	3.5	3.50	3.73	1.93
ONL	12	15	-1	9.5	7.42	20.81	4.56

Modellformulierung

Mit dem Index i für die Gruppen und dem Index j für die experimentellen Einheiten in jeder Gruppe definieren wir das Zweistichproben-T-Test-Modell wie folgt.

Definition 40.2 (Zweistichproben-T-Test-Modell). Für $i = 1, 2$ und $j = 1, \dots, n_i$ seien y_{ij} Zufallsvariablen, die die $n = n_1 + n_2$ Datenpunkte eines Zweistichproben-T-Test Szenarios modellieren. Dann hat das Zweistichproben-T-Test-Modell die strukturelle Form

$$y_{ij} = \mu_i + \varepsilon_{ij} \text{ mit } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ u.i.v. mit } \mu_i \in \mathbb{R} \text{ und } \sigma^2 > 0, \quad (40.11)$$

die Datenverteilungsform

$$y_{ij} \sim N(\mu_i, \sigma^2) \text{ u.i.v. mit } \mu_i \in \mathbb{R} \text{ und } \sigma^2 > 0, \quad (40.12)$$

und für den n -dimensionalen Datenvektor definiert als

$$y := (y_{11}, \dots, y_{1n_1}, y_{21}, \dots, y_{2n_2})^T \quad (40.13)$$

die Designmatrixform

$$y = X\beta + \varepsilon \quad (40.14)$$

mit

$$X := \begin{pmatrix} 1_{n_1} & 0_{n_1} \\ 0_{n_2} & 1_{n_2} \end{pmatrix} \in \mathbb{R}^{n \times 2}, \beta := \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^2, \varepsilon \sim N(0_n, \sigma^2 I_n), \sigma^2 > 0. \quad (40.15)$$

•

Die hier gewählte Definition des Zweistichproben-T-Test-Modells in Designmatrixform ist nicht die einzig mögliche, jedoch diejenige, unter der sich am klarsten die Äquivalenz zum Zweistichproben-T-Test-Modell im ALM-freien Kontext erkennen lässt. In Kapitel 47 lernen wir eine alternative Parameterisierung auch des Zweistichproben-T-Test-Modells kennen. Wie schon beim Szenario des Einstichproben-T-Tests ergibt sich die Äquivalenz der in Definition 40.2 formulierten Modellformen mit den Ergebnissen in Kapitel 36. Die Simulation von Daten basierend auf dem Zweistichproben-T-Test-Modell ist, bis auf die Definition von Designmatrix und Betaparametervektor mit den bisher bekannten Simulationen von ALM Spezialfällen identisch, wie folgender **R** Code demonstriert.

```
# Modellformulierung
library(MASS)
n_1 = 12
n_2 = 12
n = n_1 + n_2
p = 2
X = matrix(c(rep(1,n_1), rep(0,n_1),
              rep(0,n_2), rep(1,n_2)),
           nrow = n)
I_n = diag(n)
beta = matrix(c(1,2), nrow = p)
sigsqr = 10

# Multivariate Normalverteilung
# Anzahl von Datenpunkten Gruppe 1
# Anzahl von Datenpunkten Gruppe 2
# Gesamtanzahl Datenpunkte
# Anzahl von Betaparameter
# Designmatrix
# n x n Einheitsmatrix
# wahrer, aber unbekannter, Betaparameter
# wahrer, aber unbekannter, Varianzparameter

# Datenrealisierung
y = mvrnorm(1, X %*% beta, sigsq*I_n) # eine Realisierung eines n-dimensionalen ZVs
```

Modellschätzung

Die beiden Betaparameterkomponenten des Zweistichproben-T-Test-Modells in Designmatrixform werden wenig überraschend durch die entsprechenden Gruppenstichprobenmittel geschätzt. Für den Varianzparameterschätzer ergibt sich die sogenannte *gepoolte Stichprobenvarianz*. Dies sind die beiden Kernaussagen folgenden Theorems.

Theorem 40.3 (Parameterschätzung im Zweistichproben-T-Test-Modell). *Gegeben sei die Designmatrixform des Zweistichproben-T-Test-Modells. Dann ergeben sich für den Betaparameterschätzer*

$$\hat{\beta} = \begin{pmatrix} \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\ \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} \end{pmatrix} =: \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \quad (40.16)$$

und für den Varianzparameterschätzer

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} =: s_{12}^2 \quad (40.17)$$

◦

Beweis. Für $i = 1, 2$ sei $y_i := (y_{i1}, \dots, y_{in_i})^T$. Dann ergibt sich für den Betaparameterschätzer

$$\begin{aligned}
 \hat{\beta} &= (X^T X)^{-1} X^T y \\
 &= \left(\begin{pmatrix} 1_{n_1} & 0_{n_2} \\ 0_{n_1} & 1_{n_2} \end{pmatrix} \begin{pmatrix} 1_{n_1} & 0_{n_1} \\ 0_{n_2} & 1_{n_2} \end{pmatrix} \right)^{-1} \begin{pmatrix} 1_{n_1} & 0_{n_2} \\ 0_{n_1} & 1_{n_2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
 &= \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{pmatrix} \\
 &= \begin{pmatrix} n_1^{-1} & 0 \\ 0 & n_2^{-1} \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{pmatrix} \\
 &= \begin{pmatrix} \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\ \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} \end{pmatrix} \\
 &=: \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix}
 \end{aligned} \tag{40.18}$$

Gleichsam ergibt sich für Varianzparameterschätzer mit $n = n_1 + n_2$ und $p = 2$

$$\begin{aligned}
 \hat{\sigma}^2 &= \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p} \\
 &= \frac{1}{n_1 + n_2 - 2} \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} 1_{n_1} & 0_{n_1} \\ 0_{n_2} & 1_{n_2} \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \right)^T \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} 1_{n_1} & 0_{n_1} \\ 0_{n_2} & 1_{n_2} \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \right) \\
 &= \frac{1}{n_1 + n_2 - 2} \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1n_1} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ \vdots \\ y_{2n_2} - \bar{y}_2 \end{pmatrix}^T \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1n_1} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ \vdots \\ y_{2n_2} - \bar{y}_2 \end{pmatrix} \\
 &= \frac{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} \\
 &=: s_{12}^2.
 \end{aligned} \tag{40.19}$$

□

Man beachte, dass sich die Stichprobenvarianz s_y^2 der Komponenten von y im Allgemeinen von der gepoolten Stichprobenvarianz s_{12}^2 unterscheidet. Dies ist nicht zuletzt dadurch bedingt, dass die Stichprobenvarianz basierend auf dem Gesamtstichprobenmittel $\bar{y}$, die gepoolte Stichprobenvarianz dagegen basierend auf den gruppenspezifischen Stichprobenmittel $\bar{y}_1$ und $\bar{y}_2$ ermittelt wird. Wir wollen das Konzept der gepoolten Stichprobenvarianz hier aber nicht weiter vertiefen.

Modellevaluation

Basierend auf Theorem 38.1 formulieren wir nun die T-Teststatistik für das in Definition 40.2 in Designmatrixform definierte Zweistichproben-T-Test-Modell und geben ihre frequentistische Verteilung an.

Theorem 40.4 (T-Teststatistik des Zweistichproben-T-Tests). *Gegeben sei die Designmatrixform des Zweistichproben-T-Tests. Dann ergibt sich für die T-Teststatistik mit*

$$c := (1, -1)^T \text{ und } c^T \beta_0 =: \mu_0, \quad (40.20)$$

dass

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{y}_1 - \bar{y}_2 - \mu_0}{s_{12}} \right) \quad (40.21)$$

und es gilt

$$T \sim t(\delta, n_1 + n_2 - 2) \text{ mit } \delta = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\mu_1 - \mu_2 - \mu_0}{\sigma} \right). \quad (40.22)$$

◦

Beweis. Mit Theorem 38.1 gilt zunächst für die Zähler von T und δ , dass

$$c^T \hat{\beta} - c^T \beta_0 = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} - \mu_0 = \bar{y}_1 - \bar{y}_2 - \mu_0 \quad (40.23)$$

und

$$c^T \beta - c^T \beta_0 = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} - \mu_0 = \mu_1 - \mu_2 - \mu_0 \quad (40.24)$$

respektive. Weiterhin gilt für die Nenner von T und δ , dass

$$c^T (X^T X)^{-1} c = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} n_1^{-1} & 0 \\ 0 & n_2^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} n_1^{-1} & -n_2^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{n_1} + \frac{1}{n_2} \quad (40.25)$$

Außerdem gilt

$$\left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-\frac{1}{2}} = \left(\frac{n_2}{n_1 n_2} + \frac{n_1}{n_1 n_2} \right)^{-\frac{1}{2}} = \left(\frac{n_1 + n_2}{n_1 n_2} \right)^{-\frac{1}{2}} = \left(\frac{n_1 n_2}{n_1 + n_2} \right)^{\frac{1}{2}} \quad (40.26)$$

Zusammengenommen folgt direkt, dass

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{y}_1 - \bar{y}_2 - \mu_0}{s_{12}} \right) \text{ und } \delta = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\mu_1 - \mu_2 - \mu_0}{\sigma} \right). \quad (40.27)$$

□

Die Formen der T-Teststatistik und ihre Verteilung im Zweistichproben-T-Test Modell in Designmatrixform sind also wiederum natürlicherweise mit den entsprechenden Formen im ALM-freien Kontext identisch. Die entsprechende zur Kontrolle des Testumfangs bei Zweistichproben-T-Tests sowie der Gebrauch der Testgütefunktion zur Evaluation der Testtrennschärfe (Power) folgt also analog.

Anwendungsbeispiel

Folgender **R** Code demonstriert die Evaluation von 95%-Konfidenzintervallen für die Erwartungswertparameter μ_1 und μ_2 sowie Durchführung eines zweiseitigen Zweistichproben-T-Tests mit Nullhypothese

$$\Theta_0 := \left\{ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^2 \mid \mu_1 = \mu_2 \right\} \quad (40.28)$$

und Signifikanzlevel $\alpha_0 := 0.05$ für das oben skizzierte Anwendungsbeispiel.


```

# Dateneinlesen
D = read.csv("../data/507-t-tests.csv")
y_1 = D$BDI[D$COND == "F2F"]
y_2 = D$BDI[D$COND == "ONL"]

# Modellformulierung
n_1 = length(y_1)
n_2 = length(y_2)
n = n_1 + n_2
y = matrix(c(y_1, y_2), nrow = n)
p = 2
X = matrix(c(rep(1, n_1), rep(0, n_2),
              rep(0, n_1), rep(1, n_2)),
            nrow = n)

# Parameterschätzung
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
eps_hat = y - X %*% beta_hat
sigsqr_hat = (t(eps_hat) %*% eps_hat) / (n - p)

# Konfidenzintervall
delta = 0.95
t_delta = qt((1 + delta) / 2, n - 1)
lambda = diag(solve(t(X) %*% X))
kappa = matrix(rep(NA, p * 2), nrow = p)
for(j in 1:p){
  kappa[j,1] = beta_hat[j] - sqrt(sigsqr_hat * lambda[j]) * t_delta
  kappa[j,2] = beta_hat[j] + sqrt(sigsqr_hat * lambda[j]) * t_delta
}

# Hypothesentest
c = matrix(c(1, -1), nrow = 2)
mu_0 = 0
alpha_0 = 0.05
k_alpha_0 = qt(1 - (alpha_0 / 2), n - 1)
t_num = t(c) %*% beta_hat - mu_0
t_den = sqrt(sigsqr_hat * t(c) %*% solve(t(X) %*% X) %*% c)
t = t_num / t_den
if(abs(t) >= k_alpha_0){phi = 1} else {phi = 0}
pval = 2 * (1 - pt(abs(t), n_1 + n_2 - 2))

# Dataframe
# BDI Differenzwerte in der F2F Gruppe
# BDI Differenzwerte in der ONL Gruppe

# Anzahl Datenpunkte Gruppe 1 (F2F)
# Anzahl Datenpunkte Gruppe 2 (ONL)
# Gesamtanzahl Datenpunkte
# Datenvektor
# Anzahl Betaparameter
# Zweistichproben-T-Test Designmatrix

# Betaparameterschätzer
# Residuenvektor
# Varianzparameterschätzer

# Konfidenzbedingung
# \Psi^{-1}((1 + \delta) / 2, n - 1)
# \lambda_j Werte
# \beta_j Konfidenzintervall array
# Iteration über \beta_j
# untere KI Grenze
# obere KI Grenze

# Kontrastgewichtsvektor
# Nullhypothese H_0
# Signifikanzniveau
# kritischer Wert
# T-Teststatistik Zähler
# T-Teststatistik Nenner
# T-Teststatistik
# Test 1_{|T(X)| >= k_alpha_0}
# p-Wert

```

```

Betaparameterschätzer      : 3.5 7.42
95%-Konfidenzintervalle   : 1.41 5.32 5.59 9.51
Varianzparameterschätzer  : 12.27
alpha_0                   : 0.05
Kritischer Wert           : 2.07
Einstichproben-T-Teststatistik : -2.74
phi                        : 1
p-Wert                    : 0.01

```

Die Nullhypothese würde in diesem Fall bei einem kritischen Wert von $k_{0.05} = 2.07$ und einem Wert der T-Statistik von $T = -033$ nicht verworfen werden. Inferenzstatistisch besteht also keine Evidenz dafür, dass sich der wahre, aber unbekannte Erwartungswertparameter im F2F Therapiesetting vom wahren, aber unbekannten Erwartungswertparameter im ONL Therapiesetting unterscheidet. Die 95%Konfidenzintervalle für die wahren, aber unbekannten, Erwartungswertparameter μ_1 und μ_2 sind $[2.60, 7.74]$ und $[3.18, 8.32]$, respektive.

40.4. Literaturhinweise

Obwohl die frequentistische Literatur der ersten Hälfte des 20. Jahrhunderts von der Äquivalenz regressions- und varianzanalytischer linearer Modelle durchdrungen ist, fällt es schwer eine definite Quelle anzugeben, die hinsichtlich der Beschreibung von T-Tests als Spezialfälle des ALM Priorität hätte. Es sei hier deshalb eher allgemein auf Fisher (1925c) und Fisher (1935) verwiesen.

41. Einfaktorielle Varianzanalyse

41.1. Anwendungsszenario

Das Anwendungsszenario einer einfaktoriellen Varianzanalyse ist durch das Vorliegen von n univariaten Datenpunkten von zwei oder mehr Gruppen randomisierter experimenteller Einheiten gekennzeichnet, die sich hinsichtlich der Level eines experimentellen Faktors unterscheiden. Ist die Anzahl an Datenpunkten in jeder Gruppe gleich, so spricht man auch von einem balancierten einfaktoriellen Varianzanalysedesign. Von den Datenpunkten der i ten Gruppe bzw. des i ten Faktorlevels wird dabei angenommen, dass sie Realisierungen von jeweils n_i unabhängigen und identisch normalverteilten Zufallsvariablen sind, deren wahre, aber unbekannte, Erwartungswertparameter sich potentiell über die Gruppen hinweg unterscheiden und deren wahrer, aber unbekannter, Varianzparameter über Gruppen hinweg identisch ist. In diesen Grundannahmen handelt es sich bei dem Szenario der einfaktoriellen Varianzanalyse also um eine direkte Generalisierung des Einstich- und Zweistichproben-T-Test Szenarios zu (potentiell) mehr als zwei Gruppen. Umgekehrt können die Einstichproben- und Zweistichproben-T-Test Szenarien natürlich auch als einfaktorielle Varianzanalyseszenarien betrachtet werden, bei denen der experimentelle Faktor (nur) ein oder zwei Level, respektive, aufweist. Schließlich wird wie im Falle der T-Test Szenarien meist vorausgesetzt, dass ein Interesse an einem inferentiellen Vergleich der wahren, aber unbekannten, faktorlevelspezifischen Erwartungswertparameter besteht.

41.2. Anwendungsbeispiel

Als konkretes Anwendungsbeispiel betrachten wir die Analyse von Pre-Post-Interventions-BDI-Differenzwerten von drei Gruppen von jeweils 12 Patient:innen, die unterschiedliche Therapiesettings (Face-to-Face und Online) bzw. eine Wartelistenkontrollbedingung durchlaufen haben, wie in Tabelle 41.1 exemplarisch dargestellt. Die erste Spalte der Tabelle (COND) listet das patient:innenspezifische Therapiesetting (F2F: face-to-face, ONL: online, WLC: waitlist control) für jeweils drei Patient:innen jeder Studiengruppe auf. Die zweite Spalte der Tabelle (dBDI) list die entsprechenden patient:innenspezifischen Pre-PostInterventions-BDI-Differenzwerte. Positive Werte entsprechen hier wieder einer Abnahme der Depressionssymptomatik, negative Werte einer Zunahme der Depressionssymptomatik.

Tabelle 41.1. Exemplarische Pre-Post-Intervention-BDI-Differenzwerte des Beispieldatensatzes

	COND	dBDI
1	F2F	9
2	F2F	7
3	F2F	10

Tabelle 41.1. Exemplarische Pre-Post-Intervention-BDI-Differenzwerte des Beispieldatensatzes

	COND	dBDI
4	F2F	11
13	ONL	2
14	ONL	7
15	ONL	9
16	ONL	9
17	ONL	8
25	WLC	-1
26	WLC	2
27	WLC	-3
28	WLC	-1
29	WLC	0

Abbildung 41.1 zeigt eine Visualisierung des gesamten Datensatzes. Die Balken repräsentieren die gruppenspezifischen Stichprobenmittelwerte, die zugehörigen Fehlerbalken die gruppenspezifischen Stichprobenstandardabweichungen. Die Punktwolken repräsentieren die gruppenspezifischen Datenpunkte. In den F2F und ONL Gruppen ist die Veränderung des BDI Wertes stärker ausgeprägt als in der WLC Gruppe.

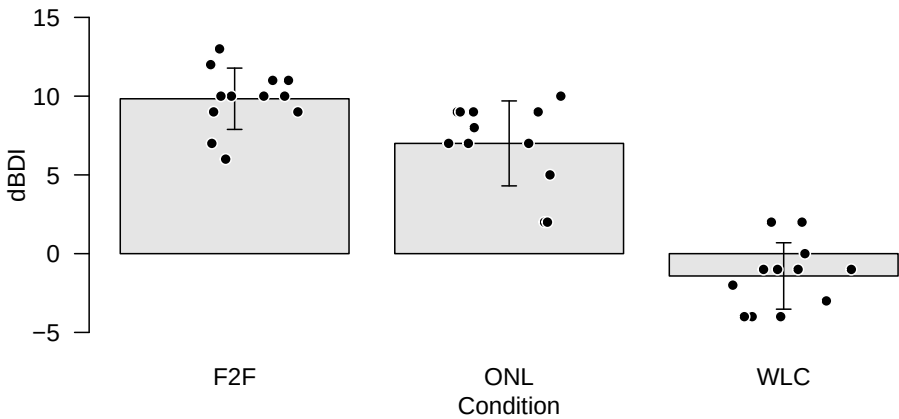


Abbildung 41.1. Datendarstellung des Anwendungsbeispiels zur Einfaktoriellen Varianzanalyse.

In Tabelle 41.2 fassen wir die Deskriptivstatistiken des Beispieldatensatzes, aufgeschlüsselt nach Therapiebedingungen, zusammen.

Tabelle 41.2. Deskriptivstatistiken der Pre-Post BDI Differenzwerte

	n	Max	Min	Median	Mean	Var	Std
F2F	12	13	6	10.0	9.83	3.79	1.95
ONL	12	10	2	7.5	7.00	7.27	2.70

Tabelle 41.2. Deskriptivstatistiken der Pre-Post BDI Differenzwerte

	n	Max	Min	Median	Mean	Var	Std
WLC	12	2	-4	-1.0	-1.42	4.45	2.11

41.3. Modellformulierung

Wir definieren das Modell der EVA zunächst in Erwartungswertparameterdarstellung. Dabei nutzen wir den Index i um die experimentellen Gruppen zu indizieren und den Index j um die experimentellen Einheiten innerhalb der Gruppen zu indizieren.

Definition 41.1 (EVA-Modell in Erwartungswertparameterdarstellung). Für $i = 1, \dots, p$ und $j = 1, \dots, n_i$ seien y_{ij} Zufallsvariablen, die die $n := \sum_{i=1}^p n_i$ Datenpunkte eines EVA Szenarios modellieren. Dann hat das EVA-Modell in Erwartungswertparameterdarstellung die strukturelle Form

$$y_{ij} = \mu_i + \varepsilon_{ij} \text{ mit } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ u.i.v. mit } \mu_i \in \mathbb{R}, \sigma^2 > 0, \quad (41.1)$$

die Datenverteilungsform

$$y_{ij} \sim N(\mu_i, \sigma^2) \text{ u.v. mit } \mu_i \in \mathbb{R}, \sigma^2 > 0. \quad (41.2)$$

und für den n -dimensionalen Datenvektor definiert als

$$y := (y_{11}, \dots, y_{1n_1}, y_{21}, \dots, y_{2n_2}, \dots, y_{p1}, \dots, y_{pn_p})^T \quad (41.3)$$

die Designmatrixform

$$y = X\beta + \varepsilon \quad (41.4)$$

mit

$$X := \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_p} & 0_{n_p} & \cdots & 1_{n_p} \end{pmatrix}, \beta := \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \in \mathbb{R}^p, \varepsilon \sim N(0_n, \sigma^2 I_n), \sigma^2 > 0. \quad (41.5)$$

Für $n_i := m$ für alle $i = 1, \dots, p$ heißt das Modell balanciert.

•

Der Vergleich mit der Definition des Modells des Zweistichproben-T-Tests in Definition 40.2 zeigt, dass es sich bei dieser Formulierung des EVA-Modells um die direkte Generalisierung des Zweistichproben-T-Test Modells von $p = 2$ für ein beliebiges $p \in \mathbb{N}$ handelt.

Motivation der Effektdarstellung

Das EVA-Modell in Erwartungswertparameterdarstellung ist ein valides Modell, auf dessen Grundlage sowohl Parameterschätzung als auch Parameter- und Modellinferenz für das EVA-Szenario entwickelt werden können (vgl. Georgii (2009)). Im Sinne der Konsistenz mit den Modellen der mehrfaktoriellen Varianzanalyse bietet sich jedoch eine Reparametrisierung des Betaparametervektors an. Kern dieser Reparametrisierung ist es, den Erwartungswertparameter der i ten Gruppe als Summe eines gruppenübergreifenden Erwartungswertparameters $\mu_0 \in \mathbb{R}$ und eines gruppenspezifischen Effektparameters $\alpha_i \in \mathbb{R}$ zu modellieren,

$$\mu_i := \mu_0 + \alpha_i \text{ für } i = 1, \dots, p. \quad (41.6)$$

Dabei modelliert α_i die Differenz zwischen dem i ten Erwartungswertparameter μ_i und dem gruppenübergreifenden Erwartungswertparameter μ_0 ,

$$\alpha_i = \mu_i - \mu_0 \text{ für } i = 1, \dots, p. \quad (41.7)$$

Allerdings hat die in dieser Form vorgenommene Reparametrisierung einen entscheidenden Nachteil: es werden p Erwartungswertparameter $\mu_i, i = 1, \dots, p$ durch die $p + 1$ Parameter μ_0 und $\alpha_i, i = 1, \dots, p$ dargestellt. Diese Darstellung ist im Allgemeinen nicht eindeutig. Zum Beispiel können die Erwartungswertparameter $\mu_1 = 3, \mu_2 = 5, \mu_3 = 6$ sowohl durch den gruppenspezifischen Erwartungswertparameter $\mu_0 = 0$ und die gruppenunspezifischen Effektparameter $\alpha_1 = 3, \alpha_2 = 5, \alpha_3 = 6$ als auch durch den gruppenunspezifischen Erwartungswertparameter $\mu_0 = 1$ und die gruppenspezifischen Effektparameter $\alpha_1 = 2, \alpha_2 = 4, \alpha_3 = 5$ dargestellt werden. Man sagt in diesem Kontext auch, dass das EVA-Modell in der Form von Gleichung 41.6 überparametrisiert ist.

Datenanalytisch hat die Überparametrisierung eines Varianzanalysemodells den Nachteil, dass aus p geschätzten Erwartungswertparametern $p + 1$ Betaparameterschätzer bestimmt werden müssten, was wie oben gesehen nicht eindeutig erfolgen kann. Um diese Probleme in der Effektparameterdarstellung des EVA-Modells zu umgehen und diese konsistent auf mehrfaktorielle Varianzanalysemodelle zu übertragen, bietet sich die Einführung der Nebenbedingung

$$\alpha_1 := 0 \quad (41.8)$$

an. Es wird also ein Effektparameter von vornherein als identisch Null angenommen. Für die gruppenspezifischen Erwartungswertparameter ergibt sich damit

$$\begin{aligned} \mu_1 &:= \mu_0 \\ \mu_i &:= \mu_0 + \alpha_i \text{ für } i = 2, \dots, p. \end{aligned} \quad (41.9)$$

Hierbei wird die erste Gruppe nun als Referenzgruppe bezeichnet und die α_i modellieren die Differenz zwischen dem Erwartungswertparameter der i ten Gruppe und dem Erwartungswertparameter der ersten Gruppe:

$$\alpha_i = \mu_i - \mu_0 = \mu_i - \mu_1 \text{ für } i = 1, \dots, p. \quad (41.10)$$

μ_0 ist, unter der Nebenbedingung $\alpha_1 := 0$ also kein gruppenübergreifender Erwartungswertparameter mehr, sondern identisch mit dem Erwartungswertparameter der ersten Gruppe. Welche tatsächliche experimentelle Gruppe dabei als "erste Gruppe" definiert wird, ist datenanalytisch unerheblich. Datenanalytisch entscheidend dagegegen ist, dass der entsprechenden Erwartungswertparameterschätzer $\hat{\mu}_0$ korrekt als Erwartungswertparameterschätzer der Referenzgruppe und die $\hat{\alpha}_i$ für $i = 2, \dots, p$ korrekt als geschätzte

Erwartungswertparameterdifferenzen zwischen dem Erwartungswertparameter der Referenzgruppe und dem Erwartungswertparameter der i ten Gruppe verstanden werden.

Wir formalisieren das oben Gesagte in folgendem Theorem.

Theorem 41.1 (EVA-Modell in Effektdarstellung mit Referenzgruppe). *Gegeben sei das EVA-Modell in Erwartungswertparameterdarstellung. Dann können die Zufallsvariablen, die die Datenpunkte des EVA-Szenarios modellieren, äquivalent in der strukturellen Form*

$$\begin{aligned} y_{1j} &= \mu_0 + \varepsilon_{1j} \quad \text{mit } \varepsilon_{1j} \sim N(0, \sigma^2) \text{ u.i.v. für } j = 1, \dots, n_1 \\ y_{ij} &= \mu_0 + \alpha_i + \varepsilon_{ij} \quad \text{mit } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ u.i.v. für } i = 2, \dots, p, j = 1, \dots, n_i \end{aligned} \quad (41.11)$$

mit

$$\alpha_i := \mu_i - \mu_1 \quad \text{für } i = 2, \dots, p \quad (41.12)$$

und in der entsprechenden Datenverteilungsform

$$\begin{aligned} y_{1j} &\sim N(\mu_0, \sigma^2) \quad \text{u.i.v. für } j = 1, \dots, n_1 \text{ mit } \mu_1 \in \mathbb{R}, \sigma^2 > 0 \\ y_{ij} &\sim N(\mu_0 + \alpha_i, \sigma^2) \quad \text{u.v. für } i = 2, \dots, p, j = 1, \dots, n_i \text{ mit } \alpha_i \in \mathbb{R}, \sigma^2 > 0 \end{aligned} \quad (41.13)$$

geschrieben werden.

◦

Beweis. Nach Definition gilt

$$\mu_i = \mu_0 + \mu_i - \mu_0. \quad (41.14)$$

Die Parametrisierungen mit μ_i und mit $\mu_0 + \mu_i - \mu_0$ sind also gleich und damit äquivalent. Dann folgt aber auch

$$\mu_i = \mu_0 + (\mu_i - \mu_0) =: \mu_0 + \alpha_i \quad \text{für } i = 1, \dots, p. \quad (41.15)$$

Mit $\alpha_1 := 0$ gilt dann $\mu_1 = \mu_0$ und $\mu_i = \mu_0 + \alpha_i$ für $i = 2, \dots, p$, wie im Theorem behauptet. $\square$

Basierend auf Theorem 41.1 definieren wir nun das Modell der EVA in Effektdarstellung.

Definition 41.2 (EVA-Modell in Effektdarstellung mit Referenzgruppe). Für $i = 1, \dots, p$ und $j = 1, \dots, n_i$ seien y_{ij} Zufallsvariablen, die die $n := \sum_{i=1}^p n_i$ Datenpunkte eines EVA-Szenarios modellieren. Dann hat das EVA-Modell in Effektdarstellung die strukturelle Form

$$\begin{aligned} y_{1j} &= \mu_0 + \varepsilon_{1j} \quad \text{mit } \varepsilon_{1j} \sim N(0, \sigma^2) \text{ u.i.v. für } j = 1, \dots, n_1 \\ y_{ij} &= \mu_0 + \alpha_i + \varepsilon_{ij} \quad \text{mit } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ u.i.v. für } i = 2, \dots, p, j = 1, \dots, n_i \end{aligned} \quad (41.16)$$

mit

$$\alpha_i := \mu_i - \mu_1 \quad \text{für } i = 2, \dots, p, \quad (41.17)$$

die Datenverteilungsform

$$\begin{aligned} y_{1j} &\sim N(\mu_0, \sigma^2) \quad \text{u.i.v. für } j = 1, \dots, n_1 \text{ mit } \mu_0 \in \mathbb{R}, \sigma^2 > 0 \\ y_{ij} &\sim N(\mu_0 + \alpha_i, \sigma^2) \quad \text{u.v. für } i = 2, \dots, p, j = 1, \dots, n_i \text{ mit } \alpha_i \in \mathbb{R}, \sigma^2 > 0 \end{aligned} \quad (41.18)$$

und die Designmatrixform

$$y = X\beta + \varepsilon \quad \text{mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (41.19)$$

und

$$n := \sum_{i=1}^p n_i, y := \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{p1} \\ \vdots \\ y_{pn_p} \end{pmatrix} \in \mathbb{R}^n, X := \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 1_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 1_{n_p} & 0_{n_p} & \cdots & 1_{n_p} \end{pmatrix} \in \mathbb{R}^{n \times p}, \beta := \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} \in \mathbb{R}^p \text{ und } \sigma^2 > 0. \quad (41.20)$$

•

Beispiel

Um die Unterschiede und Gemeinsamkeiten zwischen der Erwartungswertparameterdarstellung und der Effektparameterdarstellung des EVA-Modells in ihrer Designmatrixform zu verdeutlichen, betrachten wir ein Beispielszenario mit $n_i := 4$ und $p = 3$ für $i = 1, \dots, p$, also $n = 12$. Für die Erwartungswertparameterdarstellung gilt dann

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_{12}, \sigma^2 I_{12}) \quad (41.21)$$

mit

$$y := \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{41} \\ y_{42} \\ y_{43} \end{pmatrix}, X := \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{12 \times 3}, \beta := \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \in \mathbb{R}^3 \text{ und } \sigma^2 > 0. \quad (41.22)$$

Für die Effektparameterdarstellung dagegen gilt

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_{12}, \sigma^2 I_{12}) \quad (41.23)$$

mit

$$y := \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{41} \\ y_{42} \\ y_{43} \end{pmatrix}, X := \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{12 \times 3}, \beta := \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \in \mathbb{R}^3 \text{ und } \sigma^2 > 0 \quad (41.24)$$

Folgender **R** Code demonstriert die Realisierung von Daten in einem EVA-Szenario mit eben dieser Effektparameterdarstellung.

```
# Modellformulierung
library(MASS)
m = 4 # Multivariate Normalverteilung
p = 3 # Anzahl von Datenpunkten der iten Gruppe
n = p*m # Anzahl Gruppen
Xt = cbind( # Gesamtanzahl Datenpunkte
  matrix(1,nrow = n, ncol = 1), # Designmatrix
  kronecker(diag(p), matrix(1,nrow = m,ncol = 1)))
X = Xt[,-2]
I_n = diag(n) # n x n Einheitsmatrix
beta = matrix(c(10,-3,-12), nrow = p) # \beta = (\mu_0, \alpha_2, \alpha_3, \alpha_4)
sigsqr = 14 # \sigma^2
y = mvrnorm(1, X %*% beta, sigsqr*I_n) # eine Realisierung eines n-dimensionalen ZVs
print(X)
```

```
      [,1] [,2] [,3]
[1,]  1    0    0
[2,]  1    0    0
[3,]  1    0    0
[4,]  1    0    0
[5,]  1    1    0
[6,]  1    1    0
[7,]  1    1    0
[8,]  1    1    0
[9,]  1    0    1
[10,] 1    0    1
[11,] 1    0    1
[12,] 1    0    1
```

41.4. Modellschätzung

Wir betrachten nun die Betaparameterschätzung in der Effektparameterdarstellung des EVA-Modells mit Referenzgruppe. Entsprechend der Interpretation der Betaparameterkomponenten werden dabei μ_0 durch das Stichprobenmittel der Referenzgruppe und die $\alpha_2, \dots, \alpha_p$ durch die Differenzen des jeweiligen Gruppenstichprobenmittels und des Referenzgruppenstichprobenmittels geschätzt. Auf die Schätzung des Varianzparameters, der sich wie im Zweistichproben-T-Test Modell zu einer *gepoolten Stichprobenvarianz* ergibt, wollen wir hier nicht weiter eingehen.

Theorem 41.2 (Betaparameterschätzung im EVA-Modell). *Gegeben sei die Designmatrixform des EVA in Effektdarstellung mit Referenzgruppe. Dann ergibt sich für den Betaparameterschätzer*

$$\hat{\beta} := \begin{pmatrix} \hat{\mu}_0 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_p \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \vdots \\ \bar{y}_p - \bar{y}_1 \end{pmatrix} \quad (41.25)$$

wobei

$$\bar{y}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad (41.26)$$

das Stichprobenmittel der i ten Gruppe bezeichnet.

◦

Beweis. Wir halten zunächst fest, dass

$$\begin{aligned} X^T X &= \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & \dots & 1 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & & & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} n & n_2 & n_3 & \dots & n_p \\ n_2 & n_2 & 0 & \dots & 0 \\ n_3 & 0 & n_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_p & 0 & 0 & \dots & n_p \end{pmatrix}. \end{aligned} \quad (41.27)$$

Die Inverse von $X^T X$ ist

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{n_1+n_2}{n_1 n_2} & \dots & \frac{1}{n_1} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n_1} & \frac{1}{n_1} & \dots & \frac{n_1+n_p}{n_1 n_p} \end{pmatrix} \quad (41.28)$$

So gilt zum Beispiel für $p = 3$, dass

$$X^T X = \begin{pmatrix} n & n_2 & n_3 \\ n_2 & n_2 & 0 \\ n_3 & 0 & n_3 \end{pmatrix} \quad \text{und} \quad (X^T X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & -\frac{1}{n_1} & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{n_1+n_2}{n_1 n_2} & \frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{1}{n_1} & \frac{n_1+n_3}{n_1 n_3} \end{pmatrix} \quad (41.29)$$

Wir halten weiterhin fest, dass

$$\begin{aligned}
 X^T y &= \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & \dots & 1 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ & \vdots & & & \vdots & & \vdots & & \vdots & \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{p1} \\ \vdots \\ y_{pn_p} \end{pmatrix} \\
 &= \begin{pmatrix} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} \\ \sum_{j=1}^{n_2} y_{2j} \\ \vdots \\ \sum_{j=1}^{n_p} y_{pj} \end{pmatrix}.
 \end{aligned} \tag{41.30}$$

Es ergibt sich also

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} \frac{1}{n_1} & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{n_1+n_2}{n_1 n_2} & \dots & \frac{1}{n_1} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n_1} & \frac{1}{n_1} & \dots & \frac{n_1+n_p}{n_1 n_p} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} \\ \sum_{j=1}^{n_2} y_{2j} \\ \vdots \\ \sum_{j=1}^{n_p} y_{pj} \end{pmatrix} \tag{41.31}$$

Für die erste Komponente von $\hat{\beta}$ ergibt sich damit

$$\begin{aligned}
 \hat{\beta}_0 &= \frac{1}{n_1} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} - \frac{1}{n_1} \sum_{j=1}^{n_2} y_{2j} - \dots - \frac{1}{n_1} \sum_{j=1}^{n_p} y_{pj} \\
 &= \frac{1}{n_1} \left(\left(\sum_{j=1}^{n_1} y_{1j} + \sum_{j=1}^{n_2} y_{2j} + \dots + \sum_{j=1}^{n_p} y_{pj} \right) - \sum_{j=1}^{n_2} y_{2j} - \dots - \sum_{j=1}^{n_p} y_{pj} \right) \\
 &= \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\
 &= \bar{y}_1.
 \end{aligned} \tag{41.32}$$

Für die zweite Komponente von $\hat{\beta}$ und analog für alle weiteren ergibt sich

$$\begin{aligned}
 \hat{\beta}_2 &= -\frac{1}{n_1} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} + \frac{n_1+n_2}{n_1 n_2} \sum_{j=1}^{n_2} y_{2j} + \frac{1}{n_1} \sum_{j=1}^{n_3} y_{3j} + \dots + \frac{1}{n_1} \sum_{j=1}^{n_p} y_{pj} \\
 &= \frac{n_1+n_2}{n_1 n_2} \sum_{j=1}^{n_2} y_{2j} - \frac{1}{n_1} \left(\left(\sum_{j=1}^{n_1} y_{1j} + \sum_{j=1}^{n_2} y_{2j} + \dots + \sum_{j=1}^{n_p} y_{pj} \right) - \sum_{j=1}^{n_3} y_{3j} - \dots - \sum_{j=1}^{n_p} y_{pj} \right) \\
 &= \frac{n_1+n_2}{n_1 n_2} \sum_{j=1}^{n_2} y_{2j} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} - \frac{1}{n_1} \sum_{j=1}^{n_2} y_{2j} \\
 &= \frac{n_1+n_2}{n_1 n_2} \sum_{j=1}^{n_2} y_{2j} - \frac{n_2}{n_1 n_2} \sum_{j=1}^{n_2} y_{2j} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\
 &= \frac{n_1}{n_1 n_2} \sum_{j=1}^{n_2} y_{2j} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\
 &= \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\
 &= \bar{y}_2 - \bar{y}_1.
 \end{aligned} \tag{41.33}$$

□

Folgender **R** Code implementiert die Parameterschätzung des EVA-Modells für den Beispieldatensatz. Neben der mithilfe des Betaparameterschätzers gewonnenen Schätzwerte für den Referenzgruppen- und die Effektparameter evaluiert der Code auch die stichprobenmittel(differenzen)basierten Schätzer aus Theorem 41.2. Weiterhin evaluiert der Code neben dem Varianzparameterschätzer auch die gepoolte Stichprobenvarianz sowie die Gesamtstichprobenvarianz, die sich deutlich unterscheiden.

```
# Datenmanagement
D = read.csv("./_data/508-einfaktorielle-varianzanalyse.csv") # Dataframe
y = D$dBDI # Datenvektor
y_1 = D$dBDI[D$COND == "F2F"] # Daten F2F
y_2 = D$dBDI[D$COND == "ONL"] # Daten ONL
y_3 = D$dBDI[D$COND == "WLC"] # Daten WLC

# Modellformulierung
p = 3 # drei Gruppen
m = length(y_1) # balancierter Design
n = p*m # Datenvektordimension
Xt = cbind( # Designmatrix
  matrix(1, nrow = n, ncol = 1),
  kronecker(diag(p), matrix(1, nrow = m, ncol = 1)))
X = Xt[,-2]

# Modellschätzung
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y # Betaparameterschätzer
eps_hat = y - X %*% beta_hat # Residuenvektor
sigsqr_hat = (t(eps_hat) %*% eps_hat) / (n-p) # Varianzparameterschätzer
s_sqr_123 = ((m-1)*var(y_1) + # gepoolte Stichprobenvarianz
  (m-1)*var(y_2) +
  (m-1)*var(y_3)) / (m+m+m-p)

hat{beta} : 9.83 -2.83 -11.25
bar{y}_1, bar{y}_2, bar{y}_3 : 9.83 7 -1.42
bar{y}_1, bar{y}_2 - bar{y}_1, bar{y}_3 - bar{y}_1 : 9.83 -2.83 -11.25
hat{sigsqr} : 5.17
s_123^2 : 5.17
s_y^2 : 28.35
```

41.5. Modellevaluation

Prinzipiell sind alle Parameterschätzwerte in einem EVA-Modell von Interesse und können mithilfe von T-Statistiken im Sinne von Konfidenzintervallen oder Hypothesentests evaluiert werden. Traditionell steht im EVA-Szenario allerdings häufig die Evaluation der Nullhypothese, dass die wahren, aber unbekannten, Erwartungswertparameter aller Gruppen identisch sind, im Vordergrund. Vor dem Hintergrund der EVA Effektdarstellung mit Referenzgruppe entspricht dies der Nullhypothese, dass die Effektparameter gleich Null sind, formal

$$\Theta_0 := \left\{ \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} \in \mathbb{R}^p \mid \alpha_i = 0 \text{ für } i = 2, \dots, p \right\} = \mathbb{R} \times \{0_{p-1}\} \quad (41.34)$$

und

$$\Theta_1 := \left\{ \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} \in \mathbb{R}^p \mid \alpha_i \neq 0 \text{ für mindestens ein } i = 2, \dots, p \right\} = \mathbb{R}^p \setminus \Theta_0. \quad (41.35)$$

Zur Evaluation der Nullhypothese wird dabei im Allgemeinen eine F-Statistik genutzt. Im Folgenden wollen wir zunächst diese F-Statistik anhand einer Quadratsummenzerlegung der Datenvariabilität in einem EVA-Szenario entwickeln. In diesem Kontext führen wir mit η^2 (Eta-Quadrat) auch ein zu dem aus Kapitel 35 bekannten Bestimmtheitsmaß R^2 analoges Effektstärkenmaß ein. Ausgestattet mit der speziellen Form der F-Statistik für das EVA-Modell diskutieren wir dann den traditionellen Test der EVA Nullhypothese.

41.5.1. Quadratsummenzerlegung und Bestimmtheitsmaß η^2

Die Variabilität der Daten eines EVA-Szenarios lässt sich wie in folgendem Theorem dargestellt im Sinne einer Quadratsummenzerlegung schreiben.

Theorem 41.3 (Quadratsummenzerlegung bei einfaktorieller Varianzanalyse). *Für $i = 1, \dots, p$ und $j = 1, \dots, n_i$ sei y_{ij} die j te Datenvariable in der i ten Gruppe eines EVASzenarios. Weiterhin seien mit $n := \sum_{i=1}^p n_i$*

$$\bar{y} := \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} \text{ und } \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad (41.36)$$

das Gesamtstichprobenmittel und i te Stichprobenmittel, respektive. Schließlich seien

- *die Total-Sum-of-Squares definiert als*

$$SQT := \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \quad (41.37)$$

- *die Between-Sum-of-Squares definiert als*

$$SQB := \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 \quad (41.38)$$

und die Within-Sum-of-Squares definiert als

$$SQW := \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad (41.39)$$

Dann gilt

$$SQT = SQB + SQW. \quad (41.40)$$

◦

Beweis. Es gilt

$$\begin{aligned}
\text{SQT} &= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \\
&= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^p \sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}))^2 \\
&= \sum_{i=1}^p \sum_{j=1}^{n_i} \left((y_{ij} - \bar{y}_i)^2 + 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) + (\bar{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{j=1}^{n_i} 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) + \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) + n_i (\bar{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} \left(y_{ij} - \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \right) + n_i (\bar{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y}) \left(\sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \right) \right) + n_i (\bar{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y}) \left(\sum_{j=1}^{n_i} y_{ij} - \frac{n_i}{n_i} \sum_{j=1}^{n_i} y_{ij} \right) + n_i (\bar{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y}) \left(\sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} y_{ij} \right) + n_i (\bar{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + n_i (\bar{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 \\
&= \text{SQW} + \text{SQB}
\end{aligned} \tag{41.41}$$

und damit direkt

$$\text{SQT} = \text{SQB} + \text{SQW} \tag{41.42}$$

□

Die Quadratsummenzerlegung im Szenario der EVA ist offenbar analog zur Quadratsummenzerlegung bei einer Ausgleichsgerade (vgl. Kapitel 34.1). Dementsprechend wird die Within-Sum-of-Squares der EVA auch häufig als Residual-Sum-of-Squares bezeichnet. Die Begriffsbildungen von Theorem 41.3 ergeben sich intuitiv wiederum wie folgt:

- SQT repräsentiert die Gesamtvariabilität der Daten y_{ij} um das Gesamtstichprobenmittel $\bar{y}$.
- SQB repräsentiert die anhand des jeweiligen Gruppenumfangs n_i gewichtete Variabilität der Gruppenstichprobenmittel um das Gesamtstichprobenmittel. Große Werte von SQB implizieren also eine große Abhängigkeit der Gruppenstichprobenmittel von dem jeweils betrachteten Faktorlevel i , kleine Werte von SQB dagegen eine geringe Abhängigkeit der Gruppenstichprobenmittel von dem jeweils betrachteten Faktorlevel. In diesem Sinne repräsentiert SQB also die durch die Betrachtung des jeweiligen Faktorlevels erklärte Datenvariabilität und ist analog zur Explained-Sum-of-Squares (SQE) der Quadratsummenzerlegung bei einer Ausgleichsgerade.

- SQW repräsentiert die über Faktorlevel summierte Datenvariabilität die nach Erklärung der Datenvariabilität in der i ten Gruppe durch ihr jeweiliges Stichprobenmittel verbleibt. Damit ist die SQW analog zur Residual-Sum-of-Squares der Quadratsummenzerlegung bei Ausgleichsgerade.

In der Zusammenschau quantifiziert SQB also die Stärke der Unterschiede zwischen den Faktorleveln. Das Effektstärkenmaß η^2 und die F-Statistik der EVA setzen die SQB nun jeweils in ein anderes Verhältnis: η^2 vergleicht die SQB mit der SQT und betrachtet damit den Anteil der Variabilität zwischen den Faktorleveln an der Gesamtvariabilität der Daten. Die F-Statistik dagegen vergleicht die SQB mit der SQW und setzt damit den Einfluß der Faktorenlevel in das Verhältnis mit der nicht erklärten Datenvariabilität nach Subtraktion dieses Einflusses. Wir betrachten an dieser Stelle zunächst das Effektstärkemaß η^2 mithilfe folgender Definition.

Definition 41.3 (Effektstärkenmaß η^2). Für ein EVA-Szenario seien die Between-Sum-of-Squares SQB und die Total-Sum-of-Squares SQT definiert wie in Theorem 41.3. Dann ist das Effektstärkenmaß η^2 definiert als

$$\eta^2 := \frac{\text{SQB}}{\text{SQT}} \quad (41.43)$$

•

Der Vergleich mit Definition 35.4, dass η^2 analog zum Bestimmtheitsmaß R^2 definiert ist. Wie oben beschrieben gibt η^2 den Anteil der Datenvariabilität zwischen den Faktorleveln an der Gesamtdatenvariabilität an. Schließlich folgt mit

$$\text{SQT} = \text{SQB} + \text{SQW} \quad (41.44)$$

analog zu R^2 sofort, dass für $\text{SQT} \neq 0$ gilt, dass $\eta^2 \in [0, 1]$, weil einerseits

$$\text{SQB} = 0 \Rightarrow \text{SQT} = \text{SQW} \text{ und } \eta^2 = 0 \quad (41.45)$$

und andererseits

$$\text{SQW} = 0 \Rightarrow \text{SQT} = \text{SQB} \text{ und } \eta^2 = 1. \quad (41.46)$$

41.5.2. F-Teststatistik

Wir wollen nun zeigen, dass für die Designmatrixform der Effektdarstellung mit Referenzgruppe des EVA-Modells die F-Statistik bei der Partitionierung $p_0 := 1$ und $p_1 := p - 1$ (vgl. Kapitel 39) als Verhältnis skalierten Versionen der SQB und SQW geschrieben werden kann. Dabei impliziert $p_0 := 1$ hier insbesondere, dass das betrachtete reduzierte EVA-Modell die Designmatrix $X_0 := 1_n$ und den Betaparameter $\beta := \mu_0$ hat, und damit insbesondere auch, dass das reduzierte EVA-Modell keine Effektparameter hat. Die dabei relevanten Skalierungsfaktoren beziehen die SQB und SQW jeweils auf die Anzahl der Effektparameter bzw. die Differenz aus Gesamtanzahl der Datenpunkte und Anzahl der Betaparameter, respektive. Es gilt folgendes Theorem.

Theorem 41.4 (F-Statistik der einfaktoriellen Varianzanalyse). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (41.47)$$

die Designmatrixform der Effektdarstellung mit Referenzgruppe des EVA-Modells. Weiterhin sei dieses Modell im Sinne von Definition 39.2 partitioniert mit $p_0 := 1$ und $p_1 := p - 1$. Schließlich seien

- *die Mean-Between-Sum-of-Squares definiert als*

$$MSB := \frac{SQB}{p - 1} \quad (41.48)$$

- *und die Mean-Within-Sum-of-Squares definiert als*

$$MSW := \frac{SQW}{n - p} \quad (41.49)$$

wobei $p - 1$ auch als Between-Freiheitsgrade und $n - p$ auch als Within-Freiheitsgrade bezeichnet werden. Dann gilt mit der Definition der F-Statistik (Definition 39.3), dass

$$F = \frac{MSB}{MSW}. \quad (41.50)$$

◦

Beweis. Wir halten zunächst fest, dass für den Betaparameterschätzer des reduzierten Modells gilt, dass

$$\hat{\beta}_0 = (X_0^T X_0)^{-1} X_0^T y = (1_n^T 1_n)^{-1} 1_n^T y = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} = \bar{y} \quad (41.51)$$

Weiterhin ergibt sich

$$\hat{\varepsilon}_0^T \hat{\varepsilon}_0 = (y - X_0 \hat{\beta}_0)^T (y - X_0 \hat{\beta}_0) = (y - 1_n \bar{y})^T (y - 1_n \bar{y}) = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \text{SQT}. \quad (41.52)$$

Der Betaparameterschätzer des vollständigen Modells ergibt sich wie oben gesehen zu

$$\hat{\beta} = \begin{pmatrix} \hat{\mu}_0 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_m \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\ \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \\ \vdots \\ \frac{1}{n_m} \sum_{j=1}^{n_m} y_{mj} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \vdots \\ \bar{y}_p - \bar{y}_1 \end{pmatrix}, \quad (41.53)$$

so dass

$$\begin{aligned}
 \hat{\varepsilon}^T \hat{\varepsilon} &= (y - X\hat{\beta})^T (y - X\hat{\beta}) \\
 &= \left(\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{p1} \\ \vdots \\ y_{pn_p} \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & & 0 \\ 1 & 1 & & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & & 1 \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \bar{y}_p - \bar{y}_1 \end{pmatrix} \right)^T \left(\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{p1} \\ \vdots \\ y_{pn_p} \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & & 0 \\ 1 & 1 & & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & & 1 \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \bar{y}_p - \bar{y}_1 \end{pmatrix} \right) \\
 &= \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1n_1} - \bar{y}_1 \\ y_{21} - \bar{y}_1 - \bar{y}_2 + \bar{y}_1 \\ \vdots \\ y_{2n_2} - \bar{y}_1 - \bar{y}_2 + \bar{y}_1 \\ \vdots \\ y_{p1} - \bar{y}_1 - \bar{y}_p + \bar{y}_1 \\ \vdots \\ y_{pn_p} - \bar{y}_1 - \bar{y}_p + \bar{y}_1 \end{pmatrix}^T \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1n_1} - \bar{y}_1 \\ y_{21} - \bar{y}_1 - \bar{y}_2 + \bar{y}_1 \\ \vdots \\ y_{2n_2} - \bar{y}_1 - \bar{y}_2 + \bar{y}_1 \\ \vdots \\ y_{p1} - \bar{y}_1 - \bar{y}_p + \bar{y}_1 \\ \vdots \\ y_{pn_p} - \bar{y}_1 - \bar{y}_p + \bar{y}_1 \end{pmatrix} \\
 &= \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1n_1} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ \vdots \\ y_{2n_2} - \bar{y}_2 \\ \vdots \\ y_{p1} - \bar{y}_p \\ \vdots \\ y_{pn_p} - \bar{y}_p \end{pmatrix}^T \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1n_1} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ \vdots \\ y_{2n_2} - \bar{y}_2 \\ \vdots \\ y_{p1} - \bar{y}_p \\ \vdots \\ y_{pn_p} - \bar{y}_p \end{pmatrix} \\
 &= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \\
 &= \text{SQW}.
 \end{aligned} \tag{41.54}$$

Mit dem Theorem zur Quadratsummenzerlegung bei einfaktorieller Varianzanalyse

$$\text{SQT} = \text{SQB} + \text{SQW} \Leftrightarrow \text{SQB} = \text{SQT} - \text{SQW} \tag{41.55}$$

folgt sofort, dass

$$\text{SQB} = \text{SQT} - \text{SQW} = \hat{\varepsilon}_0^T \hat{\varepsilon}_0 - \hat{\varepsilon}^T \hat{\varepsilon}. \tag{41.56}$$

Dann aber folgt auch direkt, dass

$$\frac{\text{MSB}}{\text{MSW}} = \frac{\frac{\text{SQB}}{p-1}}{\frac{\text{SQW}}{n-p}} = \frac{\frac{\hat{\varepsilon}_0^T \hat{\varepsilon}_0 - \hat{\varepsilon}^T \hat{\varepsilon}}{p-1}}{\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n-p}} = F \tag{41.57}$$

□

Gegeben die sehr ähnlichen Definitionen des Effektstärkemaßes η^2 und der F-Statistik der EVA ist folgendes Resultat naheliegend.

Theorem 41.5 (Effektstärkenmaß η^2 und F-Teststatistik.). *Für ein EVA-Szenario mit p Gruppen und Gesamtdatenpunktzahl n seien das Effektstärkenmaß η^2 und die F-Statistik der einfaktoriellen Varianzanalyse gegeben. Dann gilt*

$$\eta^2 = \frac{F(p-1)}{F(p-1) + (n-p)} \quad (41.58)$$

◦

Beweis. Wir halten zunächst fest, dass

$$F = \frac{SQB}{SQW} \cdot \frac{n-p}{p-1} \Leftrightarrow SQB = \frac{p-1}{n-p} \cdot SQW \cdot F. \quad (41.59)$$

Damit folgt dann

$$\begin{aligned} \eta^2 &= \frac{SQB}{SQT} \\ &= \frac{SQB}{SQB + SQW} \\ &= \frac{\frac{p-1}{n-p} \cdot SQW \cdot F}{\frac{p-1}{n-p} \cdot SQW \cdot F + SQW} \\ &= \frac{\frac{F(p-1)}{n-p} \cdot SQW}{\frac{F(p-1)}{n-p} \cdot SQW + SQW} \\ &= \frac{\frac{F(p-1)}{n-p} \cdot SQW}{\left(\frac{F(p-1)}{n-p} + 1\right) \cdot SQW} \\ &= \frac{\frac{F(p-1)}{n-p}}{\frac{F(p-1)}{n-p} + \frac{n-p}{n-p}} \\ &= \frac{\frac{F(p-1)}{n-p}}{\frac{F(p-1) + (n-p)}{n-p}} \\ &= \frac{F(p-1)}{F(p-1) + (n-p)} \end{aligned} \quad (41.60)$$

□

Intuitiv ist dabei das Verhältnis von η^2 und F-Statistik analog zum Verhältnis von Cohen's d und der T-Statistik bei Einstich- und Zweistichproben-T-Tests. Insbesondere ist die gleichzeitige Angabe von η^2 und der F-Statistik bei bekannten Gruppengrößen redundant.

41.5.3. F-Test der einfaktoriellen Varianzanalyse

Wir wollen nun den Gebrauch der F-Statistik zur Durchführung eines Tests der Nullhypothese (Gleichung 41.34) diskutieren. Wir erinnern daran, dass diese Nullhypothese intuitiv besagt, dass die Erwartungswerte über alle Level des Faktors identisch sind bzw. dass alle Effektparameter gleich Null sind. Das Verwerfen dieser Nullhypothese impliziert also, dass inferentielle Evidenz dahingehend besteht, dass zumindest ein Effektparameter von Null verschieden ist. Allerdings impliziert das Verwerfen der Nullhypothese des F-Tests der einfaktoriellen Varianzanalyse keine Aussage darüber, um welchen Effektparameter es

sich dabei genau handelt. Wir definieren den F-Test der einfaktoriellen Varianzanalyse wie folgt.

Definition 41.4 (F-Test der einfaktoriellen Varianzanalyse). Gegeben sei das EVA-Modell in Effektparameterdarstellung mit Referenzgruppe sowie die Null- und Alternativhypothesen

$$\Theta_0 := \left\{ \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} \in \mathbb{R}^p \mid \alpha_i = 0 \text{ für } i = 2, \dots, p \right\} \text{ und } \Theta_1 := \mathbb{R}^p \setminus \Theta_0. \quad (41.61)$$

Weiterhin sei die F-Teststatistik definiert durch

$$F = \frac{\text{MSB}}{\text{MSW}} \quad (41.62)$$

mit der Mean-Between-Sum-of-Squares MSB und der Mean-Within-Sum-of-Squares definiert wie in Theorem 41.4. Dann ist der F-Test der einfaktoriellen Varianzanalyse definiert als der kritische Wert-basierte Test

$$\phi(y) := 1_{\{F \geq k\}} := \begin{cases} 1 & F \geq k \\ 0 & F < k \end{cases} \quad (41.63)$$

•

Der Kontrolle der Typ-I-Fehlerwahrscheinlichkeit liegt die f -Verteilung der F-Statistik zugrunde. Dies ist die Kernaussage folgenden Theorems.

Theorem 41.6 (Testumfangkontrolle des F-Tests der einfaktoriellen Varianzanalyse). ϕ sei der F-Test der einfaktoriellen Varianzanalyse. Dann ist ϕ ein Level- α_0 -Test mit Testumfang α_0 , wenn der kritische Wert definiert ist durch

$$k_{\alpha_0} := \varphi^{-1}(1 - \alpha_0; p - 1, n - p), \quad (41.64)$$

wobei $\varphi^{-1}(\cdot; p - 1, n - p)$ die inverse KVF der f -Verteilung mit Freiheitsgradparametern $p - 1$ und $n - p$ ist.

◦

Beweis. Die Testgütefunktion des betrachteten Tests im vorliegenden Testszenario ist definiert als

$$q : \mathbb{R} \rightarrow [0, 1], \beta \mapsto q_\phi(\beta) := \mathbb{P}_\beta(\phi = 1). \quad (41.65)$$

Wir haben in Kapitel 39 gesehen, dass die F-Statistik für $p_1 = p - 1$ nach einer nichtzentralen f -Verteilung verteilt ist,

$$F \sim f(\delta, p - 1, n - p). \quad (41.66)$$

Weiterhin ist der Ablehnungsbereich des hier betrachteten Tests gegeben als $[k, \infty[$. Für die funktionale Form der Testgütefunktion ergibt sich also

$$\begin{aligned} \mathbb{P}_\beta(\phi = 1) &= \mathbb{P}_\beta(F \in [k, \infty[) \\ &= \mathbb{P}_\beta(F \geq k) \\ &= 1 - \mathbb{P}_\beta(F \leq k) \\ &= 1 - \varphi(k; \delta, p - 1, n - p), \end{aligned} \quad (41.67)$$

wobei $\varphi(k; \delta, p-1, n-p)$ den Wert der KVF der nichtzentralen f -Verteilung an der Stelle k und mit Nichtzentralitätsparameter δ sowie Freiheitsgradparametern $p-1$ und $n-p$ bezeichnet (vgl. sec-f-statistik). Damit der betrachtete Test ein Level- α_0 -Test ist, muss bekanntlich gelten, dass

$$q_\phi(\beta) \leq \alpha_0 \text{ für alle } \beta \in \Theta_0 \text{ mit } \Theta_0 = \mathbb{R} \times \{0_{p-1}\}. \quad (41.68)$$

Mit der Form des Nichtzentralitätsparameters gegeben durch (vgl. Theorem 39.3)

$$\delta = \frac{1}{\sigma^2} c^T \beta \left(c^T (X^T X)^{-1} c \right)^{-1} c^T \beta \quad (41.69)$$

folgt mit $\beta \in \Theta_0$ aus

$$c = \begin{pmatrix} 0 \\ 1_{p-1} \end{pmatrix} \in \mathbb{R}^p \text{ und } \beta = \begin{pmatrix} \mu_0 \\ 0_{p-1} \end{pmatrix} \in \mathbb{R}^p \quad (41.70)$$

dann aber $\delta = 0$ und somit

$$q_\phi(\beta) = 1 - \varphi(k; p-1, n-p) \text{ für alle } \beta \in \Theta_0. \quad (41.71)$$

wobei $\varphi(k; p-1, n-p)$ den Wert der KVF der f -Verteilung an der Stelle k mit Freiheitsgradparametern $p-1$ und $n-p$ bezeichnet. Der Testumfang des betrachteten Tests ergibt sich als

$$\alpha = \max_{\beta \in \Theta_0} q_\phi(\beta) = 1 - \varphi(k; p-1, n-p) \quad (41.72)$$

da $q_\phi(\beta)$ für $\beta \in \Theta_0$ nicht von μ_0 abhängt. Wir müssen also lediglich zeigen, dass die Wahl von k_{α_0} wie im Theorem garantiert, dass ϕ den Testumfang α_0 hat. Sei also $k := k_{\alpha_0}$. Dann gilt für alle $\beta \in \Theta_0$

$$q_\phi(\beta) = 1 - \varphi(\varphi^{-1}(1 - \alpha_0; p-1, n-p); p-1, n-p) = 1 - 1 - \alpha_0 = \alpha_0 \quad (41.73)$$

und damit ist alles gezeigt. $\square$

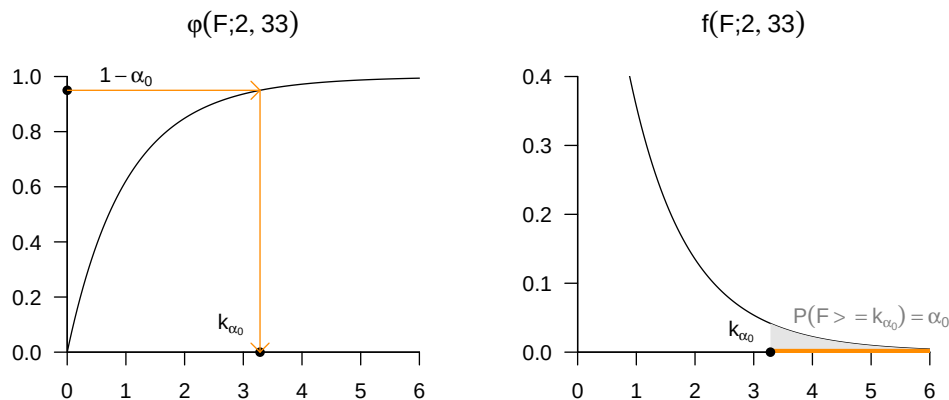


Abbildung 41.2. Exemplarischer kritischer Wert und Ablehnungsbereich bei einem F-Test der einfaktoriellen Varianzanalyse.

In Abbildung 41.2 visualisieren wir die Wahl des kritischen Wertes k_{α_0} zur Kontrolle des Testumfangs mithilfe der KVF der f -Verteilung sowie den resultierenden Ablehnungsbereich für $\alpha_0 := 0.05$, einem Faktor mit drei Levels $p = 3$ und einem balancierten Design mit Gruppenumfang $m = 12$, also $n = 3 \cdot 12 = 36$. In diesem Fall würde die Nullhypothese bei einem Wert der F-Statistik von größer als 3.28 verworfen werden.

p-Wert

Es sei f ein vorliegender Wert der F -Statistik im Kontext einer einfaktoriellen Varianzanalyse. Dann ergibt sich der zu f gehörige p-Wert anhand folgenden Theorems.

Theorem 41.7 (p-Wert der F-Statistik bei einfaktorieller Varianzanalyse). *Gegeben sei das in Definition 41.4 spezifizierte Szenario eines F-Tests bei einfaktorieller Varianzanalyse. Dann ergibt sich der zu einem vorliegenden Wert f der F-Statistik zugehörige p-Wert als*

$$\text{p-Wert} = \mathbb{P}(F \geq f) = 1 - \varphi(f; p-1, n-p) \quad (41.74)$$

◦

Beweis. Nach Definition ist der p-Wert das kleinste Signifikanzlevel α_0 , bei welchem man die Nullhypothese basierend auf einem vorliegenden Wert der Teststatistik ablehnen würde. Bei $F = f$ würde H_0 für jedes α_0 mit $f \geq \psi^{-1}(1 - \alpha_0; p-1, n-p)$ abgelehnt werden. Für ein solches α_0 gilt aber

$$\alpha_0 \geq \mathbb{P}(F \geq f), \quad (41.75)$$

denn

$$\begin{aligned} f &\geq \psi^{-1}(1 - \alpha_0; p-1, n-p) \\ \Leftrightarrow \psi(f; p-1, n-p) &\geq \psi(\psi^{-1}(1 - \alpha_0; p-1, n-p), p-1, n-p) \\ \Leftrightarrow \psi(f; p-1, n-p) &\geq 1 - \alpha_0 \\ \Leftrightarrow \mathbb{P}(F \leq f) &\geq 1 - \alpha_0 \\ \Leftrightarrow \alpha_0 &\geq 1 - \mathbb{P}(F \leq f) \\ \Leftrightarrow \alpha_0 &\geq \mathbb{P}(F \geq f) \end{aligned} \quad (41.76)$$

Das kleinste $\alpha_0 \in [0, 1]$ mit $\alpha_0 \geq \mathbb{P}(F \geq f)$ ist dann $\alpha_0 = \mathbb{P}(F \geq f)$, also folgt

$$\text{p-Wert} = \mathbb{P}(F \geq f) = 1 - \varphi(f; p-1, n-p) \quad (41.77)$$

□

Praktisches Vorgehen

Aus dem in diesem Abschnitt Gesagten ergibt sich folgendes Vorgehen zur Evaluation eines EVA-Modells mithilfe eines F-Tests: Man nimmt zunächst einmal an, dass ein vorliegender Datensatz von p Gruppendatensätzen

$$y_{11}, \dots, y_{1n_1}, y_{21}, \dots, y_{2n_2}, \dots, y_{p1}, \dots, y_{pn_p} \quad (41.78)$$

Realisationen von

$$y_{1j} \sim N(\mu_0, \sigma^2) \quad (41.79)$$

und

$$y_{ij} \sim N(\mu_0 + \alpha_i, \sigma^2) \quad (41.80)$$

für $i = 2, \dots, p$ mit wahren, aber unbekannten, Parametern $\mu_0, \alpha_i, i = 2, \dots, p$ und $\sigma^2 > 0$ sind. Weiterhin nimmt man an, dass man entscheiden möchte, ob die Nullhypothese identischer wahrer, aber unbekannter, Erwartungswertparameter über Faktorlevel bzw. zu Null identischer wahrer, aber unbekannter, Effektparameter eher zutrifft oder eher nicht (vgl. Gleichung 41.34). Dazu wählt man zunächst ein Signifikanzniveau α_0 und bestimmt den zugehörigen Freiheitsgradparameter-abhängigen kritischen Wert k_{α_0} . Zum Beispiel gilt bei Wahl von $\alpha_0 := 0.05, p = 3, m = 12, i = 1, 2, 3$ und somit $n = 36$, dass $k_{\alpha_0} = \varphi^{-1}(1 - 0.05; 2, 33) \approx 3.28$ ist. Man nutzt dann den vorliegenden Datensatz zur Berechnung der MSB und MSW und bestimmt damit den realisierten Wert der F-Statistik f . Wenn f größer gleich k_{α_0} ist, lehnt man die Nullhypothese ab, andernfalls nicht. Insgesamt garantiert die in diesem Abschnitt entwickelte Theorie dann, dass man im Mittel in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt. Schließlich mag man den zu dem Wert f assoziierten p-Wert als $1 - \varphi(f; p-1, n-p)$ bestimmen und in der Dokumentation der Analyse vermerken.

41.5.4. Anwendungsbeispiel

Folgender **R** Code implementiert das obige praktische Vorgehen für den Beispieldatensatz.

```
# Datenmanagement
D      = read.csv("./_data/508-einfaktorielle-varianzanalyse.csv")
y      = D$dBDI
n      = length(y)
p      = 3
m      = n/p

# Datensatz
# Datenvektor
# Gesamtdatenumfang
# Anzahl Gruppen
# Anzahl Datenpunkte pro Gruppe

# Modellformulierung
Xt      = cbind(
  matrix(1, nrow = n, ncol = 1),
  kronecker(diag(p),
            matrix(1, nrow = m, ncol = 1)))
# Designmatrix vollständiges Modell

X      = Xt[,-2]
X_0    = X[,1]
# Designmatrix reduziertes Modell

# F-Teststatistikevaluation
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
beta_hat_0 = solve(t(X_0) %*% X_0) %*% t(X_0) %*% y
eps_hat   = y - X %*% beta_hat
eps_hat_0 = y - X_0 %*% beta_hat_0
SQT       = t(eps_hat_0) %*% eps_hat_0
SQW       = t(eps_hat) %*% eps_hat
SQB       = SQT - SQW
DFB       = p - 1
DFW       = n - p
DFB       = p - 1
MSB       = SQB/DFB
MSW       = SQW/DFW
Eff       = MSB/MSW
pW       = 1 - pf(Eff, p-1, n-p)

# Betaparameterschätzer vollständiges Modell
# Betaparameterschätzer reduziertes Modell
# Residuenvektor vollständiges Modell
# Residuenvektor reduziertes Modell
# Sum of Squares Total
# Sum of Squares Within
# Sum of Squares Between
# Between Degrees of Freedom
# Within Degrees of Freedom
# Between Degrees of Freedom
# Mean Sum of Squares Between
# Mean Sum of Squares Within
# F-Teststatistik
# p-Wert

# F-Test Evaluation
alpha_0  = 0.05
k_alpha_0 = qf(1 - alpha_0, p-1, n-p)
if(Eff > k_alpha_0){phi = 1} else {phi = 0}
# Signifikanzlevel
# kritischer Wert
# Testwert
```

```
DFB : 2
DFW : 33
SQB : 821.72
SQW : 170.58
MSB : 410.86
MSW : 5.17
F : 79.48
p : 0
phi : 1 2
```

Im vorliegenden Fall würde die Nullhypothese

$$\Theta_0 := \left\{ \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \in \mathbb{R}^3 \mid \alpha_i = 0 \text{ für } i = 2, 3 \right\} \quad (41.81)$$

verworfen werden.

Folgender **R** Code demonstriert die Durchführung und Dokumentation der gleichen Analyse mithilfe der **R** Funktion `aov()`.

```
D      = read.csv("./_data/508-einfaktorielle-varianzanalyse.csv")
res.aov = aov(D$dBDI ~ D$COND, data = D)
summary(res.aov)

# Daten
# Modellformulierung und Modellschätzung
# Modellevaluation
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)
D$COND      2  821.7    410.9    79.48 2.41e-13 ***
Residuals   33  170.6       5.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

41.6. Literaturhinweise

Die Popularität varianzanalytischer Verfahren wird im Allgemeinen auf Fisher (1925a) und Fisher (1935) zurückgeführt. Everitt & Howell (2005) und Stigler (1986) geben einen kurzen und einen ausführlichen historischen Überblick, respektive.

42. Zweifaktorielle Varianzanalyse

42.1. Anwendungsszenario

Das klassische Anwendungsszenario einer zweifaktoriellen Varianzanalyse (ZVA) ist ein randomisiertes zweifaktorielles vollständig gekreuztes Studiendesign bestehend aus einer univariaten abhängigen Variable bestimmt an randomisierten experimentellen Einheiten und zwei unabhängigen Variablen, die in der Regel jeweils mindestens zweistufig sind. Im Kontext der Varianzanalyse werden die unabhängigen Variablen bekanntlich auch Faktoren und ihre Ausprägungsstufen Faktorlevel genannt. Im Falle eines vollständig gekreuzten Designs wird jedes Level eines Faktors mit allen Level des anderen Faktors kombiniert. Die Kombination zweier spezifischer Faktorlevel wird dann auch Zelle des Designs genannt. Da den Zellen eines ZVA Designs üblicherweise Gruppen von experimentellen Einheiten zugeordnet werden, werden sie auch oft Gruppen oder experimentelle Bedingungen genannt.

Spezielle zweifaktorielle Studiendesigns werden üblicherweise anhand ihrer Faktorlevel bezeichnet. Ein $I \times J$ Studiendesign und seine entsprechende $I \times J$ zweifaktorielle Varianzanalyse impliziert also, dass I Faktorlevel des ersten Faktors, im Folgenden als Faktor A bezeichnet, mit J Faktorleveln des zweiten Faktors, im Folgenden als Faktor B bezeichnet, gekreuzt werden. Folgende Tabelle gibt für einige der möglichen ZVA Designs die Bezeichnung und die Level der jeweiligen Faktoren an.

2×2 ZVA	$\Rightarrow$	Faktor A mit Level 1,2	Faktor B mit Level 1,2
2×3 ZVA	$\Rightarrow$	Faktor A mit Level 1,2	Faktor B mit Level 1,2,3
4×2 ZVA	$\Rightarrow$	Faktor A mit Level 1,2,3,4	Faktor B mit Level 1,2
3×1 ZVA	$\Rightarrow$	Faktor A mit Level 1,2,3	Faktor B mit Level 1

Generell sind 2×2 ZVA Designs sehr populär und ermöglichen es, die wesentlichen Charakteristika zweifaktorieller Varianzanalysedesigns zu diskutieren. Wir fokussieren im Folgenden daher weitgehend auf diesen Fall. In Abbildung 42.1 A visualisieren wir das konzeptuelle Design einer 2×2 ZVA, wobei wir die Zellen des Designs hier und im Folgenden mit A1B1, A1B2, A2B1 und A2B2 bezeichnen.

Abbildung 42.1 B visualisiert, dass zu jeder Zelle des Designs eine Gruppen von Datenpunkten korrespondiert. Wir nutzen dabei folgende Indexkonvention: y_{ijk} bezeichnet den Datenwert der k ten experimentellen Einheit im i ten Level von Faktor A und j ten Level von Faktor B, wobei $k = 1, \dots, n_{ij}$ und im vorliegenden Fall $i = 1, 2$ sowie $j = 1, 2$. Ist die Anzahl an Datenwerten n_{ij} in jeder Zelle identisch, so spricht man wiederum von einem balancierten Design.

Haupteffekte und Interaktionen

Intuitiv ist man in ZVA Designs an Haupteffekten und Interaktionen interessiert, die sich zunächst rein deskriptiv auf das Muster der zellspezifischen Gruppenmittelwerte beziehen.

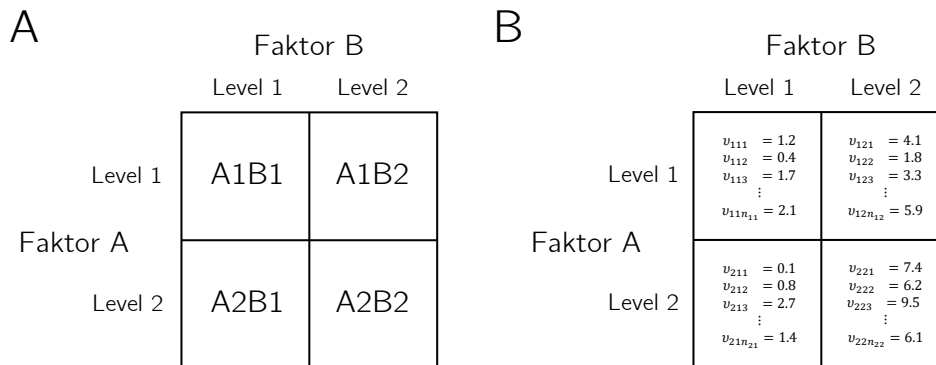


Abbildung 42.1. Konzeptuelles Design und Datennotation einer 2×2 ZVA.

Für den Fall einer 2×2 ZVA spricht man dabei intuitiv vom Vorliegen eines *Haupteffekts von Faktor A*, wenn sich die Gruppenmittelwerte zwischen Level 1 und Level 2 von Faktor A, jeweils gemittelt über die zwei Level von Faktor B, unterscheiden. Weiterhin spricht man vom Vorliegen eines *Haupteffekts von Faktor B*, wenn sich die Gruppenmittelwerte zwischen Level 1 und Level 2 von Faktor B, jeweils gemittelt über die zwei Level von Faktor A, unterscheiden.

Vom Vorliegen einer *Interaktion der Faktoren A und B* schließlich spricht man, wenn der Unterschied der Gruppenmittelwerte von Faktor A zwischen Level 1 und 2 unterschiedlich für Level 1 und Level 2 von Faktor B ausgeprägt ist bzw. wenn der Unterschied der Gruppenmittelwerte von Faktor B zwischen Level 1 und 2 unterschiedlich für Level 1 und Level 2 von Faktor A ausgeprägt ist.

Haupteffekte beziehen sich intuitiv also auf (marginale) Unterschiede (Differenzen), während sich Interaktionen auf Unterschiede von Unterschieden (Differenzen von Differenzen) beziehen. Dabei besagt das Vorliegen einer Interaktion also lediglich, dass sich die Unterschiede der Gruppenmittelwerte zwischen den Levels eines experimentellen Faktors in Abhängigkeit von den Levels des anderen experimentellen Faktors ändern, impliziert aber keine Aussage darüber, warum dies so ist. In anderen Worten sind Haupteffekte und Interaktionen bei Varianzanalysen Datenmuster, aber keine wissenschaftlichen Theorien. Die frequentistischer-inferenzstatistische Absicherung dieser Datenmuster ist das Thema dieses Kapitels.

Anwendungsbeispiel

Für ein konkretes Anwendungsbeispiel aus dem Bereich der Interventionsevaluation betrachten wir eine BDI-Pre-Postinterventionsdifferenzwertanalyse für je zwei Settings (Face-to-Face und Online) und Varianten (Mindfulness und Exercise) der kognitiven Verhaltenstherapie (CBT, Abbildung 47.1). Setting sei Faktor (A) mit Levels (1) Face-to-face (F2F) und (2) Online (ONL) und Variante sei Faktor (B) mit Levels (1) Mindfulness (MND) und

(2) Exercise (EXC). Jeder der durch Kreuzung der Faktorlevel entstehenden Interventionsbedingungen sei eine Gruppe (Stichprobe) von $n_{ij} := 12$ Patient:innen zugeordnet und für jede Patient:in sei die Differenz zwischen Preinterventions und Postinterventions BDI Wert bestimmt worden und mit dBDI bezeichnet.

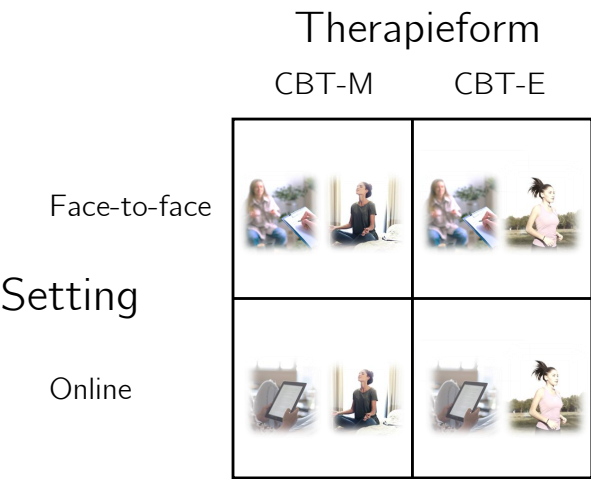


Abbildung 42.2. Konzeptuelles Design des Anwendungsbeispiels.

Tabelle 42.2. Pre-Post-BDI-Differenzwerte der Faktorlevelkombinationen

Setting	Variant	dBDI
F2F	MND	12
F2F	MND	11
F2F	MND	8
F2F	MND	8
F2F	MND	12
F2F	MND	12
F2F	MND	9
F2F	MND	10
F2F	MND	12
F2F	MND	13
F2F	MND	10
F2F	MND	9
F2F	EXC	14
F2F	EXC	11
F2F	EXC	15
F2F	EXC	16
F2F	EXC	15
F2F	EXC	19
F2F	EXC	16
F2F	EXC	14
F2F	EXC	11
F2F	EXC	15
F2F	EXC	15
F2F	EXC	17
ONL	MND	4
ONL	MND	10
ONL	MND	12
ONL	MND	13
ONL	MND	12
ONL	MND	12
ONL	MND	13
ONL	MND	10
ONL	MND	10
ONL	MND	13
ONL	MND	4
ONL	MND	8
ONL	EXC	16
ONL	EXC	19

Tabelle 42.2. Pre-Post-BDI-Differenzwerte der Faktorlevelkombinationen

Setting	Variant	dBDI
ONL	EXC	14
ONL	EXC	17
ONL	EXC	17
ONL	EXC	16
ONL	EXC	13
ONL	EXC	16
ONL	EXC	20
ONL	EXC	13
ONL	EXC	16
ONL	EXC	13

Tabelle 42.2 zeigt einen Beispieldatensatz. Dabei repräsentiert jede Zeile eine Patient:in, die Spalte Setting bezeichnet das dieser Patient:in entsprechende Level des Therapiesettingfaktors und die Spalte Variant das dieser Patient:in entsprechende Level des Therapievariantenfaktors. Der entsprechende dBDI Zeilenwert bildet den Pre-Post-InterventionsBDI-Differenzwert der entsprechenden Patient:in ab. Hohe Werte zeigen dabei eine starke Verbesserung, geringe Werte eher eine leichte Verbesserung der Depressionssymptomatik an.

Abbildung 42.3 schließlich zeigt zwei häufig genutzte Visualisationsformen für Deskriptivstatistiken (hier Gruppenmittelwerte und Standardabweichungen) in 2×2 ZVA Szenarien. Abbildung 42.3 A zeigt die faktorlevelkombinationsspezifischen Gruppenmittelwerte und ihre Standardabweichungen als Balkendiagramm. Abbildung 42.3 B zeigt die gleichen Daten als Liniendiagramm mit Fehlerbalken. Dabei entsprechen die zwei Linien den beiden Leveln des Settingfaktors F2F (dunkelgraue Linie) und ONL (hellgraue Linie). Die auf der x-Achse abgetragenen Werte dagegen entsprechen den beiden Leveln des Variantenfaktors MND und EXC. Für das Verständnis dieser Art der Visualisierung ist es sicherlich hilfreich, die Korrespondenz zwischen den in Abbildung 42.3 A und Abbildung 42.3 B einmal explizit durch nachzeichnen herzustellen.

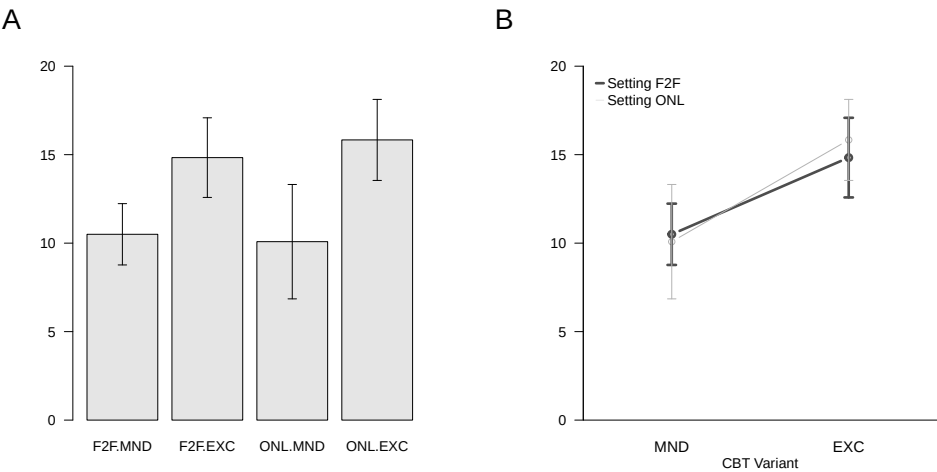


Abbildung 42.3. Visualisierung des Beispieldatensatzes als Balken- und Liniendiagramm.

42.2. Modellformulierung

Um nun den Spezialfall des ALMs für das ZVA Szenario zu diskutieren, gehen wir in zwei Schritten vor. In einem ersten Schritt führen wir mit dem *Modell der additiven ZVA* ein Modell ein, das lediglich die Haupteffekte der Faktoren, nicht aber ihre Interaktion modellieren kann. Entsprechend können auf Grundlage dieses Modells Haupteffektparameter geschätzt und inferenzstatistisch evaluiert werden. In einem zweiten Schritt führen wir mit dem Modell der ZVA mit Interaktion ein Modell ein, das sowohl die Haupteffekte der Faktoren als auch ihre Interaktion modelliert und auf dessen Grundlage entsprechend sowohl Haupteffekts- als auch Interaktionsparameter geschätzt und inferenzstatistisch evaluiert werden können. In beiden Fällen ist das zentrale Thema der Modellformulierung wie bei der EVA von einer Intuition zur Modellierung der Erwartungswertparameter der Zellen des ZVA Szenarios zu einer nicht überparameterisierten Darstellung mithilfe von Effektparametern zu kommen.

42.2.1. Modell der additiven ZVA

In Analogie zur EVA möchte man im Modell der additiven ZVA die Gruppenerwartungswerte μ_{ij} mit $i = 1, \dots, I$ für die Level von Faktor A und $j = 1, \dots, J$ für die Level von Faktor B als Summe eines gruppenunspezifischen Erwartungswertes und den Effekten der Level von Faktor A und der Level von Faktor B modellieren.

Wir bezeichnen dabei den gruppenunspezifischen Erwartungswertparameter mit μ_0 , den Effekt von Level i von Faktor A mit α_i und den Effekt von Level j von Faktor B mit β_j (β_j bezeichnet hier also nicht den j ten Eintrag des Betaparametervektors). Dann ergibt sich zum Beispiel für $I := J := 2$

$$\begin{array}{c|c} \mu_{11} := \mu_0 + \alpha_1 + \beta_1 & \mu_{12} := \mu_0 + \alpha_1 + \beta_2 \\ \hline \mu_{21} := \mu_0 + \alpha_2 + \beta_1 & \mu_{22} := \mu_0 + \alpha_2 + \beta_2 \end{array} \quad (42.1)$$

Wie im Falle der EVA ist diese Darstellung der Gruppenerwartungswerte μ_{ij} allerdings überparameterisiert, speziell werden die vier Erwartungswertparameter $\mu_{11}, \mu_{12}, \mu_{21}$ und μ_{22} in obiger Formulierung durch die fünf Parameter $\mu_0, \alpha_1, \alpha_2, \beta_1$ und β_2 dargestellt. Um eine eindeutige Darstellung der μ_{ij} zu gewährleisten, bietet sich auch hier die Restriktion an, den Effekt des ersten Levels jedes Faktors als Null zu definieren

$$\alpha_1 := \beta_1 := 0 \quad (42.2)$$

und damit die Faktorlevelkombination A1B1 als Referenzgruppe zu etablieren. Es ergibt sich somit zum Beispiel für $I := J := 2$

$$\begin{array}{c|c} \mu_{11} := \mu_0 & \mu_{12} := \mu_0 + \beta_2 \\ \hline \mu_{21} := \mu_0 + \alpha_2 & \mu_{22} := \mu_0 + \alpha_2 + \beta_2 \end{array} \quad (42.3)$$

Hier werden nun also die vier Erwartungswertparameter $\mu_{11}, \mu_{12}, \mu_{21}$ und μ_{22} durch nur drei Effektparameter μ_0, α_2 und β_2 dargestellt (in diesem Sinne ist das Modell der additiven ZVA sogar unterparameterisiert, was im folgenden Abschnitt die Einführung eines Interaktionsparameters erlaubt). Allerdings ändern sich bei dieser *Effektdarstellung des*

Modells der additiven 2×2 ZVA mit Referenzgruppe analog zur EVA die Interpretationen der Parameter μ_0, α_2, β_2 im Vergleich zum überparameterisierten Fall ohne Referenzgruppe: μ_0 entspricht dem Erwartungswert der Faktorlevelkombination A1B1, α_2 der Differenz beim Übergang von Level 1 zu Level 2 von Faktor A und β_2 der Differenz beim Übergang von Level 1 zu Level 2 von Faktor B. Wir fassen obige Überlegungen in folgender Definition zusammen.

Definition 42.1 (Modell der additiven ZVA mit Referenzgruppe). y_{ijk} mit $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij}$ sei die Zufallsvariable, die den k ten Datenpunkt zum i ten Level von Faktor A und dem j ten Level von Faktor B in einem ZVA Anwendungsszenario modelliert. Dann hat das Modell der additiven ZVA mit Referenzgruppe die strukturelle Form

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij} \quad (42.4)$$

und die Datenverteilungsform

$$y_{ijk} \sim N(\mu_{ij}, \sigma^2) \text{ u.v. für } i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij} \quad (42.5)$$

mit

$$\mu_{ij} := \mu_0 + \alpha_i + \beta_j \text{ für } i = 1, \dots, I, j = 1, \dots, J \text{ mit } \alpha_1 := \beta_1 := . \quad (42.6)$$

und $\sigma^2 > 0$.

•

Wir verzichten in Definition 42.1 auf die Angabe einer allgemeinen Designmatrixform, die wir im Folgenden lediglich für den 2×2 ZVA Spezialfall betrachten.

Die Expressivität des Modells der additiven ZVA ist, wie schon oben betont, auf das Abbilden von Haupteffekten beschränkt. Wir verdeutlichen dies in folgenden Parameterbeispielen und der zugehörigen fig-zva-add-beispiele.

Beispiel (1)

Es seien $\mu_0 := 1, \alpha_2 := 1$ und $\beta_2 := 0$, der Effektparameter für den Haupteffekt von Faktor A sei also von Null verschieden, der Effektparameter für den Haupteffekt von Faktor B dagegen gleich Null. Dann ergibt sich für die Gruppenerwartungswerte

$$\begin{array}{l|l} \mu_{11} = \mu_0 + \alpha_1 + \beta_1 = 1 + 0 + 0 = 1 & \mu_{12} = \mu_0 + \alpha_1 + \beta_2 = 1 + 0 + 0 = 1 \\ \mu_{21} = \mu_0 + \alpha_2 + \beta_1 = 1 + 1 + 0 = 2 & \mu_{22} = \mu_0 + \alpha_2 + \beta_2 = 1 + 1 + 0 = 2 \end{array} \quad (42.7)$$

fig-zva-add-beispiele A zeigt Visualisierungen dieses Erwartungswertmusters in der Balkendiagramm- und Liniendiagrammform von Abbildung 42.3.

Beispiel (2)

Es seien $\mu_0 := 1, \alpha_2 := 0$ und $\beta_2 := 1$, der Effektparameter für den Haupteffekt von Faktor A sei also gleich Null, der Effektparameter für den Haupteffekt von Faktor B dagegen von Null verschieden. Dann ergibt sich für die Gruppenerwartungswerte

$$\begin{array}{c|c} \mu_{11} = \mu_0 + \alpha_1 + \beta_1 = 1 + 0 + 0 = 1 & \mu_{12} = \mu_0 + \alpha_1 + \beta_2 = 1 + 0 + 1 = 2 \\ \mu_{21} = \mu_0 + \alpha_2 + \beta_1 = 1 + 0 + 0 = 1 & \mu_{22} = \mu_0 + \alpha_2 + \beta_2 = 1 + 0 + 1 = 2 \end{array} \quad (42.8)$$

fig-zva-add-beispiele B zeigt Visualisierungen dieses Erwartungswertmusters in der Balkendiagramm- und Liniendiagrammform von Abbildung 42.3.

Beispiel (3)

Es seien $\mu_0 := 1, \alpha_2 := 1$ und $\beta_2 := 1$, die Effektparameter für die Haupteffekte beider Faktoren seien also von Null verschieden. Dann ergibt sich für die Gruppenerwartungswerte

$$\begin{array}{c|c} \mu_{11} = \mu_0 + \alpha_1 + \beta_1 = 1 + 0 + 0 = 1 & \mu_{12} = \mu_0 + \alpha_1 + \beta_2 = 1 + 0 + 1 = 2 \\ \mu_{21} = \mu_0 + \alpha_2 + \beta_1 = 1 + 1 + 0 = 2 & \mu_{22} = \mu_0 + \alpha_2 + \beta_2 = 1 + 1 + 1 = 3 \end{array} \quad (42.9)$$

Abbildung 42.4 C zeigt Visualisierungen dieses Erwartungswertmusters in der Balkendiagramm- und Liniendiagrammform von Abbildung 42.3.

pdf

2

In Hinblick auf die Liniendiagrammform der Gruppenerwartungswerte der additiven ZVA fällt auf, dass die den Levels von Faktor A entsprechenden Linien immer parallel sind.

Für den Fall der 2×2 ZVA gibt Definition 42.2 abschließend das Modell der additiven ZVA mit Referenzgruppe inklusive seiner Designmatrixform an.

Definition 42.2 (Modell der additiven 2×2 ZVA mit Referenzgruppe). y_{ijk} mit $i = 1, 2, j = 1, 2, k = 1, \dots, n_{ij}$ sei die Zufallsvariable, die den k ten Datenpunkt zum i ten Level von Faktor A und dem j ten Level von Faktor B in einem 2×2 ZVA Anwendungsszenario modelliert. Dann hat mit $\sigma^2 > 0$ das *Modell der additiven 2×2 ZVA mit Referenzgruppe* die strukturelle Form

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \text{ mit } \varepsilon_{ijk} \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, 2, j = 1, 2, k = 1, \dots, n_{ij}, \quad (42.10)$$

die Datenverteilungsform

$$y_{ijk} \sim N(\mu_{ij}, \sigma^2) \text{ u.v. für } i = 1, 2, j = 1, 2, k = 1, \dots, n_{ij} \quad (42.11)$$

mit

$$\mu_{ij} := \mu_0 + \alpha_i + \beta_j \text{ für } i = 1, 2, j = 1, 2 \text{ mit } \alpha_1 := \beta_1 := 0, \quad (42.12)$$

sowie die Designmatrixform

$$y \sim N(X\beta, \sigma^2 I_n) \quad (42.13)$$

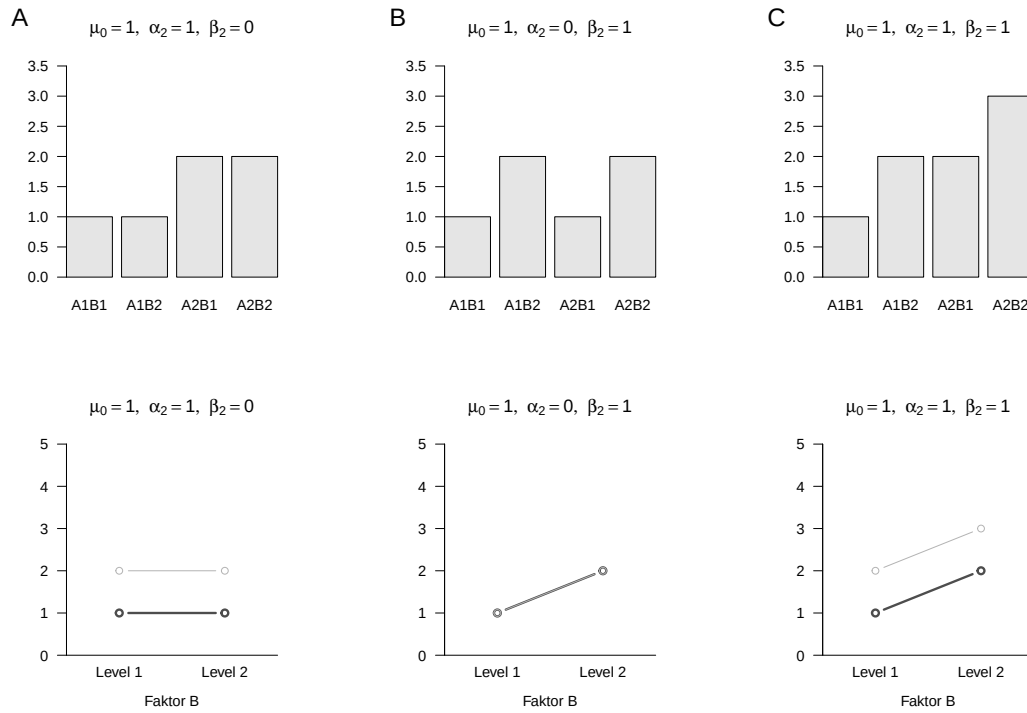


Abbildung 42.4. Gruppenerwartungswerte der Beispielparameterszenarien in Balken- und Liniendiagrammform im Modell der additiven ZVA.

wobei

$$y := \begin{pmatrix} y_{111} \\ \vdots \\ y_{11n_{11}} \\ y_{121} \\ \vdots \\ y_{12n_{12}} \\ y_{211} \\ \vdots \\ y_{21n_{21}} \\ y_{221} \\ \vdots \\ y_{22n_{22}} \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{n \times 3}, \beta := \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \beta_2 \end{pmatrix} \in \mathbb{R}^3 \text{ und } \sigma^2 > 0. \quad (42.14)$$

42.2.2. Modell der ZVA mit Interaktion

In der ZVA mit Interaktion möchte man die Gruppenerwartungswerte μ_{ij} mit $i = 1, \dots, I$ für die Level von Faktor A und $j = 1, \dots, J$ für die Level von Faktor B als Summe eines

gruppenunspezifischen Erwartungswertes, der Effekte der Level von Faktor A und Faktor B und der Interaktion der Level der Faktoren modellieren. Wir bezeichnen dazu den gruppenunspezifischen Erwartungswertparameter mit μ_0 , den Effekt von Level i von Faktor A mit α_i , den Effekt von Level j von Faktor B mit β_j , und die Interaktion von Level i von Faktor A mit Level j von Faktor B mit γ_{ij} . Dann ergibt sich zum Beispiel für $I := J := 2$

$$\begin{array}{c|c} \mu_{11} := \mu_0 + \alpha_1 + \beta_1 + \gamma_{11} & \mu_{12} := \mu_0 + \alpha_1 + \beta_2 + \gamma_{12} \\ \hline \mu_{21} := \mu_0 + \alpha_2 + \beta_1 + \gamma_{21} & \mu_{22} := \mu_0 + \alpha_2 + \beta_2 + \gamma_{22} \end{array} \quad (42.15)$$

Wie in der additiven ZVA ist diese Darstellung der Gruppenerwartungswerte μ_{ij} multipel überparameterisiert, speziell werden die vier Erwartungswertparameter $\mu_{11}, \mu_{12}, \mu_{21}$ und μ_{22} in obiger Formulierung durch die neun Parameter $\mu_0, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_{11}, \gamma_{12}, \gamma_{21}$ und γ_{22} dargestellt. Um eine eindeutige Darstellung der μ_{ij} zu gewährleisten, bietet sich auch hier die Restriktion an, den Effekt des ersten Levels jedes Faktors und jeder Interaktion als Null zu definieren

$$\alpha_1 := \beta_1 := \gamma_{11} := \gamma_{1j} := 0 \text{ für } i = 1, \dots, I, j = 1, \dots, J \quad (42.16)$$

und damit die Faktorlevelkombination A1B1 als Referenzgruppe zu etablieren. Es ergibt sich somit zum Beispiel für $I := J := 2$:

$$\begin{array}{c|c} \mu_{11} := \mu_0 & \mu_{12} := \mu_0 + \beta_2 \\ \hline \mu_{21} := \mu_0 + \alpha_2 & \mu_{22} := \mu_0 + \alpha_2 + \beta_2 + \gamma_{22} \end{array} \quad (42.17)$$

Hier werden nun also die vier Erwartungswertparameter $\mu_{11}, \mu_{12}, \mu_{21}$ und μ_{22} durch die vier Effektparameter μ_0, α_2, β_2 und γ_{22} dargestellt. Natürlich ändern sich auch bei dieser Effektdarstellung des Modells der 2×2 ZVA mit Interaktion und Referenzgruppe die Interpretationen der Parameter $\mu_0, \alpha_2, \beta_2, \gamma_{22}$ im Vergleich zum überparameterisierten Fall ohne Referenzgruppe: μ_0 entspricht dem Erwartungswert der Faktorlevelkombination A1B1, α_2 der Differenz beim Übergang von Level 1 zu Level 2 von Faktor A, β_2 der Differenz beim Übergang von Level 1 zu Level 2 von Faktor B und γ_{22} der Differenz beim Übergang von Level 1 zu Level 2 von Faktor B im Unterschied zum Übergang von Level 1 zu Level 2 von Faktor A. Wir fassen obige Überlegungen in folgender Definition zusammen, wobei wir wiederum auf die Angabe einer allgemeinen Designmatrixform verzichten wollen.

Definition 42.3 (Modell der ZVA mit Interaktion und Referenzgruppe). y_{ijk} mit $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij}$ sei die Zufallsvariable, die den k ten Datenpunkt zum i ten Level von Faktor A und dem j ten Level von Faktor B in einem ZVA Anwendungsszenario modelliert. Dann hat das Modell der ZVA mit Interaktion und Referenzgruppe die strukturelle Form

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij} \quad (42.18)$$

und die Datenverteilungsform

$$y_{ijk} \sim N(\mu_{ij}, \sigma^2) \text{ u.i.v. für } i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij} \quad (42.19)$$

mit

$$\mu_{ij} := \mu_0 + \alpha_i + \beta_j + \gamma_{ij} \quad (42.20)$$

sowie

$$\alpha_1 := \beta_1 := \gamma_{i1} := \gamma_{1j} := 0 \text{ für } i = 1, \dots, I, j = 1, \dots, J \quad (42.21)$$

und $\sigma^2 > 0$.

•

Die Expressivität des Modells der ZVA mit Interaktion entspricht nun dem Betrachten von Haupteffekten und Interaktion wie einleitend zum Anwendungsszenario der ZVA diskutiert. Wir wollen dies mithilfe folgender Parameterbeispiele und der zugehörigen Abbildung 42.5 verdeutlichen.

Beispiel (1)

Es seien $\mu_0 := 1, \alpha_2 := 0, \beta_2 := 0$ und $\gamma_{22} = 2$, die Effektparameter für die Haupteffekte beider Faktoren seien also gleich Null, aber der Effektparameter für die Interaktion der Faktoren positiv. Dann ergibt sich für die Gruppenerwartungswerte

$$\begin{array}{l|l} \mu_{11} = \mu_0 + \alpha_1 + \beta_1 + \gamma_{11} = 1 + 0 + 0 + 0 = 1 & \mu_{12} = \mu_0 + \alpha_1 + \beta_2 + \gamma_{12} = 1 + 0 + 0 + 0 = 1 \\ \mu_{21} = \mu_0 + \alpha_2 + \beta_1 + \gamma_{21} = 1 + 0 + 0 + 0 = 1 & \mu_{22} = \mu_0 + \alpha_2 + \beta_2 + \gamma_{22} = 1 + 0 + 0 + 2 = 3 \end{array} \quad (42.22)$$

Abbildung 42.5 A zeigt Visualisierungen dieses Erwartungswertmusters in Balkendiagramm- und Liniendiagrammform.

Beispiel (2)

Es seien $\mu_0 := 1, \alpha_2 := 1, \beta_2 := 1$ und $\gamma_{22} = -2$. Die Effektparameter für die Haupteffekte beider Faktoren seien also gleich und der Effektparameter für die Interaktion der Faktoren negativ. Dann ergibt sich für die Gruppenerwartungswerte

$$\begin{array}{l|l} \mu_{11} = \mu_0 + \alpha_1 + \beta_1 + \gamma_{11} = 1 + 0 + 0 + 0 = 1 & \mu_{12} = \mu_0 + \alpha_1 + \beta_2 + \gamma_{12} = 1 + 0 + 1 + 0 = 2 \\ \mu_{21} = \mu_0 + \alpha_2 + \beta_1 + \gamma_{21} = 1 + 0 + 1 + 0 = 2 & \mu_{22} = \mu_0 + \alpha_2 + \beta_2 + \gamma_{22} = 1 + 1 + 1 - 2 = 1 \end{array} \quad (42.23)$$

Abbildung 42.5 B zeigt Visualisierungen dieses Erwartungswertmusters in Balkendiagramm- und Liniendiagrammform.

Beispiel (3)

Es seien $\mu_0 := 1, \alpha_2 := 1, \beta_2 := 0$ und $\gamma_{22} = 1$, der Effektparameter für Faktor A sei also positiv, der Effektparameter für Faktor B gleich Null und der Effektparameter für die Interaktion der Faktoren positiv. Dann ergibt sich für die Gruppenerwartungswerte

$$\begin{array}{l|l} \mu_{11} = \mu_0 + \alpha_1 + \beta_1 + \gamma_{11} = 1 + 0 + 0 + 0 = 1 & \mu_{12} = \mu_0 + \alpha_1 + \beta_2 + \gamma_{12} = 1 + 0 + 0 + 0 = 1 \\ \mu_{21} = \mu_0 + \alpha_2 + \beta_1 + \gamma_{21} = 1 + 1 + 0 + 0 = 2 & \mu_{22} = \mu_0 + \alpha_2 + \beta_2 + \gamma_{22} = 1 + 1 + 0 + 1 = 3 \end{array} \quad (42.24)$$

Abbildung 42.5 C zeigt Visualisierungen dieses Erwartungswertmusters in Balkendiagramm- und Liniendiagrammform.

Beispiel (4)

Es seien $\mu_0 := 1, \alpha_2 := 0, \beta_2 := 1$ und $\gamma_{22} = 1$, der Effektparameter für Faktor A sei also gleich Null, der Effektparameter für Faktor B positiv und der Effektparameter für die Interaktion der Faktoren wiederum positiv. Dann ergibt sich für die Gruppenerwartungswerte

$$\begin{array}{l|l} \mu_{11} = \mu_0 + \alpha_1 + \beta_1 + \gamma_{11} = 1 + 0 + 0 + 0 = 1 & \mu_{12} = \mu_0 + \alpha_1 + \beta_2 + \gamma_{12} = 1 + 0 + 1 + 0 = 2 \\ \mu_{21} = \mu_0 + \alpha_2 + \beta_1 + \gamma_{21} = 1 + 0 + 0 + 0 = 1 & \mu_{22} = \mu_0 + \alpha_2 + \beta_2 + \gamma_{22} = 1 + 0 + 1 + 1 = 3 \end{array} \quad (42.25)$$

Abbildung 42.5 D zeigt Visualisierungen dieses Erwartungswertmusters in Balkendiagramm- und Liniendiagrammform.

Die obigen Beispiele zeigen, dass die Expressivität der ZVA mit Interaktion im Sinne der durch das Modell abzubildenden Datenmuster ungleich höher als die Expressivität des rein additiven Modells der ZVA ist. Insbesondere bilden obige vier Beispiele bei Weitem nicht den gesamten kombinatorischen qualitativen Parameterraum der ZVA mit Interaktion ab. Hinsichtlich der Liniendiagrammvisualisation der Gruppenerwartungswertparameter bei ZVA mit Interaktion fällt auf, dass ein von Null verschiedener Interaktionsterm immer impliziert, dass die die Level von Faktor A repräsentierenden Linien immer nicht parallel verlaufen.

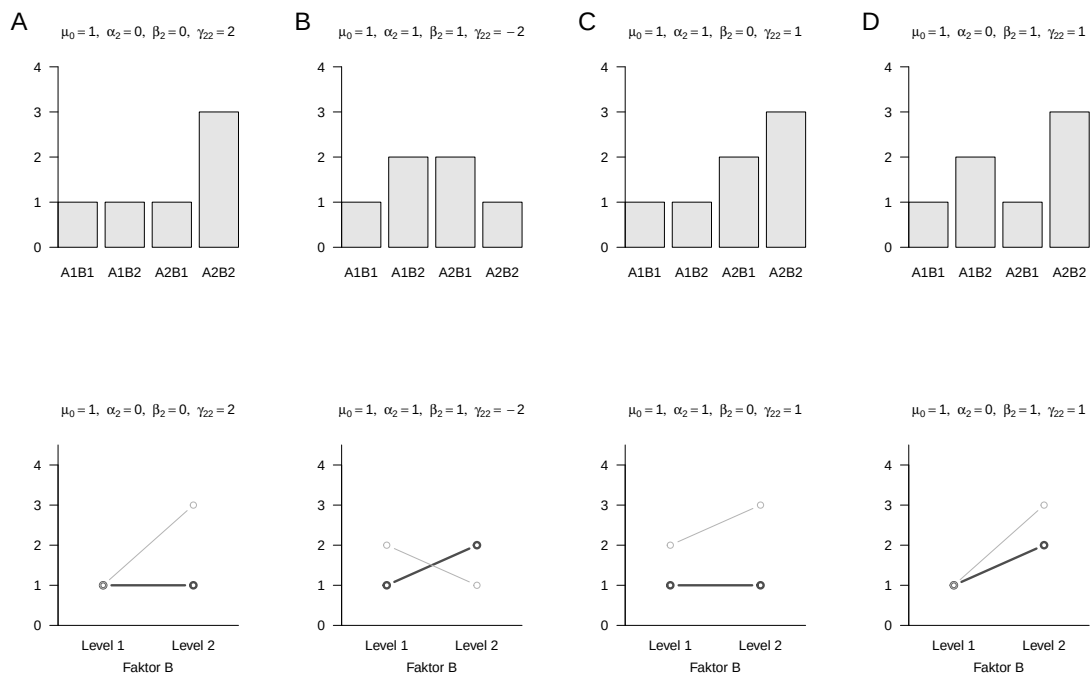


Abbildung 42.5. Gruppenerwartungswerte der Beispielparameterszenarien in Balken- und Liniendiagrammform im Modell der ZVA mit Interaktion.

Abschließend geben wir in Definition 42.4 den Spezialfall des Modells der 2×2 ZVA mit Interaktion und Referenzgruppe inklusiver seiner Designmatrixform an.

Definition 42.4 (Modell der 2×2 ZVA mit Interaktion und Referenzgruppe.). y_{ijk} mit $i = 1, 2, j = 1, 2, k = 1, \dots, n_{ij}$ sei die Zufallsvariable, die den k ten Datenpunkt zum i ten Level

von Faktor A und dem j ten Level von Faktor B in einem 2×2 ZVA Anwendungsszenario modelliert. Dann hat das Modell der 2×2 ZVA mit Interaktion und Referenzgruppe für $\sigma^2 > 0$ die strukturelle Form

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \text{ mit } \varepsilon_{ijk} \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, 2, j = 1, 2, k = 1, \dots, n_{ij}, \quad (42.26)$$

die Datenverteilungsform

$$y_{ijk} \sim N(\mu_{ij}, \sigma^2) \text{ u.i.v. für } i = 1, 2, j = 1, 2, k = 1, \dots, n_{ij} \quad (42.27)$$

mit

$$\mu_{ij} := \mu_0 + \alpha_i + \beta_j + \gamma_{ij} \quad (42.28)$$

sowie

$$\alpha_1 := \beta_1 := \gamma_{i1} := \gamma_{1j} := 0 \text{ für } i = 1, \dots, I, j = 1, \dots, J \quad (42.29)$$

und für $n := \sum_{i=1}^2 \sum_{j=1}^2 n_{ij}$ die Designmatrixform

$$y \sim N(X\beta, \sigma^2 I_n), \text{ mit} \quad (42.30)$$

$$y := \begin{pmatrix} y_{111} \\ \vdots \\ y_{11n_{11}} \\ y_{121} \\ \vdots \\ y_{12n_{12}} \\ y_{211} \\ \vdots \\ y_{21n_{21}} \\ y_{221} \\ \vdots \\ y_{22n_{22}} \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{n \times 4}, \beta := \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \beta_2 \\ \gamma_{22} \end{pmatrix} \in \mathbb{R}^4 \text{ und } \sigma^2 > 0. \quad (42.31)$$

•

Beispiel

Als konkretes Beispiel für die Designmatrixform des in Definition 42.4 definierten Modells betrachten wir das Szenario $n_{ij} := 4$, also $n = 16$. Dann gilt

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_{16}, \sigma^2 I_{16}) \quad (42.32)$$

mit

$$y := \begin{pmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{114} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{124} \\ y_{211} \\ y_{212} \\ y_{213} \\ y_{214} \\ y_{221} \\ y_{222} \\ y_{223} \\ y_{224} \end{pmatrix}, X := \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{16 \times 4}, \beta := \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \beta_2 \\ \gamma_{22} \end{pmatrix} \in \mathbb{R}^4 \text{ und } \sigma^2 > 0. \quad (42.33)$$

Folgender **R** Code erlaubt die Realisierung von Daten im Falle dieses Beispiels.

```
library(MASS)
I = 2
J = 2
n_ij = 4
n = I*J*n_ij
p = 1 + (I-1)+(J-1)+(I*J-3)
D = matrix(c(1,0,0,0,
             1,0,1,0,
             1,1,0,0,
             1,1,1,1),
           nrow = p,
           byrow = TRUE)
C = matrix(rep(1,n_ij),nrow = n_ij)
X = kronecker(D,C)
I_n = diag(n)
beta = matrix(c(1,1,1,1), nrow = p)
sigsqr = 10
y = mvrnorm(1, X %*% beta, sigsqr*I_n)
```

Multivariate Normalverteilung
Anzahl Level Faktor A
Anzahl Level Faktor B
Anzahl von Datenpunkten der i,jten Gruppe
Anzahl Datenpunkte
Anzahl Parameter
Prototypische Designmatrix für balancierte Designs
Prototypischer Zellenvektor für balancierte Designs
Kroneckerprodukt Designmatrix Erzeugung
n x n Einheitsmatrix
\beta = (\mu_0, \alpha_2, \alpha_3, \alpha_4)
\sigma^2
eine Realisierung eines n-dimensionalen ZVs

42.3. Modellschätzung

Wir betrachten im Folgenden die Betaparameterschätzer im additiven 2×2 ZVA Modell mit Referenzgruppe und im 2×2 ZVA mit Interaktion und Referenzgruppe. Dabei stellt sich wenig überraschend heraus, dass sich die Schätzer für die Komponenten des Betaparameters der in Definition 42.2 und Definition 42.4 formulierten Modelle als gewichtete Summen der gruppenspezifischen Stichprobenmittel ergeben. Die genaue Form dieser gewichteten Summen ist Inhalt der beiden folgenden Theoreme. Der Einfachheit halber gehen wir dabei jeweils von einem balancierten Design aus. Wir beginnen mit der Betaparameterschätzung im additiven 2×2 ZVA Modell mit Referenzgruppe.

Theorem 42.1 (Betaparameterschätzung im additiven 2×2 ZVA Modell mit Referenzgruppe). *Gegeben sei die Designmatrixform eines balancierten additiven 2×2 ZVA Modells mit Referenzgruppe. Dann ergibt sich für den Betaparameterschätzer*

$$\hat{\beta} := \begin{pmatrix} \hat{\mu}_0 \\ \hat{\alpha}_2 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \frac{3}{4}\bar{y}_{11} + \frac{1}{4}(\bar{y}_{12} + \bar{y}_{21}) - \frac{1}{4}\bar{y}_{22} \\ \frac{1}{2}(\bar{y}_{21} + \bar{y}_{22}) - \frac{1}{2}(\bar{y}_{11} + \bar{y}_{12}) \\ \frac{1}{2}(\bar{y}_{12} + \bar{y}_{22}) - \frac{1}{2}(\bar{y}_{11} + \bar{y}_{21}) \end{pmatrix}, \quad (42.34)$$

wobei

$$\bar{y}_{ij} := \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} \text{ für } 1 \leq i, j \leq 2 \quad (42.35)$$

das Stichprobenmittel der i, j ten Gruppe des 2×2 ZVA Designs bezeichnet.

◦

Beweis. Wir bestimmen zunächst $X^T y$, $X^T X$ und $(X^T X)^{-1}$ bei konstantem n_{ij} für $1 \leq i, j \leq 2$. Es ergeben sich

$$X^T y = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_{111} \\ \vdots \\ y_{11n_{11}} \\ y_{121} \\ \vdots \\ y_{12n_{12}} \\ y_{211} \\ \vdots \\ y_{21n_{21}} \\ y_{221} \\ \vdots \\ y_{22n_{22}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^2 \sum_{k=1}^2 \sum_{j=1}^{n_{ij}} y_{ijk} \\ \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} \\ \sum_{i=1}^2 \sum_{k=1}^{n_{i2}} y_{i2k} \end{pmatrix}, \quad (42.36)$$

$$X^T X = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 1 \end{pmatrix} = n_{ij} \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}, \quad (42.37)$$

sowie

$$(X^T X)^{-1} = n_{ij} \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}^{-1} = \frac{1}{n_{ij}} \begin{pmatrix} \frac{3}{4} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix}. \quad (42.38)$$

Es ergibt sich also

$$\hat{\beta} = \begin{pmatrix} \hat{\mu}_0 \\ \hat{\alpha}_2 \\ \hat{\beta}_2 \end{pmatrix} = \frac{1}{n_{ij}} \begin{pmatrix} \frac{3}{4} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} y_{ijk} \\ \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} \\ \sum_{i=1}^2 \sum_{k=1}^{n_{i2}} y_{ik2} \end{pmatrix}. \quad (42.39)$$

Damit ergibt sich dann

$$\begin{aligned} \hat{\mu}_0 &= \frac{1}{n_{ij}} \left(\frac{3}{4} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{2} \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} - \frac{1}{2} \sum_{i=1}^2 \sum_{k=1}^{n_{i2}} y_{ik2} \right) \\ &= \frac{1}{n_{ij}} \left(\frac{3}{4} \sum_{k=1}^{n_{11}} y_{11k} + \frac{3}{4} \sum_{k=1}^{n_{12}} y_{12k} + \frac{3}{4} \sum_{k=1}^{n_{21}} y_{21k} + \frac{3}{4} \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &\quad + \frac{1}{n_{ij}} \left(-\frac{1}{2} \sum_{k=1}^{n_{21}} y_{21k} - \frac{1}{2} \sum_{k=1}^{n_{22}} y_{22k} - \frac{1}{2} \sum_{k=1}^{n_{12}} y_{12k} - \frac{1}{2} \sum_{k=1}^{n_{11}} y_{11k} \right) \\ &= \frac{1}{n_{ij}} \left(\frac{3}{4} \sum_{k=1}^{n_{11}} y_{11k} + \frac{1}{4} \sum_{k=1}^{n_{12}} y_{12k} + \frac{1}{4} \sum_{k=1}^{n_{21}} y_{21k} - \frac{1}{4} \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &= \frac{3}{4} \bar{y}_{11} + \frac{1}{4} (\bar{y}_{12} + \bar{y}_{21}) - \frac{1}{4} \bar{y}_{22} \end{aligned} \quad (42.40)$$

sowie

$$\begin{aligned} \hat{\alpha}_2 &= \frac{1}{n_{ij}} \left(-\frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} y_{ijk} + \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} \right) \\ &= \frac{1}{n_{ij}} \left(-\frac{1}{2} \sum_{k=1}^{n_{11}} y_{11k} - \frac{1}{2} \sum_{k=1}^{n_{12}} y_{12k} - \frac{1}{2} \sum_{k=1}^{n_{21}} y_{21k} - \frac{1}{2} \sum_{k=1}^{n_{22}} y_{22k} + \sum_{k=1}^{n_{21}} y_{21k} + \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &= \frac{1}{n_{ij}} \left(-\frac{1}{2} \sum_{k=1}^{n_{11}} y_{11k} - \frac{1}{2} \sum_{k=1}^{n_{12}} y_{12k} + \frac{1}{2} \sum_{k=1}^{n_{21}} y_{21k} + \frac{1}{2} \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &= \frac{1}{2} (\bar{y}_{21} + \bar{y}_{22}) - \frac{1}{2} (\bar{y}_{11} + \bar{y}_{12}) \end{aligned} \quad (42.41)$$

und analog für $\hat{\beta}_2$. □

Beispiel

Wir verdeutlichen den in Theorem 42.1 formulierten Zusammenhang zwischen den Beta-parameterschätzerkomponenten und den gruppenspezifischen Stichprobenmitteln mithilfe folgenden **R** Codes am Beispieldatensatz.

```
# Datenreformatierung
D = read.csv("./_data/509-zweifaktorielle-varianzanalyse.csv")
A1B1 = D$dBDI[D$Setting == "F2F" & D$Variant == "MND" ]
A1B2 = D$dBDI[D$Setting == "F2F" & D$Variant == "EXC" ]
A2B1 = D$dBDI[D$Setting == "ONL" & D$Variant == "MND" ]
A2B2 = D$dBDI[D$Setting == "ONL" & D$Variant == "EXC" ]

# Datensatz
# Face-to-face, mindfulness
# Face-to-face, exercise
# Online, mindfulness
# Online, exercise

# Datenmatrix für Gruppenmittelwerte
n_ij = length(A1B1)
Y = matrix(c(A1B1, A1B2, A2B1, A2B2), nrow = n_ij)
bar_y = colMeans(Y)

# Anzahl Datenpunkte pro Gruppe
# Datenmatrix
# Zellenmittelwerte

# Modellschätzung
I = 2
J = 2
n = I*J*n_ij
p = 1 + (I-1)*(J-1) + (I*J-3)
D = matrix(c(1,0,0,
             1,0,1,
             1,1,0,
             1,1,1), nrow = p, byrow = TRUE)

# Anzahl Level Faktor A (Therapiesetting)
# Anzahl Level Faktor B (Therapievariante)
# Anzahl Datenpunkte
# Anzahl Parameter
# Prototypische Designmatrix für balancierte Designs

C = matrix(rep(1,n_ij),nrow = n_ij)
X = kronecker(D,C)
y = matrix(c(A1B1, A1B2, A2B1, A2B2), nrow = n)
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
eps_hat = y - X %*% beta_hat
sigsqr_hat = (t(eps_hat) %*% eps_hat) / (n-p)

# Prototypischer Zellenvektor für balancierte Designs
# Kroneckerprodukt Designmatrix
# Datenvektor
# Betaparameterschätzer
# Residuenvektor
# Varianzparameterschätzer
```

```

hat{beta} : 10.15 0.29 5.04
hat{sigsqr} : 6.07
bar{y}_11, bar{y}_12, bar{y}_21, bar{y}_22 : 10.5 14.83 10.08 15.83
3/4bar{y}_11 + 1/4(bar{y}_12 + bar{y}_21) - 1/4bar{y}_22 : 10.15
1/2(bar{y}_21 + bar{y}_22) - 1/2(bar{y}_11 + bar{y}_12) : 0.29
1/2(bar{y}_12 + bar{y}_22) - 1/2(bar{y}_11 + bar{y}_21) : 5.04

```

Für die Betaparameterschätzung im 2×2 ZVA Modell mit Interaktion und Referenzgruppe ergibt sich folgendes Theorem.

Theorem 42.2 (Betaparameterschätzung im 2×2 ZVA Modell mit Interaktion und Referenzgruppe). *Gegeben sei die Designmatrixform eines balancierten 2×2 ZVA Modells mit Interaktion und Referenzgruppe. Dann ergibt sich für den Betaparameterschätzer*

$$\hat{\beta} := \begin{pmatrix} \hat{\mu}_0 \\ \hat{\alpha}_2 \\ \hat{\beta}_2 \\ \hat{\gamma}_{22} \end{pmatrix} = \begin{pmatrix} \bar{y}_{11} \\ \bar{y}_{11} - \bar{y}_{11} \\ \bar{y}_{12} - \bar{y}_{11} \\ (\bar{y}_{11} - \bar{y}_{12}) - (\bar{y}_{21} - \bar{y}_{22}) \end{pmatrix} \quad (42.42)$$

wobei

$$\bar{y}_{ij} := \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} \text{ für } 1 \leq i, j \leq 2 \quad (42.43)$$

das Stichprobenmittel der i,j ten Gruppe des 2×2 ZVA Designs bezeichnet.

◦

Beweis. Wir bestimmen zunächst $X^T y$, $X^T X$ und $(X^T X)^{-1}$ bei konstantem n_{ij} für $1 \leq i, j \leq 2$. Es ergeben sich

$$X^T y = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 & 1 & \dots & 1 \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_{111} \\ \vdots \\ y_{11n_{11}} \\ y_{121} \\ \vdots \\ y_{12n_{12}} \\ y_{211} \\ \vdots \\ y_{21n_{21}} \\ y_{221} \\ \vdots \\ y_{22n_{22}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} y_{ijk} \\ \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} \\ \sum_{i=1}^2 \sum_{k=1}^n y_{i2k} \\ \sum_{k=1}^{n_{22}} y_{22k} \end{pmatrix}, \quad (42.44)$$

$$X^T X = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 & 1 & \dots & 1 \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \end{pmatrix} = n_{ij} \begin{pmatrix} 4 & 2 & 2 & 1 \\ 2 & 2 & 1 & 1 \\ 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad (42.45)$$

sowie

$$(X^T X)^{-1} = n_{ij} \begin{pmatrix} 4 & 2 & 2 & 1 \\ 2 & 2 & 1 & 1 \\ 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}^{-1} = \frac{1}{n_{ij}} \begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 2 & 1 & -2 \\ -1 & 1 & 2 & -2 \\ 1 & -2 & -2 & 4 \end{pmatrix} \quad (42.46)$$

Es ergibt sich also

$$\hat{\beta} = \begin{pmatrix} \hat{\mu}_0 \\ \hat{\alpha}_2 \\ \hat{\beta}_2 \\ \hat{\gamma}_{22} \end{pmatrix} = \frac{1}{n_{ij}} \begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 2 & 1 & -2 \\ -1 & 1 & 2 & -2 \\ 1 & -2 & -2 & 4 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} y_{ijk} \\ \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} \\ \sum_{i=1}^2 \sum_{k=1}^{n_{i2}} y_{i2k} \\ \sum_{k=1}^{n_{22}} y_{22k} \end{pmatrix} \quad (42.47)$$

Damit ergibt sich dann

$$\begin{aligned} \hat{\mu}_0 &= \frac{1}{n_{ij}} \left(\sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} - \sum_{i=1}^2 \sum_{k=1}^{n_{i2}} y_{i2k} + \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &= \frac{1}{n_{ij}} \left(\sum_{k=1}^{n_{11}} y_{11k} + \sum_{k=1}^{n_{12}} y_{12k} + \sum_{k=1}^{n_{21}} y_{21k} + \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &\quad + \frac{1}{n_{ij}} \left(- \sum_{k=1}^{n_{21}} y_{21k} - \sum_{k=1}^{n_{22}} y_{22k} - \sum_{k=1}^{n_{12}} y_{12k} - \sum_{k=1}^{n_{22}} y_{22k} + \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &= \frac{1}{n_{11}} \sum_{k=1}^{n_{11}} y_{11k} \\ &= \bar{y}_{11} \end{aligned} \quad (42.48)$$

sowie

$$\begin{aligned} \hat{\alpha}_2 &= \frac{1}{n_{ij}} \left(- \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} y_{ijk} + 2 \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} + 1 \sum_{i=1}^2 \sum_{k=1}^{n_{i2}} y_{i2k} - 2 \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &= \frac{1}{n_{ij}} \left(- \sum_{k=1}^{n_{11}} y_{11k} - \sum_{k=1}^{n_{12}} y_{12k} - \sum_{k=1}^{n_{21}} y_{21k} - \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &\quad + \frac{1}{n_{ij}} \left(2 \sum_{k=1}^{n_{21}} y_{21k} + 2 \sum_{k=1}^{n_{22}} y_{22k} + \sum_{k=1}^{n_{12}} y_{12k} + \sum_{k=1}^{n_{22}} y_{22k} - 2 \sum_{k=1}^{n_{22}} y_{22k} \right) \\ &= \frac{1}{n_{ij}} \left(\sum_{k=1}^{n_{21}} y_{21k} - \sum_{k=1}^{n_{11}} y_{11k} \right) \\ &= \bar{y}_{21} - \bar{y}_{11} \end{aligned} \quad (42.49)$$

und analog für $\hat{\beta}_2$. Schließlich ergibt sich

$$\begin{aligned}
 \hat{\gamma}_{22} &= \frac{1}{n_{ij}} \left(\sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} y_{ijk} - 2 \sum_{j=1}^2 \sum_{k=1}^{n_{2j}} y_{2jk} - 2 \sum_{i=1}^2 \sum_{k=1}^{n_{i2}} y_{i2k} + 4 \sum_{k=1}^{n_{22}} y_{22k} \right) \\
 &= \frac{1}{n_{ij}} \left(\sum_{k=1}^{n_{11}} y_{11k} + \sum_{k=1}^{n_{12}} y_{12k} + \sum_{k=1}^{n_{21}} y_{21k} + \sum_{k=1}^{n_{22}} y_{22k} \right) \\
 &\quad + \frac{1}{n_{ij}} \left(-2 \sum_{k=1}^{n_{21}} y_{21k} - 2 \sum_{k=1}^{n_{22}} y_{22k} - 2 \sum_{k=1}^{n_{12}} y_{12k} - 2 \sum_{k=1}^{n_{22}} y_{22k} + 4 \sum_{k=1}^{n_{22}} y_{22k} \right) \\
 &= \frac{1}{n_{ij}} \left(\sum_{k=1}^{n_{11}} y_{11k} + \sum_{k=1}^{n_{22}} y_{22k} - \sum_{k=1}^{n_{12}} y_{12k} - \sum_{k=1}^{n_{21}} y_{21k} \right) \\
 &= \bar{y}_{11} + \bar{y}_{22} - \bar{y}_{12} - \bar{y}_{21} \\
 &= (\bar{y}_{11} - \bar{y}_{12}) - (\bar{y}_{21} - \bar{y}_{22}).
 \end{aligned} \tag{42.50}$$

□

Auch hier verdeutlichen wir den in obigem Theorem formulierten Zusammenhang zwischen den Betaparameterschätzerkomponenten und den gruppenspezifischen Stichprobenmitteln anhand des Beispieldatensatzes mithilfe folgenden **R** Codes.

```

# Datenreformatierung
D = read.csv("./_data/509-zweifaktorielle-varianzanalyse.csv")
A1B1 = D$BDI[D$Setting == "F2F" & D$Variant == "MND" ]
A1B2 = D$BDI[D$Setting == "F2F" & D$Variant == "EXC" ]
A2B1 = D$BDI[D$Setting == "ONL" & D$Variant == "MND" ]
A2B2 = D$BDI[D$Setting == "ONL" & D$Variant == "EXC" ]

# Datensatz
# Face-to-face, mindfulness
# Face-to-face, exercise
# Online, mindfulness
# Online, exercise

# Datenmatrix für Gruppenmittelwerte
n_ij = length(A1B1)
Y = matrix(c(A1B1,A1B2,A2B1,A2B2), nrow = n_ij)
bar_y = colMeans(Y)

# Anzahl Datenpunkte pro Gruppe
# Datenmatrix
# Zellenmittelwerte

# Modellschätzung
I = 2
J = 2
n = I*J*n_ij
p = 1 + (I-1)+(J-1)+(I*J-3)
D = matrix(c(1,0,0,0,
             1,0,1,0,
             1,1,0,0,
             1,1,1,1), nrow = p, byrow = TRUE)

# Anzahl Level Faktor A (Therapiesetting)
# Anzahl Level Faktor B (Therapievariante)
# Anzahl Datenpunkte
# Anzahl Parameter
# Prototypische Designmatrix

C = matrix(rep(1,n_ij),nrow = n_ij)
X = kronecker(D,C)
y = matrix(c(A1B1,A1B2,A2B1,A2B2), nrow = n)
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
eps_hat = y - X %*% beta_hat
sigsqr_hat = (t(eps_hat) %*% eps_hat) / (n-p)

# Prototypischer Zellenvektor
# Kroneckerprodukt Designmatrix
# Datenvektor
# Betaparameterschätzer
# Residuenvektor
# Varianzparameterschätzer

hat{beta} : 10.5 -0.42 4.33 1.42
hat{sigsqr} : 5.94
bar{y}_11,bar{y}_12,bar{y}_21,bar{y}_22 : 10.5 14.83 10.08 15.83
bar{y}_11 : 10.5
bar{y}_21 - bar{y}_11 : -0.42
bar{y}_12 - bar{y}_11 : 4.33
bar{y}_11 + bar{y}_22 - bar{y}_12 + bar{y}_21 : 1.42

```

42.4. Modellevaluation

Die inferenzstatistische Evaluation von Haupteffekt- und Interaktionsparametern in ZVA Modellen in allgemeiner Form kann aus mindestens drei Blickwinkeln betrachtet werden. Analog zur EVA können auch für die ZVA Quadratsummenzerlegungen entwickelt werden und zur Definition entsprechender F-Statistiken eingesetzt werden. Wir folgen hier dem alternativen Ansatz, F-Statistiken für Haupteffekte und Interaktionen vor dem Hintergrund von Modellvergleichen in den entsprechenden ZVA Modellen zu formulieren. Einen integrativen Blick auf beide Ansätze bietet die Theorie der Allgemeinen Linearen Hypothese,

auf deren Diskussion wir hier aber verzichten. Dementsprechend beschränken wir uns hier auf (1) die Evaluation der Haupteffekte im additiven Modell der 2×2 ZVA mit Referenzgruppe und (2) die Evaluation der Interaktion im Modell der 2×2 ZVA mit Interaktion und Referenzgruppe. Für beide Evaluationen formulieren wir (1) die zugrundeliegenden Modelle und Teststatistiken, (2) die Testhypothesen und Tests und (3) die Prozeduren zur Testumfangkontrolle und Auswertung des p-Wertes.

Evaluation von Haupteffekten im additiven Modell der 2×2 ZVA

Theorem 42.3 (Teststatistiken für Haupteffekte). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (42.51)$$

die Designmatrixform des additiven Modells der 2×2 ZVA mit Referenzgruppe, wobei die $n \times 1$ Spalten von X bezeichnet seien durch

$$X := \begin{pmatrix} X_{\mu_0} & X_{\alpha_2} & X_{\beta_2} \end{pmatrix} \in \mathbb{R}^{n \times 3}. \quad (42.52)$$

Dann gelten

(A) Eine F-Teststatistik für den Haupteffekt von Faktor A ist die F-Statistik des ALMs mit

$$X_A := \begin{pmatrix} X_{\mu_0} & X_{\beta_2} & X_{\alpha_2} \end{pmatrix} \in \mathbb{R}^{n \times 3}, \beta_A := \begin{pmatrix} \mu_0 \\ \beta_2 \\ \alpha_2 \end{pmatrix} \in \mathbb{R}^3 \text{ und } p_0 := 2, p_1 := 1 \quad (42.53)$$

(B) Eine F-Teststatistik für den Haupteffekt von Faktor B ist die F-Statistik des ALMs mit

$$X_B := \begin{pmatrix} X_{\mu_0} & X_{\alpha_2} & X_{\beta_2} \end{pmatrix} \in \mathbb{R}^{n \times 3}, \beta_B := \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \beta_2 \end{pmatrix} \in \mathbb{R}^3 \text{ und } p_0 := 2, p_1 := 1 \quad (42.54)$$

◦

Definition 42.5 (Testhypothesen und Tests für Haupteffekte). Gegeben sei das Modell der additiven 2×2 ZVA und die F-Teststatistiken für die Haupteffekte von Faktor A und B seien mit F_A und F_B bezeichnet und wie oben definiert. Dann gelten folgende Definitionen:

(A) Der kritische Wert-basierte Test

$$\phi_A(y) := 1_{\{F_A \geq k\}} \text{ mit Nullhypothese } H_0^A : \alpha_2 = 0 \quad (42.55)$$

definiert den F-Test des Haupteffekts von Faktor A im Modell mit X_A, β_A .

(B) Der kritische Wert-basierte Test

$$\phi_B(y) := 1_{\{F_B \geq k\}} \text{ mit Nullhypothese } H_0^B : \beta_2 = 0 \quad (42.56)$$

definiert den F-Test des Haupteffekts von Faktor B im Modell mit X_B, β_B .

•

Theorem 42.4 (Testumfangkontrolle und p-Werte für Haupteffekte). *Mit obigen Definitionen und der KVF φ der f -Verteilung gelten:*

(A) ϕ_A ist ein Level- α_0 -Test mit Testumfang α_0 , wenn der kritische Wert definiert ist durch

$$k_{\alpha_0}^A := \varphi^{-1}(1 - \alpha_0; 1, n - 3). \quad (42.57)$$

Der zu einem beobachteten Wert f_A von F_A assoziierte p-Wert ist gegeben durch

$$p_A - \text{Wert} := 1 - \varphi(f_A; 1, n - 3). \quad (42.58)$$

(B) ϕ_B ist ein Level- α_0 -Test mit Testumfang α_0 , wenn der kritische Wert definiert ist durch

$$k_{\alpha_0}^B := \varphi^{-1}(1 - \alpha_0; 1, n - 3). \quad (42.59)$$

Der zu einem beobachteten Wert f_B von F_B assoziierte p-Wert ist gegeben durch

$$p_B - \text{Wert} := 1 - \varphi(f_B; 1, n - 3). \quad (42.60)$$

◦

Anwendungsbeispiel

Die Anwendung des oben formulierten Evaluationsverfahrens für die Haupteffekte im Modell der additiven 2×2 ZVA mit Referenzgruppe anhand des Beispieldatensatzes demonstriert folgender **R** Code.

```
D = read.csv("../data/509-zweifaktorielle-varianzanalyse.csv") # Datensatz
A1B1 = D$BDI[D$Setting == "F2F" & D$Variant == "MND" ] # Face-to-face, mindfulness
A1B2 = D$BDI[D$Setting == "F2F" & D$Variant == "EXC" ] # Face-to-face, exercise
A2B1 = D$BDI[D$Setting == "ONL" & D$Variant == "MND" ] # Online, mindfulness
A2B2 = D$BDI[D$Setting == "ONL" & D$Variant == "EXC" ] # Online, exercise
I = 2 # Anzahl Level Faktor Setting
J = 2 # Anzahl Level Faktor Variante
n_ij = length(A1B1) # balanciertes ANOVA Design
n = I*J*n_ij # Anzahl Datenpunkte
p = 3 # Anzahl Parameter vollständiges Modell
y = matrix(c(A1B1,A1B2,A2B1,A2B2), nrow = n) # Datenvektor
D = matrix(c(1,0,0,1,0,1,0,1,1,0,1,1,1), # Prototypische Designmatrix
           nrow = I*J,byrow=TRUE)
C = matrix(rep(1,n_ij),nrow = n_ij) # Prototypischer Zellenvektor
X = kronecker(D,C) # ZVA Kroneckerprodukt Designmatrix
XH = list(X[,c(1,3,2)], X) # Modellvarianten
alpha_0 = 0.05 # Signifikanzlevel
Eff = rep(NaN,2) # F-Teststatistik Arrayinitialisierung
k_alpha_0 = rep(NaN,2) # Kritischer Wert Arrayinitialisierung
phi = rep(NaN,2) # Testwert Arrayinitialisierung
p_vals = rep(NaN,2) # p-Wert Arrayinitialisierung
for(i in 1:2){ # Iteration über Modellvarianten
  X = XH[[i]] # Designmatrix vollständiges Modell
  X_0 = X[,-3] # Designmatrix reduziertes Modell
  p = ncol(X) # Anzahl Parameter vollständiges Modell
  p_0 = ncol(X_0) # Anzahl Parameter reduziertes Modell
  p_1 = p - p_0 # Anzahl zusätzlicher Parameter im vollst. Modell
  beta_hat_0 = solve(t(X_0)%*%X_0)%*%t(X_0)%*%y # Betaparameterschätzer reduziertes Modell
  beta_hat = solve(t(X) %*%X )%*%t(X) %*%y # Betaparameterschätzer vollständiges Modell
  eps_hat_0 = y - X_0%*%beta_hat_0 # Residuenvektor reduziertes Modell
  eps_hat = y - X%*%beta_hat # Residuenvektor vollständiges Modell
  eh0_eh0 = t(eps_hat_0) %*% eps_hat_0 # RQS reduziertes Modell
  eh_eh = t(eps_hat) %*% eps_hat # RQS vollständiges Modell
  sigsq_r_hat = eh_eh/(n-p) # Varianzparameterschätzer vollst. Modell
  Eff[i] = ((eh0_eh0-eh_eh)/(p_1))/sigsq_r_hat # F-Statistik
  k_alpha_0[i] = qf(1-alpha_0, p_1, n-p) # kritischer Wert
  if(Eff[i] >= k_alpha_0[i]){ phi[i] = 1 } # H_0 Ablehnen
  else { phi[i] = 0 } # H_A Annehmen
  p_vals[i] = 1 - pf(Eff[i], p_1,n-p) # p-Wert
}
```

	f	k	phi	p.Wert
Setting	0.172	4.057	0	0.68
Variant	51.356	4.057	1	0.00

Für den Haupteffekt des Faktors Settings ergibt sich dabei ein Wert der F -Statistik von $F_A = 0.17$ bei einem kritischen Wert von $k_{0.05} = 4.06$. Die Nullhypothese eines wahren, aber unbekannten, Effektparameterwertes von Null für den Faktor Setting würde im Lichte der in Abbildung 42.3 visualisierten Daten also nicht verworfen werden. Für den Haupteffekt des Faktors Variant dagegen ergibt ein Wert der F -Statistik von $F_B = 51.36$ bei dem gleichen kritischen Wert von $k_{0.05} = 4.06$. Die Nullhypothese eines wahren, aber unbekannten, Effektparameterwertes von Null für den Faktor Variant würde im Lichte der in Abbildung 42.3 visualisierten Daten mit einem Signifikanzlevel von $\alpha_0 := 0.05$ also verworfen werden.

Folgender **R** Code demonstriert die gleiche Analyse mithilfe der **R** Funktion `aov()`.

```
D = read.csv("./_data/509-zweifaktorielle-varianzanalyse.csv") # Datensatz
res.aov = aov(dBDI ~ Setting + Variant, data = D) # Modellformulierung und Modellschätzung
summary(res.aov) # Modellevaluation
```

```
      Df Sum Sq Mean Sq F value    Pr(>F)
Setting  1  1.02    1.02    0.172      0.68
Variant  1 305.02   305.02   51.356 5.78e-09 ***
Residuals 45 267.27    5.94
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

normalsize

Folgender **R** Code schließlich demonstriert die gleiche Analyse mithilfe der **R** Funktionen `lm()` und `anova()`.

```
D = read.csv("./_data/509-zweifaktorielle-varianzanalyse.csv") # Datensatz
glm = lm(dBDI ~ Setting + Variant, data = D) # Modellformulierung und Modellschätzung
anova(glm) # Modellevaluation
```

Analysis of Variance Table

```
Response: dBDI
      Df Sum Sq Mean Sq F value    Pr(>F)
Setting  1  1.021    1.021    0.1719      0.6804
Variant  1 305.021   305.021   51.3559 5.779e-09 ***
Residuals 45 267.271    5.939
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Evaluation der Interaktion im Modell der 2×2 ZVA mit Interaktion

Theorem 42.5 (Teststatistik für die Interaktion). *Es sei*

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \quad (42.61)$$

die Designmatrixform des 2×2 ZVA Modells mit Interaktion und Referenzgruppe. Dann ist die F -Statistik mit $p_0 := 3$ und $p_1 := 1$ eine F -Teststatistik für die Interaktion von Faktor A und Faktor B.

◦

Definition 42.6 (Testhypothese und Test für die Interaktion). Die F -Teststatistik für die Interaktion von Faktor A und Faktor B sei mit $F_{A \times B}$ bezeichnet und wie oben definiert. Dann definiert der kritische Wert-basierte Test

$$\phi_{A \times B}(y) := 1_{\{F_{A \times B} \geq k\}} \text{ mit Nullhypothese } H_0^{A \times B} : \gamma_{22} = 0 \quad (42.62)$$

den F -Test der Interaktion von Faktor A und Faktor B.

•

Theorem 42.6 (Testumfangkontrolle und p-Wert für die Interaktion). *Mit obigen Definition und der KVF φ der f -Verteilung gilt, dass $\phi_{A \times B}$ ein Level- α_0 -Test mit Testumfang α_0 ist, wenn der kritische Wert definiert ist durch*

$$k_{\alpha_0}^{A \times B} := \varphi^{-1}(1 - \alpha_0; 1, n - 4). \quad (42.63)$$

Der zu einem beobachteten Wert $f_{A \times B}$ von $F_{A \times B}$ assoziierte p-Wert ist gegeben durch

$$p_{A \times B} - \text{Wert} := 1 - \varphi(f_{A \times B}; 1, n - 4). \quad (42.64)$$

◦

Anwendungsbeispiel

Die Anwendung des oben formulierten Evaluationsverfahrens für die Interaktion der Faktoren im Modell der 2×2 ZVA mit Interaktion und Referenzgruppe anhand des Beispieldatensatzes demonstriert folgender **R** Code.

```
D = read.csv("./_data/509-zweifaktorielle-varianzanalyse.csv") # Datensatz
A1B1 = D$BDI[D$Setting == "F2F" & D$Variant == "MND" ] # Face-to-face, mindfulness
A1B2 = D$BDI[D$Setting == "F2F" & D$Variant == "EXC" ] # Face-to-face, exercise
A2B1 = D$BDI[D$Setting == "ONL" & D$Variant == "MND" ] # Online, mindfulness
A2B2 = D$BDI[D$Setting == "ONL" & D$Variant == "EXC" ] # Online, exercise
I = 2 # Anzahl Level Faktor Setting
J = 2 # Anzahl Level Faktor Variante
n_ij = length(A1B1) # balanciertes ANOVA Design
n = I*J*n_ij # Anzahl Datenpunkte
p = 4 # Anzahl Parameter vollständiges Modell
y = matrix(c(A1B1,A1B2,A2B1,A2B2), nrow = n) # Datenvektor
D = matrix(c(1,0,0,1,0,1,0, 1,1,0,0,1,1,1), nrow = I*J, byrow=TRUE) # Prototypische Designmatrix
C = matrix(rep(1,n_ij), nrow = n_ij) # Prototypischer Zellenvektor
X = kronecker(D,C) # ZVA Kroneckerprodukt Designmatrix
XH = list(X[,c(1,3,2)], X) # Modellvarianten
alpha_0 = 0.05 # Signifikanzlevel
X = XH[[i]] # Designmatrix vollständiges Modell
X_0 = X[,~4] # Designmatrix reduziertes Modell
p = ncol(X) # Anzahl Parameter vollständiges Modell
p_0 = ncol(X_0) # Anzahl Parameter reduziertes Modell
p_1 = p - p_0 # Anzahl zusätzlicher Parameter im vollst. Modell
beta_hat_0 = solve(t(X_0)%*%X_0)%*%t(X_0)%*%y # Betaparameterschätzer reduziertes Modell
beta_hat = solve(t(X)%*%X)%*%t(X)%*%y # Betaparameterschätzer vollständiges Modell
eps_hat_0 = y - X_0%*%beta_hat_0 # Residuenvektor reduziertes Modell
eps_hat = y - X%*%beta_hat # Residuenvektor vollständiges Modell
eh0_eh0 = t(eps_hat_0)%*%eps_hat_0 # RQS reduziertes Modell
eh_eh = t(eps_hat)%*%eps_hat # RQS vollständiges Modell
sigsqr_hat = eh_eh/(n-p) # Varianzparameterschätzer vollst. Modell
f = ((eh0_eh0-eh_eh)/p_1)/sigsqr_hat # F-Statistik
k_alpha_0 = qf(1-alpha_0, p_1, n-p) # kritischer Wert
if(f >= k_alpha_0){phi = 1} else {phi = 0} # Test
p_val = 1 - pf(f, p_1, n-p) # p-Wert
```

```
f k phi p.Wert
Setting x Variant 1.014 4.062 0 0.319
```

Es ergibt sich dabei ein Wert der F -Statistik von $F_{A \times B} = 1.01$ bei einem kritischen Wert von $k_{0.05} = 4.06$. Die Nullhypothese eines wahren, aber unbekannten, Effektparameterwertes von Null für den Interaktionseffekt würde im Lichte der in Abbildung 42.3 visualisierten Daten also nicht verworfen werden.

Folgender **R** Code demonstriert die gleiche Analyse mithilfe der **R** Funktion `aov()`.

```
D = read.csv("./_data/509-zweifaktorielle-varianzanalyse.csv") # Datensatz
res.aov = aov(dBDI ~ Setting + Variant + Setting:Variant, data = D) # Modellformulierung und Modellschätzung
summary(res.aov) # Modellevaluation
```

```
Df Sum Sq Mean Sq F value Pr(>F)
Setting 1 1.02 1.02 0.172 0.680
Variant 1 305.02 305.02 51.372 6.5e-09 ***
Setting:Variant 1 6.02 6.02 1.014 0.319
Residuals 44 261.25 5.94
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Folgender **R** Code schließlich demonstriert die gleiche Analyse mithilfe der **R** Funktionen `lm` und `anova()`.

```
D = read.csv("./_data/509-zweifaktorielle-varianzanalyse.csv") # Datensatz
glm = lm(dBDI ~ Setting + Variant + Setting:Variant, data = D) # Modellformulierung und Modellschätzung
anova(glm) # Modellevaluation
```

Analysis of Variance Table

Response: dBDI

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Setting	1	1.021	1.021	0.1719	0.6804
Variant	1	305.021	305.021	51.3719	6.502e-09 ***
Setting:Variant	1	6.021	6.021	1.0140	0.3194
Residuals	44	261.250	5.938		

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

42.5. Literaturhinweise

Die Popularität varianzanalytischer Verfahren wird im Allgemeinen auf Fisher (1925a) und Fisher (1935) zurückgeführt. Everitt & Howell (2005) und Stigler (1986) geben einen kurzen und einen ausführlichen historischen Überblick, respektive.

43. Partielle Korrelation

43.1. Motivation

Zur Motivation des Begriffs der partiellen Korrelation betrachten wir zunächst den in Abbildung 43.1 visualisierten Beispieldatensatz zum Zusammenhang von Eiskonsum und Sonnenbrandinzidenz. Wir stellen uns vor, dass jeder der in Abbildung 43.1 abgebildeten Datenpunkte ein Wertepaar aus einem durchschnittlichen und normalisierten Eiskonsum und einer durchschnittlichen und normalisierten Sonnenbrandinzidenz eines Landes über einen gewissen Erhebungszeitraum ist. Visuell betrachtet sieht man eine Tendenz dafür, dass hohe Werte des Eiskonsums mit eher hohen Werten der Sonnenbrandinzidenz auftreten, während niedrige Werte des Eiskonsums mit eher niedrigen Werten der Sonnenbrandinzidenz zusammen auftreten. Die Bestimmung des Stichprobenkorrelationskoeffizienten zu diesem Datensatz ergibt mit $r = 0.46$ eine mittelstarke positive Korrelation.

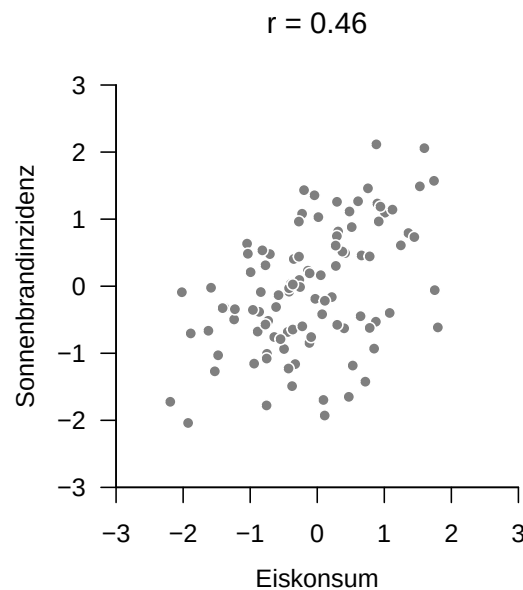


Abbildung 43.1. Beispielszenario zur Evaluation bedingter und partieller Korrelationen.

Intuitiv ist es jedoch eher unplausibel, dass Eiskonsum ursächlich Sonnenbrand hervorruft bzw. das Sonnenbrand den Eiskonsum erhöht (allerdings sind diese Szenarien auch nicht gänzlich auszuschließen: ein bestimmter Eiskonsum könnte eine allergische Reaktion hervorrufen mit Symptomen, die dem Sonnenbrand sehr ähnlich sind, andersherum wäre es denkbar, dass bei Sonnenbrand zur Abkühlung gerne Eis konsumiert wird. Wir wollen diese eher unplausiblen Erklärungsansätze hier jedoch nicht weiter verfolgen). Der in Abbildung 43.1 dargestellte Datensatz ist also ein Beispiel dafür, dass Korrelation als Maß für

den linearen Zusammenhang zweier Zufallsvariablen lediglich ein Maß für die Koinzidenz bestimmter Datenwerte ist, jedoch keine Kausalerklärung der Werte einer abhängigen Variable aus den Werten einer unabhängigen Variable impliziert. In Kurzform hat sich zur Beschreibung dieser Tatsache seit Beginn der modernen Inferenzstatistik am Anfang des 20. Jahrhunderts der Leitsatz “Correlation is not causation” eingebürgert.

Basierend auf dem negativen Ergebnis, dass eine mittelstarke Korrelation wie im Beispiel von Eiskonsum und Sonnenbrandinzidenz nur sehr unplausibel durch eine direkte kausale Beziehung der beiden Variablen zu erklären ist, stellt sich die Frage, inwieweit andere datenanalytische Verfahren hier Abhilfe schaffen können. Dabei stellt sich natürlich zunächst das philosophische Problem, was Kausalität eigentlich bedeuten soll und als nächstes die Frage, wie ein solcher Begriff mit den Mitteln der Wahrscheinlichkeitstheorie und Inferenzstatistik evaluiert werden könnte. Diesen Ansatz verfolgt das Gebiet der Kausalen Inferenz, wie zum Beispiel durch die Arbeiten von Pearl (2000) und Imbens & Rubin (2015) repräsentiert. Wir wollen an dieser Stelle diesen Ansatz nicht vertiefen, sondern stattdessen fragen, wie im obigen Beispiel anhand weiterer Daten die beobachtete Korrelation von Eiskonsum und Sonnenbrandinzidenz so aufgeklärt werden kann, dass die statistische Beschreibung des in Abbildung 43.1 dargestellten Datensatzes plausibler erscheint. Dies ist das zentrale Thema der partiellen Korrelation.

Dazu nehmen wir an, dass der Zusammenhang von Eiskonsum und Sonnenbrandinzidenz (Abbildung 43.2 A) plausibel durch die Kovariation beider Variablen mit einer dritten Variable, nämlich der Anzahl der im Erhebungszeitraum und Land auftretenden Anzahl an Sommertagen, d.h. Tagen mit einer maximalen Temperatur von über 25° Celsius, erklärt werden kann (Abbildung 43.2 B).

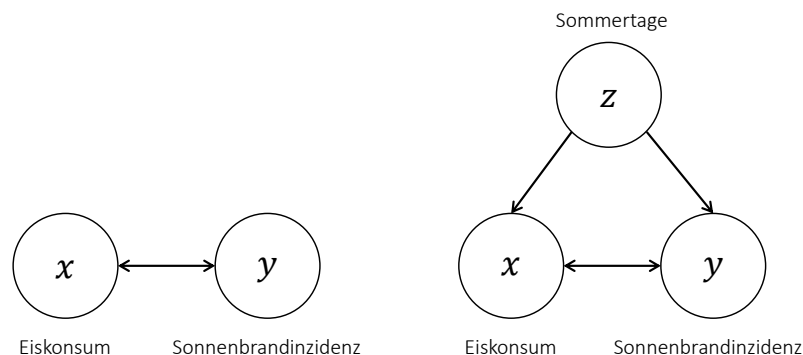


Abbildung 43.2. Zufallsvariablen im Beispielszenario.

Intuitiv erklärt sich die positive Korrelation von Eiskonsum und Sonnenbrandinzidenz dann wie folgt. Treten im Erhebungszeitraum in einem Land mehr Sommertage auf, so steigt in diesem Land sowohl der Eiskonsum als auch die Sonnenbrandinzidenz, treten dagegen weniger Sommertage auf, so fallen in diesem Land sowohl der Eiskonsum als auch die Sonnenbrandinzidenz. Lässt man die Anzahl der Sommertage außer Acht, so treten also hohe Werte von Eiskonsum und Sonnenbrandinzidenz als auch niedrige Werte von Eiskonsum und Sonnenbrandinzidenz häufig zusammen auf und es ergibt sich die in Abbildung 43.1 implizierte positive Korrelation.

Die entscheidende Frage in diesem Kontext ist also, ob bei gleicher Anzahl von Sommertagen Evidenz für eine Korrelation von Eiskonsum und Sonnenbrandinzidenz besteht oder

nicht. In diesem Fall würde die Kovarianz von Eiskonsum und Sonnenbrandinzidenz also bedingt auf einer konstanten Anzahl von Sommertagen betrachtet werden. Die datenanalytischen Werkzeuge, bei Vorliegen von Realisationen von drei Zufallsvariablen eben diese Form einer bedingten Korrelation zu evaluieren, stellen der Begriff der bedingten Korrelation und der eng verwandte Begriff der partiellen Korrelation bereit. Intuitiv handelt es sich dabei um die Korrelation zweier Zufallsvariablen (z.B. Eiskonsum und Sonnenbrandinzidenz) nachdem aus beiden Zufallsvariablen der Einfluß einer dritten Zufallsvariable (z.B. Anzahl an Sommertagen) “herausgerechnet” wurde. Die Begriffe der bedingten und partiellen Korrelation sind dabei nicht auf das Szenario von drei Zufallsvariablen beschränkt, sondern können für beliebig viele Zufallsvariablen generalisiert werden. Wir beschränken uns in diesem Abschnitt allerdings auf das Szenario dreier Zufallsvariablen um die Grundlagen der Theorie zu verdeutlichen.

Wir gehen dabei wie folgt vor. In Kapitel 12.2 führen wir mit der bedingten Kovarianz und der bedingten Korrelation zunächst allgemeine Maße für den auf den Werten einer dritten Zufallsvariable bedingten linear-affinen Zusammenhang zweier Zufallsvariablen ein, verdeutlichen dann die Begriffe anhand des Szenarios dreier gemeinsam multivariat normalverteilter Zufallsvariablen und diskutieren schließlich den Zusammenhang zwischen bedingter Korrelation und paarweisen (unbedingten) Korrelation. In Kapitel 12.3 führen wir mit der partiellen Korrelation dann ein regressionsbasiertes Maß für den bedingten Zusammenhang zweier Zufallsvariablen ein. Dabei ergibt sich insbesondere, dass im Falle von gemeinsam multivariat normalverteilten Zufallsvariablen bedingte und partielle Korrelation identisch sind. Wir schließen diesen Abschnitt mit der Evaluation der partiellen Korrelation von Eiskonsum und Sonnenbrandinzidenz im Lichte des Wissens um die Anzahl an Sommertagen für den in Abbildung 43.1 visualisierten Beispieldatensatz.

43.2. Bedingte Korrelation

Wir definieren zunächst die bedingte Kovarianz und die bedingte Korrelation zweier Zufallsvariablen gegeben eine dritte Zufallsvariable.

Definition 43.1 (Bedingte Kovarianz und bedingte Korrelation). Gegeben seien drei Zufallsvariablen x, y, z einer gemeinsamen Verteilung $\mathbb{P}_{x,y,z}(x, y, z)$. Weiterhin sei $\mathbb{P}_{x,y|z}(x, y)$ die bedingte Verteilung von x und y gegeben z . Dann heißt die Kovarianz von x und y in der Verteilung $\mathbb{P}_{x,y|z}(x, y)$ die bedingte Kovarianz von x und y gegeben z und wird mit $\mathbb{C}(x, y|z)$ bezeichnet. Weiterhin seien $\mathbb{P}_{x,y|z}(x)$ und $\mathbb{P}_{x,y|z}(y)$ die marginalen Verteilungen von x und y gegeben z , respektive, und $\mathbb{S}(x|z), \mathbb{S}(y|z)$ die Standardabweichungen von x und y hinsichtlich $\mathbb{P}_{x,y|z}(y)$ und $\mathbb{P}_{x,y|z}(x)$, respektive. Dann heißt die Korrelation von x und y in der Verteilung $\mathbb{P}_{x,y|z}(x, y)$,

$$\rho(x, y|z) := \frac{\mathbb{C}(x, y|z)}{\mathbb{S}(x|z)\mathbb{S}(y|z)} \quad (43.1)$$

die *bedingte Korrelation* von x und y gegeben z

•

Die bedingte Kovarianz zweier Zufallsvariablen ist also definiert als die Kovarianz zweier Zufallsvariablen in einer auf einer dritten Zufallsvariable bedingten Verteilung. Gleiches

gilt für die bedingte Korrelation zweier Zufallsvariablen. Durch Vertauschen in obiger Definition kann man analog $\rho(y, z|x)$ und $\rho(x, z|y)$ definieren. Wir verdeutlichen Definition 43.1 als nächstes an einem Beispiel.

Beispiel

Die Zufallsvariablen x, y, z seien multivariat normalverteilt, d.h. für $\gamma := (x, y, z)^T$ gelte, dass

$$\gamma \sim N(\mu, \Sigma) \quad (43.2)$$

mit

$$\mu := \begin{pmatrix} \mu_y \\ \mu_x \\ \mu_z \end{pmatrix} \text{ und } \Sigma := \begin{pmatrix} \sigma_x^2 & \sigma_{x,y}^2 & \sigma_{x,z}^2 \\ \sigma_{y,x}^2 & \sigma_y^2 & \sigma_{y,z}^2 \\ \sigma_{z,x}^2 & \sigma_{z,y}^2 & \sigma_z^2 \end{pmatrix} \quad (43.3)$$

Wir nehmen an, dass wir die bedingte Korrelation von x und y gegeben z bestimmen wollen und wenden uns entsprechend der bedingten Verteilung von x und y gegeben z zu. Nach Theorem 29.6 wissen wir, dass diese bedingte Verteilung ebenfalls eine Normalverteilung ist, deren Kovarianzmatrixparameter wir aus dem Kovarianzmatrixparameter der gemeinsamen Verteilung von x, y, z bestimmen können. Wir definieren zu diesem Zweck zunächst

$$\Sigma_{x,y} := \begin{pmatrix} \sigma_x^2 & \sigma_{x,y}^2 \\ \sigma_{y,x}^2 & \sigma_y^2 \end{pmatrix}, \Sigma_z := (\sigma_z^2) \text{ und } \Sigma_{(x,y),z} := \Sigma_{z,(x,y)}^T := \begin{pmatrix} \sigma_{x,z}^2 \\ \sigma_{y,z}^2 \end{pmatrix} \quad (43.4)$$

so dass für den Kovarianzmatrixparameter der gemeinsamen Verteilung von x, y, z gilt, dass

$$\Sigma = \begin{pmatrix} \Sigma_{x,y} & \Sigma_{(x,y),z} \\ \Sigma_{z,(x,y)} & \Sigma_z \end{pmatrix} \quad (43.5)$$

Mit Theorem 4.8 ergibt sich der Kovarianzmatrixparameter des Zufallsvektors $(x, y)^T$ dann zu

$$\Sigma_{x,y|z} = \Sigma_{x,y} - \Sigma_{(x,y),z} \Sigma_z^{-1} \Sigma_{z,(x,y)} \quad (43.6)$$

Mit den Eigenschaften von multivariaten Normalverteilungen gilt dann, dass die Diagonaleinträge von $\Sigma_{x,y|z}$ den bedingten Varianzen von x und y gegeben z entsprechen und dass der Nichtdiagonaleintrag von $\Sigma_{x,y|z}$ die bedingte Kovarianz von x und y gegeben z ist. In anderen Worten gilt

$$\Sigma_{x,y|z} = \begin{pmatrix} \mathbb{C}(x, x|z) & \mathbb{C}(x, y|z) \\ \mathbb{C}(y, x|z) & \mathbb{C}(y, y|z) \end{pmatrix} \quad (43.7)$$

Die bedingte Korrelation $\rho(x, y|z)$ von x und y gegeben z ergibt sich dann aus den Einträgen von $\Sigma_{x,y|z}$ gemäß

$$\rho(x, y|z) = \frac{\mathbb{C}(x, y|z)}{\sqrt{\mathbb{C}(x, x|z)} \sqrt{\mathbb{C}(y, y|z)}} \quad (43.8)$$

Sei konkret etwa der Kovarianzmatrixparameter von $(x, y, z)^T$ gegeben als

$$\Sigma := \begin{pmatrix} 1.0 & 0.5 & 0.9 \\ 0.5 & 1.0 & 0.5 \\ 0.9 & 0.5 & 1.0 \end{pmatrix} \quad (43.9)$$

Dann ergibt sich

$$\rho(x, y) = 0.50 \text{ und } \rho(x, y|z) \approx 0.13 \quad (43.10)$$

Folgender **R** Code demonstriert die Auswertung dieser bedingten Korrelation.

```
# Bedingte Korrelation bei Normalverteilung
S = matrix(c( 1,.5,.9,
             .5, 1,.5,
             .9,.5, 1), nrow = 3, byrow = TRUE)
rho_xy = S[1,2]/(sqrt(S[1,1])*sqrt(S[2,2]))
S_xy_z = S[1:2,1:2] - S[1:2,3] %*% solve(S[3,3]) %*% S[3,1:2]
rho_xy_z = S_xy_z[1,2]/(sqrt(S_xy_z[1,1])*sqrt(S_xy_z[2,2]))

# \Sigma
# \rho(x,y)
# \Sigma_{x,y|z}
# \rho(x,y|z)

rho(x,y) : 0.5
rho(x,y|z) : 0.13
```

43.3. Bedingte Korrelation bei Normalverteilung

Für den Fall dreier gemeinsam normalverteilter Zufallsvariablen eröffnet folgendes Theorem eine Möglichkeit, die bedingte Korrelation zweier dieser Zufallsvariablen gegeben die dritte auf Grundlage der (unbedingten) paarweisen Korrelationen der Zufallsvariablen zu bestimmen. So kann bei gemeinsamer Normalverteilung von x, y, z zum Beispiel $\rho(x, y|z)$ aus den Korrelationen $\rho(x, y)$, $\rho(x, z)$, und $\rho(y, z)$ bestimmt werden. Speziell gilt folgendes Theorem.

Theorem 43.1 (Bedingte Korrelation und Korrelationen bei Normalverteilung). x, y, z seien drei gemeinsam multivariat normalverteilte Zufallsvariablen. Dann gilt

$$\rho(x, y|z) = \frac{\rho(x, y) - \rho(x, z)\rho(y, z)}{\sqrt{(1 - \rho(x, z)^2)}\sqrt{(1 - \rho(y, z)^2)}} \quad (43.11)$$

◦

Beweis. Ohne Beschränkung der Allgemeinheit betrachten wir den Fall eines standardisierten multivariat normalverteilten Zufallsvektors $\gamma := (x, y, z)^T$ mit Kovarianzmatrixparameter

$$\Sigma := \begin{pmatrix} 1 & \rho(x, y) & \rho(x, z) \\ \rho(y, x) & 1 & \rho(y, z) \\ \rho(z, x) & \rho(z, y) & 1 \end{pmatrix} \quad (43.12)$$

Wir definieren nun zunächst

$$\Sigma_{x,y} := \begin{pmatrix} 1 & \rho(x, y) \\ \rho(y, x) & 1 \end{pmatrix}, \Sigma_z := (1) \text{ und } \Sigma_{(x,y),z} := \Sigma_{z,(x,y)}^T := \begin{pmatrix} \rho(x, z) \\ \rho(y, z) \end{pmatrix}, \quad (43.13)$$

so dass

$$\Sigma = \begin{pmatrix} \Sigma_{x,y} & \Sigma_{(x,y),z} \\ \Sigma_{z,(x,y)} & \Sigma_z \end{pmatrix}. \quad (43.14)$$

Mit dem Theorem 29.6 ist dann die Kovarianzmatrix des Zufallsvektors (x, y) gegeben durch

$$\Sigma_{x,y|z} = \Sigma_{x,y} - \Sigma_{(x,y),z} \Sigma_z^{-1} \Sigma_{z,(x,y)} \quad (43.15)$$

Es ergibt sich also

$$\begin{aligned}
 \begin{pmatrix} \sigma_{x,x|z}^2 & \sigma_{x,y|z}^2 \\ \sigma_{y,x|z}^2 & \sigma_{y,y|z}^2 \end{pmatrix} &= \begin{pmatrix} 1 & \rho(x,y) \\ \rho(y,x) & 1 \end{pmatrix} - \begin{pmatrix} \rho(x,z) \\ \rho(y,z) \end{pmatrix} (1)^{-1} \begin{pmatrix} \rho(x,z) & \rho(y,z) \end{pmatrix} \\
 &= \begin{pmatrix} 1 & \rho(x,y) \\ \rho(y,x) & 1 \end{pmatrix} - \begin{pmatrix} \rho(x,z)\rho(x,z) & \rho(x,z)\rho(y,z) \\ \rho(y,z)\rho(x,z) & \rho(y,z)\rho(y,z) \end{pmatrix} \\
 &= \begin{pmatrix} 1 - \rho(x,z)^2 & \rho(x,y) - \rho(x,z)\rho(y,z) \\ \rho(y,x) - \rho(y,z)\rho(x,z) & 1 - \rho(y,z)^2 \end{pmatrix}.
 \end{aligned} \tag{43.16}$$

Es ergibt sich also

$$\rho(x,y|z) = \frac{\sigma_{x,y|z}^2}{\sqrt{\sigma_{x,x|z}^2} \sqrt{\sigma_{y,y|z}^2}} = \frac{\rho(x,y) - \rho(x,z)\rho(y,z)}{\sqrt{1 - \rho(x,z)^2} \sqrt{1 - \rho(y,z)^2}} \tag{43.17}$$

Im Falle des Vorliegens von Realisierungen von x, y, z ergibt sich ein entsprechender Schätzer für $\rho(x,y|z)$ mit den Stichprobenkorrelationen $r_{x,y}, r_{x,z}, r_{y,z}$ dann zu

$$r_{x,y|z} = \frac{r_{x,y} - r_{x,z}r_{y,z}}{\sqrt{(1 - r_{x,z}^2)} \sqrt{(1 - r_{y,z}^2)}} \tag{43.18}$$

□

43.4. Partielle Korrelation

Wir definieren als nächstes die partielle Korrelation zweier Zufallsvariablen gegeben eine dritte Zufallsvariable.

Definition 43.2 (Partielle Korrelation.). x, y, z seien Zufallsvariablen mit linear-affinen Abhängigkeiten zwischen x und z sowie zwischen y und z ,

$$\begin{aligned}
 x &:= \beta_0^{x,z} + \beta_1^{x,z} z \\
 y &:= \beta_0^{y,z} + \beta_1^{y,z} z
 \end{aligned} \tag{43.19}$$

mit Residualvariablen

$$\begin{aligned}
 \varepsilon^{x,z} &:= x - \beta_0^{x,z} - \beta_1^{x,z} z \\
 \varepsilon^{y,z} &:= y - \beta_0^{y,z} - \beta_1^{y,z} z
 \end{aligned} \tag{43.20}$$

Dann ist die partielle Korrelation von x und y mit auspartialisiertem z definiert als

$$\rho(x,y|z) := \rho(\varepsilon^{x,z}, \varepsilon^{y,z}). \tag{43.21}$$

•

Intuitiv entsprechen in obiger Definition die Zufallsvariable $\varepsilon^{x,z}$ der Zufallsvariable x , aus der der Einfluss von z “herausgerechnet” wurde, und die Zufallsvariable $\varepsilon^{y,z}$ der Zufallsvariable y , aus der der Einfluss von z “herausgerechnet” wurde. Damit entspricht $\rho(x,y|z)$ dann intuitiv der Korrelation von x und y , aus denen jeweils der Einfluss von z “herausgerechnet” wurde. Wir geben als nächstes einen Schätzer für die partielle Korrelation zweier Zufallsvariablen gegeben eine dritte Zufallsvariable an.

Definition 43.3 (Partielle Stichprobenkorrelation). x, y, z seien Zufallsvariablen mit linear-affinen Abhängigkeiten zwischen y und z sowie zwischen x und z wie in der Definition der partiellen Korrelation. Weiterhin seien

- $\{(x_i, y_i, z_i)\}_{i=1, \dots, n}$ eine Menge von Realisierungen des Zufallsvektors $(x, y, z)^T$,
- $\hat{\beta}_0^{x,z}, \hat{\beta}_1^{x,z}$ die Ausgleichsgeradenparameter für $\{(x_i, z_i)\}_{i=1, \dots, n}$,
- $\hat{\beta}_0^{y,z}, \hat{\beta}_1^{y,z}$ die Ausgleichsgeradenparameter für $\{(y_i, z_i)\}_{i=1, \dots, n}$.

Schließlich seien für $i = 1, \dots, n$

- $e_i^{x,z} := x_i - \hat{\beta}_0^{x,z} + \hat{\beta}_1^{x,z} z_i$
- $e_i^{y,z} := y_i - \hat{\beta}_0^{y,z} + \hat{\beta}_1^{y,z} z_i$

die Residualwerte der jeweiligen Ausgleichsgeraden. Dann heißt die Stichprobenkorrelation der Wertemenge $\{(e_i^{y,z}, e_i^{x,z})\}_{i=1, \dots, n}$ *partielle Stichprobenkorrelation der x_i und y_i mit auspartialisierten z_i* .

•

Für den Fall, dass x, y, z multivariat normalverteilt sind, gibt folgendes Theorem, auf dessen Beweis wir hier verzichten wollen, den Zusammenhang zwischen bedingter und partieller Korrelation an.

Theorem 43.2 (Bedingte und Partielle Korrelation bei Normalverteilung). x, y, z seien drei gemeinsam multivariat normalverteilte Zufallsvariablen. Dann gilt

$$\rho(x, y|z) = \rho(x, y \setminus z) \quad (43.22)$$

◦

Man beachte, dass obiges Theorem im Falle dreier multivariat normalverteilter Zufallsvariablen gilt. Im Allgemeinen, also für beliebige Verteilungen der drei Zufallsvariablen gilt die Identität von bedingter und partieller Korrelationen nicht. Weitere Details in diesem Zusammenhang diskutieren zum Beispiel Lawrance (1976) und Baba et al. (2004).

Aus Theorem 43.2 folgt mit Theorem 43.2 dann unmittelbar, dass bei gemeinsamer Normalverteilung von x, y, z die partielle Korrelation $\rho(x, y|z)$ ebenso wie die bedingte Korrelation $\rho(x, y \setminus z)$ basierend auf den (unbedingten) Korrelationen $\rho(x, y)$, $\rho(x, z)$ und $\rho(y, z)$ bestimmt werden kann, bzw. im Falle der jeweiligen Stichprobenäquivalente durch diese geschätzt werden kann.

Folgender **R** Code demonstriert die Auswertung der partiellen Stichprobenkorrelation basierend auf einem simulierten Datensatz dreier multivariat normalverteilter Zufallsvariablen. Dabei bestimmen wir die partielle Korrelation einmal basierend aus den Residualstichprobenkorrelation wie in Definition 43.3 und einmal basierend auf den paarweisen Stichprobenkorrelationen anhand von Theorem 43.2. Schließlich stellt das **R** Paket `ppcor` mit `pcor()` eine Funktion zur automatisierten Auswertung partieller Stichprobenkorrelationen bereit, auch ihre Anwendung demonstrieren wir untenstehend. Das Resultat ist natürlich in allen drei Fällen identisch.

```

# Modellformulierung und Datenrealisierung
library(MASS)
set.seed(1)
S = matrix(c( 1,.5,.9,
             .5, 1,.5,
             .9,.5, 1),nrow=3,byrow=TRUE)

n = 1e6
xyz = mvrnorm(n,rep(0,3),S)

# Partielle Stichprobenkorrelation als Residualstichprobenkorrelation
bars = apply(xyz, 2, mean)
s = apply(xyz, 2, sd)
c = cov(xyz)
b_xz1 = c[1,3]/c[3,3]
b_xz0 = bars[1] - b_xz1*bars[3]
b_yz1 = c[2,3]/c[3,3]
b_yz0 = bars[2] - b_yz1*bars[3]
e_xz = xyz[,1] - b_xz1*xyz[,3] - b_xz0
e_yz = xyz[,2] - b_yz1*xyz[,3] - b_yz0
pr_e = cor(e_xz,e_yz)

# Partielle Stichprobenkorrelation aus Stichprobenkorrelationen
r = cor(xyz)
pr_r_n = r[1,2]-r[1,3]*r[2,3]
pr_r_d = sqrt((1-r[1,3]^2)*(1-r[2,3]^2))
pr_r = pr_r_n/pr_r_d

# partielle Stichprobenkorrelation aus Toolbox
library(ppcor)
pr_t = ppcor(xyz)

# Multivariate Normalverteilung
# reproduzierbare Daten
# Kovarianzmatrixparameter \Sigma

# Anzahl Realisierungen
# Realisierungen

# Stichprobenmittel
# Stichprobenstandardabweichungen
# Stichprobenkovarianzen
# beta_1 (x,z)
# beta_0 (x,z)
# beta_1 (y,z)
# beta_0 (y,z)
# Residualwerte e^{x,z}
# Residualwerte e^{y,z}
# \rho(x,y|z)

# Stichprobenkorrelationsmatrix
# \rho(x,y|z) Formel Zähler
# \rho(x,y|z) Formel Nenner
# \rho(x,y|z)

# Laden der Toolbox
# \rho(x,y|z), \rho(x,z|y), \rho(y,z|x)

```

```

r(x,y) : 0.5
r(x,y/z) aus Residuenkorrelation : 0.13
r(x,y/z) aus Korrelationen : 0.13
r(x,y/z) aus Toolbox : 0.13

```

Anwendungsbeispiel

Mithilfe oben eingeführten **R** Codes wenden wir uns nun abschließend dem eingangs diskutierten Beispiel zum Zusammenhang von Eiskonsum und Sonnenbrandinzidenz zu. Wir nehmen an, dass zu jedem Wertepaar von Eiskonsum (x_i) und Sonnenbrandinzidenz (y_i) der korrespondierende Wert der Anzahl der Sommertage (z_i) im betrachteten Erhebungszeitraum und Land verfügbar ist. Dann eröffnet obige Theorie die Möglichkeit, die partielle Korrelation von Eiskonsum und Sonnenbrandinzidenz nach Korrektur für die Anzahl der Sommertage zu bestimmen.

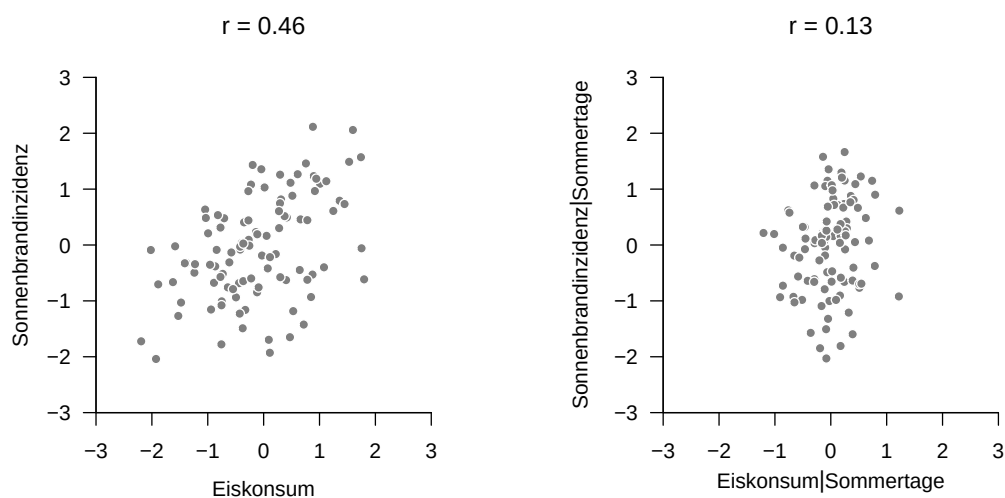


Abbildung 43.3. Evaluation der partiellen Korrelation im Beispielszenario.

Abbildung 43.3

Dazu stellt Abbildung 43.3 mit der Achsenbeschriftung **Eiskonsum | Sommertage** die Residualwerte

$$e_i^{x,z} := x_i - \hat{\beta}_0^{x,z} + \hat{\beta}_1^{x,z} z_i \quad (43.23)$$

und mit der Achsenbeschriftung **Sonnenbrandinzidenz | Sommertage** die Residualwerte

$$e_i^{y,z} := y_i - \hat{\beta}_0^{y,z} + \hat{\beta}_1^{y,z} z_i \quad (43.24)$$

dar. Man erkennt, dass kein systematischer Zusammenhang hoher bzw. niedriger Werte von **Eiskonsum | Sommertage** mit hohen bzw. niedrigen Werten von **Sonnenbrandinzidenz | Sommertage** besteht. Die Korrelation dieser Residualwerte beträgt dementsprechend auch nur $r = 0.17$ und nicht, wie im Falle der nicht für die Kovariation mit der Anzahl der Sommertage korrigierten Werte von Eiskonsum und Sonnenbrandinzidenz, $r = 0.46$ (vgl. Abbildung 43.1). Der bei der nicht durch die Anzahl der Sommertage informierten Korrelationsanalyse implizierte Zusammenhang von Eiskonsum und Sonnenbrandinzidenz lässt sich also durch die Kovariation beider Variablen mit der Drittvariable Sommertage aufklären bzw. “wegerklären”.

43.5. Literaturhinweise

Die Theorie partielle und bedingter Korrelationen findet spätestens seit Beginn der modernen Korrelationsanalyse zu Beginn des 20. Jahrhunderts Beachtung, man vergleiche hierzu zum Beispiel Pearson (1920), Yule (1907) oder Fisher (1924).

44. Multiple Regression

44.1. Anwendungsszenario

Bei der multiplen Regression handelt es sich um die Generalisierung der einfachen linearen Regression im Kontext von mehr als einer kontinuierlichen unabhängigen Variable. Wie bei der einfachen linearen Regression betrachtet man eine univariate abhängige Variable, deren Werte an randomisierten experimentellen Einheiten bestimmt werden. Die unabhängigen Variablen werden bei der multiplen Regression je nach Kontext Regressoren, Prädiktoren, Kovariaten, Features oder einfach Spalten der Designmatrix genannt. Ähnlich wie bei der einfachen linearen Regression ist das Ziel einer multiplen Regressionsanalyse, das Erklärungspotential für Variationen der abhängigen Variablen durch Variation der unabhängigen Variablen zu quantifizieren. In Erweiterung der einfachen linearen Regression liegt ein Augenmerk dabei insbesondere darauf, den Einfluss einzelner unabhängiger Variablen auf die abhängige Variable im Kontext der Variation anderer unabhängiger Variablen zu ermitteln. Darüber hinaus mag ein Ziel auch die Prädiktion der Werte der abhängigen Variablen für Werte der unabhängigen Variable nach Schätzung der entsprechenden Wichtungsbetaparameter anhand eines “Trainingsdatensatzes” sein.

Anwendungsbeispiel

Um einige Grundprinzipien der multiplen Regressionsanalyse zu verdeutlichen, beschränken wir uns in diesem Abschnitt meist auf den Fall zweier kontinuierlicher unabhängiger Variablen. Als Anwendungsbeispiel betrachten wir dazu den Effekt einer Psychotherapie auf die Depressionssymptomatik in Abhängigkeit vom Alter der Patient:innen und der Dauer der Therapie. Tabelle 44.1 zeigt einen Beispieldatensatz bestehend aus Datenwerten von $n = 20$ Patient:innen, wobei Alter das Patient:innenalter, Dauer die Therapiedauer und dBDI die Pre-Post-Interventions-BDI-Differenzwerte bezeichnen.

Tabelle 44.1. Pre-Post-BDI-Differenzwerte in Abhängigkeit von Patient:innenalter und Therapiedauer

Alter	Dauer	dBDI
36	23	35.4
42	15	17.5
54	20	17.8
74	14	-5.2
32	15	17.7
74	17	-2.4
77	12	-9.7
60	17	10.2
58	22	19.7
24	16	29.3

Tabelle 44.1. Pre-Post-BDI-Differenzwerte in Abhängigkeit von Patient:innenalter und Therapiedauer

Alter	Dauer	dBDI
32	18	26.3
31	19	26.0
61	18	5.8
43	14	11.0
66	22	15.8
50	20	22.0
63	22	11.2
80	13	-8.8
43	21	28.0
67	17	8.4

Abbildung 44.1 visualisiert den Beispieldatensatz, wobei die dBDI Werte der Höhe gemäß im dreidimensionalen Raum anhand ihrer Koordinaten in der Alter-Dauer-Ebene abgetragen sind. Die Visualisierung der multiple Regression, insbesondere für mehr als zwei unabhängige Variablen, stößt sehr schnell an ihre Grenzen.

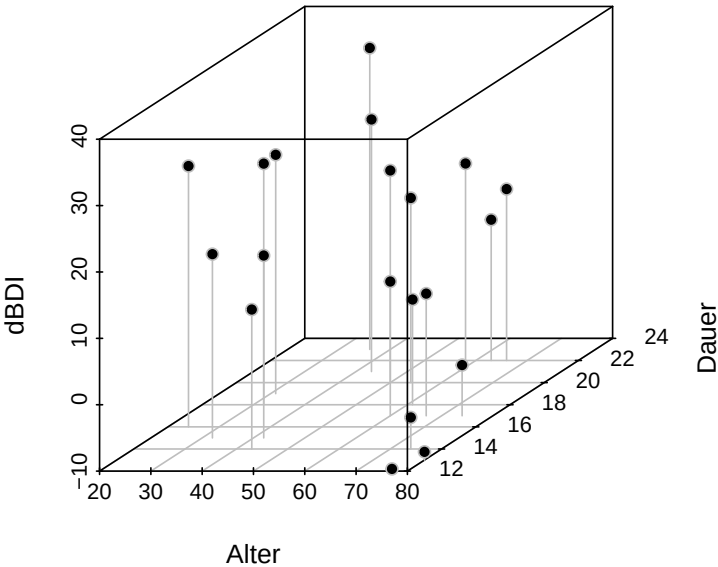


Abbildung 44.1. Visualisierung des Beispieldatensatzes als Punktwolke

44.2. Modellformulierung

Das Modell der multiplen Regression ist mit der allgemeinen Form des Allgemeinen Linearen Modells identisch. Der Vollständigkeit halber geben wir folgende Definition.

Definition 44.1 (Modell der multiplen Regression). y_i mit $i = 1, \dots, n$ sei die Zufallsvariable, die den i ten Wert einer abhängigen Variable modelliert. Dann hat das *Modell der multiplen Regression* die strukturelle Form

$$y_i = x_{i1}\beta_1 + \dots + x_{ip}\beta_p + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \dots, n \text{ und } \sigma^2 > 0,$$

wobei $x_{ij} \in \mathbb{R}$ mit $1 \leq i \leq n$ und $1 \leq j \leq p$ den i ten Wert der j ten unabhängigen Variable bezeichnet. Die unabhängigen Variablen werden auch *Regressoren*, *Prädiktoren*, *Kovariaten* oder *Features* genannt. Mit

$$x_i := (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p \text{ und } \beta := (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$$

hat das Modell der multiplen Regression die Datenverteilungsform

$$y_i \sim N(\mu_i, \sigma^2) \text{ u.v. für } i = 1, \dots, n, \text{ wobei } \mu_i := x_i^T \beta.$$

In diesem Zusammenhang wird $x_i \in \mathbb{R}^p$ auch als *iter Featurevektor* bezeichnet. Die Designmatrixform des Modells der multiplen Regression schließlich ist gegeben durch

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n)$$

mit

$$y := (y_1, \dots, y_n)^T, X := (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p} \in \mathbb{R}^{n \times p}, \beta := (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p \text{ und } \sigma^2 > 0.$$

•

Folgender **R** Code nutzt das Modell der multiplen Regression auf Definition 13.1, um den in Tabelle 44.1 dargestellten Beispieldatensatz zu erzeugen.

```
library(MASS)
set.seed(1)
n      = 20
p      = 3
x_1    = round(runif(n,20,80))
x_2    = round(runif(n,12,24))
X      = matrix(c(rep(1,n),x_1,x_2), nrow = n)
I_n    = diag(n)
beta   = matrix(c(5,-.5,2), nrow = p)
sigsqr = 10
y      = mvrnorm(1, X %*% beta, sigsqr*I_n)
D      = data.frame(Alter = x_1, Dauer = x_2, dBDI = y)
write.csv(D, "../data/511-multiple-regression.csv", row.names = FALSE)
```

Multivariate Normalverteilung
reproduzierbare Daten
Anzahl Datenpunkte
Anzahl Parameter
Regressorwerte Alter
Regressorwerte Therapiedauer
Designmatrix
Identitätsmatrix
Betaparametervektor
Varianzparameter
eine Realisierung eines n-dimensionalen ZVs
Dataframeformatierung
Datenspeichern

44.3. Modellschätzung

Das Modell der multiplen Regression erlaubt eine allgemeine Einsicht in die Bedeutung des Betaparameterschätzers

$$\hat{\beta} := (X^T X)^{-1} X^T y$$

Insbesondere wird bei der Beschäftigung mit dem Betaparameterschätzer im Kontext der multiplen Regression deutlich, dass $X^T y$ die Kovariation der Regressoren mit den Daten und $X^T X$ die Kovariation der Regressoren untereinander quantifiziert, so dass für den Betaparameterschätzer eine Intuition als “regressorkovariationsnormalisierte Regressordatenkovariation” ergibt. Wir wollen dies im Folgenden am Beispiel einer multiplen Regression mit einem Interzeptregressor und zwei Regressoren für zwei unabhängige Variablen vertiefen. Dabei zeigen wir einmal die Form des Betaparameterschätzers in Form von Stichprobenkorrelationen (Theorem 13.1) und einmal in Form von partiellen Stichprobenkorrelationen (Theorem 13.2).

Theorem 44.1 (Betaparameterschätzer und Korrelationen). *Gegeben sei ein multiples Regressionsmodell der Form*

$$y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Dann gilt

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\ \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2} s_y}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}} \\ \frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2} s_y}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_2}} \end{pmatrix},$$

wobei für die y_i, x_{i1} und x_{i2} mit $i = 1, \dots, n$, $\bar{\cdot}$, $s_{\cdot}$ und $r_{\cdot,\cdot}$, die entsprechenden Stichprobenmittel, Stichprobenstandardabweichungen, und Stichprobenkorrelationen bezeichnen.

◦

Beweis. Wir erinnern zunächst daran, dass die Form des Betaparameterschätzers bekanntlich zum System der Normalgleichungen äquivalent ist (vgl. Kapitel 37)

$$\hat{\beta} = (X^T X)^{-1} X^T y \Leftrightarrow X^T X \hat{\beta} = X^T y.$$

Ausschreiben des Normalgleichungssystems für den hier betrachteten Spezialfall ergibt dann zunächst

$$\begin{aligned} X^T X \hat{\beta} &= X^T y \\ \Leftrightarrow \begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ \Leftrightarrow \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1} x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2} x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \\ \sum_{i=1}^n y_i x_{i2} \end{pmatrix} \\ \Leftrightarrow \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1} x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2} x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \\ \sum_{i=1}^n y_i x_{i2} \end{pmatrix} \end{aligned}$$

und damit

$$\begin{aligned} X^T X \hat{\beta} &= X^T y \\ \Leftrightarrow \begin{pmatrix} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} \\ \hat{\beta}_0 \sum_{i=1}^n x_{i2} + \hat{\beta}_1 \sum_{i=1}^n x_{i1} x_{i2} + \hat{\beta}_2 \sum_{i=1}^n x_{i2}^2 \end{pmatrix} &= \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \\ \sum_{i=1}^n y_i x_{i2} \end{pmatrix}. \end{aligned}$$

Aus der Gleichung der ersten Vektorkomponenten folgt dann direkt die Form von $\hat{\beta}_0$ mit

$$\begin{aligned}\sum_{i=1}^n y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} \\ \Leftrightarrow \frac{1}{n} \sum_{i=1}^n y_i &= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \frac{1}{n} \sum_{i=1}^n x_{i2} \\ \Leftrightarrow \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2.\end{aligned}$$

Einsetzen dieser Form von $\hat{\beta}_0$ in die Gleichung der zweiten Vektorkomponenten ergibt dann

$$\begin{aligned}\hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} &= \sum_{i=1}^n y_i x_{i1} \Leftrightarrow \\ (\bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2) \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} &= \sum_{i=1}^n y_i x_{i1} \\ \Leftrightarrow \bar{y} \sum_{i=1}^n x_{i1} - \hat{\beta}_1 \bar{x}_1 \sum_{i=1}^n x_{i1} - \hat{\beta}_2 \bar{x}_2 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} &= \sum_{i=1}^n y_i x_{i1} \\ \Leftrightarrow \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 - \hat{\beta}_1 \bar{x}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} - \hat{\beta}_2 \bar{x}_2 \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n y_i x_{i1} - \bar{y} \sum_{i=1}^n x_{i1} \\ \Leftrightarrow \hat{\beta}_1 \left(\sum_{i=1}^n x_{i1}^2 - \bar{x}_1 \sum_{i=1}^n x_{i1} \right) + \hat{\beta}_2 \left(\sum_{i=1}^n x_{i1}x_{i2} - \bar{x}_2 \sum_{i=1}^n x_{i1} \right) &= \sum_{i=1}^n y_i x_{i1} - \bar{y} \sum_{i=1}^n x_{i1}\end{aligned}$$

Im Beweis des Theorems zur Ausgleichsgerade (vgl. Kapitel 34) haben wir gesehen, dass

$$\begin{aligned}\sum_{i=1}^n x_{i1}^2 - \bar{x}_1 \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1) \\ \sum_{i=1}^n x_{i1}x_{i2} - \bar{x}_2 \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \\ \sum_{i=1}^n y_i x_{i1} - \bar{y} \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n (y_i - \bar{y})(x_{i1} - \bar{x}_1)\end{aligned}$$

Es ergibt sich also, dass

$$\begin{aligned}\hat{\beta}_1 \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1) + \hat{\beta}_2 \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) &= \sum_{i=1}^n (y_i - \bar{y})(x_{i1} - \bar{x}_1) \\ \Leftrightarrow \hat{\beta}_1 \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1)}{n-1} + \hat{\beta}_2 \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{n-1} &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_{i1} - \bar{x}_1)}{n-1}\end{aligned}$$

Mit den Definitionen von Stichprobenstandardabweichung und -korrelation folgt dann weiter

$$\begin{aligned}\hat{\beta}_1 s_{x_1} s_{x_1} + \hat{\beta}_2 c_{x_1, x_2} &= c_{y, x_1} \\ \Leftrightarrow \hat{\beta}_1 \frac{s_{x_1} s_{x_1}}{s_y s_{x_1}} + \hat{\beta}_2 \frac{c_{x_1, x_2}}{s_y s_{x_1}} &= \frac{c_{y, x_1}}{s_y s_{x_1}} \\ \Leftrightarrow \hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{c_{x_1, x_2}}{s_y s_{x_1}} &= r_{y, x_1} \\ \Leftrightarrow \hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{c_{x_1, x_2} s_{x_2}}{s_y s_{x_1} s_{x_2}} &= r_{y, x_1} \\ \Leftrightarrow \hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{s_{x_2}}{s_y} r_{x_1, x_2} &= r_{y, x_1}\end{aligned}$$

Definition von

$$b_j := \frac{s_{x_j}}{s_y}, j = 1, 2$$

erlaubt dann die Schreibweise

$$b_1 + b_2 r_{x_1, x_2} = r_{y, x_1}$$

Schließlich folgt analog durch Vertauschen der Subskripte aus der Gleichung der dritten Vektorkomponenten

$$b_1 r_{x_1, x_2} + b_2 = r_{y, x_2}$$

Insgesamt haben wir also gesehen, dass die Definition des Betaparameterschätzers Spezialfall des Allgemeinen Linearen Modells ergibt, dass mit

$$\hat{\beta}_j = b_j \frac{s_y}{s_{x_j}}, j = 1, 2$$

gilt, dass

$$r_{y,x_1} = b_1 + b_2 r_{x_1,x_2}$$

$$r_{y,x_2} = b_1 r_{x_1,x_2} + b_2$$

Damit folgt aus der zweiten Gleichung dann sofort

$$b_2 = r_{y,x_2} - b_1 r_{x_1,x_2}.$$

Einsetzen in die erste Gleichung ergibt dann

$$\begin{aligned} b_1 + (r_{y,x_2} - b_1 r_{x_1,x_2}) r_{x_1,x_2} &= r_{y,x_1} \\ \Leftrightarrow b_1 + r_{y,x_2} r_{x_1,x_2} - b_1 r_{x_1,x_2}^2 &= r_{y,x_1} \\ \Leftrightarrow r_{y,x_2} r_{x_1,x_2} + b_1 (1 - r_{x_1,x_2}^2) &= r_{y,x_1} \\ \Leftrightarrow b_1 (1 - r_{x_1,x_2}^2) &= r_{y,x_1} - r_{y,x_2} r_{x_1,x_2} \\ \Leftrightarrow b_1 &= \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \end{aligned}$$

Für b_2 ergibt sich damit weiterhin

$$\begin{aligned} b_2 &= r_{y,x_2} - b_1 r_{x_1,x_2} \\ \Leftrightarrow b_2 &= r_{y,x_2} - \left(\frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \right) r_{x_1,x_2} \\ \Leftrightarrow b_2 &= \frac{r_{y,x_2} (1 - r_{x_1,x_2}^2)}{1 - r_{x_1,x_2}^2} - \frac{r_{y,x_1} r_{x_1,x_2} - r_{y,x_2} r_{x_1,x_2}^2}{1 - r_{x_1,x_2}^2} \\ \Leftrightarrow b_2 &= \frac{r_{y,x_2} - r_{y,x_2} r_{x_1,x_2}^2 - r_{y,x_1} r_{x_1,x_2} + r_{y,x_2} r_{x_1,x_2}^2}{1 - r_{x_1,x_2}^2} \\ \Leftrightarrow b_2 &= \frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \end{aligned}$$

Damit folgen dann aber

$$\begin{aligned} \hat{\beta}_1 &= b_1 \frac{s_y}{s_{x_1}} = \left(\frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \right) \frac{s_y}{s_{x_1}} \\ \hat{\beta}_2 &= b_2 \frac{s_y}{s_{x_2}} = \left(\frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \right) \frac{s_y}{s_{x_2}} \end{aligned}$$

und es ist alles gezeigt. $\square$

Exemplarisch wollen wir die Stichprobenkorrelationsform des Betaparameterschätzers für β_1

$$\hat{\beta}_1 = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}}$$

betrachten. Man erkennt unter anderem:

- Nur im Fall $r_{x_1,x_2} = 0$ und $s_y = s_{x_1}$ gilt $\hat{\beta}_1 = r_{y,x_1}$. Die Betaparameterschätzer der multiplen Regression sind also im Allgemeinen nicht mit den Stichprobenkorrelationen zwischen dem entsprechenden Regressor und den Daten identisch.
- Im Fall $r_{x_1,x_2} = \pm 1$ ist $\hat{\beta}_1$ nicht definiert. Vollständig korrelierte Regressoren ergeben also ein nicht schätzbares Modell.
- Je größer $|r_{x_1,x_2}|$, desto größer der von r_{y,x_1} subtrahierte Term $r_{y,x_2} r_{x_1,x_2}$. Die paarweise Korrelation von Regressoren untereinander reduziert also den Betaparameterschätzerwert für einen Regressor.

- Je größer $|r_{y,x_2}|$, desto größer der von r_{y,x_1} subtrahierte Term $r_{y,x_2}r_{x_1,x_2}$. Neben der paarweisen Korrelation von Regressoren untereinander reduziert also auch eine hohe Stichprobenkorrelation eines anderen Regressors mit den Daten den Wert eines Betaparameterschätzers.
- Bei identischen Korrelationen und gleich bleibender Regressorstandardabweichung steigt $\hat{\beta}_1$ mit s_y . Intuitiv erhält also ein Regressor einen höheren Betaparameterschätzerwert, wenn er bei gleich Korrelationsstruktur mehr Datenvariabilität erklärt.

Zusammengefasst verdeutlichen obige Punkte, dass der Wert des Betaparameterschätzers eines Regressors im Modell der multiplen Regression allein im Kontext der Korrelationsstruktur der anderen Regressoren mit sich selbst und den Daten betrachtet werden kann und damit keinen modellunabhängigen Absolutbeitrag eines spezifischen Regressors zur Erklärung der Daten abbildet.

Folgender **R** Code demonstriert die Äquivalenz der Matrixschreibweise des Betaparameterschätzers und seiner Darstellung durch Stichprobenmittel, Stichprobenkorrelationen, und Stichprobenstandardabweichungen.

Folgendes Theorem stellt den Bezug zwischen multipler Regression und partiellen Korrelationen zwischen unabhängigen und abhängigen Variablen für das betrachtete Szenario her.

Theorem 44.2 (Betaparameterschätzer und partielle Korrelationen). *Gegeben sei ein multiples Regressionsmodell der Form*

$$y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Dann gilt

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\ r_{y,x_1 \setminus x_2} \sqrt{\frac{1-r_{y,x_2}^2}{1-r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}} \\ r_{y,x_2 \setminus x_1} \sqrt{\frac{1-r_{y,x_1}^2}{1-r_{x_2,x_1}^2}} \frac{s_y}{s_{x_2}} \end{pmatrix},$$

wobei für $1 \leq k, l \leq 2$ und $i = 1, \dots, n$

- $r_{y,x_k \setminus x_l}$ die partielle Stichprobenkorrelation der y_i und x_{ik} gegeben die x_{il} ist,
- r_{y,x_k} die Stichprobenkorrelation der y_i und x_{ik} ist, und
- r_{x_k,x_l} die Stichprobenkorrelation der x_{ik} und x_{il} ist.

◦

Beweis. Wir betrachten $\hat{\beta}_1$, das Resultat für $\hat{\beta}_2$ folgt dann durch Vertauschen der Indizes. Wir haben in vorherigem Theorem gesehen, dass

$$\hat{\beta}_1 = \frac{r_{y,x_1} - r_{y,x_2}r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}}$$

Weiterhin haben wir in Kapitel 43 gesehen, dass unter der Annahme der multivariaten Normalverteilung von y, x_1, x_2 ein Schätzer für die partielle Korrelation von y und x_1 gegeben x_2 durch

$$r_{y,2} = \frac{r_{y,x_1} - r_{y,x_2}r_{x_1,x_2}}{\sqrt{1-r_{y,x_2}^2}\sqrt{1-r_{x_1,x_2}^2}}$$

gegeben ist. Für $\hat{\beta}_1$ ergibt sich somit

$$\begin{aligned}\hat{\beta}_1 &= \frac{r_{y,x_1} - r_{y,x_2}r_{x_1,x_2}}{1-r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}} \\ \Leftrightarrow (1-r_{x_1,x_2}^2)\hat{\beta}_1 &= (r_{y,x_1} - r_{y,x_2}r_{x_1,x_2}) \frac{s_y}{s_{x_1}} \\ \Leftrightarrow \frac{1-r_{x_1,x_2}^2}{\sqrt{1-r_{y,x_2}^2}\sqrt{1-r_{x_1,x_2}^2}}\hat{\beta}_1 &= \frac{r_{y,x_1} - r_{y,x_2}r_{x_1,x_2}}{\sqrt{1-r_{y,x_2}^2}\sqrt{1-r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}} \\ \Leftrightarrow \frac{1-r_{x_1,x_2}^2}{\sqrt{1-r_{y,x_2}^2}\sqrt{1-r_{x_1,x_2}^2}}\hat{\beta}_1 &= r_{y,x_1|x_2} \frac{s_y}{s_{x_1}}\end{aligned}$$

und damit weiter

$$\begin{aligned}\hat{\beta}_1 &= r_{y,x_1|x_2} \frac{\sqrt{1-r_{y,x_2}^2}\sqrt{1-r_{x_1,x_2}^2}}{1-r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}} \\ \Leftrightarrow \hat{\beta}_1 &= r_{y,x_1|x_2} \frac{\sqrt{1-r_{y,x_2}^2}\sqrt{1-r_{x_1,x_2}^2}}{\left(\sqrt{1-r_{x_1,x_2}^2}\right)^2} \frac{s_y}{s_{x_1}} \\ \Leftrightarrow \hat{\beta}_1 &= r_{y,x_1|x_2} \frac{\sqrt{1-r_{y,x_2}^2}}{\sqrt{1-r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}} \\ \Leftrightarrow \hat{\beta}_1 &= r_{y,x_1|x_2} \sqrt{\frac{1-r_{y,x_2}^2}{1-r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}}\end{aligned}$$

□

Im Allgemeinen gilt für $1 \leq i, l \leq k$, also dass

$$\hat{\beta}_k \neq r_{y,x_k|x_l}.$$

Betaparameterschätzer sind also im Allgemeinen keine partiellen Stichprobenkorrelationen. Allerdings gilt

$$\hat{\beta}_k = r_{y,x_k|x_l} \text{ für } 1 \leq i, l \leq k$$

genau dann, wenn

$$s_y = s_{x_1} = s_{x_2}$$

und außerdem

$$r_{y,x_k} = r_{x_k,x_l} = 0$$

wenn also die Stichprobenkorrelationen der Daten und der Werte des zweiten Regressors, sowie die Stichprobenkorrelation der Werte der beiden Regressoren gleich Null sind. Dies kann der Fall sein, wenn einer der Regressoren die Daten “sehr gut erklärt” und der andere Regressor von dem ersten “sehr verschieden” ist. Schließlich gilt obige Identität von Betaparameterschätzerkomponente und partieller Stichprobenkorrelation auch dann, wenn

$$|r_{y,x_l}| = |r_{x_k,x_l}|,$$

wenn also die obigen Stichprobenkorrelationen dem Betrage nach gleich sind. Dies ist in der Anwendung aber vermutlich selten der Fall.

Folgender **R** Code demonstriert die Äquivalenz der Matrixschreibweise des Betaparameterschätzers und seiner Darstellung durch Stichprobenmittel, Stichprobenkorrelationen, Stichprobenstandardabweichungen und partielle Stichprobenkorrelationen.

```
D = read.csv("./_data/511-multiple-regression.csv") # Dateneinlesen
y = D$dBDI # Abhängige Variable
n = length(y) # Anzahl Datenpunkte
X = matrix(c(rep(1,n), D$Alter, D$Dauer), nrow = n) # Designmatrix
p = ncol(X) # Anzahl Parameter
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y # Betaparameterschätzer
eps_hat = y - X %*% beta_hat # Residuenvektor
sigsqr_hat = (t(eps_hat) %*% eps_hat) / (n-p) # Varianzparameterschätzer

# Betaparameterschätzer aus partiellen Korrelationen und Korrelationen
library(ppcor) # partielle Korrelationentoolbox
y12 = cbind(y,X[,2:3]) # y,x_1,x_2 Matrix
bars = apply(y12, 2, mean) # Stichprobenmittel
s = apply(y12, 2, sd) # Stichprobenstandardabweichungen
r = cor(y12) # Stichprobenkorrelationen
pr = pcor(y12) # partielle Stichprobenkorrelationen
pr = pr$estimate # partielle Stichprobenkorrelationen
beta_hat_1 = pr[1,2]*sqrt((1-r[1,3]^2)/(1-r[2,3]^2))*(s[1]/s[2]) # \hat{\beta}_1
beta_hat_2 = pr[1,3]*sqrt((1-r[1,2]^2)/(1-r[3,2]^2))*(s[1]/s[3]) # \hat{\beta}_2
beta_hat_0 = bars[1] - beta_hat_1*bars[2] - beta_hat_2*bars[3] # \hat{\beta}_0

Korrelationen r(dPOMS,AGE),r(dPOMS,BRI),r(AGE,BRI) : -0.87 0.64 -0.22
Partielle Korrelationen r(dPOMS,AGE|BRI), r(dPOMS,BRI|AGE) : -0.97 0.92
beta_hat ALM Schätzer : 11.44 -0.57 1.85
beta_hat aus partieller Korrelation : 11.44 -0.57 1.85
```

Abbildung 44.2 visualisiert den Beispieldatensatz zusammen mit der durch den Betaparameterschätzer definierten Regressionsebene.

$$f_{\hat{\beta}} : \mathbb{R}^2 \rightarrow \mathbb{R}, (x_1, x_2) \mapsto f_{\hat{\beta}}(x_1, x_2) := \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

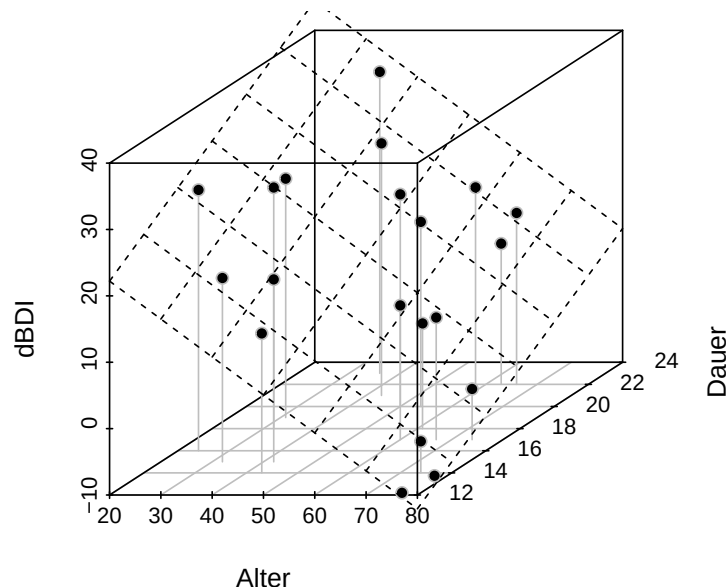


Abbildung 44.2. Visualisierung von Beispieldatensatz und Regressionsebene

44.4. Modellevaluation

Die allgemeine Theorie der T- und F-Statistiken bietet eine Vielzahl von Möglichkeiten, verschiedenste Hypothesen im Kontext der multiplen Regression inferenzstatistisch zu evaluieren. Für das Anwendungsbeispiel könnte zum Beispiel folgende Auswahl von Kontrastgewichtsvektoren und Null- und Alternativhypothesen zunächst von Interesse sein:

$$c := \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \Leftrightarrow H_0 : \beta_2 = 0, H_A : \beta_2 \neq 0 \text{ und } c := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \Leftrightarrow H_0 : \beta_3 = 0, H_A : \beta_3 \neq 0$$

Dabei würde das Ablehnen der jeweiligen Nullhypothese jeweils inferenzstatistische Evidenz für einen Effekt des Patient:innenalters bzw. der Therapiedauer auf die Pre-PostBDI-Differenz im Kontext der Präsenz der jeweils anderen unabhängigen Variable und des Interzeptterms implizieren.

Weiterhin könnte folgender Kontrastgewichtsvektor mit folgenden Null- und Alternativhypothesen von Interesse sein:

$$c := \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \Leftrightarrow H_0 : \beta_2 - \beta_3 = 0, H_A : \beta_2 - \beta_3 \neq 0$$

In diesem Fall würde das Ablehnen der Nullhypothese inferenzstatistische Evidenz für einen differentiellen Einfluss von Patient:innenalter und Therapiedauer implizieren, je nach Vorzeichen der T-Statistik dabei einen stärkeren Effekt des Patient:innenalters oder der Therapiedauer.

Folgender **R** Code evaluiert die angesprochenen T-Tests.

```
D = read.csv("./_data/511-multiple-regression.csv") # Datensatz
y = D$dBDI # Abhängige Variable
n = length(y) # Anzahl Datenpunkte
X = matrix(c(rep(1,n), D$Alter, D$Dauer), nrow = n) # Designmatrix
p = ncol(X) # Anzahl Parameter
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y # Betaparameterschätzer
eps_hat = y - X %*% beta_hat # Residuenvektor
sigsqr_hat = (t(eps_hat) %*% eps_hat) / (n-p) # Varianzparameterschätzer
C = cbind(diag(p), matrix(c(0,1,-1), nrow = 3)) # Kontrastgewichtsvektoren
ste = rep(NA, ncol(C)) # Kontraststandardfehler
tee = rep(NA, ncol(C)) # T-Statistiken
pvals = rep(NA, ncol(C)) # p-Werte
for(i in 1:ncol(C)){ # Kontrastiterationen
  c = C[,i] # Kontrastgewichtsvektor
  t_num = t(c) %*% beta_hat # Zähler der T-Statistik
  ste[i] = sqrt(sigsqr_hat * t(c) %*% solve(t(X) %*% X) %*% c) # Kontraststandardfehler/Nenner der T-Statistik
  tee[i] = t_num / ste[i] # T-Statistik
  pvals[i] = 2 * (1 - pt(abs(tee[i]), n-p)) # p-Wert
}
```

	Estimate	Std. Error	t value	Pr(> t)
(Interzept)	11.44	4.20	2.73	0.01
Alter	-0.57	0.04	-16.00	0.00
Dauer	1.85	0.19	9.97	0.00
Alter-Dauer	-2.42	0.18	-13.37	0.00

Der Beispieldatensatz liefert also inferenstatistische Inferenz für einen negativen Zusammenhang zwischen Therapieerfolg und Patient:innenalter und einen positiven Zusammenhang zwischen Therapieerfolg und Therapiedauer. Die Differenz der beiden Effekte ist dabei deutlich ausgeprägt.

Etwas globaler könnte man einen F-Test basierend auf einer Modellpartition mit $p_0 := 1$ nutzen, um etwa zu klären, ob die Daten die Annahme, dass Patient:innenalter und Therapie überhaupt zur Erklärung der Pre-Post-BDI-Differenzwertvariation, bestätigen. Folgender **R** Code evaluiert den angesprochenen F-Test.

```
D = read.csv("./_data/511-multiple-regression.csv") # Datensatz
y = D$dBDI # Abhängige Variable
n = length(y) # Anzahl Datenpunkte
X = matrix(c(rep(1,n), D$Alter, D$Dauer), nrow = n) # Designmatrix vollständiges Modell
p = ncol(X) # Anzahl Parameter vollständiges Modell
p_0 = 1 # Anzahl Parameter reduziertes Modell
p_1 = p - p_0 # Anzahl zusätzlicher Parameter im vollst. Modell
X_0 = X[,1:p_0] # Designmatrix reduzierten Modells
beta_hat_0 = solve(t(X_0)%*%X_0)%*%t(X_0)%*%y # Betaparameterschätzer reduziertes Modell
beta_hat = solve(t(X)%*%X)%*%t(X)%*%y # Betaparameterschätzer vollständiges Modell
eps_hat_0 = y - X_0%*%beta_hat_0 # Residuenvektor reduziertes Modell
eps_hat = y - X%*%beta_hat # Residuenvektor vollständiges Modell
eh0_eh0 = t(eps_hat_0)%*%eps_hat_0 # RQS reduziertes Modell
eh_eh = t(eps_hat)%*%eps_hat # RQS vollständiges Modell
sigsqr_hat = eh_eh/(n-p) # Varianzparameterschätzer vollst. Modell
f = ((eh0_eh0-eh_eh)/(p_1))/sigsqr_hat # F-Statistik
pval = 1 - pf(f,p_1,n-p) # p-Wert
```

F-statistic: 223.7 on 2 and 17 DF p-value: 6.17950135506362e-13

Der Beispieldatensatz enthält also deutliche Evidenz dafür, dass sowohl Patient:innenalter als auch Therapiedauer zur Erklärung der Pre-Post-BDI-Differenzwertvariation über Patient:innen beiträgt.

Eine direkte Implementation obiger Analyse erlaubt das Zusammenspiel der **R** Funktionen `lm()` und `summary()`, wie untenstehender **R** Code demonstriert.

```
D = read.csv("./_data/511-multiple-regression.csv") # Datensatz
alm = lm(dBDI ~ Alter + Dauer, data = D) # Modellformulierung und Modellschätzung
summary(alm) # Modellevaluation
```

```
Call:
lm(formula = dBDI ~ Alter + Dauer, data = D)

Residuals:
    Min       1Q   Median       3Q      Max
-4.9474 -2.0549  0.6748  1.9590  3.7926

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.43949    4.19776   2.725  0.0144 *
Alter        -0.57162    0.03573  -15.997 1.11e-11 ***
Dauer         1.85132    0.18577   9.966 1.63e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.612 on 17 degrees of freedom
Multiple R-squared:  0.9634,    Adjusted R-squared:  0.9591
F-statistic: 223.7 on 2 and 17 DF,  p-value: 6.179e-13
```

44.5. Literaturhinweise

Die moderne Geschichte der multiplen Regression wird oft in den Arbeiten von Legendre (1805) und Gauss (1809) zur Berechnung von Planetenbahnen verankert. Die heutige Theorie der multiplen Regression sehr breit gefächert, so dass wir in diesem Abschnitt nur einen ersten Eindruck von ihrem Wesen vermitteln können. Weiterführende Einsichten bieten zum Beispiel Draper & Smith (1998), Hocking (2003) und Fox & Tishby (2016).

Teil VI.

Multivariate Inferenz

45. Multivariate Deskriptivstatistiken

45.1. Datenanalyseszenarien

Wir wollen hier zunächst die im folgenden zu betrachtenden Datenanalyseverfahren anhand der Dimensionalität ihrer unabhängigen Variablen und abhängigen Variablen klassifizieren. Dazu bezeichnen wir wie üblich eine unabhängige Variable mit x und eine abhängige Variable mit y . Weiterhin sei mit den Subskripten i und j bei x_{ij} und y_{ij} der Wert der j ten univariaten Komponente der jeweiligen Variable, beispielsweise ein Testwert, bei der i ten experimentellen Einheit, beispielsweise einer Proband:in, bezeichnet. Die Gesamtzahl experimenteller Einheiten sei mit n bezeichnet.

Tabelle 45.1 zeigt das Szenario einer univariaten unabhängigen Variable und einer univariaten abhängigen Variable. Typische in diesem Szenario genutzte Inferenzverfahren sind die Bestimmung der *Korrelation* von x_1 und y_1 , die Durchführung einer *einfachen linearen Regression* von y_1 auf x_1 und, wenn x_1 eine kategoriale Faktorvariable ist, *T-Tests* und *Varianzanalysen*.

Tabelle 45.1. Univariete unabhängige Variable x und univariate abhängige Variable y

x_1	y_1
x_{11}	y_{11}
$\vdots$	$\vdots$
x_{n1}	y_{n1}

Tabelle 45.2 zeigt das Szenario einer multivariaten unabhängigen Variablen und einer univariaten abhängigen Variablen. Typische in diesem Szenario eingesetzte Inferenzverfahren sind die Bestimmung von *multiplen und partiellen Korrelationen* zwischen $x_1, \dots, x_m$ und y_1 , die Durchführung von *multiplen Regressionsanalysen* und *Kovarianzanalysen* und generell alle datenanalytischen Spezialfälle des *Allgemeinen Linearen Modells*.

Tabelle 45.2. Multivariate unabhängige Variable x und univariate abhängige Variable y

x_1	$\cdots$	x_m	y_1
x_{11}	$\cdots$	x_{1m}	y_{11}
$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{n1}	$\cdots$	x_{nm}	y_{nm}

Tabelle 45.3 zeigt das im Kontext von *Einstichproben- T^2 -Tests*, der *einfaktoriiellen multivariaten Varianzanalyse* und vielen Szenarien der *prädiktiven Modellierung* relevante Szenario. In diesem Fall ist die unabhängige Variable univariat und kodiert kategorial ein Faktorlevel bzw. eine Gruppenzugehörigkeit, während die abhängige Variable multivariat ist. Insbesondere in prädiktiven Modellierung wird die unabhängige Variable in diesem Kontext auch als *Targetvariable* oder *Label* und die Komponenten der abhängigen Variable als *Features* bezeichnet.

Tabelle 45.3. Univariante unabhängige Variable x und multivariate abhängige Variable y

x_1	y_1	$\cdots$	y_m
x_{11}	y_{11}	$\cdots$	y_{1m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{n1}	y_{n1}	$\cdots$	y_{nm}

Tabelle 45.4 schließlich zeigt das Szenario einer multivariaten unabhängigen Variablen und einer multivariaten abhängigen Variablen. Dies ist das datenanalytische Szenario, das wir im Rahmen der *Kanonischen Korrelationsanalyse* genauer betrachten wollen und das generell durch das *Multivariate Allgemeine Lineare Modell* abgebildet und datenanalytisch behandelt werden kann.

Tabelle 45.4. Multivariate unabhängige Variable x und multivariate abhängige Variable y

x_1	$\cdots$	x_{m_x}	y_1	$\cdots$	y_{m_y}
x_{11}	$\cdots$	x_{1m_x}	y_{11}	$\cdots$	y_{1m_y}
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_{n1}	$\cdots$	x_{nm_x}	y_{n1}	$\cdots$	y_{nm_y}

45.2. Deskriptivstatistiken

Wir wollen nun einige Standarddeskriptivstatistiken zur Beschreibung multivariater Datensätze diskutieren. Dazu verallgemeinern wir zunächst die aus der univariaten Deskriptivstatistik bekannten Begriffe des Stichprobenmittels und der Stichprobenvarianz und betrachten dann mit den Mahalanobis-Distanzen multivariate Maße für das Verhältnis von Signal zu Rauschen. Wie immer entwickeln sich diese Begriffe vor dem Hintergrund der Annahme, dass es sich bei beobachteten Daten um Realisierungen entsprechender Zufallsvektoren handelt. Im Gegensatz zu der in Kapitel 45.1 betrachteten und aus dem empirischen Kontext bekannten Organisation von Daten experimenteller Einheiten in Zeilen und ihrer jeweiligen abhängigen Variablenkomponenten in Spalten ist dabei eine Organisation der zu einer experimentellen Einheit gehörenden Variablenkomponenten in Spaltenform zielführender und mit den Schreibweisen des univariaten Falles konsistenter.

Stichprobenmittel, -kovarianzmatrix, und -korrelationsmatrix

Definition 45.1 (Stichprobenmittel, -kovarianzmatrix und -korrelationsmatrix). $y_1, \dots, y_n$ sei eine Menge von m -dimensionalen Zufallsvektoren, genannt *Stichprobe*.

- Das *Stichprobenmittel* der $y_1, \dots, y_n$ ist definiert als der m -dimensionale Vektor

$$\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i. \quad (45.1)$$

- Die *Stichprobenkovarianzmatrix* der $y_1, \dots, y_n$ ist definiert als die $m \times m$ Matrix

$$C := \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T. \quad (45.2)$$

- Die *Stichprobenkorrelationsmatrix* der $y_1, \dots, y_n$ ist definiert als die $m \times m$ Matrix

$$D := \left(\frac{(C)_{ij}}{\sqrt{(C)_{ii}} \sqrt{(C)_{jj}}} \right)_{1 \leq i, j \leq m}. \quad (45.3)$$

•

Ohne Beweis halten wir fest, dass analog zum univariaten Fall das Stichprobenmittel bei unabhängig und identisch verteilten Zufallsvektoren $y_1, \dots, y_n$ ein unverzerrter Schätzer des Stichprobenvariablen Erwartungswerts $\mathbb{E}(y_i) \in \mathbb{R}^m, i = 1, \dots, n$ ist. Ebenso ist in diesem Fall die Stichprobenkovarianzmatrix ein unverzerrter Schätzer der Stichprobenvariablenkovarianzmatrix $\mathbb{C}(y_i) \in \mathbb{R}^m, i = 1, \dots, n$. Zur konkreten Berechnung von Stichprobenmittel, Stichprobenkovarianzmatrix und Stichprobenkorrelationsmatrix basierend auf einem Datensatz bieten sich die Aussagen des folgenden Theorems an.

Theorem 45.1 (Datenmatrix und Stichprobenstatistiken).

Es sei

$$y := \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix} \quad (45.4)$$

eine $m \times n$ Datenmatrix, die durch die spaltenweise Konkatination von n m -dimensionalen Zufallsvektoren $y_1, \dots, y_n$ gegeben sei. Dann ergeben sich

- für das Stichprobenmittel

$$\bar{y} = \frac{1}{n} y 1_n, \quad (45.5)$$

- für die Stichprobenkovarianzmatrix

$$C = \frac{1}{n-1} \left(y \left(I_n - \frac{1}{n} 1_{nn} \right) y^T \right), \quad (45.6)$$

- und für Stichprobenkorrelationsmatrix mit

$$D := \text{diag} \left(\sqrt{C_{y_{ii}}}^{-1}, i = 1, \dots, m \right), \quad (45.7)$$

dass

$$R = DCD. \quad (45.8)$$

◦

Beweis. Die Darstellung des Stichprobenmittels ergibt sich aus

$$\begin{aligned} \bar{y} &:= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n y_{i1} \\ \vdots \\ \sum_{i=1}^n y_{im} \end{pmatrix} \\ &= \frac{1}{n} \left(\begin{pmatrix} y_{11} & \cdots & y_{n1} \\ \vdots & \ddots & \vdots \\ y_{1m} & \cdots & y_{nm} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right) \\ &= \frac{1}{n} y 1_n. \end{aligned} \quad (45.9)$$

Hinsichtlich der Darstellung der Stichprobenkovarianzmatrix halten wir zunächst fest, dass nach Definition gilt, dass

$$\begin{aligned} C &:= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T \\ &= \frac{1}{n-1} \sum_{i=1}^n (y_i y_i^T - y_i \bar{y}^T - \bar{y} y_i^T + \bar{y} \bar{y}^T) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n y_i y_i^T - \sum_{i=1}^n y_i \bar{y}^T - \sum_{i=1}^n \bar{y} y_i^T + \sum_{i=1}^n \bar{y} \bar{y}^T \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n y_i y_i^T - n \bar{y} \bar{y}^T - n \bar{y} \bar{y}^T + n \bar{y} \bar{y}^T \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n y_i y_i^T - n \bar{y} \bar{y}^T \right). \end{aligned} \quad (45.10)$$

Mit $1_n 1_n^T = 1_{nn}$ ergibt sich dann weiterhin

$$\begin{aligned} y \left(I_n - \frac{1}{n} 1_{nn} \right) y^T &= \left(y I_n - \frac{1}{n} y 1_{nn} \right) y^T \\ &= y y^T - \frac{1}{n} y 1_{nn} y^T \\ &= \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix} \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix} - \frac{1}{n} y 1_n 1_n^T y^T \\ &= \sum_{i=1}^n y_i y_i^T - n \left(\frac{1}{n} y 1_n \right) \left(\frac{1}{n} 1_n^T y^T \right) \\ &= \sum_{i=1}^n y_i y_i^T - n \bar{y} \bar{y}^T \\ &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T \\ &= C. \end{aligned} \quad (45.11)$$

Hinsichtlich der Korrelationsmatrix ergibt sich nach Definition und für ein beliebiges Indexpaar i, j mit $1 \leq i, j \leq m$ schließlich, dass

$$\begin{aligned} R_{y_{ij}} &= \frac{(C)_{ij}}{\sqrt{(C)_{ii}} \sqrt{(C)_{jj}}} \\ &= \frac{1}{\sqrt{(C)_{ii}}} (C)_{ij} \frac{1}{\sqrt{(C)_{jj}}} \\ &= (DCD)_{ij}. \end{aligned} \quad (45.12)$$

□

Folgender **R** Code wendet die in Theorem 45.1 diskutierten Resultate auf einen Beispieldatensatz mit Datendimensionalität $m = 4$ und Anzahl experimenteller Einheiten $n := 12$ an.

```
# Simulation eines Datensatzes
library(MASS)
m = 4
n = 12
mu = matrix(c(1,2,3,4))
Sigma = diag(4)
Y = t(mvrnorm(n, mu, Sigma))

# multivariate Normalverteilung
# Datenvektordimension
# Anzahl Datenvektoren
# Erwartungswertparameter
# Kovarianzmatrixparameter
# Datensatzrealisierung

[1,] 0.380757 3.206102 0.7449730 -0.4244947 0.8556004 1.207538 3.3079784
[2,] 1.681932 1.070638 0.5125397 0.9248077 3.0000288 1.378733 0.6155732
[3,] 1.085641 4.176583 1.3350276 2.5364696 1.8840799 2.249181 5.0871665
[4,] 3.318340 3.675730 4.0601604 3.4111055 4.5314962 2.481606 4.3065579
[1,] 1.105802 1.456999 0.9228471 0.6659992 0.965274
[2,] 3.869291 2.425100 1.7613529 3.0584830 2.886423
[3,] 3.017396 1.713699 1.3593945 3.4501871 2.981440
[4,] 2.463550 3.699024 3.4717201 3.3479052 3.943103

# Auswertung von Deskriptivstatistiken
n = ncol(Y)
I_n = diag(n)
J_n = matrix(rep(1,n^2), nrow = n)
y_bar = (1/n) * Y %*% J_n[,1]
C = (1/(n-1)) * (Y %*% (I_n - (1/n)*J_n) %*% t(Y))
D = diag(1/sqrt(diag(C)))
R = D %*% C %*% D

# Anzahl Datenvektorealisierungen
# Einheitsmatrix I_n
# 1_{nn}
# Stichprobenmittel
# Stichprobenkovarianzmatrix
# Kov-Korr-Transformationsmatrix
# Stichprobenkorrelationsmatrix

[1,] 1.199615
[2,] 1.932075
[3,] 2.573022
[4,] 3.559191

[1,] 1.1450788 -0.29291622 0.91837375 0.16929650
[2,] -0.2929162 1.20170927 -0.09630563 -0.16923746
[3,] 0.9183737 -0.09630563 1.50573951 0.08718263
[4,] 0.1692965 -0.16923746 0.08718263 0.40266165

[1,] 1.0000000 -0.24970431 0.69940198 0.2493218
[2,] -0.2497043 1.00000000 -0.07159407 -0.2432913
[3,] 0.6994020 -0.07159407 1.00000000 0.1119657
[4,] 0.2493218 -0.24329134 0.11196568 1.0000000
```

Mahalanobis-Distanzen

Abschließend wollen wir mit den sogenannten *Mahalanobis-Distanzen* multivariate Generalisierungen von aus der univariaten Anwendung bekannten Signal-zu-Rauschen-Maßen betrachten. Wir definieren den Begriff der Mahalanobis-Distanz wie folgt.

Definition 45.2 (Mahalanobis-Distanzen). x_1 sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel, x_2 sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel und X sei eine Kovarianzmatrix oder eine Stichprobenkovarianzmatrix. Dann heißt

$$D = (x_1 - x_2)^T X^{-1} (x_1 - x_2) \quad (45.13)$$

Mahalanobis-Distanz von x_1 und x_2 hinsichtlich X .

•

Eine Mahalanobis-Distanz ist damit eine durch eine Kovarianzmatrix normalisierte quadrierte Euklidische Distanz (vgl. Kapitel 8.2). Ähnliche Maße für das Verhältnis eines Abstandes und einer Variabilität sind bekanntlich die z -Transformation $z = (y - \mu)/\sigma$ für $y \in \mathbb{R}$ und die Parameter $\mu, \sigma^2 > 0$ einer univariaten Normalverteilung sowie Cohen's $d = (\bar{y}_1 - \bar{y}_2)/s_{12}$ für zwei Stichprobenmittel $\bar{y}_1$ und $\bar{y}_2$ und ihre korrespondierende gepoolte Stichprobenstandardabweichung s_{12} . Im Unterschied zur Mahalanobis-Distanz sind diese Maße allerdings nicht quadriert und damit Vorzeichen behaftet. In Analogie zur z -Transformation oder zu Cohen's d wird allerdings auch bei Mahalanobis-Distanzen ein Abstand in Einheiten von Variabilität gemessen. Bei Cohen's d bedeutet ja ein Wert von $d = 1$ gerade, dass der Abstand von $\bar{y}_1$ und $\bar{y}_2$ eine gepoolte Standardabweichung beträgt. Ebenso verhält es sich mit den Mahalanobis-Distanzen.

Anhand von Abbildung 45.1 und Abbildung 45.2 wollen wir den Einfluss der Varianz und der Kovarianz von Komponenten der x_1 und x_2 auf ihre Mahalanobis-Distanz noch etwas genauer betrachten. Die Titel der Unterabbildungen von Abbildung 45.1 zeigen die Mahalanobis-Distanzen der Vektoren $x_1 := (-1, -1)^T$ und $x_2 := (1, 1)^T$ bei Kovarianzmatrizen von

$$\Sigma_1 := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}, \Sigma_2 := \begin{pmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{pmatrix} \text{ und } \Sigma_3 := \begin{pmatrix} 1.5 & 0.0 \\ 0.0 & 1.5 \end{pmatrix}, \quad (45.14)$$

die mithilfe von Normalverteilungsisokonturen dargestellt sind. Für Σ_1 entspricht die Mahalanobis-Distanz dabei der quadrierten Euklidischen Distanz von x_1 und x_2 . An der Darstellung zu Σ_2 erkennt man, dass im Fall sphärischer Kovarianzmatrizen eine geringere Komponentenvarianz von x_1 und x_2 zu einer größeren Mahalanobis-Distanz führt. Umgekehrt erkennt man an der Darstellung zu Σ_3 , dass im Fall sphärischer Kovarianzmatrizen eine höhere Komponentenvarianz von x_1 und x_2 in einer kleineren Mahalanobis-Distanz resultiert. Intuitiv nähert die Komponentenvarianz die Komponenten also an.

Die Titel der Unterabbildungen von Abbildung 45.2 zeigen die Mahalanobis-Distanzen derselben Vektoren bei Kovarianzmatrizen von

$$\Sigma_1 := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}, \Sigma_2 := \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix} \text{ und } \Sigma_3 := \begin{pmatrix} 1.0 & -0.9 \\ -0.9 & 1.0 \end{pmatrix}, \quad (45.15)$$

Für Σ_1 entspricht dabei wiederum die die Mahalanobis-Distanz der quadrierten Euklidischen Distanz von x_1 und x_2 . An der Darstellung zu Σ_2 erkennt man, dass eine stark positive Kovarianz der Komponenten von x_1 und x_2 in einer kleineren Mahalanobis-Distanz

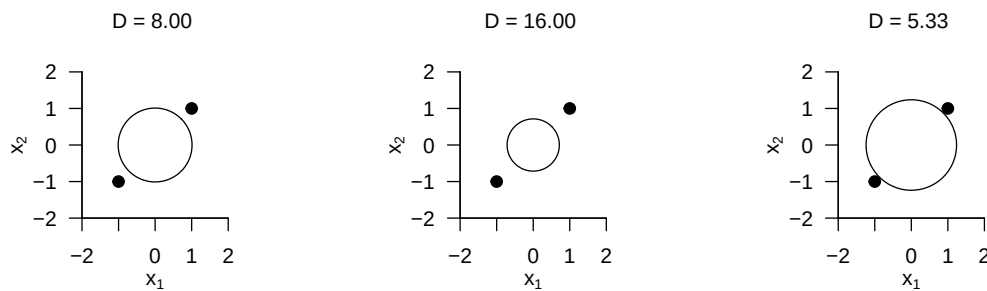


Abbildung 45.1. Mahalanobis-Distanzen als Funktion von Komponentenvarianzen

resultiert. Umgekehrt erkennt man an der Darstellung zu Σ_3 , dass eine stark negative Kovarianz der Komponenten von x_1 und x_2 zu einer größeren Mahalanobis-Distanz führt. Intuitiv nähert also auch die Kovarianz von Komponenten x_1 und x_2 an. Alternativ kann man die Höhe einer Mahalanobis-Distanz dabei auch als ein Maß für die Unwahrscheinlichkeit der Realisierung zweier Werte eines Zufallsvektors bei einer gegebenen Kovarianzmatrix verstehen.

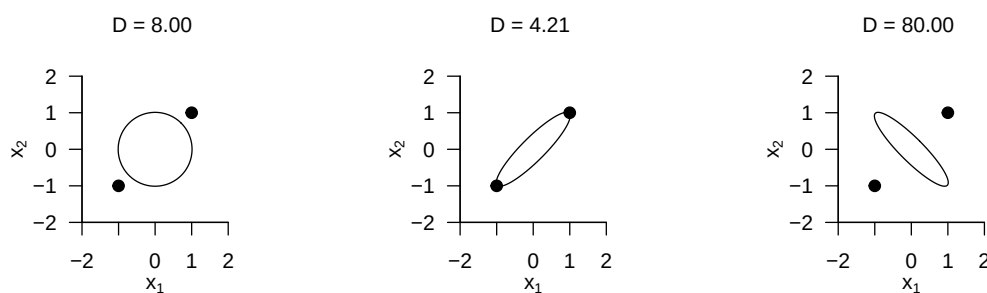


Abbildung 45.2. Mahalanobis-Distanzen als Funktion von Komponentenkovarianzen

45.3. Literaturhinweise

Die Resultate zur Matrixdarstellung von Stichprobenmittel, Stichprobenkovarianzmatrix und Stichprobenkorrelationsmatrix folgen Rencher & Christensen (2012). Der Begriff der Mahalanobis-Distanz geht zurück auf Mahalanobis (1936).

46. Einstichproben-T²-Tests

46.1. Anwendungsszenario

Wie im univariaten Fall ist das Anwendungsszenario eines Einstichproben-T²-Tests dadurch gekennzeichnet, dass n Datenpunkte einer Stichprobe (Gruppe) randomisierter experimenteller Einheiten betrachtet werden. In Generalisierung des univariaten Falls sind die n Datenpunkte allerdings multivariat, jeder Datenpunkt besteht also aus zwei oder mehr Zahlen und kann als Vektor in $\mathbb{R}^m$ mit $m > 1$ betrachtet werden. In Analogie zum univariaten Fall wird von den n Datenpunkten angenommen, dass sie Realisierungen von n unabhängigen und identisch multivariat normalverteilten Zufallsvektoren sind. Hinsichtlich der identischen multivariaten Normalverteilung $N(\mu, \Sigma)$ dieser Zufallsvektoren wird angenommen, dass sowohl der wahre Erwartungswertparameter μ als auch der wahre Kovarianzmatrixparameter Σ unbekannt sind. Schließlich wird vorausgesetzt, dass ein Interesse an einem inferentiellen Vergleich des wahren, aber unbekannten, Erwartungswertparameters μ mit einem vorgegebenen Wert μ_0 , beispielsweise $\mu_0 := 0_m$, besteht. Wie im univariaten Fall ergeben sich auch in diesem Anwendungsszenario eine Reihe möglicher Hypothesenszenarien mit jeweils unterschiedlichen Testgütefunktionen und damit Herangehensweisen an Testumfangkontrolle und Stichprobengrößenoptimierung. Wir wollen im Rahmen dieser Einführung nur das Szenario einer einfachen Nullhypothese und einer zusammengesetzten Alternativhypothese,

$$H_0 : \mu = \mu_0 \Leftrightarrow \Theta_0 := \{\mu_0\} \text{ und } H_1 : \mu \neq \mu_0 \Leftrightarrow \Theta_1 := \mathbb{R}^m \setminus \{\mu_0\} \quad (46.1)$$

genauer untersuchen.

Anwendungsbeispiel

Für ein konkretes Anwendungsbeispiel betrachten wir die Analyse simulierter Prä-Post-Interventions-Differenzwerte von BDI Scores (dBDI) und Glukokortikoidplasmaleveln (dGLU), die, wie in Tabelle 47.1 dargestellt, an einer Gruppe von $n = 20$ Patient:innen erhoben worden sein könnten. Positive Werte von dBDI und dGLU entsprächen dabei einer Reduktion der Depressionssymptomatik, negative Werte zeigen eine Verschlechterung des Depressionszustandes an.

Bei der Anwendung eines Einstichproben-T²-Tests auf die Daten dieses simulierten Datensatzes nehmen wir an, dass die zweidimensionalen Datenvektoren (dBDI, dGLU) Realisierungen von $n = 20$ unabhängig normalverteilten zweidimensionalen Zufallsvektoren $y_i \sim N(\mu, \Sigma)$ sind. Wir nehmen weiterhin an, dass wir daran interessiert sind, unsere Unsicherheit beim inferentiellen Vergleich des wahren, aber unbekannten, Erwartungswertparameters $\mu \in \mathbb{R}^2$ mit einem Vergleichswert $\mu_0 \in \mathbb{R}^2$, etwa einem Therapieerfolgswert, zu quantifizieren.

Tabelle 46.1. Prä-Post-Interventions-Differenzwerte von BDI Scores und Glukokortikoidplasmaleveln von $n = 20$ Patient:innen

dBDI	dGLU
35	6.1
25	4.0
20	1.7
29	2.6
29	1.9
17	0.9
33	2.0
28	4.1
26	3.9
31	3.8
14	2.1
18	2.0
19	5.0
28	2.6
20	2.1
35	4.4
28	4.0
32	3.9
32	1.0
25	1.9

Unabhängig von diesem inferenzstatistischen Vorgehen betrachten wir zunächst zu diesem Datensatz einige Deskriptivstatistiken wie durch folgenden **R** Code ausgewertet und in Abbildung 46.1 dargestellt. Verglichen mit einem Therapienormwert von $\mu_0 := (30, 3.5)^T$ fallen die Komponenten des Stichprobenmittels mit $\bar{y} = (26.3, 3.0)^T$ etwas geringer aus, allerdings bei einer nicht zu vernachlässigen Datenvariabilität, die sich in einer Mahalanobisdistanz von $D = 0.4$ des Stichprobenmittels vom Therapienormwert in bezug auf die Stichprobenkovarianzmatrix des Datensatzes widerspiegelt.

```
# Datendeskription
D = read.csv("_data/702-einstichproben-t2-tests.csv")
Y = rbind(D$y_1i, D$y_2i)
mu_0 = matrix(c(30,3.5), nrow = 2)
n = ncol(Y)
j_n = matrix(rep(1,n), nrow = n)
I_n = diag(n)
J_n = matrix(rep(1,n^2), nrow = n)
Y_bar = (1/n)*(Y %*% j_n)
C = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))
D = t(Y_bar - mu_0) %*% solve(C) %*% (Y_bar - mu_0)

# Daten einlesen
# Datenmatrix
# Normwert
# Anzahl Datenpunkte
# i_n
# I_n
# 1_{nn}
# Stichprobenmittel
# Stichprobenkovarianzmatrix
# Mahalanobis Distanz

Y_bar = 26.25615 2.991039
D = 0.3773184
```

46.2. Modellformulierung und Modellevaluation

Wir definieren zunächst das Einstichproben-T²-Test-Modell wie folgt.

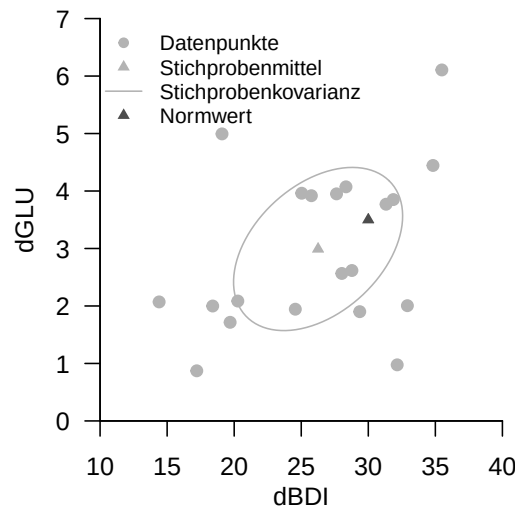


Abbildung 46.1. Deskriptivstatistiken der dBDI, dGLU Daten des Beispieldatensatzes. Jeder Punkt visualisiert die Daten einer Patient:in, die Stichprobenkovarianz ist durch die 0.4 Isokontur einer zweidimensionalen Normalverteilung mit Erwartungswertparameter und Kovarianzmatrixparameter entsprechend dem Stichprobenmittel und der Stichprobenkovarianz dargestellt

Definition 46.1 (Einstichproben- T^2 -Test-Modell). Für $i = 1, \dots, n$ seien y_i m -dimensionale Zufallsvektoren, die die n Datenpunkte eines Einstichproben- T^2 -Test Szenarios modellieren. Dann hat das *Einstichproben- T^2 -Test-Modell* die strukturelle Form

$$y_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0_m, \Sigma) \text{ u.i.v. für } i = 1, \dots, n \text{ mit } \mu \in \mathbb{R}^m, \Sigma \in \mathbb{R}^{m \times m} \text{ pd} \quad (46.2)$$

und die Datenverteilungsform

$$y_i \sim N(\mu, \Sigma) \text{ u.i.v. für } i = 1, \dots, n \text{ mit } \mu \in \mathbb{R}^m, \Sigma \in \mathbb{R}^{m \times m} \text{ pd.} \quad (46.3)$$

•

Die Äquivalenz von struktureller Form und Datenverteilungsform des Einstichproben- T^2 -Test-Modells folgt dabei direkt mit Theorem 29.8 durch Transformation der Zufallsvektoren ε_i anhand von Gleichung 46.2.

46.3. Modellevaluation

Teststatistik und Test

Wir definieren als nächstes eine Teststatistik für das Einstichproben- T^2 -Test Szenario.

Definition 46.2 (Einstichproben- T^2 -Teststatistik). Gegeben seien das Einstichproben- T^2 -Test-Modell und ein Nullhypothesenparameter $\mu_0 \in \mathbb{R}^m$. Dann ist die Einstichproben- T^2 -Teststatistik definiert als

$$T^2 := n(\bar{y} - \mu_0)^T C^{-1}(\bar{y} - \mu_0), \quad (46.4)$$

wobei $\bar{y}$ und C das Stichprobenmittel und die Stichprobenkovarianzmatrix der $y_1, \dots, y_n$ bezeichnen.

•

Die Einstichproben- T^2 -Teststatistik ist offenbar die mit dem Stichprobenumfang n skalierte Mahalanobis-Distanz von $\bar{y}$ und μ_0 hinsichtlich C (vgl. Kapitel 45.2). Damit gilt entsprechend, dass bei konstanter Stichprobenkovarianzmatrix die Einstichproben- T^2 -Teststatistik T^2 größere Werte für eine größere Euklidische Distanz von $\bar{y}$ und μ_0 annimmt und bei konstanter Euklidischer Distanz von $\bar{y}$ und μ_0 der Wert der Teststatistik T^2 von der Höhe der Datenvariabilität abhängt. Hinsichtlich der Verteilung der Einstichproben- T^2 -Teststatistik halten wir zunächst folgendes Theorem fest, das wir nicht beweisen wollen.

Theorem 46.1 (Verteilung der skalierten Einstichproben- T^2 -Teststatistik). *Es seien $y_1, \dots, y_n \sim N(\mu, \Sigma)$ mit $\mu \in \mathbb{R}^m$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd,*

$$\nu := \frac{n - m}{(n - 1)m} \quad (46.5)$$

und für $\mu \in \mathbb{R}^m$ sei die Einstichproben- T^2 -Teststatistik definiert als

$$T^2 := n(\bar{y} - \mu_0)^T C^{-1}(\bar{y} - \mu_0). \quad (46.6)$$

Dann gilt

$$\nu T^2 \sim f(\delta, m, n - m), \quad (46.7)$$

wobei $f(\delta, m, n - m)$ die nichtzentrale f -Verteilung mit Nichtzentralitätsparameter

$$\delta := n(\mu - \mu_0)^T \Sigma^{-1}(\mu - \mu_0) \quad (46.8)$$

sowie mit Freiheitsgradparametern m und $n - m$ bezeichnet.

◦

Für einen Beweis von Theorem 46.1 verweisen wir auf Hotelling (1931) und Anderson (2003). Wir erinnern in diesem Zusammenhang an die Begriffe der f -Zufallsvariable und der nichtzentralen f -Zufallsvariable, für die wir exemplarische WDFen in Abbildung 46.2 und Abbildung 46.3 darstellen.

Definition 46.3 (f -Zufallsvariable). ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}_{>0}$ und WDF

$$p_\xi : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto p_\xi(x) := \nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}} \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right)} \frac{x^{\frac{\nu_1}{2} - 1}}{(\nu_1 x + \nu_2)^{\frac{\nu_1 + \nu_2}{2}}}, \quad (46.9)$$

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass ξ einer f -Verteilung mit Freiheitsgradparametern ν_1 und ν_2 unterliegt und nennen ξ eine f -Zufallsvariable mit

Freiheitsgradparametern ν_1 und ν_2 . Wir kürzen dies mit $\xi \sim f(\nu_1, \nu_2)$ ab. Die WDF einer f -Zufallsvariable bezeichnen wir mit $f(x; \nu_1, \nu_2)$, die KVF einer f -Zufallsvariable bezeichnen wir mit $F(x; \nu_1, \nu_2)$, und die inverse KVF einer f -Zufallsvariable bezeichnen wir mit $F^{-1}(x; \nu_1, \nu_2)$.

•

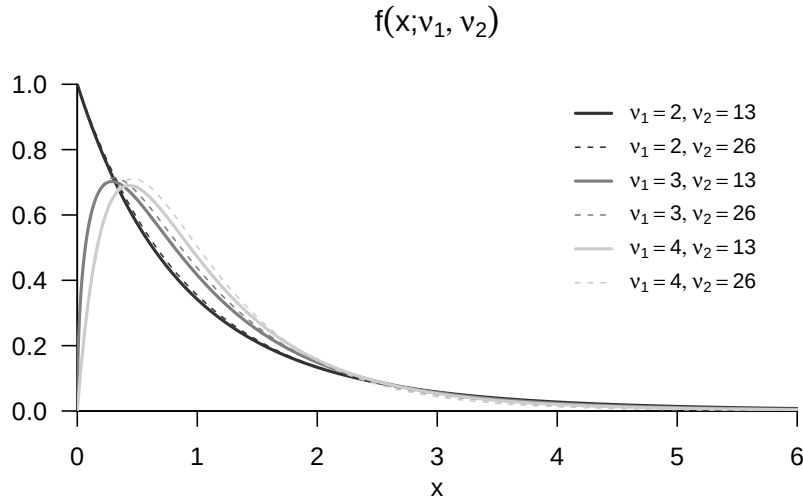


Abbildung 46.2. Exemplarische WDFen einer f -Zufallsvariable

Definition 46.4 (Nichtzentrale f -Zufallsvariable). ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}_{>0}$ und WDF

$$p_\xi : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto p_\xi(x) := \sum_{k=0}^{\infty} \frac{e^{-\delta/2} (\delta/2)^k}{\frac{\Gamma(\nu_2/2) \Gamma(\nu_1/2+k)}{\Gamma(\nu_2/2+\nu_1/2+k)} k!} \left(\frac{\nu_1}{\nu_2} \right)^{\nu_1/2+k} \left(\frac{\nu_2}{\nu_2 + \nu_1 x} \right)^{(\nu_1+\nu_2)/2+k} x^{\nu_1/2-1+k} \quad (46.10)$$

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass ξ einer nichtzentralen f -Verteilung mit Nichtzentralitätsparameter δ und Freiheitsgradparametern ν_1 und ν_2 unterliegt und nennen ξ eine nichtzentrale f -Zufallsvariable mit Nichtzentralitätsparameter δ und Freiheitsgradparametern ν_1 und ν_2 . Wir kürzen dies mit $\xi \sim f(\delta, \nu_1, \nu_2)$ ab. Die WDF einer f -Zufallsvariable bezeichnen wir mit $f(x; \delta, \nu_1, \nu_2)$, die KVF einer nichtzentralen f -Zufallsvariable bezeichnen wir mit $F(x; \delta, \nu_1, \nu_2)$, und die inverse KVF einer nichtzentralen f -Zufallsvariable bezeichnen wir mit $F^{-1}(x; \delta, \nu_1, \nu_2)$.

•

Im univariaten Fall sind bekanntlich die F -Statistiken der Varianzanalyse bei Zutreffen der Nullhypothese f -verteilt und bei Zutreffen der Alternativhypothese nichtzentral- f -verteilt. Für den Fall $\mu = \mu_0$, dass also der wahre, aber unbekannte Erwartungswertparameter mit dem Nullhypothesenparameter identisch ist, gilt nach Gleichung 46.8 $\delta = 0$ und $f(\delta, m, n-m)$ entspricht der f -Verteilung $f(m, n-m)$.

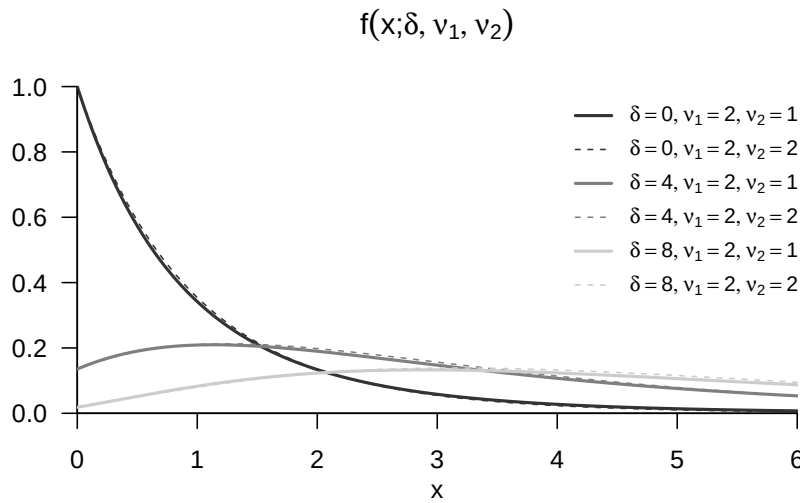


Abbildung 46.3. Exemplarische WDFen einer nichtzentralen f -Zufallsvariable

Wir halten weiterhin fest, dass aus Theorem 46.1 im univariaten Fall $m := 1$ aus Gleichung 46.5 folgt, dass

$$\nu = \frac{n-1}{(n-1) \cdot 1} = 1 \quad (46.11)$$

und mit der Stichprobenvarianz S^2 einer univariaten Stichprobe entsprechend folgt, dass

$$T^2 = n \frac{(\bar{y} - \mu_0)^2}{S^2} = \left(\sqrt{n} \frac{\bar{y} - \mu_0}{S} \right)^2. \quad (46.12)$$

Dies ist offenbar das Quadrat der bekannten univariaten Einstichproben-T-Test Teststatistik. Damit ist nach Theorem 46.1 das Quadrat der univariaten Einstichproben-T-Test Teststatistik nach $f(\delta, 1, n-1)$ verteilt. Intuitiv und verkürzt ausgedrückt ist also eine quadrierte t -Zufallsvariable eine f -Zufallsvariable.

Aus Theorem 46.1 folgen nun für Einstichproben- T^2 -Teststatistik unmittelbar folgende Formen für ihre WDF und KVF.

Theorem 46.2 (WDF und KDF der Einstichproben- T^2 -Teststatistik). *Im Einstichproben- T^2 -Test Szenario sei*

$$\nu := \frac{n-m}{(n-1)m} \quad (46.13)$$

Dann ist eine WDF der Einstichproben- T^2 -Teststatistik gegeben durch

$$p_{T^2} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}, t^2 \mapsto p_{T^2}(t^2) := \nu f(\nu t^2; \delta, m, n-m) \quad (46.14)$$

und eine KVF der Einstichproben- T^2 -Teststatistik ist gegeben durch

$$P_{T^2} : \mathbb{R}_{\geq 0} \rightarrow [0, 1], t^2 \mapsto P_{T^2}(t^2) := F(\nu t^2; \delta, m, n-m) \quad (46.15)$$

◦

Beweis. Wir halten zunächst fest, dass das Theorem zur univariaten WDF Transformation bei linear-affinen Abbildungen besagt, dass für eine Zufallsvariable ξ mit WDF p_ξ und der Definition $y = f(\xi)$ mit $f(\xi) := a\xi + b$ für $a \neq 0$ eine WDF von y definiert ist durch $p_y(y) := (1/|a|)p_\xi((y-b)/a)$. Im vorliegenden Fall ist $\xi = \nu T^2$ mit WDF $f(\delta, m, n-m)$ und $y := T^2 = \frac{1}{\nu}\nu T^2$, also $a = 1/\nu$ und $b = 0$. Mit $\nu > 0$ ergibt sich (46.14) also aus

$$p_{T^2}(t^2) = \frac{1}{a} p_{\nu T^2} \left(\frac{t^2}{a} \right) = \nu f(\nu t^2; m, n-m). \quad (46.16)$$

(46.15) folgt dann damit, dass WDFen bei kontinuierlichen Zufallsvariablen die Ableitungen der entsprechenden KVFen sind. Mit der Kettenregel der Differentiation ergibt sich

$$\begin{aligned} \frac{d}{dt^2} P_{T^2}(t^2) &= \frac{d}{dt^2} (F(\nu t^2; m, n-m, \delta)) \\ &= \frac{d}{dt^2} F(\nu t^2; m, n-m, \delta) \frac{d}{dt^2} (\nu t^2) \\ &= \nu f(\nu t^2; m, n-m, \delta) \\ &= p_{T^2}(t^2). \end{aligned} \quad (46.17)$$

□

Wir halten fest, dass die skalierte Einstichproben- T^2 -Test Teststatistik νT^2 nach $f(\delta, m, n-m)$ nichtzentral f -verteilt ist, die WDF der Einstichproben- T^2 -Test Teststatistik T^2 selbst dagegen durch $\nu f(\nu t^2; \delta, m, n-m)$ gegeben ist. Wir simulieren diese Verteilung mithilfe folgenden **R** Codes und visualisieren diese Simulation in Abbildung 46.4.

```
# Modellparameter
m      = 2                                # Dimensionalität der Zufallsvektoren/Daten
n      = 15                                # Anzahl der Datenpunkte
mu_0   = matrix(c(1,1), nrow = 2)         # Nullhypothesenparameter
mu     = matrix(c(2,2), nrow = 2)         # wahrer, aber unbekannter, Erwartungswertparameter
Sigma  = matrix(c(0.5,0.3, 0.3,0.5), nrow = 2, byrow = TRUE) # wahrer, aber unbekannter, Kovarianzmatrixparameter

# Simulation
library(MASS)
nsim   = 1e4                               # R Paket für multivariate Normalverteilungen
Yb     = matrix(rep(NA,n*nsim), nrow = 2) # Anzahl Simulationen/Datensatzrealisierungen
T2     = rep(NA,nsim)                      # Stichprobenmittellarray
j_n    = matrix(rep(1,n), nrow = n)       # Einstichproben-T^2-Teststatistik Array
I_n    = diag(n)                          # i_n
J_n    = matrix(rep(1,n^2), nrow = n)     # i_{nn}
for(s in 1:nsim){
  Y      = t(mvrnorm(n,mu,Sigma))          # Simulationsiterationen
  Y_bar  = (1/n)*(Y %*% j_n)               # y_i \sim N(\mu,\Sigma), i = 1,...,n
  C      = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y)) # Stichprobenmittel
  T2[s]  = n*t(Y_bar - mu_0) %*% solve(C) %*% (Y_bar - mu_0) # Stichprobenkovarianzmatrix
  Yb[,s] = Y_bar                          # Einstichproben-T^2-Teststatistik
}                                           # Stichprobenmittel für Visualisierung
```

Den Einstichproben- T^2 -Test definieren wir schließlich als einen kritischen Wert-basierten Test wie folgt.

Definition 46.5 (Einstichproben- T^2 -Test). Gegeben seien das Einstichproben- T^2 -Modell und die Einstichproben- T^2 -Teststatistik. Dann ist für einen kritischen Wert $k \geq 0$ der Einstichproben- T^2 -Test definiert als der kritische Wert-basierte Test

$$\phi(y) := 1_{\{T^2 > k\}} := \begin{cases} 1 & T^2 > k \\ 0 & T^2 \leq k \end{cases}. \quad (46.19)$$

•

In Definition 46.5 repräsentiert wie üblich $\phi(y) = 1$ den Vorgang des Ablehnens von H_0 und $\phi(y) = 0$ den Vorgang des Nichtablehnens von H_0 .

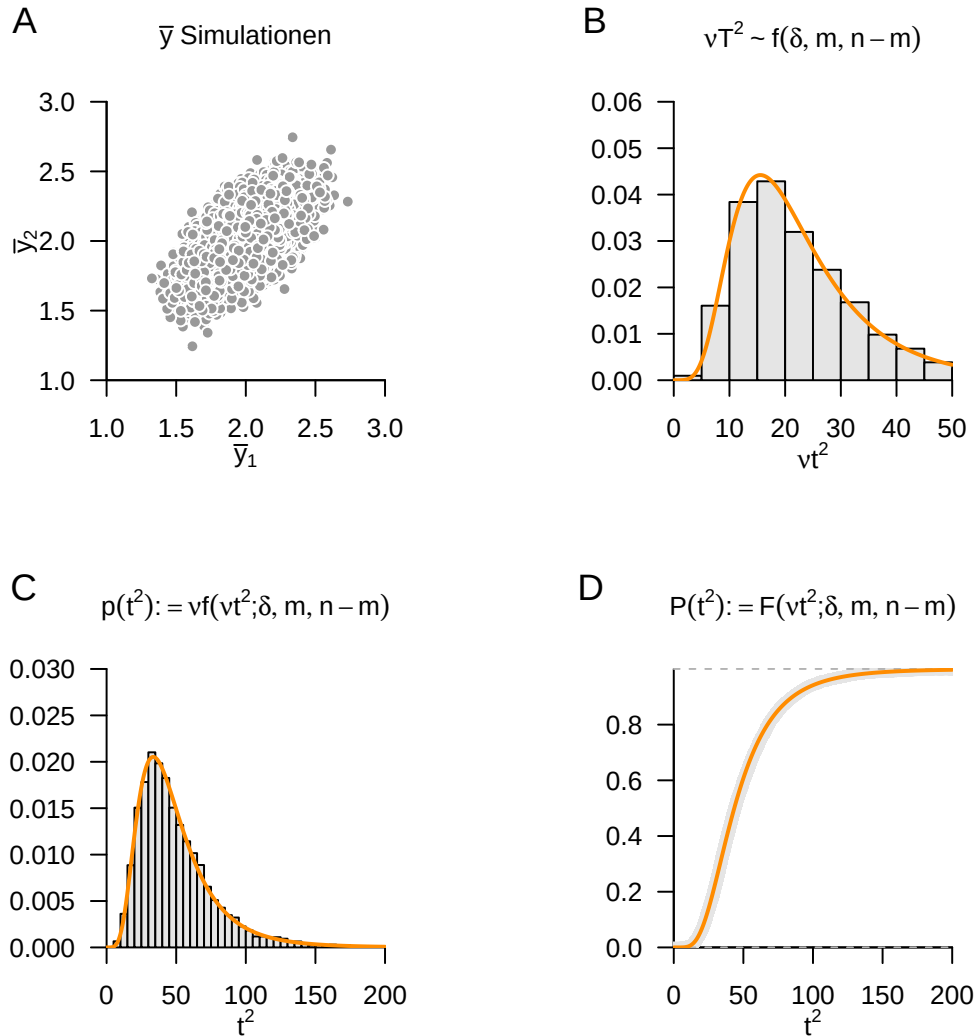


Abbildung 46.4. Verteilung der Einstichproben- T^2 -Teststatistik. **A** Stichprobenmittel von 10.000 Realisierungen eines Einstichproben- T^2 -Test-Modells mit $m := 2$ und $n := 15$ und wahren, aber unbekannten, Parametern

$$\mu := (2, 2)^T \text{ und } \Sigma = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}. \quad (46.18)$$

B Histogramm der entsprechenden Realisierungen der skalierten Einstichproben- T^2 -Teststatistik für $\mu_0 := (1, 1)^T$ (grau) und analytische Form dieser Verteilung (orange). **C** Histogramm der entsprechenden Realisierungen der Einstichproben- T^2 -Teststatistik (grau) und ihre analytische Form (orange). **D** Empirische KVF der entsprechenden Realisierungen der Einstichproben- T^2 -Teststatistik (grau) und ihre analytische Form (orange).

Analyse der Testgütefunktion

Um für den in Definition 46.5 definierten Test Prozeduren zur Testumfangkontrolle (Typ I Fehlerbegrenzung) und zur Stichprobengrößenoptimierung (Typ II Fehlerbegrenzung) zu entwickeln, betrachten wir zunächst seine Testgütefunktion. Es gilt folgendes Theorem.

Theorem 46.3 (Testgütefunktion des Einstichproben- T^2 -Tests). *ϕ sei der Einstichproben- T^2 -Test. Dann ist die Testgütefunktion von ϕ gegeben durch*

$$q_\phi : \mathbb{R}^m \rightarrow [0, 1], \mu \mapsto q_\phi(\mu) := 1 - F(\nu k; \delta_\mu, m, n - m) \quad (46.20)$$

wobei $F(\cdot; \delta_\mu, m, n - m)$ die KVF der nichtzentralen f -Verteilung mit Freiheitsgradparametern m und $n - m$ sowie mit Nichtzentralitätsparameter

$$\delta_\mu := n(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) \quad (46.21)$$

bezeichnet.

◦

Beweis. Die Testgütefunktion des betrachteten Tests ist definiert als

$$q_\phi : \mathbb{R}^m \rightarrow [0, 1], \mu \mapsto q_\phi(\mu) := \mathbb{P}_\mu(\phi = 1). \quad (46.22)$$

Da die Wahrscheinlichkeiten für $\phi = 1$ und dafür, dass die zugehörige Teststatistik im Ablehnungsbereich des Tests liegt, gleich sind, benötigen wir also zunächst die Verteilung der Teststatistik. Wir haben oben aber bereits gesehen, dass

$$\frac{n - m}{m(n - 1)} T^2 \sim f(m, n - m, \delta_\mu) \text{ mit } \delta_\mu := n(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) \quad (46.23)$$

gilt. Der Ablehnungsbereich des betrachteten Tests ist $A :=]k, \infty[$. Also ergibt sich

$$\begin{aligned} q_\phi(\mu) &= \mathbb{P}_\mu(\phi = 1) \\ &= \mathbb{P}_\mu(T^2 \in]k, \infty[) \\ &= \mathbb{P}_\mu(T^2 > k) \\ &= 1 - \mathbb{P}_\mu(T^2 \leq k) \\ &= 1 - F(\nu k; \delta_\mu, m, m - n) \end{aligned} \quad (46.24)$$

□

Wir wollen diese Testgütefunktion beispielhaft für zwei Szenarien mit $m := 2$ und $n := 15$ in Abhängigkeit des kritischen Wertes k betrachten. Abbildung 46.5 und Abbildung 46.6 visualisieren q_ϕ in diesen Szenarien für einen Nullhypothese parameter $\mu_0 := (1, 1)^T$ und die wahren, aber unbekannten, Kovarianzmatrixparameter

$$\Sigma_1 := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \text{ und } \Sigma_2 := \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix}, \quad (46.25)$$

respektive. In beiden Fällen und unabhängig von k resultiert eine größere Distanz des wahren, aber unbekannten, Erwartungswertparameters μ vom Nullhypothese parameter μ_0 in einer höheren Wahrscheinlichkeit dafür, dass der Test ϕ den Wert 1 annimmt, also die Nullhypothese abgelehnt wird. Die Zunahme dieser Wahrscheinlichkeit ist im ersten Szenario isotropisch, im zweiten dagegen aufgrund der Form des wahren, aber unbekannten,

Kovarianzparameters nicht. Bei einem kleinen kritischen Wert k werden hohe Wahrscheinlichkeiten für eine Ablehnung der Nullhypothese schon bei geringen Distanzen zwischen μ und μ_0 erreicht, bei einem größeren kritischen Wert k dagegen erst für größere Distanzen. Untenstehender **R** Code demonstriert das Vorgehen zur Evaluation dieser Testgütefunktionen

```
# Modellparameter
m      = 2
n      = 15
nu     = (n-m)/((n-1)*m)
Sigma  = diag(m)
iSigma = solve(Sigma)

# Testparameter
mu_0   = matrix(c(1,1), nrow = 2)
k_all  = c(2,4,6)
n_k    = length(k_all)

# q_\phi(\mu) Evaluation
mu_min = 0
mu_max = 2
mu_res = 1e3
mu_i    = seq(mu_min, mu_max, len = mu_res)
q_phi   = array(dim = c(mu_res, mu_res, length(k_all)))
for(k in 1:n_k){
  for(i in 1:mu_res){
    for(j in 1:mu_res){
      mu      = matrix(c(mu_i[i], mu_i[j]), nrow = 2)
      delta_mu = n*t(mu - mu_0) %*% iSigma %*% (mu - mu_0)
      q_phi[i,j,k] = 1 - pf(nu*k_all[k], m, n-m, delta_mu)}
    }
  }
}

# m
# n
# nu
# \Sigma = I_2
# \Sigma^{-1}

# \mu_0
# k <-> \phi
# Anzahl k Werte/Tests

# \mu_i Minimum
# \mu_i Maximum
# \mu_i Auflösung
# mu_i
# q_\phi Array

# \mu
# \delta_\mu
# q_\phi(\mu)
```

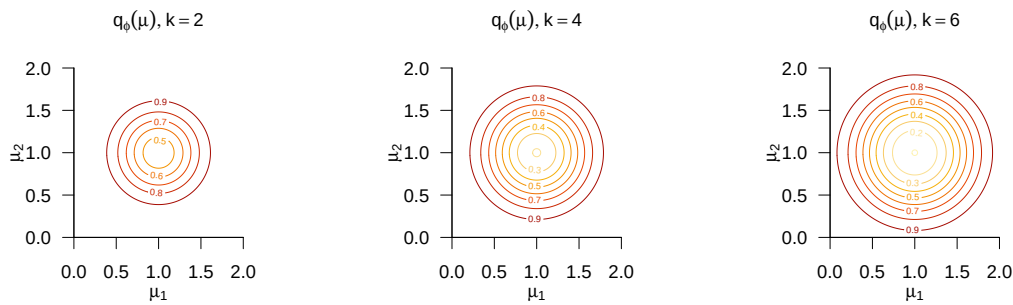


Abbildung 46.5. Einstichproben- T^2 -Test Testgütefunktionen für kritische Werte $k = 2$, $k = 4$ und $k = 6$ im Szenario

$$m := 2, n := 15, \mu_0 := (1, 1)^T, \Sigma_1 := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \quad (46.26)$$

Testumfangkontrolle

Bekanntlich erlaubt die Testumfangkontrolle die Begrenzung der größtmöglichen Wahrscheinlichkeit für einen Typ I Fehler. Im aktuellen Testszenario haben wir folgendes Theorem.

Theorem 46.4 (Testumfangkontrolle des Einstichproben- T^2 - Tests). *ϕ sei der im obigen Testszenario definierte Test. Dann ist ϕ ein Level- α_0 -Test mit Testumfang α_0 , wenn der kritische Wert definiert ist durch*

$$k_{\alpha_0} := \nu^{-1} F^{-1}(1 - \alpha_0; m, n - m), \quad (46.28)$$

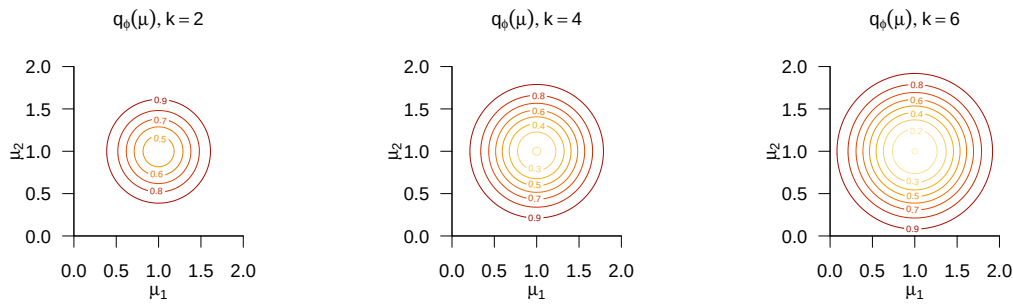


Abbildung 46.6. Einstichproben-T²-Test Testgütefunktionen für kritische Werte $k = 2$, $k = 4$ und $k = 6$ im Szenario

$$m := 2, n := 15, \mu_0 := (1, 1)^T, \Sigma_1 := \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix} \quad (46.27)$$

wobei $\nu := (n - m) / ((n - 1)m)$ und $F^{-1}(\cdot; m, n - m)$ die inverse KVF der f -Verteilung mit Freiheitsgradparametern m und $n - m$ ist.

◦

Beweis. Damit der betrachtete Test ein Level- α_0 -Test ist, muss bekanntlich $q_\phi(\mu) \leq \alpha_0$ für alle $\mu \in \{\mu_0\}$, also hier $q_\phi(\mu_0) \leq \alpha_0$ gelten. Weiterhin ist der Testumfang des betrachteten Tests durch $\alpha = \max_{\mu \in \{\mu_0\}} q_\phi(\mu)$, also hier durch $\alpha = q_\phi(\mu_0)$ gegeben. Wir müssen also zeigen, dass die Wahl von k_{α_0} garantiert, dass ϕ ein Level- α_0 -Test mit Testumfang α_0 ist. Dazu merken wird zunächst an, dass für $\mu = \mu_0$ gilt, dass

$$q_\phi(\mu_0) = 1 - F(\nu k; \delta, m, n - m) = 1 - F(\nu k; 0, m, n - m) = 1 - F(\nu k; m, n - m), \quad (46.29)$$

wobei $F(\nu k; \delta, m, n - m)$ und $F(\nu k; m, n - m)$ die KVF der nichtzentralen f -Verteilung mit Nichtzentralitätsparameter δ und Freiheitsgradparametern m und $n - m$ sowie der f -Verteilung mit Freiheitsgradparametern m und $n - m$, respektive, bezeichnen. Sei nun also $k := k_{\alpha_0}$. Dann gilt

$$\begin{aligned} q_\phi(\mu_0) &= 1 - F(\nu k_{\alpha_0}; m, n - m) \\ &= 1 - F(\nu \nu^{-1} F^{-1}(1 - \alpha_0; m, n - m); m, n - m) \\ &= 1 - F(F^{-1}(1 - \alpha_0; m, n - m); m, n - m) \\ &= 1 - (1 - \alpha_0) = \alpha_0. \end{aligned} \quad (46.30)$$

Es folgt also direkt, dass bei der Wahl von $k = k_{\alpha_0}$, $q_\phi(\mu_0) \leq \alpha_0$ gilt und der betrachtete Test somit ein Level- α_0 -Test ist. Weiterhin folgt direkt, dass der Testumfang des betrachteten Tests bei Wahl von $k = k_{\alpha_0}$ gleich α_0 ist. $\square$

Wir visualisieren die Wahl von

$$k_{\alpha_0} = \nu^{-1} F^{-1}(1 - \alpha_0; m, n - m) \quad (46.31)$$

für den Fall $m = 2, n = 15$ und ein Signifikanzlevel von $\alpha_0 := 0.05$ in [Abbildung 46.7](#). Untenstehender **R** Code simuliert die Testumfangkontrolle für ein Einstichproben-T²-Test Szenario mit

$$m := 2, n := 15, \mu := \mu_0 := \begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ und } \Sigma := \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}. \quad (46.32)$$

Der auf Grundlage von 10^4 Datensatzrealisationen geschätzte Testumfang stimmt gut mit dem Signifikanzlevel überein.

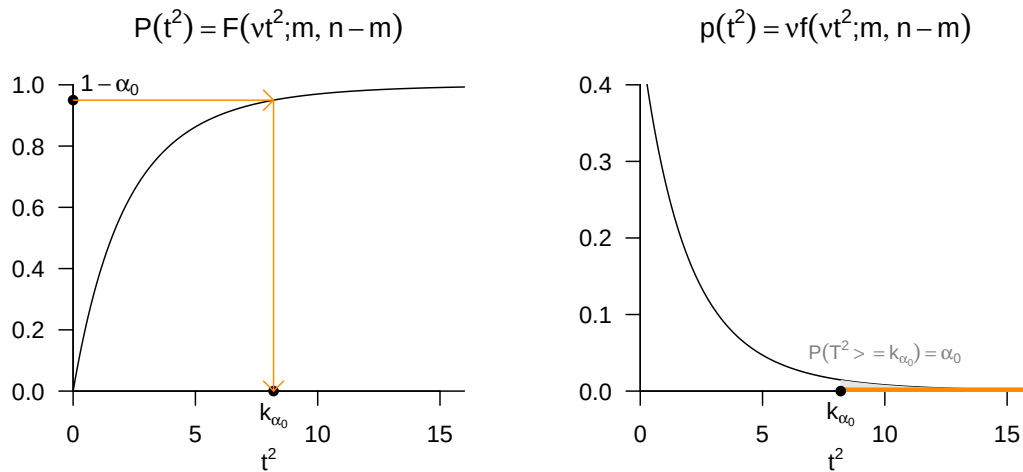


Abbildung 46.7. Testumfangkontrolle durch Selektion eines α_0 -abhängigen kritischen Wertes für den Einstichproben- T^2 -Test anhand von KVF und WDF der Einstichproben- T^2 -Teststatistik.

```
# Modellparameter
m      = 2
n      = 15
nu     = (n-m)/(m*(n-1))
mu_0   = matrix(c(1,1), nrow = 2)
mu     = mu_0
Sigma  = matrix(c(0.5,0.3, 0.3,0.5), nrow = 2, byrow = TRUE)

# Testparameter
alpha_0 = 0.05
k_alpha_0 = (1/nu)*qf(1-alpha_0, m,n-m)

# Simulation der Testumfangkontrolle
library(MASS)
nsim    = 1e4
phi     = rep(NA,n)
j_n     = matrix(rep(1,n), nrow = n)
I_n     = diag(n)
J_n     = matrix(rep(1,n^2), nrow = n)
for(s in 1:nsim){
  Y      = t(mvnrnorm(n,mu,Sigma))
  Y_bar  = (1/n)*(Y %*% j_n)
  C      = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))
  T2     = n*t(Y_bar - mu_0) %*% solve(C) %*% (Y_bar - mu_0)
  if(T2 > k_alpha_0){
    phi[s] = 1
  } else {
    phi[s] = 0
  }
}

# Dimensionalität der Zufallsvektoren/Daten
# Anzahl der Datenpunkte
# Parameter
# Nullhypothese parameter
# w.a.u. Erwartungsparameter bei Zutreffen von H0
# wahrer, aber unbekannter, Kovarianzmatrixparameter

# Signifikanzlevel
# kritischer Wert

# R Paket für multivariate Normalverteilungen
# Testentscheidungsarray
# Testentscheidungsarray
# 1_n
# I_n
# 1_{nn}
# Simulationsiterationen
# Y_i \sim N(\mu, \Sigma), i = 1, \dots, n
# Stichprobenmittel
# Stichprobenkovarianzmatrix
# Einstichproben-T^2-Test Statistik
# Test 1_{T^2} >= k_alpha_0
# Ablehnen von H_0

# Nicht Ablehnen von H_0
```

Kritischer Wert = 8.196602
Geschätzter Testumfang alpha = 0.0511

In der Praxis entsprechen obige Ergebnisse dann folgendem Vorgehen bei der Durchführung eines Einstichproben- T^2 -Tests. Man unterstellt, dass ein vorliegender Datensatz von m -dimensionalen Datenvektoren eine Realisation von n u.i.v. m -dimensionalen Zufallsvektoren $y_1, \dots, y_n \sim N(\mu, \Sigma)$ mit unbekannten Parametern $\mu \in \mathbb{R}^m$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd ist und möchte entscheiden, ob für ein $\mu_0 \in \mathbb{R}^m$ eher die Nullhypothese $H_0 : \mu = \mu_0$ oder die Alternativhypothese $H_1 : \mu \neq \mu_0$ zutrifft. Zu diesem Zweck wählt man zunächst ein Signifikanzlevel α_0 und bestimmt dann den zugehörigen kritischen Wert k_{α_0} . Beispielsweise gilt für $m = 2$ und $n = 15$ bei Wahl von $\alpha_0 := 0.05$, dass $k_{0.05} = \nu^{-1} F^{-1}(1 - 0.05; 2, 13) \approx 8.2$ ist. Anhand von m, n, μ_0 , dem Stichprobenmittel $\bar{y}$ und der Stichprobenkovarianzmatrix C berechnet man dann die Realisierung der Einstichproben- T^2 -Teststatistik

$$T^2 := n(\bar{y} - \mu_0)^T C^{-1}(\bar{y} - \mu_0). \quad (46.33)$$

Wenn das berechnete T^2 größer als k_{α_0} ist, lehnt man die Nullhypothese ab, andernfalls nicht. Die oben entwickelte Theorie zur Testumfangkontrolle des Einstichproben- T^2 -Test garantiert dann, dass man in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt.

p-Wert

Wir erinnern daran, dass per Definition der p-Wert das kleinste Signifikanzlevel α_0 ist, bei welchem man die Nullhypothese basierend auf einem vorliegenden Wert der Teststatistik ablehnen würde. Wir haben folgendes Theorem

Theorem 46.5. *Für den p-Wert des in Definition 46.5 definierten Test gilt*

$$p\text{-Wert} = \mathbb{P}(T^2 \geq t^2) = 1 - F(\nu t^2; m, n - m). \quad (46.34)$$

◦

Beweis. Bei einem beobachteten Wert t^2 der Einstichproben- T^2 -Teststatistik T^2 würde H_0 für jedes α_0 mit $t^2 \geq \nu^{-1} F^{-1}(1 - \alpha_0; m, n - m)$ abgelehnt werden. Für diese α_0 gilt, wie unten gezeigt

$$\alpha_0 \geq \mathbb{P}(T^2 \geq t^2). \quad (46.35)$$

Das kleinste $\alpha_0 \in [0, 1]$ mit $\alpha_0 \geq \mathbb{P}(T^2 \geq t^2)$ ist dann $\alpha_0 = \mathbb{P}(T^2 \geq t^2)$, also folgt

$$p\text{-Wert} = \mathbb{P}(T^2 \geq t^2) = 1 - F(\nu t^2; m, n - m). \quad (46.36)$$

Es bleibt zu zeigen, dass gilt

$$\begin{aligned} t^2 &\geq \nu^{-1} F^{-1}(1 - \alpha_0; m, n - m) \\ \Leftrightarrow \nu t^2 &\geq F^{-1}(1 - \alpha_0; m, n - m) \\ \Leftrightarrow \alpha_0 &\geq \mathbb{P}(T^2 \geq t^2). \end{aligned} \quad (46.37)$$

Dies aber folgt aus

$$\begin{aligned} t^2 &\geq \nu^{-1} F^{-1}(1 - \alpha_0; m, n - m) \\ \nu t^2 &\geq F^{-1}(1 - \alpha_0; m, n - m) \\ F(\nu t^2; m, n - m) &\geq F(F^{-1}(1 - \alpha_0; m, n - m); m, n - m) \\ F(\nu t^2; m, n - m) &\geq 1 - \alpha_0 \\ \mathbb{P}(T^2 \leq t^2) &\geq 1 - \alpha_0 \\ \alpha_0 &\geq 1 - \mathbb{P}(T^2 \leq t^2). \end{aligned} \quad (46.38)$$

□

Zum Beispiel ergeben sich bei $m = 2$ und $n = 15$ der p-Wert für $t^2 = 7.00$ zu 0.071 und bei $m = 4$ und $n = 15$ der p-Wert für $t^2 = 7.00$ zu 0.304. Die gleiche Anzahl an Datenpunkten resultiert bei höherer Datendimensionalität also in einem höheren p-Wert. Weiterhin ergeben sich bei $m = 2$ und $n = 15$ der p-Wert für $t^2 = 9.00$ zu 0.040 und bei $m = 2$ und $n = 99$ der p-Wert für $t^2 = 7.00$ zu 0.035. Geringere Verhältnisse von geschätzter Nullhypothesenabweichung und geschätzter Daten(ko)varianz können also hinsichtlich des p-Wertes durch eine höhere Anzahl an Datenpunkten ausgeglichen werden.

Analyse der Powerfunktion

Bekanntlich ist man manchmal an der Optimierung der Stichprobengröße vor der Durchführung einer Studie interessiert. Zu diesem Zweck betrachtet man die Testgütefunktion

$$q_\phi : \mathbb{R}^m \rightarrow [0, 1], \mu \mapsto q_\phi(\mu) := 1 - F(\nu k; \delta_\mu, m, n - m) \quad (46.39)$$

bei kontrolliertem Testumfang, also für

$$k_{\alpha_0} := \nu^{-1} F^{-1}(1 - \alpha_0; m, n - m) \quad (46.40)$$

mit festem α_0 als Funktion des Nichtzentralitätsparameters, also der wahren, aber unbekannten, Effektstärke und des Stichprobenumfangs. Insbesondere hängt hier k_{α_0} auch von n ab. Es ergibt sich dabei die bivariate reellwertige Funktion

$$\pi : \mathbb{R} \times \mathbb{N} \rightarrow [0, 1], (\delta_\mu, n) \mapsto \pi(\delta_\mu, n) := 1 - F(\nu k_{\alpha_0}; \delta_\mu, m, n - m). \quad (46.41)$$

Bei festgelegtem α_0 hängt diese sogenannten *Powerfunktion des Einstichproben- T^2 -Tests* also vom wahren, aber unbekannten, Nichtzentralitätsparameter δ_μ , der Datendimensionalität m und von der Stichprobengröße n ab. Wir evaluieren diese Abhängigkeiten mithilfe untenstehenden **R** Codes und visualisieren sie exemplarisch in Abbildung 46.8.

```
# Szenariospezifikationen
a_0_all = c(0.05, 0.01)
d_mu_min = 0
d_mu_max = 20
d_mu_res = 30
d_mu_all = seq(d_mu_min, d_mu_max, len = d_mu_res)
n_min = 5
n_max = 20
n_res = 30
n_all = seq(n_min, n_max, len = n_res)
m_all = c(2, 4)

# \alpha_0 Raum
# \delta_\mu Minimum
# \delta_\mu Maximum
# \delta_\mu Auflösung
# \delta_\mu d Raum
# n Minimum
# n Maximum
# n Auflösung
# n Raum
# m Raum

# Evaluation der Powerfunktion
pi = array(dim = c(d_mu_res, n_res, 2, 2))
for (a in 1:length(a_0_all)){
  for (l in 1:length(m_all)){
    for (i in 1:length(d_mu_all)){
      for (j in 1:length(n_all)){
        m = m_all[l]
        n = n_all[j]
        d_mu = d_mu_all[i]
        nu = (n-m)/(m*(n-1))
        alpha_0 = a_0_all[a]
        k_alpha_0 = (1/nu)*qf(1-alpha_0, m, n-m)
        pi[i, j, l, a] = 1 - pf(nu*k_alpha_0, m, n-m, d_mu)}}}

# Powerfunktionsarray
# m Iterationen
# \delta_\mu Iterationen
# n Iterationen
# Datendimensionalität
# Stichprobenumfang
# wahrer, aber unbekannter, Parameter
# Parameter
# Signifikanzlevel
# kritischer Wert
# Powerfunktionswert
```

Generell lässt sich aus der Perspektive der Anwendung festhalten, dass π als Funktion von n monoton steigt. Ein größerer Stichprobenumfang resultiert damit also im Allgemeinen in einer kleineren Wahrscheinlichkeit für einen Typ II Fehler. Dabei bleiben allerdings mögliche weitere Kosten für die Erhöhung des Stichprobenumfangs unberücksichtigt. Weiterhin hängen die Werte der Powerfunktion π offensichtlich vom wahren, aber unbekannten, Nichtzentralitätsparameterwert

$$\delta_\mu = n(\mu - \mu_0)^T \Sigma^{-1}(\mu - \mu_0) \quad (46.42)$$

ab. Würde man diesen Wert schon mit großer Präzision kennen, so gäbe es keinen Grund eine Studie und ihren Stichprobenumfang zu planen. Es wird deshalb zur Stichprobengrößenoptimierung im Vorfeld einer Studie im Allgemeinen folgendes Vorgehen favorisiert:

- (1) Man legt zunächst das Signifikanzlevel α_0 zur Kontrolle der Wahrscheinlichkeit eines Typ 1 Fehlers fest und evaluiert die entsprechende Powerfunktion.

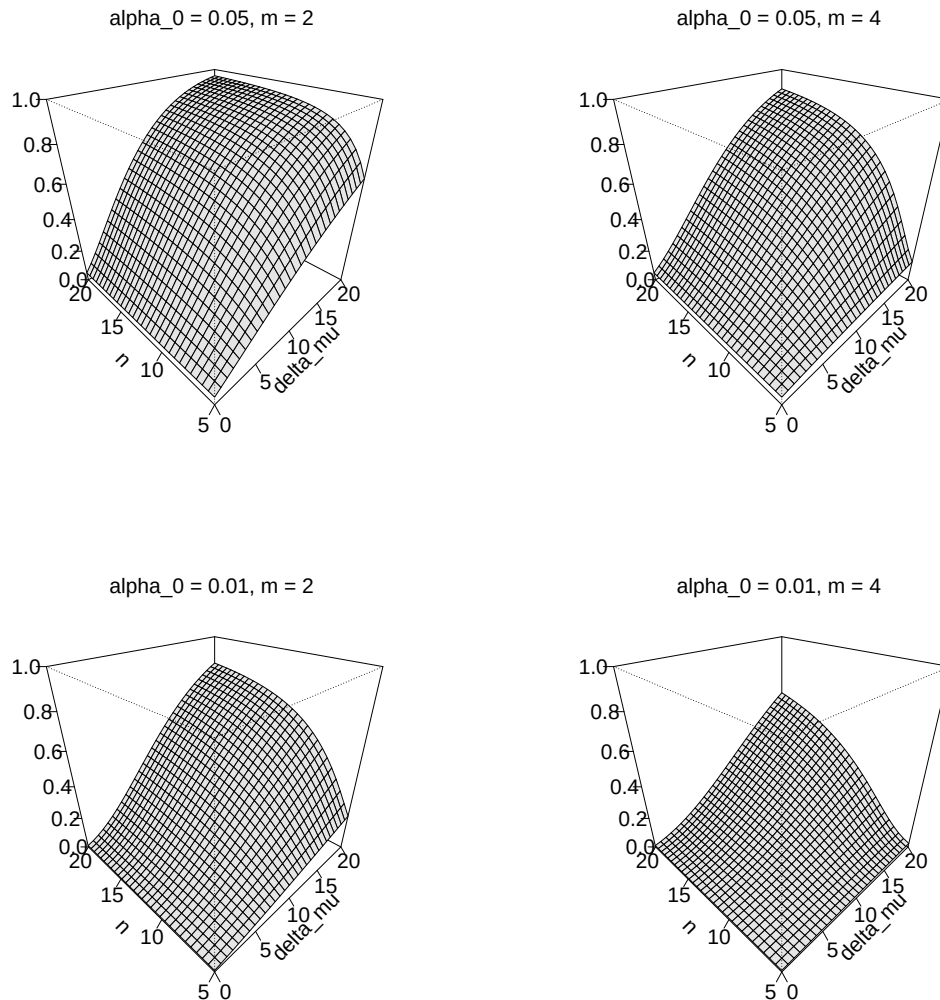


Abbildung 46.8. Powerfunktionen des Einstichproben- T^2 -Tests. Die Abbildungen zeigen die Wahrscheinlichkeit dafür, dass der Einstichproben- T^2 -Test den Wert 1 annimmt, dass also die Nullhypothese abgelehnt wird, als Funktion des wahren, aber unbekannten, Nichtzentralitätsparameters δ_μ sowie des Stichprobenumfangs n . Dabei bilden die Abbildungen der ersten Zeile die Powerfunktionen des Einstichproben- T^2 -Tests für das Signifikanzlevel $\alpha_0 = 0.05$ und die Abbildungen der zweiten Zeile die Powerfunktionen des Einstichproben- T^2 -Tests für das Signifikanzlevel $\alpha_0 = 0.01$ ab. Ein niedrigeres Signifikanzlevel resultiert wie üblich in einer geringeren Wahrscheinlichkeit für die Ablehnung der Nullhypothese über den gesamten Bereich von δ_μ und n . Die erste Spalte der Abbildung zeigt die beiden Signifikanzlevelszenarien für eine Datendimensionalität von $m := 2$, die zweite Spalte für eine Datendimensionalität von $m := 4$. Eine Erhöhung der Datendimensionalität führt, bei Konstanthalten aller weiteren Parameter, zu einer Reduktion der Wahrscheinlichkeit für das Ablehnen der Nullhypothese. Äquivalent sind bei höherer Datendimensionalität höhere Stichprobenumfänge nötig um bei vergleichbarem Nichtzentralitätsparameter die Nullhypothese gleichwahrscheinlich abzulehnen.

- (2) Man wählt einen Mindestparameterwert δ_μ^* , den man mit einer Wahrscheinlichkeit von

$$\pi(\delta_\mu, n) = \beta \quad (46.43)$$

detektieren möchte, bei dem man also die Nullhypothese ablehnen möchte. Der Wert von δ_μ^* ergibt sich dabei aus problemspezifischen Überlegungen, wie zum Beispiel der Frage nach einem klinisch bedeutsamen Wert. Ein konventioneller Wert für die gewünschte Detektionswahrscheinlichkeit ist $\beta := 0.8$.

- (3) Basierend auf der evaluierten Powerfunktion liest man die für

$$\pi(\delta_\mu = \delta_\mu^*, n) = \beta \quad (46.44)$$

minimal nötige Stichprobengröße n ab. Größere Stichprobengrößen führen aufgrund der Monotonie von π als Funktion von n sicher zu einer gleichen oder höheren Wahrscheinlichkeit für das Ablehnen der Nullhypothese.

Für eine Datendimensionalität von $m := 2$ und Mindestparameterwert von $\delta_\mu^* = 12$ evaluiert untenstehender **R** Code wie in Abbildung 46.9 dargestellt die minimale Stichprobengröße um mit einer Wahrscheinlichkeit von $\beta = 0.8$ die Nullhypothese abzulehnen.

```
# Szenariospezifikation
n_min = 5
n_max = 20
n_res = 1e2
n = seq(n_min, n_max, len = n_res)
alpha_0 = 0.05

# Poweranalyse
m = 2
d_mu_fix = 12
nu = (n-m)/(m*(n-1))
k_alpha_0 = (1/nu)*qf(1-alpha_0, m, n-m)
pi_n = 1 - pf(nu*k_alpha_0, m, n-m, d_mu_fix)
beta = 0.8
i = 1
n_min = NaN
while(pi_n[i] < beta){
  n_min = n[i]
  i = i + 1
}
```

n Minimum
n Maximum
n Auflösung
n Raum
Signifikanzlevel

Datendimensionalität
fester Nichtzentralitätsparameter
Parameter
kritischer Wert
Powerfunktionswert
gewünschter Powerfunktionswert
Indexinitialisierung
minimales n Initialisierung
Solange \pi(\delta_\mu, n) < \beta
Aufnahme des minimal nötigen ns
und Erhöhung des Indexes

Minimal nötiges n = 17

46.4. Anwendungsbeispiel

Wir betrachten das eingangs diskutierte Anwendungsbeispiel eines simulierten zweidimensionalen Datensatzes dreier Studiengruppen. Wir wollen abschließen für diesen Datensatz die Nullhypothese z für ein Abweichen des wahren, aber unbekannten, Erwartungswertparameters der Daten von μ_0 ist. Wir betrachten also weiterhin die einfache Nullhypothese $H_0 : \mu = \mu_0$ und die zusammengesetzte Alternativhypothese $H_1 : \mu \neq \mu_0$. Folgender **R** Code implementiert das praktische Vorgehen für ein Signifikanzlevel von $\alpha_0 := 0.05$.

```
# Datenbereitstellung
D = read.csv("../data/702-einstichproben-t2-tests.csv")
Y = rbind(D$y_1i, D$y_2i)

# Testparameter
m = nrow(Y)
n = ncol(Y)
nu = (n-m)/(m*(n-1))
mu_0 = matrix(c(30, 3.5), nrow = 2)
alpha_0 = 0.05
k_alpha_0 = (1/nu)*qf(1-alpha_0, m, n-m)
```

Datensatzeinlesen
Datenmatrix

Dimensionalität der Zufallsvektoren/Daten
Anzahl der Datenpunkte
Parameter
H0 Hypothesenparameter ("Normwert")
Signifikanzlevel
kritischer Wert

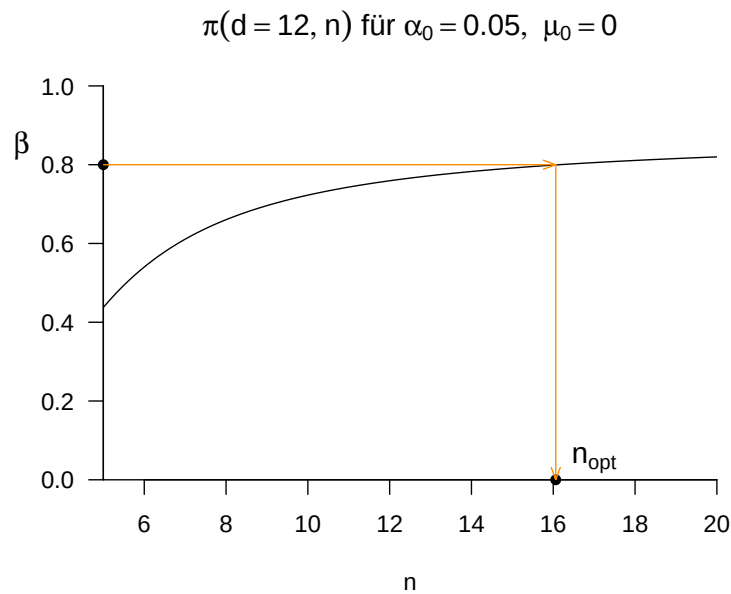


Abbildung 46.9. Bestimmung eines minimalen Stichprobenumfangs zur Detektion eines Nichtzentralitätsmindestparameterwert von $\delta_\mu = 12$. Die Abbildung zeigt die entsprechende Schnittfunktion der in Abbildung 46.8 dargestellten Funktion.

```
# Testevaluation
j_n = matrix(rep(1,n), nrow = n)
I_n = diag(n)
J_n = matrix(rep(1,n^2), nrow = n)
Y_bar = (1/n)*(Y %*% j_n)
C = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))
T2 = n*t(Y_bar - mu_0) %*% solve(C) %*% (Y_bar - mu_0)
if(T2 > k_alpha_0){
  phi = 1
} else {
  phi = 0
}
p = 1 - pf(nu*T2,m,n-m)

# 1_n
# I_n
# 1_{nn}
# Stichprobenmittel
# Stichprobenkovarianzmatrix
# T^2 Statistik
# Test 1_{T^2 >= k_alpha_0}
# Ablehnen von H_0

# Nicht Ablehnen von H_0
# p-Wert

Y_bar = 26.25615 2.991039
C = 38.8981 3.549813 3.549813 1.972143
T^2 = 7.546368
alpha_0 = 0.05
k = 7.504065
phi = 1
p = 0.04928746
```

Im vorliegenden Fall nimmt die Einstichproben- T^2 -Teststatistik einen größeren Wert als der kritische Wert an, es gilt damit $\phi(\Upsilon) = 1$ und man lehnt die Nullhypothese ab. Der korrespondierende p-Wert ist durch 0.049 gegeben.

46.5. Literaturhinweise

Die Theorie des Einstichproben- T^2 -Tests geht zurück auf Hotelling (1931).

47. Einfaktorielle Varianzanalyse

47.1. Anwendungsszenario

Das Anwendungsszenario einer einfaktoriellen multivariaten Varianzanalyse ist durch das Vorliegen von multivariaten Datenpunkten von zwei oder mehr Gruppen randomisierter experimenteller Einheiten gekennzeichnet, die sich hinsichtlich der Level eines experimentellen Faktors unterscheiden. Ist die Anzahl an Datenpunkten in jeder Gruppe gleich, so spricht von einem balancierten einfaktoriellen multivariaten Varianzanalysedesign. Von den Datenpunkten der i ten Gruppe bzw. des i ten Faktorlevels wird angenommen, dass sie Realisierungen von jeweils n_i unabhängigen und identisch multivariat normalverteilten Zufallsvektoren sind, deren wahre, aber unbekannte, Erwartungswertparameter sich potentiell über die Gruppen hinweg unterscheiden und deren wahrer, aber unbekannter, Kovarianzmatrixparameter über die Gruppen hinweg identisch ist. In diesen Grundannahmen handelt es beim Anwendungsszenario der einfaktoriellen multivariaten Varianzanalyse also um die Generalisierung des Einstichproben- T^2 -Test Szenarios zu zwei oder mehr Gruppen experimenteller Einheiten. Grundlegend wird vorausgesetzt, dass ein Interesse am einem inferentiellen Vergleich, der wahren, aber unbekannten, faktorlevelspezifischen Erwartungswertparameter besteht.

Anwendungsbeispiel

Als konkretes Anwendungsbeispiel betrachten wir die Analyse von Prä-Post-Interventions-BDI-Score und Prä-Post-Interventions-Glukokortikoidplasmalevel Differenzenwerten von drei Gruppen von jeweils 15 Patient:innen, die unterschiedliche Psychotherapiesettings (Face-to-face und Online) bzw. eine Wartelistenkontrollbedingung durchlaufen haben. Wir stellen dazu in Tabelle 47.1 einen simulierten Beispieldatensatz dar. Die erste Spalte von Tabelle 47.1 (COND) listet das spezifische Therapiesetting (F2F: Face-to-face, ONL: online, WLC: waitlist control) der Patient:innen auf. Die zweite Spalte (dBDI) listet die entsprechenden BDI-Score-Differenzwerte und die dritte Spalte (dGLU) die entsprechenden Glukokortikoidplasmalevel-Differenzwerte auf. In beiden Fällen zeigen positive Werte eine Abnahme der Depressionssymptomatik, negative Werte dagegen einer Zunahme der Depressionssymptomatik an. Abbildung 47.1 visualisiert diesen Datensatz sowie die gruppenspezifischen Stichprobenmittel und Stichprobenkovarianzen als Normalverteilungskonturen.

Tabelle 47.1. Prä-Post-Interventions-BDI-II-Score und -Glukokortikoidplasmalevel Differenzenwerte von drei Studiengruppen (F2F: Face-to-face, ONL: online, WLC: waitlist control) jeweils 15 Patient:innen

COND	dBDI	dGLU
F2F	11	4.3
F2F	10	3.9
F2F	12	3.5
F2F	7	2.6
F2F	10	3.3
F2F	12	3.5
F2F	9	3.1
F2F	9	3.6
F2F	9	5.6
F2F	11	3.6
F2F	7	3.4
F2F	9	4.0
F2F	11	5.6
F2F	14	5.3
F2F	8	3.2
ONL	6	3.1
ONL	8	2.7
ONL	7	2.1
ONL	8	3.1
ONL	11	2.8
ONL	9	2.8
ONL	9	3.7
ONL	8	2.7
ONL	6	3.6
ONL	7	1.4
ONL	9	1.3
ONL	8	3.4
ONL	7	3.7
ONL	7	2.3
ONL	9	3.4
WLC	-2	0.7
WLC	2	1.4
WLC	1	1.0
WLC	2	0.9
WLC	3	1.6
WLC	2	1.6
WLC	5	1.1
WLC	-1	-0.2
WLC	2	2.5
WLC	-2	0.4
WLC	1	0.7
WLC	3	-0.1
WLC	1	1.3
WLC	4	0.6
WLC	4	0.7

Folgender **R** Code demonstriert die Auswertung gruppenspezifischer Deskriptivstatistiken für diesen Datensatz.

```
# Studiengruppenspezifische Deskriptivstatistiken
D = read.csv("../data/703-einfaktorielle-varianzanalyse.csv")
m = 2
p = 3
k = 15
Y = array(dim = c(m,k,p))
Y[,1] = rbind(D$dBDI[D$COND == "F2F"],
              D$dGLU[D$COND == "F2F"])
Y[,2] = rbind(D$dBDI[D$COND == "ONL"],
              D$dGLU[D$COND == "ONL"])
Y[,3] = rbind(D$dBDI[D$COND == "WLC"],
              D$dGLU[D$COND == "WLC"])
y_bar_i = array(dim = c(m,p))
C_i = array(dim = c(m,m,p))
j_k = matrix(rep(1,k), nrow = k)
I_k = diag(k)
J_k = matrix(rep(1,k^2), nrow = k)
for (i in 1:p){
  y_bar_i[,i] = (1/k)*(Y[,i] %*% j_k)
  C_i[,i] = (1/(k-1))*(Y[,i] %*% (I_k-(1/k)*J_k) %*% t(Y[,i]))}

# Dateneinlesen
# Datendimension von Interesse
# Anzahl Gruppen
# Anzahl Datenpunkte pro Gruppe
# Datenarrayinitialisierung
# F2F dBDI Werte
# F2F dGLU Werte
# ONL dBDI Werte
# ONL dGLU Werte
# WLC dBDI Werte
# WLC dGLU Werte
# Stichprobenmittelarray
# Stichprobenkovarianzmatrixarray
# 1_{1}
# Einheitsmatrix I_1
# 1_{11}
# Gruppeniterationen
# Stichprobenmittel \bar{y}_i
# Stichprobenkovarianzmatrix C_i
```

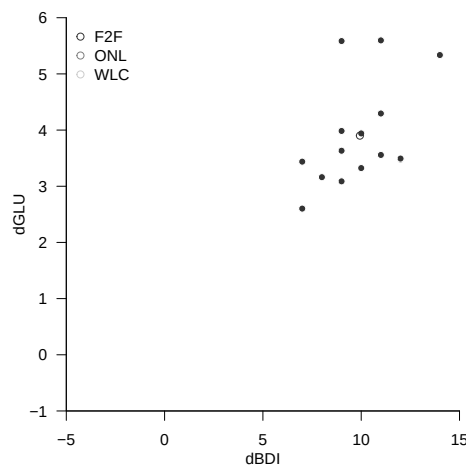


Abbildung 47.1. Deskriptivstatistiken der dBDI, dGLU Daten des Beispieldatensatzes. Jeder Punkt visualisiert die Daten einer Patient:in. Die Stichprobenkovarianz ist durch die 0.7 Isokontur einer zweidimensionalen Normalverteilung mit Erwartungswertparameter und Kovarianzmatrixparameter entsprechend dem Stichprobenmittel und der Stichprobenkovarianz der jeweiligen Gruppe dargestellt.

47.2. Modellformulierung und Modellschätzung

Wir definieren das Modell einfaktoriellen multivariaten Varianzanalyse wie folgt.

Definition 47.1 (Modell der einfaktoriellen multivariaten Varianzanalyse). Für $i = 1, \dots, p$ und $j = 1, \dots, n_i$ seien y_{ij} m -dimensionale Zufallsvektoren, die die $n := \sum_{i=1}^p n_i$ m -dimensionalen Datenpunkte eines einfaktoriellen multivariaten Varianzanalyseszenarios modellieren. Dann hat das Modell der einfaktoriellen multivariaten Varianzanalyse die strukturelle Form

$$y_{ij} = \mu_i + \varepsilon_{ij} \text{ mit } \varepsilon_{ij} \sim N(0_m, \Sigma) \text{ u.i.v. mit } \mu_i \in \mathbb{R}^m \text{ und } \Sigma \in \mathbb{R}^{m \times m} \text{ pd} \quad (47.1)$$

und die Datenverteilungsform

$$y_{ij} \sim N(\mu_i, \Sigma) \text{ u.v. mit } \mu_i \in \mathbb{R}^m \text{ und } \Sigma \in \mathbb{R}^{m \times m} \text{ pd.} \quad (47.2)$$

•

In Definition 47.1 bezeichnet n_i die Anzahl der Zufallsvektoren y_{ij} der i ten von p Gruppen experimenteller Einheiten. Im Falle eines balancierten Designs gilt offenbar $n_1 = \dots = n_p$. In diesem Fall setzen wir der Einfachheit halber $k := n_i$ für $i = 1, \dots, p$. In diesem Fall gilt für die Gesamtanzahl an Zufallsvektoren dann $n = pk$. Die Äquivalenz von struktureller Form und Datenverteilungsform der einfaktoriellen multivariaten Varianzanalyse ergibt sich mit Theorem 29.8 durch Transformation der ε_{ij} unter Multiplikation mit der Einheitsmatrix und unter Addition der jeweiligen gruppenspezifischen Erwartungswertparameter. Die wahren, aber unbekannten, Parameter $\mu_i, i = 1, \dots, p$ und Σ des einfaktoriellen multivariaten Varianzanalysemodells können anhand der in folgendem Theorem definierten Schätzer geschätzt werden.

Theorem 47.1 (Parameterschätzer der einfaktoriellen multivariaten Varianzanalyse). *Gegeben sei das Modell der einfaktoriellen multivariaten Varianzanalyse. Dann ist für $i = 1, \dots, p$*

$$\hat{\mu}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad (47.3)$$

ein unverzerrte Schätzer des gruppenspezifischen Erwartungswertparameters μ_i und

$$\hat{\Sigma} := \frac{1}{n-p} \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)(y_{ij} - \hat{\mu}_i)^T \quad (47.4)$$

ein unverzerrter Schätzer des Kovarianzmatrixparameters Σ .

◦

In Theorem 47.1 ist $\hat{\mu}_i$ offenbar das Stichprobenmittel der Zufallsvektoren der i ten Gruppe. $\hat{\Sigma}$ ist die gruppenunspezifische Stichprobenkovarianzmatrix aller Zufallsvektoren und entspricht der mit $1/(n-p)$ skalierten *Within-Group Sum-of-Squares Matrix*, die wir in Theorem 47.2 einführen werden. Folgender **R** demonstriert die Evaluation der Parameterschätzer mithilfe einer **R** Funktion.

```
estimate = function(Y){
  # Diese Funktion evaluiert die Parameterschätzer einer einfaktoriellen
  # multivariaten Varianzanalyse basierend auf einen m x k x p Datensatz Y.
  #
  # Input
  #   Y           : m x k x p Datenarray
  #
  # Output
  #   $mu_hat     : m x p \mu_i Parameterschätzer
  #   $Sigma_hat  : m x m \Sigma Parameterschätzer
  # -----
  # Dimensionsparameter
  d = dim(Y)
  m = d[1]
  k = d[2]
  p = d[3]
  # Datensatzdimensionen
  # Datendimension
  # Anzahl Datenpunkte pro Gruppe
  # Anzahl Gruppen

  # Erwartungswertparameterschätzer
  mu_hat_i = matrix(apply(Y,3,rowMeans), nrow = m)

  # Kovarianzmatrixparameterschätzer
  Sigma_hat = matrix(rep(0,m*m), nrow = m)
```

```

for(i in 1:p){
  for(j in 1:k){
    Sigma_hat = Sigma_hat + (1/(k*p-p))*(Y[,j,i] - mu_hat_i[,i]) %*% t(Y[,j,i] - mu_hat_i[,i])
  }
}

# Outputspezifikation
return(list(mu_hat_i = mu_hat_i, Sigma_hat = Sigma_hat))

```

Anstelle eines Beweises validieren wir die Aussage von Theorem 47.1 beispielhaft mithilfe folgender **R** Simulation, in der wir die Erwartungswerte der Schätzer durch ihre Stichprobenmittelwerte über Datensatzrealisierungen hinweg approximieren.

```

# Modellparameter
library(MASS)
p = 3 # multivariate Normalverteilungen
k = 15 # Anzahl Gruppen
m = 2 # Anzahl Datenpunkte pro Gruppe
mu_i = matrix(c(1,2,2,1,3,2,5), ncol = p) # Datendimension
Sigma = matrix(c(1,.5,.5,1), ncol = m) # Erwartungswertparameter
# Kovarianzmatrixparameter

# Simulationsparameter und Arrays
nsm = 1e2 # Anzahl Simulation
mu_hat_is = array(dim = c(m,p,nsm)) # \hat{\mu}_i Array
Sigma_hats = array(dim = c(m,m,nsm)) # \hat{\Sigma} Array

# Simulationen
for(s in 1:nsm){
  # Datengeneration
  Y = array(dim = c(m,k,p)) # Datenarray
  for(i in 1:p){
    Y[,i] = t(mvrnorm(k,mu_i[,i],Sigma)) # Datengeneration
  }
  S = estimate(Y) # Parameterschätzung
  mu_hat_is[,s] = S$mu_hat_i # \hat{\mu}_i
  Sigma_hats[,s] = S$Sigma_hat # \hat{\Sigma}
}

# Schätzererwartungswertschätzung
E_hat_mu_i_hat = apply(mu_hat_is, c(1,2), mean)
E_hat_Sigma_hat = apply(Sigma_hats, c(1,2), mean)

```

Wie in Abbildung 47.2 ergibt sich hier auch schon bei einem recht geringen Simulationsaufwand von 100 Datensatzrealisierungen eine gute Korrespondenz zwischen wahren, aber unbekannten, Parameterwerten $\mu_i, i = 1, \dots, p$ und Σ und den approximierten Erwartungswertparametern $\mathbb{E}(\hat{\mu}_i)$ für $i = 1, \dots, p$ und $\mathbb{E}(\hat{\Sigma})$.

Die Anwendung der Parameterschätzung auf die Daten des Beispieldatensatzes in Tabelle 47.1 ergibt folgende Resultate.

```

S = estimate(Y) # Parameterschätzung

      [,1]      [,2]      [,3]
[1,] 9.933333 7.933333 1.666667
[2,] 3.900525 2.803179 0.9594878

      [,1]      [,2]
[1,] 3.3142857 0.4022937
[2,] 0.4022937 0.6381336

```

47.3. Modellevaluation

Primäres Ziel einer einfaktoriellen multifaktoriellen Varianzanalyse ist meist das Testen der Nullhypothese

$$H_0 : \mu_1 = \dots = \mu_p. \quad (47.5)$$

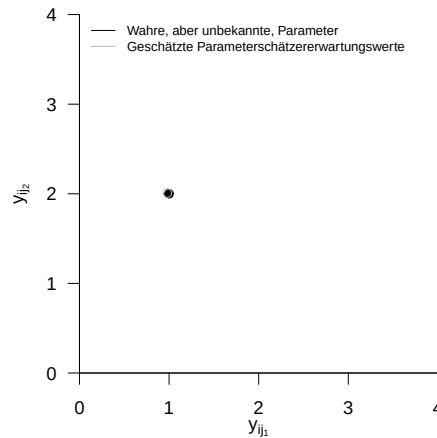


Abbildung 47.2. Simulationsbasierte Validierung des Theorems zu den Parameterschätzern der einfaktoriellen multivariaten Varianzanalyse am Beispiel von $m := 2, p := 3, k := 15$ und 100 Realisierungen der entsprechenden multivariat normalverteilten Zufallsvektoren

Diese Nullhypothese besagt, dass keine Unterschiede zwischen den wahren, aber unbekannten, Erwartungswertparametern der Stichprobengruppen bestehen. Die Alternativhypothese lautet somit

$$H_1 : \mu_{i_l} \neq \mu_{j_l} \text{ für mindestens ein Paar } i, j \text{ mit } i \neq j, 1 \leq i, j \leq p \quad (47.6)$$

und mindestens ein l mit $1 \leq l \leq m$.

Die Alternativhypothese besagt also, dass sich mindestens zwei wahre, aber unbekannten, Erwartungswertparameter in mindestens einer ihrer Komponenten unterscheiden. Wie aus dem Kontext der univariaten einfaktoriellen Varianzanalyse bekannt impliziert das Ablehnen der Nullhypothese auch hier keine Aussage über die genaue Form des inferierten Erwartungswertparameterunterschiedes.

Im Rahmen einfaktoriellen multivariaten Varianzanalyse können Tests der Nullhypothese mit verschiedenen Teststatistiken konstruiert werden. Diesen Teststatistiken ist gemein, dass sie auf eine Generalisierung der aus dem univariaten Fall bekannten Quadratsummenzerlegung der einfaktoriellen Varianzanalyse zurückgehen. Wir führen im nächsten Abschnitt zunächst diese sogenannte *Kreuzproduktsummenmatrizenzerlegung* der einfaktoriellen multivariaten Varianzanalyse ein. Nachfolgend betrachten wir dann die Modellevaluation mithilfe der *Wilks'-Λ-Statistik*.

Kreuzproduktsummenmatrizenzerlegung

Folgendes Theorem generalisiert die Quadratsummenzerlegung der einfaktoriellen Varianzanalyse auf das multivariate Anwendungsszenario.

Theorem 47.2 (Kreuzproduktsummenmatrizenzerlegung).

Gegeben sei das Modell der einfaktoriellen multivariaten Varianzanalyse. Weiterhin seien

$$\bar{y} := \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} \text{ und } \bar{y}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad (47.7)$$

das Gesamtstichprobenmittel und das i te Gruppenstichprobenmittel, respektive. Schließlich seien

$$T := \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})(y_{ij} - \bar{y})^T \quad \text{die Totale Sum-of-Squares Matrix}$$

$$B := \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \quad \text{die Between-Group Sum-of-Squares Matrix}$$

$$W := \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T \quad \text{die Within-Group Sum-of-Squares Matrix.}$$

Dann gilt

$$T = B + W. \quad (47.8)$$

◦

Beweis. Es gilt

$$\begin{aligned} T &= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})(y_{ij} - \bar{y})^T \\ &= \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})(y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^T \\ &= \sum_{i=1}^p \sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}))((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}))^T \\ &= \sum_{i=1}^p \sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T + 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y})^T + (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T) \\ &= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T + \sum_{j=1}^{n_i} 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y})^T + \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \right) \\ &= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T + 2 \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) \right) (\bar{y}_i - \bar{y})^T + n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \right) \\ &= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T + 2 \left(\sum_{j=1}^{n_i} \left(y_{ij} - \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \right) \right) (\bar{y}_i - \bar{y})^T + n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \right) \\ &= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T + 2 \left(\sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i} y_{ij} \right) (\bar{y}_i - \bar{y})^T + n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \right) \\ &= \sum_{i=1}^p \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T + n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T \right) \\ &= \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T + \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T \\ &= B + W. \end{aligned} \quad (47.9)$$

□

Die intuitive Interpretation der Totalen, Between-Group, und Within-Group Sum-of-Squares Matrizen ist analog zu den aus dem univariaten Szenario bekannten Begriffen:

Die Matrix T repräsentiert die totale Variabilität der Datenvektoren um das Gesamtstichprobenmittel, die Matrix B repräsentiert die Variabilität der Gruppenstichprobenmittel um das Gesamtstichprobenmittel und die Matrix W repräsentiert die Variabilität der Datenvektoren um ihre jeweiligen Gruppenstichprobenmittel. Wie im univariaten Fall wird auch hier also die Gesamtdatenvariabilität additiv in zwei unabhängige Beiträge zerlegt. Dabei kann die Matrix W auch als Maß für die Residualvariabilität verstanden werden, weil sie die verbleibende Variabilität nach Schätzung der Gruppenerwartungswertparameter quantifiziert. Offenbar gilt für den Schätzer $\hat{\Sigma}$ des gemeinsamen Stichprobenkovarianzmatrixparameters aus Theorem 47.1

$$W = (n - p)\hat{\Sigma}. \quad (47.10)$$

Folgender **R** demonstriert die Evaluation der in Theorem 47.2 definierten Matrizen mithilfe einer **R** Funktion.

```
sos = function(Y){
  # Diese Funktion evaluiert die Kreuzproduktsummenmatrizen T,B,W einer
  # einfaktoriellen Varianzanalyse basierend auf einem m x k x p Datensatz Y.
  #
  # Input
  #   Y      : m x k x p Datenarray
  #
  # Output
  #   $y_bar   : m x 1 Gesamtmittelwert
  #   $y_bar_i : m x p Gruppenmittelwerte
  #   $T       : m x m Total Sum of Squares Matrix
  #   $B       : m x m Between-Group Sum-of-Squares Matrix
  #   $W       : m x m Within Group Sum of Squares Matrix
  # -----
  d = dim(Y)
  m = d[1]
  k = d[2]
  p = d[3]
  # Datensatzdimensionen
  # Datendimension
  # Anzahl Datenpunkte pro Gruppe
  # Anzahl Gruppen

  # Mittelwerte
  y_bar_i = matrix(apply(Y,3,rowMeans), nrow = m)
  y_bar = matrix(rowMeans(y_bar_i), nrow = m)
  # Gruppenstichprobenmittel
  # Gesamtstichprobenmittel

  # Totale Sum-of-Squares Matrix
  T = matrix(rep(0,m*m), nrow = m)
  for(i in 1:p){
    for(j in 1:k){
      T = T + (Y[,j,i] - y_bar) %*% t(Y[,j,i] - y_bar)}

  # Between Sum of Squares Matrix
  B = matrix(rep(0,m*m), nrow = m)
  for(i in 1:p){
    B = B + k*(y_bar_i[,i] - y_bar) %*% t(y_bar_i[,i] - y_bar)}

  # Within Sum of Squares Matrix
  W = matrix(rep(0,m*m), nrow = m)
  for(i in 1:p){
    for(j in 1:k){
      W = W + (Y[,j,i] - y_bar_i[,i]) %*% t(Y[,j,i] - y_bar_i[,i])}}

  # Outputspezifikation
  return(list(y_bar_i = y_bar_i, y_bar = y_bar, T = T, B = B, W = W))}
```

Modellevaluation mit der Wilks'- Λ -Statistik

Basierend auf der Kreuzproduktsummenmatrizenzerlegung in Theorem 47.2 wurde eine Reihe von Teststatistiken für die einfaktorielle multivariate Varianzanalyse vorgeschlagen (Wilks (1932), Pillai (1955), Roy & Bose (1953), Hotelling (1951)). Wir betrachten hier exemplarisch lediglich die *Wilks'- Λ -Statistik* nach Wilks (1932). Im Gegensatz zur F -Teststatistik der univariaten einfaktoriellen Varianzanalyse sind die Frequentistischen Verteilungen von der Wilks'- Λ -Statistik bei Zutreffen der Nullhypothese nur für bestimmte Anwendungsszenarien, insbesondere bei kleinen Werten der Datendimension m und der Gruppenanzahl p , analytisch exakt zu bestimmen. In diesen Szenarien ist man auf f -Verteilungen mit m - und p -abhängigen Freiheitsgradparametern geführt. Für Anwendungsszenarien mit größeren Werten von m und/oder p existieren lediglich Approximationen der

Frequentistischen Verteilungen der Wilks'- Λ -Statistik, die nur asymptotisch für unendlich große Stichprobenumfänge $n \rightarrow \infty$ exakt sind. Auch diese Approximationen sind wiederum durch f -Verteilungen mit m - und p -abhängigen Freiheitsgradparametern gegeben. In der Anwendung unterscheiden sich Testentscheidungen basierend auf exakten oder approximativen Verteilungen der verschiedenen Teststatistiken meist nicht. Zur Absicherung dieser Aussage mögen im konkreten Fall von Datendimension und Gruppengröße Simulationen helfen, mögliche Unterschiede zwischen der Wilks'- Λ -Statistik und anderen Teststatistiken sowie der Approximation ihrer Verteilungen abzuschätzen.

Wir definieren zunächst die Wilks'- Λ -Statistik.

Definition 47.2 (Wilks'- Λ -Statistik). Gegeben sei das Modell der einfaktoriellen multivariaten Varianzanalyse sowie die Between-Group Sum-of-Squares Matrix B und die Within-Group Sum-of-Squares Matrix W . Dann ist *Wilks'- Λ -Statistik* definiert als

$$\Lambda := \frac{|W|}{|T|} = \frac{|W|}{|B + W|}. \quad (47.11)$$

•

Intuitiv misst Λ das Verhältnis von Residualdatenvariabilität, repräsentiert durch die Determinante von W , und Gesamtdatenvariabilität, repräsentiert durch die Determinante von $T = B + W$. Λ ist damit also analog zum Effektstärkenmaß η^2 der univariaten einfaktoriellen Varianzanalyse definiert. Im Falle der Gleichheit der gruppenspezifischen Stichprobenmittel gilt für Λ insbesondere

$$\bar{y}_1 = \dots = \bar{y}_p = \bar{y} \Rightarrow B = 0_{mm} \Rightarrow \Lambda = \frac{|W|}{|T|} = \frac{|W|}{|B + W|} = \frac{|W|}{|0_{mm} + W|} = \frac{|W|}{|W|} = 1. \quad (47.12)$$

Für ansteigende Unterschiede zwischen den gruppenspezifischen Stichprobenmittel $\bar{y}_i$ nimmt $|B + W|$ gegenüber $|W|$ zu und Λ somit ab. Ohne Beweis wollen wir dabei festhalten, dass $0 \leq \Lambda \leq 1$. Im Gegensatz zu den meisten bekannten Teststatistiken sprechen hier also *kleine Werte der Teststatistik Λ für eine Abweichung von der Nullhypothese*.

Äquivalent kann Λ auch in Form der Eigenwerte der Matrix $W^{-1}B$ angegeben werden. Dabei ist die Matrix $W^{-1}B$ das multivariate Analogon zum dem aus der univariaten einfaktoriellen Varianzanalyse bekanntem Verhältnis von Between-Group Sum-of-Squares und Within-Group Sum-of-Squares, welches die Grundlage für F -Teststatistik im univariaten Fall bildet. Es gilt folgendes Theorem.

Theorem 47.3 (Eigenwertform von der Wilks'- Λ -Statistik). *Es seien das Modell der einfaktoriellen multivariaten Varianzanalyse, die Between-Group Sum-of-Squares Matrix B , die Within-Group Sum-of-Squares Matrix W und die Wilks'- Λ -Statistik definiert wie oben. Weiterhin seien $\lambda_1, \dots, \lambda_s$ die Eigenwerte von $W^{-1}B$. Dann gilt*

$$\Lambda = \prod_{i=1}^s \frac{1}{1 + \lambda_i}. \quad (47.13)$$

◦

Wir verzichten auf einen Beweis.

Für spezielle, durch die Datendimension m und die Gruppenanzahl p charakterisierte Anwendungsszenarien stellt Wilks (1932) analytische Formen für die Verteilungen von Transformationen von Λ unter der Nullhypothese bereit. Dabei entsprechen diese analytischen Formen den Verteilungen von f -Zufallsvariablen mit bestimmten, durch die Datendimension und Gruppenanzahl des Anwendungsszenario vorgegebenen Freiheitsgradparametern. Wir fassen die Anwendungsszenarien, die betreffende Transformation von Λ und die analytischen Formen ihrer Frequentistischen Verteilungen bei Zutreffen der Nullhypothese in folgendem Theorem zusammen.

Theorem 47.4 (Spezielle H_0 Verteilungen von Transformationen der Wilks'- Λ -Statistik).

Es seien das Modell der einfaktoriellen Varianzanalyse und die Wilks'- Λ -Statistik definiert wie oben und es gelte außerdem die Nullhypothese

$$H_0 : \mu_1 = \dots = \mu_p. \quad (47.14)$$

Dann sind für die in den ersten beiden Tabellenspalten aufgeführten Spezialfällen die in der dritten Tabellenspalte aufgeführten Statistiken f -Zufallsvariablen und zwar mit den in der vierten Tabellenspalte aufgeführten Freiheitsgradparametern.

Datendimension m	Gruppenanzahl p	Statistik	f -Verteilungsparameter
Beliebig	2	$\frac{1-\Lambda}{\Lambda} \frac{n-p-m+1}{m}$	$m, n-p-m+1$
Beliebig	3	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{n-p-m+1}{m}$	$2m, 2(n-p-m+1)$
1	Beliebig	$\frac{1-\Lambda}{\Lambda} \frac{n-p}{p-1}$	$p-1, n-p$
2	Beliebig	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{n-p-1}{p-1}$	$2(p-1), 2(n-p-1)$

◦

In Abbildung 47.3 visualisieren wir exemplarische simulationsbasierte Validierungen der in Theorem 47.4 angegebenen Verteilungen.

Rao (1951) hat für allgemeine Anwendungsszenarien die in folgendem Theorem angegebenen Approximationen von Verteilungen von Transformationen der Wilks'- Λ -Statistik vorgeschlagen. Man beachte, dass sich die in diesem Theorem definierte Teststatistik τ wie in Abbildung 47.4 gezeigt zu Λ reziprok verhält. Geringe Werte von Λ als Evidenz gegen die Nullhypothese entsprechen also hohen Werten von τ .

Theorem 47.5 (Approximative H_0 Verteilungen von Transformationen der Wilks'- Λ -Statistik).

Es seien das Modell der einfaktoriellen Varianzanalyse und die Wilks'- Λ -Statistik definiert wie oben und es gelte außerdem die Nullhypothese

$$H_0 : \mu_1 = \dots = \mu_p. \quad (47.15)$$

Dann ist die Statistik

$$\tau := \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{\nu_2}{\nu_1} \quad (47.16)$$

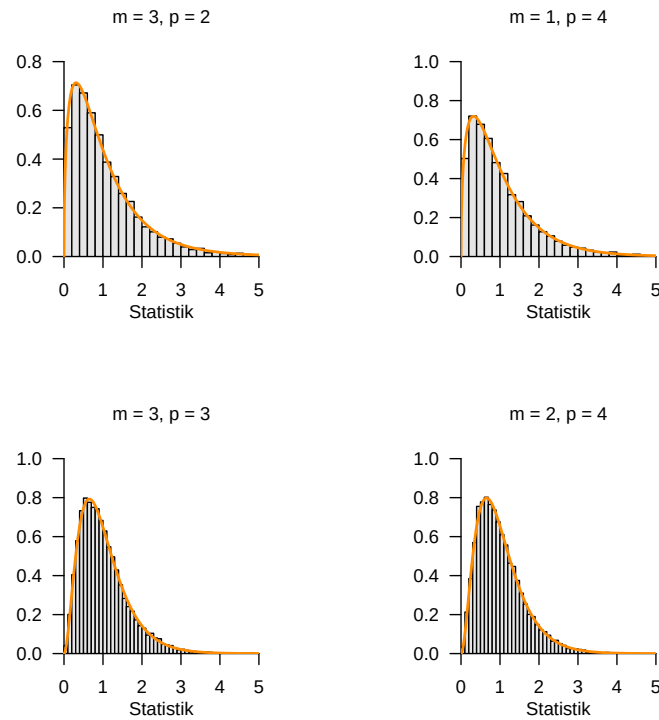


Abbildung 47.3. Simulation spezieller Verteilungen der Wilks'- Λ -Statistik Transformationen bei Zutreffen der Nullhypothese.

mit

$$\nu_1 := m(p-1) \text{ und } \nu_2 := wt - \frac{1}{2}(m(p-1) - 2) \quad (47.17)$$

sowie

$$w := n - 1 - \frac{1}{2}(m+p) \text{ und } t := \sqrt{\frac{m^2(p-1)^2 - 4}{m^2 + (p-1)^2 - 5}} \quad (47.18)$$

approximativ f -verteilt mit Freiheitsgradparametern ν_1 und ν_2 .

◦

In Abbildung 47.5 visualisieren wir exemplarische simulationsbasierte Validierungen der in Theorem 47.5 angegebenen Verteilungen.

Mithilfe der von Rao (1951) bestimmten approximativen Verteilungen der transformierten Wilks'- Λ -Statistik τ bei Zutreffen der Nullhypothese können wir nun einen Hypothesentest für die einfaktorielle multivariate Varianzanalyse angeben.

Theorem 47.6 (Wilks'- Λ -basierter Test, Testumfangkontrolle, p-Wert). *Gegeben seien Modell der einfaktoriellen multivariaten Varianzanalyse und die transformierte Wilks'- Λ -Statistik τ mit Verteilungsparametern ν_1, ν_2 wie oben definiert. Weiterhin sei für $Y = (y_1, \dots, y_n)$ der kritische Wert-basierte Test*

$$\phi(Y) := 1_{\{\tau > k\}} := \begin{cases} 1 & \tau > k \\ 0 & \tau \leq k \end{cases} \quad (47.19)$$

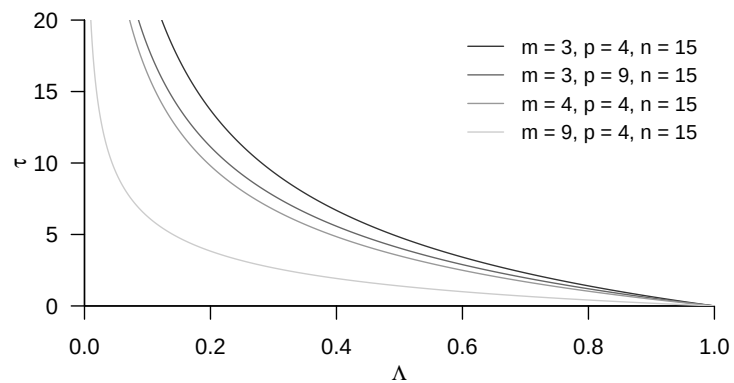


Abbildung 47.4. Beziehung der Wilks'- Λ -Statistik und der von Rao (1951) betrachteten τ Statistik in exemplarischen Anwendungsszenarien.

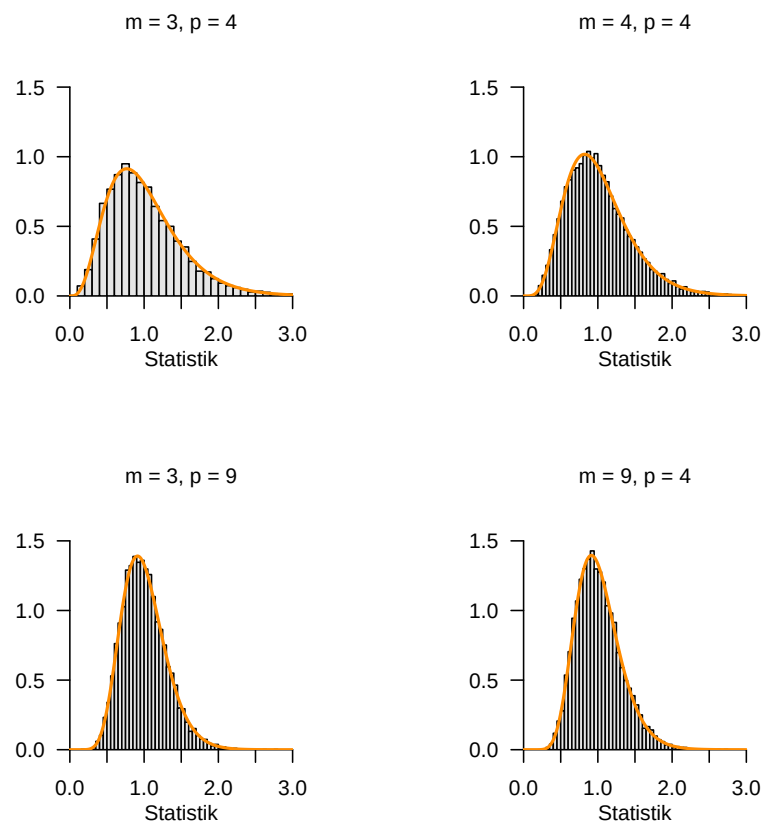


Abbildung 47.5. Simulation approximativer Verteilungen der Wilks'- Λ -Statistik Transformationen bei Zutreffen der Nullhypothese.

definiert. Dann ist ϕ genau dann ein Level- α_0 -Test mit Testumfang α , wenn

$$k := k_{\alpha_0} := F^{-1}(1 - \alpha_0; \nu_1, \nu_2) \quad (47.20)$$

ist und der p -Wert einer realisierten τ -Teststatistik $\tilde{\tau}$ ergibt sich zu

$$p\text{-Wert} = \mathbb{P}(\tau \geq \tilde{\tau}) = 1 - F(\tilde{\tau}; \nu_1, \nu_2) \quad (47.21)$$

◦

Wir verzichten auf einen Beweis von Theorem 47.6, welcher analog zu den entsprechenden Beweisen zum Beispiel beim Einstichproben- T^2 -Test geführt werden kann. Die in Theorem 47.6 implizite Wahl eines kritischen Wertes zur Testumfangkontrolle im Anwendungsszenario $m := 3, p := 4$ und $n_i = 15$ für $i = 1, \dots, p$ und damit $\nu_1 = 9$ und $\nu_2 = 132$ bei einem Signifikanzlevel von $\alpha_0 = 0.05$ visualisieren wir in Abbildung 47.6.

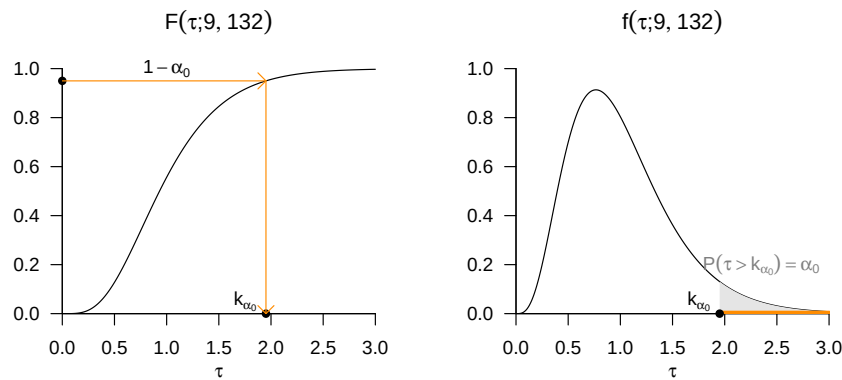


Abbildung 47.6. Testumfangkontrolle durch Selektion eines α_0 -abhängigen kritischen Wertes für den Wilks' Λ -Statistik-basierten Test anhand von KVF und WDF der transformierten Wilks' Λ -Statistik τ .

Praktisches Vorgehen

In der Praxis entsprechen obige Ergebnisse dann folgendem Vorgehen bei der Durchführung einer einfaktoriellen multivariaten Varianzanalyse: Man unterstellt, dass ein vorliegender Datensatz von $i = 1, \dots, p$ Gruppen von m -dimensionalen Datenvektoren für jeweils $j = 1, \dots, n_i$ eine Realisation von Zufallsvektoren $y_{ij} \sim N(\mu_i, \Sigma)$ mit unbekannten, gruppenspezifischen Erwartungswertparametern $\mu_i \in \mathbb{R}^m$ und gruppenunspezifischem Kovarianzmatrixparameter $\Sigma \in \mathbb{R}^{m \times m}$ pd ist. Man möchte entscheiden, ob eher die Nullhypothese $H_0 : \mu_1 = \dots = \mu_p$ identischer wahrer, aber unbekannter, Erwartungswertparameter zutrifft oder eher nicht. Zu diesem Zweck wählt man zunächst ein Signifikanzlevel α_0 und bestimmt dann den zugehörigen Freiheitsgradparameter-abhängigen kritischen Wert k_{α_0} . Zum Beispiel gilt bei Wahl von $\alpha_0 := 0.05$ im Szenario von dreidimensionalen Datenvektoren ($m = 3$), vier Gruppen ($p = 4$) und $n_i = 15$ experimentellen Einheiten pro Gruppe und damit einer Gesamtzahl von $n = 60$ Datenpunkten, dass $\nu_1 = 9$ und $\nu_2 = 132$ sind und sich der in Theorem 47.6 definierte kritische Wert zu $k_{\alpha_0} = F^{-1}(1 - 0.05; 9, 132) \approx 1.95$ ergibt. Basierend auf den realisierten Datensatz berechnet man dann zunächst die Wilk's- Λ -Statistik und den resultierenden, m, p, n -abhängigen realisierten Wert der transformierten

Wilk's- Λ -Statistik τ . Wenn der berechnete Wert von τ größer als k_{α_0} ist, lehnt man die Nullhypothese ab, andernfalls nicht. Die oben entwickelte Theorie zur Testumfangkontrolle bei der einfaktoriellen multivariaten Varianzanalyse auf Grundlage von Wilks'- Λ -Statistik garantiert dann, dass man im Mittel in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt.

Folgender **R** Code demonstriert die Anwendung des in Theorem 47.6 definierten Hypothesentests auf bei Zutreffen der Nullhypothese unter dem Modell der einfaktoriellen multivariaten Varianzanalyse generierte Daten und validiert seine Kontrolle des Testumfangs für die in Abbildung 47.5 skizzierten Anwendungsszenarien.

```
# Szenarioparameter
library(MASS)
nsm      = 1e4
M        = c(3,3,4,9)
P        = c(4,9,4,4)
k        = 15
N        = k*P
alpha_0  = 0.05
nsc      = length(M)
TAU      = matrix(rep(NaN,nsm*nsc), ncol = nsc)
NU       = matrix(rep(NaN,2*nsc), ncol = nsc)
KA       = rep(NaN, nsc)
PHI      = matrix(rep(0,nsm*nsc), ncol = nsc)

# Multivariate Normalverteilungen
# Datensimulationsanzahl
# Datendimension
# Gruppenanzahl
# Datenpunkte pro Gruppe
# Gesamtanzahl Datenpunkte
# \alpha_0
# Szenarienzahl
# Statistik Array
# Parameter Array
# Kritische Werte
# Testarray

# Simulationen
for(sc in 1:nsc){
  # Szenarioiterationen

  # Modellparameter
  m      = M[sc]
  p      = P[sc]
  n      = N[sc]
  mu_i   = matrix(rep(0,m), nrow = m)
  Sigma  = diag(m)

  # Varianzanalyse Parameter
  w      = n-1-(1/2)*(m+p)
  t      = sqrt((m^2*(p-1)^2-4)/(m^2+(p-1)^2-5))
  nu_1   = m*(p-1)
  nu_2   = w*t-(1/2)*(m*(p-1)-2)
  KA[sc] = qf(1-alpha_0,nu_1,nu_2)

  # w
  # t
  # \nu_1
  # \nu_2
  # kritischer Wert

  # Datensimulationen
  for(sm in 1:nsm){
    Y      = array(dim = c(m,k,p))
    for(i in 1:p){
      Y[,i] = t(mvrnorm(k,mu_i,Sigma))}
    S      = sos(Y)
    L      = det(S$W)/det(S$W + S$B)
    tau    = ((1-L^(1/t))/L^(1/t))*(nu_2/nu_1)
    PHI[sm,sc] = tau > KA[sc]}

    # Datenarrayinitialisierung
    # Gruppeniterationen
    # Datensimulation
    # Stichprobenmittel und Sum of Squares Matrizen
    # Wilks' Lambda
    # Statistik
    # Test
  }
}
```

```
Kritische Werte      : 1.951729 1.547641 1.821661 1.565152
Geschätzte Testumfänge: 0.0472 0.0496 0.0522 0.0488
```

47.4. Anwendungsbeispiel

Wir betrachten das eingangs diskutierte Anwendungsbeispiel eines simulierten zweidimensionalen Datensatzes dreier Therapiegruppen. Wir wollen nun mithilfe einer einfaktoriellen Varianzanalyse für diesen Datensatz die Nullhypothese identischer Gruppenerwartungswertparameter überprüfen. Folgender **R** Code implementiert das praktische Vorgehen für ein Signifikanzlevel von $\alpha_0 := 0.05$.

```
# Einlesen und Präprozessierung des Datensatzes
D      = read.csv("./_data/703-einfaktorielle-varianzanalyse.csv")
m      = 2
p      = 3
k      = 15
n      = p*k
Y      = array(dim = c(m,k,p))
Y[,1]  = rbind(D$dBDI[D$COND == "F2F"],
               D$dGLU[D$COND == "F2F"])
Y[,2]  = rbind(D$dBDI[D$COND == "ONL"],
               D$dGLU[D$COND == "ONL"])

# Dateneinlesen
# Datendimension von Interesse
# Anzahl Gruppen
# Anzahl Datenpunkte pro Gruppe
# Gesamtanzahl Datenpunkte
# Datenarrayinitialisierung
# F2F dBDI Werte
# F2F dGLU Werte
# ONL dBDI Werte
# ONL dGLU Werte
```



```

Y[,3]      = rbind(D$dBDI[D$COND == "WLC"],          # WLC dBDI Werte
                  D$dGLU[D$COND == "WLC"])          # WLC dGLU Werte

# Einfaktorielle Varianzanalyse
alpha_0    = 0.05                                  # Signifikanzlevel
S           = sos(Y)                                # Sum of Squares Matrizen
L           = det(S$W)/det(S$W + S$B)               # Wilks' Lambda
w           = n-1-(1/2)*(m+p)                        # w
t           = sqrt((m^2*(p-1)^2-4)/(m^2+(p-1)^2-5))  # t
nu_1        = m*(p-1)                                # \nu_1
nu_2        = w*t-(1/2)*(m*(p-1)-2)                 # \nu_2
tau         = ((1-L^(1/t))/L^(1/t))*(nu_2/nu_1)      # Teststatistik
k_alpha_0   = qf(1-alpha_0,nu_1,nu_2)              # kritischer Wert
phi         = as.numeric(tau > k_alpha_0)           # Nullhypothesentest
P           = 1-pf(tau,nu_1,nu_2)                   # Überschreitungswahrscheinlichkeit

Wilks' Lambda    0.1569049
tau              31.25301
nu_1             4
nu_2            82
phi             1
P(tau > tau_tilde) 8.881784e-16

```

Im vorliegenden Fall wird die Nullhypothese identischer Gruppenerwartungswertparameter also verworfen. Schließlich validieren wir obige Analyse im Sinne eines Black-Box-Verfahrens mithilfe der **R** `lm()` und `Manova()` Funktionen.

```
library(car)
```

```
Warning: package 'car' was built under R version 4.4.3
```

```
Warning: package 'carData' was built under R version 4.4.3
```

```

D           = read.csv("./_data/703-einfaktorielle-varianzanalyse.csv") # Dateneinlesen
model       = lm(cbind(D$dBDI,D$dGLU) ~ D$COND, D)                     # Modellspezifikation
Manova(model, test.statistic = "Wilks")                               # Einfaktorielle Varianzanalyse

```

```

Type II MANOVA Tests: Wilks test statistic
      Df test stat approx F num Df den Df    Pr(>F)
D$COND 2    0.1569   31.253      4    82 8.346e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

47.5. Literaturhinweise

Die Theorie der einfaktoriellen multivariaten Varianzanalyse geht zurück auf Wilks (1932). Anderson (2003) gibt eine Einführung in die Approximationstheorie für multivariate Modelle, das Wissen um die exakten Verteilungen der Teststatistiken um 1970 wird von Rao (1972) zusammengefasst.

48. Kanonische Korrelationsanalyse

48.1. Anwendungsszenario und Grundzüge

Ausgangspunkt einer kanonischen Korrelationsanalyse ist die exploratorische Frage nach der Stärke des Zusammenhangs einer multivariaten unabhängigen Variablen x und einer multivariaten abhängigen Variable y . Im Kontext der kanonischen Korrelationsanalyse werden diese Variablen auch oft als *Prädiktoren* und *Kriterien*, respektive, bezeichnet. Dabei werden vorliegende Daten von sowohl der Prädiktoren als auch der Kriterien als Realisierungen von unabhängig und identisch verteilten m_x - bzw. m_y -dimensionalen Zufallsvektoren betrachtet. Um die Frage nach der Stärke des Zusammenhangs von Prädiktoren und Kriterien zu beantworten, betrachtet die kanonische Korrelationsanalyse Linearkombinationen der Komponenten der Prädiktoren und Kriterien. Wir bezeichnen die Linearkombinationen von x und y mit Parametervektoren $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$ im Folgenden mit

$$\xi := a^T x \text{ und } v := b^T y \quad (48.1)$$

ξ und v sind dann als Linearkombinationen von Zufallsvariablen selbst Zufallsvariablen, deren Korrelation $\rho(\xi, v)$ berechnet werden kann. Die kanonische Korrelationsanalyse bestimmt dann die Parametervektoren a und b so, dass die Korrelation von ξ und v so groß wie möglich wird. Ist eine solche Parametervektorkombination und ihre entsprechende Korrelation, die dann als kanonische Korrelation bezeichnet gefunden, so kann ξ als bester Prädiktor und v als “am besten prädictierbares Kriterium” interpretiert werden.

Für Skalare α und β sind die Korrelationen $\rho(\xi, v)$ und $\rho(\alpha\xi, \beta v)$ allerdings, wie im Theorem zu Kovarianz und Korrelation bei linear-affinen Transformationen gezeigt, identisch. Die kanonische Korrelationsanalyse sucht deshalb speziell Parametervektoren a und b , für die $\rho(\xi, v)$ einerseits maximal ist und für die ξ und v gleichzeitig jeweils eine standardisierte Varianz von 1 haben. Da aufgrund des Theorem zu Kovarianz und Korrelation bei linear-affinen Transformationen die Varianzen zu verschiedenen skalaren Vielfachen von ξ und v verschieden sind, legt diese die Parametervektorwerte a und b , für die $\rho(\xi, v)$ maximal ist, eindeutig fest. Zur Bestimmung einer kanonischen Korrelation und der Parametervektoren von a und b ist man also auf ein Optimierungsproblem mit Nebenbedingungen geführt.

In unserer Darstellung kanonischen Korrelationsanalyse folgen wir Mardia et al. (1979). Dabei werden die Prädiktor- und Kriterienzufallsvektoren x und y in einem Zufallsvektor

$$z := \begin{pmatrix} x \\ y \end{pmatrix} \quad (48.2)$$

zusammengefasst, für den wir durchgängig annehmen wollen, dass $\mathbb{E}(z) = 0_m$ mit $m = m_x + m_y$ gilt. Im Anwendungskontext entspricht dies der Annahme, dass vor Durchführung der kanonischen Korrelationsanalyse die Stichprobenmittel des Stichprobenmittels von den beobachteten Daten vor Durchführung der kanonischen Korrelationsanalyse subtrahiert

werden. Da, wie wir sehen werden, die Schätzung der kanonischen Korrelation allerdings lediglich auf den Stichprobenkovarianzmatrizen beruht ist dieser Schritt verzichtbar.

Der mathematische Fokus der Entwicklung der kanonischen Korrelationsanalyse nach Mardia et al. (1979) liegt damit auf der Kovarianzmatrix $\mathbb{C}(z)$. Speziell ergeben sich die Kovarianzen von Linearkombinationen von x und y aus Matrixprodukten von $\mathbb{C}(z)$ und es können einige Matrixtheoreme, die in Kapitel 22 und Kapitel 9.7 dokumentiert sind angewendet werden. Generell wird in der Entwicklung nach Mardia et al. (1979) ein Optimierungsansatz mithilfe von Lagrangefunktionen, wie in den Originalarbeiten von Hotelling (1935) und Hotelling (1936) gewählt, zugunsten der Eigenanalyse von Matrixprodukten supprimiert. Für die Entwicklung mit einem Lagrangeansatz verweisen wir auf zum Beispiel Anderson (2003).

Im Folgenden diskutieren wir in Kapitel 48.2 zunächst zwei Theoreme, die direkt durch die kanonische Korrelationsanalyse motiviert sind und die, zusammen mit den oben erwähnten Theoremen in Kapitel 22 und Kapitel 9.7 das analytische und probabilistische Fundament der kanonischen Korrelationsanalyse bilden. In Kapitel 48.3 führen wir mit den *kanonischen Korrelationen*, den *kanonischen Variaten* und den *kanonischen Koeffizientenvektoren* dann die zentralen Begriffe der kanonischen Korrelationsanalyse ein. In Kapitel 48.4 diskutieren wir dann, wie diese Größen auf Basis der Stichprobenkovarianzmatrix eines Datensatzes von Prädiktoren- und Kriterienrealisierungen geschätzt werden können. In Kapitel 48.5 schließlich wenden wir die kanonische Korrelationsanalyse auf das unten skizzierte Anwendungsbeispiel an.

Anwendungsbeispiel

Als konkretes Anwendungsbeispiel für eine kanonische Korrelationsanalyse betrachten wir einen simulierten Datensatz zur Effektivität einer Psychotherapie bei Depressionen. Dabei seien pro Patient:in vier Variablen verfügbar: Zum einen als Maße für die Therapiequalität die Dauer der Psychotherapie (DUR) und die klinische Erfahrung der behandelnden Psychotherapeut:in (EXP); zum anderen als Maße für die Reduktion der Depressionssymptomatik sowohl BDI Score und Glukokortikoidplasmalevel Differenzwerte zwischen Beginn und Ende der Therapie (dBDI und dGLU), wobei positive Werte eine Verringerung der Depressionssymptomatik anzeigen sollen. Wir stellen uns vor, dass man im Sinne einer exploratorischen Analyse daran interessiert ist, inwieweit die Variablen DUR (x_1) und EXP (x_2) als Prädiktoren (unabhängige Variablen) mit den Variablen dBDI (y_1) und dGLU (y_2) als Kriterien (abhängige Variablen) zusammenhängen. Im Sinne obiger Bezeichner gilt hier also $m_x = m_y = 2$. Tabelle 48.1 zeigt dazu einen Beispieldatensatz für $n = 20$ Patient:innen.

Tabelle 48.1. Beispielhafte Prädiktoren- und Kriterienrealisierungen im Kontext der Psychotherapie bei Depressionen

DUR	EXP	dBDI	dGLU
27.9	7.8	35.5	6.1
15.3	9.3	25.0	4.0
17.4	2.1	19.7	1.7
21.5	6.5	28.8	2.6
28.2	1.3	29.4	1.9
14.0	2.7	17.2	0.9
28.0	3.9	32.9	2.0
28.9	0.1	28.3	4.1
23.2	3.8	25.8	3.9
22.6	8.7	31.3	3.8
11.2	3.4	14.4	2.1
14.1	4.8	18.4	2.0
13.5	6.0	19.1	5.0
23.7	4.9	28.0	2.6
17.7	1.9	20.3	2.1
25.4	8.3	34.8	4.4
20.0	6.7	27.6	4.0
24.4	7.9	31.9	3.9
29.8	1.1	32.2	1.0
17.6	7.2	24.6	1.9

48.2. Mathematischer Hintergrund

Folgendes Theorem bildet das analytische Fundament der kanonischen Korrelationsanalyse.

Theorem 48.1 (Maximierung quadratischer Formen mit Nebenbedingungen). *$A \in \mathbb{R}^{m \times m}, B \in \mathbb{R}^{m \times m}$ seien symmetrische Matrizen und λ_1 sei der größte Eigenwert von $B^{-1}A$ mit assoziiertem Eigenvektor $v_1 \in \mathbb{R}^m$. Dann ist λ_1 eine Lösung des Optimierungsproblems*

$$\max_x x^T A x \text{ unter der Nebenbedingung } x^T B x = 1.$$

(48.3)

◦

Beweis. $B^{1/2}$ sei die symmetrische Quadratwurzel von B und es sei

$$y := B^{1/2}x \Leftrightarrow x = B^{-1/2}y$$

(48.4)

Dann kann mit der symmetrischen Matrix

$$K := B^{-1/2}AB^{-1/2} \in \mathbb{R}^{m \times m}$$

(48.5)

das Optimierungsproblem Gleichung 48.3 geschrieben werden als

$$\max_y y^T K y \text{ unter der Nebenbedingung } y^T y = 1.$$

(48.6)

Dies gilt, weil

$$\max_x x^T A x \Leftrightarrow \max_y (B^{-1/2} y)^T A (B^{-1/2} y) \Leftrightarrow \max_y y^T B^{-1/2} A B^{-1/2} y \Leftrightarrow \max_y y^T K y \quad (48.7)$$

und

$$x^T B x = 1 \Leftrightarrow y^T B^{-1/2} B B^{-1/2} y = 1 \Leftrightarrow y^T y = 1. \quad (48.8)$$

Weil K eine symmetrische Matrix ist, existiert die Orthonormalzerlegung (vgl. Kapitel 9.7)

$$K = Q \Lambda Q^T, \quad (48.9)$$

wobei die Spalten der orthogonalen Matrix Q die Eigenvektoren von K und die Diagonalelemente von Λ die zugehörigen Eigenwerte von K sind. Mit der orthogonalen Matrix Q aus obiger Orthonormalzerlegung sei nun

$$z := Q^T y \Leftrightarrow y := Q z. \quad (48.10)$$

Dann kann das Optimierungsproblem Gleichung 48.6 geschrieben werden als

$$\max_z \sum_{i=1}^m \lambda_i z_i^2 \text{ unter der Nebenbedingung } z^T z = 1, \quad (48.11)$$

weil

$$\max_y y^T K y \Leftrightarrow \max_z (Q z)^T K (Q z) \Leftrightarrow \max_z z^T Q^T K Q z \Leftrightarrow \max_z z^T \Lambda z \Leftrightarrow \max_z \sum_{i=1}^m \lambda_i z_i^2 \quad (48.12)$$

und

$$y^T y = 1 \Leftrightarrow (Q z)^T Q z = 1 \Leftrightarrow z^T Q^T Q z = 1 \Leftrightarrow z^T z = 1. \quad (48.13)$$

Die Eigenwerte von K seien nun absteigend sortiert, also $\lambda_1 \geq \dots \geq \lambda_m$. Dann gilt für das Optimierungsproblem Gleichung 48.11, dass

$$\max_z \sum_{i=1}^m \lambda_i z_i^2 \leq \lambda_1, \quad (48.14)$$

weil

$$\max_z \sum_{i=1}^m \lambda_i z_i^2 \leq \max_z \sum_{i=1}^m \lambda_1 z_i^2 = \lambda_1 \max_z \sum_{i=1}^m z_i^2 = \lambda_1 \quad (48.15)$$

wobei sich die letzte Gleichung aus der Nebenbedingung $z^T z = 1$ ergibt. Schließlich gilt

$$\max_z \sum_{i=1}^m \lambda_i z_i^2 = \lambda_1, \quad (48.16)$$

für $z := e_1 = (1, 0, \dots, 0)^T$. Zusammenfassend heißt das, dass $z = e_1$ eine Lösung des Optimierungsproblem Gleichung 48.11 ist und das λ_1 das entsprechende Maximum ist. Damit ergibt sich aber sofort, dass dann

$$y = Q z = Q e_1 = q_1 \text{ und } x = B^{-1/2} q_1 \quad (48.17)$$

Lösungen der äquivalenten Optimierungsprobleme Gleichung 48.6 und Gleichung 48.3, respektive, sind. Nach Konstruktion ist q_1 ein Eigenvektor von $B^{-1/2} A B^{-1/2}$ und nach obigem Theorem zu Eigenwerten und Eigenvektoren von Matrixprodukten damit auch ein Eigenvektor von

$$B^{-1/2} B^{-1/2} A = B^{-1} A \quad (48.18)$$

und die zugehörigen Eigenwerte sind gleich. Damit aber folgt, dass der größte Eigenwert von $B^{-1} A$ und sein assoziierter Eigenvektor eine Lösung von

$$\max_x x^T A x \text{ unter der Nebenbedingung } x^T B x = 1. \quad (48.19)$$

ist. □

Nach Wortlaut des Theorems gilt also für die Funktion

$$f : \mathbb{R}^m \rightarrow \mathbb{R}, x \mapsto f(x) := x^T A x, \quad (48.20)$$

dass

$$v_1 = \operatorname{argmax}_x x^T A x \text{ unter der Nebenbedingung } x^T B x = 1 \quad (48.21)$$

und dass

$$\lambda_1 = \max_x x^T A x \text{ unter der Nebenbedingung } x^T B x = 1. \quad (48.22)$$

Das folgende Theorem setzt die für die kanonischen Korrelationsanalyse zentralen Größen der Varianzen von Linearkombinationen von Zufallsvektoren und der Korrelation von Linearkombinationen von Zufallsvektoren in Bezug zur Kovarianzmatrix der gemeinsamen Verteilung der Zufallsvektoren und bildet das probabilistische Fundament der kanonischen Korrelationsanalyse.

Theorem 48.2 (Linearkombinationen von Zufallsvektorpartitionen).

Es sei

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \text{ mit } \mathbb{E}(z) = 0_m \text{ und } \mathbb{C}(z) = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \quad (48.23)$$

ein m -dimensionaler partitionierter Zufallsvektor sowie sein Erwartungswertvektor und seine Kovarianzmatrix, respektive. Weiterhin seien für $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$ die Zufallsvariablen

$$\xi := a^T x \text{ und } v := b^T y \quad (48.24)$$

als Linearkombinationen der Komponenten von x und y definiert. Dann gelten

- (1) $\mathbb{V}(\xi) = a^T \Sigma_{xx} a$
- (2) $\mathbb{V}(v) = b^T \Sigma_{yy} b$
- (2) $\rho(\xi, v) = a^T \Sigma_{xy} b$, wenn $\mathbb{V}(\xi) = 1$ und $\mathbb{V}(v) = 1$.

◦

Beweis. Wir betrachten zunächst die Varianz von ξ . Mit dem Varianzverschiebungssatz gilt

$$\begin{aligned} \mathbb{V}(\xi) &= \mathbb{E}(\xi\xi) - \mathbb{E}(\xi)\mathbb{E}(\xi) \\ &= \mathbb{E}((a^T x)(a^T x)) - \mathbb{E}(a^T x)\mathbb{E}(a^T x) \\ &= \mathbb{E}((a^T x)(a^T x)^T) - \mathbb{E}(a^T x)\mathbb{E}(a^T x) \\ &= \mathbb{E}(a^T x x^T a) - \mathbb{E}(a^T x)\mathbb{E}(a^T x) \\ &= a^T \mathbb{E}(x x^T) a - a^T \mathbb{E}(x) a^T \mathbb{E}(x) \\ &= a^T \mathbb{E}(x x^T) a - a^T 0_{m_x} a^T 0_{m_x} \\ &= a^T \Sigma_{xx} a. \end{aligned} \quad (48.25)$$

Der Beweis zur Varianz von v folgt dann analog. Mit der Definition der Korrelation von Zufallsvariablen und mit $\mathbb{V}(\xi) = \mathbb{V}(v) = 1$ und dem Kovarianzverschiebungssatz gilt

$$\begin{aligned} \rho(\xi, v) &= \frac{\mathbb{C}(\xi, v)}{\sqrt{\mathbb{V}(\xi)}\sqrt{\mathbb{V}(v)}} \\ &= \frac{\mathbb{C}(\xi, v)}{\sqrt{1}\sqrt{1}} \\ &= \mathbb{C}(\xi, v) \\ &= \mathbb{E}(\xi v) - \mathbb{E}(\xi)\mathbb{E}(v) \\ &= \mathbb{E}((a^T x)(b^T y)) - \mathbb{E}(a^T x)\mathbb{E}(b^T y) \\ &= \mathbb{E}((a^T x)(b^T y)^T) - \mathbb{E}(a^T x)\mathbb{E}(b^T y) \\ &= \mathbb{E}(a^T x y^T b) - \mathbb{E}(a^T x)\mathbb{E}(b^T y) \\ &= a^T \mathbb{E}(x y^T) b - a^T \mathbb{E}(x) b^T \mathbb{E}(y) \\ &= a^T \mathbb{E}(x y^T) b - a^T 0_{m_x} b^T 0_{m_y} \\ &= a^T \Sigma_{xy} b. \end{aligned} \quad (48.26)$$

□

Die Varianz der Zufallsvariable $a^T x$ ergibt sich also als die “quadrierte” Linearkombination von Σ_{xx} und die Varianz der Zufallsvariable $b^T y$ ergibt sich als die “quadrierte” Linearkombination von Σ_{yy} . Die Korrelation der Zufallsvariablen $a^T x$ und $b^T y$ schließlich ergibt sich “quadrierte” Linearkombination von Σ_{xy} .

48.3. Modellformulierung

Mit den beiden in Kapitel 48.2 diskutierten Theoremen ist es nun möglich, die Begriffe der *kanonischen Koeffizientenvektoren*, der *kanonischen Variate*, und schließlich der *kanonischen Korrelationen* zu definieren.

Definition 48.1 (Kanonische Koeffizientenvektoren, Variate, Korrelationen).

Es seien

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \text{ mit } \mathbb{E}(z) := 0_m \text{ und } \mathbb{C}(z) := \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \quad (48.27)$$

ein m -dimensionaler partitionierter Zufallsvektor sowie sein Erwartungswert und seine Kovarianzmatrix, respektive. Weiterhin sei

$$K := \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} \in \mathbb{R}^{m_x \times m_y} \quad (48.28)$$

mit der Singulärwertzerlegung

$$K = A \Lambda B^T, \quad (48.29)$$

wobei

$$A := (\alpha_1 \quad \dots \quad \alpha_k) \in \mathbb{R}^{m_x \times m_y} \text{ und } B := (\beta_1 \quad \dots \quad \beta_k) \in \mathbb{R}^{m_y \times m_y} \quad (48.30)$$

die orthogonale Matrix der Eigenvektoren von KK^T und die orthogonale Matrix der Eigenvektoren von $K^T K$, respektive, bezeichnen und

$$\Lambda := \text{diag}(\lambda_1^{1/2}, \dots, \lambda_k^{1/2}) \in \mathbb{R}^{m_y \times m_y}, \quad (48.31)$$

die Diagonalmatrix der Quadratwurzeln der zugehörigen absteigend geordneten Eigenwerte bezeichnet. Schließlich seien für $i = 1, \dots, k$

$$a_i := \Sigma_{xx}^{-1/2} \alpha_i \in \mathbb{R}^{m_x} \text{ und } b_i := \Sigma_{yy}^{-1/2} \beta_i \in \mathbb{R}^{m_y}. \quad (48.32)$$

Dann heißen für $i = 1, \dots, k$

- (1) $a_i \in \mathbb{R}^{m_x}$ und $b_i \in \mathbb{R}^{m_y}$ die *iten kanonischen Koeffizientenvektoren*,
- (2) die Zufallsvektoren $\xi_i := a_i^T x$ und $v_i := b_i^T y$ die *iten kanonischen Variaten* und
- (3) $\rho_i := \lambda_i^{1/2}$ die *ite kanonische Korrelation*.

•

Ihre Bedeutsamkeit erlangen diese Begriffe durch ihre Eigenschaften, die wir im folgenden Theorem zusammenfassen.

Theorem 48.3 (Eigenschaften kanonischer Korrelationen und Variaten). *Es seien*

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \text{ mit } \mathbb{E}(z) := 0_m \text{ und } \mathbb{C}(z) := \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \quad (48.33)$$

ein m -dimensionaler partitionierter Zufallsvektor sowie sein Erwartungswert und seine Kovarianzmatrix, respektive. Weiterhin seien für $i = 1, \dots, k$ die kanonischen Koeffizientenvektoren a_i, b_i , die kanonischen Variaten ξ, v_i und die kanonischen Korrelationen ρ_i definiert wie oben. Dann gilt, dass für $1 \leq r \leq k$ das Maximum des r ten restringierten Optimierungsproblems

$$\phi_r = \max_{a,b} a^T \Sigma_{xy} b \quad (48.34)$$

unter den Nebenbedingungen

$$a^T \Sigma_{xx} a = 1, \quad b^T \Sigma_{yy} b = 1, \quad a_i^T \Sigma_{xx} a = 0 \text{ für } i = 1, \dots, r-1 \quad (48.35)$$

(1) den Wert $\phi_r = \rho_r$ hat und (2) bei $a = a_r$ und $b = b_r$ angenommen wird.

◦

Beweis. Wir betrachten das restringierte Optimierungsproblem

$$\phi_r^2 = \max_{a,b} (a^T \Sigma_{xy} b)^2 \text{ u.d.N. } a^T \Sigma_{xx} a = 1, b^T \Sigma_{yy} b = 1, a_i^T \Sigma_{xx} a = 0, i = 1, \dots, r-1 \quad (48.36)$$

Wir folgen Mardia et al. (1979), S. 284 und gehen schrittweise vor, d.h. wir lösen das restringierte Optimierungsproblem

$$\phi_r^2 = \max_a \left(\max_b (a^T \Sigma_{xy} b)^2 \text{ u.d.N. } b^T \Sigma_{yy} b = 1 \right) \text{ u.d.N. } a^T \Sigma_{xx} a = 1, a_i^T \Sigma_{xx} a = 0, i = 1, \dots, r-1 \quad (48.37)$$

von innen nach außen.

Schritt (1)

Wir wählen wir zunächst ein festes $a \in \mathbb{R}^m$ und betrachten das restringierte Optimierungsproblem

$$\max_b (a^T \Sigma_{xy} b)^2 \text{ u.d.N. } b^T \Sigma_{yy} b = 1 \quad (48.38)$$

Dieses Optimierungsproblem kann geschrieben werden als

$$\max_b b^T \Sigma_{yx} a a^T \Sigma_{xy} b \text{ u.d.N. } b^T \Sigma_{yy} b = 1, \quad (48.39)$$

weil gilt, dass

$$(a^T \Sigma_{xy} b)^2 = (a^T \Sigma_{xy} b) (a^T \Sigma_{xy} b) = (a^T \Sigma_{xy} b)^T a^T \Sigma_{xy} b = b^T \Sigma_{yx} a a^T \Sigma_{xy} b. \quad (48.40)$$

Das Optimierungsproblem Gleichung 48.39 kann nun mithilfe des Theorems zur Maximierung quadratischer Formen mit Nebenbedingungen gelöst werden. Im Sinne dieses Theorems setzen wir dazu

$$A := \Sigma_{yx} a a^T \Sigma_{xy} \text{ und } B := \Sigma_{yy}. \quad (48.41)$$

Dann hat Gleichung 48.39 die Form

$$\max_b b^T A b \text{ unter der Nebenbedingung } b^T B b = 1, \quad (48.42)$$

Das Maximum von Gleichung 48.42 entspricht nach dem Theorem zur Maximierung quadratischer Formen mit Nebenbedingungen dem größten Eigenwert von

$$B^{-1} A = \Sigma_{yy}^{-1} \Sigma_{yx} a a^T \Sigma_{xy} \quad (48.43)$$

Der größte Eigenwert von $\Sigma_{yy}^{-1} \Sigma_{yx} a a^T \Sigma_{xy}$ wiederum kann mithilfe des Theorems zum Eigenwert und Eigenvektor eines Matrixvektorprodukts bestimmt werden. Im Sinne dieses Theorems setzen wir dazu

$$A := \Sigma_{yy}^{-1} \Sigma_{yx}, \quad b := a, \quad B := \Sigma_{xy} \quad (48.44)$$

und erhalten für den betreffenden Eigenwert

$$\lambda_a = b^T B A a = a^T \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} a. \quad (48.45)$$

als Lösung (Maximum) des restringierten Optimierungsproblems

$$\max_b (a^T \Sigma_{xy} b)^2 \text{ u.d.N. } b^T \Sigma_{yy} b = 1 \quad (48.46)$$

Schritt (2)

Basierend auf Schritt (1) verbleibt die Lösung des restringierten Optimierungsproblem

$$\phi_r^2 = \max_a a^T \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} a \text{ u.d.N. } a^T \Sigma_{xx} a = 1, a_i^T \Sigma_{xx} a = 0, i = 1, \dots, r-1 \quad (48.47)$$

Dazu halten wir zunächst fest, dass (**kka_opt_3?**) mit den Definitionen von α_i und K in der Definition der Kanonischen Koeffizientenvektoren, Variaten, und Korrelationen geschrieben werden kann als

$$\phi_r^2 = \max_{\alpha} \alpha^T K K^T \alpha \text{ u.d.N. } \alpha^T \alpha = 1, \alpha_i^T \alpha = 0, i = 1, \dots, r-1, \quad (48.48)$$

denn

$$\begin{aligned} \phi_r^2 &= \max_a a^T \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} a \text{ u.d.N. } a^T \Sigma_{xx} a = 1, a_i^T \Sigma_{xx} a = 0 \Leftrightarrow \\ \phi_r^2 &= \max_{\alpha} a^T \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} a \text{ u.d.N. } \alpha^T \Sigma_{xx}^{-1/2} \Sigma_{xx} \Sigma_{xx}^{-1/2} \alpha = 1, \alpha_i^T \Sigma_{xx}^{-1/2} \Sigma_{xx} \Sigma_{xx}^{-1/2} \alpha = 0 \\ \phi_r^2 &= \max_{\alpha} \alpha^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \Sigma_{xx}^{-1/2} \alpha \text{ u.d.N. } \alpha^T \alpha = 1, \alpha_i^T \alpha = 0 \\ \phi_r^2 &= \max_{\alpha} \alpha^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} \Sigma_{yy}^{-1/2} \Sigma_{yx} \Sigma_{xx}^{-1/2} \alpha \text{ u.d.N. } \alpha^T \alpha = 1, \alpha_i^T \alpha = 0 \\ \phi_r^2 &= \max_{\alpha} \alpha^T K K^T \alpha \text{ u.d.N. } \alpha^T \alpha = 1, \alpha_i^T \alpha = 0 \end{aligned} \quad (48.49)$$

Dabei sind nach der betreffenden Definition die α_i die Eigenvektoren von $K K^T$ mit den $i = 1, \dots, r-1$ größten Eigenwerten. Nach dem Theorem zur Maximierung quadratischer Formen mit Nebenbedingungen ist die Lösung von Gleichung 48.48 der größte Eigenwert von $K K^T$ mit seinem assoziierten Eigenvektor. Die Nebenbedingung $\alpha_i^T \alpha = 0$ schränkt diese Wahl auf den r -größten Eigenwert und seinen assoziierten Eigenvektor α_r ein. Mit der Definition von Eigenwerten und Eigenvektoren gilt also

$$\phi_r^2 = \alpha_r^T K K^T \alpha_r = \alpha_r^T \lambda_r \alpha_r = \lambda_r \alpha_r^T \alpha_r = \lambda_r. \quad (48.50)$$

Wir haben also gezeigt, dass das restringierte Optimierungsproblem des Theorems den Maximumwert $\phi_r = \lambda_r^{1/2}$ hat. Es bleibt zu zeigen, dass dieser Maximumwert für a_r und b_r angenommen wird.

Schritt (3)

Einsetzen von a_r und b_r in $a^T \Sigma_{xy} b$ ergibt mit

$$K = A \Lambda B^T \Leftrightarrow K B = A \Lambda B^T B \Leftrightarrow K B = A \Lambda \Leftrightarrow K \beta_r = \alpha_r \lambda_r^{1/2} \quad (48.51)$$

dass

$$a_r^T \Sigma_{xy} b_r = \alpha_r^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} \beta_r = \alpha_r^T K \beta_r = \alpha_r^T \alpha_r \lambda_r^{1/2} = \rho_r \quad (48.52)$$

Also nimmt $a^T \Sigma_{xy} b$ bei a_r und b_r seinen restringierten Maximalwert λ_r an. $\square$

ϕ_1 ist also die größtmögliche Korrelation von

$$\xi = a^T x \text{ und } v = b^T y \quad (48.53)$$

unter den Nebenbedingungen

$$\mathbb{V}(\xi) = 1 \text{ und } \mathbb{V}(v) = 1 \quad (48.54)$$

und erfüllt damit die Forderungen an *die* kanonische Korrelatione. ϕ_r mit $r > 1$ ist die größtmögliche Korrelation von

$$\xi = a^T x \text{ und } v = b^T y \quad (48.55)$$

unter den Nebenbedingungen

$$\mathbb{V}(\xi) = 1, \mathbb{V}(v) = 1 \text{ und } \mathbb{C}(\xi_i, \xi) = 0 \text{ für die kanonischen Variate } \xi_i \text{ mit } i = 1, \dots, r-1. \quad (48.56)$$

Simulationsbeispiel

Wir betrachten das Beispiel (vgl. Uurtio et al. (2018))

$$p(x) = N(x; 0_4, I_4) \text{ und } p(y|x) = N(y; Lx, G) \quad (48.57)$$

mit

$$L := \begin{pmatrix} 0.0 & 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -1.0 \end{pmatrix} \text{ und } G := \begin{pmatrix} 0.2 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.3 \end{pmatrix} \quad (48.58)$$

Hier gilt offenbar $m_x = 4, m_y = 3, m = 7$ und

$$\begin{aligned} y_1 &= x_3 + \varepsilon_1 \\ y_2 &= x_1 + \varepsilon_2 \\ y_3 &= -x_4 + \varepsilon_3 \end{aligned} \quad (48.59)$$

mit

$$x_1 \sim N(0, 1), x_3 \sim N(0, 1), x_4 \sim N(0, 1) \quad (48.60)$$

und

$$\varepsilon_1 \sim N(0, 0.2), \varepsilon_2 \sim N(0, 0.4), \varepsilon_3 \sim N(0, 0.3) \quad (48.61)$$

Mit dem Theorem zu gemeinsamen Normalverteilungen (vgl. Kapitel 29) ergibt sich, dass

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N(0_7, \Sigma) \quad (48.62)$$

mit

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \quad (48.63)$$

wobei

$$\Sigma_{xx} = I_4, \quad \Sigma_{xy} = L^T, \quad \Sigma_{yx} = L \text{ und } \Sigma_{yy} = G + LL^T. \quad (48.64)$$

Explizit ergibt sich also

$$\Sigma = \begin{pmatrix} I_4 & L^T \\ L & G + LL^T \end{pmatrix} = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & -1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 1.2 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.4 & 0.0 \\ 0.0 & 0.0 & 0.0 & -1.0 & 0.0 & 0.0 & 1.3 \end{pmatrix} \quad (48.65)$$

Folgender **R** Code definiert zunächst den Kovarianzmatrixparameter der gemeinsamen Verteilung von x und y .

```
# R Pakete für Matrizenrechnung
library(expm)

# Modellparameter
L = matrix(c(0,0,1, 0,
             1,0,0, 0,
             0,0,0,-1),
           nrow = 3,
           byrow = T)
G = diag(c(0.2,0.4,0.3))

# Kovarianzmatrixpartition
Sigma_xx = diag(4)
Sigma_xy = t(L)
Sigma_yx = L
Sigma_yy = G + L %*% t(L)
Sigma = rbind(cbind(Sigma_xx, Sigma_xy), cbind(Sigma_yx, Sigma_yy))
print(Sigma)
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  1    0    0    0  0.0  1.0  0.0
[2,]  0    1    0    0  0.0  0.0  0.0
[3,]  0    0    1    0  1.0  0.0  0.0
[4,]  0    0    0    1  0.0  0.0 -1.0
[5,]  0    0    1    0  1.2  0.0  0.0
[6,]  1    0    0    0  0.0  1.4  0.0
[7,]  0    0    0   -1  0.0  0.0  1.3
```

Anhand von Definition 48.1 bestimmt folgender **R** Code dann basierend auf obigem Kovarianzmatrixparameter die kanonischen Korrelationen und kanonischen Koeffizientenvektoren.

```
# Evaluation der iten kanonischen Koeffizientenvektoren und Korrelationen
K = sqrtm(solve(Sigma_xx)) %*% Sigma_xy %*% sqrtm(solve(Sigma_yy)) # K
ALB = svd(K) # K = A\Lambda V
A = ALB$u # A
Lambda = ALB$d # Lambda
B = ALB$v # B
rho = Lambda # \rho_i = \lambda_i^{1/2}
a = sqrtm(solve(Sigma_xx)) %*% A # a_i = \Sigma_{xx}^{-1/2} \alpha_i
b = sqrtm(solve(Sigma_yy)) %*% B # b_i = \Sigma_{yy}^{-1/2} \beta_i

rho_1 = 0.9128709 , a_1^T = ( 0 0 -1 0 ), b_1^T = ( -0.9128709 0 0 )
rho_2 = 0.877058 , a_2^T = ( 0 0 0 1 ), b_2^T = ( 0 0 -0.877058 )
rho_3 = 0.8451543 , a_3^T = ( -1 0 0 0 ), b_3^T = ( 0 -0.8451543 0 )
```

48.4. Modellschätzung

Zur Schätzung einer kanonischen Korrelationsanalyse wird die Kovarianzmatrix $\mathbb{C}(z)$ des gemeinsamen Zufallsvektors z von Prädiktoren und Kriterien durch ihr Stichprobenäquivalent C ersetzt. Dies ist die Aussage folgender Definition.

Definition 48.2 (Schätzer der kanonischen Korrelationsanalyse). Für $i = 1, \dots, n$ seien

$$z_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix} \text{ mit } \mathbb{E}(z_i) := 0_m \text{ und } \mathbb{C}(z_i) := \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \quad (48.66)$$

unabhängig und identisch verteilte m -dimensionale partitionierte Zufallsvektoren sowie ihr Erwartungswert und ihre Kovarianzmatrix, respektive, und

$$C := \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \quad (48.67)$$

sei ihre Stichprobenkovarianzmatrix. Dann sind für $i = 1, \dots, k := \min\{m_x, m_y\}$

$$\hat{a}_i := C_{xx}^{-1/2} \hat{\alpha}_i \in \mathbb{R}^{m_x}, \quad \hat{b}_i := C_{yy}^{-1/2} \hat{\beta}_i \in \mathbb{R}^{m_y} \text{ und } \hat{\rho}_i := \hat{\lambda}_i^{1/2} \quad (48.68)$$

Schätzer der i ten kanonischen Koeffizientenvektoren und kanonischen Korrelationen, respektive. Dabei sind mit

$$\hat{K} := C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2} \in \mathbb{R}^{m_x \times m_y} \quad (48.69)$$

$\hat{\alpha}_i$ und $\hat{\lambda}_i$ der i te Eigenvektor und sein zugehöriger Eigenwert von $\hat{K}\hat{K}^T$ und $\hat{\beta}_i$ der entsprechende Eigenvektor von $\hat{K}^T\hat{K}$.

•

Wir verzichten auf eine Diskussion der Güte dieser Schätzung.

Simulationsbeispiel

Mithilfe folgenden **R** Codes verdeutlichen wir uns die Schätzung kanonischer Korrelationen und Koeffizientenvektoren in dem oben betrachteten Beispiel. Dazu generieren wir Realisierungen der Prädiktoren und Kriterien bei Stichprobenumfängen zwischen $n = 100$ und $n = 1000$. Abbildung 48.1 visualisiert die wahren, aber unbekannten, kanonischen Korrelationen ρ_1, ρ_2, ρ_3 und ihre für jede Simulation basierend auf der Stichprobenkovarianzmatrix des realisierten Datensatzes geschätzte Äquivalente $\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3$. Die Variabilität der Schätzung nimmt mit zunehmenden Stichprobenumfang ab. Abbildung 48.2 visualisiert die Absolutwerte des wahren, aber unbekannten, ersten kanonischen Koeffizientenvektors a_1 sowie ihre entsprechenden Schätzungen als $\hat{a}_1$. Auch hier nimmt die Variabilität der Schätzung mit zunehmenden Stichprobenumfang ab. Allerdings ist zu beachten, dass es sich hierbei um die Absolutwerte des kanonischen Koeffizientenvektors handelt und das Vorzeichen abhängig von der Schätzung auch im Widerspruch zum wahren, aber unbekannten, Wert des kanonischen Koeffizientenvektors stehen kann.

```
# R Pakete
library(MASS)
library(expm)

# Modellparameter
m_x = 4
m_y = 3
k = min(m_x, m_y)
L = matrix(c(0, 0, 1, 0, 0, 0, 0, 0, 0, -1), nrow = 3, byrow = 3)
G = diag(c(0.2, 0.4, 0.3))
Sigma_xx = diag(4)
Sigma_xy = t(L)
Sigma_yx = L
Sigma_yy = G + L %*% t(L)
Sigma = rbind(cbind(Sigma_xx, Sigma_xy), cbind(Sigma_yx, Sigma_yy))
K = sqrtm(solve(Sigma_xx)) %*% Sigma_xy %*% sqrtm(solve(Sigma_yy))
ALB = svd(K)
A = ALB$u
Lambda = ALB$d
B = ALB$v
rho = Lambda
a = sqrtm(solve(Sigma_xx)) %*% A
b = sqrtm(solve(Sigma_yy)) %*% B
```

```
# Simulationen
n = 1e1:1e3
rho_hat = matrix(rep(NA, length(n)*k), nrow = k)
a_1_hat = matrix(rep(NA, length(n)*m_x), nrow = m_x)
for(i in 1:length(n)){

  # Datengeneration
  Y = t(mvnorm(n[i], rep(0, m_x+m_y), Sigma))
  I_n = diag(n[i])
  J_n = matrix(rep(1, n[i]^2), nrow = n[i])

  # Stichprobenkovarianzmatrixpartition
  C = (1/(n[i]-1)) * (Y %*% (I_n - (1/n[i]) * J_n) %*% t(Y))
  C_xx = C[1:m_x, 1:m_x]
  C_xy = C[1:m_x, (m_x+1):(m_x+m_y)]
  C_yx = C[(m_x+1):(m_x+m_y), 1:m_x]
  C_yy = C[(m_x+1):(m_x+m_y), (m_x+1):(m_x+m_y)]
```

```
# Kanonische Korrelationsanalyse
K_hat = sqrtm(solve(C_xx)) %*% C_xy %*% sqrtm(solve(C_yy))
ALB_hat = svd(K_hat)
A_hat = ALB_hat$u
Lambda_hat = ALB_hat$d
B_hat = ALB_hat$v
a_hat = sqrtm(solve(C_xx)) %*% A_hat
b_hat = sqrtm(solve(C_yy)) %*% B_hat
rho_hat[,i] = as.matrix(Lambda_hat)
a_i_hat[,i] = a_hat[,i]
}
```

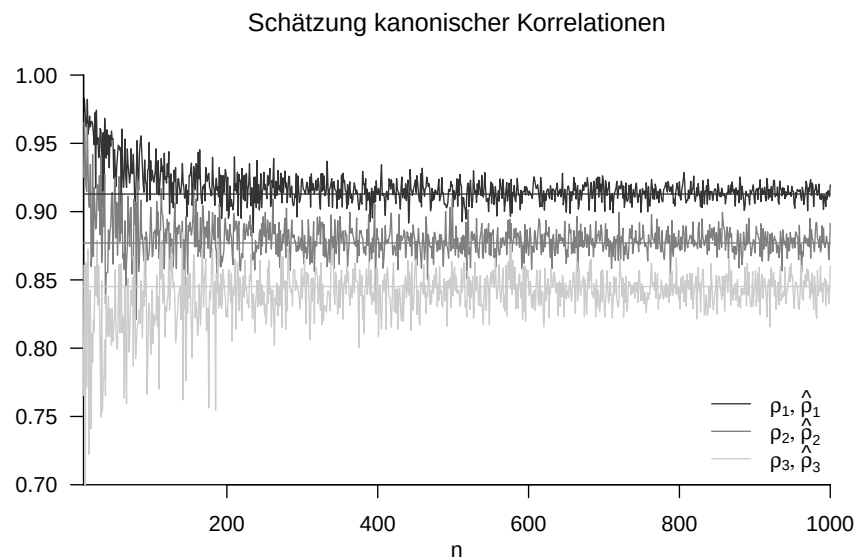


Abbildung 48.1. Schätzung kanonischer Korrelationen im Simulationsbeispiel

48.5. Anwendungsbeispiel

Zuletzt wollen wir die konkrete Berechnung einer kanonischen Korrelationsanalyse im Kontext des in Kapitel 48.1 diskutierten Anwendungsbeispiels demonstrieren. Folgender **R** Code implementiert die Berechnung der kanonischen Korrelationen und kanonischen Koefizientenvektoren für diesen Datensatz.

```
# R Paket
library(expm)

# Datenpräprozessierung
fname = "../data/704-kanonische-korrelationsanalyse.csv"
D = read.table(fname, sep = ",", header = TRUE)
x = as.matrix(cbind(D$DUR, D$EXP))
y = as.matrix(cbind(D$dBDI, D$dGLU))
n = nrow(x)
m_x = ncol(x)
m_y = ncol(y)
Y = t(cbind(x,y))

# Stichprobenkovarianzmatrixpartition
I_n = diag(n)
J_n = matrix(rep(1,n^2), nrow = n)
C = (1/(n-1))*Y %*% (I_n - (1/n)*J_n) %*% t(Y)
C_xx = C[1:m_x, 1:m_x]
C_xy = C[1:m_x, (m_x+1):(m_x+m_y)]
C_yx = C[(m_x+1):(m_x+m_y), 1:m_x]
C_yy = C[(m_x+1):(m_x+m_y), (m_x+1):(m_x+m_y)]

# Kanonische Korrelationsanalyse
K_hat = sqrtm(solve(C_xx)) %*% C_xy %*% sqrtm(solve(C_yy))
```

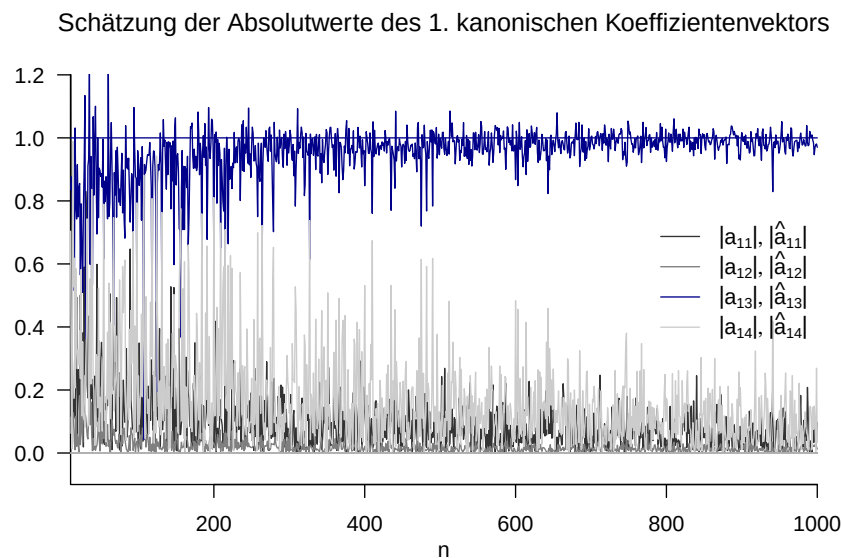


Abbildung 48.2. Schätzung der Absolutwerte des ersten kanonischen Koeffizientenvektors im Simulationsbeispiel

```
ALB_hat = svd(K_hat)
A_hat   = ALB_hat$u
Lambda_hat = ALB_hat$d
B_hat   = ALB_hat$v
a_hat   = sqrtm(solve(C_xx)) %*% A_hat
b_hat   = sqrtm(solve(C_yy)) %*% B_hat
rho_hat = as.matrix(Lambda_hat)
```

```
rho_hat_1 : 0.9950575
a_hat_1   : -0.1623409 -0.173979
b_hat_1   : -0.1554175 -0.05025419
rho_hat_2 : 0.5010358
a_hat_2   : -0.06026274 0.3118808
b_hat_2   : -0.08128072 0.7773036
```

Neben der Implementation mithilfe einer Singulärwertzerlegung bietet **R** auch eine direkte Bestimmung mithilfe der `cancor()` Funktion an. Folgender **R** Code demonstriert das entsprechende Vorgehen.

```
# Datenpräprozessierung
fname = "../data/704-kanonische-korrelationsanalyse.csv"
D      = read.table(fname, sep = ",", header = TRUE)
x      = as.matrix(cbind(D$DUR, D$EXP))
y      = as.matrix(cbind(D$dBDI, D$dGLU))
cca    = cancor(x,y)
```

```
rho_hat_1 : 0.9950575
rho_hat_2 : 0.5010358
```

Man findet also, dass die geschätzte maximale Korrelation einer Linearkombinationen der Prädiktorvariablen DUR und EXP mit einer Linearkombination der Kriterien dBDI und dGLU mit $\hat{\rho}_1 = 0.99$ sehr hoch ist. Man kann daraus schließen, dass in diesem Fall die Prädiktorvariablen gemeinschaftlich hoch mit den Kriterien assoziiert sind. Es ergeben sich hier insbesondere die Linearkombinationen

$$\xi = 0.16 \text{ DUR} + 0.17 \text{ EXP und } v = 0.15 \text{ dBDI} + 0.05 \text{ dGLU} \quad (48.70)$$

als bester Prädiktor und als am besten prädizierbares Kriterium, respektive. Die Dauer der Psychotherapie und die Erfahrung der behandelnden Psychotherapeut:in scheinen, bei der aktuellen Datenskalisierung zur bestmöglichen Prädiktion der Therapiegüte also in etwa gleichbedeutend, bei dem bestprädizierbarem Kriterium der Therapieeffizienz trägt bei der aktuellen Datenskalisierung die BDI Score Reduktion etwas mehr bei als die Glukokortikoidplasmalevel Reduktion bei.

48.6. Literaturhinweise

Die kanonische Korrelationsanalyse geht zurück auf Hotelling (1935) und Hotelling (1936).

Teil VII.

Psychometrische Verfahren

49. Psychologische Tests

Tests und Fragebögen sind in der alltäglichen Erfahrung unterschiedlicher Natur. Mit dem Begriff *Test* verbindet man vermutlich am ehesten kleinere Prüfungen, bei denen es korrekte Antworten gibt und eine Leistungsfähigkeit überprüft wird. Mit dem Begriff des *Fragebogens* verbindet man vermutlich mehr oder weniger interessante Sammlungen von Fragen, bei denen man eventuell zu einem wahrheitsgemäßen Antwortverhalten aufgefordert wird, bei denen es aber auch klar ist, dass es keine richtigen oder falschen Antworten gibt. Obwohl mit Tests und Fragebögen in der alltäglichen Erfahrung also durchaus unterschiedliche Verfahren verbunden sind, unterscheidet die psychologische Testtheorie zwischen diesen nicht grundlegend. Historisch ist die Analyse von Testdaten sicherlich zunächst durch die Analyse von Leistungstestdaten geprägt, allerdings werden heutzutage die dort entwickelten Konzepte der Testtheorie auf die gleiche Art und Weise zur Analyse von klinischen Fragebogendaten genutzt. In diesem Abschnitt wollen wir zunächst einen kurzen Überblick zum Begriff des *Psychologischen Tests* geben, die sogenannten *Testgütekriterien* auflisten, einige gängige *Normierungen* von Testrohdaten vorstellen und schließlich kurz die drei wesentlichen *Testtheorien* skizzieren.

49.1. Zum Begriff des psychologische Tests

Um uns dem Testbegriff zu nähern, betrachten wir zunächst die Definitionen eines Tests nach Moosbrugger & Kelava (2012), Bühner (2010) und Krauth (1995).

Nach Moosbrugger & Kelava (2012) ist ein psychologischer Test “ein wissenschaftliches Routineverfahren zur Erfassung eines oder mehrerer empirisch abgrenzbarer psychologischer Merkmale mit dem Ziel einer möglichst genauen quantitativen Aussage über den Grad der individuellen Merkmalsausprägung”.

Nach Bühner (2010) ist ein “psychometrischer Test ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale (vgl. Lienert & Raatz (1998), S.1). Das Ziel eines psychometrischen Tests besteht darin, die absolute oder relative Ausprägung einer Eigenschaft, einer Fähigkeit oder eines Zustands bei einer oder mehreren Personen zu messen oder eine qualitative Aussage zu treffen, welcher Personenklasse Personen zugeordnet werden können (vgl. Rost (2004)). Psychometrische Tests sind nach der Klassischen oder Probabilistischen Testtheorie entwickelt, sind theoretisch fundiert und genügen genau definierten Gütekriterien (Haupt- und Nebengütekriterien).”

Nach Krauth (1995) schließlich “besteht ein psychologischer Test aus einer Menge von Reizen mit den zugehörigen zugelassenen Reaktionen, also aus manifesten Variablen. Eine Skala ordnet den Reaktionsmustern der manifesten Variablen die Ausprägungen einer oder mehrerer latenter Variablen zu. Somit fungiert ein Test als Messinstrument zur Erfassung nicht direkt beobachtbarer (latenter) Variablen, deren Existenz für Personen und gelegentlich auch für Tiere postuliert wird.”

Nach diesen Definitionen ist psychologischen Tests also gemein, dass sie Verfahren zur Messung psychologischer Eigenschaften darstellen und sich dadurch auszeichnen, dass sie einer wissenschaftlichen Qualitätssicherung unterliegen.

Generell lassen sich psychologische Tests in verschiedene Kategorien unterteilen:

- *Leistungstests*. Dazu gehören Entwicklungstests, Intelligenztests, allgemeine Leistungstests, Schultests sowie spezielle Funktionsprüfungs- und Eignungstests.
- *Psychometrische Persönlichkeitstests*. Hierunter fallen klinische Tests, Persönlichkeitsstrukturtests, Einstellungstests und Interessentests.
- *Persönlichkeitsentfaltungs-Verfahren*. Diese umfassen Formdeutungsverfahren, verbal-thematische Verfahren sowie zeichnerische und gestalterische Verfahren.

Im Kontext der Klinischen Psychologie und Psychotherapie sind sicherlich klinische Tests von vorherrschendem Interesse. Beispiele für klinische Tests sind etwa

- *BDI-II* (Beck Depressions-Inventar II, Beck et al. (1996), Hautzinger et al. (2006)), ein klinischer Test zur Messung der Schwere einer Depression. Er basiert auf den diagnostischen Kriterien einer Major Depression nach dem DSM-IV und fragt typische depressive Symptome ab.
- *BSI* (Brief Symptom Inventory, Derogatis & Melisaratos (1983)), ein klinischer Test, der psychische Belastungen und Symptome erfasst. Der BSI deckt neun Symptomdimensionen ab: Somatisierung, Zwanghaftigkeit, Unsicherheit im Sozialkontakt, Depressivität, Ängstlichkeit, Aggressivität/Feindseligkeit, phobische Angst, paranoides Denken und Psychotizismus.
- *BSL-23* (Borderline-Symptomliste 23, Bohus et al. (2009)), ein klinischer Test zur Erfassung der Schwere von Borderline-spezifischen Symptomen. Er misst typische emotionale, kognitive und Verhaltensmuster bei Borderline-Persönlichkeitsstörung.

Psychologische Test bestehen dabei in der Regel aus *Items*, die als Grundbausteine des Tests fungieren. Items stellen allgemein Reize dar, auf die eine Reaktion erwartet und registriert wird. Das Standardbeispiel für ein Item ist eine Frage mit entsprechenden Antwortmöglichkeiten. Beispielsweise umfasst der BDI-II 21 Items zur Bewertung des Schweregrads einer Depression auf einer Skala von 0 bis 3. Das Item zum Thema *Traurigkeit* dabei lautet:

Suchen Sie eine Aussage heraus, die am besten beschreibt, wie Sie sich in den letzten zwei Wochen, einschließlich heute, gefühlt haben:

- (0) Ich bin nicht traurig.
- (1) Ich bin oft traurig.
- (2) Ich bin ständig traurig.
- (3) Ich bin so traurig oder unglücklich, dass ich es nicht aushalte.

49.2. Testgütekriterien

Im wissenschaftlichen Diskurs haben sich eine Reihe von Qualitätsanforderungen an psychologische Tests herausgebildet, die ein Test, um wissenschaftlich anerkannt zu sein, idealerweise erfüllt (vgl. Moosbrugger & Kelava (2012)). Grundlegend sind die Kriterien der Objektivität, Reliabilität und Validität

1. *Objektivität*. Ein Test ist objektiv, wenn seine Durchführung, Auswertung und Interpretation unabhängig von den Testleitenden ist. Die Objektivität eines Testverfahrens wird meist durch eine entsprechende Manualisierung sichergestellt.
2. *Reliabilität*. Ein Test ist reliabel, wenn er das zu messende Merkmal exakt und ohne Messfehler erfasst. Wichtige Unterarten der Reliabilität sind Retest-Reliabilität, Paralleltest-Reliabilität, Testhalbierungs-Reliabilität und interne Konsistenz. Reliabilitäten werden in der Regel durch Korrelationen erfasst. Die Klassische Testtheorie stellt einen theoretischen Überbau zur Einordnung und quantitativen Definition von Retest-, Paralleltest, und Testhalbierungsreliabilität dar und entwickelt beispielsweise mit Cronbach's α (Cronbach (1951)) Maße für die interne Konsistenz von psychologischen Tests.
3. *Validität*. Ein Test ist valide, wenn er tatsächlich das misst, was er messen soll. Unterschieden werden Inhaltsvalidität, Augenscheinvalidität, Kriteriumsvalidität, Konstruktvalidität und, im Kontext der Testentwicklung faktorielle Validität. Inhalts- und Augenscheinvalidität werden in der Regel durch Expert:inneneinschätzungen belegt, Kriteriumsvalidität durch Korrelation mit einem geeigneten externen Kriterium, z.B. einer klinischen Diagnose. Die Etablierung der Konstruktvalidität eines Tests ist eine weitreichende Forderung, die vermutlich selten erfüllt wird (vgl. Borsboom et al. (2004)). Belege für die faktorielle Validität schließlich tauchen häufig in Textmanualen auf und beziehen sich auf die Ergebnisse von Faktorenanalysen.

Neben den obigen Gütekriterien erfüllt ein wissenschaftlich anerkanntes Testverfahren idealerweise auch noch folgende sogenannten *Nebengütekriterien*:

4. *Skalierung*. Die Testwerte sollten die qualitativen Merkmalsrelationen im Sinne der Repräsentationstheorie des Messen adäquat abbilden.
5. *Normierung*. Das Bezugssystem ermöglicht die Vergleichbarkeit der Testwerte mit einer Eichstichprobe.
6. *Testökonomie*. Ein Test sollte mit möglichst geringem Zeit- und Ressourcenaufwand auswertbar sein. Die Digitalisierung der Answererhebung und automatisierte Auswertung mithilfe von Computersoftware wären ein erster Schritt in dieser Richtung.
7. *Nützlichkeit*. Der Test sollte praktische Relevanz besitzen und einen positiven Nutzen bieten.
8. *Zumutbarkeit*. Die Belastung der Testperson sollte in einem angemessenen Verhältnis zum Nutzen stehen.
9. *Unverfälschbarkeit*. Das Testergebnis sollte nicht durch bewusstes Verhalten verfälscht werden können. Dies ist bei den meisten klinischen Verfahren, die auf einer Selbsteinschätzung beruhen, sicherlich nicht der Fall.
10. *Fairness*. Der Test darf keine systematische Benachteiligung bestimmter Gruppen verursachen. Dieses Kriterium ist im klinischen Kontext vermutlich eher von untergeordneter Relevanz.

49.3. Normwerte

Primäres quantitatives Ergebnis eines psychologischen Tests oder auch eines einzelnen Items eines Tests ist zunächst einmal ein *Rohwert*. Beispielsweise sind die möglichen Rohwerte bei Beantwortung des oben aufgeführten Traurigkeitsitems des BDI-II die Zahlen 0,1,2, und 3. Der Rohwert eines Testergebnisses über mehrere Items ergibt sich häufig als die Summe der Rohwerte der einzelnen Items als sogenannter *Summenscore*. So variieren die Summenscores des BDI-II beispielsweise zwischen 0 und 63, wobei Werte zwischen 0 und 13 einer *minimalen*, Werte zwischen 14 und 19 einer *leichten*, Werte zwischen 20 und 28 einer *mäßigen* und Werte zwischen 29 und 63 einer *schweren Depressionssymptomatik* entsprechen. Die Rohwerte des BSI und des BSL-23 ergeben sich durch Summation der einzelnen Itemwerte, mit den möglichen Ausprägungen 0,1,2,3,4 und Division durch die beantwortete Itemanzahl, wodurch sich dann ein *Mittelwertscore* im Bereich 0 bis 4 ergibt.

Um ein gewisses Maß an Vergleichbarkeit zwischen den Wertausprägungen verschiedener klinischer Testverfahren zu gewährleisten und ein einzelnes Testergebnis vor dem Hintergrund der üblichen Verteilungen von Testergebnissen einzuordnen, werden Testergebnisse oft *normiert* (vgl. De Beurs et al. (2022)). Dazu ist es zunächst einmal nötig, Testergebnisse von einer hinreichend großen Stichprobe von gesunden oder klinisch auffälligen Proband:innen zu erheben. Diese Testergebnisse seien im Folgenden als Testrohwerte $y^{(i)}$ für $i = 1, \dots, n$ bezeichnet, wobei n der Stichprobenumfang ist. Die Resultate solcher groß angelegter Studien sind in der Regel in den entsprechenden Testmanualen dokumentiert. Legt man dann die Annahme zugrunde, dass die Testergebnisse in der entsprechenden Stichprobe einer Normalverteilung folgen, deren Erwartungswertparameter man durch das entsprechende Stichprobenmittel und deren Varianzparameter man durch die entsprechende Stichprobenvarianz schätzen kann, so ergeben sich, basierend auf dem Theorem zur linearen Transformation normalverteilter Zufallsvariablen eine Reihe von Möglichkeiten, die Testrohwerte zu standardisieren. Üblicherweise trifft man in diesem Zusammenhang auf *Z Scores* und *T Scores*, seltener auf *Stanines* oder *Stens*.

Um die Bedeutung dieser standardisierten Testwerte zu verstehen, erinnern wir zunächst an die Verteilung der Wahrscheinlichkeitsmasse einer normalverteilten Zufallsvariable. Bekanntlich gelten (vgl. Abbildung 49.1, nach Seashore (1955) und De Beurs et al. (2022))

$$\begin{aligned} \int_{-1\sigma}^{1\sigma} N(x; \mu, \sigma^2) dx &= \Phi(1\sigma; \mu, \sigma^2) - \Phi(-1\sigma; \mu, \sigma^2) \approx 0.68 \\ \int_{-2\sigma}^{2\sigma} N(x; \mu, \sigma^2) dx &= \Phi(2\sigma; \mu, \sigma^2) - \Phi(-2\sigma; \mu, \sigma^2) \approx 0.95 \\ \int_{-3\sigma}^{3\sigma} N(x; \mu, \sigma^2) dx &= \Phi(3\sigma; \mu, \sigma^2) - \Phi(-3\sigma; \mu, \sigma^2) \approx 0.99 \end{aligned} \quad (49.1)$$

Innerhalb einer Standardabweichung um den Erwartungswertparameter liegen bei einer normalverteilten Zufallsvariable also 68% der Wahrscheinlichkeitsmasse, innerhalb von zwei Standardabweichungen 95% und innerhalb von drei Standardabweichungen mit 99% fast die gesamte Wahrscheinlichkeitsmasse. Diese Verteilung der Wahrscheinlichkeitsmasse gilt offensichtlich unabhängig von den konkreten Werten von μ und σ^2 und somit für alle normalverteilten Zufallsvariablen. Weiterhin erinnern wir daran, dass das Theorem zur

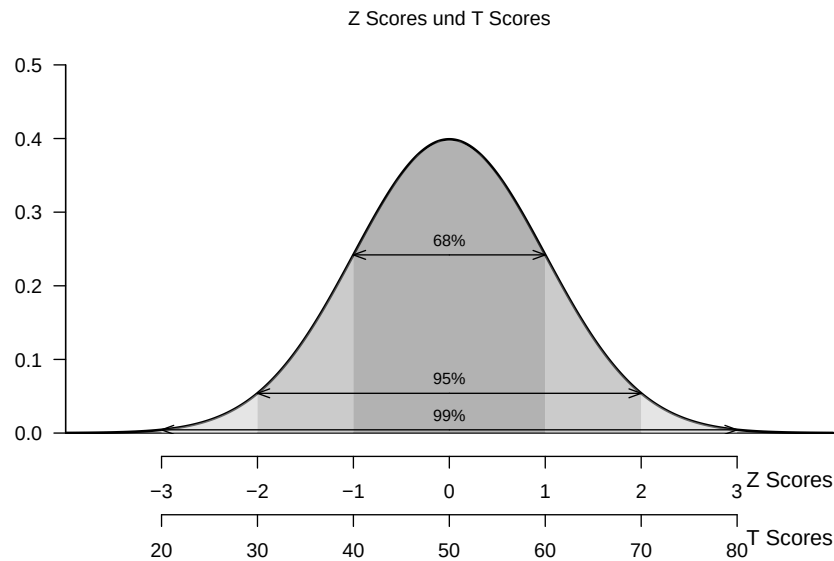


Abbildung 49.1. Z Scores und T Scores.

linearen Transformation einer normalverteilten Zufallsvariable besagt, dass

$$y \sim N(\mu, \sigma^2) \text{ und } z := ay + b \Rightarrow z \sim N(a\mu + b, a^2\sigma^2). \quad (49.2)$$

Seien also m und s^2 die Stichprobenschätzer für den Erwartungswertparameter μ und den Varianzparameter σ^2 normalverteilter Testrohwerte. Dann gilt für

$$z := \frac{y - m}{s} = \frac{1}{s}y - \frac{1}{s}m \approx \frac{1}{\sigma}y - \frac{1}{\sigma}\mu \quad (49.3)$$

näherungsweise

$$z \sim N\left(\frac{1}{\sigma}\mu - \frac{1}{\sigma}\mu, \frac{1}{s^2}s^2\right) = N(0, 1). \quad (49.4)$$

Berechnet man zu jedem Testrohwert $x^{(i)}$ also den sogenannten *Z Score*

$$z^{(i)} := \frac{x^{(i)} - m}{s}, \quad (49.5)$$

so sind die so generierten Werte, unabhängig von den Stichproben-spezifischen Werten von m und s immer näherungsweise standardnormalverteilt und damit über verschiedene Rohwertebereiche verschiedener Tests vergleichbar (vgl. Abbildung 49.2).

Nun hat eine standardnormalverteilte Zufallsvariable die Eigenschaft, mit höchster Wahrscheinlichkeit einen Wert zwischen -3 und 3 anzunehmen. Dies mag manchmal unpraktisch erscheinen, da in diesem Fall ein einzelnes Testergebnis, wenn es unter dem Stichprobenmittelwert liegt, negativ ausfallen kann. Weiterhin erfordert die Angabe eines Ergebnisses mit hinreichender Genauigkeit die Angabe von Nachkommastellen. McCall (1922) hat deshalb vorgeschlagen, eine weitere lineare Transformation der berechneten Z Scores durchzuführen, die die Z Scores in einen zugänglicheren Wertebereich transformiert. Die so transformierten Z Scores bezeichnet McCall (1922) zu Ehren von Edward Lee Thorndike,

Lewis Terman und Louis Leon Thurstone als *T Scores*. Die Bezeichnung *T Scores* geht also keinesfalls auf die *T*-Statistiken der Frequentistischen Inferenz und ihre entsprechenden Verteilungen zurück. Die von McCall (1922) vorgeschlagene Transformation ist durch

$$t^{(i)} := 10z^{(i)} + 50 \quad (49.6)$$

gegeben. Mit der näherungsweisen Standardnormalverteilung von z ergibt sich entsprechend

$$t \sim N(0 + 50, 10^2 \cdot 1) = N(50, 100). \quad (49.7)$$

T Scores haben näherungsweise also immer einen Erwartungswert von 50, eine Varianz von 100 und eine Standardabweichung von 10. Damit liegen dann im *T Score* Wertebereich von 40 bis 60 etwa 68% der Stichprobenergebnisse, im Wertebereich von 30 bis 70 etwa 95% der Stichprobenergebnisse und im Wertebereich von 20 bis 80 etwa 99% der Stichprobenergebnisse (vgl. Abbildung 49.1).

Die seltener anzutreffenden *Stanine* (standard nine) und *Sten* (standard ten) Werte folgen in ihrer Berechnung der gleichen Logik wie die *T Scores*, zielen aber auf den Wertebereich 1 bis 9 bzw. 1 bis 10. Sie ergeben sich aus *Z Scores* durch die Transformation

$$s^{(i)} := 2z^{(i)} + 5 \quad (49.8)$$

und anschließende Rundung auf Werte von 1 bis 9 bzw. 1 bis 10.

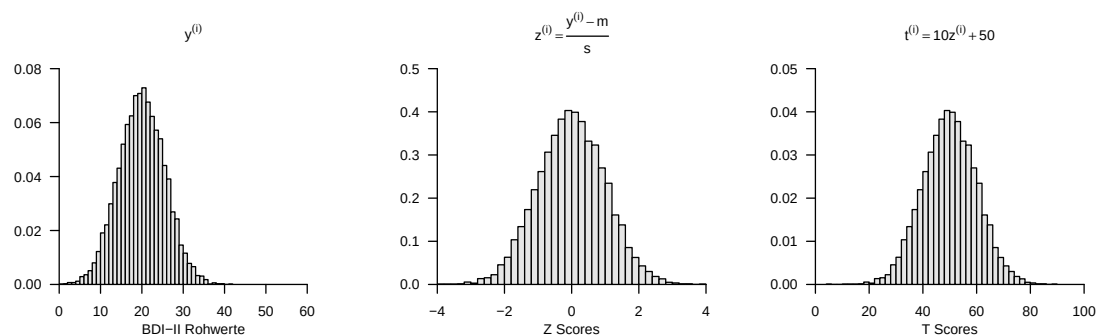


Abbildung 49.2. Simulierte Stichproben-BDI-II Rohwerte, Z Scores und T Scores

49.4. Testtheorien

Generell kann man mit der *Klassischen Testtheorie*, der *Faktorenanalyse* und der *Item-Response-Theorie* drei Theorien unterscheiden, die der Analyse von Testdaten zugrunde liegen. Allen drei Theorien ist gemein, dass es sich bei ihnen um probabilistische Modelle von beobachtbaren Item- und Testwerten handelt. Weiterhin ist ihnen gemein, dass sie neben der Modellierung von beobachtbaren Testwerten durch Zufallsvariablen auch latente (nicht direkt beobachtbare) Zufallsvariablen zur Erklärung beobachtbarer Daten heranziehen. In der wissenschaftlichen Betrachtung klinisch-diagnostischer Verfahren spielen insbesondere die Klassische Testtheorie bezüglich der Reliabilität und der internen Konsistenz von Testverfahren sowie die Faktorenanalyse bezüglich der faktoriellen Validität von Testverfahren zentrale Rollen. Die Item-Response-Theorie spielt derzeit in Bezug

auf klinisch-diagnostische Verfahren eher eine geringe Rolle (vgl. Reise & Waller (2009)). Wir wollen diese Hauptströmungen der Testtheorie im Folgenden kurz skizzieren.

Klassische Testtheorie

Die Klassische Testtheorie bildet die Grundlage für die Evaluation des Testgütekriteriums der Reliabilität. Die Klassische Testtheorie besteht im Kern aus einem probabilistischen Modell für Item- oder Testergebnisse eines oder mehrerer Individuen für einen oder mehrere Tests. Zentrale Aspekte der Klassischen Testtheorie gehen auf die Arbeiten von Charles Spearman Anfang des 20. Jahrhunderts zurück und wurden unter anderem durch Gulliksen (1950) und Lord & Novick (1968) weiter etabliert. Grundlegendes Konzept der Klassischen Testtheorie ist ein Messfehlermodell der Form

$$\text{Beobachteter Wert} = \text{Wahrer Wert} + \text{Messfehler} \quad (49.9)$$

wie in der Form “ $y = \mu + \varepsilon$ ” aus der Frequentistischen Inferenztheorie bekannt. Allerdings treten dabei im Rahmen der Klassischen Testtheorie einige Besonderheiten auf. Zum einen basieren die Modelle der Klassischen Testtheorie in der Regel *nicht* auf Normalverteilungsannahmen, was dem diskreten Charakter der beobachteten Testergebnisse geschuldet sein mag, sondern beschränken sich auf die Modellierung durch allgemeine Zufallsvariablen und die Analyse ihrer Erwartungswerte, Varianzen und Kovarianzen. Zum anderen werden “wahre Werte” im Gegensatz zur Frequentistischen Inferenz nicht als feste Werte, sondern auch als Realisierungen von Zufallsvariablen modelliert. Schließlich tritt insbesondere in der Formulierung nach Lord & Novick (1968) noch der Versuch zutage, “wahre Werte” möglichst nicht naiv-realistisch bei der Analyse von Testergebnissen zu unterstellen, sondern ihnen eine frequentistisch-propensitäre Interpretation zu geben, um sie aus naturwissenschaftlicher Sicht weniger angreifbar zu machen. Die Klassische Testtheorie liefert unter anderem die Grundlage für die in der Anwendung häufig berichteten Korrelationen zur Retestreliabilität und Cronbach’s α als Maß für die interne Konsistenz eines Tests. Wir orientieren uns in der folgenden Darstellung der Klassischen Testtheorie überwiegend an Krauth (1995).

Faktorenanalyse

Grundlage der Faktorenanalyse ist ein probabilistisches Modell der Kovariationsbeziehungen von Testitems und bildet in der Anwendung die Grundlage für das Testgütekriterium der *faktoriellen Validität*. Bedeutende klassische Beiträge zur Faktorenanalyse stammen von Spearman (1904), Hotelling (1933) und Lawley (1940). Wie die Klassische Testtheorie modelliert auch die Faktorenanalyse beobachtete Testergebnisse, hier meist auf der Ebene von Items, durch ein latentes Variablenmodell

$$\text{Beobachteter Wert} = \text{Faktorladungen} \cdot \text{Wahre Faktorenwerte} + \text{Messfehler}. \quad (49.10)$$

Die “wahren Werte” der Klassischen Testtheorie und die “Wahren Faktorwerte” der Faktorenanalyse sind von ähnlichem Charakter, wobei in der Faktorenanalyse typischerweise mehr als ein einzelner wahrer Wert angenommen wird. Unter anderem aufgrund der Tatsache, dass mit dem beobachteten Wert in der Faktorenanalyse ein einzelnes Datum durch das Zusammenspiel dreier nicht beobachtbarer Werte (Faktorladung, Faktorwert, Messfehler) erklärt wird, ist die eindeutige Schätzung von Faktorladungen und Faktorwerten nicht ohne weiteres möglich. Diese Tatsache hat klassischerweise dazu geführt, dass aus einer Vielzahl möglicher Lösungen eines Faktorenanalysemodells ein spezielles mithilfe

heuristischer Nebenkriterien ausgewählt wird, ein Prozess, der sich in den *Rotationsverfahren* der *explorativen Faktorenanalyse* niederschlägt. Allerdings kann das Faktorenanalysenmodell durch geeignete Annahmen, wie zum Beispiel die Normalverteilung von wahren Faktorwerten und Messfehlern und die Nicht-Effizienz ausgewählter Faktorenwerte zumindest in Ansätzen identifizierbarer gestaltet werden und zu einem validen Inferenzmodell ausgebaut werden. Dies ist das Thema der *konfirmatorischen Faktorenanalyse*. Mit den *Strukturgleichungsmodellen* (vgl. Bollen (1989)) schließlich wurde die Faktorenanalyse im Computerzeitalter weiter verallgemeinert und zu einem äußerst flexiblen Modellierungsansatz ausgebaut. Wir orientieren uns in der Darstellung der Faktorenanalyse überwiegend an Rencher & Christensen (2012).

Item-Response-Theorie

Die Item-Response-Theorie (IRT) modelliert die Wahrscheinlichkeit einer bestimmten Testantwort als Funktion der Fähigkeit oder des Zustands der Testperson. Wichtige Beiträge zur IRT stammen von Rasch (1960), Novick (1966) und Lord (1980). Die IRT verwendet im Ansatz logistische Regressionsmodelle und findet vor allem bei Leistungstest Anwendung. Sie bezieht sich dabei explizit auf die Messtheorie nach Stevens (1946), hat jedoch im klinischen Kontext eine geringere bis keine Bedeutung (vgl. Reise & Waller (2009)).

50. Klassische Testtheorie

50.1. Modelle multipler Testmessungen

Die Klassische Testtheorie nimmt ihren Ausgang von Modellen für beobachtete Werte von mehreren Testmessungen von mehreren Personen. Um für die folgende Modellformulierung ein konkretes Datenbeispiel vor Augen zu haben, aufgrund dessen dann Parameter des Modells der Klassischen Testtheorie geschätzt werden können, gehen wir davon aus, dass wir einen Datensatz wie in untenstehender Tabelle gezeigt vorliegen haben. Dieser Datensatz soll die T-Scores von $n = 20$ Personen abbilden, für die jeweils $m = 10$ Tests zur Messung der Depressionssymptomatik durchgeführt wurden. Speziell stellen wir uns vor, dass für jede Person der T-Score des BDI-II, des PHQ-9, der HDRS, der CES-D, und der MADRS zu jeweils zwei Zeitpunkten (T0 und T1) bestimmt wurde. Im Folgenden werden wir Personen mit dem Index i bezeichnen, haben hier also $i = 1, \dots, 20$ und Testmessungen mit dem Index j bezeichnen, haben hier also $j = 1, \dots, 10$.

Tabelle 50.1. T-Score Beispieldatensatz für $n = 20$ Personen und $m = 10$ Testmessungen

Person	BDI-II T0	BDI-II T1	PHQ-9 T0	PHQ-9 T1	HDRS T0	HDRS T1	CES-D T0	CES-D T1	MADRS T0	MADRS T1
1	53.2	56.8	46.9	51.8	61.4	58.6	46.3	51.3	55.6	52.6
2	58.2	55.4	58.3	59.0	68.6	43.5	53.8	63.8	64.3	57.6
3	54.6	58.8	51.3	64.2	51.1	50.3	61.0	52.8	58.7	55.7
4	43.0	45.6	43.9	55.6	47.1	50.6	40.2	49.0	53.0	41.5
5	54.0	53.0	58.5	51.8	47.5	45.9	47.2	50.0	48.2	59.7
6	59.3	57.5	50.8	56.9	69.8	55.0	62.9	63.8	57.0	60.5
7	37.8	20.7	24.7	33.2	22.3	20.2	23.1	26.3	27.9	30.3
8	33.8	35.7	42.5	53.3	38.5	36.6	40.0	33.7	48.6	32.4
9	51.3	48.2	48.8	50.0	43.5	53.2	43.6	54.4	48.5	51.9
10	60.1	58.2	61.6	61.4	65.5	61.2	60.8	60.6	68.8	59.1
11	44.4	47.9	43.8	43.6	43.2	44.2	41.9	52.8	42.9	42.0
12	62.1	55.3	58.0	57.0	62.5	49.8	54.1	48.5	53.5	50.0
13	52.3	48.3	51.7	48.1	57.7	63.0	58.4	60.8	61.1	51.9
14	53.9	54.4	50.9	58.0	54.2	55.6	44.5	58.6	51.8	49.5
15	37.5	41.0	40.6	38.3	42.1	34.0	35.7	43.8	36.9	43.1
16	62.4	51.8	58.7	41.6	52.0	47.1	51.2	54.6	50.8	63.6
17	61.8	65.5	57.0	60.6	72.9	70.0	56.6	64.3	59.3	64.6
18	45.1	39.9	44.5	42.4	36.8	36.9	43.4	44.1	39.1	39.6
19	53.7	55.5	49.4	47.6	47.5	42.5	49.4	51.7	49.2	56.8
20	44.6	39.9	34.7	41.5	42.1	45.6	32.0	42.8	46.5	47.2

50.1.1. Das Modell multipler Testmessungen

Die Klassische Testtheorie in der Formalisierung nach Novick (1966) und Lord & Novick (1968) nimmt ihren Ausgang von der Definition des *wahren Werts* einer Testmessung einer Person, die die Definitionen des *beobachteten Werts* und des *Messfehlers* impliziert. Die Definition nutzt das Konzept eines *bedingten Erwartungswerts* (vgl. ?@sec-theoretische-ergaenzungen) und hat folgende Form.

Definition 50.1 (Wahrer Wert, beobachteter Wert und Messfehler). Für $i = 1, \dots, n$ und $j = 1, \dots, m$ ist der *wahre Wert* (*true score*) t_{ij} einer Person i für eine Testmessung j

definiert als der bedingte Erwartungswert des *beobachteten Werts* (*observed score*) der Person für diese Testmessung

$$t_{ij} := \mathbb{E}(y_{ij} | \tau_{ij} = t_{ij}). \quad (50.1)$$

Der *Messfehler* (*error score*) einer Person ist definiert als die Zufallsvariable

$$\varepsilon_{ij} := y_{ij} - t_{ij}. \quad (50.2)$$

•

Mit dem Begriff einer *Testmessung* ist hier etwas unspezifisch und je nach Anwendung entweder eine Messung mithilfe eines einzelnen Items oder mithilfe der Summe mehrerer Items gemeint. An späterer Stelle werden wir die Unterscheidung von *Itemscores* und *Testsummenscores* explizit machen. Letztere induzieren im Rahmen der Klassischen Testtheorie den Begriff der *m-Komponententestmodelle* (vgl. Kapitel 50.3). Man beachte, dass in Definition 50.1 der Definition bedingter Erwartungswerte gemäß τ_{ij} , y_{ij} und ε_{ij} Zufallsvariablen sind und $t_{ij} \in \mathbb{R}$ eine Konstante ist.

Die Definition des wahren Werts t_{ij} in Definition 50.1 ist etwas speziell (um nicht zu sagen zirkulär bis tautologisch), da t_{ij} mithilfe von t_{ij} definiert wird. Die zentrale Motivation von Novick (1966) und Lord & Novick (1968) den wahren Wert t_{ij} als bedingten Erwartungswert, anstelle von zum Beispiel einfach einer Realisierung der Zufallsvariable τ_{ij} zu definieren, war es, sich einer Diskussion zur metaphysischen Bedeutung eines wahren Werts zu entziehen. Hätten Novick (1966) und Lord & Novick (1968) beispielsweise den wahren Wert als Realisierung einer latenten Variable definiert, die einer Person für eine bestimmte Testmessung eigen sein soll, so hätte dies in der zeitgenössischen Diskussion eindeutig den Charakter einer angreifbaren metaphysischen Aussage. Stattdessen versuchen Novick (1966) und Lord & Novick (1968) in ihrer Definition möglichst operationalistisch vorzugehen und den wahren Wert einer Person für eine Testmessung als “Durchschnitt der beobachteten Werte” einer Person über “wiederholte Testmessungen unter identischen Bedingungen” darzustellen. Zur Bedeutung des wahren Werts zitieren Lord & Novick (1968), S. 29 - 30 Lazarsfeld (1959):

“Angenommen, wir fragen eine Person, Herrn Brown, wiederholt, ob er die Vereinten Nationen befürwortet; nehmen wir weiter an, dass wir ihm nach jeder Frage „das Gehirn waschen“ und ihm dann dieselbe Frage erneut stellen. Da Herr Brown unsicher ist, wie er zu den Vereinten Nationen steht, wird er manchmal eine befürwortende und manchmal eine ablehnende Antwort geben. Nachdem wir dieses Verfahren viele Male durchgeführt haben, berechnen wir anschließend den Anteil der Male, in denen Herr Brown die Vereinten Nationen befürwortet hat.”

Diese Proportion wollen Novick (1966) und Lord & Novick (1968) dann als wahren Wert verstehen (vgl. auch Borsboom et al. (2004)). Natürlich ist dieser Versuch einer anti-metaphysischen Grundhaltung nicht durchhaltbar: zum einen handelt es sich bei der “wiederholten Testmessung unter identischen Bedingungen” um ein idealisiertes, in der Realität nicht durchführbares, Gedankenexperiment. Zum anderen definieren Novick (1966) und Lord & Novick (1968) den wahren Wert auch gerade nicht als (endlichen) Mittelwert einer Messreihe, sondern als idealisierten Erwartungswert einer Zufallsvariable. Als wirklich operationalistische Definition überzeugt Definition 50.1 also nicht, verkompliziert die Entwicklung einer Standardmessfehlertheorie, wie sie beispielsweise die klassische Formalisierung des Allgemeinen Linearen Modells darstellt, für die Analyse von Tests und Fragebögen

aber beträchtlich. Dies mag einer der Gründe sein, warum die Klassische Testtheorie bis heute lediglich in der Psychologie und wenig darüberhinaus von Bedeutung ist.

Ähnlich gelagert ist die Bezeichnung der bedingten Verteilung des beobachteten Werts $\mathbb{P}(y_{ij}|\tau_{ij} = t_{ij})$ als *Propensitätsverteilung* durch Lord & Novick (1968), welche die intraindividuelle Variabilität des beobachteten Werts bei festem wahren Wert modelliert. Dabei klingt an, dass Lord & Novick (1968) eine Propensitätsinterpretation von Wahrscheinlichkeiten als kausal bedingte “Verwirklichungstendenzen”, die, im Gegensatz zur Frequentistischen Interpretation auch im Einzelfall Sinn ergeben, implizieren. Allerdings unterstellen Propensitätsverteilungen kausale Prozesse, die in der Regel nicht spezifiziert und damit auch nicht beobachtbar sind, und führen damit letztlich auch wieder auf metaphysische Aussagen (vgl. auch Borsboom et al. (2004) und Borsboom (2009)).

In unserer Darstellung wollen wir dem modell-basierten realistischen Ansatz folgen und wählen mit Definition 50.2 deshalb eine Formulierung des Modells multipler Testmessungen der Klassischen Testtheorie, die mit Definition 50.1 und damit natürlich auch den theoretischen Ergebnissen der Klassischen Testtheorie kongruent ist, aber nicht versucht, ihren Modellcharakter zu verschleiern. Insbesondere betont unserer Ansatz dabei das Gesamtziel der Modellierung einer Menge von m Testmessungen von n Personen als eines Datensatzes von nm Datenpunkten. Wir definieren das *Modell multipler Testmessungen* daher wie folgt.

Definition 50.2 (Modell multipler Testmessungen). Für $i = 1, \dots, n$ und $j = 1, \dots, m$ seien τ_{ij} eine Zufallsvariable, die den *wahren Wert* der i ten Person in der j ten Testmessung modelliert und y_{ij} eine Zufallsvariable, die den *beobachteten Wert* der i ten Person in der j ten Testmessung modelliert. Dann nennen wir die gemeinsame Verteilung der τ_{ij} und y_{ij} mit der Faktorisierungseigenschaft

$$\mathbb{P}(\tau_{11}, y_{11}, \dots, \tau_{nm}, y_{nm}) := \prod_{i=1}^n \mathbb{P}(\tau_{i1}, \dots, \tau_{im}) \prod_{j=1}^m \mathbb{P}(y_{ij}|\tau_{ij}) \quad (50.3)$$

das *Modell multipler Testmessungen*, wenn gilt, dass

$$\mathbb{P}(\tau_{11}, \dots, \tau_{1m}) \prod_{j=1}^m \mathbb{P}(y_{1j}|\tau_{1j}) = \dots = \mathbb{P}(\tau_{n1}, \dots, \tau_{nm}) \prod_{j=1}^m \mathbb{P}(y_{nj}|\tau_{nj}). \quad (50.4)$$

•

Die Definition des Modells multipler Testmessungen bildet einige grundlegende Annahmen zur Unabhängigkeit und Identität von Verteilungen in der Klassischen Testtheorie ab. Zunächst einmal wird angenommen, dass die gemeinsame Verteilung der τ_{ij} und y_{ij} über $i = 1, \dots, n$ faktorisiert, dass die Verteilungen der wahren Werte und beobachteten Werte also über Personen unabhängig sind. Wissen um wahre oder beobachtete Werte einer Person ändert die angenommenen Verteilungen der wahren und beobachteten Werte anderer Personen also nicht. Im Gegensatz dazu faktorisiert für jedes $i = 1, \dots, n$ die gemeinsame Verteilung der $\tau_{i1}, \dots, \tau_{im}$ über Testmessungen $j = 1, \dots, m$ nicht notwendigerweise. Die wahren Werte einer Person können also abhängig sein und damit Wissen um die Ausprägung einer Testmessung bei einer Person die Verteilung des wahren Werts bei anderen Testmessungen derselben Person informieren. In der Klassischen Testtheorie werden verschiedene Arten dieser Form von Abhängigkeiten unterschieden und, wie wir später sehen

werden, beispielsweise als *Parallelität*, τ -Äquivalenz oder *essentielle τ -Äquivalenz* bezeichnet. Weiterhin wird für jede Person $i = 1, \dots, n$ angenommen, dass die beobachteten Werte y_{ij} für $j = 1, \dots, m$ gegeben τ_{ij} bedingt unabhängig sind. Dies impliziert einerseits, dass für eine Person der wahre Wert in Testmessung $k \neq j$ den beobachteten Wert in Testmessung j nicht beeinflusst und andererseits, dass der beobachtete Wert in Testmessung $k \neq j$ den beobachteten Wert in Testmessung j nicht beeinflusst.

Schließlich wird angenommen, dass die Marginalverteilungen

$$\mathbb{P}(\tau_{i1}, y_{i1}, \dots, \tau_{im}, y_{im}) = \mathbb{P}(\tau_{i1}, \dots, \tau_{im}) \prod_{j=1}^m \mathbb{P}(y_{ij} | \tau_{ij}) \quad (50.5)$$

über Personen $i = 1, \dots, n$ identisch sind. Man mag sich die Realisierungen von wahren Werten und beobachteten Werten also als unabhängige und identische Realisierungen aus einer “Populationsverteilung”

$$\mathbb{P}(\tau_{\bullet 1}, y_{\bullet 1}, \dots, \tau_{\bullet m}, y_{\bullet m}) = \mathbb{P}(\tau_{\bullet 1}, \dots, \tau_{\bullet m}) \prod_{j=1}^m \mathbb{P}(y_{\bullet j} | \tau_{\bullet j}) \quad (50.6)$$

vorstellen, wobei das Subskript $\bullet$ die Unspezifität dieser Verteilung bezüglich einer Person symbolisieren soll.

Betrachtet man den Spezialfall einer einzelnen Testmessung bei n Personen, so ergibt sich eine vereinfachte Form von Definition 50.2, auf die man häufig in der psychologischen Literatur trifft. Nach Definition 50.2 gilt für eine Testmessung, also $m = 1$,

$$\mathbb{P}(\tau_{11}, y_{11}, \dots, \tau_{n1}, y_{n1}) := \prod_{i=1}^n \mathbb{P}(\tau_{i1}) \mathbb{P}(y_{i1} | \tau_{i1}), \quad (50.7)$$

wobei nach Annahme von Definition 50.2 die gemeinsamen Marginalverteilungen $\mathbb{P}(\tau_{i1}, y_{i1})$ über Personen $i = 1, \dots, n$ identisch sind. Wie oben kann man sich die wahren und beobachteten Werte für eine Testmessung also als unabhängige Realisierungen der “Populationsverteilung”

$$\mathbb{P}(\tau_{\bullet 1}) \mathbb{P}(y_{\bullet 1} | \tau_{\bullet 1}) \quad (50.8)$$

vorstellen. Verzichtet man nun noch auf die Subskripte $\bullet 1$, gelangt man zu folgender vereinfachter Definition des Modells der Klassischen Testtheorie.

Definition 50.3 (Vereinfachtes Modell der Klassischen Testtheorie). τ sei eine Zufallsvariable, die die Verteilung der *wahren Werte* einer Testmessung in einer Population von Individuen beschreibt und y sei eine Zufallsvariable, die die Verteilung der *beobachteten Werte* dieser Testmessung beschreibt. Dann heißt die gemeinsame Verteilung von τ und y

$$\mathbb{P}(\tau, y) = \mathbb{P}(\tau) \mathbb{P}(y | \tau) \quad (50.9)$$

das *Vereinfachte Modell der Klassischen Testtheorie*.

•

Definition 50.3 hat gegenüber Definition 50.2 den Vorteil, dass weniger Zufallsvariablen und Indizes auftreten und auf die Redundanz der unterschiedlichen Bezeichnung vieler

gleicher Verteilungen verzichtet werden kann. Im Sinne Frequentistischer Produktmodelle mag man bezüglich der Daten von n Personen hier auch

$$(\tau_1, y_1), \dots, (\tau_n, y_n) \sim \mathbb{P}(\tau, y) \quad (50.10)$$

schreiben, wobei nur die $y_1, \dots, y_n$ beobachtete, die $\tau_1, \dots, \tau_n$ dagegen latente Zufallsvariablen sind. Weiterhin gilt, dass man viele wichtige Eigenschaften des Modells der Klassischen Testtheorie schon basierend auf den Eigenschaften von $\mathbb{P}(\tau, y)$ begründen kann, wie wir unten sehen werden. Generell ist das vereinfachte Modell der Klassischen Testtheorie einfacher zu handhaben als das Modell multipler Testmessungen. Problematisch wird es allerdings, sobald mehrere Testmessungen ins Spiel kommen, beispielsweise bei Abhängigkeitsbetrachtungen zwischen zwei Tests oder zwei Items eines Tests. Dies ist allerdings bei den meisten Aussagen der Klassischen Testtheorie der Fall. Außerdem gilt auch, dass Schätzer der Modellparameter immer auf allen Werten der beobachteten Zufallsvariablen $y_{1j}, \dots, y_{nj}$ beruhen, diese in Definition 50.3 aber überhaupt nicht auftreten. Definition 50.3 ist für theoretische Betrachtungen also oft einfacher zu handhaben als Definition 50.2, der Anwendung der Klassischen Testtheorie in der Testdatenanalyse liegt im Allgemeinen aber Definition 50.2 zugrunde. Wir werden je nach Bedarf zwischen beiden Modellformulierungen hin und her wechseln, halten aber fest, dass die Modellformulierung im Sinne von Definition 50.2 unser Standardfall ist.

Eigenschaften des Modells multipler Testmessungen

Eigenschaften bezüglich einer Testmessung

Das Modell multipler Testmessungen nach Definition 50.2 hat zunächst eine Reihe von Eigenschaften bezüglich einer und damit jeder Testmessung, die für die Anwendung der Klassischen Testtheorie grundlegend sind. Wir fassen fünf dieser Eigenschaften in folgendem Theorem zusammen.

Theorem 50.1 (Eigenschaften bezüglich einer Testmessung). *Gegeben sei das Modell multipler Testmessungen. Dann gelten für alle $i = 1, \dots, n$ und alle $j = 1, \dots, m$*

- (1) $\mathbb{E}(\varepsilon_{ij} | \tau_{ij} = t_{ij}) = 0$
- (2) $\mathbb{E}(\varepsilon_{ij}) = 0$
- (3) $\mathbb{C}(\tau_{ij}, \varepsilon_{ij}) = 0$
- (4) $\mathbb{V}(y_{ij}) = \mathbb{V}(\tau_{ij}) + \mathbb{V}(\varepsilon_{ij})$
- (5) $\mathbb{C}(y_{ij}, \tau_{ij}) = \mathbb{V}(\tau_{ij})$

◦

Beweis. Zur Vereinfachung der Notation im Sinne des einfachen Modells der klassischen Testtheorie nach Definition 50.3 setzen wir zunächst für alle $i = 1, \dots, n$ und $j = 1, \dots, m$

$$y := y_{ij} \text{ und } \tau := \tau_{ij} \text{ mit Ergebnisräumen } Y := Y_{ij} \text{ und } T := T_{ij}. \quad (50.11)$$

Weiterhin bezeichnen wir die Werte von y mit $\tilde{y}$ und die Werte von τ mit t . Schließlich betrachten wir nur den diskreten Fall, setzen also die Existenz einer WMF $p : Y \times T \rightarrow [0, 1]$ der Form

$$p(t, \tilde{y}) = p(\tilde{y}|t)p(t) \quad (50.12)$$

voraus. Der kontinuierliche Fall oder gemischt diskret-kontinuierliche Fall folgt dann jeweils analog.

(1) Es gilt

$$\begin{aligned}
 \mathbb{E}(\varepsilon|\tau = t) &:= \mathbb{E}(y - \tau|\tau = t) \\
 &= \sum_{\tilde{y} \in Y} (\tilde{y} - t) p(\tilde{y}|t) \\
 &= \sum_{\tilde{y} \in Y} \tilde{y} p(\tilde{y}|t) - \sum_{\tilde{y} \in Y} t p(\tilde{y}|t) \\
 &= \mathbb{E}(y|\tau = t) - t \sum_{\tilde{y} \in Y} p(\tilde{y}|t) \\
 &= t - t \cdot 1 \\
 &= 0.
 \end{aligned} \tag{50.13}$$

(2) Es gilt

$$\begin{aligned}
 \mathbb{E}(\varepsilon) &:= \mathbb{E}(y - \tau) \\
 &= \sum_{t \in T} \sum_{\tilde{y} \in Y} (\tilde{y} - t) p(t, \tilde{y}) \\
 &= \sum_{t \in T} \sum_{\tilde{y} \in Y} (\tilde{y} - t) p(\tilde{y}|t) p(t) \\
 &= \sum_{t \in T} \sum_{\tilde{y} \in Y} \tilde{y} p(\tilde{y}|t) p(t) - \sum_{t \in T} \sum_{\tilde{y} \in Y} t p(\tilde{y}|t) p(t) \\
 &= \sum_{t \in T} \sum_{\tilde{y} \in Y} p(t) (\tilde{y} p(\tilde{y}|t) - t p(\tilde{y}|t)) \\
 &= \sum_{t \in T} p(t) \left(\sum_{\tilde{y} \in Y} \tilde{y} p(\tilde{y}|t) - t \sum_{\tilde{y} \in Y} p(\tilde{y}|t) \right) \\
 &= \sum_{t \in T} p(t) (t - t \cdot 1) \\
 &= \sum_{t \in T} p(t) \cdot 0 \\
 &= 0.
 \end{aligned} \tag{50.14}$$

(3) Es gilt

$$\begin{aligned}
 \mathbb{C}(\tau, \varepsilon) &= \mathbb{E}((\tau - \mathbb{E}(\tau))(\varepsilon - \mathbb{E}(\varepsilon))) \\
 &= \mathbb{E}((\tau - \mathbb{E}(\tau))\varepsilon) \\
 &= \mathbb{E}((\tau - \mathbb{E}(\tau))(y - \tau)) \\
 &= \sum_{t \in T} \sum_{\tilde{y} \in Y} ((t - \mathbb{E}(\tau))(\tilde{y} - t)) p(t, \tilde{y}) \\
 &= \sum_{t \in T} \sum_{\tilde{y} \in Y} (t - \mathbb{E}(\tau))(\tilde{y} - t) p(\tilde{y}|t) p(t) \\
 &= \sum_{t \in T} p(t)(t - \mathbb{E}(\tau)) \sum_{\tilde{y} \in Y} (\tilde{y} - t) p(\tilde{y}|t) \\
 &= \sum_{t \in T} p(t)(t - \mathbb{E}(\tau)) \sum_{\tilde{y} \in Y} (\tilde{y} p(\tilde{y}|t) - t p(\tilde{y}|t)) \\
 &= \sum_{t \in T} p(t)(t - \mathbb{E}(\tau)) \left(\sum_{\tilde{y} \in Y} \tilde{y} p(\tilde{y}|t) - t \sum_{\tilde{y} \in Y} p(\tilde{y}|t) \right) \\
 &= \sum_{t \in T} p(t)(t - \mathbb{E}(\tau)) (t - t \cdot 1) \\
 &= \sum_{t \in T} p(t)(t - \mathbb{E}(\tau)) \cdot 0 \\
 &= 0.
 \end{aligned} \tag{50.15}$$

(4) Mit dem Theorem zu Varianzen von Summen und Differenzen von Zufallsvariablen (Theorem 25.7) sowie Aussage (3) des vorliegenden Theorems gilt

$$\mathbb{V}(y) = \mathbb{V}(\tau + \varepsilon) = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon) + 2\mathbb{C}(\tau, \varepsilon) = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon) + 2 \cdot 0 = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon) \tag{50.16}$$

(5) Mit dem Kovarianzverschiebungssatz (Theorem 25.2), der Linearkombinationseigenschaft des Erwartungswerts (Theorem 23.5), Aussage (2) des vorliegenden Theorems und der Tatsache, dass mit Aussage (4) des vorliegenden Theorems außerdem folgt, dass

$$\mathbb{E}(\varepsilon\tau) = \mathbb{E}(\tau\varepsilon) = \mathbb{C}(\tau, \varepsilon) + \mathbb{E}(\tau)\mathbb{E}(\varepsilon) = 0 + \mathbb{E}(\tau) \cdot 0 = 0, \quad (50.17)$$

gilt

$$\begin{aligned} \mathbb{C}(y, \tau) &= \mathbb{E}(y\tau) - \mathbb{E}(y)\mathbb{E}(\tau) \\ &= \mathbb{E}((\tau + \varepsilon)\tau) - \mathbb{E}(\tau + \varepsilon)\mathbb{E}(\tau) \\ &= \mathbb{E}(\tau^2 + \varepsilon\tau) - (\mathbb{E}(\tau) + \mathbb{E}(\varepsilon))\mathbb{E}(\tau) \\ &= \mathbb{E}(\tau^2) + \mathbb{E}(\varepsilon\tau) - \mathbb{E}(\tau)^2 - \mathbb{E}(\varepsilon)\mathbb{E}(\tau) \\ &= \mathbb{E}(\tau^2) + \mathbb{E}(\varepsilon\tau) - \mathbb{E}(\tau)^2 - 0 \cdot \mathbb{E}(\tau) \\ &= \mathbb{E}(\tau^2) + 0 - \mathbb{E}(\tau)^2 - 0 \cdot \mathbb{E}(\tau) \\ &= \mathbb{E}(\tau^2) - \mathbb{E}(\tau)^2 \\ &= \mathbb{V}(\tau). \end{aligned} \quad (50.18)$$

□

Wie die Subskripte in Theorem 50.1 verdeutlichen, beziehen sich die Aussagen von Theorem 50.1 auf die Zufallsvariablen zur Modellierung der Daten einer Person i und einer Testmessung j und gelten gleichermaßen für alle $i = 1, \dots, n$ und $j = 1, \dots, m$. Aussage (1) von Theorem 50.1 betrifft den Erwartungswert des Messfehlers bedingt auf einem festen Wert des wahren Werts, in diesem Fall ist der wahre Wert also *keine* Zufallsvariable und die Verteilung von Interesse ist $\mathbb{P}(\varepsilon_{ij} | \tau_{ij} = t_{ij})$. Aussagen (2) bis (5) beziehen sich auf Eigenschaften der (gemeinsamen) Marginalverteilungen von y_{ij} , τ_{ij} und ε_{ij} . Im Sinne des vereinfachten Modells der Klassischen Testtheorie (Definition 50.3) werden obige Eigenschaften oft auch als

- (1) $\mathbb{E}(\varepsilon | \tau = t) = 0$
- (2) $\mathbb{E}(\varepsilon) = 0$
- (3) $\mathbb{C}(\tau, \varepsilon) = 0$
- (4) $\mathbb{V}(y) = \mathbb{V}(\tau) + \mathbb{V}(\varepsilon)$
- (5) $\mathbb{C}(y, \tau) = \mathbb{V}(\tau)$

geschrieben.

Spezieller Weise *folgt* nach Aussage (1) von Theorem 50.1 für das Modell multipler Messungen, dass der bedingte Erwartungswert des Messfehlers $\mathbb{E}(\varepsilon | \tau = t)$ gleich Null ist aus der Definition des wahren Werts. Dies steht im direkten Gegenentwurf zu typischen Annahmen über Messfehler, die üblicherweise einen Messfehler mit einem (bedingten) Erwartungswert von Null *definieren*, da sie von Null verschiedene Beiträge zu einem Datenpunkt als Teil der durch sie repräsentierten Theorie konzeptualisieren. Weil im Modell multipler Testmessungen der bedingte Erwartungswert des Messfehlers für *jeden* wahren Wert t gleich Null, folgt dann nach (2) von Theorem 50.1, dass auch der marginale Erwartungswert des Messfehlers $\mathbb{E}(\varepsilon)$ gleich Null ist.

Aussage (3) von Theorem 50.1, dass im Modell multipler Testmessungen also gilt, dass die Kovarianz der wahren Werte und der Messfehler gleich Null ist, besagt bekanntlich, dass hohe oder niedrige wahre Werte nicht systematisch mit hohen oder niedrigen Messfehlern assoziiert sind. Die daraus folgende Aussagen (4) und (5) von Theorem 50.1, dass also die Varianz der beobachteten Werte additiv in Varianzbeiträge der wahren Wert und der Messfehler zerlegt werden kann und dass die Kovarianz der beobachteten Werte und der wahren Werte einer Testmessung der Varianz der wahren Werte entspricht, werden an späterer Stelle essentiell für die Eigenschaften der Reliabilität von Testmessungen sein.

Beispiel

Im Folgenden wollen wir Theorem 50.1 noch an einem konkreten ersten Beispiel nach Lord & Novick (1968), Exercise 2.17 für ein Modell multipler Testmessungen veranschaulichen, wobei wir hierbei natürlich auf eine einzige Testmessung fokussieren.

Theorem 50.2 (Normalverteilungsbeispiel). *Es seien $i = 1, \dots, n$ und $m := 1$ mit*

$$\mathbb{P}(\tau_{i1}) := N(\mu, 1) \text{ und } \mathbb{P}(y_{i1}|\tau_{i1}) := N(\tau_{i1}, 1) \quad (50.19)$$

Dann gelten

- (1) $\mathbb{P}(y_{i1}) = N(\mu, 2)$
- (2) $\mathbb{P}(\varepsilon_{i1}|\tau_{i1}) = N(0, 1)$
- (3) $\mathbb{P}(\varepsilon_{i1}) = N(0, 1)$
- (4) $\mathbb{C}(\tau_{i1}, \varepsilon_{i1}) = 0$

◦

Beweis. Zur Vereinfachung der Notation setzen wir $\tau := \tau_{i1}$, $y := y_{i1}$, $\varepsilon := \varepsilon_{i1}$.

(1) Wir betrachten die durch

$$\mathbb{P}(\tau) = N(\mu, 1) \text{ und } \mathbb{P}(y|\tau) = N(\tau, 1) \quad (50.20)$$

induzierte gemeinsame Verteilung von τ und y , wobei offenbar

$$\mathbb{P}(y|\tau) = N(a \cdot \mu + b, 1) \text{ mit } a := 1 \text{ und } b := 0 \quad (50.21)$$

gilt. Aus dem Theorem zu gemeinsamen Normalverteilungen (vgl. Theorem 29.5) ergibt sich dann zunächst, dass

$$\begin{pmatrix} \tau \\ y \end{pmatrix} \sim N \left(\begin{pmatrix} \mu \\ 1 \cdot \mu + 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \cdot 1 \\ 1 \cdot 1 & 1 + 1 \cdot 1 \cdot 1 \end{pmatrix} \right) = N \left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \right) \quad (50.22)$$

Aus dem Theorem zu marginalen Normalverteilungen (vgl. Theorem 29.4) ergibt sich dann durch Ablesen $y \sim N(\mu, 2)$.

(2) Wir betrachten $\mathbb{P}(\varepsilon|\tau = t)$ für einen beliebigen Wert $t \in \mathbb{R}$. Dann gilt

$$\varepsilon := y - t \text{ mit } y \sim N(t, 1) \quad (50.23)$$

Mit dem Theorem zu linear-affinen Transformation normalverteilter Zufallsvariablen (vgl. Theorem 29.8) gilt dann

$$\varepsilon \sim N(t - t, 1^2 \cdot 1) = N(0, 1) \quad (50.24)$$

Die Tatsache, dass dies für alle möglichen Werte von τ gilt, ist gerade Aussage (2).

(3) und (4) Wir betrachten die durch

$$\mathbb{P}(\tau) = N(\mu, 1) \text{ und } \mathbb{P}(\varepsilon|\tau) = N(0, 1) \quad (50.25)$$

induzierte gemeinsame Verteilung von τ und ε , wobei offenbar

$$\mathbb{P}(\varepsilon|\tau) = N(a \cdot \mu + b, 1) \text{ mit } a := 0 \text{ und } b := 0 \quad (50.26)$$

gilt. Aus dem Theorem zu gemeinsamen Normalverteilungen (vgl. Theorem 29.5) ergibt sich dann zunächst, dass

$$\begin{pmatrix} \tau \\ \varepsilon \end{pmatrix} \sim N \left(\begin{pmatrix} \mu \\ 0 \cdot \mu + 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \cdot 0 \\ 0 \cdot 1 & 1 + 0 \cdot 1 \cdot 0 \end{pmatrix} \right) = N \left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \quad (50.27)$$

Aus dem Theorem zu marginalen Normalverteilungen (vgl. Theorem 29.4) ergeben sich dann durch Ablesen

$$\varepsilon \sim N(0, 1) \text{ und } \mathbb{C}(\tau, \varepsilon) = 0. \quad (50.28)$$

□

Das Beispiel zeigt insbesondere, wie die Spezifika des Modells multipler Testmessungen der Klassischen Testtheorie äquivalent durch Definition eines zeitgemäßen probabilistischen Modells erzeugt werden können. Hier induzieren die Definition der marginalen Verteilung des wahren Werts und die Definition der bedingten Verteilung des beobachteten Werts zunächst die gemeinsame Normalverteilung von τ_{i1} und y_{i1} . Die Definition des Messfehlers als Differenz zwischen beobachtetem Wert und dem bedingten Erwartungswert des wahren Werts ergibt dann die bedingte Verteilung des Messfehlers. Diese und die Definitionen des Beispiels induzieren dann eine gemeinsame Normalverteilung von τ_{i1} und ε_{i1} . Die weiteren Eigenschaften im Sinne von Theorem 50.1 ergeben sich in diesem Beispiel dann mit den Eigenschaften gemeinsamer Normalverteilungen.

Mithilfe folgender Simulation wollen wir das hier diskutierte Beispiel noch weiter verdeutlichen. Wir setzen dafür $\mu := 1$ und generieren 10^4 Realisierungen der marginalen Verteilung von τ_{i1} und der bedingten Verteilung von y_{i1} .

```
n = 1e4 # Personenanzahl
m = 1 # Testmessungsanzahl
mu = 1 # Erwartungswertparameter Wahrer Werte
T = matrix(rep(NaN, n*m), nrow = n) # Wahrer Wert Array
Y = matrix(rep(NaN, n*m), nrow = n) # Beobachteter Wert Array
E = matrix(rep(NaN, n*m), nrow = n) # Messfehler Array
for(i in 1:n){
  for(j in 1:m){
    T[i,j] = rnorm(1,mu,1) # Wahrer Wert Realisierung
    Y[i,j] = rnorm(1,T[i,j],1) # Beobachteter Wert Realisierung
    E[i,j] = Y[i,j] - T[i,j]} # Messfehler Realisierung
e_hat_es = mean(E[,1]) # Erwartungswertschätzung Messfehler
c_hat_ts_es = cov(T[,1],E[,1]) # Kovarianzschätzung Wahren Wert, Messfehler
v_hat_os = var(Y[,1]) # Varianzschätzung Beobachteter Wert
v_hat_ts = var(T[,1]) # Varianzschätzung Wahrer Wert
v_hat_es = var(E[,1]) # Varianzschätzung Messfehler
c_hat_os_ts = cov(Y[,1],E[,1]) # Kovarianzschätzung Beobachteter Wert, Wahrer Wert
```

Abbildung 50.1 A und B zeigen die resultierenden marginale Verteilungen von y_{i1} und ε_{i1} mit ihren theoretischen Entsprechungen laut Theorem 50.2. Abbildung 50.1 C zeigt die gemeinsame Verteilung von τ_{i1} und ε_{i1} mit ihrer theoretischen Entsprechung. Abbildung 50.1 D schließlich zeigt die auf Grundlage der Simulation gewonnenen Schätzer relevanter Erwartungswerte, Varianzen und Kovarianzen in diesem Modell, die die theoretischen Einsichten nach Theorem 50.2 bestätigen.

Eigenschaften bezüglich zweier Testmessungen

Bisher haben wir fünf Eigenschaften des Modells multipler Testmessungen kennengelernt, die für den wahren Wert τ_{ij} , den beobachteten Wert y_{ij} und den Messfehler ε_{ij} einer (und damit jeder) Person i und einer (und damit jeder) Testmessung j gelten. Im Folgenden beschäftigen wir uns mit Eigenschaften des Modells multipler Testmessungen, die für die wahren Werte τ_{ij} und τ_{ik} , beobachteten Werte y_{ij} und y_{ik} und Messfehler ε_{ij} und ε_{ik} einer (und damit jeder) Person i hinsichtlich *zweier* Testmessungen j und k gelten. Wir fassen diese Eigenschaften, die manchmal als *lokale Unkorreliertheit* des Modells multipler Testmessungen bezeichnet werden (vgl. Krauth (1995)), in folgende Theorem zusammen.

Theorem 50.3 (Eigenschaften bezüglich zweier Testmessungen). *Gegeben sei das Modell multipler Testmessungen. Dann gelten für alle $i = 1, \dots, n$ und alle j und k mit $1 \leq j, k \leq m$ und $j \neq k$, dass*

- (1) $\mathbb{C}(y_{ij}, y_{ik} | \tau_{ij} = t_{ij}, \tau_{ik} = t_{ik}) = 0$
- (2) $\mathbb{C}(\varepsilon_{ij}, \varepsilon_{ik} | \tau_{ij} = t_{ij}, \tau_{ik} = t_{ik}) = 0$

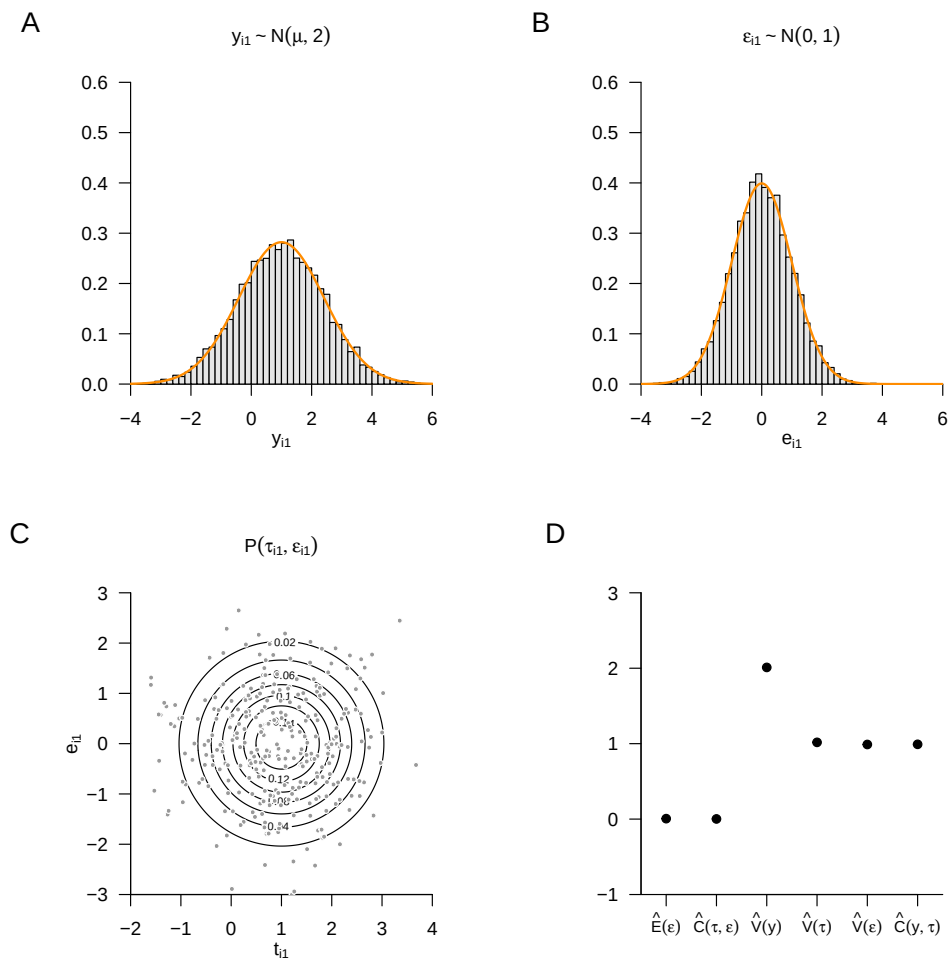


Abbildung 50.1. Normalverteilungsbeispiel bezüglich einer Testmessung.

- (3) $\mathbb{C}(\varepsilon_{ij}, \varepsilon_{ik}) = 0$
- (4) $\mathbb{C}(\tau_{ij}, \varepsilon_{ik}) = 0$
- (5) $\mathbb{C}(y_{ij}, y_{ik}) = \mathbb{C}(\tau_{ij}, \tau_{ik})$

◦

Beweis. Zur Vereinfachung der Notation verzichten wir in den Beweisen auf das i Subskript. Wir betrachten weiterhin nur den diskreten Fall und setzen die Existenz der marginalen WMF

$$p(t_j, \tilde{y}_j, t_k, \tilde{y}_k) = p(t_j, t_k)p(\tilde{y}_j|t_j)p(\tilde{y}_k|t_k) \quad (50.29)$$

und folglich auch der bedingten Wahrscheinlichkeitsmassefunktion

$$p(\tilde{y}_j, \tilde{y}_k|t_j, t_k) = \frac{p(t_j, \tilde{y}_j, t_k, \tilde{y}_k)}{p(t_j, t_k)} = \frac{p(t_j, t_k)p(\tilde{y}_j|t_j)p(\tilde{y}_k|t_k)}{p(t_j, t_k)} = p(\tilde{y}_j|t_j)p(\tilde{y}_k|t_k) \quad (50.30)$$

voraus. Der kontinuierliche Fall folgt dann wieder analog.

(1) Es gilt

$$\begin{aligned} \mathbb{C}(y_j, y_k|\tau_j = t_j, \tau_k = t_k) &= \sum_{\tilde{y}_j \in Y_j} \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_j - \mathbb{E}(y_j|\tau_j = t_j))(\tilde{y}_k - \mathbb{E}(y_k|\tau_k = t_k))p(\tilde{y}_j|t_j)p(\tilde{y}_k|t_k) \\ &= \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t_j)p(\tilde{y}_j|t_j) \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_k - t_k)p(\tilde{y}_k|t_k) \\ &= \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t_j)p(\tilde{y}_j|t_j) \left(\sum_{\tilde{y}_k \in Y_k} \tilde{y}_k p(\tilde{y}_k|t_k) - t_k \sum_{\tilde{y}_k \in Y_k} p(\tilde{y}_k|t_k) \right) \\ &= \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t_j)p(\tilde{y}_j|t_j) (t_k - t_k \cdot 1) \\ &= \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t_j)p(\tilde{y}_j|t_j) \cdot 0 \\ &= 0. \end{aligned} \quad (50.31)$$

(2) Wir bestimmen zunächst $\mathbb{E}(\varepsilon_j \varepsilon_k|\tau_j = t_j, \tau_k = t_k)$. Es gilt

$$\begin{aligned} \mathbb{E}(\varepsilon_j \varepsilon_k|\tau_j = t_j, \tau_k = t_k) &= \mathbb{E}((y_j - \tau_j)(y_k - \tau_k)|\tau_j = t_j, \tau_k = t_k) \\ &= \sum_{\tilde{y}_j \in Y_j} \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_j - t_j)(\tilde{y}_k - t_k)p(\tilde{y}_j|t_j)p(\tilde{y}_k|t_k) \\ &= \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t_j)p(\tilde{y}_j|t_j) \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_k - t_k)p(\tilde{y}_k|t_k) \\ &= \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t_j)p(\tilde{y}_j|t_j) \left(\sum_{\tilde{y}_k \in Y_k} \tilde{y}_k p(\tilde{y}_k|t_k) - t_k \sum_{\tilde{y}_k \in Y_k} p(\tilde{y}_k|t_k) \right) \\ &= \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t_j)p(\tilde{y}_j|t_j) (t_k - t_k \cdot 1) \\ &= \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t_j)p(\tilde{y}_j|t_j) \cdot 0 \\ &= 0. \end{aligned} \quad (50.32)$$

Mit dem Verschiebungssatz der bedingten Kovarianz (Theorem 25.8) und Aussage (1) des Theorems zu den Eigenschaften bezüglich einer Testmessung (Theorem 50.1) folgt dann

$$\mathbb{C}(\varepsilon_j, \varepsilon_k|\tau_j = t_j, \tau_k = t_k) = \mathbb{E}(\varepsilon_j \varepsilon_k|\tau_j = t_j, \tau_k = t_k) - \mathbb{E}(\varepsilon_j|\tau_j = t_j)\mathbb{E}(\varepsilon_k|\tau_k = t_k) = 0 - 0 \cdot 0 = 0. \quad (50.33)$$

(3) Wir bestimmen zunächst $\mathbb{E}(\varepsilon_j \varepsilon_k)$. Mit dem Beweis von Aussage (2) des vorliegenden Theorems ergibt

sich

$$\begin{aligned}
\mathbb{E}(\varepsilon_j \varepsilon_k) &= \mathbb{E}((y_j - \tau_j)(y_k - \tau_k)) \\
&= \sum_{t_j \in T_j} \sum_{t_k \in T_k} \sum_{\tilde{y}_j \in Y_j} \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_j - t_j)(\tilde{y}_k - t_k) p(\tilde{y}_j | t_j) p(\tilde{y}_k | t_k) p(t_j, t_k) \\
&= \sum_{t_j \in T_j} \sum_{t_k \in T_k} \sum_{\tilde{y}_j \in Y_j} \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_j - \mathbb{E}(y_j | \tau_j = t_j))(\tilde{y}_k - \mathbb{E}(y_k | \tau_k = t_k)) p(\tilde{y}_j | t_j) p(\tilde{y}_k | t_k) p(t_j, t_k) \\
&= \sum_{t_j \in T_j} \sum_{t_k \in T_k} \mathbb{E}((y_j - \tau_j)(y_k - \tau_k) | \tau_j = t_j, \tau_k = t_k) p(t_j, t_k) \\
&= \sum_{t_j \in T_j} \sum_{t_k \in T_k} \mathbb{E}(\varepsilon_j \varepsilon_k | \tau_j = t_j, \tau_k = t_k) p(t_j, t_k) \\
&= \sum_{t_j \in T_j} \sum_{t_k \in T_k} 0 \cdot p(t_j, t_k) \\
&= 0.
\end{aligned} \tag{50.34}$$

Mit dem Kovarianzverschiebungssatz (Theorem 25.2) und Aussage (2) des Theorems zu den Eigenschaften bezüglich einer Testmessung (Theorem 50.1) ergibt sich dann

$$\mathbb{C}(\varepsilon_j, \varepsilon_k) = \mathbb{E}(\varepsilon_j \varepsilon_k) - \mathbb{E}(\varepsilon_j) \mathbb{E}(\varepsilon_k) = 0 - 0 \cdot 0 = 0. \tag{50.35}$$

(4) Mit dem Kovarianzverschiebungssatz (Theorem 25.2) und Aussage (2) des Theorems zu den Eigenschaften bezüglich einer Testmessung (Theorem 50.1) gilt

$$\begin{aligned}
\mathbb{C}(\tau_j, \varepsilon_k) &= \mathbb{E}(\tau_j \varepsilon_k) - \mathbb{E}(\tau_j) \mathbb{E}(\varepsilon_k) \\
&= \mathbb{E}(\tau_j \varepsilon_k) - \mathbb{E}(\tau_j) \cdot 0 \\
&= \mathbb{E}(\tau_j (\tilde{y}_k - \tau_k)) \\
&= \sum_{t_j \in T_j} \sum_{t_k \in T_k} \sum_{\tilde{y}_k \in Y_k} t_j (y_k - t_k) p(\tilde{y}_k | t_k) p(t_j, t_k) \\
&= \sum_{t_j \in T_j} t_j \sum_{t_k \in T_k} p(t_j, t_k) \sum_{\tilde{y}_k \in Y_k} (y_k - t_k) p(\tilde{y}_k | t_k) \\
&= \sum_{t_j \in T_j} t_j \sum_{t_k \in T_k} p(t_j, t_k) \left(\sum_{\tilde{y}_k \in Y_k} y_k p(\tilde{y}_k | t_k) - t_k \sum_{\tilde{y}_k \in Y_k} p(\tilde{y}_k | t_k) \right) \\
&= \sum_{t_j \in T_j} t_j \sum_{t_k \in T_k} p(t_j, t_k) (t_k - t_k \cdot 1) \\
&= \sum_{t_j \in T_j} t_j \sum_{t_k \in T_k} p(t_j, t_k) \cdot 0 \\
&= 0.
\end{aligned} \tag{50.36}$$

(5) Wir halten zunächst fest, dass mit Aussage (4) durch Vertauschen der Indizes und der Symmetrie der Kovarianz auch

$$\mathbb{C}(\tau_k, \varepsilon_j) = \mathbb{C}(\varepsilon_j, \tau_k) = 0 \tag{50.37}$$

gilt. Mit dem Theorem zur paarweisen Addition von Zufallsvariablen (Theorem 25.5) und Aussage (3) des vorliegenden Theorems gilt dann

$$\begin{aligned}
\mathbb{C}(y_j, y_k) &= \mathbb{C}(\tau_j + \varepsilon_j, \tau_k + \varepsilon_k) \\
&= \mathbb{C}(\tau_j, \tau_k) + \mathbb{C}(\tau_j, \varepsilon_k) + \mathbb{C}(\varepsilon_j, \tau_k) + \mathbb{C}(\varepsilon_j, \varepsilon_k) \\
&= \mathbb{C}(\tau_j, \tau_k) + 0 + 0 + 0 \\
&= \mathbb{C}(\tau_j, \tau_k).
\end{aligned} \tag{50.38}$$

□

Im Sinne des vereinfachten Modells der Klassischen Testtheorie nach Definition 50.3 werden die Aussagen von Theorem 50.3 oft auch als

$$(1) \quad \mathbb{C}(y_j, y_k | \tau_j = t_j, \tau_k = t_k) = 0$$

- (2) $\mathbb{C}(\varepsilon_j, \varepsilon_k | \tau_j = t_j, \tau_k = t_k) = 0$
- (3) $\mathbb{C}(\varepsilon_j, \varepsilon_k) = 0$
- (4) $\mathbb{C}(\tau_j, \varepsilon_k) = 0$
- (5) $\mathbb{C}(y_j, y_k) = \mathbb{C}(\tau_j, \tau_k)$

geschrieben. Die ersten beiden Aussagen von Theorem 50.3 besagen, dass, bedingt auf den jeweiligen wahren Werten, die Kovarianzen von beobachteten Werten sowie die Kovarianzen der Messfehler einer Person zwischen zwei Testmessungen gleich Null sind. Für die Messfehler gilt dies nach Aussage (3) von Theorem 50.3 auch im Sinne der unbedingten Marginalverteilung. Dies entspricht damit der paarweisen Unabhängigkeit von Messfehlern über Testmessungen im Modell multipler Testmessungen. Weiterhin ist nach Aussage (4) die Kovarianz des wahren Werts bei einer Testmessung mit dem Messfehler einer anderen Testmessung gleich Null. Schließlich gilt nach Aussage (5), dass die Kovarianz der beobachteten Werte zweier Testmessungen gleich der Kovarianz der wahren Werte ist.

Beispiel

Als erstes Beispiel für ein Modell multipler Testmessungen mit $m > 1$ setzen wir das Beispiel aus Theorem 50.2 fort und betrachten den Fall zweier Testmessungen $j = 1, 2$. Für $i = 1, \dots, n$ seien entsprechend

$$\mathbb{P}(\tau_{i1}, \tau_{i2}) = \mathbb{P}(\tau_{i2} | \tau_{i1}) \mathbb{P}(\tau_{i1}) \quad (50.39)$$

mit

$$\mathbb{P}(\tau_{i1}) := N(1, 1) \text{ und } \mathbb{P}(\tau_{i2} | \tau_{i1}) := N(\tau_{i1} + 1, 1) \quad (50.40)$$

Die Verteilung des True-Score von Person i in Testmessung $j = 2$ hängt in diesem Beispiel also explizit von der Verteilung des wahren Werts von Person i in Testmessung $j = 1$ ab. Weiterhin seien

$$\mathbb{P}(y_{i1} | \tau_{i1}) := N(\tau_{i1}, 1) \text{ und } \mathbb{P}(y_{i2} | \tau_{i2}) := N(\tau_{i2}, 2) \quad (50.41)$$

Die Propensitätsverteilungen von Person i in Testmessung $j = 1$ unterscheide sich also von der von Person i in Testmessung $j = 2$. Folgender **R** Code generiert 10^5 Realisierungen dieses Modells und schätzt die in Theorem 50.3 betrachteten marginalen Kovarianzen $\mathbb{C}(\varepsilon_{i1}, \varepsilon_{i2})$, $\mathbb{C}(\tau_{i1}, \varepsilon_{i2})$, $\mathbb{C}(y_{i1}, y_{i2})$ und $\mathbb{C}(\tau_{i1}, \tau_{i2})$.

```
n = 1e5 # Personenanzahl
m = 2 # Testmessungsanzahl
mu = 1 # Wahrer Werte Erwartungswertparameter
T = matrix(rep(NaN, n*m), nrow = n) # Wahrer Wert Array
Y = matrix(rep(NaN, n*m), nrow = n) # Beobachteter Wert Array
E = matrix(rep(NaN, n*m), nrow = n) # Messfehler Array
for(i in 1:n){
  T[i,1] = rnorm(1,1,1) # Iteration über Personen
  Y[i,1] = rnorm(1,T[i,1],1) # Wahrer Wert Realisierung für j = 1
  E[i,1] = Y[i,1] - T[i,1] # Beobachteter Wert Realisierung für j = 1
  T[i,2] = rnorm(1,T[i,1] + 1,.5) # Messfehler Realisierung für j = 1
  Y[i,2] = rnorm(1,T[i,2],.5) # Wahrer Wert Realisierung für j = 2
  E[i,2] = Y[i,2] - T[i,2] # Beobachteter Wert Realisierung für j = 2
  # Messfehler Realisierung für j = 2
  c_hat_e1_e2 = cov(E[,1],E[,2]) # Kovarianzschätzung Messfehler 1, Messfehler 2
  c_hat_t1_e2 = cov(T[,1],E[,2]) # Kovarianzschätzung Wahrer Wert 1, Messfehler 2
  c_hat_y1_y2 = cov(Y[,1],Y[,2]) # Kovarianzschätzung Beobachteter Wert 1, Beobachteter Wert 2
  c_hat_t1_t2 = cov(T[,1],T[,2]) # Kovarianzschätzung Wahrer Wert 1, Wahrer Wert 2
}
```

Abbildung 50.2 visualisiert 500 der so generierten Zufallsvariablenrealisierungen und dokumentiert die resultierenden Kovarianzschätzer. Abbildung 50.2 A zeigt die marginale Unkorreliertheit der Messfehler bezüglich zweier Testmessungen (Aussage (2) von von Theorem 50.3), Abbildung 50.2 B zeigt die marginale Unkorreliertheit des wahren Wertes und des Messfehlers bezüglich zweier Testmessungen (Aussage (3) von von Theorem 50.3) und

Abbildung 50.2 C und Abbildung 50.2 D zeigen die Gleichheit der marginalen Kovarianzen von beobachteten und wahren Werten bezüglich zweier Testmessungen.

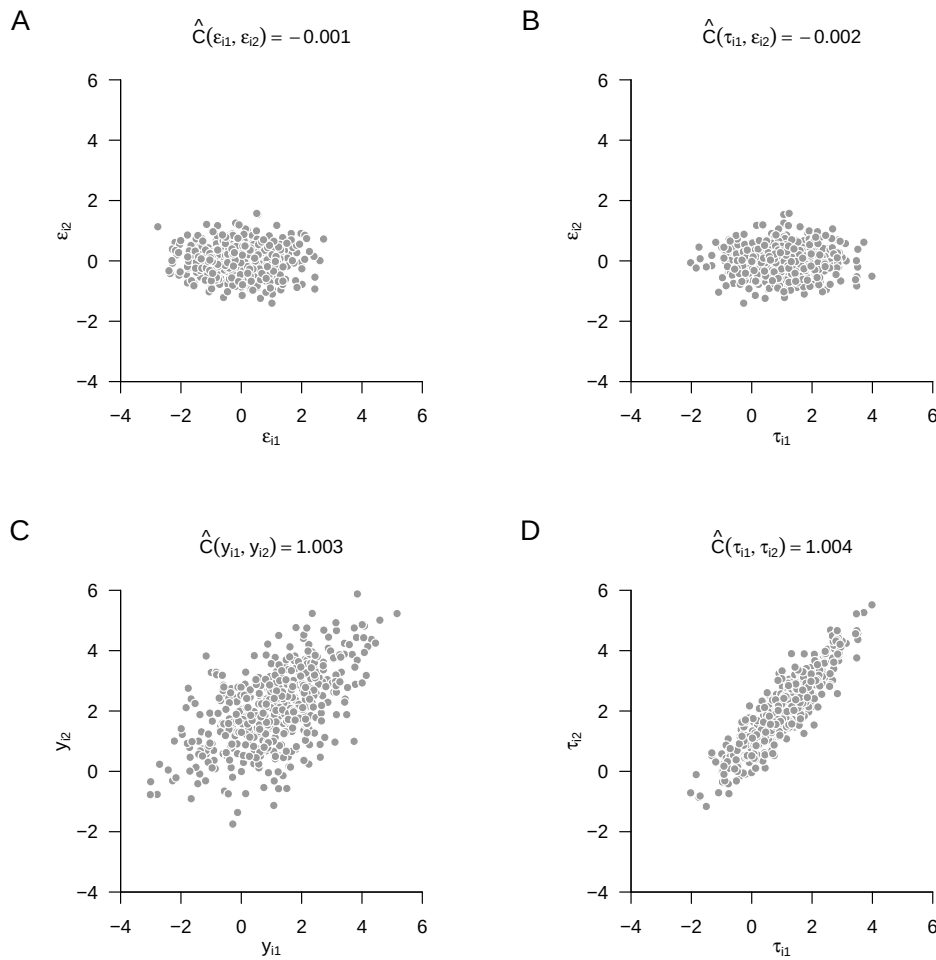


Abbildung 50.2. Normalverteilungsbeispiel bezüglich zweier Testmessungen.

50.1.2. Das Modell paralleler Testmessungen

Bisher haben wir im Modell der multiplen Testmessungen keine Aussage zu den Verhältnissen der wahren Werte über verschiedene Testmessungen hinweg gemacht. Wir haben einerseits angenommen, dass für die Marginalverteilung der Testmessungen bei einer Person i keine Unabhängigkeit gelten muss, dass also im Allgemeinen gilt, dass für $i = 1, \dots, n$

$$\mathbb{P}(\tau_{i1}, \dots, \tau_{im}) \neq \prod_{j=1}^m \mathbb{P}(\tau_{ij}). \quad (50.42)$$

Andererseits haben wir die Form möglicher Abhängigkeiten zwischen den wahren Werten $\tau_{i1}, \dots, \tau_{im}$ bislang nicht genauer spezifiziert. Die Klassische Testtheorie betrachtet in dieser Hinsicht einige Spezialfälle, die sich allgemein durch funktionale Abhängigkeiten zwischen τ_{i1} und $\tau_{i2}, \dots, \tau_{im}$ der Form

$$\tau_{ij} = f(\tau_{i1}) \text{ für } f: \mathbb{R} \rightarrow \mathbb{R} \text{ und } j = 2, \dots, m. \quad (50.43)$$

ausdrücken lassen. Wir betrachten im Folgenden den Fall, dass $f := \text{id}_{\mathbb{R}}$, dass also insbesondere für Realisierungen t_{ij} von τ_{ij} gilt, dass

$$t_{ij} = \text{id}_{\mathbb{R}}(t_{i1}) = t_{i1} \text{ für } j = 2, \dots, m, \quad (50.44)$$

dass also die Werte der wahren Werte einer Person über Testmessungen identisch sind. Die Klassische Testtheorie bezeichnet solche Testmessungen als *parallele Testmessungen*. Eine weitere Form der funktionalen Abhängigkeit, die wir hier nicht weiter vertiefen wollen, ist der Fall, dass es sich bei f um eine linear-affine Funktion handelt, dass also

$$\tau_{ij} = f(\tau_{i1}) = a\tau_{i1} + b \text{ für } a, b \in \mathbb{R} \text{ und } j = 2, \dots, m. \quad (50.45)$$

Die Klassische Testtheorie bezeichnet solche Testmessungen als *wesentlich τ -äquivalente Testmessungen*.

Definition 50.4 (Modell paralleler Testmessungen). Für $i = 1, \dots, n$ und $j = 1, \dots, m$ seien τ_i eine Zufallsvariable, die den wahren Wert der i ten Person in jeder Testmessung $j = 1, \dots, m$ modelliere, y_{ij} eine Zufallsvariable, die den beobachteten Wert der i ten Person in der j ten Testmessung modelliere und $\varepsilon_{ij} := y_{ij} - \tau_i$ die Zufallsvariable, die den Messfehler der i ten Person in der j ten Testmessung modelliere. Dann heißt die gemeinsame Verteilung der τ_i und y_{ij} mit den Faktorisierungseigenschaften

$$\mathbb{P}(\tau_1, y_{11}, \dots, y_{1m}, \dots, \tau_n, y_{n1}, \dots, y_{nm}) := \prod_{i=1}^n \mathbb{P}(\tau_i) \prod_{j=1}^m \mathbb{P}(y_{ij}|\tau_i) \quad (50.46)$$

das *Modell paralleler Testmessungen*, wenn gilt, dass

- (1) $\mathbb{P}(\tau_1) = \dots = \mathbb{P}(\tau_n)$
- (2) $\mathbb{P}(y_{1j}|\tau_1) = \dots = \mathbb{P}(y_{nj}|\tau_n)$ für alle $1 \leq j \leq m$
- (3) $\mathbb{E}(y_{ij}|\tau_i = t_i) = \mathbb{E}(y_{ik}|\tau_i = t_i) := t_i$ für alle $1 \leq i \leq n, 1 \leq j, k \leq m$
- (4) $\mathbb{V}(y_{ij}|\tau_i = t_i) = \mathbb{V}(y_{ik}|\tau_i = t_i)$ für alle $1 \leq i \leq n, 1 \leq j, k \leq m$

•

Insbesondere werden also im Modell paralleler Testmessungen die Werte des wahren Werts einer Person werden über Testmessungen hinweg als identisch angenommen. Damit geht dann die Varianz der beobachteten Werte einer Person zwischen zwei Testmessungen allein auf die Propensitätsverteilung zurück. Aus generativer Sichtweise entstehen Werte der beobachteten Werte also wie folgt: Zunächst wird für die i te Person und Testmessungen $j = 1, \dots, m$ ein True-Score t_i von gemäß $\mathbb{P}(\tau_i)$ realisiert. Dann wird für die i te und die Testmessung $j = 1, \dots, m$ ein Observed-Score y_{ij} anhand von $\mathbb{P}(y_{ij}|\tau_i = t_i)$ realisiert.

Eigenschaften des Modells paralleler Testmessungen

Betrachtet man nur eine einzige Testmessung, so hat das Modell paralleler Testmessungen natürlich die gleiche Form wie das Modell multipler Testmessungen. Damit gilt dann aber auch Theorem 50.1 analog für das Modell paralleler Testmessungen. Betrachtet man allerdings mehr als eine Testmessung, so ergeben sich für das Modell paralleler Testmessungen speziellere Eigenschaften, die wir in folgendem Theorem festhalten

Theorem 50.4 (Eigenschaften des Modells paralleler Testmessungen bezüglich zweier Testmessungen). *Gegeben sei das Modell paralleler Testmessungen. Dann gelten für alle $i = 1, \dots, n$ und alle j, k mit $1 \leq j, k \leq m$ und $j \neq k$, dass*

- (1) $\mathbb{E}(y_{ij}) = \mathbb{E}(y_{ik})$
- (2) $\mathbb{V}(y_{ij}) = \mathbb{V}(y_{ik})$
- (3) $\mathbb{C}(\varepsilon_{ij}, \varepsilon_{ik}) = 0$
- (4) $\mathbb{C}(\tau_i, \varepsilon_{ik}) = 0$
- (5) $\mathbb{C}(y_{ij}, y_{ik}) = \mathbb{V}(\tau_i)$

◦

Beweis. Zum Beweis setzen zur Vereinfachung der Notation zunächst

$$y_j := y_{ij}, y_k := y_{ik}, \tau := \tau_i, \tilde{y}_j := \tilde{y}_{ij}, \tilde{y}_k := \tilde{y}_{ik}, t := t_i, Y_j := Y_{ij}, Y_k := Y_{ik} \text{ und } T := T_i \quad (50.47)$$

für alle $i = 1, \dots, n$. Wir betrachten weiterhin nur den diskreten Fall, setzen also die Existenz einer Wahrscheinlichkeitsmassenfunktion der Form

$$p(t, \tilde{y}_j, \tilde{y}_k) = p(\tilde{y}_j|t)p(\tilde{y}_k|t)p(t) \quad (50.48)$$

voraus. Der kontinuierliche Fall folgt dann analog.

(1) Mit der Gleichheit der bedingten Erwartungswerte im Falle paralleler Testmessungen gilt

$$\mathbb{E}(y_j) = \sum_{t \in T} \sum_{\tilde{y}_j \in Y_j} \tilde{y}_j p(\tilde{y}_j|t)p(t) = \sum_{t \in T} \mathbb{E}(y_j|\tau = t)p(t) = \sum_{t \in T} \sum_{\tilde{y}_k \in Y_k} \tilde{y}_k p(\tilde{y}_k|t)p(t) = \mathbb{E}(y_k). \quad (50.49)$$

(2) Mit dem Satz von der iterierten Varianz (Theorem 24.8) gilt

$$\mathbb{V}(y_j) = \mathbb{V}(\mathbb{E}(y_j|\tau)) + \mathbb{E}(\mathbb{V}(y_j|\tau)) = \mathbb{V}(\mathbb{E}(y_k|\tau)) + \mathbb{E}(\mathbb{V}(y_k|\tau)) = \mathbb{V}(y_k). \quad (50.50)$$

(3) Wir bestimmen zunächst $\mathbb{E}(\varepsilon_j \varepsilon_k)$. Mit Aussage (2) von Theorem 50.1 gilt dann

$$\begin{aligned} \mathbb{E}(\varepsilon_j \varepsilon_k) &:= \mathbb{E}((y_j - \tau)(y_k - \tau)) \\ &= \sum_{t \in T} \sum_{\tilde{y}_j \in Y_j} \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_j - t)(\tilde{y}_k - t)p(t)p(\tilde{y}_j|t)p(\tilde{y}_k|t) \\ &= \sum_{t \in T} p(t) \sum_{\tilde{y}_j \in Y_j} \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_j - t)(\tilde{y}_k - t)p(\tilde{y}_j|t)p(\tilde{y}_k|t) \\ &= \sum_{t \in T} p(t) \sum_{\tilde{y}_j \in Y_j} (\tilde{y}_j - t)p(\tilde{y}_j|t) \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_k - t)p(\tilde{y}_k|t) \\ &= \sum_{t \in T} p(t) \left(\sum_{\tilde{y}_j \in Y_j} \tilde{y}_j p(\tilde{y}_j|t) - t \sum_{\tilde{y}_j \in Y_j} p(\tilde{y}_j|t) \right) \left(\sum_{\tilde{y}_k \in Y_k} \tilde{y}_k p(\tilde{y}_k|t) - t \sum_{\tilde{y}_k \in Y_k} p(\tilde{y}_k|t) \right) \\ &= \sum_{t \in T} p(t) (t - t) (t - t) \\ &= \sum_{t \in T} p(t) \cdot 0 \cdot 0 \\ &= 0. \end{aligned} \quad (50.51)$$

Mit dem Kovarianzverschiebungssatz (Theorem 25.2) und wiederum mit Aussage (2) von Theorem 50.1 folgt dann

$$\mathbb{C}(\varepsilon_j, \varepsilon_k) = \mathbb{E}(\varepsilon_j \varepsilon_k) - \mathbb{E}(\varepsilon_j)\mathbb{E}(\varepsilon_k) = 0 - 0 \cdot 0 = 0 \quad (50.52)$$

(4) Mit dem Kovarianzverschiebungssatz (Theorem 25.2) und Aussage (2) von Theorem 50.1 gilt

$$\begin{aligned}
 \mathbb{C}(\tau, \varepsilon_k) &= \mathbb{E}(\tau \varepsilon_k) - \mathbb{E}(\tau) \mathbb{E}(\varepsilon_k) \\
 &= \mathbb{E}(\tau \varepsilon_k) - \mathbb{E}(\tau) \cdot 0 \\
 &= \mathbb{E}(\tau(y_k - \tau)) \\
 &= \sum_{t \in T} \sum_{\tilde{y}_k \in Y_k} t(\tilde{y}_k - t) p(t) p(\tilde{y}_k | t) \\
 &= \sum_{t \in T} t p(t) \sum_{\tilde{y}_k \in Y_k} (\tilde{y}_k - t) p(\tilde{y}_k | t) \\
 &= \sum_{t \in T} t p(t) \left(\sum_{\tilde{y}_k \in Y_k} \tilde{y}_k p(\tilde{y}_k | t) - t \sum_{\tilde{y}_k \in Y_k} p(\tilde{y}_k | t) \right) \\
 &= \sum_{t \in T} t p(t) (t - t) \\
 &= \sum_{t \in T} t p(t) \cdot 0 \\
 &= 0.
 \end{aligned} \tag{50.53}$$

(5) Wir halten zunächst fest, dass mit Aussage (2) des Theorems neben $\mathbb{C}(\tau, \varepsilon_k) = 0$ durch Austausch des Index und der Symmetrie der Kovarianz auch

$$\mathbb{C}(\tau, \varepsilon_j) = \mathbb{C}(\varepsilon_j, \tau) = 0 \tag{50.54}$$

gilt. Mit dem Theorem zur Kovarianz bei paarweiser Addition von Zufallsvariablen (Theorem 25.5) und Aussagen (3) und (4) des vorliegenden Theorems gilt dann

$$\begin{aligned}
 \mathbb{C}(y_j, y_k) &= \mathbb{C}(\tau + \varepsilon_j, \tau + \varepsilon_k) \\
 &= \mathbb{C}(\tau, \tau) + \mathbb{C}(\tau, \varepsilon_k) + \mathbb{C}(\varepsilon_j, \tau) + \mathbb{C}(\varepsilon_j, \varepsilon_k) \\
 &= \mathbb{C}(\tau, \tau) + 0 + 0 + 0 \\
 &= \mathbb{C}(\tau, \tau) \\
 &= \mathbb{V}(\tau).
 \end{aligned} \tag{50.55}$$

□

Aussage (1) von Theorem 50.4 besagt, dass bei Paralleltestmessungen alle Erwartungswerte der beobachteten Werte identisch sind und Aussage (2) des Theorems besagt, dass bei Paralleltestmessungen alle Varianzen der beobachteten Werte identisch sind. Die Aussagen (3) und (4) von Theorem 50.4 sind analog zu den lokalen Unkorreliertheitseigenschaften des Modells multipler Testmessungen. Aussage (5) schließlich besagt insbesondere, dass $\mathbb{C}(y_{ij}, y_{ik})$ für beliebige j und k identisch zu $\mathbb{V}(\tau_i)$ sind. Zusammen mit Aussage (2) des Theorems sind im Modell paralleler Testmessungen also alle paarweisen Korrelationen verschiedener Testmessungen identisch.

Beispiel

Wir betrachten den Fall zweier Testmessungen $j = 1, 2$ im Modell paralleler Testmessungen. Für $i = 1, \dots, n$ seien

$$\mathbb{P}(\tau_i) = N(1, 1) \text{ und } \mathbb{P}(y_{i1} | \tau_i) := \mathbb{P}(y_{i2} | \tau_i) := N(\tau_i, 1) \tag{50.56}$$

Für Person i gibt es also nur eine True-Score Zufallsvariable für alle Testmessungen und die Propensitätsverteilungen unterscheiden sich zwischen Testmessungen nicht.

```

n      = 1e5                                # Personenanzahl
m      = 2                                  # Testmessungsanzahl
T      = matrix(rep(NaN, n), nrow = n)      # Wahrer Wert Array
Y      = matrix(rep(NaN, n*m), nrow = n)    # Beobachteter Wert Array
E      = matrix(rep(NaN, n*m), nrow = n)    # Messfehler Array
for(i in 1:n){
  T[i] = rnorm(1,1,1)                        # Personeniterationen
  for(j in 1:m){
    Y[i,j] = rnorm(1,T[i],1)                # Wahrer Wert Realisierung für j = 1,2
    E[i,j] = Y[i,j] - T[i]}                # Testmessungsiterationen
  }
e_hat_o1_o2 = apply(Y, 2, mean)             # Beobachteter Wert Realisierung f
v_hat_o1_o2 = apply(Y, 2, var)              # Messfehler Realisierung
c_hat_e1_e2 = cov(E[,1],E[,2])             # Erwartungswertschätzung Beobachteter Wert 1, Beobachteter Wert 2
c_hat_t_e2 = cov(T, E[,2])                 # Varianzschätzung Beobachteter Wert 1, Beobachteter Wert 2
c_hat_y1_y2 = cov(Y[,1],Y[,2])             # Kovarianzschätzung Messfehler 1, Messfehler 2
v_hat_t     = var(T)                       # Kovarianzschätzung Wahrer Wert 1, Messfehler 2
v_hat_y1_y2 = cov(Y[,1],Y[,2])             # Kovarianzschätzung Beobachteter Wert 1, Beobachteter Wert 2
v_hat_t     = var(T)                       # Varianzschätzung Wahrer Wert

```

Abbildung 50.3 visualisiert 500 der so generierten Zufallsvariablenrealisierungen und dokumentiert die resultierenden Kovarianzschätzer. Abbildung 50.3 A zeigt die Unkorreliertheit der Messfehler bezüglich zweier paralleler Testmessungen (Aussage (3) von Theorem 50.4), Abbildung 50.3 B zeigt die Unkorreliertheit des wahren Wertes und des Messfehlers bezüglich zweier paralleler Testmessungen (Aussage (4) von Theorem 50.4) und Abbildung 50.3 C zeigt die durch die identischen wahren Werte induzierte Korrelation zwischen den marginalen Kovarianzen der beobachteten Werten zweier paralleler Testmessungen.

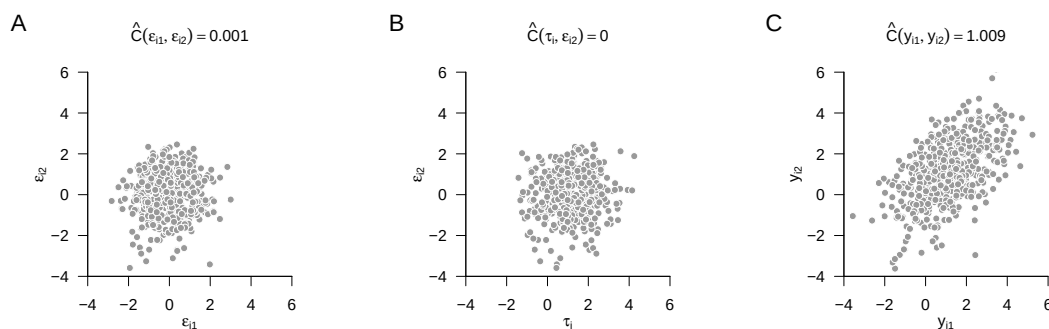


Abbildung 50.3. Normalverteilungsbeispiel bezüglich zweier paralleler Testmessungen.

50.2. Reliabilität

Die *Reliabilität* einer Testmessung ist das zentrale Konzept der Klassischen Testtheorie. Wir folgen hier dem Ansatz nach Lord & Novick (1968), wonach die Reliabilität einer Testmessung als quadrierte Korrelation von beobachtetem und wahrem Wert definiert ist. Basierend auf den Eigenschaften des Modells multipler Testmessungen ergeben sich dann verschiedene Möglichkeiten, diese Korrelation darzustellen und damit verschiedene Möglichkeiten, die Reliabilität einer Testmessung zu interpretieren. Allerdings ergibt sich dabei keine Möglichkeit, die Reliabilität von Testmessungen empirisch zu messen. Zentral ist dann die Übertragung des Konzepts der Reliabilität in das Modell paralleler Testmessungen. Die so definierte *Paralleltestreliabilität* ist dann empirisch schätzbar.

50.2.1. Reliabilität einer Testmessung

Wir beginnen zunächst mit der Definition der Reliabilität einer Testmessung im Modell multipler Testmessungen.

Definition 50.5 (Reliabilität einer Testmessung). Gegeben sei das Modell multipler Testmessungen für eine beliebige Testmessung j mit $1 \leq j \leq m$,

$$\mathbb{P}(\tau_{1j}, y_{1j}, \dots, \tau_{nj}, y_{nj}) := \prod_{i=1}^n \mathbb{P}(\tau_{ij}, y_{ij}) := \prod_{i=1}^n \mathbb{P}(y_{ij} | \tau_{ij}) \mathbb{P}(\tau_{ij}) \quad (50.57)$$

wobei nach Definition des Modells gilt, dass

$$\mathbb{P}(\tau_{1j}, y_{1j}) = \dots = \mathbb{P}(\tau_{nj}, y_{nj}). \quad (50.58)$$

Die *Reliabilität der Testmessung j* ist dann definiert als

$$R_j := \rho(y_{ij}, \tau_{ij})^2 \text{ für ein beliebiges } 1 \leq i \leq n. \quad (50.59)$$

•

Vor dem Hintergrund des vereinfachten Modells der Klassischen Testtheorie Definition 50.3 schreibt man auch

$$R = \rho(y, \tau)^2. \quad (50.60)$$

Per Definition ist die Reliabilität einer Testmessung also die quadrierte Korrelation von beobachtetem und wahren Wert. Mit

$$-1 \leq \rho(y_{ij}, \tau_{ij}) \leq 1 \quad (50.61)$$

folgt direkt, dass für die Reliabilität einer Testmessung gilt, dass

$$0 \leq R_j \leq 1 \quad (50.62)$$

Die Aussage $R_j = 0$ impliziert dann $\rho(y_{ij}, \tau_{ij}) = 0$, also die linear-affine Unabhängigkeit von beobachtetem und wahren Wert, und damit den Umstand, dass der beobachtete Wert hinsichtlich des wahren Werts nicht aussagekräftig ist. $R_j = 1$ dagegen impliziert $\rho(y_{ij}, \tau_{ij}) = \pm 1$, also die deterministische linear-affine Abhängigkeit von beobachtetem und wahren Wert und damit, dass der beobachtete Wert hinsichtlich des wahren Werts vollständig aussagekräftig ist.

Folgendes Theorem zeigt weitere Möglichkeiten auf, die Reliabilität einer Testmessung äquivalent zu formulieren und damit zu interpretieren.

Theorem 50.5 (Eigenschaften der Reliabilität einer Testmessung). Gegeben sei das Modell multipler Testmessungen für eine beliebige Testmessung j mit $1 \leq j \leq m$,

$$\mathbb{P}(\tau_{1j}, y_{1j}, \dots, \tau_{nj}, y_{nj}) := \prod_{i=1}^n \mathbb{P}(y_{ij} | \tau_{ij}) \mathbb{P}(\tau_{ij}). \quad (50.63)$$

Dann gelten für die Reliabilität R_j der Testmessung j , dass

- (1) $R_j = \frac{\mathbb{V}(\tau_{ij})}{\mathbb{V}(y_{ij})}$
- (2) $R_j = 1 - \frac{\mathbb{V}(\varepsilon_{ij})}{\mathbb{V}(y_{ij})}$

◦

Beweis. (1) Mit Aussage (5) von Theorem 50.1 gilt

$$R_j = \rho(y_{ij}, \tau_{ij})^2 = \left(\frac{\mathbb{C}(y_{ij}, \tau_{ij})}{\mathbb{S}(y_{ij})\mathbb{S}(\tau_{ij})} \right)^2 = \frac{\mathbb{C}(y_{ij}, \tau_{ij})^2}{\mathbb{V}(y_{ij})\mathbb{V}(\tau_{ij})} = \frac{\mathbb{V}(\tau_{ij})^2}{\mathbb{V}(y_{ij})\mathbb{V}(\tau_{ij})} = \frac{\mathbb{V}(\tau_{ij})}{\mathbb{V}(y_{ij})}. \quad (50.64)$$

(2) Mit Aussage (4) von Theorem 50.1 gilt dann weiter

$$R_j = \frac{\mathbb{V}(\tau_{ij})}{\mathbb{V}(y_{ij})} = \frac{\mathbb{V}(y_{ij}) - \mathbb{V}(\varepsilon_{ij})}{\mathbb{V}(y_{ij})} = \frac{\mathbb{V}(y_{ij})}{\mathbb{V}(y_{ij})} - \frac{\mathbb{V}(\varepsilon_{ij})}{\mathbb{V}(y_{ij})} = 1 - \frac{\mathbb{V}(\varepsilon_{ij})}{\mathbb{V}(y_{ij})}. \quad (50.65)$$

□

Da τ_{ij} und ε_{ij} weiterhin latent und damit nur indirekt beobachtbar sind, sind die Darstellungen nach Theorem 50.5 nur von theoretischem Interesse. Die erste Aussage besagt dabei insbesondere, dass die Reliabilität einer Testmessung der Anteil der Varianz der wahren Werte an der Varianz der beobachteten Werte ist,

$$R_j = \frac{\mathbb{V}(\tau_{ij})}{\mathbb{V}(y_{ij})} = \frac{\mathbb{V}(\tau_{ij})}{\mathbb{V}(\tau_{ij}) + \mathbb{V}(\varepsilon_{ij})}. \quad (50.66)$$

Gilt dabei, dass $\mathbb{V}(\tau_{ij}) = 0$, so ist auch $R_j = 0$, ist dagegen $\mathbb{V}(\varepsilon_{ij}) = 0$ so ist $R_j = 1$. Eine Reliabilität $R_j > 0$ impliziert also immer eine von Null verschiedene Varianz der wahren Werte.

50.2.2. Paralleltestreliabilität

Um das Konzept der Reliabilität nun anhand empirischer Daten von beobachteten Werten schätzbar zu machen, bedarf es seiner Übertragung in das Modell paralleler Testmessungen. Wir definieren daher zunächst explizit die Reliabilität einer Paralleltestmessung.

Definition 50.6 (Reliabilität einer Paralleltestmessung). Gegeben sei das Modell paralleler Testmessungen für eine beliebige Testmessung j mit $1 \leq j \leq m$,

$$\mathbb{P}(\tau_1, y_{1j}, \dots, \tau_n, y_{nj}) := \prod_{i=1}^n \mathbb{P}(\tau_i, y_{ij}) = \prod_{i=1}^n \mathbb{P}(\tau_i) \mathbb{P}(y_{ij} | \tau_i) \quad (50.67)$$

wobei nach Definition des Modells gilt, dass

$$\mathbb{P}(\tau_1, y_{1j}) = \dots = \mathbb{P}(\tau_n, y_{nj}). \quad (50.68)$$

Die *Reliabilität der Paralleltestmessung* j ist dann definiert als

$$R_j := \rho(y_{ij}, \tau_i)^2 \text{ für ein beliebiges } 1 \leq i \leq n. \quad (50.69)$$

•

Vor dem Hintergrund des vereinfachten Modells der Klassischen Testtheorie (Definition 50.3) schreibt man auch hier

$$R = \rho(y, \tau)^2. \quad (50.70)$$

Praktisch bedeutsam ist nun folgendes Theorem.

Theorem 50.6 (Paralleltestreliabilität). *Gegeben sei das Modell paralleler Testmessungen*

$$\mathbb{P}(\tau_1, y_{1j}, \dots, \tau_n, y_{nj}) := \prod_{i=1}^n \mathbb{P}(\tau_i, y_{ij}) = \prod_{i=1}^n \mathbb{P}(\tau_i) \mathbb{P}(y_{ij} | \tau_i) \quad (50.71)$$

Dann gelten

- (1) $R_j = \frac{\mathbb{V}(\tau_i)}{\mathbb{V}(y_{ij})}$ für alle $1 \leq j \leq m$.
- (2) $R_j = 1 - \frac{\mathbb{V}(\varepsilon_{ij})}{\mathbb{V}(y_{ij})}$ für alle $1 \leq j \leq m$.
- (3) $R_j = \rho(y_{ij}, y_{ik}) = R_k$ für alle $1 \leq j, k \leq m$.

◦

Beweis. (1) Mit Aussage (5) von Theorem 50.1 gilt

$$R_j = \rho(y_{ij}, \tau_i)^2 = \left(\frac{\mathbb{C}(y_{ij}, \tau_i)}{\mathbb{S}(y_{ij})\mathbb{S}(\tau_i)} \right)^2 = \frac{\mathbb{C}(y_{ij}, \tau_i)^2}{\mathbb{V}(y_{ij})\mathbb{V}(\tau_i)} = \frac{\mathbb{V}(\tau_i)^2}{\mathbb{V}(y_{ij})\mathbb{V}(\tau_i)} = \frac{\mathbb{V}(\tau_i)}{\mathbb{V}(y_{ij})}. \quad (50.72)$$

(2) Mit Aussage (4) von Theorem 50.1 gilt dann weiter

$$R_j = \frac{\mathbb{V}(\tau_i)}{\mathbb{V}(y_{ij})} = \frac{\mathbb{V}(y_{ij}) - \mathbb{V}(\varepsilon_{ij})}{\mathbb{V}(y_{ij})} = \frac{\mathbb{V}(y_{ij})}{\mathbb{V}(y_{ij})} - \frac{\mathbb{V}(\varepsilon_{ij})}{\mathbb{V}(y_{ij})} = 1 - \frac{\mathbb{V}(\varepsilon_{ij})}{\mathbb{V}(y_{ij})}. \quad (50.73)$$

(3) Mit Aussagen (2) und (5) von Theorem 50.3 gilt dann weiter

$$R_j = \frac{\mathbb{V}(\tau_i)}{\mathbb{V}(y_{ij})} = \frac{\mathbb{C}(y_{ij}, y_{ik})}{\sqrt{\mathbb{V}(y_{ij})}\sqrt{\mathbb{V}(y_{ik})}} = \frac{\mathbb{C}(y_{ij}, y_{ik})}{\sqrt{\mathbb{V}(y_{ij})}\sqrt{\mathbb{V}(y_{ik})}} = \rho(y_{ij}, y_{ik}) \quad (50.74)$$

und dass ebenso

$$R_k = \frac{\mathbb{V}(\tau_i)}{\mathbb{V}(y_{ik})} = \frac{\mathbb{C}(y_{ij}, y_{ik})}{\sqrt{\mathbb{V}(y_{ik})}\sqrt{\mathbb{V}(y_{ik})}} = \frac{\mathbb{C}(y_{ij}, y_{ik})}{\sqrt{\mathbb{V}(y_{ij})}\sqrt{\mathbb{V}(y_{ik})}} = \rho(y_{ij}, y_{ik}). \quad (50.75)$$

□

Aussagen (1) und (2) von Theorem 50.6 sind analog zu den Aussagen von Theorem 50.5. Aussage (3) von Theorem 50.6 begründet das praktische Vorgehen zur Reliabilitätsschätzung mithilfe von Parallel- oder Retestverfahren. Zur Bestimmung der Reliabilität eines Tests nutzt man entsprechend eine Schätzung der Korrelation zweier paralleler Testmessungen und die Klassische Testtheorie begründet dieses Vorgehen vor der Annahme von latenten wahren Werten und Messfehlern. Vereinfacht wird Aussage (3) von Theorem 50.6 oft dadurch ausgedrückt, dass man sagt, dass “die Korrelation paralleler Testmessungen gleich ihrer Reliabilität ist”.

Beispiel

Wir betrachten den Fall zweier Testmessungen $j = 1, 2$ im Modell paralleler Testmessungen. Für alle $i = 1, \dots, n$ seien, wobei wir der notationellen Einfachheit halber auf das i Subskript verzichten wollen,

$$p(t) = N(t; 0, \sigma_\tau^2) \text{ und } p(\tilde{y}_1 | t) := N(\tilde{y}_1; t, \sigma_\varepsilon^2) \text{ und } p(\tilde{y}_2 | t) := N(\tilde{y}_2; t, \sigma_\varepsilon^2) \quad (50.76)$$

Für Person i gibt es also nur eine Zufallsvariablen für den wahren Wert für alle Testmessungen und die Propensitätsverteilungen unterscheiden sich zwischen Testmessungen nicht. Dann

gilt zunächst mit dem Theorem zu gemeinsamen Normalverteilungen (Theorem 29.5) mit $A := 1$ und $b := 0$

$$p(t)p(\tilde{y}_1|t) = p(t, \tilde{y}_1) = N \left(\begin{pmatrix} t \\ \tilde{y}_1 \end{pmatrix}; \begin{pmatrix} \mu_\tau \\ \mu_\tau \end{pmatrix}, \begin{pmatrix} \sigma_\tau^2 & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma_\varepsilon^2 \end{pmatrix} \right) \quad (50.77)$$

und weiterhin mit $A := \begin{pmatrix} 1 & 0 \end{pmatrix}$ und $b := 0$

$$p(t, \tilde{y}_1)p(\tilde{y}_2|t) = p(t, \tilde{y}_1, \tilde{y}_2) = N \left(\begin{pmatrix} t \\ \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix}; \begin{pmatrix} \mu_\tau \\ \mu_\tau \\ \mu_\tau \end{pmatrix}, \begin{pmatrix} \sigma_\tau^2 & \sigma_\tau^2 & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma_\varepsilon^2 & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 & \sigma_\tau^2 + \sigma_\varepsilon^2 \end{pmatrix} \right) \quad (50.78)$$

Damit gilt dann durch Ablesen am Kovarianzmatrixparameter von $p(t, \tilde{y}_1, \tilde{y}_2)$, dass

$$\rho(y_1, y_2) = \frac{\mathbb{C}(y_1, y_2)}{\sqrt{\mathbb{V}(y_1)}\sqrt{\mathbb{V}(y_2)}} = \frac{\sigma_\tau^2}{\sqrt{\sigma_\tau^2 + \sigma_\varepsilon^2}\sqrt{\sigma_\tau^2 + \sigma_\varepsilon^2}} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}. \quad (50.79)$$

Insbesondere gilt also auch für $j = 1, 2$

$$R_j = \rho(y_j, \tau)^2 = \left(\frac{\mathbb{C}(y_j, \tau)}{\mathbb{S}(y_j)\mathbb{S}(\tau)} \right)^2 = \frac{\mathbb{C}(y_j, \tau)^2}{\mathbb{V}(y_j)\mathbb{V}(\tau)} = \frac{(\sigma_\tau^2)^2}{(\sigma_\tau^2 + \sigma_\varepsilon^2)\sigma_\tau^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2} = \rho(y_1, y_2). \quad (50.80)$$

Gilt also beispielsweise $\sigma_\tau^2 := 1.0$ und $\sigma_\varepsilon^2 := 0.2$, so ergibt sich für die Paralleltestreliabilität

$$R_j = \rho(y_j, \tau)^2 = \rho(y_1, y_2) = \frac{1.0}{1.0 + 0.2} \approx 0.83. \quad (50.81)$$

50.2.3. Schätzung der Paralleltestreliabilität

Wie üblich wird die Korrelation zweier paralleler Testmessungen in der Anwendung durch eine Stichprobenkorrelation geschätzt. Wir formulieren dies für das vorliegende Szenario mithilfe folgender Definition.

Definition 50.7 (Paralleltestreliabilitätsschätzer). Gegeben sei das Modell paralleler Testmessungen für n Personen und zwei Testmessungen $j = 1, 2$,

$$\mathbb{P}(\tau_1, y_{11}, y_{12}, \dots, \tau_n, y_{n1}, y_{n2}) = \prod_{i=1}^n \mathbb{P}(\tau_i) \mathbb{P}(y_{i1}|\tau_i) \mathbb{P}(y_{i2}|\tau_i). \quad (50.82)$$

Dann wird der mit den Stichprobenmitteln

$$\bar{y}_1 := \frac{1}{n} \sum_{i=1}^n y_{i1} \text{ und } \bar{y}_2 := \frac{1}{n} \sum_{i=1}^n y_{i2} \quad (50.83)$$

definierte Stichprobenkorrelationskoeffizient

$$r_{12} := \frac{\frac{1}{n-1} \sum_{i=1}^n (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_{i1} - \bar{y}_1)^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_{i2} - \bar{y}_2)^2}} \quad (50.84)$$

Paralleltestreliabilitätsschätzer genannt.

•

Beispiel

Bekanntlich bietet **R** mit der `cor()` Funktion eine Möglichkeit, Stichprobenkorrelationskoeffizienten zu berechnen. Wir demonstrieren dies an einem Simulationsbeispiel. Dazu seien

$$p(t_i) = N(t_i; 0, \sigma_\tau^2) \text{ und } p(y_{ij}|t_i) := N(y_{ij}; t_i, \sigma_\varepsilon^2) \quad (50.85)$$

für $j = 1, 2$ mit $\sigma_\tau^2 := 1.0$ und $\sigma_\varepsilon^2 := 0.2$ für $i = 1, \dots, 30$. Die wahre, aber unbekannte, Paralleltestreliabilität ergibt sich hier zu

$$R_{12} = \frac{1.0}{1.0 + 0.2} \approx 0.833. \quad (50.86)$$

Anhand von $n = 30$ simulierten Datenpunkten wird diese in folgender Simulation als $r_{12} = 0.857$ geschätzt.

```
set.seed(1)
n      = 30                                # Personenanzahl
m      = 2                                # Testmessungsanzahl
sigsqr_tau = 1                             # Wahrer Wert Varianz
sigsqr_eps = .2                            # Beobachteter Wert Varianz
R_12    = sigsqr_tau/(sigsqr_tau+sigsqr_eps) # Paralleltestreliabilität
T       = matrix(rep(NaN, n), nrow = n)    # Wahrer Wert Array
Y       = matrix(rep(NaN, n*m), nrow = n)   # Beobachteter Wert Array
E       = matrix(rep(NaN, n*m), nrow = n)   # Messfehler Array
for(i in 1:n){
  T[i] = rnorm(1,1,sqrt(sigsqr_tau))         # Personeniterationen
  for(j in 1:m){
    Y[i,j] = rnorm(1,T[i],sqrt(sigsqr_eps))  # Wahrer Wert Realisierung für j = 1,2
    E[i,j] = Y[i,j] - T[i]}                 # Testmessungsiterationen
  }                                           # Beobachteter Wert Realisierung
r_12    = cor(Y[,1],Y[,2])                  # Messfehler Realisierung
cat("Paralleltestreliabilität R_12          : ", round(R_12,4), # Paralleltestreliabilitätsschätzer
    "\nParalleltestrelibabilitätsschätzer r_12 : ", round(r_12,4)) # Ausgabe w.a.u. Paralleltestreliabilität
# Ausgabe geschätzte Paralleltestreliabilität
```

```
Paralleltestreliabilität R_12          : 0.8333
Paralleltestrelibabilitätsschätzer r_12 : 0.8569
```

Um neben einem Punktschätzer auch ein Konfidenzintervall für die Paralleltestreliabilität anzugeben, ist eine Annahme zur empirischen Verteilung des Reliabilitätsschätzers erforderlich. Dazu nutzt man üblicherweise folgende Aussage zur asymptotischen Verteilung einer Stichprobenkorrelation.

Theorem 50.7 (Approximative Verteilung der Fishertransformation einer Stichprobenkorrelation). *Gegeben sei eine Stichprobe $(y_{11}, y_{12}), \dots, (y_{n1}, y_{n2}) \sim \mathbb{P}(y_1, y_2)$ von Beobachtungen zweier Zufallsvariablen y_1 und y_2 mit Korrelation $\rho := \rho(y_1, y_2)$ und Stichprobenkorrelation r . Weiterhin bezeichne*

$$\tilde{r} := \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \quad (50.87)$$

die Fishertransformation von r . Dann ist $\tilde{r}$ asymptotisch normalverteilt mit

$$\tilde{r} \stackrel{a}{\sim} N \left(\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right), (n-3)^{-1} \right). \quad (50.88)$$

◦

Wir verzichten auf einen Beweis von Theorem 50.7 und verweisen für eine ausführliche Darstellung auf Kapitel 32 in Johnson et al. (1994). Theorem 50.7 bildet die Grundlage zur Konstruktion des in folgendem Theorem angegebenen approximativen Konfidenzintervalls für eine Korrelation.

Theorem 50.8 (Approximatives Konfidenzintervall einer Korrelation). *Gegeben sei eine Stichprobe $(y_{11}, y_{12}), \dots, (y_{n1}, y_{n2}) \sim \mathbb{P}(y_1, y_2)$ von Beobachtungen zweier Zufallsvariablen y_1 und y_2 mit Korrelation ρ , Stichprobenkorrelation r und Fishertransformation $\tilde{r}$. Ferner sei $\delta \in]0, 1[$ ein Konfidenzlevel und es sei*

$$z_\delta := \phi^{-1} \left(\frac{1 + \delta}{2} \right) \quad (50.89)$$

mit der inversen KVF ϕ^{-1} einer standardnormalverteilten Zufallsvariable. Schließlich seien

$$\tilde{r}_u := \tilde{r} - z_\delta (\sqrt{n-3})^{-1} \quad \text{und} \quad \tilde{r}_o := \tilde{r} + z_\delta (\sqrt{n-3})^{-1}. \quad (50.90)$$

Dann gilt für $n \rightarrow \infty$ für das Intervall

$$\kappa(r) := \left[\frac{\exp(2\tilde{r}_u) - 1}{\exp(2\tilde{r}_u) + 1}, \frac{\exp(2\tilde{r}_o) - 1}{\exp(2\tilde{r}_o) + 1} \right], \quad (50.91)$$

dass

$$\mathbb{P}_\rho(\kappa(r) \ni \rho) = \delta. \quad (50.92)$$

◦

Beweis. Mit Theorem 50.7 gilt durch Z -Transformation von r zunächst für $n \rightarrow \infty$, dass

$$\tilde{r}_z := \left(\tilde{r} - \frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right) \right) \sqrt{(n-3)} \stackrel{a}{\sim} N(0, 1) \quad (50.93)$$

und damit asymptotisch, dass

$$\mathbb{P}(-z_\delta \leq \tilde{r}_z \leq z_\delta) = \delta. \quad (50.94)$$

Also gilt asymptotisch auch, dass

$$\begin{aligned} \delta &= \mathbb{P}(-z_\delta \leq \tilde{r}_z \leq z_\delta) \\ &= \mathbb{P}\left(-z_\delta \leq \left(\tilde{r} - \frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right)\right) \sqrt{(n-3)} \leq z_\delta\right) \\ &= \mathbb{P}\left(-z_\delta (\sqrt{(n-3)})^{-1} \leq \tilde{r} - \frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right) \leq z_\delta (\sqrt{(n-3)})^{-1}\right) \\ &= \mathbb{P}\left(-z_\delta (\sqrt{(n-3)})^{-1} - \tilde{r} \leq -\frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right) \leq z_\delta (\sqrt{(n-3)})^{-1} - \tilde{r}\right) \\ &= \mathbb{P}\left(\tilde{r} + z_\delta (\sqrt{(n-3)})^{-1} \geq \frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right) \geq \tilde{r} - z_\delta (\sqrt{(n-3)})^{-1}\right) \\ &= \mathbb{P}\left(2 \left(\tilde{r} + z_\delta (\sqrt{(n-3)})^{-1}\right) \geq \ln \left(\frac{1 + \rho}{1 - \rho} \right) \geq 2 \left(\tilde{r} - z_\delta (\sqrt{(n-3)})^{-1}\right)\right) \\ &= \mathbb{P}\left(2 \left(\tilde{r} - z_\delta (\sqrt{(n-3)})^{-1}\right) \leq \ln \left(\frac{1 + \rho}{1 - \rho} \right) \leq 2 \left(z_\delta (\sqrt{(n-3)})^{-1} + \tilde{r}\right)\right) \\ &= \mathbb{P}\left(2\tilde{r}_u \leq \ln \left(\frac{1 + \rho}{1 - \rho} \right) \leq 2\tilde{r}_o\right) \\ &= \mathbb{P}\left(\exp(2\tilde{r}_u) \leq \frac{1 + \rho}{1 - \rho} \leq \exp(2\tilde{r}_o)\right) \\ &= \mathbb{P}(\exp(2\tilde{r}_u)(1 - \rho) \leq 1 + \rho \leq \exp(2\tilde{r}_o)(1 - \rho)) \\ &= \mathbb{P}(\exp(2\tilde{r}_u) - \exp(2\tilde{r}_u)\rho \leq 1 + \rho \leq \exp(2\tilde{r}_o) - \exp(2\tilde{r}_o)\rho) \\ &= \mathbb{P}(\exp(2\tilde{r}_u) - \exp(2\tilde{r}_u)\rho - 1 \leq \rho \leq \exp(2\tilde{r}_o) - \exp(2\tilde{r}_o)\rho - 1). \end{aligned} \quad (50.95)$$

Weiterhin gilt aber

$$\begin{aligned} \exp(2\tilde{r}_u) - \exp(2\tilde{r}_u)\rho - 1 &\leq \rho \\ &\Leftrightarrow \exp(2\tilde{r}_u) - 1 \leq \exp(2\tilde{r}_u)\rho + \rho \\ &\Leftrightarrow \exp(2\tilde{r}_u) - 1 \leq (\exp(2\tilde{r}_u) + 1)\rho \\ &\Leftrightarrow \frac{\exp(2\tilde{r}_u) - 1}{\exp(2\tilde{r}_u) + 1} \leq \rho, \end{aligned} \quad (50.96)$$

also

$$\delta = \mathbb{P} \left(\frac{\exp(2\tilde{r}_u) - 1}{\exp(2\tilde{r}_u) + 1} \leq \rho \leq \exp(2\tilde{r}_o) - \exp(2\tilde{r}_o)\rho - 1 \right). \quad (50.97)$$

Analog gilt

$$\begin{aligned} \rho &\leq \exp(2\tilde{r}_o) - \exp(2\tilde{r}_o)\rho - 1 \\ &\Leftrightarrow \exp(2\tilde{r}_o)\rho + \rho \leq \exp(2\tilde{r}_o) \\ &\Leftrightarrow (\exp(2\tilde{r}_o) + 1)\rho \leq \exp(2\tilde{r}_o) - 1 \\ &\Leftrightarrow \rho \leq \frac{\exp(2\tilde{r}_o) - 1}{\exp(2\tilde{r}_o) + 1} \end{aligned} \quad (50.98)$$

und damit schließlich

$$\begin{aligned} \delta &= \mathbb{P} \left(\frac{\exp(2\tilde{r}_u) - 1}{\exp(2\tilde{r}_u) + 1} \leq \rho \leq \frac{\exp(2\tilde{r}_o) - 1}{\exp(2\tilde{r}_o) + 1} \right) \\ &= \mathbb{P} \left(\left[\frac{\exp(2\tilde{r}_u) - 1}{\exp(2\tilde{r}_u) + 1}, \frac{\exp(2\tilde{r}_o) - 1}{\exp(2\tilde{r}_o) + 1} \right] \ni \rho \right) \\ &= \mathbb{P}_\rho (\kappa(r) \ni \rho). \end{aligned} \quad (50.99)$$

□

Beispiel

Wir demonstrieren Theorem 50.7 und Theorem 50.8 mithilfe eines Simulationsbeispiels. Dazu seien wie oben für $i = 1, \dots, 30$

$$p(t_i) := N(t_i; 0, \sigma_\tau^2) \text{ und } p(y_{ij}|t_i) := N(y_{ij}; t_i, \sigma_\varepsilon^2) \quad (50.100)$$

für $j = 1, 2$ mit $\sigma_\tau^2 := 1.0$ und $\sigma_\varepsilon^2 := 0.2$. Folgender **R** Code generiert 10^5 Datensätze dieses Modells und evaluiert die entsprechenden Fisher-transformierten Stichprobenkorrelationen und Konfidenzintervalle für ein Konfidenzlevel von $\delta = 0.95$. Abbildung 50.4 A zeigt den Vergleich zwischen der so empirisch generierten Frequentistischen Verteilung der Fisher-transformierten Stichprobenkorrelationen und Abbildung 50.4 B zeigt die ersten 100 simulierten Konfidenzintervalle der Korrelation. In der vorliegenden Simulation zeigt sich im Vergleich zur ihrer analytischen Approximation empirisch eine leichte Verschiebung der Fisher-transformierten Stichprobenkorrelationen zu höheren Werten. Die betrachteten Konfidenzintervalle überdecken die wahre, aber unbekannte, Korrelation in zwei von 100 Fällen nicht.

```
set.seed(0)
nsim      = 1e5
n         = 30
m         = 2
sigsqr_tau = 1
sigsqr_eps = .2
R         = sigsqr_tau/(sigsqr_tau+sigsqr_eps)
delta     = 0.95
z_delta   = qnorm((1+delta)/2)
r         = rep(NA, nsim)
r_til     = rep(NA, nsim)
r_til_u   = rep(NA, nsim)
r_til_o   = rep(NA, nsim)
kappa     = matrix(rep(NA, 2*nsim), ncol= 2)
for(s in 1:nsim){
  T      = matrix(rep(NA, n), nrow = n)
  Y      = matrix(rep(NA, n*m), nrow = n)
  E      = matrix(rep(NA, n*m), nrow = n)
  for(i in 1:n){
    T[i] = rnorm(1,1,sqrt(sigsqr_tau))
    for(j in 1:m){
      Y[i,j] = rnorm(1,T[i],sqrt(sigsqr_eps))
      E[i,j] = Y[i,j] - T[i]}
    r[s] = cor(Y[,1],Y[,2])
    r_til[s] = 1/2*log((1+r[s])/(1-r[s]))
    r_til_u[s] = r_til[s] - z_delta*(1/sqrt(n-3))
  }
  kappa[s,1] = R
  kappa[s,2] = delta
}
```

```

r_til_o[s] = r_til[s] + z_delta*(1/sqrt(n-3)) # oberer Konfidenzintervallparameter
kappa[s,1] = (exp(2*r_til_u[s])-1)/(exp(2*r_til_u[s])+1) # untere Konfidenzintervallgrenze
kappa[s,2] = (exp(2*r_til_o[s])-1)/(exp(2*r_til_o[s])+1) # obere Konfidenzintervallgrenze
}

```

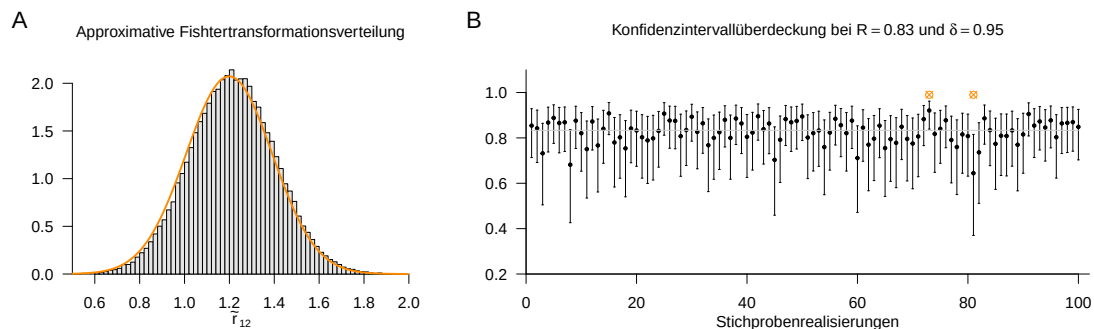


Abbildung 50.4. A Simulation der Verteilung der Fisher-transformierten Stichprobenkorrelation und ihre analytische Approximation B Simulation der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für die Stichprobenkorrelation bei einer wahren, aber unbekannten, Korrelation von $R = 0.833$ und einer gewünschten Überdeckungswahrscheinlichkeit von $\delta := 0.95$. Die Abbildung zeigt für jede Stichprobenrealisierung das Konfidenzintervall und die entsprechende Stichprobenkorrelation. In der vorliegenden Simulation überdecken die Konfidenzintervalle, die durch eine graue Linie eingezeichnete immer gleichen wahren, aber unbekannten, Korrelation $R := 0.833$ in 98 von 100 Fällen. Die Stichprobenrealisierungen, für die dies nicht der Fall ist, sind mit einem orangen Kreis markiert.

50.3. Interne Konsistenz

50.3.1. m -Komponententestmodelle

In den vorherigen Abschnitten bezeichnete y_{ij} die Zufallsvariable zur Modellierung des beobachteten Werts der j ten Testmessung der i ten Person, wobei wir das Vorliegen von m Testmessungen von n Personen angenommen haben. Wir haben dabei aber zunächst offen gelassen, ob mit der j ten Testmessung das summative Ergebnis eines Tests oder das Ergebnis ein einzelnes Testitems gemeint ist, um die Theorie für beide Anwendungsfälle zu entwickeln. Als Grundlage der Bestimmung der internen Konsistenz eines Tests identifizieren wir in diesem Abschnitt nun die j te Testmessung mit dem j ten Item eines Tests. y_{ij} bezeichnet also im Folgenden die Zufallsvariable zur Modellierung des Beobachteten Werts des j ten Items der i ten Person in einem Test, wobei wir weiterhin m Items und n Personen annehmen. Weiterhin gehen wir in diesem Zusammenhang davon aus, dass für jede Person ein *Gesamt-Beobachteter-Wert* durch Summation über die Items eines Test gebildet wird. Die Zufallsvariable zur Modellierung dieses Gesamt-Beobachteten-Werts bezeichnen wir mit

$$y_i := \sum_{j=1}^m y_{ij}. \quad (50.101)$$

Wir fassen diese Vorüberlegungen in folgender Definition des m -Komponententestmodells zusammen.

Definition 50.8 (m -Komponententestmodell). Gegeben sei das Modell multipler Testmessungen für $i = 1, \dots, n$ Personen und $j = 1, \dots, m$ Testmessungen. Für $i = 1, \dots, n$ seien

- $y_i := \sum_{j=1}^m y_{ij}$ die Summe der Beobachteten Werte und
- $\tau_i := \sum_{j=1}^m \tau_{ij}$ die Summe der Wahren Werte.

Dann heißt die gemeinsame Verteilung der y_i und τ_i für $i = 1, \dots, n$ mit der Faktorisierungseigenschaft

$$\mathbb{P}(\tau_1, \dots, \tau_n, y_1, \dots, y_n) := \prod_{i=1}^n \mathbb{P}(\tau_i, y_i) = \prod_{i=1}^n \mathbb{P}(y_i | \tau_i) \mathbb{P}(\tau_i) \quad (50.102)$$

das m -Komponententestmodell, wenn gilt dass

$$\mathbb{P}(\tau_1, y_1) = \dots = \mathbb{P}(\tau_n, y_n). \quad (50.103)$$

•

Zum Verständnis dazu, wie Cronbach's α die interne Konsistenz eines Tests misst, bietet es sich zunächst an, für das m -Komponententestmodell die Reliabilität im Sinne der Formulierung von Aussage (1) in Theorem 50.5 zu definieren.

Definition 50.9 (Reliabilität von m -Komponententestmodellen). Gegeben sei ein m -Komponententestmodell mit der Summe der Beobachteten Werte y_i und der Summe der Wahren Werte τ_i für $i = 1, \dots, n$ Personen. Dann ist die Reliabilität des Modells definiert als

$$R := \frac{\mathbb{V}(\tau_i)}{\mathbb{V}(y_i)} \text{ für ein beliebiges } 1 \leq i \leq n. \quad (50.104)$$

•

50.3.2. Cronbach's α

Vor dem Hintergrund des m -Komponentenmodells definieren wir *Cronbach's α* wie folgt.

Definition 50.10 (Cronbach's α). Gegeben sei ein m -Komponententestmodell. Dann heißt

$$\alpha := \frac{m}{m-1} \left(1 - \frac{\sum_{j=1}^m \mathbb{V}(y_{ij})}{\mathbb{V}(y_i)} \right) \quad (50.105)$$

Cronbach's α oder *Koeffizient α* .

•

Man beachte, dass in Definition 50.10 $\mathbb{V}(y_{ij})$ die Varianzen der Beobachteten Werte für Personen und Items und $\mathbb{V}(y_i)$ die Varianzen der Summen der Beobachteten Werte für Personen bezeichnen. Da nach Annahme des m -Komponententestmodells die Verteilungen der y_i identisch sind, wird Cronbach's α auch häufig in der Form

$$\alpha = \frac{m}{m-1} \left(1 - \frac{\sum_{j=1}^m \mathbb{V}(y_j)}{\mathbb{V}(y)} \right) \quad (50.106)$$

geschrieben, wobei $\mathbb{V}(y_j)$ die Varianzen der Items und $\mathbb{V}(y)$ die Varianz der Summen der Beobachten Werte bezeichnen.

Intuitiv nimmt Cronbach's α einen Wert nahe 1 an, wenn die Anzahl der Items m hoch und damit $\frac{m}{m-1} \approx 1$ ist und gleichzeitig das Verhältnis der summierten Itemvarianzen $\sum_{j=1}^m \mathbb{V}(y_j)$ zur Varianz der Summe der Itemwerte $\mathbb{V}(y)$ sehr gering und damit $\sum_{j=1}^m \mathbb{V}(y_j)/\mathbb{V}(y) \approx 0$ ist. Seine zentrale Bedeutung in der Klassischen Testtheorie bekommt Cronbach's α durch seinen Zusammenhang mit der Reliabilität eines m -Komponententests, welchen wir in unserem Theorem darstellen.

Theorem 50.9 (Cronbach's α und Reliabilität). *Gegeben sei ein m -Komponententestmodell mit Reliabilität R . Dann gilt für Cronbach's α , dass*

$$\alpha \leq R \quad (50.107)$$

und Gleichheit tritt insbesondere dann ein, wenn die Testmessungen des m -Komponententestmodell parallel sind.

◦

Beweis. Zur Vereinfachung der Notation verzichten wir auf die explizite Auszeichnung der Personen $i = 1, \dots, n$ und setzen

$$y_j := y_{ij}, \tau_j := \tau_{ij}, y := \sum_{j=1}^m y_j \text{ und } \tau := \sum_{j=1}^m \tau_j, \quad (50.108)$$

sowie

$$R := \frac{\mathbb{V}(\tau)}{\mathbb{V}(y)} \text{ und } \alpha := \frac{m}{m-1} \left(1 - \frac{\sum_{j=1}^m \mathbb{V}(y_j)}{\mathbb{V}(y)} \right). \quad (50.109)$$

Wir gehen in vier Schritten vor.

(1) (*Summendarstellung*) Wir halten zunächst folgende Darstellung der Summe von m Zahlen fest: für Zahlen $x_1, \dots, x_m$ gilt, dass

$$\sum_{j=1}^m x_j = \frac{1}{m-1} \sum_{j=1}^{m-1} \sum_{k=j+1}^m (x_j + x_k). \quad (50.110)$$

Anstelle eines Beweises betrachten wir den Fall $m := 4$. Dann gilt

$$\begin{aligned} \frac{1}{3} \sum_{j=1}^3 \sum_{k=j+1}^4 (x_j + x_k) &= \frac{1}{3} \left(\sum_{k=1+1}^4 (x_1 + x_k) + \sum_{k=2+1}^4 (x_2 + x_k) + \sum_{k=3+1}^4 (x_3 + x_k) \right) \\ &= \frac{1}{3} \left(\sum_{k=2}^4 (x_1 + x_k) + \sum_{k=3}^4 (x_2 + x_k) + \sum_{k=4}^4 (x_3 + x_k) \right) \\ &= \frac{1}{3} ((x_1 + x_2) + (x_1 + x_3) + (x_1 + x_4) + (x_2 + x_3) + (x_2 + x_4) + (x_3 + x_4)) \\ &= \frac{1}{3} ((x_1 + x_1 + x_1) + (x_2 + x_2 + x_2) + (x_3 + x_3 + x_3) + (x_4 + x_4 + x_4)) \\ &= \frac{1}{3} (3x_1 + 3x_2 + 3x_3 + 3x_4) \\ &= x_1 + x_2 + x_3 + x_4 \\ &= \sum_{j=1}^4 x_j \end{aligned} \quad (50.111)$$

(2) (*True-Score-Kovarianzungleichung*) Wir leiten nun im Modell multipler Testmessungen eine Ungleichung her. Dazu betrachten wir in diesem Modell die Varianz der Differenz zweier wahrer Werte τ_j und τ_k . Mit der Nicht-Negativität der Varianz (Theorem 24.4) und dem Theorem zur Varianz spezieller Linearkombinationen von Zufallsvariablen (Theorem 25.7) ergibt sich

$$\mathbb{V}(\tau_j - \tau_k) \geq 0 \Leftrightarrow \mathbb{V}(\tau_j) + \mathbb{V}(\tau_k) - 2\mathbb{C}(\tau_j, \tau_k) \geq 0 \Leftrightarrow \mathbb{V}(\tau_j) + \mathbb{V}(\tau_k) \geq 2\mathbb{C}(\tau_j, \tau_k) \quad (50.112)$$

Weiterhin ergibt sich für beliebige $1 \leq j, k \leq m$ bei parallelen Testmessungen, dass

$$\mathbb{V}(\tau_j - \tau_k) = \mathbb{V}(f_j(\tau_1) - f_k(\tau_1)) = \mathbb{V}(\tau_1 - \tau_1) = \mathbb{V}(0) = 0. \quad (50.113)$$

Im Fall paralleler Testmessungen ergibt sich in obiger Ungleichung und ihrer Anwendung im Folgenden also Gleichheit.

(3) (*Summen-True-Score-Varianzungleichung*) Wir betrachten nun die Varianz der True-Score Summe im m -Komponententestmodell. Mit dem Theorem zur Varianz einer Linearkombination von Zufallsvariablen (Theorem 25.6), der Summendarstellung aus (1) und der True-Score-Kovarianzungleichung aus (2) ergibt sich zunächst

$$\begin{aligned} \mathbb{V}(\tau) &= \mathbb{V}\left(\sum_{j=1}^m \tau_j\right) \\ &= \sum_{j=1}^m \mathbb{V}(\tau_j) + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^m \mathbb{C}(\tau_j, \tau_k) \\ &= \frac{1}{m-1} \sum_{j=1}^{m-1} \sum_{k=j+1}^m (\mathbb{V}(\tau_j) + \mathbb{V}(\tau_k)) + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^m \mathbb{C}(\tau_j, \tau_k) \\ &\geq \frac{1}{m-1} \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\mathbb{C}(\tau_j, \tau_k) + \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\mathbb{C}(\tau_j, \tau_k) \\ &= \left(\frac{1}{m-1} + 1\right) \sum_{j=1}^{m-1} \sum_{k=j+1}^m \mathbb{C}(\tau_j, \tau_k) \\ &= \left(\frac{1}{m-1} + \frac{m-1}{m-1}\right) \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\mathbb{C}(\tau_j, \tau_k) \\ &= \frac{1+m-1}{m-1} \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\mathbb{C}(\tau_j, \tau_k) \\ &= \frac{m}{m-1} \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\mathbb{C}(\tau_j, \tau_k) \end{aligned} \quad (50.114)$$

Mit Aussage (5) von Theorem 50.3 und wiederum mit Theorem 25.6 gilt dann

$$\begin{aligned} \mathbb{V}(\tau) &\geq \frac{m}{m-1} \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\mathbb{C}(\tau_j, \tau_k) \\ &= \frac{m}{m-1} \sum_{j=1}^{m-1} \sum_{k=j+1}^m 2\mathbb{C}(y_j, y_k) \\ &= \frac{m}{m-1} \left(\mathbb{V}\left(\sum_{j=1}^m y_j\right) - \sum_{j=1}^m \mathbb{V}(y_j) \right) \\ &= \frac{m}{m-1} \left(\mathbb{V}(y) - \sum_{j=1}^m \mathbb{V}(y_j) \right). \end{aligned} \quad (50.115)$$

(4) (*Reliabilität*) Wir betrachten schließlich die Reliabilität im m -Komponententestmodell. Es ergibt sich

$$\begin{aligned} R &= \frac{\mathbb{V}(\tau)}{\mathbb{V}(y)} \\ &\geq \frac{\frac{m}{m-1} (\mathbb{V}(y) - \sum_{j=1}^m \mathbb{V}(y_j))}{\mathbb{V}(y)} \\ &= \frac{m}{m-1} \left(\frac{\mathbb{V}(y)}{\mathbb{V}(y)} - \frac{\sum_{j=1}^m \mathbb{V}(y_j)}{\mathbb{V}(y)} \right) \\ &= \frac{m}{m-1} \left(1 - \frac{\sum_{j=1}^m \mathbb{V}(y_j)}{\mathbb{V}(y)} \right) \\ &=: \alpha. \end{aligned} \quad (50.116)$$

□

Cronbach's α ist also eine untere Grenze für die Reliabilität eines m -Komponententestmodells, die Reliabilität eines m -Komponententestmodells also ist mindestens so groß wie Cronbach's α , kann aber größer sein. Für parallele Testmessungen, also parallele Items, ist die Reliabilität eines m -Komponententestmodells sogar gleich α .

50.3.3. Schätzung von Cronbach's α

Die Schätzung von Cronbach's α greift auf die Stichprobenvarianzen der beobachteten Itemscores und der beobachteten Testsummenscores zurück. Für Daten von Personen $i = 1, \dots, n$ ist ein Schätzer für die Varianz des j ten Itemscores durch

$$S_j^2 := \frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2 \quad \text{mit} \quad \bar{y}_j := \frac{1}{n} \sum_{i=1}^n y_{ij} \quad (50.117)$$

und ein Schätzer für die Varianz der Testsummenscores durch

$$S^2 := \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{mit} \quad \bar{y} := \frac{1}{n} \sum_{i=1}^n y_i \quad (50.118)$$

gegeben. Ein Schätzer für α ergibt sich entsprechend zu

$$\hat{\alpha} = \frac{m}{m-1} \left(1 - \frac{\sum_{j=1}^m S_j^2}{S^2} \right). \quad (50.119)$$

Wir demonstrieren die Evaluation dieses Schätzers untenstehend anhand einer Simulation im Modell paralleler Testmessungen für $n := 30$ und $m := 21$ mit

$$\mathbb{P}(\tau_i) := N(1, 1) \quad \text{und} \quad \mathbb{P}(y_{ij}|\tau_i) := N(\tau_i, 4) \quad \text{für} \quad i = 1, \dots, n, j = 1, \dots, m. \quad (50.120)$$

```
library(psych)
set.seed(0)
n = 30
m = 21
mu = 1
T = matrix(rep(NaN, n), nrow = n)
Y = matrix(rep(NaN, n*m), nrow = n)
for(i in 1:n){
  T[i] = rnorm(2,mu,sqrt(1))
  for(j in 1:m){
    Y[i,j] = rnorm(1,T[i],sqrt(4))}
vsi = var(apply(Y,1,sum))
siv = sum(apply(Y,2,var))
a = (m/(m-1))*(1-(siv/vsi))
ap = alpha(Y,warnings = F)
```

R Paket zur Testanalyse
Reproduzierbarkeit
Personenanzahl
Itemanzahl
Wahrer Wert Erwartungswertparameter
Wahrer Wert Array
Beobachteter Wert Array
Iteration über Personen
Wahrer Wert Realisierung
Iteration über Items
Beobachteter Wert Realisierung
Stichprobenvarianz der Beobachten Wert Summen
Summe der Item-Stichprobenvarianzen
Direkte Berechnung von Cronbach's alpha
Berechnung von Cronbach's alpha mit psych

```
Cronbach's alpha (manuell) : 0.823
Cronbach's alpha (psych)  : 0.823
```

Die Frequentistische Verteilungs- und Inferenztheorie zu diesem Schätzer ist von Kristof (1963), Feldt (1965), Feldt et al. (1987), Van Zyl et al. (2000) und Yuan & Bentler (2002) ausgearbeitet worden.

51. Faktoranalyse

Mit dem Begriff *Faktoranalyse* werden verschiedene Inferenzverfahren bezeichnet, bei denen den zugrundeliegenden Modellen gemeinsam ist, dass sie mithilfe linear-affiner Transformationen nicht-beobachteter Zufallsvektoren die Kovarianzstruktur eines beobachtbaren Zufallsvektors generieren. Dabei sind in der psychologischen Anwendung Realisierungen des beobachtbaren Zufallsvektors typischerweise die Itemwerte einer Gruppe von Proband:innen in einem psychologischen Test. Ziel einer Faktoranalyse ist es dann, die geschätzte Item $\times$ Item Kovarianzmatrix mithilfe eines Faktoranalysemodells von möglichst einfacher Struktur zu erklären. Stilprägend sind in dieser Hinsicht die Arbeit von Spearman (1904), in der versucht wird, die Kovarianzstruktur verschiedener Leistungstestwerte mithilfe eines *generellen Intelligenzfaktors* zu erklären und Arbeiten zur Analyse des 16PF-Fragebogens (Cattell (1957)) im Sinne der *Big Five* Faktoren der Persönlichkeit. Im Kontext der Fragebogenentwicklung hat es sich eingebürgert, reproduzierbare Ergebnisse von Faktoranalysen als Ausdruck der *faktoriellen Validität* eines Fragebogens zu interpretieren.

Anwendungsbeispiel

Wir betrachten einen simulierten Datensatz nach den Ergebnissen einer Studie zur faktoriellen Validität der deutschen Version des BDI-II (Hautzinger et al. (2006), Keller et al. (2008)). Keller et al. (2008) betrachten dabei eine Stichprobe von $n = 266$ depressiven Patient:innen, die zum Zeitpunkt der Bearbeitung des BDI-II die Kriterien einer depressiven Störung nach ICD-10/DSM-IV erfüllten. Untenstehende Tabelle zeigt die BDI-II-Item-Scores der ersten 12 Patient:innen des simulierten Datensatzes, Abbildung 51.1 zeigt den gesamten Datensatz als kolorierte Datensatzmatrix. Abbildung 51.2 zeigt die Item $\times$ Item Kovarianz- und Korrelationsmatrizen des Datensatzes, die die primären Explananda einer faktoranalytischen Modellierung bilden.

Tabelle 51.1. BDI-II Item-Scores von 12 von 266 Patient:innen

	1	2	3	4	5	6	7	8	9	10	11	12
Traurigkeit	4	3	0	2	2	4	2	0	3	4	1	3
Pessimismus	2	2	1	2	3	3	3	0	3	2	2	1
Versagensgefühle	2	2	0	2	3	3	3	0	2	2	2	1
VerlustAnFreude	2	2	1	2	3	3	3	0	2	2	2	1
Schuldgefühle	3	3	0	2	2	3	1	1	3	2	0	4
Bestrafungsgefühle	3	3	0	2	3	3	1	1	2	3	0	4
Selbstablehnung	3	3	0	2	2	3	1	1	2	3	1	4
Selbstkritik	3	3	0	2	3	4	2	1	2	3	0	3
Suizidgedanken	3	3	1	2	2	3	1	2	2	3	0	4
Weinen	3	3	0	2	3	4	2	0	3	3	1	2
Unruhe	2	2	0	2	3	3	3	0	2	2	3	2
Interessenverlust	2	2	1	2	3	4	3	0	3	2	2	1
Entschlusslosigkeit	3	2	1	2	3	4	3	0	2	2	2	1
Wertlosigkeit	4	3	1	2	2	3	1	2	2	2	0	4
Energieverlust	2	2	1	2	3	3	3	0	3	2	2	1
Schlaf	2	2	1	2	3	3	3	0	2	3	2	1
Reizbarkeit	2	2	2	2	3	3	3	0	2	2	3	1
Appetitveränderung	2	2	1	2	3	3	3	0	2	2	2	1
Konzentrationsschwierigkeiten	2	2	1	2	3	3	3	0	2	2	2	1

	1	2	3	4	5	6	7	8	9	10	11	12
Ermüdung	3	2	1	2	3	3	3	0	2	2	3	1
VerlustSexuellenInteresses	3	2	1	2	3	3	3	0	2	2	2	1

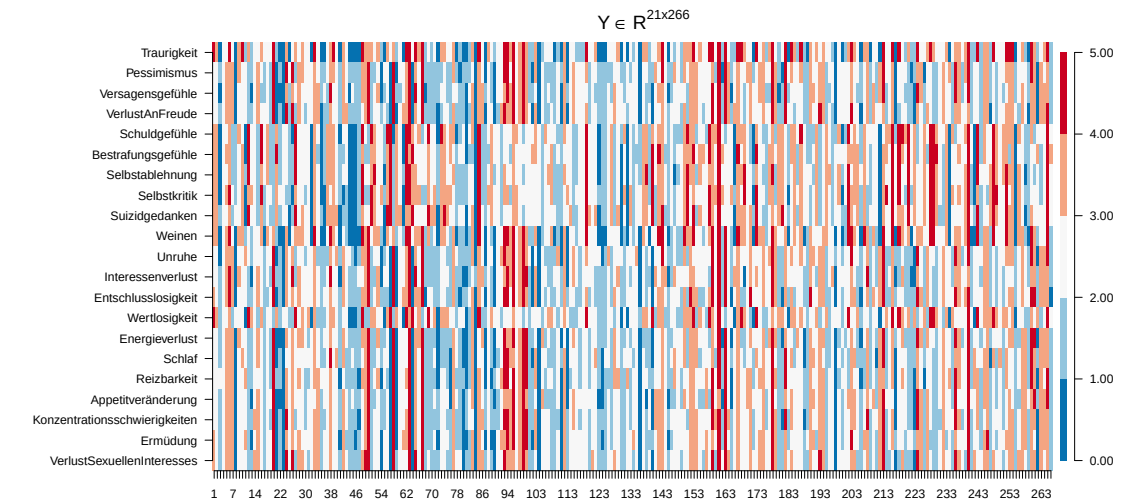


Abbildung 51.1. Simulierter BDI-II Datensatz nach Keller et al. (2008).

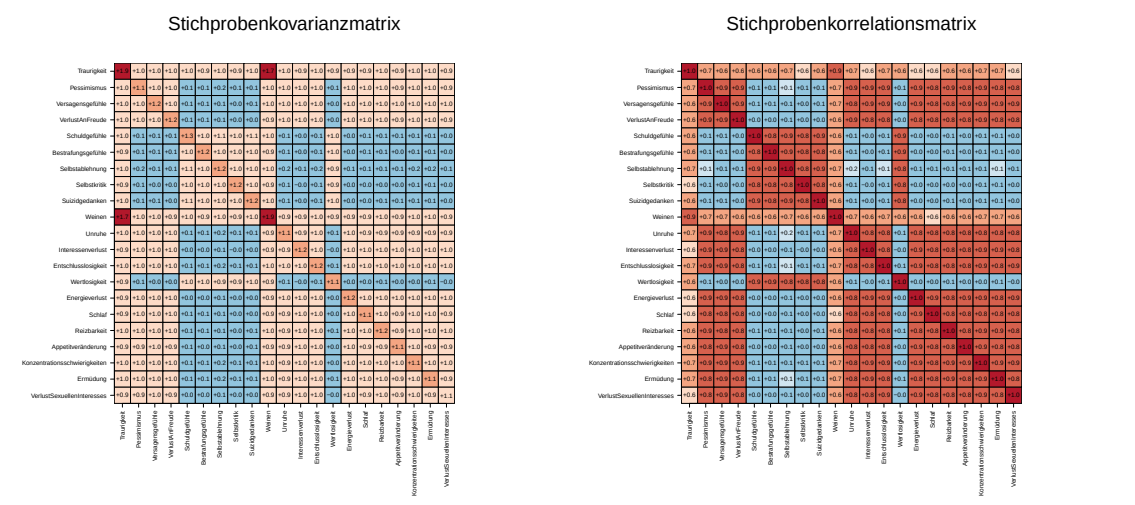


Abbildung 51.2. Stichprobenkovarianz- und Stichprobenkorrelationsmatrix des BDI-II Datensatz nach Keller et al. (2008).

51.1. Faktoranalysemodelle

Wir beginnen mit folgender Definition.

Definition 51.1 (Faktoranalysemodelle in struktureller Form). Es sei

$$y = Lf + \varepsilon \tag{51.1}$$

wobei für $m > k$

- y ein m -dimensionaler beobachtbarer Zufallsvektor ist, der *Daten* genannt wird, ist,
- $L = (l_{ij}) \in \mathbb{R}^{m \times k}$ eine Matrix ist, die *Faktorladungsmatrix* genannt wird,
- f ein k -dimensionaler nicht-beobachtbarer Zufallsvektor ist, dessen Komponenten *Faktoren* genannt werden und für den gilt, dass

$$\mathbb{E}(f) = 0_k \text{ und } \mathbb{C}(f) = I_k, \quad (51.2)$$

- ε ein m -dimensionaler nicht-beobachtbarer und von f unabhängiger Zufallsvektor ist, der *Beobachtungsfehler* genannt wird und für den gilt, dass

$$\mathbb{E}(\varepsilon) = 0_m \text{ und } \mathbb{C}(\varepsilon) = \text{diag}(\psi_1, \dots, \psi_m) =: \Psi \text{ mit } \psi_i > 0 \text{ für } i = 1, \dots, m. \quad (51.3)$$

Dann heißt Gleichung 51.1 *Faktoranalysemodell in struktureller Form*. Gelten darüberhinaus, dass

- $f \sim N(0_k, \Phi)$ mit $\Phi \in \mathbb{R}^{k \times k}$ p.d. und
- $\varepsilon \sim N(0_m, \Psi)$ mit $\Psi := \text{diag}(\psi_1, \dots, \psi_m)$ für $\psi_i > 0$ und $i = 1, \dots, m$,

dann heißt Gleichung 51.1 *Normalverteilungsmodell der Faktoranalyse in struktureller Form*.

•

Die von den Komponenten des Datenvektors y modellierten Daten beziehen sich in der Analyse von psychologischen Testdaten in der Regel auf die Itemscores eines Tests. Die von den Komponenten des Vektors f modellierten *Faktoren* werden manchmal auch *gemeinsame Faktoren* (*common factors*) genannt und gegenüber den Komponenten des Zufallsvektors ε abgegrenzt, die als *spezifische Faktoren* (*unique factors*) bezeichnet werden. Gemäß des Matrixproduktes Lf in (??) wird der Eintrag l_{ij} für $i = 1, \dots, m, j = 1, \dots, k$ der Faktorladungsmatrix L für $i = 1, \dots, m$ und $j = 1, \dots, k$ die *Faktorladung des j ten Faktors auf die i te Datenkomponente* genannt. Wir schreiben das Faktoranalysemodell in der Folge meist in der Kurzform

$$y = Lf + \varepsilon \text{ mit } f \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi) \quad (51.4)$$

Dabei soll die Notation $\zeta \sim (\mu, \Sigma)$ ausdrücken, dass $\mathbb{E}(\zeta) = \mu$ und $\mathbb{C}(\zeta) = \Sigma$ sind und f und ε sollen unabhängige Zufallsvektoren sein.

Verteilungsform des Faktoranalysemodells

In Verteilungsform kann das Faktoranalysemodell äquivalent geschrieben werden als

$$\mathbb{P}(f, y) = \mathbb{P}(f)\mathbb{P}(y|f) \text{ mit } f \sim (0_k, I_k) \text{ und } y|f \sim (Lf, \Psi). \quad (51.5)$$

Dabei erschließt sich die Bedingtheit der Verteilung von y auf f aus datengenerativer Sicht wie folgt: Es wird zunächst ein Wert von f anhand von $\mathbb{P}(f)$ realisiert, dieser Wert wird dann durch L transformiert und bildet den Erwartungswert der bedingten Verteilung $\mathbb{P}(y|f)$ von y gegeben f . Dann wird ein Wert von ε realisiert und schließlich werden die Werte von Lf und ε werden zu einem Wert von y addiert. Entsprechend gilt für das Normalverteilungsmodell der Faktoranalyse die Wahrscheinlichkeitsdichte

$$p(f, y) = p(f)p(y|f) \text{ mit } p(f) = N(0_k, \Phi) \text{ und } p(y|f) = N(Lf, \Psi). \quad (51.6)$$

Datensatzmodell der Faktoranalyse

Im datenanalytischen Kontext betrachtet man die vorliegenden Itemscores einer Proband:in j , die einen psychologischen Fragebogen oder Test bearbeitet hat, als Realisierung eines Proband:innen-spezifischen und anhand der Verteilung des Faktoranalysemodells verteilten Zufallsvektors y_j . Gleichzeitig unterstellt man, dass diesem Zufallsvektor eine Realisierung des nicht-beobachtbaren Zufallsvektors f_j , also der Proband:innen-spezifischen Faktorwerte, entspricht. Weiterhin nimmt man an, dass die Realisierungen von beobachtbaren Zufallsvektoren $y_1, \dots, y_n$ und ihren entsprechenden nicht-beobachtbaren Zufallsvektoren $f_1, \dots, f_n$ über Proband:innen unabhängig und identisch nach einem zugrundeliegenden proband:innenübergreifenden Faktoranalysemodell verteilt sind. Formal drückt man dies entweder als Faktorisierung der gemeinsamen Verteilung der beobachtbaren und nicht-beobachtbaren Zufallsvektoren in der Form

$$\mathbb{P}(f_1, y_1, \dots, f_n, y_n) = \prod_{j=1}^n \mathbb{P}(f_j, y_j) \text{ mit } \mathbb{P}(f_j, y_j) = \mathbb{P}(f_k, y_k) \text{ für } 1 \leq j, k \leq n \quad (51.7)$$

oder im Sinne unabhängig und identisch verteilter Zufallsvektoren in der Form

$$(f_1, y_1), \dots, (f_n, y_n) \sim \mathbb{P}(f, y) \quad (51.8)$$

aus. Entscheidend ist, dass man bei entsprechendem Vorliegen einer Datenmatrix $Y \in \mathbb{R}^{m \times n}$ wie in Abbildung 51.1 gezeigt, unterstellt, dass zu jedem Spalteneintrag ein virtueller, nicht-beobachteter, proband:innenspezifischer, Faktorenvektorwert korrespondiert. Inferenz bezüglich proband:innenspezifischer Faktorwerte, also die Modellschätzer-basierte Angabe der wahrscheinlichsten Werte des proband:innenspezifischen Faktorvektors, wird im Kontext der Faktoranalyse als *Evaluation von Factorscores* bezeichnet, steht in der Anwendung allerdings meist nicht im Vordergrund.

Beispiel (1) Spearmans g-Faktor Modell

Grundlage von Spearman's Intelligenzmodell ist die Korrelationsstruktur von Leistungswerten in $m = 6$ Themenbereichen (Classics, French, English, Mathematics, Pitch discrimination, Musical talent) in einer Stichprobe von ungefähr 30 Kindern im Alter von 9 und 13 Jahren (Yanai & Ichikawa (2007)). Spearman (1904) erklärt die entsprechenden Leistungswerte y_i mithilfe eines Faktoranalysemodells der Form

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \\ l_6 \end{pmatrix} f + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} \Leftrightarrow y_i = l_i f + \varepsilon_i \text{ für } i = 1, \dots, 6. \quad (51.9)$$

Dabei bezeichnet die latente Zufallsvariable f den *Allgemeinen Faktor der Intelligenz* und wird traditionell als *g-Faktor* bezeichnet. Für die Faktorladungen $l_i, i = 1, \dots, 6$ fordert Spearman (1904) $|l_i| < 1$. Die Komponenten ε_i werden im Kontext von Spearman (1904) als Themenbereichs-spezifische Faktoren bezeichnet. Entsprechend wird auch oft von *Spearman's Zweifaktoren Modell* gesprochen, wobei es im Sinne der modernen Faktoranalyse allerdings nur einen Faktor und einen Vektor von Zufallsfehlern in diesem Modell gibt.

Beispiel (2) Thurstones' Multifaktorenmodell

Thurstone (1947) schlägt mit dem Multifaktorenmodell ein generelles Modell von k Faktoren vor, um die Kovarianzstruktur eines m -dimensionalen Zufallsvektors, der m Leistungstestwerte einer Gruppe von Proband:innen modelliert, zu erklären. Dabei soll $k < m$ gelten. In struktureller Form hat dieses Modell die Form

$$y_i = l_{i1}f_1 + l_{i2}f_2 + \dots + l_{ik}f_k + \varepsilon_i \text{ für } i = 1, \dots, m \quad (51.10)$$

und entspricht damit der allgemeinen Form eines Faktoranalysemodells, wobei hier die Faktoren $f_1, \dots, f_k$ als *gemeinsame Faktoren* (*common factors*) und die Zufallsfehler $\varepsilon_1, \dots, \varepsilon_m$ als *spezifische Faktoren* (*unique factors*) bezeichnet werden. In Thurstone (1936) beispielsweise wird versucht zu zeigen, dass zur Erklärung der beobachteten Kovarianzstruktur von Leistungstestdaten von 240 Proband:innen sieben Faktoren (Memory, Word fluency, Verbal relation, Number, Perceptual Speed, Visualization, Reasoning), benötigt werden.

Beispiel (3) Das Modell paralleler Testmessungen

$y_1, \dots, y_m$ seien die beobachtbaren Itemwerte eines psychologischen Tests und sei das Modell paralleler Testmessungen für den wahren Wert τ in der Form

$$y_i = \tau + \varepsilon_i \text{ für } i = 1, \dots, m \quad (51.11)$$

mit Messfehler ε_i . Dann entspricht dieses Modell dem Faktoranalysemodell

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tau + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix} \quad (51.12)$$

Insbesondere entspricht der Faktor f hier also dem wahren Wert τ und die Faktorladungsmatrix hat die als bekannt vorausgesetzte Form $L := 1_m$.

Die zentrale Eigenschaft des Faktoranalysemodells ist die Beschaffenheit seiner marginalen Datenkovarianzmatrix, die wir im folgenden Theorem festhalten.

Theorem 51.1 (Marginale Datenkovarianzmatrix des Faktoranalysemodells). *Gegeben sei ein Faktoranalysemodell*

$$y = Lf + \varepsilon \text{ mit } f \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \quad (51.13)$$

Dann gilt für die marginale Kovarianzmatrix des Datenvektors

$$\mathbb{C}(y) = LL^T + \Psi. \quad (51.14)$$

◦

Beweis. Mit Theorem 25.9 gilt aufgrund der Unabhängigkeit von f und ε

$$\mathbb{C}(y) = L\mathbb{C}(f)L^T + \mathbb{C}(\varepsilon) = LI_kL^T + \Psi = LL^T + \Psi. \quad (51.15)$$

□

Die Diagonaleinträge der marginalen Datenkovarianzmatrix des Faktoranalysemodells lassen sich additiv durch die Einträge der Faktorladungsmatrix und der Kovarianzmatrix des Beobachtungsfehlers darstellen. Dies ist der Inhalt folgenden Theorems.

Theorem 51.2 (Varianzzerlegung der faktoranalytischen Datenkomponenten). *Gegeben sei ein Faktoranalysemodell*

$$y = Lf + \varepsilon \text{ mit } f \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_k, \Psi). \quad (51.16)$$

Dann ist für $i = 1, \dots, m$ die Varianz der i ten Komponente von y gegeben durch

$$\mathbb{V}(y_i) = \sum_{j=1}^k l_{ij}^2 + \psi_i. \quad (51.17)$$

◦

Beweis. Mit Theorem 51.1 gilt

$$\begin{aligned} \mathbb{C}(y) &= LL^T + \Psi \\ &= \begin{pmatrix} l_{11} & \dots & l_{1k} \\ l_{21} & \dots & l_{2k} \\ \vdots & \ddots & \vdots \\ l_{m1} & \dots & l_{mk} \end{pmatrix} \begin{pmatrix} l_{11} & \dots & l_{m1} \\ l_{12} & \dots & l_{m2} \\ \vdots & \ddots & \vdots \\ l_{1k} & \dots & l_{mk} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \psi_m \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^k l_{1j}l_{1j} & \sum_{j=1}^k l_{1j}l_{2j} & \dots & \sum_{j=1}^k l_{1j}l_{mj} \\ \sum_{j=1}^k l_{2j}l_{1j} & \sum_{j=1}^k l_{2j}l_{2j} & \dots & \sum_{j=1}^k l_{2j}l_{mj} \\ \vdots & \dots & \ddots & \vdots \\ \sum_{j=1}^k l_{mj}l_{1j} & \sum_{j=1}^k l_{mj}l_{2j} & \dots & \sum_{j=1}^k l_{mj}l_{mj} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \psi_m \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^k l_{1j}^2 + \psi_1 & \sum_{j=1}^k l_{1j}l_{2j} & \dots & \sum_{j=1}^k l_{1j}l_{mj} \\ \sum_{j=1}^k l_{2j}l_{1j} & \sum_{j=1}^k l_{2j}^2 + \psi_2 & \dots & \sum_{j=1}^k l_{2j}l_{mj} \\ \vdots & \dots & \ddots & \vdots \\ \sum_{j=1}^k l_{mj}l_{1j} & \sum_{j=1}^k l_{mj}l_{2j} & \dots & \sum_{j=1}^k l_{mj}^2 + \psi_m \end{pmatrix}. \end{aligned} \quad (51.18)$$

Für den i ten Diagonaleintrag von $\mathbb{C}(y)$ aber gilt dann bekanntlich $\mathbb{C}(y_i, y_i) = \mathbb{V}(y_i)$. □

Die Terme in der Varianzdarstellung der faktoranalytischen Datenkomponenten von Theorem 51.2 erhalten im Kontext der Faktoranalyse spezielle Bezeichnungen.

Definition 51.2 (Kommunalität und Spezifität). Gegeben sei ein Faktoranalysemodell

$$y = Lf + \varepsilon \text{ mit } f \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_k, \Psi). \quad (51.19)$$

Dann werden in

$$\mathbb{V}(y_i) = \sum_{j=1}^k l_{ij}^2 + \psi_i \quad (51.20)$$

- $h_i^2 := \sum_{j=1}^k l_{ij}^2$ die *Kommunalität* von y_i
- ψ_i die *Spezifität* von y_i

genannt.

•

Die *Kommunalität* einer Datenkomponente y_i ist also der Varianzanteil von y_i , der durch die Faktorladungen erklärt wird. Die *Spezifität* einer Datenkomponente y_i dagegen ist der Varianzanteil von y_i , der nicht durch die Faktorladungen erklärt wird und damit spezifisch für diese Datenkomponente ist. Mnemonisch gilt für jede Datenkomponente eines Faktoranalysemodells also

$$\text{Varianz} = \text{Kommunalität} + \text{Spezifität}. \quad (51.21)$$

Auch die Summe der Varianzen der Datenkomponenten eines Faktoranalysemodells erhält eine eigene Bezeichnung.

Definition 51.3 (Gesamtvarianz). Gegeben sei ein Faktoranalysemodell

$$y = Lf + \varepsilon \text{ mit } f \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \quad (51.22)$$

Dann wird

$$G := \sum_{i=1}^m \mathbb{V}(y_i) = \sum_{i=1}^m \sum_{j=1}^k l_{ij}^2 + \sum_{i=1}^m \psi_i \quad (51.23)$$

die *Gesamtvarianz von y* genannt.

•

Die Gesamtvarianz des beobachtbaren Datenvektors ist also definiert als die Summe der Varianzen der Datenkomponenten und entspricht damit der Spur der marginalen Datenkovarianzmatrix. Für die Gesamtvarianz gilt mnemonisch also

$$\text{Gesamtvarianz} = \text{Summe der Kommunalitäten} + \text{Summe der Spezifitäten}. \quad (51.24)$$

Wie wir später sehen werden kann die entsprechende Zerlegung der Gesamtstichprobenvarianz als Grundlage zur Evaluation der Modellgüte einer Faktoranalyse dienen.

Nichtidentifizierbarkeit

Eine fundamentale Eigenschaft des Faktoranalysemodells ist seine *Nichtidentifizierbarkeit*. Damit wird ausgedrückt, dass verschiedene Kombinationen von Faktorwerten und Faktorladungsmatrizen in der gleichen marginalen Datenkovarianzmatrix resultieren und somit anhand einer Datenkovarianzmatrix bzw. ihrem Stichprobenäquivalent nicht eindeutig identifiziert werden können. Um diese Eigenschaft formal zu beschreiben, definieren wir zunächst den Begriff der orthogonalen Transformation eines Faktoranalysemodells

Definition 51.4 (Orthogonale Transformation eines Faktoranalysemodells). Gegeben sei ein Faktoranalysemodell

$$y = Lf + \varepsilon \text{ mit } f \sim (0_k, I_k) \text{ mit } \varepsilon \sim (0_m, \Psi) \quad (51.25)$$

und es sei $Q \in \mathbb{R}^{k \times k}$ eine orthogonale Matrix. Dann nennen wir

$$\tilde{y} = \tilde{L}\tilde{f} + \varepsilon \text{ mit } \tilde{L} := LQ \text{ und } \tilde{f} := Q^T f \quad (51.26)$$

eine *orthogonale Transformation des Faktoranalysemodells*.

•

Die orthogonale Transformation eines Faktoranalysemodells lässt den Datenvektor und seine Kovarianzmatrix unberührt. Dies ist die Aussage folgenden Theorems.

Theorem 51.3 (Nichtidentifizierbarkeit und Kovarianzinvarianz der Faktoranalyse). *Gegeben sei ein Faktoranalysemodell*

$$y = Lf + \varepsilon \text{ mit } f \sim (0_k, I_k) \text{ mit } \varepsilon \sim (0_m, \Psi) \quad (51.27)$$

sowie eine seiner orthogonale Transformationen

$$\tilde{y} = \tilde{L}\tilde{f} + \varepsilon \text{ mit } \tilde{L} := LQ \text{ und } \tilde{f} := Q^T f \quad (51.28)$$

für eine orthogonale Matrix $Q \in \mathbb{R}^{k \times k}$. Dann gelten

$$y = \tilde{y} \text{ und } \mathbb{C}(\tilde{y}) = \mathbb{C}(y) \quad (51.29)$$

◦

Beweis. Es gilt zum einen

$$\tilde{y} = \tilde{L}\tilde{x} + \varepsilon = LQQ^T f + \varepsilon = LI_k f + \varepsilon = Lf + \varepsilon = y. \quad (51.30)$$

Es gilt zum anderen

$$\mathbb{C}(\tilde{y}) = LQ(LQ)^T + \Psi = LQQ^T L^T + \Psi = LI_k L^T + \Psi = LL^T + \Psi = \mathbb{C}(y). \quad (51.31)$$

□

Mit

$$y = Lf + \varepsilon = \tilde{L}\tilde{f} + \varepsilon \quad (51.32)$$

folgt also unmittelbar, dass für einen festen Wert von y die Faktorladungsmatrix L und der Faktorvektorwert von f nicht eindeutig bestimmt sind. Verschiedene Faktorladungsmatrizen und Faktorwerte können also die gleichen Daten erklären und aus einer gegebenen Stichprobenkovarianzmatrix kann nicht eindeutig auf L und f geschlossen werden. Weiterhin folgt mit

$$\mathbb{C}(y) = LL^T + \Psi = \tilde{L}\tilde{L}^T + \Psi \quad (51.33)$$

aber auch, dass sich die Gesamtvarianz und die Kommunalitäten bei orthogonaler Transformation nicht ändern. Fundamental ist die Nichtidentifizierbarkeit des Faktoranalysemodells dadurch bedingt, dass nach Annahme weder die Faktorladungsmatrix, noch die Werte der Faktoren, noch die Beobachtungsfehler bekannt sind und Daten daher durch das Zusammenbeispiel dreier unbekannter Entitäten generiert werden.

Identifizierbarkeitsbedingungen

Zumindest im Kontext des Normalverteilungsmodells der Faktoranalyse kann man versuchen, die Identifizierbarkeit der Parameter dieses Modells durch eine Reihe von Nebenbedingungen zu gewährleisten. Allerdings sind auch in diesem Modell allgemeine hinreichende und notwendige Bedingungen für die Modellidentifizierbarkeit nicht vollumfänglich bekannt, die Modellidentifizierbarkeit im Bereich der Faktoranalyse und dem eng verwandten Feld der Strukturgleichungsmodelle bleibt also in aktives Forschungsfeld (vgl. Hyvärinen et al. (2024)). In der Anwendung sind deshalb simulationsbasierte Parameterrecoverystudien

sicherlich eine gute Forschungspraxis. In diesem Abschnitt wollen wir die Identifizierbarkeit des Normalverteilungsmodells der Faktoranalyse genauer betrachten. Dazu definieren wir zunächst den Begriff der Identifizierbarkeit für das Normalverteilungsmodell der Faktoranalyse und führen dann mit der *Ordnungsbedingung* eine häufig verwendete Heuristik zur Modellidentifizierbarkeit ein, die die Intuition formalisiert, dass ein Modell im Sinne der datenanalytischen Datenreduktion nicht mehr Parameter haben sollte als Datenstatistiken betrachtet werden. Dazu zählen wir zunächst die Anzahl der unikal Parameter und der Statistiken des Normalverteilungsmodells der Faktoranalyse.

Definition 51.5 (Identifizierbares Normalverteilungsmodell der Faktoranalyse). Gegeben sei ein Normalverteilungsmodell der Faktoranalyse

$$y = Lf + \varepsilon \text{ mit } f \sim N(0_k, \Phi) \text{ und } \varepsilon \sim N(0_m, \Psi). \quad (51.34)$$

Weiterhin sei der *Parametervektor* dieses Modells definiert als die spaltenweise Konkatenation der Parametermatrizen L, Φ, Ψ , also

$$\theta := \text{vec}(L, \Phi, \Psi), \quad (51.35)$$

so dass die marginale Datenkovarianz geschrieben werden kann als

$$\Sigma_\theta := L\Phi L^T + \Psi. \quad (51.36)$$

Dann heißen das Normalverteilungsmodell der Faktoranalyse und der Parametervektor θ *identifizierbar*, wenn für alle $\theta_1, \theta_2 \in \Theta$ gilt, dass

$$\Sigma_{\theta_1} = \Sigma_{\theta_2} \Leftrightarrow \theta_1 = \theta_2. \quad (51.37)$$

Wenn für $\theta_1 \neq \theta_2$ gilt, dass $\Sigma_{\theta_1} = \Sigma_{\theta_2}$, so heißen das Modell und θ *nicht identifizierbar*.

•

Bei nicht identifizierbaren Modellen ergeben unterschiedliche Parameterwerte also die gleiche Datenverteilung und es kann folglich aus Daten nicht eindeutig auf Parameterwerte zurückgeschlossen werden. Allerdings gibt es bisher keine allgemein gültigen notwendigen und hinreichenden Bedingungen für die Identifizierbarkeit von Normalverteilungsmodell der Faktoranalysen. Die unten vorgestellte *Ordnungsbedingung* ist lediglich eine notwendige Bedingung für die Identifizierbarkeit. Um sie diskutieren zu können, zählen wir zunächst die Anzahl an einzelnen (unikalen) skalaren Parametern und Stichprobenstatistiken eines Normalverteilungsmodell der Faktoranalysen. Dabei ist zentral, dass die Symmetrie von Kovarianzmatrizen bedeutet, dass nur die Einträge über und inklusive ihrer Hauptdiagonale unikal sind.

Theorem 51.4 (Anzahl unikal Parameter und Statistiken). Gegeben sei ein Normalverteilungsmodell der Faktoranalyse

$$y = Lf + \varepsilon \text{ mit } f \sim N(0_k, \Phi) \text{ und } \varepsilon \sim N(0_m, \Psi). \quad (51.38)$$

ein konfirmatorisches Faktoranalysemodell. Dann gilt für die Anzahl an unikal Parameter des Modells

$$n_\theta = mk + \frac{k(k+1)}{2} + m \quad (51.39)$$

Weiterhin sei C sei die Stichprobenkovarianzmatrix eines Datensatzes von n unabhängigen Beobachtungen von y . Dann gilt für die Anzahl an unikalenen skalaren Stichprobenstatistiken des Normalverteilungsmodell der Faktoranalysen

$$n_c = \frac{m(m+1)}{2}. \quad (51.40)$$

◦

Beweis. Die Anzahl der Einträge der Faktorladungsmatrix $L \in \mathbb{R}^{m \times k}$ ist mk .

Die Anzahl der Einträge einer symmetrischen Matrix $S \in \mathbb{R}^{k \times k}$ ist k^2 . Aufgrund der Symmetrie von S sind dabei allerdings nur die k Einträge der Hauptdiagonale und die Einträge oberhalb (oder unterhalb) der Hauptdiagonalen unikal. Die Anzahl der Einträge oberhalb (oder unterhalb) der Hauptdiagonalen ist die Hälfte aller $k^2 - k$ nicht-diagonalen Einträge von S , also $(k^2 - k)/2$. Zusammen mit den Einträgen auf der Hauptdiagonalen ergibt sich damit für die Anzahl unikatler Einträge einer symmetrischen Matrix

$$\frac{k^2 - k}{2} + k = \frac{k^2 - k}{2} + \frac{2k}{2} = \frac{k^2 + k}{2} = \frac{k(k+1)}{2}. \quad (51.41)$$

Als Kovarianzmatrix ist $\Phi \in \mathbb{R}^{k \times k}$ symmetrisch und hat damit $\frac{k(k+1)}{2}$ unikale Einträge. Die Anzahl der von Null verschiedenen Einträge von $\Psi \in \mathbb{R}^{m \times m}$ ist m .

Für die Anzahl an unikalenen skalaren Parametern des Normalverteilungsmodell der Faktoranalysen ergibt sich also zusammenfassend

$$n_\theta = mk + \frac{k(k+1)}{2} + m. \quad (51.42)$$

Die Stichprobenkovarianzmatrix $C \in \mathbb{R}^{m \times m}$ eines Datensatzes ist symmetrisch. Mit obigen Überlegungen zu den unikalenen Einträgen einer symmetrischen Matrix ergibt sich also direkt

$$n_c = \frac{m(m+1)}{2}. \quad (51.43)$$

□

Nach Theorem 51.4 bezeichnet n_θ also die Dimension des unikalenen Parametervektors θ , es gilt also $\theta \in \mathbb{R}^{n_\theta}$ und n_c ist die Anzahl der unikalenen skalaren Einträge von C . Das Verhältnis von n_θ und n_c ist die Grundlage der Ordnungsbedingung zur Identifizierbarkeit von Normalverteilungsmodell der Faktoranalysen.

Definition 51.6 (Ordnungsbedingung). Gegeben sei ein Normalverteilungsmodell der Faktoranalyse

$$y = Lf + \varepsilon \text{ mit } f \sim N(0_k, \Phi) \text{ und } \varepsilon \sim N(0_m, \Psi) \quad (51.44)$$

Dann sagt man, dass das Modell der *Ordnungsbedingung* genügt, wenn für die Anzahl n_θ der unikalenen skalaren Parameter und für die Anzahl n_c der unikalenen skalaren Statistiken gilt, dass

$$n_\theta \leq n_c. \quad (51.45)$$

•

Die Ordnungsbedingung besagt also, dass die Anzahl der unbekannten und damit zu schätzenden Parameter des betrachteten Modells kleiner oder gleich der unikalenen Einträge der Stichprobenkovarianzmatrix ist. Softwarelösungen wie Lavaan implementieren die Ordnungsbedingung meist per default (vgl. Rosseel (2021)).

51.2. Schätzverfahren

Zur Schätzung der Parameter eines Faktoranalysemodells stehen verschiedene Schätzverfahren zur Verfügung die in unterschiedlichen Softwareumgebungen unterschiedlich implementiert sein können und so zu numerisch unterschiedlichen Schätzern für die Parameter des Faktoranalysemodells führen können. Gemeinsam ist den hier betrachteten Schätzverfahren, dass sie von einem Faktoranalysemodell fester Ordnung, d.h. mit einer prädefinierten Anzahl k von Faktoren ausgehen.

51.2.1. Hauptkomponentenschätzung

Mit der *Hauptkomponentenschätzung* wollen wir zunächst ein grundlegendes Verfahren zur Schätzung der Parameter eines Faktoranalysemodells diskutieren. Zentrale Motivation ist dabei die Approximation der Stichprobenkovarianzmatrix eines durch ein Faktoranalysemodell generierten Datensatzes anhand von Theorem 51.1 durch

$$C \approx \hat{L}\hat{L}^T + \hat{\Psi}. \quad (51.46)$$

Die Hauptkomponentenschätzung vernachlässigt dabei zunächst $\hat{\Psi}$ und nutzt die Orthonormalzerlegung

$$C = Q\Lambda Q^T = Q\Lambda^{1/2}\Lambda^{1/2}Q^T = (Q\Lambda^{1/2})(Q\Lambda^{1/2})^T \quad (51.47)$$

zur Darstellung von C . Die Hauptkomponentenschätzung vernachlässigt dann weiterhin die $k+1, \dots, m$ Spalten von Q sowie die $k+1, \dots, m$ Zeilen und Spalten von Λ und setzt

$$\hat{L}\hat{L}^T = Q_k\Lambda_k^{1/2}(Q_k\Lambda_k^{1/2})^T \text{ mit } \hat{L} \in \mathbb{R}^{m \times k}. \quad (51.48)$$

Für die Diagonalelemente c_{ii} , $\hat{h}_i^2$ und $\hat{\psi}_i$ von C , $\hat{L}\hat{L}^T$ bzw. $\hat{\Psi}$ folgt dann, dass

$$c_{ii} = \sum_{j=1}^k \hat{l}_{ij}^2 + \hat{\psi}_i \Leftrightarrow \hat{\psi}_i = c_{ii} - \sum_{j=1}^k \hat{l}_{ij}^2, \quad (51.49)$$

worauf sich die entsprechenden Spezifitätsschätzer bei Hauptkomponentenschätzung gründen. Wir fassen das skizzierte Vorgehen in folgenden Definitionen zusammen.

Definition 51.7 (Hauptkomponentenschätzer k ter Ordnung von L und Ψ). Gegeben sei ein Datensatz $Y \in \mathbb{R}^{m \times n}$ von n unabhängigen Beobachtungen eines Faktoranalysemodell

$$y = Lf + \varepsilon \text{ mit } f \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \quad (51.50)$$

$C \in \mathbb{R}^{m \times m}$ sei die Stichprobenkovarianzmatrix von Y und

$$C = Q\Lambda Q^T \quad (51.51)$$

sei ihre Orthonormalzerlegung mit spaltenweise der Größe nach sortierten Eigenwerten und zugehörigen Eigenvektoren. Dann sind die *Hauptkomponentenschätzer k ter Ordnung von L und Ψ* definiert als

$$\hat{L} := Q_k\Lambda_k^{1/2} \in \mathbb{R}^{m \times k} \text{ und } \hat{\Psi} := \text{diag}(\hat{\psi}_1, \dots, \hat{\psi}_m), \quad (51.52)$$

wobei $Q_k \in \mathbb{R}^{m \times k}$ die Matrix bestehend aus den ersten k Spalten von $Q \in \mathbb{R}^{m \times m}$, $\Lambda_k \in \mathbb{R}^{k \times k}$ die Matrix bestehend aus den ersten k Zeilen und k Spalten von $\Lambda \in \mathbb{R}^{m \times m}$ und für $i = 1, \dots, m$

$$\hat{\psi}_i := c_{ii} - \sum_{j=1}^k \hat{l}_{ij}^2 \quad (51.53)$$

mit den Diagonaleinträgen c_{ii} von C bezeichnen.

•

Die Selektion der ersten k Spalten von C und der ersten k Zeilen und Spalten von Λ in Definition 51.7 impliziert ein Faktoranalysemodell mit k Faktoren. Vor dem Hintergrund der Bezeichnungen von Definition 51.2 und Definition 51.3 ergeben sich auf Grundlage von Definition 51.7 folgende weitere Schätzer.

Definition 51.8 (Varianz-, Kommunalitäts-, Spezifitäts- und Gesamtvarianzschätzer). Für einen Datensatz $Y \in \mathbb{R}^{m \times n}$ von n unabhängigen Beobachtungen und ein festes $k < m$ seien $\hat{L} = (\hat{l}_{ij})_{1 \leq i \leq m, 1 \leq j \leq k} \in \mathbb{R}^{m \times k}$ und $\hat{\Psi} = \text{diag}(\hat{\psi}_1, \dots, \hat{\psi}_m) \in \mathbb{R}^{m \times m}$ die auf der Stichprobenkovarianzmatrix $C = (c_{ij})_{1 \leq i, j \leq m}$ basierenden Hauptkomponentenschätzer k ter Ordnung. Dann ergeben sich für $i = 1, \dots, m$,

- c_{ii} als Schätzer von $\mathbb{V}(y_i)$ (*Varianzschätzer*),
- $\hat{h}_i^2 := \sum_{j=1}^k \hat{l}_{ij}^2$ als Schätzer von h_i^2 (*Kommunalitätsschätzer*),
- $\hat{\psi}_i$ als Schätzer von ψ_i (*Spezifitätsschätzer*),
- $G := \text{tr}(C)$ als Schätzer von $\mathbb{G}$ (*Gesamtvarianzschätzer*).

•

Anwendungsbeispiel

Folgender **R** Code demonstriert die Hauptkomponentenschätzung für ein Faktoranalysemodell mit $k = 2$ anhand des in Abbildung 51.1 dargestellten Datensatzes nach Keller et al. (2008). Abbildung 51.3 zeigt die resultierenden Faktorladungsmatrix- und Spezifitätschätzer.

```
Y      = as.matrix(read.csv("./data/903-fa-keller.csv"))      # Y \in \mathbb{R}^{m \times n}
m      = nrow(Y)                                             # Datendimension
n      = ncol(Y)                                             # Datenpunkanzahl
k      = 2                                                  # Faktoranzahl
I_n    = diag(n)                                             # Einheitsmatrix I_n
J_n    = matrix(rep(1,n^2), nrow = n)                       # 1_{nn}
C      = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))        # Stichprobenkovarianzmatrix
EA     = eigen(C)                                           # Eigenanalyse von R
lambda_k = EA$values[1:k]                                   # k größte Eigenwerte von R
Q_k    = EA$vectors[,1:k]                                   # k zugehörige Eigenvektoren von R
L_hat  = Q_k %*% diag(sqrt(lambda_k))                      # Faktorladungsmatrixschätzer
Psi_hat = diag(diag(C) - diag(L_hat %*% t(L_hat)))         # Beobachtungsauschenkovarianzmatrixschätzer
V_i_hat = diag(C)                                           # Varianzschätzer
h2_i_hat = rowSums(L_hat^2)                                  # Kommunalitätsschätzer
psi_i_hat = diag(Psi_hat)                                   # Spezifitätsschätzer
```


•

Im Initialisierungsschritt (S0) ist $\varepsilon_{\min}$ eine Schranke für das Konvergenzkriterium, wie weiter unten erläutert, $i_{\max}$ ist die maximale Anzahl an Iterationen, die durchgeführt werden sollen, $R^{(0)}$ ist die iterierte Korrelationsmatrix, die mit der experimentell beobachteten Korrelationsmatrix R initialisiert wird, $h^{(0)}$ ist der Anfangswert der Summe der Kommunalitätsschätzungen, also die Summe der Diagonalelemente von $R^{(0)}$, $\varepsilon^{(1)}$ ist das Konvergenzkriterium und i ist der Iterationszähler. Die Iterationen werden solange durchgeführt, bis entweder das Konvergenzkriterium $\varepsilon^{(i)}$ einen Wert annimmt, der kleiner oder gleich der Schranke $\varepsilon_{\min}$ ist, oder mehr Iterationen als die maximale Iterationsanzahl $i_{\max}$ durchgeführt wurden.

Im ersten Iterationsschritt (S1) wird die Eigenwertzerlegung der iterierten Korrelationsmatrix durchgeführt, und die Schätzung der Faktorladungsmatrix k -ter Ordnung wird basierend auf den k Eigenvektoren mit den k größten Eigenwerten der iterierten Korrelationsmatrix berechnet. Genauer gesagt sind $Q_k^{(i)} \in \mathbb{R}^{m \times k}$ und $\Lambda_k^{(i)} \in \mathbb{R}^{m \times k}$ die Matrizen, die aus den ersten k Spalten von $Q^{(i)} \in \mathbb{R}^{m \times m}$ bzw. $\Lambda^{(i)} \in \mathbb{R}^{m \times m}$ bestehen, wobei angenommen wird, dass die Spalten entsprechend der absteigenden Reihenfolge der Eigenwerte sortiert sind. Entscheidend ist, dass im zweiten Iterationsschritt (S2) die Diagonaleinträge der iterierten Korrelationsmatrix durch die aus der aktuellen Schätzung der Faktorladungsmatrix resultierenden Kommunalitätsschätzungen ersetzt werden. Dadurch entsteht eine aktualisierte, iterierte Korrelationsmatrix, die die Grundlage für die nächste Schätzung der Faktorladungsmatrix bildet. Im Iterationsschritt (S3) wird die Summe der Kommunalitätsschätzungen neu berechnet, das Konvergenzkriterium als absoluter Unterschied zwischen den Summen der aktuellen und vorherigen Iteration bestimmt, und der Iterationszähler wird erhöht. Der Algorithmus konvergiert also, wenn sich die Kommunalitätsschätzungen zwischen zwei aufeinanderfolgenden Iterationen nicht wesentlich ändern oder die maximale Anzahl erlaubter Iterationen erreicht wurde.

Abschließend werden angesichts der fundamentalen Unbestimmtheit der Parameter des Faktorenanalysemodells im Schritt (S4) die Vorzeichen der Einträge der geschätzten Faktorladungsmatrix so gewählt, dass die Spaltensummen der Faktorladungsmatrix größer oder gleich null sind. Zu diesem Zweck wird die bei Konvergenz erhaltene Faktorladungsmatrix $\hat{L}^{(i)}$ von rechts mit einer Diagonalmatrix multipliziert, deren Diagonalelemente den Werten der Signumfunktion sgn der Spaltensummen von $\hat{L}^{(i)}$ entsprechen, wobei $\text{sgn}(x) = +1$ für $x \geq 0$ und $\text{sgn}(x) = -1$ für $x < 0$ gilt.

Anwendungsbeispiel

Folgender **R** Code demonstriert die iterierte Hauptachsenfaktorisation zur Schätzung der Parameter eines Faktorenanalysemodells mit Faktorenanalysemodell mit $k = 2$ anhand des in Abbildung 51.1 dargestellten Datensatzes nach Keller et al. (2008).

```
Y      = as.matrix(read.csv("./_data/903-fa-keller.csv"))      # Y \in \mathbb{R}^{m \times n}
R      = cor(Y)                                                # m x m Korrelationsmatrix
k      = 2                                                      # Faktorenzahl
eps_min = 0.01                                                  # Konvergenzkriterium
i_max  = 25                                                      # Maximale Anzahl an IPAF Iterationen
i      = 0                                                      # Iterationszähler
R_i    = R                                                      # Initiale Korrelationsmatrix R^{(0)}
h_i    = sum(diag(R_i))                                          # Summe der Kommunalitätsschätzer
eps_i  = h_i                                                    # Fehlerinitialisierung
while(eps_i > eps_min && i <= i_max){
  QLQ_i = eigen(R_i)                                            # Iterationen
  L_hat_i = QLQ_i$vectors[,1:k] %*% diag(sqrt(QLQ_i$values[1:k])) # Eigenanalyse
  diag(R_i) = diag(L_hat_i %*% t(L_hat_i))                    # Hauptkomponentenschätzer kter Ordnung
  h_ii    = sum(diag(R_i))                                     # Kommunalitätsschätzer Update
  eps_i   = abs(h_ii - h_i)                                    # Kommunalitätsschätzersummen Update
  h_i     = h_ii                                               # Fehlerkonvergenzkriterium
  h_i     = h_ii                                               # Kommunalitätsschätzersummen Update
  i       = i + 1
}
```

```

i      = i + 1}
D      = diag(sign(colSums(L_hat_i)))
L_hat  = L_hat_i %*% D
# Iterationszähler Update
# Vorzeichenfinalisierung
# IHAF Faktorladungsmatrixschätzer

```

51.2.3. Maximum-Likelihood-Schätzung

Im Normalverteilungsmodell der Faktoranalyse können die Parameter mithilfe des Maximum-Likelihood-Verfahrens geschätzt werden. Traditionell werden die Parameter dabei durch Minimierung der sogenannten *Diskrepanzfunktion* geschätzt (vgl. Lawley (1940), Jöreskog (1967) und Jöreskog (1969)). Die funktionale Form der Diskrepanzfunktion ist dabei durch ein Log-Likelihood-Kriterium bei Betrachtung der Frequentistischen Verteilung der Stichprobenkovarianz motiviert. Diese ist nach Wishart (1928) benannt. Modernere Sichtweisen im Kontext von Strukturgleichungsmodellen motivieren die funktionale Form der Diskrepanzfunktion direkt durch ein Log-Likelihood-Kriterium bei Betrachtung der multivariaten Datennormalverteilung (vgl. Bollen (1989) und Rosseel (2021)). Wir wollen hier diesen modernen Weg nachzeichnen und dabei auf eine Einführung der Wishart-Verteilung verzichten. Zu diesem Zweck nehmen wir durchgängig die *Zentrierung* des betrachteten Datensatzes $Y \in \mathbb{R}^{m \times n}$, also $\bar{y} = 0_m$, sowie die Identifizierbarkeit des Modells an. Für einen alternativen zeitgemäßen Zugang zur Parameterschätzung im Normalverteilungsmodell der Faktoranalyse mithilfe des Expectation-Maximization Algorithmus der variationalen Inferenz, siehe z.B. Rubin & Thayer (1982), Roweis & Ghahramani (1999) und Ghojogh et al. (2022). Die genauen Bezüge und qualitativen Eigenschaften der traditionellen und modernen Faktoranalyseschätzverfahren sind dabei eine offene Forschungsfrage.

Zur Diskussion des Maximum-Likelihood-Verfahrens im Normalverteilungsmodell der Faktoranalyse gehen wir wie folgt vor: Wir evaluieren zunächst die Log-Likelihood-Funktion des Normalverteilungsmodells der Faktoranalyse und definieren dann die funktionale Form der traditionellen Diskrepanzfunktion nach Lawley (1940). Wir zeigen dann, dass die Minimumstellen der Diskrepanzfunktion Maximum-Likelihood-Schätzer für das Normalverteilungsmodell der Faktoranalyse sind. Die Bestimmung von Minimumstellen der Diskrepanzfunktion mithilfe von Standardverfahren der nichtlinearen Optimierung (vgl. Rosseel (2012), Rosseel (2021)) ist also äquivalent zur Bestimmung von Maximumstellen der Log-Likelihood-Funktion des Normalverteilungsmodells der Faktoranalyse. Eine weitere Motivation für die Betrachtung der Diskrepanzfunktion ergibt sich im Kontext von Likelihood-Ration-Kriterium-basierten Modellvergleichen, wie in Kapitel 51.3.2 diskutiert.

Für die Log-Likelihood-Funktion des Normalverteilungsmodells der Faktoranalyse gilt zunächst folgendes Theorem.

Theorem 51.5 (Log-Likelihood-Funktion und Maximum-Likelihood-Schätzer des Normalverteilungsmodells der Faktoranalyse). *Gegeben sei ein Faktoranalysemodell der Form*

$$y = Lf + \varepsilon \text{ mit } f \sim N(0_k, \Phi) \text{ und } \varepsilon \sim N(0_m, \Psi) \quad (51.54)$$

mit Parametervektor

$$\theta := \text{vec}(L, \Phi, \Psi), \quad (51.55)$$

und marginaler Datenkovarianzmatrix

$$\Sigma_\theta := L\Phi L^T + \Psi. \quad (51.56)$$

Weiterhin sei $Y := (y_1, \dots, y_n) \in \mathbb{R}^{m \times n}$ ein zentrierter Datensatz von n unabhängigen Beobachtungen von y , C seine Stichprobenkovarianzmatrix und

$$S := \frac{n-1}{n}C \quad (51.57)$$

seine verzerrte Stichprobenkovarianzmatrix. Dann kann die Log-Likelihood-Funktion von Y geschrieben werden als

$$\ell_Y : \Theta \rightarrow \mathbb{R}, \theta \mapsto \ell_Y(\theta) := -\frac{n}{2} \ln |\Sigma_\theta| - \frac{n}{2} \text{tr}(S \Sigma_\theta^{-1}) - \frac{nm}{2} \ln(2\pi) \quad (51.58)$$

und eine Maximumstelle von ℓ_Y , also ein Wert

$$\hat{\theta}_{ML} = \operatorname{argmax}_{\theta \in \Theta} \ell_Y(\theta), \quad (51.59)$$

heißt Maximum-Likelihood-Schätzer für θ .

◦

Beweis. Mit der Definition der Log-Likelihood-Funktion als und der marginalen Datenverteilung des Normalverteilungsmodells der Faktorenanalyse ergibt sich

$$\begin{aligned} \ell_Y(\theta) &= \ln \prod_{i=1}^n p_\theta(y_i) \\ &= \prod_{i=1}^n \ln N(y_i; 0_m, L \Phi L^T + \Psi) \\ &= \ln \left(\prod_{i=1}^n (2\pi)^{-m/2} |\Sigma_\theta|^{-1/2} \exp \left(-\frac{1}{2} (y_i - 0_m)^T \Sigma_\theta^{-1} (y_i - 0_m) \right) \right) \\ &= -\frac{mn}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma_\theta| - \frac{1}{2} \sum_{i=1}^n y_i^T \Sigma_\theta^{-1} y_i. \end{aligned} \quad (51.60)$$

Um die Gleichheit des letzten Terms auf der rechten Seite mit dem letzten Term in der postulierten Funktionsform zu zeigen, halten wir zunächst fest, dass mit elementaren Eigenschaften der Matrixspur gilt, dass

$$\operatorname{tr} \left(\sum_{i=1}^n y_i^T \Sigma_\theta^{-1} y_i \right) = \sum_{i=1}^n \operatorname{tr} (y_i^T \Sigma_\theta^{-1} y_i) = \sum_{i=1}^n \operatorname{tr} (\Sigma_\theta^{-1} y_i y_i^T) = \operatorname{tr} \left(\Sigma_\theta^{-1} \sum_{i=1}^n y_i y_i^T \right). \quad (51.61)$$

Weiterhin gilt mit dem Binomischen Lehrsatz

$$\begin{aligned} \sum_{i=1}^n y_i^T \Sigma_\theta^{-1} y_i &= \sum_{i=1}^n (y_i - \bar{y} + \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y} + \bar{y}) \\ &= \sum_{i=1}^n (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^n \bar{y}^T \Sigma_\theta^{-1} \bar{y} + 2 \sum_{i=1}^n (y_i - \bar{y})^T \Sigma_\theta^{-1} \bar{y} \\ &= \sum_{i=1}^n (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^n \bar{y}^T \Sigma_\theta^{-1} \bar{y} + 2 \left(\sum_{i=1}^n \left(y_i - \frac{1}{n} \sum_{i=1}^n y_i \right)^T \right) \Sigma_\theta^{-1} \bar{y} \\ &= \sum_{i=1}^n (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^n \bar{y}^T \Sigma_\theta^{-1} \bar{y} + 2 \left(\sum_{i=1}^n y_i^T - \frac{n}{n} \sum_{i=1}^n y_i^T \right) \Sigma_\theta^{-1} \bar{y} \\ &= \sum_{i=1}^n (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^n \bar{y}^T \Sigma_\theta^{-1} \bar{y} + 2 (0_m^T \Sigma_\theta^{-1} \bar{y}) \\ &= \sum_{i=1}^n (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^n \bar{y}^T \Sigma_\theta^{-1} \bar{y} \end{aligned} \quad (51.62)$$

Also folgt mit obiger Eigenschaft der Matrixspur, der Zentrierung des Datensatzes $\bar{y} = 0_m$ und der Definition von S

$$\begin{aligned}
 \sum_{i=1}^n y_i^T \Sigma_{\theta}^{-1} y_i &= \sum_{i=1}^n (y_i - \bar{y})^T \Sigma_{\theta}^{-1} (y_i - \bar{y}) + \sum_{i=1}^n \bar{y}^T \Sigma_{\theta}^{-1} \bar{y} \\
 &= \sum_{i=1}^n y_i^T \Sigma_{\theta}^{-1} y_i \\
 &= \text{tr} \left(\sum_{i=1}^n y_i^T \Sigma_{\theta}^{-1} y_i \right) \\
 &= \text{tr} \left(\Sigma_{\theta}^{-1} \sum_{i=1}^n y_i y_i^T \right) \\
 &= \text{tr} (\Sigma_{\theta}^{-1} nS) \\
 &= n \text{tr} (S \Sigma_{\theta}^{-1})
 \end{aligned} \tag{51.63}$$

Substitution in $\ell_Y(\theta)$ von oben ergibt dann die postulierte funktionale Form der Log-Likelihood Funktion. $\square$

Die Diskrepanzfunktion der Faktoranalyse und basierend auf ihr definierte Parameterschätzer sind nun wie folgt definiert.

Definition 51.10 (Diskrepanzfunktion der Faktoranalyse). Gegeben sei ein Faktoranalysemodell der Form

$$y = Lf + \varepsilon \text{ mit } f \sim N(0_k, \Phi) \text{ und } \varepsilon \sim N(0_m, \Psi) \tag{51.64}$$

mit Parametervektor

$$\theta := \text{vec}(L, \Phi, \Psi), \tag{51.65}$$

und marginaler Datenkovarianzmatrix

$$\Sigma_{\theta} := L\Phi L^T + \Psi. \tag{51.66}$$

Weiterhin sei für einen Datensatz $Y \in \mathbb{R}^{m \times n}$ von n unabhängigen Beobachtungen von y S die verzerrte Stichprobenkovarianzmatrix von Y . Dann heißt die Funktion

$$F_Y : \Theta \rightarrow \mathbb{R}, \theta \mapsto F_Y(\theta) := n \ln |\Sigma_{\theta}| + n \text{tr} (S \Sigma_{\theta}^{-1}) - n \ln |S| - nm \tag{51.67}$$

die *Diskrepanzfunktion der Faktoranalyse*. Weiterhin heißt ein Wert $\hat{\theta} \in \Theta$ mit

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmin}} F_Y(\theta), \tag{51.68}$$

also eine Minimumstelle von F_Y , ein *diskrepanzfunktionsbasierte Faktoranalyseparameterschätzer*. $\rightarrow$

•

```

library(psych)
Y      = as.matrix(read.csv("./data/903-fa-keller.csv"))
M      = fa(r = Y, nfactors = 2, rotate = "none", fm = "ml")
L_hat  = M$loadings[,1:2]
# Y \in \mathbb{R}^{m \times n}
# Modellformulierung und -schätzung
# Faktorladungsmatrixschätzer

```

Der Vergleich mit Theorem 51.5 zeigt, dass die Log-Likelihood-Funktion und die Diskrepanzfunktion des Normalverteilungsmodells der Faktoranalyse nicht identisch sind. Allerdings kann man zeigen, dass Minimumstellen der Diskrepanzfunktion Maximumstellen der Log-Likelihood-Funktion sind und damit die Minimierung der Diskrepanzfunktion

Maximum-Likelihood-Schätzer für die Parameter des Faktoranalysenormalverteilungsmodells ergibt. Die Motivation dafür, die Diskrepanzfunktion in modernen Betrachtungen der Faktoranalyse beizubehalten und nicht zugunsten der Log-Likelihood-Funktion aufzugeben (vgl. Rosseel (2021)) erschließt sich dann vor dem Hintergrund der Modellevaluation.

Theorem 51.6 (Äquivalenz von diskrepanzfunktionsbasierten und Maximum-Likelihood-Schätzern). *Jede Minimumstelle der Diskrepanzfunktion der Faktoranalyse maximiert die Log-Likelihood-Funktion des Normalverteilungsmodells der Faktoranalyse.*

◦

Beweis. Wir halten zunächst fest, dass mit von θ unabhängigen Konstanten $a, b \in \mathbb{R}$ gilt, dass

$$\ell_Y(\theta) = a(-F_Y(\theta)) + b. \quad (51.69)$$

Die Log-Likelihood-Funktion ist also eine linear-affine und damit insbesondere monotone Transformation von F_Y . Da monotone Transformationen Extremstellen unverändert lassen und ein negatives Vorzeichen eine Minimumstelle in eine Maximumstelle transformiert, ergibt sich das Theorem direkt. $\square$

Minima von F_Y werden in populären Analyseprogrammen zur Faktorenanalyse mit iterativen Standardverfahren der nichtlinearen Optimierung bestimmt. Allgemein gehen diese Verfahren von einem Startwert $\hat{\theta}^{(0)}$ aus und evaluieren weitere Iteranden rekursiv durch

$$\hat{\theta}^{(i+1)} = g(\hat{\theta}^{(i)}, Y) \quad \text{für } i = 0, 1, \dots \quad (51.70)$$

mit einer entsprechend gewählten Funktion g des vorherigen Iteranden $\hat{\theta}^{(i)}$ und des Datensatzes Y solange, bis ein entsprechend gewähltes Abbruchkriterium erfüllt ist. Einen Überblick über die zum Beispiel im populären **R** Paket [lavaan](#) implementierten Optimierungsalgorithmen geben Rosseel (2012) und Rosseel (2021). Wir wollen die numerische Minimierung der Diskrepanzfunktion hier nicht weiter vertiefen.

Anwendungsbeispiel

Folgender **R** Code demonstriert die Maximum-Likelihood-Schätzung der Parameter eines Faktoranalysenormalverteilungsmodells mit Faktoranalysemodell mit $k = 2$ anhand des in Abbildung 51.1 dargestellten Datensatzes nach Keller et al. (2008) mithilfe des **psych** Paketes nach Revelle (2024). Abbildung 51.4 stellt die Ergebnisse von Hauptkomponentenschätzung, iterierter Hauptachsenfaktorisierung und Maximum-Likelihood-Schätzung der Faktorladungsmatrix nebeneinander dar. Es ergeben sich qualitativ ähnliche Ladungen der beiden Faktoren auf die Items des BDI-II, allerdings mit feineren quantitativen Unterschieden.

51.3. Modellevaluation

Grundlegende Fragen der Modellevaluation im Kontext der Faktoranalyse betreffen die Frage der Anzahl der zur Erklärung eines Datensatzes benötigten Faktoren, sowie die Überlegenheit eines spezifizierten Modells gegenüber einem parametrischen Nullmodell. Erstere Frage behandeln wir aus varianzanalytischer Sicht in Kapitel 51.3.1. Zweitere Frage behandeln wir aus Likelihood-Ratio-Sicht in Kapitel 51.3.2.

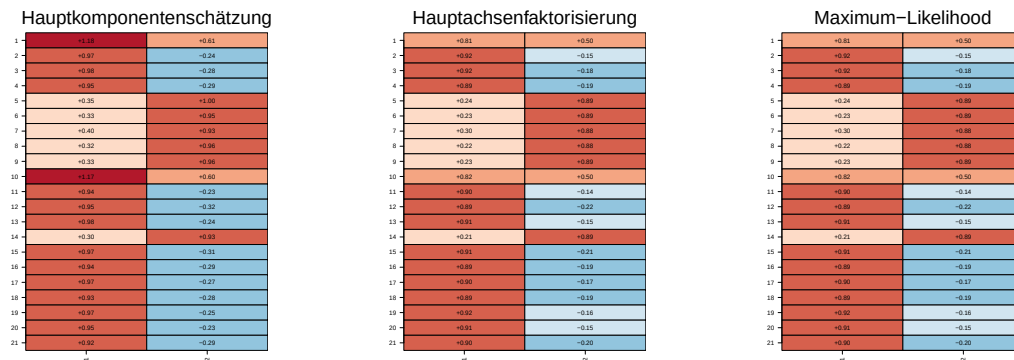


Abbildung 51.4. Faktorladungsmatrixschätzer für den BDI-II Datensatz nach Keller et al. (2008) basierend auf Hauptkomponenten-, iterierter Hauptachsenfaktorisierung- und Maximum-Likelihood-Schätzung

51.3.1. Stichprobenvarianzbasierte Wahl der Faktoranzahl

Die quantitative Grundlage der stichprobenvarianzbasierten Faktorzahlwahl ist die additive Zerlegung der *Gesamtstichprobenvarianz* in eine *faktorenbasierte Stichprobenvarianz* und eine *fehlerbasierte Stichprobenvarianz*. Man bestimmt die Faktorenanzahl dann anhand des Prinzips, dass bei möglichst geringer Faktorenzahl das Verhältnis von faktorenbasierter Stichprobenvarianz und Fehlervarianz möglichst groß ist. Traditionell gibt es zu diesem Zweck eine Reihe von Heuristiken. Im Folgenden zeigen zunächst die Validität der skizzierten additiven Stichprobenvarianzzerlegung und diskutieren dann einige Möglichkeiten zur Wahl der Faktorenzahl.

Definition 51.11 (Stichprobenvarianzzerlegung der Faktoranalyse). $Y \in \mathbb{R}^{m \times n}$ sei ein Datensatz von n unabhängigen Beobachtungen eines Faktoranalysemodells, $C \in \mathbb{R}^{m \times m}$ sei seine Stichprobenkovarianzmatrix und $\hat{L} \in \mathbb{R}^{m \times k}$ und $\hat{\Psi} \in \mathbb{R}^{m \times m}$ seien die Hauptkomponentenschätzer k ter Ordnung. Dann werden

- die Summe der Diagonalelemente von C als *Gesamtstichprobenvarianz*,
- die Summe der Diagonalelemente von $\hat{L}\hat{L}^T$ als *faktorenbasierte Stichprobenvarianz* und
- die Summe der Diagonalelemente von $\hat{\Psi}$ als *fehlerbasierte Stichprobenvarianz*

bezeichnet.

•

Es gilt nun folgendes Theorem.

Theorem 51.7 (Stichprobenvarianzzerlegung der Faktoranalyse). *Basierend auf n unabhängigen Beobachtungen eines Faktoranalysemodells seien*

- $C = (c_{ij})_{1 \leq i, j \leq m} \in \mathbb{R}^{m \times m}$ die Stichprobenkovarianzmatrix,
- $\hat{L} = (\hat{l}_{ij})_{1 \leq i \leq m, 1 \leq j \leq k} \in \mathbb{R}^{m \times k}$ der Hauptkomponentenschätzer k ter Ordnung von L ,
- $\hat{\Psi} = \text{diag}(\psi_1, \dots, \psi_m) \in \mathbb{R}^{m \times m}$ der Hauptkomponentenschätzer k ter Ordnung von Ψ ,

sowie

- $G := \sum_{i=1}^m c_{ii}$ die Gesamtstichprobenvarianz,

- $F := \sum_{i=1}^m \sum_{j=1}^k \hat{l}_{ij}^2$ die Faktorbasierte Stichprobenvarianz,
- $R := \sum_{i=1}^m \hat{\psi}_i$ die Beobachtungsrauschenbasierte Stichprobenvarianz.

Dann gilt

$$G = F + R. \quad (51.71)$$

Außerdem gilt mit den Eigenwerten $\lambda_1, \dots, \lambda_k$ von C , dass

$$F = \sum_{j=1}^k \lambda_j, \text{ wobei } \lambda_j = \sum_{i=1}^m \hat{l}_{ij}^2 \quad (51.72)$$

für $j = 1, \dots, k$ der Anteil des j ten Faktors an F ist.

◦

Beweis. Wir erinnern zunächst daran, dass die Diagonalelemente von $\hat{L}\hat{L}^T$ durch

$$\sum_{j=1}^k \hat{l}_{ij}^2 \quad (51.73)$$

gegeben sind, wovon man sich durch Betrachtung der Einträge von $\hat{L}\hat{L}^T$ überzeugt:

$$\begin{aligned} \hat{L}\hat{L}^T &= \begin{pmatrix} \hat{l}_{11} & \dots & \hat{l}_{1k} \\ \hat{l}_{21} & \dots & \hat{l}_{2k} \\ \vdots & \ddots & \vdots \\ \hat{l}_{m1} & \dots & \hat{l}_{mk} \end{pmatrix} \begin{pmatrix} \hat{l}_{11} & \dots & \hat{l}_{m1} \\ \hat{l}_{12} & \dots & \hat{l}_{m2} \\ \vdots & \ddots & \vdots \\ \hat{l}_{1k} & \dots & \hat{l}_{mk} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^k \hat{l}_{1j}^2 & \sum_{j=1}^k \hat{l}_{1j}\hat{l}_{2j} & \dots & \sum_{j=1}^k \hat{l}_{1j}\hat{l}_{mj} \\ \sum_{j=1}^k \hat{l}_{2j}\hat{l}_{1j} & \sum_{j=1}^k \hat{l}_{2j}^2 & \dots & \sum_{j=1}^k \hat{l}_{2j}\hat{l}_{mj} \\ \vdots & \dots & \ddots & \vdots \\ \sum_{j=1}^k \hat{l}_{mj}\hat{l}_{1j} & \sum_{j=1}^k \hat{l}_{mj}\hat{l}_{2j} & \dots & \sum_{j=1}^k \hat{l}_{mj}^2 \end{pmatrix} \end{aligned} \quad (51.74)$$

Die Identität von G und $F + R$ folgt dann direkt aus der Identität der Diagonalelemente von C , $\hat{L}\hat{L}^T$ und $\hat{\Psi}$, die im Rahmen der Hauptkomponentenschätzung mithilfe von

$$\hat{\psi}_i := c_{ii} - \sum_{j=1}^k \hat{l}_{ij}^2 \text{ für } i = 1, \dots, m \quad (51.75)$$

konstruiert wird. Um als nächstes

$$F = \sum_{j=1}^k \lambda_j \quad (51.76)$$

zu zeigen halten zunächst fest, dass mit der Definition des Hauptkomponentenschätzer $\hat{L}$ die Summe der quadrierten Einträge in der j ten Spalte von $\hat{L}$ gleich der Summe der quadrierten Einträge in der j ten Spalte von $Q_k \Lambda_k^{1/2}$ ist. Dies mag man sich zum Beispiel für $m = 5$ und $k = 2$ verdeutlichen:

$$\hat{L} = Q_k \Lambda_k^{1/2} \Leftrightarrow \begin{pmatrix} \hat{l}_{11} & \hat{l}_{12} \\ \hat{l}_{21} & \hat{l}_{22} \\ \hat{l}_{31} & \hat{l}_{32} \\ \hat{l}_{41} & \hat{l}_{42} \\ \hat{l}_{51} & \hat{l}_{52} \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \\ q_{31} & q_{32} \\ q_{41} & q_{42} \\ q_{51} & q_{52} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} q_{11} & \sqrt{\lambda_2} q_{12} \\ \sqrt{\lambda_1} q_{21} & \sqrt{\lambda_2} q_{22} \\ \sqrt{\lambda_1} q_{31} & \sqrt{\lambda_2} q_{32} \\ \sqrt{\lambda_1} q_{41} & \sqrt{\lambda_2} q_{42} \\ \sqrt{\lambda_1} q_{51} & \sqrt{\lambda_2} q_{52} \end{pmatrix} \quad (51.77)$$

Weiterhin halten wir fest, dass, wenn q_j für $j = 1, \dots, k$ die j te Spalte von Q_k bezeichnet aufgrund der Orthonormalität von Q folgt, dass

$$q_j^T q_j = \sum_{i=1}^m q_{ij}^2 = 1. \quad (51.78)$$

Dann ergibt sich für die Summe der Diagonalelemente von $\hat{L}\hat{L}^T$ aber

$$F = \sum_{i=1}^m \sum_{j=1}^k \hat{l}_{ij}^2 = \sum_{j=1}^k \sum_{i=1}^m \hat{l}_{ij}^2 = \sum_{j=1}^k \sum_{i=1}^m (\sqrt{\lambda_j} q_{ij})^2 = \sum_{j=1}^k \lambda_j \sum_{i=1}^m q_{ij}^2 = \sum_{j=1}^k \lambda_j \quad (51.79)$$

Die Tatsache, dass der j te Eigenwert λ_j von C dabei der Anteil der durch den j ten Faktor erklärten Gesamtstichprobenvarianz ist ergibt sich dabei durch die Einsicht, dass der Beitrag des j ten Faktors in der j ten Spalte von $\hat{L}$ enkodiert ist und obige Gleichungskette impliziert, dass

$$\sum_{i=1}^m \hat{l}_{ij}^2 = \lambda_j \text{ für } j = 1, \dots, k. \quad (51.80)$$

□

Anwendungsbeispiel

Folgender **R** Code bestimmt die Gesamtstichprobenvarianz, faktorbasierte Stichprobenvarianz und fehlerbasierte Stichprobenvarianz sowie ihre Beziehungen nach Theorem 51.2 für ein Faktoranalysemodell mit $k = 2$ anhand des in Abbildung 51.1 dargestellten Datensatzes nach Keller et al. (2008).

```
# EFA mit Hauptkomponentenschätzung für k = 2
YT = read.csv("../data/903-fa-keller.csv")
Y = as.matrix(t(YT))
m = nrow(Y)
n = ncol(Y)
k = 2
I_n = diag(n)
J_n = matrix(rep(1,n^2), nrow = n)
C = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))
EA = eigen(C)
lambda_k = EA$values[1:k]
Q_k = EA$vectors[,1:k]
L_hat = Q_k %*% diag(sqrt(lambda_k))
Psi_hat = diag(diag(C) - diag(L_hat %*% t(L_hat)))
GG = sum(diag(C))
FF = sum(diag(L_hat %*% t(L_hat)))
RR = sum(diag(Psi_hat))
FF_lambda = sum(lambda_k)

# Y^T \in \mathbb{R}^{n \times m}
# Y \in \mathbb{R}^{m \times n}
# Datendimension
# Datenpunktzahl
# Faktoranzahl
# Einheitsmatrix I_n
# 1_{nn}
# Stichprobenkovarianzmatrix
# Eigenanalyse von R
# k größte Eigenwerte von R
# k zugehörige Eigenvektoren von R
# Faktorladungsmatrixschätzer
# Beobachtungsrauschenkovarianzmatrixschätzer
# Gesamtstichprobenvarianz
# Faktorenbasierte Stichprobenvarianz
# Beobachtungsrauschenbasierter Stichprobenvarianz
# Summe der Eigenwerte \lambda_1, \dots, \lambda_k
```

```
G = 25.84
F = 22.51
R = 3.33
F+B = 25.84
```

```
Faktorbasierte Stichprobenvarianz F = 22.51
Summe der Eigenwerte lambda_1, ..., lambda_k = 22.51
```

Basierend auf obiger Stichprobenvarianzzerlegung kann man nun die Faktorzahl k so wählen, dass ein vorgegebener Anteil der Gesamtstichprobenvarianz durch das entsprechende Faktoranalysemodell erklärt wird. Obige Stichprobenvarianzzerlegung besagt ja gerade, dass der durch den j ten Faktor erklärte Anteil an der Gesamtstichprobenvarianz G durch λ_j , der durch den j ten Faktor erklärte relative Anteil an der Gesamtstichprobenvarianz demnach durch λ_j/G und der durch die $j = 1, \dots, k$ Faktoren erklärte relative Anteil an der Gesamtstichprobenvarianz damit durch $\sum_{j=1}^k \lambda_j/G$ gegeben ist. Es macht also Sinn, λ_j , λ_j/G und $\sum_{j=1}^k \lambda_j/G$ zu untersuchen und zu visualisieren. Basierend auf einer solchen Darstellung mag man dann k so wählen, dass k möglichst klein und $\sum_{j=1}^k \lambda_j/G$ möglichst groß ist. Die Visualisierung der λ_j wird in diesem Kontext *Scree-Plot* genannt, wobei sich das englische Wort *Scree* auf die geologische Formation einer Geröllhalde oder eines Talus bezieht und den charakteristischen zunächst schnell und dann langsam abfallenden

Anteil an erklärter Stichprobenvarianz durch die der Größe nach geordneten Eigenwerte beschreibt.

Awendungsbeispiel

```
# FA mit Hauptkomponentenschätzung für k = 5
YT = read.csv("../data/903-fa-keller.csv")
Y = as.matrix(t(YT))
m = nrow(Y)
n = ncol(Y)
k = 5
I_n = diag(n)
J_n = matrix(rep(1,n^2), nrow = n)
C = (1/(n-1))*(Y - I_n*(1/n)*J_n) %>% t(Y)
EA = eigen(C)
lambda_k = EA$values[1:k]
Q_k = EA$vectors[,1:k]
L_hat = Q_k %>% diag(sqrt(lambda_k))
Psi_hat = diag(diag(C) - diag(L_hat %>% t(L_hat)))
G = sum(diag(C))

# Y^T \in \mathbb{R}^{n \times m}
# Y \in \mathbb{R}^{m \times n}
# Datendimension
# Datenpunkanzahl
# Faktoranzahl
# Einheitsmatrix I_n
# 1_{nn}
# Stichprobenkovarianzmatrix
# Eigenanalyse von R
# k größte Eigenwerte von R
# k zugehörige Eigenvektoren von R
# Faktorladungsmatrixschätzer
# Beobachtungsauschenkovarianzmatrixschätzer
# Gesamtstichprobenvarianz
```

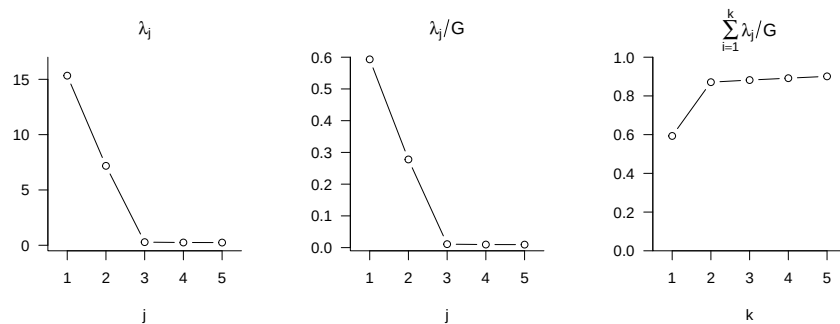


Abbildung 51.5. Scree-Plot und relative und kumulative Gesamtstichprobenvarianz durch steigende Faktorenanzahlen für den BDI-II Datensatz nach Keller et al. (2008).

Aus Abbildung 51.5 liest man ab, dass mit einer Faktoranzahl von $k = 1$ 56 % der Gesamtstichprobenvarianz, mit $k = 2$ können 87 % der Gesamtstichprobenvarianz und mit $k = 3$ 88 % der Gesamtstichprobenvarianz erklärt werden können. Der relative Anteil der durch Hinzunahme des dritten Faktors ist mit 1% also gering, so dass im Sinne der Modellparsimonität die Wahl von $k = 2$ Faktoren sinnvoll erscheint.

51.3.2. Likelihood-basierter Nullmodellvergleich

Die Normalverteilungsannahme im Faktoranalysemodell ermöglicht es, strukturierte Modellvergleiche mithilfe eines Log-Likelihood-Ratio-Kriteriums durchzuführen. Zu dessen Erläuterung seien $Y := (y_1, \dots, y_n) \in \mathbb{R}^{m \times n}$ sei ein Datensatz von n unabhängigen Beobachtungen eines Faktoranalysemodells mit $\bar{y} = 0_m$ und $\text{uvec}(A, B, \dots)$ sei die konkatenierte Vektorisierung der unikalen Werte der Matrizen $A, B, \dots$ sowie $\text{uvec}^{-1}(A, B, \dots)$ ihre Umkehrung. Weiterhin seien M_1 ein Normalverteilungsmodell der Faktoranalyse, das die Ordnungsrelation erfüllt und identifizierbar ist,

$$y \sim N(0, \Sigma_\theta) \quad (51.81)$$

mit

$$\Sigma_\theta = L\Phi L^T + \Psi, \theta = \text{uvec}(L, \Phi, \Psi) \in \Theta, \Theta \subset \mathbb{R}^p \text{ und } p \leq m(m+1)/2 \quad (51.82)$$

und M2 ein multivariates Normalverteilungsmodell mit beliebigem Kovarianzmatrixparameter,

$$y \sim N(0, \Sigma_\gamma) \text{ mit } \Sigma_\gamma = \text{uvec}^{-1}(\gamma), \gamma \in \Gamma \text{ und } \Gamma \subset \mathbb{R}^{m(m+1)/2}. \quad (51.83)$$

Das Log-Likelihood-Ratio-Kriterium zum Vergleich von M1 und M2 setzt bekanntlich die maximierten Wahrscheinlichkeitsdichten von Y unter M1 und M2 ins Verhältnis. Im vorliegenden Fall hat es die Form

$$\Lambda_Y := \ln \left(\frac{\max_{\theta \in \Theta} \prod_{i=1}^n N(y_i; 0_m, \Sigma_\theta)}{\max_{\gamma \in \Gamma} \prod_{i=1}^n N(y_i; 0_m, \Sigma_\gamma)} \right). \quad (51.84)$$

Große Werte von Λ_Y bedeuten, dass Y unter M1 eine größere Wahrscheinlichkeitsdichte besitzt als unter M2. Dies wird allgemein als Evidenz dafür verstanden, dass Y eher von M1 als von M2 generiert wurden. Dabei ist Λ_Y letztlich die zentrale Motivation für die funktionale Form der Diskrepanzfunktion (cf. Lawley (1940)). Folgendes Theorem zeigt nun zunächst den Zusammenhang zwischen dem Likelihood-Ratio-Kriterium Λ_Y und der im Kontext der Maximum-Likelihood-Schätzung der Faktoranalyse eingeführten Diskrepanzfunktion.

Theorem 51.8 (Diskrepanzfunktion). *Für einen Datensatz $Y := (y_1, \dots, y_n) \in \mathbb{R}^{m \times n}$ mit $\bar{y} = 0_m$ von n unabhängigen Beobachtungen eines Zufallsvektors y sei das Log-Likelihood-Ratio-Kriterium der Faktoranalyse gegeben durch*

$$\Lambda_Y := \ln \left(\frac{\max_{\theta \in \Theta} \prod_{i=1}^n N(y_i; 0_m, \Sigma_\theta)}{\max_{\gamma \in \Gamma} \prod_{i=1}^n N(y_i; 0_m, \Sigma_\gamma)} \right), \quad (51.85)$$

wobei

$$\Sigma_\theta = L\Phi L^T + \Psi, \theta = \text{uvec}(L, \Phi, \Psi) \in \Theta, \Theta \subset \mathbb{R}^p \text{ und } p \leq m(m+1)/2 \quad (51.86)$$

und

$$\Sigma_\gamma = \text{uvec}^{-1}(\gamma), \gamma \in \Gamma \text{ und } \Gamma \subset \mathbb{R}^{m(m+1)/2} \quad (51.87)$$

seien. Weiterhin sei für die verzerrte Stichprobenkovarianzmatrix S von Y

$$F_Y(\theta) := n \ln |\Sigma_\theta| + n \text{tr}(S \Sigma_\theta^{-1}) - n \ln |S| - nm \quad (51.88)$$

die Diskrepanzfunktion und $\hat{\theta}$ eine Minimumstelle von F_Y . Dann gilt

$$-2\Lambda_Y = F_Y(\hat{\theta}) \quad (51.89)$$

◦

Beweis. Wir halten zunächst fest, dass

$$\max_{\theta \in \Theta} \prod_{i=1}^n N(y_i; 0_m, \Sigma_\theta) = \prod_{i=1}^n N(y_i; 0_m, \Sigma_{\hat{\theta}}) \quad (51.90)$$

weil eine Minimumstelle $\hat{\theta}$ von F_Y wie in Theorem 51.6 gesehen die Log-Likelihood-Funktion und damit auch die Likelihood-Funktion eines Normalverteilungsmodells der Faktoranalyse maximiert. Weiterhin halten wir fest, dass

$$\max_{\gamma \in \Gamma} \prod_{i=1}^n N(y_i; 0_m, \Sigma_\gamma) = \prod_{i=1}^n N(y_i; 0_m, S) \quad (51.91)$$

weil die verzerrte Stichprobenkovarianz der Maximum-Likelihood-Schätzer des Kovarianzmatrixparameters einer multivariaten Normalverteilung ist. Die Logarithmuseigenschaften ergeben dann

$$\Lambda_Y = \ln \left(\frac{\prod_{i=1}^n N(y_i; 0_m, \Sigma_{\hat{\theta}})}{\prod_{i=1}^n N(y_i; 0_m, S)} \right) = \sum_{i=1}^n \ln N(y_i; 0_m, \Sigma_{\hat{\theta}}) - \sum_{i=1}^n \ln N(y_i; 0_m, S) \quad (51.92)$$

Substitution der funktionalen Form der Log-Likelihood-Funktion der konfirmatorischen Faktoranalyse und der funktionalen Form der WDF der multivariaten Normalverteilung ergibt

$$\begin{aligned} \Lambda_Y &= \sum_{i=1}^n \ln N(y_i; 0_m, \Sigma_{\hat{\theta}}) - \sum_{i=1}^n \ln N(y_i; 0_m, S) \\ &= -\frac{mn}{2} \ln(2\pi) - \frac{n}{2} \ln |\Sigma_{\theta}| - \frac{n}{2} \text{tr}(S \Sigma_{\theta}^{-1}) - \frac{n}{2} \bar{y}^T \Sigma_{\theta}^{-1} \bar{y} \\ &\quad + \frac{mn}{2} \ln(2\pi) + \frac{n}{2} \ln |S| + \frac{n}{2} \text{tr}(S S^{-1}) + \frac{n}{2} \bar{y}^T S^{-1} \bar{y} \\ &= -\frac{n}{2} \ln |\Sigma_{\theta}| - \frac{n}{2} \text{tr}(S \Sigma_{\theta}^{-1}) - \frac{n}{2} 0_m^T \Sigma_{\theta}^{-1} 0_m \\ &\quad + \frac{n}{2} \ln |S| + \frac{n}{2} \text{tr}(S S^{-1}) + \frac{n}{2} 0_m^T S^{-1} 0_m \\ &= -\frac{n}{2} \ln |\Sigma_{\theta}| - \frac{n}{2} \text{tr}(S \Sigma_{\theta}^{-1}) + \frac{n}{2} \ln |S| + \frac{mn}{2} \end{aligned} \quad (51.93)$$

Multiplikation mit -2 ergibt schließlich

$$-2\Lambda_Y = n \ln |\Sigma_{\theta}| + n \text{tr}(S \Sigma_{\theta}^{-1}) - n \ln |S| - mn = F_Y(\theta). \quad (51.94)$$

□

51.4. Modellinterpretation

Da die Werte des beobachtbaren Zufallsvektors eines Faktorenanalyse-Modells als Punkte in einem kanonischen Koordinatensystem aufgefasst werden, entspricht der j -te Eintrag in der i -ten Zeile einer Faktorenladungsmatrix dem Wert der beobachteten Variable y_i für f , unter der Annahme, dass f den Wert des j -ten kanonischen Einheitsvektors annimmt. Wie in Theorem 51.3 gesehen, bleibt das grundlegende Theorem der Faktorenanalyse jedoch erfüllt, wenn sowohl die Faktorkoordinaten als auch die kanonischen Beobachtungen in den Zeilen der Faktorenladungsmatrix auf eine beliebige andere orthogonale Basis projiziert werden (Lawley & Maxwell (1971)). Mit anderen Worten: In Bezug auf die Korrelationsmatrix des beobachtbaren Zufallsvektors ist der Parameter der Faktorenladungsmatrix eben nur bis zu einer beliebigen orthogonalen Koordinatentransformation eindeutig bestimmt — im Jargon der Faktorenanalyse als eine *orthogonale Rotation* bezeichnet. Diese Eigenschaft, die die Parameter des Faktorenanalysemodells nicht-identifizierbar macht, hat Forschende dazu angeregt, Kriterien zur Auswahl einer bestimmten Faktorenladungsmatrix aus ihrer unendlichen Menge gleichermaßen gültiger Werte zu entwickeln.

Varimax-Rotation

Eine Kriterium, das eine *einfache Struktur* (*simple structure*) nach Thurstone (1937), der geschätzten Faktorladungsmatrix begünstigen soll, ist das Varimax-Kriterium nach Kaiser (1958). Im Wesentlichen formuliert das Varimax-Kriterium eine Zielfunktion, um die Summe der Varianzen der quadrierten Einträge der Spalten der Faktorenladungsmatrix zu maximieren. Auf diese Weise sollen Spalten bevorzugt werden, deren Einträge entweder nahe am Spaltenmittelwert der Quadrate oder stark positiv bzw. negativ davon abweichend sind und damit ein Muster bevorzugt werden, das in modernen Begriffen als

sparse matrix bezeichnet wird (Rohe & Zeng (2023)). Entscheidend ist, dass das Varimax-Kriterium eine Funktion der auf die ursprüngliche Faktorenladungsmatrix angewandten orthogonalen Rotation ist, dargestellt durch die Multiplikation mit einer orthogonalen Matrix (Deisenroth et al. (2020)). Das Varimax-Problem besteht somit darin, die orthogonale Matrix zu identifizieren, die das Varimax-Kriterium für eine gegebene Anfangsfaktorenladungsmatrix maximiert. Eine so erhaltene Version der Faktorenladungsmatrix ist hoffentlich leicht interpretierbar, da sie direkt Gruppen von beobachtbaren Variablen mit kanonischen Einheitsvektor-Faktorwerten verknüpft und somit den Faktoren im Hinblick auf die von ihnen getragenen beobachtbaren Variablen Bedeutung verleiht. Der spezifische Varimax-Ansatz der **R** Funktion `varimax()` implementiert einen Algorithmus, der als *simultane Faktoren-Varimax-Lösung* bezeichnet wird und von Horst (1965) und Lawley & Maxwell (1971) entwickelt wurde.

Im Folgenden geben wir daher zunächst einen kurzen Überblick über die Varimax-Zielfunktion, die verwendet wird, um eine einfache Struktur der Faktorenladungsmatrix zu erreichen, betrachten dann das daraus resultierende Optimierungsproblem mit Nebenbedingungen, gegen dann die notwendige Bedingung für ein Maximum der Lagrange-Funktion dieses Problems an und betrachten schließlich über den iterativen Algorithmus, der in `varimax()` implementiert ist, um die notwendige Bedingung für eine orthogonale Koordinatentransformation zu lösen. Zu diesem Zweck sei durchgängig

$$\hat{L}_M = (\hat{l}_{M_{ij}}) \quad 1 \leq i \leq m, 1 \leq j \leq k := \hat{L}M \quad (51.95)$$

die koordinatentransformierte Version einer Schätzung der Faktorenladungsmatrix $\hat{L} \in \mathbb{R}^{m \times k}$, sodass die Koordinaten der Zeilenvektoren von $\hat{L}$ bezüglich der Basisvektoren ausgedrückt werden, die die Spalten der orthogonalen Matrix $M \in \mathbb{R}^{k \times k}$ bilden. Im Jargon der Faktorenanalyse wird $\hat{L}_M$ üblicherweise als die *rotierte Faktorenladungsmatrix* bezeichnet und M als die *Rotationsmatrix*.

Definition 51.12. Gegeben sei ein Schätzer $\hat{L}$ für die Faktorladungsmatrix eines Faktoranalysemodells. Dann ist die Varimax-Zielfunktion definiert als

$$v : \mathbb{R}_O^{k \times k} \rightarrow \mathbb{R}, M \mapsto v(M) := \sum_{j=1}^k \sum_{i=1}^m \left(\hat{l}_{M_{ij}}^2 - \frac{1}{m} \sum_{r=1}^m \hat{l}_{M_{rj}}^2 \right)^2 = \sum_{j=1}^k \sum_{i=1}^m \left(\hat{l}_{M_{ij}}^4 - \frac{1}{m} \left(\sum_{r=1}^m \hat{l}_{M_{rj}}^2 \right)^2 \right), \quad (51.96)$$

wobei $\mathbb{R}_O^{k \times k}$ die Menge aller orthogonalen $k \times k$ -Matrizen bezeichnen soll.

•

Für eine gegebene Faktorenladungsmatrix bewertet die Varimax-Zielfunktion v also die Summe der Abweichungsquadrate der quadrierten Einträge von $\hat{L}_M$, wobei jede Abweichung vom entsprechenden Spaltenmittelwert als Funktion der Rotationsmatrix M gemessen wird. Es sei angemerkt, dass die Gleichheit auf der rechten Seite von Gleichung 51.96 einige algebraische Umformungen erfordert, die hier nicht ausgeführt werden.

Die Grundidee des Varimax-Ansatzes besteht nun darin, v bezüglich der orthogonalen Matrix M zu maximieren. Dies führt auf das nichtlineare Optimierungsproblem

$$\max v(M) \text{ unter der Nebenbedingung } M^T M = I_k. \quad (51.97)$$

Mittels der Differentialrechnung für Matrizen konnte gezeigt werden (vgl. Horst (1965), Lawley & Maxwell (1971), Neudecker (1969), Magnus & Neudecker (1989)), dass die notwendige Bedingung für ein Maximum der Lagrange-Funktion des Optimierungsproblems Gleichung 51.97 in der Gleichung

$$MA_M = B_M \quad (51.98)$$

ausgedrückt werden kann, wobei $A_M \in \mathbb{R}^{k \times k}$ eine symmetrische und positiv definite Matrix unbekannter Lagrange-Multiplikatoren ist und

$$B_M := \hat{L}^T \left(C_M - \frac{1}{m} \hat{L}_M D_M \right) \quad (51.99)$$

mit

$$\hat{L}_M := \hat{L} M, C_M := \left(\hat{l}_{M_{ij}}^3 \right)_{1 \leq i \leq m, 1 \leq j \leq k} \text{ und } D_M := \text{diag} \left(\sum_{i=1}^m \hat{l}_{M_{i1}}^2, \dots, \sum_{i=1}^m \hat{l}_{M_{ik}}^2 \right) \quad (51.100)$$

ist. Bemerkenswert ist, dass Gleichung 51.98 lediglich eine implizite Definition des optimalen M liefert, da beide Seiten der Gleichung von M abhängen. Um dieses Gleichungssystem iterativ nach M zu lösen, geht der in `varimax()` implementierte Ansatz der simultanen Faktoren-Varimax-Lösung nach folgendem Algorithmus vor.

Definition 51.13 (Simultane-Faktoren-Varimax-Lösungsalgorithmus). Gegeben sei ein Schätzer $\hat{L}$ für die Faktorladungsmatrix eines Faktoranalysemodells. Dann hat der Algorithmus der **R** Funktion `varimax()` die folgende Form:

Initialisierung

$$(SN) \text{ Setze } \hat{L}_N := N \hat{L} \text{ mit } N := \text{diag} \left(\left(\sum_{j=1}^k \hat{l}_{1j}^2 \right)^{-\frac{1}{2}}, \dots, \left(\sum_{j=1}^k \hat{l}_{mj}^2 \right)^{-\frac{1}{2}} \right)$$

$$(S0) \text{ Setze } M^{(0)} := I_k, \delta^{(0)} := 0, \delta^{(1)} := 1 \text{ und definiere } \varepsilon^{(0)}$$

Iterationen

Für $i = 1, \dots, 1000$ und solange $d^{(i)} \geq d^{(i-1)}(1 + \varepsilon)$

$$(S1) \text{ Setze } \hat{L}_{M^{(i)}} = \hat{L}_N M^{(i-1)}$$

$$(S2) \text{ Setze } B_{M^{(i)}} = \hat{L}_N^T \left(C_{M^{(i)}} - \frac{1}{m} \hat{L}_{M^{(i)}} D_{M^{(i)}} \right)$$

$$(S3) \text{ Evaluiere } B_{M^{(i)}} = U^{(i)} \Delta^{(i)} V^{(i)T}$$

$$(S4) \text{ Setze } M^{(i)} = U^{(i)} V^{(i)T}$$

$$(S5) \text{ Evaluiere } d^{(i)} := \text{tr}(\Delta^{(i)})$$

Finalisierung

$$(S6) \text{ Setze } \hat{L}_{NM} := \hat{L}_N M^{(i)} \text{ und } \hat{L}_M := N^{-1} \hat{L}_{NM}$$

•

In Definition 51.13 normalisiert der Initialisierungsschritt (SN) jede Zeile der Faktorladungsmatrixschätzung $\hat{L}$ auf die euklidische Einheitslänge durch Multiplikation mit einer $m \times m$ diagonalen Normalisierungsmatrix N , deren Diagonalelemente den reziproken euklidischen Längen der Zeilen von $\hat{L}$ entsprechen. Obwohl nicht unumstritten (vgl. Kaiser (1970), Kaiser & Rice (1974), Rohe & Zeng (2023)), wurde diese zeilenweise Normalisierung bei der Einführung der Varimax-Rotation durch Kaiser (1958) vorgeschlagen und bildet das Standardverfahren in der **R** Implementierung der Varimax-Prozedur. Die Rotationsmatrix wird dann als $k \times k$ Einheitsmatrix initialisiert, die Konvergenzvariablen δ werden auf 0 gesetzt und ein geeigneter Wert für das Konvergenzkriterium ε wird definiert. Die Iterationsschritte (S1) und (S2) evaluieren dann die relevanten Iterandenmatrizen für die rechte Seite von Gleichung 51.98, $L_{M^{(i)}}$, $D_{M^{(i)}}$, $C_{M^{(i)}}$, und $B_{M^{(i)}}$, wodurch die rekursive Gleichung

$$M^{(i+1)} A_{M^{(i)}} = B_{M^{(i)}} \quad (51.101)$$

implizit definiert wird. Diese implizite rekursive Gleichung wird dann für $M^{(i+1)}$ mittels einer Singulärwertzerlegung von $B_{M^{(i)}}$ gelöst, wobei $M^{(i+1)}$ in den Iterationsschritten (S3) und (S4) auf $U^{(i)} V^{(i)}$ gesetzt wird. Hierbei sind $U^{(i)}$ und $V^{(i)}$ die Matrizen der linken bzw. rechten Singulärvektoren von $B_{M^{(i)}}$, und wir begründen diese Lösung untenstehend. Die Konvergenz des Algorithmus wird dann anhand der Spur der Singulärwertmatrix $\Delta^{(i)}$ im Iterationsschritt (S5) bewertet. Schließlich wird nach der Konvergenz die optimierte Transformationsmatrix $M^{(i)}$ auf die normalisierte Faktorladungsmatrixschätzung angewendet und die Normalisierungsskalierung in Schritt (S6) rückgängig gemacht.

Wir wollen schließlich den Iterationsschritt (S4) begründen, welche die notwendige Bedingung für ein Maximum der Lagrange-Funktion des restringierten Optimierungsproblems Gleichung 51.97 in der Form von Gleichung 51.98 iterativ mittels einer Singulärwertzerlegung der Matrix auf der rechten Seite von Gleichung 51.98 löst. Dieser Ansatz nutzt die Eigenschaften der beteiligten Matrizen und deren Beziehungen. Zur Vereinfachung der Notation verzichten wir im Folgenden auf die Matrixabhängigkeit von M sowie auf den Iterations-Superskript und betrachten die Lösung von $MA = B$ für bekanntes B und unbekannte (aber positiv-definite und symmetrische) Lagrange-Multiplikator-Matrix A . Zunächst stellen wir fest, dass die Prämultiplikation von $MA = B$ mit der Transponierten von B ergibt, dass

$$MA = B \Leftrightarrow B^T MA = B^T B \Leftrightarrow (MA)^T MAB^T B \Leftrightarrow A^T M^T MA = B^T B \Leftrightarrow A^2 = B^T B, \quad (51.102)$$

wobei die letzte Äquivalenz aus der Orthogonalität von M und der Symmetrie von A folgt.

Basierend auf Gleichung 51.102 formulieren wir als Nächstes zwei Ausdrücke für die unbekannte Matrix A^2 . Erstens betrachten wir die Singulärwertzerlegung von $B = USV^T$, wobei U die orthogonale Matrix der linken Singulärvektoren, S die Diagonalmatrix der Singulärwerte und V die orthogonale Matrix der rechten Singulärvektoren ist. Wir erhalten dann:

$$A^2 = B^T B = (USV^T)^T USV^T = VSU^T USV^T = VS^2 V^T. \quad (51.103)$$

Da A eine positiv-definite und symmetrische Matrix ist, kann sie als Zerlegung in ihre orthogonale Matrix von Eigenvektoren Q und ihre Diagonalmatrix von Eigenwerten Λ geschrieben werden. Wir haben also auch:

$$A^2 = AA = Q\Lambda Q^T Q\Lambda Q^T = Q\Lambda^2 Q^T \quad (51.104)$$

und wir erhalten:

$$A^2 = VS^2V^T = QA^2Q^T = A^2. \quad (51.105)$$

Ohne Beschränkung der Allgemeinheit können wir $V = Q$ annehmen, sodass $S^2 = \Lambda^2$ und damit $S = \Lambda$ bis auf Vorzeichenpermutationen gilt. Wir können nun Gleichung 51.98 wie folgt lösen:

$$MA = B \Leftrightarrow M = BA^{-1} \Leftrightarrow M = USV^TQA^{-1}Q^T \Leftrightarrow M = USQ^TQS^{-1}V^T \Leftrightarrow M = UV^T, \quad (51.106)$$

was schließlich den Iterationsschritt (S4) rechtfertigt.

Anwendungsbeispiel

Folgender **R** Code demonstriert die Implementnation von Definition 51.13.

```
N = solve(diag(sqrt(diag(L_hat %*% t(L_hat))))) # Zeilennormalisierende Matrix
L_hat_N = N %*% L_hat # zeilennormalisierte Faktorladungsmatrixschätzer
m = nrow(L_hat_N) # Zeilenanzahl
k = ncol(L_hat_N) # Spaltenanzahl
Mi = diag(k) # Transformationsmatrixinitialisierung
d = 0 # Konvergenzvariableninitialisierung
eps = 1e-05 # Konvergenzkriterium (change factor)
for (i in 1L:100L){ # Iterationen
  L_hat_Mi = L_hat_N %*% Mi # \hat{L}_M^{(i)}
  Ci = L_hat_Mi^3 # C^{(i)}
  Di = diag(drop(rep(1, m) %*% L_hat_Mi^2)) # D^{(i)}
  Bi = t(L_hat_N) %*% (Ci - (1/m) * L_hat_Mi %*% Di) # B^{(i)}
  UDVi = svd(Bi) # Singulärwertzerlegung von B^{(i)}
  Mi = UDVi$u %*% t(UDVi$v) # Transformationsmatrix Update M^{(i)}
  dd = sum(UDVi$d) # Konvergenzkriterium Update
  if (dd < d * (1 + eps)){break} # Konvergenztest
  d = dd # Konvergenzvariablen Update
  L_hat_NM = L_hat_N %*% Mi # Transformierter Faktorladungsmatrixschätzer
  L_hat_M = solve(N) %*% L_hat_NM # Zeilennormalisierungsaufhebung
```

Referenzen

- Abbott, S. (2015). *Understanding Analysis*. Springer New York. <https://doi.org/10.1007/978-1-4939-2712-8>
- Aitken, A. C. (1936). IV.—On Least Squares and Linear Combination of Observations. *Proceedings of the Royal Society of Edinburgh*, 55, 42–48. <https://doi.org/10.1017/S0370164600014346>
- Aldrich, J. (1997). R.A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, 12(3), 162–176. <https://doi.org/10.1214/ss/1030037906>
- Amrhein, V., & Greenland, S. (2018). Remove, Rather than Redefine, Statistical Significance. *Nature Human Behaviour*, 2(1), 4–4. <https://doi.org/10.1038/s41562-017-0224-0>
- Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed). Wiley-Interscience.
- Arens, T., Hettlich, F., Karpfinger, C., Kockelkorn, U., Lichtenegger, K., & Stachel, H. (2018). *Mathematik*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-56741-8>
- Baba, K., Shibata, R., & Sibuya, M. (2004). Partial Correlation and Conditional Correlation as Measures of Conditional Independence. *Australian & New Zealand Journal of Statistics*, 46(4), 657–664. <https://doi.org/10.1111/j.1467-842X.2004.00360.x>
- Bärwolff, G. (2017). *Höhere Mathematik für Naturwissenschaftler und Ingenieure*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-55022-9>
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53, 370–418. <https://doi.org/10.1098/rstl.1763.0053>
- Beck, A. T. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6), 561. <https://doi.org/10.1001/archpsyc.1961.01710120031004>
- Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients. *Journal of Personality Assessment*, 67(3), 588–597. https://doi.org/10.1207/s15327752jpa6703_13
- Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics* (Reprint). Chapman & Hall.
- Bertram, M. G., Sundin, J., Roche, D. G., Sánchez-Tójar, A., Thoré, E. S. J., & Brodin, T. (2023). Open Science. *Current Biology*, 33(15), R792–R797. <https://doi.org/10.1016/j.cub.2023.05.036>
- Bohus, M., Kleindienst, N., Limberger, M. F., Stieglitz, R.-D., Domsalla, M., Chapman, A. L., Steil, R., Philipsen, A., & Wolf, M. (2009). The Short Version of the Borderline Symptom List (BSL-23): Development and Initial Data on Psychometric Properties. *Psychopathology*, 42(1), 32–39. <https://doi.org/10.1159/000173701>
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley New York.
- Borel, É. (1898). *Leçons Sur La Théorie Des Fonctions*. Paris: Gauthier Villars.
- Borsboom, D. (2009). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge University Press.
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The Concept of Validity.

- Psychological Review*, 111(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Box, J. F. (1981). Gosset, Fisher, and the t Distribution. *The American Statistician*, 35(2), 61. <https://doi.org/10.2307/2683142>
- Bravais, A. (1844). *Analyse Mathématique : Sur Les Probabilités Des Erreurs de Situation d'un Point*.
- Bühner, M. (2010). *Einführung in die Test- und Fragebogenkonstruktion* (2., aktualisierte und erw. Aufl., [Nachdr.]). Pearson Studium.
- Burden, R. L., Faires, J. D., & Burden, A. M. (2016). *Numerical Analysis* (Tenth edition). Cengage Learning.
- Cantor, G. (1874). Über Eine Eigenschaft Des Inbegriffes Aller Reellen Algebraischen Zahlen. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 1.
- Cantor, G. (1891). Über Eine Elementare Frage Der Mannigfaltigkeitslehre. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 1, 75–78.
- Cantor, G. (1895). Beiträge Zur Begründung Der Transfiniten Mengenlehre. *Mathematische Annalen*, 46(4), 481–512. <https://doi.org/10.1007/BF02124929>
- Casella, G., & Berger, R. (2002). *Statistical Inference* (2. Aufl.). Thomson Learning.
- Cattell, R. B. (1957). *Personality and Motivation - Structure and Measurement*.
- Chiossi, S. G. (2021). *Essential Mathematics for Undergraduates: A Guided Approach to Algebra, Geometry, Topology and Analysis*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-87174-1>
- Christensen, R. (2011). *Plane Answers to Complex Questions*. Springer New York. <https://doi.org/10.1007/978-1-4419-9816-3>
- Cohen, J. (1968). Multiple Regression as a General Data-Analytic System. *Psychological Bulletin*, 70(6, Pt.1), 426–443. <https://doi.org/10.1037/h0026714>
- Cotton, R. (2013). *Learning R* (First Edition). O'Reilly.
- Cronbach, L. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3), 297–334.
- Dahm, M. (2005). *Grundlagen Der Mensch-Computer-Interaktion*. Pearson Studium.
- De Beurs, E., Boehnke, J. R., & Fried, E. I. (2022). Common Measures or Common Metrics? A Plea to Harmonize Measurement Results. *Clinical Psychology & Psychotherapy*, 29(5), 1755–1767. <https://doi.org/10.1002/cpp.2742>
- De Finetti, B. (1975). *Theory of Probability*. John Wiley & Sons.
- DeGroot, M. H., & Schervish, M. J. (2012). *Probability and Statistics* (4th ed). Addison-Wesley.
- Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for Machine Learning* (1. Aufl.). Cambridge University Press. <https://doi.org/10.1017/9781108679930>
- Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An Introductory Report. *Psychological Medicine*, 13(3), 595–605. <https://doi.org/10.1017/S003329170048017>
- Draper, N., & Smith, H. (1998). *Applied Regression Analysis*. Wiley-Interscience.
- Edgeworth, F. Y. (1892). The Law of Error and Correlated Averages. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 34(210), 429–438. <https://doi.org/10.1080/14786449208620355>
- Everitt, B., & Howell, D. C. (Hrsg.). (2005). *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons.
- Feldt, L. S. (1965). The Approximate Sampling Distribution of Kuder-Richardson Reliability Coefficient Twenty. *Psychometrika*, 30(3), 357–370. <https://doi.org/10.1007/BF02289499>

- Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical Inference for Coefficient Alpha. *Applied Psychological Measurement*, 11(1), 93–103. <https://doi.org/10.1177/014662168701100107>
- Fischer, H. (2011). *A History of the Central Limit Theorem*. Springer New York. <https://doi.org/10.1007/978-0-387-87857-7>
- Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594–604), 309–368. <https://doi.org/10.1098/rsta.1922.0009>
- Fisher, R. A. (1924). The Distribution of the Partial Correlation Coefficient. *Metron*, 3, 329–332.
- Fisher, R. A. (1925a). Applications of "Student's" Distribution. *Metron*, 5, 90–104.
- Fisher, R. A. (1925b). *Statistical Methods for Research Workers*. Oliver & Boyd.
- Fisher, R. A. (1925c). Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5), 700–725. <https://doi.org/10.1017/S030500410009580>
- Fisher, R. A. (1935). *The Design of Experiments* (1. ed). Hafner Press.
- Foulley, J. (1993). *A Simple Argument Showing How to Derive Restricted Maximum Likelihood*.
- Fox, R., & Tishby, N. (2016). Minimum-Information LQG Control Part I: Memoryless Controllers. 5610–5616. <https://doi.org/10.1109/CDC.2016.7799131>
- Friston, K. (2005). A Theory of Cortical Responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 815–836. <https://doi.org/10.1098/rs.tb.2005.1622>
- Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246. <https://doi.org/10.2307/2841583>
- Galton, F. (1890). Kinship and Correlation. *Statistical Science*, 4(2). <https://doi.org/10.1214/ss/1177012581>
- Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Cambridge University Press.
- Georgii, H.-O. (2009). *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik* (4., überarb. und erw. Aufl). de Gruyter.
- Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2022). *Factor Analysis, Probabilistic Principal Component Analysis, Variational Inference, and Variational Autoencoder: Tutorial and Survey* (arXiv:2101.00734). arXiv. <https://doi.org/10.48550/arXiv.2101.00734>
- Gigerenzer, G. (2004). Mindless Statistics. *The Journal of Socio-Economics*, 33(5), 587–606. <https://doi.org/10.1016/j.socec.2004.09.033>
- Golub, G. H., & Van Loan, C. F. (2013). *Matrix Computations* (Fourth edition). The Johns Hopkins University Press.
- Gray, R. (1994). Georg Cantor and Transcendental Numbers. *The American Mathematical Monthly*, 101(9), 819–832. <https://doi.org/10.1080/00029890.1994.11997035>
- Gulliksen, H. (1950). *Theory of Mental Tests*. John Wiley & Sons Inc. <https://doi.org/10.1037/13240-000>
- Hájek, A. (2019). Interpretations of Probability. In E. N. Zalta (Hrsg.), *The Stanford Encyclopedia of Philosophy* (Fall 2019). Metaphysics Research Lab, Stanford University.
- Hald, A. (1990). *A History of Probability and Statistics and Their Applications before 1750*. Wiley.

- Hald, A. (2007). *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. Springer.
- Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72(358), 320. <https://doi.org/10.2307/2286796>
- Hautzinger, M., Keller, F., & Kühner, C. (2006). *BDI-II Beck Depressions-Inventar*. Pearson.
- Held, L., & Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-37887-4>
- Henze, N. (2018). *Stochastik für Einsteiger*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-22044-0>
- Herzog, S., & Ostwald, D. (2013). Sometimes Bayesian Statistics Are Better. *Nature*, 494(7435), 35–35. <https://doi.org/10.1038/494035b>
- Hesse, C. (2009). *Wahrscheinlichkeitstheorie* (2. Aufl.). Vieweg + Teubner.
- Hocking, R. (2003). *Methods and Applications of Linear Models - Regression and the Analysis of Variance*. Wiley.
- Horst, P. (1965). *Factor Analysis of Data Matrices*. Holt, Rinehart and Winston, Inc.
- Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2(3), 360–378. <https://doi.org/10.1214/aoms/1177732979>
- Hotelling, H. (1933). Analysis of Complex Variables into Principal Components. *Journal of Educational Psychology*, 24, 417-441 and 498-520.
- Hotelling, H. (1935). The Most Predictable Criterion. *Journal of Educational Psychology*, 26(2), 139–142. <https://doi.org/10.1037/h0058165>
- Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3/4), 321. <https://doi.org/10.2307/2333955>
- Hotelling, H. (1951). *A Generalized T Test and Measure of Multivariate Dispersion*.
- Huber, P. J. (1981). *Robust Statistics*. Wiley.
- Hyndman, R. J., & Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, 50(4), 361. <https://doi.org/10.2307/2684934>
- Hyvärinen, A., Khemakhem, I., & Monti, R. (2024). Identifiability of Latent-Variable and Structural-Equation Models: From Linear to Nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1), 1–33. <https://doi.org/10.1007/s10463-023-00884-4>
- Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3), 2999–2314.
- Imbens, G., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Academic Press.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 2* (2nd ed). Wiley.
- Johnson, N. L., & Welch, B. L. (1940). Applications of the Non-Central t-Distribution. *Biometrika*, 31(3/4), 362. <https://doi.org/10.2307/2332616>
- Jöreskog, K. G. (1967). *Some Contributions to Maximum Likelihood Factor Analysis*.
- Jöreskog, K. G. (1969). *A General Approach to Confirmatory Maximum Likelihood Factor Analysis*.
- Kaiser, H. F. (1958). The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika*, 23(3), 187–200. <https://doi.org/10.1007/BF02289233>
- Kaiser, H. F. (1970). A Second Generation Little Jiffy. *Psychometrika*, 35(4), 401–415. <https://doi.org/10.1007/BF02291817>
- Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, 34(1), 111–117. <https://doi.org/10.1177/001316447403400115>

- Keller, F., Hautzinger, M., & Kühner, C. (2008). Zur faktoriellen Struktur des deutschsprachigen BDI-II. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 37(4), 245–254. <https://doi.org/10.1026/1616-3443.37.4.245>
- Kochenderfer, M. J., Wheeler, T. A., & Wray, K. H. (2022). *Algorithms for Decision Making*. The MIT Press.
- Kolmogoroff, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-49888-6>
- Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Beltz, Psychologie Verl.-Union.
- Kristof, W. (1963). The Statistical Theory of Stepped-up Reliability Coefficients When a Test Has Been Divided into Several Equivalent Parts. *Psychometrika*, 28(3), 221–2238. <https://doi.org/10.1007/BF02289571>
- Lawley, D. (1940). The Estimation of Factor Loadings by the Method of Maximum Likelihood. *Proceedings of the Royal Society of Edinburgh. Section B: Biological Sciences*.
- Lawley, D., & Maxwell, A. (1971). *Factor Analysis as a Statistical Method* (2. Aufl.). London Butterworths.
- Lawrance, A. J. (1976). On Conditional and Partial Correlation. *The American Statistician*, 30(3), 146–149. <https://doi.org/10.1080/00031305.1976.10479163>
- Lazarsfeld, P. (1959). Latent Structure Analysis. In *Psychology: A Study of a Science, Vol. 3*. McGraw-Hill.
- Legendre, A. M. (1805). *Nouvelles Methodes Pour La Determination Des Orbites Des Cometes*. Didot Paris.
- Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Wiley Series in Probability and Statistics.
- Lehmann, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*. Springer New York. <https://doi.org/10.1007/978-1-4419-9500-1>
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Auflage). Beltz, Psychologie Verlags Union.
- Lindholm, M., & Wahl, F. (2020). On the Variance Parameter Estimator in General Linear Models. *Metrika*, 83(2), 243–254. <https://doi.org/10.1007/s00184-019-00751-4>
- Lord, F. M. (1980). *Applications of Item Response Theory To Practical Testing Problems* (0. Aufl.). Routledge. <https://doi.org/10.4324/9780203056615>
- Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores* (Nachdr. der Ausg. Reading, Mass. [u.a.], 1968). Information Age Publ.
- Magnus, J. R., & Neudecker, H. (1989). Matrix Differential Calculus with Applications in Statistics and Econometrics. *Journal of the American Statistical Association*, 84(408), 1103. <https://doi.org/10.2307/2290109>
- Mahalanobis, S. A. (1936). Reprint of: Mahalanobis, P.C. (1936) "On the Generalised Distance in Statistics.". 80(S1), 1–7. <https://doi.org/10.1007/s13171-019-00164-5>
- Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.
- McCall, W. A. (1922). *How to Measure in Education*. MacMillan Co. <https://doi.org/10.1037/13551-000>
- McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician*, 32(1), 12. <https://doi.org/10.2307/2683468>
- McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, 73(sup1), 235–245. <https://doi.org/10.1080/00031305.2018.1527253>
- Meintrup, D., & Schäffler, S. (2005). *Stochastik: Theorie und Anwendungen*. Springer.
- Miedema, F. (2022). *Open Science: The Very Idea*. Springer Netherlands. <https://doi.org/10.1007/978-94-024-2115-6>

- Moosbrugger, H., & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion: mit 66 Abbildungen und 41 Tabellen* (2., aktualisierte und überarbeitete Auflage). Springer.
- Neudecker, H. (1969). Some Theorems on Matrix Differentiation with Special Reference to Kronecker Matrix Products. *Journal of the American Statistical Association*, 64.
- Newton, I. (1687). *Philosophiae Naturalis Principia Mathematica*. Royal Society.
- Neyman, J. (1935). On the Problem of Confidence Intervals. *The Annals of Mathematical Statistics*, 6(3), 111–116. <https://doi.org/10.1214/aoms/1177732585>
- Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Statistical Stimulation*.
- Neyman, J., & Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, 20A(1/2), 175. <https://doi.org/10.2307/2331945>
- Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Phil. Trans. R. Soc. Lond. A*, 231(694-706), 289–337.
- Novick, M. R. (1966). The Axioms and Principal Results of Classical Test Theory. *Journal of Mathematical Psychology*, 3(1), 1–18. [https://doi.org/10.1016/0022-2496\(66\)90002-2](https://doi.org/10.1016/0022-2496(66)90002-2)
- Patterson, H. D., & Thompson, R. (1971). Recovery of Inter-Block Information When Block Sizes Are Unequal. *Biometrika*, 58(3), 545–554. <https://doi.org/10.1093/biomet/58.3.545>
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, 5, 240–242. <https://www.jstor.org/stable/115794>
- Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 18, 253–318. <https://www.jstor.org/stable/90707>
- Pearson, K. (1900). On the Criterion That a given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157–175. <https://doi.org/10.1080/14786440009463897>
- Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572. <https://doi.org/10.1080/14786440109462720>
- Pearson, K. (1920). Notes on the History of Correlation. *Biometrika*, 13(1), 25–45. <https://doi.org/10.1093/biomet/13.1.25>
- Philip, S., Kew, S., Van Oldenborgh, G. J., Otto, F., Vautard, R., Van Der Wiel, K., King, A., Lott, F., Arrighi, J., Singh, R., & Van Aalst, M. (2020). A Protocol for Probabilistic Extreme Event Attribution Analyses. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(2), 177–203. <https://doi.org/10.5194/ascmo-6-177-2020>
- Pillai, K. C. S. (1955). Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, 26(1), 117–121. <https://doi.org/10.1214/aoms/1177728599>
- Plackett, R. L. (1949). A Historical Note on the Method of Least Squares. *Biometrika*, 36(3/4), 458. <https://doi.org/10.2307/2332682>
- Pratt, J., Raiffa, H., & Schlaifer, R. (1995). *Statistical Decision Theory*. MIT Press.
- Puterman, M. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

- Rao, C. R. (1951). An Asymptotic Expansion of the Distribution of Wilk's Criterion. *Bulletin of the International Statistical Institute*, 33(2), 177–180.
- Rao, C. R. (1972). Recent Trends of Research Work in Multivariate Analysis. *Biometrics*, 28(1), 3. <https://doi.org/10.2307/2528958>
- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests* (Expanded ed). University of Chicago Press.
- Rat für Informationsinfrastrukturen. (2017). *Datenschutz Und Forschungsdaten - Aktuelle Empfehlungen*.
- Reise, S. P., & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology*, 5(1), 27–48. <https://doi.org/10.1146/annurev.clinpsy.032408.153553>
- Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis* (Third Edition). Wiley.
- Rencher, A. C., & Schaalje, G. B. (2008). *Linear Models in Statistics* (2nd ed). Wiley-Interscience.
- Revelle, W. (2024). *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University.
- Richter, T., & Wick, T. (2017). *Einführung in die Numerische Mathematik: Begriffe, Konzepte und zahlreiche Anwendungsbeispiele*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-54178-4>
- Riley, J. (2017). *Understanding Metadata: What Is Metadata, and What Is It For*. National Information Standards Organization.
- Rohe, K., & Zeng, M. (2023). Vintage Factor Analysis with Varimax Performs Statistical Inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4), 1037–1060. <https://doi.org/10.1093/jrsssb/qkad029>
- Rosenblatt, B. (1997). The Digital Object Identifier: Solving the Dilemma of Copyright Protection Online. *Journal of Electronic Publishing*, 3(2). <https://doi.org/10.3998/3336451.0003.204>
- Rosseel, Y. (2012). **Lavaan** : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2). <https://doi.org/10.18637/jss.v048.i02>
- Rosseel, Y. (2021). Evaluating the Observed Log-Likelihood Function in Two-Level Structural Equation Modeling with Missing Data: From Formulas to R Code. *Psych*, 3(2), 197–232. <https://doi.org/10.3390/psych3020017>
- Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* (2., vollständig überarbeitete und erweiterte Aufl). H. Huber.
- Roweis, S., & Ghahramani, Z. (1999). A Unifying Review of Linear Gaussian Models. *Neural Computation*, 11(2), 305–345. <https://doi.org/10.1162/089976699300016674>
- Roy, S. N., & Bose, R. C. (1953). Simultaneous Confidence Interval Estimation. *The Annals of Mathematical Statistics*, 24(4), 513–536. <https://doi.org/10.1214/aoms/1177728912>
- Rubin, D. B., & Thayer, D. T. (1982). EM Algorithms for ML Factor Analysis. *Psychometrika*, 47(1), 69–76. <https://doi.org/10.1007/BF02293851>
- Scheffé, H. (1959). *The Analysis of Variance* (Wiley classics library ed). Wiley-Interscience Publication.
- Schmidt, K. D. (2009). *Maß und Wahrscheinlichkeit*. Springer.
- Scott, D. W. (1979). *On Optimal and Data-Based Histograms*. 6.
- Seal, H. L. (1967). Studies in the History of Probability and Statistics. XV: The Historical Development of the Gauss Linear Model. *Biometrika*, 54(1/2), 1. <https://doi.org/10.2307/2333849>
- Searle, S. (1982). *Matrix Algebra Useful for Statistics*. Wiley-Interscience.

- Searle, S. R. (1971). *Linear Models*. Wiley.
- Searle, S. R., & Gruber, M. H. J. (2017). *Linear Models* (Second edition). Wiley.
- Seashore, H. G. (1955). Methods of Expressing Test Scores. *Test Service Bulletin*, 48, 7–10.
- Seber, G. A. F., & Lee, A. J. (2003). *Linear Regression Analysis* (2nd ed). Wiley-Interscience.
- Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201. <https://doi.org/10.2307/1412107>
- Spivak, M. (2008). *Calculus* (4. Aufl.). Publish or Perish, Inc.
- Starke, L., & Ostwald, D. (2017). Variational Bayesian Parameter Estimation Techniques for the General Linear Model. *Frontiers in Neuroscience*, 11. <https://doi.org/10.3389/fnins.2017.00504>
- Steele, J. M. (2006). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities* (repr). Cambridge University Press [u.a.].
- Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science, New Series*, 103(2684), 677–680. <https://www.jstor.org/stable/1671815>
- Stigler, S. M. (1981). Gauss and the Invention of Least Squares. *The Annals of Statistics*, 9(3). <https://doi.org/10.1214/aos/1176345451>
- Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press.
- Strang, G. (2009). *Introduction to Linear Algebra*. Cambridge University Press.
- Student. (1908). The Probable Error of a Mean. *Biometrika*, 6(1), 1–25.
- Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, 21(153), 65–66. <https://doi.org/10.1080/01621459.1926.10502161>
- Thurstone, L. (1936). The Factorial Isolation of Primary Abilities. *Psychometrika*, 1(3), 175–182. <https://doi.org/10.1007/BF02288363>
- Thurstone, L. (1937). Current Misuse of the Factorial Methods. *Psychometrika*, 2(2), 73–76. <https://doi.org/10.1007/BF02288060>
- Thurstone, L. (1947). *Multiple Factor Analysis*. University of Chicago Press.
- Toelch, U., & Ostwald, D. (2018). Digital Open Science—Teaching Digital Tools for Reproducible and Transparent Research. *PLOS Biology*, 16(7), e2006022. <https://doi.org/10.1371/journal.pbio.2006022>
- Tong, Y. L. (1990). *Multivariate Normal Distribution*. Springer.
- Ulrich, H., Kock-Schoppenhauer, A.-K., Deppenwiese, N., Gött, R., Kern, J., Lablans, M., Majeed, R. W., Stöhr, M. R., Stausberg, J., Varghese, J., Dugas, M., & Ingenerf, J. (2022). Understanding the Nature of Metadata: Systematic Review. *Journal of Medical Internet Research*, 24(1), e25440. <https://doi.org/10.2196/25440>
- Unger, L. (2000). *Grundkurs Mathematik*.
- Uurtio, V., Monteiro, J. M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., & Rousu, J. (2018). A Tutorial on Canonical Correlation Methods. *ACM Computing Surveys*, 50(6), 1–33. <https://doi.org/10.1145/3136624>
- Vaart, A. W. van der. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Van Rooij, I., & Wareham, T. (2012). Intractability and Approximation of Optimization Theories of Cognition. *Journal of Mathematical Psychology*, 56(4), 232–247. <https://doi.org/10.1016/j.jmp.2012.05.002>
- Van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the Distribution of the Maximum Likelihood Estimator of Cronbach's Alpha. *Psychometrika*, 65(3), 271–280. <https://doi.org/10.1007/BF02296146>
- Verbyla, A. P. (1990). A Conditional Derivation of Residual Maximum Likelihood. *Australian Journal of Statistics*, 32(2), 227–230. <https://doi.org/10.1111/j.1467-842X.199>

- 0.tb01015.x
- Von Neumann, J. (1945). *First Draft of a Report on the EDVAC*. Moore School of Electrical Engineering, University of Pennsylvania. <https://doi.org/10.5479/sil.538961.39088011475779>
- Von Plato, J. (1994). *Creating Modern Probability: Its Mathematics, Physics, and Philosophy in Historical Perspective*. Cambridge University Press.
- Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities* (1st ed). Chapman and Hall.
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond „ $p < 0.05$ “. *The American Statistician*, 73(sup1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>
- Wickham, H. (2019). *Advanced R, Second Edition*. CRC Press.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3(1), 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wilks, S. S. (1932). Certain Generalizations in the Analysis of Variance. *Biometrika*, 24(3-4), 471–494. <https://doi.org/10.1093/biomet/24.3-4.471>
- Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62. <https://doi.org/10.1214/aoms/1177732360>
- Wishart, J. (1928). The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population. *Biometrika*, 20A(1/2), 32. <https://doi.org/10.2307/2331939>
- Wright, S. (1921). Correlation and Causation. *Journal of Agriculture Research*, 20(7), 557–585.
- Yanai, H., & Ichikawa, M. (2007). Factor Analysis. In *Handbook of Statistics, Vol 26* (1st ed). Elsevier.
- Yuan, K.-H., & Bentler, P. M. (2002). On Robusiness of the Normal-Theory Based Asymptotic Distributions of Three Reliability Coefficient Estimates. *Psychometrika*, 67(2), 251–259. <https://doi.org/10.1007/BF02294845>
- Yule, G. U. (1907). On the Theory of Correlation for Any Number of Variables, Treated by a New System of Notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 79(529), 182–193. <https://www.jstor.org/stable/92723>
- Zabell, S. L. (2008). On Student’s 1908 Article „The Probable Error of a Mean“. *Journal of the American Statistical Association*, 103(481), 1–7. <https://doi.org/10.1198/016214508000000030>