Wahrscheinlichkeitstheorie und Frequentistische Inferenz

Dirk Ostwald

2024-01-21

Inhaltsverzeichnis

Teil I.

Mathematische Grundlagen

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)). Gleichermaßen 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), Friston et al. (2023), Ostwald et al. (2014), Blei et al. (2017)). 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 einfache Transferierbarkeit von Erkenntnissen in andere Kontexte interessiert. Um dies zu ermöglichen, versucht die Mathematik möglichst genau und verständlich, also im Sinne präziser Begriffsbildungen zu arbeiten. Sie geht dabei insbesondere streng hierarchisch vor, so dass an späterer Stelle eingeführte Begrifflichkeiten oft ein gutes Verständnis der ihnen zugrundeliegenden und an früherer Stelle eingeführten Begrifflichkeiten voraussetzen.

Die Genauigkeit der mathematischen Sprache impliziert dabei eine hohe Informationsdichte. Sie ist daher eher nüchtern und lässt überflüssiges weg, so dass in mathematischen Texten im besten Fall *alles* für die Kommunikation einer Idee relevant ist. Als Rezipient:in mathematischer Texte nimmt man die Informationsdichte mathematischer Texte anhand des hohen Verbrauchs an kognitiver Energie beim Lesen eines 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 nachzubauen.

Möchte man sich also die Welt der naturwissenschaftliche 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 ständige Exposition mit mathematischen Ausdrucksweisen, sicherlich auch zunächst einmal das Auswendiglernen von Begriffen und das aktive Lesen 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 mathematischer Kommunikation

In diesem Abschnitt 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 eine Grundannahme eines mathematischen Systems, die 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 $\bullet$.

Theorem

Ein *Theorem* ist eine mathematische Aussage, die mittels eines Beweises als wahr (richtig) erkannt werden kann. Dass 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 auch oft als *Sätze* bezeichnet. In der angewandten, datenanalytischen Mathematik sind Theoreme oft 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 also zum Beispiel der Ausdruck "$a = 2$" in einem gegebenen Kontext, dass aufgrund bestimmter Voraussetzungen das Symbol oder die Variable a den Wert zwei hat. Theoreme enden in diesem Text immer mit dem Symbol $\circ$.

Beweis

Ein *Beweis* ist eine logische Argumentationskette, die auf bekannte Definitionen und Theoreme zurückgreift, um die Wahrheit (Richtigkeit) 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 entsprechende mathematische Gebäude abzukommen. Beweise enden in diesem Text immer mit dem Symbol □.

Neben den oben vorgestellten Begriffen gibt es mit *Axiomen*, *Lemmata*, *Korollaren* und *Vermutungen* noch weitere typische Grundbausteine mathematischer Texte. Wir werden diese Begriff 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 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 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 Modellbildung kennengelernt haben, wollen wir uns mit der *Aussagenlogik* einem einfachem System nähern, das es erlaubt, Beziehungen zwischen mathematischen Aussagen herzustellen und zu formalisieren. Im Folgenden spielt die Aussagenlogik zum Beispiel in der Definition von Mengenoperationen, bei Optimierungsbedingungen von Funktionen und in vielen Beweisen einen tragende Rolle. In der mathematischen Anwendung ist Aussagenlogik die Grundlage der Booleschen Logik der Programmierung. In der mathematischen Psychologie ist die Aussagenlogik zum Beispiel 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.

A	$\neg A$
w	f
f	w

Tabellen dieser Form nennt man *Wahrheitstafeln*. Sie sind ein beliebtes Hilfsmittel in der Aussagenlogik. Möchte man zwei Aussagen logisch verbinden, so bieten sich zunächst die Begriffe der *Konjunktion* und *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.

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 nun 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

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 wiederrum 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".

•

Man mag $\Rightarrow$ auch als "daraus folgt" lesen. Die Definition der Implikation impliziert folgende Wahrheitstafel.

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

Ein 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. Betrachtet man obige Wahrheitstafel unter diesem Gesichtspunkt, 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 (dass also gilt, wenn A wahr ist, dass

$A \Rightarrow B$ 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. Aus einer wahren Voraussetzung folgt also nur bei wahrer Implikation eine wahre Konklusion. Insbesondere genügt die Definition der Implikation damit der Forderung "Aus Falschem folgt beliebiges (ex falso sequitur quodlibet)". Aus falschen Aussagen kann man also mithilfe der Implikation nichts richtiges folgern.

Im Kontext der Implikation ergeben sich die Begriffe der *hinreichenden* und der *notwendigen Aussagen (Bedingungen)*. Diese sind definiert wie folgt: wenn $A \Rightarrow B$ wahr ist, sagt man "A ist *hinreichend* für B" und "B ist *notwendig* für A". Diese Sprachregelung erklärt sich 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.

Eine sehr häufig autretender 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

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 Wahrheitstafeln gleich sind.

•

Als Beispiele für logische Äquivalenzen, die häufig in Beweisargumentationen genutzt werden, zeigen wir folgendes Theorem.

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 Begriffs der logischen Äquivalenz müssen wir zeigen, dass die Wahrheitstafeln 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 $A \Leftrightarrow$ 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.

□

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 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*.

1.4. Beweistechniken

Im letzten Abschnitt wollen wir mit den Begriffen der *direkten* und *indirekten Beweise* sowie des *Beweises durch Widerspruch* kurz drei Beweistechniken skizzieren, von denen vor allem die erste in diesem Text immer wieder zur Begründung von Theoremen herangezogen wird. Dabei haben typische Theoreme die Form $A \Rightarrow B$ für Aussagen A und B.

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.

Um diese Techniken an einem Beispiel zu erläutern, erinnern wir kurz an folgende *Äquivalenzumformungen von Gleichungen*:

- Addition oder Subtraktion einer Zahl auf beiden Seiten der Gleichung, zum Beispiel

$$2x + 4 = 10 \Leftrightarrow 2x = 6, \tag{1.3}$$

- Multiplikation mit einer oder Division durch eine von Null verschiedene Zahl auf beiden Seiten der Gleichung, zum Beispiel

$$2x = 6 \Leftrightarrow x = 3, \tag{1.4}$$

- Anwendung einer injektiven Funktion auf beiden Seiten der Gleichung, zum Beispiel

$$\exp(x) = 2 \Leftrightarrow x = \ln(2), \tag{1.5}$$

sowie an folgende elementaren *Äquivalenzumformungen von Ungleichungen*:

- Addition oder Subtraktion einer Zahl auf beiden Seiten der Ungleichung, zum Beispiel

$$-2x + 4 \geq 10 \Leftrightarrow -2x \geq 6, \tag{1.6}$$

- Multiplikation mit einer Zahl oder Division durch eine von Null verschiedene Zahl auf beiden Seiten der Ungleichung, wobei die Multiplikation oder Division mit einer negativen Zahl die Umkehrung der Ungleichung impliziert, zum Beispiel

$$-2x \geq 6 \Leftrightarrow x \leq -3, \tag{1.7}$$

- Anwendung monotoner Funktionen auf beiden Seiten der Ungleichung

$$\exp(x) \geq 2 \Leftrightarrow x \geq \ln(2). \tag{1.8}$$

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, dazu, 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

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

□

1.5. Selbstkontrollfragen

1. Erläutern Sie die Besonderheiten der mathematischen Sprache.
2. Was sind wesentliche Tätigkeiten zum Erlernen einer Sprache?
3. Erläutern Sie den Begriff der Definition.
4. Erläutern Sie den Begriff des Theorems.
5. Erläutern Sie den Begriff des Beweises.
6. Geben Sie die Definition einer mathematischen Aussage wieder.
7. Geben Sie die Definition der Negation einer mathematischen Aussage wieder.
8. Geben Sie die Definition der Konjunktion zweier mathematischer Aussagen wieder.
9. Geben Sie die Definition der Disjunktion zweier mathematischer Aussagen wieder.
10. Geben Sie die Definition der Implikation wieder.
11. Geben Sie die Definition der Äquivalenz wieder.
12. Geben Sie die Definition der logischen Äquivalenz wieder.
13. Erläutern Sie die Begriffe des direkten Beweises, des indirekten Beweises und des Beweises durch Widerspruch.
14. Beweisen Sie, dass gilt

$$x^2 + px + q = 0 \Leftrightarrow x = -\frac{p}{2} - \sqrt{\left(\frac{p}{2}\right)^2 - q} \vee x = -\frac{p}{2} + \sqrt{\left(\frac{p}{2}\right)^2 - q}. \tag{1.14}$$

2. Mengen

2.1. Grundlegende Definitionen

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

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

$$m \in M \text{ bzw. } m \notin M \tag{2.1}$$

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 definieren 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, dass heißt die Reihenfolge der Auflistung der Elemente einer Menge spielt keine Rolle. 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)**.** A und B seien zwei Mengen.

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

  $$A \subseteq B \tag{2.2}$$

  und nennt A *Untermenge* von B und B *Obermenge* von A.
- Eine Menge A heißt *echte Teilmenge* einer Menge B, wenn für jedes Element $a \in A$ gilt, dass auch $a \in B$, es aber zumindest ein Element $b \in B$ gibt, für das gilt $b \notin A$. Ist A eine echte Teilmenge von B, so schreibt man

  $$A \subset B. \tag{2.3}$$

- Zwei Mengen A und B heißen *gleich*, wenn für jedes Element $a \in A$ gilt, dass auch $a \in B$, und wenn für jedes Element $b \in B$ gilt, dass auch $b \in A$. Sind die Mengen A und B gleich, so schreibt man
$$A = B. \tag{2.4}$$

•

Betrachten wir zum Beispiel die Mengen $A := \{1\}$, $B := \{1, 2\}$, und $C := \{1, 2\}$. Dann gilt mit obigen Definitionen, dass $A \subset B$, weil $1 \in A$ und $1 \in B$, aber $2 \in B$ und $2 \notin A$. Weiterhin gilt, dass $B \subseteq C$, weil $1 \in B$ und $1 \in C$ sowie $2 \in B$ und $2 \in C$ und es kein Element von C gibt, welches nicht in B ist. Ebenso gilt $C \subseteq B$, weil $1 \in C$ und $1 \in B$ sowie $2 \in C$ und $2 \in B$ und es kein Element von B gibt, welches nicht in C ist. Schließlich gilt sogar $B = C$, weil für jedes Element $b \in B$ gilt, dass auch $b \in C$, und gleichzeitig für jedes Element $c \in C$ gilt, dass auch $c \in B$.

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 $A := \{1, 2, 3\}$, $B = \{a, b, c, d\}$ und $C := \{\,\}$. Dann gelten $|A| = 3$, $B = 4$ und $|C| = 0$.

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 vier Beispiele zum Begriff der Potenzmenge.

- $M_0 := \emptyset$ sei die leere Menge. Dann gilt
$$\mathcal{P}(M_0) = \emptyset. \tag{2.5}$$

- M_1 sei die einelementige Menge $M_1 := \{a\}$. Dann gilt
$$\mathcal{P}(M_1) = \{\emptyset, \{a\}\}. \tag{2.6}$$

- Es sei $M_2 := \{a, b\}$. Dann hat M_2 sowohl ein- als auch zweielementige Teilmengen und es gilt
$$\mathcal{P}(M_2) = \{\emptyset, \{a\}\,\{b\}, \{a, b\}\}. \tag{2.7}$$

- Schließlich sei $M_3 := \{a, b, c\}$. Dann hat M ein-, zwei-, als auch dreielementige Teilmengen und es gilt
$$\mathcal{P}(M_3) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}. \tag{2.8}$$

Hinsichtlich der Kardinalitäten einer Menge und ihrer Potenzmenge kann man beweisen, dass aus $|M| = n$ mit $n > 0$ folgt, dass die Kardinalität der Potenzmenge $|\mathcal{P}(M)| = 2^n$ ist. In den obigen Beispielen haben wir die Fälle $|M_1| = 1$ und somit $|\mathcal{P}(M_1)| = 2^1 = 2$, $|M_2| = 2$ und somit $|\mathcal{P}(M_3)| = 2^2 = 4$ und schließlich $|M_3| = 3$ und somit $|\mathcal{P}(M_3)| = 2^3 = 8$, wovon man sich durch Nachzählen schnell überzeugt.

2.2. Verknüpfungen von Mengen

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
$$M \cup N := \{x | x \in M \vee x \in N\}, \tag{2.9}$$
wobei $\vee$ wie immer im inklusiven Sinne als und/oder zu verstehen ist.

- Der *Durchschnitt von M und N* ist definiert als die Menge
$$M \cap N := \{x | x \in M \wedge x \in N\}. \tag{2.10}$$
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
$$M \setminus N := \{x | x \in M \wedge x \notin N\}. \tag{2.11}$$
Die Differenz M und N heißt, insbesondere bei $M \subseteq N$, auch das *Komplement von N bezüglich M* und wird mit N^c bezeichnet.

- Die *symmetrische Differenz von M und N* ist definiert als die Menge
$$M \Delta N := \{x | (x \in M \vee x \in N) \wedge x \notin M \cap N\}, \tag{2.12}$$
Die symmetrische Differenz kann also als *exklusives oder* verstanden werden.

•

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 $5 \notin \{2, 3\}$.

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

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

$$M = \cup_{i=1}^{n} N_i \wedge N_i \cap N_j = \emptyset \text{ für } i = 1, ..., n, j = 1, ..., n, i \neq j. \tag{2.13}$$

Dann heißt P eine *Partition von* M.

•

Intuitiv entspricht die Partition einer Menge also dem Aufteilen der Menge in disjunkte Teilmengen. Partitionen sind generell nicht eindeutig, d.h. es gibt meist verschiedene Möglichkeiten eine gegebene Menge zu partitionieren. Betrachten wir zum Beispiel 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 mögliche Partitionen von M.

2.3. Spezielle Mengen

In der Naturwissenschaft versucht man, Phänomene der Welt mit Zahlen zu beschreiben. Je nach Phänomen bieten sich dazu *diskrete* oder *kontinuierliche* Zahlenmengen an. Die Mathematik stellt dazu unter anderem die in folgender Definition gegebenen Zahlenmengen bereit.

Definition 2.8 (Zahlenmengen)**.** Es bezeichnen

- $\mathbb{N} := \{1, 2, 3, ...\}$ die *natürlichen Zahlen*,
- $\mathbb{N}_n := \{1, 2, 3, ..., n\}$ die *natürlichen Zahlen der Ordnung* n,
- $\mathbb{N}^0 := \mathbb{N} \cup \{0\}$ die *natürlichen Zahlen* und Null,
- $\mathbb{Z} := \{..., -3, -2, -1, 0, 1, 2, 3...\}$ die *ganzen Zahlen*,
- $\mathbb{Q} := \{\frac{p}{q} | p, q \in \mathbb{Z}, q \neq 0\}$ 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 ganzen Zahlen eignen sich insbesondere 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, wie oben definiert, durch Brüche ganzer Zahlen ausdrücken lassen. Dies sind alle ganzen Zahlen sowie die negativen und positiven Dezimalzahlen wie z.B. $-\frac{9}{10} = -0.9$, $\frac{1}{3} = 1.3\overline{3}$, 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 kann man beweisen (Cantor (1892)), 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. Insbesondere gilt

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}. \tag{2.14}$$

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 kann. Sie gehören auch nicht zu den in obiger Definition gegebenen Zahlenmengen, es gilt also sowohl $\infty \notin \mathbb{R}$ als auch $-\infty \notin \mathbb{R}$.

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 insbesondere bei der Modellierung physikalischer Phänomene und im Bereich der Fourieranalyse zum Einsatz. Wir vertiefen die Theorie komplexer Zahlen an dieser Stelle nicht.

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

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

- das *abgeschlossene Intervall*

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

- das *offene Interval*

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

- und die *halboffenen Intervalle*

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

•

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 $\mathbb{R} =]-\infty, \infty[$.

Oft möchte man mehrere 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 auch *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

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

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

$$M^2 := M \times M. \tag{2.19}$$

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

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

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

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

•

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

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* 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

$$\mathbb{R}^n := \prod_{i=1}^{n} \mathbb{R} := \{x := (x_1, , ..., x_n) | x_i \in \mathbb{R}\} \tag{2.22}$$

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

$$x := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{2.23}$$

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

•

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

$$x = \begin{pmatrix} 0.16 \\ 1.76 \\ 0.23 \\ 7.11 \end{pmatrix}. \tag{2.24}$$

2.4. Selbstkontrollfragen

1. Geben Sie die Definition einer Menge nach Cantor (1895) wieder.
2. Nennen Sie drei Möglichkeiten zur Definition einer Menge.
3. Erläutern Sie die Ausdrücke $m \in M, m \notin N, M \subseteq N, M \subset N$ für zwei Mengen M und N.
4. Geben Sie die Definition der Kardinalität einer Menge wieder.
5. Geben Sie die Definition der Potenzmenge einer Menge wieder.
6. Es sei $M := \{1, 2\}$. Bestimmen Sie $\mathcal{P}(M)$.
7. Es seien $M := \{1, 2\}, N := \{1, 4, 5\}$. Bestimmen Sie $M \cup N, M \cap N, M \setminus N, M \Delta N$.
8. Erläutern Sie die Symbole $\mathbb{N}$, $\mathbb{N}_n$, und $\mathbb{N}^0$.
9. Erläutern Sie die Unterschiede zwischen $\mathbb{N}$ und $\mathbb{Z}$ und zwischen $\mathbb{R}$ und $\mathbb{Q}$.
10. Geben Sie die Definition abgeschlossener, offener, und halboffener Intervalle wieder.
11. Es seien M und N Mengen. Erläutern Sie die Notation $M \times N$.
12. Geben Sie die Definition von $\mathbb{R}^n$ wieder.

3. Summen, Produkte, Potenzen

3.1. Summen

Diese Einheit führt einige Schreibweisen für die Grundrechenarten ein.

Definition 3.1 (Summenzeichen)**.** Es bezeichnet

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

Dabei stehen

- Σ für das griechische *S*igma, mnemonisch für*S*umme,
- das Subskript $i = 1$ für den *Laufindex* und den *Startindex*,
- das Superskript n für den *Endindex* und
- $x_1, x_2, ..., x_n$ für die *Summanden*.

•

Für die sinnvolle Benutzung des Summenzeichens ist es essentiell, dass mit mithilfe des Subskripts und des Superskripts Anfang und Ende der Summation festgelegt werden. Die genaue Bezeichnung des Laufindexes ist dagegen für den Wert der Summe irrelevant, es gilt

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j. \tag{3.2}$$

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

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

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
$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 2 + 10 - 4 = 8. \tag{3.4}$$
- *Summation natürlicher Zahlen.* Es gilt
$$\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5 = 15. \tag{3.5}$$

- *Summation gerader natürlicher Zahlen.* Es gilt

$$\sum_{i=1}^{5} 2i = 2 \cdot 1 + 2 \cdot 2 + 2 \cdot 3 + 2 \cdot 4 + 2 \cdot 5 = 2 + 4 + 6 + 8 + 10 = 30. \tag{3.6}$$

- *Summation ungerader natürlicher Zahlen.* Es gilt

$$\sum_{i=1}^{5} (2i-1) = 2 \cdot 1 - 1 + 2 \cdot 2 - 1 + 2 \cdot 3 - 1 + 2 \cdot 4 - 1 + 2 \cdot 5 - 1 = 1 + 3 + 5 + 7 + 9 = 25. \tag{3.7}$$

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

Theorem 3.1 (Rechenregeln für Summen)**.**

(1) Summen gleicher Summanden

$$\sum_{i=1}^{n} x = nx \tag{3.8}$$

(2) Assoziativität bei Summen gleicher Länge

$$\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (x_i + y_i) \tag{3.9}$$

(3) Distributivität bei Multiplikation mit einer Konstante

$$\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i \tag{3.10}$$

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

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i \tag{3.11}$$

(5) Umindizierung

$$\sum_{i=0}^{n} x_i = \sum_{j=m}^{n+m} x_{j-m} \tag{3.12}$$

∘

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

$$\begin{aligned} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i &= x_1 + x_2 + \cdots + x_n + y_1 + y_2 + \cdots + y_n \\ &= x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n \\ &= \sum_{i=1}^{n} (x_i + y_i). \end{aligned} \tag{3.13}$$

Hinsichtlich letzterer gilt

$$\begin{aligned}\sum_{i=1}^{n} ax_i &= ax_1 + ax_2 + \cdots + ax_n \\ &= a(x_1 + x_2 + \cdots + x_n) \\ &= a\sum_{i=1}^{n} x_i.\end{aligned} \tag{3.14}$$

□

Als Beispiel für die Anwendung einer Rechenregel betrachten wir die Auswertung eines *Mittelwertes* (manchmal auch *Durchschnitt* genannt). Dazu seien $x_1, x_2, ..., x_n$ reelle Zahlen. Der Mittelwert dieser Zahlen entspricht der Summe von $x_1, x_2, ..., x_n$ geteilt durch die Anzahl der Zahlen n. Dabei ist es nach obiger Rechenregel (3) 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 Ergebenisse dann aufaddiert werden. Genauer gilt durch Anwendung von Rechenregel (3) mit $a = 1/n$, dass

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{x_i}{n}. \tag{3.15}$$

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

$$\frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1}{4}(1+4+2+1) = \frac{8}{4} = 2 = \frac{8}{4} = \frac{1}{4} + \frac{4}{4} + \frac{2}{4} + \frac{1}{4} = \sum_{i=1}^{4} \frac{x_i}{4}. \tag{3.16}$$

3.2. Produkte

Eine analoge Schreibweise zum Summenzeichen bietet das Produktzeichen für Produkte.

Definition 3.2 (Produktzeichen)**.** Es bezeichnet

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \cdots \cdot x_n. \tag{3.17}$$

Dabei stehen

- $\prod$ für das griechische *P*i, mnemonisch für *P*rodukt,
- das Subskript $i = 1$ für den *Laufindex* und den *Startindex*,
- das Superskript n für den *Endindex*,
- $x_1, x_2, ..., x_n$ 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 Laufindizes ist wiederum irrelevant, es gilt

$$\prod_{i=1}^{n} x_i = \prod_{j=1}^{n} x_j. \tag{3.18}$$

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

$$\prod_{j \in J} x_j := x_0 \cdot x_1 \cdot x_2. \tag{3.19}$$

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

$$a^0 := 1 \text{ und } a^{n+1} := a^n \cdot a. \tag{3.20}$$

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

$$a^{-n} := (a^n)^{-1} := \frac{1}{a^n}. \tag{3.21}$$

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 3.2 (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:*

$$a^n a^m = a^{n+m} \tag{3.22}$$
$$(a^n)^m = a^{nm} \tag{3.23}$$
$$(ab)^n = a^n b^n \tag{3.24}$$

∘

Wir verzichten auf einen Beweis. Beispielsweise gelten also

$$2^2 \cdot 2^3 = (2 \cdot 2) \cdot (2 \cdot 2 \cdot 2) = 2^5 = 2^{2+3}, \tag{3.25}$$

$$(3^2)^3 = (3 \cdot 3)^3 = (3 \cdot 3) \cdot (3 \cdot 3) \cdot (3 \cdot 3) = 3^6 = 3^{2 \cdot 3}, \tag{3.26}$$

und

$$(2 \cdot 4)^2 = (2 \cdot 4) \cdot (2 \cdot 4) = (2 \cdot 2) \cdot (4 \cdot 4) = 2^2 \cdot 4^2. \tag{3.27}$$

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 Zahl r, so dass

$$r^n = a. \tag{3.28}$$

•

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

Theorem 3.3 (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*

$$r = a^{\frac{1}{n}} \tag{3.29}$$

∘

Beweis. Es gilt

$$\left(a^{\frac{1}{n}}\right)^n = a^{\frac{1}{n}} \cdot a^{\frac{1}{n}} \cdot \cdots \cdot a^{\frac{1}{n}} = a^{\sum_{i=1}^{n} \frac{1}{n}} = a^1 = a. \tag{3.30}$$

Also gilt mit der Definition der n-ten Wurzel, dass $r = a^{\frac{1}{n}}$.

□

Das Rechnen mit Quadratwurzeln wird durch die Potenzschreibweise $\sqrt{x} = x^{\frac{1}{2}}$ sehr erleichtert. Zum Beispiel gilt

$$\frac{2\pi}{\sqrt{2\pi}} = \frac{2\pi}{(2\pi)^{\frac{1}{2}}} = (2\pi)^1 \cdot (2\pi)^{-\frac{1}{2}} = (2\pi)^{1-\frac{1}{2}} = (2\pi)^{\frac{1}{2}} = \sqrt{2\pi}. \tag{3.31}$$

3.4. Selbstkontrollfragen

1. Geben Sie die Definition des Summenzeichens wieder.
2. Berechnen Sie die Summen $\sum_{i=1}^{3} 2$, $\sum_{i=1}^{3} i^2$, und $\sum_{i=1}^{3} \frac{2}{3} i$.
3. Schreiben Sie die Summe $1 + 3 + 5 + 7 + 9 + 11$ mithilfe des Summenzeichens.
4. Schreiben Sie die Summe $0 + 2 + 4 + 6 + 8 + 10$ mithilfe des Summenzeichens.
5. Geben Sie die Definition des Produktzeichens wieder.
6. Geben Sie die Definition der n-ten Potenz von $a \in \mathbb{R}$ wieder.
7. Berechnen Sie $2^2 \cdot 2^3$ und 2^5 und geben Sie die zugehörige Potenzregel wieder.
8. Berechnen Sie 6^2 und $2^2 \cdot 3^2$ und geben Sie die zugehörige Potenzregel wieder.
9. Begründen Sie, warum die n-te Wurzel von a als $a^{\frac{1}{n}}$ geschrieben werden kann.
10. Berechnen Sie $(\sqrt{2})^{\frac{2}{3}}, 9^{\frac{1}{2}}$, und $4^{-\frac{1}{2}}$.

4. Funktionen

Funktionen bilden zusammen mit den Mengen die Grundpfeiler mathematischer Modellierung. In dieser Einheit definieren wir den Begriff der Funktion, führen erste Eigenschaften von Funktionen ein und geben eine Übersicht über einige elementare Funktionen. Funktionen werden äquivalent auch als Abbildungen bezeichnet.

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 Zielmenge Z zuordnet. D wird dabei *Definitionsmenge von* f und Z wird *Zielmenge von* f genannt. Wir schreiben

$$f : D \to Z, x \mapsto f(x), \tag{4.1}$$

wobei $f : D \to Z$ 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 $\to$ 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

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

die Definition der Potenz genutzt.

Funktionen sind immer *eindeutig*, in dem Sinne dass sie jedem $x \in D$ bei jeder Anwendung der Funktion immer dasselbe $f(x) \in Z$ zuordnen. Funktionen setzen dabei Elemente von Mengen miteinander in Beziehung. Die Mengen dieser Elemente erhalten spezielle Bezeichnungen.

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

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

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

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

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

•

Man beachte, dass der Wertebereich $f(D)$ von f und die Zielmenge Z von f sind nicht notwendigerweise identisch sein müssen. Grundlegende Eigenschaften von Funktionen werden in folgender Definition festgelegt.

Definition 4.3 (Injektivität, Surjektivität, Bijektivität)**.** $f : D \to Z, x \mapsto f(x)$ 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.

•

Abbildung 4.1 verdeutlicht diese Definitionen anhand dreier (Gegen)beispiele.

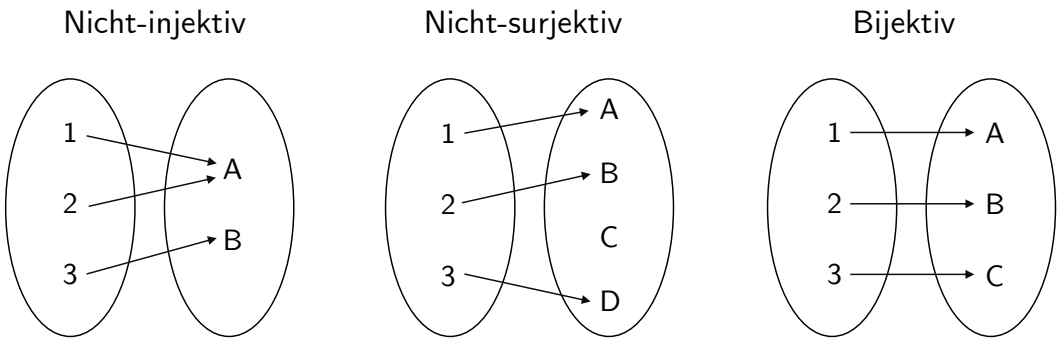


Abbildung 4.1. Injektivität, Surjektivität, Bijektivität.

Abbildung 4.1 A visualisiert die *nicht-injektive* Funktion

$$f : \{1,2,3\} \to \{A,B\}, x \mapsto f(x) := \begin{cases} f(1) & := A \\ f(2) & := A \\ f(3) & := B \end{cases}. \tag{4.5}$$

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 1 und 2.

Abbildung 4.1 B visualisiert die nicht-surjektive Funktion

$$g : \{1,2,3\} \to \{A,B,C,D\}, x \mapsto g(x) := \begin{cases} g(1) & := A \\ g(2) & := B \\ g(3) & := D \end{cases}. \tag{4.6}$$

Die Funktion ist nicht surjektiv, weil das Element D in der Zielmenge von f kein Urbild in der Definitionsmenge von f hat. Abbildung 4.1 C schließlich visualisiert die bijektive Funktion

$$h : \{1,2,3\} \to \{A,B,C\}, x \mapsto g(x) := \begin{cases} h(1) & := A \\ h(2) & := B \\ h(3) & := C \end{cases}. \tag{4.7}$$

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

Als weiteres Beispiel betrachten wir die Funktion

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := x^2 \tag{4.8}$$

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

$$\tilde{f} : [0,\infty[\to [0,\infty[, x \mapsto \tilde{f}(x) := x^2, \tag{4.9}$$

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

4.2. Funktionentypen

Durch Verkettung lassen sich aus Funktionen weitere Funktionen bilden.

Definition 4.4 (Verkettung von Funktionen)**.** Es seien $f : D \to Z$ und $g : Z \to S$ zwei Funktionen, wobei die Wertemenge von f mit der Definitionsmenge von g übereinstimmen sollen. Dann ist durch

$$g \circ f : D \to S, x \mapsto (g \circ f)(x) := g(f(x)) \tag{4.10}$$

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 verkette Funktion und $(g \circ f)(x)$ ein Element in der Zielmenge der verketten 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.

Als Beispiel für die Verkettung zweier Funktionen betrachten wir

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := -x^2 \tag{4.11}$$

und

$$g : \mathbb{R} \to \mathbb{R}, x \mapsto g(x) := \exp(x). \tag{4.12}$$

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

$$g \circ f : \mathbb{R} \to \mathbb{R}, x \mapsto (g \circ f)(x) := g(f(x)) = \exp\left(-x^2\right). \tag{4.13}$$

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

Definition 4.5 (Inverse Funktion)**.** Es sei $f : D \to Z, x \mapsto f(x)$ eine bijektive Funktion. Dann heißt die Funktion f^{-1} mit

$$f^{-1} \circ f : D \to D, x \mapsto (f^{-1} \circ f)(x) := f^{-1}(f(x)) = x \tag{4.14}$$

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. Betrachtet man 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. Betrachten wir zum Beispiel die Funktion

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := 2x =: y. \tag{4.15}$$

Dann ist die inverse Funktion von f gegeben durch

$$f^{-1} : \mathbb{R} \to \mathbb{R}, y \mapsto f^{-1}(y) := \frac{1}{2}y, \tag{4.16}$$

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

$$(f^{-1} \circ f)(x) := f^{-1}(f(x)) = f^{-1}(2x) = \frac{1}{2} \cdot 2x = x. \tag{4.17}$$

Eine wichtige Klasse von Funktionen sind *lineare Abbildungen*.

Definition 4.6 (Lineare Abbildung)**.** Eine Abbildung $f : D \to Z, x \mapsto f(x)$ heißt *lineare Abbildung*, wenn für $x, y \in D$ und einen Skalar c gelten, dass

$$f(x + y) = f(x) + f(y) f(cx) = cf(x) \qquad \text{(Additivität)}$$

und

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

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)**.** *$f : D \to Z$ sei eine lineare Abbildung. Dann gilt*

$$f(0) = 0. \tag{4.18}$$

∘

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

$$f(0) = f(0 + 0) = f(0) + f(0). \tag{4.19}$$

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

$$\begin{aligned} f(0) - f(0) &= f(0) + f(0) - f(0) \\ 0 &= f(0) \end{aligned} \tag{4.20}$$

und damit ist alles gezeigt.

□

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

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

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := ax \tag{4.21}$$

  eine lineare Abbildung, weil gilt, dass

$$f(x+y) = a(x+y) = ax+ay = f(x)+f(y) \text{ und } f(cx) = acx = cax = cf(x). \tag{4.22}$$

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

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := ax + b \tag{4.23}$$

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

$$f(x + y) = 1(x + y) + 1 = x + y + 1 \neq x + 1 + y + 1 = f(x) + f(y). \tag{4.24}$$

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 4.7 (Funktionenarten)**.** Wir unterscheiden

- *univariate reellwertige Funktionen* der Form

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x), \tag{4.25}$$

- *multivariate reellwertige Funktionen* der Form

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) = f(x_1, ..., x_n), \tag{4.26}$$

- und *multivariate vektorwertige Funktionen* der Form

$$f : \mathbb{R}^n \to \mathbb{R}^m, x \mapsto f(x) = \begin{pmatrix} f_1(x_1, ..., x_n) \\ \vdots \\ f_m(x_1, ..., x_n) \end{pmatrix}, \tag{4.27}$$

  wobei $f_i, i = 1, ..., m$ 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 zum Beispiel auch *matrixvariate 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 wesentliche Eigenschaften dieser Funktionen und ihre Graphen an. Für Beweise der Eigenschaften der hier vorgestellten F unktionen verweisen wir auf die weiterführende Literatur.

Definition 4.8 (Polynomfunktionen)**.** Eine Funktion der Form

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := \sum_{i=0}^{k} a_i x^i = a_0 + a_1 x^1 + a_2 x^2 + \cdots + a_k x^k \tag{4.28}$$

heißt *Polynomfunktion* k-ten Grades mit *Koeffizienten* $a_0, a_1, ..., a_k \in \mathbb{R}$.

•

Einige ausgewählte Polynomfunktionen sind in Tabelle 4.1 aufgelistet, Abbildung 4.2 zeigt die enstprechende Graphen.

Tabelle 4.1. Ausgewählte Polynomfunktionen

Name	Funktionale 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	$f(x) = x^2$	$a_0 := 0, a_1 := 0, a_2 := 1, a_i := 0, i > 2$

Ein wichtiges Funktionenpaar sind die Exponentialfunktion und die Logarithmusfunktion. Die Graphen der Exponential- und Logarithmusfunktion sind in Abbildung 4.3 abgebildet.

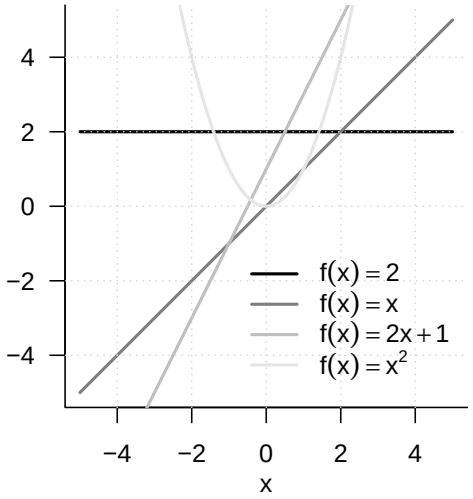


Abbildung 4.2. Ausgewählte Polynomfunktionen

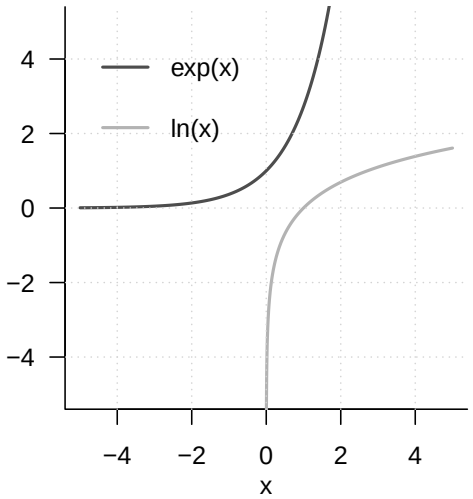


Abbildung 4.3. Exponentialfunktion und Logarithmusfunktion

Definition 4.9 (Exponentialfunktion)**.** Die *Exponentialfunktion* ist definiert als

$$\exp : \mathbb{R} \to \mathbb{R}, x \mapsto \exp(x) := e^x := \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots . \tag{4.29}$$

•

Die Exponentialfunktion hat unter anderem folgende Eigenschaften.

Wertebereich der Exponentialfunktion

- $x \in]-\infty, 0[\Rightarrow \exp(x) \in]0, 1[$
- $x \in]0, \infty[\quad \Rightarrow \exp(x) \in]1, \infty[$

Insbesondere nimmt die Exponentialfunktion also nur positive Werte an.

Monotonieeigenschaft der Exponentialfunktion

- $x < y \Rightarrow \exp(x) < \exp(y)$

Spezielle Werte der Exponentialfunktion

- $\exp(0) = 1$
- $\exp(1) = e \approx 2.71$

Die Logarithmusfunktion schneidet die y-Achse also bei 0. Die Zahl e heißt *Eulersche Zahl*.

Summationseigenschaft und Subtraktionseigenschaft der Exponentialfunktion

- $\exp(x + y) = \exp(x) \exp(y)$
- $\exp(x - y) = \frac{\exp(x)}{\exp(y)}$

Mit den speziellen Werten der Exponentialfunktion gilt dann insbesondere auch

$$\exp(x) \exp(-x) = \exp(x - x) = \exp(0) = 1. \tag{4.30}$$

Definition 4.10 (Logarithmusfunktion)**.** Die *Logarithmusfunktion* ist definiert als inverse Funktion der Exponentialfunktion,

$$\ln :]0, \infty[\to \mathbb{R}, x \mapsto \ln(x) \text{ mit } \ln(\exp(x)) = x \text{ für alle } x \in \mathbb{R}. \tag{4.31}$$

•

Die Logarithmusfunktion hat unter anderem folgende Eigenschaften.

Wertebereich der Logarithmusfunktion

- $x \in]0, 1[\Rightarrow \ln(x) \in]-\infty, 0[$
- $x \in]1, \infty[\Rightarrow \ln(x) \in]0, \infty[$

Die Logarithmusfunktion nimmt also sowohl negative als auch positive Werte an.

Monotonie der Logarithmusfunktion

- $x < y \Rightarrow \ln(x) < \ln(y)$

Spezielle Werte der Logarithmusfunktion

- $\ln(1) = 0$ und $\ln(e) = 1$.

Die Logarithmusfunktion schneidet die x-Achse also bei 1.

Produkteigenschaft, Potenzeigenschaft und Divisionseigenschaft der Logarithmusfunktion

- $\ln(xy) = \ln(x) + \ln(y)$
- $\ln(x^c) = c \ln(x)$
- $\ln\left(\frac{1}{x}\right) = -\ln(x)$

Letztere Eigenschaft sind beim Rechnen Logarithmusfunktionen zentral. Man merkt sie sich intuitiv als "Die Logarithmusfunktion wandelt Produkte in Summen und Potenzen in Produkte um."

Ein häufiger Begleiter in der Wahrscheinlichkeitstheorie ist die *Gammafunktion*. Ein Auschnitt des Graphen der Gammafunktion ist in Abbildung 4.4 dargestellt.

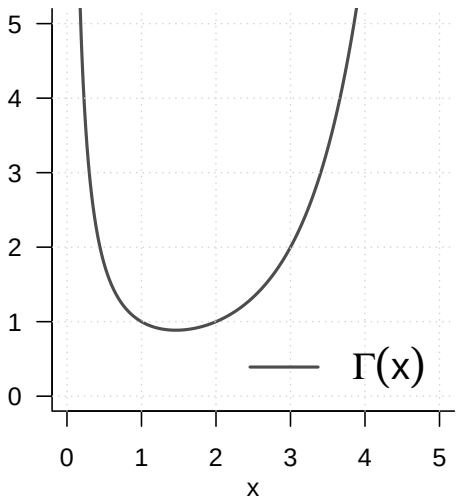


Abbildung 4.4. Gammafunktion

Definition 4.11 (Gammafunktion)**.** Die *Gammafunktion* ist definiert durch

$$\Gamma : \mathbb{R} \to \mathbb{R}, x \mapsto \Gamma(x) := \int_0^\infty \xi^{x-1} \exp(-\xi)\, d\xi \tag{4.32}$$

•

Die Gammafunktion hat folgende Eigenschaften:

Spezielle Werte der Gammafunktion

- $\Gamma(1) = 1$
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
- $\Gamma(n) = (n-1)!$ für $n \in \mathbb{N}$.

Rekursionseigenschaft der Gammafunktion

- Für $x > 0$ gilt $\Gamma(x+1) = x\Gamma(x)$

4.4. Selbstkontrollfragen

1. Geben Sie die Definition einer Funktion wieder.
2. Geben Sie die Definition der Begriffe Bildmenge, Wertebereich, und Urbildmenge wieder.
3. Geben Sie die Definitionen der Begriffe Surjektivität, Injektivität, und Bijektivität wieder.
4. Erläutern Sie, warum $f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := x^2$ weder injektiv noch surjektiv ist.
5. Erläutern Sie, warum $f : [0, \infty[\to [0, \infty[, x \mapsto f(x) := x^2$ bijektiv ist.
6. Geben Sie die Definition der Verkettung von Funktionen wieder.
7. Geben Sie die Definition des Begriffs der inversen Funktion wieder.
8. Geben Sie die inverse Funktion von x^2 auf $[0, \infty[$ an.
9. Geben Sie die Definition des Begriffs der linearen Abbildung wieder.
10. Geben Sie die Definitionen der Begriffe der univariat-reellwertigen, multivariat-reellwertigen und multivariat-vektorwertigen Funktion wieder.
11. Skizzieren Sie die Identitätsfunktion und die konstante Funktion für $a := 1$.
12. Skizzieren Sie die linear-affine Funktion $f(x) = ax + b$ für $a = 2$ und $b = 3$.
13. Skizzieren Sie die Funktionen $f(x) := (x-1)^2$ und $g(x) := (x+3)^2$.
14. Skizzieren Sie die Exponential- und Logarithmusfunktionen.
15. Geben Sie die Summations- und Subtraktionseigenschaften der Exponentialfunktion an.
16. Geben Sie die Produkt-, Potenz- und Divisionseigenschaften der Logarithmusfunktion an.

5. Differentialrechnung

Die Differentialrechnung befasst sich mit der Änderung von Funktionen. Sie bildet einerseits die Grundlage für die mathematische Modellierung mithilfe von *Differentialgleichungen*, also der Beschreibung von Funktionen anhand ihrer Änderungsraten. Zum anderen bildet die Differentialrechnung die Grundlage der *Optimierung*, also des Bestimmens von Extremstellen von Funktionen. In Kapitel 5.1 führen wir zunächst den Begriff der Ableitung und mit ihm verbundene elementare Rechenregeln ein. In Kapitel 5.2 widmen wir uns dann der Frage, wie man mithilfe von Ableitungen Extremstellen von Funktionen bestimmen kann.

5.1. Definitionen und Rechenregeln

Wir beginnen mit folgender Definition.

Definition 5.1 (Differenzierbarkeit und Ableitung)**.** Es sei $I \subseteq \mathbb{R}$ ein Intervall und

$$f : I \to \mathbb{R}, x \mapsto f(x) \tag{5.1}$$

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

$$f'(a) := \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} \tag{5.2}$$

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

$$f' : I \to \mathbb{R}, x \mapsto f'(x) \tag{5.3}$$

heißt *Ableitung von f*.

•

Für $h > 0$ heißt der Ausdruck

$$\frac{f(a+h) - f(a)}{h} \tag{5.4}$$

Newtonscher Differenzquotient. Der Newtonsche Differenzquotient misst die Änderung $f(a+h) - f(a)$ von f 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 $h \to 0$ misst der Newtonsche Differenzquotient die instantane Änderungsrate von f in a, also im Beispiel die Geschwindigkeit des Objektes zu einem Zeitpunkt a.

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 5.2 (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 $\dot{f}(x)$, sowie
- die *Euler-Notation* Df und $Df(x)$,

respektive

•

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 adapatierte 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 dann üblicherweise mit t für "time" bezeichnet wird. $\dot{f}(t)$ bezeichnet dann die Änderungsrate von f zum Zeitpunkt t. Die Euler-Notation ist vor allem im Kontext multivariater reell- oder vektorwertiger Funktionen nützlich.

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

Definition 5.3 (Höhere Ableitungen)**.** Es sei f eine univariate reellwertige Funktion und

$$f^{(1)} := f' \tag{5.5}$$

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

$$f^{(k)} := \left(f^{(k-1)}\right)' \text{ für } k \geq 0, \tag{5.6}$$

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

$$f'' := (f')' \tag{5.7}$$

•

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 $f^{(3)}$.

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 5.1 (Rechenregeln für Ableitungen)**.** *Für $i = 1, ..., n$ seien g_i reellwertige univariate differenzierbare Funktionen. Dann gelten folgende Rechenregeln:*

(1) Summenregel

$$\text{Für } f(x) := \sum_{i=1}^{n} g_i(x) \text{ gilt } f'(x) = \sum_{i=1}^{n} g_i'(x). \tag{5.8}$$

(2) Produktregel

$$\text{Für } f(x) := g_1(x)g_2(x) \text{ gilt } f'(x) = g_1'(x)g_2(x) + g_1(x)g_2'(x). \tag{5.9}$$

(3) Quotientenregel

$$\text{Für } f(x) := \frac{g_1(x)}{g_2(x)} \text{ gilt } f'(x) = \frac{g_1'(x)g_2(x) - g_1(x)g_2'(x)}{g_2^2(x)}. \tag{5.10}$$

(4) Kettenregel

$$\text{Für } f(x) := g_1(g_2(x)) \text{ gilt } f'(x) = g_1'(g_2(x))g_2'(x). \tag{5.11}$$

∘

Erste Beispiele für die Anwendung obiger Rechenregeln lernen wir im Abschnitt Kapitel 5.2 kennen. Wir setzen eine Reihe von Ableitungen elementarer Funktionen als bekannt voraus, diese sind in Tabelle 5.1 zusammengestellt. Für Beweise verweisen wir wiederum auf die weiterführende Literatur.

Tabelle 5.1. Ableitungen elementarer Funktionen

Name	Definition	Ableitung
Polynomfunktion	$f(x) := \sum_{i=0}^{n} a_i x^i$	$f'(x) = \sum_{i=1}^{n} i a_i x^{i-1}$
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 5.1 visualisieren wir die Identitätsfunktion, eine linearen Funktion und die Quadratfunktion zusammen mit ihrer jeweiligen Ableitung. In Abbildung Abbildung 5.2 visualisieren wir die Exponential- und Logarithmusfunktionen zusammen mit ihrer jeweiligen Ableitung.

5.2. Analytische Optimierung

Eine wichtige Anwendung der Differentialrechnung 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

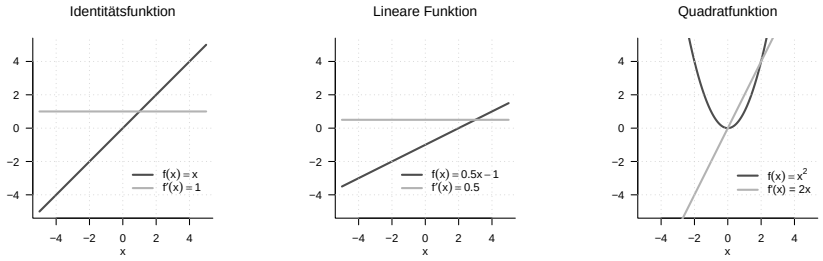


Abbildung 5.1. Ableitungen dreier elementarer Funktionen

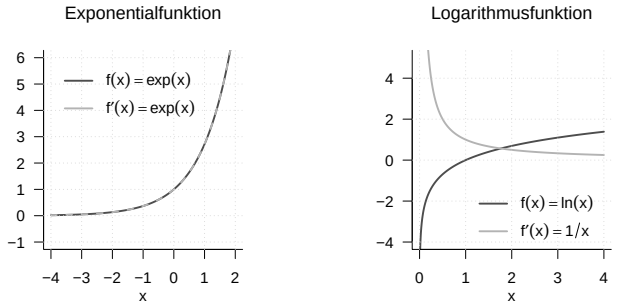


Abbildung 5.2. Ableitungen von Exponentialfunktion und Logarithmusfunktion

analytisches Vorgehen zur Optimierung von Funktionen meist nicht möglich und es werden Computeralgorithmen 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 gehen dabei eher informell vor. Einen formaleren Zugang geben wir an späterer Stelle im Kontext der nichtlinearen Optimierung. Wir beginnen damit, die Begriffe der erwähnten Maxima und Minima von univariaten reellwertigen Funktionen zu präzisieren.

Definition 5.4 (Extremstellen und Extremwerte)**.** Es seien $U \subseteq \mathbb{R}$ und $f : U \to \mathbb{R}$ 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

$$f(x_0) \leq f(x) \text{ für alle } x \in I \cap U, \tag{5.12}$$

- ein *globales Minimum*, wenn gilt, dass

$$f(x_0) \leq f(x) \text{ für alle } x \in U, \tag{5.13}$$

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

$$f(x_0) \geq f(x) \text{ für alle } x \in I \cap U, \tag{5.14}$$

- ein *lokales Maximum*, wenn gilt, dass

$$f(x_0) \geq f(x) \text{ für alle } x \in U. \tag{5.15}$$

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

$$\underset{x \in I \cap U}{\operatorname{argmin}} f(x) \text{ oder } \underset{x \in I \cap U}{\operatorname{argmax}} f(x) \tag{5.16}$$

bezeichnet und Extremwerte von Funktionen werden häufig mit

$$\min_{x \in I \cap U} f(x) \text{ oder } \max_{x \in I \cap U} f(x) \tag{5.17}$$

bezeichnet.

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

$$x_0 \textit{ ist Extremstelle von } f \Rightarrow f'(x_0) = 0. \tag{5.18}$$

◦

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 eine lokaler 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 5.3 (Hinreichende Bedingungen für lokale Extrema)**.** *f sei eine zweimal differenzierbare univariate reellwertige Funktion.*

- *Wenn für $x_0 \in U \subseteq \mathbb{R}$*
$$f'(x_0) = 0 \textit{ und } f''(x_0) > 0 \tag{5.19}$$
gilt, dann hat f an der Stelle x_0 ein Minimum.
- *Wenn für $x_0 \in U \subseteq \mathbb{R}$*
$$f'(x_0) = 0 \textit{ und } f''(x_0) < 0 \tag{5.20}$$
gilt, dann hat f an der Stelle x_0 ein Maximum.

◦

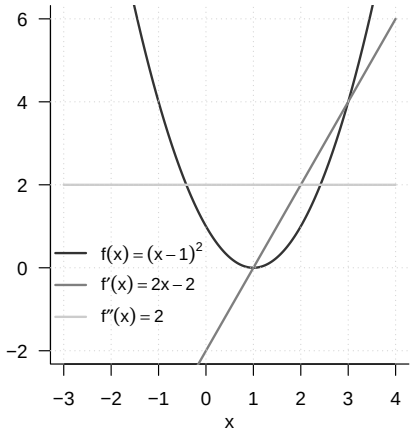


Abbildung 5.3. Analytische Optimierung von $f(x) := (x-1)^2$

Wir verzichten wiederum auf einen Beweis und verdeutlichen uns die Bedingung an dem in Abbildung 5.3 gezeigtem 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 hinreichende Bedingung für das Vorliegen von Extremstellen legt folgendes *Standardverfahren* zur Bestimmung von lokalen Extremstellen nahe.

Theorem 5.4 (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.*
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.*

∘

Anstelle eines Beweises betrachten wir beispielhaft die Funktion

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := (x-1)^2. \tag{5.21}$$

aus Abbildung 5.3. Die erste Ableitung von f ergibt sich mit der Kettenregel zu

$$f'(x) = \frac{d}{dx}\left((x-1)^2\right) = 2(x-1) \cdot \frac{d}{dx}(x-1) = 2x - 2. \tag{5.22}$$

Die zweite Ableitung von f ergibt sich zu

$$f''(x) = \frac{d}{dx}f'(x) = \frac{d}{dx}(2x-2) = 2 > 0 \text{ für alle } x \in \mathbb{R}. \tag{5.23}$$

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

$$f'(x^*) = 0 \Leftrightarrow 2x^* - 2 = 0 \Leftrightarrow 2x^* = 2 \Leftrightarrow x^* = 1. \tag{5.24}$$

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

5.3. Differentialrechnung multivariater reellwertiger Funktionen

Wir erinnern zunächst an den Begriff der multivariaten reellwertigen Funktion.

Definition 5.5 (Multivariate reellwertige Funktion)**.** Eine Funktion der Form

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) = f(x_1, ..., x_n) \tag{5.25}$$

heißt *multivariate reellwertiger* Funktion.

•

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

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2 \tag{5.26}$$

Wir visualisieren diese Funktion in Abbildung 5.4. 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.

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

Definition 5.6 (Partielle Differenzierbarkeit und partielle Ableitung)**.** Es sei $D \subseteq \mathbb{R}^n$ eine Menge und

$$f : D \to \mathbb{R}, x \mapsto f(x) \tag{5.27}$$

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

$$\frac{\partial}{\partial x_i}f(x) := \lim_{h \to 0} \frac{f(a + he_i) - f(a)}{h} \tag{5.28}$$

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

$$\frac{\partial}{\partial x_i}f : D \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_i}f(x) \tag{5.29}$$

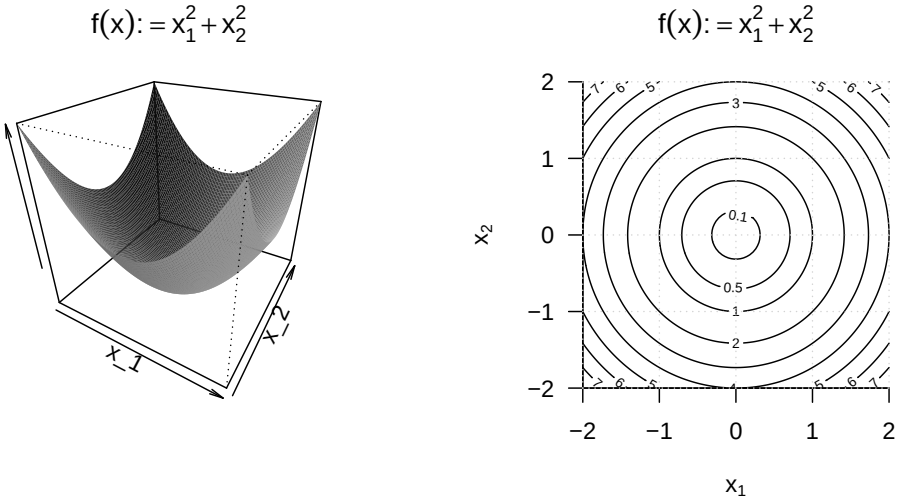


Abbildung 5.4. Visualisierungen einer bivariaten Funktion.

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

•

In Definition 5.6 bezeichnet $e_i \in \mathbb{R}^n$ bezeichnet den $i$ten kanonischen Einheitsvektor, für den gilt, dass $e_{i_j} = 1$ für $i = j$ und $e_{i_j} = 0$ für $i \neq j$ mit $j = 1, ..., n$ (vgl. Definition 8.14). In Analogie und Verallgemeinerung zum Newtonschen Differenzquotienten misst der hier auftretende Differenzquotient

$$\frac{f(x + he_i) - f(x)}{h} \tag{5.30}$$

die Änderung $f(x + he_i) - f(x)$ von f pro Strecke h in Richtung e_i. Für $h \to 0$ 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

$$\frac{d}{dx_i} \tilde{f}_{x_1,...x_{i-1},x_{i+1},...,x_n}(x_i) \tag{5.31}$$

der univariaten reellwertigen Funktion

$$\tilde{f} : \mathbb{R} \to \mathbb{R}, x_i \mapsto \tilde{f}_{x_1,...x_{i-1},x_{i+1},...,x_n}(x_i) := f(x_1, ..., x_i, ..., x_n). \tag{5.32}$$

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

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \tag{5.33}$$

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

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) \text{ und } \frac{\partial}{\partial x_2} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x). \tag{5.34}$$

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

$$f_{x_2} : \mathbb{R} \to \mathbb{R}, x_1 \mapsto f_{x_2}(x_1) := x_1^2 + x_2^2, \tag{5.35}$$

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},

$$f'_{x_2}(x) = 2x_1. \tag{5.36}$$

Es ergibt sich also

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) = \frac{\partial}{\partial x_1}(x_1^2 + x_2^2) = f'_{x_2}(x) = 2x_1. \tag{5.37}$$

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

$$\frac{\partial}{\partial x_2} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x) = \frac{\partial}{\partial x_2}(x_1^2 + x_2^2) = f'_{x_1}(x) = 2x_2. \tag{5.38}$$

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

Definition 5.7 (Zweite partielle Ableitungen)**.** $f : \mathbb{R}^n \to \mathbb{R}$ 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

$$\frac{\partial^2}{\partial x_j x_i} f(x) := \frac{\partial}{\partial x_j} \left(\frac{\partial}{\partial x_i} f \right). \tag{5.39}$$

•

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

Theorem 5.5 (Satz von Schwarz)**.** $f : \mathbb{R}^n \to \mathbb{R}$ *sei eine partiell differenzierbare multivariate reellwertige Funktion. Dann gilt*

$$\frac{\partial^2}{\partial x_j \partial x_i} f(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x) \textit{ für alle } 1 \leq i, j \leq n. \tag{5.40}$$

◦

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

Beispiel (1)

Wir wollen die partiellen Ableitungen zweiter Ordnung der Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \tag{5.41}$$

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

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 x_1} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_1}(2x_1) = 2 \\
\frac{\partial^2}{\partial x_1 x_2} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_1}(2x_2) = 0 \\
\frac{\partial^2}{\partial x_2 x_1} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_2}(2x_1) = 0 \\
\frac{\partial^2}{\partial x_2 x_2} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_2}(2x_2) = 2
\end{aligned} \tag{5.42}$$

Offenbar gilt

$$\frac{\partial^2}{\partial x_1 x_2} f(x) = \frac{\partial^2}{\partial x_2 x_1} f(x). \tag{5.43}$$

Beispiel (2)

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

$$f : \mathbb{R}^3 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_1 x_2 + x_2\sqrt{x_3}. \tag{5.44}$$

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

$$\begin{aligned}
\frac{\partial}{\partial x_1} f(x) &= \frac{\partial}{\partial x_1}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = 2x_1 + x_2, \\
\frac{\partial}{\partial x_2} f(x) &= \frac{\partial}{\partial x_2}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = x_1 + \sqrt{x_3}, \\
\frac{\partial}{\partial x_3} f(x) &= \frac{\partial}{\partial x_3}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = \frac{x_2}{2\sqrt{x_3}}.
\end{aligned} \tag{5.45}$$

Für die zweiten partiellen Ableitungen hinsichtlich x_1 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_1} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_1}(2x_1 + x_2) = 2, \\
\frac{\partial^2}{\partial x_2 \partial x_1} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_2}(2x_1 + x_2) = 1, \\
\frac{\partial^2}{\partial x_3 \partial x_1} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_3}(2x_1 + x_2) = 0.
\end{aligned} \tag{5.46}$$

Für die zweiten partiellen Ableitungen hinsichtlich x_2 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_1}\left(x_1 + \sqrt{x_3}\right) = 1, \\
\frac{\partial^2}{\partial x_2 \partial x_2} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_2}\left(x_1 + \sqrt{x_3}\right) = 0, \\
\frac{\partial^2}{\partial x_3 \partial x_2} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_3}\left(x_1 + \sqrt{x_3}\right) = \frac{1}{2\sqrt{x_3}}.
\end{aligned} \tag{5.47}$$

Beispiel (2) Für die zweiten partiellen Ableitungen hinsichtlich x_3 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_1}\left(\frac{x_2}{2}\sqrt{x_3}\right) = 0, \\
\frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_2}\left(\frac{x_2}{2\sqrt{x_3}}\right) = \frac{1}{2\sqrt{x_3}}, \\
\frac{\partial^2}{\partial x_3 \partial x_3} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_3}\left(x_2 \frac{1}{2} x_3^{-\frac{1}{2}}\right) = -\frac{1}{4} x_2 x_3^{-\frac{3}{2}}.
\end{aligned} \tag{5.48}$$

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

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial^2}{\partial x_2 \partial x_1} f(x) = 1, \\
\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_1} f(x) = 0, \\
\frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_2} f(x) = \frac{1}{2\sqrt{x_3}}.
\end{aligned} \tag{5.49}$$

Wie oben gesehen gibt es für eine multivariate reellwertige Funktion $f : \mathbb{R}^n \to \mathbb{R}$ 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 5.8 (Gradient)**.** $f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann ist der *Gradient* $\nabla f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix} \in \mathbb{R}^n. \tag{5.50}$$

•

Man beachte, dass Gradienten multivariate vektorwertige Funktionen der

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto \nabla f(x) \tag{5.51}$$

sind. Für $n = 1$ gilt $\nabla f(x) = f'(x)$. Eine wichtige Eigenschaften 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 Beispiel (1) betrachtete Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} \in \mathbb{R}^2. \tag{5.52}$$

In Abbildung 5.5 visualisieren wir ausgewählte Werte dieses Gradienten für

$$x := \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix}, \quad , x := \begin{pmatrix} -0.3 \\ 0.1 \end{pmatrix}, \quad x := \begin{pmatrix} -0.5 \\ -0.4 \end{pmatrix}, \quad x := \begin{pmatrix} 0.1 \\ -1.0 \end{pmatrix}$$

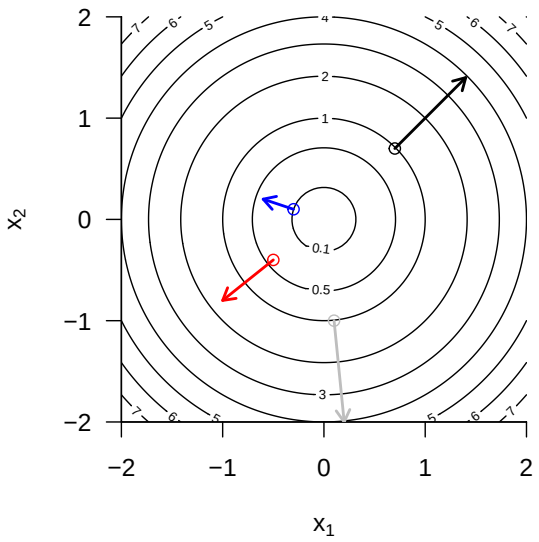


Abbildung 5.5. Exemplarische Gradientenwerte der bivariaten Funktion $f(x) = x_1^2 + x_2^2$.

Beispiel (2)

Für die in Beispiel (2) betrachtete Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \frac{\partial}{\partial x_3} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 + x_2 \\ x_1 + \sqrt{x_3} \\ \frac{x_2}{2\sqrt{x_3}} \end{pmatrix} \in \mathbb{R}^3. \tag{5.53}$$

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

Definition 5.9 (Hesse-Matrix)**.** $f : \mathbb{R}^n \to \mathbb{R}$ sei ein multivariate reellwertige Funktion. Dann ist die *Hesse-Matrix* $\nabla^2 f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 x_1} f(x) & \frac{\partial^2}{\partial x_1 x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_1 x_n} f(x) \\ \frac{\partial^2}{\partial x_2 x_1} f(x) & \frac{\partial^2}{\partial x_2 x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_2 x_n} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n x_1} f(x) & \frac{\partial^2}{\partial x_n x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_n x_n} f(x) \end{pmatrix} \in \mathbb{R}^{n \times n}. \tag{5.54}$$

•

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

$$\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n \times n}, x \mapsto \nabla^2 f(x) \tag{5.55}$$

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

$$\frac{\partial^2}{\partial x_i \partial x_j} f(x) = \frac{\partial^2}{\partial x_j \partial x_i} f(x) \text{ für } 1 \leq i, j \leq n \tag{5.56}$$

dass die Hesse-Matrix symmetrisch ist, dass also

$$\left(\nabla^2 f(x)\right)^T = \nabla^2 f(x) \tag{5.57}$$

gilt.

Beispiel (1)

Für die in Beispiel (1) betrachtete Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 x_1} f(x) & \frac{\partial^2}{\partial x_1 x_2} f(x) \\ \frac{\partial^2}{\partial x_2 x_1} f(x) & \frac{\partial^2}{\partial x_2 x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \in \mathbb{R}^{2 \times 2} \tag{5.58}$$

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

Beispiel (2)

Für die in Beispiel (2) betrachtete Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 x_1} f(x) & \frac{\partial^2}{\partial x_1 x_2} f(x) & \frac{\partial^2}{\partial x_1 x_3} f(x) \\ \frac{\partial^2}{\partial x_2 x_1} f(x) & \frac{\partial^2}{\partial x_2 x_2} f(x) & \frac{\partial^2}{\partial x_2 x_3} f(x) \\ \frac{\partial^2}{\partial x_3 x_1} f(x) & \frac{\partial^2}{\partial x_3 x_2} f(x) & \frac{\partial^2}{\partial x_3 x_3} f(x) \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & \frac{1}{2\sqrt{3}} \\ 0 & \frac{1}{2\sqrt{3}} & -\frac{1}{4} x_2 x_3^{-3/2} \end{pmatrix}. \tag{5.59}$$

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.

5.4. Selbstkontrollfragen

1. Geben Sie die Definition des Begriffs der Ableitung $f'(a)$ einer Funktion f an einer Stelle a wieder.
2. Geben Sie die Definition des Begriffs der Ableitung f' einer Funktion f.
3. Erläutern Sie die Symbole $f'(x), \dot{f}(x), \frac{df(x)}{dx}$, und $\frac{d}{dx}f(x)$.
4. Geben Sie die Definition des Begriffs der zweiten Ableitung f'' einer Funktion f wieder.
5. Geben Sie die Summenregel für Ableitungen wieder.
6. Geben Sie die Produktregel für Ableitungen wieder.
7. Geben Sie die Quotientenregel für Ableitungen wieder.
8. Geben Sie die Kettenregel für Ableitungen wieder.
9. Bestimmen Sie die erste Ableitung der Funktion $f(x) := 3x^2 + \exp(-x^2) - x\ln(x)$.
10. Bestimmen Sie die erste Ableitung der Funktion $f(x) := \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$ für $\mu \in \mathbb{R}$.
11. Geben Sie die Definition der Begriffe des globalen und lokalen Maximums/Minimums einer univariaten reellwertigen Funktion wieder.
12. Geben Sie die notwendige Bedingung für ein Extremum einer Funktion wieder.
13. Geben Sie die hinreichende Bedingung für ein lokales Extremum einer Funktion wieder.
14. Geben Sie das Standardverfahren der analytischen Optimierung wieder.
15. Bestimmen Sie einen Extremwert von $f(x) := \exp\left(-\frac{1}{2}(x-\mu)^2\right)$ für $\mu \in \mathbb{R}$.
16. Berechnen Sie die partiellen Ableitungen der Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := \exp\left(-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right). \tag{5.60}$$

17. Berechnen Sie die zweiten partiellen Ableitungen obiger Funktion f.
18. Geben Sie den Satz von Schwarz wieder.
19. Geben Sie die Definition des Gradienten einer multivariaten reellwertigen Funktion wieder.
20. Geben Sie den Gradienten der Funktion in Gleichung 5.60 an und werten Sie ihn in $x = (1, 2)^T$ aus.
21. Geben Sie die Definition der Hesse-Matrix einer multivariaten reellwertigen Funktion wieder.
22. Geben Sie die Hesse-Matrix der Funktion in Gleichung 5.60 an und werten Sie sie in $x = (1, 2)^T$ aus.

6. Folgen, Grenzwerte, Stetigkeit

Die in diesem Kapitel behandelten Themen sind in der probabilistischen Datenanayse nicht zentral, sondern bilden Grundpfeiler der reellen Analysis. Durch die enge Verschränkung der modernen Wahrscheinlichkeitstheorie mit analytischen Ansätzen dienen sie jedoch dem Verständnis von zum Beispiel dem Zentralen Grenzwertsatz, der eine Hauptgrundlage für die weit verbreitete Normalverteilungsannahme in der probabilistischen Datenanalyse darstellt. In aller Kürze ist der Zentrale Grenzwertsatz eine Aussage über die Grenzfunktion einer Funktionenfolge, nämlich einer Folge von Zufallsvariablen. Das Wissen um das Wesen von Folgen, Funktionenfolgen und ihren Grenzwerten erlaubt also ein tieferes Verständnis wichtiger Grundannahmen der probabilistischen Datenanalyse. Weiterhin ermöglichen die in diesem Kapitel behandelten Themen zumindest einen ersten Einstieg in das Verständnis der Stetigkeit und Glattheit von Funktionen, die insbesondere in der nichtlinearen Optimierung zu Bestimmung von Parameterschätzern in probabilistischen Modellen wichtige Grundkonzepte bilden.

6.1. Folgen

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

Definition 6.1 (Reelle Folge)**.** Eine *reelle Folge* ist eine Funktion der Form

$$f : \mathbb{N} \to \mathbb{R}, n \mapsto f(n) \tag{6.1}$$

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

$$(x_1, x_2, ...) \text{ oder } (x_n)_{n=1}^{\infty} \text{ oder } (x_n)_{n\in\mathbb{N}} \text{ oder } (x_n). \tag{6.2}$$

•

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 $(x_1, x_2, ...)$ bewusst machen. Wir wollen zwei Standardbeispiele für reelle Folgen betrachten.

Beispiele für reelle Folgen

(1) Reelle Folgen der Form

$$f : \mathbb{N} \to \mathbb{R}, n \mapsto f(n) := \left(\frac{1}{n}\right)^{\frac{p}{q}} \text{ mit } p, q \in \mathbb{N} \tag{6.3}$$

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

$$\left(\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, ...\right). \tag{6.4}$$

(2) Reelle Folgen der Form

$$f : \mathbb{N} \to \mathbb{R}, n \mapsto f(n) := q^n \text{ mit } q \in]-1, 1[\tag{6.5}$$

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

$$\begin{aligned}\left(\left(\frac{1}{2}\right)^1, \left(\frac{1}{2}\right)^2, \left(\frac{1}{2}\right)^3, ...\right) &= \left(\left(\frac{1}{2}\right)^1, \left(\frac{1}{2}\right)^2, \left(\frac{1}{2}\right)^3, ...\right) \\ &= \left(\frac{1^1}{2^1}, \frac{1^2}{2^2}, \frac{1^3}{2^3}...\right) \\ &= \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}...\right)\end{aligned} \tag{6.6}$$

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

Definition 6.2 (Funktionenfolge)**.** Es sei ϕ eine Menge univariater reellwertiger Funktionen mit Definitionsmenge $D \subseteq \mathbb{R}$. Dann ist eine Funktionenfolge eine Funktion der Form

$$F : \mathbb{N} \to \phi, n \mapsto F(n). \tag{6.7}$$

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

$$(f_1, f_2, ...) \text{ oder } (f_n)_{n=1}^{\infty} \text{ oder } (f_n)_{n\in\mathbb{N}} \text{ oder } (f_n). \tag{6.8}$$

•

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 Funktionenfolgen dagegen univariate reellwertige Funktionen sind. Auch hier wollen wir zwei Standardbeispiel diskutieren.

Beispiele für Funktionenfolgen

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

$$\phi := \{f_n | f_n : [0, 1] \to \mathbb{R}, x \mapsto f_n(x) := x^n \text{ für } n \in \mathbb{N}\} \tag{6.9}$$

Dann definiert

$$F : \mathbb{N} \to \phi, n \mapsto F(n) \tag{6.10}$$

eine Funktionenfolge. Für die Funktionswerte der Folgenglieder von F gilt

$$f_1(x) := x^1, f_2(x) := x^2, f_3(x) := x^3, ... \tag{6.11}$$

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

$$\phi := \{f_n | f_n : [-a, a] \to \mathbb{R}, x \mapsto f_n(x) := \sum_{k=0}^{n} \frac{x^k}{k!} \text{ für } n \in \mathbb{N}\} \tag{6.12}$$

Dann definiert

$$F : \mathbb{N} \to \phi, n \mapsto F(n) \tag{6.13}$$

eine Funktionenfolge. Für die Funktionswerte der Folgenglieder von F gilt

$$f_1(x) := \sum_{k=0}^{1} \frac{x^k}{k!}, f_2(x) := \sum_{k=0}^{2} \frac{x^k}{k!}, f_3(x) := \sum_{k=0}^{3} \frac{x^k}{k!}, ... \tag{6.14}$$

6.2. 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 6.3 (Grenzwert einer 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

$$|x_n - x| < \epsilon \text{ für alle } n \geq m. \tag{6.15}$$

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

$$\lim_{n\to\infty} x_n = x \text{ oder } x_n \to x \text{ für } n \to \infty \text{ oder } x_n \xrightarrow{n\to\infty} x. \tag{6.16}$$

•

Der Grenzwert einer Folge kann also, aber muss nicht existieren. So hat zum Beispiel die Folge

$$f : \mathbb{N} \to \mathbb{R}, n \mapsto f(n) := n \tag{6.17}$$

keinen Grenzwert, da hier sowohl n als auch $f(n)$ unendlich groß werden. Die oben betrachteten Beispiele für reelle Folgen dagegen haben Grenzwert. Dies ist Inhalt folgender Beispiele

Beispiele

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

$$\lim_{n\to\infty} \left(\frac{1}{n}\right)^{\frac{p}{q}} = 0. \tag{6.18}$$

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

$$\lim_{n\to\infty} q^n = 0. \tag{6.19}$$

Man nennt die harmonischen und geometrischen Folgen entsprechend auch *Nullfolgen*. Für Beweise von Gleichung 6.18 und Gleichung 6.19 verweisen wir auf die weiterführende Literatur. Tatsächlich sind diese Beweise nicht trivial und rühren an die Grundannahmen über das Wesen der reellen Zahlen. Wir visualisieren die ersten zehn Folgenglieder sowie die Grenzwerte der harmonischen Folge fr $p := q := 1$ und der geometrischen Folge für $q := 1/2$ in Abbildung 6.1.

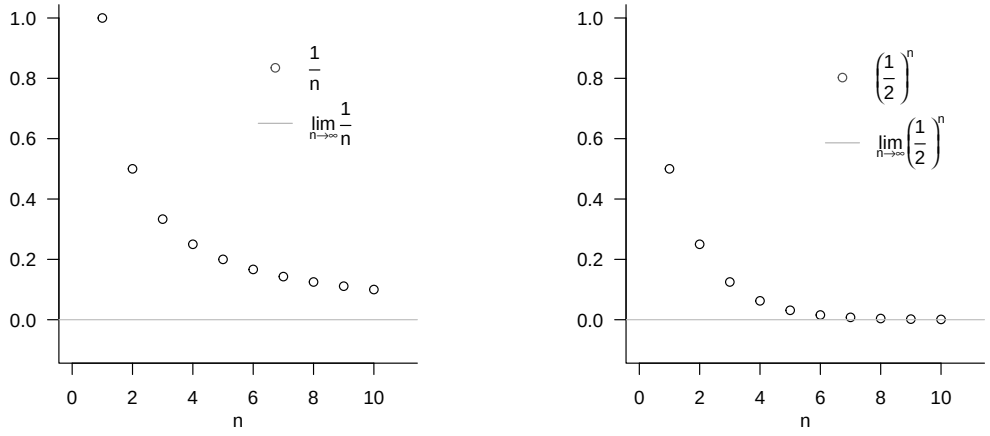


Abbildung 6.1. Beispiele für Grenzwerte reeller Folgen.

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

Definition 6.4 (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

$$f : D \to \mathbb{R}, x \mapsto f(x) := \lim_{n\to\infty} f_n(x). \tag{6.20}$$

•

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 ebenfalls auf die weiterführende Literatur verweisen.

Beispiele

(1) Wir betrachten die Funktionenfolge

$$F : \mathbb{N} \to \phi, n \mapsto F(n) \tag{6.21}$$

mit

$$\phi := \{f_n | f_n : [0,1] \to \mathbb{R}, x \mapsto f_n(x) := x^n \text{ für } n \in \mathbb{N}\} \tag{6.22}$$

Dann ist F punktweise konvergent mit Grenzfunktion

$$f : [0,1] \to \mathbb{R}, x \mapsto f(x) := \begin{cases} 0, & \text{für } x \in [0,1[\\ 1, & \text{für } x = 1 \end{cases} \tag{6.23}$$

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

$$F : \mathbb{N} \to \phi, n \mapsto F(n) \tag{6.24}$$

mit

$$\phi := \{f_n | f_n : [-a,a] \to \mathbb{R}, x \mapsto f_n(x) := \sum_{k=0}^{n} \frac{x^k}{k!} \text{ für } n \in \mathbb{N}\} \tag{6.25}$$

Dann ist F punktweise konvergent mit Grenzfunktion

$$f : [-a,a] \to \mathbb{R}, x \mapsto f(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!} =: \exp(x) \tag{6.26}$$

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

$$\exp(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{6.27}$$

definiert.

6.3. 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 nur zu 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 6.5 (Grenzwert einer Funktion)**.**

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

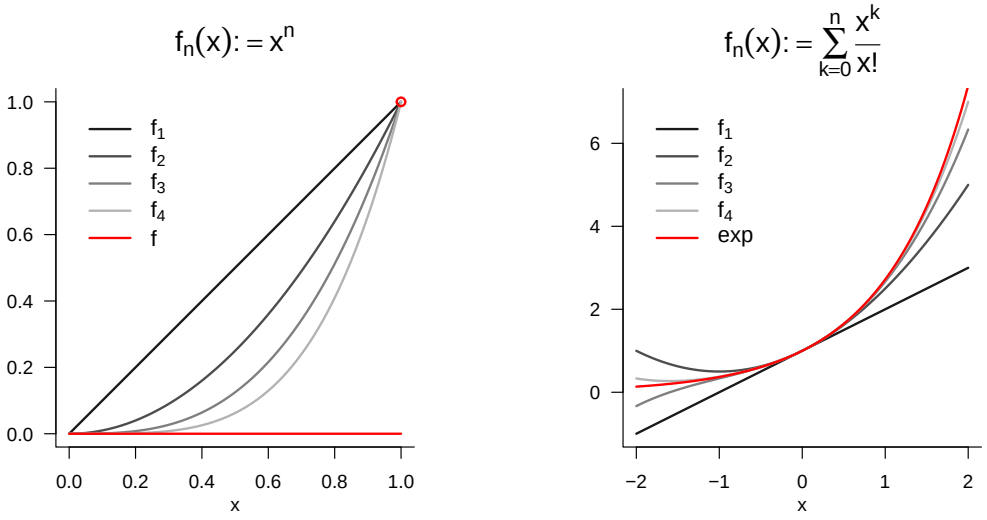


Abbildung 6.2. Beispiele für Grenzwerte von Funktionenfolgen

(1) es eine reelle Folge $(x_n)_{n=1}^{\infty}$ mit Folgengliedern in D mit Grenzwert a gibt, also $\lim_{n\to\infty} x_n = a$ 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 $\lim_{n\to\infty} f(x_n) = b$ gilt.

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

•

In Abbildung 6.3 visualisieren wir den Grenzwert der Exponential funktion in $a = 1$ durch Darstellung von Folgenglieder $x_n \to 1$ und den entsprechenden Folgengliedern $f(x_n)$. Offenbar gilt $\lim_{x\to 1} \exp(x) = e$.

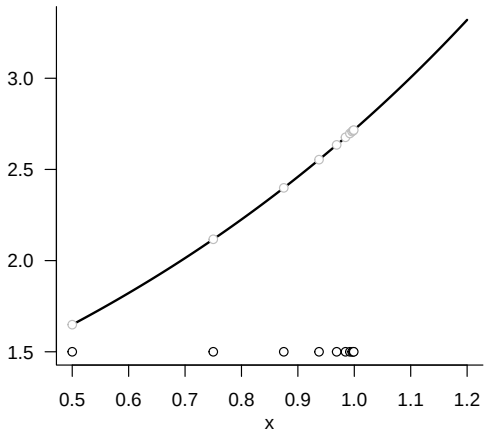


Abbildung 6.3. Beispiele für einen Grenzwert einer Funktion

Wir können nun den Begriff der Stetigkeit einer Funktion definieren.

Definition 6.6 (Stetigkeit einer Funktion)**.** Eine Funktion $f : D \to Z$ mit $D \subseteq \mathbb{R}, Z \subseteq \mathbb{R}$ heißt *stetig in* $a \in D$, wenn

$$\lim_{x \to a} f(x) = f(a). \tag{6.28}$$

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

$$\lim_{x \to a} f(x) = f\left(\lim_{x \to a} x\right) \tag{6.29}$$

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

6.4. Selbstkontrollfragen

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 unbestimmen Integrals und dem Begriff der Stammfunktion.

Definition 7.1 (Unbestimmtes Integral und Stammfunktion)**.** Für ein Intervall $I \subseteq \mathbb{R}$ sei $f : I \to \mathbb{R}$ eine univariate reellwertige Funktion. Dann heißt eine differenzierbare Funktion $F : I \to \mathbb{R}$ 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 reine 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 Differentialrechnung. 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	$f(x) := \sum_{i=0}^{n} a_i x^i$	$F(x) = \sum_{i=0}^{n} \frac{a_i}{i+1} x^{i+1}$
Konstante Funktion	$f(x) := a$	$F(x) = ax$

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

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

(2) Partielle Integration

$$\int f'(x)g(x)\,dx = f(x)g(x) - \int f(x)g'(x)\,dx \tag{7.4}$$

(3) Substitionsregel

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

∘

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

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x). \tag{7.6}$$

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

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

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

$$(F(g(x)))' = F'(g(x))g'(x) = f(g(x))g'(x). \tag{7.8}$$

Integration beider Seiten der Gleichung

$$(F(g(x)))' = f(g(x))g'(x) \tag{7.9}$$

ergibt dann

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

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.

□

Unbestimmte Integrale nehmen in der Lösung von Differentialgleichungen 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 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.

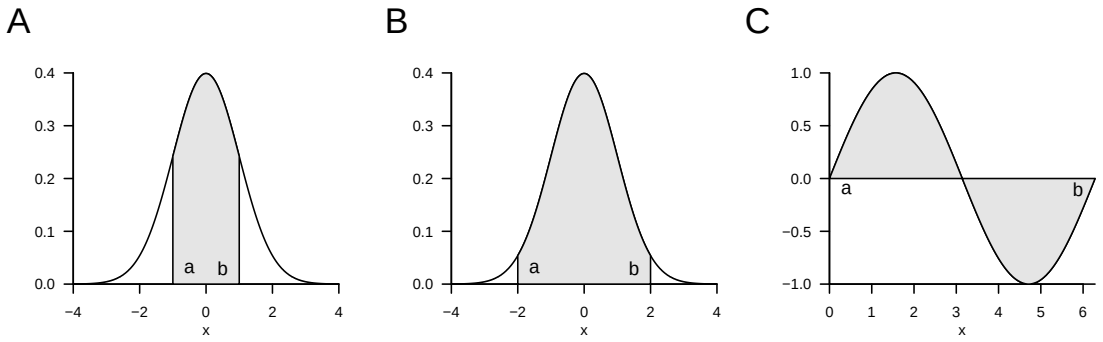


Abbildung 7.1. Beispiele bestimmter Integrale

Um den Begriff des *bestimmten Integrals* im Sinne des *Riemannschen 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 $x_0, x_1, x_2, ..., x_n \in [a, b]$ eine Menge von Punkten mit

$$a =: x_0 < x_1 < x_2 \cdots < x_n := b \tag{7.11}$$

und

$$\Delta x_i := x_i - x_{i-1} \text{ für } i = 1, ..., n. \tag{7.12}$$

Dann heißt die Menge

$$Z := \{[x_0, x_1], [x_1, x_2], ..., [x_{n-1}, x_n]\} \tag{7.13}$$

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

$$Z_{\max} := \max_{i \in n} \Delta x_i, \tag{7.14}$$

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)**.** $f : [a, b] \to \mathbb{R}$ 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 $i = 1, ..., n$. Weiterhin sei ξ_i für $i = 1, ..., n$ ein beliebiger Punkt im Teilintervall $[x_{i-1}, x_i]$ der Zerlegung Z. Dann heißt

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

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*,

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

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

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

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

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

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

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

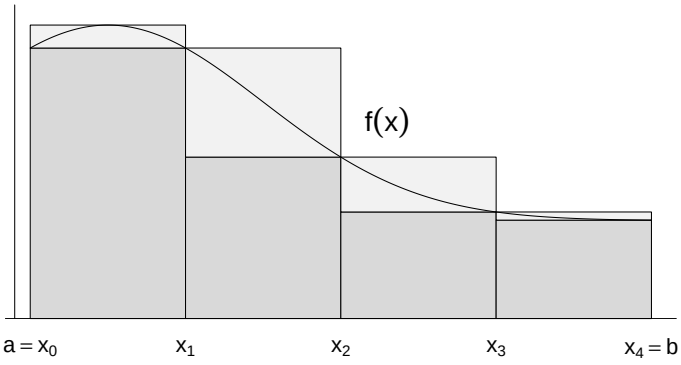


Abbildung 7.2. Riemannsche Summen

Stellt man sich nun vor, dass man Δx_i für alle $i = 1, ..., n$ 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.

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

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

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 allerding um die Hauptsätze der Differential- und Integralrechnung erweitern, so dass 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 Bezeichungen *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

$$\int_a^b (c_1 f(x) + c_2 g(x))\, dx = c_1 \int_a^b f(x)\, dx + c_2 \int_a^b f(x)\, dx. \tag{7.21}$$

(2) Additivität. Für $a < c < b$ gilt

$$\int_a^b f(x)\, dx = \int_a^c f(x)\, dx + \int_c^b f(x)\, dx. \tag{7.22}$$

(3) Vorzeichenwechsel bei Umkehrung der Integralgrenzen

$$\int_a^b f(x)\, dx = -\int_b^a f(x)\, dx. \tag{7.23}$$

(4) Unabhängigkeit von der Wahl der Integrationsvariable

$$\int_a^b f(x)\, dx = \int_a^b f(y)\, dy. \tag{7.24}$$

(5) Unabhängigkeit des Integrals von Art des Intervalls. Es gilt

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

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

∘

Eine graphische Darstellung der Rechenregel der Additivität findet sich in 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 $\int_a^b f(x)\, dx$.

Die in der Nachfolge vermerkten Hauptsätze der Differential- und Integralrechnung schließlich, ermöglichen es, bestimmte Integrale einer Funktion f direkt mithilfe der Stammfunktion F von f zu berechnen.

Theorem 7.3 (Erster Hauptsatz der Differential- und Integralrechnung)**.** *Ist $f : I \to \mathbb{R}$ eine auf dem Intervall $I \subset \mathbb{R}$ stetige Funktion, dann ist die Funktion*

$$F : I \to \mathbb{R}, x \mapsto F(x) := \int_a^x f(t)\, dt \text{ mit } x, a \in I \tag{7.26}$$

eine Stammfunktion von f.

∘

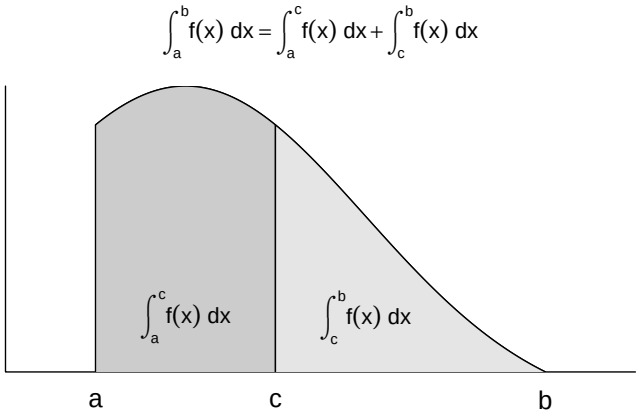


Abbildung 7.3. Additivität bestimmter Integrale

Beweis. Wir betrachten den Differenzquotienten

$$\frac{1}{h}(F(x+h) - F(x)) \tag{7.27}$$

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

$$\frac{1}{h}(F(x+h) - F(x)) = \frac{1}{h}\left(\int_a^{x+h} f(t)\,dt - \int_a^x f(t)\,dt\right) = \frac{1}{h}\int_x^{x+h} f(t)\,dt \tag{7.28}$$

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

$$\frac{1}{h}(F(x+h) - F(x)) = f(\xi) \tag{7.29}$$

Grenzwertbildung ergibt dann

$$\lim_{h\to 0}\frac{1}{h}(F(x+h) - F(x)) = \lim_{h\to 0} f(\xi) \text{ für } \xi \in]x, x+h[\Leftrightarrow F'(x) = f(x). \tag{7.30}$$

□

Für den Beweis des Ersten Hauptsatzes der Differential- und Integralrechnung benötigen wir offenbar den Mittelwertsatz der Integralrechnung, welchen wir hier ohne Beweis wiedergeben und in Abbildung 7.4 veranschaulichen.

Theorem 7.4 (Mittelwertsatz der Integralrechnung)**.** *Für eine stetige Funktion* $f : [a, b] \to \mathbb{R}$ *existiert ein* $\xi \in]a, b[$ *mit*

$$\int_a^b f(x)\,dx = f(\xi)(b-a) \tag{7.31}$$

∘

Der Mittelwertsatz der Integralrechnung garantiert die Existenz eines $\xi \in [a, b]$, so dass das bestimmte Integral $\int_a^b f(x)\,dx$ gleich dem Produkt aus der "Rechteckhöhe" $f(\xi)$ und

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 nachvollziebaren 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.

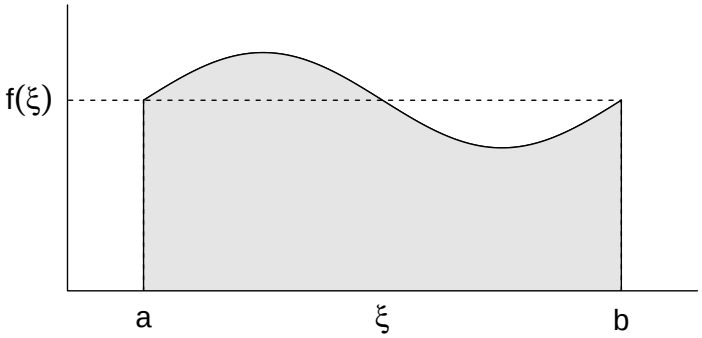


Abbildung 7.4. Zum Mittelwertsatz der Integralrechnung

Der Zweite Hauptsatz der Differential- und Integralrechnung schließlich besagt, wie man mithilfe der Stammfunktion ein bestimmtes Integral berechnet.

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

$$\int_a^b f(x)\,dx = F(b) - F(a) =: F(x)|_a^b \tag{7.32}$$

∘

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

$$F(b) - F(a) = \int_\alpha^b f(t)\,dt - \int_\alpha^a f(t)\,dt = \int_a^b f(x)\,dx \tag{7.33}$$

□

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

Beispiel (1)

Wir betrachten die Identitätsfunktion

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := x \tag{7.34}$$

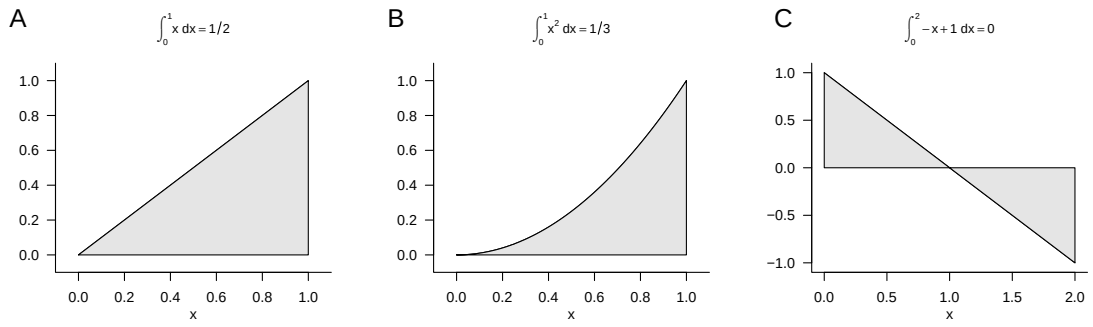


Abbildung 7.5. Beispiele zum Zweiten Hauptsatz der Differential- und Integralrechnung

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

$$\int_0^1 f(x)\,dx = \int_0^1 x\,dx \tag{7.35}$$

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

$$F : \mathbb{R} \to \mathbb{R}, x \mapsto F(x) := \frac{1}{2}x^2 \tag{7.36}$$

gegeben ist, weil

$$F'(x) = \frac{d}{dx}\left(\frac{1}{2}x^2\right) = 2 \cdot \frac{1}{2}x^{2-1} = x. \tag{7.37}$$

Einsetzen in den Zweiten Hauptsatz der Differential- und Integralrechnung ergibt dann sofort

$$\int_0^1 x\,dx = \frac{1}{2}1^2 - \frac{1}{2}0^2 = \frac{1}{2}. \tag{7.38}$$

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

Beispiel (2)

Als nächstes betrachten wird die Quadratfunktion

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := x^2 \tag{7.39}$$

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

$$\int_0^1 f(x)\,dx = \int_0^1 x^2\,dx \tag{7.40}$$

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

$$F : \mathbb{R} \to \mathbb{R}, x \mapsto F(x) := \frac{1}{3}x^3 \tag{7.41}$$

gegeben ist, weil

$$F'(x) = \frac{d}{dx}\left(\frac{1}{3}x^3\right) = 3 \cdot \frac{1}{3}x^{3-1} = x^2. \tag{7.42}$$

Einsetzen in den Zweiten Hauptsatz der Differential- und Integralrechnung ergibt dann sofort

$$\int_0^1 x^2\,dx = \frac{1}{3}1^3 - \frac{1}{3}0^3 = \frac{1}{3}. \tag{7.43}$$

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 lineare Funktion

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := -x + 1 \tag{7.44}$$

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

$$\int_0^2 f(x)\,dx = \int_0^2 -x + 1\,dx \tag{7.45}$$

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

$$F : \mathbb{R} \to \mathbb{R}, x \mapsto F(x) := -\frac{1}{2}x^2 + x \tag{7.46}$$

gegeben ist, weil

$$F'(x) = \frac{d}{dx}\left(-\frac{1}{2}x^2 + x\right) = -2 \cdot \frac{1}{2}x^{2-1} + 1 \cdot x^{1-1} = -x + 1. \tag{7.47}$$

Einsetzen in den Zweiten Hauptsatz der Differential- und Integralrechnung ergibt dann sofort

$$\int_0^2 -x + 1\,dx = \left(-\frac{1}{2}2^2 + 2\right) - \left(-\frac{1}{2}0^2 + 0\right). = -2 + 2 - 0 = 0. \tag{7.48}$$

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 ist, sondern $-\infty$ oder ∞. Wir beleuchten die Natur uneigentlicher Integrale mit folgender Definition und einem Beispiel.

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

$$\int_{-\infty}^b f(x)\,dx := \lim_{a \to -\infty} \int_a^b f(x)\,dx \text{ und } \int_a^\infty f(x)\,dx := \lim_{b \to \infty} \int_a^b f(x)\,dx \tag{7.49}$$

und der Additivität von Integralen

$$\int_{-\infty}^\infty f(x)\,dx = \int_{-\infty}^b f(x)\,dx + \int_b^\infty f(x)\,dx \tag{7.50}$$

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

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x)\frac{1}{x^2} \tag{7.51}$$

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

$$\int_1^\infty \frac{1}{x^2}\, dx. \tag{7.52}$$

Nach den Festlegungen in der Definition uneigentlicher Integrale gilt

$$\int_1^\infty \frac{1}{x^2}\, dx = \lim_{b\to\infty} \int_1^b \frac{1}{x^2}\, dx. \tag{7.53}$$

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

$$\int_1^b \frac{1}{x^2}\, dx = F(b) - F(1) = -\frac{1}{b} - \left(-\frac{1}{1}\right) = -\frac{1}{b} + 1. \tag{7.54}$$

Es ergibt sich also

$$\int_1^\infty \frac{1}{x^2}\, dx = \lim_{b\to\infty} \int_1^b \frac{1}{x^2}\, dx = \lim_{b\to\infty} \left(-\frac{1}{b} + 1\right) = -\lim_{b\to\infty} \frac{1}{b} + \lim_{b\to\infty} 1 = 0 + 1 = 1. \tag{7.55}$$

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 zum Beispiel schon im zweidimensionalen arbiträr geformte zweidimensionale 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 Integrale)**.** $f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann heißen Integrale der Form

$$\int\limits_{[a_1,b_1]\times\cdots\times[a_n,b_n]} f(x)\, dx = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, ..., x_n)\, dx_1 ... dx_n \tag{7.56}$$

mehrdimensionale bestimmte Integrale auf Hyperrechtecken. Weiterhin heißen Integrale der Form

$$\int_{\mathbb{R}^n} f(x)\, dx = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_n)\, dx_1 ... dx_n \tag{7.57}$$

mehrdimensionale uneigentliche Integrale.

•

Wie schon erwähnt kann man multivariate reellwertige Funktion nicht nur auf Hyperrechtecken, sondern im Prinzip auf beliebigen Hyperflächen integrieren. Dies kann sich jedoch oft schwierig gestalten.

Als Beispiel betrachten wir das zweidimensionale bestimmte Integral der Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, (x_1, x_2) \mapsto f(x_1, x_2) := x_1^2 + 4x_2 \tag{7.58}$$

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

$$\int_{a_1}^{b_1} \left(\int_{a_2}^{b_2} f(x_1, x_2)\, dx_2 \right) dx_1 = \int_{a_2}^{b_2} \left(\int_{a_1}^{b_1} f(x_1, x_2)\, dx_1 \right) dx_2. \tag{7.59}$$

In diesem Sinne betrachten wir für das Beispiel

$$\int_0^1 \int_0^1 x_1^2 + 4x_2\, dx_1\, dx_2 = \int_0^1 \left(\int_0^1 x_1^2 + 4x_2\, dx_1 \right) dx_2 \tag{7.60}$$

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 $G(x_1) = \frac{1}{3}x_1^3 + 4x_2x_1$, wie man sich durch Ableiten von G überzeugt. Es ergibt sich also für das innere Integral

$$\begin{aligned} \int_0^1 x_1^2 + 4x_2\, dx_1 &= G(1) - G(0) \\ &= \frac{1}{3} \cdot 1^3 + 4x_2 \cdot 1 - \frac{1}{3} \cdot 0^3 - 4x_2 \cdot 0 \\ &= \frac{1}{3} + 4x_2. \end{aligned} \tag{7.61}$$

Betrachten des äußeren Integrals ergibt dann mit der Stammfunktion

$$H(x_2) = \frac{1}{3}x_2 + 2x_2^2 \tag{7.62}$$

von

$$h(x_2) := \frac{1}{3} + 4x_2, \tag{7.63}$$

dass

$$\begin{aligned} \int_0^1 \int_0^1 x_1^2 + 4x_2\, dx_1\, dx_2 &= \int_0^1 \frac{1}{3} + 4x_2\, dx_2 \\ &= H(1) - H(0) \\ &= \frac{1}{3} \cdot 1 + 4 \cdot 1^2 - \frac{1}{3} \cdot 0 + 4 \cdot 0^2 \\ &= \frac{13}{3}. \end{aligned} \tag{7.64}$$

7.5. Selbstkontrollfragen

1. Geben Sie die Definition des Begriffs der Stammfunktion wieder.

2. Geben Sie die Definition des Begriffs des unbestimmten Integrals wieder.
3. Erläutern Sie die intuitive Bedeutung des Begriff des Riemannschen Integrals.
4. Geben Sie den ersten Hauptsatz der Differential- und Integralrechnung wieder.
5. Geben Sie den zweiten Hauptsatz der Differential- und Integralrechnung wieder.
6. Erläutern Sie den Begriff des uneigentlichen Integrals.
7. Erläutern Sie den Begriff des mehrdimensionalen Integrals.

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, z.B. einen Selbstauskunftscore, einen Biomarker und eine 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 **??** 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

$$+ : V \times V \to V, (v_1, v_2) \mapsto +(v_1, v_2) =: v_1 + v_2, \tag{8.1}$$

genannt *Vektoraddition*, definiert. Schließlich sei eine Abbildung

$$\cdot : S \times V \to V, (s, v) \mapsto \cdot(s, v) =: sv, \tag{8.2}$$

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*

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

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

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

(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 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

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

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

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

und die *Skalarmultiplikation* durch

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

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

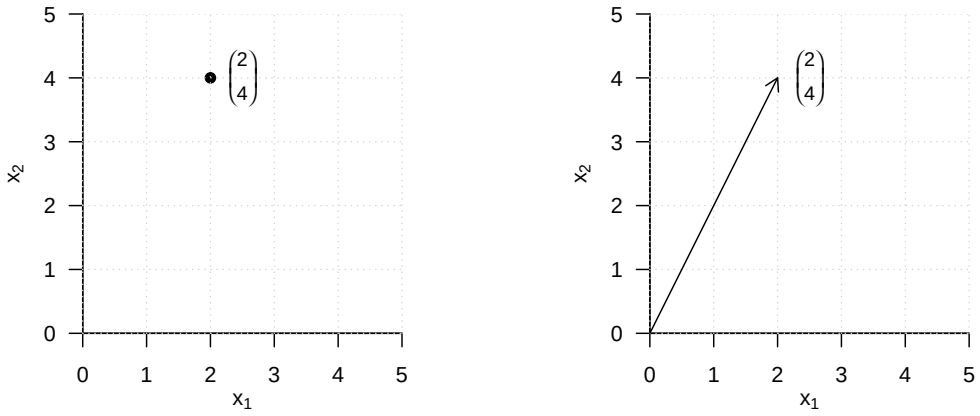


Abbildung 8.1. Visualisierung von Vektoren in $\mathbb{R}^2$

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

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}. \tag{8.6}$$

Der Summenvektor entspricht dabei der Diagonale des von den beiden Summanden aufgespannten Parallelogramms.

Abbildung 8.3 visualisiert die Vektorsubtraktion

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} -3 \\ -1 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \tag{8.7}$$

Der resultierende Vektor entspricht dabei der Diagonale des von dem ersten Vektors und dem entgegensetzten Vektor des zweiten Vektors aufgespannten Parallelogramms.

Abbildung 8.4 schließlich visualisiert die Skalarmultiplikation

$$3 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \tag{8.8}$$

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*

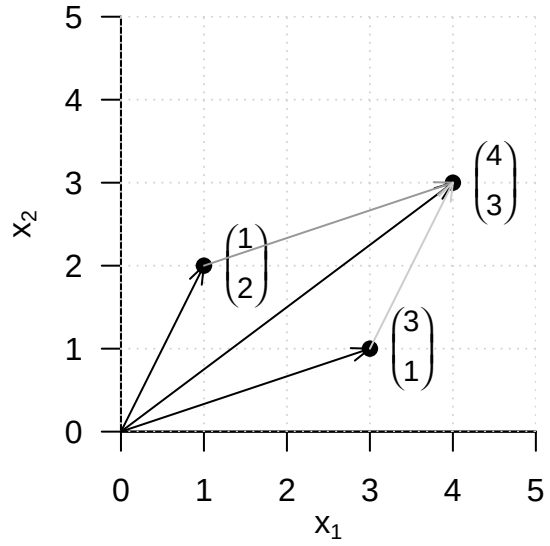


Abbildung 8.2. Vektoraddition in $\mathbb{R}^2$

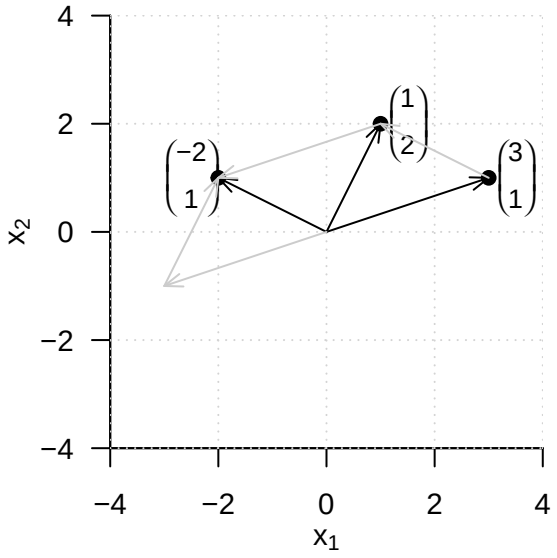


Abbildung 8.3. Vektorsubtraktion in $\mathbb{R}^2$

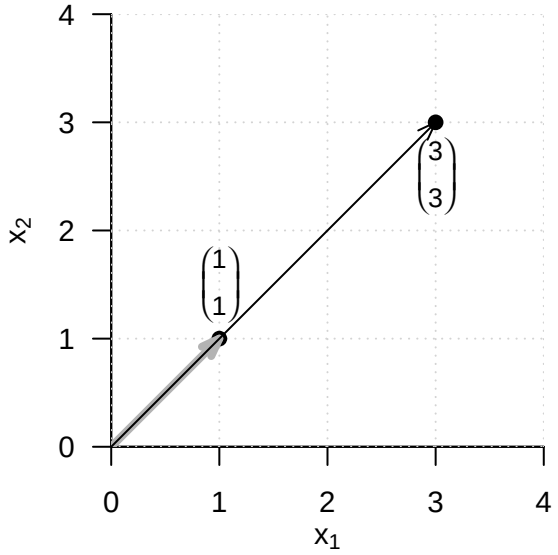


Abbildung 8.4. Skalarmultiplikation in $\mathbb{R}^2$

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

$$\langle\rangle : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, (x, y) \mapsto \langle (x, y) \rangle := \langle x, y \rangle := \sum_{i=1}^{m} x_i y_i. \tag{8.9}$$

•

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

Es seien

$$x := \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \text{ und } y := \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} \tag{8.10}$$

Dann ergibt sich

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + x_3 y_3 = 1 \cdot 2 + 2 \cdot 0 + 3 \cdot 1 = 2 + 0 + 3 = 5. \tag{8.11}$$

In **R** gibt es verschiedene Möglichkeiten, ein Skalarprodukt auszuwerten. Wir führen zwei von ihnen für das gebebene Beispiel untenstehend auf.

```r
# 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
```

```r
# 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* erweiter 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

$$\|x\| := \sqrt{\langle x, x\rangle}. \tag{8.12}$$

(2) Der *Abstand* zweier Vektoren $x, y \in \mathbb{R}^m$ ist definiert als

$$d(x, y) := \|x - y\|. \tag{8.13}$$

(3) Der *Winkel* α zwischen zwei Vektoren $x, y \in \mathbb{R}^m$ mit $x, y \neq 0$ ist definiert durch

$$0 \leq \alpha \leq \pi \text{ und } \cos\alpha := \frac{\langle x, y\rangle}{\|x\|\|y\|} \tag{8.14}$$

•

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)

$$\left\| \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right\| = \sqrt{\left\langle \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right\rangle} = \sqrt{2^2 + 0^2} = \sqrt{4} = 2.00 \tag{8.15}$$

```
norm(matrix(c(2,0),nrow = 2), type = "2")              # Vektorlänge = l_2 Norm
```

```
[1] 2
```

Beispiel (2)

$$\left\| \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\| = \sqrt{\left\langle \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\rangle} = \sqrt{2^2 + 2^2} = \sqrt{8} \approx 2.83 \tag{8.16}$$

```
norm(matrix(c(2,2),nrow = 2), type = "2")              # Vektorlänge = l_2 Norm
```

```
[1] 2.828427
```

Beispiel (3)

$$\left\| \begin{pmatrix} 2 \\ 4 \end{pmatrix} \right\| = \sqrt{\left\langle \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix} \right\rangle} = \sqrt{2^2 + 4^2} = \sqrt{20} \approx 4.47 \tag{8.17}$$

```
norm(matrix(c(2,4),nrow = 2), type = "2")              # Vektorlänge = l_2 Norm
```

```
[1] 4.472136
```

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

$$d(x, y) \geq 0, d(x, x) = 0 \text{ und } d(x, y) = d(y, x) \tag{8.18}$$

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,

$$d(x, y) \leq d(x, z) + d(z, y). \tag{8.19}$$

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.

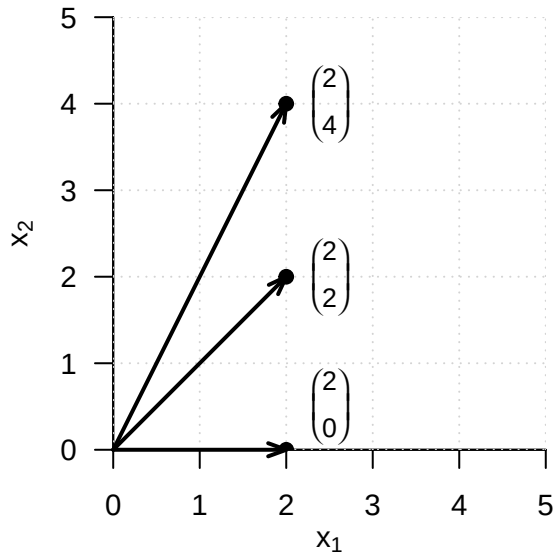


Abbildung 8.5. Vektorlänge in $\mathbb{R}^2$

Beispiel (1)

$$d\left(\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}2\\2\end{pmatrix}\right)=\left\|\begin{pmatrix}1\\1\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right\|=\left\|\begin{pmatrix}-1\\-1\end{pmatrix}\right\|=\sqrt{(-1)^2+(-1)^2}=\sqrt{2}\approx 1.41 \tag{8.20}$$

```
norm(matrix(c(1,1),nrow = 2) - matrix(c(2,2),nrow = 2), type = "2")
```

```
[1] 1.414214
```

Beispiel (2)

$$d\left(\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}4\\1\end{pmatrix}\right)=\left\|\begin{pmatrix}1\\1\end{pmatrix}-\begin{pmatrix}4\\1\end{pmatrix}\right\|=\left\|\begin{pmatrix}-3\\0\end{pmatrix}\right\|=\sqrt{(-3)^2+0^2}=\sqrt{9}=3 \tag{8.21}$$

```
norm(matrix(c(1,1),nrow = 2) - matrix(c(1,4),nrow = 2), type = "2")
```

```
[1] 3
```

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 *acos*, der Arkuskosinusfunktion, ist. Auch für den

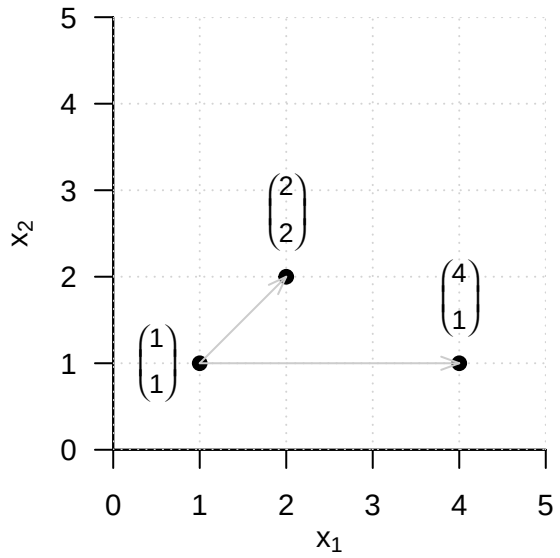


Abbildung 8.6. Vektorabstände in $\mathbb{R}^2$

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)

$$\text{acos}\left(\frac{\left\langle \begin{pmatrix}3\\0\end{pmatrix}, \begin{pmatrix}3\\3\end{pmatrix} \right\rangle}{\left\| \begin{pmatrix}3\\0\end{pmatrix} \right\| \left\| \begin{pmatrix}3\\3\end{pmatrix} \right\|}\right) = \text{acos}\left(\frac{3 \cdot 3 + 3 \cdot 0}{\sqrt{3^2 + 0^2} \cdot \sqrt{3^2 + 3^2}}\right) = \text{acos}\left(\frac{9}{3 \cdot \sqrt{18}}\right) = \frac{\pi}{4} \approx 0.785 \tag{8.22}$$

Die Umrechnung in Grad ergibt dann

$$0.785 \cdot \frac{180°}{\pi} = 45° \tag{8.23}$$

In **R** implementiert man dies wie folgt.

```
x = matrix(c(3,0), nrow = 2)                                  # Vektor 1
y = matrix(c(3,3), 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)

$$\alpha = \text{acos}\left(\frac{\left\langle \begin{pmatrix}3\\0\end{pmatrix}, \begin{pmatrix}0\\3\end{pmatrix} \right\rangle}{\left\| \begin{pmatrix}3\\0\end{pmatrix} \right\| \left\| \begin{pmatrix}0\\3\end{pmatrix} \right\|}\right) = \text{acos}\left(\frac{3 \cdot 0 + 0 \cdot 3}{\sqrt{3^2 + 0^2} \cdot \sqrt{0^2 + 3^2}}\right) = \text{acos}\left(\frac{0}{3 \cdot 3}\right) = \frac{\pi}{2} \approx 1.57 \quad (8.24)$$

Die Umrechnung in Grad ergibt dann

$$\frac{\pi}{2} \cdot \frac{180°}{\pi} = 90° \quad (8.25)$$

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
```

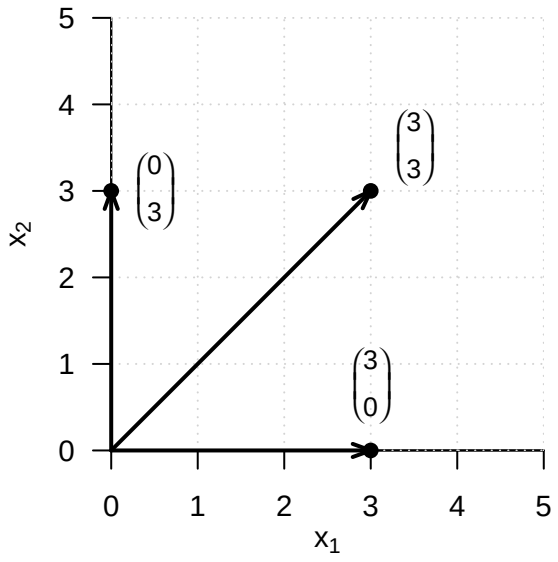


Abbildung 8.7. Winkel in $\mathbb{R}^2$

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

$$\langle x, y\rangle = 0 \tag{8.26}$$

(2) Zwei Vektoren $x, y \in \mathbb{R}^m$ heißen *orthonormal*, wenn gilt, dass

$$\langle x, y\rangle = 0 \text{ und } \|x\| = \|y\| = 1. \tag{8.27}$$

•

Für orthogonale und orthonormale Vektoren gilt also insbesondere auch

$$\cos\alpha = \frac{\langle x, y\rangle}{\|x\|\|y\|} = \frac{0}{\|x\|\|y\|} = 0, \tag{8.28}$$

also

$$\alpha = \frac{\pi}{2} = 90°. \tag{8.29}$$

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)**.** $\{v_1, v_2, ..., v_k\}$ sei eine Menge von k Vektoren eines Vektorraums V und $a_1, a_2, ..., a_k$ seien Skalare. Dann ist die *Linearkombination* der Vektoren in $\{v_1, v_2, ..., v_k\}$ mit den *Koeffizienten* $a_1, a_2, ..., a_k$ definiert als der Vektor

$$w := \sum_{i=1}^{k} a_i v_i \in V. \tag{8.30}$$

•

Beispiel

Es seien

$$v_1 := \begin{pmatrix}2\\1\end{pmatrix}, v_2 := \begin{pmatrix}1\\1\end{pmatrix}, v_3 := \begin{pmatrix}0\\1\end{pmatrix} \text{ und } a_1 := 2, a_2 := 3, a_3 := 0. \tag{8.31}$$

Dann ergibt sich die Linearkombination von v_1, v_2, v_3 mit den Koeffizienten a_1, a_2, a_3 zu

$$\begin{aligned} w &= a_1 v_1 + a_2 v_2 + a_3 v_3 \\ &= 2 \cdot \begin{pmatrix}2\\1\end{pmatrix} + 3 \cdot \begin{pmatrix}1\\1\end{pmatrix} + 0 \cdot \begin{pmatrix}0\\1\end{pmatrix} \\ &= \begin{pmatrix}4\\2\end{pmatrix} + \begin{pmatrix}3\\3\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix} \\ &= \begin{pmatrix}7\\5\end{pmatrix}. \end{aligned} \tag{8.32}$$

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 $W := \{w_1, w_2, ..., w_k\}$ 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*

$$0 = a_1 w_1 + a_2 w_2 + \cdots + a_k w_k \text{ mit } a_1 = a_2 = \cdots = a_k = 0 \tag{8.33}$$

ist. Wenn die Menge W nicht linear unabhängig ist, dann heißt sie *linear abhängig*.

•

Um zu prüfen, ob eine gegeben 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

$$v_1 = \lambda v_2 \text{ mit } \lambda \neq 0. \tag{8.34}$$

Dann gilt

$$v_1 - \lambda v_2 = 0. \tag{8.35}$$

Dies aber entspricht der Linearkombination

$$a_1 v_1 + a_2 v_2 = 0 \tag{8.36}$$

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 $w_1, ..., w_k \in V$ sei eine Menge von Vektoren in V. Wenn einer der Vektoren w_i mit $i = 1, ..., k$ eine Linearkombination der anderen Vektoren ist, dann ist die Menge der Vektoren linear abhängig.*

∘

Beweis. Die Vektoren $w_1, ..., w_k$ 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

$$0 = \sum_{i=1}^{k} a_i w_i = \sum_{i=1, i\neq j}^{k} a_i w_i + a_j w_j \tag{8.37}$$

Also folgt

$$a_j w_j = -\sum_{i=1, i\neq j}^{k} a_i w_i \tag{8.38}$$

und damit

$$w_j = -a_j^{-1} \sum_{i=1, i\neq j}^{k} a_i w_i = -\sum_{i=1, i\neq j}^{k} (a_j^{-1} a_i) w_i \tag{8.39}$$

Also ist w_j eine Linearkombination der w_i für $i = 1, ..., k$ 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 $W := \{w_1, ..., w_k\} \subset V$. Dann ist die *lineare Hülle* von W definiert als die Menge aller Linearkombinationen der Elemente von W,

$$\text{Span}(W) := \left\{ \sum_{i=1}^{k} a_i w_i | a_1, ..., a_k \text{ sind skalare Koeffizienten} \right\} \tag{8.40}$$

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 unendliche 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)**.** $B := \{b_1, ..., b_m\}$ sei eine Basis eines m-dimensionalen Vektorraumes V und es sei $v \in V$. Dann heißt die Linearkombination

$$v = \sum_{i=1}^{m} c_i b_i \tag{8.41}$$

die *Darstellung von v bezüglich der Basis B* und die Koeffizienten $c_1, ..., c_m$ 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

$$\begin{aligned} v &= a_1 b_1 + \cdots + a_m b_m \\ v &= c_1 b_1 + \cdots + c_m b_m \end{aligned} \tag{8.42}$$

Subtraktion der unteren von dern oberen Gleichung ergibt

$$0 = (a_1 - c_1) b_1 + \cdots + (a_m - c_m) b_m \tag{8.43}$$

Weil die $b_1, ..., b_m$ linear unabhängig sind, gilt aber, dass $(a_i - c_i) = 0$ für alle $i = 1, ..., m$ und somit sind die beiden Darstellungen von v bezüglich der Basis B identisch.

□

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 $v_1, ..., v_m \in \mathbb{R}^m$ heißt *Orthonormalbasis* von $\mathbb{R}^m$, wenn $v_1, ..., v_m$ jeweils die Länge 1 haben und wechselseitig orthogonal sind, also wenn

$$\langle v_i, v_j \rangle = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases}. \tag{8.44}$$

•

Wir wollen zunächst ein Beispiel für eine Orthonormalbasis betrachten.

Beispiel (1)

Es ist

$$B_1 := \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \tag{8.45}$$

eine Orthonormalbasis von $\mathbb{R}^2$, denn B_1 besteht aus zwei Vektoren und es gelten

$$\left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\rangle = 1 \cdot 1 + 0 \cdot 0 = 1 + 0 = 1 \tag{8.46}$$

sowie

$$\left\langle \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\rangle = 0 \cdot 0 + 1 \cdot 1 = 0 + 1 = 1 \tag{8.47}$$

und

$$\left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\rangle = 1 \cdot 0 + 0 \cdot 1 = 0 + 0 = 0 \tag{8.48}$$

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

$$B := \left\{ e_1, ..., e_m | e_{i_j} = 1 \text{ für } i = j \text{ und } e_{i_j} = 0 \text{ für } i \neq j \right\} \subset \mathbb{R}^m \tag{8.49}$$

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

$$B := \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}. \tag{8.50}$$

Allerdings gibt es auch nicht kanonische Orthonormalbasen. Dazu betrachten wir ein weiteres Beispiel

Beispiel (2)

Es ist auch

$$B_2 := \left\{ \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\} \tag{8.51}$$

eine Orthonormalbasis von $\mathbb{R}^2$, denn B_2 besteht aus zwei Vektoren und es gelten

$$\left\langle \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} + \frac{1}{2} = 1, \tag{8.52}$$

sowie

$$\left\langle \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = \left(-\frac{1}{\sqrt{2}}\right) \cdot \left(-\frac{1}{\sqrt{2}}\right) + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} + \frac{1}{2} = 1 \tag{8.53}$$

und

$$\left\langle \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\rangle = -\frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = -\frac{1}{2} + \frac{1}{2} = 0 \tag{8.54}$$

Wir visualisieren die beiden Orthonormalbasen B_1 und B_2 von $\mathbb{R}^2$ in Abbildung 8.8.

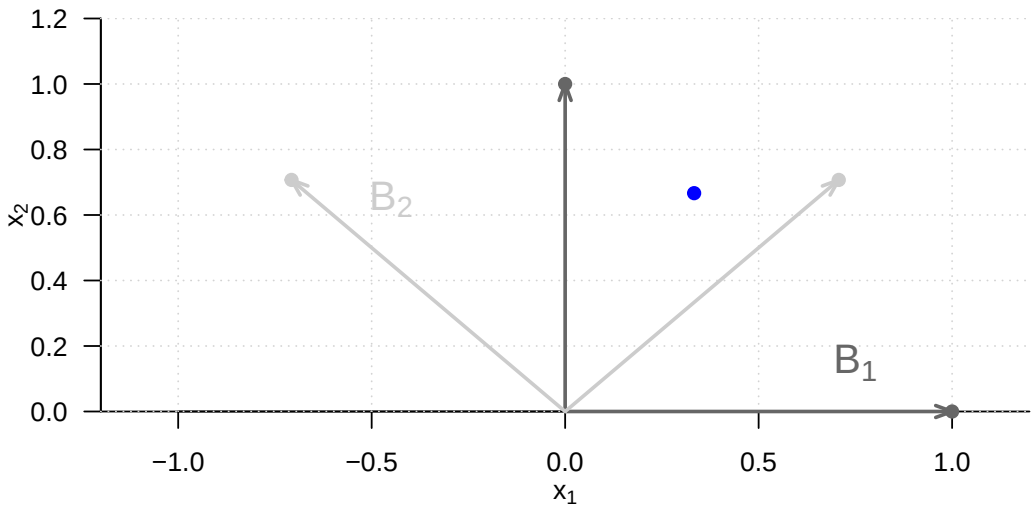


Abbildung 8.8. Zwei Basen von $\mathbb{R}^2$

8.5. Selbstkontrollfragen

1. Geben Sie die Definition eines Vektorraums wieder.
2. Geben Sie die Definition des reellen Vektorraums wieder.
3. Es seien
$$x := \begin{pmatrix} 2 \\ 1 \end{pmatrix}, y := \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ und } a := 2. \tag{8.55}$$
Berechnen Sie
$$v = a(x + y) \text{ und } w = \frac{1}{a}(y - x) \tag{8.56}$$
4. Geben Sie die Definition des Skalarproduktes auf $\mathbb{R}^m$ wieder.
5. Für
$$x := \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}, y := \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, z := \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} \tag{8.57}$$
berechnen Sie
$$\langle x, y \rangle, \langle x, z \rangle, \langle y, z \rangle \tag{8.58}$$
6. Geben Sie die Definition des Euklidischen Vektorraums wieder.
7. Geben Sie die Definition der Länge eines Vektors im Euklidischen Vektorraum wieder,
8. Berechnen Sie die Längen der Vektoren x, y, z aus Gleichung 8.57.
9. Geben Sie Definition des Abstands zweier Vektoren im Euklidischen Vektorraum wieder.
10. Berechnen Sie $d(x, y), d(x, z)$ und $d(y, z)$ für x, y, z aus Gleichung 8.57.
11. Geben Sie die Definition des Winkels zwischen zwei Vektoren im Euklidischen Vektorraum wieder.
12. Berechnen Sie die Winkel zwischen den Vektoren x und y, x und z, sowie y und z aus Gleichung 8.57.
13. Geben Sie die Definitionen der Orthogonalität und Orthonormalität von Vektoren wieder.
14. Geben Sie die Definition der Linearkombination von Vektoren wieder.
15. Geben Sie die Definition der linearen Unabhängigkeit von Vektoren wieder.

16. Woran kann man erkennen, ob zwei reelle Vektoren linear abhängig sind oder nicht?
17. Geben Sie die Definition der linearen Hülle einer Menge von Vektoren wieder.
18. Geben Sie die Definition der Basis eines Vektorraums wieder.
19. Geben Sie das Theorem zu den Eigenschaften von Vektorraumbasen wieder.
20. Geben Sie die Definition der Basisdarstellung eines Vektors wieder.
21. Geben Sie die Definition eienr Orthonormalbasis von $\mathbb{R}^m$ wieder.
22. Geben Sie die Definition der kanonischen Basis von $\mathbb{R}^m$ wieder.

Teil II.

Wahrscheinlichkeitstheorie

Vorbemerkungen

Die Wahrscheinlichkeitstheorie ist ein mathematisches Modell zur Beschreibung von und zum quantitativen Schlussfolgern über *Zufallsvorgänge* der Wirklichkeit (Abbildung 8.9). 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.

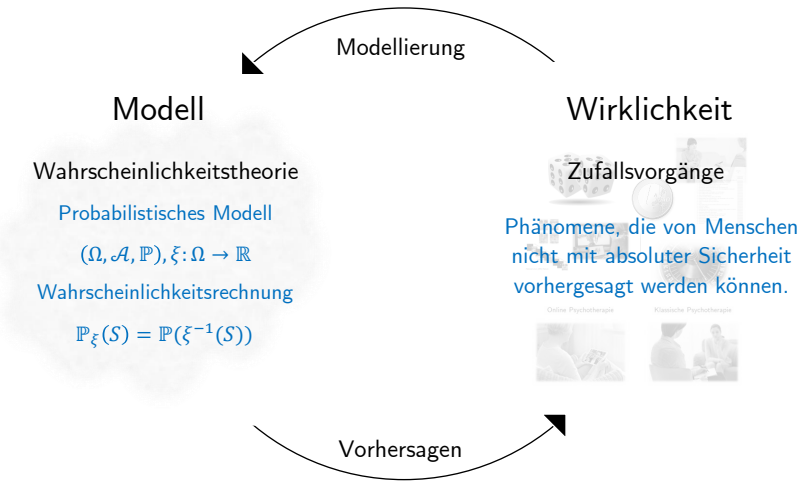


Abbildung 8.9. *Wahrscheinlichkeitstheorie als Modell von Zufallsvorgängen.* Ausgangspunkt der Wahrscheinlichkeitstheorie ist die Absicht, über einen Zufallsvorgang, also ein mit Unsicherheit behaftetes Phänomen der Wirklichkeit, logisch-quantitative Schlüsse zu ziehen. Die Repräsentation zentraler Aspekte des Zufallsvorgang 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 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 9 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 einzutreten 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 wiederrum 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 A zumisst. 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 Wahrscheinlichkeiten 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.

9. 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 9.1) und geben dann mithilfe von Wahrscheinlichkeitsfunktionen (Kapitel 9.2) erste Beispiele für die Modellierung von Zufallsvorgängen durch Wahrscheinlichkeitsräume (Kapitel 9.3).

9.1. 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 9.1 (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 $A^c := \Omega \setminus A \in \mathcal{A}$ ist,
  - aus $A_1, A_2, ... \in \mathcal{A}$ folgt, dass auch $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$ ist,

  ist, *σ-Algebra auf Ω* genannt wird und *Ereignissystem* heißt,
- $\mathbb{P}$ eine Abbildung der Form $\mathbb{P} : \mathcal{A} \to [0, 1]$ mit den Eigenschaften
  - $\mathbb{P}(A) \geq 0$ für alle $A \in \mathcal{A}$ (*Nicht-Negativität*),
  - $\mathbb{P}(\Omega) = 1$ (*Normiertheit*) und
  - $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ 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 $\omega \in \Omega$ selbt 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 Ergeignisse 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 allgemeinster Form, dass aus $A_1, A_2, ... \in \mathcal{A}$ folgen soll, dass auch $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$ 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 9.1 (Ereignissystem bei endlicher Ergebnismenge)**.** $\Omega := \{\omega_1, \omega_2, ..., \omega_n\}$ *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 Ereignisssytem 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 n Teilmengen $A_1, A_2, ..., A_n \subseteq \Omega$. Dann ist $\cup_{i=1}^{n} A_i$ die Menge der $\omega \in \Omega$ für die gilt, dass $\omega \in A_1$ und/oder $\omega \in A_2$ … und/oder $\omega \in A_n$. Da für alle diese ω gilt, dass $\omega \in \Omega$ ist also auch $\cup_{i=1}^{n} A_i$ eine Teilmenge von Ω und damit auch die dritte σ-Algebra Eigenschaft erfüllt. Die Potenzmenge erfüllt also die geforderten Eigenschaften an ein Ereignissystem.

□

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 $\mathbb{R}^n$ 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 Wahrscheinlichkeitsraummmodells. 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

$$\mathbb{P} : \mathcal{A} \to [0, 1], A \mapsto \mathbb{P}(A) \tag{9.1}$$

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 $\Omega, \mathcal{A}$ 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-Addivitität 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 9.2 (Wahrscheinlichkeit des unmöglichen Ereignisses)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ *sei ein Wahrscheinlichkeitsraum. Dann gilt*

$$\mathbb{P}(\emptyset) = 0. \tag{9.2}$$

∘

Beweis. Für $i = 1, 2, \ldots$ sei $A_i := \emptyset$. Dann ist $A_1, A_2, \ldots$ 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

$$\mathbb{P}(\emptyset) = \mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(\emptyset\right). \tag{9.3}$$

Das unendliche Aufaddieren der Zahl $\mathbb{P}(\emptyset) \in [0, 1]$ soll also wieder $\mathbb{P}(\emptyset)$ ergeben. Dies ist aber nur möglich, wenn $\mathbb{P}(\emptyset) = 0$.

□

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-Additivät, die in der Definition des Wahrscheinlichkeitsraums (nur) für die Vereinigung unendlich vieler disjunkter Ereignisse definiert ist, die σ-Additivät endlich vieler disjunkter Ereignisse, wie sie in in der Anwendung oft vorkommen, impliziert.

Theorem 9.3 (σ-Additivität bei endlichen Folgen disjunkter Ereignisse)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ *sei ein Wahrscheinlichkeitsraum und* $A_1, ..., A_n$ *sei eine endliche Folge paarweise disjunkter Ereignisse. Dann gilt*

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i). \tag{9.4}$$

∘

Beweis. Wir betrachten eine unendliche Folge von paarweise disjunkten Ereignissen $A_1, A_2, ...$ 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

$$\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(A_i\right) = \sum_{i=1}^{n} \mathbb{P}\left(A_i\right) + \sum_{i=n+1}^{\infty} \mathbb{P}\left(A_i\right). \tag{9.5}$$

Mit $\mathbb{P}\left(A_i\right) = \mathbb{P}(\emptyset) = 0$ für $i = n+1, n+2, ...$ folgt dann direkt

$$\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}\left(A_i\right) + 0 = \sum_{i=1}^{n} \mathbb{P}\left(A_i\right). \tag{9.6}$$

□

9.2. Wahrscheinlichkeitsfunktionen

In diesem Abschnitt wollen wir mit den *Wahrscheinlichkeitsfunktionen* eine erste Möglichkeit kennenlernen, für Wahrscheinlichkeitsräume mit endlicher Ergebnismenge Wahrscheinlichkeitsmaße festzulegen. In Kapitel 9.3 nutzen wir dieses Hilfsmittel intensiv, um erste Beispiele für die Modellierung von Zufallsvorgängen mithilfe von Wahrscheinlichkeitsräumen geben zu können. Wir definieren den Begriff der Wahrscheinlichkeitsfunktion wie folgt.

Definition 9.2 (Wahrscheinlichkeitsfunktion)**.** Ω sei eine endliche Menge. Dann heißt eine Funktion $\pi : \Omega \to [0, 1]$ *Wahrscheinlichkeitsfunktion*, wenn gilt, dass

$$\sum_{\omega \in \Omega} \pi(\omega) = 1. \tag{9.7}$$

Sei weiterhin $\mathbb{P}$ ein Wahrscheinlichkeitsmaß. Dann heißt die durch

$$\pi : \Omega \to [0, 1], \omega \mapsto \pi(\omega) := \mathbb{P}(\{\omega\}) \tag{9.8}$$

definierte Funktion *Wahrscheinlichkeitsfunktion* von $\mathbb{P}$ auf Ω.

•

Wir merken an, dass weil $\mathbb{P}$ per Definition σ-additiv ist, insbesondere auch gilt, dass

$$\mathbb{P}(\Omega) = \mathbb{P}(\cup_{\omega\in\Omega}\{\omega\}) = \sum_{\omega\in\Omega}\mathbb{P}(\{\omega\}) = \sum_{\omega\in\Omega}\pi(\omega) = 1. \tag{9.9}$$

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 9.4. *$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum mit endlicher Ergebnismenge und $\pi : \Omega \to [0,1]$ sei eine Wahrscheinlichkeitsfunktion. Dann existiert ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf Ω mit π als Wahrscheinlichkeitsfunktion von $\mathbb{P}$. Dieses Wahrscheinlichkeitsmaß ist definiert als*

$$\mathbb{P} : \mathcal{A} \to [0,1], A \mapsto \mathbb{P}(A) := \sum_{\omega\in A}\pi(\omega). \tag{9.10}$$

∘

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 $A_1, A_2, \ldots \in \mathcal{A}$. Dann gilt

$$\mathbb{P}\left(\cup_{i=1}^{\infty}A_i\right) = \sum_{\omega\in\cup_{i=1}^{\infty}A_i}\pi(\omega) = \sum_{i=1}^{\infty}\sum_{\omega\in A_i}\pi(\omega) = \sum_{i=1}^{\infty}\mathbb{P}(A_i). \tag{9.11}$$

und damit die σ-Addivität von $\mathbb{P}$.

□

Definiert man also für eine gegebene Ergebnismenge Ω eine Funktion $\pi : \Omega \to [0,1]$ 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.

9.3. 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 geignetes Ereignissystem; im Falle einer endlichen Ergebnismenge bietet sich die Potenzmenge $\mathcal{P}(\Omega)$ an, 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 inbesondere durch Definition der Wahrscheinlichkeiten der Elementarereignisse. In der Folge verdeutlichen wir dieses Vorgehen anhand von Beispielen. Im Falle der überabzählbaren 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 untenstehender Tabelle listen wir sechs dieser 64 Ereignisse in ihrer verbalen Beschreibung und als Teilmenge A von Ω auf.

Tabelle 9.1. 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 9.2 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

$$\mathbb{P}(\{\omega\}) := \frac{1}{|\Omega|} := 1/6 \text{ für alle } \omega \in \Omega \tag{9.12}$$

wählen. Für ein Modell eines verfälschten Würfels, der das Werfen einer Sechs bevorzugt, könnte man zum Beispiel definieren, dass

$$\mathbb{P}(\{\omega\}) := \frac{1}{8} \text{ für } \omega \in \{1, 2, 3, 4, 5\} \text{ und } \mathbb{P}(\{6\}) := \frac{3}{8}. \tag{9.13}$$

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

$$\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(\{2\} \cup \{4\} \cup \{6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}. \tag{9.14}$$

Im Fall des obigen Modells eines verfälschten Würfels ergibt sich für das gleiche Ereignis die Wahrscheinlichkeit zu

$$\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(\{2\} \cup \{4\} \cup \{6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{8} + \frac{1}{8} + \frac{3}{8} = \frac{5}{8}. \tag{9.15}$$

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 blauem und einem roten Würfel

Wir wollen nun das gleichzeitige Werfen eines blauen und eines roten Würfels modellieren. Dazu ist es sinnvoll, die Ergebnismenge als

$$\Omega := \{(r, b) | r \in \{1, 2, 3, 4, 5, 6\}, b \in \{1, 2, 3, 4, 5, 6\}\} \tag{9.16}$$

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 untenstehender Tabelle listen wir sechs dieser Ereignisse in ihrer verbalen Beschreibung und als Teilmenge A von Ω auf.

Tabelle 9.2. 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 gefallenen 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

$$\mathbb{P}(\{\omega\}) := \frac{1}{|\Omega|} = \frac{1}{36} \text{ für alle } \omega \in \Omega \tag{9.17}$$

wählen. Unter diesem Wahrscheinlichkeitsmaß ergibt sich dann beispielsweise die Wahrscheinlichkeit für das Ereignis “Die Summe der gefallenen Zahlen ist Vier” mit der σ-Additivät von $\mathbb{P}$ zu

$$\begin{aligned}
\mathbb{P}\left(\{(1,3),(3,1),(2,2)\}\right) &= \mathbb{P}\left(\{(1,3)\} \cup \{(3,1)\} \cup \{(2,2)\}\right) \\
&= \mathbb{P}\left(\{(1,3)\}\right) + \mathbb{P}\left(\{(3,1)\}\right) + \mathbb{P}\left(\{(2,2)\}\right) \\
&= 1/36 + 1/36 + 1/36 \\
&= 1/12.
\end{aligned}$$

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 untenstehender Tabelle gezeigt, komplett auflisten.

Tabelle 9.3. 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

$$\mathbb{P}(\Omega) = 1 \Leftrightarrow \mathbb{P}(\{H\}) + \mathbb{P}(\{T\}) = 1 \Leftrightarrow \mathbb{P}(\{T\}) = 1 - \mathbb{P}(\{H\}). \tag{9.18}$$

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

$$\mathbb{P}(\emptyset) = 0, \mathbb{P}(\{H\}) = 1/2, \mathbb{P}(\{T\}) = 1/2 \text{ und } \mathbb{P}(\{H, T\}) = 1. \tag{9.19}$$

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 9.4. 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

$$\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) := \frac{1}{4} \tag{9.20}$$

definieren.

9.4. 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 Zufallvorgängen 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).

9.5. Selbstkontrollfragen

1. Geben Sie die Definition des Begriffs der σ-Algebra wieder.
2. Geben Sie die Definition des Begriffs des Wahrscheinlichkeitsmaßes wieder.
3. Geben Sie die Definition des Begriffs des Wahrscheinlichkeitsraums wieder.
4. Erläutern Sie den Begriff der Ergebnismenge Ω.
5. Erläutern Sie die stillschweigende Mechanik des Wahrscheinlichkeitsraummodells.
6. Erläutern Sie den Begriff eines Ereignisses $A \in \mathcal{A}$.
7. Erläutern Sie den Begriff des Ereignissystems $\mathcal{A}$.
8. Welche σ-Algebra wählt man sinnvoller Weise für einen Wahrscheinlichkeitsraum mit endlicher Ergebnismenge?
9. Erläutern Sie den Begriff des Wahrscheinlichkeitsmaßes $\mathbb{P}$.
10. Geben Sie die Definition des Begriffs der Wahrscheinlichkeitsfunktion wieder.
11. Warum ist der Begriff der Wahrscheinlichkeitsfunktion bei der Modellierung eines Zufallsvorgans durch einen Wahrscheinlichkeitsraums mit endlicher Ergebnismenge hilfreich?
12. Erläutern Sie die Modellierung des Werfens eines Würfels mithilfe eines Wahrscheinlichkeitsraums.
13. Erläutern Sie die Modellierung des gleichzeitigen Werfens eines roten und eines blauen Würfels mithilfe eines Wahrscheinlichkeitsraums.

10. Elementare 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.

10.1. Gemeinsame Wahrscheinlichkeiten

Der Begriff der gemeinsamen Wahrscheinlichkeit zweier Ereignisse ist wie folgt definiert.

Definition 10.1 (Gemeinsame Wahrscheinlichkeit)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und es seien $A, B \in \mathcal{A}$. Dann heißt

$$\mathbb{P}(A \cap B) \tag{10.1}$$

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

$$A \cap B = \{2, 4, 6\} \cap \{4, 5, 6\} = \{4, 6\}. \tag{10.2}$$

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

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(\{2,4,6\} \cap \{4,5,6\}) \\ &= \mathbb{P}(\{4,6\}) \\ &= \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) \\ &= \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{3}.\end{aligned} \tag{10.3}$$

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 *inklusiven 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

$$A \cup B = \{2,4,6\} \cup \{4,5,6\} = \{2,4,5,6\} \tag{10.4}$$

mit der Interpretation "Es fällt eine gerade Augenzahl und/oder es fällt eine Augenzahl größer als Drei". Für die entsprechende Wahrscheinlichkeit ergibt sich

$$\mathbb{P}(\{2,4,5,6\}) = \frac{2}{3}, \tag{10.5}$$

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 10.1 (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. $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

∘

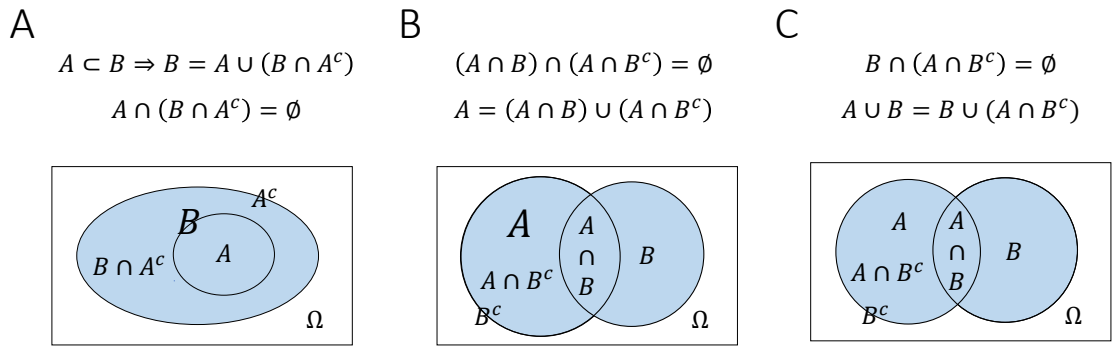


Abbildung 10.1. Venn-Diagramme zum Beweis von Theorem 10.1

Beweis. Die zweite, dritte, und vierte Aussage dieses Theorems basieren auf elementaren mengentheoretischen Aussagen und der σ-Addivität 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 10.1 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 Nomiertheit und der σ-Additivität von $\mathbb{P}$ folgt dann

$$\mathbb{P}(\Omega) = 1 \Leftrightarrow \mathbb{P}(A^c \cup A) = 1 \Leftrightarrow \mathbb{P}(A^c) + \mathbb{P}(A) = 1 \Leftrightarrow \mathbb{P}(A^c) = 1 - \mathbb{P}(A). \tag{10.6}$$

Zu 2.: Zunächst gilt (vgl. Abbildung A)

$$A \subset B \Rightarrow B = A \cup (B \cap A^c) \text{ mit } A \cap (B \cap A^c) = \emptyset. \tag{10.7}$$

Mit der σ-Additivät von $\mathbb{P}$ folgt dann aber

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c). \tag{10.8}$$

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)

$$(A \cap B) \cap (A \cap B^c) = \emptyset \text{ und } A = (A \cap B) \cup (A \cap B^c). \tag{10.9}$$

Mit der σ-Addivität von $\mathbb{P}$ folgt dann

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B^c) \tag{10.10}$$

Zu 4.: Zunächst gilt (vgl. Abbildung C)

$$B \cap (A \cap B^c) = \emptyset \text{ und } A \cup B = B \cup (A \cap B^c). \tag{10.11}$$

Mit der σ-Addivität von $\mathbb{P}$ folgt dann

$$\mathbb{P}(A \cup B) = \mathbb{P}(B) + \mathbb{P}(A \cap B^c). \tag{10.12}$$

Mit 3. folgt dann

$$\mathbb{P}(A \cup B) = \mathbb{P}(B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) \tag{10.13}$$

Zu 5.: Die Aussage folgt direkt aus 4. mit $\mathbb{P}(A \cap B) = \emptyset$ und $\mathbb{P}(\emptyset) = 0$. □

□

10.2. Bedingte Wahrscheinlichkeiten

Wir wenden uns nun dem Begriff der bedingten Wahrscheinlichkeit zu.

Definition 10.2 (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

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{10.14}$$

Weiterhin heißt das für ein festes $B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$ definierte Wahrscheinlichkeitsmaß

$$\mathbb{P}(\ |B) : \mathcal{A} \to [0,1], A \mapsto \mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{10.15}$$

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
- $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i|B\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$ für paarweise disjunkte $A_1, A_2, ... \in \mathcal{A}$.

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

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ und } \mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \tag{10.16}$$

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 10.2 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

$$\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} \tag{10.17}$$

und dass

$$\mathbb{P}(\{4, 5, 6\}) = \mathbb{P}(\{4\}) + \mathbb{P}(\{5\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}. \tag{10.18}$$

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

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{2, 4, 6\} \cap \{4, 5, 6\}) = \mathbb{P}(\{4, 6\}) = \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} \tag{10.19}$$

hat. Nach Definition der bedingten Wahrscheinlichkeit ergibt sich dann direkt

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{4, 6\})}{\mathbb{P}(\{4, 5, 6\})} = \frac{2}{6} \cdot \frac{6}{3} = \frac{2}{3}. \tag{10.20}$$

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 10.2 (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* $P(B) > 0$*. Dann gilt*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A). \tag{10.21}$$

∘

Beweis. Mit der Definition der jeweiligen bedingten Wahrscheinlichkeit folgen direkt

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \tag{10.22}$$

und

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A). \tag{10.23}$$

□

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 10.3 (Gesetz der totalen Wahrscheinlichkeit)**.** *$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $A_1, ..., A_k$ sei eine Partition von Ω. Dann gilt für jedes $B \in \mathcal{A}$, dass*

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B \cap A_i) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i). \tag{10.24}$$

∘

Beweis. Für $i = 1, ..., k$ 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 10.2 mithilfe eines Venn-Diagramms.

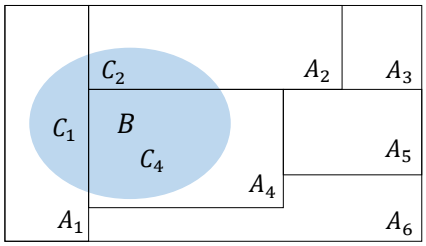


Abbildung 10.2. Venn-Diagramm zum Beweis von Theorem 10.3.

Also gilt

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(C_i) = \sum_{i=1}^{k} \mathbb{P}(B \cap A_i) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i). \tag{10.25}$$

□

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 $i = 1, .., k$ sind.

Schließlich betrachten wir mit dem *Bayesschen Theorem* eine Formel zur alternativen Berechnung von bedingten Wahrscheinlichkeiten.

Theorem 10.4 (Bayessches Theorem)**.** *$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $A_1, ..., A_k$ sei eine Partition von Ω mit $\mathbb{P}(A_i) > 0$ für alle $i = 1, ..., k$. Wenn $\mathbb{P}(B) > 0$ gilt, dann gilt für jedes $i = 1, ..., k$, dass*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}. \tag{10.26}$$

∘

Beweis. Mit der Definition der bedingten Wahrscheinlichkeit und dem Gesetz der totalen Wahrscheinlichkeit gilt

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}. \tag{10.27}$$

□

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.

10.3. 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 10.5. Wir betrachten zunächst die Definition.

Definition 10.3 (Unabhängige Ereignisse)**.** Zwei Ereignisse $A \in \mathcal{A}$ and $B \in \mathcal{A}$ heißen *unabhängig*, wenn

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{10.28}$$

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

$$\mathbb{P}\left(\cap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j). \tag{10.29}$$

•

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 10.3 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 verschiedenenn 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 10.5 (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*

$$\mathbb{P}(A|B) = \mathbb{P}(A). \tag{10.30}$$

∘

Beweis. Unter den Annahmen des Theorems gilt

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A). \tag{10.31}$$

□

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.

10.4. 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.

10.5. Selbstkontrollfragen

1. Geben Sie die Definition der gemeinsamen Wahrscheinlichkeit zweier Ereignisse wieder.
2. Erläutern Sie die intuitive Bedeutung der gemeinsamen Wahrscheinlichkeit zweier Ereignisse.
3. Geben Sie das Theorem zu weiteren Eigenschaften von Wahrscheinlichkeiten wieder.
4. Geben Sie die Definition der bedingten Wahrscheinlichkeit eines Ereignisses wieder.
5. Geben Sie die Definition der bedingten Wahrscheinlichkeit wieder.
6. Geben Sie das Theorem zu gemeinsamen und bedingten Wahrscheinlichkeiten wieder.
7. Geben Sie das Gesetz von der totalen Wahrscheinlichkeit wieder.
8. Geben Sie das Theorem von Bayes wieder.
9. Geben Sie den Beweis des Theorems von Bayes wieder.
10. Geben Sie die Definition der Unabhängigkeit zweier Ereignisse wieder.
11. Geben Sie das Theorem zur bedingten Wahrscheinlichkeit unter Unabhängigkeit wieder.
12. Geben Sie den Beweis des Theorems zur bedingten Wahrscheinlichkeit unter Unabhängigkeit wieder.
13. Erläutern Sie das Theorem zur bedingten Wahrscheinlichkeit unter Unabhängigkeit.

11. 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.

11.1. Konstruktion, Definition und Intuition

Wir skizzieren zunächst die Konstruktion einer Zufallsvariable und ihrer Verteilung anhand von Abbildung 11.1. Dazu sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und

$$\xi : \Omega \to \mathcal{X}, \omega \mapsto \xi(\omega) \tag{11.1}$$

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)

$$\xi^{-1}(S) := \{\omega \in \Omega | \xi(\omega) \in S\}. \tag{11.2}$$

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

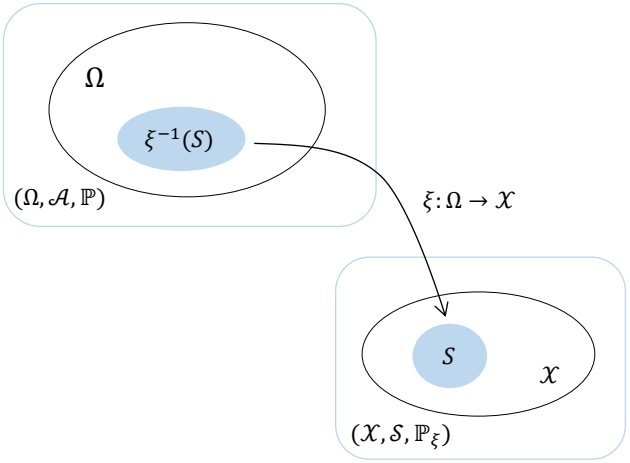
$$\mathbb{P}_\xi : \mathcal{S} \to [0,1], S \mapsto \mathbb{P}_\xi(S) := \mathbb{P}\left(\xi^{-1}(S)\right) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) \in S\}\right) \tag{11.3}$$

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 11.1 (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 $\xi : \Omega \to \mathcal{X}$ mit der *Messbarkeitseigenschaft*

$$\{\omega \in \Omega | \xi(\omega) \in S\} \in \mathcal{A} \text{ für alle } S \in \mathcal{S}. \tag{11.4}$$

•



$$\mathbb{P}\left(\xi^{-1}(S)\right) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \in S\}) =: \mathbb{P}_\xi(S)$$

Abbildung 11.1. Konstruktion von Zufallsvariable und Verteilung.

Nach Definition 11.1 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 modellierten Zufallsvorgangs wird dabei ein ω anhand von $\mathbb{P}$ realisiert und dann (deterministisch) auf $\xi(\omega)$ abgebildet. Wir definieren dementsprechend die Begriff des *Ergebnisraums* und der *Realisierung* einer Zufallsvariable.

Definition 11.2 (Realisierung einer Zufallsvariable)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{S})$ sei ein Messraum und $\xi : \Omega \to \mathcal{X}$ sei eine Zufallsvariable. Dann heißt $\mathcal{X}$ der *Ergebnisraum der Zufallsvariable* ξ und ein $\xi(\omega) \in \mathcal{X}$ heißt eine *Realisierung der Zufallvariable* ξ.

•

Beispiel

Aufbauend auf dem in Kapitel 9.3 betrachteten Beispiel eines Wahrscheinlichkeitsraums zur Modellierung des geichzeitigen Würfelns mit einem blauem und einem roten Würfel wollen wir mit der Summe der Würfelaugenzahlen ein erstes Beispiel für eine Zufallsvariable und ihre Verteilung betrachten. Wir haben in Kapitel 9.3 gesehen, dass ein sinnvolles Wahrscheinlichkeitsraummodell für das geichzeitige Würfeln mit einem blauem und einem roten Würfel durch $(\Omega, \mathcal{A}, \mathbb{P})$ mit

- $\Omega := \{(r, b) | r \in \mathbb{N}_6, b \in \mathbb{N}_6\}$,
- $\mathcal{A} := \mathcal{P}(\Omega)$ und
- $\mathbb{P} : \mathcal{A} \to [0, 1]$ mit $\mathbb{P}(\{(r, b)\}) = 1/36$ für alle $(r, b) \in \Omega$.

gegeben ist. Die Auswertung der Summe der beiden Würfelaugenzahlen wird dann sinnvoller Weise durch die Abbildung

$$\xi : \Omega \to \mathcal{X}, (r, b) \mapsto \xi((r, b)) := r + b, \tag{11.5}$$

beschrieben, wobei offenbar $\mathcal{X} := \{2, 3, ..., 12\}$ gelten muss. Der Ergebnisraum der Zufallsvariable ist also wiederrum endlich und $\mathcal{S} := \mathcal{P}(\mathcal{X})$ ist eine sinnvolle σ-Algebra auf $\mathcal{X}$. Mithilfe der σ-Addivität 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 untenstehender Tabelle gezeigt.

$$
\begin{array}{llll}
\mathbb{P}_\xi(\{2\}) & = \mathbb{P}\left(\xi^{-1}(\{2\})\right) & = \mathbb{P}\left(\{(1,1)\}\right) & = \frac{1}{36} \\
\mathbb{P}_\xi(\{3\}) & = \mathbb{P}\left(\xi^{-1}(\{3\})\right) & = \mathbb{P}\left(\{(1,2),(2,1)\}\right) & = \frac{2}{36} \\
\mathbb{P}_\xi(\{4\}) & = \mathbb{P}\left(\xi^{-1}(\{4\})\right) & = \mathbb{P}\left(\{(1,3),(3,1),(2,2)\}\right) & = \frac{3}{36} \\
\mathbb{P}_\xi(\{5\}) & = \mathbb{P}\left(\xi^{-1}(\{5\})\right) & = \mathbb{P}\left(\{(1,4),(4,1),(2,3),(3,2)\}\right) & = \frac{4}{36} \\
\mathbb{P}_\xi(\{6\}) & = \mathbb{P}\left(\xi^{-1}(\{6\})\right) & = \mathbb{P}\left(\{(1,5),(5,1),(2,4),(4,2),(3,3)\}\right) & = \frac{5}{36} \\
\mathbb{P}_\xi(\{7\}) & = \mathbb{P}\left(\xi^{-1}(\{7\})\right) & = \mathbb{P}\left(\{(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)\}\right) & = \frac{6}{36} \\
\mathbb{P}_\xi(\{8\}) & = \mathbb{P}\left(\xi^{-1}(\{8\})\right) & = \mathbb{P}\left(\{(2,6),(6,2),(3,5),(5,3),(4,4)\}\right) & = \frac{5}{36} \\
\mathbb{P}_\xi(\{9\}) & = \mathbb{P}\left(\xi^{-1}(\{9\})\right) & = \mathbb{P}\left(\{(3,6),(6,3),(4,5),(5,4)\}\right) & = \frac{4}{36} \\
\mathbb{P}_\xi(\{10\}) & = \mathbb{P}\left(\xi^{-1}(\{10\})\right) & = \mathbb{P}\left(\{(4,6),(6,4),(5,5)\}\right) & = \frac{3}{36} \\
\mathbb{P}_\xi(\{11\}) & = \mathbb{P}\left(\xi^{-1}(\{11\})\right) & = \mathbb{P}\left(\{(5,6),(6,5)\}\right) & = \frac{2}{36} \\
\mathbb{P}_\xi(\{12\}) & = \mathbb{P}\left(\xi^{-1}(\{12\})\right) & = \mathbb{P}\left(\{(6,6)\}\right) & = \frac{1}{36}
\end{array}
$$

Die Wahrscheinlichkeiten der Elementarereignisse in $\mathcal{S}$ wiederrum erlauben mithilfe des Begriffs der Wahrscheinlichkeitsmassefunktion (vgl. Kapitel 9.2) das Berechnen beliebiger Ereigniswahrscheinlichkeiten hinsichtlich der Würfelaugenzahlsumme. 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()                                # Ergebnisrauminitialisierung
idx      = 1                                     # Ergebnisindexinitialisierung
for(r in 1:6){                                   # Ergebnisse roter  Würfel
    for(b in 1:6){                               # Ergebnisse blauer Würfel
        Omega[[idx]] = c(r,b)                    # \omega \in \Omega
        idx          = idx + 1 }}                # Ergebnisindexupdate
K        = length(Omega)                         # Kardinalität von \Omega
pi       = rep(1/K,1,K)                          # 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     :  3 3
xi(omega) :  6
```

Im Kontext des Vermessens zufällig ausgewählter experimenteller Einheiten dienen Zufallsvariablen oft als Modelle für Messvorgänge. Betrachten wie beispielsweise die Bestimmung des Wertes eines Intelligenztests an zufällig ausgewählten Proband:innen (Abbildung 11.2), 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 Probandin 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.

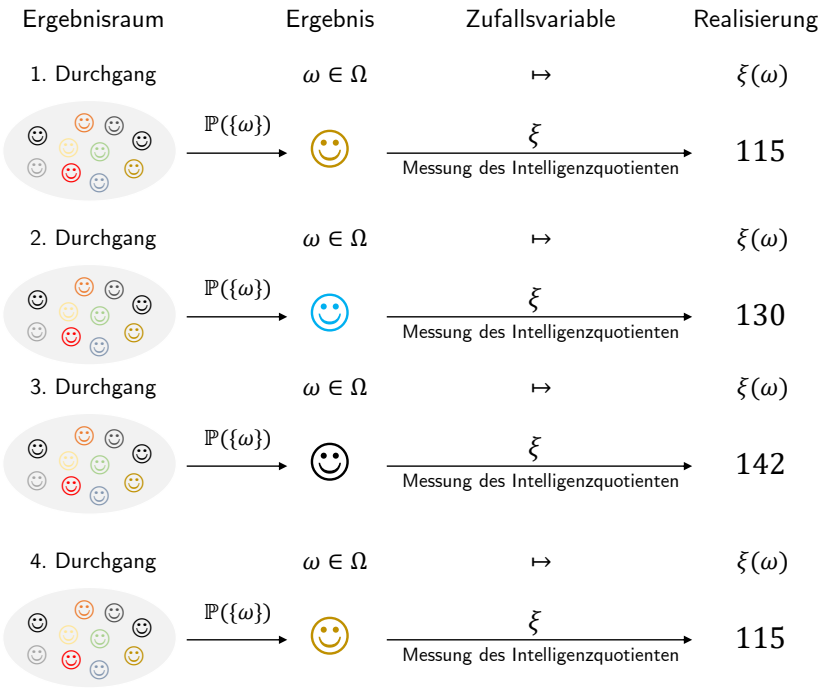


Abbildung 11.2. Zufallsvariable als Modell eines Messvorgangs.

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 11.3 (Notation für Zufallsvariablen)**.** Es seien $(\Omega, \mathcal{A}, \mathbb{P})$ und $(\mathcal{X}, \mathcal{S}, \mathbb{P}_\xi)$ Wahrscheinlichkeitsräume und $\xi : \Omega \to \mathcal{X}$ sei eine Zufallsvariable. Dann gelten mit $S \in \mathcal{S}$ und $x \in \mathcal{X}$ folgende Notationskonventionen für Ereignisse in $\mathcal{A}$:

$$
\begin{aligned}
\{\xi \in S\} &:= \{\omega \in \Omega | \xi(\omega) \in S\} \\
\{\xi = x\} &:= \{\omega \in \Omega | \xi(\omega) = x\} \\
\{\xi \leq x\} &:= \{\omega \in \Omega | \xi(\omega) \leq x\} \\
\{\xi < x\} &:= \{\omega \in \Omega | \xi(\omega) < x\} \\
\{\xi \geq x\} &:= \{\omega \in \Omega | \xi(\omega) \geq x\} \\
\{\xi > x\} &:= \{\omega \in \Omega | \xi(\omega) > x\}
\end{aligned}
$$

Aus diesen Konventionen ergeben sich folgende Konventionen für Wahrscheinlichkeiten von Verteilungen

$$\begin{aligned}
\mathbb{P}_\xi(\xi \in S) &= \mathbb{P}(\{\xi \in S\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \in S\}) \\
\mathbb{P}_\xi(\xi = x) &= \mathbb{P}(\{\xi = x\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) = x\}) \\
\mathbb{P}_\xi(\xi \leq x) &= \mathbb{P}(\{\xi \leq x\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \leq x\}) \\
\mathbb{P}_\xi(\xi < x) &= \mathbb{P}(\{\xi < x\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) < x\}) \\
\mathbb{P}_\xi(\xi \geq x) &= \mathbb{P}(\{\xi \geq x\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \geq x\}) \\
\mathbb{P}_\xi(\xi > x) &= \mathbb{P}(\{\xi < x\}) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) > x\}).
\end{aligned}$$

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 Verwechselungsmöglichkeit der beiden Wahrscheinlichkeitsmaße ergeben kann.

•

Wir wollen diesen Abschnitt mit einem technischem Theorem zum Rechnen mit Zufallsvariablen abschließen.

Theorem 11.1 (Arithmetik reeller Zufallsvariablen)**.** *$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ sei der reelle Messraum, $\xi : \Omega \to \mathbb{R}$, $\upsilon : \Omega \to \mathbb{R}$ seien reellwertige Zufallsvariablen und $c \in \mathbb{R}$ sei eine Konstante. Weiterhin seien*

$$\begin{aligned}
\xi + c : \Omega \to \mathbb{R}, \omega \mapsto (\xi + c)(\omega) &:= \xi(\omega) + c \text{ für } c \in \mathbb{R} \\
c\xi : \Omega \to \mathbb{R}, \omega \mapsto (c\xi)(\omega) &:= c\xi(\omega) \text{ für } c \in \mathbb{R} \\
\xi + \upsilon : \Omega \to \mathbb{R}, \omega \mapsto (\xi + \upsilon)(\omega) &:= \xi(\omega) + \upsilon(\omega) \\
\xi\upsilon : \Omega \to \mathbb{R}, \omega \mapsto (\xi\upsilon)(\omega) &:= \xi(\omega)\upsilon(\omega)
\end{aligned} \tag{11.6}$$

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 $\xi + c$, $c\xi$, $\xi + \upsilon$ und $\xi\upsilon$ reellwertige Zufallsvariablen.

◦

Für einen Beweis dieses Theorems verweisen wir auf die weiterführende Literatur, beispielsweise Hesse (2009). Intuitiv besagt Theorem 11.1, 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.

11.2. 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 11.4 (Diskrete Zufallsvariable und Wahrscheinlichkeitsmassefunktion)**.** Eine Zufallsvariable ξ heißt *diskret*, wenn ihr Ergebnisraum $\mathcal{X}$ endlich oder abzählbar ist und eine Funktion der Form

$$p_\xi : \mathcal{X} \to \mathbb{R}_{\geq 0}, x \mapsto p_\xi(x) \tag{11.7}$$

existiert, für die gilt

(1) $\sum_{x \in \mathcal{X}} p_\xi(x) = 1$ und
(2) $\mathbb{P}_\xi(\xi = x) = p_\xi(x)$ für alle $x \in \mathcal{X}$.

Eine entsprechende Funktion p_ξ 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. Der notationellen Einfachheit halber verzichtet man wie bei den Bildmaßen auch bei WMFen meist auf das Subskript ξ, schreibt also einfach $p(x)$ anstelle von $p_\xi(x)$, wenn aus dem Kontext klar ist, auf welche Zufallsvariable sich die WMF bezieht. Ohne Beweis halten wir fest, dass jede Funktion $p : \mathcal{X} \to \mathbb{R}_{\geq 0}$, die die Normiertheiteigenschaft $\sum_{x \in \mathcal{X}} p(x) = 1$ 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 11.5 (Bernoulli Zufallsvariable)**.** Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} = \{0, 1\}$ und WMF

$$p : \mathcal{X} \to [0, 1], x \mapsto p(x) := \mu^x (1 - \mu)^{1-x} \text{ mit } \mu \in [0, 1]. \tag{11.8}$$

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

$$\text{Bern}(x; \mu) := \mu^x (1 - \mu)^{1-x}. \tag{11.9}$$

•

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 Münzwurfes auf $\{0, 1\}$ ab, also definiert etwa $0 :=$ Heads und $1 :=$ 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

$$\mathbb{P}(\xi = 1) = \mu^1(1 - \mu)^{1-1} = \mu. \tag{11.10}$$

Wir visualisieren die WMFen von Bernoulli-Zufallsvariablen für $\mu := 0.1, \mu := 0.5$ und $\mu := 0.7$ in Abbildung 11.3.

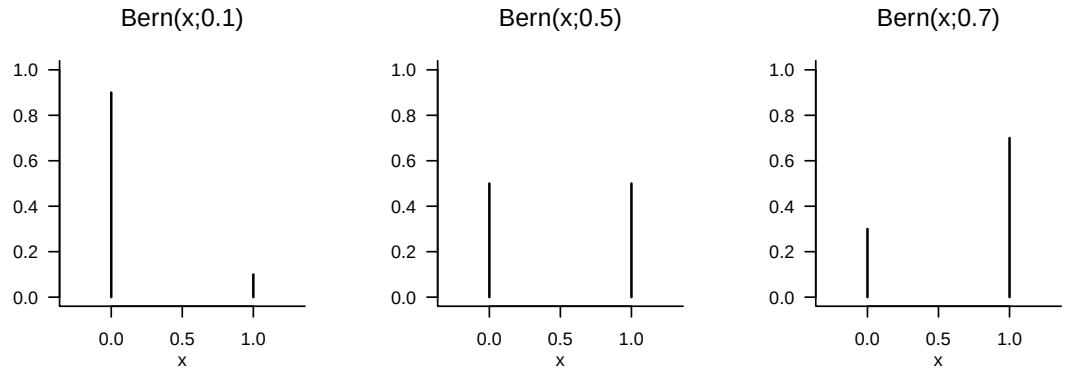


Abbildung 11.3. WMFen von Bernoulli-Zufallsvariablen.

Definition 11.6 (Binomialzufallsvariable)**.** Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := \mathbb{N}_n^0$ und WMF

$$p : \mathcal{X} \to [0, 1], x \mapsto p(x) := \binom{n}{x} \mu^x (1 - \mu)^{n-x} \text{ für } \mu \in [0, 1]. \tag{11.11}$$

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

$$\text{Bin}(x; \mu, n) := \binom{n}{x} \mu^x (1 - \mu)^{n-x}. \tag{11.12}$$

•

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 11.4.

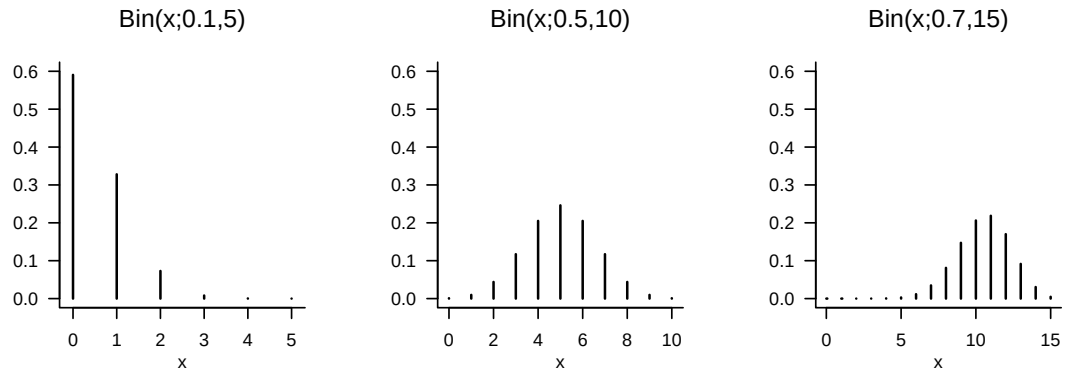


Abbildung 11.4. WMFen von Binomial-Zufallsvariablen.

Definition 11.7 (Diskret-gleichverteilte Zufallsvariable)**.** Es sei ξ eine diskrete Zufallsvariable mit endlichem Ergebnisraum $\mathcal{X}$ und WMF

$$p : \mathcal{X} \to \mathbb{R}_{\geq 0}, x \mapsto p(x) := \frac{1}{|\mathcal{X}|}. \tag{11.13}$$

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

$$U(x; |\mathcal{X}|) := \frac{1}{|\mathcal{X}|}. \tag{11.14}$$

•

Diskrete-gleichverteilte Zufallsvariable können offenbar immer dann zur probabilistischen Modellierung genutzt werden, wenn die möglichen diskreten Ergebnisse des modellierten Phänomenens 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 $\mathcal{X} := \{0, 1\}$.

$$U(x; |\mathcal{X}|) = \text{Bern}(x; 0.5) = \text{Bin}(x; 1, 0.5) \tag{11.15}$$

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 11.5. Man beachte, dass Definition 11.7 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.

11.3. Wahrscheinlichkeitsdichtefunktionen

In diesem Abschnitt führen wir mit den *Wahrscheinlichkeitsdichtefunktionen (WDFen)* ein Hilfsmittel ein, um Verteilungen von Zufallsvariablen mit *kontinuierlichem* (genauer

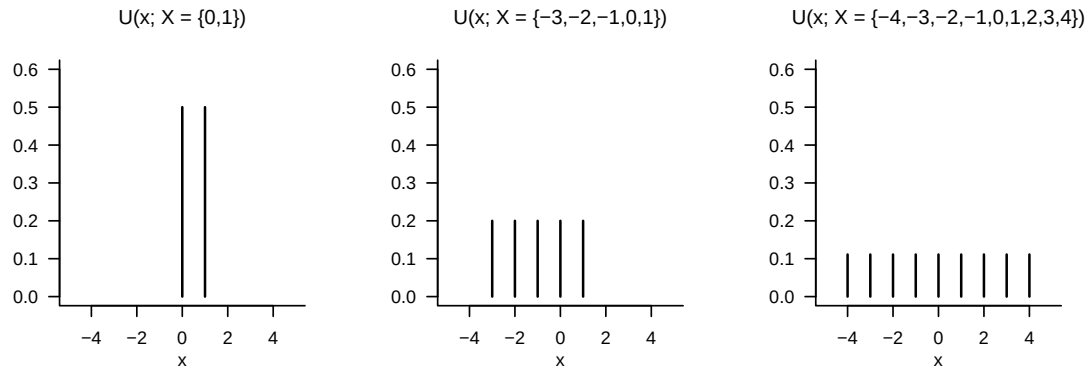


Abbildung 11.5. WMFen von diskret-gleichverteilten Zufallsvariablen.

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 diskret-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 11.8 (Kontinuierliche Zufallsvariable und Wahrscheinlichkeitsdichtefunktion)**.** Eine Zufallsvariable ξ heißt *kontinuierlich*, wenn $\mathbb{R}$ der Ergebnisraum von ξ ist und eine Funktion

$$p_\xi : \mathbb{R} \to \mathbb{R}_{\geq 0}, x \mapsto p_\xi(x) \tag{11.16}$$

existiert, für die gilt

(1) $\int_{-\infty}^{\infty} p_\xi(x)\, dx = 1$ und
(2) $\mathbb{P}_\xi(\xi \in [a, b]) = \int_a^b p_\xi(x)\, dx$ für alle $a, b \in \mathbb{R}$ mit $a \leq b$.

Eine entsprechende Funktion p_ξ heißt *Wahrscheinlichkeitsdichtefunktion (WDF)* von ξ.

•

Im Englischen nennt man WDFen *probability density functions (PDFs)*. Der notationellen Einfachheit halber verzichtet man wie bei den Bildmaßen und den WMFen auch bei WDFen meist auf das Subskript ξ, schreibt also einfach $p(x)$ anstelle von $p_\xi(x)$, wenn aus dem Kontext klar ist, auf welche Zufallsvariable sich die WDF bezieht. Ohne Beweis halten wir fest, dass jede Funktion $p : \mathbb{R} \to \mathbb{R}_{\geq 0}$, für deren uneigentliches Integral $\int_{-\infty}^{\infty} p_\xi(x)\, dx = 1$ 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 *Dichteeigenschaft* einer WDF immer bewusst machen: Die Werte einer WDF stellen keine Wahrscheinlichkeiten, sondern Wahrscheinlichkeitsdichten dar, Wahrscheinlichkeiten werden aus WDFen durch Integration berechnet. Wie im physikalischen Sinn ergibt sich die einem reellen Intervall zugeordnete Wahrscheinlichkeit(smasse) also erst durch "Multiplikation" mit dem enstprechenden "Intervallvolumen". Man denke hierzu auch

an die Approximation des bestimmten Integrals $\int_a^b p_\xi(x)\,dx$ durch einen Riemannschen Summenterm (vgl. Definition 7.3). Intuitiv gilt also

$$\text{(Wahrscheinlichkeits)Masse} = \text{(Wahrscheinlichkeits)Dichte} \cdot \text{(Mengen)Volumen}, \quad (11.17)$$

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,

$$\mathbb{P}_\xi(\xi = a) = \int_a^a p(x)\,dx = 0. \quad (11.18)$$

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 mithile einer WDF nun die wichtigste Zufallsvariable der probabilistischen Modellbildung einführen.

Definition 11.9 (Normalverteilte Zufallsvariable)**.** Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \quad (11.19)$$

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

$$N(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \quad (11.20)$$

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 11.6. 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.

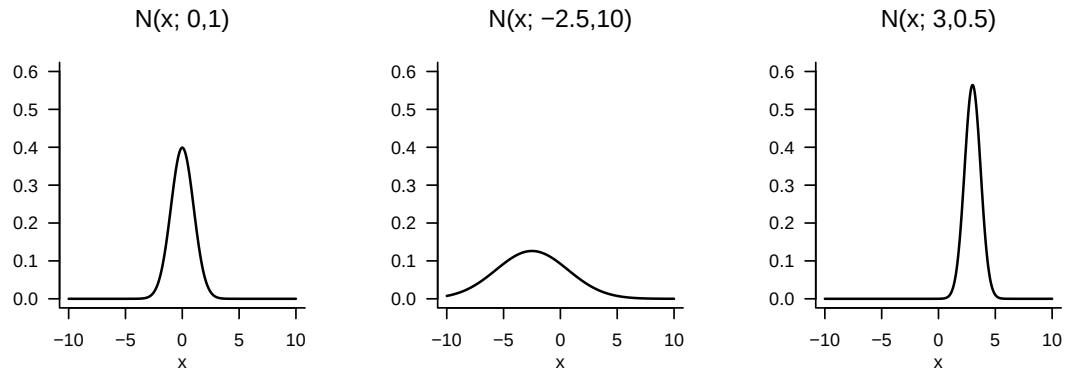


Abbildung 11.6. WDFen von normalverteilten Zufallsvariablen.

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 11.10 (Gamma-Zufallsvariable)**.** Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := \mathbb{R}_{>0}$ und WDF

$$p : \mathbb{R}_{>0} \to \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right), \tag{11.21}$$

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 gammaverteilen Zufallsvariable bezeichnen wir mit

$$G(x; \alpha, \beta) := \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right). \tag{11.22}$$

•

Die spezielle Gamma-Zufallsvariable mit WDF $G\left(x; \frac{n}{2}, 2\right)$ 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 11.7.

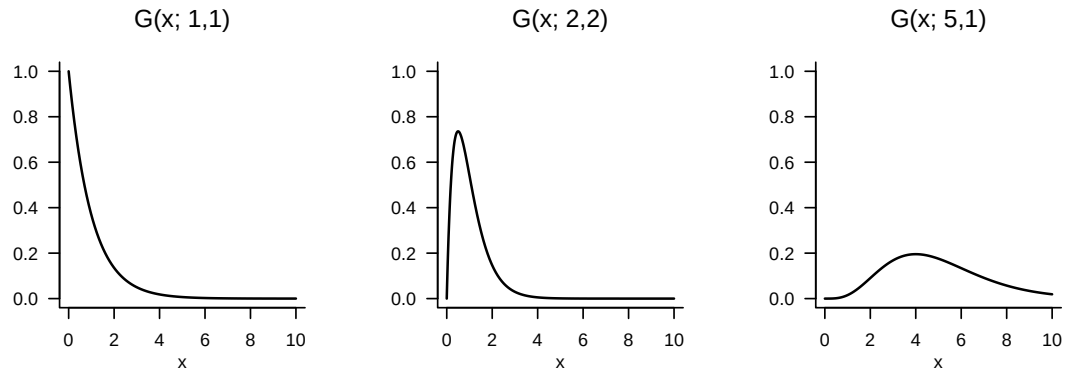


Abbildung 11.7. WDFen von Gamma-Zufallsvariablen.

Definition 11.11 (Beta-Zufallvariable)**.** Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := [0, 1]$ und WDF

$$p : \mathcal{X} \to [0, 1], x \mapsto p(x) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \text{ mit } \alpha, \beta \in \mathbb{R}_{>0}, \tag{11.23}$$

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-verteilte Zufallsvariable. Wir kürzen dies mit $\xi \sim \text{Beta}(\alpha, \beta)$ ab. Die WDF einer Beta-verteilten Zufallsvariable bezeichnen wir mit

$$\text{Beta}(x; \alpha, \beta) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}. \tag{11.24}$$

•

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 11.8.

Mit den gleichverteilten Zufallsvariablen betrachten wir abschließend noch das Analogon zu diskret-gleichverteilten Zufallsvariablen für den Fall kontinuierlicher Zufallsvariablen.

Definition 11.12 (Gleichverteilte Zufallsvariable)**.** Es sei ξ eine kontinuierliche Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \to \mathbb{R}_{\geq 0}, x \mapsto p(x) := \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}. \tag{11.25}$$

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

$$U(x; a, b) := \frac{1}{b - a}. \tag{11.26}$$

•

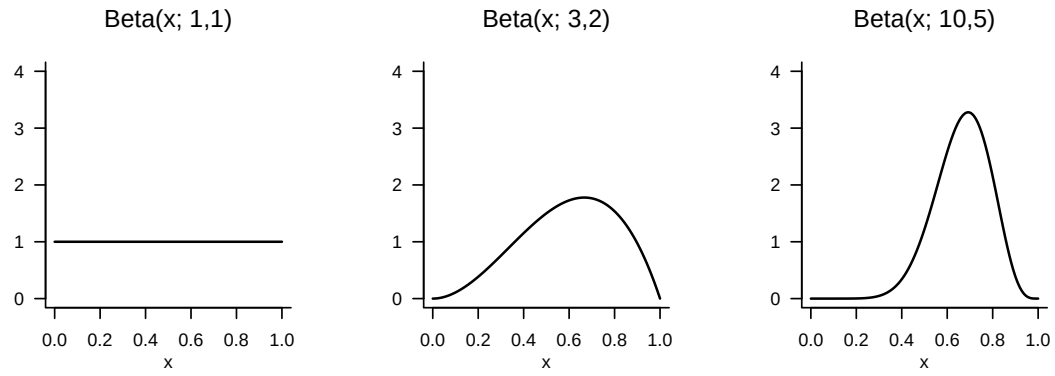


Abbildung 11.8. WDFen von Beta-Zufallsvariablen.

Wir visualisieren die WDFen von gleichverteilten Zufallsvariablen für $(a, b) := (0, 1), (a, b) := (-3, 1)$ und $(a, b) := (-4, 4)$ in Abbildung 11.9.

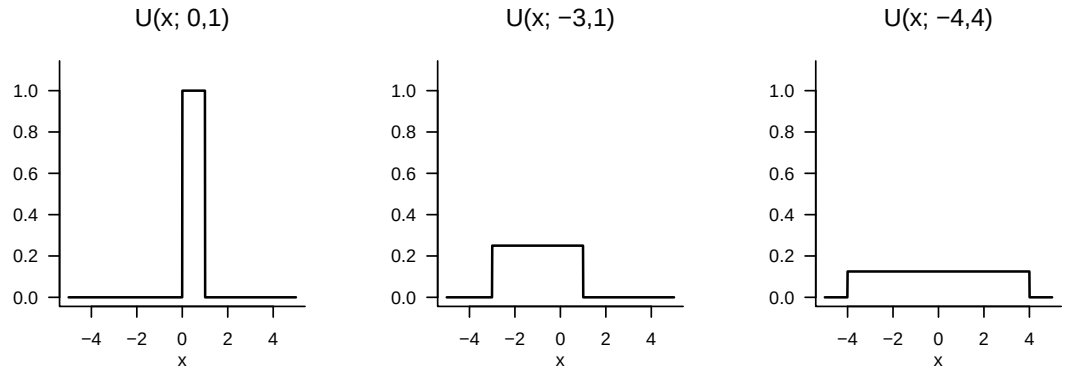


Abbildung 11.9. WDFen von gleichverteilten-Zufallsvariablen.

11.4. 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 nauch 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 kontiniuerlichen Zufallsvariablen im einzelnen zu. Für eine beliebige Zufallsvariable definieren wir den Begriff der KVF wie folgt.

Definition 11.13 (Kumulative Verteilungsfunktion)**.** Die *kumulative Verteilungsfunktion (KVF)* einer Zufallsvariable ξ ist definiert als

$$P_\xi : \mathbb{R} \to [0,1], x \mapsto P_\xi(x) := \mathbb{P}_\xi(\xi \leq x). \tag{11.27}$$

•

Man beachte, dass $P_\xi(x)$ ist für jedes $x \in \mathbb{R}$ definiert ist, auch wenn für ein gegebenes $x \in \mathbb{R}$ gilt, dass $x \notin \mathcal{X}$. Wie bereits für Wahrscheinlichkeitsverteilungen, WMFen und WDFen gesehen verzichtet man meist auf das Subskript $_\xi$, wenn aus dem Kontext klar ist, auf welche Zufallsvariable sich eine gegebene KVF bezieht. Mithilfe von KVFen können zum Beispiel *Überschreitungswahrscheinlichkeiten* und *Intervallwahrscheinlichkeiten* von Zufallsvariablen direkt ausgewertet werden. Dies ist der Inhalt folgender Theoreme.

Theorem 11.2 (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*

$$\mathbb{P}(\xi > x) = 1 - P(x) \text{ für alle } x \in \mathcal{X}. \tag{11.28}$$

∘

Beweis. Die Ereignisse $\{\xi > x\}$ und $\{\xi \leq x\}$ sind disjunkt und

$$\Omega = \{\omega \in \Omega | \xi(\omega) > x\} \cup \{\omega \in \Omega | \xi(\omega) \leq x\} = \{\xi > x\} \cup \{\xi \leq x\}. \tag{11.29}$$

Mit der σ-Additivität von $\mathbb{P}$ folgt dann

$$\begin{aligned} \mathbb{P}(\Omega) &= 1 \\ \Leftrightarrow \mathbb{P}(\{\xi > x\} \cup \{\xi \leq x\}) &= 1 \\ \Leftrightarrow \mathbb{P}(\{\xi > x\}) + \mathbb{P}(\{\xi \leq x\}) &= 1 \\ \Leftrightarrow \mathbb{P}(\{\xi > x\}) &= 1 - \mathbb{P}(\{\xi \leq x\}) \\ \Leftrightarrow \mathbb{P}(\{\xi > x\}) &= 1 - P(x). \end{aligned} \tag{11.30}$$

□

Theorem 11.3 (Intervallwahrscheinlichkeiten)**.** *Es sei ξ eine Zufallsvariable mit Ereignisraum $\mathcal{X}$ und P ihre KVF. Dann gilt für die* Intervallwahrscheinlichkeit $\mathbb{P}(\xi \in]x_1, x_2])$*, dass*

$$\mathbb{P}(\xi \in]x_1, x_2]) = P(x_2) - P(x_1) \text{ für alle } x_1, x_2 \in \mathcal{X} \text{ mit } x_1 < x_2. \tag{11.31}$$

∘

Beweis. Wir betrachten die Ereignisse $\{\xi \leq x_1\}$,$\{x_1 < \xi \leq x_2\}$ und $\{\xi \leq x_2\}$, wobei

$$\{\xi \leq x_1\} \cap \{x_1 < \xi \leq x_2\} = \emptyset \text{ und } \{\xi \leq x_1\} \cup \{x_1 < \xi \leq x_2\} = \{\xi \leq x_2\}. \tag{11.32}$$

gelten. Mit der σ-Additivität von $\mathbb{P}$ gilt dann

$$\begin{aligned} \mathbb{P}(\{\xi \leq x_1\} \cup \{x_1 < \xi \leq x_2\}) &= \mathbb{P}(\{\xi \leq x_2\}) \\ \Leftrightarrow \mathbb{P}(\{\xi \leq x_1\}) + \mathbb{P}(\{x_1 < \xi \leq x_2\}) &= \mathbb{P}(\{\xi \leq x_2\}) \\ \Leftrightarrow \mathbb{P}(\{x_1 < \xi \leq x_2\}) &= \mathbb{P}(\{\xi \leq x_2\}) - \mathbb{P}(\{\xi \leq x_1\}) \\ \Leftrightarrow \mathbb{P}(\{x_1 < \xi \leq x_2\}) &= P(x_2) - P(x_1) \\ \Leftrightarrow \mathbb{P}(\xi \in]x_1, x_2]) &= P(x_2) - P(x_1). \end{aligned} \tag{11.33}$$

□

Folgendes Theorem gibt drei zentrale Eigenschaften von KVFen 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 KVFen, das heißt, jede Funktion P, die die Eigenschaften von Theorem 11.4 erfüllt, kann als eine KVF einer Zufallsvariable interpretiert werden. Für einen Beweis dieser Tatsache verweisen wir auf die weiterführende Literatur.

Theorem 11.4 (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) $\lim_{x\to-\infty} P(x) = 0$ und $\lim_{x\to\infty} P(x) = 1$.
(3) P ist rechtsseitig stetig*, d.h., $P(x) = P(x^+) = \lim_{y\to x, y>x} P(y)$ 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 $x_1 < x_2$,

$$\{\xi \leq x_1\} = \{\omega \in \Omega | \xi(\omega) \leq x_1\} \subset \{\omega \in \Omega | \xi(\omega) \leq x_2\} = \{\xi \leq x_2\}. \tag{11.34}$$

Also gilt

$$\mathbb{P}(\{\xi \leq x_1\}) \leq \mathbb{P}\{\xi \leq x_2\} \Rightarrow P(x_1) \leq P(x_2). \tag{11.35}$$

(2) Für einen Beweis verweisen wir auf die weiterführende Literatur.

(3) Wir definieren

$$P(x^+) = \lim_{y\to x, y>x} P(y). \tag{11.36}$$

Seien nun $y_1 > y_2 > \cdots$ so, dass $\lim_{n\to\infty} y_n = x$. Dann gilt

$$\{\xi \leq x\} = \cap_{n=1}^{\infty} \{\xi \leq y_n\}. \tag{11.37}$$

Es gilt also

$$P(x) = \mathbb{P}(\{\xi \leq x\}) = \mathbb{P}(\cap_{n=1}^{\infty} \{\xi \leq y_n\}) = \lim_{n\to\infty} \mathbb{P}(\{\xi \leq y_n\}) = P(x^+), \tag{11.38}$$

wobei wir die dritte Gleichung unbegründet stehen lassen.

□

KVFen von diskreten Zufallsvariablen

Anhand von Abbildung 11.10 und Abbildung 11.11 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: Wennn $a < b$ 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

$$P : \mathbb{R} \to [0, 1], x \mapsto P(x) := \sum_{k=0}^{\lfloor x \rfloor} \mathbb{P}(\xi = k) \tag{11.39}$$

gegeben, wobei $\lfloor x \rfloor$ die Abrundungsfunktion bezeichnet.

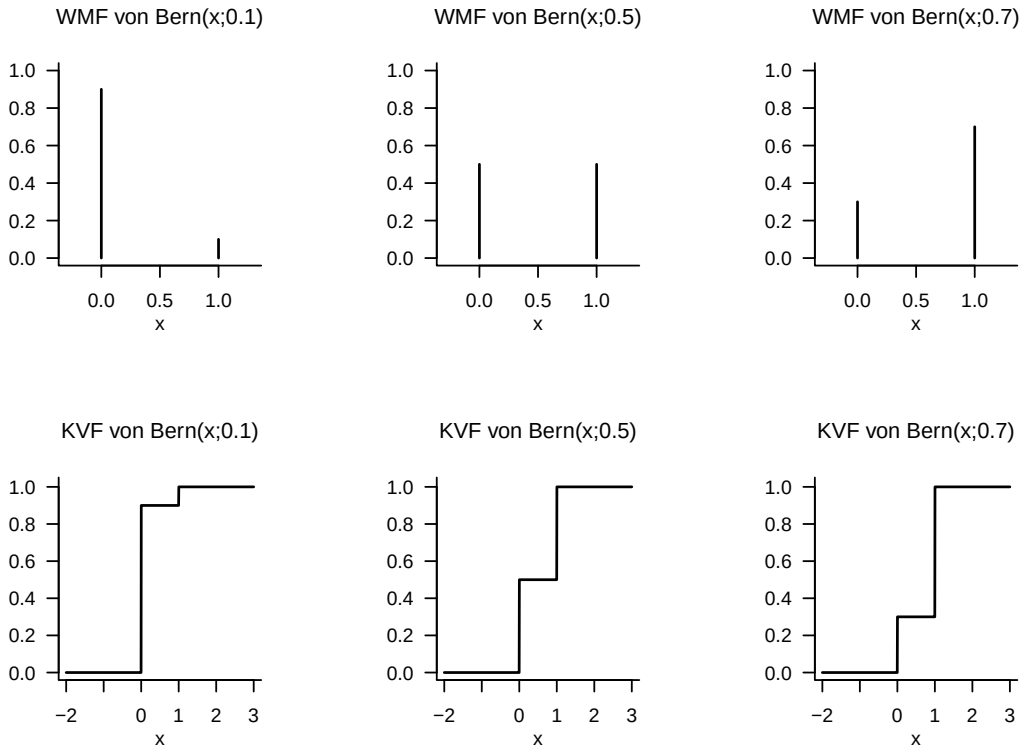


Abbildung 11.10. WMFen und KVFen von Bernoulli-Zufallsvariablen.

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 haben zunächst folgendes, vielleicht etwas überraschendes Theorem.

Theorem 11.5 (Kumulative Verteilungsfunktionen von kontinuierlichen Zufallsvariablen)**.** *ξ sei eine kontinuierliche Zufallsvariable mit WDF p und KVF P. Dann gilt*

$$P(x) = \int_{-\infty}^{x} p(s)\, ds \text{ und } p(x) = P'(x). \tag{11.40}$$

∘

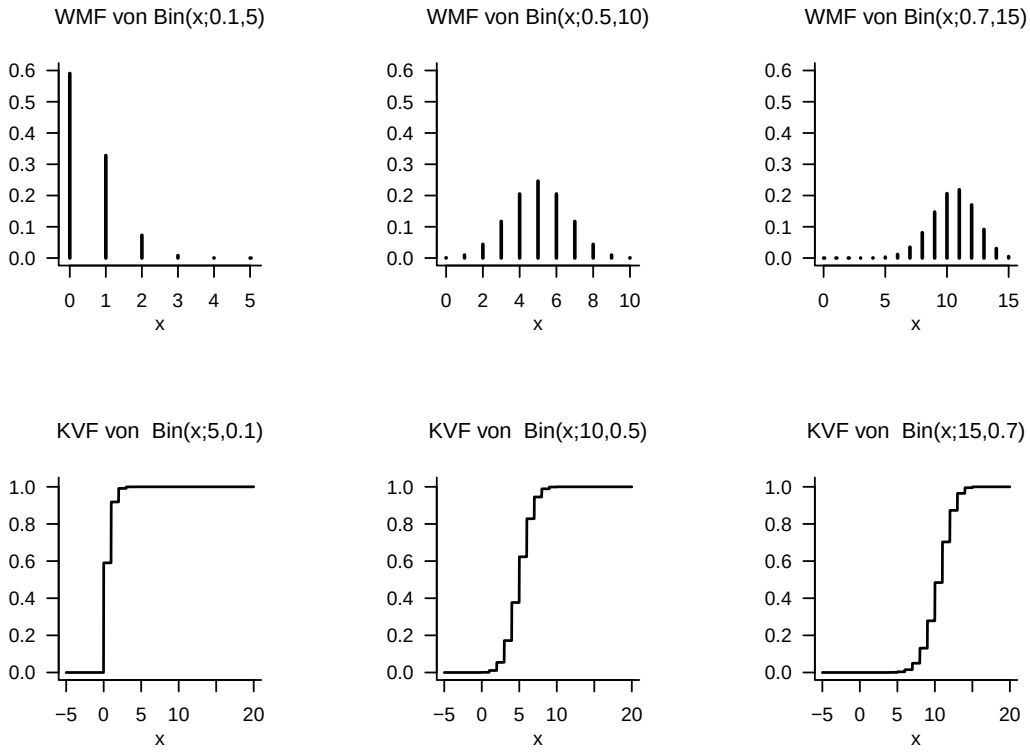


Abbildung 11.11. WMFen und KVFen von Binomial-Zufallsvariablen.

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 der 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, Theorem 7.3.

□

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

$$p : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{11.41}$$

gegeben. Für die KVF von ξ folgt entsprechend, dass

$$P : \mathbb{R} \to]0,1[, x \mapsto P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2\sigma^2}(s-\mu)^2\right) ds. \tag{11.42}$$

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 11.12.

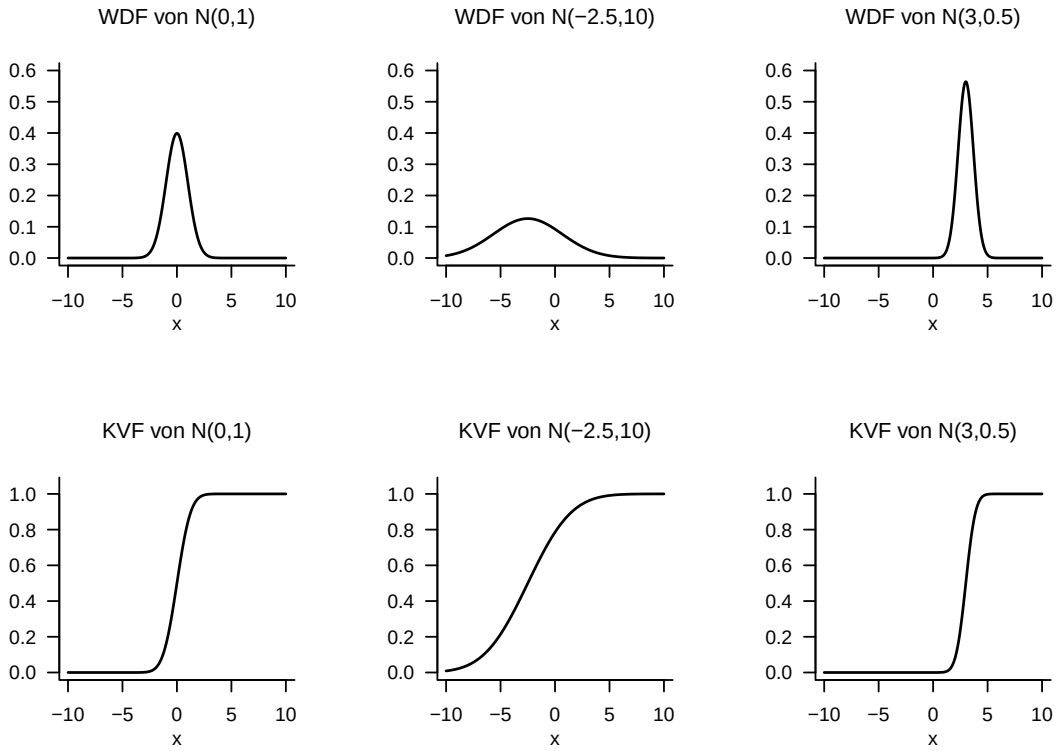


Abbildung 11.12. WDFen und KVFen von normalverteilten Zufallsvariablen.

Inverse Kumulative Verteilungsfunktion

Wir beschließen dieses Kapitel 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 11.14 (Inverse Kumulative Verteilungsfunktion)**.** ξ sei eine kontinuierliche Zufallsvariable mit KVF P. Dann heißt die Funktion

$$P^{-1} :]0,1[\to \mathbb{R}, q \mapsto P^{-1}(q) := \{x \in \mathbb{R} | P(x) = q\} \tag{11.43}$$

die *inverse kumulative Verteilungsfunktion von* ξ.

•

Nach Definition 11.14 gilt offenbar, dass die Funktion P^{-1} die Umkehrfunktion von P ist, also

$$P^{-1}(P(x)) = x. \tag{11.44}$$

Da bekanntlich gilt, dass

$$P(x) = q \Leftrightarrow \mathbb{P}(\xi \leq x) = q \text{ für } q \in]0,1[, \tag{11.45}$$

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 11.13. Im Falle einer normalverteilten Zufallsvariable $\xi \sim N(0, 1)$ gilt zum Beispiel, dass

$$P(1.645) = 0.950 \Leftrightarrow P^{-1}(0.950) = 1.645, \tag{11.46}$$

und

$$P(1.906) = 0.975 \Leftrightarrow P^{-1}(0.975) = 1.960. \tag{11.47}$$

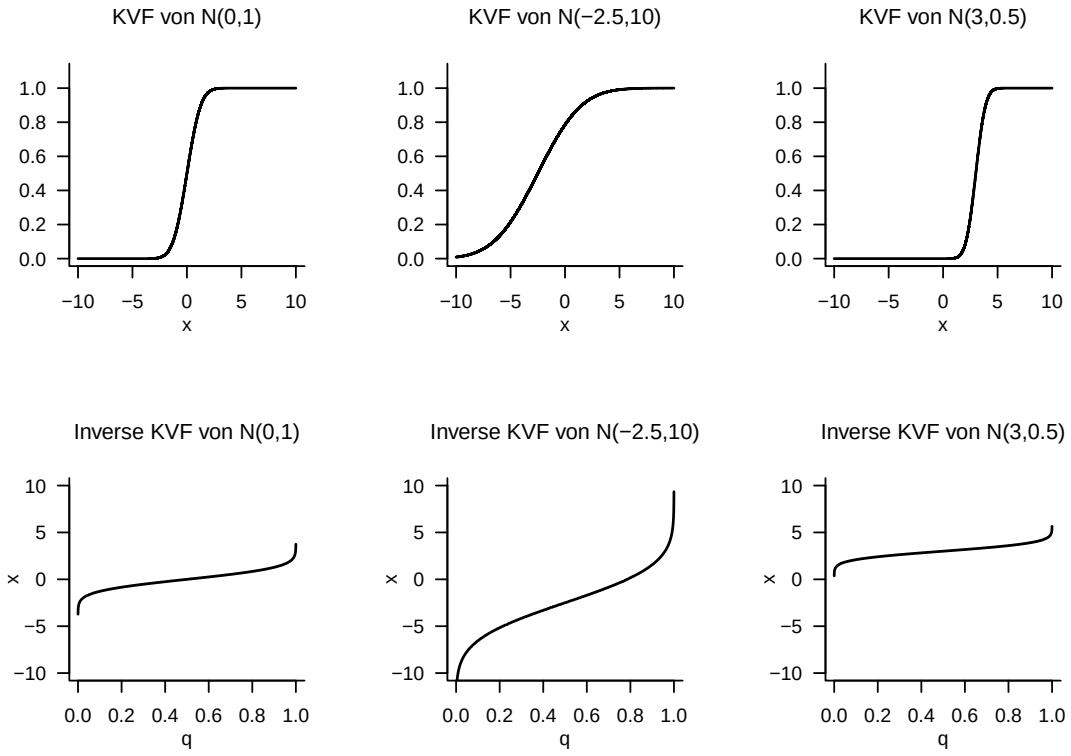


Abbildung 11.13. KVFen und inverse KVFen von normalverteilten Zufallsvariablen.

11.5. Zufallsergebnisse und Zufallsvariablen

Mit den in Kapitel 9 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 11.9) 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

$$\xi : \mathbb{R} \to \mathbb{R}, \omega \mapsto \xi(\omega) := \omega := x \tag{11.48}$$

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 Identititätsabbildung impliziert, über dies hinausgehend aber die Möglichkeit eröffnet, durch von der Identitätsabbildung verschiedene Zufallsvariablen die Transformation von Wahrscheinlichkeitsmaßen auf unterschiedlichen Messräumen zu formalisieren.

11.6. 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.

11.7. Selbstkontrollfragen

1. Geben Sie die Definition des Begriffs der Zufallsvariable wieder.
2. Erläutern Sie die Gleichung $\mathbb{P}_\xi(\xi = x) = \mathbb{P}(\{\xi = x\})$.
3. Erläutern Sie die Bedeutung von $\mathbb{P}(\xi = x)$.
4. Geben Sie die Definition des Begriffs der Wahrscheinlichkeitsmassefunktion wieder.
5. Geben Sie die Definition des Begriffs der Wahrscheinlichkeitsdichtefunktion wieder.
6. Geben Sie die Definition des Begriffs der kumulativen Verteilungsfunktion wieder.
7. Schreiben sie die Intervallwahrscheinlichkeit einer Zufallsvariable mithilfer ihrer KVF.
8. Geben Sie die Definition der WDF einer normalverteilten Zufallsvariable wieder.
9. Geben Sie die Definition der KVF einer normalverteilten Zufallsvariable wieder.
10. Schreiben Sie den Wert $P(x)$ der KVF einer Zufallsvariable mithilfe ihrer WDF.
11. Schreiben Sie den Wert $p(x)$ der WDF einer Zufallsvariable mithilfe ihrer KVF.
12. Geben Sie die Definition des Begriffs der inversen Verteilungsfunktion wieder.

12. 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 12.1). Mit den *marginalen* und *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, das einen Spezialfall der gemeinsamen Verteilung des durch die Zufallsvariablen konstituierten Zufallsvektors darstellt.

12.1. 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 12.1 (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

$$\xi : \Omega \to \mathcal{X}, \omega \mapsto \xi(\omega) := \begin{pmatrix} \xi_1(\omega) \\ \vdots \\ \xi_n(\omega) \end{pmatrix} \tag{12.1}$$

mit der *Messbarkeitseigenschaft*

$$\{\omega \in \Omega | \xi(\omega) \in S\} \in \mathcal{A} \text{ für alle } S \in \mathcal{S}. \tag{12.2}$$

•

Das Standardbeispiel für den Ergebisraum 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-dimensionale 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

$$\xi_i : \Omega \to \mathbb{R}, \omega \mapsto \xi_i(\omega). \tag{12.3}$$

Ohne Beweis halten wir fest, dass der Zufallsvektor ξ messbar ist, wenn für alle $i = 1, ..., n$ 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 Allgemeinem 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 $\xi := (\xi_1, ..., \xi_n)$, notieren.

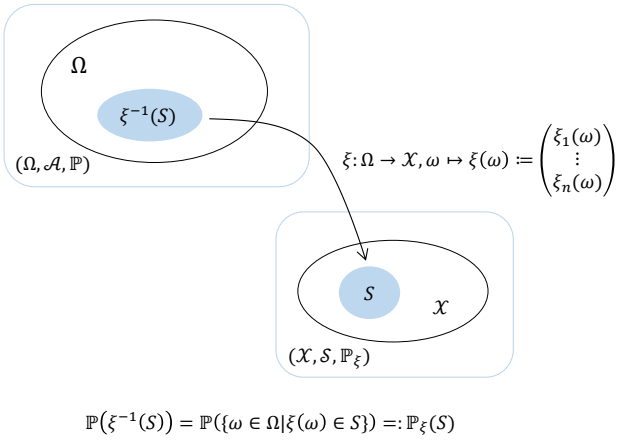


Abbildung 12.1. Konstruktion von Zufallsvektor und multivariater Verteilung.

Das durch die Konstruktion eines Zufallsvektors definierte Bildmaß heißt die *multivariate Verteilung des Zufallsvektors*, wie in folgender Definition ausgeführt (Abbildung 12.1).

Definition 12.2 (Multivariate Verteilung)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{S})$ sei ein n-dimensionaler Messraum und

$$\xi : \Omega \to \mathcal{X}, \omega \mapsto \xi(\omega) \tag{12.4}$$

sei ein Zufallsvektor. Dann heißt das Wahrscheinlichkeitsmaß $\mathbb{P}_\xi$, definiert durch

$$\mathbb{P}_\xi : \mathcal{S} \to [0, 1], S \mapsto \mathbb{P}_\xi(S) := \mathbb{P}(\xi^{-1}(S)) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) \in S\}\right) \tag{12.5}$$

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 11.3 gelten für Zufallsvektoren analog. Zum Beispiel gelten

$$\begin{aligned}
\mathbb{P}_\xi(\xi \in S) &:= \mathbb{P}\left(\{\xi \in S\}\right) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) \in S\}\right) \\
\mathbb{P}_\xi(\xi = x) &:= \mathbb{P}\left(\{\xi = x\}\right) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) = x\}\right) \\
\mathbb{P}_\xi(\xi \leq x) &:= \mathbb{P}\left(\{\xi \leq x\}\right) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) \leq x\}\right) \\
\mathbb{P}_\xi(x_1 \leq \xi \leq x_2) &:= \mathbb{P}\left(\{x_1 \leq \xi \leq x_2\}\right) = \mathbb{P}\left(\{\omega \in \Omega | x_1 \leq \xi(\omega) \leq x_2\}\right)
\end{aligned} \tag{12.6}$$

wobei die Relationsoperatoren $<, \leq, >, \geq$ werden hier *komponentenweise* verstanden werden. So heißt beispielsweise $x \leq y$ für $x, y \in \mathbb{R}^n$, dass für alle Komponenten $x_i, y_i, i = 1, ..., n$ gilt, dass $x_i \leq y_i$. Eben dieser Konvention folgt auch die Definition der *multivariaten kumulativen Verteilungsfunktion* in Generalisierung von Definition 11.13.

Definition 12.3 (Multivariate kumulative Verteilungsfunktionen)**.** ξ sei ein Zufallsvektor mit Ergebnisraum $\mathcal{X}$. Dann heißt eine Funktion der Form

$$P_\xi : \mathcal{X} \to [0, 1],\ x \mapsto P_\xi(x) := \mathbb{P}_\xi(\xi \leq x) \tag{12.7}$$

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 (vgl. Definition 11.4 und Definition 11.8) wie folgt.

Definition 12.4 (Diskreter Zufallsvektor und multivariate Wahrscheinlichkeitsmassefunktion)**.** Ein Zufallsvektor ξ heißt *diskret*, wenn sein Ergebnisraum $\mathcal{X}$ endlich oder abzählbar ist und eine Funktion

$$p_\xi : \mathcal{X} \to [0, 1], x \mapsto p_\xi(x) \tag{12.8}$$

existiert, für die gilt

(1) $\sum_{x \in \mathcal{X}} p(x) = 1$ und
(2) $\mathbb{P}_\xi(\xi = x) = p(x)$ für alle $x \in \mathcal{X}$.

Ein entsprechende Funktion p_ξ 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 Subscript, 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 Zufallsvektors, der 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 untenstehender Tabelle spezifizierten Menge an Tupeln (x_1, x_2).

(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

$$p_\xi : \{1, 2, 3\} \times \{1, 2, 3, 4\} \to [0, 1], (x_1, x_2) \mapsto p_\xi(x_1, x_2) \tag{12.9}$$

ist dann durch nachfolgende Tabelle definiert:

$p_\xi(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$x_1 = 1$	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 spezifierte Funktion p_ξ den Normiertheits- und Nichtnegativitätsansprüchen an eine WMF genügt. Insbesondere gilt hier

$$\sum_{x \in \mathcal{X}} p_\xi(x) = \sum_{x_1=1}^{3} \sum_{x_2=1}^{4} p_\xi(x_1, x_2) = 1. \tag{12.10}$$

Den Begriff des kontinuierlichen Zufallsvektors und der multivariaten Wahrscheinlichkeitsdichtefunktion definieren wir wie folgt.

Definition 12.5 (Kontinuierlicher Zufallsvektor und multivariate Wahrscheinlichkeitdichtefunktion)**.** Ein Zufallsvektor ξ heißt *kontinuierlich*, wenn sein Ergebnisraum durch $\mathbb{R}^n$ gegeben ist und eine Funktion

$$p_\xi : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, x \mapsto p_\xi(x), \tag{12.11}$$

existiert, für die gilt, dass

(1) $\int_{\mathbb{R}^n} p_\xi(x)\, dx = 1$ und
(2) $\mathbb{P}_\xi(x_1 \leq \xi \leq x_2) = \int_{x_{1_1}}^{x_{2_1}} \cdots \int_{x_{1_n}}^{x_{2_n}} p_\xi(s_1, ..., s_n)\, ds_1 \cdots ds_n$.

Eine entsprechende Funktion p_ξ heißt *multivariate Wahrscheinlichkeitsdichtefunktion (WDF) von ξ.*

•

Offenbar ist der 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 identifizieren Subskripte. Wie für kontinuierliche Zufallsvariablen gilt für kontinuierliche Zufallsvektoren

$$\mathbb{P}_\xi(\xi = x) = \mathbb{P}_\xi(x \leq \xi \leq x) = \int_{x_1}^{x_1} \cdots \int_{x_n}^{x_n} p_\xi(s_1, ..., s_n)\, ds_1 \cdots ds_n = 0. \tag{12.12}$$

Das Standardbeispiel für eine multivariate WDF ist die *multivariate Normalverteilung*, welcher wir mit **?@sec-normalverteilungen** ein eigenens Kapitel widmen.

12.2. 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 12.6 (Univariate Marginalverteilung)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{S})$ sei ein n-dimensionaler Messraum, $\xi : \Omega \to \mathcal{X}$ sei ein Zufallsvektor, $\mathbb{P}_\xi$ 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

$$\mathbb{P}_{\xi_i} : \mathcal{S}_i \to [0,1], S \mapsto \mathbb{P}_\xi\left(\mathcal{X}_1 \times \cdots \times \mathcal{X}_{i-1} \times S \times \mathcal{X}_{i+1} \times \cdots \times \mathcal{X}_n\right) \text{ für } S \in \mathcal{S}_i \tag{12.13}$$

definierte Verteilung die *$i$te 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 12.1 (Marginale Wahrscheinlichkeitsmasse und Wahrscheinlichkeitsdichtefunktionen)**.** *(1) $\xi = (\xi_1, ..., \xi_n)$ sei ein n-dimensionaler diskreter Zufallsvektor mit Wahrscheinlichkeitsmassefunktion p_ξ und Komponentenergebnisräumen $\mathcal{X}_1, ..., \mathcal{X}_n$. Dann ergibt sich die Wahrscheinlichkeitsmassefunktion der $i$ten Komponente ξ_i von ξ als*

$$p_{\xi_i} : \mathcal{X}_i \to [0,1], x_i \mapsto$$
$$p_{\xi_i}(x_i) := \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p_\xi(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n). \tag{12.14}$$

(2) $\xi = (\xi_1, ..., \xi_n)$ *sei ein* n*-dimensionaler kontinuierlicher Zufallsvektor mit Wahrscheinlichkeitsdichtefunktion* p_ξ *und Komponentenergebnisraum* $\mathbb{R}$*. Dann ergibt sich die Wahrscheinlichkeitsdichtefunktion der* i*ten Komponente* ξ_i *von* ξ *als*

$$p_{\xi_i} : \mathbb{R} \to \mathbb{R}_{\geq 0}, x_i \mapsto$$
$$p_{\xi_i}(x_i) := \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_n} p_\xi(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n)\, dx_1 ... dx_{i-1}\, dx_{i+1} ... dx_n. \tag{12.15}$$

∘

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 12.1 verweisen wir auf die weiterführende Literatur.

Beispiel

In Fortführung des in Kapitel 12.1 betrachteten Beispiels eines zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ ergeben sich für die dort definierte WMF für die marginalen WMFen p_{ξ_1} und p_{ξ_2} die an den Rändern der unten spezifizierter Tabelle aufgelisteten WMFen anhand von

$$p_{\xi_1}(x_1) = \sum_{x_2=1}^{4} p_\xi(x_1, x_2) \text{ und } p_{\xi_2}(x_2) = \sum_{x_1=1}^{3} p_\xi(x_1, x_2) \tag{12.16}$$

zu

$p_\xi(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$p_{\xi_1}(x_1)$
$x_1 = 1$	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
$p_{\xi_2}(x_2)$	0.2	0.3	0.3	0.2	

Für die Werte von p_{ξ_1} werden die entsprechenden Werte von p_ξ also zeilenweise und für die Werte von p_{ξ_2} spaltenweise addiert. Man beachte, dass aus der Normiertheit von p_ξ die Normiertheit von p_{ξ_1} und p_{ξ_2} direkt folgt, da sich die Gesamtsumme an Wahrscheinlichkeitsmasse nicht ändert:

$$1 = \sum_{x_1=1}^{3} \sum_{x_2=1}^{4} p_\xi(x_1, x_2) = \sum_{x_1=1}^{3} p_{\xi_1}(x_1) = \sum_{x_2=1}^{4} p_{\xi_2}(x_2). \tag{12.17}$$

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 $p_\xi(x_1, x_2)$ 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 $p_\xi(3, 2) \approx 12/100 = 0.12$ 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 $p_{\xi_2}(2)$ natürlicherweise durch

$$\frac{0 + 22 + 12}{100} = \frac{0}{100} + \frac{22}{100} + \frac{12}{100} = 0.00 + 0.22 + 0.12 = 0.34 \tag{12.18}$$

schätzen. Anstelle der Wahrscheinlichkeiten $p_\xi(1, 2)$, $p_\xi(2, 2)$, $p_\xi(3, 2)$ addiert man hier also die entsprechenden relativen Häufigkeiten.

Marginale Verteilungen im Fall von kontinuierlichen Zufallsvektoren behandeln wir am Standardbeispiel der multivariaten Normalverteilung in **?@sec-normalverteilungen**.

12.3. 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 10.2) 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

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{12.19}$$

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

$$A := \{\xi_1 \in S_1\} \text{ und } B := \{\xi_2 \in S_2\} \tag{12.20}$$

definiert. So ergibt sich zum Beispiel die bedingte Wahrscheinlichkeit, dass $\xi_1 \in S_1$ gegeben, dass $\xi_2 \in S_2$ unter der Annahme, dass $\mathbb{P}(\{\xi_2 \in S_2\}) > 0$, zu

$$\mathbb{P}(\{\xi_1 \in S_1\}|\{\xi_2 \in S_2\}) = \frac{\mathbb{P}(\{\xi_1 \in S_1\} \cap \{\xi_2 \in S_2\})}{\mathbb{P}(\{\xi_2 \in S_2\})}. \tag{12.21}$$

Wir betrachten zunächst die Definition der bedingten Verteilungen von diskreten Zufallsvektoren, die lediglich aus zwei Zufallsvariablen bestehen.

Definition 12.7 (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 $p_\xi = p_{\xi_1, \xi_2}$ und marginalen WMFen p_{ξ_1} und p_{ξ_2}. Die bedingte WMF von ξ_1 gegeben $\xi_2 = x_2$ ist dann für $p_{\xi_2}(x_2) > 0$ definiert als

$$p_{\xi_1|\xi_2=x_2} : \mathcal{X}_1 \to [0, 1], x_1 \mapsto p_{\xi_1|\xi_2=x_2}(x_1|x_2) := \frac{p_{\xi_1,\xi_2}(x_1, x_2)}{p_{\xi_2}(x_2)} \tag{12.22}$$

Analog ist für $p_{\xi_1}(x_1) > 0$ die bedingte WMF von ξ_2 gegeben $\xi_1 = x_1$ definiert als

$$p_{\xi_2|\xi_1=x_1} : \mathcal{X}_2 \to [0,1], x_2 \mapsto p_{\xi_2|\xi_1=x_2}(x_1|x_2) := \frac{p_{\xi_1,\xi_2}(x_1,x_2)}{p_{\xi_1}(x_1)} \tag{12.23}$$

Die bedingten Verteilungen mit WMFen $p_{\xi_1|\xi_2=x_2}$ und $p_{\xi_2|\xi_1=x_1}$ heißen dann die *diskreten bedingten Verteilungen von* ξ_1 *gegeben* $\xi_2 = x_2$ *und* ξ_2 *gegeben* $\xi_1 = x_1$, respektive.

•

In Analogie zur Definition der bedingten Wahrscheinlichkeit von Ereignissen gilt also

$$p_{\xi_1|\xi_2}(x_1|x_2) = \frac{p_{\xi_1,\xi_2}(x_1,x_2)}{p_{\xi_2}(x_2)} = \frac{\mathbb{P}(\{\xi_1 = x_1\} \cap \{\xi_2 = x_2\})}{\mathbb{P}(\{\xi_2 = x_2\})}. \tag{12.24}$$

Es ist dabei entscheidend zu erkennen, dass bedingte Verteilungen lediglich normalisierte gemeinsame Verteilungen sind.

Beispiel

In Fortführung des in Kapitel 12.1 betrachteten Beispiels eines zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ ergeben und seiner in Kapitel 12.2 bestimmten Marginalverteilungen ergeben sich folgende bedingte WMFen für $p_{\xi_2|\xi_1=x_1}$:

$p_{\xi_2\|\xi_1}(x_2\|x_1)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$p_{\xi_2\|\xi_1=1}(x_2\|x_1 = 1)$	$\frac{0.1}{0.4} = 0.25$	$\frac{0.0}{0.4} = 0.00$	$\frac{0.2}{0.4} = 0.50$	$\frac{0.1}{0.4} = 0.25$
$p_{\xi_2\|\xi_1=2}(x_2\|x_1 = 2)$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.2}{0.3} = 0.6\bar{6}$	$\frac{0.0}{0.3} = 0.00$	$\frac{0.0}{0.3} = 0.00$
$p_{\xi_2\|\xi_1=3}(x_2\|x_1 = 3)$	$\frac{0.0}{0.3} = 0.00$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.1}{0.3} = 0.3\bar{3}$

Man beachte, dass zum einen gilt, dass

$$\sum_{x_2=1}^{4} p_{\xi_2|\xi_1=x_1}(x_2|x_1) = 1 \text{ für alle } x_1 \in \mathcal{X}_1, \tag{12.25}$$

die bedingten WMFen sind also normiert. Zum anderen beachte man die qualitative Ähnlichkeit der WMFen $p_{\xi_1,\xi_2}(x_1,x_2)$ und $p_{\xi_2|\xi_1}(x_2|x_1)$, die sich einfach daraus ergibt, dass $p_{\xi_1,\xi_2}(x_1,x_2)$ und $p_{\xi_2|\xi_1}(x_2|x_1)$ für alle $x_1 \in \mathcal{X}_1$ bis auf den gemeinsamen Skalierungsfaktor $1/p_{\xi_1}(x_1)$ identisch sind.

Im Fall eines kontinuierlichen Zufallsvektors sind die analogen bedingten WDFen definiert wie folgt.

Definition 12.8 (Bedingte Wahrscheinlichkeitsdichtefunktion und kontinuierliche bedingte Verteilung)**.** $\xi := (\xi_1, \xi_2)$ sei ein kontinuierlicher Zufallsvektor mit Ergebnisraum $\mathbb{R}^2$, WDF $p_\xi = p_{\xi_1,\xi_2}$ und marginalen WDFen p_{ξ_1} und p_{ξ_2}. Die bedingte WDF von ξ_1 gegeben $\xi_2 = x_2$ ist dann für $p_{\xi_2}(x_2) > 0$ definiert als

$$p_{\xi_1|\xi_2=x_2} : \mathbb{R} \to \mathbb{R}_{\geq 0}, x_1 \mapsto p_{\xi_1|\xi_2=x_2}(x_1|x_2) := \frac{p_{\xi_1,\xi_2}(x_1,x_2)}{p_{\xi_2}(x_2)} \tag{12.26}$$

Analog ist für $p_{\xi_1}(x_1) > 0$ die bedingte WMF von ξ_2 gegeben $\xi_1 = x_1$ definiert als

$$p_{\xi_2|\xi_1=x_1} : \mathbb{R} \to \mathbb{R}_{\geq 0}, x_2 \mapsto p_{\xi_2|\xi_1=x_1}(x_2|x_1) := \frac{p_{\xi_1,\xi_2}(x_1, x_2)}{p_{\xi_1}(x_1)} \tag{12.27}$$

Die Verteilungen mit WDFen $p_{\xi_1|\xi_2=x_2}$ und $p_{\xi_2|\xi_1=x_1}$ 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 $p_\xi(x) = 0$ gilt. Die bedingten Verteilungen multivariater Normalverteilungen diskutieren wir in **?@sec-normalverteilungen**.

12.4. 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 12.9 (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

$$\mathbb{P}_\xi(\xi_1 \in S_1, \xi_2 \in S_2) = \mathbb{P}_{\xi_1}(\xi_1 \in S_1)\mathbb{P}_{\xi_2}(\xi_2 \in S_2). \tag{12.28}$$

•

Definition 12.9 besagt, dass die Ereignisse $\{\xi_1 \in S_1\}$ und $\{\xi_2 \in S_2\}$ unabhängig sind. Es gilt also auch, dass

$$\mathbb{P}(\{\xi_1 \in S_1\})|\{\xi_2 \in S_2\}) = \mathbb{P}(\{\xi_1 \in S_1\}) \tag{12.29}$$

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 12.2 (Unabhängigkeit und Faktorisierung)**.**

(1) $\xi := (\xi_1, \xi_2)$ *sei ein diskreter Zufallsvektor mit Ergebnisraum* $\mathcal{X}_1 \times \mathcal{X}_2$*, WMF* p_ξ *und marginalen WMFen* p_{ξ_1}, p_{ξ_2}*. Dann gilt*

$$\xi_1 \text{ und } \xi_2 \text{ sind unabhängige Zufallsvariablen} \Leftrightarrow$$
$$p_\xi(x_1, x_2) = p_{\xi_1}(x_1)p_{\xi_2}(x_2) \text{ für alle } (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2. \tag{12.30}$$

(2) $\xi := (\xi_1, \xi_2)$ *sei ein kontinuierlicher Zufallsvektor mit Ergebnisraum* $\mathbb{R}^2$*, WDF* p_ξ *und marginalen WDFen* p_{ξ_1}, p_{ξ_2}*. Dann gilt*

$$\xi_1 \text{ und } \xi_2 \text{ sind unabhängige Zufallsvariablen } \Leftrightarrow$$
$$p_\xi(x_1, x_2) = p_{\xi_1}(x_1)p_{\xi_2}(x_2) \text{ für alle } (x_1, x_2) \in \mathbb{R}^2. \tag{12.31}$$

∘

Generell ist die Unabhängigkeit zweier Zufallsvariablen also äquivalent zur Faktorisierung ihrer gemeinsamen WMF oder WDF. Für einen Beweis von Theorem 12.2 verweisen wir auf die weiterführende Literatur. Nichtsdestotrotz ist Theorem 12.2 für weite Aspekte der probabilistischen Modellierung grundlegend.

Beispiel

Wir betrachten erneut den zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ aus Kapitel 12.1, dessen gemeinsame und marginale WMFen bekanntlich die untenstehende Form haben

$p_\xi(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$p_{\xi_1}(x_1)$
$x_1 = 1$	0.10	0.00	0.20	0.10	0.40
$x_1 = 2$	0.10	0.20	0.00	0.00	0.30
$x_1 = 3$	0.00	0.10	0.10	0.10	0.30
$p_{\xi_2}(x_2)$	0.20	0.30	0.30	0.20	

Wir fragen zunächst, ob ξ_1 und ξ_2 wohl unabhängig sind. Dies ist nicht der Fall, da hier gilt, dass

$$p_\xi(1,1) = 0.10 \neq 0.08 = 0.40 \cdot 0.20 = p_{\xi_1}(1) p_{\xi_2}(1). \tag{12.32}$$

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 $p_\xi(\xi_1, \xi_2)$ aus dem jeweiligen Produkt der Marginalwahrscheinlichkeiten ergeben. Die gemeinsame Verteilung von ξ_1 und ξ_2 unter der Annahme der Unabhängigkeit von ξ_1 und ξ_2 bei gleichen Marginalverteilungen wie im obigen Fall ergibt sich also zu

$p_\xi(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$p_{\xi_1}(x_1)$
$x_1 = 1$	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
$p_{\xi_2}(x_2)$	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 $p_{\xi_2|\xi_1}$ zu wie folgt:

$p_{\xi_1\|\xi_2}(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$p_{\xi_2\|\xi_1=1}(x_2 \| x_1 = 1)$	$\frac{0.08}{0.40} = 0.2$	$\frac{0.12}{0.40} = 0.3$	$\frac{0.12}{0.40} = 0.3$	$\frac{0.08}{0.40} = 0.2$
$p_{\xi_2\|\xi_1=2}(x_2 \| x_1 = 2)$	$\frac{0.06}{0.30} = 0.2$	$\frac{0.09}{0.30} = 0.3$	$\frac{0.09}{0.30} = 0.3$	$\frac{0.06}{0.30} = 0.2$
$p_{\xi_2\|\xi_1=3}(x_2 \| x_1 = 3)$	$\frac{0.06}{0.30} = 0.2$	$\frac{0.09}{0.30} = 0.3$	$\frac{0.09}{0.30} = 0.3$	$\frac{0.06}{0.30} = 0.2$

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.

Wir wollen den Begriff der unabhängigen Zufallsvariablen nun für mehr als zwei Zufallsvariablen definieren.

Definition 12.10 (n unabhängige Zufallsvariablen)**.** $\xi := (\xi_1, ..., \xi_n)$ sei ein n-dimensionaler Zufallsvektor mit Ergebnisraum $\mathcal{X} = \times_{i=1}^n \mathcal{X}_i$. Die n Zufallsvariablen $\xi_1, ..., \xi_n$ heißen *unabhängig*, wenn für alle $S_i \in \mathcal{X}_i, i = 1, ..., n$ gilt, dass

$$\mathbb{P}_\xi(\xi_1 \in S_1, ..., \xi_n \in S_n) = \prod_{i=1}^n \mathbb{P}_{\xi_i}(\xi_i \in S_i). \tag{12.33}$$

Wenn der Zufallsvektor eine n-dimensionale WMF oder WDF p_ξ mit marginalen WMFen oder WDFen $p_{\xi_i}, i = 1, ..., n$ besitzt, dann ist die Unabhängigkeit von $\xi_1, ..., \xi_n$ gleichbedeutend mit der Faktorisierung der gemeinsamen WMF oder WDF, also mit

$$p_\xi(\xi_1, ..., \xi_n) = \prod_{i=1}^n p_{\xi_i}(x_i). \tag{12.34}$$

•

Es handelt bei Definition 12.10 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 12.11 (Unabhängig und identisch verteilte Zufallsvariablen)**.** n Zufallsvariablen $\xi_1, ..., \xi_n$ heißen *unabhängig und identisch verteilt (u.i.v.)*, wenn

(1) $\xi_1, ..., \xi_n$ unabhängige Zufallsvariablen sind, und
(2) die Marginalverteilungen der ξ_i übereinstimmen, also gilt, dass

$$\mathbb{P}_{\xi_i} = \mathbb{P}_{\xi_j} \text{ für alle } 1 \leq i, j \leq n. \tag{12.35}$$

Wenn die Zufallsvariablen $\xi_1, ..., \xi_n$ unabhängig und identisch verteilt sind und die $i$te Marginalverteilung $\mathbb{P}_\xi := \mathbb{P}_{\xi_i}$ ist, so schreibt man auch

$$\xi_1, ..., \xi_n \sim \mathbb{P}_\xi. \tag{12.36}$$

•

Man sagt kurz, dass "$\xi_1, ..., \xi_n$ 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 Zufallsvektoren werden als

$$\xi_1, ..., \xi_n \sim N(\mu, \sigma^2) \tag{12.37}$$

geschrieben werden. In **?@sec-normalverteilungen** zeigen wir, wie genau die gemeinsame Verteilung von n u.i.v. normalverteilte Zufallsvektoren beschaffen ist.

12.5. Selbstkontrollfragen

1. Geben Sie die Definition des Begriffs des Zufallsvektors wieder.
2. Geben Sie die Definition des Begriffs der multivariaten Verteilung eines Zufallsvektors wieder.
3. Geben Sie die Definition des Begriffs der multivariaten WMF wieder.
4. Geben Sie die Definition des Begriffs der multivariaten WDF wieder.
5. Geben Sie die Definition des Begriffs der univariaten Marginalverteilung eines Zufallsvektors wieder
6. Wie berechnet man die WMF der $i$ten Komponente eines diskreten Zufallsvektors?
7. Wie berechnet man die WDF der $i$ten Komponente eines kontinuierlichen Zufallsvektors?
8. Geben Sie die Definition des Begriffs der Unabhängigkeit zweier Zufallsvariablen wieder.
9. Wie erkennt man an der gemeinsamen WMF oder WDF eines zweidimensionalen Zufallsvektors, ob die Komponenten des Zufallsvektors unabhängig sind oder nicht?
10. Geben Sie die Definition des Begriffs der Unabhängigkeit von n Zufallsvariablen wieder.
11. Geben Sie die Definition des Begriffs der n unabhängig und identisch verteilten Zufallsvariablen wieder.

13. Erwartungswerte

In diesem Kapitel führen wir mit den Begriffen des *Erwartungswerts* und der *Varianz* einer Zufallsvariable skalare Zusammenfassungen von Verteilungen ein, die häufig als charakteristische Kennzahlen von Wahrscheinlichkeitsverteilungen dienen. Dabei ist der Erwartungswert als ein Maß der "durchschnittlichen Realisierung" und die Varianz als Maß der "durchschnittlichen Variabilität" einer Wahrscheinlichkeitsverteilung zu verstehen. Weiterhin führen wir mit der *Kovarianz* zweier Zufallsvariablen ein Maß für linear-affine Abhängigkeiten zwischen Zufallsvariablen ein. Wir ergänzen diese Begriffe um ihre Analoga in Bezug auf Zufallsvektoren (*Erwartungswert* und *Kovarianzmatrix*) und ihre deskriptiv-statistischen Äquivalente, das *Stichprobenmittel*, die *Stichprobenvarianz*, und die *Stichprobenkovarianz*.

13.1. Erwartungswert

Definition 13.1 (Erwartungswert)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und ξ sei eine Zufallsvariable. Dann ist der *Erwartungswert von* ξ definiert als

- $\mathbb{E}(\xi) := \sum_{x \in \mathcal{X}} x\, p_\xi(x)$, wenn $\xi : \Omega \to \mathcal{X}$ diskret mit WMF p_ξ ist,
- $\mathbb{E}(\xi) := \int_{-\infty}^{\infty} x\, p_\xi(x)\, dx$, wenn $\xi : \Omega \to \mathbb{R}$ kontinuierlich mit WDF p_ξ ist.

Man sagt, dass der Erwartungswert einer Zufallsvariable *existiert*, wenn er endlich ist.

•

Der Erwartungswert ist also eine skalare Zusammenfassung der Verteilung einer Zufallsvariable. Eine integrierte 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 dahingehend 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, also etwa

$$\mathbb{E}(\xi) \approx \frac{1}{n} \sum_{i=1}^{n} \xi_i \tag{13.1}$$

für eine große Zahl n von Kopien ξ_i von ξ. Wir werden diese Intuition im Kontext der Gesetze der großen Zahl in Kapitel 15 weiter ausarbeiten.

Beispiele

Mit dem Erwartungswert einer Bernoulli-Zufallsvariable und dem Erwartungswert einer normalverteilten Zufallsvariable wollen wir nun zwei erste Beispiele für den Erwartungswert einer diskreten und einer kontinuierlichen Zufallsvariable betrachten.

Theorem 13.1 (Erwartungswert einer Bernoulli Zufallsvariable)**.** *Es sei* $\xi \sim Bern(\mu)$*. Dann gilt* $\mathbb{E}(\xi) = \mu$*.*

∘

Beweis. ξ ist diskret mit $\mathcal{X} = \{0, 1\}$. Also gilt

$$\begin{aligned}\mathbb{E}(\xi) &= \sum_{x \in \{0,1\}} x \operatorname{Bern}(x; \mu) \\ &= 0 \cdot \mu^0 (1-\mu)^{1-0} + 1 \cdot \mu^1 (1-\mu)^{1-1} \\ &= 1 \cdot \mu^1 (1-\mu)^0 \\ &= \mu.\end{aligned} \tag{13.2}$$

□

Es ergibt sich hier also, dass der Parameter $\mu \in [0, 1]$ der Verteilung einer Bernoulli-Zufallsvariable gleichzeitig auch ihr Erwartungswert ist.

Theorem 13.2 (Erwartungswert einer normalverteilten Zufallsvariable)**.** *Es sei* $\xi \sim N(\mu, \sigma^2)$*. Dann gilt* $\mathbb{E}(\xi) = \mu$*.*

∘

Beweis. Die Herleitung des Erwartungswerts einer normalverteilten Zufallsvariable ist überraschend aufwändig. Wir müssen in diesem Fall einige grundlegende Eigenschaften der Exponentialfunktion als gegeben annehmen. Dazu halten wir zunächst ohne Beweis fest, dass

$$\int_{-\infty}^{\infty} \exp\left(-x^2\right)\, dx = \sqrt{\pi} \tag{13.3}$$

und dass

$$\lim_{x \to -\infty} \exp\left(-x^2\right) = 0 \text{ und } \lim_{x \to \infty} \exp\left(-x^2\right) = 0. \tag{13.4}$$

Gleichung 13.3 ist unter der Bezeichnung *Gauss-* oder *Euler-Poisson-Integral* bekannt. Mit der Definition des Erwartungswerts für kontinuierliche Zufallsvariablen gilt dann zunächst

$$\mathbb{E}(\xi) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\, dx. \tag{13.5}$$

Mit der allgemeinen Substitutionsregel (vgl. Theorem 7.1)

$$\int_{g(a)}^{g(b)} f(x)\, dx = \int_a^b f(g(x)) g'(x)\, dx \tag{13.6}$$

und der Definition von

$$g : \mathbb{R} \to \mathbb{R}, x \mapsto g(x) := \sqrt{2\sigma^2} x + \mu \text{ with } g'(x) = \sqrt{2\sigma^2}, \tag{13.7}$$

gilt dann

$$\begin{aligned}
\mathbb{E}(\xi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\, dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x + \mu) \exp\left(-\frac{1}{2\sigma^2}\left(\left(\sqrt{2\sigma^2}x+\mu\right)-\mu\right)^2\right) \sqrt{2\sigma^2}\, dx \\
&= \frac{\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x + \mu) \exp\left(-x^2\right)\, dx \\
&= \frac{1}{\sqrt{\pi}} \left(\sqrt{2\sigma^2} \int_{-\infty}^{\infty} x \exp\left(-x^2\right)\, dx + \mu \int_{-\infty}^{\infty} \exp\left(-x^2\right)\, dx\right) \\
&= \frac{1}{\sqrt{\pi}} \left(\sqrt{2\sigma^2} \int_{-\infty}^{\infty} x \exp\left(-x^2\right)\, dx + \mu\sqrt{\pi}\right).
\end{aligned} \tag{13.8}$$

Eine Stammfunktion von $x \exp\left(-x^2\right)$ ist $-\frac{1}{2}\exp\left(-x^2\right)$, weil

$$\frac{d}{dx}\left(-\frac{1}{2}\exp\left(-x^2\right)\right) = -\frac{1}{2}\frac{d}{dx}\exp\left(-x^2\right) = -\frac{1}{2}\exp\left(-x^2\right)(-2x) = x\exp\left(-x^2\right) \tag{13.9}$$

Mit Gleichung 13.4 und der Definition des uneigentlichen Integrals (vgl. Definition 7.5) verschwindet der Integralterm $\int_{-\infty}^{\infty} x \exp\left(-x^2\right)\, dx$ damit und wir erhalten

$$\mathbb{E}(\xi) = \frac{1}{\sqrt{\pi}}\left(\mu\sqrt{\pi}\right) = \mu. \tag{13.10}$$

□

Der Erwartungswert einer univariaten Normalverteilung ist also durch ihren Parameter $\mu \in \mathbb{R}$ gegeben.

In Verallgemeinerung von Definition 13.1 geben wir folgende Definition für den Erwartungswert einer Funktion einer Zufallsvariable

Definition 13.2 (Erwartungswert einer Funktion einer Zufallsvariable)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, ξ sei eine Zufallsvariable mit Ergebnisraum $\mathcal{X}$ und $f : \mathcal{X} \to \mathcal{Z}$ sei eine Funktion mit Zielmenge $\mathcal{Z}$. Dann ist der *Erwartungswert der Funktion f der Zufallsvariable ξ* definiert als

- $\mathbb{E}(f(\xi)) := \sum_{x\in\mathcal{X}} f(x)\, p_\xi(x)$, wenn $\xi : \Omega \to \mathcal{X}$ diskret mit WMF p_ξ ist,
- $\mathbb{E}(f(\xi)) := \int_{-\infty}^{\infty} f(x)\, p_\xi(x)\, dx$, wenn $\xi : \Omega \to \mathbb{R}$ kontinuierlich mit WDF p_ξ ist.

•

Der Erwartungswert einer Zufallsvariable ergibt sich anhand von Definition 13.2 als der Spezialfall, in dem gilt, dass

$$f : \mathcal{X} \to \mathcal{Z}, x \mapsto f(x) := x. \tag{13.11}$$

In der englischsprachigen Literatur ist Definition 13.2 auch als “Law of the unconscious statistician” bekannt und wird oft auch direkt zur Definition des Erwartungswertes herangezogen.

Weiterhin ist man wie im univariaten Fall manchmal darum bemüht, die Verteilung eines Zufallsvektors mit einigen wenigen Maßzahlen zu charakterisieren. Das multivariate Analogon des des Erwartungswerts einer Zufallsvariablen ist der *Erwartungswert eines Zufallsvektors*, der wie folgt definiert ist.

Definition 13.3 (Erwartungswert eines Zufallsvektors)**.** ξ sei ein n-dimensionaler Zufallvektor. Dann ist der *Erwartungwert* von ξ definiert als der n-dimensionale reelle Vektor

$$\mathbb{E}(\xi) := \begin{pmatrix} \mathbb{E}(\xi_1) \\ \vdots \\ \mathbb{E}(\xi_n) \end{pmatrix} \tag{13.12}$$

•

Der Erwartungswert eines Zufallsvektors ξ ist also der Vektor der Erwartungswerte der Komponenten $\xi_1, ..., \xi_n$ von ξ, ist also direkt im Sinne von Erwartungswerten von Zufallsvariablen definiert. In Analogie zu Definition 13.2 definiert man für die Funktion eines Zufallsvektors den Erwartungswert dieser Transformation wie folgt.

Definition 13.4 (Erwartungswert einer Funktion eines Zufallsvektors)**.** $(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, ξ sei ein Zufallsvektor mit Ergebnisraum $\mathcal{X}$ und $f : \mathcal{X} \to \mathcal{Z}$ sei eine Funktion mit Zielmenge $\mathcal{Z}$. Dann ist der *Erwartungswert der Funktion f des Zufallsvektors ξ* definiert als

- $\mathbb{E}(f(\xi)) := \sum_{x \in \mathcal{X}} f(x)\, p_\xi(x)$, wenn $\xi : \Omega \to \mathcal{X}$ diskret mit WMF p_ξ ist,
- $\mathbb{E}(f(\xi)) := \int_{-\infty}^{\infty} f(x)\, p_\xi(x)\, dx$, wenn $\xi : \Omega \to \mathbb{R}$ kontinuierlich mit WDF p_ξ ist.

•

Folgendes Theorem gibt nun einige Rechenregeln im Umgang mit Erwartungswerten an, die uns an vielen Stellen begegnen werden. Diese Rechenregeln folgen direkt aus der Summen- bzw. Integraldefinition des Erwartungswertes, im Beweis des Theorems betrachten wir dementsprechend lediglich den Fall kontinuierlicher Zufallsvariablen.

Theorem 13.3 (Eigenschaften des Erwartungswerts)**.**

(1) (Linear-affine Transformation) Für eine Zufallsvariable ξ und $a, b \in \mathbb{R}$ gilt

$$\mathbb{E}(a\xi + b) = a\mathbb{E}(\xi) + b. \tag{13.13}$$

(2) (Linearkombination) Für Zufallsvariablen $\xi_1, ..., \xi_n$ und $a_1, ..., a_n \in \mathbb{R}$ gilt

$$\mathbb{E}\left(\sum_{i=1}^{n} a_i \xi_i\right) = \sum_{i=1}^{n} a_i \mathbb{E}(\xi_i). \tag{13.14}$$

(3) (Faktorisierung bei Unabhängigkeit) Für unabhängige Zufallsvariablen $\xi_1, ..., \xi_n$ gilt

$$\mathbb{E}\left(\prod_{i=1}^{n} \xi_i\right) = \prod_{i=1}^{n} \mathbb{E}(\xi_i). \tag{13.15}$$

∘

Beweis. Eigenschaft (1) folgt aus den Linearitätseigenschaften von Summen und Integralen. Wir betrachten nur den Fall einer kontinuierlichen Zufallsvariable ξ mit WDF p_ξ genauer und definieren zunächst $\upsilon := a\xi + b$. Dann gilt

$$\begin{aligned}
\mathbb{E}(\upsilon) &= \mathbb{E}(a\xi + b) \\
&= \int_{-\infty}^{\infty} (ax + b) p_\xi(x)\, dx \\
&= \int_{-\infty}^{\infty} axp_\xi(x) + bp_\xi(x)\, dx \\
&= a\int_{-\infty}^{\infty} xp_\xi(x)\, dx + b\int_{-\infty}^{\infty} p_\xi(x)\, dx \\
&= a\mathbb{E}(\xi) + b.
\end{aligned} \tag{13.16}$$

Eigenschaft (2) folgt gleichfalls aus den Linearitätseigenschaften von Summen und Integralen. Wir wollen nur den Fall von zwei kontinuierlichen Zufallsvariablen ξ_1 und ξ_2 mit bivariater WDF p_{ξ_1,ξ_2} genauer betrachten. In diesem Fall gilt

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{2} a_i\xi_i\right) &= \mathbb{E}(a_1\xi_1 + a_2\xi_2) \\
&= \iint_{\mathbb{R}^2} (a_1x_1 + a_2x_2) p_{\xi_1,\xi_2}(x_1, x_2)\, dx_1\, dx_2 \\
&= \iint_{\mathbb{R}^2} a_1x_1p_{\xi_1,\xi_2}(x_1, x_2) + a_2x_2p_{\xi_1,\xi_2}(x_1, x_2)\, dx_1\, dx_2 \\
&= a_1 \iint_{\mathbb{R}^2} x_1p_{\xi_1,\xi_2}(x_1, x_2)\, dx_1\, dx_2 + a_2 \iint_{\mathbb{R}^2} x_2p_{\xi_1,\xi_2}(x_1, x_2)\, dx_1\, dx_2 \\
&= a_1 \int_{-\infty}^{\infty} x_1 \left(\int_{-\infty}^{\infty} p_{\xi_1,\xi_2}(x_1, x_2)\, dx_2\right) dx_1 + a_2 \int_{-\infty}^{\infty} x_2 \left(\int_{-\infty}^{\infty} p_{\xi_1,\xi_2}(x_1, x_2)\, dx_1\right) dx_2 \\
&= a_1 \int_{-\infty}^{\infty} x_1p_{\xi_1}(x_1)\, dx_1 + a_2 \int_{-\infty}^{\infty} x_2p_{\xi_2}(x_2)\, dx_2 \\
&= a_1\mathbb{E}(\xi_1) + a_2\mathbb{E}(\xi_2) \\
&= \sum_{i=1}^{2} a_i\mathbb{E}(\xi_i).
\end{aligned} \tag{13.17}$$

Ein Induktionsbeweis erlaubt dann die Generalisierung vom bivariaten auf den n-variaten Fall.

Zu Eigenschaft (3) betrachten wir den Fall von n kontinuierlichen Zufallsvariablen mit gemeinsamer WDF $p_{\xi_1,\ldots,\xi_n}$. Weil als $\xi_1, \ldots, \xi_n$ unabhängig vorausgesetzt sind, gilt

$$p_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{\xi_i}(x_i). \tag{13.18}$$

Weiterhin gilt also

$$\begin{aligned}
\mathbb{E}\left(\prod_{i=1}^{n} \xi_i\right) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} x_i\right) p_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{n} x_i \prod_{i=1}^{n} p_{\xi_i}(x_i)\, dx_1 \ldots dx_n \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{n} x_ip_{\xi_i}(x_i)\, dx_1 \ldots dx_n \\
&= \prod_{i=1}^{n} \int_{-\infty}^{\infty} x_ip_{\xi_i}(x_i)\, dx_i \\
&= \prod_{i=1}^{n} \mathbb{E}(\xi_i).
\end{aligned} \tag{13.19}$$

□

13.2. Varianz und Standardabweichung

Häufig genutzte Maße für die Streuung von Verteilungen von Zufallsvariablen sind die Varianz und die Standardabweichung. Diese sind wie folgt definiert.

Definition 13.5 (Varianz und Standardabweichung)**.** ξ sei eine Zufallsvariable mit existierendem Erwartungswert $\mathbb{E}(\xi)$.

- Die *Varianz von* ξ ist definiert als

$$\mathbb{V}(\xi) := \mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right). \tag{13.20}$$

- Die *Standardabweichung von* ξ ist definiert als

$$\mathbb{S}(\xi) := \sqrt{\mathbb{V}(\xi)}. \tag{13.21}$$

•

Inwiefern die Varianz und ihre Quadratwurzel als Maße für die Streuung einer Zufallsvariable dienen, werden wir in Theorem 14.2 begründen. Die Quadrierung der Abweichung der Zufallsvariable von ihrem Erwartungswert in der Definition der Varianz ist nötig, da andernfalls mit Theorem 13.3 immer gelten würde, dass

$$\mathbb{E}(\xi - \mathbb{E}(\xi)) = \mathbb{E}(\xi) - \mathbb{E}(\xi) = 0. \tag{13.22}$$

Allerdings gibt es neben der Varianz durchaus weitere Maße der Streuung von Zufallsvariablen, hier seien beispielsweise die erwartete absolute Abweichung einer Zufallsvariable von ihrem Erwartungswert, $\mathbb{E}(|\xi - \mathbb{E}(\xi)|)$ und die sogenannte *Entropie* $-\mathbb{E}(\ln p_\xi)$ genannt. Im Sinne von Definition 13.2 ist die Varianz der Zufallsvariable $\xi : \Omega \to \mathcal{X}$ der Erwartungswert der Funktion

$$f : \mathcal{X} \to \mathcal{Z}, x \mapsto f(x) := (x - \mathbb{E}(\xi))^2. \tag{13.23}$$

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 13.4 (Varianzverschiebungssatz)**.** *ξ sei eine Zufallsvariable. Dann gilt*

$$\mathbb{V}(\xi) = \mathbb{E}\left(\xi^2\right) - \mathbb{E}(\xi)^2. \tag{13.24}$$

∘

Beweis. Mit der Definition der Varianz und der Linearität des Erwartungswerts gilt

$$\begin{aligned}
\mathbb{V}(\xi) &= \mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right) \\
&= \mathbb{E}\left(\xi^2 - 2\xi\mathbb{E}(\xi) + \mathbb{E}(\xi)^2\right) \\
&= \mathbb{E}(\xi^2) - 2\mathbb{E}(\xi)\mathbb{E}(\xi) + \mathbb{E}\left(\mathbb{E}(\xi)^2\right) \\
&= \mathbb{E}(\xi^2) - 2\mathbb{E}(\xi)^2 + \mathbb{E}(\xi)^2 \\
&= \mathbb{E}(\xi^2) - \mathbb{E}(\xi)^2.
\end{aligned} \tag{13.25}$$

□

Wie für den Erwartungswert gibt es auch für die Varianz einige Rechenregeln, die den Umgang mit ihr oft erleichtern. Wir fassen sie in folgendem Theorem zusammen.

Theorem 13.5 (Eigenschaften der Varianz)**.**

(1) (Linear-affine Transformation) Für eine Zufallsvariable ξ und $a, b \in \mathbb{R}$ gelten

$$\mathbb{V}(a\xi + b) = a^2\mathbb{V}(\xi) \text{ und } \mathbb{S}(a\xi + b) = |a|\mathbb{S}(\xi). \tag{13.26}$$

(2) (Linearkombination bei Unabhängigkeit) Für unabhängige Zufallsvariablen $\xi_1, ..., \xi_n$ und $a_1, ..., a_n \in \mathbb{R}$ gilt

$$\mathbb{V}\left(\sum_{i=1}^{n} a_i\xi_i\right) = \sum_{i=1}^{n} a_i^2\mathbb{V}(\xi_i). \tag{13.27}$$

∘

Beweis. Um Eigenschaft (1) zu zeigen, definieren wir zunächst $\upsilon := a\xi + b$ und halten fest, dass $\mathbb{E}(\upsilon) = a\mathbb{E}(\xi) + b$. Für die Varianz von υ ergibt sich dann

$$\begin{aligned}
\mathbb{V}(\upsilon) &= \mathbb{E}\left((\upsilon - \mathbb{E}(\upsilon))^2\right) \\
&= \mathbb{E}\left((a\xi + b - a\mathbb{E}(\xi) - b)^2\right) \\
&= \mathbb{E}\left((a\xi - a\mathbb{E}(\xi))^2\right) \\
&= \mathbb{E}\left((a(\xi - \mathbb{E}(\xi))^2\right) \\
&= \mathbb{E}\left(a^2(\xi - \mathbb{E}(\xi))^2\right) \\
&= a^2\mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right) \\
&= a^2\mathbb{V}(\xi)
\end{aligned} \tag{13.28}$$

Wurzelziehen ergibt dann das Resultat für die Standardabweichung.

Für Eigenschaft (2) betrachten wir den Fall zweier unabhängiger Zufallsvariablen ξ_1 und ξ_2 genauer. Wir halten zunächst fest, dass in diesem Fall gilt, dass

$$\mathbb{E}\left(a_1\xi_1 + a_2\xi_2\right) = a_1\mathbb{E}(\xi_1) + a_2\mathbb{E}(\xi_2). \tag{13.29}$$

Es ergibt sich also

$$\begin{aligned}
\mathbb{V}\left(\sum_{i=1}^{2} a_i\xi_i\right) &= \mathbb{V}(a_1\xi_1 + a_2\xi_2) \\
&= \mathbb{E}\left((a_1\xi_1 + a_2\xi_2 - \mathbb{E}\left(a_1\xi_1 + a_2\xi_2\right))^2\right) \\
&= \mathbb{E}\left((a_1\xi_1 + a_2\xi_2 - a_1\mathbb{E}(\xi_1) - a_2\mathbb{E}(\xi_2))^2\right) \\
&= \mathbb{E}\left((a_1\xi_1 - a_1\mathbb{E}(\xi_1) + a_2\xi_2 - a_2\mathbb{E}(\xi_2))^2\right) \\
&= \mathbb{E}\left(((a_1(\xi_1 - \mathbb{E}(\xi_1)) + (a_2(\xi_2 - \mathbb{E}(\xi_2)))^2\right) \\
&= \mathbb{E}\left((a_1(\xi_1 - \mathbb{E}(\xi_1)))^2 + 2(a_1(\xi_1 - \mathbb{E}(\xi_1))(a_2(\xi_2 - \mathbb{E}(\xi_2)) + (a_2(\xi_2 - \mathbb{E}(\xi_2)))^2\right) \\
&= \mathbb{E}\left((a_1^2(\xi_1 - \mathbb{E}(\xi_1))^2 + 2a_1a_2(\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2)) + a_2^2(\xi_2 - \mathbb{E}(\xi_2))^2\right) \\
&= a_1^2\mathbb{E}\left((\xi_1 - \mathbb{E}(\xi_1))^2\right) + 2a_1a_2\mathbb{E}\left((\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2))\right) + a_2^2\mathbb{E}\left((\xi_2 - \mathbb{E}(\xi_2))^2\right) \\
&= a_1^2\mathbb{V}(\xi_1) + 2a_1a_2\mathbb{E}\left((\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2))\right) + a_2^2\mathbb{V}(\xi_2) \\
&= \sum_{i=1}^{2} a_i^2\mathbb{V}(\xi_i) + 2a_1a_2\mathbb{E}\left((\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2))\right).
\end{aligned} \tag{13.30}$$

Weil ξ_1 und ξ_2 unabhängig sind, ergibt sich mit den Eigenschaften des Erwartungswerts für unabhängige Zufallsvariablen, dass

$$\begin{aligned}
\mathbb{E}\left((\xi_1 - \mathbb{E}(\xi_1))(\xi_2 - \mathbb{E}(\xi_2))\right) &= \mathbb{E}\left((\xi_1 - \mathbb{E}(\xi_1))\right)\mathbb{E}\left((\xi_2 - \mathbb{E}(\xi_2))\right) \\
&= (\mathbb{E}(\xi_1) - \mathbb{E}(\xi_1))(\mathbb{E}(\xi_2) - \mathbb{E}(\xi_2)) \\
&= 0
\end{aligned} \tag{13.31}$$

ist. Damit folgt also

$$\mathbb{V}\left(\sum_{i=1}^{2} a_i \xi_i\right) = \sum_{i=1}^{2} a_i^2 \mathbb{V}(\xi_i). \tag{13.32}$$

Ein Induktionsbeweis erlaubt dann die Generalisierung vom bivariaten zum n-variaten Fall.

□

Beispiele

Mit der Varianz einer Bernoulli-Zufallsvariable und der Varianz einer normalverteilten Zufallsvariable wollen wir auch hier zwei erste Beispiele für die Varianz einer diskreten und einer kontinuierlichen Zufallsvariable betrachten.

Theorem 13.6 (Varianz einer Bernoulli Zufallsvariable)**.** *Es sei $\xi \sim Bern(\mu)$. Dann ist die Varianz von ξ gegeben durch*

$$\mathbb{V}(\xi) = \mu(1-\mu). \tag{13.33}$$

∘

Beweis. ξ ist eine diskrete Zufallsvariable und es gilt $\mathbb{E}(\xi) = \mu$. Also gilt

$$\begin{aligned}
\mathbb{V}(\xi) &= \mathbb{E}\left((\xi-\mu)^2\right) \\
&= \sum_{x\in\{0,1\}} (x-\mu)^2 \text{Bern}(x;\mu) \\
&= (0-\mu)^2\mu^0(1-\mu)^{1-0} + (1-\mu)^2\mu^1(1-\mu)^{1-1} \\
&= \mu^2(1-\mu) + (1-\mu)^2\mu \\
&= \left(\mu^2 + (1-\mu)\mu\right)(1-\mu) \\
&= \left(\mu^2 + \mu - \mu^2\right)(1-\mu) \\
&= \mu(1-\mu).
\end{aligned} \tag{13.34}$$

□

Theorem 13.7 (Varianz einer normalverteilten Zufallsvariable)**.** *Es sei $\xi \sim N(\mu, \sigma^2)$. Dann ist die Varianz von ξ gegeben durch*

$$\mathbb{V}(\xi) = \sigma^2. \tag{13.35}$$

∘

Beweis. Die Herleitung der Varianz einer normalverteilten Zufallsvariable ist nicht unaufwändig, so dass wir hier auch wieder unbewiesen die Gültigkeit von Gleichung 13.3 und Gleichung 13.4 sowie weiterhin von

$$\int_{-\infty}^{\infty} x \exp(-x^2)\, dx = 0 \tag{13.36}$$

annehmen wollen. Wir halten zunächst fest, dass mit dem Varianzverschiebungssatz gilt, dass

$$\mathbb{V}(\xi) = \mathbb{E}(\xi^2) - \mathbb{E}(\xi)^2 = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx - \mu^2. \tag{13.37}$$

Mit der allgemeinen Substitutionsregel (Theorem 7.1)

$$\int_a^b f(g(x))g'(x)\,dx = \int_{g(a)}^{g(b)} f(x)\,dx \tag{13.38}$$

und der Definition von

$$g : \mathbb{R} \to \mathbb{R}, x \mapsto \sqrt{2\sigma^2}x + \mu, g(-\infty) := -\infty, g(\infty) := \infty, \text{ mit } g'(x) = \sqrt{2\sigma^2} \tag{13.39}$$

kann das Integral auf der rechten Seite von Gleichung (13.37) dann als

$$\begin{aligned}\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\,dx &= \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x+\mu)^2 \exp\left(-\frac{1}{2\sigma^2}((\sqrt{2\sigma^2}x+\mu)-\mu)^2\right)\sqrt{2\sigma^2}\,dx \\ &= \sqrt{2\sigma^2}\int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x+\mu)^2 \exp\left(-\frac{2\sigma^2x^2}{2\sigma^2}\right)\,dx \\ &= \sqrt{2\sigma^2}\int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x+\mu)^2 \exp\left(-x^2\right)\,dx\end{aligned} \tag{13.40}$$

geschrieben werden. Also gilt

$$\begin{aligned}\mathbb{V}(\xi) &= \frac{\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x+\mu)^2 \exp\left(-x^2\right)\,dx - \mu^2 \\ &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x)^2 + 2\sqrt{2\sigma^2}x\mu + \mu^2) \exp\left(-x^2\right)\,dx - \mu^2 \\ &= \frac{1}{\sqrt{\pi}}\left(2\sigma^2\int_{-\infty}^{\infty} x^2 \exp\left(-x^2\right)\,dx + 2\sqrt{2\sigma^2}\mu\int_{-\infty}^{\infty} x\exp\left(-x^2\right)\,dx + \mu^2\int_{-\infty}^{\infty}\exp\left(-x^2\right)\,dx\right) - \mu^2.\end{aligned} \tag{13.41}$$

Mit Gleichung 13.36 ergibt sich dann

$$\begin{aligned}\mathbb{V}(\xi) &= \frac{1}{\sqrt{\pi}}\left(2\sigma^2\int_{-\infty}^{\infty} x^2 \exp\left(-x^2\right)\,dx + \mu^2\sqrt{\pi}\right) - \mu^2 \\ &= \frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} x^2 \exp\left(-x^2\right)\,dx + \mu^2 - \mu^2 \\ &= \frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} x^2 \exp\left(-x^2\right)\,dx.\end{aligned} \tag{13.42}$$

Mit der allgemeinen Form der partiellen Integrationsregel (Theorem 7.1)

$$\int_a^b f'(x)g(x)\,dx = f(x)g(x)|_a^b - \int_a^b f(x)g'(x)\,dx \tag{13.43}$$

und der Definition von

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x) := \exp\left(-x^2\right) \text{ mit } f'(x) = -2\exp\left(-x^2\right) \tag{13.44}$$

und

$$g : \mathbb{R} \to \mathbb{R}, x \mapsto g(x) := -\frac{1}{2}x \text{ mit } g'(x) = -\frac{1}{2}, \tag{13.45}$$

so dass

$$f'(x)g(x) = -2\exp\left(-x^2\right)\left(-\frac{1}{2}x\right) = x^2\exp\left(-x^2\right), \tag{13.46}$$

gilt, ergibt sich dann

$$\begin{aligned}\mathbb{V}(\xi) &= \frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} x^2\exp\left(-x^2\right)\,dx \\ &= \frac{2\sigma^2}{\sqrt{\pi}}\left(-\frac{1}{2}x\exp\left(-x^2\right)|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\exp\left(-x^2\right)\left(-\frac{1}{2}\right)\,dx\right) \\ &= \frac{2\sigma^2}{\sqrt{\pi}}\left(-\frac{1}{2}x\exp\left(-x^2\right)|_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty}\exp\left(-x^2\right)\,dx\right).\end{aligned} \tag{13.47}$$

Aus Gleichung 13.4 schließen wir dann, dass der erste Term in den Klammern auf der rechten Seite der obigen Gleichung gleich 0 ist. Schließlich ergibt sich damit

$$\mathbb{V}(\xi) = \frac{2\sigma^2}{\sqrt{\pi}} \left(\frac{1}{2} \int_{-\infty}^{\infty} \exp\left(-x^2\right)\, dx \right) = \frac{\sigma^2}{\sqrt{\pi}} \sqrt{\pi} = \sigma^2. \tag{13.48}$$

□

Allgemein ergeben sich die Erwartungswerte und Varianzen parametrischer Verteilungen als Funktionen ihrer Parameter. Wir fassen die Erwartungswerte uns bekannter Verteilungen in Theorem 13.8 zusammen.

Theorem 13.8 (Erwartungswerte und Varianzen einiger Wahrscheinlichkeitsverteilungen)**.**

Zufallsvariable	*Erwartungswert*	*Varianz*
$\xi \sim B(\mu)$	μ	$\mu(1-\mu)$
$\xi \sim Bin(\mu, n)$	$n\mu$	$n\mu(1-\mu)$
$\xi \sim N(\mu, \sigma^2)$	μ	σ^2
$\xi \sim G(\alpha, \beta)$	$\alpha\beta$	$\alpha\beta^2$
$\xi \sim Beta(\alpha, \beta)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
$\xi \sim U(a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

∘

Wir verzichten auf einen Beweis.

13.3. Kennzahlen univariater Stichproben

Wie wir Kapitel 18 noch ausführlich diskutieren, werden ist eine Charakteristikum der probabilistischen Modellierung, beobachtete Daten als Realisierungen von Zufallsvariablen zu verstehen. Hat meine Menge $\xi_1, ..., \xi_n$ von Zufallsvariablen, so nennt man diese auch eine *Stichprobe*. Basierend auf einer Stichprobe kann man nun Kennzahlen berechnen, die auf den ersten Blick den Begriffen von Erwartungswert, Varianz und Standardabweichung ähneln, mit diesen aber keinesfalls zu verwechseln sind. Defacto dienen die in folgender Definition aufgeführten Stichprobenkennzahlen oft als *Schätzer* für die Kennzahlen von Zufallsvariablen, wie wir in **?@sec-parameterschätzung** ausführlich darlegen wollen. Gewissermaßen Vorgriff zur Abgrenzung der Begrifflichkeiten und auch als Grundlage für Kapitel 15 definieren wir hier einige deskriptive Stichprobenkennzahlen.

Definition 13.6 (Stichprobenmittel, Stichprobenvarianz, Stichprobenstandardabweichung)**.** $\xi_1, ..., \xi_n$ sei eine Menge von Zufallsvariablen, genannt *Stichprobe*.

- Das *Stichprobenmittel* von $\xi_1, ..., \xi_n$ ist definiert als

$$\bar{\xi} := \frac{1}{n} \sum_{i=1}^{n} \xi_i. \tag{13.49}$$

- Die *Stichprobenvarianz* von $\xi_1, ..., \xi_n$ ist definiert als

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2. \tag{13.50}$$

- Die *Stichprobenstandardabweichung* ist definiert als

$$S := \sqrt{S^2}. \tag{13.51}$$

•

Zur Abgrenzung erinnern wir noch einmal daran, dass Erwartungswert $\mathbb{E}(\xi)$, Varianz $\mathbb{V}(\xi)$ und Standardabweichung $\mathbb{S}(\xi)$ Kennzahlen einer Zufallsvariable ξ sind, wohingegen $\bar{\xi}, S^2$, und S Kennzahlen einer Stichprobe $\xi_1, ..., \xi_n$ sind.

Beispiel

Wir wollen die Bestimmung der in Definition 13.6 eingeführten Stichprobenkennzahlen an einem Beispiel erläutern. Dazu halten wir nochmals fest, dass $\bar{\xi}, S^2$, S Zufallsvariablen sind und wollen ihre Realisationen im Folgenden mit $\bar{x}, s^2$ und s bezeichnen. Nehmen wir also an, wir haben für $n := 10$ die in folgender Tabelle gezeigten Realisationen von u.i.v. nach $N(1, 2)$ verteilten Zufallsvariable $\xi_1, ..., \xi_{10}$, wobei für $i = 1, ..., 10$ die Realisation von ξ_i mit x_i bezeichnen ist:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.54	1.01	-3.28	0.35	2.75	-0.51	2.32	1.49	0.96	1.25

Nach Definition 13.6 ist die Stichprobenmittelrealisation dann gegeben durch

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{6.88}{10} = 0.68, \tag{13.52}$$

die Stichprobenvarianzrealisation gegeben durch

$$s^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - 0.68)^2 = \frac{25.37}{9} = 2.82. \tag{13.53}$$

und die Stichprobenstandardabweichungrealisation gegeben durch

$$s = \sqrt{s^2} = \sqrt{2.82} = 1.68. \tag{13.54}$$

13.4. Kovarianz und Korrelation

Häufig genutzte Maße für den Zusammenhang zweier Zufallsvariablen sind die *Kovarianz* und die *Korrelation*. Diese sind wie folgt definiert.

Definition 13.7 (Kovarianz und Korrelation)**.** Die *Kovarianz* zweier Zufallsvariablen ξ und υ ist definiert als

$$\mathbb{C}(\xi, \upsilon) := \mathbb{E}\left((\xi - \mathbb{E}(\xi))\,(\upsilon - \mathbb{E}(\upsilon))\right). \tag{13.55}$$

Die *Korrelation* zweier Zufallsvariablen ξ und υ ist definiert als

$$\rho(\xi, \upsilon) := \frac{\mathbb{C}(\xi, \upsilon)}{\sqrt{\mathbb{V}(\xi)}\sqrt{\mathbb{V}(\upsilon)}} = \frac{\mathbb{C}(\xi, \upsilon)}{\mathbb{S}(\xi)\mathbb{S}(\upsilon)}. \tag{13.56}$$

•

Die Kovarianz einer Zufallsvariable ξ mit sich entspricht ihrer Varianz, da

$$\mathbb{C}(\xi, \xi) = \mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right) = \mathbb{V}(\xi). \tag{13.57}$$

Im Gegensatz zur Varianz kann die Kovarianz aber auch negative Werte annehmen.

Beispiel

Wir wollen beispielgebend für zwei Zufallsvariablen mit gemeinsamer diskreter Verteilung ihre Kovarianz berechnen. Dazu sei $\zeta := (\xi, \upsilon)$ ein Zufallsvektor mit WMF $p_{\xi,\upsilon}$ definiert durch

$p_{\xi,\upsilon}(x,y)$	$y=1$	$y=2$	$y=3$	$p_\xi(x)$
$x=1$	0.10	0.05	0.15	0.30
$x=2$	0.60	0.05	0.05	0.70
$p_\upsilon(y)$	0.70	0.10	0.20	

und damit ξ, υ zwei Zufallsvariablen mit einer bekannten bivariaten Verteilung. Um $\mathbb{C}(\xi, \upsilon)$ zu berechnen, halten wir zunächst fest, dass

$$\mathbb{E}(\xi) = \sum_{x=1}^{2} x p_\xi(x) = 1 \cdot 0.3 + 2 \cdot 0.7 = 1.7 \tag{13.58}$$

und

$$\mathbb{E}(\upsilon) = \sum_{y=1}^{3} y p_\upsilon(y) = 1 \cdot 0.7 + 2 \cdot 0.1 + 3 \cdot 0.2 = 1.5. \tag{13.59}$$

Mit der Definition der Kovarianz von ξ und υ gilt dann

$$
\begin{aligned}
\mathbb{C}(\xi,\upsilon) &= \mathbb{E}((\xi-\mathbb{E}(\xi))(\upsilon-\mathbb{E}(\upsilon))) \\
&= \sum_{x=1}^{2}\sum_{y=1}^{3}(x-\mathbb{E}(\xi))(y-\mathbb{E}(\upsilon))p_{\xi,\upsilon}(x,y) \\
&= \sum_{x=1}^{2}\sum_{y=1}^{3}(x-1.7)(y-1.5)p_{\xi,\upsilon}(x,y) \\
&= \sum_{x=1}^{2}(x-1.7)(1-1.5)p_{\xi,\upsilon}(x,1)+(x-1.7)(2-1.5)p_{\xi,\upsilon}(x,2)+(x-1.7)(3-1.5)p_{\xi,\upsilon}(x,3) \\
&= \quad (1-1.7)(1-1.5)p_{\xi,\upsilon}(1,1)+(1-1.7)(2-1.5)p_{\xi,\upsilon}(1,2)+(1-1.7)(3-1.5)p_{\xi,\upsilon}(1,3) \\
&\quad +(2-1.7)(1-1.5)p_{\xi,\upsilon}(2,1)+(2-1.7)(2-1.5)p_{\xi,\upsilon}(2,2)+(2-1.7)(3-1.5)p_{\xi,\upsilon}(2,3) \\
&= \quad (-0.7)\cdot(-0.5)\cdot 0.10+(-0.7)\cdot 0.5\cdot 0.05+(-0.7)\cdot 1.5\cdot 0.15 \\
&\quad +0.3\cdot(-0.5)\cdot 0.60+0.3\cdot 0.5\cdot 0.05+0.3\cdot 1.5\cdot 0.05 \\
&= 0.035-0.0175-0.1575-0.09+0.0075+0.0225 \\
&= -0.2.
\end{aligned}
\tag{13.60}
$$

Die Kovarianz der Zufallsvariablen ξ und υ mit der in obiger Tabelle festgelegter Verteilung ist also $\mathbb{C}(\xi,\upsilon) = -0.2$.

Die Korrelation $\rho(\xi,\upsilon)$ zweier Zufallsvariablen entspricht ihrer anhand der Standardabweichungen der jeweiligen Zufallsvariablen standardisierten Kovarianz und wird manchmal auch als *Korrelationskoeffizient* von ξ und υ bezeichnet. Ist die Korrelation $\rho(\xi,\upsilon)=0$, so werden ξ und υ *unkorreliert* genannt. Insbesondere ist die Korrelation im Gegensatz zur Kovarianz *normalisiert*, d.h. es gilt, wie wir an späterer Stellte mithilfe der Cauchy-Schwarz Ungleichung (Theorem 14.3) zeigen gilt

$$-1 \leq \rho(\xi,\upsilon) \leq 1. \tag{13.61}$$

Man sagt in diesem Kontext auch, dass die Korrelation im Gegensatz zur Kovarianz maßstabsunabhängig sei: wendet man auf eine Zufallsvariable eine linear-affine Transformation an, so ändert sich die Kovarianz der Zufallsvariablen, nicht aber ihre Korrelation. Das ist die Kernaussage folgenden Theorems.

Theorem 13.9 (Kovarianz und Korrelation bei linear affinen Transformationen von Zufallsvariablen)**.** *ξ und υ seien Zufallsvariablen und es seien $a,b,c,d \in \mathbb{R}$. Dann gelten*

$$\mathbb{C}(a\xi+b, c\upsilon+d) = ac\mathbb{C}(\xi,\upsilon) \tag{13.62}$$

und

$$\rho(a\xi+b, c\upsilon+d) = \rho(\xi,\upsilon). \tag{13.63}$$

∘

Beweis. Es gilt zunächst

$$
\begin{aligned}
\mathbb{C}(a\xi+b, c\upsilon+d) &= \mathbb{E}((a\xi+b-\mathbb{E}(a\xi+b))(c\upsilon+d-\mathbb{E}(c\upsilon+d))) \\
&= \mathbb{E}((a\xi+b-a\mathbb{E}(\xi)-b)(c\upsilon+d-c\mathbb{E}(c\upsilon)-d)) \\
&= \mathbb{E}(a(\xi-\mathbb{E}(\xi))(c(\upsilon-c\mathbb{E}(\upsilon))) \\
&= \mathbb{E}(ac((\xi-\mathbb{E}(\xi))(\upsilon-c\mathbb{E}(\upsilon)))) \\
&= ac\mathbb{C}(\xi,\upsilon)
\end{aligned}
\tag{13.64}
$$

Also folgt

$$\begin{aligned}\rho(a\xi + b, c\upsilon + d) &= \frac{\mathbb{C}(a\xi + b, c\upsilon + d)}{\sqrt{\mathbb{V}(a\xi + b)}\sqrt{\mathbb{V}(c\upsilon + d)}} \\ &= \frac{ac\mathbb{C}(\xi, \upsilon)}{\sqrt{a^2\mathbb{V}(\xi)}\sqrt{c^2\mathbb{V}(\upsilon)}} \\ &= \frac{ac\mathbb{C}(\xi, \upsilon)}{a\mathbb{S}(\xi)c\mathbb{S}(\upsilon)} \\ &= \frac{\mathbb{C}(\xi, \upsilon)}{\mathbb{S}(\xi)\mathbb{S}(\upsilon)} \\ &= \rho(\xi, \upsilon).\end{aligned} \tag{13.65}$$

□

Wie das Berechnen von Varianzen wird auch das Berechnen von Kovarianzen manchmal durch folgendes Theorem, den sogenannten *Kovarianzverschiebungssatz* erleichtert.

Theorem 13.10 (Kovarianzverschiebungssatz)**.** *ξ und υ seien Zufallsvariablen. Dann gilt*

$$\mathbb{C}(\xi, \upsilon) = \mathbb{E}(\xi\upsilon) - \mathbb{E}(\xi)\mathbb{E}(\upsilon). \tag{13.66}$$

∘

Beweis. Mit der Definition der Kovarianz gilt

$$\begin{aligned}\mathbb{C}(\xi, \upsilon) &= \mathbb{E}\left((\xi - \mathbb{E}(\xi))(\upsilon - \mathbb{E}(\upsilon))\right) \\ &= \mathbb{E}\left(\xi\upsilon - \xi\mathbb{E}(\upsilon) - \mathbb{E}(\xi)\upsilon + \mathbb{E}(\xi)\mathbb{E}(\upsilon)\right) \\ &= \mathbb{E}(\xi\upsilon) - \mathbb{E}(\xi)\mathbb{E}(\upsilon) - \mathbb{E}(\xi)\mathbb{E}(\upsilon) + \mathbb{E}(\xi)\mathbb{E}(\upsilon) \\ &= \mathbb{E}(\xi\upsilon) - \mathbb{E}(\xi)\mathbb{E}(\upsilon).\end{aligned} \tag{13.67}$$

□

Natürlich ist Theorem 13.10 nur dann wirklich nützlich, wenn $\mathbb{E}(\xi\upsilon)$ leicht zu berechnen sind. Der Kovarianzverschiebungssatz in Theorem 13.4 ergibt sich aus Theorem 13.10 im Spezialfall, dass $\upsilon := \xi$, da dann gilt

$$\mathbb{V}(\xi) = \mathbb{C}(\xi, \xi) = \mathbb{E}(\xi\xi) - \mathbb{E}(\xi)\mathbb{E}(\xi) = \mathbb{E}(\xi^2) - \mathbb{E}(\xi)\mathbb{E}(\xi) \tag{13.68}$$

Mithilfe des Begriffes des Kovarianz ist es möglich eine stärkere Aussage über die Varianzen von Summen und Differenzen von Zufallsvariablen zu treffen als es in Theorem 13.5 der Fall war, wo lediglich *unabhängige* Zufallsvariablen betrachtetet wurden. Folgende Aussagen gelten generell.

Theorem 13.11 (Varianzen von Summen und Differenzen von Zufallsvariablen)**.** *ξ und υ seien zwei Zufallsvariablen und es seien $a, b, c \in \mathbb{R}$. Dann gilt*

$$\mathbb{V}(a\xi + b\upsilon + c) = a^2\mathbb{V}(\xi) + b^2\mathbb{V}(\upsilon) + 2ab\mathbb{C}(\xi, \upsilon). \tag{13.69}$$

Speziell gelten

$$\mathbb{V}(\xi + \upsilon) = \mathbb{V}(\xi) + \mathbb{V}(\upsilon) + 2\mathbb{C}(\xi, \upsilon) \tag{13.70}$$

und

$$\mathbb{V}(\xi - \upsilon) = \mathbb{V}(\xi) + \mathbb{V}(\upsilon) - 2\mathbb{C}(\xi, \upsilon) \tag{13.71}$$

∘

Beweis. Wir halten zunächst fest, dass

$$\mathbb{E}(a\xi + b\upsilon + c) = a\mathbb{E}(\xi) + b\mathbb{E}(\upsilon) + c. \tag{13.72}$$

Es ergibt sich also

$$\begin{aligned}
&\mathbb{V}(a\xi + b\upsilon + c)\\
&= \mathbb{E}\left((a\xi + b\upsilon + c - a\mathbb{E}(\xi) - b\mathbb{E}(\upsilon) - c)^2\right)\\
&= \mathbb{E}\left((a(\xi - \mathbb{E}(\xi)) + b(\upsilon - \mathbb{E}(\upsilon)))^2\right)\\
&= \mathbb{E}\left(a^2(\xi - \mathbb{E}(\xi))^2 + 2ab(\xi - \mathbb{E}(\xi))(\upsilon - \mathbb{E}(\upsilon))) + b^2(\upsilon - \mathbb{E}(\upsilon))^2\right)\\
&= a^2\mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right) + b^2\mathbb{E}\left((\upsilon - \mathbb{E}(\upsilon))^2\right) + 2ab\mathbb{E}\left((\xi - \mathbb{E}(\xi))(\upsilon - \mathbb{E}(\upsilon))\right)\\
&= a^2\mathbb{V}(\xi) + b^2\mathbb{V}(\upsilon) + 2ab\mathbb{C}(\xi, \upsilon)
\end{aligned} \tag{13.73}$$

Die Spezialfälle folgen dann direkt mit $a := b := 1$ und $a := 1, b := -1$, respektive.

□

Im Gegensatz zu Erwartungswerten addieren sich die Varianzen von Zufallsvariablen also nicht einfach, sondern die Varianz der Summe zweier Zufallsvariablen hängt von ihrer Kovarianz ab. Ist diese zum Beispiel im Fall der Summe zweier Zufallsvariablen positiv, so verstärkt sie die Varianz der Zufallsvariable, die sich aus der Addition der Zufallsvariablen ergibt. Intuitiv führt hierbei die Realisierung eines Extremwertes einer der Zufallvariablen häufigt auch zu der Realisierung eines Extremwertes der anderen Zufallsvariablen, so dass die Variabilität der Summe der Zufallsvariablen überproportional verstärkt wird.

Schließlich wollen wir mit Theorem 13.12 einen ersten Eindruck zum Zusammenhang von Kovarianz und Korrelation mit dem Begriff der Unabhängigkeit von Zufallsvariablen erlangen. Es zeigt sich, dass Kovarianz und Korrelation lediglich für bestimmte Formen der Abhängigkeit von Zufallsvariablen sensitiv sind und insbesondere, dass von einer Kovarianz von Null *nicht* auf die Unabhängigkeit der Zufallsvariablen geschlossen werden kann. Anderseits impliziert die Unabhängigkeit zweier Zufallsvariablen immer, dass ihre Kovarianz Null und sie damit unkorreliert sind. Abhängigkeit und Unabhängigkeit von Zufallsvariablen sind also sehr viel allgemeinere Begrifflichkeiten zur Beschreibung des Zusammenhangs von Zufallsvariablen als Kovarianz und Korrelation.

Theorem 13.12 (Korrelation und Unabhängigkeit)**.** *ξ und υ seien zwei Zufallsvariablen. Wenn ξ und υ unabhängig sind, dann ist $\mathbb{C}(\xi, \upsilon) = 0$ und ξ und υ sind unkorreliert. Ist dagegen $\mathbb{C}(\xi, \upsilon) = 0$ und sind ξ und υ somit unkorreliert, dann sind ξ und υ nicht notwendigerweise unabhängig.*

∘

Beweis. Wir zeigen zunächst, dass aus der Unabhängigkeit von ξ und υ $\mathbb{C}(\xi, \upsilon) = 0$ folgt. Hierzu halten wir zunächst fest, dass für unabhängige Zufallsvariablen gilt, dass

$$\mathbb{E}(\xi\upsilon) = \mathbb{E}(\xi)\mathbb{E}(\upsilon). \tag{13.74}$$

Mit dem Kovarianzverschiebungssatz folgt dann

$$\mathbb{C}(\xi, \upsilon) = \mathbb{E}(\xi\upsilon) - \mathbb{E}(\xi)\mathbb{E}(\upsilon) = \mathbb{E}(\xi)\mathbb{E}(\upsilon) - \mathbb{E}(\xi)\mathbb{E}(\upsilon) = 0. \tag{13.75}$$

Mit der Definition des Korrelationskoeffizienten folgt dann unmittelbar, dass $\rho(\xi, \upsilon) = 0$ und ξ und υ somit unkorreliert sind.

Wir zeigen nun durch Angabe eines Beispiels, dass die Kovarianz von abhängigen Zufallsvariablen ξ und υ Null sein kann. Zu diesem Zweck betrachten wir den Fall zweier diskreter Zufallsvariablen ξ und υ mit

Ergebnisräumen $\mathcal{X} = \{-1, 0, 1\}$ und $\upsilon = \{0, 1\}$, marginaler WMF von ξ gegeben durch $p_\xi(x) := 1/3$ für $x \in \mathcal{X}$ und der Definition $\upsilon := \xi^2$. Wir halten dann zunächst fest, dass

$$\mathbb{E}(\xi) = \sum_{x \in \mathcal{X}} x p_\xi(x) = -1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0 \tag{13.76}$$

und

$$\mathbb{E}(\xi\upsilon) = \mathbb{E}(\xi\xi^2) = \mathbb{E}(\xi^3) = \sum_{x \in \mathcal{X}} x^3 p_\xi(x) = -1^3 \cdot \frac{1}{3} + 0^3 \cdot \frac{1}{3} + 1^3 \cdot \frac{1}{3} = 0. \tag{13.77}$$

Mit dem Kovarianzverschiebungssatz ergibt sich dann

$$\mathbb{C}(\xi, \upsilon) = \mathbb{E}(\xi\upsilon) - \mathbb{E}(\xi)\mathbb{E}(\upsilon) = \mathbb{E}(\xi^3) - \mathbb{E}(\xi)\mathbb{E}(\upsilon) = 0 - 0 \cdot \mathbb{E}(\upsilon) = 0. \tag{13.78}$$

Die Kovarianz von ξ und υ ist also Null. Wie unten gezeigt faktorisiert die gemeinsame WMF von ξ und υ jedoch nicht, und somit sind ξ und υ nicht unabhängig. Wir halten zunächst fest, dass die Definition von $\upsilon := \xi^2$ die folgende bedingte WMF von υ gegeben ξ impliziert:

$p_{\upsilon\|\xi}(y\|x)$	$x = -1$	$x = 0$	$x = 1$
$y = 0$	0	1	0
$y = 1$	1	0	1

Die marginale WMF p_ξ und die bedingte WMF $p_{\upsilon|\xi}$ implizieren wiederum die gemeinsame WMF

$p_{\xi,\upsilon}(x, y)$	$x = -1$	$x = 0$	$x = 1$	$p_\upsilon(y)$
$y = 0$	0	1/3	0	1/3
$y = 1$	1/3	0	1/3	2/3
$p_\xi(x)$	1/3	1/3	1/3	

von ξ und υ. Es gilt also zum Beispiel

$$p_{\xi,\upsilon}(-1, 0) = 0 \neq \frac{1}{9} = \frac{1}{3} \cdot \frac{1}{3} = p_\xi(-1)p_\upsilon(0) \tag{13.79}$$

und damit sind ξ und υ nicht unabhängig.

□

13.5. 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 13.8 (Kovarianzmatrix eines Zufallsvektors)**.** ξ sei ein n-dimensionaler Zufallvektor. Dann ist die *Kovarianzmatrix* von ξ definiert als die $n \times n$ Matrix

$$\mathbb{C}(\xi) := \mathbb{E}\left((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right). \tag{13.80}$$

•

Die Kovarianzmatrix ist in Definition 13.8 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 13.13 (Kovarianzmatrix eines Zufallsvektors)**.** *ξ sei ein n-dimensionaler Zufallvektor und $\mathbb{C}(\xi)$ sei seine Kovarianzmatrix. Dann gilt*

$$\mathbb{C}(\xi) = \left(\mathbb{C}(\xi_i, \xi_j)\right)_{1 \leq i,j \leq n} = \begin{pmatrix} \mathbb{C}(\xi_1, \xi_1) & \mathbb{C}(\xi_1, \xi_2) & \cdots & \mathbb{C}(\xi_1, \xi_n) \\ \mathbb{C}(\xi_2, \xi_1) & \mathbb{C}(\xi_2, \xi_2) & \cdots & \mathbb{C}(\xi_2, \xi_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}(\xi_n, \xi_1) & \mathbb{C}(\xi_n, \xi_2) & \cdots & \mathbb{C}(\xi_n, \xi_n) \end{pmatrix}. \tag{13.81}$$

∘

Beweis. Es gilt

$$\mathbb{C}(\xi) := \mathbb{E}\left((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right) \tag{13.82}$$

$$= \mathbb{E}\left(\left(\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} - \begin{pmatrix} \mathbb{E}(\xi_1) \\ \vdots \\ \mathbb{E}(\xi_n) \end{pmatrix}\right)\left(\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} - \begin{pmatrix} \mathbb{E}(\xi_1) \\ \vdots \\ \mathbb{E}(\xi_n) \end{pmatrix}\right)^T\right) \tag{13.83}$$

$$= \mathbb{E}\left(\begin{pmatrix} \xi_1 - \mathbb{E}(\xi_1) \\ \vdots \\ \xi_n - \mathbb{E}(\xi_n) \end{pmatrix}\begin{pmatrix} \xi_1 - \mathbb{E}(\xi_1) \\ \vdots \\ \xi_n - \mathbb{E}(\xi_n) \end{pmatrix}^T\right) \tag{13.84}$$

$$= \mathbb{E}\left(\begin{pmatrix} \xi_1 - \mathbb{E}(\xi_1) \\ \vdots \\ \xi_n - \mathbb{E}(\xi_n) \end{pmatrix}\begin{pmatrix} \xi_1 - \mathbb{E}(\xi_1) & \dots & \xi_n - \mathbb{E}(\xi_n) \end{pmatrix}\right) \tag{13.85}$$

$$= \mathbb{E}\begin{pmatrix} (\xi_1 - \mathbb{E}(\xi_1))(\xi_1 - \mathbb{E}(\xi_1)) & \dots & (\xi_1 - \mathbb{E}(\xi_1))(\xi_n - \mathbb{E}(\xi_n) \\ \vdots & \ddots & \vdots \\ (\xi_n - \mathbb{E}(\xi_n))(\xi_1 - \mathbb{E}(\xi_1)) & \dots & (\xi_n - \mathbb{E}(\xi_n))(\xi_n - \mathbb{E}(\xi_n)) \end{pmatrix} \tag{13.86}$$

$$= \left(\mathbb{E}\left((\xi_i - \mathbb{E}(\xi_i))(\xi_j - \mathbb{E}(\xi_j))\right)\right)_{1 \leq i,j \leq n} \tag{13.87}$$

$$= \left(\mathbb{C}(\xi_i, \xi_j)\right)_{1 \leq i,j \leq n}. \tag{13.88}$$

$$\tag{13.89}$$

□

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 Zufallvektorprodukten an, die zum Beispiel im Rahmen der Kanonischen Korrelationsanalyse hilfreich ist.

Theorem 13.14 (Kovarianzmatrizen von Zufallsvektoren)**.** *Es seien*

$$\zeta = \begin{pmatrix} \xi \\ \upsilon \end{pmatrix} \text{ mit } \mathbb{E}(\zeta) := 0_m \tag{13.90}$$

ein $m_\xi + m_\upsilon$-dimensionaler Zufallsvektor und sein Erwartungswertvektor, respektive. Dann kann die $m \times m$ Kovarianzmatrix von ζ geschrieben werden als

$$\mathbb{C}(\zeta) = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\upsilon} \\ \Sigma_{\upsilon\xi} & \Sigma_{\upsilon\upsilon} \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{13.91}$$

wobei

$$\begin{aligned} \Sigma_{\xi\xi} &:= \mathbb{E}\left(\xi\xi^T\right) \in \mathbb{R}^{m_\xi \times m_\xi} \\ \Sigma_{\xi\upsilon} &:= \mathbb{E}\left(\xi\upsilon^T\right) \in \mathbb{R}^{m_\xi \times m_\upsilon} \\ \Sigma_{\upsilon\xi} &:= \mathbb{E}\left(\upsilon\xi^T\right) \in \mathbb{R}^{m_\upsilon \times m_\xi} \\ \Sigma_{\upsilon\upsilon} &:= \mathbb{E}\left(\upsilon\upsilon^T\right) \in \mathbb{R}^{m_\xi \times m_\upsilon} \end{aligned} \tag{13.92}$$

∘

Beweis. Nach Definition der Kovarianzmatrix eines Zufallsvektors gilt

$$\begin{aligned} \mathbb{C}(z) &= \mathbb{E}\left((\zeta - \mathbb{E}(\zeta))(\zeta - \mathbb{E}(\zeta))^T\right) \\ &= \mathbb{E}\left((\zeta - 0_m)(\zeta - 0_m)^T\right) \\ &= \mathbb{E}\left(\zeta\zeta^T\right) \\ &= \mathbb{E}\left(\begin{pmatrix} \xi \\ \upsilon \end{pmatrix}\begin{pmatrix} \xi^T & \upsilon^T \end{pmatrix}\right) \\ &= \mathbb{E}\left(\begin{pmatrix} \xi\xi^T & \xi\upsilon^T \\ \upsilon\xi^T & \upsilon\upsilon^T \end{pmatrix}\right) \\ &= \begin{pmatrix} \mathbb{E}\left(\xi\xi^T\right) & \mathbb{E}\left(\xi\upsilon^T\right) \\ \mathbb{E}\left(\upsilon\xi^T\right) & \mathbb{E}\left(\upsilon\upsilon^T\right) \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\upsilon} \\ \Sigma_{\upsilon\xi} & \Sigma_{\upsilon\upsilon} \end{pmatrix} \end{aligned} \tag{13.93}$$

□

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 13.9 (Korrelationsmatrix)**.** ξ sei ein n-dimensionaler Zufallsvektor. Dann ist die *Korrelationsmatrix* von ξ definiert als die $n \times n$ Matrix

$$\mathbb{R}(\xi) := \left(\rho_{ij}\right)_{1 \leq i,j \leq n} = \left(\frac{\mathbb{C}(\xi_i, \xi_j)}{\sqrt{\mathbb{V}(\xi_i)}\sqrt{\mathbb{V}(\xi_j)}}\right)_{1 \leq i,j \leq n}. \tag{13.94}$$

•

Da es sich bei den Varianzen der Komponenten von ξ um die Diagonalelement der Kovarianzmatrix von ξ handelt, ist die Korrelationsmatrix natürlich in der Kovarianzmatrix imoplizit. Weiterhin gelten, wie immer für Korrelationen, für die Einträge $\rho_{ij}, 1 \leq i, j \leq n$ der Korrelationsmatrix, dass

$$\rho_{ij} \in [-1, 1] \text{ für } 1 \leq i, j \in n \text{ und } \rho_{ii} = 1 \text{ für } 1 \leq i \leq n. \tag{13.95}$$

13.6. Stichprobenkennzahlen von Zufallsvektoren

Die Begriffe des Stichprobenmittels, der Stichprobenvarianz und der Stichprobenkovarianz lassen sich auch auf den Fall multivariater Stichproben übertragen. Wir nutzen folgende Definition.

Definition 13.10 (Stichprobenmittel, -kovarianmatrix und -korrelationsmatrix)**.** $\upsilon_1, ..., \upsilon_n$ sei eine Menge von m-dimensionalen Zufallsvektoren, genannt *Stichprobe*.

- Das *Stichprobenmittel+ der $\upsilon_1, ..., \upsilon_n$ ist definiert als der m-dimensionale Vektor

$$\bar{\upsilon} := \frac{1}{n} \sum_{i=1}^{n} \upsilon_i. \tag{13.96}$$

- Die *Stichprobenkovarianzmatrix* der $\upsilon_1, ..., \upsilon_n$ ist definiert als die $m \times m$ Matrix

$$C := \frac{1}{n-1} \sum_{i=1}^{n} (\upsilon_i - \bar{\upsilon})(\upsilon_i - \bar{\upsilon})^T. \tag{13.97}$$

- Die *Stichprobenkorrelationsmatrix* der $\upsilon_1, ..., \upsilon_n$ ist definiert als die $m \times m$ Matrix

$$D := \left(\frac{(C)_{ij}}{\sqrt{(C)_{ii}}\sqrt{(C)_{jj}}} \right)_{1 \leq i,j \leq m}. \tag{13.98}$$

•

Zur konkreten Berechnung von Stichprobenmittel, Stichprobenkovarianzmatrix und Stichprobenkorrrelationsmatrix basierend auf einem multivariaten Datensatz bieten sich die Aussagen des folgenden Theorems an.

Theorem 13.15 (Datenmatrix und Stichprobenstatistiken)**.**

Es sei

$$\Upsilon := \begin{pmatrix} \upsilon_1 & \cdots & \upsilon_n \end{pmatrix} \tag{13.99}$$

eine $m \times n$ Datenmatrix, die durch die spaltenweise Konkatenation von n m-dimensionalen Zufallvektoren $\upsilon_1, ..., \upsilon_n$ gegeben sei. Dann ergeben sich

- *für das Stichprobenmittel*

$$\bar{\upsilon} = \frac{1}{n} \Upsilon 1_n, \tag{13.100}$$

- *für die Stichprobenkovarianzmatrix*

$$C = \frac{1}{n-1}\left(\Upsilon\left(I_n - \frac{1}{n}1_{nn}\right)\Upsilon^T\right), \tag{13.101}$$

- *und für Stichprobenkorrelationsmatrix mit*

$$D := diag\left(\sqrt{C_{y_{ii}}}^{-1}, i = 1, ..., m\right), \tag{13.102}$$

dass

$$R = DCD. \tag{13.103}$$

∘

Beweis. Die Darstellung des Stichprobenmittels ergibt sich aus

$$\begin{aligned}
\bar{\upsilon} &:= \frac{1}{n}\sum_{i=1}^{n}\upsilon_i \\
&= \frac{1}{n}\begin{pmatrix}\sum_{i=1}^{n}\upsilon_{i1}\\ \vdots \\ \sum_{i=1}^{n}\upsilon_{im}\end{pmatrix} \\
&= \frac{1}{n}\left(\begin{pmatrix}\upsilon_{11} & \cdots & \upsilon_{n1}\\ \vdots & \ddots & \vdots \\ \upsilon_{1m} & \cdots & \upsilon_{nm}\end{pmatrix}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right) \\
&= \frac{1}{n}\Upsilon 1_n.
\end{aligned} \tag{13.104}$$

Hinsichtlich der Darstellung der Stichprobenkovarianzmatrix halten wir zunächst fest, dass nach Definition gilt, dass

$$\begin{aligned}
C &:= \frac{1}{n-1}\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon})(\upsilon_i - \bar{\upsilon})^T \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\upsilon_i\upsilon_i^T - \upsilon_i\bar{\upsilon}^T - \bar{\upsilon}\upsilon_i^T + \bar{\upsilon}\bar{\upsilon}^T) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - \sum_{i=1}^{n}\upsilon_i\bar{\upsilon}^T - \sum_{i=1}^{n}\bar{\upsilon}\upsilon_i^T + \sum_{i=1}^{n}\bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - n\bar{\upsilon}\bar{\upsilon}^T - n\bar{\upsilon}\bar{\upsilon}^T + n\bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - n\bar{\upsilon}\bar{\upsilon}^T\right).
\end{aligned} \tag{13.105}$$

Mit $1_n 1_n^T = 1_{nn}$ ergibt sich dann weiterhin

$$\begin{aligned}
\Upsilon \left(I_n - \frac{1}{n} 1_{nn} \right) \Upsilon^T &= \left(\Upsilon I_n - \frac{1}{n} \Upsilon 1_{nn} \right) \Upsilon^T \\
&= \Upsilon \Upsilon^T - \frac{1}{n} \Upsilon 1_{nn} \Upsilon^T \\
&= \begin{pmatrix} \upsilon_1 & \cdots & \upsilon_n \end{pmatrix} \begin{pmatrix} \upsilon_1^T \\ \vdots \\ \upsilon_n^T \end{pmatrix} - \frac{1}{n} \Upsilon 1_n 1_n^T \Upsilon^T \\
&= \sum_{i=1}^{n} \upsilon_i \upsilon_i^T - n \left(\frac{1}{n} \Upsilon 1_n \right) \left(\frac{1}{n} 1_n^T \Upsilon^T \right) \\
&= \sum_{i=1}^{n} \upsilon_i \upsilon_i^T - n \bar{\upsilon} \bar{\upsilon}^T \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (\upsilon_i - \bar{\upsilon})(\upsilon_i - \bar{\upsilon})^T \\
&= C.
\end{aligned} \tag{13.106}$$

Hinsichtlich der Korrelationsmatrix ergibt sich nach Definition und für ein beliebiges Indexpaar i, j mit $1 \leq i, j \leq m$ schließlich, dass

$$\begin{aligned}
R_{y_{ij}} &= \frac{(C)_{ij}}{\sqrt{(C)_{ii}} \sqrt{(C)_{jj}}} \\
&= \frac{1}{\sqrt{(C)_{ii}}} (C)_{ij} \frac{1}{\sqrt{(C)_{jj}}} \\
&= (DCD)_{ij}.
\end{aligned} \tag{13.107}$$

□

13.7. Selbstkontrollfragen

1. Geben Sie die Definition des Erwartungswerts einer Zufallsvariable wieder.
2. Geben Sie die Interpretation der Erwartungswerts einer Zufallsvariable wieder
3. Berechnen Sie den Erwartungswert einer Bernoulli Zufallsvariable.
4. Geben Sie das Theorem zu den Eigenschaften des Erwartungswerts wieder.
5. Geben Sie die Definition der Varianz und der Standardabweichung einer Zufallsvariable wieder.
6. Geben Sie die Interpretation der Varianz einer Zufallsvariable wieder.
7. Berechnen Sie die Varianz einer Bernoulli Zufallsvariable.
8. Geben Sie das Theorem zum Varianzverschiebungssatz wieder.
9. Geben Sie das Theorem zu den Eigenschaften der Varianz wieder.
10. Geben Sie die Definition des Begriffs einer Stichprobe wieder.
11. Geben Sie die Definitionen von Stichprobenmittel, -varianz und -standardabweichung wieder.
12. Geben Sie die Definition von Kovarianz und Korrelation zweier Zufallsvariablen wieder.
13. Geben Sie das Theorem zum Kovarianzverschiebungssatz wieder.
14. Geben Sie das Theorem zu Varianzen von Summen und Differenzen von Zufallsvariablen wieder.
15. Geben Sie das Theorem zur Korrelation und Unabhängigkeit zweier Zufallsvariablen wieder.

14. 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 *Chebychev Ungleichung*, Kapitel 14.1) und *Erwartungswertungleichungen* (*Cauchy-Schwarz Ungleichung*, *Korrelationsungleichung* und *Jensensche Ungleichung*, Kapitel 14.2).

14.1. Wahrscheinlichkeitsungleichungen

Die *Markov Ungleichung* stellt einen Bezug zwischen den Überschreitungswahrscheinlichkeiten (vgl. Theorem 11.2) 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 14.1 (Markov Ungleichung)**.** *ξ sei eine Zufallsvariable mit $\mathbb{P}(\xi \geq 0) = 1$. Dann gilt für alle $x \in \mathbb{R}$, dass*

$$\mathbb{P}(\xi \geq x) \leq \frac{\mathbb{E}(\xi)}{x}. \tag{14.1}$$

∘

Beweis. Wir betrachten den Fall einer kontinuierlichen Zufallsvariable ξ mit WDF p. Wir halten zunächst fest, dass

$$\mathbb{E}(\xi) = \int_{-\infty}^{\infty} s\, p(s)\, ds = \int_{0}^{\infty} s\, p(s)\, ds = \int_{0}^{x} s\, p(s)\, ds + \int_{x}^{\infty} s\, p(s)\, ds, \tag{14.2}$$

weil ξ nicht-negativ ist. Es folgt dann

$$\mathbb{E}(\xi) \geq \int_{x}^{\infty} s\, p(s)\, ds \geq \int_{x}^{\infty} x\, p(s)\, ds = x \int_{x}^{\infty} p(s)\, ds = x\, \mathbb{P}(\xi \geq x). \tag{14.3}$$

Dabei gilt die erste Ungleichung, weil

$$\int_{0}^{x} s\, p(s)\, ds \geq 0, \tag{14.4}$$

und die zweite Ungleichung gilt, weil $x \leq \xi$ für $\xi \in [x, \infty[$. Es folgt also, dass

$$\mathbb{E}(\xi) \geq x\, \mathbb{P}(\xi \geq x) \Leftrightarrow \mathbb{P}(\xi \geq x) \leq \frac{\mathbb{E}(\xi)}{x}. \tag{14.5}$$

□

Gilt beispielweise für eine nichtnegative Zufallsvariable ξ, dass $\mathbb{E}(\xi) = 1$ ist, dann folgt aus der Markov Ungleichung, dass

$$\mathbb{P}(\xi \geq 100) \leq 0.01. \tag{14.6}$$

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 11.10) und wir haben gesehen, dass für den Erwartungswert einer Gamma-Zufallsvariable $\mathbb{E}(\xi) = \alpha\beta$ gilt (vgl. Theorem 13.8). 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 14.1 A stellen wir die KVF P dieser Zufallsvariable dar, in Abbildung 14.1 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.

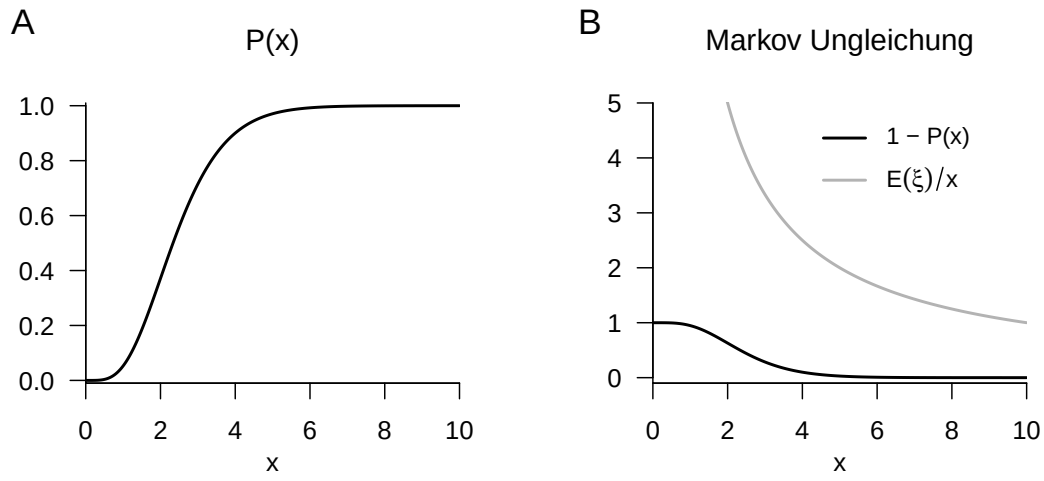


Abbildung 14.1. Markov Ungleichung am Beispiel einer Gamma-Zufallsvariable.

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 13.5 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 14.2 (Chebyshev Ungleichung)**.** *Es sei ξ eine Zufallsvariable mit Varianz $\mathbb{V}(\xi)$. Dann gilt für alle $x \in \mathbb{R}$*

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq x) \leq \frac{\mathbb{V}(\xi)}{x^2}. \tag{14.7}$$

∘

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

$$a^2 \geq b^2 \Rightarrow \sqrt{a^2} \geq \sqrt{b^2} \Rightarrow a \geq b \Rightarrow |a| \geq b. \tag{14.8}$$

(2) $a^2 \geq b^2$ für $a \leq 0$ und $b \geq 0$. Dann gilt

$$a^2 \geq b^2 \Rightarrow \sqrt{a^2} \geq \sqrt{b^2} \Rightarrow -a \geq b \Rightarrow |a| \geq b. \tag{14.9}$$

(3) $a^2 \geq b^2$ für $a \geq 0$ und $b \leq 0$. Dann gilt

$$a^2 \geq b^2 \Rightarrow \sqrt{a^2} \geq \sqrt{b^2} \Rightarrow a \geq -b \geq b \Rightarrow |a| \geq b. \tag{14.10}$$

(4) $a^2 \geq b^2$ für $a \leq 0$ und $b \leq 0$. Dann gilt

$$a^2 \geq b^2 \Rightarrow \sqrt{a^2} \geq \sqrt{b^2} \Rightarrow -a \geq -b \geq b \Rightarrow |a| \geq b. \tag{14.11}$$

Als nächstes definieren wir $\upsilon := (\xi - \mathbb{E}(\xi))^2$. Dann gilt offenbar $\upsilon \geq 0$ und es folgt aus der Markov Ungleichung

$$\begin{aligned} &\mathbb{P}\left(\upsilon \geq x^2\right) \leq \frac{\mathbb{E}(\upsilon)}{x^2} \\ \Leftrightarrow\ &\mathbb{P}\left((\xi - \mathbb{E}(\xi))^2 \geq x^2\right) \leq \frac{\mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right)}{x^2} \\ \Leftrightarrow\ &\mathbb{P}(|\xi - \mathbb{E}(\xi)| \geq x) \leq \frac{\mathbb{V}(\xi)}{x^2}. \end{aligned} \tag{14.12}$$

□

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

$$\mathbb{P}\left(|\xi - \mathbb{E}(\xi)| \geq 2\sqrt{\mathbb{V}(\xi)}\right) \leq \frac{\mathbb{V}(\xi)}{\left(2\sqrt{\mathbb{V}(\xi)}\right)^2} = \frac{1}{4} \tag{14.13}$$

und

$$\mathbb{P}\left(|\xi - \mathbb{E}(\xi)| \geq 3\sqrt{\mathbb{V}(\xi)}\right) \leq \frac{\mathbb{V}(\xi)}{\left(3\sqrt{\mathbb{V}(\xi)}\right)^2} = \frac{1}{9}. \tag{14.14}$$

14.2. 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 14.3 (Cauchy-Schwarz Ungleichung)**.** *ξ und υ seien zwei Zufallsvariablen und $\mathbb{E}(\xi\upsilon)$ sei endlich. Dann gilt*

$$\mathbb{E}(\xi\upsilon)^2 \leq \mathbb{E}\left(\xi^2\right)\mathbb{E}\left(\upsilon^2\right). \tag{14.15}$$

∘

Für einen Beweis verweisen wir auf den Beweis von Theorem 4.6.2 in DeGroot & Schervish (2012). Analog zu Theorem 14.3 gilt zum Beispiel für Vektoren $x, y \in \mathbb{R}^n$, dass

$$\langle x, y\rangle^2 \leq \langle x, x\rangle\langle x, y\rangle. \tag{14.16}$$

Im Kontext der probabilistischen Datenanalyse ist die Anwendung der Cauchy-Schwarz Ungleichung vor allem im Beweis der sogenannten *Korrelationsungleichung* von Relevanz.

Theorem 14.4 (Korrelationsungleichung)**.** *ξ und υ seien Zufallsvariablen mit $\mathbb{V}(\xi), \mathbb{V}(\upsilon) > 0$. Dann gelten*

$$\frac{\mathbb{C}(\xi, \upsilon)^2}{\mathbb{V}(\xi)\mathbb{V}(\upsilon)} \leq 1 \text{ und } -1 \leq \rho(\xi, \upsilon) \leq 1. \tag{14.17}$$

∘

Beweis. Mit der Cauchy-Schwarz-Ungleichung für zwei Zufallsvariablen α und β gilt, dass

$$\mathbb{E}(\alpha\beta)^2 \leq \mathbb{E}\left(\alpha^2\right)\mathbb{E}\left(\beta^2\right). \tag{14.18}$$

Wir definieren nun $\alpha := \xi - \mathbb{E}(\xi)$ und $\beta := \upsilon - \mathbb{E}(\upsilon)$. Dann besagt die Cauchy-Schwarz Ungleichung gerade, dass

$$\mathbb{E}\left((\xi - \mathbb{E}(\xi))(\upsilon - \mathbb{E}(\upsilon))\right)^2 \leq \mathbb{E}\left((\xi - \mathbb{E}(\xi))^2\right)\mathbb{E}\left((\upsilon - \mathbb{E}(\upsilon))^2\right). \tag{14.19}$$

Also gilt

$$\mathbb{C}(\xi, \upsilon)^2 \leq \mathbb{V}(\xi)\mathbb{V}(\upsilon) \Leftrightarrow \frac{\mathbb{C}(\xi, \upsilon)^2}{\mathbb{V}(\xi)\mathbb{V}(\upsilon)} \leq 1. \tag{14.20}$$

Weiterhin folgt aus der Definition der Korrelation dann sofort, dass auch

$$\rho(\xi, \upsilon)^2 \leq 1. \tag{14.21}$$

Dann gilt aber auch

$$|\rho(\xi, \upsilon)^2| \leq 1 \Leftrightarrow -1 \leq \rho(\xi, \upsilon) \leq 1, \tag{14.22}$$

denn

$$\rho(\xi, \upsilon)^2 \leq 1 \Rightarrow \sqrt{\rho(\xi, \upsilon)^2} \leq \sqrt{1} \Rightarrow \quad \rho(\xi, \upsilon) \leq 1 \Rightarrow |\rho(\xi, \upsilon)| \leq 1 \text{ für } \rho(\xi, \upsilon) \geq 0 \tag{14.23}$$

und

$$\rho(\xi, \upsilon)^2 \leq 1 \Rightarrow \sqrt{\rho(\xi, \upsilon)^2} \leq \sqrt{1} \Rightarrow -\rho(\xi, \upsilon) \leq 1 \Rightarrow |\rho(\xi, \upsilon)| \leq 1 \text{ für } \rho(\xi, \upsilon) \leq 0 \tag{14.24}$$

□

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 13.4).

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. **?@sec-parameterschätzung**) 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, wohingegehn 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 14.2.

Theorem 14.5 (Jensensche Ungleichung)**.** *ξ sei eine Zufallsvariable und $g : \mathbb{R} \to \mathbb{R}$ eine konvexe Funktion, d.h.*

$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda)g(x_2) \tag{14.25}$$

für alle $x_1, x_2 \in \mathbb{R}, \lambda \in [0,1]$. Dann gilt

$$\mathbb{E}(g(\xi)) \geq g(\mathbb{E}(\xi)). \tag{14.26}$$

Analog sei $g : \mathbb{R} \to \mathbb{R}$ eine konkave Funktion, d.h.

$$g(\lambda x_1 + (1-\lambda)x_2) \geq \lambda g(x_1) + (1-\lambda)g(x_2) \tag{14.27}$$

für alle $x_1, x_2 \in \mathbb{R}, \lambda \in [0,1]$. Dann gilt

$$\mathbb{E}(g(\xi)) \leq g(\mathbb{E}(\xi)). \tag{14.28}$$

∘

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

$$g(x) \geq t(x) := g(x_0) + g'(x_0)(x - x_0) \tag{14.29}$$

Wir setzen nun $x := \xi$ und $x_0 := \mathbb{E}(\xi)$. Dann gilt mit obiger Ungleichung, dass

$$g(\xi) \geq g(\mathbb{E}(\xi)) + g'(\mathbb{E}(\xi))(\xi - \mathbb{E}(\xi)) \tag{14.30}$$

Erwartungswertbildung ergibt dann

$$\begin{aligned} &\mathbb{E}(g(\xi)) \geq \mathbb{E}(g(\mathbb{E}(\xi))) + \mathbb{E}(g'(\mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))) \\ \Leftrightarrow\ &\mathbb{E}(g(\xi)) \geq g(\mathbb{E}(\xi)) + g'(\mathbb{E}(\xi))\mathbb{E}((\xi - \mathbb{E}(\xi))) \\ \Leftrightarrow\ &\mathbb{E}(g(\xi)) \geq g(\mathbb{E}(\xi)) + g'(\mathbb{E}(\xi))(\mathbb{E}(\xi) - \mathbb{E}(\xi)) \\ \Leftrightarrow\ &\mathbb{E}(g(\xi)) \geq g(\mathbb{E}(\xi)). \end{aligned} \tag{14.31}$$

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

$$\begin{aligned} &\mathbb{E}(-g(\xi)) \geq -g(\mathbb{E}(\xi)) \\ \Leftrightarrow\ &-\mathbb{E}(g(\xi)) \geq -g(\mathbb{E}(\xi)) \\ \Leftrightarrow\ &\mathbb{E}(g(\xi)) \leq g(\mathbb{E}(\xi)). \end{aligned} \tag{14.32}$$

□

Im Kontext der Variationalen Bayesianischen Inferenz ist grundlegend, dass der Logarithmus ist eine konkave Funktion ist und damit für eine beliebige Zufallsvariable ξ gilt, dass

$$\mathbb{E}(\ln \xi) \leq \ln \mathbb{E}(\xi). \tag{14.33}$$

14.3. Selbstkontrollfragen

1. Geben Sie die Markov Ungleichung wieder.
2. Geben Sie die Chebyshev Ungleichung wieder.
3. Geben Sie die Cauchy-Schwarz Ungleichung wieder.
4. Geben Sie die Korrelationsungleichung wieder.
5. Geben Sie die Jensensche Ungleichung wider.

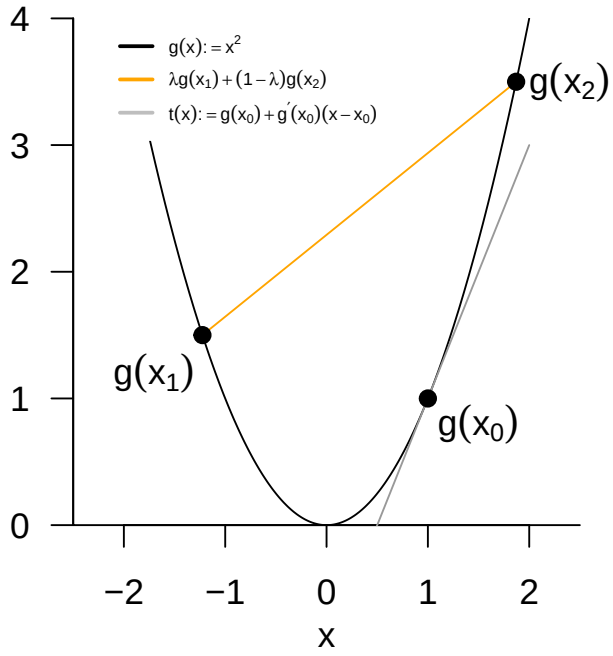


Abbildung 14.2. Darstellung der konvexen Funktion $g(x) := x^2$ mit $x_1 := -\sqrt{1.5}, x_2 := \sqrt{3.5}, \lambda \in [0, 1]$ und $x_0 := 1$.

15. 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 15.1) 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 15.2). 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 6.

15.1. 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 15.1 (Konvergenz in Wahrscheinlichkeit)**.** Eine Folge von Zufallsvariablen $\xi_1, \xi_2, ...$ *konvergiert gegen eine Zufallsvariable* ξ *in Wahrscheinlichkeit*, wenn für jedes noch so kleine $\epsilon > 0$ gilt, dass

$$\lim_{n\to\infty} \mathbb{P}(|\xi_n - \xi| < \epsilon) = 1 \Leftrightarrow \lim_{n\to\infty} \mathbb{P}(|\xi_n - \xi| \geq \epsilon) = 0. \tag{15.1}$$

Die Konvergenz von $\xi_1, \xi_2, ...$ gegen ξ in Wahrscheinlichkeit wird geschrieben als

$$\xi_n \xrightarrow[n\to\infty]{\mathrm{P}} \xi. \tag{15.2}$$

•

$\xi_n \xrightarrow[n\to\infty]{\mathrm{P}} \xi$ heißt also, dass sich die Wahrscheinlichkeit dafür, dass ξ_n in dem zufälligen Intervall

$$]\xi - \epsilon, \xi + \epsilon[\tag{15.3}$$

liegt, unabhängig davon, wie klein dieses Intervall sein mag, 1 nähert, wenn n gegen Unendlich geht. Intuitiv heißt das, dass sich für $n \to \infty$ 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 15.1 (Schwaches Gesetz der Großen Zahlen)**.** *$\xi_1, ..., \xi_n$ seien unabhängig und gleichverteilte Zufallsvariablen mit $\mathbb{E}(\xi_i) = \mu$ für alle $i = 1, ..., n$. Weiterhin bezeichne*

$$\bar{\xi}_n := \frac{1}{n} \sum_{i=1}^{n} \xi_i \tag{15.4}$$

das Stichprobenmittel der $\xi_i, i = 1, ..., n$. Dann konvergiert $\bar{\xi}_n$ in Wahrscheinlichkeit gegen μ,

$$\bar{\xi}_n \xrightarrow[n\to\infty]{P} \mu. \tag{15.5}$$

∘

Für einen Beweis dieses Theorems verweisen wir auf die weiterführende Literatur, so zum Beispiel auf Abschnitt 5.1 in Georgii (2009). Intuitiv heißt

$$\bar{\xi}_n \xrightarrow[n\to\infty]{\mathrm{P}} \mu \tag{15.6}$$

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 $\xi_1, ..., \xi_n \sim N(0, 1)$. Abbildung 15.1 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 $\xi_i, i = 1, ..., n$ liegen. Basierend auf diesen Stichprobenmittelrealisationen zeigt Abbildung 15.1 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 Abwecihung 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.

15.2. 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 υ also als Summe eines deterministischen Einflusses μ und der Summe

$$\varepsilon := \sum_{i=1}^{n} \xi_i \tag{15.7}$$

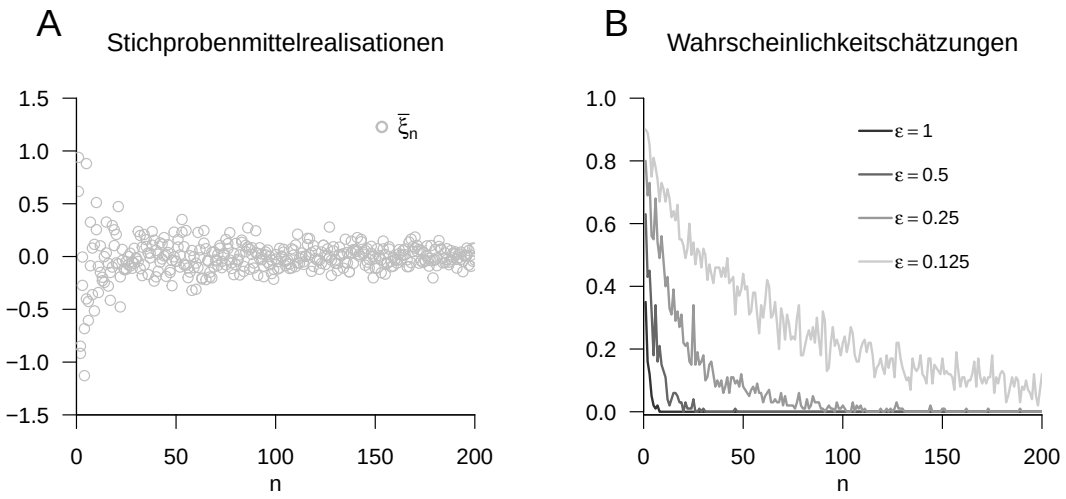


Abbildung 15.1. Simulation des schwachen Gesetz der Großen Zahlen.

einer Vielzahl von unabhängigen Zufallsvariablen $\xi_i, i = 1, ..., n$, die unbekannte Störeinflüsse modellieren sollen, so ist für großes n die Annahme

$$\upsilon = \mu + \varepsilon \text{ mit } \varepsilon \sim N(0, \sigma^2) \tag{15.8}$$

mathematisch gerechtfertigt. Wie wir später sehen werden, liegt die Annahme in Gleichung Gleichung 15.8 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 voraussetzt. In beiden Formulierungen des Zentralen Grenzwertsatzes ist die betrachtete Konvergenz von Zufallsvariablen die *Konvergenz in Verteilung*, welche wir zunächst einführen.

Definition 15.2 (Konvergenz in Verteilung)**.** Eine Folge $\xi_1, \xi_2, ...$ von Zufallsvariablen *konvergiert in Verteilung gegen eine Zufallsvariable* ξ, wenn

$$\lim_{n\to\infty} P_{\xi_n}(x) = P_\xi(x) \tag{15.9}$$

für alle ξ an denen P_ξ stetig ist. Die Konvergenz in Verteilung von $\xi_1, \xi_2, ...$ gegen ξ wird geschrieben als

$$\xi_n \xrightarrow[n\to\infty]{\text{D}} \xi. \tag{15.10}$$

Gilt $\xi_n \xrightarrow[n\to\infty]{\text{D}} \xi$, dann heißt die Verteilung von ξ die *asymptotische Verteilung der Folge* $\xi_1, \xi_2, ...$

•

Die Konvergenz in Verteilung ist also eine Aussage zur Konvergenz von Funktionenfolgen, speziell von KVFen. 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 15.2 (Zentraler Grenzwertsatz nach Lindenberg und Lévy)**.** *$\xi_1, ..., \xi_n$ seien unabhängig und identisch verteilte Zufallsvariablen mit*

$$\mathbb{E}(\xi_i) := \mu \text{ und } \mathbb{V}(\xi_i) := \sigma^2 > 0 \text{ für alle } i = 1, ..., n. \tag{15.11}$$

Weiterhin sei ζ_n die Zufallsvariable definiert als

$$\zeta_n := \sqrt{n}\left(\frac{\bar{\xi}_n - \mu}{\sigma}\right). \tag{15.12}$$

Dann gilt für alle $z \in \mathbb{R}$

$$\lim_{n\to\infty} P_{\zeta_n}(z) = \Phi(z), \tag{15.13}$$

wobei Φ die kumulative Verteilungsfunktion der Standardnormalverteilung bezeichnet.

∘

Wir zeigen an späterer Stelle, dass damit für $n \to \infty$ auch gilt, dass

$$\sum_{i=1}^{n} \xi_i \sim N(\mu, n\sigma^2) \text{ und } \bar{\xi}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{15.14}$$

Simulation

Wir betrachten den Fall von u.i.v. χ^2-Zufallsvariablen $\xi_1, ..., \xi_n \sim \chi^2(3)$. 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 11.3) Nichtsdestotrotz resultiert ihre standardisierte Summe asymptotisch in einer Normalverteilung, wie in Abbildung 15.2 visualisiert. Dazu nutzen wir auf Ebene der Implementation die Tatsache, für die χ^2-Zufallsvariablen $\xi_i, i = 1, ..., n$ mit Freiheitsgradparameter 3 bekanntlich gilt (vgl. Kapitel 13.1 und Kapitel 13.2)

$$\mathbb{E}(\xi_i) = 3 \text{ und } \mathbb{V}(\xi_i) = 6 \tag{15.15}$$

Die Abbildungen in Abbildung 15.2 A zeigen Histogrammschätzer der Wahrscheinlichkeitsdichte von

$$\zeta_n := \sqrt{n}\left(\frac{\bar{\xi}_n - \mu}{\sigma}\right) \tag{15.16}$$

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 Realisiationen von ζ_2 der Standardnormalverteilung noch sehr unähnlich, wohingegen sich die Verteilung der Realisationen von ζ_{200} der Standardnormalverteilung schon annähert. Abbildung 15.2 B zeigt die entsprechenden geschätzten KVFen über die Theorem 15.2 formal eine Aussage trifft.

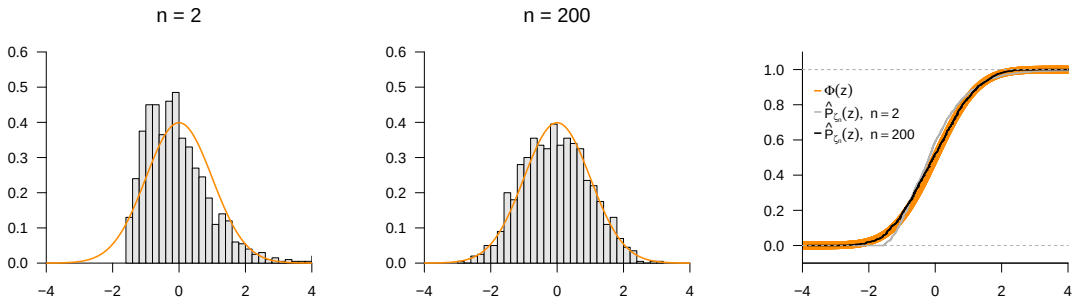


Abbildung 15.2. Simulation des Zentralen Grenzwertsatzes nach Lindenberg und Lévy.

Theorem 15.3 (Zentraler Grenzwertsatz nach Liapounov)**.** *$\xi_1, ..., \xi_n$ seien unabhängige aber nicht notwendigerweise identisch verteilten Zufallsvariablen mit*

$$\mathbb{E}(\xi_i) := \mu_i \text{ und } \mathbb{V}(\xi_i) := \sigma_i^2 > 0 \text{ für alle } i = 1, ..., n. \tag{15.17}$$

Weiterhin sollen für $\xi_1, ..., \xi_n$ folgende Eigenschaften gelten:

$$\mathbb{E}(|\xi_i - \mu_i|^3) < \infty \text{ und } \lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}\left(|\xi_i - \mu_i|^3\right)}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0. \tag{15.18}$$

Dann gilt für die Zufallsvariable ζ_n definiert als

$$\zeta_n := \frac{\sum_{i=1}^n \xi_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}, \tag{15.19}$$

für alle $z \in \mathbb{R}$, dass

$$\lim_{n\to\infty} P_{\zeta_n}(z) = \Phi(z), \tag{15.20}$$

wobei Φ KVF der Standardnormalverteilung bezeichnet.

∘

Wir zeigen an späterer Stelle, dass damit für $n \to \infty$ auch gilt, dass

$$\sum_{i=1}^n \xi_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right) \tag{15.21}$$

15.3. Literaturhinweise

Zur mathematik-geschichtlichen Genese der Zentralen Grenzwertsätze siehe z.B. Fischer (2011).

15.4. Selbstkontrollfragen

1. Definieren Sie den Begriff der Konvergenz in Wahrscheinlichkeit.
2. Definieren Sie den Begriff der Konvergenz in Verteilung.
3. Geben Sie das Schwache Gesetz der Großen Zahlen wieder.
4. Erläutern Sie den Zentralen Grenzwertsatz nach Lindenberg und Lévy.
5. Erläutern Sie den Zentralen Grenzwertsatz nach Liapunov.
6. Warum sind die Zentralen Grenzwertsätze für die probabilistische Modellbildung wichtig?

16. Transformationstheoreme

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 vogegebene Verteilung hat, wie ist dann eine Zufallsvariable υ, 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 $\upsilon := \xi^2$ 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 υ. In diesem Beispiel fällt auf, dass υ keine negativen Werte annimmt, die Verteilung von υ ordnet negativen Werten daher Wahrscheinlichkeitsdichten von 0 zu. Untenstehender **R** Code simuliert diese Überlegungen. Abbildung 16.1 zeigt das Histogramm der gewonnenen Realisierungen der hier betrachteten Zufallsvariable ξ und Abbildung 16.2 zeigt das Histogramm der quadrierten Realisierungen von ξ, also unabhängig und identisch verteilte Realisierungen von υ.

```r
# 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 \ups

# Ausgabe der ersten acht Werte der Realisierungen von \xi und von \upsilon
print(x[1:8], digits = 2)
```

```
[1]  2.934  0.131 -0.025  1.641  1.213 -0.622 -0.819  1.028
```

```r
print(y[1:8], digits = 2)
```

```
[1] 8.6057 0.0171 0.0006 2.6922 1.4719 0.3864 0.6704 1.0565
```

Grundlegend für die nachfolgenden Betrachtungen ist folgendes Theorem, das wir nicht beweisen wollen.

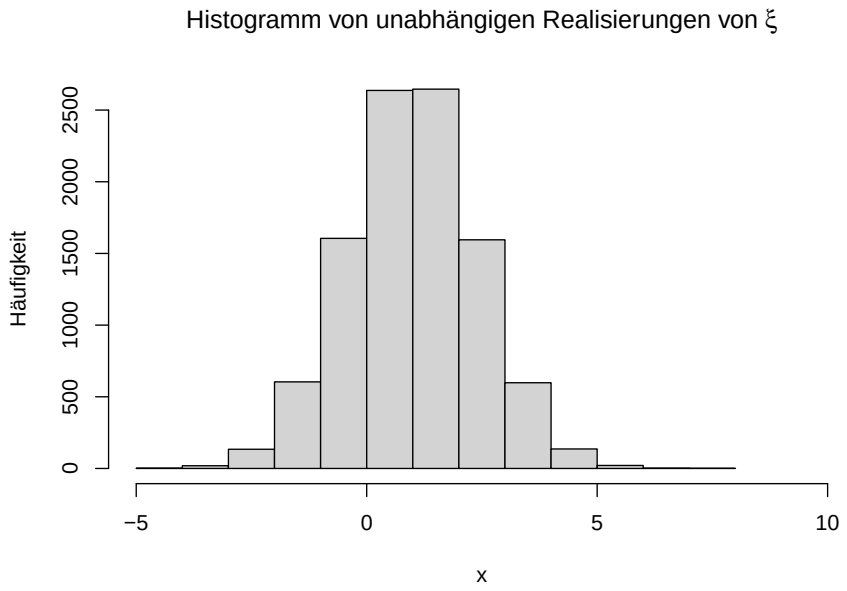


Abbildung 16.1. Histogramm von 10.000 Realisierungen unabhängig und identisch normalverteilter Zufallsvariablen.

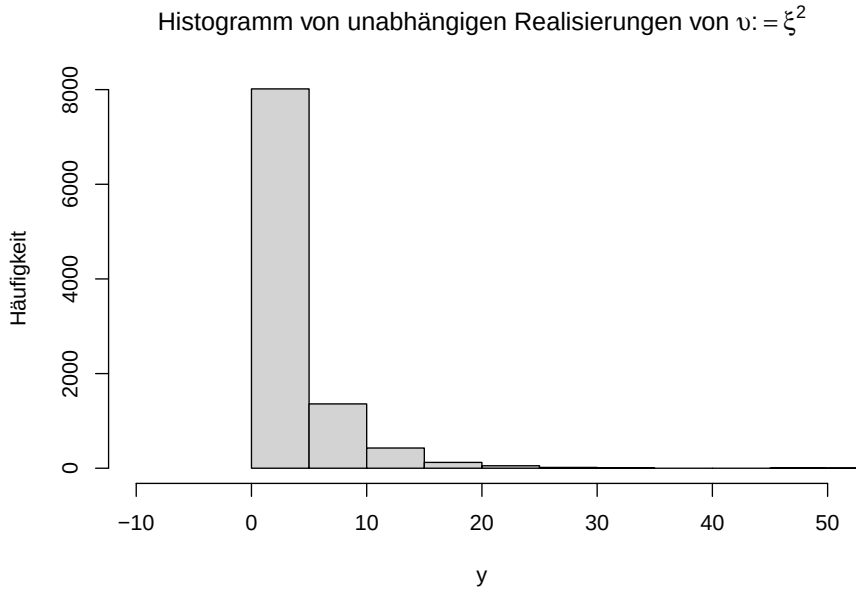


Abbildung 16.2. Histogramm von 10.000 quadrierten Realisierungen unabhängig und identisch normalverteilter Zufallsvariablen.

Theorem 16.1 (Transformation eines Zufallsvektors)**.** *$\xi : \Omega \to \mathcal{X}$ sei ein Zufallsvektor und $f : \mathcal{X} \to \mathbb{R}^m$ sei eine multivariate vektorwertige Funktion. Dann ist*

$$\upsilon : \Omega \to \mathbb{R}, \omega \mapsto \upsilon(\omega) := (f \circ \xi)(\omega) := f(\xi(\omega)) \tag{16.1}$$

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 16.1 müsste die Messbarkeit von υ als Folge der Messbarkeit von ξ nachgewiesen werden. Im Folgenden ist oft $\mathcal{X} := \mathbb{R}$ und $f : \mathbb{R} \to \mathbb{R}$. Wir schreiben in diesem Fall in der Regel einfach $\upsilon := f(\xi)$ und nennen υ die *transformierte Zufallsvariable*. In Kapitel 16.1 betrachten wir drei Theorem zur Transformation von Zufallsvariablen in Abhängigkeit von der Art der betrachteten Funktion f. In Kapitel 16.2 betrachten wir ein Theorem zur Transformation von Zufallsvektoren bei bijektiver Transformationsfunktion. In Kapitel 16.3 schließlich betrachten wir, wie der Verteilungen von Verknüpfungen von Zufallsvariablen analytisch bestimmt werden können.

16.1. Univariate Transformationstheoreme

Dabei liefert Theorem 16.2 eine Formel zur Berechnung der WDF p_υ von $\upsilon := f(\xi)$, wenn ξ eine Zufallsvariable mit WDF p_ξ ist und f eine bijektive Funktion ist. Theorem 16.3 gibt weiterhin eine vereinfachte Formel zur Berechnung der WDF p_υ von $\upsilon := f(\xi)$ an, wenn f speziell eine linear-affine Funktion ist. Theorem 16.4 schließlich gibt eine Formel zur Berechnung der WDF p_υ von $\upsilon := f(\xi)$ an, wenn f zumindest in Teilen bijektiv ist.

Theorem 16.2 (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*

$$\upsilon := f(\xi), \tag{16.2}$$

wobei die univariate reellwertige Funktion $f :]a, b[\to \mathbb{R}$ 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 Umkehrunktion 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 υ gegeben durch

$$p_\upsilon : \mathbb{R} \to \mathbb{R}_{\geq 0}, y \mapsto p_\upsilon(y) := \begin{cases} \frac{1}{|f'(f^{-1}(y))|} p_\xi\left(f^{-1}(y)\right) & \text{für } y \in f(]a, b[) \\ 0 & \text{für } y \in \mathbb{R} \setminus f(]a, b[). \end{cases} \tag{16.3}$$

∘

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

$$P_\upsilon(y) = \mathbb{P}(\upsilon \leq y) = \mathbb{P}\left(f(\xi) \leq y\right) = \mathbb{P}\left(f^{-1}(f(\xi)) \leq f^{-1}(y)\right) = \mathbb{P}\left(\xi \leq f^{-1}(y)\right) = P_\xi\left(f^{-1}(y)\right).$$

P_υ 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_υ sich ergibt wie folgt:

$$p_\upsilon(y) = \frac{d}{dy} P_\upsilon(y) = \frac{d}{dy} P_\xi\left(f^{-1}(y)\right) = p_\xi\left(f^{-1}(y)\right) \frac{d}{dy} f^{-1}(y) = \frac{1}{f'\left(f^{-1}(y)\right)} p_\xi\left(f^{-1}(y)\right),$$

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

$$P_\upsilon(y) = \mathbb{P}(f(\xi) \leq y) = \mathbb{P}\left(f^{-1}(f(\xi)) \geq f^{-1}(y)\right) = \mathbb{P}\left(\xi \geq f^{-1}(y)\right) = 1 - P_\xi\left(f^{-1}(y)\right),$$

Mit der Kettenregel und dem Satz von der Umkehrabbildung folgt dann

$$p_\upsilon(y) = \frac{d}{dy}(1 - P_\upsilon(y)) = -\frac{d}{dy} P_\xi\left(f^{-1}(y)\right) = -p_\xi\left(f^{-1}(y)\right) \frac{d}{dy} f^{-1}(y) = -\frac{1}{f'\left(f^{-1}(y)\right)} p_\xi\left(f^{-1}(y)\right).$$

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.

□

Ein wichtiger Anwendungsfall von Theorem 16.2 ist Theorem 16.3.

Theorem 16.3 (Univariates WDF Transformationstheorem bei linear-affinen Abbildungen)**.** *ξ sei eine Zufallsvariable mit WDF p_ξ und es sei*

$$\upsilon = f(\xi) \text{ mit } f(\xi) := a\xi + b \text{ für } a \neq 0. \tag{16.4}$$

Dann ist die WDF von υ gegeben durch

$$p_\upsilon : \mathbb{R} \to \mathbb{R}_{\geq 0}, y \mapsto p_\upsilon(y) := \frac{1}{|a|} p_\xi\left(\frac{y-b}{a}\right). \tag{16.5}$$

∘

Beweis. Wir halten zunächst fest, dass

$$f^{-1} : \mathbb{R} \to \mathbb{R}, y \mapsto f^{-1}(y) = \frac{y-b}{a} \tag{16.6}$$

ist, weil dann $f \circ f^{-1} = \mathrm{id}_\mathbb{R}$ gilt, wie man anhand von

$$f(f^{-1}(x)) = a\left(\frac{x-b}{a}\right) + b = x - b + b = x \text{ für alle } x \in \mathbb{R} \tag{16.7}$$

einsieht. Wir halten weiterhin fest, dass

$$f' : \mathbb{R} \to \mathbb{R}, x \mapsto f'(x) = \frac{d}{dx}(ax + b) = a. \tag{16.8}$$

Also folgt mit Theorem 16.2, dass

$$\begin{aligned} p_\upsilon : \mathbb{R} \to \mathbb{R}_{\geq 0}, y \mapsto p_\upsilon(y) &= \frac{1}{|f'\left(f^{-1}(y)\right)|} p_\xi\left(f^{-1}(y)\right) \\ &= \frac{1}{|a|} p_\xi\left(\frac{y-b}{a}\right). \end{aligned} \tag{16.9}$$

□

Ein wichtiger Anwendungsfall dieses Theorems ist die in Kapitel 17 betrachtete *Z-Transformation*. Das folgende Theorem, dass wir nicht beweisen wollen, verallgemeinert Theorem 16.2 auf den Fall nur stückweise bijektiver Abbildungen.

Theorem 16.4 (Univariate WDF Transformation bei stückweise bijektiven Abbildungen)**.** *ξ sei eine Zufallsvariable mit Ergebnisraum $\mathcal{X}$ und WDF p_ξ. Weiterhin sei*

$$\upsilon = f(\xi), \tag{16.10}$$

wobei f so beschaffen sei, dass der Ergebnisraum von ξ in eine endliche Anzahl von Mengen $\mathcal{X}_1, ..., \mathcal{X}_k$ mit einer entsprechenden Anzahl von Mengen $\mathcal{Y}_1 := f(\mathcal{X}_1), ..., \mathcal{Y}_k := f(\mathcal{X}_k)$ im Ergebnisraum $\mathcal{Y}$ von υ 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 $\mathcal{X}_1, ..., \mathcal{X}_k$ bijektiv ist (d.h. f ist eine stückweise bijektive Abbildung). Für $i = 1, ..., k$ bezeichne f_i^{-1} die Umkehrfunktion von f auf $\mathcal{Y}_i$. Schließlich nehmen wir an, dass die Ableitungen f_i' für alle $i = 1, ..., k$ existieren und stetig sind. Dann ist eine WDF von υ durch

$$p_\upsilon : \mathcal{Y} \to \mathbb{R}_{\geq 0}, y \mapsto p_\upsilon(y) := \sum_{i=1}^{k} 1_{\mathcal{Y}_i}(y) \frac{1}{|f_i'(f_i^{-1}(y))|} p_\xi\left(f_i^{-1}(y)\right). \tag{16.11}$$

gegeben.

∘

Ein wichtiger Anwendungsfall ist die in Kapitel 17 betrachtete χ^2*-Transformation.*

16.2. Multivariate WDF Transformationstheoreme

Theorem 16.5 liefert eine Formel zur Berechnung der WDF p_υ von $\upsilon := f(\xi)$, wenn ξ ein Zufallsvektor mit WDF p_ξ ist und f eine bijektive multivariate vektorwertigeFunktion ist. Es handelt sich dabei um eine direkte Generalisierung von Theorem 16.2 und wir verzichten auf einen Beweis.

Theorem 16.5 (Multivariate WDF Transformation bei bijektiven Abbildungen)**.** *ξ sei ein n-dimensionaler Zufallsvektor mit Ergebnisraum $\mathbb{R}^n$ und WDF p_ξ. Weiterhin sei*

$$\upsilon := f(\xi), \tag{16.12}$$

wobei die multivariate vektorwertige Funktion $f : \mathbb{R}^n \to \mathbb{R}^n$ differenzierbar und bijektiv auf $]a, b[$ sei. Schließlich seien

$$J^f(x) = \left(\frac{\partial}{\partial x_j} f_i(x)\right)_{1 \leq i \leq n, 1 \leq j \leq n} \in \mathbb{R}^{n \times n} \tag{16.13}$$

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 υ durch

$$p_\upsilon : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, y \mapsto p_\upsilon(y) := \begin{cases} \frac{1}{|J^f(f^{-1}(y))|} p_\xi\left(f^{-1}(y)\right) & \text{for } y \in f(\mathbb{R}^n) \\ 0 & \text{for } y \in \mathbb{R}^n \setminus f(\mathbb{R}^n) \end{cases} \tag{16.14}$$

gegeben.

∘

Wichtige Anwendungsfälle sind die in Kapitel 17 betrachteten T- und F*-Transformationen.*

16.3. Operationstheoreme

Das folgende sogenannte *Konvolutionstheorem* liefert eine Formel zur Berechnung der WDF p_υ von $\upsilon := \xi_1 + \xi_2$, wenn ξ_1 und ξ_2 zwei Zufallsvariablen mit WDFen p_{ξ_1} und p_{ξ_2} sind.

Theorem 16.6 (Summe unabhängiger Zufallsvariablen (Konvolution))**.** *ξ_1 und ξ_2 seien zwei kontinuierliche unabhängige Zufallsvariablen mit WDF p_{ξ_1} und p_{ξ_2}, respektive. $\upsilon := \xi_1 + \xi_2$ sei die Summe von ξ_1 und ξ_2. Dann ergibt sich eine WDF der Verteilung von υ als*

$$p_\upsilon(y) = \int_{-\infty}^{\infty} p_{\xi_1}(y - x_2) p_{\xi_2}(x_2)\, dx_2 = \int_{-\infty}^{\infty} p_{\xi_1}(x_1) p_{\xi_2}(y - x_1)\, dx_1 \tag{16.15}$$

Die Formel für die WDF p_υ 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

$$f : \mathbb{R}^2 \to \mathbb{R}^2, x \mapsto f(x) := \begin{pmatrix} x_1 + x_2 \\ x_2 \end{pmatrix} := \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \tag{16.16}$$

Die inverse Funktion von f ist dann gegeben durch

$$f : \mathbb{R}^2 \to \mathbb{R}^2, z \mapsto f(z) := \begin{pmatrix} z_1 - x_2 \\ z_2 \end{pmatrix} \tag{16.17}$$

weil dann $f \circ f^{-1} = \mathrm{id}_{\mathbb{R}^2}$ gilt, wie man anhand von

$$f^{-1}\left(f(x)\right) = f^{-1}\begin{pmatrix} x_1 + x_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 - x_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{16.18}$$

einsieht. Die Jacobimatrix von f ergibt sich zu

$$J^f(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} f_1(x) & \frac{\partial}{\partial x_2} f_1(x) \\ \frac{\partial}{\partial x_1} f_2(x) & \frac{\partial}{\partial x_2} f_2(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial x_1}(x_1 + x_2) & \frac{\partial}{\partial x_2}(x_1 + x_2) \\ \frac{\partial}{\partial x_1} x_2 & \frac{\partial}{\partial x_2} x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{16.19}$$

und die Jacobideterminante damit zu $|J^f(x)| = 1$. Wir halten weiterhin fest, dass die Unabhängigkeit von ξ_1 und ξ_2 impliziert, dass

$$p_{\xi_1,\xi_2}(x_1, x_2) = p_{\xi_1}(x_1) p_{\xi_2}(x_2) \tag{16.20}$$

impliziert. Einsetzen und Integration hinsichtlich x_2 ergibt dann ergibt dann für $z \in f(\mathbb{R}^2)$

$$\begin{aligned} p_\zeta(z) &= \frac{1}{|J^f\left(f^{-1}(z)\right)|} p_\xi\left(f^{-1}(z)\right) \\ &= \frac{1}{1} p_{\xi_1,\xi_2}\left(z_1 - x_2, x_2\right) \\ &= p_{\xi_1}(z_1 - x_2) p_{\xi_2}(x_2) \end{aligned} \tag{16.21}$$

Integration über x_2 ergibt dann eine WDF für die marginale Verteilung von ζ_1

$$p_{\zeta_1}(z_1) = \int_{-\infty}^{\infty} p_{\xi_1}(z_1 - x_2) p_{\xi_2}(x_2)\, dx_2 \tag{16.22}$$

Mit $\zeta_1 = \xi_1 + \xi_2 = \upsilon$ ergibt sich dann die erste Form des Konvlutionstheorems zu

$$p_\upsilon(y) = \int_{-\infty}^{\infty} p_{\xi_1}(y - x_2) p_{\xi_2}(x_2)\, dx_2. \tag{16.23}$$

□

Wichtige in Kapitel 17 betrachtete Anwendungsfällte sind die *Summentransformation* und die *Mittelwerttransformation*.

17. Transformationen der Normalverteilung

In diesem Kapitel diskutieren wir sechs Transformationen normalverteilter Zufallsvariablen, die in der Frequentistischen Inferenz zentrale Rollen spielen und bei denen es sich um Anwendungen von den in Kapitel 16 eingeführten Transformationtheore handelt. Unsere Aussagen sind dabei von der allgemeinen Form "Wenn $\xi_i, i = 1, ..., n$ unabhängig und identisch normalverteilte Zufallsvariablen sind und $\upsilon := f(\xi_1, ..., \xi_n)$ eine Transformation dieser Zufallsvariablen ist, dann ist die WDF von υ durch die Formel $p_\upsilon := \{\text{Formel}\}$ gegeben und man nennt die Verteilung von υ *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 Kapitel 18 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.

17.1. Summentransformation und Mittelwertstransformation

In diesem Abschnitt betrachten wir die resultierenden Verteilung bei Summation und Mittelwertbildung von unabhängig und identisch normalverteilten Zufallsvariablen. Speziell besagt das Theorem 17.1 besagt, dass die Summe unabhängig normalverteilter Zufallsvariablen wiederum normalverteilt ist und gibt die Parameter dieser Verteilung an, während Theorem 17.2 besagt, dass das Stichprobenmittel unabhängig normalverteilter Zufallsvariablen wiederum normalverteilt ist und gibt die Parameter dieser Verteilung an.

Theorem 17.1 (Summationstransformation)**.** *Für $i = 1, ..., n$ seien $\xi_i \sim N(\mu_i, \sigma_i^2)$ unabhängige normalverteilte Zufallsvariablen. Dann gilt für die Summe $\upsilon := \sum_{i=1}^n \xi_i$, dass*

$$\upsilon \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right) \tag{17.1}$$

Für unabhängige und identisch normalverteilte Zufallsvariablen $\xi_i \sim N(\mu, \sigma^2)$ gilt folglich

$$\upsilon \sim N(n\mu, n\sigma^2). \tag{17.2}$$

∘

Beweis. Wir skizzieren mithilfe von Theorem 16.6, dass für $\xi_1 \sim N(\mu_1, \sigma_1^2)$, $\xi_2 \sim N(\mu_2, \sigma_2^2)$, und $\upsilon := \xi_1 + \xi_2$ gilt, dass $\upsilon \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Für $n > 2$ folgt das Theorem dann durch Iteration. Mit der Definition der WDF der Normalverteilung erhalten wir zunächst

$$\begin{aligned} p_\upsilon(y) &= \int_{-\infty}^{\infty} p_{\xi_1}(x_1) p_{\xi_2}(y - x_1)\, dx_1 \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{y - x_1 - \mu_2}{\sigma_2}\right)^2\right) dx_1 \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - \frac{1}{2}\left(\frac{y - x_1 - \mu_2}{\sigma_2}\right)^2\right) dx_1. \end{aligned} \tag{17.3}$$

Mit einigem algebraischen Aufwand erhält man die Identität

$$-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - \frac{1}{2}\left(\frac{y - x_1 - \mu_2}{\sigma_2}\right)^2 = -\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)} - \frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2\sigma_1^2 - \mu_1\sigma_2^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)}, \tag{17.4}$$

so dass weiterhin gilt, dass

$$\begin{aligned} p_\upsilon(y) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)} - \frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2\sigma_1^2 - \mu_1\sigma_2^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)}\right) dx_1 \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \exp\left(-\frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2\sigma_1^2 - \mu_1\sigma_2^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)}\right) dx_1 \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2\sigma_1^2 - \mu_1\sigma_2^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)}\right) dx_1. \end{aligned} \tag{17.5}$$

Für das verbleibende Integral zeigt man mithilfe der Integration durch Substitution, dass

$$\int_{-\infty}^{\infty} \exp\left(-\frac{((\sigma_1^2 + \sigma_2^2)x_1 - \sigma_1^2 y + \mu_2\sigma_1^2 - \mu_1\sigma_2^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2)}\right) dx_1 = \frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \tag{17.6}$$

Es ergibt sich also

$$\begin{aligned} p_\upsilon(y) &= \frac{1}{2\pi\sigma_1\sigma_2} \frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \\ &= \frac{(2\pi)^{-1}(2\pi)^2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(y - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right). \end{aligned} \tag{17.7}$$

Schließlich folgt, dass

$$p_\upsilon(y) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left(-\frac{1}{2(\sigma_1^2 + \sigma_2^2)}\left(y - (\mu_1 + \mu_2)\right)^2\right) = N(y; \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \tag{17.8}$$

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.

□

Ein wichtiger Anwendungsfall von Theorem 17.1 ist das nachfolgende Theorem 17.2 sowie die in Gleichung 15.14 und Gleichung 15.21 erwähnten Generalisierungen der Zentralen Grenzwertsätze. Wir visualisieren Theorem 17.1 exemplarisch in Abbildung 17.1.

Theorem 17.2 (Mittelwertstransformation)**.** *Für $i = 1, ..., n$ 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*

$$\bar{\xi}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{17.9}$$

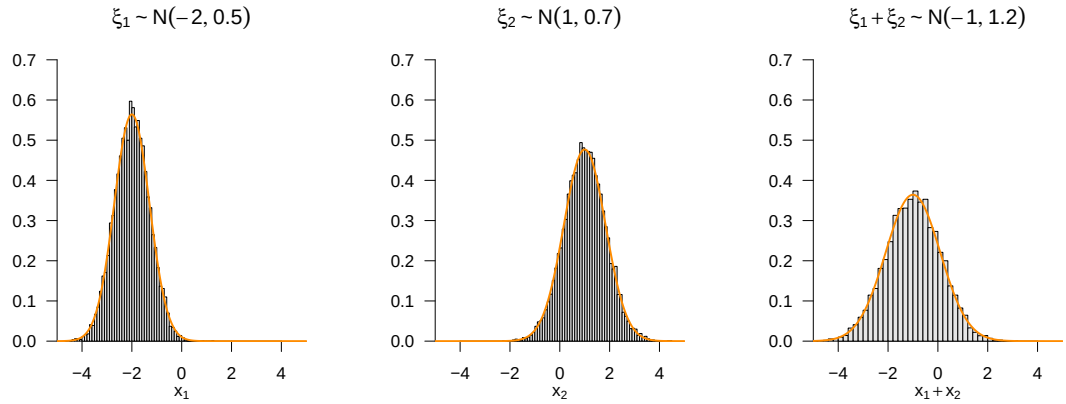


Abbildung 17.1. Summation normalverteilter Zufallsvariablen.

∘

Beweis. Wir halten zunächst fest, dass mit dem Theorem zur Summe von unabhängig normalverteilten Zufallsvariablen gilt, dass $\bar{\xi}_n = \frac{1}{n}\upsilon$ mit $\upsilon := \sum_{i=1}^{n} \xi_i \sim N(n\mu, n\sigma^2)$. Einsetzen in Theorem 16.3 ergibt dann

$$
\begin{aligned}
p_{\bar{\xi}_n}(\bar{x}_n) &= \frac{1}{|1/n|} N\left(n\bar{x}_n; n\mu, n\sigma^2\right) \\
&= \frac{n}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{1}{2n\sigma^2}\left(n\bar{x}_n - n\mu\right)^2\right) \\
&= \frac{n}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{1}{2n\sigma^2}\left(n\bar{x}_n - n\mu\right)^2\right) \\
&= nn^{-\frac{1}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(n\bar{x}_n)^2}{2n\sigma^2} + \frac{2(n\bar{x}_n)(n\mu)}{2n\sigma^2} - \frac{(n\mu)^2}{2n\sigma^2}\right) \\
&= \sqrt{n}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{n\bar{x}_n^2}{2\sigma^2} + \frac{2n\bar{x}_n\mu}{2\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right) \\
&= \frac{1}{1/\sqrt{n}}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\bar{x}_n^2}{2(\sigma^2/n)} + \frac{2\bar{x}_n\mu}{2(\sigma^2/n)} - \frac{\mu^2}{2(\sigma^2/n)}\right) \\
&= \frac{1}{\sqrt{2\pi(\sigma^2/n)}} \exp\left(-\frac{1}{2(\sigma^2/n)}(\bar{x}_n - \mu)^2\right) \\
&= N\left(\bar{x}_n; \mu, \sigma^2/n\right)
\end{aligned}
\tag{17.10}
$$

□

Wichtige Anwendungsfälle von Theorem 17.2 sind die Analyse von Erwartungswertschätzern in **?@sec-punktschätzung** sowie die sowie die in Gleichung 15.14 erwähnte Generalisierung des Zentralen Grenzwertsatzes nach Lindenberg-Lévy. Wir visualisieren Theorem 17.2 exemplarisch in Abbildung 17.2.

17.2. Z-Transformation

Das Theorem 17.3 besagt, dass Subtraktion des Erwartungswertparameters und gleichzeitige Division mit der Wurzel des Varianzsparameters die Verteilung einer normalverteilten Zufallsvariable in eine Standardnormalverteilung transformiert.

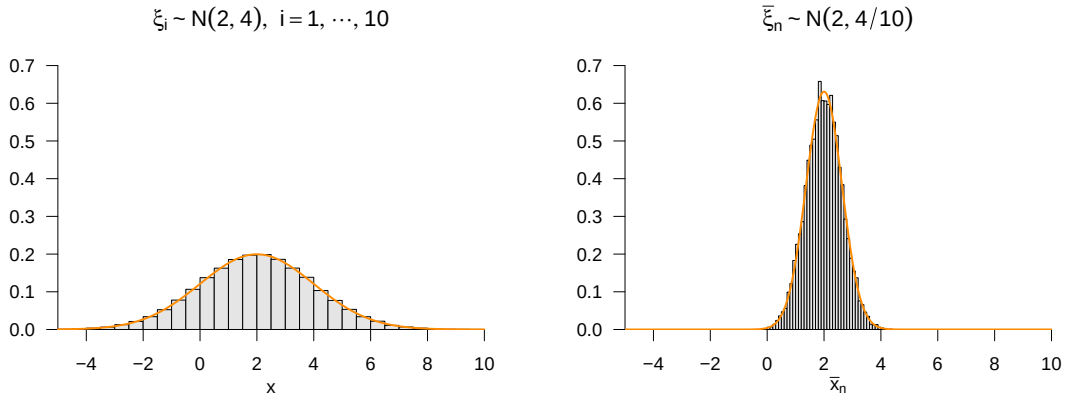


Abbildung 17.2. Mittelwertbildung bei normalverteilten Zufallsvariablen.

Theorem 17.3 (Z-Transformation)**.** *Es sei* $\upsilon \sim N(\mu, \sigma^2)$ *eine normalverteilte Zufallsvariable. Dann ist die Zufallsvariable*

$$Z := \frac{\upsilon - \mu}{\sigma} \tag{17.11}$$

eine standardnormalverteilte Zufallsvariable, es gilt also $Z \sim N(0, 1)$*.*

∘

Beweis. Wir nutzen Theorem 16.3. Dazu halten wir zunächst fest, dass die Z-Transformation einer Funktion der Form

$$f(\upsilon) := \frac{\upsilon - \mu}{\sigma} =: Z \tag{17.12}$$

entspricht. Wir stellen weiterhin fest, dass die Umkehrfunktion von f durch

$$f^{-1}(Z) := \sigma Z + \mu \tag{17.13}$$

gegeben ist, da für alle $z \in \mathbb{R}$ mit $z = y - \mu\sigma$ gilt, dass

$$\zeta^{-1}(z) = \zeta^{-1}\left(\frac{y-\mu}{\sigma}\right) = \frac{\sigma(y-\mu)}{\sigma} + \mu = y - \mu + \mu = y. \tag{17.14}$$

Schließlich stellen wir fest, dass für die Ableitung f' von f gilt, dass

$$f'(y) = \frac{d}{dy}\left(\frac{y-\mu}{\sigma}\right) = \frac{d}{dy}\left(\frac{y}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}. \tag{17.15}$$

Einsetzen in das univariate WDF Transformationstheorem für lineare Funktionen ergibt dann

$$\begin{aligned} p_Z(z) &= \frac{1}{|1/\sigma|} N\left(\sigma z + \mu; \mu, \sigma^2\right) \\ &= \frac{1}{1/\sqrt{\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sigma z + \mu - \mu\right)^2\right) \\ &= \frac{\sqrt{\sigma^2}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\sigma^2 z^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \\ &= N(z; 0, 1) \end{aligned} \tag{17.16}$$

also, dass $Z \sim N(0, 1)$.

□

Wichtige Anwendungsfälle von Theorem 17.3 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 17.3 exemplarisch in Abbildung 17.3.

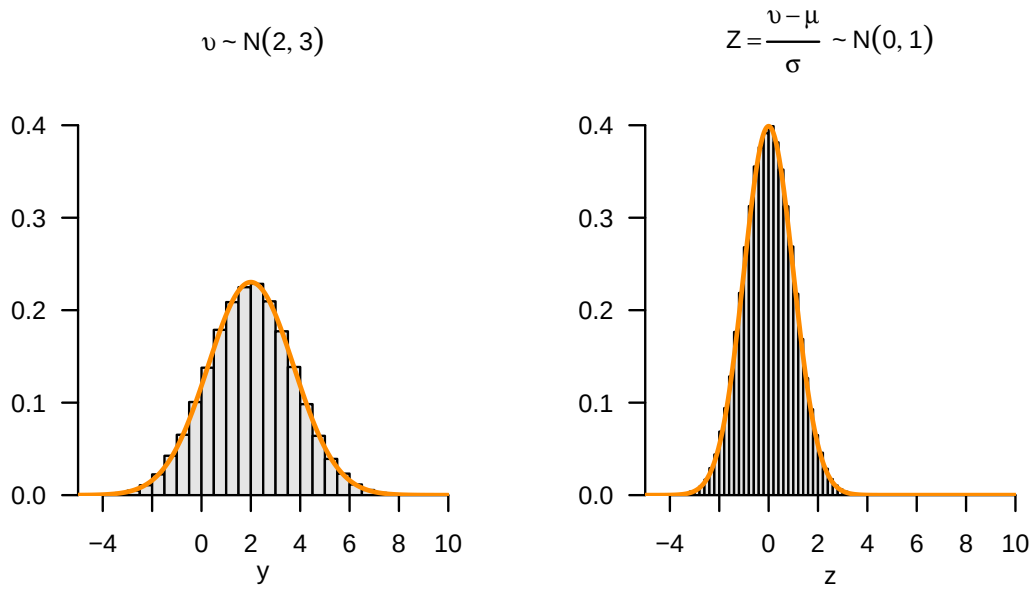


Abbildung 17.3. Z-Transformation normalverteilter Zufallsvariablen.

17.3. χ^2-Transformation

Mit der χ^2-Transformation führen wir nun eine erste Transformation unabhängig und identisch normalverteilter Zufallsvariablen ein, die *nicht* wiederrum auf eine Normalverteilung führt. Speziell besagt Theorem 17.4, 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 11 betrachteten Gammazufallsvariablen (vgl. Definition 11.10).

Definition 17.1 (χ^2-Zufallsvariable)**.** U sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}_{>0}$ und WDF

$$p : \mathbb{R}_{>0} \to \mathbb{R}_{>0}, u \mapsto p(u) := \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} u^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}u\right), \tag{17.17}$$

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

$$\chi^2(u; n) := \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} u^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}u\right). \tag{17.18}$$

•

Wir erinnern daran, dass die WDF der χ^2-Verteilung der WDF $G\left(u; \frac{n}{2}, 2\right)$ einer Gammaverteilung entspricht. In Abbildung 17.4 visualisieren wir exemplarisch einige WDFen von χ^2-Zufallsvariablen. Wir beobachten, dass mit ansteigendem n sich $\chi^2(u; n)$ verbreiter und Wahrscheinlichkeitsmasse zur größeren Werten von u verschoben wird.

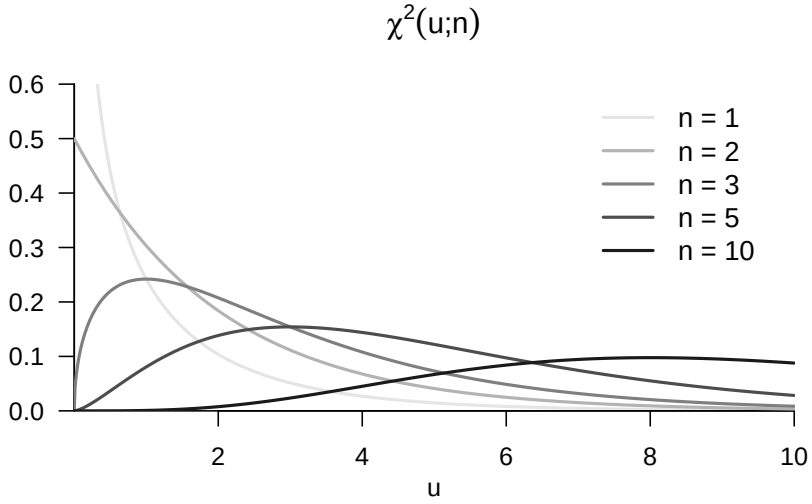


Abbildung 17.4. WDFen von χ^2 Zufallsvariablen.

Theorem 17.4 (χ^2-Transformation)**.** *$Z_1, ..., Z_n \sim N(0,1)$ seien unabhängig und identisch standardnormalverteilte Zufallsvariablen. Dann ist die Zufallsvariable*

$$U := \sum_{i=1}^{n} Z_i^2 \tag{17.19}$$

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 16.4. Danach ist die WDF einer Zufallsvariable $U := f(Z)$, welche aus der Transformation einer Zufallsvariable Z mit WDF p_ζ durch eine stückweise bijektive Abbildung hervorgeht, gegeben durch

$$p_U(u) = \sum_{i=1}^{k} 1_{\mathcal{U}_i} \frac{1}{|f_i'(f_i^{-1}(u))|} p_\zeta\left(f_i^{-1}(u)\right). \tag{17.20}$$

Wir definieren

$$\mathcal{U}_1 :=]-\infty, 0[, \mathcal{U}_2 :=]0, \infty[, \text{ und } \mathcal{U}_i := \mathbb{R}_{>0} \text{ für } i = 1, 2, \tag{17.21}$$

sowie

$$f_i : \mathcal{Z}_i \to \mathcal{U}_i, x \mapsto f_i(z) := z^2 =: u \text{ für } i = 1, 2. \tag{17.22}$$

Die Ableitung und die Umkehrfunktion der f_i ergeben sich zu

$$f_i' : \mathcal{Z}_i \to \mathcal{Z}_i, x \mapsto f_i'(z) = 2z \text{ für } i = 1, 2, \tag{17.23}$$

und

$$f_1^{-1} : \mathcal{U}_1 \to \mathcal{U}_1, u \mapsto f_1^{-1}(u) = -\sqrt{u} \text{ und } f_2^{-1} : \mathcal{U}_2 \to \mathcal{U}_2, u \mapsto f_2^{-1}(u) = \sqrt{u}, \tag{17.24}$$

respektive. Einsetzen in Gleichung (17.20) ergibt dann

$$\begin{aligned} p_U(u) &= 1_{\mathcal{U}_1}(u)\frac{1}{|f_1'(f_1^{-1}(u))|}p_\zeta\left(f_1^{-1}(u)\right) + 1_{\mathcal{U}_2}(u)\frac{1}{|f_2'(f_2^{-1}(u))|}p_\zeta\left(f_2^{-1}(u)\right) \\ &= \frac{1}{|2(-\sqrt{u})|}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(-\sqrt{u})^2\right) + \frac{1}{|2(\sqrt{u})|}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\sqrt{u})^2\right) \\ &= \frac{1}{2\sqrt{u}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u\right) + \frac{1}{2\sqrt{u}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u\right) \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{u}}\exp\left(-\frac{1}{2}u\right). \end{aligned} \tag{17.25}$$

Andererseits gilt, dass mit $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, die PDF einer χ^2-Zufallsvariable U mit $n = 1$ durch

$$\frac{1}{\Gamma\left(\frac{1}{2}\right)2^{\frac{1}{2}}}u^{\frac{1}{2}-1}\exp\left(-\frac{1}{2}u\right) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{u}}\exp\left(-\frac{1}{2}u\right) \tag{17.26}$$

gegeben ist. Also gilt, dass wenn $Z \sim N(0,1)$ ist, dann ist $U := Z^2 \sim \chi^2(1)$.

□

Wichtige Anwendungsfälle sind die U-Konfidenzintervallstatistik sowie die im folgenden eingeführten t- und f-Zufallsvariablen. Wir visualisieren Theorem 17.4 exemplarisch in Abbildung 17.5.

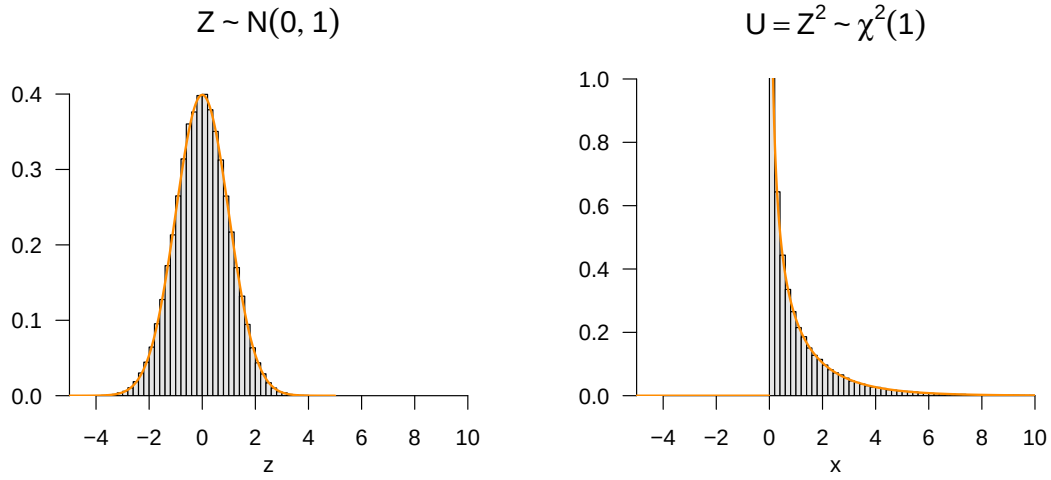


Abbildung 17.5. χ^2-Transformation normalverteilter Zufallsvariablen.

17.4. 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 17.5 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 17.2 (t-Zufallsvariable)**.** T sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \to \mathbb{R}_{>0}, t \mapsto p(t) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \tag{17.27}$$

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

$$T(t; n) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}. \tag{17.28}$$

•

In Abbildung 17.6 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 $T(t; n) \approx N(0, 1)$.

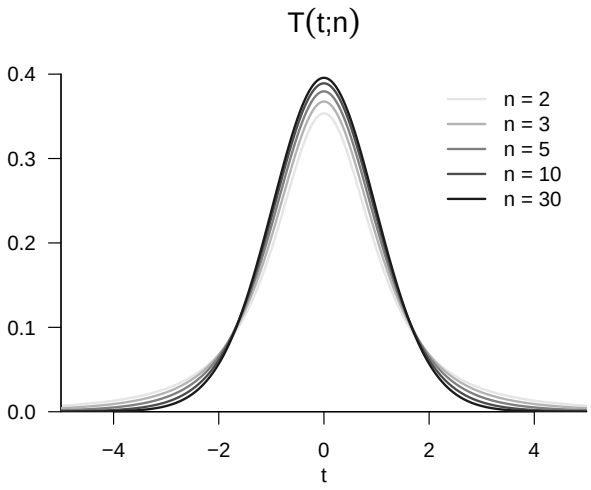


Abbildung 17.6. WDFen von T-Zufallsvariablen.

Theorem 17.5 (T-Transformation)**.** *$Z \sim N(0, 1)$ sei eine standarnormalverteilte Zufallsvariable, $U \sim \chi^2(n)$ sei eine χ^2-Zufallsvariable mit Freiheitsgradparameter n, und Z und U seien unabhängig. Dann ist die Zufallsvariable*

$$T := \frac{Z}{\sqrt{U/n}} \tag{17.29}$$

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

$$p_{Z,U}(z,u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} u^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}u\right). \tag{17.30}$$

gegeben ist. Wir betrachten dann die multivariate vektorwertige Abbildung

$$f : \mathbb{R}^2 \to \mathbb{R}^2, (z,u) \mapsto f(z,u) := \left(\frac{z}{\sqrt{u/n}}, u\right) =: (t,w) \tag{17.31}$$

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 $\upsilon := f(\xi)$ für eine differenzierbare und bijektive Abbildung $f : \mathbb{R}^n \to \mathbb{R}^n$ ist, die WDF des Zufallsvektors υ durch

$$p_\upsilon : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, y \mapsto p_\upsilon(y) := \frac{1}{|J^f(f^{-1}(y))|} p_\xi\left(f^{-1}(y)\right) \tag{17.32}$$

gegeben ist. Für die im vorliegenden Fall betrachtete Abbildung halten wir zunächst fest, dass

$$f^{-1} : \mathbb{R}^2 \to \mathbb{R}^2, (t,w) \mapsto f^{-1}(t,w) := \left(\sqrt{w/n}t, w\right). \tag{17.33}$$

Dies ergibt sich direkt aus

$$f^{-1}(f(z,u)) = f^{-1}\left(\frac{z}{\sqrt{u/n}}, u\right) = \left(\frac{\sqrt{u/n}z}{\sqrt{u/n}}, u\right) = (z,u) \text{ für alle } (z,u) \in \mathbb{R}^2. \tag{17.34}$$

Wir halten dann fest, dass die Determinante der Jacobi-Matrix von f an der Stelle (z,u) durch

$$|J^f(z,u)| = \begin{vmatrix} \frac{\partial}{\partial z}\left(\frac{z}{\sqrt{u/n}}\right) & \frac{\partial}{\partial u}\left(\frac{z}{\sqrt{u/n}}\right) \\ \frac{\partial}{\partial z}u & \frac{\partial}{\partial u}u \end{vmatrix} = \left(\frac{\upsilon}{n}\right)^{-1/2}, \tag{17.35}$$

gegeben ist, sodass folgt, dass

$$\frac{1}{|J^f(f^{-1}(z,u))|} = \left(\frac{w}{n}\right)^{1/2}. \tag{17.36}$$

Einsetzen in Gleichung (17.32) ergibt dann

$$p_{T,W}(t,w) = \left(\frac{w}{n}\right)^{1/2} p_{Z,V}\left(\sqrt{w/n}t, w\right), \tag{17.37}$$

Es folgt also

$$\begin{aligned}
p_T(t) &= \int_0^\infty p_{T,W}(t,w)\,dw \\
&= \int_0^\infty \left(\frac{w}{n}\right)^{1/2} p_{Z,V}\left(\sqrt{w/n}t, w\right)\,dw \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{w/n}t)^2\right) \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} w^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}w\right) \left(\frac{w}{n}\right)^{1/2} dw \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\frac{w}{n}t^2\right) w^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}w\right) w^{1/2}\,dw \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\frac{w}{n}t^2 - \frac{1}{2}w\right) w^{\frac{n}{2}-1} w^{\frac{1}{2}}\,dw \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{w}{n}t^2 + w\right)\right) w^{\frac{n+1}{2}-1}\,dw \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\left(1 + \frac{t^2}{n}\right)\right) w^{\frac{n+1}{2}-1}\,dw
\end{aligned} \tag{17.38}$$

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:

$$\begin{aligned}\Gamma(w; \alpha, \beta) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} w^{\alpha-1} \exp\left(-\frac{w}{\beta}\right) \\ \Rightarrow \Gamma\left(w; \frac{n+1}{2}, \frac{2}{1+\frac{t^2}{n}}\right) &= \frac{1}{\Gamma(\frac{n+1}{2})\left(\frac{2}{1+\frac{t^2}{n}}\right)^{\frac{n+1}{2}}} w^{\frac{n+1}{2}-1} \exp\left(-\frac{w}{\frac{2}{1+\frac{t^2}{n}}}\right) \\ &= \frac{1}{\Gamma(\frac{n+1}{2})\left(\frac{2}{1+\frac{t^2}{n}}\right)^{\frac{n+1}{2}}} \exp\left(-\frac{1}{2}\left(1+\frac{t^2}{n}\right)\right) w^{\frac{n+1}{2}-1}.\end{aligned}$$

Es ergibt sich also

$$p_T(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}} n^{\frac{1}{2}}} \int_0^\infty \Gamma\left(w; \frac{n+1}{2}, \frac{2}{1+\frac{t^2}{n}}\right) dw. \tag{17.39}$$

Schließlich stellen wir fest, dass der Integralterm in obiger Gleichung dem Normalisierungsterm einer Gamma WDF entspricht. Abschließend ergibt sich also

$$p_T(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}} n^{\frac{1}{2}}} \Gamma\left(\frac{n+1}{2}\right) \left(\frac{2}{1+\frac{t^2}{n}}\right)^{\frac{n+1}{2}}. \tag{17.40}$$

Die Verteilung von $Z/\sqrt{U/n}$ hat also die WDF einer t-Zufallsvariable.

□

Wichtige Anwendungsfälle sind die T-Konfidenzintervallstatistik sowie die T-Teststatistiken der Theorie von Hypothesentests im Kontext des Allgemeinen Linearen Modells. Wir visualisieren Theorem 17.5 exemplarisch in Abbildung 17.7.

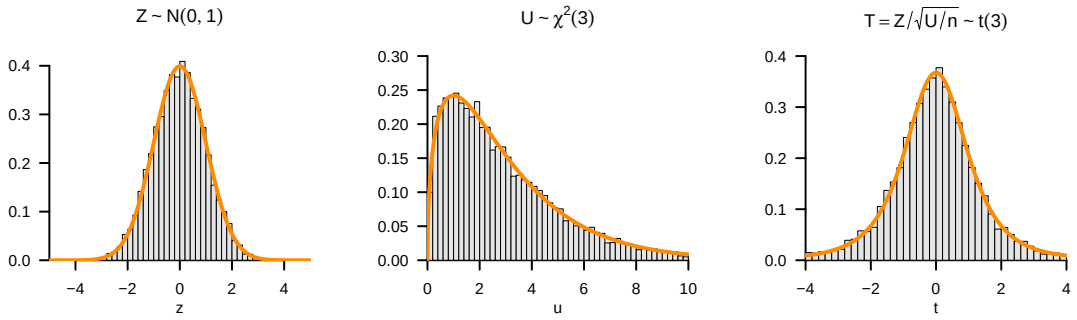


Abbildung 17.7. T-Transformation normalverteilter Zufallsvariablen.

17.5. Nichtzentrale T-Transformation

In diesem Abschnitt betrachten wir den Fall, dass der Erwartungswertparameter der Zählervariable der in Theorem 17.5 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)).

17.6. Nichtzentrale t-Zufallsvariable

T sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \to \mathbb{R}_{>0}, t \mapsto p(t) := \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n}{2}\right)(n\pi)^{\frac{1}{2}}} \times \int_0^\infty \tau^{\frac{n-1}{2}} \exp\left(-\frac{\tau}{2}\right) \exp\left(-\frac{1}{2}\left(t\left(\frac{\tau}{n}\right)^{\frac{1}{2}} - \delta\right)^2\right) d\tau. \quad (17.41)$$

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 enstpricht, es gelten also

$$t(T; 0, n) = t(T; n) \quad (17.42)$$

sowie

$$\Psi(T; 0, n) = \Psi(T; n) \text{ und } \Psi^{-1}(T; 0, n) = \Psi^{-1}(T; n). \quad (17.43)$$

In Abbildung 17.8 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.

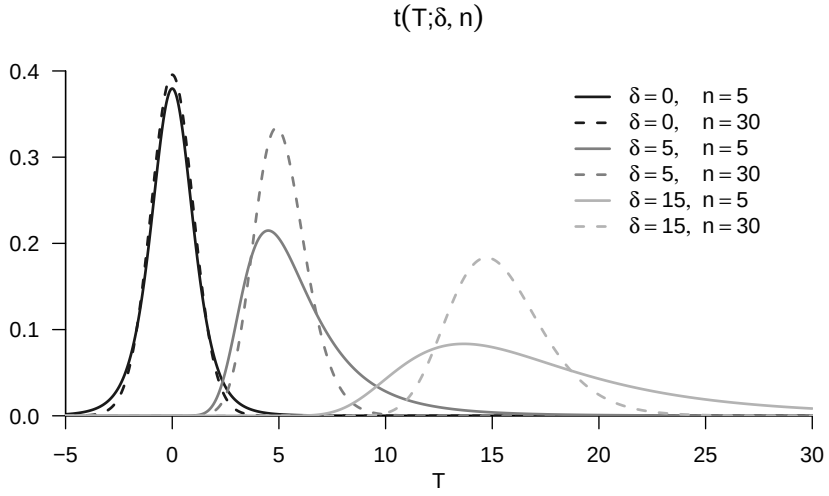


Abbildung 17.8. WDFen von Nichtzentralen-T-Zufallsvariablen.

Eine nichtzentrale t-Zufallsvariable ist das Resultat einer nichtzentralen T-Transformation, wie folgendes Theorem besagt.

Theorem 17.6 (Nichtzentrale T-Transformation)**.** *$\upsilon \sim N(\mu, 1)$ sei eine normalverteilte Zufallsvariable, $U \sim \chi^2(n)$ sei eine χ^2 Zufallsvariable mit Freiheitsgradparameter n, und υ und U seien unabhängige Zufallsvariablen. Dann ist die Zufallsvariable*

$$T := \frac{\upsilon}{\sqrt{U/n}} \tag{17.44}$$

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 17.6 exemplarisch in Abbildung 17.9.

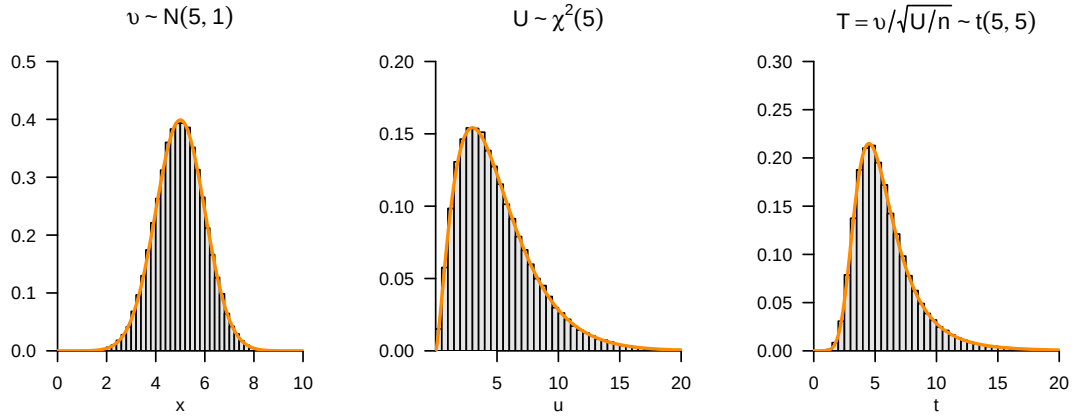


Abbildung 17.9. Nichtzentrale T-Transformation normalverteilter Zufallsvariablen.

17.7. F-Transformation

Das in diesem Abschnitt zentrale Theorem 17.7 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.

Definition 17.3 (f-Zufallsvariable)**.** F sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}_{>0}$ und WDF

$$p_F : \mathbb{R} \to \mathbb{R}_{>0}, f \mapsto p_F(f) := m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{f^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}f\right)^{\frac{m+n}{2}}}, \tag{17.45}$$

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

$$F(f; n, m) := m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{f^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n}f\right)^{\frac{m+n}{2}}}. \tag{17.46}$$

•

In Abbildung 17.10 visualisieren wir exemplarisch einige WDFen von f-Zufallsvariablen. Wir beobachten, dass die Form der WDFen zunächt primär durch den Freiheitsgradparameter n und dann sekundär durch den Freiheitsgradparameter m bestimmt werden.

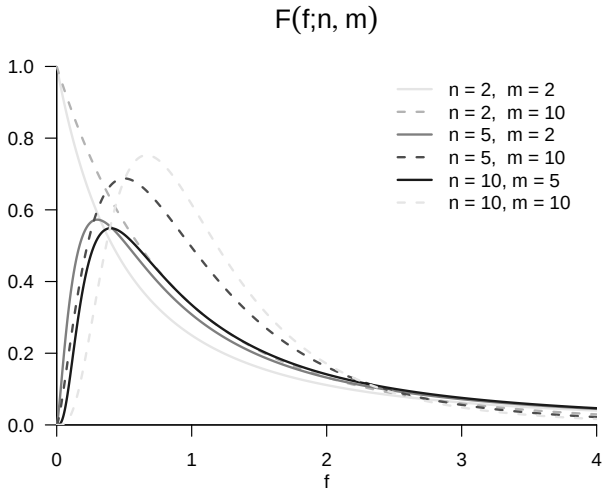


Abbildung 17.10. WDFen von f-verteilten Zufallsvariablen.

Theorem 17.7 (F-Transformation)**.** *$V \sim \chi^2(n)$ und $W \sim \chi^2(m)$ seien zwei unabhängige χ^2-Zufallfsvariablen mit Freiheitsgradparametern n und m, respektive. Dann ist die Zufallsvariable*

$$F := \frac{V/n}{W/m} \tag{17.47}$$

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 16.1 und Marginalisierung herleitet und dieses Theorem dann auf die WDF von χ^2-verteilten Zufallsvariablen anwendet. Wir visualisieren Theorem 17.7 exemplarisch in Abbildung 17.11. Wichtige Anwendungsfälle von Theorem 17.7 sind die im Rahmen der Theorie des Allgemeinen Linearen Modells betrachteten F-Statistiken.

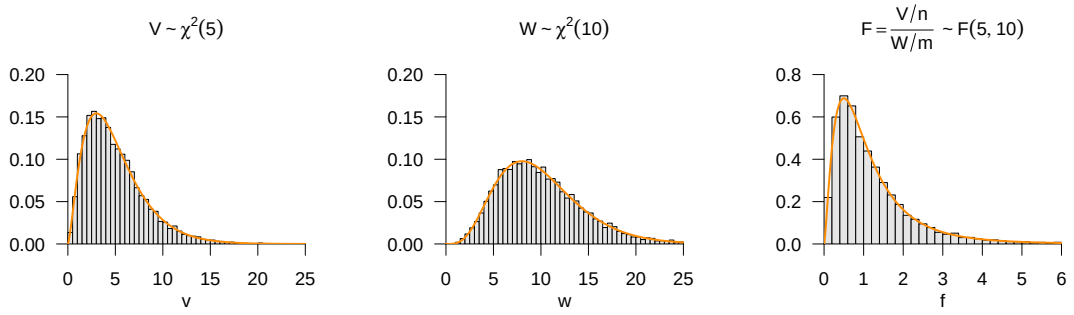


Abbildung 17.11. F-Transformation normalverteilter Zufallsvariablen.

17.8. Selbstkontrollfragen

1. Erläutern Sie die Bedeutung der in diesem Abschnitt betrachteten Transformationen von normalverteilten Zufallsvariablen für die Frequentistische Inferenz.
2. Geben Sie das Theorem zur Summentransformation wieder.
3. Geben Sie das Theorem zur Mittelwerttransformation wieder.
4. Geben Sie das Theorem zur Z-Transformation wieder.
5. Geben Sie das Theorem zur χ^2-Transformation wieder.
6. Beschreiben Sie die WDF der t-Verteilung in Abhängigkeit ihrer Freiheitsgradparameter.
7. Geben Sie das Theorem zur T-Transformation wieder.
8. Geben Sie das Theorem zur F-Transformation wieder.

Teil III.

Frequentistische Inferenz

18. Grundbegriffe Frequentistischer Inferenz

18.1. Frequentistische Inferenzmodelle

Mit folgender Definition wollen wir zunächst einige grundlegende Begrifflichkeiten bei der Betrachtung Frequentistischer Inferenzmodelle einführen.

Definition 18.1 (Frequentistische Inferenzmodelle)**.** Ein *Frequentistisches Inferenzmodell* ist ein Tupel

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \tag{18.1}$$

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 $\mathcal{M}_0 := (\mathcal{Y}_0, \mathcal{A}_0, \{\mathbb{P}_\theta^0 | \theta \in \Theta\})$ wrid 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 wird durch einen Zufallsvektor υ, 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 Frequentistische Inferenzmodellen also explizit zwei oder mehr Wahrscheinlichkeitsmaße, die die Verteilung von υ mutmaßlich bestimmen. Eine Realisierung von υ, also konkret vorliegende Datenwerte $y \in \mathcal{Y}$, nennt man *Datensatz*, *Beobachtungswert*, *Messwert* oder *Stichprobenwert*. Erwartungswerte und (Ko)Varianzen von υ bezüglich $\mathbb{P}_\theta$ schreibt man meist als $\mathbb{E}_\theta(\upsilon)$, $\mathbb{V}_\theta(\upsilon)$ und $\mathbb{C}_\theta(\upsilon)$. Frequentistische Produktmodelle modellieren die n-fache unabhängige Wiederholung eines Zufallsvorgangs. Die entsprechende Menge von Zufallsvektoren $\upsilon_1, ..., \upsilon_n$ 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 $y_1, ..., y_n$ von $\upsilon_1, ..., \upsilon_n$ 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 unbekannten, 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 Auszeichung des in einem Anwendungskontext unterstellten wahren, aber unbekannten, Parameterwerts.

Beispiele

Mit dem univariten Normalverteilungsmodell und dem Bernoullimodell wollen wir zwei erste Beispiel für Frequentistische Inferenzmodelle geben.

Definition 18.2 (Normalverteilungsmodell)**.** Das univariate parametrische Produktmodell

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \tag{18.2}$$

mit

$$\mathcal{Y} := \mathbb{R}^n, \mathcal{A} := \mathcal{B}(\mathbb{R}^n), \theta := (\mu, \sigma^2), \Theta := \mathbb{R} \times \mathbb{R}_{>0}, \tag{18.3}$$

also

$$\{\mathbb{P}_\theta | \theta \in \Theta\} := \left\{ \prod_{i=1}^{n} N(\mu, \sigma^2) | (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \right\}, \tag{18.4}$$

und damit

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \tag{18.5}$$

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 15 schon kurz angerissen haben und an späterer Stelle vertiefen werden.

Definition 18.3 (Bernoullimodell)**.** Das univariate parametrische Produktmodell

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \tag{18.6}$$

mit

$$\mathcal{Y} := \{0, 1\}^n, \mathcal{A} := \mathcal{P}\left(\{0, 1\}^n\right), \theta := \mu, \Theta :=]0, 1[, \tag{18.7}$$

also

$$\{\mathbb{P}_\theta|\theta \in \Theta\} := \left\{ \prod_{i=1}^{n} \text{Bern}(\mu)|\mu \in]0,1[\right\}, \tag{18.8}$$

und damit

$$\upsilon_1, ..., \upsilon_n \sim \text{Bern}(\mu) \text{ mit } \mu \in]0,1[, \tag{18.9}$$

heißt *Bernoullimodell*.

•

18.2. 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 18.4 (Statistik)**.** $\mathcal{M}$ sei ein Frequentistisches Inferenzmodell und $(\Sigma, \mathcal{S})$ sei ein Messraum. Dann ist eine *Statistik* ein Zufallsvektor der Form

$$S : \mathcal{Y} \to \Sigma. \tag{18.10}$$

•

Sowohl Daten als auch Statistiken werden in der Frequentistischen Inferenz also durch Zufallsvektoren (im univariaten Fall entsprechend durch Zufallsvariablen) modelliert. Allerdings unterscheiden sich diese Zufallvektoren 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 11.1).

Beispiele

$\mathcal{M}$ sei das Normalverteilungsmodell. Dann sind zum Beispiel folgende Zufallsvariablen Statistiken:

- Das *Stichprobenmittel*

$$\bar{y} : \mathbb{R}^n \to \mathbb{R}, y \mapsto \bar{y}(y) := \frac{1}{n} \sum_{i=1}^{n} y_i, \tag{18.11}$$

- Die *Stichprobenvarianz*

$$s^2 : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, y \mapsto s^2(y) := \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}(y))^2, \tag{18.12}$$

- Die *Stichprobenstandardabweichung*

$$s : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, y \mapsto s(y) := \sqrt{s^2(y)}, \tag{18.13}$$

Oft bleibt wie hier das Wesen von Statistiken als Zufallsvariablen oder Zufallsvektoren notationell eher implizit. Dies ändert allerdings nichts an der fundamental zu beachtetenden Tatsache, dass Statistiken als Funktionen von vom Zufall abhängigen Werten selbst wiederrum Zufallsvariablen oder Zufallsvektoren sind.

Definition 18.5 (Schätzer)**.** $\mathcal{M}$ sei ein Frequentistisches Inferenzmodell, $(\Sigma, \mathcal{S})$ sei ein Messraum und $\tau : \Theta \to \Sigma$ sei eine Abbildung, die jedem $\theta \in \Theta$ eine Kenngröße $\tau(\theta) \in \Sigma$ zuordnet. Dann heißt eine Statistik

$$\hat{\tau} : \mathcal{Y} \to \Sigma \tag{18.14}$$

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 *perse* ein guter Schätzer. In der Frequentistischen Inferenz definiert man deshalb zusätzlich *Schätzgütekriterien*, wie in Kapitel 19 ausführlich dargestellt.

Beispiel

$\mathcal{M}$ sei das Normalverteilungsmodell. Dann ist zum Beispiel das Stichprobenmittel $\bar{y} : \mathbb{R}^n \to \mathbb{R}$ ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}, (\mu, \sigma^2) \mapsto \tau(\mu, \sigma^2) := \mu. \tag{18.15}$$

Ebenso ist $\bar{y}$ ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}, (\mu, \sigma_2) \mapsto \tau(\mu, \sigma^2) := \mathbb{E}_{\mu,\sigma^2}(y_1). \tag{18.16}$$

Weiterhin ist die konstante Funktion

$$\hat{\tau} : \mathbb{R}^n \to \mathbb{R}, y \mapsto \hat{\tau}(y) := 42 \tag{18.17}$$

ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}_{>0}, (\mu, \sigma_2) \mapsto \tau(\mu, \sigma^2) := \sigma^2. \tag{18.18}$$

Dass eine Funktion $\hat{\tau} : \mathcal{Y} \to \Sigma$ ein Schätzer ist impliziert also keinesfalls, dass sie ein guter Schätzer ist.

18.3. 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 18.1). Dazu sei $\mathcal{M}$ ein univariates parametrisches Frequentistisches Produktmodell und es seien $\upsilon_1, ..., \upsilon_n \sim p_\theta$ die Zufallsvariablen der Stichprobe, die wir etwa in einem Zufallvektor $\upsilon := (\upsilon_1, ..., \upsilon_n)$ zusammenfassen können. Von einem konkret vorliegenden Datensatz $y_1, ..., y_n$ mit $y_i \in \mathbb{R}$, $i = 1, ..., n$, den wir etwa in einem n-dimensionalen Vektor $y := (y_1, ..., y_n)^T \in \mathbb{R}^n$ zusammenfassen können, *wird dann angenommen, dass er eine der möglichen Realisierungen von υ auf Grundlage einer Verteilung $\mathbb{P}_\theta$ mit wahrem, 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 $y^{(1)} := \left(y_1^{(1)}, ..., y_n^{(1)}\right)$ mit $\bar{y}^{(1)} = \frac{1}{n}\sum_{i=1}^n y_i^{(1)}$

Datenrealisierung $y^{(2)} := \left(y_1^{(2)}, ..., y_n^{(2)}\right)$ mit $\bar{y}^{(2)} = \frac{1}{n}\sum_{i=1}^n y_i^{(2)}$

Datenrealisierung $y^{(3)} := \left(y_1^{(3)}, ..., y_n^{(3)}\right)$ mit $\bar{y}^{(3)} = \frac{1}{n}\sum_{i=1}^n y_i^{(3)}$

Datenrealisierung $y^{(4)} := \left(y_1^{(4)}, ..., y_n^{(4)}\right)$ mit $\bar{y}^{(4)} = \frac{1}{n}\sum_{i=1}^n y_i^{(4)}$

Datenrealisierung $y^{(5)} := \left(y_1^{(5)}, ..., y_n^{(4)}\right)$ mit $\bar{y}^{(5)} = \frac{1}{n}\sum_{i=1}^n y_i^{(5)}$

...

Vor diesem Hintegrund behandelt die behandelt die Frequentistische Inferenz dann üblicheweise 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 $\upsilon = (\upsilon_1, ..., \upsilon_n) \sim p_\theta$. 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 $\bar{\upsilon}_n$. Wenn ein Inferenzverfahren auf Grundlage dieser Annahmen für "gut" befunden wir, so heißt das also insbesondere nur, dass das Verfahren bei häufiger Anwendung "im Mittel gut" ist. Im Einzelfall, also im Normalfall nur eines vorliegenden

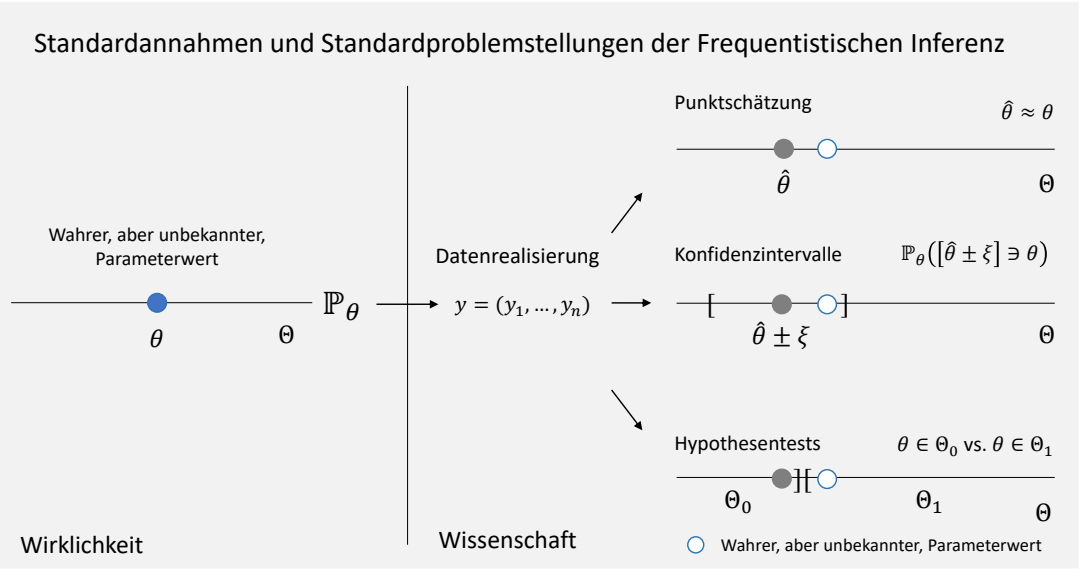


Abbildung 18.1. Standardannahmen und Standardproblemstellungen Frequentistischer Inferenz. Die Frequentistische Inferenz unterstellt, dass es in der Wirklichkeit einen wahren, aber unbekannten, Parameterwert des Wahrscheinlichkeitsmaßes $\mathbb{P}_\theta$ gibt, dass als Modell für die Erhebung eines Datensatzes dient. Ein konkret vorliegender Datenzsatz $y = (y_1, \ldots, y_n)$ ist dann eine (und insbesondere *nur* eine) der möglichen Realisierungen des anhand $\mathbb{P}_\theta$ verteilten Zufallsvektors $\upsilon := (\upsilon_1, \ldots, \upsilon_n)$. Auf Grundage 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 19, Kapitel 20 und Kapitel 21 diskutiert werden soll. Der tatsächliche wahre, aber unbekannte, Parameterwert aber bleibt auch nach Abschluss eines Inferenzverfahrens immer unbekannt.

Datensatzes, kann sie auch "schlecht" sein. Wir werden diese Denkweise insbesondere im Kontext der Punktschätzung (Kapitel 19) vertiefen. Ebenso beurteilt die Frequentistische Inferenz die *Stärke empirischer Evidenz* vor dem Hintegrund 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 20) und Hypothesentests (Kapitel 21).

18.3.1. Anwendungsbeispiel

Für ein erstes Anwendungsbeispiel Frequentistischer Inferenzverfahren betrachten wir die evidenzbasierte Evaluation einer Psychotherapie bei Depression. Dazu sei der in Tabelle 18.1 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. (2009)). Die `dBDI` Werte sollen dabei die Reduktion des BDI-II Scores der Patient:innen über den Zeitraum der Therapie wiederspiegeln. 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 18.1. 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

$$v_i := \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, ..., n \tag{18.19}$$

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 Zufallvorgänge, 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 18.19 behaftet ist.

Aus Gleichung 18.19 folgt dann direkt

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2), \tag{18.20}$$

denn für $i = 1, ..., n$ und mit

$$\upsilon_i = f(\varepsilon_i) \text{ mit } f : \mathbb{R} \to \mathbb{R}, e_i \mapsto f(e_i) := e_i + \mu. \tag{18.21}$$

gilt für die WDFen der υ_i, dass

$$\begin{aligned} p_{\upsilon_i}(y_i) &= \frac{1}{|1|} p_{\varepsilon_i}\left(\frac{y_i - \mu}{1}\right) \\ &= N\left(y_i - \mu; 0, \sigma^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu - 0)^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\ &= N(y_i; \mu, \sigma^2) \end{aligned} \tag{18.22}$$

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?

18.4. Selbstkontrollfragen

1. Geben Sie die Definition des Begriffs des parametrischen Frequentistischen Inferenzmodells wieder.
2. Erläutern Sie den Begriff des parametrischen Frequentistischen Inferenzmodells.
3. Geben Sie die Definition des Begriffs des parametrischen Frequentistischen Produktmodells wieder.
4. Erläutern Sie den Begriff des parametrischen Frequentistischen Produktmodells.
5. Was ist der Unterschied zwischen univariaten und multivariaten Frequentistischen Inferenzmodellen?
6. Geben Sie die Definition des Begriffs des Normalverteilungsmodells wieder.
7. Geben Sie die Definition des Begriffs des Bernoullimodells wieder.
8. Geben Sie die Definition des Begriffs der Statistik wieder.
9. Erläutern Sie den Begriff der Statistik.
10. Geben Sie die Definition des Begriffs des Schätzers wieder.
11. Erläutern Sie den Begriff des Schätzers.
12. Erläutern Sie die datenanalytischen Standardannahmen der Frequentistischen Inferenz.

19. Punktschätzung

In diesem Kapitel gehen wir immer im Sinne der in Kapitel 18 eingeführten Begrifflichkeiten immer von einem parametrischem Produktmodell

$$\mathcal{M} := \{\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}\} \tag{19.1}$$

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. $\upsilon := (\upsilon_1, ..., \upsilon_n)$ bezeichnet die zu $\mathcal{M}$ gehörende Stichprobe unabhängig und identisch verteilter Zufallsvariablen, es gilt also durchgängig

$$\upsilon_1, ..., \upsilon_n \sim \mathbb{P}_\theta. \tag{19.2}$$

Wesen und Ziel der hier behandelten *Punkt*schätzung 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 20 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, Parametert eines parametrischen Produktmodells oder auch allgemein eines Frequentistischen Inferenzmodells zu schätzen, nutzt man in der Frequentistischen Inferenz sogenannte *Parameterpunktschätzer*.

Definition 19.1. $\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\})$ sei ein Frequentistisches Inferenzmodell, $(\Theta, \mathcal{S})$ sei ein Messraum und $\hat{\theta} : \mathcal{Y} \to \Theta$ sei eine Abbildung. Dann nennt man $\hat{\theta}$ einen *Parameterpunktschätzer* für θ.

•

Parameterpunktschätzer werden meist auch einfach als *Parameterschätzer* bezeichnet. Im Sinne von Definition 18.5 sind Parameterpunktschätzer Schätzer mit $\tau := \mathrm{id}_\Theta$. 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 19.1 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 gegebenenes Frequentistisches Inferenzmodell Parameterschätzer zu bestimmen, die, wie wir an späterer Stelle sehen werden, garantiert bestimmte wünschenswerte Eigenschaften haben (Kapitel 19.4). 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 19.2 und Kapitel 19.3 ein.

19.1. 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 19.2 (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

$$L : \Theta \to [0, \infty[, \theta \mapsto L(\theta) := \prod_{i=1}^{n} p_\theta(y_i) \tag{19.3}$$

und die *Log-Likelihood-Funktion* ist definiert als

$$\ell_n : \Theta \to \mathbb{R}, \theta \mapsto \ell(\theta) := \ln L(\theta). \tag{19.4}$$

•

Die Likelihood-Funktion ist also eine Funktion des Parameters und ihre Funktionswerte sind die Werte der gemeinsamen WMF bzw. WDF beobachteter Datenwerte $y_1, ..., y_n$. 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 19.3 (Maximum-Likelihood Schätzer)**.** $\mathcal{M}$ sei ein parametrisches Produktmodell mit Parameter $\theta \in \Theta$. Ein *Maximum-Likelihood Schätzer* von θ ist definiert als

$$\hat{\theta}^{\text{ML}} : \mathcal{Y} \to \Theta, y \mapsto \hat{\theta}^{\text{ML}}(y) := \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \ell(\theta) \tag{19.5}$$

•

Man beachte bei Definition 19.3, 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 19.3, dass Definition 19.2 impliziert, dass

$$\hat{\theta}^{\mathrm{ML}}(y) = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} p_\theta(y_i) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \ln p_\theta(y_i) \tag{19.6}$$

was die Abhängigkeit eines Maximum-Likelihood Schätzers von den Daten verdeutlicht.

Mit Definition 19.3 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 5.2). 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 19.1 zeigen wir, dass der Maximum-Likelihood Schätzer für den Parameter des Bernoullimodells aus Definition 18.3 durch das entsprechende Stichprobenmittel gegeben ist und in Theorem 19.2 zeigen wir, dass die Maximum-Likelihood Schätzer für den Erwartungswert- und Varianzparameter des Normalverteilungsmodells aus Definition 18.2 durch das Stichprobenmittel und eine modifizierte Stichprobenvarianz, respektive, gegeben sind.

Beispiele

Theorem 19.1 (Maximum-Likelihood Schätzer des Bernoullimodells)**.** *$\mathcal{M}$ sei das Bernoullimodell, es gelte also $\upsilon_1, ..., \upsilon_n \sim Bern(\mu)$. Dann ist*

$$\hat{\mu}^{ML} : \{0,1\}^n \to [0,1], y \mapsto \hat{\mu}^{ML}(y) := \frac{1}{n}\sum_{i=1}^{n} y_i \tag{19.7}$$

ein Maximum-Likelihood Schätzer von μ

∘

Beweis. Wir formulieren zunächst die Log-Likelihood-Funktion. Für die Likelihood-Funktion gilt

$$L :]0,1[\to]0,1[, \mu \mapsto L(\mu) := \prod_{i=1}^{n} \mu^{y_i}(1-\mu)^{1-y_i} = \mu^{\sum_{i=1}^n y_i}(1-\mu)^{n-\sum_{i=1}^n y_i}. \tag{19.8}$$

Logarithmieren ergibt

$$\ell :]0,1[\to \mathbb{R}, \mu \mapsto \ell(\mu) = \ln\mu \sum_{i=1}^{n} y_i + \ln(1-\mu)\left(n - \sum_{i=1}^{n} y_i\right). \tag{19.9}$$

Wir werten dann die Ableitung der Log-Likelihood-Funktion aus. Es gilt

$$\begin{aligned}\frac{d}{d\mu}\ell(\mu) &= \frac{d}{d\mu}\left(\ln\mu\sum_{i=1}^{n}y_i + \ln(1-\mu)\left(n-\sum_{i=1}^{n}y_i\right)\right)\\ &= \frac{d}{d\mu}\ln\mu\sum_{i=1}^{n}y_i + \frac{d}{d\mu}\ln(1-\mu)\left(n-\sum_{i=1}^{n}y_i\right)\\ &= \frac{1}{\mu}\sum_{i=1}^{n}y_i - \frac{1}{1-\mu}\left(n-\sum_{i=1}^{n}y_i\right).\end{aligned} \tag{19.10}$$

Nullsetzen ergibt dann folgende *Maximum-Likelihood-Gleichung* als notwendige Bedingung für einen Maximum-Likelihood Schätzer im Bernoullimodell:

$$\frac{1}{\hat{\mu}^{\mathrm{ML}}}\sum_{i=1}^{n}y_i - \frac{1}{1-\hat{\mu}^{\mathrm{ML}}}\left(n-\sum_{i=1}^{n}y_i\right) = 0. \tag{19.11}$$

Auflösen der Maximum-Likelihood-Gleichung nach $\hat{\mu}^{\mathrm{ML}}$ ergibt dann

$$\begin{aligned}&\frac{1}{\hat{\mu}^{\mathrm{ML}}}\sum_{i=1}^{n}y_i - \frac{1}{1-\hat{\mu}^{\mathrm{ML}}}\left(n-\sum_{i=1}^{n}y_i\right) = 0\\ \Leftrightarrow\ &\hat{\mu}^{\mathrm{ML}}(1-\hat{\mu}^{\mathrm{ML}})\left(\frac{1}{\hat{\mu}^{\mathrm{ML}}}\sum_{i=1}^{n}y_i - \frac{1}{1-\hat{\mu}^{\mathrm{ML}}}\left(n-\sum_{i=1}^{n}y_i\right)\right) = 0\\ \Leftrightarrow\ &\sum_{i=1}^{n}y_i - \hat{\mu}^{\mathrm{ML}}\sum_{i=1}^{n}y_i - n\hat{\mu}^{\mathrm{ML}} + \hat{\mu}^{\mathrm{ML}}\sum_{i=1}^{n}y_i = 0\\ \Leftrightarrow\ &n\hat{\mu}^{\mathrm{ML}} = \sum_{i=1}^{n}y_i\\ \Leftrightarrow\ &\hat{\mu}^{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}y_i.\end{aligned} \tag{19.12}$$

$\hat{\mu}^{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}y_i$ 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.

□

Theorem 19.2 (Maximum-Likelihood Schätzer des Normalverteilungsmodells)**.** *$\mathcal{M}$ sei das Normalverteilungsmodell, es gelt also $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$. Dann sind*

$$\hat{\mu}^{ML} : \mathbb{R}^n \to \mathbb{R}, y \mapsto \hat{\mu}^{ML}(y) := \frac{1}{n}\sum_{i=1}^{n}y_i \tag{19.13}$$

und

$$\hat{\sigma}^{2^{ML}} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}, y \mapsto \hat{\sigma}^{2^{ML}}(y) := \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\mu}^{ML}\right)^2. \tag{19.14}$$

Maximum-Likelihood Schätzer für μ und σ^2, respektive.

∘

Beweis. Wir formulieren zunächst die Log-Likelihood-Funktion. Für die Likelihood-Funktion ergibt sich

$$\begin{aligned}L : \mathbb{R}\times\mathbb{R}_{>0} \to \mathbb{R}_{>0}, (\mu,\sigma^2) \mapsto L(\mu,\sigma^2) &:= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y_i-\mu)^2\right)\\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right).\end{aligned} \tag{19.15}$$

Logarithmieren ergibt dann

$$\ell : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}, (\mu, \sigma^2) \mapsto \ell_n(\mu, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2. \tag{19.16}$$

Die Auswertung der partiellen Ableitungen der Log-Likelihood-Funktion ergeben dann

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma^2) = -\frac{\partial}{\partial \mu} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial}{\partial \mu} (y_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \tag{19.17}$$

und

$$\frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2) = -\frac{n}{2} \frac{\partial}{\partial \sigma^2} \ln \sigma^2 - \frac{\partial}{\partial \sigma^2} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2. \tag{19.18}$$

Das System der Maximum-Likelihood Gleichungen als Ausdruck der notwendigen Bedingungen für Extremstellen der Log-Likelihood-Funktion hat in diesem Fall also die Form

$$\sum_{i=1}^{n} (y_i - \hat{\mu}^{\mathrm{ML}}) = 0 \text{ und } -\frac{n}{2\hat{\sigma}^{2\mathrm{ML}}} + \frac{1}{2\hat{\sigma}^{4\mathrm{ML}}} \sum_{i=1}^{n} (y_i - \mu)^2 = 0. \tag{19.19}$$

Lösen des Systems der Maximum-Likelihood Gleichungen ergibt dann zunächst

$$\sum_{i=1}^{n} (y_i - \hat{\mu}^{\mathrm{ML}}) = 0 \Leftrightarrow \sum_{i=1}^{n} y_i = n\hat{\mu}^{\mathrm{ML}} \Leftrightarrow \hat{\mu}^{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{19.20}$$

Damit ist

$$\hat{\mu}^{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{19.21}$$

ein potentieller Maximum-Likelihood Schätzer von μ. Einsetzen dieses Schätzers in die zweite Maximum-Likelihood Gleichung ergibt dann

$$\begin{aligned} -\frac{n}{2\hat{\sigma}^{2\mathrm{ML}}} + \frac{1}{2\hat{\sigma}^{4\mathrm{ML}}} \sum_{i=1}^{n} (y_i - \hat{\mu}^{\mathrm{ML}})^2 = 0 \\ \Leftrightarrow -n\hat{\sigma}^{2\mathrm{ML}} + \sum_{i=1}^{n} (y_i - \hat{\mu}^{\mathrm{ML}})^2 = 0 \\ \Leftrightarrow \hat{\sigma}^{2\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu}^{\mathrm{ML}})^2. \end{aligned} \tag{19.22}$$

Also ist

$$\hat{\sigma}^{2\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu}^{\mathrm{ML}})^2 \tag{19.23}$$

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.

□

Man beachte bei Theorem 19.2, dass $\hat{\mu}^{\mathrm{ML}}$ mit dem Stichprobenmittel $\bar{v}$ identisch ist, aber $\hat{\sigma}^{2\mathrm{ML}}$ 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 19.2 im Kontext des Anwendungsbeispiels aus Kapitel 18.3.1 betrachten. Wir hatten dort den beobachteten `dBDI` Werten das Normalverteilungsmodell

$$v_1, ..., v_n \sim N(\mu, \sigma^2) \tag{19.24}$$

zugrundegelegt. Die Maximum-Likelihood Schätzer für die Parameter dieses Modells lassen sich dann anhand von Theorem 19.2 mithilfe der **R** Stichprobenmittel- und Stichprobenvarianzfunktionen `mean()` und `var()` und unter Beachtung der Identität

$$\frac{n-1}{n}s^2 = \frac{n-1}{n} \cdot \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}^{\text{ML}})^2 = \hat{\sigma}^{2^{\text{ML}}} \tag{19.25}$$

wie in folgendem **R** Code auswerten.

```
D           = read.csv("./_data/302-Punktschätzung.csv") # Datensatzeinlesen
y           = D$dBDI                                     # 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

$$\hat{\mu}^{\text{ML}} = 3.17 \text{ und } \hat{\sigma}^{2^{\text{ML}}} = 12.6 \tag{19.26}$$

Tipps für die wahren, aber unbekannten, Parameter des Modells.

19.2. 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 19.3 und betreffen die Eigenschaften eines Schätzers im Grenzfall $n \to \infty$ von großen Stichprobenumfängen.

Es sei zunächst (Σ, S) ein Messraum und $\hat{\tau} : \mathcal{Y} \to \Sigma$ ein Schätzer von $\tau : \Theta \to \Sigma$ (vgl. Definition 18.5). In der Folge betrachten wir neben Parameterschätzern der Form

$$\tau : \Theta \to \Sigma, \tau(\theta) := \theta \tag{19.27}$$

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 $\upsilon_1, ..., \upsilon_n$ identisch sind, handelt es sich dabei um Schätzer der Form

$$\tau : \Theta \to \Sigma,\ \theta \mapsto \tau(\theta) \text{ mit } \tau(\theta) := \mathbb{E}_\theta(\upsilon_1), \tau(\theta) := \mathbb{V}_\theta(\upsilon_1), \text{ und } \tau(\theta) := \mathbb{S}_\theta(\upsilon_1). \tag{19.28}$$

Speziell wollen wir in diesem Abschnitt vier Aspekte von Schätzereigenschaften bei endlichen Stichproben beleuchten. In Kapitel 19.2.1 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 19.2.2 führen wir mit den Begriffen der *Varianz* und des *Standardsfehlers* eines Schätzers als Bezeichungen für die Varianz der Zufallsvariable $\hat{\tau}(\upsilon)$ und die Standardabweichung der Zufallsvariable $\hat{\tau}(\upsilon)$ zwei Maße für die Frequentistische Variabilität von Schätzern ein. Mit dem *mittlere quadratischen Fehler* eines Schätzers $\hat{\tau}$ als Erwartungswert der quadrierten Abweichung von $\hat{\tau}(\upsilon)$ von $\tau(\theta)$ führen wir dann in Kapitel 19.2.3 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 19.2.4 disktutierte *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.

19.2.1. Erwartungstreue

Der Begriff der Erwartungstreue eines Schätzers ergibt sich im Kontext des *Fehlers* und des *systematischen Fehlers* eines Schätzers wie folgt.

Definition 19.4 (Fehler, Systematischer Fehler und Erwartungstreue)**.** υ sei eine Stichprobe eines Frequentischen Inferenzmodells und $\hat{\tau}$ sei ein Schätzer für τ.

- Der *Fehler* von $\hat{\tau}$ ist definiert als

$$\hat{\tau}(\upsilon) - \tau(\theta). \tag{19.29}$$

- Der *systematische Fehler* (engl. *Bias*) von $\hat{\tau}$ ist definiert als

$$\mathrm{B}(\hat{\tau}) := \mathbb{E}_\theta(\hat{\tau}(\upsilon)) - \tau(\theta). \tag{19.30}$$

- Der Schätzer $\hat{\tau}$ heißt *erwartungstreu* (engl. *unbiased*), wenn

$$\mathrm{B}(\hat{\tau}) = 0 \Leftrightarrow \mathbb{E}_\theta(\hat{\tau}(\upsilon)) = \tau(\theta) \text{ für alle } \theta \in \Theta \text{ und alle } n \in \mathbb{N}. \tag{19.31}$$

  Andernfalls heißt $\hat{\tau}$ *verzerrt (engl. biased).*

•

Man beachte, dass in Definition 19.4 der Fehler eines Schätzers von der spezifischen Realisation der Stichprobe υ 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 19.4, dass er erwartungstreu ist, wenn gilt, dass

$$\mathbb{E}_\theta(\hat{\theta}(\upsilon)) = \theta. \tag{19.32}$$

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 19.3 (Erwartungstreue von Stichprobenmittel und Stichproenvarianz)**.** *$\upsilon := (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells. Dann gelten*

(1) Das Stichprobenmittel

$$\bar{\upsilon} := \frac{1}{n}\sum_{i=1}^{n} \upsilon_i \tag{19.33}$$

ist ein erwartungstreuer Schätzer des Erwartungswerts $\mathbb{E}_\theta(\upsilon_1)$.

(2) Die Stichprobenvarianz

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n} (\upsilon_i - \bar{\upsilon})^2 \tag{19.34}$$

ist ein erwartungstreuer Schätzer der Varianz $\mathbb{V}_\theta(\upsilon_1)$.

∘

Beweis. (1) Die Erwartungstreue des Stichprobenmittels ergibt mit den Eigenschaften des Erwartungswerts (vgl. Theorem 13.3) aus

$$\mathbb{E}_\theta(\bar{\upsilon}) = \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} \upsilon_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta\left(\upsilon_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta\left(\upsilon_1\right) = \frac{1}{n} n \mathbb{E}_\theta\left(\upsilon_1\right) = \mathbb{E}_\theta\left(\upsilon_1\right). \tag{19.35}$$

(2) Um die Erwartungstreue der Stichprobenvarianz zu zeigen, halten wir zunächst fest, dass mit den Eigenschaften der Varianz gilt, dass (vgl. Theorem 13.5)

$$\mathbb{V}_\theta(\bar{\upsilon}) = \mathbb{V}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} \upsilon_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}_\theta\left(\upsilon_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}_\theta\left(\upsilon_1\right) = \frac{1}{n^2} n \mathbb{V}_\theta\left(\upsilon_1\right) = \frac{\mathbb{V}_\theta\left(\upsilon_1\right)}{n}. \tag{19.36}$$

Weiterhin gilt für den Term der summierten quadratischen Abweichungen in der Stichprobenvarianz, dass

$$\sum_{i=1}^{n} \left(\upsilon_i - \bar{\upsilon}\right)^2 = \sum_{i=1}^{n} (\upsilon_i - \mathbb{E}_\theta(\upsilon_1))^2 - n(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2, \tag{19.37}$$

weil

$$\begin{aligned}
\sum_{i=1}^{n} \left(\upsilon_i - \bar{\upsilon}\right)^2 &= \sum_{i=1}^{n} \left(\upsilon_i - \mathbb{E}_\theta(\upsilon_1) - \bar{\upsilon} + \mathbb{E}_\theta(\upsilon_1)\right)^2 \\
&= \sum_{i=1}^{n} \left((\upsilon_i - \mathbb{E}_\theta(\upsilon_1)) - (\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))\right)^2 \\
&= \sum_{i=1}^{n} (\upsilon_i - \mathbb{E}_\theta(\upsilon_1))^2 - 2(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1)) \left(\sum_{i=1}^{n} (\upsilon_i - \mathbb{E}_\theta(\upsilon_1))\right) + \sum_{i=1}^{n} (\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2 \\
&= \sum_{i=1}^{n} (\upsilon_i - \mathbb{E}_\theta(\upsilon_1))^2 - 2(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1)) \left(\sum_{i=1}^{n} \upsilon_i - n\mathbb{E}_\theta(\upsilon_1)\right) + n(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2 \\
&= \sum_{i=1}^{n} (\upsilon_i - \mathbb{E}_\theta(\upsilon_1))^2 - 2(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1)) \left(n\left(\frac{1}{n}\sum_{i=1}^{n} \upsilon_i\right) - n\mathbb{E}_\theta(\upsilon_1)\right) + n(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2 \\
&= \sum_{i=1}^{n} (\upsilon_i - \mathbb{E}_\theta(\upsilon_1))^2 - 2n(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2 + n(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2 \\
&= \sum_{i=1}^{n} (\upsilon_i - \mathbb{E}_\theta(\upsilon_1))^2 - n(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2.
\end{aligned} \tag{19.38}$$

Zusammen ergibt sich also

$$\mathbb{E}_\theta\left((n-1)S^2\right) = \mathbb{E}_\theta\left(\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon})^2\right) \tag{19.39}$$

$$= \mathbb{E}_\theta\left(\sum_{i=1}^{n}(\upsilon_i - \mathbb{E}_\theta(\upsilon_1))^2 - n(\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2\right) \tag{19.40}$$

$$= \sum_{i=1}^{n}\mathbb{E}_\theta\left((\upsilon_i - \mathbb{E}_\theta(\upsilon_1))^2\right) - n\mathbb{E}_\theta\left((\bar{\upsilon} - \mathbb{E}_\theta(\upsilon_1))^2\right) \tag{19.41}$$

$$= n\mathbb{V}_\theta(\upsilon_1) - n\mathbb{V}_\theta(\bar{\upsilon}) \tag{19.42}$$

$$= n\mathbb{V}_\theta(\upsilon_1) - n\frac{\mathbb{V}_\theta(\upsilon_1)}{n} \tag{19.43}$$

$$= n\mathbb{V}_\theta(\upsilon_1) - \mathbb{V}_\theta(\upsilon_1) \tag{19.44}$$

$$= (n-1)\mathbb{V}_\theta(\upsilon_1). \tag{19.45}$$

Schließlich ergibt sich dann

$$\mathbb{E}_\theta\left(S^2\right) = \mathbb{E}_\theta\left(\frac{1}{n-1}(n-1)S^2\right) = \frac{1}{n-1}\mathbb{E}_\theta\left((n-1)S^2\right) = \frac{1}{n-1}(n-1)\mathbb{V}_\theta(\upsilon_1) = \mathbb{V}_\theta(\upsilon_1) \tag{19.46}$$

und damit die Erwartungstreue der Stichprobenvarianz als Schätzer der Varianz.

□

Natürlich sind in Theorem 19.3 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 υ_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 19.4 (Verzerrtheit der Stichprobenstandardabweichung)**.** *$\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells. Dann ist die Stichprobenstandardabweichung*

$$S := \sqrt{S^2} \tag{19.47}$$

ein verzerrter Schätzer der Standardabweichung $\mathbb{S}_\theta(\upsilon_1)$.

∘

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 14.5), dass

$$\mathbb{E}_\theta(S) = \mathbb{E}_\theta\left(\sqrt{S^2}\right) < \sqrt{\mathbb{E}_\theta(S^2)} = \sqrt{\mathbb{V}_\theta(\upsilon_1)} = \mathbb{S}_\theta(\upsilon_1). \tag{19.48}$$

□

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

$$\upsilon_1, ..., \upsilon_{12} \sim N(1.7, 2) \tag{19.49}$$

Dabei werden die Erwartungswerte der Schätzer anhand ihrer Stichprobenmittel über viele Realisierungen von $\upsilon_1, ..., \upsilon_{12}$ als Funktion der Anzahl an Realsierungen (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(NaN,nsim)             # Stichprobenmittelarray
s_sqr   = rep(NaN,nsim)             # Stichprobenvarianzarray
s       = rep(NaN,nsim)             # Stichprobenstandardabweichungarray

# Simulationsiterationen
for(sim in 1:nsim){

    # Stichprobenrealisation von \ups_1,...,\ups_{12}
    y          = rnorm(n,mu,sqrt(sigsqr))

    # Erwartungswert-, Varianz-, StandardabweichungSchätzer
    y_bar[sim] = mean(y)            # Stichprobenmittel
    s_sqr[sim] = var(y)             # Stichprobenvarianz
    s[sim]     = sd(y)              # Stichprobenstandardabweichung
}

# Erwartungswertschaetzung
E_hat_y_bar = cumsum(y_bar)/(1:nsim) # \mathbb{E}(\bar{\ups}) Schaetzungen
E_hat_s_sqr = cumsum(s_sqr)/(1:nsim) # \mathbb{E}(S^2) Schaetzungen
E_hat_s     = cumsum(s)    /(1:nsim) # \mathbb{E}(S) Schaetzungen
```

Abbildung 19.1 visualisiert die Ergebnisse obiger Simulation. Gezeigt sind Schätzungen der Erwartungswerte von Stichprobenmittel, Stichprobenvarianz und Stichprobenstandardabweichung als Funktion der Anzahl an Realisierungen der Stichprobenvariablen $\upsilon_1, ..., \upsilon_{12}$ sowie die wahren, aber unbekannten, Werte des Erwartungswerts, der Varianz und der Standardabweichung der υ_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 $\upsilon_1, ..., \upsilon_{12}$ entsprechen die Stichprobenmittel von $\bar{\upsilon}$ 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 Realsierungen von $\upsilon_1, ..., \upsilon_{12}$
konstant eine zu niedrige Schätzung der wahren, aber unbekannten, Standardabweichung.

19.2.2. 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 19.5 (Varianz und Standardfehler)**.** $\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines Frequentistischen Inferenzmodells und $\hat{\tau}$ sei ein Schätzer von τ.

- Die *Varianz* von $\hat{\tau}$ ist definiert als

$$\mathbb{V}_\theta(\hat{\tau}) := \mathbb{E}_\theta\left((\hat{\tau}(\upsilon) - \mathbb{E}_\theta(\hat{\tau}(\upsilon)))^2\right). \tag{19.50}$$

- Der *Standardfehler* von $\hat{\tau}$ ist definiert als

$$\mathrm{SE}(\hat{\tau}) := \sqrt{\mathbb{V}_\theta(\hat{\tau})}. \tag{19.51}$$

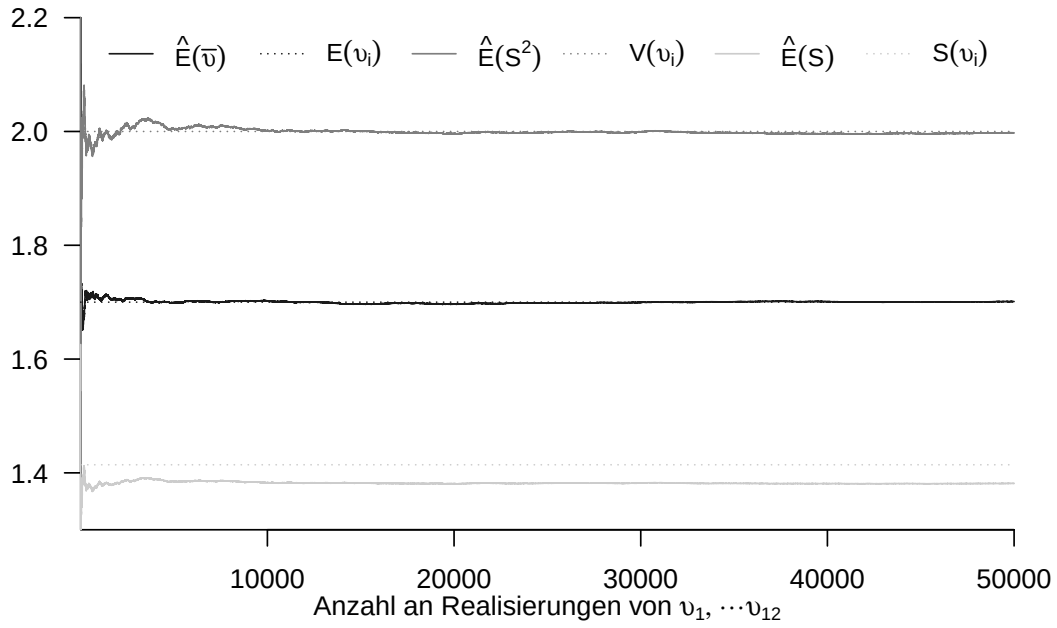


Abbildung 19.1. Simulation der Erwartungstreue von Stichprobenmittel und Stichprobenvarianz als Schätzer des Erwartungswerts und der Varianz bei normalverteilten Stichprobenvariablen und Simulation der Verzerrtheit der Stichprobenstandardabweichung als Schätzer der Standardabweichung bei normalverteilten Stichprobenvariablen

•

Die Varianz eines Schätzers $\hat{\tau}$ ist also als die Varianz der Zufallsvariable $\hat{\tau}(\upsilon)$ definiert. Der Standardfehler eines Schätzers $\hat{\tau}$ ist als die Standardabweichung von $\hat{\tau}(\upsilon)$ definiert. Als erstes Beispiel für einen Standardfehler betrachten wir den *Standardfehler des Stichprobenmittels*.

Theorem 19.5 (Standardfehler des Stichprobenmittels)**.** *$\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells. Dann ist der Standardfehler des Stichprobenmittels gegeben durch*

$$SE(\bar{\upsilon}) = \frac{\mathbb{S}_\theta(\upsilon_1)}{\sqrt{n}}. \tag{19.52}$$

.

∘

Beweis. Mit der Varianz des Stichprobenmittels ergibt sich

$$\mathrm{SE}(\bar{\upsilon}) = \sqrt{\mathbb{V}_\theta(\bar{\upsilon})} = \sqrt{\frac{\mathbb{V}_\theta(\upsilon_1)}{n}} = \frac{\mathbb{S}_\theta(\upsilon_1)}{\sqrt{n}}. \tag{19.53}$$

□

Der Standardfehler des Mittelwerts beschreibt die Variabilität des Stichprobenmittels. Da die Standardabweichung $\mathbb{S}_\theta(\upsilon_1)$ unbekannt ist, ist auch der Standardfehler $\mathrm{SE}(\bar{\upsilon})$ unbekannt, kann also nur geschätzt werden. Mit der Stichprobenstandardabweichung als verzerrter Schätzer der Standardabweichung $\mathbb{S}_\theta(\upsilon_1)$ ergibt sich ein ebenfalls verzerrter Schätzer für den Standardfehler des Stichprobenmittels zu

$$\hat{\mathrm{SE}}(\bar{\upsilon}) = \frac{S}{\sqrt{n}}. \tag{19.54}$$

Als zweites Beispiel wollen wir den Standardfehler des Maximum-Likelihood Schätzers für den Parameter eines Bernoulli-Modells betrachten.

Theorem 19.6 (Standardfehler des Maximum-Likelihood Schätzers des Bernoullimodellparameters)**.** *Es sei $\upsilon = (\upsilon_1, ..., \upsilon_n)$ 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*

$$SE(\hat{\mu}^{ML}) = \sqrt{\frac{\mu(1-\mu)}{n}}. \tag{19.55}$$

∘

Beweis. Es gilt

$$\mathrm{SE}\left(\hat{\mu}^{\mathrm{ML}}\right) = \sqrt{\mathbb{V}_\mu\left(\hat{\mu}^{\mathrm{ML}}\right)} = \sqrt{\mathbb{V}_\mu\left(\frac{1}{n}\sum_{i=1}^{n}\upsilon_i\right)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}_\mu(\upsilon_i)} = \sqrt{\frac{n\mu(1-\mu)}{n^2}} = \sqrt{\frac{\mu(1-\mu)}{n}}, \tag{19.56}$$

wobei die dritte Gleichung mit der Unabhängigkeit der υ_i und die vierte Gleichung mit der Varianz $\mathbb{V}_\mu(\upsilon_1) = \mathbb{V}_\mu(\upsilon_i) = \mu(1-\mu)$ 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 $\mathrm{SE}\,(\hat{\mu}^{\mathrm{ML}})$ ergibt sich mit dem Maximum-Likelihood Schätzer für den Bernoullimodellparameter durch

$$\widehat{\mathrm{SE}}\,(\hat{\mu}^{\mathrm{ML}}) = \sqrt{\frac{\hat{\mu}^{\mathrm{ML}}(1-\hat{\mu}^{\mathrm{ML}})}{n}}. \tag{19.57}$$

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 19.2 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 19.3 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(NaN, 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
    }
}
```

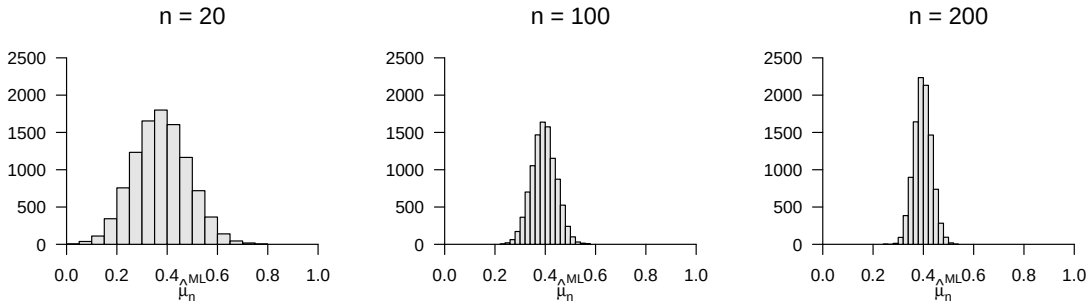


Abbildung 19.2. Simulation der Verteilung des Maximum-Likelihood Schätzers eines Bernoullimodells. Die Variabilität des Schätzers hängt dabei offenbar vom Stichprobenumfang n ab.

19.2.3. 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 (Erwatungstreue) und Variabilität (Varianz) eines Schätzer im Sinne seiner sogenannten *Bias-Varianz-Zerlegung*. Wir definieren den mittleren quadratischen Fehler eines Schätzers zunächt wie folgt.

Definition 19.6 (Mittlerer quadratischer Fehler)**.** $\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}$ ein Schätzer für τ. Dann ist der *mittlere quadratischer Fehler (engl. mean squared error)* von $\hat{\tau}$ definiert als

$$\text{MQF}(\hat{\tau}) := \mathbb{E}_\theta\left((\hat{\tau}(\upsilon) - \tau(\theta))^2\right). \tag{19.58}$$

•

Der mittlere quadratische Fehler von $\hat{\tau}$ ist also die erwartete quadrierte Abweichung von $\hat{\tau}(\upsilon)$ von $\tau(\theta)$. Man beachte, dass in Abgrenzung dazu die Varianz von $\hat{\tau}$ die erwartete quadrierte Abweichung von $\hat{\tau}$ von $\mathbb{E}_\theta(\hat{\tau}(\upsilon))$ ist. Dabei kann, wie in Kapitel 19.2.1 gesehen $\mathbb{E}_\theta(\hat{\tau}(\upsilon))$ 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 19.7 (Zerlegung des mittleren quadratischen Fehlers)**.** *$\upsilon = (\upsilon_1, ..., \upsilon_n)$ 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*

$$MQF(\hat{\tau}) = B(\hat{\tau})^2 + \mathbb{V}_\theta(\hat{\tau}). \tag{19.59}$$

∘

Beweis. Zur Vereinfachung der Notation seien $\tau := \tau(\theta)$, $\hat{\tau} := \hat{\tau}(\upsilon)$ und $\bar{\tau}_n := \mathbb{E}_\theta(\hat{\tau}(\upsilon))$. Dann gilt:

$$\begin{aligned}
\mathbb{E}_\theta\left((\hat{\tau} - \tau)^2\right) &= \mathbb{E}_\theta\left((\hat{\tau} - \bar{\tau}_n + \bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau} - \bar{\tau}_n)^2 + 2(\hat{\tau} - \bar{\tau}_n)(\bar{\tau}_n - \tau) + (\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau} - \bar{\tau}_n)^2\right) + 2\mathbb{E}_\theta\left((\hat{\tau} - \bar{\tau}_n)(\bar{\tau}_n - \tau)\right) + \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau} - \bar{\tau}_n)^2\right) + 2\mathbb{E}_\theta\left(\hat{\tau}\bar{\tau}_n - \hat{\tau}\tau - \bar{\tau}_n\bar{\tau}_n + \bar{\tau}_n\tau\right) + \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau} - \bar{\tau}_n)^2\right) + 2\left(\bar{\tau}_n\bar{\tau}_n - \bar{\tau}_n\tau\right) + \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\tau} - \bar{\tau}_n)^2\right) + 0 + \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) \\
&= \mathbb{E}_\theta\left((\bar{\tau}_n - \tau)^2\right) + \mathbb{E}_\theta\left((\hat{\tau} - \bar{\tau}_n)^2\right) \\
&= \mathbb{E}_\theta\left((\mathbb{E}_\theta(\hat{\tau}) - \tau)^2\right) + \mathbb{E}_\theta\left((\hat{\tau} - \mathbb{E}_\theta(\hat{\tau}))^2\right) \\
&= (\mathbb{E}_\theta(\hat{\tau}) - \tau)^2 + \mathbb{V}_\theta(\hat{\tau}) \\
&= \text{B}(\hat{\tau})^2 + \mathbb{V}_\theta(\hat{\tau}).
\end{aligned} \tag{19.60}$$

□

19.2.4. 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 erwartungstreuer 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 wiederrum auf dem Begriff der *Scorefunktion* eines Frequentischen 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 19.7 (Scorefunktion und Fisher-Information)**.** $\upsilon = (\upsilon_1, ..., \upsilon_n)$ 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

$$S(\theta) := \frac{d}{d\theta}\ell(\theta) \tag{19.61}$$

  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

$$I(\theta) := -\frac{d^2}{d\theta^2}\ell(\theta) \tag{19.62}$$

  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 Hintegrund 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 19.8 (Beobachtete und erwartete Fisher-Information)**.** $\upsilon = (\upsilon_1, ..., \upsilon_n)$ 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

$$I\left(\hat{\theta}^{\text{ML}}\right) := -\frac{d^2}{d\theta^2}\ell\left(\hat{\theta}^{\text{ML}}\right), \tag{19.63}$$

  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

$$J(\theta) := \mathbb{E}_\theta(I(\theta)). \tag{19.64}$$

  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 19.10) und des Normalverteilungsmodells (Theorem 19.12 und Theorem 19.11) verdeutlichen wollen, führen wir mit der *Additivität der Fisher-Information* bei parametrischen Produktmodellen (Theorem 19.8) und dem Erwartungswert und der Varianz der Scorefunktion (Theorem 19.9) noch wichtige Eigenschaften der genannten Begriffe ein, die die folgende Diskussion vereinfachen.

Theorem 19.8 (Additivität der Fisher-Information)**.** *$\upsilon = (\upsilon_1, ..., \upsilon_n)$ 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*

$$I(\theta) = nI_1(\theta) \text{ und } J(\theta) = nJ_1(\theta). \tag{19.65}$$

∘

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

$$\begin{aligned}
J(\theta) &= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ell(\theta)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ln\left(\prod_{i=1}^n p_\theta(\upsilon_i)\right)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\sum_{i=1}^n \ln p_\theta(\upsilon_i)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\sum_{i=1}^n \ln p_\theta(y_1)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ell_1(\theta)n\right) \\
&= n\mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ell_1(\theta))\right) \\
&= nJ(\theta).
\end{aligned} \tag{19.66}$$

□

Nach Theorem 19.8 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 19.9 (Erwartungswert und Varianz der Scorefunktion)**.** *Der Erwartungswert der Scorefunktion einer Zufallsvariable ist*

$$\mathbb{E}_\theta(S(\theta)) = 0 \tag{19.67}$$

und die Varianz der Scorefunktion einer Zufallsvariable ist

$$\mathbb{V}_\theta(S(\theta)) = J(\theta). \tag{19.68}$$

∘

Beweis. Wir betrachten nur den Fall, dass p_θ eine WDF ist und zeigen zunächst, dass $\mathbb{E}_\theta(S(\theta)) = 0$ ist.

$$\begin{aligned}
\mathbb{E}_\theta(S(\theta)) &= \int S(\theta) p_\theta(x)\,dx \\
&= \int \frac{d}{d\theta}\ell(\theta) p_\theta(x)\,dx \\
&= \int \frac{d}{d\theta}\ln L(\theta) p_\theta(x)\,dx \\
&= \int \frac{1}{L(\theta)}\frac{d}{d\theta} L(\theta) p_\theta(x)\,dx \\
&= \int \frac{1}{p_\theta(x)}\frac{d}{d\theta} L(\theta) p_\theta(x)\,dx \\
&= \int \frac{d}{d\theta} L(\theta)\,dx \\
&= \frac{d}{d\theta}\int p_\theta(x)\,dx \\
&= \frac{d}{d\theta} 1 \\
&= 0.
\end{aligned} \tag{19.69}$$

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.

$$\begin{aligned}
J(\theta) &= \mathbb{E}_\theta\left(-\frac{d^2}{d\theta^2}\ln L(\theta)\right) \\
&= \mathbb{E}_\theta\left(-\frac{d}{d\theta}\frac{\frac{d}{d\theta}L(\theta)}{L(\theta)}\right) \\
&= \mathbb{E}_\theta\left(-\frac{\frac{d^2}{d\theta^2}L(\theta)L(\theta) - \frac{d}{d\theta}L(\theta)\frac{d}{d\theta}L(\theta)}{L(\theta)L(\theta)}\right) \\
&= -\mathbb{E}_\theta\left(\frac{\frac{d^2}{d\theta^2}L(\theta)}{L(\theta)}\right) + \mathbb{E}_\theta\left(\frac{\left(\frac{d}{d\theta}L(\theta)\right)^2}{(L(\theta))^2}\right) \\
&= -\int \frac{\frac{d^2}{d\theta^2}L(\theta)}{L(\theta)} p_\theta(x)\,dx + \int \frac{\left(\frac{d}{d\theta}L(\theta)\right)^2}{(L(\theta))^2} p_\theta(x)\,dx \\
&= -\frac{d^2}{d\theta^2}\int p_\theta(x)\,dx + \int \left(\frac{1}{L(\theta)}\frac{d}{d\theta}L(\theta)\right)^2 p_\theta(x)\,dx \\
&= -\frac{d^2}{d\theta^2} 1 + \int \left(\frac{d}{d\theta}\ln L(\theta)\right)^2 p_\theta(x)\,dx = \mathbb{E}_\theta\left(S(\theta)^2\right).
\end{aligned} \tag{19.70}$$

□

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 19.10 (Scorefunktion und Fisher-Informationen des Bernoullimodells)**.** *Es sei* $\upsilon = (\upsilon_1, ..., \upsilon_n)$ *die Stichprobe eines Bernoullimodells mit Parameter* $\mu \in]0, 1[$*. Dann gelten:*

- *Die Scorefunktion der Stichprobe ist*

$$S :]0, 1[\to \mathbb{R}, \mu \mapsto S(\mu) := \frac{1}{\mu} \sum_{i=1}^{n} y_i - \frac{1}{1-\mu} \left(n - \sum_{i=1}^{n} y_i \right). \tag{19.71}$$

- *Die Fisher-Information der Stichprobe ist*

$$I :]0, 1[\to \mathbb{R}, \mu \mapsto I(\mu) := I(\mu) = \frac{ny}{\mu^2} + \frac{n(1-y)}{1-\mu}^2. \tag{19.72}$$

- *Die beobachtete Fisher-Information der Stichprobe ist*

$$I :]0, 1[\to \mathbb{R}, \hat{\mu}^{ML} \mapsto I\left(\hat{\mu}^{ML}\right) := \frac{ny}{\hat{\mu}_n^{ML^2}} + \frac{n(1-y)}{1-\hat{\mu}^{ML}}. \tag{19.73}$$

- *Die erwartete Fisher-Information der Stichprobe ist*

$$J :]0, 1[\to \mathbb{R}, \mu \mapsto J(\mu) := \frac{n}{\mu(1-\mu)}. \tag{19.74}$$

∘

Beweis. Die Scorefunktion wurde bereits im Kontext der Maximum-Likelihood-Schätzung von μ hergeleitet. Wir betrachten die Fisher-Information einer einzelnen Bernoulli-Zufallsvariable υ.

$$\begin{aligned}
I(\mu) &:= -\frac{d^2}{d\mu^2} \ell_1(\mu) \\
&= -\frac{d^2}{d\mu^2} \ln p_\mu(y) \\
&= -\frac{d^2}{d\mu^2} \left(y \ln \mu + (1-y) \ln(1-\mu) \right) \\
&= -\frac{\partial}{\partial \mu} \left(\frac{\partial}{\partial \mu} \left(y \ln \mu + (1-y) \ln(1-\mu) \right) \right) \\
&= -\frac{\partial}{\partial \mu} \left(\frac{y}{\mu} + \frac{(1-y)}{1-\mu} \right) \\
&= -\left(-\frac{y}{\mu^2} - \frac{(1-y)}{1-\mu}^2 \right) \\
&= \frac{y}{\mu^2} + \frac{(1-y)}{1-\mu}^2.
\end{aligned} \tag{19.75}$$

Damit ergibt sich die erwartete Fisher-Information der Zufallsvariable υ als

$$\begin{aligned} J(\mu) &= \mathbb{E}_\mu(I(\mu)) \\ &= \mathbb{E}_\mu\left(\frac{\upsilon}{\mu^2} + \frac{(1-\upsilon)}{1-\mu}^2\right) \\ &= \frac{\mathbb{E}_\mu(\upsilon)}{\mu^2} + \frac{(1-\mathbb{E}_\mu(\upsilon))}{1-\mu}^2 \\ &= \frac{\mu}{\mu^2} + \frac{(1-\mu)}{1-\mu}^2 \\ &= \frac{1}{\mu(1-\mu)}. \end{aligned} \tag{19.76}$$

Mit der Additivitätseigenschaft der Fisher-Information und der Definition der beobachteten Fisher-Information ergibt sich dann sofort

$$I(\mu) = \frac{ny}{\mu^2} + \frac{n(1-y)}{1-\mu}^2 \text{ und } J(\mu) = \frac{n}{\mu(1-\mu)}. \tag{19.77}$$

□

Die Scorefunktion und die Fisher-Informationen des Normalverteilungsmodells betrachten wir lediglich unter der zusätzlichen Annahme eines bekannten Varianzparameters (Theorem 19.11) bzw. eines bekannten Erwartungswertparameters (Theorem 19.12)

Theorem 19.11 (Scorefunktion und Fisher-Informationen des Normalverteilungsmodells bei bekanntem Varianzparameter)**.** *$\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines Normalverteilungsmodells und der Varianzparameter σ^2 sei als bekannt vorausgesetzt. Dann gelten:*

- *Die Scorefunktion der Stichprobe ist*

$$S : \mathbb{R} \to \mathbb{R}, \mu \mapsto S(\mu) := \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu). \tag{19.78}$$

- *Die Fisher-Information der Stichprobe ist*

$$I : \mathbb{R} \to \mathbb{R}, \mu \mapsto I(\mu) := \frac{n}{\sigma^2}. \tag{19.79}$$

- *Die beobachtete Fisher-Information der Stichprobe ist*

$$I(\hat{\mu}_n^{ML}) = \frac{n}{\sigma^2}. \tag{19.80}$$

- *Die erwartete Fisher-Information der Stichprobe ist*

$$J : \mathbb{R} \to \mathbb{R}, \mu \mapsto J(\mu) := \frac{n}{\sigma^2}. \tag{19.81}$$

∘

Beweis. Wir erinnern uns, dass die Log-Likelihood-Funktion eines Normalverteilungsmodells bei bekanntem Varianzparameter σ^2 durch

$$\ell : \mathbb{R} \to \mathbb{R}, \mu \mapsto \ell(\mu) := -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{19.82}$$

gegeben ist. Damit ergibt sich die Scorefunktion als

$$S(\mu) = \frac{\partial}{\partial \mu}\ell(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) \tag{19.83}$$

Die Fisher-Information der Stichprobe ergibt sich als

$$I(\mu) = -\frac{d^2}{d\mu^2}\ell(\mu) = -\frac{\partial}{\partial \mu}S(\mu) = -\frac{1}{\sigma^2}\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{n} y_i - n\mu\right) = \frac{n}{\sigma^2}. \tag{19.84}$$

Die beobachtete Fisher-Information ist die Fisher-Information an der Stelle des Maximum-Likelihood Schätzes $\hat{\mu}_n^{\mathrm{ML}}$. Die erwartete Fisher-Information schließlich ergibt sich als

$$J(\mu) = \mathbb{E}_\mu(I(\mu)) = \mathbb{E}_\mu\left(\frac{n}{\sigma^2}\right) = \frac{n}{\sigma^2}. \tag{19.85}$$

□

Theorem 19.12 (Scorefunktion und Fisher-Informationen des Normalverteilungsmodells bei bekanntem Erwartungswertparameter)**.** *$\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines Normalverteilungsmodells und der Varianzparameter σ^2 sei als bekannt vorausgesetzt. Dann gelten:*

- *Die Scorefunktion der Stichprobe ist gegeben durch*

$$S : \mathbb{R}_{>0} \to \mathbb{R}, \sigma^2 \mapsto S(\sigma^2) := -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{19.86}$$

- *Die Fisher-Information der Stichprobe ist gegeben durch*

$$I : \mathbb{R}_{>0} \to \mathbb{R}, \sigma^2 \mapsto I(\sigma^2) := \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2 - \frac{n}{2\sigma^4} \tag{19.87}$$

- *Die beobachtete Fisher-Information der Stichprobe ist gegeben durch*

$$I(\hat{\sigma}_n^{2\,ML}) = \frac{n}{2\hat{\sigma}_{ML}^4} \tag{19.88}$$

- *Die erwartete Fisher-Information der Stichprobe ist gegeben durch*

$$J : \mathbb{R}_{>0} \to \mathbb{R}, \sigma^2 \mapsto J(\sigma^2) := \frac{n}{2\sigma^4}. \tag{19.89}$$

∘

Beweis. Wir erinnern uns, dass die Log-Likelihood-Funktion der Stichprobe eines Normalverteilungsmodells bei bekanntem Erwartungswert-Parameter μ durch

$$\ell : \mathbb{R}_{>0} \to \mathbb{R}, \sigma^2 \mapsto \ell(\sigma^2) := -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2. \tag{19.90}$$

gegeben ist. Die Scorefunktion ergibt sich also als

$$S(\sigma^2) = \frac{\partial}{\partial \sigma^2}\ell(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2. \tag{19.91}$$

Die Fisher-Information der Stichprobe ergibt sich als

$$\begin{aligned} I(\sigma^2) &= -\frac{\partial}{\partial \sigma^2} S(\sigma^2) \\ &= -\frac{\partial}{\partial \sigma^2} \left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \\ &= \frac{1}{\sigma^6} \sum_{i=1}^{n} (y_i - \mu)^2 - \frac{n}{2\sigma^4}. \end{aligned} \tag{19.92}$$

Die beobachtete Fisher-Information ist die Fisher-Information an der Stelle des Maximum-Likelihood Schätzes $\hat{\sigma^2}^{\text{ML}}$,

$$\begin{aligned} I(\hat{\sigma}_n^{2\,\text{ML}}) &= \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{\left(\hat{\sigma}_n^{2\,\text{ML}}\right)^3} - \frac{n}{2\left(\hat{\sigma}_n^{2\,\text{ML}}\right)^2} \\ &= \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{\frac{1}{n^3}\left(\sum_{i=1}^{n} (y_i - \mu)^2\right)^3} - \frac{n}{2\left(\hat{\sigma}_n^{2\,\text{ML}}\right)^2} \\ &= \frac{1}{\frac{1}{n^3}\left(\sum_{i=1}^{n} (y_i - \mu)^2\right)^2} - \frac{n}{2\left(\hat{\sigma}_n^{2\,\text{ML}}\right)^2} \\ &= \frac{n}{\left(\hat{\sigma}_n^{2\,\text{ML}}\right)^2} - \frac{n}{2\left(\hat{\sigma}_n^{2\,\text{ML}}\right)^2} \\ &= \frac{n}{2\left(\hat{\sigma}_n^{2\,\text{ML}}\right)^2} \\ &= \frac{n}{2\hat{\sigma}_n^{4\,\text{ML}}}. \end{aligned} \tag{19.93}$$

Die erwartete Fisher-Information ergibt sich schließlich als

$$\begin{aligned} J(\sigma^2) &= \mathbb{E}_{\sigma^2}(I(\sigma^2)) \\ &= \mathbb{E}_{\sigma^2}\left(\frac{1}{\sigma^6} \sum_{i=1}^{n} (y_i - \mu)^2 - \frac{n}{2\sigma^4} \right) \\ &= \frac{1}{\sigma^6} \sum_{i=1}^{n} \mathbb{E}_{\sigma^2}\left((y_i - \mu)^2\right) - \frac{n}{2\sigma^4} \\ &= \frac{1}{\sigma^6} \sum_{i=1}^{n} \sigma^2 - \frac{n}{2\sigma^4} \\ &= \frac{n\sigma^2}{\sigma^6} - \frac{n}{2\sigma^4} \\ &= \frac{n}{\sigma^4} - \frac{n}{2\sigma^4} \\ &= \frac{n}{2\sigma^4}. \end{aligned} \tag{19.94}$$

□

Mit den oben diskutierten Eigenschaften der Scorefunktion können wir nun die Cramér-Rao-Ungleichung formulieren und beweisen.

Theorem 19.13 (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*

$$\mathbb{V}_\theta(\hat{\tau}) \geq \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{J(\theta)}. \tag{19.95}$$

Speziell gilt für einen Parameterschätzer mit $\tau(\theta) := \theta$ und somit

$$\hat{\tau} = \hat{\theta} \text{ und } \left(\frac{d}{d\theta}\tau(\theta)\right)^2 = 1, \tag{19.96}$$

dass

$$\mathbb{V}_\theta(\hat{\theta}) \geq \frac{1}{J(\theta)}. \tag{19.97}$$

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 14.4) und der Identität von $\mathbb{V}_\theta(S(\theta))$ und $J(\theta)$ (Theorem 19.9) gilt, dass

$$\frac{\mathbb{C}_\theta(S(\theta), \hat{\tau})^2}{\mathbb{V}_\theta(S(\theta))\mathbb{V}_\theta(\hat{\tau})} \leq 1 \Leftrightarrow \mathbb{V}_\theta(\hat{\tau}) \geq \frac{\mathbb{C}_\theta(S(\theta), \hat{\tau})^2}{J(\theta)}. \tag{19.98}$$

Mit dem Kovarianzverschiebungssatz (Theorem 13.10), der Tatsache, dass der Erwartungswert der Scorefunktion immer Null ist (Theorem 19.9) und der vorausgesetzten Erwartungstreue von $\hat{\tau}$ ergibt sich dann zunächst

$$\begin{aligned}
\mathbb{C}_\theta(S(\theta), \hat{\tau}) &= \mathbb{E}_\theta(S(\theta)\hat{\tau}) - \mathbb{E}_\theta(S(\theta))\mathbb{E}_\theta(\hat{\tau}) \\
&= \mathbb{E}_\theta(S(\theta)\hat{\tau}) \\
&= \int S(\theta)\,\hat{\tau}\,p_\theta(x)\,dx \\
&= \int \frac{d}{d\theta}\ln L(\theta)\,\hat{\tau}\,p_\theta(x)\,dx \\
&= \int \frac{\frac{d}{d\theta}L(\theta)}{L(\theta)}\,\hat{\tau}\,p_\theta(x)\,dx \\
&= \int \frac{\frac{d}{d\theta}L(\theta)}{p_\theta(x)}\,\hat{\tau}\,p_\theta(x)\,dx \\
&= \int \frac{d}{d\theta}L(\theta)\,\hat{\tau}\,dx \\
&= \frac{d}{d\theta}\int L(\theta)\,\hat{\tau}\,dx \\
&= \frac{d}{d\theta}\int \hat{\tau}\,p_\theta(x)\,dx \\
&= \frac{d}{d\theta}\mathbb{E}_\theta(\hat{\tau}) \\
&= \frac{d}{d\theta}\tau(\theta).
\end{aligned} \tag{19.99}$$

Damit folgt dann aber direkt

$$\mathbb{V}_\theta(\hat{\tau}) \geq \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{J(\theta)}. \tag{19.100}$$

□

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

$$\mathbb{V}_\theta(\hat{\theta}) = \frac{1}{J(\theta)} \tag{19.101}$$

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 19.4 zurück.

19.3. 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 19.2 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 19.3.1 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 $n \to \infty$ mit dem wahren, aber unbekannten, Wert $\tau(\theta)$ identisch ist. In Kapitel 19.3.2 führen wir den Begriff der *Konsistenz* eines Schätzers ein. Intuitiv heißt ein Schätzer $\hat{\tau}_n$ für τ *konsistent*, wenn für große Stichprobenumfänge $n \to \infty$ die Wahrscheinlichkeit dafür, dass $\hat{\tau}_n(\upsilon)$ vom wahren, aber unbekannten, Wert $\tau(\theta)$ abweicht, beliebig klein wird. Für große Stichprobenumfänge resultieren die Verteilungen von Schätzern oft in Normalverteilungen. In Kapitel 19.3.3 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 $n \to \infty$, die Verteilung von $\hat{\tau}_n$ durch eine Normalverteilung gegeben ist. In Kapitel 19.3.4 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 $n \to \infty$ die Verteilung von $\hat{\tau}_n$ durch eine Normalverteilung mit Erwartungswertparameter $\tau(\theta)$ und Varianzparameter gleich der Cramér-Rao-Schranke gegeben ist.

19.3.1. Asymptotische Erwartungstreue

Die Asymptotische Erwartungstreue eines Schätzers verfeinert den Begriff des erwartungstreuen Schätzers wie folgt.

Definition 19.9 (Asymptotische Erwartungstreue)**.** $\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}_n$ sei ein Schätzer für τ. $\hat{\tau}_n$ heißt *asymptotisch erwartungstreu*, wenn gilt, dass

$$\lim_{n\to\infty} \mathbb{E}_\theta(\hat{\tau}_n(\upsilon)) = \tau(\theta) \text{ für alle } \theta \in \Theta. \tag{19.102}$$

•

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 19.14 (Asymptotische Erwartungstreue des Varianzparameterschätzers)**.** *$\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines Normalverteilungsmodells mit Varianzparameter σ^2. Dann ist der Maximum-Likelihood Schätzer von σ^2*

$$\hat{\sigma}_n^{2^{ML}} := \frac{1}{n}\sum_{i=1}^{n}\left(\upsilon_i - \bar{\upsilon}_n\right)^2 \tag{19.103}$$

nicht erwartungstreu, aber asymptotisch erwartungstreu.

∘

Beweis. Mit der Erwartungstreue der Stichprobenvarianz ergibt sich zunächst (vgl. Theorem 19.3)

$$\mathbb{E}_{\mu,\sigma^2}\left(\hat{\sigma}_n^{2^{\mathrm{ML}}}\right) = \mathbb{E}_{\mu,\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\upsilon_i - \bar{\upsilon}_n\right)^2\right) = \frac{1}{n}\mathbb{E}_{\mu,\sigma^2}\left(\sum_{i=1}^{n}\left(\upsilon_i - \bar{\upsilon}_n\right)^2\right) = \frac{n-1}{n}\sigma^2. \tag{19.104}$$

Also gilt

$$\mathbb{E}_{\mu,\sigma^2}\left(\hat{\sigma}_n^{2^{\mathrm{ML}}}\right) \neq \sigma^2 \tag{19.105}$$

und $\hat{\sigma}_n^{2^{\mathrm{ML}}}$ kein erwartungstreuer Schätzer von σ^2. Allerdings gilt auch

$$\frac{n-1}{n} \to 1 \text{ für } n \to \infty, \tag{19.106}$$

so dass

$$\lim_{n\to\infty}\mathbb{E}_{\mu,\sigma^2}\left(\hat{\sigma}_n^{2^{\mathrm{ML}}}\right) = \lim_{n\to\infty}\frac{n-1}{n}\sigma^2 = \sigma^2 \lim_{n\to\infty}\frac{n-1}{n} = \sigma^2 \tag{19.107}$$

gilt und der Maximum-Likelihood Schätzer von σ^2 damit asymptotisch erwartungstreu ist.

□

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(NaN, length(n)*ns),ncol = length(n)) # \hat{\sigma^2}^{ML} Array

# Simulation
for(i in seq_along(n)){                                # Stichprobenumfangsiterationen
    for(s in 1:ns){                                    # Stichprobenrealisierungsiterationen
        y              = rnorm(n[i], mu, sqrt(sigsqr)) # Stichprobenrealisation
        sigsqr_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 19.3. Für kleine Stichprobenumfänge unterschätzt der Schätzer den wahren, aber unbekannten, Varianzparameter deutlich, für größere Stichprobenumfänge dagegen nicht.

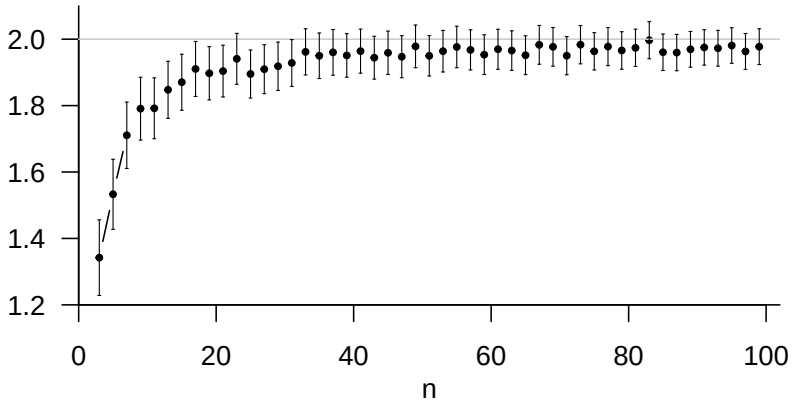


Abbildung 19.3. Simulation des Erwartungswerts des Maximum-Likelihood Schätzers für den Varianzparameter eines Normalverteilungsmodells. Bei kleinen Stichprobenumfängen unterschätzt der Maximum-Likelihood Schätzer den wahren, aber unbekannten, Varianzparameter systematisch, bei größeren Stichprobenumfängen ist der Schätzer in etwa erwartungstreu.

19.3.2. Konsistenz

Der Begriff der Konsistenz eines Schätzers verallgemeinert das Schwache Gesetz der Großen Zahlen von Stichprobenmitteln und Erwartungswerten (vgl. Theorem 15.1) auf beliebige Schätzer und wahre, aber unbekannte, Parameterwerte.

Definition 19.10 (Konsistenz)**.** $\upsilon := (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}_n$ sei ein Schätzer von τ. Dann heißt eine Folge von Schätzern $\hat{\tau}_1, \hat{\tau}_2, ...$ eine *konsistente Folge von Schätzern*, wenn für jedes noch so kleine $\epsilon > 0$ und jedes $\theta \in \Theta$ gilt, dass

$$\lim_{n \to \infty} \mathbb{P}_\theta \left(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon \right) = 0. \tag{19.108}$$

Wenn $\hat{\tau}_1, \hat{\tau}_2, ...$ eine konsistente Folge von Schätzern ist, dann heißt $\hat{\tau}_n$ ein *konsistenter Schätzer*.

•

Die Intuition zu Definition 19.10 entspricht der dem Gesetz der Großen Zahlen. Für Stichprobenumfänge mit $n \to \infty$ wird die Wahrscheinlichkeit, dass $\hat{\tau}_n(\upsilon)$ beliebig nah bei $\tau(\theta)$ liegt bzw. die Wahrscheinlichkeit, dass $\hat{\tau}_n(\upsilon)$ 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 15.1 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
            }
        }
    }
}

# Schaetzung von \mathbb{P}(|\hat{\tau}_n(\ups)-\tau(\theta)| \ge \epsilon)
P_hat       = apply(E, c(1,2), mean)
```

Wir visualisieren die Schätzung der Wahrscheinlichkeit $\mathbb{P}(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon)$ für dieses Beispiel in Abbildung 19.4 in Abhängigkeit des Stichprobenumfangs und des Kriteriumwerts 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.

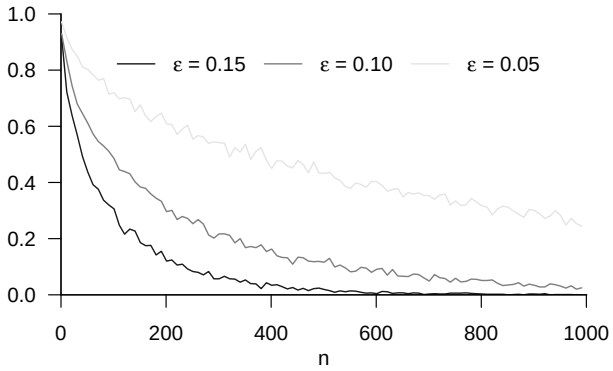


Abbildung 19.4. Simulation der Konsistenz des Stichprobenmittels als Schätzer für den Erwartungswertparameter des Normalverteilungsmodells.

Die in Theorem 19.15 und Theorem 19.16 angegebenen Kriterien für die Konsistenz von Schätzern vereinfachen den Nachweis der Konsistenz eines Schätzers mithilfe des Mittleren Quadratischen Fehlers (Definition 19.6).

Theorem 19.15 (Mittlerer Quadratischer Fehler Kriterium für Konsistenz)**.** *$\upsilon := (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}_n$ sei ein Schätzer von τ. Wenn gilt, dass*

$$\lim_{n \to \infty} MQF(\hat{\tau}_n) = 0, \tag{19.109}$$

dann ist $\hat{\tau}_n$ ein konsistenter Schätzer.

∘

Beweis. Mit der Chebychev-Ungleichung gilt, dass

$$\mathbb{P}_\theta\left(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon\right) \leq \frac{\mathbb{E}_\theta\left((\hat{\tau}_n(\upsilon) - \tau(\theta))^2\right)}{\epsilon^2} \tag{19.110}$$

Grenzwertbildung ergibt dann

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \lim_{n\to\infty} \mathbb{E}_\theta\left((\hat{\tau}_n(\upsilon) - \tau(\theta))^2\right). \tag{19.111}$$

Wenn also $\lim_{n\to\infty} \mathbb{E}_\theta\left((\hat{\tau}_n(\upsilon) - \tau(\theta))^2\right) = 0$ gilt, dann gilt mit $\mathbb{P}_\theta(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon) \geq 0$, dass

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(|\hat{\tau}_n(\upsilon) - \tau(\theta)| \geq \epsilon\right) = 0. \tag{19.112}$$

Also ist $\hat{\tau}_n$ ein konsistenter Schätzer.

□

Theorem 19.16 (Bias-Varianz-Kriterium für Konsistenz)**.** *$\upsilon := (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\tau}_n$ sei ein Schätzer von τ. Wenn*

$$\lim_{n\to\infty} B(\hat{\tau}_n) = 0 \text{ und } \lim_{n\to\infty} \mathbb{V}_\theta(\hat{\tau}_n) = 0 \tag{19.113}$$

gelten, dann ist $\hat{\tau}_n$ ein konsistenter Schätzer

∘

Beweis. Wenn $n \to \infty$, dann gilt $\mathrm{B}(\hat{\tau}_n) \to 0$, also auch $\mathrm{B}(\hat{\tau}_n)^2 \to 0$. Wenn für $n \to \infty$ sowohl $\mathrm{B}(\hat{\tau}_n)^2 \to 0$ als auch $\mathbb{V}_\theta(\hat{\tau}_n) \to 0$, dann gilt auch $\lim_{n\to\infty} \mathrm{MQF}(\hat{\theta}_n) = 0$. Also gilt mit Theorem 19.15, dass $\hat{\tau}_n$ konsistent ist.

□

Als Anwendung von Theorem 19.16 wollen wir mit folgendem Theorem die Konsistenz des Stichprobenmittels als Schätzer für den Erwartungswert bei Normalverteilung nachweisen.

Theorem 19.17 (Konsistenz des Erwartungswertschätzers bei Normalverteilung)**.** *Es sei υ die Stichprobe eines Normalverteilungsmodells. Dann ist $\bar{\upsilon}_n$ ein konsistenter Schätzer von $\mathbb{E}(\upsilon_1)$.*

∘

Beweis. Mit der Erwartungstreue des Stichprobenmittels als Schätzer für den Erwartungswert gilt zunächst

$$\lim_{n\to\infty} \mathrm{B}(\bar{\upsilon}_n) = 0. \tag{19.114}$$

Weiterhin gilt mit der Varianz des Stichprobenmittels

$$\lim_{n\to\infty} \mathbb{V}_\theta(\bar{\upsilon}_n) = \lim_{n\to\infty} \frac{1}{n}\mathbb{V}(\upsilon_1) = 0. \tag{19.115}$$

Mit dem Bias-Varianz-Kriterium folgt dann die Konsistenz von $\bar{\upsilon}_n$ als Schätzer von $\mathbb{E}(\upsilon_1)$

□

Man beachte, dass Theorem 19.17 natürlich auch schon im Schwachen Gesetz der Großen Zahlen impliziert ist (vgl. Theorem 15.1).

19.3.3. 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 19.11 (Asymptotisch normalverteilter Schätzer)**.** $\upsilon = (\upsilon_1, ..., \upsilon_n)$ sei die Stichprobe eines parametrischen Produktmodells und $\hat{\theta}_n$ sei ein Parameterschätzer für θ. Weiterhin sei

$$\tilde{\theta} \sim N(\mu, \sigma^2) \tag{19.116}$$

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

$$\lim_{n\to\infty} P_n(\hat{\theta}_n) = P(\tilde{\theta}), \tag{19.117}$$

dann heißt $\hat{\theta}_n$ *asymptotisch normalverteilt* und wir schreiben

$$\hat{\theta}_n \overset{a}{\sim} N(\mu, \sigma^2). \tag{19.118}$$

•

Als Beispiel für einen asymptotisch normalverteilten Schätzer betrachten wir in Kapitel 19.3.4 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.

19.3.4. 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 19.12. $\upsilon = (\upsilon_1, ..., \upsilon_n)$ 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 υ. Wenn gilt, dass

$$\hat{\theta}_n \overset{a}{\sim} N\left(\theta, J(\theta)^{-1}\right), \tag{19.119}$$

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

```r
# Modellformulierung
mu          = 0.4                                          # wahrer, aber unbekannter, Parameterwert
n_all       = c(1e1,5e1,1e2)                               # Stichprobenumfang
ns          = 1e4                                          # Anzahl Stichprobenrealisierungen
mu_hat   = matrix(rep(NaN, length(n_all)*ns), nrow = length(n_all))   # Maximum-Likelihood-Schätzerarray
mu_hat_r = 1e3                                             # Maximum-Likelihood Schätzerraumaufloesung
mu_hat_y = seq(0,1,len = mu_hat_r)                         # Maximum-LikelihoodSchätzerraum
mu_hat_p = matrix(rep(NaN, length(n_all)*mu_hat_r), nrow = length(n_all))   # Maximum-Likelihood WDF Array

# Stichprobenumfängeiterationen
for(i in seq_along(n_all)){

    # Simulationsiterationen
    for(s in 1:ns){
        y           = rbinom(n_all[i],1,mu)                # Stichprobenrealisation
        mu_hat[i,s] = mean(y)                              # Maximum-Likelihood-Schätzerrealisation
    }
    mu_hat_p[i,] = dnorm(mu_hat_y, mu, sqrt(mu*(1-mu)/n_all[i]))   # WDF der asymptotischen Verteilung
}
```

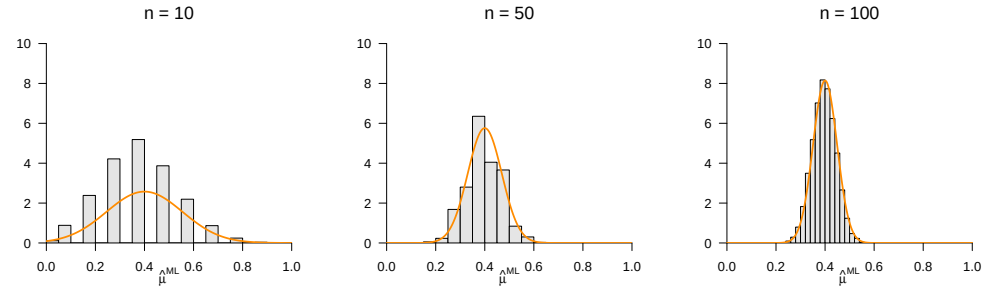


Abbildung 19.5. Simulation der asymptotischen Effizienz des Maximum-Likelihood-Parameterschätzers für den Parameter eines Bernoullimodells. Mit steigendem Stichprobenumfang gleicht sich die durch ein Histogramm dargestellte Verteilung der simulierten Schätzerwerte der in Definition 19.12 formulierten Normalverteilung

19.4. 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 19.18 (Eigenschaften von Maximum-Likelihood Schätzern)**.** *υ 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 19.18 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 19.18 Schätzergütekriterien.

19.5. Literaturhinweise

Die in diesem Kapitel vorgestellten Resultate gehen in ganz wesentlicher Weise auf Fisher (1922) zurück. Aldrich (1997) gibt dazu eine historische Einordnung.

19.6. Selbstkontrollfragen

1. Geben Sie die Definition des Begriffs eines Parameterpunktschätzers wieder.
2. Erläutern Sie den Begriff des Parameterpunktschätzers.
3. Geben Sie Definition der Begriffe der Likelihood-Funktion und der Log-Likelihood-Funktion wieder.
4. Geben Sie Definition des Begriffs des Maximum-Likelihood Schätzes wieder.
5. Erläutern Sie das Vorgehen zur Gewinnung von Maximum-Likelihood-Schätzern.
6. Geben Sie das Theorem zum Maximum-Likelihood-Schätzer des Bernoullimodellparameters wieder.
7. Geben Sie das Theorem zu den Maximum-Likelihood-Schätzern der Normalverteilungsmodellparameter wieder.
8. Geben Sie die Definition des Begriffs der Erwartungstreue eines Schätzers wieder.
9. Erläutern Sie den Begriff der Erwartungstreue eines Schätzers.
10. Geben Sie Definition der Begriffe der Varianz und des Standardfehlers eines Schätzers wieder.
11. Erläutern Sie den Begriff der asymptotischen Erwartungstreue eines Schätzers.
12. Erläutern Sie den Begriff der Konsistenz eines Schätzers.
13. Erläutern Sie den Begriff der asymptotischen Normalität eines Schätzers.
14. Geben Sie das Theorem zu den Eigenschaften von Maximum-Likelihood-Schätzern wieder.

20. 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 $\hat{\theta} = 2.14$ ist diese Schätzung zum Beispiel vor dem Hintergrund der verschwindenen Wahrscheinlichkeit einer normalverteilten Zufallsvariable wie dem Stichprobenmittel im Normalverteilungsmodell genau einen reellen Wert anzunehmen mit sehr 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 wie im folgenden thematisiert so konstruiert werden kann, dass sie mit einer hohen gewünschten Wahrscheinlichkeit, also, vor der Form eines geeigneten Frequentistischen Inferenzmodell nur mit einem geringes Maß an Unsicherheit assoziiert ist. 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 Konfidenz*intervalle*, d.h. Teilmengen von $\mathbb{R}$. Eine Generalisierung der hier vorgestellten Konzepte auf höher dimensionale Parameterräume im Sinne von Konfidenz*mengen* ist jedoch relativ unproblematisch möglich.

20.1. Definition

Definition 20.1 (δ-Konfidenzintervall)**.** Es sei υ die Stichprobe eines Frequentistischen Inferenzmodells mit wahrem, aber unbekannten Parameter, $\theta \in \Theta$, es sei $\delta \in]0, 1[$ und es seien $G_u(\upsilon)$ und $G_o(\upsilon)$. Dann heißt ein Intervall der Form

$$\kappa(\upsilon) := [G_u(\upsilon), G_o(\upsilon)], \tag{20.1}$$

so dass

$$\mathbb{P}_\theta\left(\kappa(\upsilon) \ni \theta\right) = \mathbb{P}_\theta\left(G_u(\upsilon) \leq \theta \leq G_o(\upsilon)\right) = \delta \text{ für alle } \theta \in \Theta \text{ gilt} \tag{20.2}$$

ein *δ-Konfidenzintervall für θ*. δ ist die Überdeckungswahrscheinlichkeit von $\kappa(\upsilon)$ für θ und wird meist *Konfidenzlevel* genannt. Die Statistiken $G_u(\upsilon)$ und $G_o(\upsilon)$ heißen die unteren und oberen Grenzen des Konfidenzintervalls, respektive.

•

Man beachte in Definition 20.1 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 $\kappa(\upsilon) \ni \theta$

bedeutet schlicht $\theta \in \kappa(\upsilon)$. Da $\kappa(\upsilon)$ in dem Ausdruck $\mathbb{P}_\theta(\kappa(\upsilon) \ni \theta)$ wie beschrieben die zufällige Entität ist, steht $\kappa(\upsilon)$ 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 20.1 jedoch auch für jeden beliebigen wahren, aber unbekannten, Parameterwert $\theta_i, i = 1, 2, ...$ Auch wenn man also unterschiedliche, wahre, aber unbekannte, Parameterwerte $\theta_1, \theta_2, ...$ 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 20.1 zu bestimmen, auch im 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 $G_u(\upsilon)$ und $G_o(\upsilon)$ zu konstruieren, geht man vor wie folgt. Zunächst definiert man das Frequentistische Inferenzmodell und legt damit die Verteilung der Stichprobe υ fest. In einem zweiten Schritt definiert man eine Statistik, also eine Funktion der Stichprobe, die als Grundlage für $G_u(\upsilon)$ und $G_o(\upsilon)$ 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.

20.2. 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 20.2 (T-Konfidenzintervallstatistk)**.** Gegeben sei das Normalverteilungsmodell

$$\upsilon_1, ..., \upsilon_n \sim N\left(\mu, \sigma^2\right) \tag{20.3}$$

Dann heißt die mit dem Stichprobenmittel und der Stichprobenstandardabweichung

$$\bar{\upsilon} := \frac{1}{n}\sum_{i=1}^{n}\upsilon_i \text{ und } S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon})^2}, \tag{20.4}$$

definierte Statistik

$$T := \sqrt{n}\frac{\bar{\upsilon} - \mu}{S} \tag{20.5}$$

T-Konfidenzintervallstatistik.

•

Für die Verteilung der T-Konfidenzintervallstatistik gilt folgendes Theorem.

Theorem 20.1 (Verteilung der T-Konfidenzintervallstatistik)**.** *Die T-Konfidenzintervallstatistik ist eine t-verteilte Zufallsvariable mit Parameter $n-1$, es gilt also*

$$T \sim t(n-1) \tag{20.6}$$

∘

Beweis. □

Man beachte, dass die T-Konfidenzintervallstatistik nach Definition 20.2 eine Funktion der Stichprobe ist, während ihre Verteilung nach Theorem 20.1 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 Zufallvariable mit t, die KVF einer t-verteilten Zufallvariable mit Ψ und die inverse KVF einer t-verteilten Zufallvariable mit Ψ^{-1} bezeichnen. Folgender **R** Code simuliert zunächst die Verteilung der T-Konfidenzintervallstatistitk.

```
# Modellformulierung
mu       = 10                          # wahrer, aber unbekannter, Erwartungswertparameter
sigsqr   = 4                           #  wahrer, aber unbekannter, Varianzparameter
n        = 12                          # Stichprobenumfang
ns       = 1e4                         # Anzahl Stichprobenrealisierungen
res      = 1e3                         # Ausgangsraumaufloesung

# analytische Definitionen und Resultate
yx       = seq(3,17,len = res)         # \upsilon_i Raum
ssqrx    = seq(0,20,len = res)         # S^2 Raum
tx       = seq(-4,4,len = res)         # T Raum
p_y_i    = dnorm(yx,mu,sqrt(sigsqr))   # \upsilon_i WDF
p_y_bar  = dnorm(yx,mu,sqrt(sigsqr/n)) # \upsilon_bar WDF
p_sqr    = dchisq(ssqrx,n-1)           # S^2 WDF
p_t      = dt(tx,n-1)                  # T WDF

# Simulation
y_i      = rep(NaN,ns)                 # y_i Array
y_bar    = rep(NaN,ns)                 # \bar{y} Array
S        = rep(NaN,ns)                 # S Array
TKS      = rep(NaN,ns)                 # T-Konfidenzintervallstatistik Array
for(s in 1:ns){                        # Simulationsiterationen
  y         = rnorm(n,mu,sqrt(sigsqr)) # Stichprobenrealisierung
  y_i[s]    = y[1]                     # Stichprobenrealisierung \upsilon_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
```

```
28  }
```

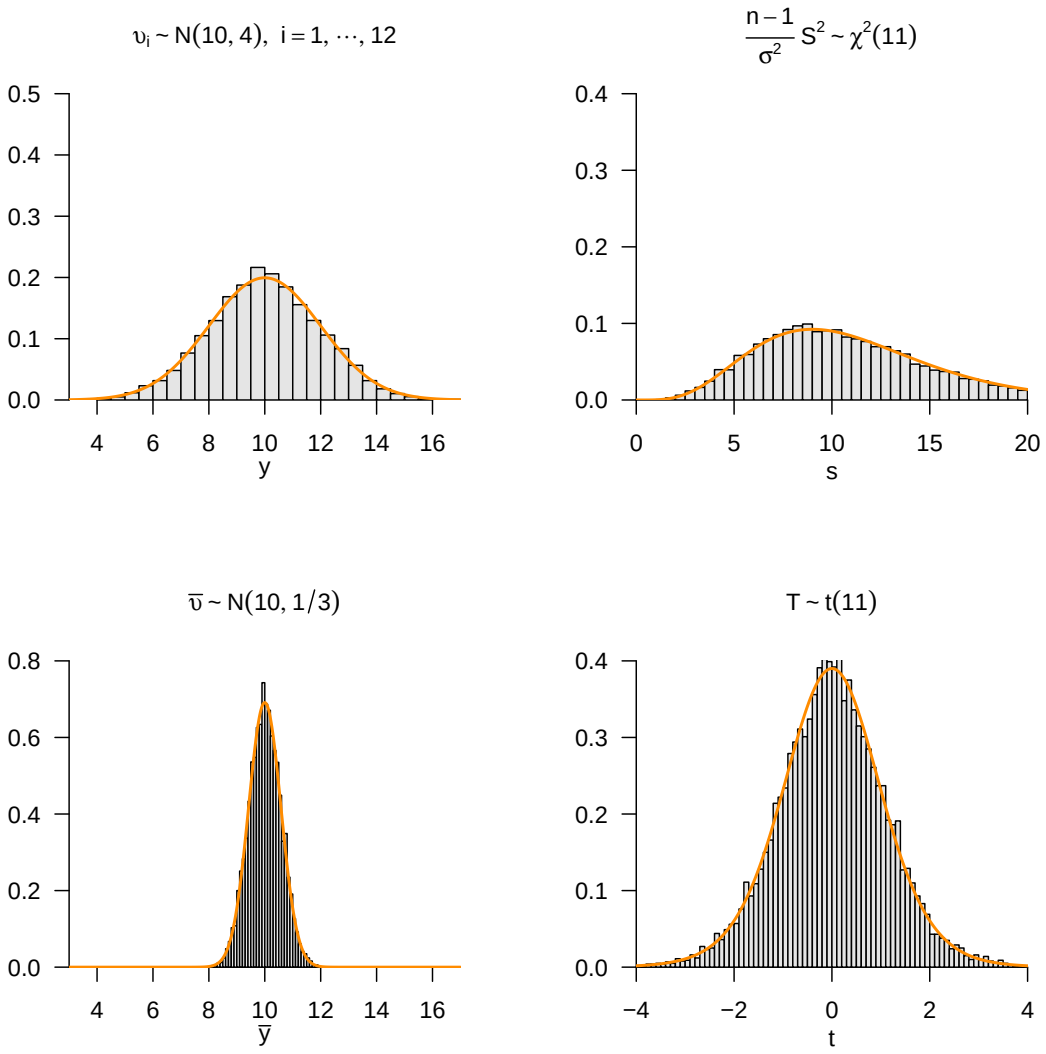


Abbildung 20.1. Simulation der Verteilung der T-Konfidenzintervallstatistik und der ihr zugrundeliegenden Verteilungen der Stichprobenvariable, des Stichprobenmittels und der Stichprobenvarianz.

In Abbildung 20.1 visualisieren wir die Verteilung der T-Konfidenzintervallstatistik als Resultat der ihr zugrundeliegenden Verteilungen der Stichprobenvariablen, des Stichprobenmittels und der Stichprobenvarianz (vgl. Kapitel 17.4). Mithilfe der Verteilung der T-Konfidenzintervallstatistik können wir jetzt folgendes Theorem zum Konfidenzintervall für den Erwartungswertparameter des Normalverteilungsmodells beweisen.

Theorem 20.2 (Konfidenzintervall für den Erwartungswertparameter des Normalverteilungsmodells)**.** *Gegeben sei das Normalverteilungsmodell*

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2) \tag{20.7}$$

mit wahren, aber unbekannten, Parametern μ und σ^2, es sei $\delta \in]0,1[$ und es sei

$$t_\delta := \Psi^{-1}\left(\frac{1+\delta}{2}; n-1\right). \tag{20.8}$$

mit der inversen KVF Ψ^{-1} einer t-verteilten Zufallsvariable. Dann gilt für das Intervall

$$\kappa(\upsilon) := \left[\bar{\upsilon} - \frac{S}{\sqrt{n}}t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}}t_\delta\right], \tag{20.9}$$

mit dem Stichprobenmittel und der Stichprobenstandardabweichung

$$\bar{\upsilon} := \frac{1}{n}\sum_{i=1}^{n}\upsilon_i \text{ und } S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon})^2}, \tag{20.10}$$

respektive, dass

$$\mathbb{P}_\mu(\kappa(\upsilon) \ni \mu) = \delta. \tag{20.11}$$

∘

Beweis. Für $\delta \in]0,1[$ seien zunächst

$$t_1 := \Psi^{-1}\left(\frac{1-\delta}{2}; n-1\right) \text{ und } t_2 := \Psi^{-1}\left(\frac{1+\delta}{2}; n-1\right) \tag{20.12}$$

definiert. Dann gilt

$$\frac{1+\delta}{2} - \frac{1-\delta}{2} = \delta \tag{20.13}$$

und weiterhin gilt mit der Symmetrie der WDF der t-Verteilung, dass

$$t_1 = -t_2. \tag{20.14}$$

Per Definition gilt dann aber mit Definition 20.2 und Theorem 20.1, dass

$$\mathbb{P}_\mu\left(-t_\delta \leq T \leq t_\delta\right) = \delta. \tag{20.15}$$

Damit folgt dann aber direkt

$$\begin{aligned}
\delta &= \mathbb{P}_\mu\left(-t_\delta \leq T \leq t_\delta\right) \\
&= \mathbb{P}_\mu\left(-t_\delta \leq \frac{\sqrt{n}}{S}(\bar{\upsilon} - \mu) \leq t_\delta\right) \\
&= \mathbb{P}_\mu\left(-\frac{S}{\sqrt{n}}t_\delta \leq \bar{\upsilon} - \mu \leq \frac{S}{\sqrt{n}}t_\delta\right) \\
&= \mathbb{P}_\mu\left(-\bar{\upsilon} - \frac{S}{\sqrt{n}}t_\delta \leq -\mu \leq -\bar{\upsilon} + \frac{S}{\sqrt{n}}t_\delta\right) \\
&= \mathbb{P}_\mu\left(\bar{\upsilon} + \frac{S}{\sqrt{n}}t_\delta \geq \mu \geq \bar{\upsilon} - \frac{S}{\sqrt{n}}t_\delta\right) \\
&= \mathbb{P}_\mu\left(\bar{\upsilon} - \frac{S}{\sqrt{n}}t_\delta \leq \mu \leq \bar{\upsilon} + \frac{S}{\sqrt{n}}t_\delta\right) \\
&= \mathbb{P}_\mu\left(\left[\bar{\upsilon} - \frac{S}{\sqrt{n}}t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}}t_\delta\right] \ni \mu\right). \\
&= \mathbb{P}_\mu\left(\kappa(\upsilon) \ni \mu\right).
\end{aligned} \tag{20.16}$$

□

Der entscheidene Schritt zur Sicherung der Überdeckungswahrscheinlichkeit δ des wahren, aber unbekannten, Erwartungswertparameters durch das in Theorem 20.2 definierte Konfidenzintervall ist die Definition von

$$t_\delta := \Psi^{-1}\left(\frac{1+\delta}{2}; n-1\right). \tag{20.17}$$

Wie im Beweis von Theorem 20.2 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 20.2. In diesem Fall ergibt sich

$$-t_\delta = \Psi^{-1}(0.025; 4) = -2.57 \text{ und } t_\delta = \Psi^{-1}(0.975; 4) = 2.57. \tag{20.18}$$

Abbildung 20.2 A zeigt diese Wahl aus Perspektive der WDF der T-Konfidenintervallstatistik. 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 20.2 B zeigt die entsprechende Perspektive der KVF der T-Konfidenintervallstatistik. 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 20.1 nicht implizit ist, sondern sich aus den Gegebenheiten der Verteilung der T-Konfidenzintervallstatistik, insbesondere ihrer Symmetrie um 0, ergibt.

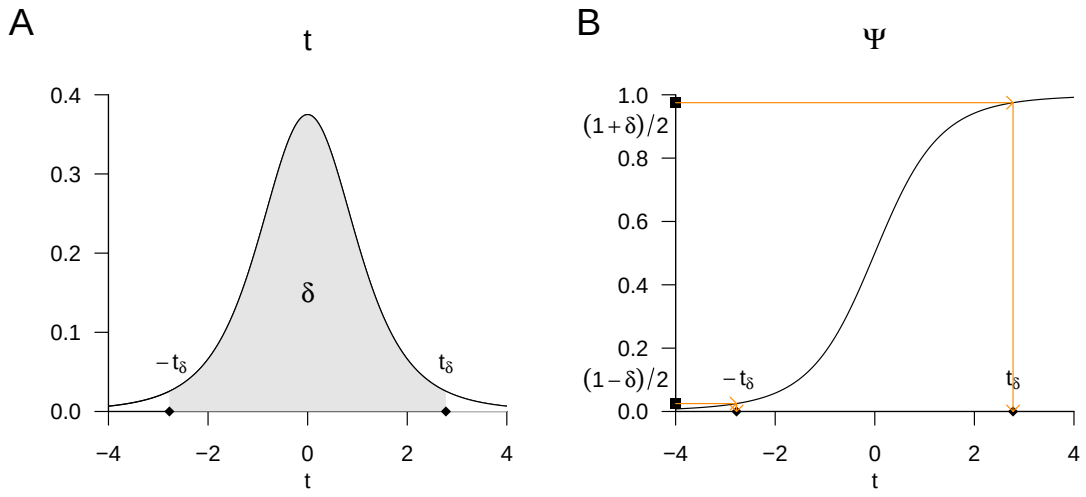


Abbildung 20.2. Sicherung der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Erwartungswertparameter des Normalverteilungsmodells für $\delta := 0.95$ und $n := 5$ aus Perspektive der WDF (A) und der KVF (B) der Verteilung der T-Konfidenzintervallstatistik.

Abschließend wollen wir die Überdeckungswahrscheinlichkeit des durch Theorem 20.2 gegebenen Konfidenzintervalls mithilfe einer Simulation demonstrieren. Wir betrachten dabei lediglich die erste Interpretation eines Konfidenzintervalls bei konstantem, wahrem, aber unbekanntem, 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(NaN,ns)                        # Stichprobenmittelarray
S       = rep(NaN,ns)                        # Standardabweichungsarray
kappa   = matrix(rep(NaN,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 20.3.

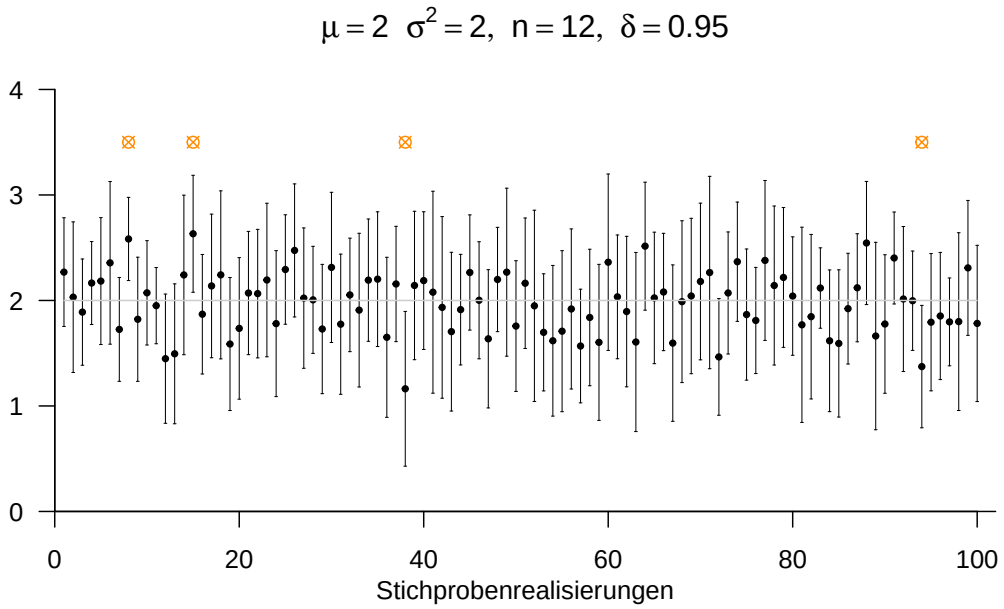


Abbildung 20.3. Simulation der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Erwartungswertparameter des Normalverteilungsmodells bei konstanten, wahren, aber unbekannten, Erwartungswertparameter $\mu := 2$ für $\sigma^2 := 2, n := 12$ und einer gewünschten Überdeckungswahrscheinlichkeit 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 einen 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(NaN,ns)                        # Stichprobenmittelarray
S       = rep(NaN,ns)                        # Standardabweichungsarray
kappa   = matrix(rep(NaN,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 20.4.

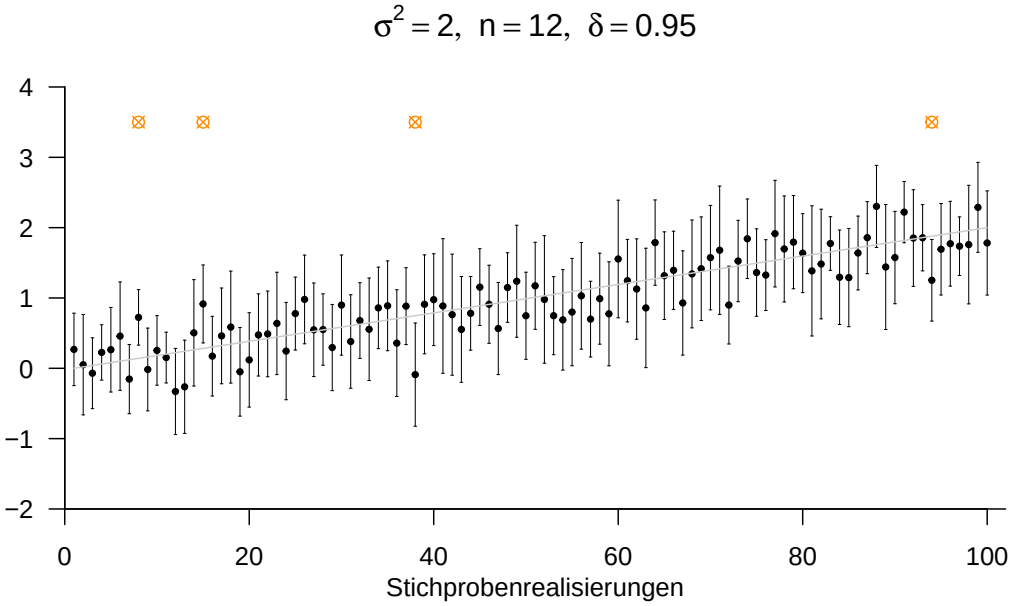


Abbildung 20.4. Simulation der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Erwartungswertparameter des Normalverteilungsmodells bei variablem, wahren, aber unbekannten, Erwartungswertparameter μ für $\sigma^2 := 2, n := 12$ und einer gewünschten Überdeckungswahrscheinlichkeit von $\delta := 0.95$. 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 sind, sind mit einen 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 20.3 (U-Konfindenzintervallstatistik)**.** Gegeben sei das Normalverteilungsmodell

$$v_1, ..., v_n \sim N(\mu, \sigma^2) \tag{20.19}$$

Dann heißt die mit der Stichprobenvarianz

$$\frac{1}{n-1} \sum_{i=1}^{n} (v_i - \bar{v})^2 \tag{20.20}$$

definierte Statistik

$$U := \frac{n-1}{\sigma^2} S^2 \tag{20.21}$$

U-Konfidenzintervallstatistik.

•

Für die Verteilung der U-Konfidenzintervallstatistik gilt folgendes Theorem.

Theorem 20.3 (Verteilung der U-Konfidenzintervallstatistik)**.** *Die U-Konfidenzintervallstatistik ist eine χ^2-verteilte Zufallsvariable mit Parameter $n-1$, es gilt also*

$$U \sim \chi^2(n-1) \tag{20.22}$$

◦

Für einen Beweis von Theorem 20.3 verrweisen wir auf Casella & Berger (2012). Wie die T-Konfidenzintervallstatistik besitzt auch die U-Konfidenzintervallstatistik die Pivoteigenschaft, da sie eine Funktion der Stichprobe ist, aber ihre Verteilung nach Theorem 20.3 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 Zufallvariable mit χ^2, die KVF einer χ^2-verteilten Zufallvariable mit Ξ und die inverse KVF einer χ^2-verteilten Zufallvariable 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                               # Ausgangsraumaufloesung

# analytische Definitionen und Resultate
yx      = seq(3,17,len = res)               # \upsilon_i Raum
ux      = seq(0,30,len = res)               # U Raum
p_y_i   = dnorm(yx,mu,sqrt(sigsqr))         # \upsilon_i WDF
p_y_bar = dnorm(yx,mu,sqrt(sigsqr/n))       # \upsilon_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){                             # Simulationsiterationen
  y         = rnorm(n,mu,sqrt(sigsqr))      # Stichprobenrealisierung
  y_i[s]    = y[1]                          # Stichprobenrealisierung \upsilon_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 20.4 (Konfidenzintervall für den Varianzparameter des Normalverteilungsmodells)**.** *Gegeben sei das Normalverteilungsmodell*

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2) \tag{20.23}$$

mit wahren, aber unbekannten, Parametern μ und σ^2, es sei $\delta \in]0,1[$ und es seien

$$u_\delta := \Xi^{-1}\left(\frac{1-\delta}{2}; n-1\right) \text{ und } u'_\delta := \Xi^{-1}\left(\frac{1+\delta}{2}; n-1\right) \tag{20.24}$$

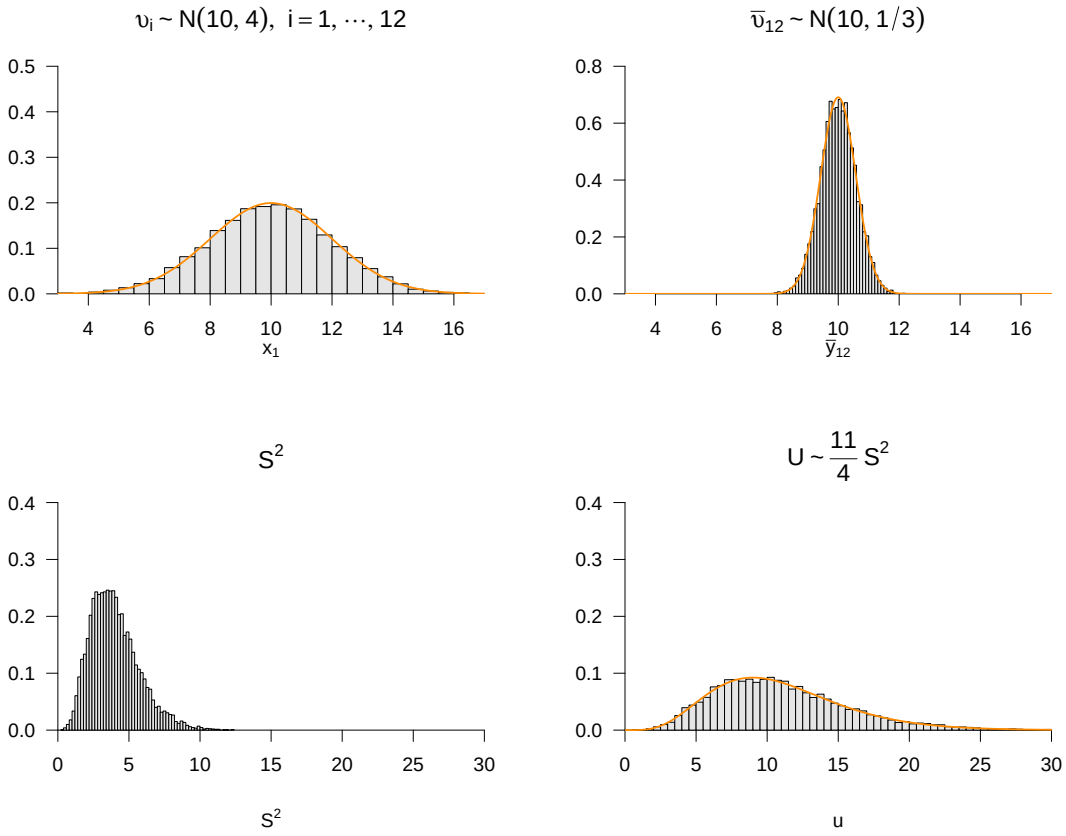


Abbildung 20.5. Simulation der Verteilung der U-Konfidenzintervallstatistik und der ihr zugrundeliegenden Verteilungen der Stichprobenvariable, des Stichprobenmittels und der Stichprobenvarianz.

mit der inversen KVF Ξ^{-1} einer χ^2-verteilten Zufallsvariable. Dann gilt für das Intervall

$$\kappa(\upsilon) := \left[\frac{(n-1)S^2}{u'_\delta}, \frac{(n-1)S^2}{u_\delta}\right]. \tag{20.25}$$

mit der Stichprobenvarianz

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon})^2, \tag{20.26}$$

dass

$$\mathbb{P}_{\sigma^2}(\kappa(\upsilon) \ni \sigma^2) = \delta. \tag{20.27}$$

∘

Beweis. Per Definition gilt mit Definition 20.3 und Theorem 20.3, dass

$$\mathbb{P}_{\sigma^2}\left(u_\delta \leq U \leq u'_\delta\right) = \delta. \tag{20.28}$$

Damit folgt dann aber direkt

$$\begin{aligned}
\delta &= \mathbb{P}_{\sigma^2}\left(u_\delta \leq U \leq u'_\delta\right) \\
&= \mathbb{P}_{\sigma^2}\left(u_\delta \leq \frac{n-1}{\sigma^2}S^2 \leq u'_\delta\right) \\
&= \mathbb{P}_{\sigma^2}\left(u_\delta^{-1} \geq \frac{\sigma^2}{(n-1)S^2} \geq u_\delta'^{-1}\right) \\
&= \mathbb{P}_{\sigma^2}\left(\frac{(n-1)S^2}{u_\delta} \geq \sigma^2 \geq \frac{(n-1)S^2}{u'_\delta}\right) \\
&= \mathbb{P}_{\sigma^2}\left(\frac{(n-1)S^2}{u'_\delta} \leq \sigma^2 \leq \frac{(n-1)S^2)}{u_\delta}\right) \\
&= \mathbb{P}_{\sigma^2}\left(\left[\frac{(n-1)S^2}{u'_\delta}, \frac{(n-1)S^2)}{u_\delta}\right] \ni \sigma^2\right).
\end{aligned} \tag{20.29}$$

□

Wie im Falle von Theorem 20.2 ist der entscheidene Schritt zur Sicherung der Überdeckungswahrscheinlichkeit δ des wahren, aber unbekannten, Varianzparameters durch das in Theorem 20.4 definierte Konfidenzintervall die Definition von

$$u_\delta := \Xi^{-1}\left(\frac{1-\delta}{2}; n-1\right) \text{ und } u'_\delta := \Xi^{-1}\left(\frac{1+\delta}{2}; n-1\right) \tag{20.30}$$

Wie im Beweis von Theorem 20.4 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 20.6. In diesem Fall ergibt sich

$$u_\delta := \Xi^{-1}(0.025; 9) = 2.70 \text{ und } u'_\delta := \Xi^{-1}(0.975; 9) = 19.0. \tag{20.31}$$

Abbildung 20.6 A zeigt diese Wahl aus Perspektive der WDF der U-Konfidenintervallstatistik. 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 20.6 B zeigt die entsprechende Perspektive der KVF der U-Konfidenintervallstatistik. 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.

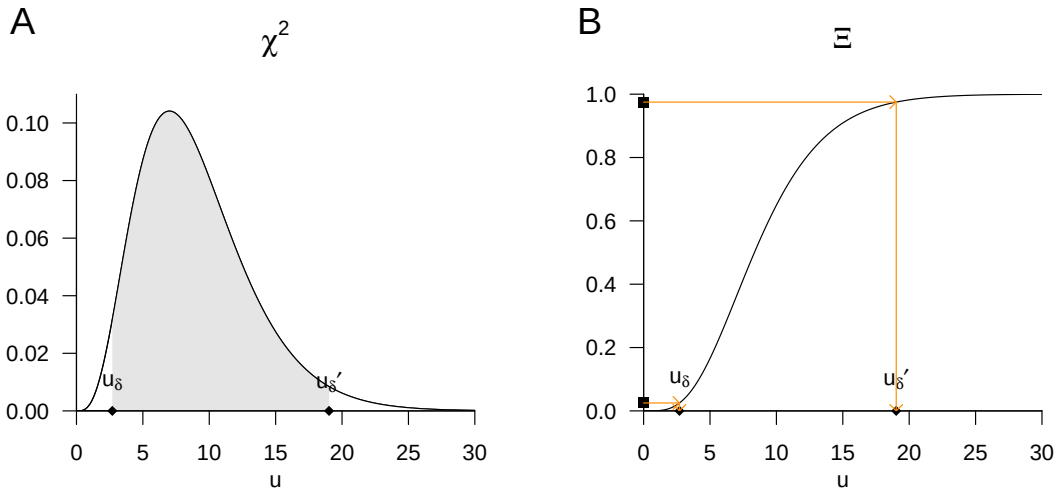


Abbildung 20.6. Sicherung der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Varianzparameter des Normalverteilungsmodells für $\delta := 0.95$ und $n := 10$ aus Perspektive der WDF (A) und der KVF (B) der Verteilung der U-Konfidenzintervallstatistik.

Abschließend wollen wir die Überdeckungswahrscheinlichkeit des durch Theorem 20.4 gegebenen Konfidenzintervalls mithilfe einer Simulation demonstrieren. Dazu betrachten wir zunächst die in Kapitel 20.1 gegebene erste Interpretation eines Konfidenzintervalls bei immer gleichem wahrem, aber unbekanntem, 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)    # \Xi^2((1-\delta)/2; n - 1)
u_delta_o   = qchisq((1+delta)/2, n - 1)    # \Xi^2((1+\delta)/2; n - 1)

# Stichprobenrealisierungen
ns          = 1e2                           # Anzahl Simulationen
y_bar       = rep(NaN,ns)                   # Stichprobenmittelarray
S2          = rep(NaN,ns)                   # Stichprobenvarianzarray
kappa       = matrix(rep(NaN,2*ns), ncol = 2)  # Konfidenzintervallarray
for(i in 1:ns){                             # Simulationsiterationen
   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
}
```

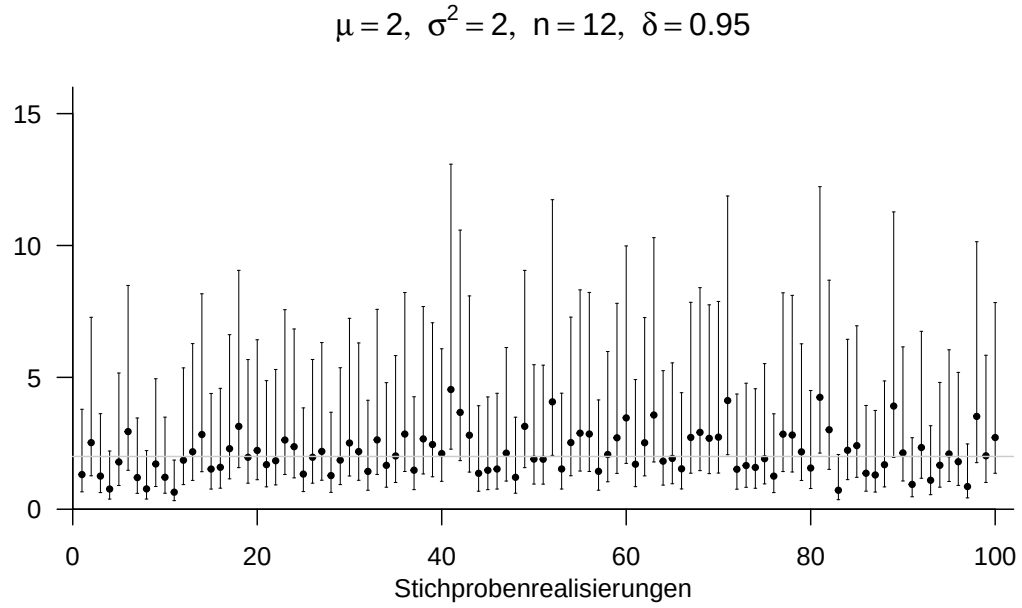


Abbildung 20.7. Simulation der Überdeckungswahrscheinlichkeit des Konfidenzintervalls für den Varianzparameter des Normalverteilungsmodells bei konstanten, wahren, aber unbekannten, Varianzparameter $\sigma^2 := 2$ für $\mu := 2, n := 12$ und einer gewünschten Überdeckungswahrscheinlichkeit von $\delta := 0.95$. 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 sind, sind mit einen orangen Kreis markiert. Man beachte, dass die Konfidenzintervalle nicht symmetrisch um den den Varianzparameterschätzer angeordnet sind

20.3. 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 18.3.1. 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/303-Konfidenzintervalle.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

$$\hat{\mu} = 3.17 \text{ und } \hat{\sigma}^2 = 13.8 \tag{20.32}$$

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

$$\kappa(y) = [0.80, 5.52]. \tag{20.33}$$

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 Varianzwertparameter
delta       =  0.95                              # Konfidenzlevel
n           = length(y)                          # Anzahl Datenpunkte
u_delta_u   = qchisq((1-delta)/2, n - 1)         # \Xi^2((1-\delta)/2; n - 1)
u_delta_o   = qchisq((1+delta)/2, n - 1)         # \Xi^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

$$\kappa(y) = [6.91, 39.74]. \tag{20.34}$$

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.

20.4. Literaturhinweise

Die in diesem Kapitel vorgestellten Ergebnisse gehen in ganz wesentlicher Weise auf Neyman (1937) zurück.

20.5. Selbstkontrollfragen

1. Geben Sie die Definition des Begriffs eines δ-Konfidenzintervalls wieder.
2. Erläutern Sie die zwei Interpretationen eines δ-Konfidenzintervalls.
3. Erläutern Sie die typischen Schritte zur Konstruktion eines δ-Konfidenzintervalls.
4. Geben Sie das Theorem zum δ-Konfidenzintervall für den Erwartungswert der Normalverteilung wieder.
5. Geben Sie das Theorem zum δ-Konfidenzintervall für den Varianzparameter der Normalverteilung wieder.

21. Hypothesentests

Die grundlegende Logik Frequentistischer Hypothesentests kann am Beispiel eines Normalverteilungsmodells für einen beobachteten univariaten Datensatz $y_1, ..., y_n$ grob wie folgt umrissen werden. Man unterstellt zunächst, dass der beobachtete Datensatz eine Realisierung der Stichprobe $\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2)$ 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 Nullmodels generieren würde. Je nach Beschaffenheit des zugrundliegenden 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 Nullmodel 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 Nullmodel 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 natürliich nicht, ob im vorliegenden Fall nun wirklich das Nullmodel 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)).

21.1. Testhypothesen und Tests

Im Kontext von Frequentistischen Hypothesentests wird der Begriff des Frequentistischen Inferenzmodells (vgl. Definition 18.1) zunächt durch die sogenannten *Testhypothesen* zu einem *Testszenario* erweitert. Wir nutzen folgende Definition.

Definition 21.1 (Testhypothesen und Testszenario)**.** Gegeben sei ein Frequentistisches Inferenzmodell mit Stichprobe υ, Ergebnisraum $\mathcal{Y}$ und Parameterraum Θ. Weiterhin sei $\{\Theta_0, \Theta_1\}$ eine Partition des Parameterraums, so dass

$$\Theta = \Theta_0 \cup \Theta_1 \text{ und } \Theta_0 \cap \Theta_1 = \emptyset. \tag{21.1}$$

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 21.2 (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 wahre, 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

$$\Theta_0 := \{0\}, \tag{21.2}$$

die entsprechend zusammengesetzte Alternativhypothese ist dann gegeben durch

$$\Theta_1 = \mathbb{R} \setminus \{0\}. \tag{21.3}$$

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 21.3 (Einseitige und zweiseitige Testhypothesen)**.** Gegeben sei ein Testszenario mit eindimensionalem Parameteraum $\Theta := \mathbb{R}$ 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 21.4 (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

$$\phi : \mathcal{Y} \to \{0, 1\}, y \mapsto \phi(y), \tag{21.4}$$

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 21.4 eine konkrete Realisierung $y \in \mathcal{Y}$ der Stichprobe υ, 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 $\phi(\upsilon)$.

In der Anwendung ist man oft an Tests interessiert, die eine bestimmte Struktur haben, wir formalisieren diese unter dem Begriff der *Standardtests*.

Definition 21.5 (Standardtest)**.** Gegeben sei ein Testszenario. Dann ist ein *Standardtest* ϕ definiert als die Verkettung einer *Teststatistik*

$$\gamma : \mathcal{Y} \to \Gamma \tag{21.5}$$

und einer *Entscheidungsregel*

$$\delta : \Gamma \to \{0, 1\} \tag{21.6}$$

kann also geschrieben werden als

$$\phi := \delta \circ \gamma : \mathcal{Y} \to \{0, 1\}. \tag{21.7}$$

•

Wie oben angemerkt gibt es auch bei Definition 21.5 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 υ ist, $\gamma(y) \in \Gamma$ eine Realisierung von $\gamma(\upsilon)$ und $(\delta \circ \gamma)(y)$ eine Realisierung von $(\delta \circ \gamma)(\upsilon)$.

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 21.6 (Kritischer Bereich eines Tests)**.** Gegeben sei ein Testszenario und ein Test ϕ. Dann heißt die Untermenge K des Ergebnisraums $\mathcal{Y}$ der Stichprobe υ, für die der Test den Wert 1 annimmt, *kritischer Bereich* des Tests, formal

$$K := \{y \in \mathcal{Y} | \phi(y) = 1\} \subset \mathcal{Y}. \tag{21.8}$$

•

Man beachte, dass vor dem Hintergrund von Definition 21.6 die zufälligen Ereignisse $\{\upsilon \in K\}$ und $\{\phi(\upsilon) = 1\}$, 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 21.7 (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

$$A := \{\gamma(y) \in \Gamma | \phi(y) = 1\} \subset \Gamma. \tag{21.9}$$

•

Wie zum Begriff des kritischen Bereichs angemerkt gilt auch hier, dass die Ereignisse $\{\phi(\upsilon) = 1\}$ und $\{\gamma(\upsilon) \in A\}$ äquivalent sind und damit insbesondere auch die gleiche Wahrscheinlichkeit besitzen. Insgesamt gelten mit Definition 21.6 und Definition 21.7 für einen Standardtest also

$$\{\upsilon \in K\} \Leftrightarrow \{\gamma(\upsilon) \in A\} \Leftrightarrow \{\phi(\upsilon) = 1\} \tag{21.10}$$

und

$$\mathbb{P}_\theta\left(\{\upsilon \in K\}\right) = \mathbb{P}_\theta\left(\{\gamma(\upsilon) \in A\}\right) = \mathbb{P}_\theta\left(\{\phi(\upsilon) = 1\}\right), \tag{21.11}$$

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 21.5 allgemein angebene 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 21.8 (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

$$\phi : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := 1_{\{\gamma(y) \geq k\}} = \begin{cases} 1 & \gamma(y) \geq k \\ 0 & \gamma(y) < k \end{cases}, \tag{21.12}$$

- ein *zweiseitiger kritischer Wert-basierter Test* von der Form

$$\phi : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := 1_{\{|\gamma(y)| \geq k\}} = \begin{cases} 1 & |\gamma(y)| \geq k \\ 0 & |\gamma(y)| < k \end{cases}. \tag{21.13}$$

•

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.

21.2. 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 21.9 (Richtige Testentscheidungen und Testfehler)**.** Gegeben seien ein Testszenario 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 Nichtablehen der Nullhypothese, wenn die Alternativhypothese $\theta \in \Theta_1$ zutrifft, heißt *Typ II Fehler*.

•

Testentscheidung

Wahrer, aber unbekannter, Parameter	$\phi(v) = 0$	$\phi(v) = 1$
$\theta \in \Theta_0$	Richtige Entscheidung	Typ I Fehler
$\theta \in \Theta_1$	Typ II Fehler	Richtige Entscheidung

Abbildung 21.1. Richtige Testentscheidungen und Typ I und Typ II Fehler

Abbildung 21.1 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 21.10 (Testgütefunktion)**.** Gegeben sei ein Testszenario und ein Test ϕ. Dann ist die *Testgütefunktion* von ϕ definiert als

$$q_\phi : \Theta \to [0,1], \theta \mapsto q_\phi(\theta) := \mathbb{P}_\theta(\phi(v) = 1). \tag{21.14}$$

Für $\theta \in \Theta_1$ heißt q_ϕ auch *Trennschärfefunktion* oder *Powerfunktion*.

•

Man beachte, dass $\mathbb{P}_\theta$ in Definition 21.10 die Verteilung der Zufallsvariable $\phi(v)$ unter der Annahme, dass die Verteilung von v durch θ festgelegt ist, bezeichnen soll. Für jedes $\theta \in \Theta$ liefert q_ϕ also die 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 21.6) und des Ablehnungsbereichs (vgl. Definition 21.7) wie bereits gesehen

$$\mathbb{P}_\theta(\phi(v) = 1) = \mathbb{P}_\theta(\gamma \in A) = \mathbb{P}_\theta(v \in K). \tag{21.15}$$

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

$$q_\phi(\theta) = \mathbb{P}_\theta(\phi(\upsilon) = 1) = 0 \text{ für } \theta \in \Theta_0 \text{ und } q_\phi(\theta) = \mathbb{P}_\theta(\phi(\upsilon) = 1) = 1 \text{ für } \theta \in \Theta_1. \quad (21.16)$$

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

$$\phi_a : \mathcal{Y} \to \{0, 1\}, y \mapsto \phi_a(y) := 0. \quad (21.17)$$

ϕ_a sei also ein Test, der die Nullhypothese unabhängig von den beobachteten Daten nie ablehnt. Für ϕ_a gilt dann

$$q_{\phi_a}(\theta) = \mathbb{P}_\theta(\phi(\upsilon) = 1) = 0 \text{ für } \theta \in \Theta_0. \quad (21.18)$$

Allerdings gilt für ϕ_a dann auch automatisch

$$q_{\phi_a}(\theta) = \mathbb{P}_\theta(\phi(\upsilon) = 1) = 0 \text{ für } \theta \in \Theta_1. \quad (21.19)$$

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

$$\phi_b : \mathcal{Y} \to \{0, 1\}, y \mapsto \phi_b(y) := 1. \quad (21.20)$$

ϕ_b sei also ein Test, der die Nullhypothese, unabhängig von den beobachteten Daten immer ablehnt. Für ϕ_b gilt dann

$$q_{\phi_b}(\theta) = \mathbb{P}_\theta(\phi(\upsilon) = 1) = 1 \text{ für } \theta \in \Theta_1. \quad (21.21)$$

ϕ_b ist also maximal sensitiv für das Zutreffen der Alternativhypothese. Allerdings gilt für ϕ_b dann auch automatisch

$$q_{\phi_b}(\theta) = \mathbb{P}_\theta(\phi(\upsilon) = 1) = 0 \text{ für } \theta \in \Theta_0, \quad (21.22)$$

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

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0, \quad (21.23)$$

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 (21.23) 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 (21.23) motiviert dann zunächst die Definition der Begriffe des *Level-α_0-Tests*, des *Signifikanzlevels* α_0 und des *Testumfangs* α.

Definition 21.11 (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

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0. \tag{21.24}$$

Wenn ϕ ein Level-α_0-Test ist, nennt man den Wert α_0 auch das *Signifikanzlevel* des Tests. Weiterhin heißt die Zahl

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) \in [0, 1] \tag{21.25}$$

der *Testumfang* von ϕ.

•

Nach Definition 21.11 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 21.2) 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

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) = \max_{\theta \in \{\theta_0\}} q_\phi(\theta) = q_\phi(\theta_0) = \mathbb{P}_{\theta_0}(\phi = 1). \tag{21.26}$$

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

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) = \max_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi = 1). \tag{21.27}$$

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 21.3 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

$$\alpha = \alpha_0. \tag{21.28}$$

Ein Test, für den der Testumfang kleiner als sein Signifikanzlevel ist, für den also gilt

$$\alpha < \alpha_0 \tag{21.29}$$

wird *konservativ* genannt. Ein Test schließlich, dessen Testumfang größer als sein Signifikanzlevel ist,

$$\alpha > \alpha_0 \tag{21.30}$$

und der damit natürlich kein Level-α_0-Test sein kann, wird *liberal* genannt.

p-Wert

Ein definierendes Charakterstikum 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 Testergbnisses 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 Testsstatistik 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 21.12 (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 21.3.1 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 Nullmodels 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 wissenschafltichen 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 fälschlichweise abzulehnen wird damit ein Typ II Fehler. Intuitiv ergibt sich also folgende Zuordnung:

Nicht-favorisierte	wissenschaftliche Hypothese	$\rightarrow$	Nullhypothese
Favorisierte	wissenschaftliche Hypothese	$\rightarrow$	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 (2005), 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 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.

21.3. Testbeispiele

21.3.1. 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 einen vorgebenenen 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,

$$H_0 : \mu = \mu_0 \text{ und } H_1 : \mu = \mu_1 \tag{21.31}$$

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

$$H_0 : \mu = \mu_0 \text{ und } H_1 : \mu \neq \mu_0 \tag{21.32}$$

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

$$H_0 : \mu \leq \mu_0 \text{ und } H_1 : \mu > \mu_0 \text{ oder } H_0 : \mu \geq \mu_0 \text{ und } H_1 : \mu < \mu_0 \tag{21.33}$$

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 21.13 (Frequentistisches Inferenzmodell des Einstichproben-T-Tests)**.** Das Frequentistische Inferenzmodell des Einstichproben-T-Tests ist gegeben durch das Normalverteilungsmodell (vgl. Definition 18.2)

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \tag{21.34}$$

•

Wir erinnern daran, dass aus generativer Sicht das Normalverteilungsmodell dem Modell

$$\upsilon_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ für } i = 1, ..., n \tag{21.35}$$

entspricht (vgl. Kapitel 18.3.1). Die Annahme unabhängig und identisch normalverteilter Zufallsvariablen als Grundlage der Modellierung der Beobachtung von n Datenpunkten ist wie in Kapitel 18.3.1 gesehen äquivalent zu der Annahme, dass sich jede einen Datenpunkt modellierende Zufallsvariable υ_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 21.14 (Einfache Nullhypothese und zusammengesetzte Alternativhypothese des Einstichproben-T-Tests)**.** Gegeben sei das Frequentistiche Inferenzmodell des Einstichproben-T-Tests

$$\upsilon_1, ..., \upsilon_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \tag{21.36}$$

und es sei $\Theta := \mathbb{R}$ der Parameterunteraum 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

$$\Theta_0 := \{\mu_0\} \Leftrightarrow H_0 : \mu = \mu_0 \text{ und } \Theta_1 := \mathbb{R} \setminus \{\mu_0\} \Leftrightarrow H_1 : \mu \neq \mu_0. \tag{21.37}$$

•

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 21.15 (Einstichproben-T-Test-Statistik)**.** Gegeben sei das Testszenario eines Einstichproben-T-Tests mit Stichprobe $\upsilon_1, ..., \upsilon_n$, Stichprobenmittel $\bar{\upsilon}$, Stichprobenstandardabweichung S und Nullhypothesenparameter μ_0. Dann ist die *Einstichproben-T-Test-Statistik* definiert als

$$T := \sqrt{n}\frac{\bar{\upsilon} - \mu_0}{S}. \tag{21.38}$$

•

Offenbar hat die Einstichproben-T-Test-Statistik eine hohe Ähnlichkeit mit der T-Konfidenzintervallstatistik (vgl. Definition 20.2). Man beachte allerdings, dass im Fall der Einstichproben-T-Test-Statistik der Nullhypothesenparameter 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 $\bar{\upsilon}$ zum angenommenen Nullhypothesenparameter μ_0 auf. Wie gesehen ist das Stichprobenmittel ein unverzerrter Schätzer des Erwartungswertparameters μ der Stichprobe. Die Differenz $\bar{\upsilon} - \mu_0$ entspricht also einer Schätzung der Abweichung des wahren, aber unbekannten, Erwartungswertsparameters vom Nullhypothesenparameter 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 $\bar{v} - \mu_0$ 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 $\bar{v} - \mu_0 = 2$ und ist $S = 1$, so beträgt die Abweichung des Stichprobenmittels vom Nullhypothesenparameter 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 $\frac{\bar{v}-\mu_0}{S}$ 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 21.1 (Verteilung der Einstichproben-T-Test-Statistik)**.** *Gegeben sei das Testszenario eines Einstichproben-T-Tests mit Stichprobe $v_1, ..., v_n$, Stichprobenmittel $\bar{v}$, Stichprobenstandard- abweichung S , Nullhypothesenparameter μ_0 und Einstichproben-T-Test-Statistik definiert als*

$$T := \sqrt{n}\frac{\bar{v} - \mu_0}{S}. \tag{21.39}$$

Dann ist T eine nichtzentrale t-Zufallsvariable mit Nichtzentralitätsparameter

$$d = \sqrt{n}\frac{\mu - \mu_0}{\sigma} \tag{21.40}$$

und Freiheitsgradparameter $n - 1$, es gilt also $T \sim t(d, n - 1)$

∘

Beweis. □

Man beachte, dass im Falle des Zutreffens der Nullhypothese der Nullhypothesenparameter 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 Nullhypothesenparameter $\mu_0 = 0$ in Abbildung 21.2 (B), Die Parameter dieser Verteilung ergeben sich dabei zu

$$d = \sqrt{n}\frac{\mu - \mu_0}{\sigma} = \sqrt{12}\frac{3-0}{\sqrt{2}} \approx 7.34 \text{ und } n - 1 = 11 \tag{21.41}$$

ergeben.

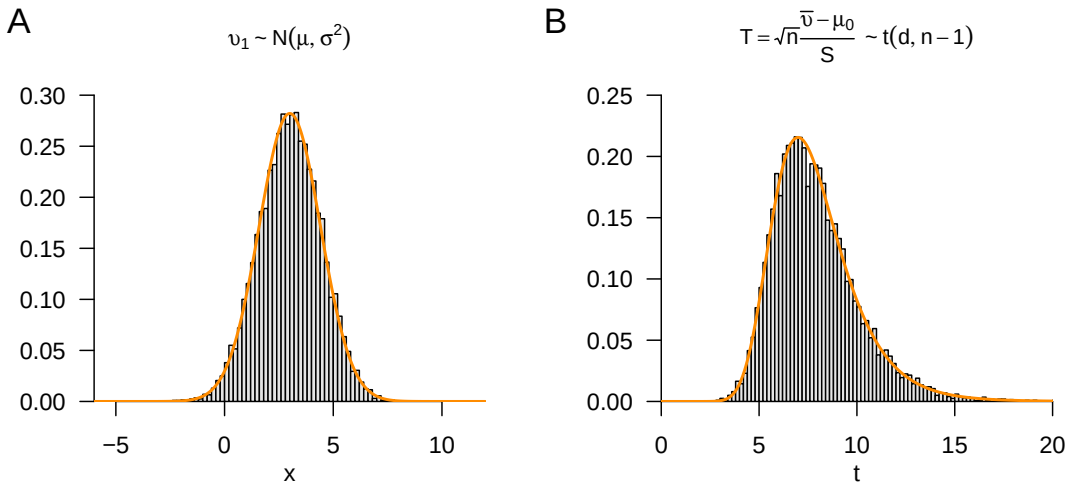


Abbildung 21.2. Verteilung der Einstichproben-T-Test-Statistik für $n = 12, \mu = 3, \sigma^2$ und $\mu_0 = 0$. (A) Verteilung der Stichprobenvariablen. (B) Verteilung der Einstichproben-T-Test-Statstik

Testdefinition

Wir können nun den zweiseitigen Einstichproben-T-Test mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese definieren und seine Testgütefunktion analysieren.

Definition 21.16 (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

$$\phi : \mathcal{Y} \to \{0, 1\}, y \mapsto \phi(y) := 1_{\{|t| \geq k\}} = \begin{cases} 1 & |t| \geq k \\ 0 & |t| < k \end{cases}. \tag{21.42}$$

•

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 21.2 (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*

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - \Psi(k; d_\mu, n-1) + \Psi(-k; d_\mu, n-1), \tag{21.43}$$

wobei $\Psi(\cdot; d_\mu, n-1)$ die KVF der nichtzentralen t-Verteilung mit Nichtzentralitätsparameter

$$d_\mu := \sqrt{n}\frac{\mu - \mu_0}{\sigma} \tag{21.44}$$

und Freiheitsgradparameter $n-1$ bezeichnet.

∘

Beweis. Die Testgütefunktion des betrachteten Test im vorliegenden Testszenario ist definiert als

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := \mathbb{P}_\mu(\phi = 1). \tag{21.45}$$

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

$$T := \sqrt{n}\frac{\bar{v} - \mu_0}{S} \tag{21.46}$$

unter der Annahme $v_1, ..., v_n \sim N(\mu, \sigma^2)$ anhand einer nichtzentralen t-Verteilung $t(d_\mu, n-1)$ mit Nichtzentralitätsparameter

$$d_\mu := \sqrt{n}\frac{\mu - \mu_0}{\sigma} \tag{21.47}$$

verteilt ist. Der Ablehnungsbereich des zweiseitigen Einstichproben-T-Tests ist

$$A =]-\infty, -k] \cup]k, \infty[. \tag{21.48}$$

Mit diesem Ablehungsbereich ergibt sich dann

$$\begin{aligned}
q_\phi(\mu) &= \mathbb{P}_\mu(\phi = 1) \\
&= \mathbb{P}_\mu\left(T \in]-\infty, -k] \cup]k, \infty[\right) \\
&= \mathbb{P}_\mu\left(T \in]-\infty, -k]\right) + \mathbb{P}_\mu\left(T \in [k, \infty[\right) \\
&= \mathbb{P}_\mu(T \leq -k) + \mathbb{P}_\mu(T \geq k) \\
&= \mathbb{P}_\mu(T \leq -k) + (1 - \mathbb{P}_\mu(T \leq k)) \\
&= 1 - \mathbb{P}_\mu(T \leq k) + \mathbb{P}_\mu(T \leq -k) \\
&= 1 - \Psi(k; d_\mu, n-1) + \Psi(-k; d_\mu, n-1),
\end{aligned} \tag{21.49}$$

wobei $\Psi(\cdot; d_\mu, n-1)$ die KVF der nichtzentralen T-Verteilung mit Nichtzentralitätsparameter d_μ und Freiheitsgradparameter $n-1$ bezeichnet.

□

In Abbildung 21.3 visualisieren wir die Testgütefunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese aus Theorem 21.2 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.

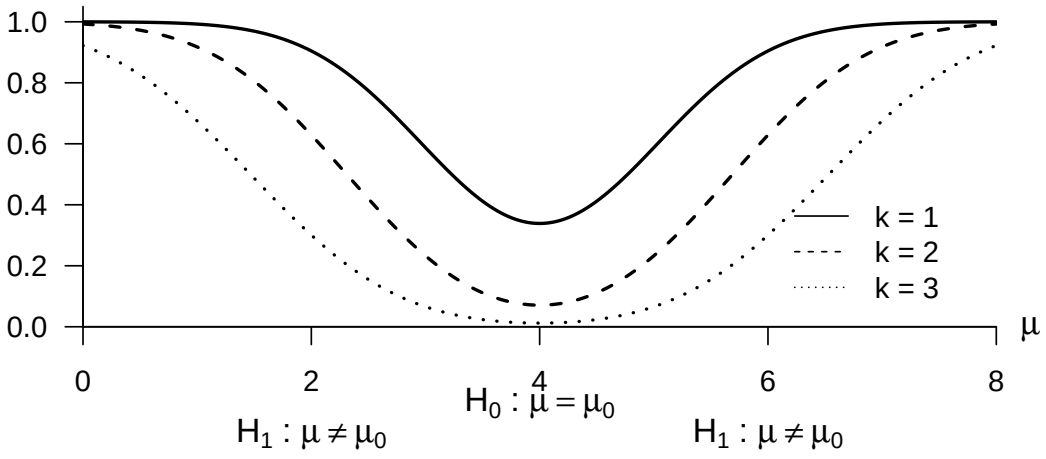


Abbildung 21.3. Testgütefunktion des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese für $\sigma^2 = 9, \mu_0 = 4, n = 12$ und $k = 1, 2, 3$.

Testumfangkontrolle

Die Werte der Testgütefunktion bei $\mu = \mu_0$ in Abbildung 21.3 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 21.3 (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

$$k_{\alpha_0} := \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right), \tag{21.50}$$

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

$$\begin{aligned} q_\phi(\mu_0) &= 1 - \Psi(k; d_{\mu_0}, n-1) + \Psi(-k; d_{\mu_0}, n-1) \\ &= 1 - \Psi(k; 0, n-1) + \Psi(-k; 0, n-1) \\ &= 1 - \Psi(k; n-1) + \Psi(-k; n-1), \end{aligned} \tag{21.51}$$

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

$$\begin{aligned} q_\phi(\mu_0) &= 1 - \Psi(k_{\alpha_0}; n-1) + \Psi(-k_{\alpha_0}; n-1) \\ &= 1 - \Psi(k_{\alpha_0}; n-1) + (1 - \Psi(k_{\alpha_0}; n-1) \\ &= 2(1 - \Psi(k_{\alpha_0}; n-1)) \\ &= 2\left(1 - \Psi\left(\Psi^{-1}\left(1 - \frac{\alpha_0}{2}, n-1\right), n-1\right)\right) \\ &= 2\left(1 - 1 + \frac{\alpha_0}{2}\right) \\ &= \alpha_0, \end{aligned} \tag{21.52}$$

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.

□

Man beachte, dass nach Theorem 21.3 der hier betrachtete Tests inbesondere exakt ist, der Testumfang also mit dem Signifikanzlevel identisch ist. In Abbildung 21.4 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

```

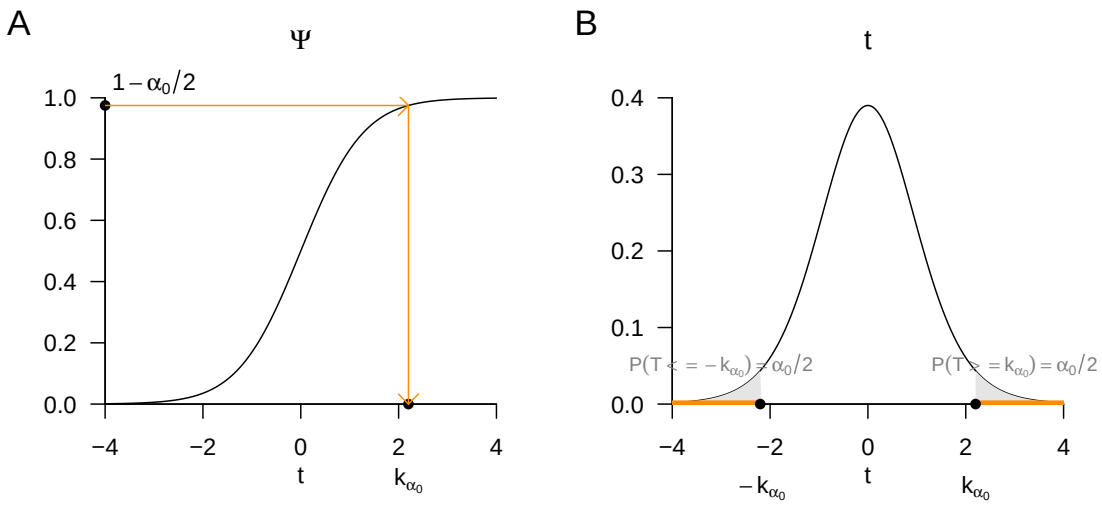


Abbildung 21.4. Bestimmung des kritischen Werts k_{α_0} für den zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese für $n = 12$ und $\alpha_0 := 0.05$ zu Kontrolle des Testumfangs (A) Für ein gewähltes α_0 ist der kritische Wert des betrachteten Tests nach Theorem 21.3 durch den Wert der 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.

```
# Testparameter
mu_0      = 0                                   # Nullhypothesenparameter, 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(NaN,nsim)                       # Testentscheidungsarray
for(j in 1:nsim){                               # Simulationsiterationen
    y      = rnorm(n,mu,sigsqr)                 # \ups_i \sim N(\mu,\Sigma), i = 1,...,n
    y_bar  = mean(y)                            # Stichprobenmittel
    s      = sd(y)                              # Stichprobenstandardabweichung
    Tee    = sqrt(n)*((y_bar - mu_0)/s)         # Einstichproben-T-Test-Statistik
    if(abs(Tee) > k_alpha_0){                   # Test 1_{\vert t \vert >= k_alpha_0}
        phi[j] = 1                              # Ablehnen der Nullhypothese
    } else {
        phi[j] = 0                              # Nichtablehnen der Nullhypothese
    }
}

# Ausgabe
cat("Kritischer Wert              =", k_alpha_0,
    "\nGeschätzter Testumfang alpha =", mean(phi))
```

```
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 21.4 (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*

$$p\text{-}Wert = 2(1 - \Psi(|t|; n-1)) \tag{21.53}$$

wobei $\Psi(\cdot; n-1)$ die KVF der t-Verteilung mit Freiheitsgradparameter $n-1$ bezeichnet.

∘

Beweis. Nach Definition 21.12 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

$$|t| \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right) \tag{21.54}$$

abgelehnt werden, vgl. Theorem 21.3. Für diese α_0 gilt, dass

$$\alpha_0 \geq 2(1 - \Psi(|t|; n-1)), \tag{21.55}$$

denn

$$\begin{aligned}
&|t| \geq \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right) \\
\Leftrightarrow\ &\Psi(|t|; n-1) \geq \Psi\left(\Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right); n-1\right) \\
\Leftrightarrow\ &\Psi(|t|; n-1) \geq 1 - \frac{\alpha_0}{2} \\
\Leftrightarrow\ &\mathbb{P}(T \leq |t|) \geq 1 - \frac{\alpha_0}{2} \\
\Leftrightarrow\ &\frac{\alpha_0}{2} \geq 1 - \mathbb{P}(T \leq |t|) \\
\Leftrightarrow\ &\frac{\alpha_0}{2} \geq \mathbb{P}(T \geq |t|) \\
\Leftrightarrow\ &\alpha_0 \geq 2\mathbb{P}(T \geq |t|) \\
\Leftrightarrow\ &\alpha_0 \geq 2(1 - \Psi(|t|; n-1)).
\end{aligned} \tag{21.56}$$

Das kleinste $\alpha_0 \in [0, 1]$ mit

$$\alpha_0 \geq 2\mathbb{P}(T \geq |t|) \tag{21.57}$$

ist dann entsprechend

$$\alpha_0 = 2(1 - \Psi(|t|; n-1)). \tag{21.58}$$

□

In Abbildung 21.5 visualisieren wir die Bestimmung von p-Werten für $t = 2.26$ und für $t = 3.81$, welche sich zu $\mathrm{p} = 0.045$ und $\mathrm{p} = 0.003$, respektive, ergeben. Man beachte, dass zum Beispiel der p-Wert zu $t = -2.26$ auch $\mathrm{p} = 0.045$ beträgt.

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

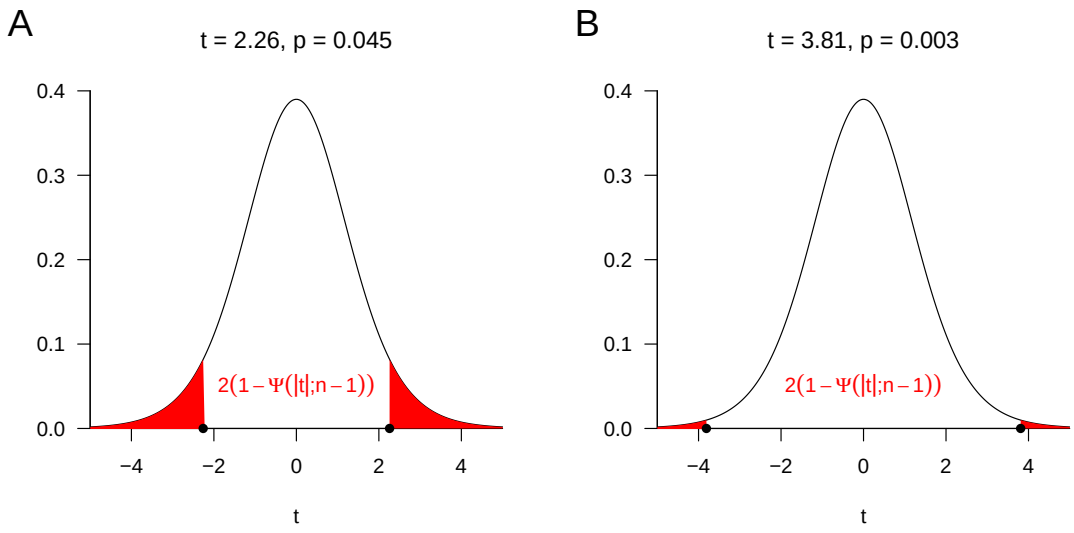


Abbildung 21.5. p-Werte für zwei mögliche Werte der Einstichproben-T-Test-Statistik bei $n = 12$. Die Bereiche im Ergebnisraum der Einstichproben-T-Test-Statistik über denen roten Flächen eingezeichnet sind, markieren die Bereiche mit $T \geq |t|$ gilt. Die summierten ihnen zugeordneten Wahrscheinlichkeiten, also die entsprechenden roten Flächen unter der WDF der t-Verteilungen entsprechen dem p-Wert

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 21.2) 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

$$q_\phi : \mathbb{R} \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - \Psi(k_{\alpha_0}; d_\mu, n-1) + \Psi(-k_{\alpha_0}; d_\mu, n-1) \tag{21.59}$$

bei kontrolliertem Testumfang, also für

$$k_{\alpha_0} := \Psi^{-1}\left(1 - \frac{\alpha_0}{2}; n-1\right) \tag{21.60}$$

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 21.17 (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

$$\pi : \mathbb{R} \times \mathbb{N} \to [0,1], (d,n) \mapsto \pi(d,n) := 1 - \Psi(k_{\alpha_0}; d, n-1) + \Psi(-k_{\alpha_0}; d, n-1) \tag{21.61}$$

•

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)                        # Nichtzentralitätsparameterraum
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(NaN, 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 21.6 und Abbildung 21.7. 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.

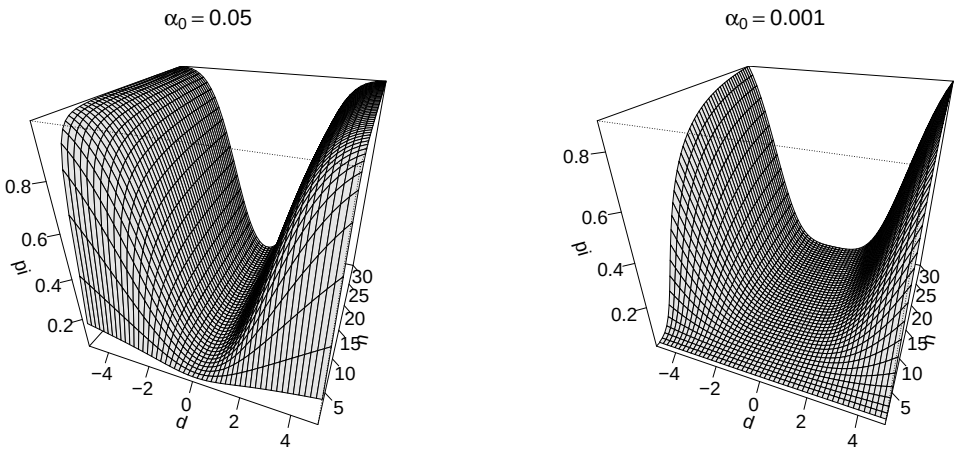


Abbildung 21.6. Powerfunktionen des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. (A) Abhängigkeit der Powerfunktion π vom Nichtzentralitätsparameter d und Stichprobenumfang n bei Wahl von $\alpha_0 := 0.05$. (B) Abhängigkeit der Powerfunktion 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

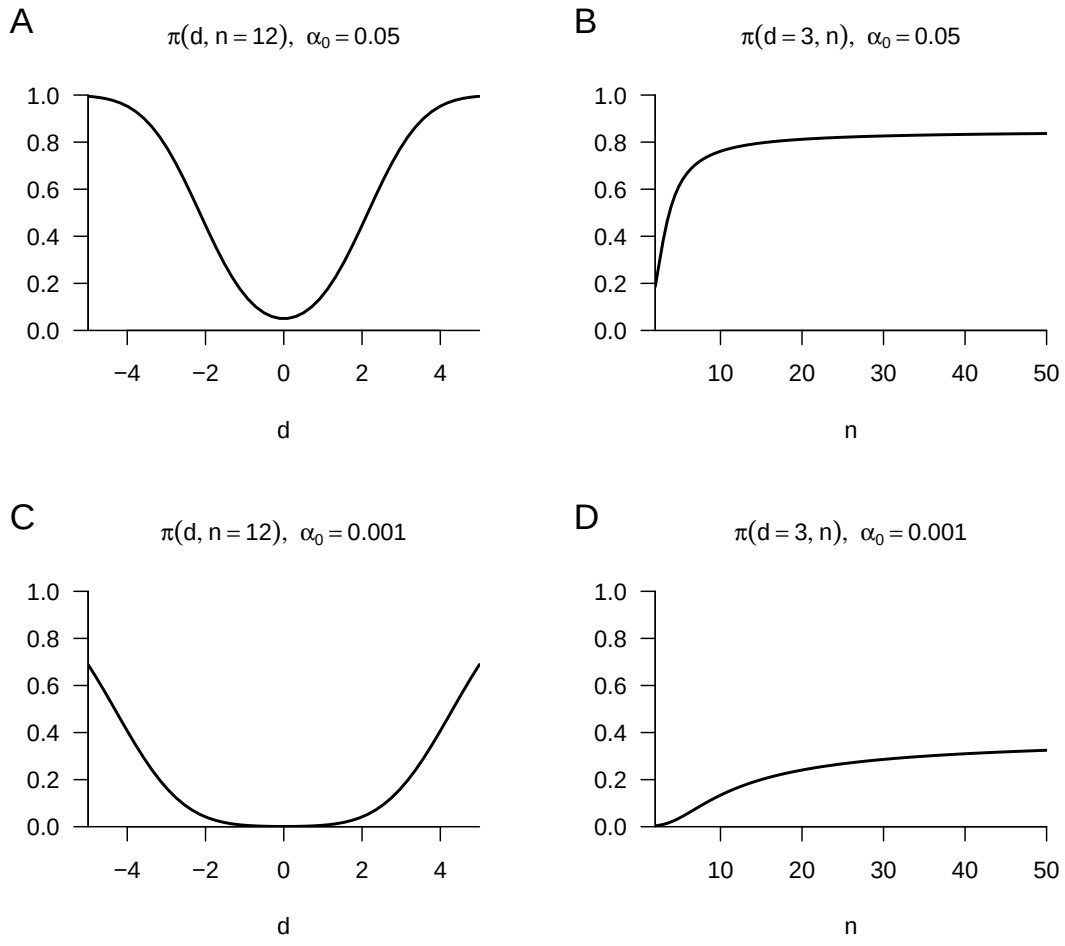


Abbildung 21.7. Schnitte der Powerfunktionen des zweiseitigen Einstichproben-T-Tests mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese. (A) Abhängigkeit der Powerfunktion π vom Nichtzentralitätsparameter d bei konstantem Stichprobenumfang $n = 12$ und Wahl von $\alpha_0 := 0.05$. (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 $\alpha_0 := 0.001$.

Alternativhypothese ergibt sich dann zunächst folgendes routiniertes Durchführen des Tests im Rahmen einer Datenanalyse.

Man nimmt an, dass ein vorliegender univariater Datensatz $y_1, ..., y_n$ eine Realisierung des Frequentistischen Inferenzmodells $v_1, ..., v_n \sim N(\mu, \sigma^2)$ 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 $\bar{v}$ und der Stichprobenstandardabweichung s berechnet man sodann den Wert der Einstichproben-T-Test-Statistik durch

$$t := \sqrt{n}\left(\frac{\bar{y} - \mu_0}{s}\right). \tag{21.62}$$

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

$$\text{p-Wert} = 2(1 - \Psi(|t|; n-1)) \tag{21.63}$$

Folgender **R** Code demonstriert dieses Vorgehen bei Annahme eines vorliegenden Datenvektors `y` der Länge `n`.

```
n          = length(y)                          # Stichprobenumfang
mu_0       = 0                                  # Nullhypothesenparameter
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 entscheided 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

$$\pi(d = d^*, n) = \beta \tag{21.64}$$

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 21.8 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 Stichprobenumfangraums
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
```

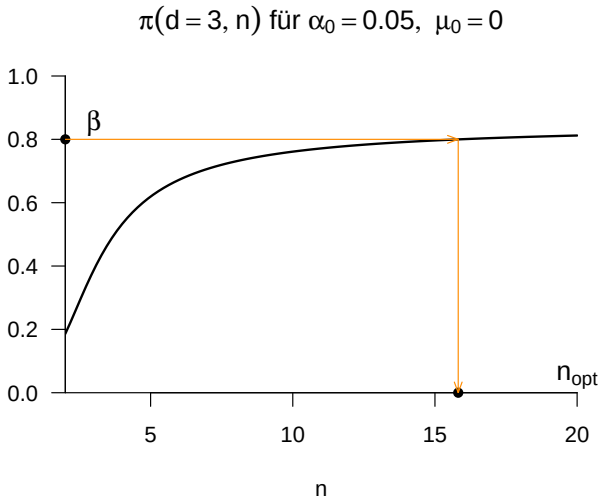


Abbildung 21.8. Praktische Durchführung einer Powerfunktionsbasiertern Stichprobenumfangoptimierung. Bei gewählten Signifikanzlevel α_0 und fest angenommenen Nichtzentralitätsparameter $d*$ evaluiert man den Stichprobenumfang für den die Testpowerfunktion einen Wert von β oder größer hat.

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 18.3.1 eingeführten Anwendungsbeispiel demonstrieren. Inhaltlich entsprich 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/304-Hypothesentests.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                                               # Nullhypothesenparameter, 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 untensstehend

```
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 vorliegen 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.

21.4. 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 21.5 (Dualität von Konfidenzintervallen und Hypothesentests)**.** *Es sei υ die Stichprobe eines Frequentistischen Inferenzmodells mit Ergebnisraum $\mathcal{Y}$ und Parameterraum Θ. Weiterhin sei für ein $\delta \in]0,1[$ mit $[G_u(\upsilon), G_o(\upsilon)]$ ein δ-Konfidenzintervall für den wahren, aber unbekannten, Parameter $\theta \in \Theta$ definiert. Dann gilt, dass der Hypothesentest definiert durch*

$$\phi_\theta : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := \begin{cases} 0, & [G_u(y), G_o(y)] \ni \theta_0 \\ 1, & [G_u(y), G_o(y)] \not\ni \theta_0 \end{cases} \tag{21.65}$$

ein Hypothesentest vom Signifikanzlevel $\alpha_0 = 1 - \delta$ für die Hypothesen

$$\Theta_0 := \{\theta_0\} \text{ und } \Theta_1 := \Theta \setminus \{\theta_0\}. \tag{21.66}$$

∘

Beweis. Aufgrund der einfachen Nullhypothese und somit $\alpha_0 = \alpha$ folgt

$$\alpha_0 = \alpha = \mathbb{P}_{\theta_0}(\phi(\upsilon) = 1) = \mathbb{P}_{\theta_0}([G_u(y), G_o(y)] \not\ni \theta) = 1 - \mathbb{P}_{\theta_0}([G_u(y), G_o(y)] \ni \theta) = 1 - \delta. \tag{21.67}$$

□

Theorem 21.5 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 21.5 für das in Kapitel 20.2 betrachtete Konfidenzintervall für den Erwartungswertparameter des Normalverteilungsmodells und den in Kapitel 21.3.1 betrachteten Einstichproben-T-Test konkretisieren.

Theorem 21.6 (Dualität von Erwartungswertkonfidenzintervall und Einstichproben-T-Test)**.** *Gegeben sei das Normalverteilungsmodell und es sei*

$$\kappa := \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right]. \tag{21.68}$$

das mithilfe von

$$t_\delta := \Psi^{-1}\left(\frac{1+\delta}{2}; n-1\right) \tag{21.69}$$

in Theorem 20.2 definierte δ-Konfidenzintervall für den Erwartungswertparameter. Dann ist mit Theorem 21.5 der Test

$$\phi : \mathcal{Y} \to \{0,1\}, y \mapsto \phi(y) := \begin{cases} 0, & \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right] \ni \mu_0 \\ 1, & \left[\bar{\upsilon} - \frac{S}{\sqrt{n}} t_\delta, \bar{\upsilon} + \frac{S}{\sqrt{n}} t_\delta\right] \not\ni \mu_0 \end{cases} \tag{21.70}$$

ein Test der einfachen Nullhypothese $H_0 : \mu = \mu_0$ und der zusammengesetzten Alternativhypothese $H_1 := \mu_0 \neq \mu$ mit Signifikanzlevel $\alpha_0 = 1 - \delta$.

∘

Beweis. Es gilt

$$\mathbb{P}_{\mu_0}\left(\phi(\upsilon)=1\right)=1-\mathbb{P}_{\mu_0}\left(\phi(\upsilon)=0\right)=1-\mathbb{P}_{\mu_0}\left(\left[\bar{\upsilon}-\frac{S}{\sqrt{n}}t_\delta,\bar{\upsilon}+\frac{S}{\sqrt{n}}t_\delta\right]\ni\mu_0\right)=1-\delta. \tag{21.71}$$

□

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                                    # Nullhypothesenparameter bei Zutreffen von H_0
set.seed(1)                                     # random number generator seed
ns      = 1e2                                   # Anzahl Simulationen
y_bar   = rep(NaN,ns)                           # Stichprobenmittelarray
s       = rep(NaN,ns)                           # Stichprobenstandardabweichungarray
kappa   = matrix(rep(NaN,2*ns), ncol = 2)       # Konfidenzintervallarray
kfn     = rep(NaN,ns)                           # Überdeckungsindikatorarray
phi     = rep(NaN,ns)                           # Testarray
for(i in 1:ns){                                 # Simulationsiterationen
  y          = rnorm(n,mu_0,sqrt(sigsqr))       # 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
  if(kappa[i,1] <= mu_0 & mu_0 <= kappa[i,2]){
      kfn[i] = 1} else{kfn[i] = 0}              # Überdeckungsindikatorevaluation
  if(kappa[i,1] <= mu_0 & mu_0 <= kappa[i,2]){
      phi[i] = 0} else{phi[i] = 1}}             # Testevaluation
cat(   "Geschätztes Konfidenzniveau =", mean(kfn),   # Ausgabe
     "\nGeschätzter Testumfang      =", mean(phi))
```

```
Geschätztes Konfidenzniveau = 0.96
Geschätzter Testumfang      = 0.04
```

Wir visualisieren die Ergebnisse dieser Simulation in Abbildung 21.9.

21.5. 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.

21.6. Selbstkontrollfragen

1. Erläutern Sie die grundlegende Logik Frequentistischer Hypothesentests.
2. Geben Sie die Definition der Begriffe der Testhypothesen und des Testszenario wieder.
3. Geben Sie die Definition der Begriffe der einfachen und zusammengesetzten Testhypothesen wieder.
4. Geben Sie die Definition der Begriffe einseitigen und zweiseitigen Testhypothesen wieder.
5. Geben Sie die Definition des Begriff des Tests wieder.
6. Geben Sie die Definition des Begriffs des Standardtests wieder.
7. Geben Sie die Definition des Begriffs des kritischen Bereichs wieder.
8. Geben Sie die Definition des Begriffs des Ablehungsbereichs wieder
9. Geben Sie die Definition des Begriffs des kritischen Wert-basierten Tests wieder.
10. Geben Sie die Definition der Begriffe der richtigen Testentscheidungen und der Testfehler wieder.
11. Geben Sie die Definition des Begriffs der Testgütefunktion wieder.

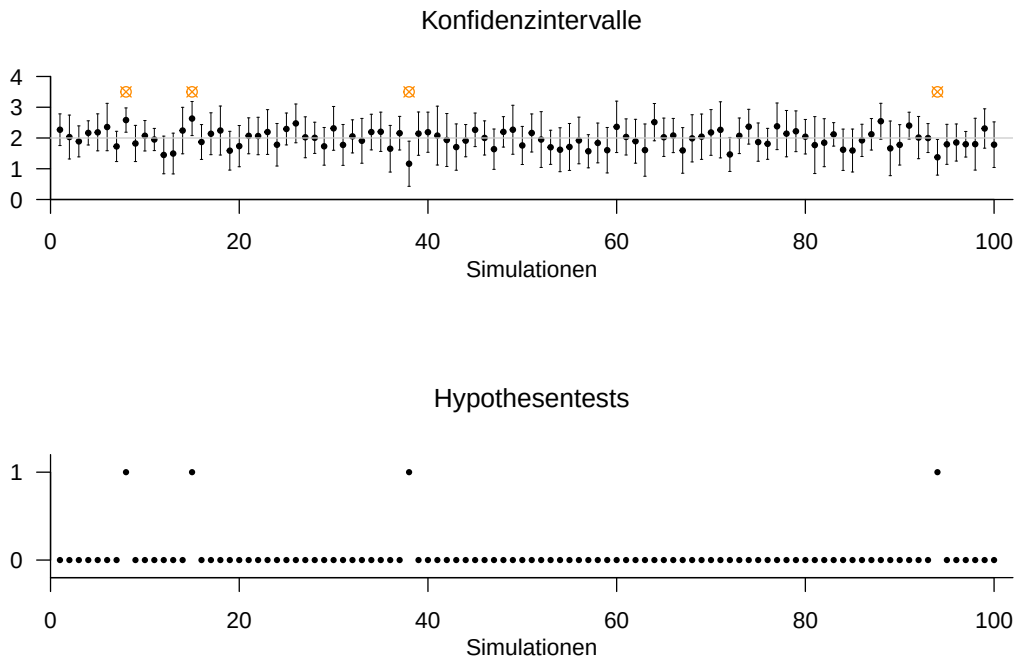


Abbildung 21.9. Dualität von Konfidenzintervall und Hypothesentest am Beispiel des Normalverteilungsmodells. Nutzt man das δ-Konfidenzintervall für den Erwartungswertparameter um bei Nichtübedeckung des Nullhypothesenparameters die Nullhypothese abzulehnen, so ergibt sich in diesem Fall ein Hypothesentest mit Signifikanzlevel $\alpha_0 = 1 - \delta$.

12. Erläutern Sie die Bedeutung der Testgütefunktion im Rahmen der Konstruktion von Hypothesentests.
13. Geben Sie die Definition der Begriffe des Level-α_0-Tests und des Signifikanzlevels α_0 wieder
14. Geben Sie die Definition des Begriffs des Testumfangs α wieder.
15. Geben Sie die Definition des Begriffs des p-Werts wieder.

Referenzen

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

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

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., & Brown, G. K. (2009). *Beck-Depressions-Inventar II. Deutsche Bearbeitung von M. Hautzinger / F. Keller / C. Kühner*. Hogrefe.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Borel, É. (1898). *Leçons Sur La Théorie Des Fonctions*. Paris: Gauthier Villars.

Cantor, G. (1892). Über Eine Eigenschaft Des Inbegriffes Aller Reellen Algebraischen Zahlen. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *1*.

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. (2012). *Statistical Inference*. Duxbury.

Christensen, R. (2011). *Plane Answers to Complex Questions*. Springer New York. https://doi.org/10.1007/978-1-4419-9816-3

De Finetti, B. (1975). *Theory of Probability*. John Wiley & Sons.

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and Statistics* (4th ed). Addison-Wesley.

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

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/rstb.2005.1622

Friston, K., Da Costa, L., Sakthivadivel, D. A. R., Heins, C., Pavliotis, G. A., Ramstead, M., & Parr, T. (2023). Path Integrals, Particular Kinds, and Strange Things. *Physics of Life Reviews*, *47*, 35–62. https://doi.org/10.1016/j.plrev.2023.08.016

Georgii, H.-O. (2009). *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und*

Statistik (4., überarb. und erw. Aufl). de Gruyter.
Gigerenzer, G. (2004). Mindless Statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033
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.
Held, L., & Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37887-4
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.
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
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
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
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.
Newton, I. (1687). *Philosophiae Naturalis Principia Mathematica*. Royal Society.
Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Statistical Stimation*.
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.
Ostwald, D., Kirilina, E., Starke, L., & Blankenburg, F. (2014). A Tutorial on Variational Bayes for Latent Linear Stochastic Time-Series Models. *Journal of Mathematical Psychology*, *60*, 1–19. https://doi.org/10.1016/j.jmp.2014.04.003
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
Pratt, J., Raiffa, H., & Schlaifer, R. (1995). *Statistical Decision Theory*. MIT Press.
Puterman, M. (2005). *Markov Decision Processes*. Wiley-Interscience.
Schmidt, K. D. (2009). *Maß und Wahrscheinlichkeit*. Springer.
Steele, J. M. (2006). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities* (repr). Cambridge University Press [u.a.].

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press.

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1–25.

Unger, L. (2000). *Grundkurs Mathematik*.

Vaart, A. W. van der. (1998). *Asymptotic Statistics*. Cambridge University Press.

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

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