

Einführung

Statistik und Wahrscheinlichkeitsrechnung

Lukas Meier

Teilweise basierend auf Vorlesungsunterlagen von Marloes Maathuis, Hansruedi
Künsch, Peter Bühlmann und Markus Kalisch.

Einführung

In den Natur- und den Ingenieurwissenschaften sind viele Phänomene mit **Unsicherheit** verbunden. Einfache Beispiele sind die jährliche maximale Wasserhöhe bei einem Fluss oder das kumulierte tägliche Verkehrsaufkommen bei einer Brücke. Auch wenn man die Zugfestigkeit von Stahl experimentell ermittelt, ist dies mit Unsicherheit verbunden. Auf der einen Seite wegen der Messungenauigkeit, auf der anderen Seite, weil es eine natürliche Variabilität zwischen Prüfkörpern gibt (keine zwei Prüfkörper sind exakt identisch). Die Unsicherheit kann auch durch fehlendes Wissen auftreten, z.B. weil wir ein Phänomen nicht genügend genau mit deterministischen Modellen beschreiben können.

Wir benötigen also Methoden, um unsichere Phänomene adäquat zu modellieren, aber auch um Daten entsprechend auszuwerten. Aus unseren Daten wollen wir nämlich (korrekte) *Rückschlüsse* ziehen und basierend auf diesen *Entscheidungen* treffen. Um dies zu können, benötigen wir die Wahrscheinlichkeitsrechnung und die Statistik.

In der **Wahrscheinlichkeitsrechnung** geht man aus von einem **Modell** (man beschreibt sozusagen einen datengenerierenden Prozess) und *leitet* davon entsprechende Eigenschaften *ab*. Wie in Abbildung 1 dargestellt, kann man sich unter einem Modell symbolisch eine Urne vorstellen, aus der man Kugeln (Daten) zieht. Wenn wir ein Modell haben für den jährlichen maximalen Wasserstand eines Flusses, so interessiert es uns zum Beispiel, was die Wahrscheinlichkeit ist, dass in einer 100-Jahr Periode der maximale Wasserstand gewisse Höhen überschreitet. Damit können wir versuchen, eine “gute” Dammhöhe zu ermitteln. “Gut” im Sinne, dass der Damm genügend Sicherheit bietet, aber gleichzeitig auch noch finanzierbar ist. Hierzu müssen wir diese Unsicherheit quantifizieren können (z.B. in einer 100-Jahr Periode), wozu wir uns auf die Wahrscheinlichkeitsrechnung stützen.

In der **Statistik** geht es darum, aus vorhandenen Daten auf den datengenerierenden Mechanismus (das Modell) zu *schliessen*. Wir denken also gerade “in die andere Richtung”. Wir sehen ein paar (wenige) Datenpunkte (z.B. Wasserstandsmessungen) und versuchen mit diesem beschränkten Wissen herauszufinden, was wohl ein gutes Modell dafür ist. Abbildung 1 illustriert diese unterschiedlichen “Denkrichtungen”. In der Statistik können wir zusätzlich auch Angaben darüber machen, wie sicher wir über unsere Rückschlüsse sind (was auf den ersten Blick erstaunlich erscheint).

Auch wenn wir Experimente durchführen, erhalten wir Daten, die entsprechend adäquat ausgewertet werden müssen. Wenn Sie also einen Fachartikel beurteilen sollen, dann kommt darin wohl fast immer auch eine Datenanalyse vor. Um entsprechende Fehlschlüsse zu durchschauen (was auch einen Grund für den schlechten Ruf der Statistik ist) benötigen Sie das nötige Rüstzeug. Dieses Skript gibt eine *Einführung* in die beiden Gebiete.

Organisatorisches

Wir beginnen mit der Wahrscheinlichkeitsrechnung, da die Statistik danach auf den entsprechenden Grundlagen aufbaut. In der Mittelschule haben Sie vermutlich Wahrscheinlichkeitsrechnung kennen gelernt durch die Kombinatorik. Das heisst es ging darum, die Anzahl “günstiger Fälle” und die Anzahl “möglicher Fälle” zu bestimmen. Dabei lag die Hauptschwierigkeit oft in der Bestimmung dieser Anzahlen (was hat man z.B. doppelt gezählt etc.). Dies hat wohl vielen unter ihnen Schwierigkeiten bereitet. Die gute Nachricht vorweg: Wir werden dies hier nur am Rande wieder antreffen.

Vielleicht auf den ersten Blick etwas exotisch in der Wahrscheinlichkeitsrechnung “eingeschoben” ist die deskriptive (beschreibende) Statistik. Dies ist einerseits wegen der Koordination mit der Analysis so (mehrdimensionale Integrale), andererseits, weil es sich auch anbietet als Übergang vom eindimensio-

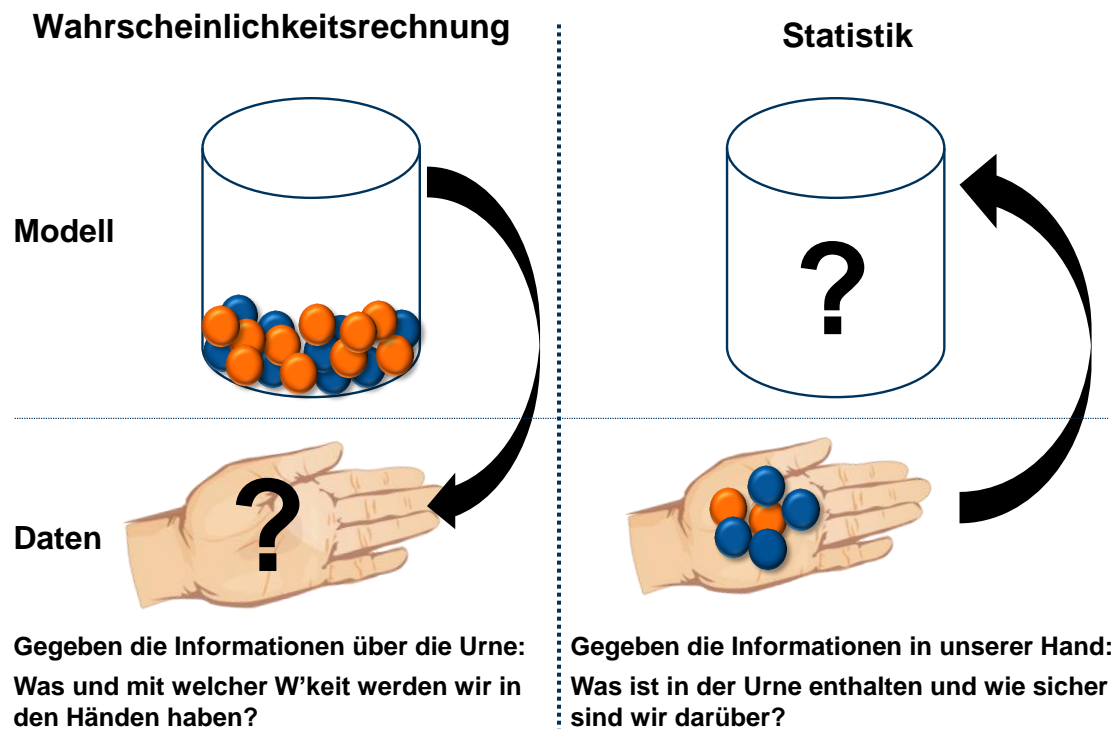


Abbildung 1: Darstellung der Konzepte der Wahrscheinlichkeitsrechnung und der Statistik. Das Modell wird hier durch eine Urne symbolisiert.

nalen zum mehrdimensionalen Fall. Im zweiten Teil folgt dann die schliessende Statistik, wo es darum geht, mit (wenigen) Daten auf den zugrundeliegenden datengenerierenden Prozess zu schliessen.

Wichtige Sachverhalte sind zur besseren Übersichtlichkeit blau hinterlegt. Beispiele sind jeweils entsprechend markiert und kursiv geschrieben. Zudem ist das Ende eines Beispiels zusätzlich mit dem Symbol “◁” hervorgehoben. Lernziele findet man vielleicht etwas unüblich am Ende der entsprechenden Kapitel. Der Grund liegt darin, dass Sie nicht zu Beginn mit den entsprechenden Fachbegriffen “erschlagen” werden sollen. Im Anhang befinden sich diverse Zusammenfassungen und Tabellen sowie einige Herleitungen.

Falls Sie Fehler entdecken oder bei gewissen Kapiteln oder Abschnitten Verständnisschwierigkeiten haben, melden Sie dies unbedingt unter <http://goo.gl/RMv7D> (anonym) bzw. normal per E-Mail an meier@stat.math.ethz.ch. Vielen Dank!

Inhaltsverzeichnis

Einführung	i
I Wahrscheinlichkeitsrechnung und Deskriptive Statistik	1
1 Grundlagen der Wahrscheinlichkeitsrechnung	3
1.1 Grundbegriffe	3
1.2 Diskrete Wahrscheinlichkeitsmodelle	6
1.3 Unabhängigkeit von Ereignissen	7
1.4 Bedingte Wahrscheinlichkeiten	8
1.4.1 Satz der totalen Wahrscheinlichkeit und Satz von Bayes	10
1.5 Review / Lernziele	14
2 Wahrscheinlichkeitsverteilungen	15
2.1 Der Begriff der Zufallsvariable	15
2.1.1 Wahrscheinlichkeitsverteilungen	15
2.2 Diskrete Verteilungen	16
2.2.1 Kennzahlen	18
2.2.2 Bernoulliverteilung [Bernoulli (p)]	19
2.2.3 Binomialverteilung [Bin (n, p)]	20
2.2.4 Geometrische Verteilung [Geom (p)]	20
2.2.5 Poissonverteilung [Pois (λ)]	23
2.3 Stetige Verteilungen	25
2.3.1 Wahrscheinlichkeitsdichte	25
2.3.2 Kennzahlen von stetigen Verteilungen	26
2.3.3 Uniforme Verteilung [Uni (a, b)]	27
2.3.4 Normalverteilung [$\mathcal{N}(\mu, \sigma^2)$]	28
2.3.5 Exponentialverteilung [Exp (λ)]	29
2.3.6 Transformationen	30
2.3.7 Simulation von Zufallsvariablen	33
2.4 Ausblick: Poissonprozesse	33
2.5 Vergleich der Konzepte: Diskrete vs. stetige Verteilungen	34
2.6 Review / Lernziele	34
3 Deskriptive Statistik	37
3.1 Einführung	37
3.2 Kennzahlen	37
3.3 Grafische Darstellungen	39
3.3.1 Histogramm	39
3.3.2 Boxplot	40
3.3.3 Empirische kumulative Verteilungsfunktion	40
3.4 Mehrere Messgrößen	41
3.5 Modell vs. Daten	44
3.6 Review / Lernziele	46

4	Mehrdimensionale Verteilungen	47
4.1	Gemeinsame, Rand- und bedingte Verteilungen	47
4.1.1	Diskreter Fall	47
4.1.2	Stetiger Fall	48
4.2	Erwartungswert bei mehreren Zufallsvariablen	50
4.3	Kovarianz und Korrelation	52
4.4	Zweidimensionale Normalverteilung	54
4.5	Dichte einer Summe von zwei Zufallsvariablen	55
4.6	Mehr als zwei Zufallsvariablen	57
4.7	Vergleich der Konzepte: Diskrete vs. stetige mehrdimensionale Verteilungen	57
4.8	Review / Lernziele	57
5	Grenzwertsätze	59
5.1	Die i.i.d. Annahme	59
5.2	Summen und arithmetische Mittel von Zufallsvariablen	59
5.3	Das Gesetz der Grossen Zahlen und der Zentrale Grenzwertsatz	60
5.4	Review / Lernziele	62
II	Schliessende Statistik	63
6	Parameterschätzungen	65
6.1	Einführung in die schliessende Statistik	65
6.1.1	Daten als Realisierungen von Zufallsvariablen	65
6.1.2	Überblick über die Konzepte	65
6.2	Wahl der Verteilungsfamilie	66
6.3	Methoden zur Parameterschätzung	69
6.3.1	Momentenmethode	71
6.3.2	Maximum-Likelihood Methode	72
6.3.3	Allgemeine Schätzer für Erwartungswert und Varianz	74
6.3.4	Genauigkeit von Schätzern – Ein erster Ansatz	75
6.4	Review / Lernziele	76
7	Statistische Tests und Vertrauensintervalle für eine Stichprobe	77
7.1	Illustration der Konzepte mit der Binomialverteilung: Binomialtest	77
7.2	Tests für eine Stichprobe bei normalverteilten Daten	80
7.2.1	Z-Test (σ bekannt)	81
7.2.2	t-Test (σ unbekannt)	82
7.3	Allgemeine Eigenschaften von statistischen Tests	84
7.3.1	Macht	84
7.3.2	P-Wert	87
7.3.3	Multiples Testen	89
7.4	Vertrauensintervalle	90
7.4.1	Statistische Signifikanz und fachliche Relevanz	93
7.5	Tests für eine Stichprobe bei nicht normalverteilten Daten	93
7.5.1	Vorzeichen-Test	93
7.5.2	Wilcoxon-Test	94
7.6	Rückblickender Überblick über Konzepte	97
7.6.1	Vorgehen und Fragen bei statistischen Tests	97
7.7	Review / Lernziele	100
8	Vergleich zweier Stichproben	101
8.1	Gepaarte und ungepaarte Stichproben	101
8.2	Grundlegende Gedanken zur Versuchsplanung	102
8.3	Gepaarte Vergleiche	104

8.4	Zwei-Stichproben Tests	104
8.5	Vergleich der Konzepte	107
8.6	Review / Lernziele	107
9	Ausblick: Lineare Regression	109
9.1	Einführung	109
9.2	Einfache lineare Regression	109
9.2.1	Modell	109
9.2.2	Parameterschätzungen	110
9.2.3	Tests und Vertrauensintervalle	113
9.2.4	Residuenanalyse	115
9.3	Multiple lineare Regression	117
9.3.1	Modell	117
9.3.2	Parameterschätzungen	119
9.3.3	Tests und Vertrauensintervalle	119
9.4	Review / Lernziele	120
III	Anhänge	121
A	Zusammenfassungen und Tabellen	123
A.1	Die wichtigsten eindimensionalen Verteilungen	123
A.2	Die wichtigsten Rechenregeln für Erwartungswert, Varianz und Kovarianz	124
A.3	Tabelle der Standardnormalverteilung	126
A.4	Quantile der t -Verteilung	127
B	Alternative Ansätze	129
B.1	Dialog: Dr. Nulli vs. Prof. Altmeier	129
C	Herleitungen	131
C.1	Herleitung der Binomialverteilung	131
C.2	Uneigentliche Integrale	133
	Literaturverzeichnis	135
	Index	137

Teil I

Wahrscheinlichkeitsrechnung und Deskriptive Statistik

1 Grundlagen der Wahrscheinlichkeitsrechnung

1.1 Grundbegriffe

Die Wahrscheinlichkeitsrechnung befasst sich mit **Zufallsexperimenten**. Bei einem Zufallsexperiment ist der Ausgang nicht (exakt) vorhersagbar. Zudem erhalten wir unter “gleichen Versuchsbedingungen” jeweils verschiedene Ergebnisse.

Für einfache Beispiele greift man oft auf Glücksspiele wie z.B. Würfel oder Roulette zurück. Es ist uns bewusst, dass diese nichts mit ihrem Fachgebiet zu tun haben. Oft eignen sie sich aber für kurze Illustrationen, insbesondere jetzt am Anfang. Daher erlauben wir uns, diese ab und zu zu verwenden.

Wenn man z.B. die Druckfestigkeit von Beton misst, ist dies auch ein Zufallsexperiment. Die Messung enthält einen Messfehler und zudem gibt es sicher eine (kleine) Variation von Prüfkörper zu Prüfkörper. Von einer Serie von 10 Prüfkörpern aus der gleichen Produktion werden wir also für jeden Prüfkörper einen (leicht) anderen Wert erhalten.

Um richtig loslegen zu können, müssen wir am Anfang viele Begriffe neu einführen. Wir werden versuchen, so wenig wie möglich “abstrakt” zu behandeln (aber so viel wie nötig) und hoffen, dass diese Durststrecke erträglich kurz bleibt.

Für ein Zufallsexperiment führen wir folgende Begriffe ein:

- **Elementarereignis** ω : Ein möglicher Ausgang des Zufallsexperiments.
- **Grundraum** Ω : Die Menge *aller* Elementarereignisse, d.h. die Menge aller möglichen Ausgänge des Zufallsexperiments.
- **Ereignis**: Eine Kollektion von *gewissen* Elementarereignissen, also eine Teilmenge $A \subseteq \Omega$. “Ereignis A tritt ein” heisst: Der Ausgang ω des Zufallsexperiments liegt in A . Oft beschreiben wir ein Ereignis auch einfach nur in Worten, siehe auch die Beispiele unten.

Wie sieht das an einem konkreten Beispiel aus?

Beispiel. *Eine Münze 2 Mal werfen*
Mit K bezeichnen wir “Kopf” und mit Z “Zahl”.

Ein Elementarereignis ist zum Beispiel $\omega = ZK$: Im ersten Wurf erscheint “Zahl” und im zweiten “Kopf”.

Es ist $\Omega = \{KK, KZ, ZK, ZZ\}$, Ω hat also 4 Elemente. Wir schreiben auch $|\Omega| = 4$.

Das Ereignis “Es erscheint genau 1 Mal Kopf” ist gegeben durch die Menge $A = \{KZ, ZK\}$. ◁

Beispiel. *Messung der Druckfestigkeit von Beton [MPa, Megapascal]*

Das Resultat ist hier eine Messgrösse. Ein Elementarereignis ist einfach eine positive reelle Zahl, z.B. $\omega = 31.2$ MPa.

Es ist also $\Omega = \mathbb{R}_+$ (die Menge der positiven reellen Zahlen).

Das Ereignis “Die Druckfestigkeit liegt zwischen 10 und 20 MPa” ist gegeben durch das Intervall $A = [10, 20]$ MPa. ◁

Oft betrachtet man mehrere Ereignisse zusammen, z.B. ein Ereignis A und ein Ereignis B . Man

interessiert sich z.B. dafür, wie wahrscheinlich es ist, dass A und B *gemeinsam* eintreten oder man interessiert sich für die Wahrscheinlichkeit, dass *mindestens eines* der beiden Ereignisse eintritt.

Für solche Fälle ist es nützlich, sich die Operationen der Mengenlehre und deren Bedeutung in Erinnerung zu rufen.

Name	Symbol	Bedeutung
Durchschnitt	$A \cap B$	“ A und B ”
Vereinigung	$A \cup B$	“ A oder B ” (“oder” zu verstehen als “und/oder”)
Komplement	A^c	“nicht A ”
Differenz	$A \setminus B = A \cap B^c$	“ A ohne B ”

Tabelle 1.1: Operationen der Mengenlehre und ihre Bedeutung.

Statt dem Wort “Durchschnitt” verwendet man manchmal auch den Begriff “Schnittmenge”.

A und B heißen **disjunkt** (d.h. A und B schliessen sich gegenseitig aus und können daher nicht zusammen eintreten), falls $A \cap B = \emptyset$, wobei wir mit \emptyset die **leere Menge** (d.h. das unmögliche Ereignis) bezeichnen.

Ferner gelten die sogenannten **De Morgan’sche Regeln**

- $(A \cap B)^c = A^c \cup B^c$
- $(A \cup B)^c = A^c \cap B^c$.

Alle diese Begriffe, Operationen und Regeln lassen sich einfach mit sogenannten Venn-Diagrammen illustrieren, siehe Abbildung 1.1.

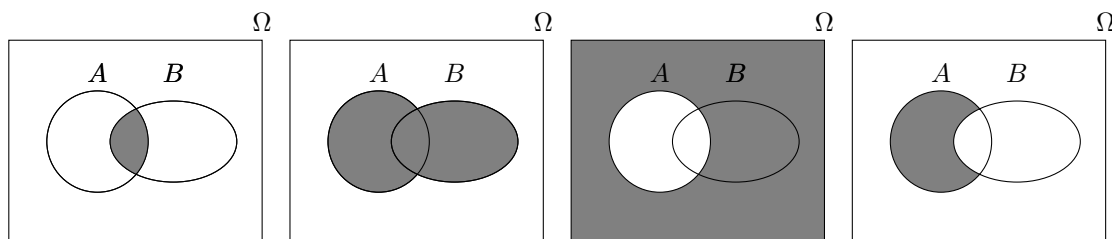


Abbildung 1.1: Illustration der Operationen der Mengenlehre an Venn-Diagrammen: $A \cap B$, $A \cup B$, A^c und $A \setminus B$ jeweils entsprechend markiert (von links nach rechts).

Beispiel. Sei A das Ereignis “Stahlträger 1 hat strukturelle Mängel” und B das entsprechende Ereignis bei Stahlträger 2. Das Ereignis $A \cup B$ bedeutet dann: “Mindestens einer der beiden Stahlträger hat strukturelle Mängel” (dies beinhaltet die Möglichkeit, dass beide Mängel haben). Die Schnittmenge $A \cap B$ ist das Ereignis “Beide Stahlträger haben strukturelle Mängel”, A^c bedeutet, dass Stahlträger 1 keine Mängel aufweist, etc. \triangleleft

Bis jetzt haben wir zwar teilweise schon den Begriff “Wahrscheinlichkeit” verwendet, diesen aber noch nicht spezifiziert.

Wir kennen also den Grundraum Ω bestehend aus Elementarereignissen ω und mögliche Ereignisse A, B, C, \dots . Jetzt wollen wir einem Ereignis aber noch eine Wahrscheinlichkeit zuordnen und schauen, wie man mit Wahrscheinlichkeiten rechnen muss.

Für ein Ereignis A bezeichnen wir mit $\mathbb{P}(A)$ die **Wahrscheinlichkeit**, dass das Ereignis A eintritt (d.h. dass der Ausgang ω des Zufallsexperiments in der Menge A liegt). Bei einem Wurf mit einer fairen Münze wäre für A =“Münze zeigt Kopf” also $\mathbb{P}(A) = 0.5$.

Es müssen die folgenden Rechenregeln (die sogenannten Axiome der Wahrscheinlichkeitsrechnung von Kolmogorov) erfüllt sein.

Axiome der Wahrscheinlichkeitsrechnung (Kolmogorov)

$$(A1) \quad 0 \leq \mathbb{P}(A) \leq 1$$

$$(A2) \quad \mathbb{P}(\Omega) = 1$$

$$(A3) \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \quad \text{für alle Ereignisse } A, B \text{ die sich gegenseitig ausschliessen (d.h. } A \cap B = \emptyset).$$

Bzw. allgemeiner:

$$\mathbb{P}(A_1 \cup A_2 \cup \dots) = \sum_{i \geq 1} \mathbb{P}(A_i) \quad \text{für } A_k \cap A_l = \emptyset, k \neq l.$$

(A1) bedeutet, dass Wahrscheinlichkeiten immer zwischen 0 und 1 liegen und (A2) besagt, dass das sichere Ereignis Ω Wahrscheinlichkeit 1 hat.

Weitere Rechenregeln werden daraus abgeleitet, z.B.

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \quad \text{für jedes Ereignis } A \quad (1.1)$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \quad \text{für je zwei Ereignisse } A \text{ und } B \quad (1.2)$$

$$\mathbb{P}(A_1 \cup \dots \cup A_n) \leq \mathbb{P}(A_1) + \dots + \mathbb{P}(A_n) \quad \text{für je } n \text{ Ereignisse } A_1, \dots, A_n \quad (1.3)$$

$$\mathbb{P}(B) \leq \mathbb{P}(A) \quad \text{für je zwei Ereignisse } A \text{ und } B \text{ mit } B \subseteq A \quad (1.4)$$

$$\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(B) \quad \text{für je zwei Ereignisse } A \text{ und } B \text{ mit } B \subseteq A \quad (1.5)$$

Wenn man sich Wahrscheinlichkeiten als Flächen im Venn-Diagramm vorstellt (die Totalfläche von Ω ist 1), so erscheinen diese Rechenregeln ganz natürlich. Verifizieren Sie dies als Übung für alle obigen Regeln.

Interpretation von Wahrscheinlichkeiten

Wir haben gesehen, welche Rechenregeln Wahrscheinlichkeiten erfüllen müssen. Doch wie interpretiert man eine Wahrscheinlichkeit überhaupt? Die beiden wichtigsten Interpretationen sind die “Idealisierung der relativen Häufigkeit bei vielen unabhängigen Wiederholungen” (die sogenannte **frequentistische Interpretation**) und das (subjektive) “Mass für den Glauben, dass ein Ereignis eintreten wird” (die sogenannte **bayes’sche Interpretation**).

Zur frequentistischen Interpretation:

Wenn ein Ereignis A eines Zufallsexperiments Wahrscheinlichkeit $1/2$ hat, so werden wir bei vielen unabhängigen Wiederholungen des Experiments bei ca. der Hälfte der Fälle sehen, dass das Ereignis eingetreten ist (eine mathematische Definition für Unabhängigkeit werden wir später sehen). Für eine unendliche Anzahl Wiederholungen würden wir exakt $1/2$ erreichen. Man denke z.B. an den Wurf mit einer Münze. Wenn man die Münze sehr oft wirft, so wird die relative Häufigkeit von “Kopf” nahe bei $1/2$ liegen, siehe Abbildung 1.2. Die frequentistische Interpretation geht also insbesondere von einer Wiederholbarkeit des Zufallsexperiments aus.

Etwas formeller: Sei $f_n(A)$ die relative Häufigkeit des Auftretens des Ereignisses A in n unabhängigen Experimenten. Dieses Mass $f_n(\cdot)$ basiert auf **Daten** oder **Beobachtungen**. Falls n gross wird, so gilt

$$f_n(A) \xrightarrow{n \rightarrow \infty} \mathbb{P}(A).$$

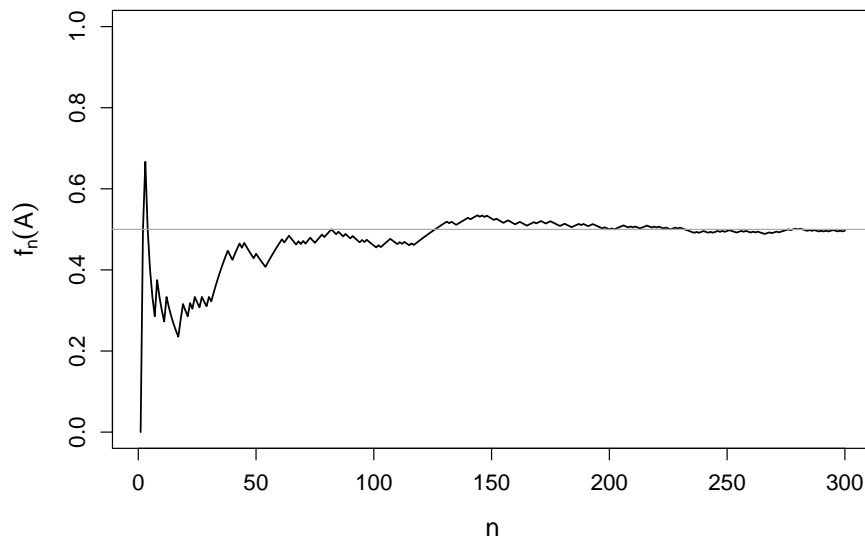


Abbildung 1.2: Relative Häufigkeiten $f_n(A)$ für das Ereignis A ="Münze zeigt Kopf" beim Wurf mit einer Münze in Abhängigkeit der Anzahl Würfe n .

Man beachte, dass $\mathbb{P}(A)$ also ein theoretisches Mass in einem **Modell** ist (wo keine Experimente oder Daten vorliegen).

Zur bayes'schen Interpretation:

Hier ist $\mathbb{P}(A)$ ein Mass für den Glauben, dass ein Ereignis eintreten wird. Sie vermuten zum Beispiel, dass mit Wahrscheinlichkeit 15% auf ihrem Grundstück Ölvorräte vorhanden sind. Dies heisst nicht, dass wenn Sie auf ihrem Grundstück viele Bohrungen machen, im Schnitt in 15% der Bohrlöcher Öl vorliegen wird. Denn: entweder ist das Öl da oder es ist nicht da.

Je nach Problemstellung eignet sich die eine oder die andere Interpretation.

1.2 Diskrete Wahrscheinlichkeitsmodelle

Für den Moment nehmen wir an, dass Ω entweder endlich viele Elemente enthält (d.h. $|\Omega| < \infty$) oder dass Ω abzählbar ist (d.h. wir können die Elemente durchnummerieren). Wir können Ω also schreiben als

$$\Omega = \{\omega_1, \omega_2, \dots\}.$$

Man spricht in diesem Fall auch von einem sogenannten **diskreten Wahrscheinlichkeitsmodell**. Das Beispiel mit dem Münzwurf passt in dieses Schema, während dies beim Beispiel mit der Druckfestigkeit des Betons *nicht* der Fall ist, da man die reellen Zahlen nicht durchnummerieren kann. Wie man mit diesem Fall umgeht, werden wir im nächsten Kapitel sehen.

Da Elementarereignisse per Definition disjunkt sind, können wir wegen (A3) die Wahrscheinlichkeit $\mathbb{P}(A)$ schreiben als

$$\mathbb{P}(A) = \sum_{k: \omega_k \in A} \mathbb{P}(\{\omega_k\}),$$

wobei wir mit $\{k: \omega_k \in A\}$ einfach alle Elementarereignisse "sammeln", die in A liegen (A ist ja eine Menge von Elementarereignissen). Wenn wir also die Wahrscheinlichkeiten der Elementarereignisse kennen, können wir die Wahrscheinlichkeit eines Ereignisses A berechnen, indem wir die entsprechenden Wahrscheinlichkeiten der passenden Elementarereignisse ganz simpel aufsummieren. Wir schreiben hier bewusst $\{\omega_k\}$ in geschweiften Klammern, um zu unterstreichen, dass wir eine *Menge* (d.h. ein

Ereignis) meinen mit *einem* Element ω_k . Ferner gilt

$$1 \stackrel{(A2)}{=} \mathbb{P}(\Omega) \stackrel{(A3)}{=} \sum_{k \geq 1} \mathbb{P}(\{\omega_k\}).$$

Die Summe der Wahrscheinlichkeiten aller Elementarereignisse muss also immer 1 ergeben.

Also: Wenn uns jemand eine “Liste” gibt mit allen Elementarereignissen und deren Wahrscheinlichkeiten, dann muss zwangsläufig die Summe von diesen Wahrscheinlichkeiten 1 ergeben und zudem dient uns diese “Liste” als Werkzeug, um die Wahrscheinlichkeit $\mathbb{P}(A)$ eines *beliebigen* Ereignisses A zu berechnen.

Woher kriegen wir diese “Liste” im Alltag? Falls Ω endlich ist, ist das einfachste Modell das **Modell von Laplace**. Dieses nimmt an, dass alle Elementarereignisse *gleich wahrscheinlich* sind. Beim Beispiel mit dem Münzwurf ist dies sicher eine sinnvolle Annahme. Bei einer fairen Münze haben wir *keine* Präferenz, dass ein möglicher Ausgang des Experiments (ein Elementarereignis) wahrscheinlicher ist als ein anderer.

Damit sich die Wahrscheinlichkeiten aller Elementarereignisse zu 1 addieren (siehe oben), haben wir hier

$$\mathbb{P}(\{\omega_k\}) = \frac{1}{|\Omega|}, \quad k \geq 1.$$

Für ein Ereignis A gilt also im Laplace-Modell

$$\mathbb{P}(A) = \sum_{k: \omega_k \in A} \mathbb{P}(\{\omega_k\}) = \sum_{k: \omega_k \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl günstige Fälle}}{\text{Anzahl mögliche Fälle}}.$$

Dies kennen Sie vermutlich aus der Mittelschule. Dort bestand dann die Wahrscheinlichkeitsrechnung in der Regel darin, durch (mühsames) Abzählen die Anzahl günstiger Fälle zu bestimmen.

Beispiel. *Münzwurf*

Für die Elementarereignisse haben wir also

$$\mathbb{P}(\{KK\}) = \mathbb{P}(\{KZ\}) = \mathbb{P}(\{ZK\}) = \mathbb{P}(\{ZZ\}) = \frac{1}{4}.$$

Für das Ereignis $A = \{KZ, ZK\}$ (genau 1 Mal Kopf) gilt demnach

$$\mathbb{P}(A) = \mathbb{P}(\{KZ\}) + \mathbb{P}(\{ZK\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \quad \triangleleft$$

Wie wir aber sehen werden, geht die Wahrscheinlichkeitsrechnung weit über das Laplace-Modell hinaus. Insbesondere ist das Laplace-Modell für viele Anwendungen ungeeignet.

1.3 Unabhängigkeit von Ereignissen

Wenn man die Wahrscheinlichkeiten $\mathbb{P}(A)$ und $\mathbb{P}(B)$ kennt, so können wir nur aus diesen Angaben allein die Wahrscheinlichkeit $\mathbb{P}(A \cap B)$ im Allgemeinen *nicht* berechnen (siehe Venn-Diagramm!). Es kann z.B. sein, dass die Schnittmenge die leere Menge ist oder dass B ganz in A liegt bzw. umgekehrt. Wir sehen anhand der einzelnen Wahrscheinlichkeiten $\mathbb{P}(A)$ und $\mathbb{P}(B)$ also nicht, was für eine Situation vorliegt und können damit $\mathbb{P}(A \cap B)$ *nicht* berechnen.

Eine Ausnahme bildet der Fall, wenn folgende Produktformel gilt

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

Man nennt dann A und B (**stochastisch**) **unabhängig**.

Man multipliziert in diesem Fall einfach die Wahrscheinlichkeiten. Wenn also A mit Wahrscheinlichkeit $1/3$ eintritt und B mit Wahrscheinlichkeit $1/6$, dann sehen wir sowohl A wie auch B (also $A \cap B$) mit Wahrscheinlichkeit $1/18$, wenn die Ereignisse unabhängig sind. Bei einer grossen Population (n gross) “sammeln” wir also zuerst alle Fälle, bei denen A eintritt (ca. $1/3$) und *davon* nochmals diejenigen, bei denen B eintritt (ca. $1/6$) und haben am Schluss so noch ca. $1/18$ der ursprünglichen Fälle. Das Ereignis B “kümmert es also nicht”, ob A schon eingetroffen ist oder nicht, die Wahrscheinlichkeit $1/6$ bleibt. Dies muss nicht immer der Fall sein, siehe auch das Beispiel unten.

Typischerweise wird die Unabhängigkeit basierend auf physikalischen und technischen Überlegungen postuliert, d.h. man *nimmt an*, dass obige Produktformel gilt.

Achtung. *Unabhängige Ereignisse sind nicht disjunkt und disjunkte Ereignisse sind nicht unabhängig (ausser wenn ein Ereignis Wahrscheinlichkeit 0 hat). Unabhängigkeit hängt ab von den Wahrscheinlichkeiten, während Disjunktheit nur ein mengentheoretischer Begriff ist.*

Beispiel. *Ein Gerät bestehe aus zwei Bauteilen und funktioniere, solange mindestens eines der beiden Bauteile noch in Ordnung ist. A_1 und A_2 seien die Ereignisse, dass Bauteil 1 bzw. Bauteil 2 defekt sind mit entsprechenden Wahrscheinlichkeiten $\mathbb{P}(A_1) = 1/100$ und $\mathbb{P}(A_2) = 1/100$. Wir wollen zudem davon ausgehen, dass die beiden Ereignisse A_1 und A_2 unabhängig voneinander sind.*

Die Ausfallwahrscheinlichkeit für das Gerät ist also wegen der Unabhängigkeit gegeben durch

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) \mathbb{P}(A_2) = \frac{1}{100} \cdot \frac{1}{100} = 10^{-4}.$$

Wir sehen also, dass durch die Annahme der Unabhängigkeit eine kleine Ausfallwahrscheinlichkeit resultiert. Wenn in Tat und Wahrheit aufgrund eines Ausfalls des einen Bauteils das andere Bauteil auch gerade ausfällt (wodurch die Unabhängigkeit nicht mehr gegeben ist), dann steigt die Ausfallwahrscheinlichkeit des Geräts auf $1/100$ an (da in diesem Fall $A_1 = A_2$ und somit $A_1 \cap A_2 = A_1 = A_2$)! \triangleleft

Wenn man also Wahrscheinlichkeiten unter der Annahme von Unabhängigkeit berechnet, diese aber in der Realität nicht erfüllt ist, so kann das Resultat um einige Grössenordnungen falsch sein!

Der Begriff der Unabhängigkeit kann auch auf mehrere Ereignisse erweitert werden: Die n Ereignisse A_1, \dots, A_n heissen **unabhängig**, wenn für jedes $k \leq n$ und alle $1 \leq i_1 < \dots < i_k \leq n$ gilt

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Dies bedeutet nichts anderes, als dass die entsprechende Produktformel für alle k -Tupel von Ereignissen gelten muss.

1.4 Bedingte Wahrscheinlichkeiten

Wenn zwei Ereignisse *nicht* unabhängig sind, können wir also durch das (Nicht-) Eintreten des einen Ereignisses etwas über das andere aussagen (oder “lernen”).

Beispiel. *Eine Konstruktion besteht aus zwei Stahlträgern. A priori nehmen wir an, dass ein Träger mit einer gewissen Wahrscheinlichkeit Korrosionsschäden aufweist. Wenn wir jetzt aber wissen, dass der erste Stahlträger Korrosionsschäden hat, werden wir vermutlich annehmen, dass in diesem Falle der zweite Träger eher auch betroffen ist (da sie z.B. aus der selben Produktion stammen und den gleichen Witterungsbedingungen ausgesetzt waren etc.). Die Wahrscheinlichkeit für Korrosionsschäden beim zweiten Träger (dessen Zustand wir noch nicht kennen) würden wir also nach Erhalt der Information über den ersten Träger höher einschätzen als ursprünglich. \triangleleft*

Dies führt zum Konzept der bedingten Wahrscheinlichkeiten. Diese treten zum Beispiel dann auf, wenn ein Zufallsexperiment aus verschiedenen Stufen besteht und man sukzessive das Resultat der entsprechenden Stufen erfährt. Oder salopper: “Die Karten (die Unsicherheit) werden sukzessive aufgedeckt”.

Die **bedingte Wahrscheinlichkeit von A gegeben B** ist definiert als

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

Die Interpretation ist folgendermassen: “ $\mathbb{P}(A | B)$ ist die Wahrscheinlichkeit für das Ereignis A , wenn wir *wissen*, dass das Ereignis B schon eingetroffen ist”.

Wie kann man die Formel verstehen? Da wir wissen, dass B schon eingetreten ist (wir haben also einen neuen Grundraum $\Omega' = B$), müssen wir von A nur noch denjenigen Teil anschauen, der sich in B abspielt (daher $A \cap B$). Dies müssen wir jetzt noch in Relation zur Wahrscheinlichkeit von B bringen: die Normierung mit $\mathbb{P}(B)$ sorgt gerade dafür, dass $\mathbb{P}(\Omega') = \mathbb{P}(B) = 1$. Dies ist auch in Abbildung 1.3 illustriert. Wenn man wieder mit Flächen denkt, dann ist die bedingte Wahrscheinlichkeit $\mathbb{P}(A | B)$ der Anteil der schraffierten Fläche an der Fläche von B .

Bemerkung: In der Definition sind wir stillschweigend davon ausgegangen, dass $\mathbb{P}(B) > 0$ gilt.

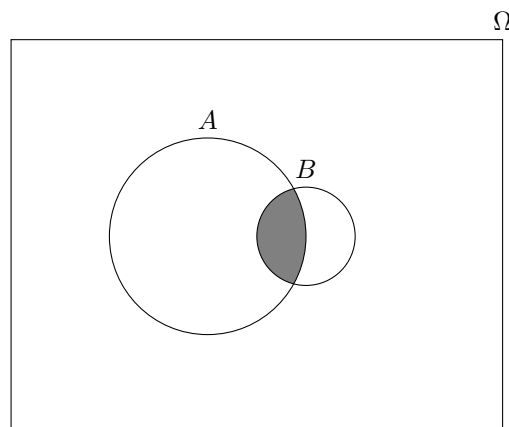


Abbildung 1.3: Hilfsillustration für bedingte Wahrscheinlichkeiten.

Beispiel. *Würfel*

Was ist die Wahrscheinlichkeit, eine 6 zu würfeln? Offensichtlich $1/6$! Was ist die Wahrscheinlichkeit, eine 6 zu haben, wenn wir wissen, dass eine gerade Zahl gewürfelt wurde?

Wir haben hier

$$\Omega = \{1, \dots, 6\}, A = \{6\} \text{ und } B = \{2, 4, 6\}.$$

Also ist $A \cap B = \{6\}$. Weiter ist $\mathbb{P}(B) = 3/6 = 1/2$. Dies liefert

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Durch die zusätzliche Information (gerade Augenzahl) hat sich die Wahrscheinlichkeit für eine 6 also geändert. \triangleleft

Bedingte Wahrscheinlichkeiten sind nichts anderes als Wahrscheinlichkeiten für spezielle Situationen. Es gelten daher wieder die von früher bekannten Rechenregeln.

Rechenregeln

$$\begin{aligned}
0 &\leq \mathbb{P}(A \mid B) \leq 1 && \text{für jedes Ereignis } A \\
\mathbb{P}(B \mid B) &= 1 \\
\mathbb{P}(A_1 \cup A_2 \mid B) &= \mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B) && \text{für } A_1, A_2 \text{ disjunkt (d.h. } A_1 \cap A_2 = \emptyset) \\
\mathbb{P}(A^c \mid B) &= 1 - \mathbb{P}(A \mid B) && \text{für jedes Ereignis } A
\end{aligned}$$

So lange man am “bedingenden Ereignis” B nichts ändert, kann man also mit bedingten Wahrscheinlichkeiten wie gewohnt rechnen. Sobald man aber das bedingende Ereignis ändert, muss man sehr vorsichtig sein (siehe unten).

Weiter gilt für zwei Ereignisse A, B mit $\mathbb{P}(A) > 0$ und $\mathbb{P}(B) > 0$:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B) \mathbb{P}(B) = \mathbb{P}(B \mid A) \mathbb{P}(A) \quad (1.6)$$

Deshalb können wir die Unabhängigkeit auch folgendermassen definieren:

$$A, B \text{ unabhängig} \iff \mathbb{P}(A \mid B) = \mathbb{P}(A) \iff \mathbb{P}(B \mid A) = \mathbb{P}(B) \quad (1.7)$$

Unabhängigkeit von A und B bedeutet also, dass sich die Wahrscheinlichkeiten *nicht* ändern, wenn wir wissen, dass das andere Ereignis schon eingetreten ist. Oder nochmals: “Wir können nichts von A über B lernen” (bzw. umgekehrt).

Achtung

Oft werden im Zusammenhang mit bedingten Wahrscheinlichkeiten falsche Rechenregeln verwendet und damit falsche Schlussfolgerungen gezogen. Man beachte, dass im Allgemeinfall

$$\begin{aligned}
\mathbb{P}(A \mid B) &\neq \mathbb{P}(B \mid A) \\
\mathbb{P}(A \mid B^c) &\neq 1 - \mathbb{P}(A \mid B).
\end{aligned}$$

Man kann also bedingte Wahrscheinlichkeiten in der Regel nicht einfach “umkehren” (erste Gleichung). Dies ist auch gut in Abbildung 1.3 ersichtlich. $\mathbb{P}(A \mid B)$ ist dort viel grösser als $\mathbb{P}(B \mid A)$.

1.4.1 Satz der totalen Wahrscheinlichkeit und Satz von Bayes

Wie wir in (1.6) gesehen haben, kann man

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B) \mathbb{P}(B)$$

schreiben, d.h. $\mathbb{P}(A \cap B)$ ist bestimmt durch $\mathbb{P}(A \mid B)$ und $\mathbb{P}(B)$. In vielen Anwendungen wird dieser Weg beschritten. Man legt die Wahrscheinlichkeiten für die erste Stufe $\mathbb{P}(B)$ und die bedingten Wahrscheinlichkeiten $\mathbb{P}(A \mid B)$ und $\mathbb{P}(A \mid B^c)$ für die zweite Stufe gegeben die erste fest (aufgrund von Daten, Plausibilität und subjektiven Einschätzungen). Dann lassen sich die übrigen Wahrscheinlichkeiten berechnen.

Beispiel. Es sei z.B. A = “Ein Unfall passiert” und B = “Strasse ist nass”. Wir nehmen an, dass wir folgendes kennen

$$\begin{aligned}
\mathbb{P}(A \mid B) &= 0.01 \\
\mathbb{P}(A \mid B^c) &= 0.001 \\
\mathbb{P}(B) &= 0.2.
\end{aligned}$$

Mit den Rechenregeln für Wahrscheinlichkeiten erhalten wir $\mathbb{P}(B^c) = 1 - \mathbb{P}(B) = 0.8$. Können wir damit die Wahrscheinlichkeit für A bestimmen? Wir können A schreiben als disjunkte Vereinigung (siehe Venn-Diagramm)

$$A = (A \cap B) \cup (A \cap B^c).$$

Daher haben wir

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \\ &= \mathbb{P}(A | B) \mathbb{P}(B) + \mathbb{P}(A | B^c) \mathbb{P}(B^c) \\ &= 0.01 \cdot 0.2 + 0.001 \cdot 0.8.\end{aligned}$$

Dies ergibt $\mathbb{P}(A) = 0.0028$. Mit der Wahrscheinlichkeit von B und den bedingten Wahrscheinlichkeiten von A gegeben B bzw. B^c können wir also die Wahrscheinlichkeit von A berechnen. \triangleleft

Wir schauen also in den einzelnen Situationen (B bzw. B^c), was die bedingte Wahrscheinlichkeit für A ist und gewichten diese mit den entsprechenden Wahrscheinlichkeiten $\mathbb{P}(B)$ bzw. $\mathbb{P}(B^c)$.

Dieses Vorgehen wird besonders anschaulich, wenn man das Zufallsexperiment als sogenannten **Wahrscheinlichkeitsbaum** darstellt, siehe Abbildung 1.4. In jeder Verzweigung ist die Summe der (bedingten) Wahrscheinlichkeiten jeweils 1. Um die Wahrscheinlichkeit für eine spezifische “Kombination” (z.B. $A^c \cap B$) zu erhalten, muss man einfach dem entsprechenden Pfad entlang “durchmultiplizieren”. Um die Wahrscheinlichkeit von A zu erhalten, muss man alle Pfade betrachten, die A enthalten und die entsprechenden Wahrscheinlichkeiten aufsummieren.

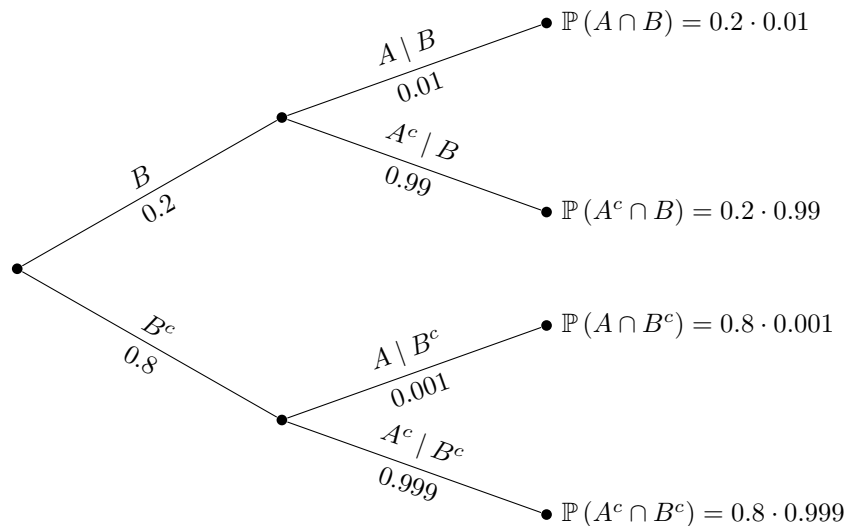


Abbildung 1.4: Wahrscheinlichkeitsbaum.

Diese Aufteilung in verschiedene sich gegenseitig ausschließende Situationen (B , B^c) funktioniert noch viel allgemeiner und führt zum Satz der totalen Wahrscheinlichkeit.

Satz der totalen Wahrscheinlichkeit

Wir nehmen an, dass wir k disjunkte Ereignisse B_1, \dots, B_k haben mit

$$B_1 \cup \dots \cup B_k = \Omega \quad (\text{“alle möglichen Fälle sind abgedeckt”})$$

Dann gilt

$$\mathbb{P}(A) \stackrel{(A3)}{=} \sum_{i=1}^k \mathbb{P}(A \cap B_i) \stackrel{(1.6)}{=} \sum_{i=1}^k \mathbb{P}(A | B_i) \mathbb{P}(B_i).$$

Dies ist genau gleich wie beim einführenden Beispiel mit der Strasse und den Unfällen (dort hatten wir $B_1 = B$ und $B_2 = B^c$). Wir haben jetzt einfach k verschiedene “Verzweigungen”. Wenn wir also die (bedingte) Wahrscheinlichkeit von A in jeder Situation B_i wissen, dann ist die Wahrscheinlichkeit von A einfach deren gewichtete Summe, wobei die Gewichte durch $\mathbb{P}(B_i)$ gegeben sind.

B_1, \dots, B_k heisst auch **Partition** von Ω . Sie deckt alle möglichen Fälle ab und zwei Ereignisse B_i und B_j können nicht zusammen eintreten. Ein Illustration einer Partition findet man in Abbildung 1.5.

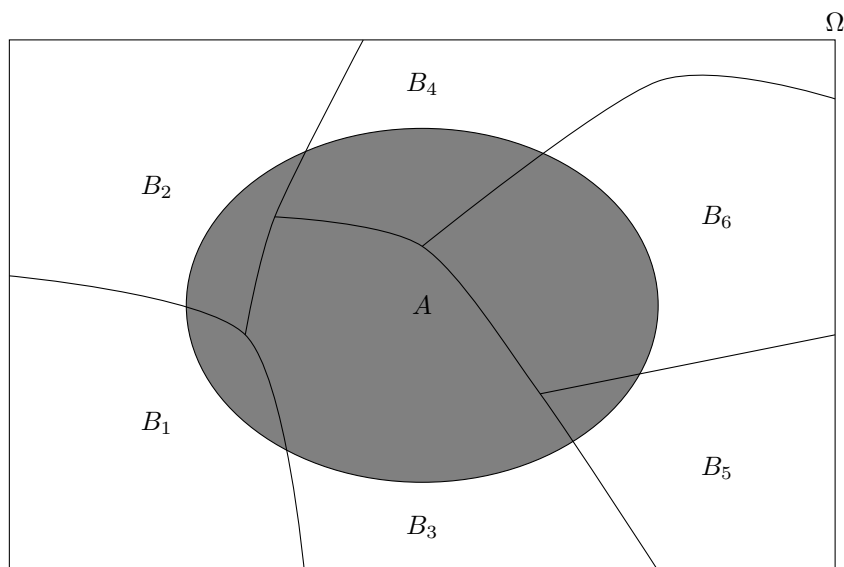


Abbildung 1.5: Illustration einer Partition von Ω (B_1, \dots, B_6).

Manchmal will man die bedingten Wahrscheinlichkeiten auch “umkehren”. Sie haben z.B. ein technisches Verfahren entwickelt, um Haarrisse in Oberflächen zu detektieren. Wir betrachten folgende Ereignisse

A = “Technisches Verfahren indiziert, dass Risse da sind”

B_1 = “Oberfläche weist in der Tat Haarrisse auf”

$B_2 = B_1^c$ = “Oberfläche weist in der Tat *keine* Haarrisse auf”

Das Verfahren arbeitet nicht ganz fehlerfrei, die Fehlerquote ist aber (auf den ersten Blick) relativ tief (fiktive Zahlen):

$$\mathbb{P}(A \mid B_1) = 0.99$$

$$\mathbb{P}(A \mid B_2) = 0.03$$

Zudem nehmen wir an, dass gilt

$$\mathbb{P}(B_1) = 0.001.$$

Wenn die Oberfläche also tatsächlich Risse hat, so weisen wir das mit Wahrscheinlichkeit 0.99 nach. Wenn keine Risse da sind, dann schlagen wir “nur” mit Wahrscheinlichkeit 0.03 fälschlicherweise Alarm. Zudem gehen wir davon aus, dass mit Wahrscheinlichkeit 0.001 überhaupt Risse vorhanden sind (a-priori, ohne einen Test gemacht zu haben).

Die Frage ist nun: Gegeben, dass das technische Verfahren Haarrisse nachweist, was ist die Wahrscheinlichkeit, dass in Tat und Wahrheit wirklich Risse da sind? Oder ausgedrückt in bedingten Wahrscheinlichkeiten: Wie gross ist $\mathbb{P}(B_1 \mid A)$? Dies können wir mit dem Satz von Bayes beantworten.

Satz von Bayes

Für zwei Ereignisse A und B mit $\mathbb{P}(A), \mathbb{P}(B) > 0$ gilt

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

In der Situation des Satzes der totalen Wahrscheinlichkeit haben wir

$$\begin{aligned} \mathbb{P}(B_i | A) &= \frac{\mathbb{P}(A | B_i) \mathbb{P}(B_i)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A | B_i) \mathbb{P}(B_i)}{\sum_{l=1}^k \mathbb{P}(A | B_l) \mathbb{P}(B_l)}. \end{aligned}$$

Oft ist das Resultat einer solchen Berechnung stark verschieden von dem, was man intuitiv erwartet.

Beispiel. *In obigem Beispiel haben wir also*

$$\begin{aligned} \mathbb{P}(B_1 | A) &= \frac{\mathbb{P}(A | B_1) \mathbb{P}(B_1)}{\mathbb{P}(A | B_1) \mathbb{P}(B_1) + \mathbb{P}(A | B_2) \mathbb{P}(B_2)} \\ &= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.03 \cdot 0.999} = 0.032. \end{aligned}$$

Obwohl die Spezifikationen von unserem Test auf den ersten Blick gut ausgesehen haben, sagt hier ein positives Testresultat nicht sehr viel aus! Oder haben wir uns nur verrechnet oder etwas falsch angewendet? Schauen wir uns die Geschichte einmal mit konkreten Anzahlen an. Wir nehmen an, dass wir $n = 100'000$ Untersuchungen machen. Davon sind im Schnitt 99'900 in der Tat in Ordnung. In der folgenden Tabelle sehen wir, wie sich die Fälle im Schnitt gemäss den Fehlerquoten des Tests aufteilen.

	B_1	B_2	Summe
A	99	2'997	3'096
A^c	1	96'903	96'904
Summe	100	99'900	100'000

Wir interessieren uns nun für die Subgruppe, die ein positives Testresultat haben (Zeile A). Es sind dies 3'096 Fälle, 99 davon sind wirklich defekt. Also ist der Anteil $99/3'096 = 0.032$. Für die Kommunikation an fachfremde Personen eignet sich eine solche Tabelle in der Regel gut. Die Anzahlen kann jeder selber rasch nachrechnen bzw. überprüfen. \triangleleft

1.5 Review / Lernziele



- ☐ Sie kennen die Grundbegriffe der Wahrscheinlichkeitsrechnung sowie die Operationen der Mengenlehre und deren Bedeutung. Sie wissen, dass man Wahrscheinlichkeiten auf verschiedene Arten interpretieren kann.
- ☐ Sie kennen die Axiome der Wahrscheinlichkeitsrechnung und die resultierenden Rechenregeln.
- ☐ Sie können in diskreten Wahrscheinlichkeitsmodellen entsprechende Berechnungen durchführen und kennen das Laplace-Modell als Spezialfall.
- ☐ Sie wissen, was unabhängige Ereignisse sind und wie man mit ihnen rechnen kann.
- ☐ Sie verstehen das Konzept und die Rechenregeln der bedingten Wahrscheinlichkeiten. Sie können Unabhängigkeit auch mit bedingten Wahrscheinlichkeiten ausdrücken.
- ☐ Sie können mit Hilfe des Satzes der totalen Wahrscheinlichkeit sowie des Satzes von Bayes entsprechende Aufgaben lösen.

2 Wahrscheinlichkeitsverteilungen

Bis jetzt haben wir ganz allgemein Zufallsexperimente angeschaut. Deren Ausgang waren entweder Zahlen (Druckfestigkeit, Augenzahl Würfel etc.) oder “abstraktere” Dinge wie eine Kombination von K und Z beim Beispiel mit dem zweimaligen Wurf mit einer Münze.

In der Praxis sind Messungen, z.B. von einem physikalischen Versuch (ein Zufallsexperiment), in der Regel Zahlen. Man führt für diesen Spezialfall den Begriff der Zufallsvariable ein. Oft weist man den verschiedenen “abstrakten” Ausgängen eines Zufallsexperiments einfach auch Zahlen zu, z.B. entsprechende Gewinne bei einem Glücksspiel. In beiden Fällen haben wir schlussendlich zufällige *Zahlen* als Ausgänge.

2.1 Der Begriff der Zufallsvariable

Eine **Zufallsvariable** X ist ein Zufallsexperiment mit möglichen Werten in \mathbb{R} , bzw. in einer Teilmenge von \mathbb{R} , z.B. $\mathbb{N}_0 = \{0, 1, \dots\}$. Wir haben also die gleiche Situation wie vorher, d.h. $\Omega = \mathbb{R}$, bzw. $\Omega = \mathbb{N}_0$ etc.; jetzt aber angereichert mit einem neuen Begriff und neuer Notation. Der Wert einer Zufallsvariablen ist insbesondere im Voraus also *nicht* bekannt. Oft schreiben wir für den Wertebereich auch W statt Ω .

Wir verwenden *Grossbuchstaben* X für die Zufallsvariable und *Kleinbuchstaben* x für die realisierten Werte. Wenn wir $\{X = x\}$ schreiben ist dies also das Ereignis, dass die *Zufallsvariable* X den *Wert* x annimmt, d.h. dass das Elementarereignis x eintritt. Unter dem Grossbuchstaben können Sie sich einfach den “Wortswall” vorstellen (z.B. “Messung der Druckfestigkeit”). Beim Kleinbuchstaben setzt man einen konkreten Wert ein, z.B. $x = 30$.

Wenn X die Druckfestigkeit ist, dann bezeichnen wir mit $\{X \leq 30\}$ das Ereignis “Druckfestigkeit ist kleiner gleich 30”. Dazu äquivalent schreiben wir manchmal auch $\{X \in (-\infty, 30]\}$.

Der Begriff der **Unabhängigkeit** ist analog wie früher definiert: Zwei Zufallsvariablen X und Y heissen unabhängig, falls für alle Mengen $A, B \subset \mathbb{R}$ gilt, dass

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B),$$

wobei wir hier mit $\{X \in A, Y \in B\}$ das Ereignis $\{X \in A\} \cap \{Y \in B\}$ meinen. Die Erweiterung auf den Fall mit mehr als zwei Zufallsvariablen ist entsprechend wie früher.

Bemerkung:

Wie in der Einleitung bereits angedeutet, können wir eine Zufallsvariable mathematisch auch interpretieren als eine Funktion $X : \Omega \rightarrow \mathbb{R}$, die jedem zufälligen $\omega \in \Omega$ eine reelle Zahl $X(\omega) \in \mathbb{R}$ zuweist. Ein einfaches Beispiel ist die Augensumme von zwei Würfeln. Die Funktion ist natürlich nicht zufällig, sehr wohl aber ihr Argument und der resultierende Funktionswert! Für unsere Betrachtungen reicht aber die “weniger mathematische” Definition oben. Wir vergessen dann sozusagen das ursprüngliche Ω .

2.1.1 Wahrscheinlichkeitsverteilungen

Von Interesse ist die Frage, mit welchen Wahrscheinlichkeiten eine Zufallsvariable in welchen Bereichen liegt. Man spricht von der sogenannten **Wahrscheinlichkeitsverteilung** bzw. kurz von der **Verteilung** von X .

Was ist z.B. die Wahrscheinlichkeit, dass die Druckfestigkeit kleiner gleich 30 MPa ist oder im Intervall $[25, 30]$ MPa liegt? Oder was ist die Wahrscheinlichkeit, dass wir in einer Lieferung von 100 Bauteilen weniger als 5 defekte Teile vorfinden?

Wenn wir die Verteilung einer Zufallsvariablen X kennen, können wir auf jede beliebige solche Frage die entsprechende Antwort geben. Wir unterscheiden dabei zwischen diskreten und stetigen Verteilungen (bzw. Zufallsvariablen).

Wie wir später sehen werden, gibt es für die Modellierung von gewissen unsicheren Phänomenen bestimmte Verteilungen, die sich speziell gut dafür eignen. Wenn man also einmal die wichtigsten Verteilungen kennt, so kann man diese Sammlung als "Toolbox" brauchen. Man muss für die Modellierung von einem Phänomen dann einfach diejenige heraus suchen, die am besten passt.

2.2 Diskrete Verteilungen

Eine Zufallsvariable X (bzw. deren Verteilung) heisst **diskret**, falls die Menge W der möglichen Werte von X (der Wertebereich) endlich oder abzählbar ist. Mögliche Wertebereiche W sind zum Beispiel $W = \{0, 1, 2, \dots, 100\}$, $W = \mathbb{N}_0 = \{0, 1, 2, \dots\}$ oder ganz allgemein $W = \{x_1, x_2, \dots\}$.

Die Augenzahl bei einem Würfel ist ein Beispiel für eine diskrete Zufallsvariable mit Wertebereich $W = \{1, 2, \dots, 6\}$. Die Anzahl defekter Teile in einer Lieferung von 100 Bauteilen ist eine diskrete Zufallsvariable mit Wertebereich $\{0, 1, \dots, 100\}$.

Wie früher können wir hier eine *Liste* von Wahrscheinlichkeiten erstellen. Damit ist die Verteilung einer diskreten Zufallsvariablen festgelegt, da wir dann alle möglichen Wahrscheinlichkeiten berechnen können.

Die Liste ist gegeben durch die sogenannte **Wahrscheinlichkeitsfunktion** $p(x_k)$, wobei

$$p(x_k) = \mathbb{P}(X = x_k), \quad k \geq 1.$$

Dies ist genau gleich wie früher. Ein Elementarereignis ist hier einfach ein Element x_k des Wertebereichs W . Die Summe aller Wahrscheinlichkeiten muss insbesondere wieder 1 ergeben, d.h.

$$\sum_{k \geq 1} p(x_k) = 1.$$

Zudem gilt für ein Ereignis $A \subset W$

$$\mathbb{P}(X \in A) = \sum_{k: x_k \in A} p(x_k).$$

Auch das ist nichts Neues, sondern einfach die alte Erkenntnis in leicht anderer Notation verpackt.

Die Verteilung einer Zufallsvariablen X kann man auch mit der **kumulativen Verteilungsfunktion** F charakterisieren. Diese ist definiert als

$$F(x) = \mathbb{P}(X \leq x)$$

für $x \in \mathbb{R}$. Die kumulative Verteilungsfunktion enthält alle Information der Verteilung von X und ist gleichzeitig einfach darstellbar.

Beispiel. Bei einem fairen Würfel haben wir

k	1	2	3	4	5	6
x_k	1	2	3	4	5	6
$p(x_k)$	1/6	1/6	1/6	1/6	1/6	1/6

Es ist z.B.

$$\begin{aligned} F(3) &= \mathbb{P}(X \leq 3) = \mathbb{P}(\{X = 1\} \cup \{X = 2\} \cup \{X = 3\}) \\ &\stackrel{(A3)}{=} \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}. \end{aligned}$$

Wir können die Verteilungsfunktion an beliebigen Stellen evaluieren, z.B.

$$\begin{aligned} F(3.5) &= \mathbb{P}(X \leq 3.5) = \mathbb{P}(\{X \leq 3\} \cup \{3 < X \leq 3.5\}) \\ &\stackrel{(A3)}{=} \mathbb{P}(X \leq 3) + \mathbb{P}(3 < X \leq 3.5) \\ &= \frac{3}{6} + 0 = \frac{3}{6}. \end{aligned}$$

Die ganze Funktion ist in Abbildung 2.1 dargestellt. ◁

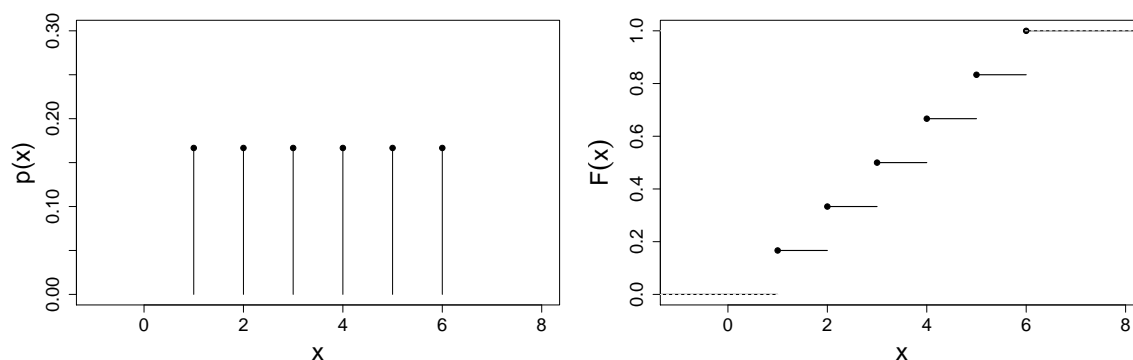


Abbildung 2.1: Wahrscheinlichkeitsfunktion (links) und kumulative Verteilungsfunktion (rechts) beim Beispiel mit dem Würfel.

Die kumulative Verteilungsfunktion ist also bei einer diskreten Zufallsvariable eine Treppenfunktion mit Sprüngen an den Stellen $x_k \in W$ mit Sprunghöhen $p(x_k)$, also insbesondere *nicht* stetig.

Rechenregeln und Eigenschaften

Es gilt (egal ob X diskret ist oder nicht)

$$\begin{aligned} \mathbb{P}(a < X \leq b) &= \mathbb{P}(X \in (a, b]) \\ &\stackrel{(1.5)}{=} \mathbb{P}(X \in (-\infty, b]) - \mathbb{P}(X \in (-\infty, a]) \\ &= F(b) - F(a) \\ \mathbb{P}(X > x) &\stackrel{(1.1)}{=} 1 - \mathbb{P}(X \leq x) = 1 - F(x) \end{aligned}$$

Die kumulative Verteilungsfunktion F erfüllt zudem immer:

- F ist monoton steigend
- $\lim_{x \rightarrow -\infty} F(x) = 0$ und $\lim_{x \rightarrow \infty} F(x) = 1$.
- F ist rechts-stetig, d.h. $\lim_{x \searrow a} F(x) = F(a)$.

2.2.1 Kennzahlen

Wir haben gesehen, dass die Verteilung einer diskreten Zufallsvariable durch eine (unendlich) lange Liste von Wahrscheinlichkeiten gegeben ist. Es stellt sich oft die Frage, ob man diese Liste durch ein paar wenige **Kennzahlen** zusammenfassen kann, um die Verteilung (grob) zu charakterisieren.

Es zeigt sich, dass hierzu Kennzahlen für die **mittlere Lage** (\rightsquigarrow Erwartungswert) und für die **Streuung** (\rightsquigarrow Varianz, Standardabweichung) geeignet sind.

Der **Erwartungswert** μ_X oder $\mathbb{E}[X]$ einer diskreten Zufallsvariable X ist definiert als

$$\mu_X = \mathbb{E}[X] = \sum_{k \geq 1} x_k p(x_k).$$

Merkregel: Man summiert über “was passiert” (x_k) \times “mit welcher Wahrscheinlichkeit passiert es” ($p(x_k)$).

Der Erwartungswert ist ein **Mass für die mittlere Lage der Verteilung**, ein sogenannter **Lageparameter**. Er wird interpretiert als das “Mittel der Werte von X bei (unendlich) vielen Wiederholungen”. D.h. er ist eine Idealisierung des arithmetischen Mittels der Werte einer Zufallsvariablen bei unendlich vielen Wiederholungen. Also: $\mathbb{E}[X]$ ist eine Kennzahl im wahrscheinlichkeitstheoretischen Modell.

Physikalisch gesehen ist der Erwartungswert nichts anderes als der Schwerpunkt, wenn wir auf dem Zahlenstrahl an den Positionen x_k die entsprechenden Massen $p(x_k)$ platzieren (der Zahlenstrahl selber hat hier keine Masse).

Beispiel. Bei einem fairen Würfel haben wir

k	1	2	3	4	5	6
x_k	1	2	3	4	5	6
$p(x_k)$	1/6	1/6	1/6	1/6	1/6	1/6

Der Erwartungswert ist demnach gegeben durch

$$\mathbb{E}[X] = \sum_{k=1}^6 k \cdot \frac{1}{6} = 3.5,$$

siehe auch der Schwerpunkt in Abbildung 2.1. Wenn wir also oft Würfeln und mitteln, dann werden wir ungefähr 3.5 erhalten. An diesem Beispiel sehen wir auch, dass der Erwartungswert gar nicht einmal im Wertebereich der Zufallsvariable liegen muss.

Was passiert, wenn wir einen “gezinkten” Würfel, der eine erhöhte Wahrscheinlichkeit für die 6 hat, verwenden?

k	1	2	3	4	5	6
x_k	1	2	3	4	5	6
$p(x_k)$	1/7	1/7	1/7	1/7	1/7	2/7

Es ist dann

$$\mathbb{E}[X] = \sum_{k=1}^5 k \cdot \frac{1}{7} + 6 \cdot \frac{2}{7} = 3.86.$$

Der Erwartungswert wird also grösser; der Schwerpunkt hat sich etwas nach rechts verschoben. \triangleleft

Manchmal betrachtet man statt der Zufallsvariablen X eine Transformation $g(X)$, wobei $g: \mathbb{R} \rightarrow \mathbb{R}$ eine Funktion ist. Für den Erwartungswert einer transformierten diskreten Zufallsvariable $Y = g(X)$ gilt

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_{k \geq 1} g(x_k) p(x_k). \quad (2.1)$$

Wieder wie vorher summiert man über “was passiert” ($g(x_k)$) \times “mit welcher Wahrscheinlichkeit passiert es” ($p(x_k)$).

Die **Varianz** $\text{Var}(X)$ oder σ_X^2 einer diskreten Zufallsvariable X ist definiert als

$$\text{Var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] \stackrel{(2.1)}{=} \sum_{k \geq 1} (x_k - \mu_X)^2 p(x_k).$$

Die Varianz ist also die mittlere quadratische Abweichung der Zufallsvariablen von ihrem Erwartungswert und somit ein **Mass für die Streuung** um die mittlere Lage, ein sogenannter **Streuungsparameter**.

Physikalisch gesehen ist die Varianz das Trägheitsmoment, wenn wir obigen Körper um die Achse drehen, die senkrecht zum Zahlenstrahl steht und durch den Schwerpunkt (Erwartungswert) geht. Je mehr Masse (Wahrscheinlichkeit) also weit weg vom Schwerpunkt (Erwartungswert) liegt, desto grösser wird die Varianz.

Für viele Berechnungen werden wir die **Standardabweichung** σ_X brauchen. Diese ist definiert als die Wurzel aus der Varianz, d.h.

$$\sigma_X = \sqrt{\text{Var}(X)}.$$

Wie der Erwartungswert hat die Standardabweichung die gleichen Einheiten wie die Zufallsvariable X (z.B. m). Dies im Gegensatz zur Varianz, die die quadrierten Einheiten hat (z.B. m^2).

Die folgenden Rechenregeln werden immer wieder gebraucht:

Rechenregeln für Erwartungswert und Varianz

Es gilt (egal ob X diskret ist oder nicht)

$$\begin{aligned}\mathbb{E}[a + bX] &= a + b \cdot \mathbb{E}[X], \quad a, b \in \mathbb{R} \\ \text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ \text{Var}(a + bX) &= b^2 \text{Var}(X), \quad a, b \in \mathbb{R} \\ \text{Var}(a) &= 0, \quad a \in \mathbb{R}.\end{aligned}$$

Falls man noch eine zweite Zufallsvariable Y hat, so gilt immer

$$\mathbb{E}[a + bX + cY] = a + b \cdot \mathbb{E}[X] + c \cdot \mathbb{E}[Y], \quad a, b, c \in \mathbb{R}.$$

Für die Varianz sieht es nicht mehr so einfach aus, mehr davon später (siehe Kapitel 4.3).

Wir wollen nun die wichtigsten diskreten Verteilungen betrachten, die wir immer wieder antreffen werden.

2.2.2 Bernoulli-Verteilung [Bernoulli(p)]

Die **Bernoulli-Verteilung** mit Parameter $p \in (0, 1)$ ist die “einfachste” diskrete Verteilung. Hier kann X nur die Werte 0 oder 1 annehmen, d.h.

$$X = \begin{cases} 1 & \text{Wahrscheinlichkeit } p \\ 0 & \text{Wahrscheinlichkeit } 1 - p \end{cases}$$

Es gilt (nachrechnen!)

$$\begin{aligned}\mathbb{E}[X] &= p \\ \text{Var}(X) &= p \cdot (1 - p).\end{aligned}$$

Wir schreiben auch $X \sim \text{Bernoulli}(p)$, wobei das Symbol “ \sim ” (Tilde) übersetzt wird als “ist verteilt wie”.

2.2.3 Binomialverteilung [Bin(n, p)]

Die **Binomialverteilung** mit den Parametern $n \in \mathbb{N}$ und $p \in (0, 1)$, ist die Verteilung der Anzahl “Erfolge” bei n (unabhängigen) Wiederholungen eines “Experiments” mit “Erfolgswahrscheinlichkeit” p . Hier ist also $W = \{0, 1, \dots, n\}$. Die Binomialverteilung kann also insbesondere aufgefasst werden als eine Summe von n unabhängigen Bernoulliverteilungen mit Parameter p .

Die Begriffe Erfolg und Experiment können hier vieles bedeuten. Die Anzahl defekter Bauteile bei einer Lieferung von $n = 10$ Bauteilen folgt einer Binomialverteilung mit Parametern $n = 10$ und p , wobei p die Wahrscheinlichkeit ist, dass ein einzelnes Bauteil defekt ist, z.B. $p = 0.05$. Hier ist ein Experiment die Überprüfung eines Bauteils und Erfolg bedeutet, dass das Bauteil defekt ist.

Man kann zeigen, dass gilt

$$\begin{aligned} p(x) &= \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in W \\ \mathbb{E}[X] &= np \\ \text{Var}(X) &= n \cdot p \cdot (1-p), \end{aligned}$$

wobei $\binom{n}{x}$ (sprich: “ n tief x ”) der sogenannte Binomialkoeffizient ist, d.h.

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

Eine Herleitung für die Wahrscheinlichkeitsfunktion findet man in Kapitel C.1. In Abbildung 2.2 sind einige Fälle mit verschiedenen Parametern dargestellt. Für grosses n hat man schon ein ziemlich “glockenförmiges” Bild, mehr dazu später (siehe Kapitel 5.3).

Den Parameter n kennt man in der Regel aus dem Kontext. Die Erfolgswahrscheinlichkeit p nehmen wir bis auf Weiteres als gegeben an. Später werden wir dann sehen, wie wir p aus Daten schätzen können.

Wenn wir erkannt haben, dass etwas binomial-verteilt ist, dann ist das Rechnen damit nicht kompliziert. Was ist z.B. die Wahrscheinlichkeit, dass von 10 Bauteilen genau 3 mangelhaft sind? Diese Wahrscheinlichkeit ist gegeben durch

$$\mathbb{P}(X = 3) = p(3) = \binom{10}{3} 0.05^3 \cdot 0.95^7 = \frac{10!}{3! \cdot 7!} \cdot 0.05^3 \cdot 0.95^7 = 0.0105.$$

Oder was ist die Wahrscheinlichkeit, dass von 10 Bauteilen mindestens eines defekt ist? Fast immer wenn wir das Wort “mindestens” hören, lohnt es sich, mit dem komplementären Ereignis zu arbeiten. Statt

$$\mathbb{P}(X \geq 1) \stackrel{(A3)}{=} \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \dots + \mathbb{P}(X = 10)$$

mühsam zu bestimmen, erhalten wir direkt mit dem komplementären Ereignis

$$\{X = 0\} = \{X \geq 1\}^c$$

dass

$$\mathbb{P}(X \geq 1) \stackrel{(1.1)}{=} 1 - \mathbb{P}(X = 0) = 1 - p(0) = 1 - 0.95^{10} = 0.401.$$

Also: Wenn wir einmal erkannt haben, dass etwas mit einer Binomialverteilung modelliert werden kann, dann können wir damit bequem alle Wahrscheinlichkeiten bestimmen. Die mühsame Abzählerei müssen wir nicht machen, alle Information steht in der Formel für $p(x)$.

2.2.4 Geometrische Verteilung [Geom(p)]

Die **geometrische Verteilung** mit Parameter $p \in (0, 1)$ tritt auf, wenn wir die **Anzahl Wiederholungen** von unabhängigen Bernoulli(p) Experimenten **bis zum ersten Erfolg** betrachten. Man

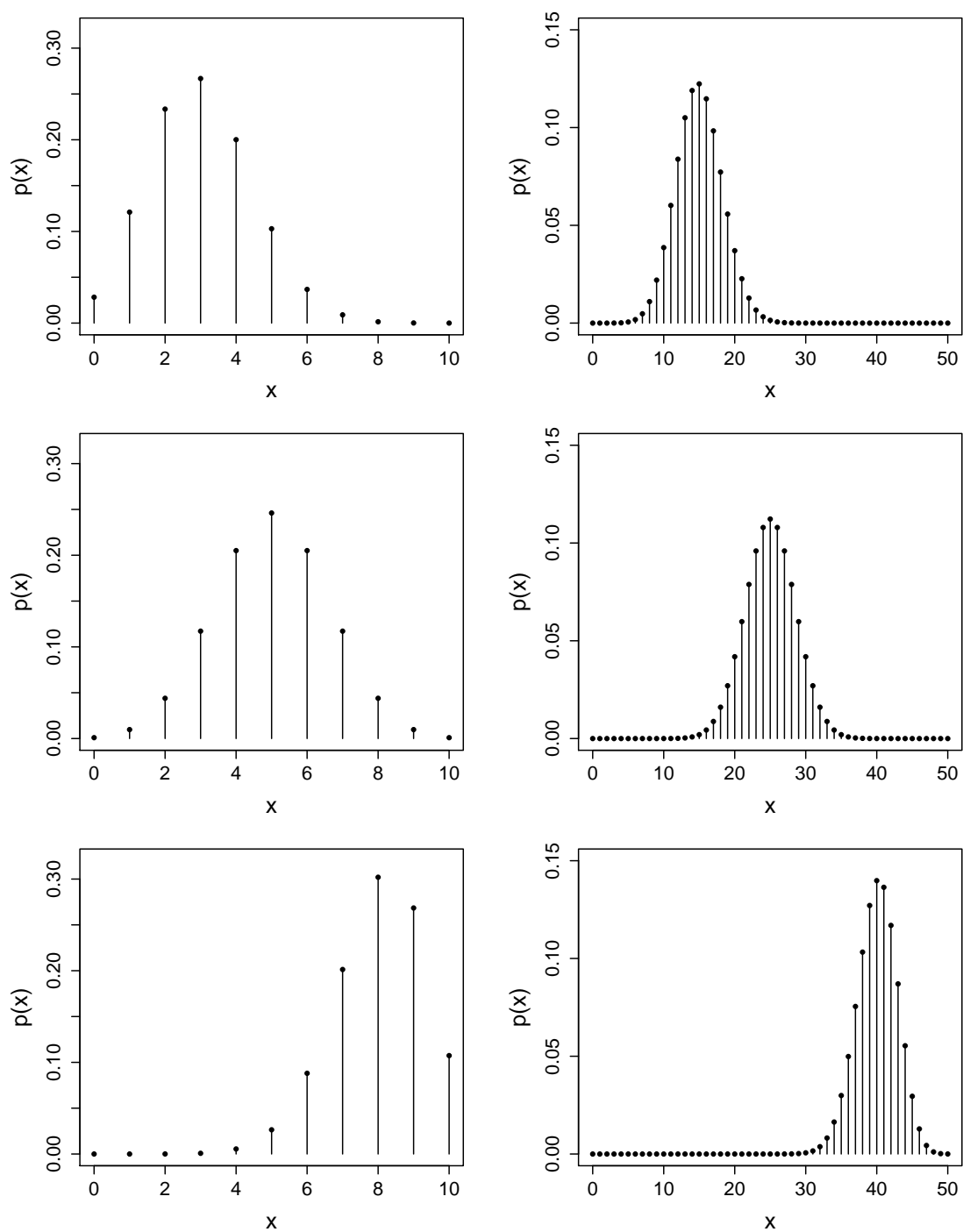


Abbildung 2.2: Wahrscheinlichkeitsfunktion der Binomialverteilung für $n = 10$ (links) und $n = 50$ (rechts) für jeweils $p = 0.3, 0.5, 0.8$ (oben nach unten).

wirft z.B. eine Münze so lange, bis das erste Mal Kopf fällt und notiert sich die Anzahl benötigter Würfe.

Hier ist $W = \{1, 2, \dots\}$ (unbeschränkt!) und

$$p(x) = p \cdot (1 - p)^{x-1}$$

$$\mathbb{E}[X] = \frac{1}{p}$$

$$\text{Var}(X) = \frac{1-p}{p^2}.$$

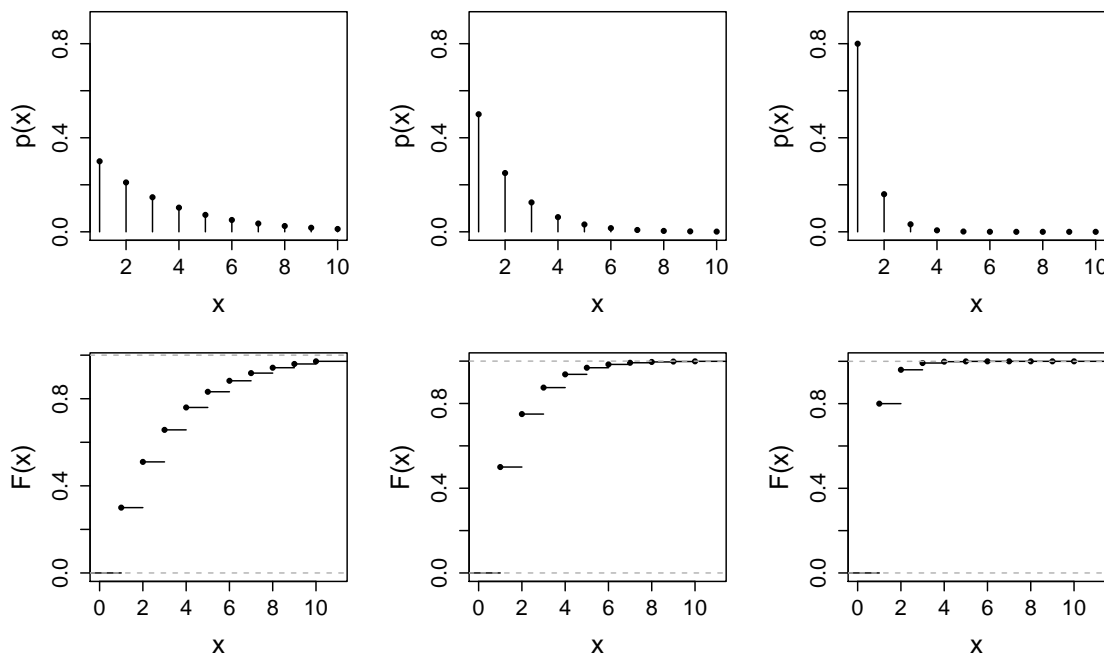


Abbildung 2.3: Wahrscheinlichkeitsfunktion (oben) und kumulative Verteilungsfunktion (unten) der geometrischen Verteilung für $p = 0.3, 0.5, 0.8$ (links nach rechts), jeweils abgeschnitten bei $x = 10$.

Wenn ein einzelner Versuch mit Wahrscheinlichkeit $p = 1/10$ erfolgreich ist, dann brauchen wir im Schnitt $\mathbb{E}[X] = 10$ Versuche, bis wir den ersten Erfolg sehen. Der Erwartungswert entspricht hier der **mittleren Wartezeit bis zum ersten Erfolg**, was auch als **Wiederkehrperiode** bezeichnet wird.

Die Verteilungsfunktion wollen wir hier einmal konkret aufschreiben. Es ist

$$F(x) = \sum_{i=1}^x p \cdot (1-p)^{i-1} \stackrel{(\text{geom. Reihe})}{=} 1 - (1-p)^x$$

für $x \in W$. Dazwischen ist F konstant, siehe auch Abbildung 2.3.

Beispiel. Man kann sich z.B. die Frage stellen, wie oft man einen Versuch mindestens durchführen muss, damit man eine 50% Chance hat, in dieser Versuchsreihe (mindestens) einmal Erfolg zu haben. Die gesuchte Anzahl Versuche wollen wir n nennen ($n \in W$). Übersetzt heisst dies nichts anderes, als dass das erste Auftreten des Erfolgs (bezeichnet mit X) mit Wahrscheinlichkeit mindestens 50% kleiner gleich n sein muss, d.h. dass gilt $\mathbb{P}(X \leq n) \geq 0.5$. Dies wiederum heisst nichts anderes, als dass wir das kleinste n suchen, so dass $F(n) \geq 0.5$ gilt, oder eingesetzt

$$1 - (1-p)^n \geq 0.5,$$

für n minimal. Aufgelöst erhält man

$$n \geq \frac{\log(0.5)}{\log(1-p)},$$

wobei wir mit \log den natürlichen Logarithmus bezeichnen. Für kleine p gilt $\log(1-p) \approx -p$. Dies führt zu approximativen Lösung

$$n \geq \frac{0.7}{p}.$$

Wir betrachten nun ein Erdbeben mit einer solchen Stärke, dass die Eintrittswahrscheinlichkeit pro Jahr $p = 1/1000$ ist. Ferner nehmen wir an, dass pro Jahr nur ein Beben vorkommen kann, und dass die Ereignisse in verschiedenen Jahren unabhängig sind. Im Schnitt warten wir also 1000 Jahre bis zum ersten Erdbeben.

Mit obiger Formel erhalten wir

$$n \geq \frac{0.7}{p} = 700.$$

Wenn man also eine 700-Jahr-Periode betrachtet, so hat man eine 50% Chance, dass (mindestens) ein Erdbeben eintritt. Insbesondere ist die Wahrscheinlichkeit in einer 500-Jahr-Periode kleiner als 50%! Wenn man nur die Hälfte der Wiederkehrperiode betrachtet, bedeutet dies also nicht, dass man eine Wahrscheinlichkeit von 50% hat, dass das Ereignis eintritt. \triangleleft

2.2.5 Poissonverteilung [Pois (λ)]

Bei der Binomialverteilung ging es um die Anzahl Erfolge in n Experimenten. Der Wertebereich war insbesondere beschränkt (nämlich durch n). Was ist, wenn man allgemein (potentiell unbeschränkte) Anzahlen betrachtet? Es zeigt sich, dass sich hierzu die sogenannte Poissonverteilung gut eignet.

Die **Poissonverteilung** mit Parameter $\lambda > 0$ ist gegeben durch

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x \in W$$

$$\mathbb{E}[X] = \lambda$$

$$\text{Var}(X) = \lambda.$$

Hier ist $W = \{0, 1, \dots\}$ (unbeschränkt). Die Poissonverteilung ist sozusagen *die* Standardverteilung für unbeschränkte Zählraten.

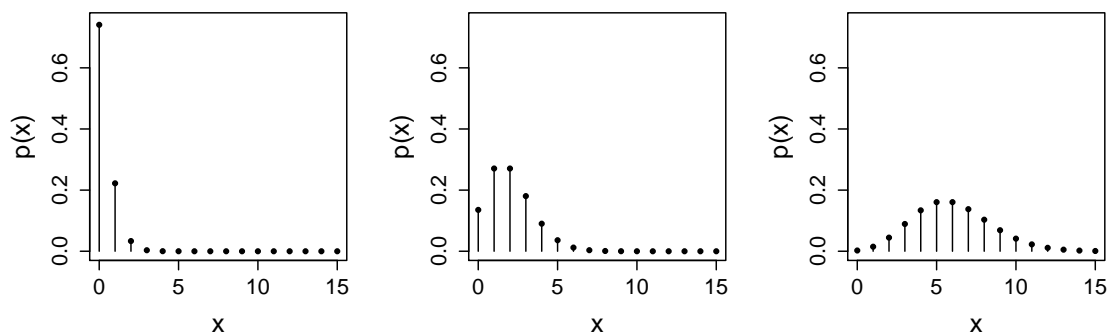


Abbildung 2.4: Wahrscheinlichkeitsfunktion der Poissonverteilung für $\lambda = 0.3, 2, 6$ (links nach rechts), jeweils abgeschnitten bei $x = 15$.

Beispiel. In einem Callcenter erwarten wir im Schnitt pro Minute 5 Anrufe. Wir modellieren die Anzahl Anrufe pro Minute (X) mit einer Poissonverteilung mit Parameter $\lambda = 5$, d.h. $X \sim \text{Pois}(\lambda)$,

$\lambda = 5$, denn so stimmt gerade der Erwartungswert. Damit können wir nun “alle” Wahrscheinlichkeiten berechnen, z.B. die Wahrscheinlichkeit, dass in einer Minute niemand anruft:

$$\mathbb{P}(X = 0) = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-5} = 0.00674. \quad \triangleleft$$

Poissonapproximation der Binomialverteilung

Man kann zeigen, dass die Poissonverteilung eine Approximation der Binomialverteilung ist für grosses n und kleines p mit $np = \lambda$. D.h. falls $X \sim \text{Bin}(n, p)$, dann gilt in diesen Situationen (n gross, p klein)

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \approx e^{-\lambda} \frac{\lambda^x}{x!}$$

für $\lambda = np$. Dies ist insbesondere nützlich, da die Berechnung der Binomialkoeffizienten für grosse n aufwendig wird. Damit kann man aber auch die Poissonverteilung interpretieren: Wir zählen die Anzahl seltener Ereignisse (Erfolge) bei vielen unabhängigen Versuchen. Betrachten wir z.B. nochmals die Anzahl Anrufe in einem Callcenter: Viele Leute können potentiell anrufen, aber die Wahrscheinlichkeit für eine einzelne Person ist sehr klein. Hier ist also n die Anzahl Personen (potentielle Anrufer) und p die Wahrscheinlichkeit, dass eine Person anruft. Also macht eine Modellierung mit einer Poissonverteilung so betrachtet durchaus Sinn.

Beispiel. Eine Fluggesellschaft überbucht einen Flug mit 200 Plätzen um 4 Plätze. Wie gross ist die Wahrscheinlichkeit, dass genügend Sitzplätze vorhanden sind, wenn ein einzelner Passagier unabhängig von den anderen mit 5% Wahrscheinlichkeit nicht erscheint?

Wir haben Total 204 verkaufte Tickets. Jedes Ticket wird mit Wahrscheinlichkeit 5% nicht “eingelöst” (d.h. der Passagier erscheint nicht). Die Anzahl Passagiere X , die nicht erscheinen, wäre unter obigen idealisierten Annahmen $\text{Bin}(204, 0.05)$ -verteilt. Diese Verteilung approximieren wir mit einer Poissonverteilung, d.h. wir verwenden

$$X \sim \text{Pois}(\lambda), \lambda = 204 \cdot 0.05 = 10.2.$$

Damit der Flug nicht überbucht ist, muss gelten $X \geq 4$, die entsprechende Wahrscheinlichkeit ist

$$\mathbb{P}(X \geq 4) = 1 - \mathbb{P}(X \leq 3) = 1 - \sum_{k=0}^3 e^{-\lambda} \frac{\lambda^k}{k!} = 0.991.$$

Wenn man mit der Binomialverteilung rechnen würde, erhielte man $\mathbb{P}(X \geq 4) = 0.992$. \triangleleft

Summen von unabhängigen Poissonverteilungen

Wenn $X \sim \text{Pois}(\lambda_1)$ und $Y \sim \text{Pois}(\lambda_2)$ mit X und Y unabhängig, dann gilt

$$X + Y \sim \text{Pois}(\lambda_1 + \lambda_2).$$

Wenn wir also unabhängige Poissonverteilungen addieren, so haben wir immer noch eine Poissonverteilung. Die Parameter müssen sich dann zwangsläufig gerade addieren wegen den Rechenregeln für den Erwartungswert.

Wenn wir aber $\frac{1}{2}(X + Y)$ betrachten, so liegt *keine* Poissonverteilung vor mit Parameter $\frac{1}{2}(\lambda_1 + \lambda_2)$. Der Grund ist ganz einfach: Nur schon der Wertebereich stimmt nicht für eine Poissonverteilung! Der Erwartungswert ist aber $\frac{1}{2}(\lambda_1 + \lambda_2)$.

2.3 Stetige Verteilungen

Eine Zufallsvariable X (bzw. deren Verteilung) heisst **stetig**, falls die Menge der möglichen Werte W aus einem oder mehreren Intervallen besteht, z.B. $W = [0, 1]$ oder $W = \mathbb{R}$. Im Gegensatz zu früher haben wir hier keine “Liste” mehr von möglichen Werten. Dies führt dazu, dass wir neue Konzepte einführen müssen, vieles können wir aber von früher wiederverwenden.

Betrachten wir zuerst ein einfaches Beispiel. Wir nehmen an, dass wir eine Zufallsvariable X haben, die Werte im Intervall $[0, 1]$ annehmen kann und die keine Regionen “bevorzugt” (eine sogenannte Uniform- oder Gleichverteilung). D.h. es soll z.B. gelten $\mathbb{P}(0.2 \leq X \leq 0.4) = \mathbb{P}(0.6 \leq X \leq 0.8)$, da die Intervalle gleich breit sind. Natürlich gilt in diesem Fall $\mathbb{P}(0 \leq X \leq 1) = 1$. Die Wahrscheinlichkeit muss also gleich der Intervallbreite sein, d.h. es gilt

$$\mathbb{P}(x \leq X \leq x + h) = h.$$

Wenn wir jetzt h klein werden lassen ($h \rightarrow 0$), dann wird auch die Wahrscheinlichkeit immer kleiner, d.h. $\mathbb{P}(x \leq X \leq x + h) \rightarrow 0$. D.h. für einen *einzelnen* Punkt x ist die Wahrscheinlichkeit $\mathbb{P}(X = x) = 0$. Dies gilt allgemein für stetige Zufallsvariablen. Wir müssen daher den neuen Begriff der Wahrscheinlichkeitsdichte einführen.

2.3.1 Wahrscheinlichkeitsdichte

Die **Wahrscheinlichkeitsdichte** (oder oft kurz einfach nur **Dichte**) einer stetigen Verteilung ist definiert als

$$f(x) = \lim_{h \rightarrow 0} \frac{\mathbb{P}(x < X \leq x + h)}{h} = \lim_{h \rightarrow 0} \frac{F(x + h) - F(x)}{h} = F'(x).$$

Dabei sind wir stillschweigend davon ausgegangen, dass die Ableitung der kumulativen Verteilungsfunktion existiert.

Es gilt daher die folgende Interpretation

$$\mathbb{P}(x < X \leq x + h) \approx hf(x)$$

für kleines h . Wenn also in einer Region die Dichte gross ist, dann ist die Wahrscheinlichkeit, in diese Region zu fallen, erhöht verglichen mit anderen Regionen. Im einführenden Beispiel wäre die Dichte konstant.

Zwischen der Dichte f und der kumulativen Verteilungsfunktion F bestehen gemäss Definition ferner folgende Beziehungen:

$$f(x) = F'(x) \qquad F(x) = \int_{-\infty}^x f(u) \, du.$$

Hat man also eine Dichte, so erhält man durch integrieren die kumulative Verteilungsfunktion. Umgekehrt erhält man durch Ableiten der kumulativen Verteilungsfunktion immer die Dichte.

Insbesondere gilt

$$\mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) \, dx.$$

Um Wahrscheinlichkeiten zu erhalten, müssen wir also einfach die Dichte über das entsprechende Gebiet integrieren. Oder anders ausgedrückt: “Die Fläche unter der Dichte entspricht der Wahrscheinlichkeit”, siehe Abbildung 2.5. Früher hatten wir statt Integrale einfach Summen.

Damit eine Funktion f als Dichte verwendet werden kann, muss gelten $f(x) \geq 0$ für alle x , sowie

$$\int_{-\infty}^{\infty} f(x) \, dx = 1.$$

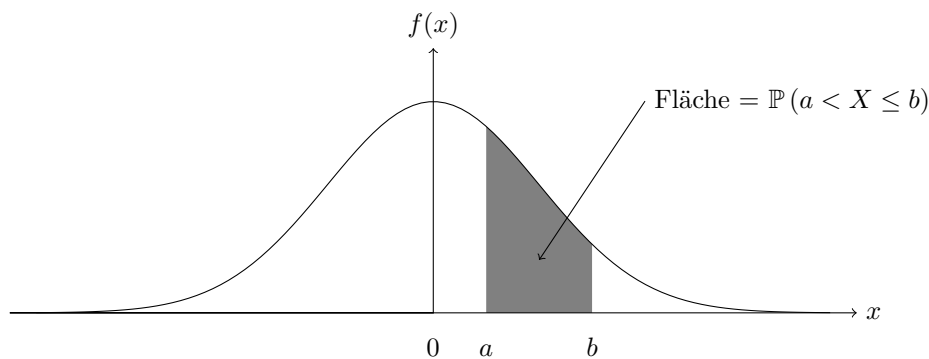


Abbildung 2.5: Illustration einer Dichte einer Zufallsvariablen und der Wahrscheinlichkeit, in das Intervall $(a, b]$ zu fallen (graue Fläche).

Dies folgt aus den ursprünglichen Axiomen. Man beachte insbesondere, dass es durchaus (kleine) Intervalle geben kann, in denen $f(x) > 1$ gilt, siehe z.B. Abbildung 2.10. Dies im Gegensatz zum diskreten Fall, wo jeweils *immer* $0 \leq p(x_k) \leq 1$ gilt.

Im stetigen Fall spielt es jeweils keine Rolle, ob wir Intervalle offen – wie (a, b) – oder geschlossen – wie $[a, b]$ – schreiben, da sich die Wahrscheinlichkeiten nicht ändern, weil die einzelnen Punkte a und b Wahrscheinlichkeit 0 haben. Achtung: Im diskreten Fall spielt dies sehr wohl eine Rolle.

2.3.2 Kennzahlen von stetigen Verteilungen

Erwartungswert und Varianz

Der Erwartungswert berechnet sich im stetigen Fall als

$$\mathbb{E}[X] = \mu_X = \int_{-\infty}^{\infty} x f(x) \, dx.$$

Für eine Transformation $g(X)$ gilt analog zu früher

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) \, dx.$$

Für die Varianz haben wir entsprechend

$$\text{Var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) \, dx.$$

Alle diese Formeln sind genau gleich wie früher: Man ersetzt die Summe durch das Integral und die Wahrscheinlichkeit $p(x)$ durch $f(x) \, dx$. Es gelten insbesondere die gleichen Rechenregeln wie im diskreten Fall. Auch die Interpretationen bleiben unverändert, sowohl die statistische wie auch die physikalische (Schwerpunkt, Trägheitsmoment).

Quantile

Das $(\alpha \times 100)\%$ -Quantil q_α für $\alpha \in (0, 1)$ ist definiert als der Wert, der mit Wahrscheinlichkeit $(\alpha \times 100)\%$ unterschritten wird, d.h. für q_α muss gelten

$$\alpha = \mathbb{P}(X \leq q_\alpha) = F(q_\alpha).$$

Es ist also

$$q_\alpha = F^{-1}(\alpha),$$

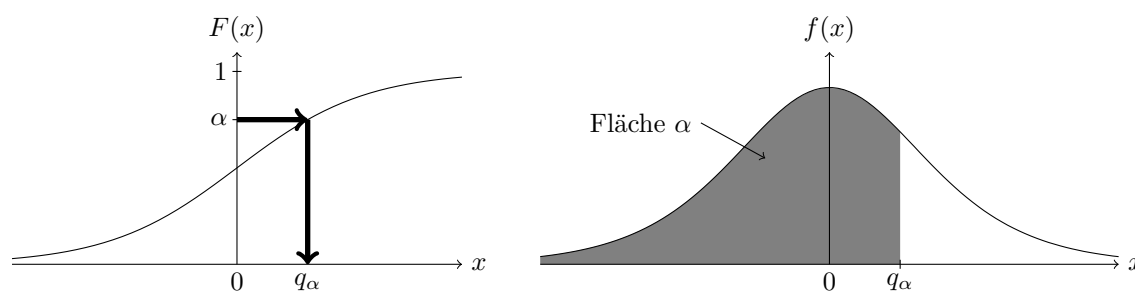


Abbildung 2.6: Illustration des Quantils q_α anhand der Verteilungsfunktion (links) und der Dichte (rechts) für $\alpha = 0.75$.

was auch in Abbildung 2.6 dargestellt ist.

Der **Median** ist das 50%-Quantil. Er teilt die Dichte in zwei flächenmässig gleich grosse Teile auf. Bei symmetrischen Dichten gilt zudem, dass der Median dem Erwartungswert und dem Symmetriepunkt entspricht, denn der Erwartungswert ist ja gerade der Schwerpunkt.

Quantile kann man auch für diskrete Verteilungen definieren. Dort “trifft” man α aber in der Regel *nicht* exakt, da die Verteilungsfunktion ja eine Stufenfunktion ist (dies haben wir nicht betrachtet).

Wie im diskreten Fall gibt es auch im stetigen Fall gewisse Verteilungen, die immer wieder gebraucht werden. Wir wollen nun die wichtigsten davon betrachten.

2.3.3 Uniforme Verteilung [Uni(a, b)]

Die **uniforme Verteilung** mit den Parametern $a, b \in \mathbb{R}$ tritt z.B. auf bei Rundungsfehlern und als Formalisierung der völligen “Ignoranz”. Sie ist die stetige Version des Laplace-Modells. Hier ist $W = [a, b]$ und

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$$

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b. \end{cases}$$

Die Dichte ist also *konstant* und die kumulative Verteilungsfunktion eine *lineare* Funktion auf dem Definitionsbereich $[a, b]$, siehe Abbildung 2.7.

Für Erwartungswert und Varianz gilt

$$\mathbb{E}[X] = \frac{a+b}{2}$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}.$$

Beispiel. Ein Computer liefert Zufallszahlen X , die uniform-verteilt auf $[0, 5]$ sind. Was ist die Wahrscheinlichkeit, dass wir eine Zahl beobachten, die im Intervall $[2, 4]$ liegt? Es ist

$$\mathbb{P}(2 \leq X \leq 4) = \frac{2}{5},$$

denn das Integral entspricht hier gerade der Rechtecksfläche.

◁

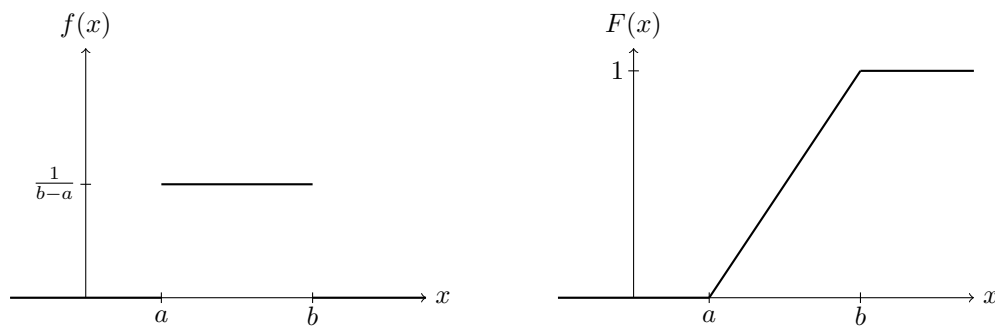


Abbildung 2.7: Dichte (links) und Verteilungsfunktion (rechts) der uniformen Verteilung.

2.3.4 Normalverteilung $[\mathcal{N}(\mu, \sigma^2)]$

Die **Normal-** oder **Gauss-Verteilung** mit den Parametern $\mu \in \mathbb{R}$ und $\sigma > 0$ ist die häufigste Verteilung für Messwerte. Hier ist $W = \mathbb{R}$ sowie

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\}, \quad x \in \mathbb{R}$$

mit

$$\begin{aligned} \mathbb{E}[X] &= \mu \\ \text{Var}(X) &= \sigma^2. \end{aligned}$$

Dies bedeutet, dass die Parameter gerade der Erwartungswert bzw. die Varianz (oder Standardabweichung) sind.

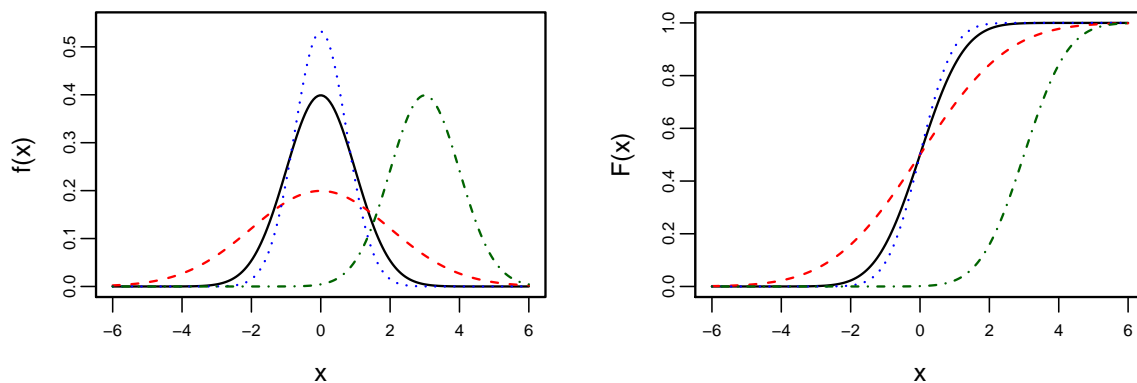


Abbildung 2.8: Dichte und Verteilungsfunktion der Normalverteilung für $\mu = 0, \sigma = 1$ (schwarz, durchgezogen), $\mu = 0, \sigma = 2$ (rot, gestrichelt), $\mu = 0, \sigma = 0.75$ (blau, gepunktet) und $\mu = 3, \sigma = 1$ (grün, strichpunktet).

Die Dichte der Normalverteilung ist symmetrisch um den Erwartungswert μ . Je grösser σ , desto flacher oder breiter wird die Dichte. Für kleine σ gibt es einen “schmalen und hohen” Gipfel. Mit μ verschieben wir einfach die Dichte nach links bzw. rechts, siehe auch Abbildung 2.8.

Die Fläche über dem Intervall $[\mu - \sigma, \mu + \sigma]$ ist ca. $2/3$. Die Fläche über dem Intervall $[\mu - 2\sigma, \mu + 2\sigma]$ ist ca. 0.95, siehe auch Abbildung 2.9. Oder ausgedrückt in Wahrscheinlichkeiten: Die Wahrscheinlichkeit, weniger als eine Standardabweichung vom Erwartungswert entfernt zu liegen, beträgt ca. 66%. Bei zwei Standardabweichungen sind es ca. 95%. Das heisst, dass nur 5% der Werte mehr als zwei Standardabweichungen vom Erwartungswert entfernt liegen.

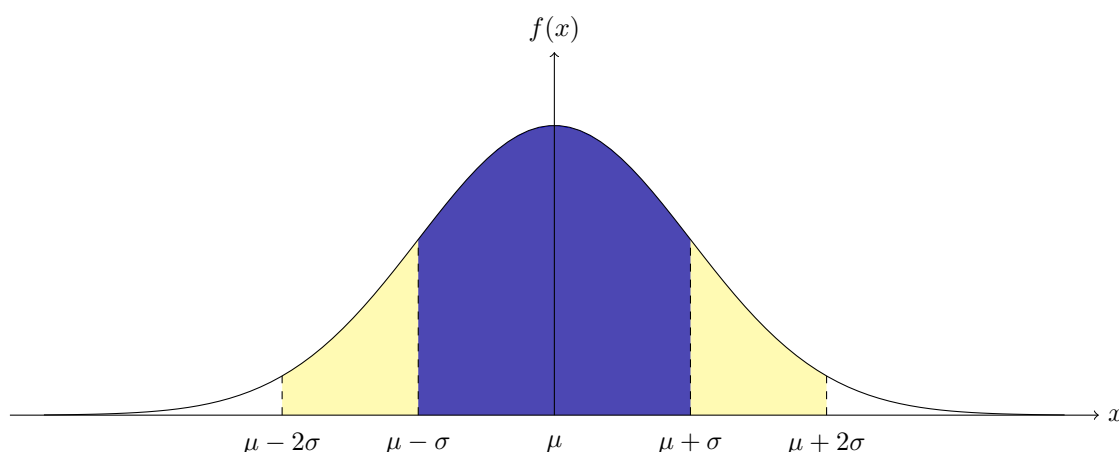


Abbildung 2.9: Dichte der Normalverteilung. Ca. 66% der Fläche befindet sich im Intervall $[\mu - \sigma, \mu + \sigma]$, ca. 95% der Fläche im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$.

Standardnormalverteilung

Die $\mathcal{N}(0, 1)$ -Verteilung, auch als **Standardnormalverteilung** bezeichnet, ist ein wichtiger Sonderfall, weshalb es für deren Dichte und Verteilungsfunktion sogar eigene Symbole gibt. Es sind dies

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

$$\Phi(x) = \int_{-\infty}^x \varphi(u) \, du.$$

Die Funktion Φ ist leider *nicht* geschlossen darstellbar. Eine Tabelle findet man in Anhang [A.3](#).

Die entsprechenden Quantile kürzen wir hier ab mit

$$z_\alpha = \Phi^{-1}(\alpha), \quad \alpha \in (0, 1).$$

Die Verteilungsfunktion F einer $\mathcal{N}(\mu, \sigma^2)$ -verteilten Zufallsvariable kann man aus der Verteilungsfunktion Φ der Standardnormalverteilung berechnen mittels der Formel

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

für $x \in \mathbb{R}$, mehr dazu in Kürze.

2.3.5 Exponentialverteilung $[\text{Exp}(\lambda)]$

Die **Exponentialverteilung** mit Parameter $\lambda > 0$ ist das einfachste Modell für Wartezeiten auf Ausfälle und eine stetige Version der geometrischen Verteilung. Hier ist $W = [0, \infty)$,

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

Das führt zu

$$\mathbb{E}[X] = 1/\lambda$$

$$\text{Var}(X) = 1/\lambda^2.$$

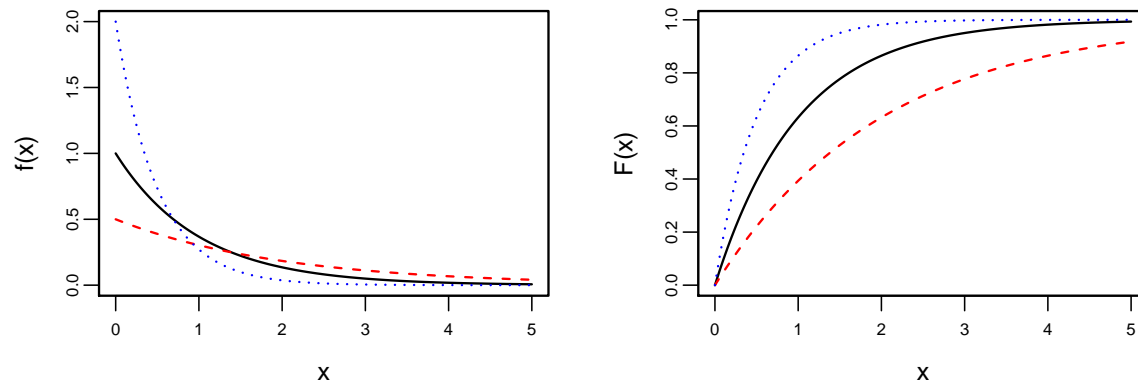


Abbildung 2.10: Dichte und Verteilungsfunktion der Exponentialverteilung für $\lambda = 1$ (schwarz, durchgezogen), $\lambda = 2$ (blau, gepunktet) und $\lambda = 1/2$ (rot, gestrichelt).

Der Parameter $\lambda > 0$ wird als **Ausfallrate** interpretiert.

Beispiel. Die Lebensdauer T eines Bauteils (in Wochen) sei exponential-verteilt mit erwarteter Lebensdauer 15 Wochen. Es ist also $T \sim \text{Exp}(\lambda)$ mit $\lambda = 1/15$. Die Wahrscheinlichkeit, dass das Bauteil in den ersten 10 Wochen ausfällt, ist in diesem Falle gegeben durch

$$\mathbb{P}(T \leq 10) = F(10) = 1 - e^{-\lambda \cdot 10} = 1 - e^{-10/15} = 0.487.$$

Die Wahrscheinlichkeit, dass das Bauteil mindestens 20 Wochen hält, ist

$$\mathbb{P}(T > 20) = 1 - F(20) = e^{-\lambda \cdot 20} = e^{-20/15} = 0.264.$$

◁

2.3.6 Transformationen

Bei stetigen Verteilungen spielen Transformationen eine wichtige Rolle. Transformationen treten bereits auf bei “simplen” Dingen wie der Änderung von Masseinheiten (z.B. Fahrenheit statt Celsius). Es kann auch sein, dass Sie die Verteilung der Dauer X einer typischen Baustelle kennen, aber sich für die Verteilung der mit der Dauer verbundenen Kosten $Y = g(X)$ interessieren, wobei die Kosten eine spezielle (monotone) Funktion der Dauer sind.

Wir betrachten also hier jeweils die neue Zufallsvariable $Y = g(X)$, wobei wir davon ausgehen, dass wir sowohl die Verteilung von X wie auch die Funktion g kennen. Das Ziel ist es, aus diesen Angaben die Verteilung von Y zu ermitteln.

Um Missverständnisse zu vermeiden, schreiben wir hier jeweils im Index der Verteilungsfunktion, des Erwartungswertes etc., um was für Zufallsvariablen es sich handelt.

Linearer Fall

Falls g **linear** ist mit $g(x) = a + bx$ für $b > 0$, dann gilt

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(a + bX \leq y) \\ &= \mathbb{P}\left(X \leq \frac{y - a}{b}\right) \\ &= F_X\left(\frac{y - a}{b}\right). \end{aligned}$$

Wir brauchen die Bedingung $b > 0$, damit das Zeichen “ \leq ” nicht umkehrt. Für den Fall $b < 0$ haben wir

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(a + bX \leq y) \\ &= \mathbb{P}\left(X \geq \frac{y-a}{b}\right) \\ &= 1 - F_X\left(\frac{y-a}{b}\right). \end{aligned}$$

Durch Ableiten erhält man dann die Dichte und damit das folgende Resultat.

Für $b \neq 0$ ist die Dichte von $Y = a + bX$ gegeben durch

$$f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right).$$

Beispiel. Wenn $X \sim \mathcal{N}(\mu, \sigma^2)$, dann gilt für $Y = a + bX$, dass $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$, denn nach obiger Transformationsformel haben wir

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma|b|} \exp\left\{-\frac{1}{2}\left(\frac{\frac{y-a}{b} - \mu}{\sigma}\right)^2\right\} = \frac{1}{\sqrt{2\pi}\sigma|b|} \exp\left\{-\frac{1}{2}\left(\frac{y-a-b\mu}{\sigma|b|}\right)^2\right\},$$

was die Dichte einer Normalverteilung mit Erwartungswert $a + b\mu$ und Varianz $b^2\sigma^2$ ist. Wir “verlassen” also die Normalverteilung nicht, wenn wir lineare Transformationen anwenden (bei der Poissonverteilung geht dies z.B. nicht). Durch Skalenänderungen (d.h. $a = 0, b > 0$) kann man auch alle Exponentialverteilungen ineinander überführen. Auch uniforme Verteilungen kann man durch lineare Transformation ineinander überführen. \triangleleft

Mit den Rechenregeln von früher haben wir zudem

$$\begin{aligned} \mathbb{E}[Y] &= a + b\mathbb{E}[X] \\ \text{Var}(Y) &= b^2 \text{Var}(X) \\ \sigma_Y &= |b|\sigma_X. \end{aligned}$$

Diese Kennzahlen müssen wir also *nicht* via Umweg über die transformierte Dichte berechnen.

Standardisierung

Wir können eine Zufallsvariable X immer so linear transformieren, dass sie Erwartungswert 0 und Varianz 1 hat, indem wir die Transformation

$$g(x) = \frac{x - \mu_X}{\sigma_X}$$

anwenden. Für $Z = g(X)$ gilt (nachrechnen!)

$$\begin{aligned} \mathbb{E}[Z] &= 0 \\ \text{Var}(Z) &= 1. \end{aligned}$$

Wir sprechen in diesem Zusammenhang von **Standardisierung**. Typischerweise verwenden wir den Buchstaben Z für standardisierte Zufallsvariablen.

Standardisierung ist z.B. bei der Normalverteilung nützlich. Sei $X \sim \mathcal{N}(\mu, \sigma^2)$. Wie gross ist dann $\mathbb{P}(X \leq 3)$? Wir haben

$$\begin{aligned}\mathbb{P}(X \leq 3) &= \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{3 - \mu}{\sigma}\right) \\ &= \mathbb{P}\left(Z \leq \frac{3 - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{3 - \mu}{\sigma}\right),\end{aligned}$$

denn $Z \sim \mathcal{N}(0, 1)$. Falls $\mu = 2$ und $\sigma = 4$ haben wir

$$\mathbb{P}(X \leq 3) = \mathbb{P}(Z \leq 0.25) = \Phi(0.25).$$

In der Tabelle in A.3 lesen wir ab, dass $\Phi(0.25) = 0.5987$ (Zeile “.2” und Spalte “.05”).

Wir können also mit diesem Trick alle Normalverteilungen zurückführen auf die Standardnormalverteilung. Dies ist auch der Grund, wieso nur diese tabelliert ist.

Allgemeiner monotoner Fall

Ist g eine (beliebige) differenzierbare, streng *monotone* Funktion, so hat $Y = g(X)$ die Dichte

$$f_Y(y) = \left| \frac{1}{g'(g^{-1}(y))} \right| f_X(g^{-1}(y)), \quad y \in W_Y.$$

Falls W_X der Wertebereich von X ist, so ist der Wertebereich von Y gegeben durch

$$W_Y = g(W_X) = \{g(x), x \in W_X\}.$$

Die Herleitung der Transformationsformel geht genau gleich wie im linearen Fall.

Beispiel. Wenn $X \sim \mathcal{N}(\mu, \sigma^2)$ normalverteilt ist, dann folgt die transformierte Zufallsvariable $Y = e^X$ einer sogenannten **Lognormalverteilung**. Eine Zufallsvariable $Y > 0$ heisst also lognormalverteilt, wenn der Logarithmus davon normalverteilt ist. Die Dichte ist gemäss obiger Transformationsformel gegeben durch

$$f_Y(y) = \begin{cases} 0 & y \leq 0 \\ \frac{1}{\sqrt{2\pi\sigma y}} \exp\left\{-\frac{1}{2}\left(\frac{\log(y)-\mu}{\sigma}\right)^2\right\} & y > 0, \end{cases}$$

denn wir haben hier $g(x) = e^x$, $g'(x) = e^x$, $g^{-1}(y) = \log(y)$ und damit $g'(g^{-1}(y)) = y$. ◁

Wie wir schon früher gesehen haben, gilt für beliebiges g immer

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Wir brauchen für den Erwartungswert von Y die transformierte Dichte f_Y also *nicht*.

Achtung: Der Erwartungswert transformiert nicht einfach mit. Falls g konvex ist (d.h. $g'' \geq 0$), so gilt die **Jensen'sche Ungleichung**

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]).$$

Beispiel. Ist Y lognormal-verteilt, so gilt

$$\mathbb{E}[Y] = e^{\mu + \sigma^2/2} > e^\mu = g(\mu),$$

wobei wir die linke Seite ohne Herleitung einfach hinschreiben. \triangleleft

Die Quantile transformieren bei monoton wachsenden Funktionen mit, d.h. das $(\alpha \times 100)\%$ -Quantil q_α von X wird zum $(\alpha \times 100)\%$ -Quantil $g(q_\alpha)$ bei Y , denn

$$\alpha = \mathbb{P}(X \leq q_\alpha) = \mathbb{P}(g(X) \leq g(q_\alpha)) = \mathbb{P}(Y \leq g(q_\alpha)).$$

Beispiel. Der Median der Lognormalverteilung ist $e^\mu = g(\mu)$. Im Gegensatz zum Erwartungswert transformiert der Median also einfach mit. \triangleleft

2.3.7 Simulation von Zufallsvariablen

Wenn U uniform auf $[0, 1]$ verteilt ist und F eine beliebige kumulative Verteilungsfunktion ist, dann ist die Verteilungsfunktion von $X = F^{-1}(U)$ gleich F , denn

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F_U(F(x)) = F(x),$$

wobei wir hier ausgenutzt haben, dass die Verteilungsfunktion (streng) monoton wachsend ist und dass $F_U(x) = x$ bei der uniformen Verteilung auf $[0, 1]$, siehe Kapitel 2.3.3.

Was bringt uns dieses Resultat? Es ist sehr nützlich, um Zufallsvariablen zu **simulieren**. So lange wir eine Implementierung der $\text{Uni}(0, 1)$ -Verteilung haben, können wir mit diesem Trick “beliebige” Verteilungen simulieren. Man geht dabei folgendermassen vor

1. Erzeuge eine Realisation u von einer uniform-verteilten Zufallsvariable $U \sim \text{Uni}(0, 1)$. Dies wird mittels einem “Standard-Paket” gemacht.
2. Berechne $x = F^{-1}(u)$. Gemäss obiger Herleitung ist dann x eine Realisation einer Zufallsvariablen X mit kumulativer Verteilungsfunktion F .

2.4 Ausblick: Poissonprozesse

Eine Verallgemeinerung der Poissonverteilung sind sogenannte **Poissonprozesse**. Ein Poissonprozess kommt zum Zug, wenn man z.B. die Anzahl Ereignisse in einem Zeitintervall zählt, wie z.B. die Anzahl Skiunfälle in einer Woche. Wenn wir das Zeitintervall verdoppeln, dann erwarten wir auch doppelt so grosse Anzahlen. Man muss also eine Rate oder **Intensität** λ spezifizieren (pro Zeiteinheit). Die Anzahl in einem Intervall der Länge t modelliert man dann mit einer Poissonverteilung mit Parameter λt . Dabei nimmt man zusätzlich an, dass Anzahlen aus disjunkten (nicht überlappenden) Zeitintervallen unabhängig sind.

Es sei also $N(t)$ die Anzahl Ereignisse im Zeitintervall $[0, t]$, $t \in \mathbb{R}$. Für einen sogenannten **homogenen Poissonprozess** gilt

$$N(t) \sim \text{Pois}(\lambda t).$$

Sei jetzt T_1 der Zeitpunkt des *ersten* Ereignisses. Es gilt

$$\{T_1 > t\} = \{\text{Kein Ereignis in } [0, t]\} = \{N(t) = 0\}.$$

Also haben wir

$$\mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t},$$

bzw.

$$\mathbb{P}(T_1 \leq t) = 1 - e^{-\lambda t}.$$

Die Zeit bis zum ersten Ereignis ist also exponential-verteilt mit Parameter λ , d.h. $T_1 \sim \text{Exp}(\lambda)$. Wegen der Annahme der Unabhängigkeit gilt allgemein, dass bei homogenen Poissonprozessen die Zeiten zwischen zwei aufeinanderfolgenden Ereignissen exponential-verteilt sind.

2.5 Vergleich der Konzepte: Diskrete vs. stetige Verteilungen

Die wichtigsten Konzepte der stetigen und diskreten Verteilungen sind in Tabelle 2.1 einander gegenüber gestellt.

2.6 Review / Lernziele



- ☐ Sie kennen den Begriff der Zufallsvariable und der dazugehörigen Verteilung.
- ☐ Sie kennen den Unterschied zwischen diskreten und stetigen Verteilungen und können die entsprechenden Konzepte einander gegenüber stellen.
- ☐ Sie können die kumulative Verteilungsfunktion einer Zufallsvariable berechnen und kennen deren Eigenschaften.
- ☐ Sie kennen die wichtigsten Kennzahlen (Erwartungswert, Varianz, ...) einer Verteilung und ihre Bedeutung. Zudem wissen sie, wie sich lineare Transformationen auf die Kennzahlen auswirken.
- ☐ Sie kennen die wichtigsten diskreten und stetigen Verteilungen, die Bedeutung deren Parameter, und Sie wissen, für welche Situationen welche Verteilungen in der Regel verwendet werden.
- ☐ Sie wissen, wie sich bei stetigen Zufallsvariablen Transformationen auf die Dichte auswirken. Sie wissen zudem, wie man eine Zufallsvariable standardisieren kann und wieso dies bei der Normalverteilung nützlich ist.
- ☐ Sie kennen den Begriff des Poissonprozesses und seine Eigenschaften.

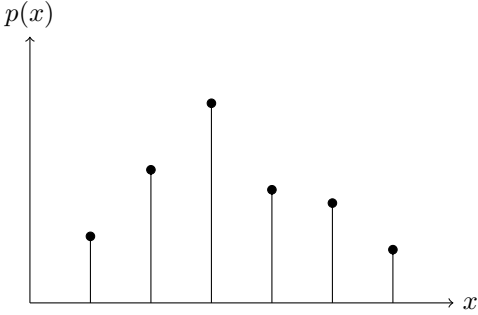
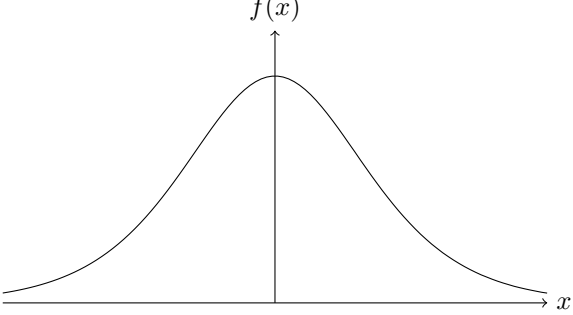
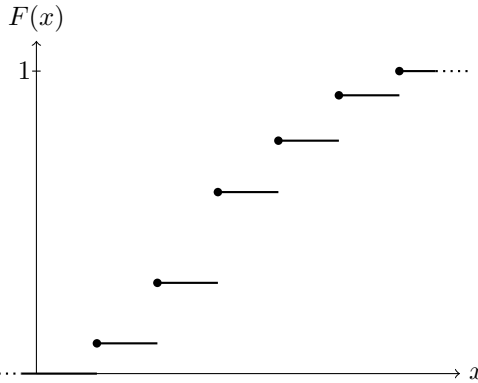
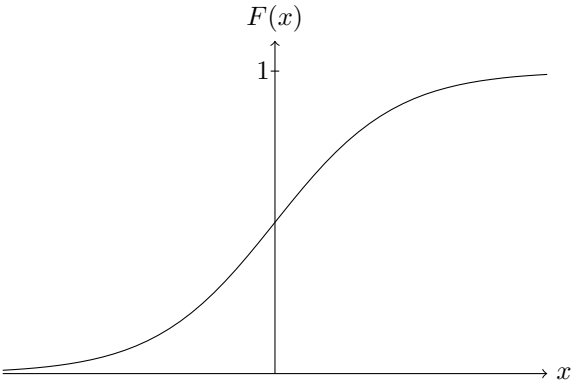
diskret	stetig
<p>Wahrscheinlichkeitsfunktion</p>  <p>$\mathbb{P}(X = x_k) = p(x_k) \in [0, 1], x_k \in W$</p>	<p>Dichte</p>  <p>$\mathbb{P}(X = x) = 0, x \in W$</p>
<p>Kumulative Verteilungsfunktion</p>  <p>$F(x) = \sum_{k: x_k \leq x} p(x_k)$</p>	<p>Kumulative Verteilungsfunktion</p>  <p>$F(x) = \int_{-\infty}^x f(u) \, du$</p>
<p>Erwartungswert</p> <p>$\mathbb{E}[X] = \sum_{k \geq 1} x_k p(x_k)$</p> <p><i>etc.</i></p>	<p>Erwartungswert</p> <p>$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) \, dx$</p>

Tabelle 2.1: Vergleich der Konzepte der diskreten und der stetigen Verteilungen.

3 Deskriptive Statistik

3.1 Einführung

In der *schliessenden* Statistik wird es später darum gehen, aus beobachteten Daten Schlüsse über den dahinterliegenden datengenerierenden Mechanismus zu ziehen. Man nimmt dabei jeweils an, dass die Daten Realisierungen von Zufallsvariablen sind, deren Verteilung man aufgrund der Daten bestimmen möchte. Hier bei der *deskriptiven* (oder *beschreibenden*) Statistik geht es in einem ersten Schritt hingegen zunächst einmal darum, die vorhandenen Daten übersichtlich *darzustellen* und *zusammenzufassen*.

Mit Grafiken können wir sehr schnell erkennen, ob unsere Daten unerwartete Strukturen und Besonderheiten aufweisen. Wenn immer man also Daten sammelt, ist es sozusagen eine Pflicht, die Daten als erstes mit geeigneten Grafiken darzustellen. Man muss sich aber auch bewusst sein, dass wenn immer man Daten zusammenfasst – sei dies durch Kennzahlen oder Grafiken – zwangsläufig auch Information verloren geht!

Unsere Daten interpretieren wir als **Stichprobe** einer (grossen) **Grundgesamtheit**. Wir können z.B. eine Stichprobe von 50 Studenten von allen an der ETH eingeschriebenen Studenten ziehen und von diesen gewisse Eigenschaften analysieren. Damit eine Stichprobe die Grundgesamtheit gut repräsentiert, muss sie idealerweise *zufällig* aus der Grundgesamtheit entnommen werden (d.h. jedes Element der Grundgesamtheit hat die gleiche Wahrscheinlichkeit, ausgewählt zu werden). Man spricht dann von einer sogenannten (einfachen) **Zufallsstichprobe**.

Es ist schwierig und aufwändig, eine gute Stichprobe zu ziehen. In der Praxis wird leider oft der Weg des kleinsten Aufwands gewählt. Man sollte daher bei Stichproben generell skeptisch sein und den Mechanismus, mit dem die Daten gewonnen wurden, kritisch hinterfragen. In dem Sinne gilt: Eine gut gewählte kleine Stichprobe ist viel aussagekräftiger als eine schlecht gewählte grosse Stichprobe!

3.2 Kennzahlen

Wir betrachten also einen Datensatz mit n Beobachtungen: x_1, x_2, \dots, x_n . Wenn wir z.B. $n = 15$ Prüfkörper bezüglich ihrer Druckfestigkeit ausmessen, dann ist x_i die Druckfestigkeit des i -ten Prüfkörpers, $i = 1, \dots, 15$.

Für die numerische Zusammenfassung von Daten gibt es diverse Kennzahlen. Das **arithmetische Mittel (Durchschnitt, Mittelwert, Stichprobenmittel)**

$$\bar{x} = \frac{1}{n}(x_1 + \dots + x_n)$$

ist eine Kennzahl für die *Lage* der Daten und entspricht gerade dem Schwerpunkt der Datenpunkte, wenn wir jeder Beobachtung das gleiche Gewicht geben. Das arithmetische Mittel ist also gerade das empirische Pendant des Erwartungswertes (**empirisch** bedeutet: experimentell beobachtet bzw. aus Daten ermittelt).

Die **empirische Standardabweichung** s ist die Wurzel aus der **empirischen Varianz**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

und eine Kennzahl für die *Streuung* der Daten. Der auf den ersten Blick gewöhnungsbedürftige Nenner $n-1$ ist mathematisch begründet und sorgt dafür, dass man keinen systematischen Fehler macht (siehe später). Auf der Modellseite entspricht der empirischen Varianz natürlich die Varianz.

Je grösser also die empirische Standardabweichung (Varianz), desto “breiter” streuen unsere Beobachtungen um das arithmetische Mittel.

Um weitere Kennzahlen zu definieren, führen wir zuerst die **geordneten Werte**

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

ein. Dies ist nichts anderes als unsere in aufsteigender Reihenfolge geordnete Stichprobe. Also: Wenn immer wir den Index einer Beobachtung in Klammern setzen, gehen wir davon aus, dass die Beobachtungen der Grösse nach aufsteigend geordnet sind.

Das **empirische** $(\alpha \times 100)\%$ -**Quantil** q_α ($0 < \alpha < 1$) ist die Beobachtung $x_{(k)}$, die die geordneten Daten (in etwa) im Verhältnis $\alpha : (1 - \alpha)$ aufteilt. D.h. ca. $(\alpha \times 100)\%$ der Beobachtungen sind kleiner als $x_{(k)}$ und $(1 - \alpha) \times 100\%$ sind grösser. Genauer: Das empirische $(\alpha \times 100)\%$ -Quantil q_α ist definiert als

$$q_\alpha = \begin{cases} \frac{1}{2} (x_{(\alpha \cdot n)} + x_{(\alpha \cdot n + 1)}) & \text{falls } \alpha \cdot n \text{ eine ganze Zahl ist} \\ x_{(\lceil \alpha \cdot n \rceil)} & \text{sonst} \end{cases}$$

Die Notation $\lceil \alpha \cdot n \rceil$ bedeutet, dass man auf die nächste grössere ganze Zahl aufrundet: $k = \lceil \alpha \cdot n \rceil$ ist die kleinste ganze Zahl, die grösser als $\alpha \cdot n$ ist. Wenn $\alpha \cdot n$ eine ganze Zahl ist, mittelt man also über zwei Beobachtungen aus, sonst nimmt man die nächste grössere ganze Zahl und betrachtet diese Beobachtung. Es gibt noch (viele) alternative Definitionen des empirischen Quantils; für grosse n wird der Unterschied zwischen den Definitionen vernachlässigbar.

Ein spezielles Quantil ist der **empirische Median** (oder **Zentralwert**). Er ist definiert als das 50%-Quantil und steht “in der Mitte” der geordneten Stichprobe. Also haben wir entsprechend obiger Definition

$$q_{0.5} = \begin{cases} \frac{1}{2} (x_{(\frac{n}{2})} + x_{(\frac{n}{2} + 1)}) & \text{falls } n \text{ gerade} \\ x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \end{cases}$$

Der empirische Median ist wie das arithmetische Mittel eine Kennzahl für die *Lage* der Datenpunkte. Im Gegensatz zum arithmetischen Mittel ist der Median “robust”: Wenn wir z.B. den grössten Wert in unserem Datensatz nochmals stark erhöhen (wenn wir z.B. bei der Datenaufnahme einen Fehler machen und eine Null zu viel schreiben), so ändert sich der Median *nicht*. Anschaulich interpretiert: Der Median schaut nur, ob links und rechts gleich viele Beobachtungen liegen, die aktuelle Lage der Beobachtungen spielt keine Rolle. Das arithmetische Mittel hingegen kann sich bei einer solchen Datenänderung *drastisch* verändern und ist demnach *nicht* robust.

Neben dem Median werden oft auch noch die **Quartile** verwendet: Das **untere Quartil** ist das empirische 25%-Quantil, das **obere Quartil** entsprechend das empirische 75%-Quantil.

Die **Quartilsdifferenz** (engl. interquartile range, IQR) ist die Differenz zwischen dem oberen und dem unteren Quartil. Sie ist eine (robuste) Kennzahl für die *Streuung* der Daten.

Beispiel. *Old Faithful Geysir*

Wir betrachten einen Auszug aus Daten des Geysirs “Old Faithful” im Yellowstone Nationalpark (USA). Notiert wurde die Dauer (in Minuten) von 10 Eruptionen.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
3.600	1.800	3.333	2.283	4.533	2.883	4.700	3.600	1.950	4.350

Die geordneten Beobachtungen sind also demnach

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
1.800	1.950	2.283	2.883	3.333	3.600	3.600	4.350	4.533	4.700

Wir haben

$$\begin{aligned}\bar{x} &= 3.3032 \\ s^2 &= 1.11605 \\ s &= 1.056433.\end{aligned}$$

Der empirische Median ist gegeben durch

$$q_{0.5} = \frac{1}{2}(3.333 + 3.600) = 3.4665.$$

Das empirische 15%-Quantil ist gegeben durch die zweitkleinste Beobachtung, denn $10 \cdot 0.15 = 1.5$ und demnach $\lceil 10 \cdot 0.15 \rceil = 2$, also

$$q_{0.15} = x_{(2)} = 1.950.$$

◁

3.3 Grafische Darstellungen

Typische grafische Darstellungen eines eindimensionalen Datensatzes sind das *Histogramm*, der *Boxplot* und die *empirische kumulative Verteilungsfunktion*. Wenn man Daten paarweise beobachtet kommen noch andere Grafiken dazu.

3.3.1 Histogramm

Beim **Histogramm** teilen wir den Wertebereich der Beobachtungen auf, d.h. wir bilden *Klassen* (Intervalle) $(c_{k-1}, c_k]$. Ferner ermitteln wir die Anzahl Beobachtungen in den entsprechenden Intervallen. Diese Anzahlen bezeichnen wir mit h_k .

Grafisch trägt man über den Intervallen Balken auf, deren Höhe *proportional* ist zu

$$\frac{h_k}{c_k - c_{k-1}}.$$

Dies führt dazu, dass die *Fläche* der Balken dann proportional zu der Anzahl Beobachtungen im entsprechenden Intervall ist. Wenn man überall die gleiche Klassenbreite $c_k - c_{k-1}$ wählt, so kann man auch direkt die Anzahlen auftragen. Eine schematische Darstellung findet man in Abbildung 3.1. Im rechten Histogramm sind zwei Klassen zusammengefasst worden. Das Histogramm ist die empirische Version der Dichte und liefert einen guten Überblick über die empirische Verteilung: Man sieht z.B. sehr einfach, wie (un)symmetrisch eine Verteilung ist, ob sie mehrere Gipfel hat etc.

Die Wahl der Anzahl Klassen ist subjektiv. Je nach Wahl der Intervalle kann es sein, dass Strukturen verschwinden. Wenn wir z.B. die Klassenbreite sehr gross wählen, kann es sein, dass mehrere Gipfel “verschmolzen” werden zu einem einzelnen Gipfel. Wenn man die Klassenbreite grösser macht, findet “Erosion” statt: Gipfel werden abgetragen und Täler werden aufgefüllt.

Eine mögliche Faustregel für die Anzahl Klassen ist die sogenannte “Sturges Rule”: Diese teilt die Spannbreite der Daten auf in $\lceil 1 + \log_2(n) \rceil$ gleich breite Intervalle. Zur Erinnerung: das Symbol $\lceil \cdot \rceil$ bedeutet, dass man auf die nächst grössere ganze Zahl aufrundet.

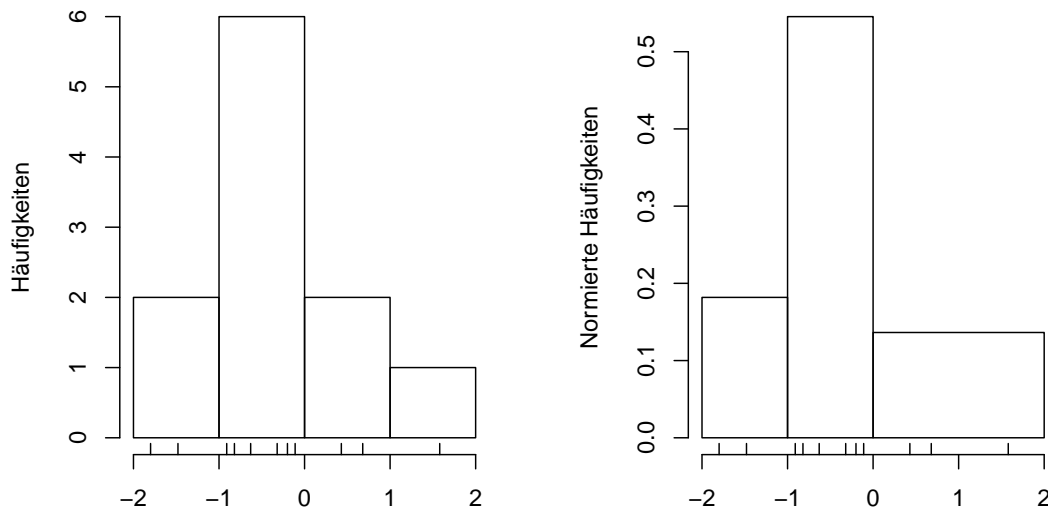


Abbildung 3.1: Schematische Darstellung von zwei Histogrammen vom gleichen Datensatz. Zur Illustration sind die einzelnen Beobachtungen mit kleinen Strichen eingezeichnet. Das rechte Histogramm ist so normiert, dass die Totalfläche 1 ergibt.

3.3.2 Boxplot

Wenn man sehr viele Verteilungen miteinander vergleichen will (z.B. wenn man eine Grösse bei verschiedenen Versuchsbedingungen oder an verschiedenen Orten misst), wird es oft schwierig Histogramme zu verwenden. Eine geeignetere Wahl sind sogenannte *Boxplots*.

Der **Boxplot** (siehe Abbildung 3.2) besteht aus einem Rechteck, das vom unteren und vom oberen Quartil begrenzt ist. Innerhalb des Rechtecks markieren wir den Median mit einem Strich. Hinzu kommen Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten “normalen” Wert gehen. Per Definition ist ein normaler Wert höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt. Beobachtungen, die weiter entfernt sind (sogenannte Ausreisser) werden zusätzlich durch Punkte eingezeichnet.

3.3.3 Empirische kumulative Verteilungsfunktion

Die **empirische kumulative Verteilungsfunktion** F_n ist die empirische Version der kumulativen Verteilungsfunktion. Sie ist definiert als

$$F_n(x) = \frac{1}{n} \text{Anzahl}\{i \mid x_i \leq x\} \in [0, 1].$$

Der Wert $F_n(2)$ gibt einem also an, wie gross im Datensatz der Anteil der Beobachtungen ist, die kleiner gleich 2 sind. Insbesondere ist also F_n eine Treppenfunktion, die an den Datenpunkten einen Sprung der Höhe $1/n$ hat (bzw. ein Vielfaches davon, wenn ein Wert mehrmals vorkommt). Links von der kleinsten Beobachtung ist die Funktion 0 und rechts von der grössten Beobachtung ist die Funktion 1. In Regionen wo viele Punkte liegen (das Histogramm hat dort einen Peak), ist die empirische kumulative Verteilungsfunktion also steil.

In Abbildung 3.3 sind Histogramme, Boxplots und empirische kumulative Verteilungsfunktionen von vier (fiktiven) Datensätzen der Grösse $n = 100$ dargestellt. Man sieht z.B., dass beim dritten Datensatz im Boxplot nicht ersichtlich ist, dass die Verteilung zwei Gipfel hat. Man spricht in diesem Fall von einer sogenannten *bimodalen* Verteilung.

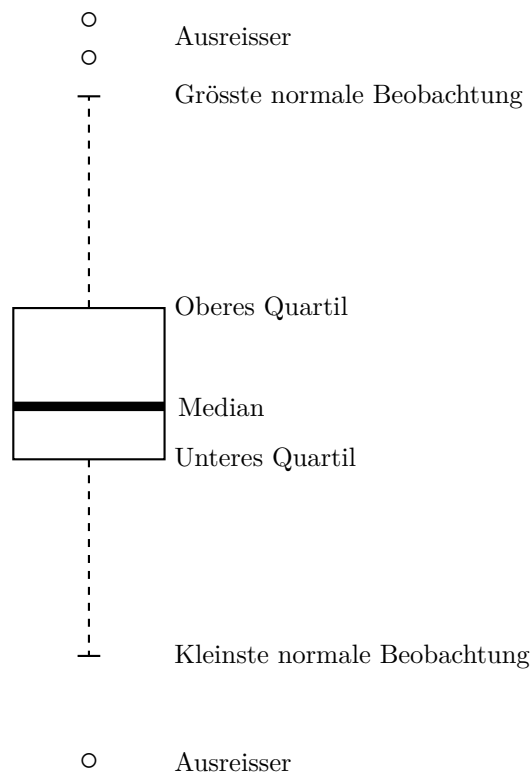


Abbildung 3.2: Schematische Darstellung eines Boxplots.

3.4 Mehrere Messgrößen

Oft liegen die Daten *paarweise* vor. Wir haben in diesem Fall n Datenpaare $(x_1, y_1), \dots, (x_n, y_n)$. So kann z.B. x_i das Verkehrsaufkommen beim Gubrist-Tunnel und y_i das Verkehrsaufkommen beim Baregg-Tunnel sein am gleichen Tag i . In der Regel interessiert man sich für die Zusammenhänge (Abhängigkeiten) zwischen den beiden Größen x_i und y_i .

Die einfachste Form der Abhängigkeit ist die *lineare* Abhängigkeit. Diese wird numerisch durch die **empirische Korrelation** r erfasst:

$$r = \frac{s_{xy}}{s_x s_y} \in [-1, 1],$$

wobei

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

die **empirische Kovarianz** zwischen x_i und y_i ist. Mit s_x und s_y bezeichnen wir die empirische Standardabweichungen der x_i bzw. y_i .

Die empirische Korrelation r ist eine *dimensionslose* Grösse. Es gilt

$$\begin{aligned} r = +1 & \quad \text{genau dann, wenn } y_i = a + bx_i \text{ für ein } a \in \mathbb{R} \text{ und ein } b > 0. \\ r = -1 & \quad \text{genau dann, wenn } y_i = a + bx_i \text{ für ein } a \in \mathbb{R} \text{ und ein } b < 0. \end{aligned}$$

D.h. das Vorzeichen von r gibt die *Richtung* und der Betrag von r die *Stärke* der linearen Abhängigkeit an. Einige Beispiele findet man in Abbildung 3.4.

Man sollte nie die Korrelation r einfach “blind” aus den Daten berechnen, ohne auch das Streudiagramm betrachtet zu haben! Ganz verschiedene Strukturen können zum gleichen Wert von r führen, siehe Abbildung 3.4 bzw. 3.5.

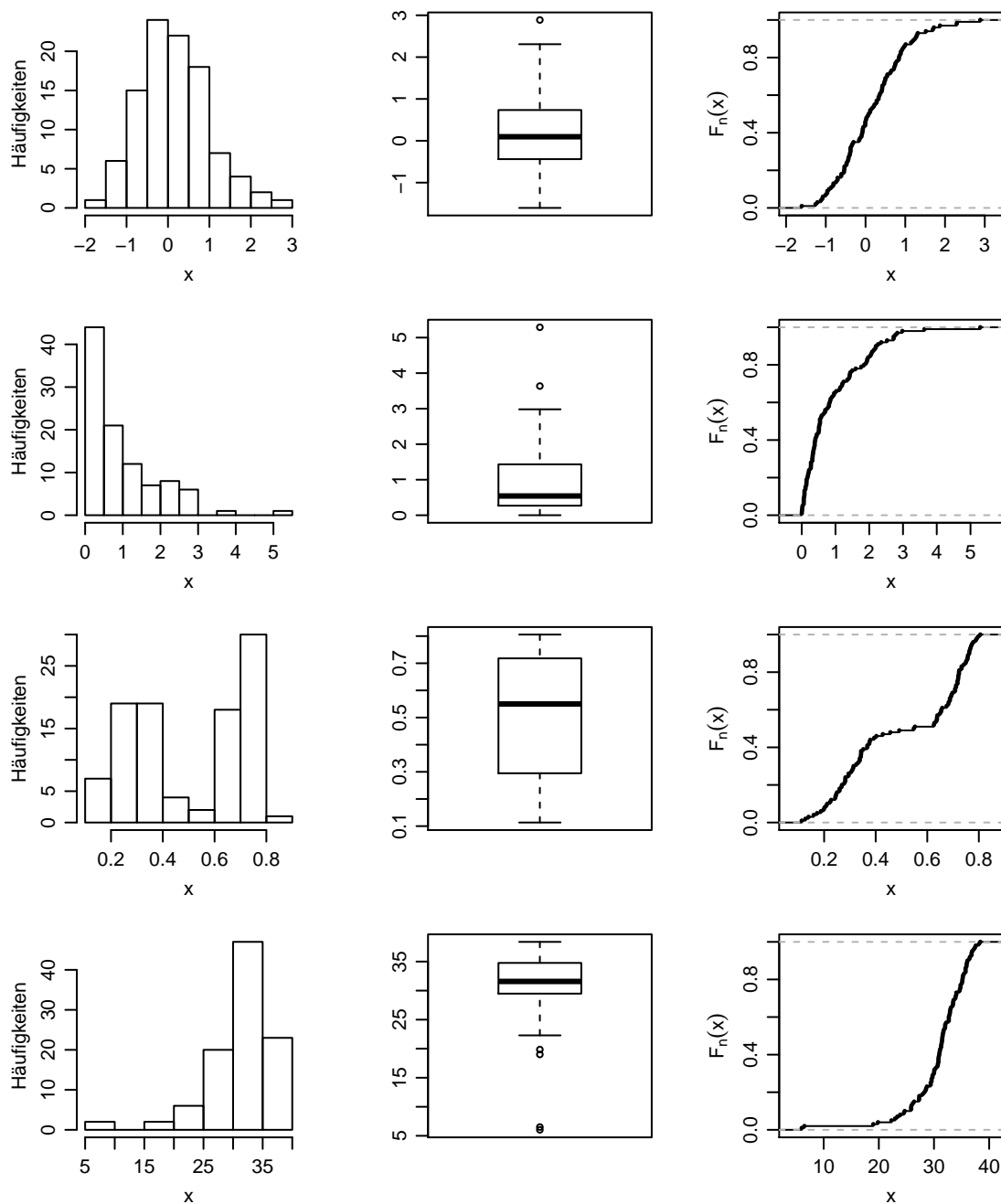


Abbildung 3.3: Histogramm (links), Boxplot (mitte) und empirische kumulative Verteilungsfunktion (rechts) von 4 Datensätzen der Grösse $n = 100$.

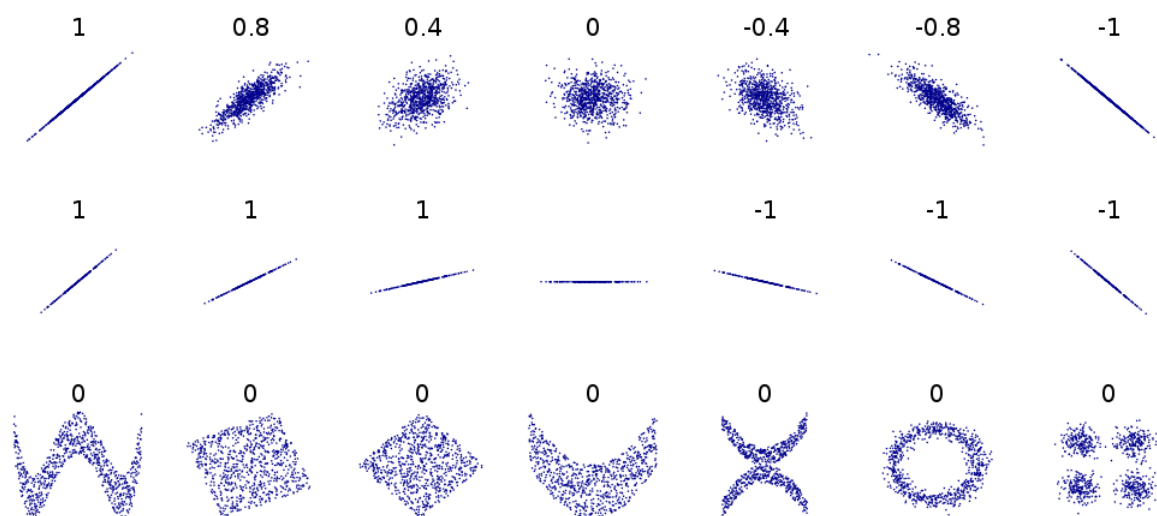


Abbildung 3.4: Empirische Korrelation bei verschiedenen Datensätzen (Quelle: Wikipedia).

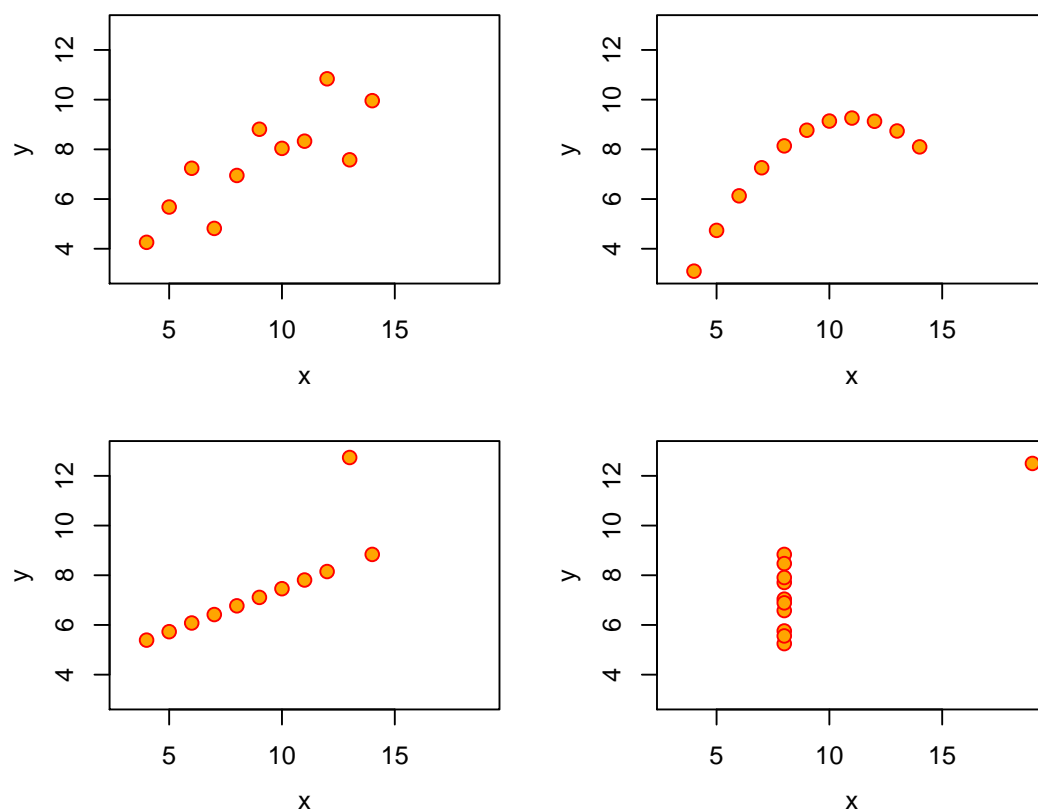


Abbildung 3.5: Vier Datensätze (von Anscombe) mit jeweils identischer empirischer Korrelation $r = 0.82$ zwischen x und y .

3.5 Modell vs. Daten

Wir haben jetzt also “beide Welten” kennen gelernt. Auf der einen Seite die Modelle (Verteilungen), auf der anderen Seite die konkret vorliegenden Daten, die wir als Realisierungen von Zufallsvariablen der entsprechenden Verteilung auffassen.

Die Kennzahlen und Funktionen bei den Modellen sind theoretische Grössen. Wenn wir (unendlich) viele Beobachtungen von einer Verteilung haben, dann entsprechen die empirischen Grössen gerade den korrespondierenden theoretischen Grössen. Oder anders herum: Für einen konkreten Datensatz kann man die empirischen Grössen auch als Schätzungen für die theoretischen Grössen betrachten. Dies werden wir in der schliessenden Statistik dann genauer betrachten. In Tabelle 3.1 sind die entsprechenden “Gegenstücke” nochmals aufgelistet, vorerst aber nur für den eindimensionalen Fall. Die Theorie für den zweidimensionalen (oder mehrdimensionalen) Fall betrachten wir im nächsten Kapitel.

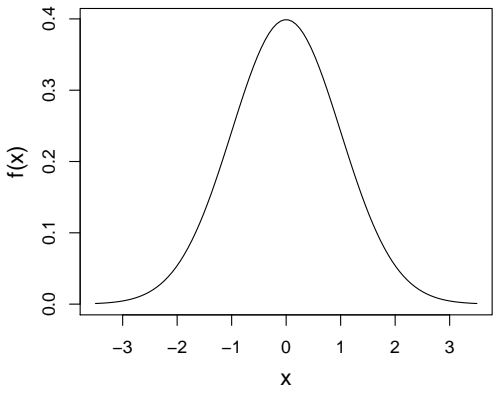
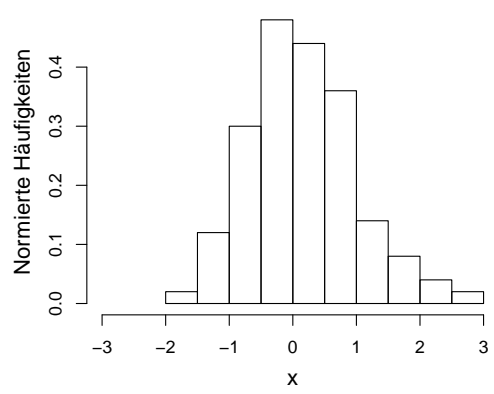
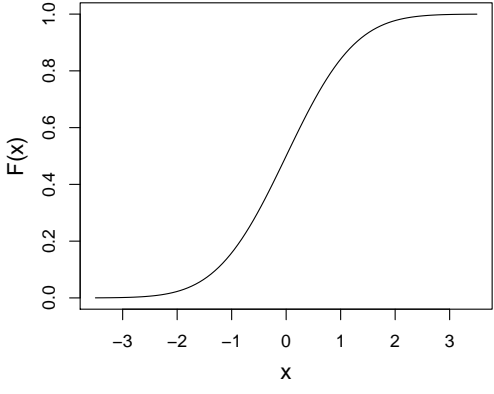
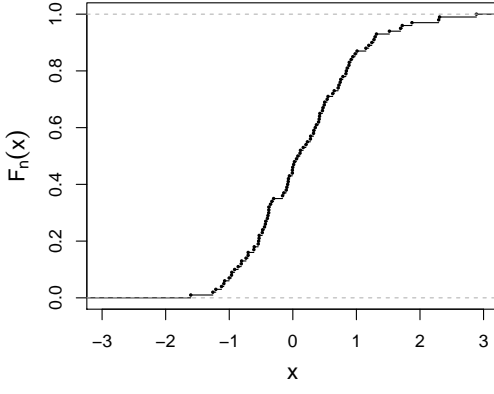
Modell	Daten
<p>Dichte</p> 	<p>Histogramm</p> 
<p>Kumulative Verteilungsfunktion</p> 	<p>Empirische kumulative Verteilungsfunktion</p> 
<p>Erwartungswert $\mathbb{E}[X]$</p>	<p>Arithm. Mittel $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$</p>
<p>Varianz $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$</p>	<p>Emp. Varianz $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$</p>
<p>Quantile q_α</p> <p>etc.</p>	<p>Emp. Quantile q_α</p>

Tabelle 3.1: Modell vs. Daten.

3.6 Review / Lernziele



- ☐ Sie kennen die wichtigsten empirischen Kennzahlen, deren Bedeutung (inkl. Gefahren) und deren Zusammenhang mit den entsprechenden Modellgrößen.
- ☐ Sie kennen die wichtigsten graphischen Darstellungsmöglichkeiten, deren Interpretation und sofern geeignet den Zusammenhang mit den entsprechenden Modellgrößen.

4 Mehrdimensionale Verteilungen

Wie wir in der deskriptiven Statistik schon kurz gesehen haben, misst man oft mehrere Grössen *gleichzeitig*, z.B. den Wasserstand an zwei verschiedenen Positionen A und B eines Flusses oder das Verkehrsaufkommen an verschiedenen Stellen einer Strasse. Oft kann man nicht von Unabhängigkeit zwischen den Messgrössen ausgehen. Wenn an Position A der Wasserstand hoch ist, dann wird dies wohl mit grosser Wahrscheinlichkeit auch an Position B der Fall sein (und umgekehrt). Für die Modellierung solcher Fälle greift man auf sogenannte *gemeinsame* Verteilungen zurück.

4.1 Gemeinsame, Rand- und bedingte Verteilungen

4.1.1 Diskreter Fall

Die **gemeinsame Verteilung** zweier diskreter Zufallsvariablen X mit Werten in W_X und Y mit Werten in W_Y ist gegeben durch die **gemeinsame Wahrscheinlichkeitsfunktion** von X und Y , d.h. die Werte

$$\mathbb{P}(X = x, Y = y), \quad x \in W_X, y \in W_Y.$$

In diesem “gemeinsamen” Zusammenhang nennt man dann die “einzelnen” Verteilungen $\mathbb{P}(X = x)$ von X und $\mathbb{P}(Y = y)$ von Y die **Randverteilungen** der gemeinsamen Zufallsvariable (X, Y) .

Die Randverteilungen lassen sich aus der gemeinsamen Verteilung berechnen durch

$$\mathbb{P}(X = x) = \sum_{y \in W_Y} \mathbb{P}(X = x, Y = y), \quad x \in W_X,$$

und analog für Y . Dies ist nichts anderes als der Satz der totalen Wahrscheinlichkeit.

Aus den Randverteilungen auf die gemeinsame Verteilung zu schliessen, geht aber *nur* im Falle der Unabhängigkeit von X und Y , denn dann gilt

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \mathbb{P}(Y = y), \quad x \in W_X, y \in W_Y.$$

In diesem Fall ist die gemeinsame Verteilung durch die Randverteilungen vollständig bestimmt und man erhält sie einfach durch Multiplikation.

Weiter definiert man die **bedingte Verteilung** von X gegeben $Y = y$ durch die Werte

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

Die Randverteilung lässt sich dann schreiben als

$$\mathbb{P}(X = x) = \sum_{y \in W_Y} \mathbb{P}(X = x \mid Y = y) \mathbb{P}(Y = y), \quad x \in W_X.$$

Diese Form kommt immer dann zum Einsatz, wenn man die Verteilung von X berechnen will, aber nur dessen bedingte Verteilung gegeben Y und die Verteilung von Y kennt.

Der **bedingte Erwartungswert** von Y gegeben $X = x$ ist gegeben durch

$$\mathbb{E}[Y \mid X = x] = \sum_{y \in W_Y} y \mathbb{P}(Y = y \mid X = x).$$

Ausser den neuen Begriffen haben wir soweit eigentlich alles schon einmal in leicht anderer Form gesehen, siehe bedingte Wahrscheinlichkeiten in Kapitel 1.

Beispiel. Zwei Wetterstationen X und Y messen die Bewölkung auf einer Skala von 1 bis 4. Die Wahrscheinlichkeiten für alle Kombinationen befinden sich in Tabelle 4.1. Es ist z.B.

$X \setminus Y$	1	2	3	4	Σ
1	0.080	0.015	0.003	0.002	0.1
2	0.050	0.350	0.050	0.050	0.5
3	0.030	0.060	0.180	0.030	0.3
4	0.001	0.002	0.007	0.090	0.1
Σ	0.161	0.427	0.240	0.172	1

Tabelle 4.1: Gemeinsame diskrete Verteilung von (X, Y) im Beispiel mit den Wetterstationen.

$$\mathbb{P}(X = 2, Y = 3) = 0.05.$$

Die Randverteilung von X befindet sich in der letzten Spalte. Es sind dies einfach die zeilenweise summierten Wahrscheinlichkeiten. Entsprechend findet man die Randverteilung von Y in der letzten Zeile. Die bedingte Verteilung von Y gegeben $X = 1$ ist gegeben durch die Wahrscheinlichkeiten

$$\frac{y}{\mathbb{P}(Y = y \mid X = 1)} \quad \begin{array}{c|c} 1 & 2 & 3 & 4 \\ \hline 0.8 & 0.15 & 0.03 & 0.02 \end{array}.$$

Dies ist die erste Zeile aus Tabelle 4.1 dividiert durch $\mathbb{P}(X = 1) = 0.1$. Wir können auch die Wahrscheinlichkeit berechnen, dass beide Stationen den gleichen Wert messen. Es ist dies die Summe der Wahrscheinlichkeiten auf der Diagonalen, d.h.

$$\mathbb{P}(X = Y) = \sum_{j=1}^4 \mathbb{P}(X = j, Y = j) = 0.08 + 0.35 + 0.18 + 0.09 = 0.7.$$

Wenn die beiden Zufallsvariablen unabhängig wären, dann wären die Einträge in der Tabelle jeweils das Produkt der entsprechenden Wahrscheinlichkeiten der Randverteilungen. Wir sehen schnell, dass das hier nicht der Fall ist. Also liegt keine Unabhängigkeit vor. \triangleleft

4.1.2 Stetiger Fall

Bei zwei oder mehreren stetigen Zufallsvariablen muss man das Konzept der Dichte auf mehrere Dimensionen erweitern.

Gemeinsame Dichte

Die **gemeinsame Dichte** $f_{X,Y}(\cdot, \cdot)$ von zwei stetigen Zufallsvariablen X und Y ist in “Ingenieurnotation” gegeben durch

$$\mathbb{P}(x \leq X \leq x + dx, y \leq Y \leq y + dy) = f_{X,Y}(x, y) dx dy.$$

Die Interpretation der Dichte ist also genau gleich wie früher. Die Darstellung als Ableitung einer geeigneten kumulativen Verteilungsfunktion ist nicht sehr instruktiv.

Die Wahrscheinlichkeit, dass der Zufallsvektor (X, Y) in $A \subset \mathbb{R}^2$ liegt, kann man dann wie im eindimensionalen Fall durch Integration der Dichte über den entsprechenden Bereich berechnen

$$\mathbb{P}((X, Y) \in A) = \iint_A f_{X,Y}(x, y) dx dy.$$

Ferner sind X und Y genau dann **unabhängig**, wenn

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad x, y \in \mathbb{R}. \quad (4.1)$$

In diesem Fall genügt das Konzept von eindimensionalen Dichten: die gemeinsame Dichte kann dann sehr einfach mittels *Multiplikation* berechnet werden.

Beispiel. Wir betrachten zwei Maschinen mit exponential-verteilten Lebensdauern $X \sim \text{Exp}(\lambda_1)$ und $Y \sim \text{Exp}(\lambda_2)$, wobei X und Y unabhängig seien. Was ist die Wahrscheinlichkeit, dass Maschine 1 länger läuft als Maschine 2? Die gemeinsame Dichte ist hier wegen der Unabhängigkeit gegeben durch

$$f_{X,Y}(x,y) = \lambda_1 e^{-\lambda_1 x} \lambda_2 e^{-\lambda_2 y}$$

für $x, y \geq 0$ (sonst ist die Dichte 0). Wir müssen das Gebiet

$$A = \{(x,y) : 0 \leq y < x\}$$

betrachten. Es sind dies alle Punkte unterhalb der Winkelhalbierenden, siehe Abbildung 4.1. Also haben wir

$$\begin{aligned} \mathbb{P}(Y < X) &= \int_0^\infty \left(\int_0^x \lambda_1 e^{-\lambda_1 x} \lambda_2 e^{-\lambda_2 y} dy \right) dx \\ &= \int_0^\infty \lambda_1 e^{-\lambda_1 x} (1 - e^{-\lambda_2 x}) dx \\ &= \int_0^\infty \lambda_1 e^{-\lambda_1 x} dx - \int_0^\infty \lambda_1 e^{-(\lambda_1 + \lambda_2)x} dx \\ &= 1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\lambda_2}{\lambda_1 + \lambda_2}. \end{aligned}$$

Das erste Integral in der zweitletzten Gleichung ist 1, weil wir über die Dichte der $\text{Exp}(\lambda_1)$ -Verteilung integrieren. \triangleleft

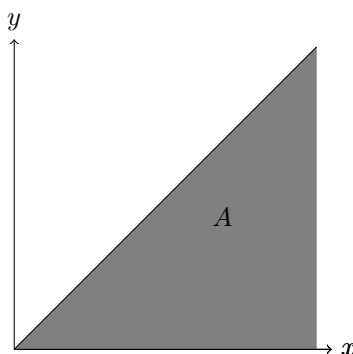


Abbildung 4.1: Integrationsbereich im Beispiel mit zwei Lebensdauern.

Randdichte und bedingte Dichte

Wie im diskreten Fall bezeichnen wir mit der **Randverteilung** die Verteilung der einzelnen Komponenten. Wir tun also so, als ob wir nur eine Komponente X bzw. Y “sehen würden”.

Aus der gemeinsamen Dichte erhält man die **Randdichte** von X bzw. Y durch “herausintegrieren” der anderen Komponente

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx.$$

Dies ist genau gleich wie im diskreten Fall, dort haben wir einfach mit Hilfe des Satzes der totalen Wahrscheinlichkeiten summiert statt integriert. Eine Illustration findet man in Abbildung 4.2.

Für die **bedingte Verteilung** von Y gegeben $X = x$ wird die **bedingte Dichte** benutzt, definiert durch

$$f_{Y|X=x}(y) = f_Y(y | X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Dies ist ein Quer- bzw. Längsschnitt der gemeinsamen Dichte. Wir halten x fest und variieren nur noch y . Der Nenner sorgt dafür, dass sich die Dichte zu 1 integriert. Im diskreten Fall haben wir einfach die entsprechende Zeile oder Spalte in der Tabelle festgehalten und umskaliert, so dass die Summe 1 ergab.

Der **bedingte Erwartungswert** von Y gegeben $X = x$ ist im stetigen Fall

$$\mathbb{E}[Y | X = x] = \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy.$$

Die Berechnung ist also wie beim “gewöhnlichen” Erwartungswert, man verwendet einfach die entsprechende bedingte Dichte.

Insbesondere folgt aus der Definition der bedingten Dichte, dass X und Y genau dann unabhängig sind, wenn gilt

$$f_{Y|X=x}(y) = f_Y(y) \text{ bzw. } f_{X|Y=y}(x) = f_X(x)$$

für alle x, y . Das bedeutet also, dass im Falle von Unabhängigkeit das Wissen von X keinen Einfluss auf die Verteilung von Y hat (bzw. umgekehrt).

Ferner können wir die gemeinsame Dichte immer schreiben als

$$f_{X,Y}(x,y) = f_{Y|X=x}(y)f_X(x) = f_{X|Y=y}(x)f_Y(y).$$

Dies ist insbesondere dann nützlich, wenn man ein Modell “stufenweise” definiert.

Aus den obigen Definitionen folgt, dass alle wahrscheinlichkeitstheoretischen Aspekte von zwei stetigen Zufallsvariablen X und Y durch deren *gemeinsame* Dichte $f_{X,Y}(\cdot, \cdot)$ vollständig bestimmt sind.

4.2 Erwartungswert bei mehreren Zufallsvariablen

Den Erwartungswert einer transformierten Zufallsvariable $Z = g(X, Y)$ mit $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ können wir berechnen als

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy.$$

Im diskreten Fall lautet die entsprechende Formel

$$\mathbb{E}[g(X, Y)] = \sum_{x \in W_X} \sum_{y \in W_Y} g(x, y) \mathbb{P}(X = x, Y = y).$$

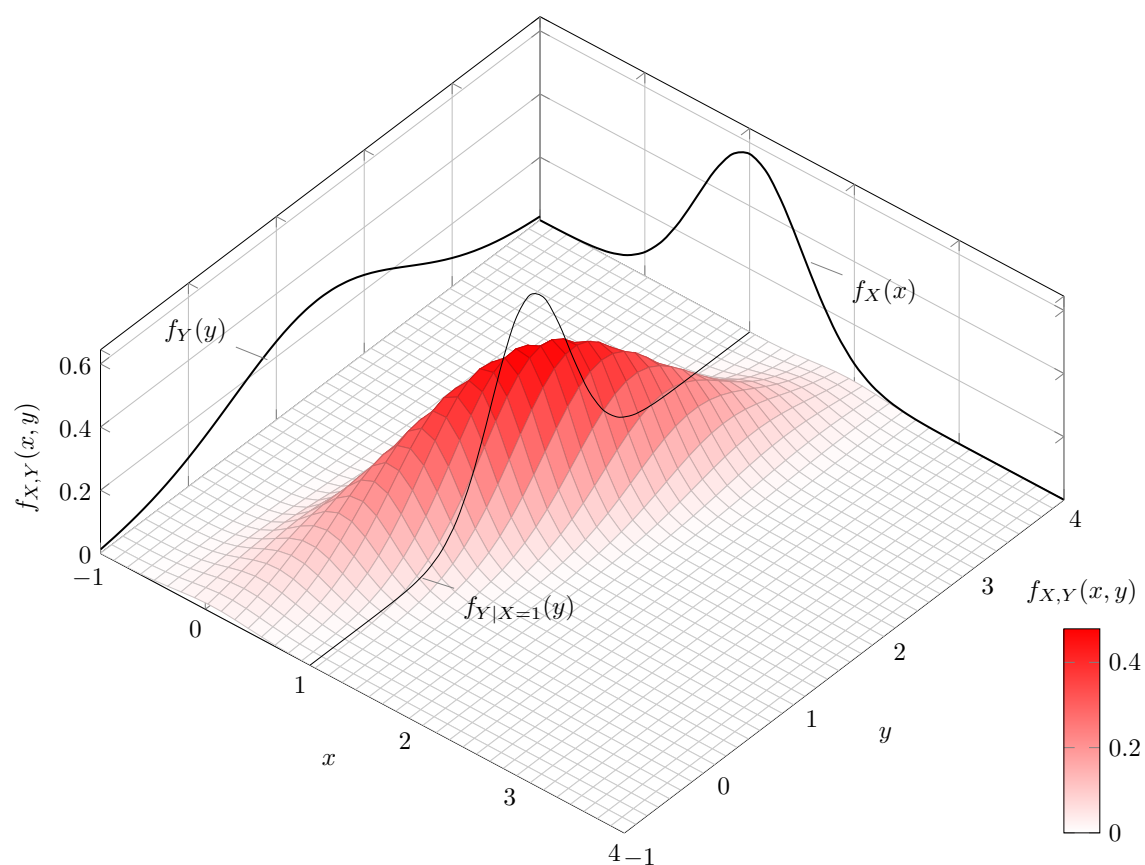


Abbildung 4.2: Illustration einer zweidimensionalen Dichte, deren Randverteilungen und der bedingten Verteilung gegeben $X = 1$. (tikZ Code von <http://tex.stackexchange.com/questions/31708/draw-a-bivariate-normal-distribution-in-tikz>).

Insbesondere gilt für eine Linearkombination

$$\mathbb{E}[a + bX + cY] = a + b \cdot \mathbb{E}[X] + c \cdot \mathbb{E}[Y], \quad a, b, c \in \mathbb{R}.$$

Dies gilt immer, egal ob die Zufallsvariablen unabhängig sind oder nicht. Wenn man mehr als zwei Zufallsvariablen betrachtet, geht alles analog, d.h.

$$\mathbb{E}\left[a_0 + \sum_{i=1}^n a_i X_i\right] = a_0 + \sum_{i=1}^n a_i \mathbb{E}[X_i],$$

wobei $a_0, a_1, \dots, a_n \in \mathbb{R}$.

4.3 Kovarianz und Korrelation

Da die gemeinsame Verteilung von abhängigen Zufallsvariablen im Allgemeinen kompliziert ist, begnügt man sich oft mit einer *vereinfachenden* Kennzahl zur Beschreibung der Abhängigkeit.

Man verwendet hierzu die **Kovarianz** bzw. die **Korrelation** zwischen X und Y . Diese sind folgendermassen definiert:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)], \quad \text{Corr}(X, Y) = \rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Die Korrelation ist nichts anderes als eine standardisierte Version der Kovarianz. Im Gegensatz zur Kovarianz ist die Korrelation also eine *dimensionslose* Grösse.

Es gilt immer

$$-1 \leq \text{Corr}(X, Y) \leq 1.$$

Die Korrelation ist ein Mass für die Stärke und Richtung der *linearen Abhängigkeit* zwischen X und Y . Es gilt

$$\text{Corr}(X, Y) = +1 \text{ genau dann, wenn } Y = a + bX \text{ für ein } a \in \mathbb{R} \text{ und ein } b > 0.$$

$$\text{Corr}(X, Y) = -1 \text{ genau dann, wenn } Y = a + bX \text{ für ein } a \in \mathbb{R} \text{ und ein } b < 0.$$

Falls also $|\text{Corr}(X, Y)| = 1$, so hat man einen *perfekten* linearen Zusammenhang zwischen X und Y . Falls $\text{Corr}(X, Y) = 0$ gilt, sagt man, dass X und Y **unkorreliert** sind. Es gibt dann *keinen* linearen Zusammenhang (es kann aber durchaus ein nichtlinearer Zusammenhang vorhanden sein).

Ferner gilt

$$\boxed{X \text{ und } Y \text{ unabhängig} \implies \text{Corr}(X, Y) = 0 \quad (\text{und damit auch } \text{Cov}(X, Y) = 0)} \quad (4.2)$$

Die Umkehrung gilt im Allgemeinen *nicht*, d.h. aus Unkorreliertheit folgt nicht Unabhängigkeit, siehe Abbildung 4.3. Ein Spezialfall, wo auch die Umkehrung gilt, wird in Kapitel 4.4 diskutiert.

In Abbildung 4.4 sind die Konturlinien (Höhenlinien) von zweidimensionalen Dichten für verschiedene Werte von ρ dargestellt. Wir sehen, dass wenn ρ betragsmässig gross wird, die Dichte immer konzentrierter “um eine Gerade herum” liegt. Das heisst, dass wir mit hoher Wahrscheinlichkeit Punkte sehen, die nahe bei dieser Geraden liegen.

Die Kovarianz können wir insbesondere zur Berechnung der Varianz von Summen von Zufallsvariablen verwenden (siehe unten).

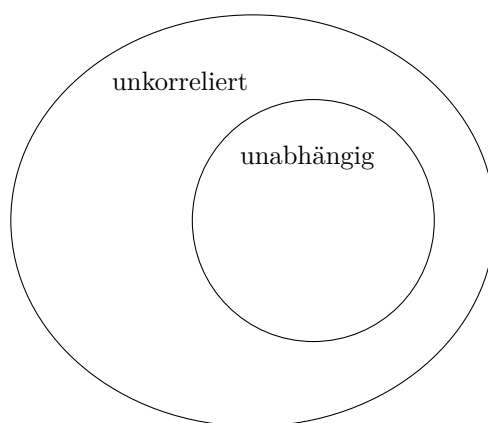


Abbildung 4.3: Zusammenhang zwischen Unkorreliertheit und Unabhängigkeit illustriert mit einem Venn-Diagramm.

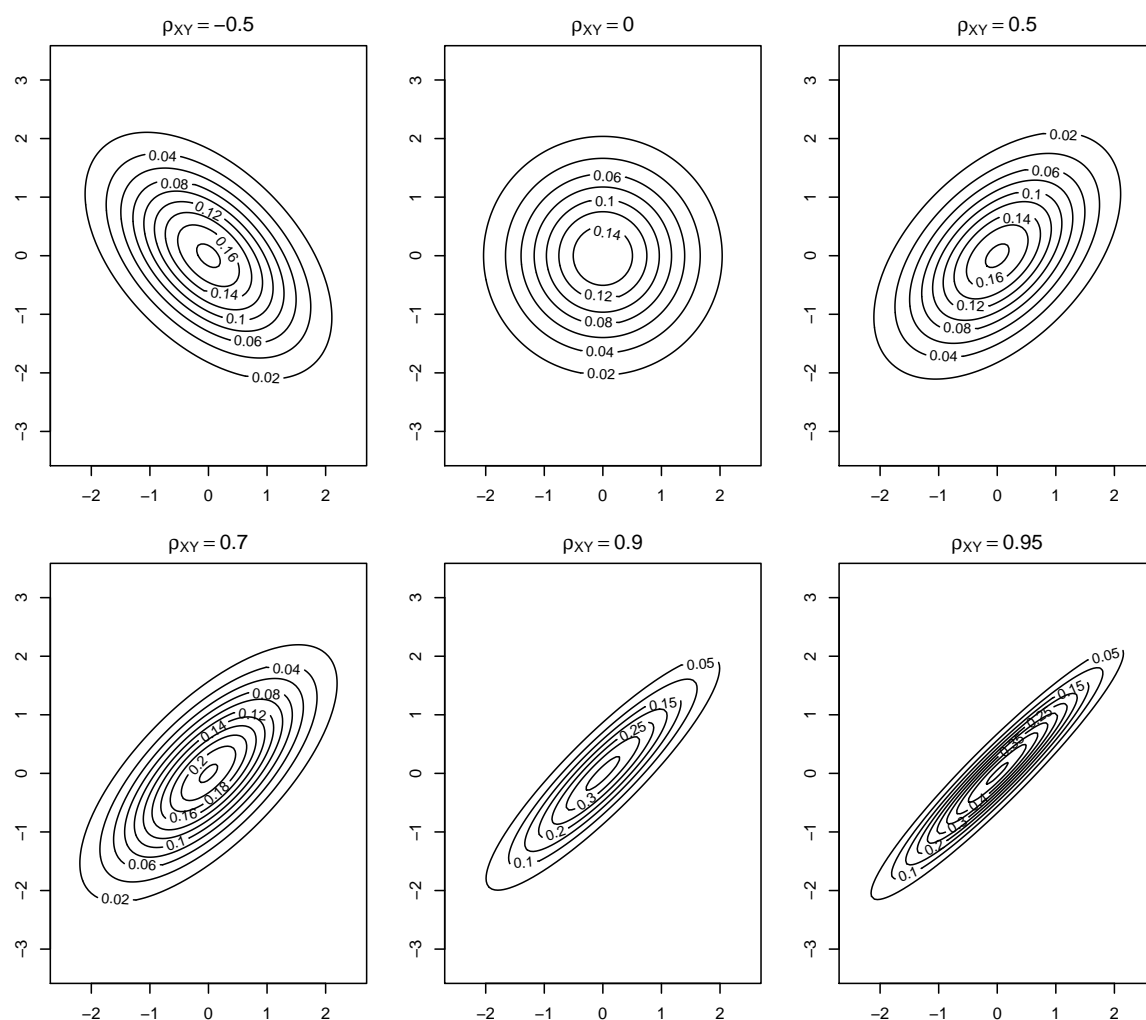


Abbildung 4.4: Konturlinien von zweidimensionalen Dichten für verschiedene Werte von ρ_{XY} . In der Tat handelt es sich hier um zweidimensionale Normalverteilungen, siehe Kapitel 4.4.

Unmittelbar aus der Definition der Kovarianz folgt sofort

$$\text{Var}(X) = \text{Cov}(X, X),$$

sowie die wichtige Formel

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \quad (4.3)$$

zur praktischen Berechnung der Kovarianz. Insbesondere gilt im Falle von Unabhängigkeit (bzw. allgemeiner Unkorreliertheit), dass

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y], \quad (X, Y \text{ unabhängig}).$$

Dies folgt sofort aus (4.3).

Ferner ist die Kovarianz *bilinear*, d.h. es gilt

$$\text{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, Y_j), \quad a_i, b_j \in \mathbb{R}.$$

und *symmetrisch*, d.h. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.

Weitere Rechenregeln sind

$$\begin{aligned} \text{Cov}(a + bX, c + dY) &= bd \text{Cov}(X, Y) \\ \text{Corr}(a + bX, c + dY) &= \text{sign}(b) \text{sign}(d) \text{Corr}(X, Y), \end{aligned}$$

wobei $\text{sign}(\cdot)$ die Vorzeichenfunktion ist.

Für die Varianz der Summe erhalten wir also

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j),$$

bzw. im Falle von nur zwei Zufallsvariablen

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y).$$

Falls die X_i unabhängig (oder noch allgemeiner unkorreliert) sind, ist die Varianz der Summe gleich der Summe der Varianzen

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n) \quad (X_1, \dots, X_n \text{ unabhängig}).$$

4.4 Zweidimensionale Normalverteilung

Die wichtigste zweidimensionale Verteilung ist die **zweidimensionale Normalverteilung**. Diese ist vollständig spezifiziert durch

- die Erwartungswerte und Varianzen der Randverteilungen: μ_X, σ_X^2 und μ_Y, σ_Y^2 ,
- sowie der Kovarianz zwischen X und Y : $\text{Cov}(X, Y)$.

Die gemeinsame Dichte ist gegeben durch die Funktion

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(x - \mu_X, y - \mu_Y) \Sigma^{-1} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}\right\},$$

wobei in Σ die Information über die Varianzen und Kovarianzen in einer Matrix “verpackt” ist. Man spricht von der sogenannten **Kovarianzmatrix**. Sie ist gegeben durch

$$\Sigma = \begin{pmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \sigma_Y^2 \end{pmatrix}.$$

In Abbildung 4.4 sind die Konturlinien für den Fall $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$ und verschiedene Werte von $\text{Cov}(X, Y)$ dargestellt. Mit $\mu_X \neq 0$ und $\mu_Y \neq 0$ würde man die Verteilung “herumschieben”.

Man kann zeigen, dass die Randverteilungen wieder normalverteilt sind: $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ und $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.

Wenn $\text{Cov}(X, Y) = 0$ gilt, so ist Σ eine Diagonalmatrix. Man kann nachrechnen, dass dann die Bedingung (4.1) gilt. Das heisst: im Falle der zweidimensionalen Normalverteilung gilt auch die Umkehrung von (4.2). Aus Unkorreliertheit folgt hier Unabhängigkeit. Im Allgemeinen gilt dies aber *nicht*, siehe Abbildung 4.3!

4.5 Dichte einer Summe von zwei Zufallsvariablen

Seien X, Y Zufallsvariablen mit gemeinsamer Dichte $f_{X,Y}$. Dann hat die neue Zufallsvariable $S = X + Y$ die Dichte (ohne Herleitung)

$$f_S(s) = \int_{-\infty}^{\infty} f_{X,Y}(x, s-x) dx.$$

Falls X und Y unabhängig sind, zerfällt die gemeinsame Dichte in das Produkt der Randdichten und wir haben

$$f_S(s) = \int_{-\infty}^{\infty} f_X(x) f_Y(s-x) dx.$$

Man spricht auch von der **Faltung** der beiden Funktionen f_X und f_Y .

Wenn wir nur am Erwartungswert (oder an der Varianz) von S interessiert sind, brauchen wir natürlich den Umweg über die Dichte nicht zu machen und können direkt wie in Kapitel 4.2 bzw. 4.3 vorgehen.

Beispiel. *Schauen wir einmal einen auf den ersten Blick “einfachen” Fall an. Wir betrachten zwei unabhängige Arbeitsprozesse. Der erste dauert zwischen 3 und 5 Minuten, der zweite zwischen 6 und 10 Minuten. Wir wollen jeweils uniforme Verteilungen annehmen. Die Frage ist, wie die totale Zeit verteilt ist.*

Es ist also $X \sim \text{Uni}(3, 5)$ und $Y \sim \text{Uni}(6, 10)$, wobei X und Y unabhängig sind. Die Dichten der beiden uniformen Verteilungen können wir auch schreiben als

$$f_X(x) = \frac{1}{2} 1_{[3,5]}(x),$$

$$f_Y(y) = \frac{1}{4} 1_{[6,10]}(y)$$

wobei $1_{[a,b]}(x)$ die sogenannte **Indikatorfunktion** ist, für die gilt

$$1_{[a,b]}(x) = \begin{cases} 1 & a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$$

Wir haben daher gemäss obiger Formel

$$f_S(s) = \frac{1}{8} \int_{-\infty}^{\infty} 1_{[3,5]}(x) \cdot 1_{[6,10]}(s-x) dx.$$

Das Integral ist nichts anderes als die Fläche des Bereichs, wo sich die beiden Indikatorfunktionen überlappen. Die zweite Indikatorfunktion können wir auch schreiben als $1_{[s-10, s-6]}(x)$, da $6 \leq s-x \leq 10$ äquivalent ist zu $s-10 \leq x \leq s-6$. Wenn wir also s grösser werden lassen, wandert diese nach rechts. Für $s < 9$ gibt es keine Überlappung. Zwischen 9 und 11 hat man eine teilweise und zwischen 11 und 13 eine volle Überlappung mit der Indikatorfunktion $1_{[3,5]}(x)$, die an Ort und Stelle stehen bleibt. Danach nimmt die Überlappung wieder ab und hört schlussendlich für $s > 15$ ganz auf. Diese Situationen sind in Abbildung 4.5 dargestellt. Dies führt zu

$$f_S(s) = \begin{cases} 0 & s < 9 \\ \frac{1}{8}(s-9) & 9 \leq s \leq 11 \\ \frac{1}{4} & 11 \leq s \leq 13 \\ \frac{1}{4} - \frac{1}{8}(s-13) & 13 \leq s \leq 15 \\ 0 & s > 15 \end{cases}$$

Die entsprechende Funktion ist in Abbildung 4.6 dargestellt. ◁

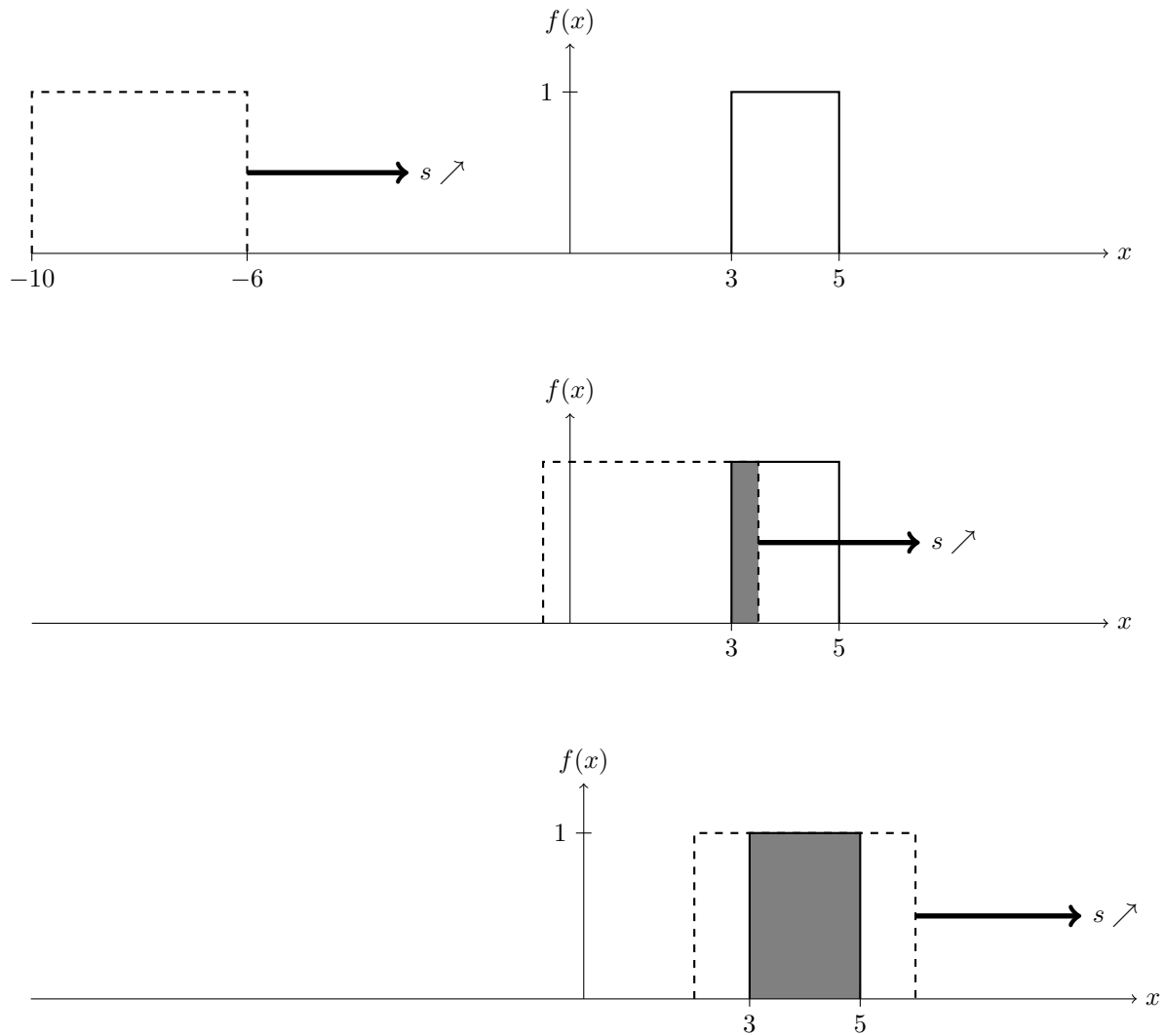


Abbildung 4.5: Illustration der beiden Indikatorfunktionen $1_{[s-10, s-6]}(x)$ (gestrichelt) und $1_{[3,5]}(x)$ (durchgezogen) für $s = 0$ (oben), $s = 9.5$ (mitte) und $s = 12$ (unten). Markiert ist die entsprechende Überlappung.

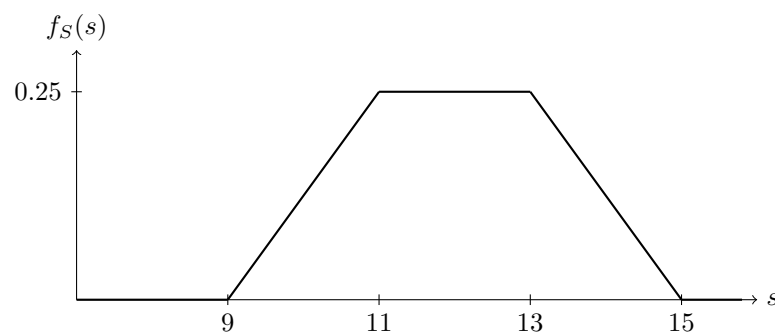


Abbildung 4.6: Dichte der Summe von zwei uniformen Verteilungen (Faltung).

Man kann so auch zeigen, dass wenn man zwei unabhängige Normalverteilungen addiert, wieder eine Normalverteilung resultiert. Das Resultat gilt auch, wenn die beiden Normalverteilungen korreliert sind, so lange sie noch zweidimensional normalverteilt sind. Man “verlässt” also in diesen Situationen die Normalverteilung nicht. Die Parameter müssen natürlich den üblichen Rechenregeln für Erwartungswert und Varianz folgen.

4.6 Mehr als zwei Zufallsvariablen

Alle diese Begriffe und Definitionen lassen sich natürlich auf mehr als zwei Zufallsvariablen verallgemeinern. Die Formeln sehen im Wesentlichen gleich aus, vor allem wenn man die Sprache der Linearen Algebra verwendet. *Ausblick:* Wenn man eine dynamische Grösse während eines Zeitintervalls misst, erhält man einen **stochastischen Prozess** $\{X(t); t \in [a, b]\}$. Die linearen Abhängigkeiten zwischen den Werten zu verschiedenen Zeitpunkten werden dann durch die sogenannte **Autokovarianzfunktion** beschrieben.

4.7 Vergleich der Konzepte: Diskrete vs. stetige mehrdimensionale Verteilungen

Die wichtigsten Konzepte der stetigen und diskreten mehrdimensionalen Verteilungen sind in Tabelle 4.2 nochmals einander gegenüber gestellt.

4.8 Review / Lernziele



- ☐ Sie verstehen das Konzept der mehrdimensionalen Verteilungen, sowohl im diskreten wie auch im stetigen Fall.
- ☐ Sie können aus der gemeinsamen Verteilung die bedingten und die Randverteilungen ermitteln (inkl. entsprechende Kennzahlen).
- ☐ Sie kennen die Korrelation als Mass für die lineare Abhängigkeit zwischen zwei Zufallsvariablen.
- ☐ Sie können die Varianz und den Erwartungswert einer Linearkombination von (abhängigen) Zufallsvariablen berechnen.

diskret	stetig
Wahrscheinlichkeitsfunktion $\mathbb{P}(X = x, Y = y)$ Kann in Form einer <i>Tabelle</i> angegeben werden.	Dichte $f_{X,Y}(x, y)$ $f_{X,Y}$ ist eine <i>Funktion</i> : $\mathbb{R}^2 \rightarrow \mathbb{R}$.
Randverteilungen: W'keitsfunktionen $\mathbb{P}(X = x) = \sum_{y \in W_Y} \mathbb{P}(X = x, Y = y)$ $\mathbb{P}(Y = y) = \sum_{x \in W_X} \mathbb{P}(X = x, Y = y)$ (Satz der totalen Wahrscheinlichkeit)	Randverteilungen: Dichten $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy$ $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx$ (Andere Komponente herausintegrieren)
Bedingte Verteilungen: W'keitsfunktionen $\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$ $\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}$ Fixierung einer Zeile bzw. Spalte in der Tabelle und Normierung der Wahrscheinlichkeiten auf Summe 1.	Bedingte Verteilungen: Dichten $f_X(x \mid Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$ $f_Y(y \mid X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$ Längs- bzw. Querschnitt der zweidimensionalen Dichte und Normierung auf Integral 1.
Erwartungswert von $g(X, Y)$ ($g: \mathbb{R}^2 \rightarrow \mathbb{R}$) $\mathbb{E}[g(X, Y)] = \sum_{x \in W_X, y \in W_Y} g(x, y) \mathbb{P}(X = x, Y = y)$	Erwartungswert von $g(X, Y)$ ($g: \mathbb{R}^2 \rightarrow \mathbb{R}$) $\mathbb{E}[g(X, Y)] = \iint_{\mathbb{R}^2} g(x, y) f_{X,Y}(x, y) \, dx \, dy$
Bedingter Erwartungswert $\mathbb{E}[X \mid Y = y] = \sum_{x \in W_X} x \mathbb{P}(X = x \mid Y = y)$ $\mathbb{E}[Y \mid X = x] = \sum_{y \in W_Y} y \mathbb{P}(Y = y \mid X = x)$	Bedingter Erwartungswert $\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_X(x \mid Y = y) \, dx$ $\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_Y(y \mid X = x) \, dy$
Unabhängigkeit zwischen X und $Y \iff$ $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \mathbb{P}(Y = y)$ $\mathbb{P}(X = x \mid Y = y) = \mathbb{P}(X = x)$ $\mathbb{P}(Y = y \mid X = x) = \mathbb{P}(Y = y)$ (jeweils für alle $x \in W_X, y \in W_Y$) Alle drei Aussagen sind äquivalent.	Unabhängigkeit zwischen X und $Y \iff$ $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ $f_X(x \mid Y = y) = f_X(x)$ $f_Y(y \mid X = x) = f_Y(y)$ (jeweils für alle $x \in W_X, y \in W_Y$) Alle drei Aussagen sind äquivalent.

Tabelle 4.2: Konzepte der diskreten (links) und der stetigen (rechts) mehrdimensionalen Verteilungen.

5 Grenzwertsätze

Wie wir schon ganz am Anfang gesehen haben, ist der Erwartungswert eine Idealisierung des arithmetischen Mittels bei unendlich vielen Wiederholungen. Dies bedeutet insbesondere, dass wir bei unendlich vielen Wiederholungen keine Varianz mehr haben, wenn wir das arithmetische Mittel betrachten. In diesem Kapitel wollen wir nun etwas genauer untersuchen, wie schnell die Varianz abfällt, und ob wir eine Aussage über die Verteilung des arithmetischen Mittels machen können (statt nur über Kennzahlen).

5.1 Die i.i.d. Annahme

Wir betrachten also n Zufallsvariablen X_1, \dots, X_n , wobei X_i die i -te Wiederholung von unserem Zufallsexperiment ist. Wir nehmen an, dass alle Zufallsvariablen die *gleiche* Verteilung haben und dass sie *unabhängig* voneinander sind, es gibt also keine Wechselwirkungen zwischen den verschiedenen Messungen. Man sagt in diesem Fall, dass die X_1, \dots, X_n **i.i.d.** sind. Die Abkürzung “i.i.d.” kommt vom Englischen: **i**ndependent and **i**dentically **d**istributed.

Die i.i.d. Annahme ist ein “Postulat”, welches in der Praxis in vielen Fällen vernünftig erscheint. Die Annahme bringt erhebliche Vereinfachungen, um mit mehreren Zufallsvariablen zu rechnen.

5.2 Summen und arithmetische Mittel von Zufallsvariablen

Ausgehend von X_1, \dots, X_n kann man neue Zufallsvariablen $Y = g(X_1, \dots, X_n)$ bilden. Hier betrachten wir die wichtigen Spezialfälle Summe

$$S_n = X_1 + \dots + X_n$$

und arithmetisches Mittel

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} S_n.$$

Wir nehmen stets an, dass die X_1, \dots, X_n i.i.d. sind.

Wenn $X_i = 1$, falls ein bestimmtes Ereignis bei der i -ten Wiederholung eintritt und $X_i = 0$ sonst, dann ist \bar{X}_n nichts anderes als die relative Häufigkeit dieses Ereignisses.

Im Allgemeinen ist es schwierig, die Verteilung von S_n exakt zu bestimmen. Es gibt aber folgende Ausnahmen:

1. Wenn $X_i \in \{0, 1\}$ wie oben, dann ist $S_n \sim \text{Bin}(n, p)$ mit $p = \mathbb{P}(X_i = 1)$.
2. Wenn $X_i \sim \text{Pois}(\lambda)$, dann ist $S_n \sim \text{Pois}(n\lambda)$.
3. Wenn $X_i \sim \mathcal{N}(\mu, \sigma^2)$, dann ist $S_n \sim \mathcal{N}(n\mu, n\sigma^2)$ und $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Einfacher sind die Berechnungen von Erwartungswert, Varianz und Standardabweichung.

$$\begin{array}{lll} \mathbb{E}[S_n] = n\mathbb{E}[X_i] & \text{Var}(S_n) = n \text{Var}(X_i) & \sigma_{S_n} = \sqrt{n} \sigma_{X_i} \\ \mathbb{E}[\bar{X}_n] = \mathbb{E}[X_i] & \text{Var}(\bar{X}_n) = \frac{1}{n} \text{Var}(X_i) & \sigma_{\bar{X}_n} = \frac{1}{\sqrt{n}} \sigma_{X_i}. \end{array}$$

Dies folgt aus den Rechenregeln von früher. Für die Varianz und die Standardabweichung ist die Unabhängigkeitsannahme zentral.

Die Standardabweichung der Summe *wächst* also, aber *langsamer* als die Anzahl Beobachtungen. D.h. auf einer *relativen* Skala haben wir eine kleinere Streuung für wachsendes n .

Die Standardabweichung des arithmetischen Mittels *nimmt ab* mit dem Faktor $1/\sqrt{n}$, da

$$\sigma_{\bar{X}_n} = \frac{1}{\sqrt{n}} \sigma_{X_i}.$$

Um die Standardabweichung zu halbieren, braucht man also *viermal* so viele Beobachtungen. Dies nennt man auch das \sqrt{n} -Gesetz.

5.3 Das Gesetz der Grossen Zahlen und der Zentrale Grenzwertsatz

Von den obigen Formeln über Erwartungswert und Varianz wissen wir, dass:

- $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X_i]$: das heisst \bar{X}_n hat den *gleichen* Erwartungswert wie ein einzelnes X_i .
- $\text{Var}(\bar{X}_n) \rightarrow 0$ ($n \rightarrow \infty$): das heisst, \bar{X}_n besitzt keine Variabilität mehr im Limes.

Diese beiden Punkte implizieren den folgenden Satz.

Gesetz der Grossen Zahlen (GGZ)

Seien X_1, \dots, X_n i.i.d. mit Erwartungswert μ . Dann gilt

$$\bar{X}_n \xrightarrow{n \rightarrow \infty} \mu.$$

Ein Spezialfall davon ist

$$f_n(A) \xrightarrow{n \rightarrow \infty} \mathbb{P}(A),$$

wobei $f_n(A)$ die relative Häufigkeit des Eintretens des Ereignisses A in n unabhängigen Experimenten ist (siehe Kapitel 1.1).

Korrekterweise müsste man den Begriff der Konvergenz für Zufallsvariablen zuerst mathematisch geeignet definieren.

Dies haben wir schon einmal gesehen, nämlich bei der Interpretation des Erwartungswertes als das Mittel bei unendlich vielen Beobachtungen, bzw. bei der Interpretation der Wahrscheinlichkeit als relative Häufigkeit bei unendlich vielen Versuchen. In Abbildung 1.2 sehen wir auch, dass die Streuung der relativen Häufigkeit mit zunehmendem n abnimmt.

Wir kennen also das Verhalten der Varianz von S_n und \bar{X}_n . Offen ist aber noch, ob wir Aussagen über die (genäherte) *Verteilung* von S_n und \bar{X}_n machen können (dies wäre eine viel stärkere Aussage als das GGZ). Dabei stützt man sich auf den folgenden berühmten Satz.

Zentraler Grenzwertsatz (ZGWS)

Seien X_1, \dots, X_n i.i.d. mit Erwartungswert μ und Varianz σ^2 , dann ist

$$\begin{aligned} S_n &\approx \mathcal{N}(n\mu, n\sigma^2) \\ \bar{X}_n &\approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \end{aligned}$$

für grosse n . Approximativ liegen also Normalverteilungen vor. Wie gut diese Approximationen für ein gegebenes n sind, hängt von der Verteilung der X_i 's ab. Für grosses n werden die Approximationen natürlich besser.

Selbst wenn wir die Verteilung der X_i *nicht* kennen, so haben wir eine Ahnung über die approximative Verteilung von S_n und \bar{X}_n ! Der Zentrale Grenzwertsatz ist mitunter ein Grund für die Wichtigkeit der Normalverteilung.

In Abbildung 5.1 sieht man den Zentralen Grenzwertsatz an einem empirischen Beispiel. Wir betrachten X_1, \dots, X_8 i.i.d. $\sim \text{Uni}(-1/2, 1/2)$. Von jeder Zufallsvariablen simulieren wir 5'000 Realisierungen. Wir betrachten die Histogramme für U_1 , $U_1 + U_2, \dots$ etc. und sehen, dass schon bei wenigen Summanden eine glockenförmige Struktur vorliegt.

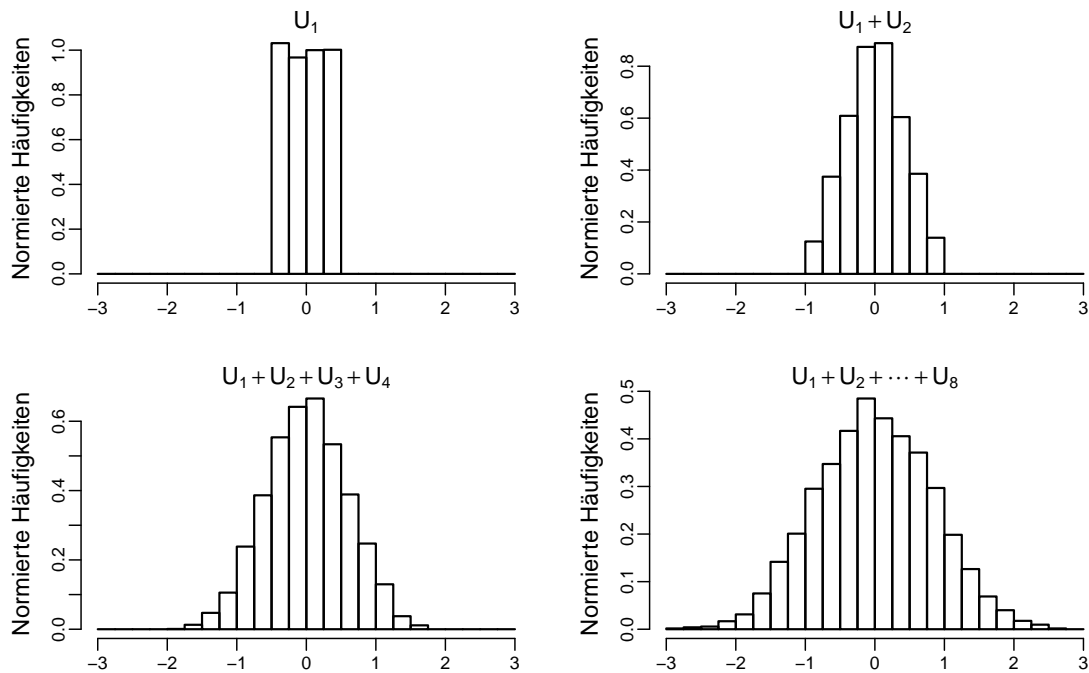


Abbildung 5.1: Histogramme der Summen von simulierten uniform-verteilten Zufallsvariablen. Die Stichprobengrösse beträgt jeweils 5'000.

Wenn wir die entsprechenden Dichten aufzeichnen würden, hätten wir qualitativ genau das gleiche Bild.

Die X_i 's können natürlich auch diskret sein. Wir haben schon bei der Binomialverteilung in Abbildung 2.2 gesehen, dass diese für n gross "glockenförmig" aussieht. Dasselbe gilt für die Poissonverteilung in Abbildung 2.4 für grösser werdendes λ .

Man kann daher die Normalverteilung verwenden, um die Binomialverteilung mit grossem n zu approximieren (denn die Binomialverteilung ist eine i.i.d. Summe von Bernoulli-Verteilungen). Man spricht dann von der sogenannten **Normalapproximation** der Binomialverteilung.

Wenn $X \sim \text{Bin}(n, p)$, dann haben wir $\mathbb{E}[X] = np$ und $\text{Var}(X) = np(1 - p)$. Für n gross können wir also X gemäss dem ZGWS approximativ als Normalverteilung mit Erwartungswert $\mu = np$ und Varianz $\sigma^2 = np(1 - p)$ behandeln. D.h. es gilt dann

$$\mathbb{P}(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1 - p)}}\right).$$

Manchmal verwendet man im Zähler noch Korrekturfaktoren (sogenannte Stetigkeitskorrekturen). Dies wollen wir hier nicht näher betrachten.

Beispiel. *Wie gross ist die Wahrscheinlichkeit, dass bei 1000 Würfeln mit einer Münze maximal 530 Mal Kopf erscheint?*

Die Anzahl Würfe X , bei denen Kopf erscheint, ist $\text{Bin}(1000, 0.5)$ -verteilt. Diese Verteilung approximieren wir mit einer Normalverteilung, d.h.

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

mit $\mu = 1000 \cdot 0.5 = 500$ und $\sigma^2 = 1000 \cdot 0.5 \cdot (1 - 0.5) = 250$.

Von Interesse ist

$$\mathbb{P}(X \leq 530) = \Phi\left(\frac{530 - 500}{\sqrt{250}}\right) = \Phi(1.897) \approx 0.97. \quad \triangleleft$$

Analog gilt für $X \sim \text{Pois}(\lambda)$ mit λ gross (was aufgefasst werden kann als i.i.d. Summe von vielen unabhängigen Poissonverteilungen mit kleinem λ)

$$\mathbb{P}(X \leq x) \approx \Phi\left(\frac{x - \lambda}{\sqrt{\lambda}}\right),$$

wobei wir genau gleich wie vorher vorgegangen sind.

Immer wenn also eine Zufallsvariable als eine Summe von vielen (unabhängigen) Effekten aufgefasst werden kann, ist sie wegen des Zentralen Grenzwertsatzes in erster Näherung normalverteilt. Das wichtigste Beispiel dafür sind Messfehler. Wenn sich die Effekte eher multiplizieren als addieren, kommt man entsprechend zur Lognormal-Verteilung.

5.4 Review / Lernziele



- ☐ Sie wissen, was unter der i.i.d. Annahme gemeint ist und wieso diese Annahme zentral für viele Berechnungen ist.
- ☐ Sie kennen das Gesetz der grossen Zahlen und wissen, dass die Standardabweichung des arithmetischen Mittels im i.i.d. Fall mit dem Faktor \sqrt{n} abnimmt (\sqrt{n} -Gesetz).
- ☐ Sie kennen den Zentralen Grenzwertsatz und können ihn auf passende Situationen anwenden.

Teil II

Schliessende Statistik

6 Parameterschätzungen

6.1 Einführung in die schliessende Statistik

6.1.1 Daten als Realisierungen von Zufallsvariablen

In der schliessenden Statistik wollen wir anhand von konkreten Daten (Beobachtungen) Aussagen über ein Wahrscheinlichkeitsmodell machen. Oder anders ausgedrückt: Wir haben ein paar wenige Beobachtungen und wollen Rückschlüsse über den zugrunde liegenden datengenerierenden Prozess ziehen. Dass man dies tun kann, scheint auf den ersten Blick erstaunlich. Man benutzt die induktive Logik, um probabilistische Aussagen (d.h. Aussagen, welche mit typischerweise hoher Wahrscheinlichkeit gelten) zu machen.

Grundlegend für die schliessende Statistik ist die Annahme, dass Daten Realisierungen von Zufallsvariablen sind. Das heisst: eine Beobachtung (oder “Messung”) x ist entstanden, indem eine Zufallsvariable “realisiert” wurde. Bei mehreren Daten geht alles analog: n Beobachtungen x_1, \dots, x_n werden aufgefasst als i.i.d. Realisierungen von Zufallsvariablen X_1, \dots, X_n , welche die Werte $X_i = x_i$, $i = 1, \dots, n$ angenommen haben.

6.1.2 Überblick über die Konzepte

Wir betrachten folgendes kleines Beispiel. Bei einer Stichprobe von 20 (unabhängigen) Bauteilen finden wir 5 mit einem Defekt. Dies können wir abstrakt als eine Realisierung einer Bin (n, p) -verteilten Zufallsvariablen X mit $n = 20$ auffassen. Wir möchten basierend auf unserer Beobachtung Rückschlüsse über den unbekannten Parameter p ziehen. Genauer geht es um folgende drei Fragestellungen:

- Welches ist der plausibelste Wert des unbekannten Parameters p ? (\rightsquigarrow Parameterschätzung)
- Ist ein bestimmter vorgegebener Parameterwert p_0 (z.B. ein Sollwert) mit der Beobachtung verträglich? (\rightsquigarrow statistischer Test)
- Was ist der Bereich von plausiblen Parameterwerten? (\rightsquigarrow Vertrauensintervall)

Für einen Parameterschätzer erscheint in diesem Beispiel intuitiv die beobachtete Ausfallhäufigkeit $\hat{p} = X/n$ sinnvoll zu sein. So lange die Daten nicht “realisiert” sind, sind die Schätzer also wiederum *Zufallsvariablen*. Die *realisierte* Schätzung ist dann $\hat{p} = x/n$, d.h. man ersetzt X durch dessen Realisierung x und erhält somit als (realisierten) Parameterschätzer $\hat{p} = 5/20$. Da wir nur wenige Daten haben, ist die realisierte Schätzung natürlich in der Regel *nicht* gleich dem unbekannten (“wahren”) Wert p , aber hoffentlich nahe daran.

Wenn wir allgemein einem Parameter “einen Hut aufsetzen”, bezeichnen wir damit den Schätzer für den entsprechenden Modellparameter (z.B. \hat{p} als Schätzer für p). Die Notation \hat{p} unterscheidet leider *nicht* zwischen dem Schätzer als Zufallsvariable und seinem realisierten Wert, welcher eine numerische Zahl ist.

Bei obigem Beispiel mit den Ausfällen der Bauteile war die Wahl der Verteilungsfamilie (Binomialverteilung) “klar” durch die Problemstellung. Insbesondere bei stetigen Zufallsvariablen ist dies in der Praxis nicht mehr so. Wir wollen uns darum jetzt zuerst der Frage widmen, wie man eine gute Verteilung für Messdaten finden bzw. verifizieren kann. Denn dies ist die Grundlage für alle weiterführenden Schritte.

6.2 Wahl der Verteilungsfamilie

Bis jetzt sind wir jeweils davon ausgegangen, dass wir die Verteilung (z.B. $\mathcal{N}(3, 2)$ -Verteilung) bzw. die Verteilungsfamilie (z.B. Normalverteilung) einer Zufallsvariable kennen. Damit haben wir dann diverse Wahrscheinlichkeiten und Kennzahlen etc. berechnet. In der Praxis ist dies leider nicht so. Basierend auf (wenigen) Daten müssen wir uns für eine Verteilung entscheiden, mit der wir etwas modellieren wollen. Nehmen wir also an, dass wir einen Datensatz x_1, \dots, x_n mit n Beobachtungen haben (“unsere n Messungen”).

Die Wahl einer Verteilungsfamilie kann einerseits durch Erfahrung (“was sich bisher bewährt hat”) oder aber auch durch physikalische Argumente (Summe von vielen Effekten sind z.B. gemäss ZGWS normalverteilt) geschehen. Ob eine Verteilungsfamilie zu einem konkreten Datensatz passt, kann man *qualitativ* gut mit grafischen Methoden überprüfen. Konzeptionell könnten wir z.B. schauen, wie gut das (normierte) Histogramm der Daten zur Dichte unserer Modellverteilung passt (z.B. eine bestimmte Normalverteilung). Es zeigt sich aber, dass man Abweichungen besser durch den Vergleich der Quantile erkennen kann.

Die Grundidee bei den QQ-Plots (Quantile-Quantile-Plots) besteht darin, dass wenn die Daten wirklich Realisierungen von einer entsprechenden Verteilung sind, die empirischen (d.h. aus unseren Daten berechneten) Quantile dann ungefähr den (theoretischen) Quantilen entsprechen sollten. Oder: “Was wir in den Daten sehen, soll ungefähr dem entsprechen, was wir vom Modell erwarten”.

Wir betrachten hierzu für

$$\alpha_k = \frac{k - 0.5}{n}, \quad k = 1, \dots, n$$

die entsprechenden empirischen $(\alpha_k \times 100)\%$ -Quantile.

Es gilt $\alpha_k \cdot n = k - 0.5$. Also ist die Beobachtung $x_{(k)}$ gerade das entsprechende empirische Quantil. Dies ist auch der Grund für die auf den ersten Blick spezielle Wahl von α_k . Das entsprechende (theoretische) $(\alpha_k \times 100)\%$ -Quantil ist gegeben durch $F^{-1}(\alpha_k)$, wobei F die kumulative Verteilungsfunktion unserer Modellverteilung ist. Für diverse α_k 's haben wir also die dazugehörigen empirischen und theoretischen Quantile. Falls die Daten wirklich von unserer Modellverteilung generiert wurden, sollten die empirischen Quantile ungefähr den theoretischen Quantilen entsprechen.

Der **QQ-Plot** besteht nun darin, dass wir die n Punkte

$$\{F^{-1}(\alpha_k), x_{(k)}\}, \quad k = 1, \dots, n$$

in einem Streudiagramm aufzeichnen (manchmal werden auch die Achsen vertauscht). Die Punkte sollten “in etwa” auf der Winkelhalbierenden liegen, falls das Modell stimmt. Beispiele wo dies der Fall ist, sieht man in Abbildung 6.1. Man kann also mit einem QQ-Plot Abweichungen der Daten von einer gewählten Modellverteilung *grafisch* überprüfen.

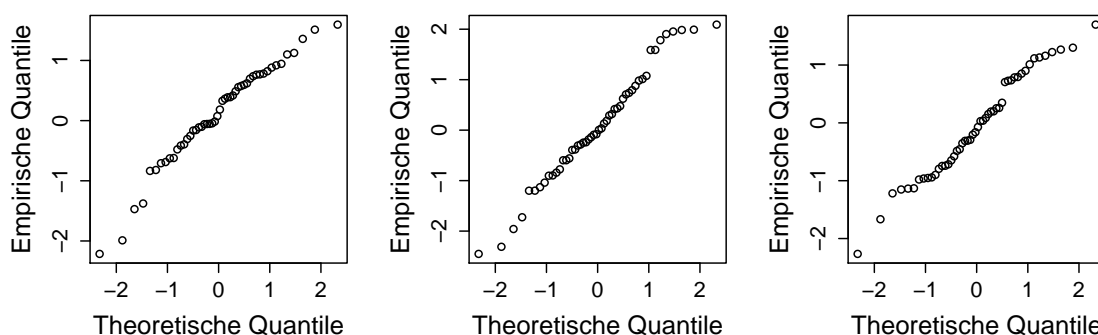


Abbildung 6.1: QQ-Plots von drei verschiedenen Datensätzen der Grösse $n = 50$.

Dieses Vorgehen hat den Nachteil, dass wir die Parameter der Modellverteilung eigentlich schon kennen müssen, sonst können wir die (theoretischen) Quantile ja gar nicht berechnen! Uns interessiert aber in der Regel zuerst die Frage, ob die Daten normalverteilt sind, und nicht, ob sie einer Normalverteilung mit *spezifischen* Parametern (μ, σ^2) folgen. Die Parameter müssen wir nämlich später noch schätzen.

Bei der Normalverteilung (und vielen anderen Verteilungen) können wir trotzdem das gleiche Vorgehen wie oben anwenden. Man verwendet hierzu einen sogenannten **Normalplot**. Dieser ist nichts anderes als ein QQ-Plot, bei dem die Modellverteilung F die Standardnormalverteilung $\mathcal{N}(0, 1)$ ist. Wenn die wahre Modellverteilung eine Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ ist, so liefert der Normalplot approximativ eine Gerade, welche jedoch im Allgemeinen *nicht* durch den Nullpunkt geht und *nicht* Steigung 45 Grad hat. Für $X \sim \mathcal{N}(\mu, \sigma^2)$ gilt nämlich, dass

$$q_\alpha = \mu + \sigma z_\alpha.$$

Falls die Daten tatsächlich von einer Normalverteilung stammen, liegen die empirischen Quantile also in etwa auf einer Geraden mit Achsenabschnitt μ und Steigung σ . Die ersten beiden Normalplots in Abbildung 6.2 zeigen in etwa eine Gerade, während dies für den letzten Plot nicht mehr zutrifft.

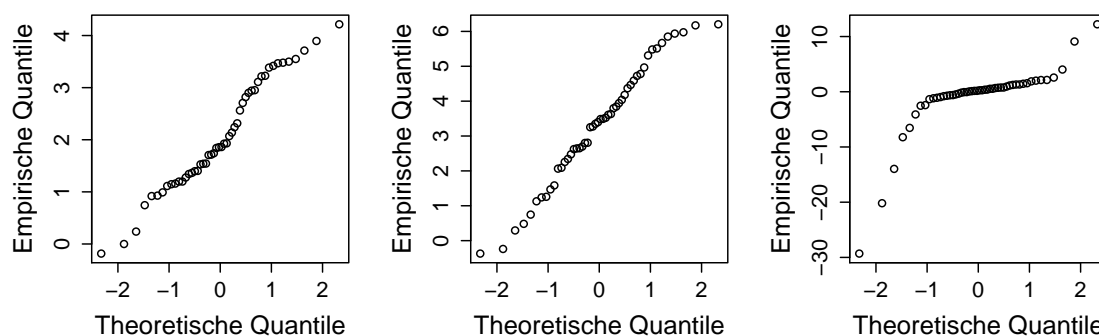


Abbildung 6.2: Normalplots von drei verschiedenen Datensätzen der Grösse $n = 50$.

Mit einem Normalplot können wir also überprüfen, ob eine Normalverteilung zur Modellierung unserer Daten geeignet ist, *ohne* uns um die Parameter μ und σ kümmern zu müssen. Es reicht, wenn man als Modellverteilung die Standardnormalverteilung verwendet. Die Punkte sollten dann in etwa auf einer (beliebigen) Geraden liegen.

Selbst wenn unser Modell stimmt, liegen die Punkte *nicht* exakt auf einer Geraden. Die Frage stellt sich, wie viel Abweichung noch “tolerierbar” ist, bzw. was wir mit dem Begriff “in etwa auf einer Geraden” meinen. Wir haben hierzu Datensätze von einer Standardnormalverteilung simuliert und die jeweiligen Normalplots gezeichnet, siehe Abbildung 6.3. In diesem Fall wissen wir, dass die Daten von einer Normalverteilung stammen (wir haben ja davon simuliert). Mit einer solchen Simulation können wir ein Gefühl dafür bekommen, wie gross die Abweichung von einer Geraden bei entsprechender Stichprobengrösse ist, wenn tatsächlich eine Normalverteilung vorliegt. Passt ein Normalplot nicht in das “Bild” der entsprechenden Stichprobengrösse, so müssen wir davon ausgehen, dass die Normalverteilung keine geeignete Verteilung ist.

Falls der Normalplot keine schöne Gerade zeigt, kann man trotzdem etwas über die zugrunde liegende Verteilung lernen. Verschiedene Situationen sind in Abbildung 6.4 dargestellt. Falls die empirischen Quantile in einem Bereich zwar auf einer Geraden liegen, dann im oberen Bereich nach oben und im unteren Bereich nach unten “davon wandern”, spricht man von einer sogenannten **langschwänzigen Verteilung** (verglichen mit einer Normalverteilung). Der QQ-Plot sieht dann aus wie ein “invertiertes” S. Das bedeutet, dass dann z.B. das empirische 99% Quantil viel grösser ist als man von der Normalverteilung erwarten würde. Oder anders ausgedrückt: die 1% grössten Werte der (empirischen) Verteilung sind grösser als von der Normalverteilung erwartet. Analog sieht es bei ganz kleinen

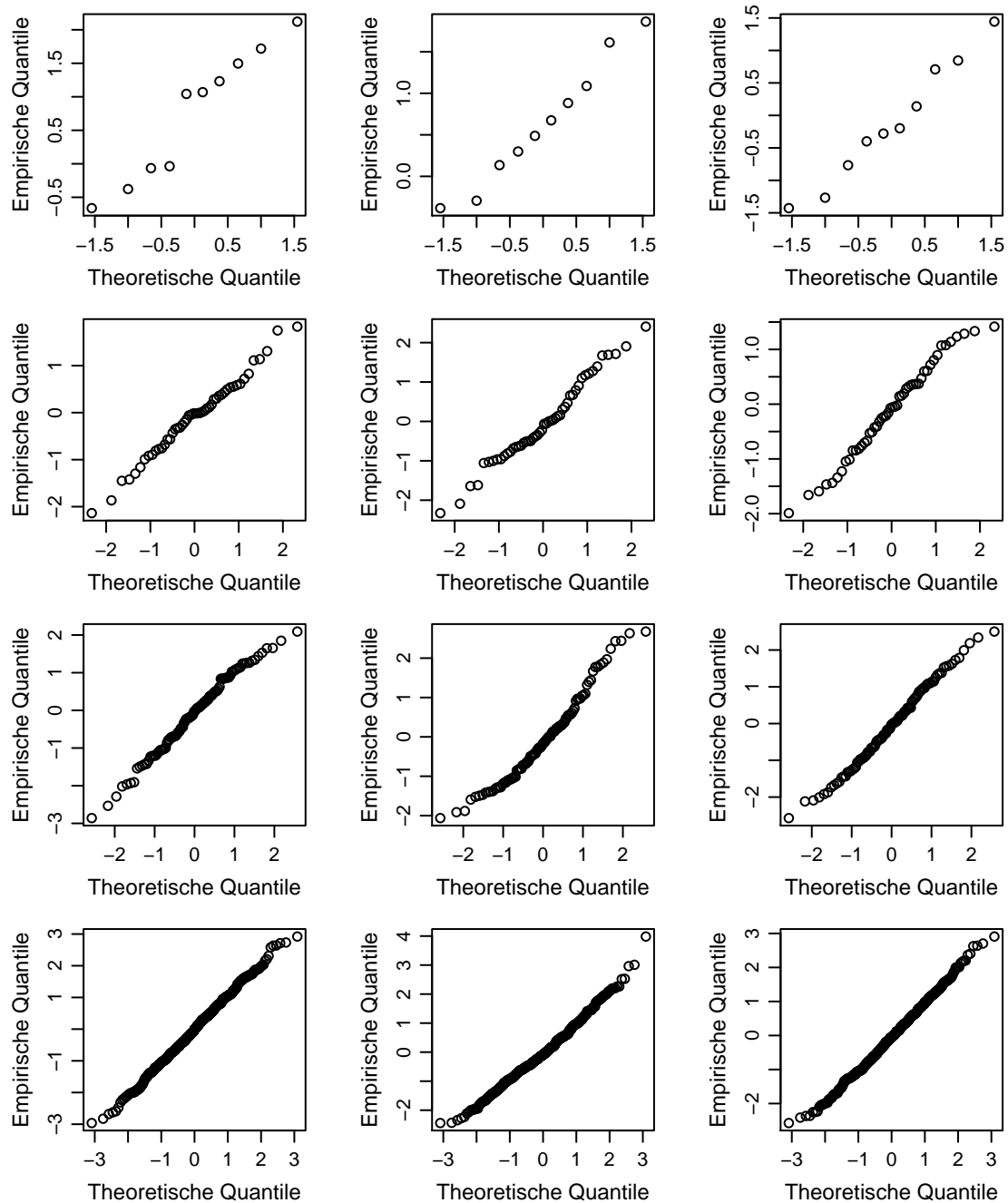


Abbildung 6.3: Normalplots von simulierten Standardnormalverteilungen für die Stichprobengrößen $n = 10, 50, 100, 500$ (oben nach unten).

Quantilen aus. Die Normalverteilung scheint nur für den mittleren Bereich passend zu sein und die empirische Verteilung hat die Tendenz, eher extreme Werte anzunehmen. Gerade umgekehrt ist es bei einer sogenannten **kurzschwänzigen** Verteilung. Der QQ-Plot zeigt dann eine “S-Form”. Schiefe Verteilungen zeigen sich durch “durchgebogene” Kurven.

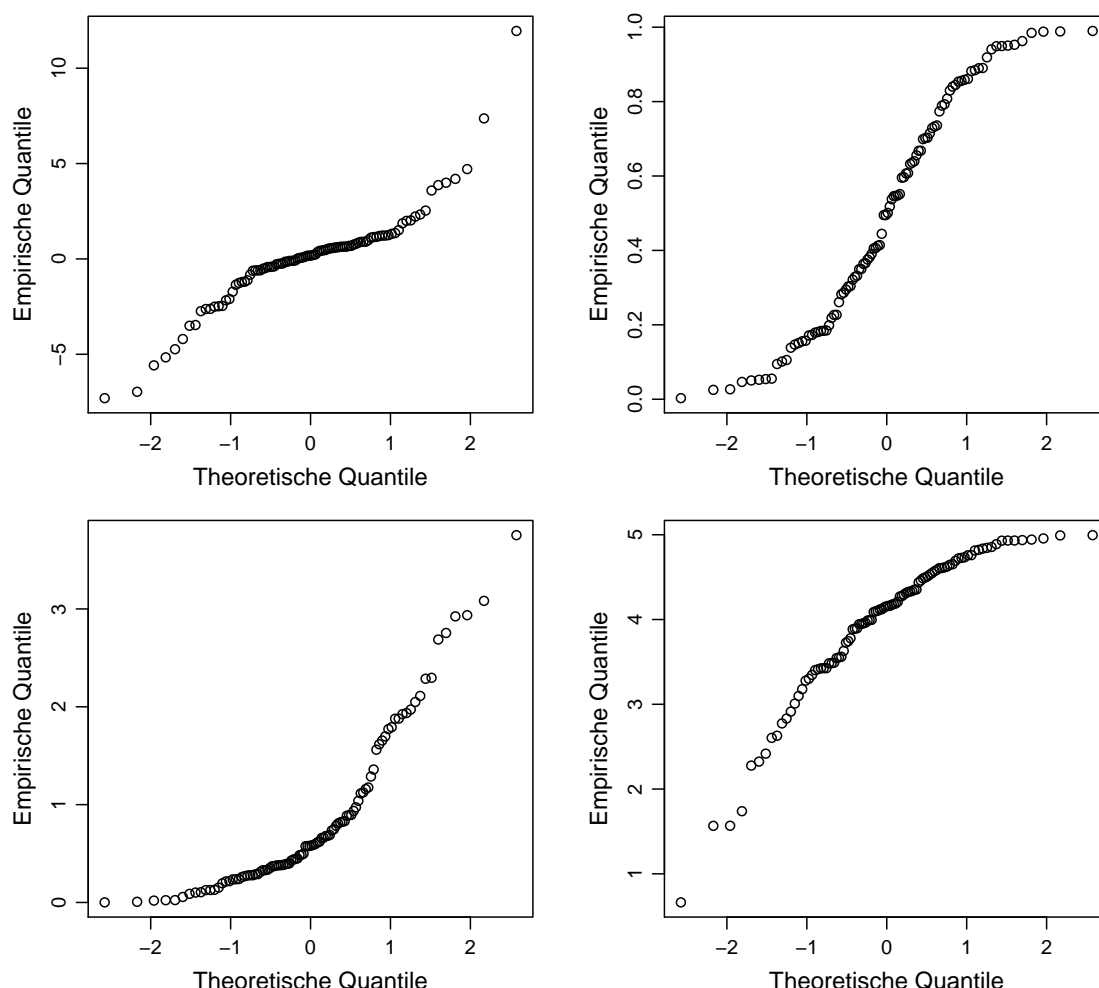


Abbildung 6.4: Beispiele für QQ-Plots von langschwänzigen (oben links), kurzschwänzigen (oben rechts) und schiefen Verteilungen (unten).

Hier sieht man auch den Vorteil von einem QQ-Plot: Die Abweichungen am Rand der Verteilung sind einfacher zu erkennen als z.B. bei einem Histogramm. Die entsprechenden Histogramme sind in [Abbildung 6.5](#) ersichtlich.

6.3 Methoden zur Parameterschätzung

Wir gehen nun davon aus, dass wir basierend auf unseren n i.i.d. Beobachtungen x_1, \dots, x_n die Verteilungsfamilie (z.B. Normalverteilung) gewählt haben. Nun müssen wir aber noch geeignete Parameter (bei der Normalverteilung μ und σ^2) finden.

Damit wir nicht für jede Verteilungsfamilie die Notation ändern müssen, schreiben wir ganz allgemein θ für den unbekannten Parameter, bzw. Parametervektor bei mehreren Parametern. Bei der Normalverteilung ist $\theta = (\mu, \sigma^2)$, bei der Poissonverteilung ist $\theta = \lambda$, etc.

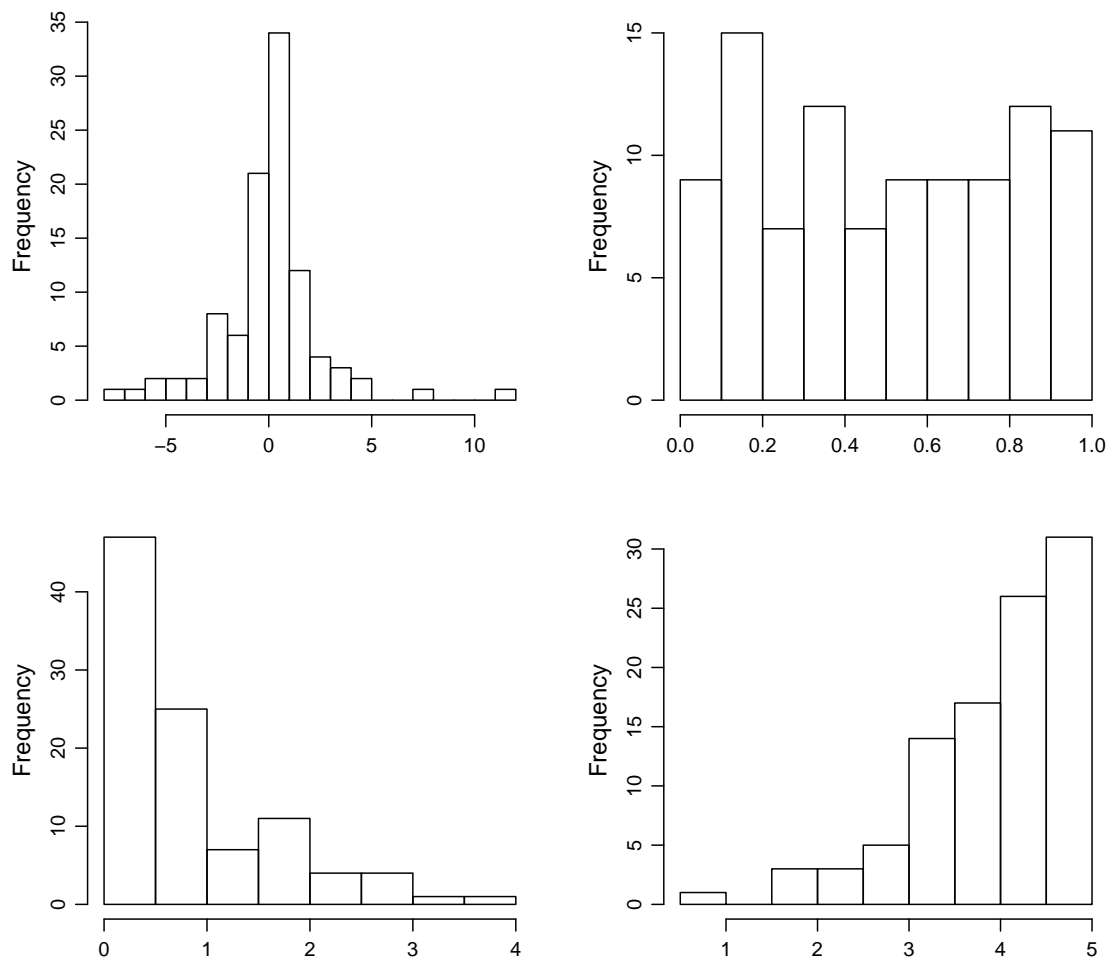


Abbildung 6.5: Histogramme der gleichen Datensätze wie in Abbildung 6.4.

Wir wollen also mit unseren Daten den unbekannten (fixen) Parameter θ schätzen. Ein Schätzer $\hat{\theta}$ nimmt unsere Daten und konstruiert damit einen möglichst “guten” Wert für den unbekannten Parameter θ . Oder formaler: Ein **Schätzer** für θ (zur Stichprobengrösse n) ist eine Funktion

$$\hat{\theta} : \mathbb{R}^n \rightarrow \mathbb{R},$$

d.h.

$$\hat{\theta} = \hat{\theta}(x_1, \dots, x_n),$$

bzw. als Zufallsvariable interpretiert

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n).$$

Man beachte, dass bei der Verwendung von griechischen Buchstaben keine Unterscheidung mehr zwischen Realisierung (Kleinbuchstaben) und Zufallsvariable (Grossbuchstaben) gemacht wird.

Schon jetzt wichtig zu wissen: Der gemäss unseren Daten berechnete Wert des Schätzers $\hat{\theta}$ entspricht in der Regel *nicht* exakt dem wahren Parameterwert θ . Wir werden aber später sehen, wie wir die Schätzungenauigkeit quantifizieren können.

Es stellt sich natürlich die Frage, wie man den Schätzer (eine Funktion) wählen soll. Hierzu gibt es verschiedene Ansätze. Die bekanntesten sind die *Momentenmethode* und die *Maximum-Likelihood Methode*.

6.3.1 Momentenmethode

Die Idee der Momentenmethode besteht darin, die Parameter so zu schätzen, dass gewisse Grössen aus dem (geschätzten) Modell mit den entsprechenden empirischen Grössen aus den Daten übereinstimmen. Die dabei verwendeten Grössen sind die sogenannten Momente.

Das k -te **Moment** von X ist definiert als

$$\mu_k = \mathbb{E}[X^k],$$

während das k -te **empirische Moment** von x_1, \dots, x_n gegeben ist durch

$$m_k = \frac{1}{n} \sum_{i=1}^n x_i^k.$$

Das erste Moment ist also gerade der Erwartungswert und das empirische Gegenstück ist das arithmetische Mittel.

Wir interessieren uns nun aber für die Parameterschätzer. Diese wählen wir so, dass die empirischen und die theoretischen Momente (gemäss unserem geschätzten Modell) übereinstimmen. Schauen wir uns dies einmal beispielhaft bei der Normalverteilung an. Dort haben wir

$$\begin{aligned}\mu_1 &= \mathbb{E}[X] = \mu \\ \mu_2 &= \mathbb{E}[X^2] = \sigma^2 + \mu^2.\end{aligned}$$

Wir sehen also insbesondere, dass wir die (theoretischen) Momente durch die Parameter ausdrücken können. Für eine Stichprobe mit $m_1 = 3.3$ und $m_2 = 11.9$ haben wir also das Gleichungssystem

$$\begin{aligned}m_1 &= 3.3 \stackrel{!}{=} \hat{\mu} \\ m_2 &= 11.9 \stackrel{!}{=} \hat{\sigma}^2 + \hat{\mu}^2.\end{aligned}$$

Aufgelöst nach $\hat{\mu}$ und $\hat{\sigma}^2$ erhalten wir als (realisierte) Parameterschätzer

$$\begin{aligned}\hat{\mu} &= 3.3 \\ \hat{\sigma}^2 &= 11.9 - (3.3)^2 = 1.01.\end{aligned}$$

Als Zufallsvariablen geschrieben haben wir

$$\begin{aligned}\hat{\mu} &= \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2.\end{aligned}$$

Etwas allgemeiner formuliert ist der Momentenschätzer $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_r)$ für den Parameter(vektor) $\theta = (\theta_1, \dots, \theta_r)$ definiert als die Lösung des Gleichungssystems

$$\begin{aligned}\mu_1(\hat{\theta}) &= m_1 \\ \mu_2(\hat{\theta}) &= m_2 \\ &\vdots \\ \mu_r(\hat{\theta}) &= m_r.\end{aligned}$$

Dabei schreiben wir $\mu_k(\hat{\theta})$, um zu verdeutlichen, dass das k -te Moment von $\hat{\theta}$ abhängt.

Beispiel. *Poissonverteilung*

Wir haben beobachtete Daten $x_1, \dots, x_5 : 3, 6, 4, 2, 3$ mit $\bar{x} = 3.6$, die wir als i.i.d. Realisierungen einer Poissonverteilung $X \sim \text{Pois}(\lambda)$ mit unbekanntem Parameter λ auffassen. Das erste Moment von X ist

$$\mu_1 = \mathbb{E}[X] = \lambda.$$

Dies setzen wir mit dem ersten empirischen Moment m_1 gleich, was uns direkt schon die Lösung liefert

$$\hat{\lambda} = m_1 = \frac{1}{n} \sum_{i=1}^n x_i = 3.6.$$

Als Zufallsvariable geschrieben erhält man einen Momentenschätzer für λ durch

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i. \quad \triangleleft$$

Der Momentenschätzer ist einfach, aber nicht immer die optimale Methode (im Sinne einer zu definierenden Genauigkeit für den unbekannten Parameter). Überdies ist der Momentenschätzer nicht unbedingt eindeutig und kann manchmal auch unsinnige Resultate liefern (falls z.B. der Parameterschätzer nicht mehr im Definitionsbereich liegt).

6.3.2 Maximum-Likelihood Methode

Die Idee der Maximum-Likelihood Methode besteht darin, die Parameter so zu schätzen, dass das beobachtete Ereignis (unsere Daten) möglichst plausibel erscheint. Oder anders herum: Wären Sie von ihrer Wahl der Parameter überzeugt, wenn jemand anderes eine Alternative vorschlägt, unter der die Daten wahrscheinlicher (und damit plausibler) sind?

Sei zunächst X diskret. Um die Abhängigkeit vom unbekannten Parameter θ zu betonen, bezeichnen wir die Wahrscheinlichkeitsfunktion $p_X(x)$ von X mit $p_X(x | \theta)$. Die Wahrscheinlichkeit, dass tatsächlich das Ereignis $\{X_1 = x_1, \dots, X_n = x_n\}$ (das wir beobachtet haben) eintritt, ist wegen der Unabhängigkeit gegeben durch

$$L(\theta) = p_{X_1, \dots, X_n}(x_1, \dots, x_n | \theta) = \prod_{i=1}^n p_X(x_i | \theta) = p_X(x_1 | \theta) \cdots p_X(x_n | \theta),$$

wenn wir annehmen, dass tatsächlich der Parameter θ gilt. Die Funktion $L(\theta)$ ist die sogenannte **Likelihoodfunktion** zur gegebenen Stichprobe x_1, \dots, x_n . Die Likelihoodfunktion ist also bei gegebenen Daten eine Funktion des Parameters θ . Das Wort Likelihood kann man z.B. mit Wahrscheinlichkeit oder aber auch mit "Mutmasslichkeit" übersetzen.

Die **Maximum-Likelihood Methode** besteht nun darin, diese Wahrscheinlichkeit zu *maximieren*, also jenen Parameter θ zu finden, für den die Wahrscheinlichkeit, dass die gegebene Stichprobe x_1, \dots, x_n eintritt, am grössten (maximal) ist. D.h. wir möchten θ so wählen, dass die Likelihood $L(\theta)$ maximal wird. Daher der Name "Maximum-Likelihood Methode".

Da der Logarithmus monoton wachsend ist, kann man äquivalent zu obiger Maximierungsaufgabe auch den Logarithmus der Likelihoodfunktion maximieren, was meist einfacher ist. Die **log-Likelihoodfunktion** ist definiert durch

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^n \log(p(x_i | \theta)) = \log(p_X(x_1 | \theta)) + \cdots + \log(p_X(x_n | \theta)),$$

wobei mit \log der *natürliche* Logarithmus gemeint ist. Die Maximierungsaufgabe löst man (in der Regel) wie aus der Analysis bekannt durch Ableiten nach dem Parameter θ und Nullsetzen. Um die zusätzliche Abhängigkeit von der Stichprobe x_1, \dots, x_n zu betonen, schreibt man auch

$$l(\theta; x_1, \dots, x_n) = \log(p_X(x_1 | \theta)) + \cdots + \log(p_X(x_n | \theta)).$$

Man muss dann die Gleichung

$$\frac{\partial}{\partial \theta} l(\theta; x_1, \dots, x_n) = 0$$

nach θ auflösen und erhält das Ergebnis

$$\hat{\theta} = \hat{\theta}(x_1, \dots, x_n),$$

bzw. als Zufallsvariable ausgedrückt

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$$

als allgemeinen Maximum-Likelihood Schätzer von θ zum Stichprobenumfang n .

Bemerkung:

Zusätzlich müssen wir (wie in der Analysis gelernt) noch die zweite Ableitung überprüfen, um sicherzustellen, dass wirklich ein Maximum und nicht ein Minimum vorliegt. Aus Platzgründen werden wir dies in der Regel weglassen, da es bei den verwendeten Funktionen oft intuitiv klar ist, dass es sich um ein Maximum handeln muss.

Beispiel. Sei $X \sim \text{Pois}(\lambda)$ mit unbekanntem Parameter λ , welchen wir mit der Maximum-Likelihood Methode schätzen wollen. Die zugehörige i.i.d. Stichprobe bezeichnen wir wie vorher mit x_1, \dots, x_n . Die Wahrscheinlichkeitsfunktion ist bei der Poissonverteilung gegeben durch

$$p_X(x | \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x \in \mathbb{N}_0.$$

Dies führt zur Likelihoodfunktion

$$L(\lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}.$$

Die log-Likelihoodfunktion ist somit

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^n \log \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) \\ &= \sum_{i=1}^n (x_i \log(\lambda) - \log(x_i!) - \lambda) \\ &= \sum_{i=1}^n (x_i \log(\lambda)) - n\lambda - \sum_{i=1}^n \log(x_i!). \end{aligned}$$

Leitet man $l(\lambda)$ nach λ ab und setzt $l'(\lambda) = 0$, so erhält man die Gleichung

$$\frac{1}{\lambda} \sum_{i=1}^n x_i - n = 0.$$

Dies führt zum Maximum-Likelihood Schätzer

$$\hat{\lambda} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Dies ist derselbe Schätzer wie bei der Momentenmethode. Die Maximum-Likelihood Methode liefert aber nicht zwangsläufig bei allen Verteilungen das gleiche Resultat wie die Momentenmethode. \triangleleft

Im stetigen Fall geht im Wesentlichen alles analog und man braucht nur den Buchstaben p durch f und “Wahrscheinlichkeitsfunktion” durch “Wahrscheinlichkeitsdichte” zu ersetzen. Es wird dann jener Parameter θ gesucht, für den die gemeinsame Dichte der X_1, \dots, X_n an der Stelle x_1, \dots, x_n am grössten ist.

Beispiel. Wir beobachten die Lebensdauern T_1, T_2, \dots, T_5 (in Wochen) von $n = 5$ Systemkomponenten. Als Modell verwenden wir eine Exponentialverteilung mit Parameter λ , d.h. T_i i.i.d. $\sim \text{Exp}(\lambda)$, $i = 1, \dots, n$. Der Parameter λ ist unbekannt. Die beobachteten Daten seien

i	1	2	3	4	5
t_i	1.2	3.4	10.6	5.8	0.9

Die Dichte der Exponentialverteilung ist gegeben durch

$$f_T(t | \lambda) = \begin{cases} 0 & t < 0 \\ \lambda e^{-\lambda t} & t \geq 0 \end{cases}$$

Die gemeinsame Dichte ist deshalb wegen der Unabhängigkeit gegeben durch

$$f_{T_1, \dots, T_n}(t_1, \dots, t_n | \lambda) = f_{T_1}(t_1 | \lambda) \cdots f_{T_n}(t_n | \lambda) = \begin{cases} \lambda^n e^{-\lambda \sum_{i=1}^n t_i} & \text{alle } t_i \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Dies ergibt die log-Likelihoodfunktion (die t_i 's sind hier per Definition grösser gleich Null)

$$l(\lambda) = n \log(\lambda) - \lambda \sum_{i=1}^n t_i.$$

Ableiten nach λ und Nullsetzen führt zur Gleichung

$$n \frac{1}{\lambda} - \sum_{i=1}^n t_i = 0.$$

Der Maximum-Likelihood Schätzer als Zufallsvariable geschrieben ist also

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^n T_i \right)^{-1}.$$

Der realisierte Wert ist dann hier

$$\hat{\lambda} = \frac{1}{4.38} = 0.228.$$

Auch hier würde die Momentenmethode das gleiche Resultat ergeben. ◁

Gegenüber der Momentenmethode hat die Maximum-Likelihood Methode den Vorteil, dass sie im Allgemeinen effizienter ist (genauere Schätzungen liefert) und zusätzlich mit weiteren theoretischen Überlegungen die Genauigkeit der Schätzer angegeben werden kann (Stichwort: Fisher-Information, was wir hier aber nicht behandeln).

6.3.3 Allgemeine Schätzer für Erwartungswert und Varianz

Wir können uns auch überlegen, ganz allgemein den Erwartungswert μ_X , bzw. die Varianz σ_X^2 (oder andere Kennzahlen) von einer Zufallsvariablen X basierend auf i.i.d. Beobachtungen x_1, \dots, x_n zu schätzen. Bei der Normalverteilung waren Erwartungswert und Varianz gerade die Parameter der Verteilung; bei anderen Verteilungen muss das nicht so sein (siehe z.B. die Exponentialverteilung). Hier wollen wir aber *keine* konkrete Verteilungsfamilie annehmen.

Bei der deskriptiven Statistik haben wir schon die empirischen Gegenstücke von Erwartungswert und Varianz kennengelernt. Es waren dies das arithmetische Mittel und die empirische Varianz. Diese

wollen wir gerade als Schätzer verwenden. Geschrieben als Zufallsvariablen haben wir

$$\hat{\mu}_X = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\hat{\sigma}_X^2 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

wobei die X_i i.i.d. sind mit der gleichen Verteilung wie X . Wir schreiben hier S_n^2 für die Streuung, obwohl wir S_n schon in Kapitel 5 angetroffen haben als die Variable für die Summe. Aus historischen Gründen hat man also den gleichen Buchstaben für verschiedene Dinge. Die Bedeutung sollte aber jeweils aus dem Kontext und durch das Quadrat klar sein.

Was haben diese Schätzer für Eigenschaften? Für den Schätzer des Erwartungswertes haben wir (nachrechnen!)

$$\mathbb{E}[\hat{\mu}_X] = \mathbb{E}[X] = \mu$$

$$\text{Var}(\hat{\mu}_X) = \frac{1}{n} \sigma_X^2.$$

Beim Schätzer für die Varianz ist es etwas komplizierter. Man kann zeigen, dass gilt

$$\mathbb{E}[\hat{\sigma}_X^2] = \sigma_X^2.$$

Im Erwartungswert ergeben unsere Schätzer also genau das Gewünschte, nämlich unsere gesuchten Kennzahlen. Man sagt auch, dass die Schätzer **erwartungstreu** seien. Dies ist auch der Grund für den Nenner $n-1$ bei der empirischen Varianz. Er sorgt gerade dafür, dass der Schätzer erwartungstreu wird für die Schätzung von σ_X^2 . Im Mittel machen wir also so keinen Fehler. Dies gilt *unabhängig* davon, was die zugrundeliegende Verteilung ist.

6.3.4 Genauigkeit von Schätzern – Ein erster Ansatz

Unsere Parameterschätzer liefern uns hoffentlich möglichst genaue Werte (d.h. Werte, die möglichst nahe bei den wahren aber unbekannten Parametern liegen).

Wenn wir aber ein Experiment wiederholen würden, dann würden wir (leicht) andere Daten und somit auch leicht andere Werte für die Parameterschätzer erhalten. Oder nach einem Zitat von John Tukey:

“The data could have been different”.

Wir sollten also nicht allzuviel Gewicht auf den konkreten Schätzwert (eine *einzelne* Zahl) legen, sondern versuchen zu quantifizieren, wie genau unsere Schätzung ist.

Hier kommen wieder unsere Modellannahmen zum Zuge. Wir betrachten dies hier kurz illustrativ an der Normalverteilung. Wir nehmen einmal an, dass unsere Daten x_1, \dots, x_n i.i.d. Realisierungen einer $\mathcal{N}(\mu, \sigma^2)$ -verteilten Zufallsvariablen sind. Als Schätzer für den Erwartungswert betrachten wir hier

$$\hat{\mu}_X = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Die Verteilung von unserem Schätzer (auch eine Zufallsvariable!) ist demnach

$$\hat{\mu}_X \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Der Schätzer fluktuiert also um den wahren Wert μ . Die Varianz wird mit grösser werdendem n kleiner und unsere Schätzung damit genauer. Die Standardabweichung des Schätzers wird allgemein auch als **Standardfehler** bezeichnet. Hier wäre also der Standardfehler von $\hat{\mu}_X$ gegeben durch σ/\sqrt{n} .

Wir nehmen hier vereinfachend einmal an, dass wir σ kennen. Mit Wahrscheinlichkeit 0.95 gilt dann, dass unser Schätzer $\hat{\mu}_X$ im Intervall

$$\mu \pm z_{0.975} \frac{\sigma}{\sqrt{n}}$$

liegt (nachrechnen!), wobei $z_{0.975}$ das 97.5%-Quantil der Standardnormalverteilung ist. Oder anders ausgedrückt: Mit Wahrscheinlichkeit 0.95 liegt unser Schätzer $\hat{\mu}_X$ weniger als $z_{0.975} \cdot \sigma/\sqrt{n}$ vom wahren Wert μ entfernt. Oder nochmals anders ausgedrückt: Mit Wahrscheinlichkeit 0.95 fangen wir den wahren Wert μ ein, wenn wir das (zufällige) Intervall

$$I = \hat{\mu}_X \pm z_{0.975} \frac{\sigma}{\sqrt{n}}$$

konstruieren (da der wahre Wert μ nicht zu weit von $\hat{\mu}_X$ entfernt sein kann). Dieses Intervall liefert uns also die möglichen “plausiblen” Werte für μ . Wenn das Intervall schmal ist, dann sind wir unserer Sache sicherer. Die Breite ist daher eine Angabe für die Genauigkeit unserer Schätzung. Man nennt ein solches Intervall auch ein 95%-**Vertrauensintervall** für den wahren Parameter μ . Wir werden später Techniken kennen lernen, wie man ein solches Intervall im Allgemeinen konstruieren kann.

Es ist $z_{0.975} = 1.96 \approx 2$. Dies führt zu

$$I \approx \hat{\mu}_X \pm 2 \frac{\sigma}{\sqrt{n}}$$

als 95%-Vertrauensintervall für μ . Die Merkregel ist: “Schätzung $\pm 2 \times$ Standardfehler”.

Erst dank unseren Modellannahmen können wir ein solches Vertrauensintervall konstruieren. Mit rein “numerischen Methoden” erhalten wir nur die Schätzung (eine *einzelne* Zahl) und können *nicht* angeben, wie genau diese ist. Mit den statistischen Methoden können wir also viel gewinnen! Alles steht und fällt natürlich mit den Modellannahmen (hier: Normalverteilung, bekannte Varianz). Diese muss man natürlich überprüfen, z.B. mit einem Normalplot.

6.4 Review / Lernziele



- ☐ Sie verstehen, wie ein QQ-Plot (bzw. Normalplot) aufgebaut ist und Sie wissen, wie man damit qualitativ abschätzen kann, wie gut eine Verteilung (bzw. Verteilungsfamilie) zu einem Datensatz passt.
- ☐ Sie können mit der Momenten- und der Maximum-Likelihood-Methode entsprechende Parameterschätzer herleiten (sowohl für stetige, wie auch für diskrete Verteilungen).
- ☐ Sie kennen die allgemeinen Schätzer für Erwartungswert und Varianz und deren Eigenschaften.
- ☐ Sie kennen die Begriffe Standardfehler und Vertrauensintervall.

7 Statistische Tests und Vertrauensintervalle für eine Stichprobe

Wir haben gesehen, wie man einen Parameter eines Modells aus Daten schätzen kann. Dies ist sozusagen der plausibelste Wert des Parameters. Oft muss man aber entscheiden, ob die Daten von einer Verteilung mit einem *bestimmten* Parameterwert generiert wurden oder nicht. Ist z.B. ein Sollwert oder ein Grenzwert bei einer Schadstoffkonzentration eingehalten? Oder handelt es sich wirklich um eine faire Münze?

Da der Parameterschätzer eine Ungenauigkeit aufweist (“the data could have been different”) können wir nicht einfach die Punktschätzung mit dem Sollwert vergleichen. Wenn wir die Ungenauigkeit aber quantifizieren können (siehe kurzes Beispiel mit Vertrauensintervallen in Kapitel 6.3.4), dann hilft uns dies schon weiter. Wir könnten z.B. schauen, ob der Sollwert in diesem Intervall enthalten ist oder nicht. Wie wir später sehen werden, ist dies genau der richtige Ansatz.

Wir wollen das Problem jetzt aber mit dem Konzept des statistischen Tests angehen. Später werden wir dann sehen, dass der Ansatz mit dem Vertrauensintervall genau äquivalent dazu ist.

7.1 Illustration der Konzepte mit der Binomialverteilung: Binomialtest

Beim Testproblem konzentrieren wir uns zuerst einmal auf die Binomialverteilung. Wir nehmen an, dass wir eine Beobachtung $x \in \{0, \dots, n\}$ haben, die wir als Realisierung einer $\text{Bin}(n, p)$ -verteilten Zufallsvariable X interpretieren, wobei n fix und bekannt ist, und p der unbekannte Parameter ist.

Diesen unbekannten Parameter wollen wir mit einem (von der Problemstellung abhängigen) Wert p_0 vergleichen (testen). Genauer gesagt wollen wir je nach Fragestellung überprüfen, ob sich entweder $p \neq p_0$, $p < p_0$ oder $p > p_0$ statistisch (aufgrund der gegebenen Beobachtung x) bestätigen lässt.

Man kann sich z.B. vorstellen, dass p_0 die von einem Hersteller angegebene Wahrscheinlichkeit ist, dass ein Bauteil Ausschussware ist, z.B. $p_0 = 0.1$. Sie vermuten aber, dass die Wahrscheinlichkeit grösser ist, d.h. dass gilt $p > p_0$. Sie wollen den Hersteller davon überzeugen, dass er zu schlechte Qualität liefert. Dieser ist natürlich gegenüber ihrer Behauptung sehr skeptisch.

Dazu nehmen wir zuerst an, es sei $p = p_0$ (d.h. wir nehmen die Position des Herstellers ein) und überprüfen, ob die Beobachtung x damit verträglich ist.

Die Annahme $p = p_0$ wird **Nullhypothese** genannt und man schreibt

$$H_0 : p = p_0.$$

Man spricht von einer Nullhypothese, weil man keine Abweichung vom Normal- oder Sollzustand hat.

Die entsprechende Vermutung (was wir persönlich zeigen wollen) wird **Alternativhypothese** genannt und mit H_A bezeichnet. Allgemein möglich sind

$$\begin{aligned} H_A : p &\neq p_0 \quad (\text{“zweiseitig”}) \\ p &> p_0 \quad (\text{“einseitig nach oben”}) \\ p &< p_0 \quad (\text{“einseitig nach unten”}). \end{aligned}$$

Wir beschränken uns nun auf den Fall $H_A : p > p_0$, d.h. wir sind nur an Abweichungen nach oben interessiert (wir wollen ja zeigen, dass die Qualität zu schlecht ist). Wir wollen den Hersteller davon überzeugen, dass seine Annahme *nicht* mit den beobachteten Daten verträglich ist. Qualitativ betrachtet scheint es plausibel, dass wir die Nullhypothese $H_0 : p = p_0$ verwerfen, wenn wir allzu viele defekte Bauteile finden. Falls die Beobachtung x also zu gross ist, d.h. $x \geq c$ für ein bestimmtes c , dann glauben wir der Nullhypothese nicht mehr. Würden Sie dem Hersteller z.B. noch glauben, wenn Sie in einer Stichprobe von 20 Bauteilen 5 mit einem Defekt finden? Bemerkung: Selbst wenn der Hersteller recht hat, kann es durchaus sein, dass wir in einer Stichprobe von 20 Bauteilen mehr als 10% defekte Teile finden. Die Frage ist nun, wieviele defekte Bauteile noch tolerierbar sind.

Da wir eine Entscheidung unter Unsicherheit treffen müssen, kann es sein, dass wir Fehlentscheide treffen. Der Zufall kann uns in die Quere kommen und dafür sorgen, dass obwohl H_0 stimmt, wir einen sehr grossen Wert für x beobachten (dies kommt vor, aber nur selten). In diesem Fall würden wir uns gemäss obigem Vorgehen aber *gegen* H_0 entscheiden. Man spricht von einem sogenannten **Fehler 1. Art**. Andererseits kann es sein, dass obwohl H_A stimmt, wir einen nicht allzu extremen Wert von x beobachten und wir uns *nicht* gegen H_0 entscheiden. Dann spricht man von einem sogenannten **Fehler 2. Art**. Die Fehlerarten bei einem statistischen Test sind in Tabelle 7.1 dargestellt.

		Entscheidung	
		H_0	H_A
Wahrheit	H_0	Kein Fehler	Fehler 1. Art
	H_A	Fehler 2. Art	Kein Fehler

Tabelle 7.1: Verschiedene Fehlerarten bei einem statistischen Test.

Wie findet man nun einen guten Wert für c ? Wir nehmen hierzu einmal an, dass die Nullhypothese stimmt; wir nehmen also die Position des Herstellers ein, der natürlich gegenüber unserer Behauptung immer noch skeptisch ist. Unter H_0 ist die Wahrscheinlichkeit, die Nullhypothese fälschlicherweise abzulehnen (d.h. die Wahrscheinlichkeit, einen Fehler 1. Art zu begehen) gegeben durch

$$\mathbb{P}_{p_0}(X \geq c) = \sum_{k=c}^n \binom{n}{k} p_0^k (1-p_0)^{n-k},$$

wobei wir mit dem Index p_0 bei $\mathbb{P}_{p_0}(X \geq c)$ nochmals betonen, dass wir die Wahrscheinlichkeit unter der Nullhypothese berechnen. Wir sollten also c nicht zu klein wählen, damit die Wahrscheinlichkeit für einen Fehler 1. Art nicht zu gross wird. Auf der anderen Seite möchten wir aber auch c nicht allzu gross wählen, weil wir sonst zu häufig einen Fehler 2. Art begehen: kein Verwerfen der Nullhypothese H_0 , obwohl sie falsch ist. Man schliesst einen Kompromiss, indem man das kleinste $c = c(\alpha)$ nimmt, so dass gilt

$$\mathbb{P}_{p_0}(X \geq c) \leq \alpha.$$

Dabei ist α eine im Voraus festgelegte (kleine) Zahl, das sogenannte **Signifikanzniveau** (oft kurz auch nur Niveau); typischerweise wählt man $\alpha = 0.05$ oder $\alpha = 0.01$. Obige Ungleichung besagt, dass die Wahrscheinlichkeit eines Fehlers 1. Art mit dem Signifikanzniveau α kontrolliert wird. Die unserer Behauptung gegenüber skeptisch eingestellte Person (hier: der Hersteller) kann also die “Spielregeln” definieren. Aus ihrer Perspektive wird nur mit Wahrscheinlichkeit α falsch entschieden.

Die Wahrscheinlichkeit für einen Fehler 2. Art ist *nicht* explizit kontrolliert, weil man nur einen Fehler direkt kontrollieren kann. Da man mit dem Test Skeptiker (bzgl. unserer Behauptung) überzeugen will, ist es wichtiger, den Fehler 1. Art zu kontrollieren (man versetzt sich sozusagen in ihre Lage, bzw. sie dürfen die “Spielregeln” definieren).

Nach obigen Überlegungen kommt man zum Rezept, dass H_0 verworfen wird, falls $x \geq c(\alpha)$. Wenn dies zutrifft, sagt man, dass man die Alternativhypothese statistisch nachgewiesen hat, und dass man

die Nullhypothese verwirft. Man sagt auch, dass die Abweichung von der Nullhypothese **signifikant** ist. Die Menge K aller Ausgänge, bei denen man H_0 zugunsten von H_A verwirft, wird **Verwerfungsbereich** genannt. Hier ist der Verwerfungsbereich K gegeben durch

$$K = \{c, c + 1, \dots, n\}.$$

Entsprechend nennt man die Werte, bei denen H_0 nicht verworfen wird, den **Annahmebereich**.

Falls wir die Nullhypothese *nicht* verwerfen können, ist das (leider) *kein* Nachweis für die Nullhypothese. Nehmen wir an, dass $p_0 = 0.1$ und wir $H_0 : p = p_0$ nicht verwerfen können. Dann bleibt $p = 0.1$ zwar ein plausibler Wert für den Parameter, aber z.B. $p = 0.11$ wäre wohl auch noch plausibel. Oder anders ausgedrückt: Nur weil wir keine Abweichung von p_0 nachweisen können, heisst dies leider noch lange nicht, dass keine Abweichung vorhanden ist! Oder besser in Englisch:

“Absence of evidence is not evidence of absence.”

Wir rechnen jetzt das kleine Beispiel einmal ganz durch.

Beispiel. Ein Hersteller von Bauteilen behauptet, dass (maximal) 10% der Teile Ausschussware sind. Sie sind aufgrund alten Beobachtungen skeptisch und vermuten, dass es mehr als 10% sind. Wir haben also

$$H_0 : p = p_0 = 0.1$$

und

$$H_A : p > 0.1.$$

In einer neuen Stichprobe von $n = 20$ Bauteilen finden wir $x = 5$ mit einem Defekt. Wir modellieren die Anzahl Bauteile X mit Defekt mit einer Binomialverteilung, d.h.

$$X \sim \text{Bin}(n, p), n = 20.$$

Unter der Nullhypothese $H_0 : p = p_0 = 0.1$ haben wir folgende Wahrscheinlichkeiten:

x	0	1	2	3	4	5	6	...
$\mathbb{P}_{p_0}(X = x)$	0.12	0.27	0.29	0.19	0.09	0.03	0.01	...
$\mathbb{P}_{p_0}(X \leq x)$	0.12	0.39	0.68	0.87	0.96	0.99	1.00	...

Es ist also

$$\begin{aligned}\mathbb{P}_{p_0}(X \geq 4) &= 1 - \mathbb{P}_{p_0}(X \leq 3) = 1 - 0.87 = 0.13 \\ \mathbb{P}_{p_0}(X \geq 5) &= 1 - \mathbb{P}_{p_0}(X \leq 4) = 1 - 0.96 = 0.04.\end{aligned}$$

Der Verwerfungsbereich K ist also auf dem 5%-Niveau gegeben durch

$$K = \{5, 6, 7, \dots, 20\}.$$

Unsere Beobachtung $x = 5$ liegt gerade knapp noch im Verwerfungsbereich. Also verwerfen wir H_0 und haben statistisch nachgewiesen, dass $p > 0.1$ gilt. Wir haben also eine signifikante Abweichung von der Nullhypothese nachweisen können. \triangleleft

Falls man nach Abweichungen nach unten interessiert ist, also $H_A : p < p_0$, geht alles analog. D.h. man sucht das grösste c , so dass gilt

$$\mathbb{P}_{p_0}(X \leq c) \leq \alpha.$$

Der Verwerfungsbereich ist dann

$$K = \{0, 1, 2, \dots, c\}.$$

Bei zweiseitiger Alternative $H_A : p \neq p_0$ verwerfen wir die Nullhypothese $H_0 : p = p_0$, wenn $x \leq c_1$ oder $x \geq c_2$. Hier wählt man c_1 möglichst gross und c_2 möglichst klein, so dass gilt

$$\mathbb{P}_{p_0}(X \leq c_1) = \sum_{k=0}^{c_1} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha/2 \quad \text{und} \quad \mathbb{P}_{p_0}(X \geq c_2) = \sum_{k=c_2}^n \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha/2.$$

Der Verwerfungsbereich ist dann gegeben durch die Vereinigung

$$K = \{0, 1, 2, \dots, c_1\} \cup \{c_2, c_2 + 1, \dots, n\}.$$

Wir schneiden also bei der Verteilung links und rechts (ca.) $\alpha/2$ ab, damit der Fehler 1. Art kontrolliert wird. Der Verwerfungsbereich hat also ganz allgemein nach Konstruktion die gleiche “Form” wie die Alternativhypothese H_A .

Der hier vorgestellte Test bei der Binomialverteilung wird auch als **Binomialtest** bezeichnet.

Zusammenfassung eines statistischen Tests

Die Durchführung eines statistischen Tests kann – zumindest teilweise – “rezeptartig” erfolgen.

1. Wähle ein geeignetes Modell für die Daten.
2. Lege die Nullhypothese $H_0 : \theta = \theta_0$ fest. θ bezeichnet hier allgemein einen Parameter in einem Modell.
3. Anhand der Problemstellung, spezifiziere die Alternative

$$\begin{aligned} H_A : \theta &\neq \theta_0 \quad (\text{“zweiseitig”}) \\ \theta &> \theta_0 \quad (\text{“einseitig nach oben”}) \\ \theta &< \theta_0 \quad (\text{“einseitig nach unten”}). \end{aligned}$$

4. Wähle das Signifikanzniveau α , typischerweise $\alpha = 0.05$ oder 0.01 .
5. Konstruiere den Verwerfungsbereich K für H_0 , so dass gilt

$$\mathbb{P}_{\theta_0}(\text{Fehler 1. Art}) \leq \alpha.$$

Die Form des Verwerfungsbereichs hängt ab von der Alternative H_A .

6. Erst jetzt: Betrachte, ob die Beobachtung x (oder eine Funktion von mehreren Beobachtungen) in den Verwerfungsbereich fällt. Falls ja, so verwerfe H_0 zugunsten von H_A . Man sagt dann auch, dass ein statistisch signifikantes Resultat vorliegt. Falls x nicht in den Verwerfungsbereich fällt, so belassen wir H_0 , was aber noch lange nicht heisst, dass deswegen H_0 statistisch nachgewiesen wurde (“absence of evidence is not evidence of absence”).

7.2 Tests für eine Stichprobe bei normalverteilten Daten

Wir betrachten hier die Situation, in der wir n voneinander unabhängige Beobachtungen x_1, \dots, x_n einer Zufallsvariable $X \sim \mathcal{N}(\mu, \sigma^2)$ haben. Als Beispiel kann man sich 10 Messungen einer Schadstoffkonzentration vorstellen.

Als Schätzer für die unbekannten Parameter der Normalverteilung betrachten wir die erwartungstreuen Schätzer

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \tag{7.1}$$

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu})^2. \tag{7.2}$$

Wir fixieren je nach Problemstellung ein $\mu_0 \in \mathbb{R}$ und wollen die Nullhypothese

$$H_0 : \mu = \mu_0$$

gegen eine der möglichen Alternativen

$$\begin{aligned} H_A : \mu &\neq \mu_0 \quad (\text{“zweiseitig”}) \\ \mu &> \mu_0 \quad (\text{“einseitig nach oben”}) \\ \mu &< \mu_0 \quad (\text{“einseitig nach unten”}) \end{aligned}$$

testen. Wir interessieren uns hier also für Tests bzgl. dem Erwartungswert und nicht bezüglich der Varianz. Dabei unterscheiden wir zwei Fälle: Bekannte Streuung σ , dann verwenden wir den sogenannten Z -Test, oder unbekannte Streuung σ (dann muss sie aus den beobachteten Daten geschätzt werden), in diesem Fall ergibt sich der sogenannte t -Test.

7.2.1 Z -Test (σ bekannt)

Wir nehmen hier an, dass σ bekannt ist. Von früher wissen wir, dass für die Verteilung des arithmetischen Mittels gilt

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n).$$

Wenn wir die Daten mitteln, dann haben wir also immer noch den gleichen Erwartungswert wie eine einzelne Messung, aber eine kleinere Varianz. Falls die Nullhypothese $H_0 : \mu = \mu_0$ stimmt, dann haben wir

$$\bar{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n).$$

Wenn der realisierte Wert \bar{x}_n also “allzu weit” von μ_0 entfernt ist, sollten wir die Nullhypothese verwerfen. Wir könnten jetzt wieder so vorgehen wie beim einführenden Beispiel, d.h. wir könnten die Quantile der $\mathcal{N}(\mu_0, \sigma^2/n)$ -Verteilung berechnen und damit je nach Alternative H_A den Verwerfungsbereich bestimmen.

Typischerweise geht man aber über zu der standardisierten **Teststatistik**

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{\text{beobachtet} - \text{erwartet}}{\text{Standardfehler}}.$$

Unter der Nullhypothese ist Z also $\mathcal{N}(0, 1)$ verteilt, denn wir verwenden hier gerade eine Standardisierung. Bemerkung: Eine Teststatistik ist nichts anders als eine (spezielle) Zufallsvariable, die dazu verwendet wird, die Testentscheidung zu treffen.

Für eine gegebene Realisierung

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$$

von Z lehnen wir je nach Alternative H_A die Nullhypothese $H_0 : \mu = \mu_0$ ab, falls

$$|z| \geq z_{1-\frac{\alpha}{2}} \iff z \in K = (-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, \infty) \quad \text{für } H_A : \mu \neq \mu_0 \quad (7.3)$$

$$z \geq z_{1-\alpha} \iff z \in K = [z_{1-\alpha}, \infty) \quad \text{für } H_A : \mu > \mu_0 \quad (7.4)$$

$$z \leq z_\alpha = -z_{1-\alpha} \iff z \in K = (-\infty, z_\alpha] = (-\infty, -z_{1-\alpha}] \quad \text{für } H_A : \mu < \mu_0 \quad (7.5)$$

Das Symbol K bezeichnet wieder den Verwerfungsbereich und z_α ist das $(\alpha \times 100)\%$ -Quantil der Standardnormalverteilung.

Die Begründung für (7.3) ist wie folgt. Die Teststatistik Z ist unter der Nullhypothese $\mathcal{N}(0, 1)$ -verteilt, woraus sich die Wahrscheinlichkeit für einen Fehler 1. Art, wie man aus Abbildung 7.1 erkennen kann, als

$$\mathbb{P}_{\mu_0}(|Z| \geq z_{1-\frac{\alpha}{2}}) = \alpha$$

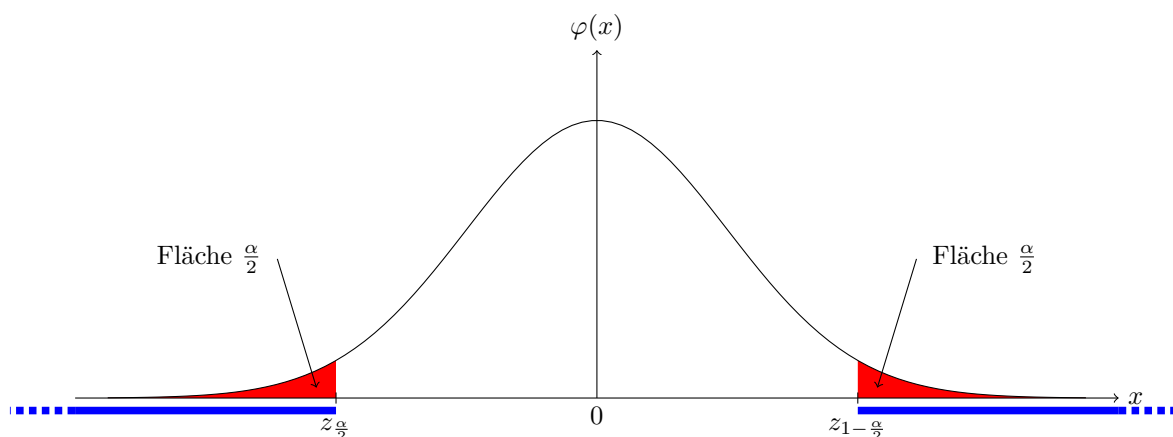


Abbildung 7.1: Dichtefunktion der Teststatistik Z mit Verwerfungsbereich (blau) des zweiseitigen Z -Tests zum Niveau α . Beachte $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$.

ergibt. Also genau so wie es sein sollte: Unter H_0 fallen wir nur mit Wahrscheinlichkeit α in den Verwerfungsbereich. Für (7.4) und (7.5) gehen die Überlegungen genau gleich.

Mit dem ursprünglichen Ansatz würden wir übrigens genau das Gleiche (d.h. den gleichen Testentscheid) erhalten, einfach auf der Skala von \bar{X}_n . D.h. wir verwerfen H_0 zugunsten von H_A , falls (nachrechnen!)

$$\begin{aligned} |\bar{X}_n - \mu_0| &\geq \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} && \text{für } H_A : \mu \neq \mu_0 \\ \bar{X}_n &\geq \mu_0 + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha} && \text{für } H_A : \mu > \mu_0 \\ \bar{X}_n &\leq \mu_0 + \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha} = \mu_0 - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha} && \text{für } H_A : \mu < \mu_0. \end{aligned}$$

7.2.2 t -Test (σ unbekannt)

Die Annahme, dass man die Standardabweichung σ kennt, ist in der Praxis meist unrealistisch. Wenn wir σ nicht kennen, ersetzen wir es durch den Schätzer S_n aus (7.2). Die Teststatistik ist dann

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} = \frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}}.$$

Da wir durch die Schätzung von σ eine zusätzliche “Variationsquelle” ins Spiel gebracht haben, wird die Streuung von T grösser sein als die Streuung von Z . Es liegt also daher unter der Nullhypothese sicher keine Standardnormalverteilung mehr vor für T .

Man kann zeigen, dass T unter H_0 einer sogenannten **t-Verteilung** mit $n - 1$ Freiheitsgraden folgt. Wir schreiben

$$T \sim t_{n-1}.$$

Die t -Verteilung ist wie die Standardnormalverteilung symmetrisch um 0, hat aber eher die Tendenz, (betragsmässig) grosse Werte anzunehmen. Man sagt auch, sie sei langschwänziger. Dies sieht man auch schön in Abbildung 7.2 im Vergleich mit der Standardnormalverteilung. Für $n \rightarrow \infty$ liegt eine Standardnormalverteilung vor. Die konkrete Dichte wollen wir hier nicht aufschreiben. Den Freiheitsgrad kann man sich als einen Parameter mit speziellem Namen vorstellen. Die Merkregel ist: Pro Beobachtung erhalten wir einen Freiheitsgrad, pro Parameter der uns interessiert, müssen wir einen bezahlen. Es verbleiben also hier $n - 1$ Freiheitsgrade, da wir n Beobachtungen haben und uns μ interessiert.

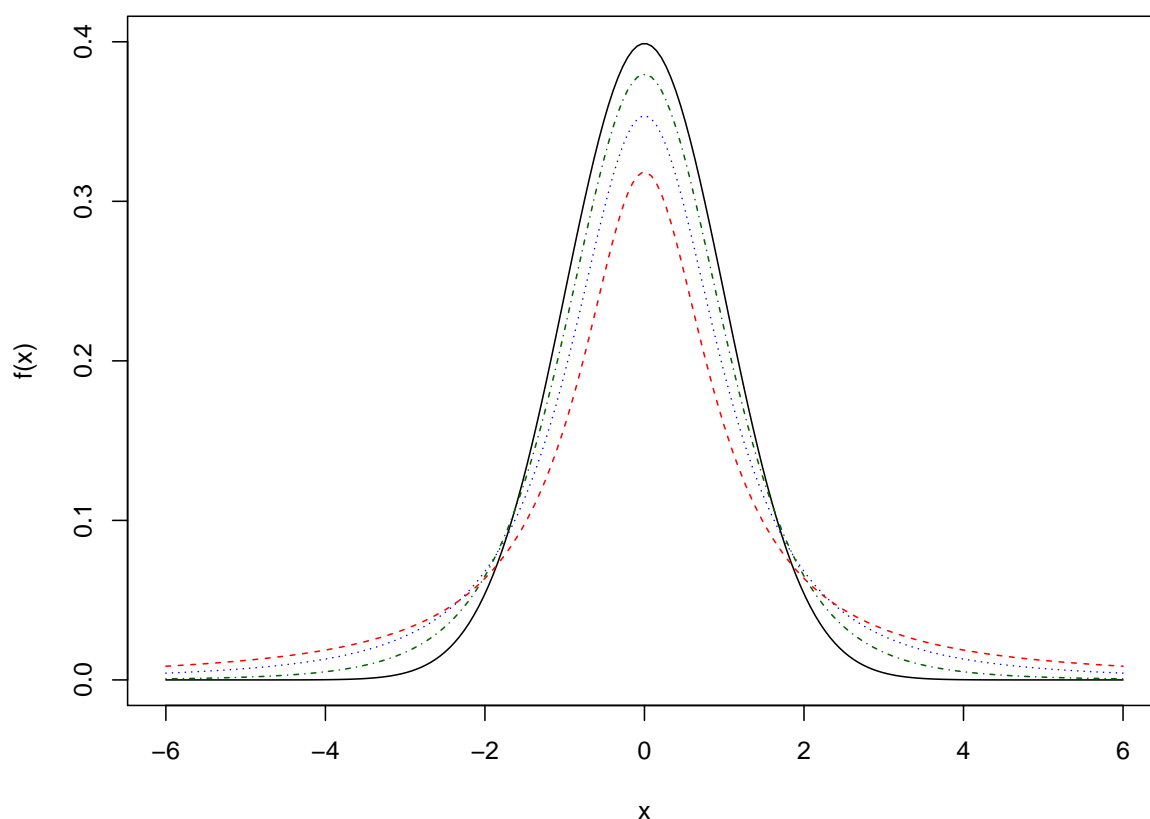


Abbildung 7.2: Dichten der t -Verteilung mit 1 (rot, gestrichelt), 2 (blau, gepunktet) und 5 (grün, strichpunktet) Freiheitsgraden. Die schwarze durchgezogene Kurve ist die Dichte der Standardnormalverteilung.

Das $(\alpha \times 100)\%$ -Quantil der t -Verteilung mit n Freiheitsgraden bezeichnen wir mit $t_{n,\alpha}$. Die Quantile sind tabelliert für kleine bis mittlere n und häufig gebrauchte Werte von α (siehe Anhang A.4), oder sie können mittels Computer numerisch berechnet werden. Für grosses n können wir auch auf die Quantile der Standardnormalverteilung zurückgreifen.

Analog wie beim Z -Test lehnen wir für eine gegebene Realisierung

$$t = \frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}}$$

von T je nach Alternative H_A die Nullhypothese $H_0 : \mu = \mu_0$ ab, falls

$$\begin{array}{lll} |t| \geq t_{n-1, 1-\frac{\alpha}{2}} & \iff & t \in K = (-\infty, -t_{n-1, 1-\frac{\alpha}{2}}] \cup [t_{n-1, 1-\frac{\alpha}{2}}, \infty) \quad \text{für } H_A : \mu \neq \mu_0 \\ t \geq t_{n-1, 1-\alpha} & \iff & t \in K = [t_{n-1, 1-\alpha}, \infty) \quad \text{für } H_A : \mu > \mu_0 \\ t \leq t_{n-1, \alpha} = -t_{n-1, 1-\alpha} & \iff & t \in K = (-\infty, t_{n-1, \alpha}] = (-\infty, -t_{n-1, 1-\alpha}] \quad \text{für } H_A : \mu < \mu_0 \end{array}$$

Beispiel. Der Sollwert einer Abfüllmaschine von Paketen beträgt 1000g. Sie haben die Vermutung, dass die Maschine falsch kalibriert ist. Als Modell für das Gewicht eines Paketes verwenden wir eine Normalverteilung. Zusätzlich nehmen wir an, dass die Gewichte der verschiedenen Pakete unabhängig voneinander sind. Wir betrachten in einer Stichprobe das Gewicht von 10 Paketen. D.h. wir haben Realisierungen x_1, \dots, x_{10} von X_1, \dots, X_{10} i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$.

Die gemessenen Werte (nicht dargestellt) liefern $\bar{x}_{10} = 1002.63$ und $s_{10} = 1.23$.

Gemäss Fragestellung ist

$$H_0 : \mu = \mu_0 = 1000,$$

$$H_A : \mu \neq 1000.$$

Unter H_0 folgt die Teststatistik T einer t_9 -Verteilung, da $n = 10$.

Auf dem 5%-Niveau ist der Verwerfungsbereich K gegeben durch die links- und rechtsseitigen 2.5% extremsten Werte der Verteilung von T unter H_0 , d.h.

$$K = (-\infty, -2.262] \cup [2.262, \infty),$$

da $t_{9,0.975} = 2.262$, siehe Tabelle in Anhang A.4.

Der realisierte Wert der Teststatistik ist

$$t = \frac{1002.63 - 1000}{1.23/\sqrt{10}} = 6.76.$$

Der realisierte Wert liegt also im Verwerfungsbereich, daher verwerfen wir die Nullhypothese. Wir haben somit statistisch nachgewiesen, dass die Maschine falsch kalibriert ist.

Wenn wir $\sigma = 1.23$ als bekannt vorausgesetzt hätten, dann wäre hier ein Z-Test angesagt gewesen. Der einzige Unterschied im Vorgehen wäre die Berechnung des Verwerfungsbereichs, der in diesem Fall dann gegeben wäre durch

$$K = (-\infty, -1.96] \cup [1.96, \infty),$$

da $z_{0.975} = 1.96$. ◁

Man kann zeigen, dass der t -Test der optimale Test (bzgl. der Macht, siehe nächstes Kapitel) unter allen möglichen Tests ist, falls die Beobachtungen normalverteilt sind. Bei nicht normalverteilten Beobachtungen können andere Tests (siehe Kapitel 7.5) sehr viel besser sein als der t -Test!

Wir haben nun sowohl im diskreten wie auch im stetigen Fall gesehen, wie man statistische Tests durchführen kann. Bevor wir noch weitere stetige Situationen anschauen, wollen wir uns in den nächsten Kapiteln allgemein mit den Eigenschaften und Besonderheiten von statistischen Tests befassen.

7.3 Allgemeine Eigenschaften von statistischen Tests

7.3.1 Macht

Ein statistischer Test kontrolliert per Konstruktion direkt die Wahrscheinlichkeit eines Fehlers 1. Art durch das Signifikanzniveau α :

$$\mathbb{P}(\text{Fehler 1. Art}) = \mathbb{P}(\text{Test verwirft } H_0, \text{ obwohl } H_0 \text{ stimmt}) \leq \alpha.$$

Bei stetigen Verteilungen ist obige Ungleichung eine Gleichung, da wir das Niveau *exakt* kontrollieren können.

Die Wahrscheinlichkeit eines Fehlers 2. Art ist hingegen eine Funktion des Parameterwerts $\theta \in H_A$, wir bezeichnen sie mit $\beta(\theta)$, d.h.

$$\beta(\theta) = \mathbb{P}(\text{Test akzeptiert } H_0, \text{ obwohl } \theta \in H_A \text{ stimmt}).$$

Die **Macht** (englisch “power”) eines Tests ist definiert als

$$1 - \beta(\theta) = \mathbb{P}(\text{Test verwirft richtigerweise } H_0 \text{ für } \theta \in H_A).$$

Die Macht können wir also nur unter einer entsprechenden Annahme für $\theta \in H_A$ berechnen.

Die Macht liefert uns die Antwort auf die Frage, wie wahrscheinlich es ist, die Alternative H_A nachzuweisen, wenn wir einen gewissen Parameterwert $\theta \in H_A$ annehmen. Während wir früher beim Signifikanzniveau den Standpunkt der Nullhypothese eingenommen haben, geht man bei der Macht sozusagen nun vom eigenen Standpunkt aus (denken Sie z.B. an das Beispiel mit den kaputten Bauteilen, dort war die Nullhypothese der Standpunkt des Herstellers). Die “Spielregeln” für den statistischen Test (d.h. der Verwerfungsbereich) werden unter H_0 bestimmt. Sie persönlich nehmen aber auch Teil am Spiel, glauben aber an ein spezifisches $\theta \in H_A$. Für Sie ist es natürlich von Interesse, wie wahrscheinlich es ist, dass Sie “gewinnen”, d.h. mit welcher Wahrscheinlichkeit die Nullhypothese verworfen wird.

Intuitiv scheint klar, dass je weiter weg das wahre θ von $H_0 : \theta = \theta_0$ in Richtung der Alternative liegt, desto wahrscheinlicher wird es sein, dass die Nullhypothese verworfen wird.

Beispiel. Wir betrachten 10 Würfe mit einer Münze. Die Wahrscheinlichkeit für Kopf sei p . Bei einer fairen Münze hätten wir $p = 0.5$. Wir vermuten, dass $p > 0.5$ ist. Also haben wir

$$H_0 : p = p_0 = 0.5,$$

$$H_A : p > 0.5.$$

Wenn wir mit X die Anzahl Würfe mit Ausgang Kopf bezeichnen, so haben wir unter H_0 , dass $X \sim \text{Bin}(10, 0.5)$ verteilt ist. Dies ergibt folgende Wahrscheinlichkeiten:

x	0	1	2	3	4	5	6	7	8	9	10
$\mathbb{P}_{p_0}(X = x)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

Der Verwerfungsbereich K ist also auf dem 5%-Niveau gegeben durch

$$K = \{9, 10\}.$$

Wir wollen nun die Macht des Tests berechnen für das Szenario $p = 0.75$. Dies ist die Wahrscheinlichkeit, dass wir H_0 verwerfen, wenn in der Tat $p = 0.75$ gilt. Unter der Annahme $p = 0.75$ ist $X \sim \text{Bin}(10, 0.75)$ -verteilt. Dies resultiert in folgenden Wahrscheinlichkeiten:

x	0	1	2	3	4	5	6	7	8	9	10
$\mathbb{P}_{p=0.75}(X = x)$	0.000	0.000	0.000	0.003	0.016	0.058	0.146	0.250	0.282	0.188	0.056

Die Macht entspricht nun der Wahrscheinlichkeit, in diesem Szenario in den Verwerfungsbereich K zu fallen. Der Verwerfungsbereich ändert sich nicht, denn dieser wird ja immer nur unter der Nullhypothese bestimmt! Also haben wir

$$\mathbb{P}_{p=0.75}(X \in K) = \mathbb{P}_{p=0.75}(X \geq 9) = 0.188 + 0.056 = 0.244.$$

Unser “Gedankenexperiment” liefert also folgendes Resultat: Wenn in Tat und Wahrheit $p = 0.75$ gilt, so werden wir (nur) mit Wahrscheinlichkeit 0.244 ein signifikantes Testresultat erhalten. Die Macht des Tests ist also 0.244 bei der Alternative $p = 0.75$. \triangleleft

Beim einseitigen Z-Test kann man die Macht schön illustrieren, siehe Abbildung 7.3. Man muss hierzu zwei Verteilungen betrachten. Auf der einen Seite die Verteilung unter der Nullhypothese; diese ist zentriert um μ_0 (z.B. ein Sollwert) und mit ihr wird der Verwerfungsbereich bestimmt. Auf der anderen Seite hat man die Verteilung unter einer Alternative μ (z.B. eine gewisse Überschreitung des Sollwertes). Die Frage ist dann, mit welcher Wahrscheinlichkeit wir in diesem Szenario die Überschreitung nachweisen können. Diese Wahrscheinlichkeit ist gegeben durch die entsprechend markierte Fläche unter der Dichte. Dies ist ganz analog zu obigem Beispiel. Dort hatte man Summen von Wahrscheinlichkeiten im entsprechenden Bereich. Je weiter weg wir μ von μ_0 platzieren, desto grösser wird die Macht.

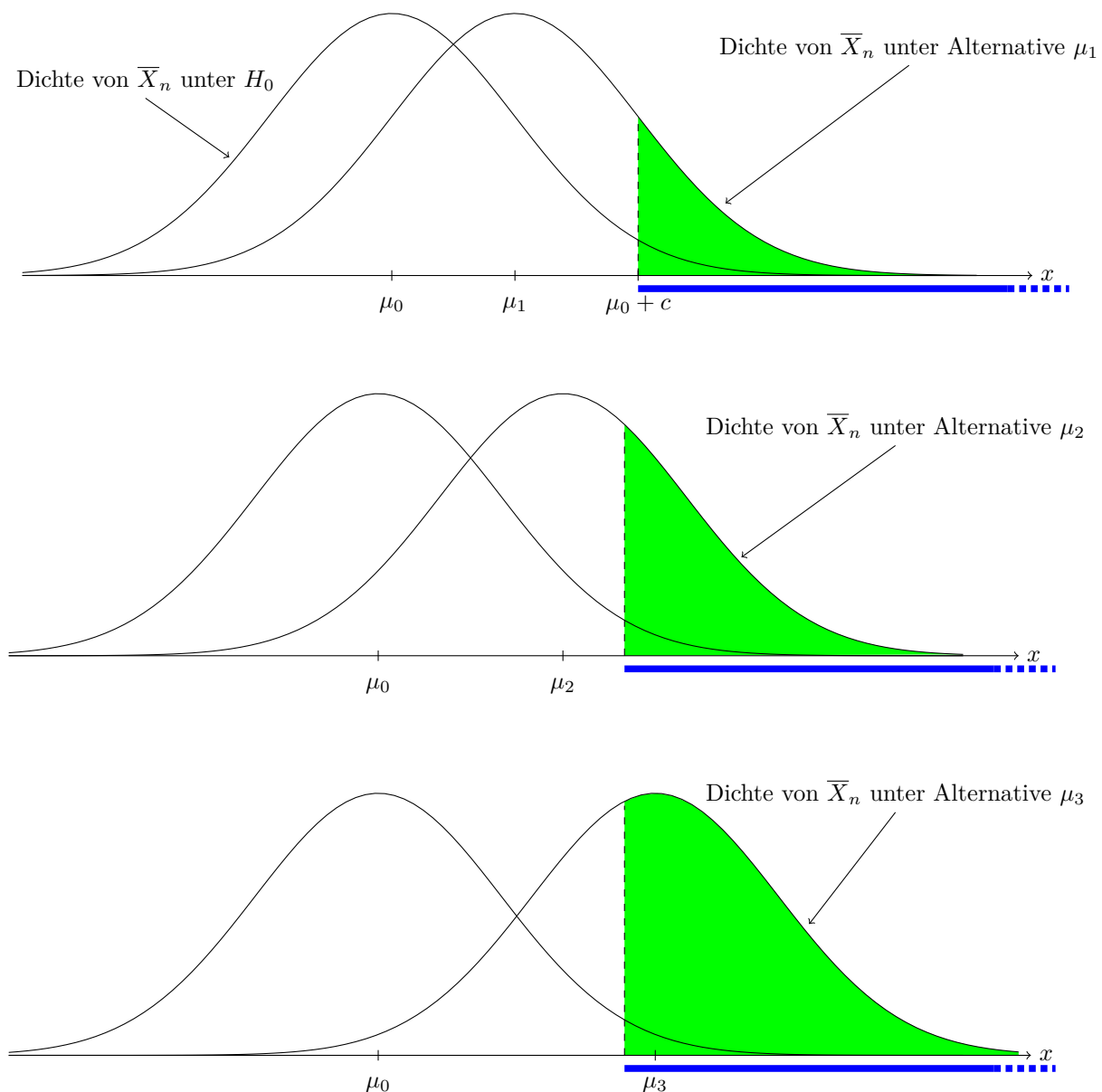


Abbildung 7.3: Illustration der Macht bei einem einseitigen Z-Test ($H_0 : \mu = \mu_0$, $H_A : \mu > \mu_0$) für drei verschiedene Szenarien. Die blaue Linie markiert den Verwerfungsbereich (bestimmt durch die Dichte von \bar{X}_n unter H_0 ; es ist $c = z_{1-\alpha} \cdot \sigma/\sqrt{n}$). Die grün markierte Fläche ist die Macht unter den entsprechenden Alternativen μ_1 bis μ_3 .

Die Berechnung der Macht ist also eine “theoretische” Angelegenheit, für die man *keine* Daten braucht, dafür eine Ahnung über mögliche Parameterwerte $\theta \in H_A$. Oft wird vor einem Experiment mittels obigen Überlegungen ermittelt, wie wahrscheinlich es ist, dass man mit dem Experiment einen gewissen Effekt nachweisen kann, d.h. wie gross die Macht für ein bestimmtes Szenario ist. Diese muss natürlich genügend gross sein, damit man ein Experiment durchführen wird (man will ja einen Erfolg sehen). Da die Macht mit zunehmender Stichprobengrösse grösser wird, kann man so auch ermitteln, wie gross die Stichprobe sein muss, dass z.B. die Macht mindestens 80% ist.

7.3.2 P-Wert

Bis jetzt haben wir jeweils basierend auf dem Signifikanzniveau α den Verwerfungsbereich ermittelt und dann geschaut, ob der realisierte Wert der Teststatistik in diesem Bereich liegt oder nicht. Den Verwerfungsbereich haben wir so konstruiert, dass er die $(\alpha \times 100)\%$ “extremsten” Werte der Teststatistik enthält (bzgl. der Verteilung unter der Nullhypothese). Beachten Sie: Das Signifikanzniveau und der Verwerfungsbereich hängen *nicht* von den beobachteten Daten ab, aber der realisierte Wert der Teststatistik sehr wohl.

Alternativ können wir auch versuchen, direkt zu schauen, wie “extrem” der beobachtete Wert der Teststatistik liegt (bzgl. der Verteilung unter der Nullhypothese). D.h. wir müssen dann den Umweg über den Verwerfungsbereich nicht machen. Dies führt uns zum sogenannten p-Wert.

Der **p-Wert** ist die Wahrscheinlichkeit, unter der Nullhypothese einen mindestens so extremen Wert der Teststatistik zu beobachten, wie der aktuell beobachtete. Dabei bestimmt die Alternativhypothese, was als extremer gilt (genau gleich wie bei der Form des Verwerfungsbereichs). Schauen wir uns dies an einem Beispiel an.

Beispiel. Bei einer Binomialverteilung mit $n = 10$ wollen wir die Nullhypothese

$$H_0 : p = p_0 = 0.5$$

gegen die Alternative

$$H_A : p > 0.5$$

testen (p ist z.B. die Wahrscheinlichkeit für Kopf bei einer Münze). Wir haben also unter H_0 die Zufallsvariable

$$X \sim \text{Bin}(10, 0.5)$$

(X ist dann die Anzahl Würfe mit Kopf bei insgesamt 10 Würfeln). Die Verteilung von X unter H_0 ist in Abbildung 7.4 dargestellt. Beobachtet wurde

$$x = 7.$$

Da $H_A : p > 0.5$ gilt, sind grosse Werte von X extrem im Sinne von H_A . Der p-Wert ist hier also die Summe aller Wahrscheinlichkeiten für X grösser gleich 7, d.h.

$$p\text{-Wert} = \mathbb{P}_{p_0}(X \geq 7) = 0.17.$$

Wenn wir zweiseitig testen würden ($H_A : p \neq 0.5$), wären sowohl sehr grosse als auch sehr kleine Werte von X extrem im Sinne von H_A . Dann müssten wir also die Wahrscheinlichkeiten “auf der anderen Seite” auch noch dazu addieren, d.h. wir hätten dann

$$p\text{-Wert} = \mathbb{P}_{p_0}(X \leq 3) + \mathbb{P}_{p_0}(X \geq 7) = 0.34.$$

als p-Wert. ◀

Wir können am p-Wert direkt den Testentscheid ablesen. Wenn der p-Wert kleiner als das Signifikanzniveau α ist, dann verwerfen wir die Nullhypothese, ansonsten nicht. Denn falls der p-Wert kleiner als α ist, dann liegt der beobachtete Wert der Teststatistik sicher im Verwerfungsbereich (zur Ermittlung des Verwerfungsbereichs verwendet man ja das gleiche “Schema” wie bei der Berechnung des p-Werts).

Bei obigem Beispiel mit der einseitigen Alternative würden wir also die Nullhypothese auf dem 5% Niveau nicht verwerfen, da $0.17 > 0.05$ gilt.

Schauen wir noch ein anderes Beispiel an.

Beispiel. Bei einem t-Test mit

$$H_0 : \mu = \mu_0$$

und

$$H_A : \mu \neq \mu_0$$

geht es konzeptionell genau gleich. Die Verteilung unter H_0 ist in Abbildung 7.5 dargestellt. Betrachten wir nun ein Beispiel, wo wir

$$t = 1.7$$

beobachtet haben (Daten nicht dargestellt). Statt Wahrscheinlichkeiten haben wir hier eine Dichte, die wir integrieren müssen. Der p -Wert ist gerade das Integral der Dichte über Werte kleiner als -1.7 bzw. grösser als 1.7 , d.h.

$$p\text{-Wert} = \mathbb{P}_{\mu_0}(T \leq -1.7) + \mathbb{P}_{\mu_0}(T \geq 1.7) = \mathbb{P}_{\mu_0}(|T| \geq 1.7).$$

Wenn wir einseitig testen würden, dann müsste man nur die Wahrscheinlichkeit betrachten, die “in Richtung der Alternative” liegt. \triangleleft

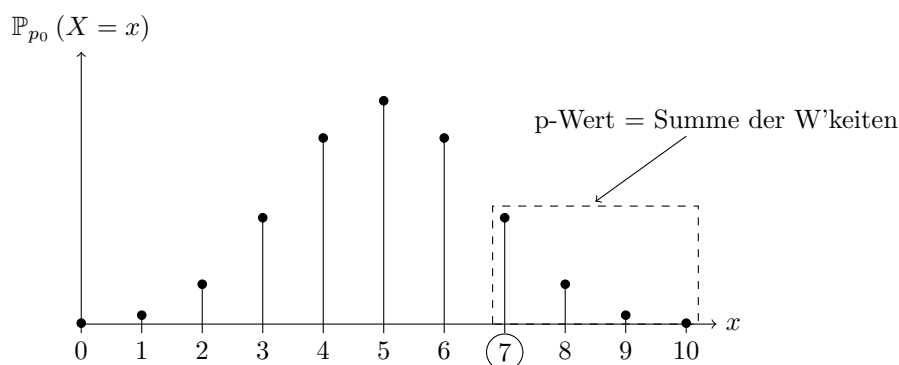


Abbildung 7.4: Illustration des p -Werts anhand einer $\text{Bin}(10, 0.5)$ -Verteilung unter $H_0 : p = 0.5$ und der Alternative $H_A : p > 0.5$. Beobachtet wurde $x = 7$. Der p -Wert ist die Summe der Wahrscheinlichkeiten für $x \geq 7$.

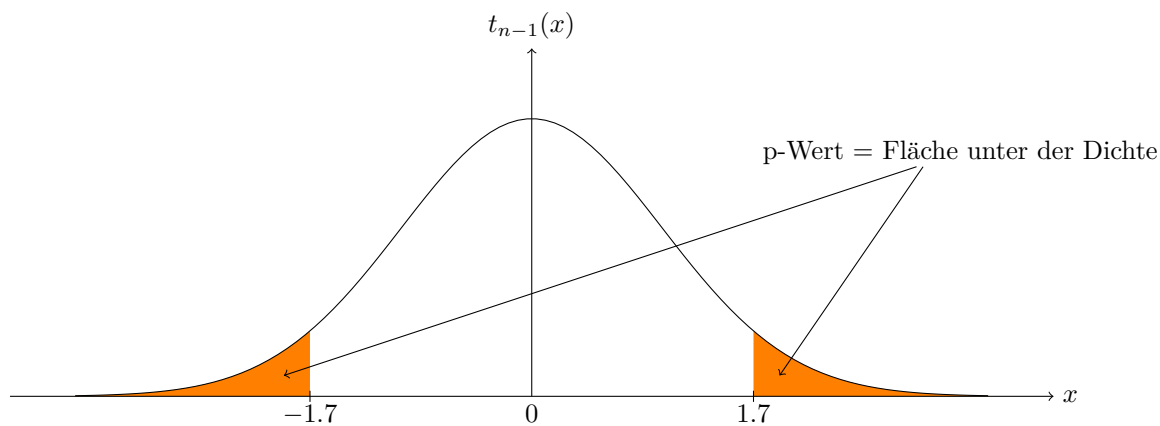


Abbildung 7.5: Illustration des p -Werts anhand eines zweiseitigen t -Tests. Beobachtet wurde $t = 1.7$. Der p -Wert ist die Fläche unter der Dichte bei “extremere” Fällen (d.h. Beobachtungen, die betragsmässig grösser als 1.7 sind).

Bei einer diskreten Verteilung ist der p -Wert also einfach die Summe der Wahrscheinlichkeiten (unter H_0) derjenigen Ausgänge, die mindestens so extrem sind (in Richtung der Alternative) wie der beobachtete Wert. Bei stetigen Verteilungen hat man einfach die entsprechenden Flächen unter der Dichte.

Der Vollständigkeit halber alles nochmals zusammengefasst.

Der p-Wert ist die Wahrscheinlichkeit, unter der Nullhypothese einen mindestens so extremen Wert der Teststatistik zu beobachten (in Richtung der Alternative), wie der aktuell beobachtete.

Man kann anhand des p-Werts direkt den Testentscheid ablesen: Wenn der p-Wert kleiner als das Signifikanzniveau α ist, so verwirft man H_0 , ansonsten nicht.

Verglichen mit dem reinen Testentscheid enthält der p-Wert aber *mehr Information*, da man direkt sieht, “wie stark” die Nullhypothese verworfen wird. Viele Computer-Pakete liefern den Testentscheid nur indirekt, indem der p-Wert ausgegeben wird. Man kann sich den p-Wert auch als “vollstandardisierte” Teststatistik vorstellen. Man kann alle Information ablesen und braucht keine Verteilungstabellen etc. mehr.

Bemerkung:

Da der p-Wert schlussendlich von den Daten abhängt, ist dieser auch zufällig. In der Tat kann einfach gezeigt werden, dass unter der Nullhypothese der p-Wert $\text{Uni}(0, 1)$ -verteilt ist.

Fehlinterpretationen und Gefahren des p-Werts

Der p-Wert wird oft falsch interpretiert. Der p-Wert ist insbesondere *nicht* die Wahrscheinlichkeit, dass die Nullhypothese stimmt (darüber können wir hier gar keine Aussagen machen, da die Parameter fix und nicht zufällig sind). Bei der Berechnung des p-Werts geht man davon aus, dass H_0 stimmt und schaut, wie extrem dann das beobachtete Ereignis liegt.

Zusätzlich bedeutet ein (sehr) kleiner p-Wert nicht zwangsläufig, dass ein fachlich relevantes Resultat gefunden wurde, da der p-Wert nichts über eine Effektgrösse aussagt. Hierzu werden wir später Vertrauensintervalle anschauen.

7.3.3 Multiples Testen

In der Praxis trifft man oft die Situation an, dass man nicht nur einen statistischen Test durchführt, sondern *mehrere*. Wir schreiben $H_{0,j}$ für die j -te Nullhypothese, $j = 1, \dots, m$. Mit m bezeichnen wir also die Anzahl Tests. Wenn wir annehmen, dass alle Nullhypothesen stimmen, und wir auf dem Signifikanzniveau α testen, dann erwarten wir in $(\alpha \times 100)\%$ der m Fälle, dass die Nullhypothese verworfen wird (oder äquivalent dazu: dass der p-Wert kleiner als α ist). Wenn wir genügend viele statistische Tests durchführen, werden wir also signifikante Resultate erhalten, selbst wenn alle Nullhypothesen stimmen.

Etwas genauer: Wenn wir annehmen, dass alle Nullhypothesen stimmen und alle Tests unabhängig voneinander sind, dann haben wir

$$\begin{aligned} \mathbb{P}(\text{Mindestens ein } H_{0,j} \text{ wird verworfen}) &= 1 - \mathbb{P}(\text{Kein } H_{0,j} \text{ wird verworfen}) \\ &= 1 - \mathbb{P}\left(\bigcap_{j=1}^m \{H_{0,j} \text{ wird nicht verworfen}\}\right) \\ &= 1 - \prod_{j=1}^m \mathbb{P}(H_{0,j} \text{ wird nicht verworfen}) \\ &= 1 - (1 - \alpha)^m. \end{aligned}$$

Für $\alpha = 0.05$ und $m = 50$ ist dies schon 0.92!

Für den allgemeineren Fall, wo wir keine Unabhängigkeit annehmen, haben wir die (grobe) Abschätzung

$$\begin{aligned}\mathbb{P}(\text{Mindestens ein } H_{0,j} \text{ wird verworfen}) &= \mathbb{P}\left(\bigcup_{j=1}^m \{H_{0,j} \text{ wird verworfen}\}\right) \\ &\stackrel{(1.3)}{\leq} \sum_{j=1}^m \mathbb{P}(H_{0,j} \text{ wird verworfen}) \\ &= \alpha \cdot m.\end{aligned}$$

Wenn wir also für jeden einzelnen Test das striktere Niveau $\alpha^* = \frac{\alpha}{m}$ verwenden, dann haben wir

$$\mathbb{P}(\text{Mindestens ein } H_{0,j} \text{ wird verworfen}) \leq \alpha.$$

Diese Korrektur nennt man auch **Bonferroni-Korrektur**. Sie ist sehr einfach und universell gültig. Der Nachteil ist, dass man ein sehr striktes Niveau verwenden muss und daher Macht verliert im Gegensatz zu anderen Korrektur-Methoden (die wir hier nicht anschauen).

In der Praxis sollte man also nur einen im Voraus definierten Test durchführen, oder falls man wirklich an mehreren Tests interessiert ist, eine entsprechende Korrektur-Methode anwenden.

7.4 Vertrauensintervalle

Wir haben bis jetzt gesehen, wie wir mit Parameterschätzern basierend auf Daten den plausibelsten Parameterwert berechnen können. Zusätzlich können wir mit statistischen Tests entscheiden, welche Parameterwerte sicher nicht mit den Daten verträglich sind (nämlich diejenigen, bei denen die Nullhypothese verworfen wird). Jetzt geht es noch darum, die Menge der plausiblen Parameterwerte zu ermitteln.

Ein **Vertrauensintervall** I für den Parameter θ zum Niveau $1 - \alpha$ (oft auch **Konfidenzintervall** genannt) besteht aus allen Parameterwerten, die im Sinne eines statistischen Tests zum Signifikanzniveau α mit der Beobachtung verträglich sind (üblicherweise nimmt man den zweiseitigen Test). Mathematisch heisst das:

$$I = \{\theta_0 : \text{Nullhypothese } H_0 : \theta = \theta_0 \text{ wird nicht verworfen}\}.$$

Das bedeutet also, dass wir sozusagen alle θ_0 “durchtesten” und diejenigen “sammeln”, bei denen die entsprechende Nullhypothese *nicht* verworfen wird (unsere Daten bleiben dabei natürlich *fix*).

Diese Beziehung stellt eine **Dualität zwischen Tests und Vertrauensintervall** dar. Wenn ein Wert θ_0 im Vertrauensintervall enthalten ist, so wissen wir, dass die entsprechende Nullhypothese *nicht* verworfen wird (sonst wäre ja der Wert nicht im Vertrauensintervall enthalten). Wir erhalten so also direkt den Testentscheid. Auf der anderen Seite kann mit Hilfe des Tests direkt das Vertrauensintervall konstruiert werden (gemäss Definition oben).

Das Vertrauensintervall ist *zufällig*, denn es hängt indirekt von unseren Beobachtungen ab, die wir als Realisierungen von *Zufallsvariablen* betrachten. Für andere Realisierungen werden wir also ein (leicht) anderes Vertrauensintervall erhalten!

Diese Überlegung führt zu einer **alternativen Interpretation**: Man kann zeigen, dass das Vertrauensintervall I den unbekannten wahren Parameter θ mit Wahrscheinlichkeit $1 - \alpha$ “einfängt”, d.h.

$$\mathbb{P}(I \ni \theta) = 1 - \alpha.$$

Hier ist I *zufällig* und θ *fix*, daher auch die etwas speziellere Schreibweise mit dem Symbol “ \ni ”, das wir mit “enthält” übersetzen. Damit haben wir auch in Kapitel 6.3.4 die Herleitung bestritten. Das heisst, wenn wir ein Experiment (oder eine Simulation) viele Male wiederholen, dann fängt das Vertrauensintervall den wahren (unbekannten) Parameter im Schnitt in $(1 - \alpha) \times 100\%$ der Fälle

ein. Dies ist in Abbildung 7.6 illustriert. Dort hat man den wahren Parameter in 3 von 100 Fällen “verpasst”.

Beide Sichtweisen führen dazu, dass wir das Vertrauensintervall für θ als denjenigen Wertebereich für unseren Modellparameter θ interpretieren, den wir aufgrund der vorliegenden Daten als plausibel betrachten.

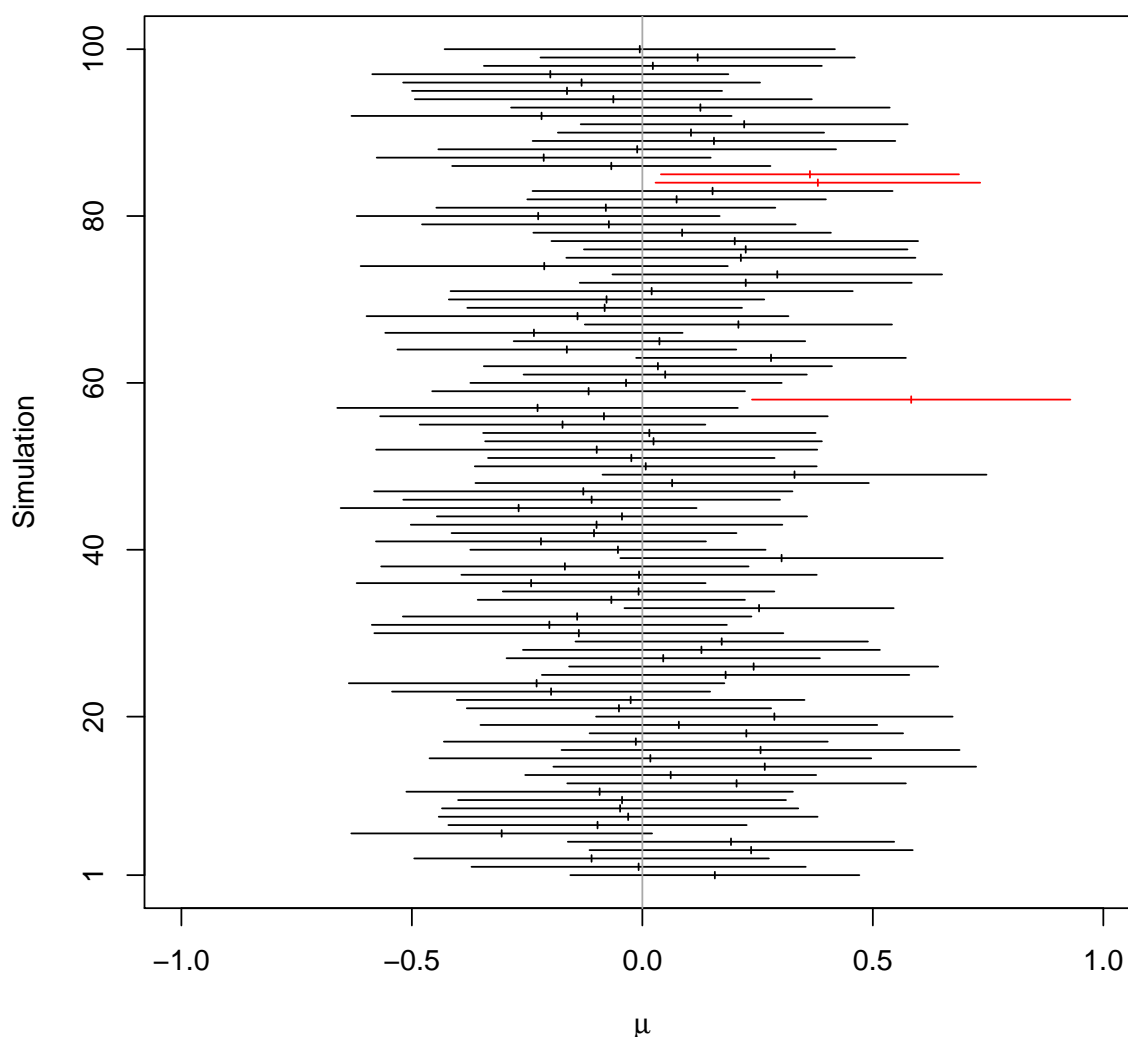


Abbildung 7.6: Illustration der Überdeckungswahrscheinlichkeit des 95%-Vertrauensintervalls für den Parameter μ einer Normalverteilung mit unbekannter Varianz. Es wurden 100 Datensätze simuliert mit wahren Parameter $\mu = 0$. Für jede Simulation ist das Vertrauensintervall mit einem horizontalen Strich dargestellt. Darin zusätzlich markiert ist der jeweilige Parameterschätzer. In 3 von 100 Fällen enthält das Vertrauensintervall für μ den wahren Wert $\mu = 0$ nicht (rot markiert).

Beispiel. Wir betrachten nochmals das Beispiel mit der Abfüllmaschine von Paketen. Wir hatten dort für $n = 10$ Beobachtungen die Kennzahlen $\bar{x}_{10} = 1002.63$ und $s_{10} = 1.23$. Jetzt wollen wir damit

ein 95%-Vertrauensintervall für μ bestimmen. Wir müssen also alle Nullhypothesen “sammeln”, die nicht verworfen werden. Diese sind gegeben durch die Menge

$$\begin{aligned} I &= \left\{ \mu_0 : \left| \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \right| \leq 2.262 \right\} \\ &= \left\{ \mu_0 : |\bar{X}_n - \mu_0| \leq \frac{S_n}{\sqrt{n}} \cdot 2.262 \right\} \\ &= \left\{ -\frac{S_n}{\sqrt{n}} \cdot 2.262 \leq \mu_0 - \bar{X}_n \leq \frac{S_n}{\sqrt{n}} \cdot 2.262 \right\} \\ &= \left\{ \bar{X}_n - \frac{S_n}{\sqrt{n}} \cdot 2.262 \leq \mu_0 \leq \bar{X}_n + \frac{S_n}{\sqrt{n}} \cdot 2.262 \right\} \\ &= \bar{X}_n \pm \frac{S_n}{\sqrt{n}} \cdot 2.262. \end{aligned}$$

Ganz genau genommen müsste man “<” statt “≤” verwenden in obigen Ungleichungen. Bei stetigen Verteilungen spielt dies aber keine Rolle und wir schreiben das Vertrauensintervall typischerweise als geschlossenes Intervall.

Wenn wir die beobachteten Werte einsetzen, so erhalten wir

$$1002.63 \pm \frac{1.23}{\sqrt{10}} \cdot 2.262 = [1001.75, 1003.51].$$

Plausible Parameterwerte für den Parameter μ liegen also zwischen 1001.75 und 1003.51. Wir sehen insbesondere auch, dass 1000 nicht im Vertrauensintervall enthalten ist. Das heisst, wir würden die entsprechende Nullhypothese verwerfen (wie wir das früher auch gemacht haben).

Wenn wir das Ganze mit einem Z-Test durchrechnen würden, dann würden wir genau das gleiche Resultat wie in Kapitel 6.3.4 erhalten (nachrechnen!). \triangleleft

Wie wir an obigem Beispiel sehen, ist in der Situation des t -Tests das (zweiseitige) $(1 - \alpha)\%$ -Vertrauensintervall für μ gegeben durch

$$I = \bar{X}_n \pm \frac{S_n}{\sqrt{n}} t_{n-1, 1-\frac{\alpha}{2}}.$$

Beim Z-Test erhält man entsprechend

$$I = \bar{X}_n \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}.$$

Die Form ist also genau gleich. Man verwendet einfach andere Quantile.

Ein Vertrauensintervall enthält sehr viel Information. Auf der einen Seite sehen wir automatisch, welche Nullhypothesen verworfen werden und welche nicht, auf der anderen Seite erhalten wir mit dem Vertrauensintervall auch eine Angabe über die **Genauigkeit** der Parameterschätzung (je schmaler, desto genauer).

Bezüglich Informationsgehalt können wir die Begriffe “Vertrauensintervall”, “P-Wert” und “Testentscheid” also “ordnen”:

$$\text{Testentscheid} \prec \text{P-Wert} \prec \text{Vertrauensintervall}$$

wobei sich die Relation “ \prec ” auf den Informationsgehalt bezieht.

Bemerkung:

Oft wird das Vertrauensintervall mit dem Annahmebereich eines Tests verwechselt. Beim Vertrauensintervall macht man basierend auf den vorliegenden Daten eine Aussage darüber, was plausible Werte

für einen Modellparameter sind. Beim Annahmebereich hingegen geht man von einer konkreten Nullhypothese aus und überlegt sich, in welchem Bereich dann die Teststatistik liegt (wozu man keine konkreten Daten braucht).

7.4.1 Statistische Signifikanz und fachliche Relevanz

Der Begriff der statistischen Signifikanz wird oft missbraucht, um gleichzeitig auch die entsprechende fachliche Relevanz zu untermauern. Diese beiden Begriffe müssen aber nicht unbedingt miteinander einhergehen. Wenn man genügend viele Beobachtungen sammelt, dann wird man *jede* Nullhypothese verwerfen können (denn diese stimmt in der Praxis nie exakt). Bedeutet dies nun, dass die vorgestellten Konzepte in der Praxis alle nutzlos sind? Die Antwort ist natürlich nein, aber man muss sie richtig verwenden.

Hierzu müssen wir das beste aus “beiden Welten” miteinander kombinieren: Entsprechendes Fachwissen und der statistische Output. Wir müssen zuerst basierend auf Fachwissen definieren, was ein **relevanter Effekt** oder Unterschied ist (die Statistik kann uns hier nicht helfen). Wenn wir dies gemacht haben, können wir die Statistik ins Spiel bringen.

Am Beispiel der Abfüllmaschine: Nehmen wir an, dass Abweichungen bis 5g vom Sollgewicht keine Rolle spielen, also nicht relevant sind. Wir haben also einen “irrelevanten Bereich”, der von 995 bis 1005g geht. Ausserhalb sprechen wir vom **Relevanzbereich**. Die Idee besteht nun darin, zu schauen, wie das Vertrauensintervall für μ liegt. Dieses war $[1001.75, 1003.51]$, was vollständig im irrelevanten Bereich liegt. Wir würden daher die Abweichung als statistisch signifikant, aber als *nicht* relevant taxieren. Andere mögliche Fälle und deren Interpretation sind in Abbildung 7.7 dargestellt.

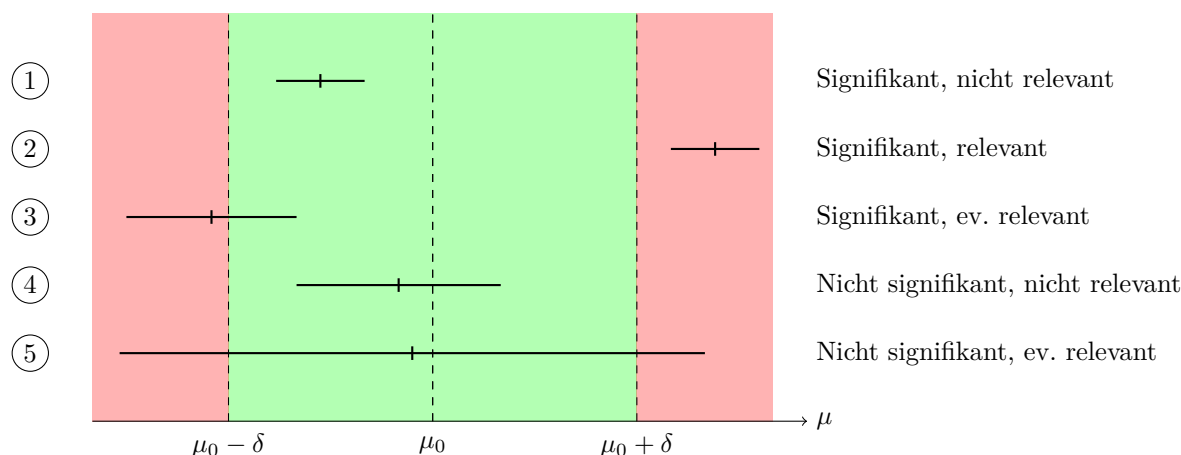


Abbildung 7.7: Verschiedene Fälle (1 bis 5) von statistischer Signifikanz und fachlicher Relevanz. Die Vertrauensintervalle für μ sind durch Striche dargestellt (Stahel, 2007). Der “irrelevante Bereich” geht von $\mu_0 - \delta$ bis zu $\mu_0 + \delta$ (grün), wobei das δ durch entsprechendes Fachwissen definiert wurde.

7.5 Tests für eine Stichprobe bei nicht normalverteilten Daten

In der Praxis sind unsere Messdaten nicht immer normalverteilt. Auch für diese Situationen gibt es entsprechende Tests mit weniger starken Annahmen.

7.5.1 Vorzeichen-Test

Wir gehen hier *nicht* mehr von einer Normalverteilung aus, sondern betrachten allgemeiner Beobachtungen x_1, \dots, x_n von i.i.d. Zufallsvariablen $X_1, \dots, X_n \sim \mathcal{F}(\mu)$, wobei \mathcal{F} eine beliebige stetige

Verteilung mit *Median* μ ist.

Wie vorher können wir wieder Null- und Alternativhypothesen aufstellen, jetzt aber bzgl. dem Median, z.B.

$$H_0 : \mu = \mu_0$$

und

$$H_A : \mu \neq \mu_0.$$

Bei symmetrischen Verteilungen entspricht der Median dem Erwartungswert und wir sind wieder in der Situation wie früher.

Wenn $\mu = \mu_0$ tatsächlich stimmt, dann beobachten wir mit 50% Wahrscheinlichkeit einen Wert grösser als μ_0 (gemäss der Definition des Medians). Wenn wir also allzu viele (oder zu wenige) Werte haben, die grösser als μ_0 sind, dann spricht dies gegen die Nullhypothese und für die Alternative. Wir können also durch reines Ermitteln der Anzahl Werte, die grösser als μ_0 sind, einen Testentscheid fällen. Dies ist die Idee des sogenannten Vorzeichen-Tests.

Beim **Vorzeichen-Test** betrachtet man die *Anzahl* positiver $X_i - \mu_0$ (man zählt also die Anzahl positiver Vorzeichen, daher auch der Name). Analog kann man natürlich einfach die Anzahl Werte grösser als μ_0 zählen. Die Anzahl positiver Vorzeichen, die wir fortan mit Q bezeichnen, folgt unter H_0 gemäss obigen Ausführungen einer $\text{Bin}(n, 0.5)$ -Verteilung. Damit kann man genau gleich wie beim einführenden Beispiel mit dem Binomialtest einen entsprechenden Test durchführen, siehe Kapitel 7.1.

Beispiel. *Wir betrachten nochmals das Beispiel mit der Abfüllmaschine. Um den Vorzeichen-Test durchführen zu können, reichen die beiden Kennzahlen \bar{x}_{10} bzw. s_{10} nicht, wir brauchen die Originaldaten. Diese sind in Tabelle 7.2 dargestellt.*

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1003.04	1004.10	1002.60	1002.50	1003.98	1003.01	1002.94	1001.89	999.69	1002.46

Tabelle 7.2: Originaldaten im Beispiel der Abfüllmaschine.

Wir sehen, dass wir 9 positive und 1 negatives Vorzeichen haben (9 Werte sind grösser als 1000 und nur 1 Wert ist kleiner). Wir wollen den Testentscheid nun mittels p -Wert bestimmen. Wenn wir zweiseitig testen, dann ist hier der p -Wert gegeben durch (nochmals zur Erinnerung: beim p -Wert des zweiseitigen Tests müssen wir die Wahrscheinlichkeiten von mindestens so extremen Ereignissen “auf beiden Seiten” berechnen)

$$\begin{aligned} p\text{-Wert} &= \mathbb{P}_{p=0.5}(Q = 0) + \mathbb{P}_{p=0.5}(Q = 1) + \mathbb{P}_{p=0.5}(Q = 9) + \mathbb{P}_{p=0.5}(Q = 10) \\ &= 2(\mathbb{P}_{p=0.5}(Q = 9) + \mathbb{P}_{p=0.5}(Q = 10)) \\ &= 2(10 \cdot 0.5^{10} + 0.5^{10}) \\ &= 0.0215, \end{aligned}$$

wobei wir ausgenutzt haben, dass die Verteilung von Q unter H_0 symmetrisch ist, da $Q \sim \text{Bin}(10, 0.5)$. Da der p -Wert kleiner als 5% ist, verwerfen wir also mit dem Vorzeichen-Test die Nullhypothese auf dem 5%-Niveau (wie wir das beim t -Test auch gemacht haben). \triangleleft

Bemerkung:

Falls es in der Praxis Beobachtungen gibt, die exakt mit μ_0 übereinstimmen, so lässt man diese weg und reduziert die Stichprobengrösse n entsprechend.

7.5.2 Wilcoxon-Test

Der **Wilcoxon-Test** ist ein Kompromiss, der weniger voraussetzt als der t -Test, aber die Information der Daten besser ausnützt als der Vorzeichen-Test.

Wir fassen hier unsere Beobachtungen x_1, \dots, x_n auf als i.i.d. Realisierungen von $X_1, \dots, X_n \sim \mathcal{F}(\mu)$, wobei $\mathcal{F}(\mu)$ eine symmetrische, stetige Verteilung mit Erwartungswert (bzw. Median) μ ist.

Wie früher haben wir also die Nullhypothese

$$H_0 : \mu = \mu_0$$

und z.B. die zweiseitige Alternative

$$H_A : \mu \neq \mu_0.$$

Zur Konstruktion einer Teststatistik ermitteln wir zuerst, wie stark die Daten betragsmässig von μ_0 abweichen. Wir betrachten also die Grössen $|X_i - \mu_0|$. Diese Werte ersetzen wir nun durch die entsprechenden Ränge. Zur Erinnerung: Dabei bedeutet

$$\text{Rang}(|X_i - \mu_0|) = k,$$

dass $|X_i - \mu_0|$ den k -ten kleinsten Wert hat unter allen $|X_1 - \mu_0|, \dots, |X_n - \mu_0|$. Wenn einzelne Werte zusammenfallen, erhalten die entsprechenden Beobachtungen den mittleren zugehörigen Rang.

Ferner sei V_i der Indikator dafür, ob $X_i - \mu_0$ positiv ist, d.h. $V_i = 1$ falls $X_i > \mu_0$ ist und $V_i = 0$ sonst. Schlussendlich verwenden wir als Teststatistik

$$W = \sum_{i=1}^n \text{Rang}(|X_i - \mu_0|) V_i,$$

d.h. wir betrachten nur die Ränge auf einer “Seite” von μ_0 . Unter H_0 erwarten wir “eine gute Mischung” von Rängen auf beiden Seiten von μ_0 . Wir verwerfen also H_0 , falls W allzu gross oder allzu klein ist (je nach Form der Alternative). Die konkreten Schranken für zu gross oder zu klein entnimmt man aus Tabellen, siehe z.B. Tabelle 7.4. In der Tat ist die Verteilung unter H_0 nichts anderes als dass wir die (fixen) Zahlen $1, \dots, n$ durchgehen (die Ränge) und jeweils eine Münze werfen, ob der entsprechende Rang “links” oder “rechts” von μ_0 liegt. Alternativ und einfacher liest man den Testentscheid direkt am p-Werts des Computer-Outputs ab.

Man kann zeigen, dass dieser Test das Niveau exakt einhält (d.h. die Wahrscheinlichkeit für einen Fehler 1. Art ist gleich α), wenn die X_i i.i.d. sind und eine um μ_0 symmetrische Dichte haben. Beim t -Test wird zwar das Niveau auch ungefähr eingehalten, falls die Daten nicht normalverteilt sind (wegen dem zentralen Grenzwertsatz), aber die Wahrscheinlichkeit eines Fehlers 2. Art ist in solchen Fällen unter Umständen beim t -Test *viel grösser* als beim Wilcoxon-Test.

In der Praxis ist der Wilcoxon-Test allermeist dem t - oder Vorzeichen-Test vorzuziehen. Nur falls die Daten sehr gut mit einer Normalverteilung beschrieben werden, ist der t -Test für eine gute Datenanalyse “vollumfänglich tauglich”: diese Annahme kann man z.B. mit dem Normalplot (siehe Kapitel 6.2) grafisch überprüfen.

Beispiel. Wir wollen zur Illustration die Teststatistik des Wilcoxon-Tests im Beispiel der Daten aus Tabelle 7.2 berechnen. Hierzu erstellen wir zuerst einmal ein “Inventar” über die benötigten Grössen, siehe Tabelle 7.3.

Der realisierte Wert der Teststatistik W ist also

$$W = 8 + 10 + 5 + 4 + 9 + 7 + 6 + 2 + 3 = 54.$$

Dieser liegt im Verwerfungsbereich, wenn wir zweiseitig auf dem 5%-Niveau testen. Ein Computer-Programm würde einen p -Wert von 0.004 liefern, was natürlich zum gleichen Testentscheid führt. \triangleleft

Bemerkung:

Auch beim Wilcoxon-Test gilt: Falls es in der Praxis Beobachtungen gibt, die exakt mit μ_0 übereinstimmen, so lässt man diese weg und reduziert die Stichprobengrösse n entsprechend.

k	x_k	$ x_k - \mu_0 $	$\text{Rang}(x_k - \mu_0)$	V_k
1	1003.04	3.04	8	1
2	1004.10	4.10	10	1
3	1002.60	2.60	5	1
4	1002.50	2.50	4	1
5	1003.98	3.98	9	1
6	1003.01	3.01	7	1
7	1002.94	2.94	6	1
8	1001.89	1.89	2	1
9	999.69	0.31	1	0
10	1002.46	2.46	3	1

Tabelle 7.3: Daten und entsprechende Ränge im Beispiel der Abfüllmaschine.

n	zweiseitig		einseitig	
	l	u	l	u
6	0	21	2	19
7	2	26	3	25
8	3	33	5	31
9	5	40	8	37
10	8	47	10	45
11	10	56	13	53
12	13	65	17	61
13	17	74	21	70
14	21	84	25	80
15	25	95	30	90
16	29	107	35	101
17	34	119	41	112
18	40	131	47	124
19	46	144	53	137
20	52	158	60	150
21	58	173	67	164
22	65	188	75	178
23	73	203	83	193
24	81	219	91	209
25	89	236	100	225
26	98	253	110	241
27	107	271	119	259
28	116	290	130	276
29	126	309	140	295
30	137	328	151	314

Tabelle 7.4: Kritische Grenzen beim Wilcoxon-Test für das 5%-Niveau. Für den zweiseitigen Test ist der Verwerfungsbereich gegeben durch $K = \{W \leq l\} \cup \{W \geq u\}$. Bei einem einseitigen Test verwendet man die entsprechenden Werte in der Spalte “einseitig”.

7.6 Rückblickender Überblick über Konzepte

7.6.1 Vorgehen und Fragen bei statistischen Tests

Das Vorgehen bei einem statistischen Test und die wichtigsten Fragen sind in Abbildung 7.8 und Tabelle 7.5 nochmals dargestellt. Ferner findet man in Abbildung 7.9 und 7.10 nochmals eine Übersicht über die wichtigsten Tests.

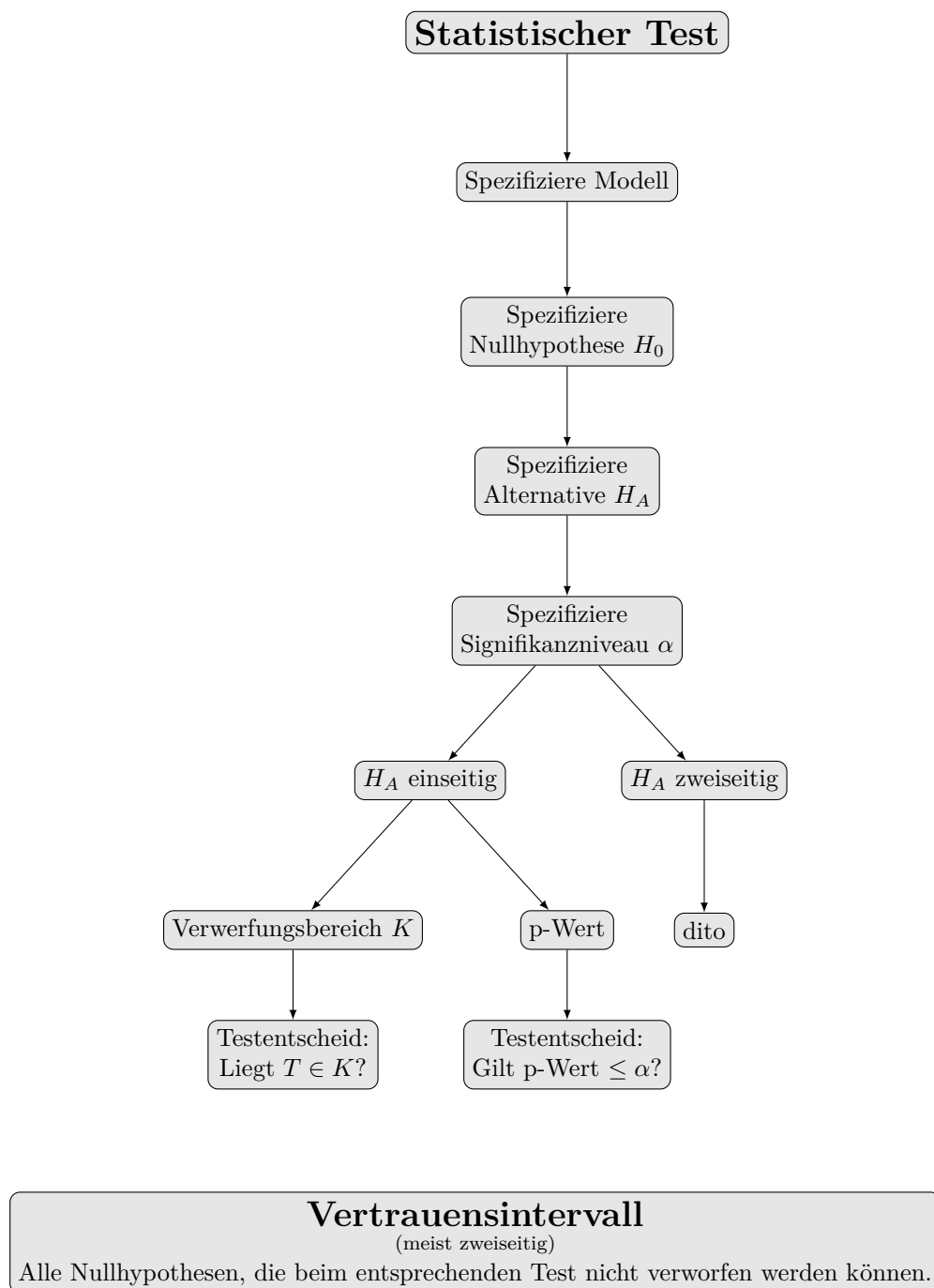


Abbildung 7.8: Vorgehen bei statistischen Tests und Zusammenhang mit dem Vertrauensintervall.

Frage	Stichwort	Berechnung / Antwort
Welche Werte sind aufgrund der vorliegenden Daten plausibel für den Modellparameter θ ?	Vertrauensintervall für θ	Alle Nullhypothesen, die nicht verworfen werden können.
Liegt ein signifikantes Testresultat vor? D.h. können wir die Nullhypothese verwerfen?	Testentscheid	Verwerfe die Nullhypothese, falls <ul style="list-style-type: none"> • Teststatistik liegt im Verwerfungsbereich • p-Wert ist kleiner als Signifikanzniveau • Vertrauensintervall enthält Nullhypothese nicht (alle äquivalent)
Wie wahrscheinlich ist es, dass wir ein signifikantes Testresultat erhalten, wenn H_0 stimmt?	Fehler 1. Art	Automatisch kontrolliert durch die Wahl des Signifikanzniveaus.
Wie wahrscheinlich ist es, dass wir kein signifikantes Testresultat erhalten, wenn H_0 nicht stimmt?	Fehler 2. Art	Berechne $\mathbb{P}_{\theta_A}(T \notin K)$ für ein $\theta_A \in H_A$. Der Parameter θ_A muss gewählt werden.
Wie wahrscheinlich ist es, dass wir ein signifikantes Testresultat erhalten, wenn H_0 nicht stimmt?	Macht $= 1 - \mathbb{P}(\text{Fehler 2. Art})$	Berechne $\mathbb{P}_{\theta_A}(T \in K)$ für ein $\theta_A \in H_A$. Der Parameter θ_A muss gewählt werden.
Ist ein signifikanter Effekt auch relevant?	Relevanzbereich	Betrachte, wie das Vertrauensintervall bzgl. dem Relevanzbereich liegt. Der Relevanzbereich wird durch Fachwissen festgelegt.

Tabelle 7.5: Typische Fragen im Zusammenhang mit statistischen Tests.

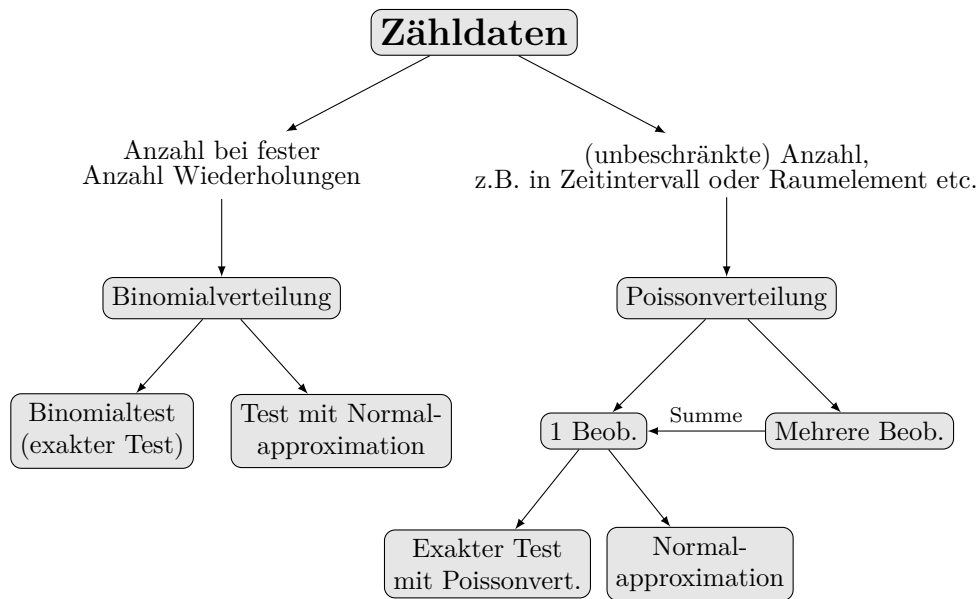


Abbildung 7.9: Statistische Tests bei Zähldaten.

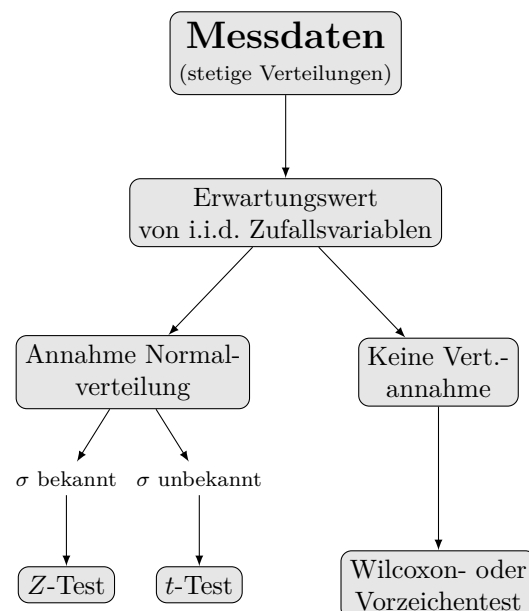


Abbildung 7.10: Statistische Tests bei Messdaten mit einer Stichprobe.

7.7 Review / Lernziele



- ☐ Sie verstehen das Konzept und die beiden Fehlerarten eines statistischen Tests.
- ☐ Sie können entscheiden, ob ein- oder zweiseitig getestet werden muss.
- ☐ Sie können auch für neue Situationen den Verwerfungsbereich eines Tests herleiten.
- ☐ Sie können den t - und den Z -Test durchführen (ein- und zweiseitig) und wissen, wann diese angebracht sind.
- ☐ Sie kennen Alternativen zum t -Test, falls die Normalverteilungsannahme nicht erfüllt ist.
- ☐ Sie kennen die Bedeutung der Begriffe Macht und p -Wert.
- ☐ Sie wissen, wie ein Vertrauensintervall aus einem Test hergeleitet werden kann (bzw. umgekehrt).
- ☐ Sie wissen, wie ein Vertrauensintervall interpretiert wird.
- ☐ Sie können mit Hilfe eines Vertrauensintervalls und Fachwissen die Relevanz eines Effektes untersuchen.

8 Vergleich zweier Stichproben

Häufige und wichtige Anwendungen der Statistik liegen im Vergleich verschiedener Verfahren oder Versuchsbedingungen. Hat z.B. Legierung *A* im Mittel eine höhere Zugfestigkeit als Legierung *B* (wie dies der Hersteller behauptet)? Oder führt ein neues technisches Verfahren zu weniger Ausschussware? Wenn Sie an ihre eigene Gesundheit denken, dann wünschen Sie wohl, dass ein neues Medikament wirksamer als das alte ist. Als einfachsten Fall behandeln wir hier den Vergleich zweier Methoden (Verfahren, Gruppen, Versuchsbedingungen, Behandlungen) bezüglich dem Erwartungswert.

8.1 Gepaarte und ungepaarte Stichproben

Wir sprechen von **gepaarten Stichproben**, wenn beide Versuchsbedingungen an *derselben* Versuchseinheit eingesetzt werden. Wir haben dann folgende Datenlage

$$\begin{aligned}x_1, \dots, x_n &\text{ unter Versuchsbedingung 1,} \\ y_1, \dots, y_n &\text{ unter Versuchsbedingung 2,}\end{aligned}$$

wobei (x_i, y_i) die Messungen an Versuchseinheit i sind. Notwendigerweise gilt dann: die Stichprobengrösse ist für beide Versuchsbedingungen dieselbe. Zwei Stichproben sind also gepaart, wenn man jede Versuchseinheit in der einen Gruppe *genau einer* Versuchseinheit in der anderen Gruppe zuordnen kann. Dies ist auch in Abbildung 8.1 illustriert. Man sagt auch, dass eine Versuchseinheit hier ein **Block** ist, bei dem wir *beide* Versuchsbedingungen anwenden.

Beispiel. Einige Fälle für gepaarte Stichproben:

- Vergleich zweier Reifentypen bzgl. Bremsweg, wobei jedes Testfahrzeug einmal mit Reifenart *A* und einmal mit Reifenart *B* ausgerüstet wurde. Die Versuchsbedingungen sind gegeben durch die Reifentypen, die Versuchseinheiten durch die Testfahrzeuge.
- Zwei Labors messen 15 Prüfkörper aus (nicht destruktiv). Jeder Prüfkörper wird von beiden Labors ausgemessen. Können wir einen Unterschied zwischen den Labors nachweisen? Hier ist ein Prüfkörper eine Versuchseinheit. Die beiden Labors sind die Versuchsbedingungen. <

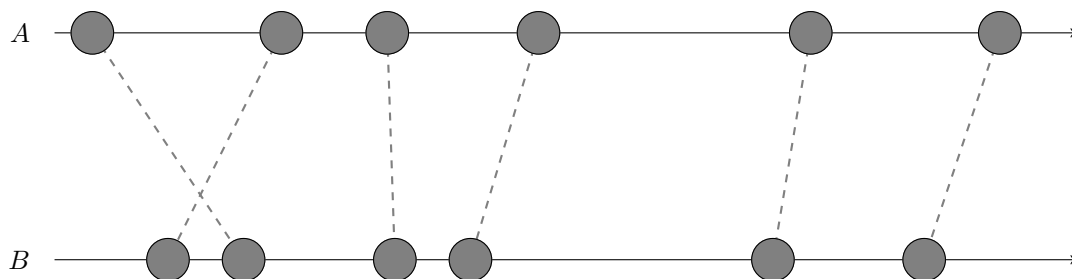


Abbildung 8.1: Illustration einer gepaarten Stichprobe. Zwei Messungen (zu Versuchsbedingung *A* und *B*) einer Versuchseinheit sind jeweils durch eine gestrichelte Linie verbunden. Die beiden Linien illustrieren die Zahlenstrahle und die Kreise die entsprechenden Messwerte.

Im Gegensatz dazu sprechen wir von **ungepaarten** oder **unabhängigen Stichproben**, wenn die Versuchseinheiten in der einen Gruppe *nichts* mit den Versuchseinheiten in der anderen Gruppe zu

tun haben. Wir haben dann Beobachtungen

$$\begin{aligned} x_1, x_2, \dots, x_n & \text{ unter Versuchsbedingung 1,} \\ y_1, y_2, \dots, y_m & \text{ unter Versuchsbedingung 2.} \end{aligned}$$

Im Allgemeinen ist $m \neq n$, aber nicht notwendigerweise.

Beispiel. *Einige Beispiele für unabhängige Stichproben:*

- *Vergleich der Zugfestigkeit von Stahldrähten aus zwei verschiedenen Werken. Aus jedem Werk wurden aus der Produktion zufällig 15 Drähte entnommen und entsprechend ausgemessen. Hier ist ein Draht eine Versuchseinheit und die Werke sind die verschiedenen Versuchsbedingungen.*
- *Zufällige Zuordnung von 100 Testpatienten zu Gruppe der Grösse 50 mit Medikamenten-Behandlung und zu anderer Gruppe der Grösse 50 mit Placebo-Behandlung. Hier ist ein Patient eine Versuchseinheit und die verschiedenen Behandlungen (mit Medikament bzw. Placebo) sind die Versuchsbedingungen.* ◀

8.2 Grundlegende Gedanken zur Versuchsplanung

Fast wichtiger als die korrekte Auswertung ist die **Versuchsplanung**. Man muss sicherstellen, dass allfällige Unterschiede zwischen den beiden Gruppen tatsächlich durch die verschiedenen Versuchsbedingungen und nicht durch eine andere Störgrösse verursacht werden. Hierzu müssen wir sicherstellen, dass der einzige *systematische* Unterschied zwischen den Messgrössen der beiden Gruppen die Versuchsbedingungen sind. Als universelles “Rezept” gelingt dies mit **Randomisierung**. Bei unabhängigen Stichproben bedeutet dies, dass man die *Zuordnung* von Versuchseinheit zu Versuchsbedingung *zufällig* wählt und auch in zufälliger Reihenfolge ausmisst. Bei gepaarten Stichproben kann man auch randomisieren, z.B. die “Reihenfolge” oder die “Platzierung” der beiden Versuchsbedingungen bei den einzelnen Versuchseinheiten wie auch die Reihenfolge der Ausmessung. Falls man dies so durchführt und die eine Versuchsbedingung als Kontrollgruppe dient, spricht man von einer sogenannten **randomisierten kontrollierten Studie** (randomized controlled trial).

Beispiel. *Bei den vorhergehenden gepaarten Stichproben bedeutet dies Folgendes:*

- *Bremsweg: Werfe eine Münze, ob ein Fahrzeug zuerst mit Reifenart A oder B ausgerüstet wird. Falls wir zuerst alle mit Reifenart A ausrüsten und testen, kann dies problematisch sein, falls es zeitliche Effekte gibt (z.B. verursacht durch das Wetter).*
- *Labors: Wie oben können wir durch den Zufall entscheiden lassen, zu welchem Labor ein Prüfkörper zuerst geht.* ◀

Beispiel. *Bei den vorhergehenden unabhängigen Stichproben haben wir:*

- *Zugfestigkeit: Die Drähte wurden schon zufällig den einzelnen Werken entnommen. Idealerweise werden diese nun auch in zufälliger Reihenfolge ausgemessen. Schlecht wäre es, wenn wir zuerst alle aus dem einen Werk ausmessen und dann alle aus dem anderen Werk. Sollte sich etwas mit der Messmethode über die Zeit hinweg ändern (Eichung, Lerneffekt des Bedieners etc.), so würden wir dies der unterschiedlichen Qualität der Werke zuordnen, was nicht korrekt ist! Schlecht wäre auch, wenn wir Drähte von Werk A von Mitarbeiter 1 ausmessen lassen würden und Drähte von Werk B von Mitarbeiter 2. Allfällige Unterschiede zwischen den Mitarbeitern würden wir auch dann den Werken zuordnen.*
- *Medikament: Wie oben.* ◀

Wieso ist Randomisierung so mächtig, dass sie immer funktioniert? Wir schauen dies anhand des Beispiels mit dem Medikamententest an (unabhängige Stichproben). Unter den 100 Patienten gibt es sicher solche, die gesünder (oder kränker) sind als andere. Wir wollen aber zwei Gruppen bilden, die möglichst *identische* Eigenschaften haben! Erschwerend kommt hinzu, dass wir nicht alle möglichen

Details über die Patienten kennen. Durch die Randomisierung hat jeder “Patiententyp” die gleiche Wahrscheinlichkeit, in die Behandlungsgruppe zu fallen. Die Randomisierung sorgt also dafür, dass wir schlussendlich in beiden Gruppen “die gleiche Mischung” von allen “Patiententypen” haben. Dies gilt für alle möglichen Eigenschaften, insbesondere auch für solche, die wir gar nicht kennen. Es gibt also keinen *systematischen* Unterschied zwischen den beiden Gruppen, denn wir haben sie ja *zufällig* gebildet! Falls wir nach Durchführung des Experiments (Einnahme des Medikamentes) einen systematischen Unterschied zwischen den beiden Gruppen feststellen können, dann können wir daraus schliessen, dass dies durch das Medikament verursacht wurde. Wir können also eine Aussage über eine **Ursache-Wirkung Beziehung** (kausaler Zusammenhang) treffen!

Bekannte Eigenschaften, von denen man im Voraus weiss (oder ahnt), dass sie einen Einfluss auf die entsprechende Messgrösse haben (z.B. Geschlecht, Spital, ...) sollen natürlich ausgenutzt werden, um homogene Gruppen zu bilden (man spricht wieder von Blockbildung). Innerhalb dieser homogenen Gruppen wird dann entsprechend in zwei Gruppen randomisiert.

Die Merkgel lautet (nach George Box):

“Block what you can, randomize what you cannot.”

In dem Sinne sind also gepaarte Stichproben (falls realisierbar) unabhängigen Stichproben vorzuziehen.

Sobald Menschen involviert sind, ist es ausserdem wichtig, dass ein Experiment wenn möglich **doppelblind** durchgeführt wird. Das heisst, dass weder die Person, welche die Behandlung durchführt oder deren Erfolg beurteilt, noch die Versuchsperson die Gruppenzugehörigkeit kennen. Dies ist wichtig, um den Effekt von Voreingenommenheit bei der Beurteilung auszuschalten.

Aus ethischen Gründen ist es nicht immer möglich, eine randomisierte kontrollierte Studie durchzuführen. Das bekannteste Beispiel ist wohl der Zusammenhang zwischen Rauchen und Lungenkrebs. Wir können nicht Leute zum Rauchen zwingen. In solchen Fällen kann man kein Experiment durchführen, sondern man muss die *vorhandenen* Daten möglichst gut ausnutzen. Man spricht von sogenannten **Beobachtungsstudien**. Während wir beim Beispiel mit dem Medikament selber entscheiden können, wer das Medikament erhält und wer das Placebo, ist dies bei Beobachtungsstudien *nicht* der Fall. Wir hätten z.B. auch bei einem Spital nachfragen können, wie sich 50 Patienten mit dem Medikament entwickelt haben und diese dann vergleichen können mit 50 Patienten ohne das entsprechende Medikament. Wahrscheinlich hätten wir dann gesehen, dass es den Patienten mit Medikament sehr viel schlechter geht! Daraus können wir aber *nicht* auf den kausalen Zusammenhang schliessen, dass das Medikament schädlich ist. Es kann neben dem Medikament durchaus einen anderen systematischen Unterschied zwischen den Gruppen geben, den wir nicht kennen und der einen Einfluss auf die Messgrösse (z.B. Überlebenszeit) hat. Hier ist es naheliegend, dass die Patienten ohne Medikament einfach gesünder sind als die anderen (sonst hätten sie wohl auch das Medikament erhalten). Gesündere Patienten leben aber auch länger. Der Gesundheitszustand ist in diesem Fall ein sogenannter **confounder** (to confound: vermengen, durcheinander bringen). Der Zusammenhang zwischen den einzelnen Variablen ist in Abbildung 8.2 dargestellt.

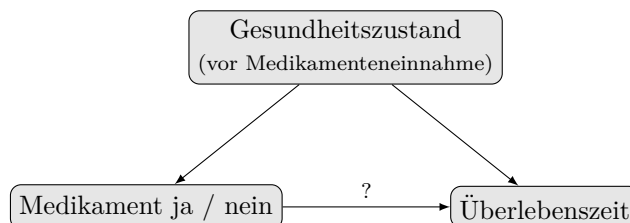


Abbildung 8.2: Zusammenhang zwischen den einzelnen Variablen. Ein Pfeil bedeutet einen kausalen Zusammenhang (Ursache-Wirkung Prinzip).

Was wir in den Daten sehen, ist eine Vermengung des Effektes des Medikaments und des Gesundheits-

zustandes. Wir sind aber nur am Effekt des Medikaments interessiert, der leider nicht mehr einfach rekonstruierbar ist. In der Praxis versucht man in solchen Fällen, die kausal relevanten Störgrößen mittels Regressionsmethoden “herauszurechnen” bzw. “dafür zu kontrollieren” (was wir hier nicht weiter besprechen). Schlussendlich kann man sich aber nie sicher sein, an alles gedacht zu haben und man kann bei Beobachtungsstudien nicht einfach auf einen kausalen Zusammenhang schliessen. Der springende Punkt ist, dass *nicht* randomisiert wurde. Wenn man randomisiert, eliminiert man automatisch alle möglichen confounder.

Eine randomisierte kontrollierte Studie ist also nicht nur viel **einfacher zum Auswerten**, sondern sie erlaubt auch viel **stärkere Schlussfolgerungen** (kausale Zusammenhänge!). Der “Preis”, den man zahlt, ist eine typischerweise aufwändige kostspielige Durchführung, da man nicht einfach Daten sammelt sondern aktiv Experimente durchführt.

8.3 Gepaarte Vergleiche

Bei der Analyse von gepaarten Stichproben arbeitet man stets mit den Differenzen innerhalb der Paare,

$$u_i = x_i - y_i, \quad i = 1, \dots, n,$$

welche wir als Realisierungen von i.i.d. Zufallsvariablen U_1, \dots, U_n auffassen. Kein Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $\mathbb{E}[U_i] = 0$. Dies kann man formal testen mit der Nullhypothese $H_0 : \mathbb{E}[U_i] = 0$ und mit der zweiseitigen (oder auch einseitigen) Alternative $H_A : \mathbb{E}[U_i] \neq 0$. Die folgenden Tests bieten sich dazu an:

1. der *t-Test*, siehe Kapitel 7.2.2.
2. der *Vorzeichen-Test*, falls die Normalverteilung nicht gerechtfertigt scheint, siehe Kapitel 7.5.1.
3. der *Wilcoxon-Test*, siehe Kapitel 7.5.2.

Da wir innerhalb eines “Blockes” (Versuchseinheit) Differenzen bilden, verschwindet die Variabilität zwischen den Blöcken und man hat ein “klareres” Bild (d.h. eine kleinere Varianz). Wenn wir z.B. zwei Größen an jeder Person (=Versuchseinheit) messen, so haben wir pro Person die Differenz der Messgrößen. Das Problem, dass die Messgrößen *zwischen* verschiedenen Personen stark unterschiedlich sein können (viel stärker als innerhalb einer Person), haben wir so elegant eliminiert. Diese Unterschiede “verschwinden” in den Differenzen. Denn: Pro Person sehen wir nur noch die Differenz, das personenspezifische “Niveau” kürzt sich automatisch weg.

8.4 Zwei-Stichproben Tests

Bei unabhängigen Stichproben kann man keine Paare bilden (da es keine Zuordnung zwischen Versuchseinheiten gibt). Man hat dann i.i.d. Zufallsvariablen

$$X_1, \dots, X_n$$

für die eine Versuchsbedingung und

$$Y_1, \dots, Y_m$$

für die andere. Ferner nimmt man an, dass alle Zufallsvariablen unabhängig sind (d.h. insbesondere X_i und Y_j). Die effektiv gemachten Beobachtungen sind wie üblich als Realisierungen von diesen Zufallsvariablen zu interpretieren. Das einfachste Problem lässt sich unter folgender Annahme lösen:

$$X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu_X, \sigma^2) \quad \text{sowie} \quad Y_1, \dots, Y_m \text{ i.i.d. } \sim \mathcal{N}(\mu_Y, \sigma^2),$$

und X_i, Y_j unabhängig. Wir nehmen also insbesondere an, dass die Varianz σ^2 in beiden Gruppen *gleich* gross ist.

Wir interessieren uns für die beiden Erwartungswerte μ_X und μ_Y . Wenn wir zweiseitig testen, haben wir die Nullhypothese

$$H_0 : \mu_X = \mu_Y \quad (\text{“Erwartungswerte der beiden Gruppen unterscheiden sich nicht”})$$

und die Alternative

$$H_A : \mu_X \neq \mu_Y \quad (\text{“Erwartungswerte der beiden Gruppen sind unterschiedlich”})$$

Durch Standardisierung können wir wieder eine Teststatistik herleiten. Wir definieren

$$Z = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Da $\text{Var}(\bar{X}_n - \bar{Y}_m) = \sigma^2(\frac{1}{n} + \frac{1}{m})$ gilt, dass $Z \sim \mathcal{N}(0, 1)$. In der Praxis kennen wir σ nicht. Wir ersetzen es durch die Schätzung S_{pool} , wobei

$$\begin{aligned} S_{pool}^2 &= \frac{1}{n+m-2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^m (Y_i - \bar{Y}_m)^2 \right) \\ &= \frac{1}{n+m-2} ((n-1)S_X^2 + (m-1)S_Y^2). \end{aligned}$$

Dies heisst nichts anderes, als dass S_{pool} ein gewichtetes Mittel der Schätzungen der Varianzen in den beiden Gruppen ist. Die Gewichte sind gegeben durch $(n-1)/(n+m-2)$ bzw. $(m-1)/(n+m-2)$.

Dies führt zur Teststatistik

$$T = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{S_{pool} \sqrt{\frac{1}{n} + \frac{1}{m}}}. \quad (8.1)$$

Man kann zeigen, dass T einer t -Verteilung mit $n+m-2$ Freiheitsgraden folgt, d.h. dass

$$T \sim t_{n+m-2}$$

gilt. Wir müssen hier 2 Freiheitsgrade abziehen, weil wir 2 Parameter μ_X und μ_Y geschätzt haben.

Wenn wir $H_0 : \mu_X = \mu_Y$ testen wollen, dann verwerfen wir für eine gegebene Realisierung

$$t = \frac{\bar{x}_n - \bar{y}_m}{s_{pool} \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

von T je nach Alternative H_A die Nullhypothese, falls

$$\begin{aligned} |t| \geq t_{n+m-2, 1-\frac{\alpha}{2}} &\iff t \in K = (-\infty, -t_{n+m-2, 1-\frac{\alpha}{2}}] \cup [t_{n+m-2, 1-\frac{\alpha}{2}}, \infty) & H_A : \mu_X \neq \mu_Y \\ t \geq t_{n+m-2, 1-\alpha} &\iff t \in K = [t_{n+m-2, 1-\alpha}, \infty) & H_A : \mu_X > \mu_Y \\ t \leq t_{n+m-2, \alpha} = -t_{n+m-2, 1-\alpha} &\iff t \in K = (-\infty, t_{n+m-2, \alpha}] = (-\infty, -t_{n+m-2, 1-\alpha}] & H_A : \mu_X < \mu_Y \end{aligned}$$

Dies ist der sogenannte **Zwei-Stichproben t-Test** für unabhängige Stichproben.

Analog wie früher kann man damit auch ein **Vertrauensintervall für die Differenz** $d = \mu_X - \mu_Y$ konstruieren, indem man alle Differenzen “durchtestet”, und diejenigen sammelt, die nicht verworfen

werden können. Dies führt zu

$$\begin{aligned} I &= \left\{ d : \left| \frac{(\bar{X}_n - \bar{Y}_m) - d}{S_{pool} \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| < t_{n+m-2, 1-\frac{\alpha}{2}} \right\} \\ &= \bar{X}_n - \bar{Y}_m \pm S_{pool} \sqrt{\frac{1}{n} + \frac{1}{m}} \cdot t_{n+m-2, 1-\frac{\alpha}{2}}. \end{aligned}$$

Die Verallgemeinerungen des Zwei-Stichproben t -Tests bei ungleichen Varianzen $\sigma_X^2 \neq \sigma_Y^2$ ist in der Literatur zu finden. Ebenfalls in der Literatur zu finden ist der Zwei-Stichproben Wilcoxon-Test, welcher ein für die Praxis guter Test für ungepaarte Stichproben ist.

Beispiel. Wir vergleichen zwei Produktionslinien (A und B) einer Aluminiumlegierung. Die Frage ist, ob entsprechende Prüfkörper im Mittel verschiedene Zugfestigkeiten aufweisen oder nicht.

Hierzu haben wir folgende Daten:

- Eine Stichprobe von $n = 10$ Prüfkörpern (x_1, \dots, x_{10}) von Produktionslinie A mit $\bar{x}_{10} = 398.9$ und $s_X = 15.9 \text{ N/mm}^2$.
- Die entsprechenden Daten B (y_1, \dots, y_8) für eine Stichprobe der Grösse $m = 8$ liefern $\bar{y}_8 = 380.8$ und $s_Y = 19.0 \text{ N/mm}^2$.

Es handelt sich um ungepaarte Stichproben, weil die Prüfkörper nichts miteinander zu tun haben.

Unser Modell für die Daten ist X_1, \dots, X_{10} i.i.d. $\sim \mathcal{N}(\mu_X, \sigma^2)$, Y_1, \dots, Y_8 i.i.d. $\sim \mathcal{N}(\mu_Y, \sigma^2)$ und X_i, Y_j unabhängig.

Für die Null- und die Alternativhypothese haben wir hier

$$\begin{aligned} H_0 : \quad & \mu_X = \mu_Y \\ H_A : \quad & \mu_X \neq \mu_Y \end{aligned}$$

Wir führen nun einen Zwei-Stichproben t -Test durch.

Es ist

$$s_{pool}^2 = \frac{1}{10 + 8 - 2} (9 \cdot 15.9^2 + 7 \cdot 19.0^2) = 300.14.$$

Dies führt zum realisierten Wert der Teststatistik

$$t = \frac{\bar{x}_{10} - \bar{y}_8}{\sqrt{300.14} \sqrt{\frac{1}{10} + \frac{1}{8}}} = 2.2$$

Der Verwerfungsbereich ist hier auf dem 5%-Niveau gegeben durch

$$K = \{|t| \geq t_{16, 0.975}\} = (-\infty, -2.12] \cup [2.12, \infty).$$

Wir können also die Nullhypothese auf dem 5%-Niveau verwerfen.

Das 95%-Vertrauensintervall für die Differenz $\mu_X - \mu_Y$ ist gegeben durch

$$\begin{aligned} I &= \bar{x}_{10} - \bar{y}_8 \pm s_{pool} \sqrt{\frac{1}{10} + \frac{1}{8}} \cdot t_{16, 0.975} \\ &= 18.1 \pm \sqrt{300.14} \sqrt{\frac{1}{10} + \frac{1}{8}} \cdot 2.12 \\ &= [0.68, 35.5] \text{ N/mm}^2. \end{aligned}$$

Wie wir schon vom Testresultat her wissen, enthält das Vertrauensintervall die Null nicht (sonst hätten wir nicht verworfen). Wir sehen aber, dass das Vertrauensintervall sehr nahe bei 0 liegt. Die Relevanz des Unterschieds zwischen den Produktionslinien ist also nicht gesichert! \triangleleft

Eine Erweiterung des Zwei-Stichproben t -Tests behandelt den Fall, bei dem die Varianzen in den beiden Gruppen nicht gleich gross sind. Man spricht vom sogenannten **Welch-Test**. Zudem gibt es auch eine Erweiterung des Wilcoxon-Tests für unabhängige Stichproben, der sogenannte **Mann-Whitney U-Test**.

8.5 Vergleich der Konzepte

Die wichtigsten Konzepte sind in Abbildung 8.3 nochmals illustriert.

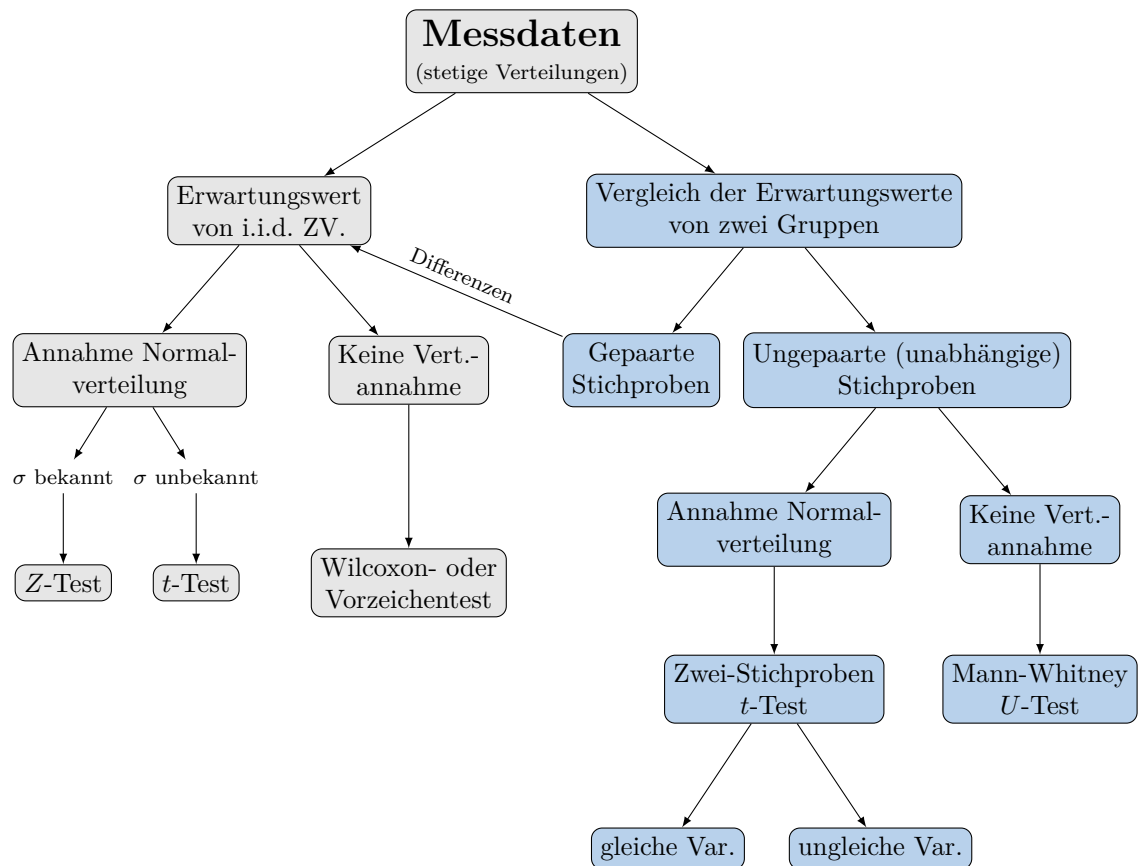


Abbildung 8.3: Statistische Tests bei stetigen Messgrößen.

8.6 Review / Lernziele



- ☐ Sie können zwischen gepaarten und ungepaarten Stichproben unterscheiden.
- ☐ Sie wissen, wie man gepaarte Stichproben mit den Methoden für eine Stichprobe behandeln kann.
- ☐ Sie kennen den t -Test für zwei unabhängige Stichproben und können diesen anwenden, auch für die Konstruktion von Vertrauensintervallen.

9 Ausblick: Lineare Regression

9.1 Einführung

Oft will man eine Grösse y durch eine andere Grösse x “erklären” oder “vorhersagen”. Können wir zum Beispiel bei einer Sprengung die (mittlere) Erschütterung (y) durch den Abstand zum Sprengzentrum (x) vorhersagen? Oder wie sieht es bei einer Bodenprobe mit der einaxialen Druckfestigkeit (y) in Abhängigkeit zur Tiefe aus?

Bei der Korrelation haben wir schon lineare Abhängigkeiten zwischen zwei Grössen untersucht. Dort hatten wir aber weder eine bevorzugte “Richtung” noch war Prognose ein Thema.

9.2 Einfache lineare Regression

Wir betrachten wieder wie früher paarweise vorliegende Daten (x_i, y_i) , $i = 1, \dots, n$.

Wie schon in Kapitel 8.2 besprochen, können wir anhand von Daten von Beobachtungsstudien im Allgemeinen *nicht* auf einen *kausalen* Zusammenhang (d.h. auf ein Ursache-Wirkungs-Prinzip) zwischen x und y schliessen. Wir können aber trotzdem versuchen, den Zusammenhang zwischen x und y zu modellieren.

9.2.1 Modell

Das **einfache lineare Regressionsmodell** lautet

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \dots, n.$$

Man bezeichnet Y_i als **Zielvariable** (response variable), während x_i die sogenannte **erklärende Variable** (explanatory variable oder predictor) ist. Die erklärende Variable wird typischerweise als nicht-zufällig aufgefasst, daher auch die Notation mit einem Kleinbuchstaben.

Bei E_i handelt es sich um einen zufälligen **Fehlerterm** (error). Man kann sich z.B. Messfehler oder nicht-systematische Effekte darunter vorstellen. Typischerweise nehmen wir für die Fehler an, dass sie normalverteilt sind, d.h.

$$E_i \text{ i.i.d. } \sim \mathcal{N}(0, \sigma^2).$$

Der Zusammenhang zwischen der Zielgrösse Y und der erklärenden Variable x ist also nicht exakt, d.h. die Werte der Zielgrösse streuen gemäss Modellgleichung um die wahre (unbekannte) Gerade herum. Die Y_i 's sind also insbesondere zufällig. Wir verwenden in der Modellgleichung für die Zielgrösse daher Grossbuchstaben (die realisierten Werte schreiben wir wie gewohnt als y_i). In der Tat gilt für Y_i , dass

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

Dies haben wir in Abbildung 9.1 illustriert. Wenn wir einen Wert für x fixieren, so erwarten wir im Mittel den Wert $\beta_0 + \beta_1 x$ für die Zielgrösse Y . Dies ist gerade der Wert der Geraden an der Stelle x . Die Streuung um die Gerade herum wird durch den Fehlerterm verursacht und ist durch die Dichte der Normalverteilung illustriert. Da $\mathbb{E}[E_i] = 0$, gibt es keine systematischen Abweichungen von der Geraden. Zusätzlich ist $\text{Var}(E_i) = \sigma^2$, d.h. die Streuung um die Gerade ist überall gleich gross.

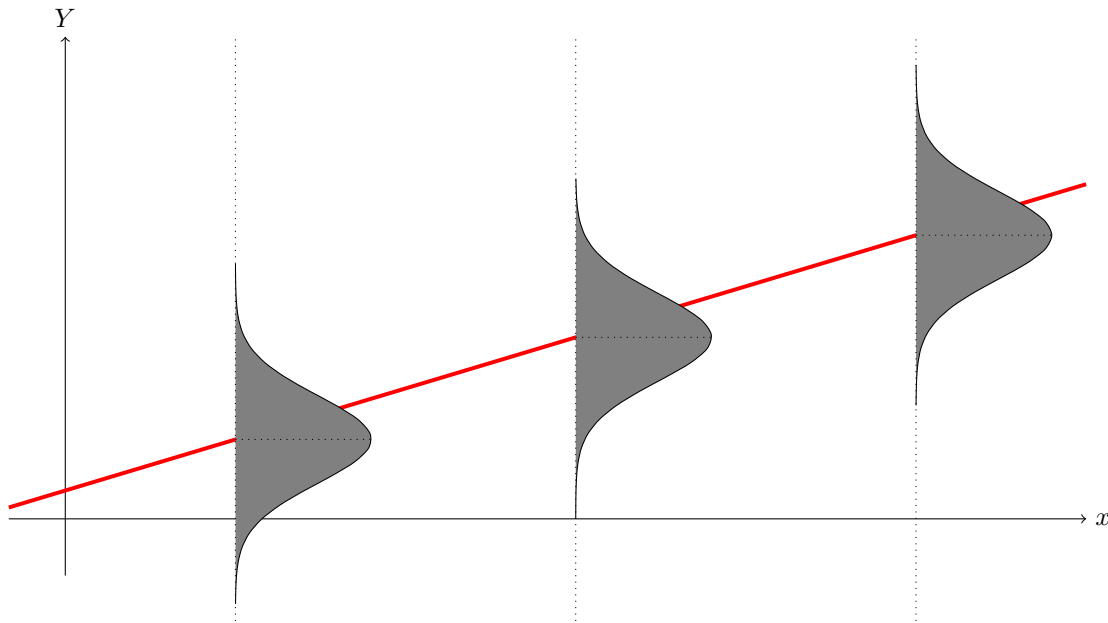


Abbildung 9.1: Illustration des datengenerierenden Prozesses bei der einfachen linearen Regression. Für drei verschiedene Werte der erklärenden Variable x ist die entsprechende Verteilung der Zielgrösse Y dargestellt. Die wahre Gerade ist als durchgezogene rote Linie eingezeichnet.

Als unbekannte Parameter haben wir also β_0 (Achsenabschnitt), β_1 (Steigung) und σ^2 (Varianz des Fehlerterms). Das Interesse liegt typischerweise bei den ersten beiden Parametern. Daher nennt man die Fehlervarianz σ^2 hier auch **Störparameter** (nuisance parameter).

Das Modell heisst “einfach”, weil nur *eine* erklärende Variable vorhanden ist. Zudem heisst das Modell “linear”, da die Parameter β_0 und β_1 linear in der Modellgleichung vorkommen. Bei der erklärenden Variable x oder auch bei der Zielgrösse Y kann es durchaus sein, dass diese durch eine Transformation einer ursprünglich gemessenen Grösse zustande gekommen sind.

9.2.2 Parameterschätzungen

Die Daten (x_i, y_i) , $i = 1, \dots, n$ liegen uns in Form einer “Punktwolke” vor. Unser Ziel ist es, die Parameter der Modellgeraden zu schätzen. Diese wählen wir so, dass die geschätzte Gerade “am Besten” durch die Punktwolke passt (siehe auch Abbildung 9.2). Als Gütekriterium verwenden wir die Summe der quadrierten (vertikalen) Abweichungen

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

D.h. wir wählen

$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Man kann nachrechnen, dass gilt

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned}$$

Wegen obiger Wahl des Gütekriteriums bezeichnet man diese auch als **Kleinste-Quadrate Schätzer**.

Wir können $\hat{\beta}_1$ auch umschreiben als

$$\begin{aligned}\hat{\beta}_1 &= \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}} \cdot \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}} \\ &= r \frac{s_y}{s_x},\end{aligned}$$

wobei r die empirische Korrelation und s_x bzw. s_y die empirischen Standardabweichungen sind (siehe Kapitel 3.4). Dies bedeutet, dass wenn wir x um eine Standardabweichung erhöhen, dann ändert sich y im Mittel um den Wert $r \cdot s_y$. Falls die Punkte nicht exakt auf einer Geraden liegen, dann gilt $|r| < 1$, und somit ändert sich y im Schnitt um *weniger* als eine Standardabweichung. Man nennt dies auch “Regression (Rückschritt) zum Mittel” (regression to the mean). Dies ist der Grund für den Namen Regression.

Die (vertikale) Abweichung eines Datenpunktes von der *geschätzten* Gerade bezeichnen wir als **Residuum** r_i (Mehrzahl: **Residuen**), d.h.

$$r_i = y_i - \hat{y}_i,$$

wobei

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

der geschätzte Wert der Geraden an der Stelle x_i ist, siehe auch Abbildung 9.2.

Die Residuen benutzen wir als Approximation für die nicht-beobachtbaren Fehlerterme E_i (diese kennen wir nicht, weil wir die wahre Gerade nicht kennen). Man kann zeigen, dass das arithmetische Mittel der r_i bei der Kleinste-Quadrate Schätzung immer 0 ist. Daher schätzen wir die Varianz des Fehlerterms mit

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n r_i^2.$$

Dies ist nichts anderes als die empirische Varianz der Residuen mit dem Faktor $n-2$ statt $n-1$. Man kann zeigen, dass die so definierten Schätzer alle erwartungstreue Schätzer für die entsprechenden Parameter sind.

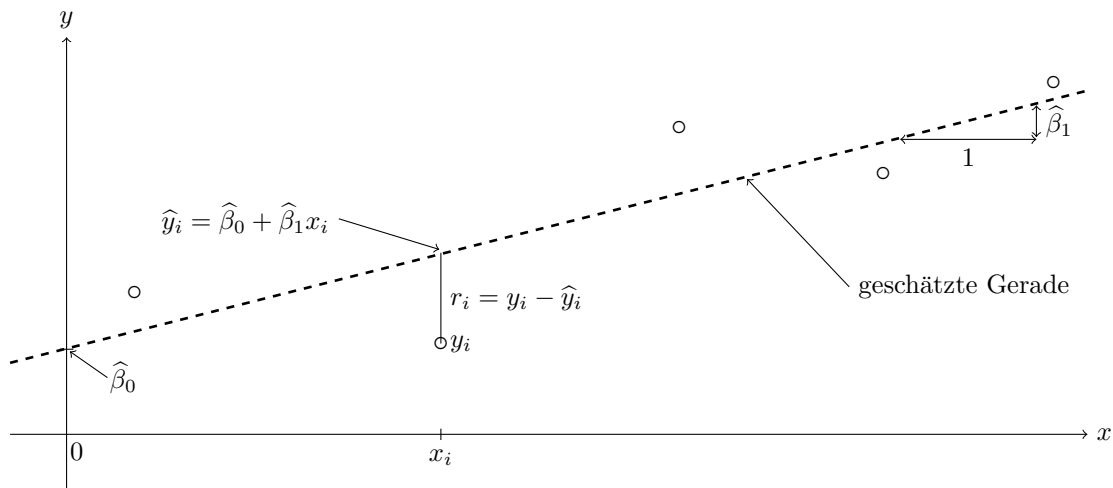


Abbildung 9.2: Datenpunkte und mit der Methode der Kleinsten-Quadraten geschätzte Gerade (gestrichelt). An der Stelle x_i sind zudem der angepasste Wert und das Residuum illustriert.

Die geschätzte Gerade entspricht *nicht* der wahren (unbekannten) Geraden (aber sie ist hoffentlich sehr ähnlich dazu). Wie früher bei den Parameterschätzern gehen wir von einem datengenerierenden Prozess aus, dessen Parameter wir nicht kennen, aber mit den vorhandenen Daten schätzen wollen.

Dies ist in Abbildung 9.3 illustriert. Für 4 verschiedene Stichprobengrößen sind jeweils 3 simulierte Datensätze dargestellt, bei denen sowohl die wahre und die geschätzte Gerade eingezeichnet sind (die wahre Gerade kennen wir hier, weil wir die Daten selber simuliert haben). Wir sehen, dass die geschätzte Gerade um die wahre Gerade fluktuiert und nicht exakt mit ihr übereinstimmt. Die Genauigkeit nimmt mit zunehmender Stichprobengröße zu.

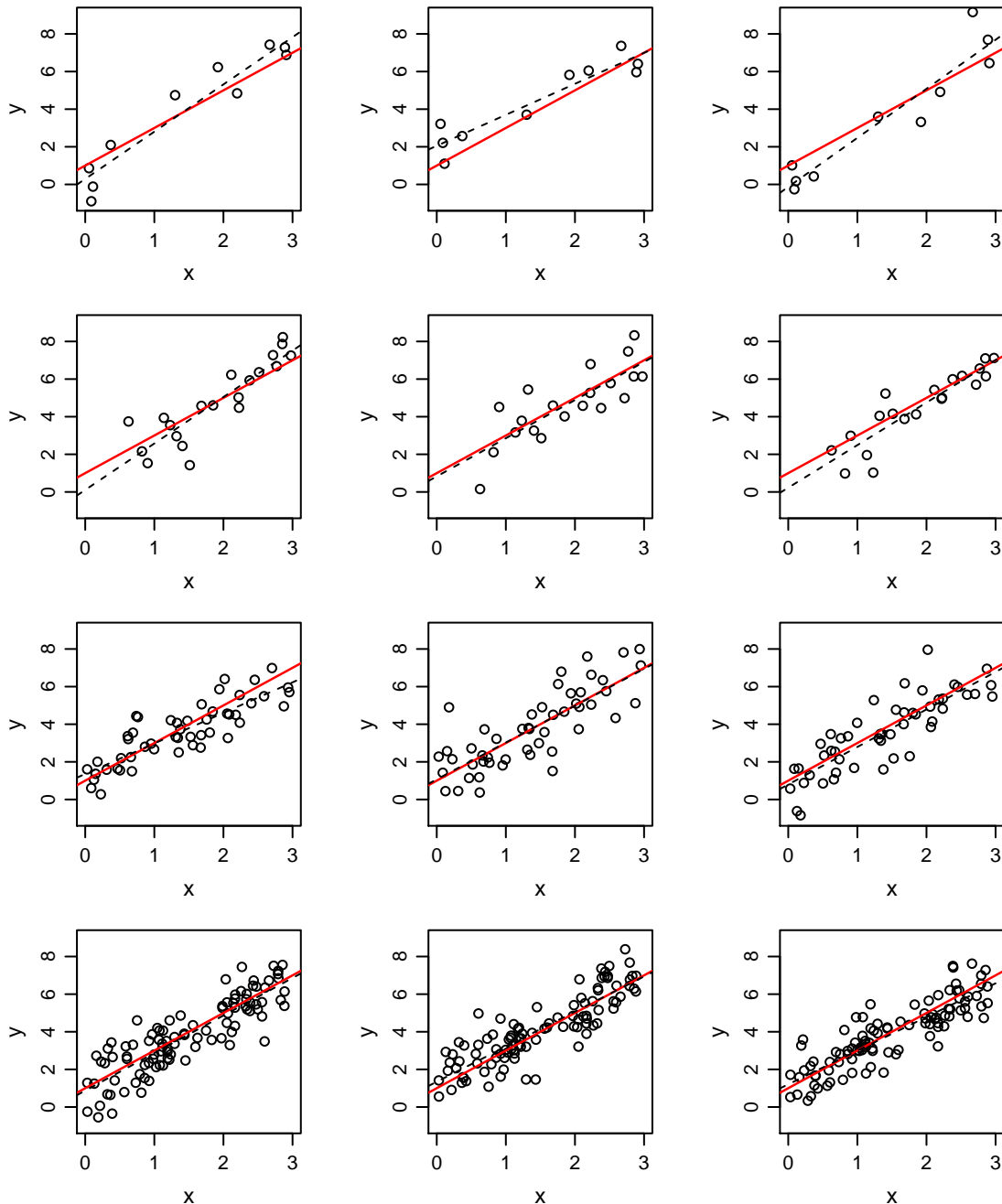


Abbildung 9.3: Simulierte Datensätze der Stichprobengröße $n = 10, 20, 50, 100$ (zeilenweise, oben nach unten). Die wahre Gerade ist gegeben durch $y = 1 + 2x$ und durchgezogen (rot) eingezeichnet. Die jeweils geschätzte Gerade ist gestrichelt dargestellt.

Maximum-Likelihood Schätzer

Man kann auch die Maximum-Likelihood Methode zur Schätzung der Parameter verwenden. Diese führt bei den getroffenen Modellannahmen zur gleichen Lösung wie die Kleinste-Quadrate Schätzer. Wir illustrieren dies hier kurz, lassen aber den Störparameter σ^2 einmal aussen vor.

Gemäss unserem Modell gilt, dass

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), Y_i \text{ unabhängig.}$$

Also ist die Likelihoodfunktion gegeben durch

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma} \right)^2 \right\}.$$

Entsprechend erhält man für die log-Likelihoodfunktion

$$l(\beta_0, \beta_1) = c - \frac{1}{2} \sum_{i=1}^n \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma} \right)^2,$$

wobei c eine von den Parametern β_0 und β_1 unabhängige Konstante ist. Die Maximierung der log-Likelihoodfunktion entspricht also gerade dem Kleinste-Quadrate Problem. Das (willkürliche) Gütekriterium “Kleinste Quadrate” kann also durch normalverteilte Fehler und den entsprechenden Maximum-Likelihood Schätzer motiviert werden.

9.2.3 Tests und Vertrauensintervalle

Verglichen mit der Numerik (oder linearer Algebra) haben wir hier in der Statistik den grossen Vorteil, dass wir die Genauigkeit von $\hat{\beta}_0$ und $\hat{\beta}_1$ angeben können. Man kann herleiten, dass gilt

$$\begin{aligned} \hat{\beta}_0 &= \mathcal{N} \left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_X} \right) \right) \\ \hat{\beta}_1 &= \mathcal{N} \left(\beta_1, \frac{\sigma^2}{SS_X} \right), \end{aligned}$$

wobei

$$SS_X = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Die Parameterschätzer sind also auch wieder Zufallsvariablen, die um die wahren (unbekannten) Parameterwerte herum streuen (genau wie früher!). Wir sehen insbesondere sofort, dass die beiden Parameterschätzer erwartungstreu sind. Die Genauigkeit ist dann gegeben durch die Varianz (oder Standardabweichung) der Verteilung der Parameterschätzer. Die Standardabweichung eines Parameterschätzers bezeichnen wir wie früher als **Standardfehler**.

Der Standardfehler von $\hat{\beta}_1$ ist also $\sigma/\sqrt{SS_X}$. Setzt man die Schätzung $\hat{\sigma}$ ein, so erhält man den **geschätzten Standardfehler**. Damit kann man analog wie beim t -Test eine Teststatistik konstruieren. Es gilt tatsächlich, dass

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{SS_X}} \sim t_{n-2}.$$

Wie früher können wir also eine Funktion der Form

$$\frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}}$$

als Teststatistik verwenden. Beim Test der Nullhypothese

$$H_0 : \beta_1 = 0 \quad (\text{“Es gibt keinen linearen Zusammenhang zwischen } x \text{ und } y\text{”})$$

vs. die Alternative

$$H_A : \beta_1 \neq 0 \quad (\text{“Es ist ein linearer Zusammenhang zwischen } x \text{ und } y \text{ vorhanden”})$$

verwerfen wir H_0 auf dem Niveau α also zu Gunsten von H_A , falls

$$\left| \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{SS_X}} \right| \geq t_{n-2, 1-\frac{\alpha}{2}}.$$

Oft wird der p-Wert des Tests automatisch von entsprechender Software geliefert.

Analog wie beim t -Test ist das $(1 - \alpha) \times 100\%$ -Vertrauensintervall für β_1 gegeben durch

$$\hat{\beta}_1 \pm t_{n-2, 1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{SS_X}}.$$

Wieder wie früher hat das Vertrauensintervall also die Form

$$\text{Schätzung} \pm \text{Quantil} \times (\text{geschätzter Standardfehler}).$$

Für $\alpha = 0.05$ gilt $t_{n-2, 1-\frac{\alpha}{2}} \approx 1.96$ für n gross. Als Faustregel kann man also für das 95%-Vertrauensintervall die Formel

$$\text{Schätzung} \pm 2 \times (\text{geschätzter Standardfehler})$$

verwenden.

Wir haben hier alles nur für die Steigung β_1 betrachtet. In den meisten Anwendungen ist die Steigung auch der Parameter von Interesse. Alle Berechnungen für den Achsenabschnitt β_0 gehen aber ganz analog.

Beispiel. Bei einer Bohrung in Permafrostboden wurden in verschiedenen Tiefen jeweils die Temperatur gemessen. Die Daten sind in folgender Tabelle aufgelistet und in Abbildung 9.4 dargestellt.

Tiefe [m]	0	0.2	0.5	0.6	0.8	0.9	1.2
Temperatur ($^{\circ}\text{C}$)	6	4.2	0.6	-2.1	-5.2	-7.3	-8.9

Ein Computer-Output liefert folgendes Resultat

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.3681	0.7191	8.856	0.000305
tiefe	-13.6373	1.0111	-13.487	4.01e-05

Die geschätzten Parameter sind also $\hat{\beta}_0 = 6.37$ und $\hat{\beta}_1 = -13.64$. Die Steigung ist auf dem 5%-Niveau signifikant von Null verschieden, denn der p -Wert des entsprechenden Tests ist mit $4.01 \cdot 10^{-5}$ kleiner als 5%, d.h. die Nullhypothese $H_0 : \beta_1 = 0$ wird deutlich verworfen zu Gunsten von $H_A : \beta_1 \neq 0$.

Das 95%-Vertrauensintervall für β_1 ist gegeben durch

$$-13.64 \pm t_{5, 0.975} \cdot 1.01 = -13.64 \pm 2.571 \cdot 1.01 = [-16.2, -11.0].$$

Gemäss unseren Daten können wir davon ausgehen, dass die wahre Steigung im Bereich $[-16.2, -11.0]$ liegt. \triangleleft

Vertrauensintervalle für den Erwartungswert und Prognoseintervalle

Wir können nicht nur für die Parameter β_0 und β_1 , sondern auch für den wahren Wert der Geraden an einer Stelle x ein Vertrauensintervall angeben. Der Wert der Modellgerade an der Stelle x ist nichts anderes als der Erwartungswert der Zielgrösse, wenn wir ein x fixieren und ist gegeben durch

$$\beta_0 + \beta_1 x.$$

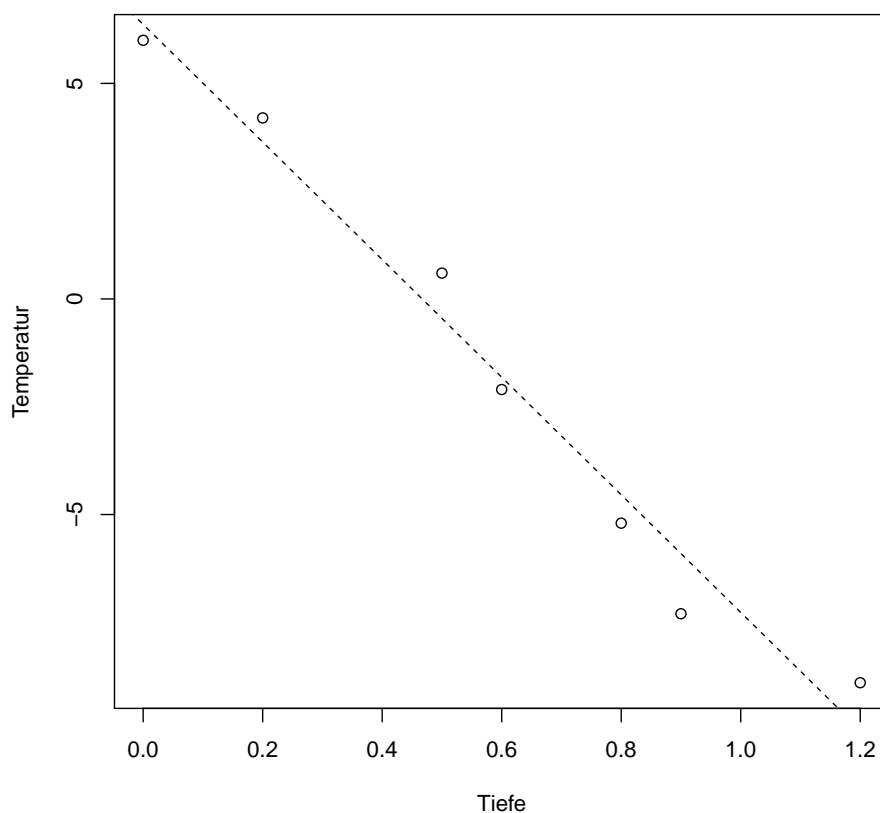


Abbildung 9.4: Daten und geschätzte Gerade im Beispiel mit der Bohrung in Permafrostboden.

Also macht man ein Vertrauensintervall für den (fixen, aber unbekannten) speziellen Modellparameter “Erwartungswert an der Stelle x ”.

Zusätzlich kann man auch ein sogenanntes **Prognoseintervall** konstruieren. Dies ist dann ein Intervall, das mit hoher Wahrscheinlichkeit eine neue (zufällige) *Beobachtung* Y an der Stelle x einfängt.

Sowohl Prognose- wie auch Vertrauensintervalle für das Beispiel mit dem Permafrostboden sind in Abbildung 9.5 als “Bänder” dargestellt. Wenn wir ein x fixieren, so ist das Prognoseintervall für die Temperatur für die fixierte Tiefe gegeben durch die Werte zwischen den zwei gepunkteten Linien (analog beim Vertrauensintervall). Das Prognoseintervall ist also immer breiter. Dies ist auch einleuchtend, denn es muss noch die Variabilität einer Beobachtung “abdecken” (wegen dem Fehlerterm E). Beide Bänder sind übrigens gekrümmt, beim Prognoseband ist dies nur viel schlechter sichtbar.

Wir verzichten hier auf Formeln, denn typischerweise erhält man die entsprechenden Angaben einfach mit entsprechender Software. So erhalten wir z.B. als Vertrauensintervall für die mittlere Temperatur an der Stelle $x = 0.6$ das Intervall $[-2.81, -0.82]$ und als Prognoseintervall für eine (einzelne) Messung in dieser Tiefe entsprechend $[-4.62, 0.99]$.

9.2.4 Residuenanalyse

Die betrachteten Tests und Vertrauensintervalle basieren auf den Annahmen des linearen Regressionsmodells. Diese kann man folgendermassen aufschlüsseln:

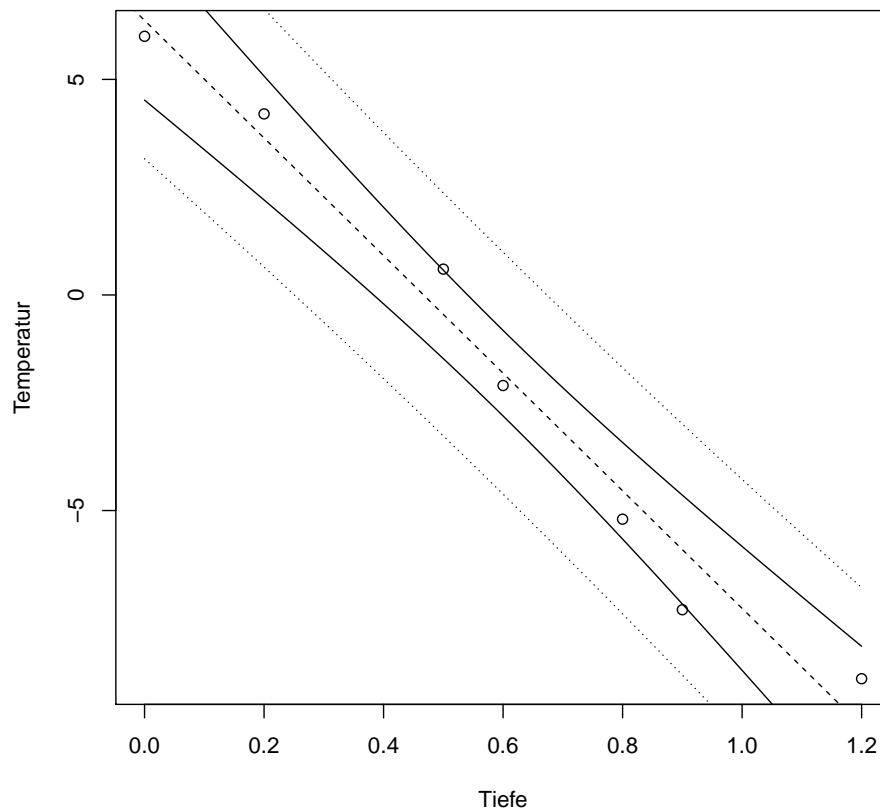


Abbildung 9.5: Daten, geschätzte Gerade (gestrichelt), Vertrauensintervalle (durchgezogen) und Prognoseintervall (gepunktet) im Beispiel mit der Bohrung in Permafrostboden.

1. Es gilt $\mathbb{E}[E_i] = 0$, d.h. es gibt keinen systematischen Fehler im Modell, oder anders ausgedrückt: die Modellgleichung ist korrekt.
2. Die E_1, \dots, E_n sind i.i.d. Die Fehler sind also unabhängig voneinander und folgen der gleichen Verteilung (insbesondere muss also auch die Varianz gleich gross sein).
3. Die E_1, \dots, E_n sind normalverteilt.

Je deutlicher die Modellannahmen verletzt sind, desto weniger vertrauenswürdig sind die Resultate (p-Werte der Tests, Vertrauensintervalle, ...). Der Übergang ist jeweils fließend. Bei einer nur “leichten” Verletzung der Modellannahmen sind die Resultate sicher noch brauchbar.

Wie früher bei den QQ-Plots untersuchen wir die Güte der Modellanpassung qualitativ mit diversen Grafiken. Deren Beurteilung erfordert einige Erfahrung.

Tukey-Anscombe Plot (TA-Plot)

Beim Tukey-Anscombe Plot zeichnet man die Residuen r_i gegen die angepassten Werte \hat{y}_i auf (bei der einfachen linearen Regression könnte man die Residuen auch gegen x_i aufzeichnen). Im Idealfall sollte man eine gleichmässige Streuung der Punkte um die x -Achse sehen.

Mögliche Szenarien sind:

- Falls systematische Abweichungen von der x -Achse erkennbar sind, so spricht dies gegen die Annahme 1 von oben. Denn in diesem Fall gibt es Bereiche, wo der Fehler im Schnitt nicht 0 ist (z.B. falls man einen quadratischen Effekt im Modell vergessen hat).
- Ist die Streuung stark unterschiedlich (z.B. trichterförmiges Bild), so spricht dies gegen Annahme 2.
- Ev. sieht man auch "Ausreisserpunkte".

Normalplot

Man erstellt einen gewöhnlichen Normalplot der Residuen. Es sollten keine groben Abweichungen von einer Geraden vorliegen, sonst wäre dies eine Verletzung der Modellannahme 3.

Serial Correlation Plot

Um die Unabhängigkeit der E_1, \dots, E_n zu überprüfen, kann man z.B. die Residuen r_i gegen die entsprechende Beobachtungsnummer i aufzeichnen. Dies ist insbesondere dann sinnvoll, wenn die Beobachtungen in dieser zeitlichen Reihenfolge aufgenommen wurden.

Im Idealfall sollte es keine Regionen geben, wo sich die Residuen ähnlich verhalten (d.h. wo z.B. alle positiv sind).

9.3 Multiple lineare Regression

In der Praxis hat man oft nicht nur eine, sondern *mehrere* erklärende Variablen $x^{(1)}, \dots, x^{(m)}$, $m > 1$.

Das einfache lineare Regressionsmodell kann man für diesen Fall erweitern. Man spricht dann vom sogenannten multiplen linearen Regressionsmodell.

9.3.1 Modell

Das **multiple lineare Regressionsmodell** ist gegeben durch

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_p x_i^{(m)} + E_i, \quad i = 1, \dots, n,$$

wobei wieder wie früher E_i i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ angenommen wird. Bei $x_i^{(j)}$ handelt es sich um die j -te erklärende Variable der i -ten Beobachtung.

Total haben wir also $p = m + 1$ verschiedene β -Parameter. Der Parameter β_j ist der Effekt der erklärenden Variable $x^{(j)}$ auf die Zielgrösse Y , wenn alle anderen erklärenden Variablen fest gehalten werden und nur $x^{(j)}$ variiert wird. Wenn wir also $x^{(j)}$ um eine Einheit erhöhen, dann erwarten wir gemäss Modell eine um β_j grössere Zielgrösse, wenn an den anderen erklärenden Variablen *nichts* geändert wird.

Das Modell heisst wieder linear, weil die Parameter linear in der Modellgleichung vorkommen. So ist z.B. das Modell

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$$

auch ein multiples lineares Regressionsmodell, weil die Parameter β_0, β_1 und β_2 linear in der Modellgleichung vorkommen. Wir sehen also, dass die erklärenden Variablen insbesondere also sogar Funktionen voneinander sein können! Die Modellklasse der multiplen linearen Regression ist also sehr gross und die geforderte Linearität ist nicht eine so grosse Einschränkung wie vielleicht ursprünglich befürchtet.

Man kann das Modell für n Beobachtungen auch in Matrix-Schreibweise darstellen. Hierzu fassen wir die verschiedenen Größen zuerst in Vektoren bzw. Matrizen zusammen:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, E = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix}.$$

Das Modell kann dann geschrieben werden als

$$Y = X\beta + E.$$

Die Matrix X heisst **Designmatrix**. In der i -ten Zeile findet man alle erklärenden Variablen der i -ten Beobachtung. In den Spalten findet man die verschiedenen erklärenden Variablen. Die erste Spalte besteht nur aus Einsen: es handelt sich um den Achsenabschnitt.

Beispiele für multiple lineare Regressionsmodelle sind unter anderem:

- **Einfache lineare Regression**

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

$$p = 2, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

- **Regression mit quadratischen erklärenden Variablen**

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$$

$$p = 3, \quad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Beachten Sie wieder, dass das Modell linear in den Parametern ist. Daher wird es ein lineares Regressionsmodell genannt.

- **Regression mit transformierten erklärenden Variablen**

$$Y_i = \beta_0 + \beta_1 \log(x_i^{(1)}) + \beta_2 \sin(\pi x_i^{(2)}) + E_i$$

$$p = 3, \quad X = \begin{pmatrix} 1 & \log(x_1^{(1)}) & \sin(\pi x_1^{(2)}) \\ 1 & \log(x_2^{(1)}) & \sin(\pi x_2^{(2)}) \\ \vdots & \vdots & \vdots \\ 1 & \log(x_n^{(1)}) & \sin(\pi x_n^{(2)}) \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Wiederum ist das Modell linear in den Parametern und wird daher lineares Modell genannt (obwohl es in den $x_i^{(j)}$ nicht linear ist).

9.3.2 Parameterschätzungen

Wie bei der einfachen linearen Regression schätzt man die Parameter mit der Methode der kleinsten Quadrate. Falls die Matrix X vollen Rang hat, so ist die Lösung geschlossen darstellbar:

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

wobei wir hier mit y den Vektor der *beobachteten* Werte der Zielgrösse bezeichnen. Für $\hat{\sigma}^2$ verwendet man

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i^{(1)} - \dots - \hat{\beta}_m x_i^{(m)} \right)^2.$$

Der Nenner bei der Schätzung von $\hat{\sigma}^2$ hat die Form “Anzahl Beobachtungen Minus Anzahl Parameter” und sorgt wieder dafür, dass der Schätzer erwartungstreu ist.

9.3.3 Tests und Vertrauensintervalle

Individuelle Tests

Da wir hier für jeden einzelnen Parameter β_j einen Test durchführen können, haben wir (potentiell) viele *individuelle* Tests

$$H_{0,j} : \beta_j = 0$$

vs.

$$H_{A,j} : \beta_j \neq 0$$

für $j = 0, \dots, m$. Wie bei der einfachen linearen Regression kann man Teststatistiken konstruieren und erhält wieder eine t -Verteilung, jetzt aber mit $n - m$ Freiheitsgraden. Die Anzahl Freiheitsgrade hat also auch hier die Form “Anzahl Beobachtungen Minus Anzahl Parameter”.

Auf konkrete Formeln verzichten wir, da die entsprechenden Werte einfach von einem Computer-Output ablesbar sind.

Ein individueller Test beantwortet die Frage, ob man eine *einzelne* erklärende Variable weglassen kann. Wenn sich zwei erklärende Variablen sehr ähnlich sind (d.h. wenn sie stark korreliert sind), so kann es sein, dass man aufgrund der individuellen Tests jeweils zum Schluss kommt, dass man beide (einzeln) weglassen kann. Die Ursache liegt darin, dass die andere Variable ja (fast) die gleiche Information liefert und daher durch den Wegfall einer der beiden Variablen kein “Verlust” entsteht. Dies bedeutet aber nicht, dass man *beide* Variablen weglassen kann.

F-Test

Wir können bei der multiplen Regression auch testen, ob es plausibel ist, dass *alle* Variablen weglassen werden können (typischerweise ausser dem Achsenabschnitt). D.h. wir haben dann die Nullhypothese

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0 \text{ (“keine Variable hat einen Einfluss”)}$$

vs. die Alternative

$$H_A : \text{Mindestens ein } \beta_j \neq 0, j \in \{1, \dots, m\}.$$

Dies ist der sogenannte **F-Test**, der auf der gleichnamigen F -Verteilung basiert. Wir verzichten auf Details. Computerprogramme liefern typischerweise direkt den p -Wert des entsprechenden Tests, den wir auch als **Globaltest** bezeichnen, weil er simultan alle erklärenden Variablen testet.

Mit ähnlichen Überlegungen können wir auch mit einer entsprechenden F -Verteilung testen, ob gewisse Gruppen von erklärenden Variablen weggelassen werden können.

9.4 Review / Lernziele



- ☐ Sie kennen das einfache und das multiple lineare Regressionsmodell und die entsprechenden Modellannahmen.
- ☐ Sie wissen, nach welchem Kriterium die Parameterschätzer ermittelt werden und wie man mit diesen Tests durchführen kann und Vertrauensintervalle ermittelt werden können.
- ☐ Sie wissen, wie ein einzelner Parameter im multiplen Regressionsmodell interpretiert wird und wie man simultan alle Koeffizienten testen kann.

Teil III

Anhänge

A Zusammenfassungen und Tabellen

A.1 Die wichtigsten eindimensionalen Verteilungen

Beachte $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$.

Verteilung	$p(x)$ bzw. $f(x)$	W_X	$\mathbb{E}[X]$	$\text{Var}(X)$
Bernoulli (p)	$p^x(1-p)^{1-x}$	$\{0, 1\}$	p	$p(1-p)$
Bin (n, p)	$\binom{n}{x} p^x (1-p)^{n-x}$	$\{0, \dots, n\}$	np	$np(1-p)$
Geom (p)	$p(1-p)^{x-1}$	$\{1, 2, \dots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Pois (λ)	$e^{-\lambda} \frac{\lambda^x}{x!}$	$\{0, 1, \dots\}$	λ	λ
Uni (a, b)	$\frac{1}{b-a}$	$[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp (λ)	$\lambda e^{-\lambda x}$	\mathbb{R}_+	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(α, λ)	$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$	\mathbb{R}_+	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	\mathbb{R}	μ	σ^2

A.2 Die wichtigsten Rechenregeln für Erwartungswert, Varianz und Kovarianz

Folgende Rechenregeln gelten sowohl für stetige wie auch für diskrete Zufallsvariablen.

1. $\mathbb{E}[a + bX] = a + b \cdot \mathbb{E}[X]$, $a, b \in \mathbb{R}$
2. $\mathbb{E}[a + bX + cY] = a + b \cdot \mathbb{E}[X] + c \cdot \mathbb{E}[Y]$, $a, b, c \in \mathbb{R}$ (egal ob X, Y unabhängig sind oder nicht)
3. $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ (hilft oft bei der Berechnung der Varianz)
4. $\text{Var}(a + bX) = b^2 \text{Var}(X)$, $a, b \in \mathbb{R}$ (konstanter Term hat keinen Einfluss, Skalierung mit b wirkt sich mit b^2 auf die Varianz aus)
5. $\sigma_{a+bX} = |b|\sigma_X$, $b \in \mathbb{R}$ (Vorzeichen spielt keine Rolle).
6. $\text{Var}(a) = 0$, $a \in \mathbb{R}$ (Varianz einer Konstanten ist 0)
7. $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
8. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
9. $\text{Cov}(X, X) = \text{Var}(X)$ (Kovarianz mit sich selber ist Varianz)
10. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$
11. $\text{Cov}(X, a) = 0$, $a \in \mathbb{R}$
12. $\text{Cov}(a + bX, c + dY) = bd \text{Cov}(X, Y)$, $a, b, c, d \in \mathbb{R}$
13. $\text{Corr}(a + bX, c + dY) = \text{sign}(b) \text{sign}(d) \text{Corr}(X, Y)$, $a, b, c, d \in \mathbb{R}$
14. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$
15. Sind X und Y **unabhängig**, so gilt
 - $\text{Cov}(X, Y) = 0$
 - $\text{Corr}(X, Y) = 0$

Achtung: Die Umkehrung gilt im Allgemeinen *nicht*! D.h. aus Unkorreliertheit folgt *nicht* Unabhängigkeit.

16. Sind X und Y **unabhängig** (oder allgemeiner: unkorreliert), so gilt
 - $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
 - $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$ (!).
 - $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Oder etwas allgemeiner für mehrere Zufallsvariablen.

17. $\mathbb{E}\left[a_0 + \sum_{i=1}^n a_i X_i\right] = a_0 + \sum_{i=1}^n a_i \mathbb{E}[X_i]$, $a_i \in \mathbb{R}$
18. $\text{Cov}\left(a_0 + \sum_{i=1}^n a_i X_i, b_0 + \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, Y_j)$, $a_i, b_j \in \mathbb{R}$
(konstanter Term fällt weg, alle Kombinationen werden aufsummiert)
19. $\text{Var}\left(a_0 + \sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j)$, $a_i \in \mathbb{R}$
(konstanter Term fällt weg, Kovarianz aller Kombinationen werden aufsummiert)

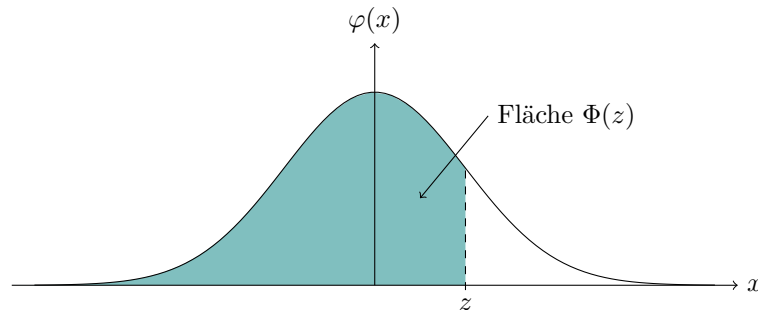
20. Sind X_1, \dots, X_n **unabhängig** (oder allgemeiner: unkorreliert), so gilt

$$\text{Var} \left(a_0 + \sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n a_i^2 \text{Var} (X_i)$$

für $a_i \in \mathbb{R}$ (konstanter Term fällt weg, es verbleiben die Summen der Varianzen)

A.3 Tabelle der Standardnormalverteilung

$$\Phi(z) = \mathbb{P}(Z \leq z), \quad Z \sim \mathcal{N}(0, 1)$$



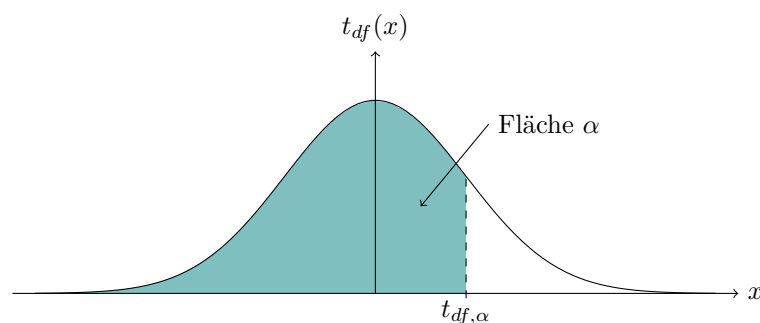
Lesebeispiel Tabelle: $\mathbb{P}(Z \leq 1.96) = 0.975$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Bemerkung:

Die Quantile der Standardnormalverteilung findet man auch direkt bei den Quantilen der t -Verteilung bei $df = \infty$, siehe Tabelle auf Seite 127.

A.4 Quantile der t -Verteilung



Lesebeispiel Tabelle: $t_{9, 0.975} = 2.262$

$df \setminus \alpha$	0.60	0.70	0.80	0.90	0.95	0.975	0.99	0.995
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
90	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

Bemerkung:

Die Zeile mit $df = \infty$ enthält gerade die Quantile der Standardnormalverteilung.

B Alternative Ansätze

B.1 Dialog: Dr. Nulli vs. Prof. Altmeier

Folgender Text orientiert sich sehr stark an einem Beispiel in [Freedman et al. \(2007\)](#).

Dr. Nulli und Prof. Altmeier haben sich mit ihrem Forschungsgeld das neueste Wunderwerk gekauft, den sogenannten Normalisator, welcher standardnormalverteilte Zufallsvariablen generieren kann. Prof. Altmeier hat auf seiner letzten Konferenz von anderen Forschern gehört, dass die Eichung der Maschine bzgl. Erwartungswert ab Werk oft mangelhaft ist¹. Dr. Nulli hingegen ist fest überzeugt davon, dass alles im Lot ist², schliesslich war es ja seine Idee, die Maschine anzuschaffen. Immerhin bei der Standardabweichung gibt es keine Streitigkeit. Jetzt wollen sie endlich Klarheit schaffen: die beiden entlocken dem Normalisator 50 Zufallszahlen. Das arithmetische Mittel davon ist 0.35.

N: Siehst Du, das arithmetische Mittel ist fast Null, genau wie ich es erwartet hatte!

A: Nein, das arithmetische Mittel weicht in der Tat von Null ab, wie *ich* es vermutet hatte!

N: Moment! Wir müssen unterscheiden zwischen dem Modell der Maschine und dem, was wir in den von ihr generierten Daten sehen. Auch wenn die Maschine richtig kalibriert ist, werden wir nie exakt ein arithmetisches Mittel von Null beobachten. Die Maschine ist auf Standardabweichung 1 eingestellt, 95% der generierten Werte liegen also zwischen -1.96 und 1.96 . Ein Wert von 0.35 ist also bestens durch Zufall erklärbar.

A: Wir schauen hier aber das arithmetische Mittel von 50 Zahlen an! Das hat eine Standardabweichung (welche hier auch Standardfehler heisst) von $1/\sqrt{50} = 0.14$, was viel kleiner ist als die Standardabweichung einer einzelnen Zufallsvariable.

N: Ah ja, das gute \sqrt{n} -Gesetz. Aber auch das arithmetische Mittel streut um den wahren Wert. Wir werden nie exakt Null beobachten, selbst wenn die Maschine richtig kalibriert ist. Ich bleibe bei meiner Erklärung durch Zufall.

A: Wenn wir aber den realisierten Wert 0.35 mit dem Standardfehler von 0.14 vergleichen, dann liegt 0.35 mehr als zwei Standardfehler von Deiner Annahme (Null) entfernt!

N: Findest Du das zu gross?

A: Ja, wenn Du daran glaubst, dass die Maschine richtig eingestellt ist, dann erwartest Du 95% der Werte im Bereich von zwei Standardabweichungen um Null³, genauer: $\pm 1.96 \cdot 0.14 = \pm 0.28$. Somit gehört unsere Beobachtung von 0.35 zu den 5% extremsten Werten. Eine solche Abweichung bei einer korrekt geeichten Maschine kann man also nicht mehr gut mit Zufall erklären, das ist schlicht und einfach ein zu seltenes Ereignis. Wir schliessen also daraus, dass die Maschine nicht richtig geeicht ist.

N: Ok, das tönt plausibel. Aber ich könnte ja auch Pech⁴ haben, dass wir jetzt gerade etwas seltenes beobachtet haben, obwohl die Maschine richtig eingestellt ist?

A: Ja, das kann in der Tat passieren, schliesslich haben wir es ja mit zufälligen Daten zu tun. Mit obiger Entscheidungsregel bist Du aber vor solchen Fehlentscheidungen geschützt im Sinne, dass dies nur mit Wahrscheinlichkeit⁵ 5% passiert.

¹Alternativhypothese H_A

²Nullhypothese H_0

³Annahmebereich

⁴Fehler 1. Art

⁵Signifikanzniveau

C Herleitungen

C.1 Herleitung der Binomialverteilung

Wir betrachten unabhängige Experimente mit Ausgang Erfolg oder Misserfolg. Die Erfolgswahrscheinlichkeit in einem Experiment sei $p \in (0, 1)$.

Frage: Was ist die Wahrscheinlichkeit, dass wir im Total x Erfolge beobachten? Z.B. $x = 3$?

Wenn wir uns festgelegt haben, bei welchen der Experimente Erfolg eintritt, so ist die Wahrscheinlichkeit für genau eine solche Auswahl

$$p^x(1-p)^{n-x}$$

da die Experimente als unabhängig angenommen wurden. In untenstehender Tabelle haben wir ein Feld eines “Experiments” mit dem Symbol \bullet markiert wenn Erfolg eintritt und sonst mit dem Symbol \circ .

1	2	3	4	5	6	$n-1$	n
\bullet	\circ	\bullet	\circ	\circ	\bullet	\circ	\circ	\circ	\circ

Um die Wahrscheinlichkeit zu berechnen, dass im Total x Erfolge eintreten, müssen wir alle “Auswahlen” betrachten, die zu diesem Ergebnis führen. Die Reihenfolge innerhalb einer Auswahl spielt keine Rolle, d.h. es interessiert uns nicht, ob zuerst Experiment 4 und erst dann Experiment 1 Erfolg hat oder umgekehrt. In der Tabelle interessieren uns daher nur die verschiedenen “Muster” und nicht, in welcher spezifischer Reihenfolge wir ein einzelnes Muster “angemalt” haben.

Um den ersten Erfolg zu platzieren, haben wir n Möglichkeiten, für den zweiten verbleiben noch $n-1$ und so weiter; bis für den letzten dann noch $n-x+1$ Möglichkeiten übrig sind. Das gibt im Total $n(n-1)\cdots(n-x+1)$ Möglichkeiten.

Hier haben wir aber jeweils stillschweigend unterschieden, in welcher Reihenfolge die Erfolge eintreten. In obenstehender Tabelle hätten wir jeweils die Auswahlen $1 \rightarrow 4 \rightarrow 6$, $1 \rightarrow 6 \rightarrow 4$, $4 \rightarrow 1 \rightarrow 6$, $4 \rightarrow 6 \rightarrow 1$, $6 \rightarrow 1 \rightarrow 4$ und $6 \rightarrow 4 \rightarrow 1$ einzeln gezählt, obwohl wir dies ja eigentlich nicht unterscheiden wollen, da alle zum selben Muster führen.

Für eine gegebene Auswahl gibt es $x!$ verschiedene mögliche Reihenfolgen, diese zu platzieren. Also haben wir genau so viel Mal zu viel gezählt.

Wenn wir dies korrigieren, erhalten wir

$$\frac{n(n-1)\cdots(n-x+1)}{x!}$$

verschiedene Möglichkeiten. Dies können wir auch schreiben als

$$\frac{n!}{x!(n-x)!}$$

was wir mit dem Binomialkoeffizienten $\binom{n}{x}$ abkürzen (“ n tief x ”).

Wir haben $\binom{n}{x}$ verschiedene Möglichkeiten, die alle zum Resultat “im Total x Erfolge” führen. Jede dieser Möglichkeiten hat die gleiche Wahrscheinlichkeit $p^x(1-p)^{n-x}$.

Die Wahrscheinlichkeit, im Total x Erfolge zu beobachten, ist also damit durch

$$\binom{n}{x} p^x (1-p)^{n-x}$$

gegeben.

C.2 Uneigentliche Integrale

In der Wahrscheinlichkeitsrechnung treten häufig Integrale auf mit einem Integrationsbereich, der von 0 nach ∞ geht oder sogar von $-\infty$ nach ∞ .

Für eine Dichte fordern wir z.B., dass

$$\int_{-\infty}^{\infty} f(x) \, dx = 1$$

gilt. Die totale Fläche unter der Kurve soll also 1 sein. Man integriert hier nicht über ein beschränktes Intervall und man spricht daher von einem *uneigentlichen Integral* (nicht zu verwechseln mit dem unbestimmten Integral).

Wir beginnen mit einem einfachen Fall. Nehmen wir z.B. die Exponentialverteilung mit Parameter $\lambda > 0$. Diese hat Dichte

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Wir wollen nun überprüfen, dass $f(x)$ überhaupt eine Dichte ist. Gemäss Definition einer Dichte muss das Integral über den Wertebereich 1 ergeben, d.h. hier

$$\int_0^{\infty} f(x) \, dx = 1.$$

Wie ist dieses Integral genau zu verstehen und wie berechnet man es? Das Integral

$$\int_0^{\infty} f(x) \, dx$$

ist ein uneigentliches Integral und ist definiert als

$$\int_0^{\infty} f(x) \, dx = \lim_{a \rightarrow \infty} \int_0^a f(x) \, dx.$$

Wenn der Grenzwert existiert, dann heisst das uneigentliche Integral konvergent und der Grenzwert stellt den Wert des uneigentlichen Integrals dar.

D.h. auf der rechten Seite liegt auch gerade der Schlüssel zur Berechnung. Wir berechnen das Integral auf dem Intervall $[0, a]$ “wie gewohnt” und ziehen dann den Limes.

Für obige Exponentialverteilung haben wir

$$\int_0^a \lambda e^{-\lambda x} \, dx = -e^{-\lambda x} \Big|_0^a = -e^{-\lambda a} + 1.$$

Wenn wir jetzt den Limes $a \rightarrow \infty$ ziehen, so haben wir

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

da $\lim_{a \rightarrow \infty} e^{-\lambda a} = 0$.

In diesem Beispiel war die untere Integrationsgrenze 0, was uns Arbeit erspart hat.

Was ist, falls dem nicht so ist? Wir teilen das Integral an einer (beliebigen) Stelle c und haben so wieder die Situation von vorher.

$$\int_{-\infty}^{\infty} f(x) \, dx = \lim_{a \rightarrow -\infty} \int_a^c f(x) \, dx + \lim_{b \rightarrow \infty} \int_c^b f(x) \, dx.$$

Das heisst implizit, dass wir die beiden Grenzen *unabhängig voneinander* nach $\pm\infty$ gehen lassen:

$$\lim_{\substack{a \rightarrow -\infty \\ b \rightarrow \infty}} \int_a^b f(x) \, dx.$$

Man darf *nicht*

$$\lim_{a \rightarrow \infty} \int_{-a}^a f(x) \, dx$$

verwenden, da dies zu falschen Resultaten führen kann. Betrachte z.B. die Funktion $f(x) = x$. Mit dieser falschen Rechnung wäre das Integral 0, obwohl die beiden uneigentlichen Integrale

$$\int_{-\infty}^0 x \, dx \quad \text{bzw.} \quad \int_0^{\infty} x \, dx$$

gar nicht existieren.

In der Praxis schreiben wir also die Stammfunktion auf und lassen zuerst die obere Grenze b nach ∞ gehen und dann entsprechend die untere Grenze a nach $-\infty$ (bzw. umgekehrt).

Betrachten wir z.B. das uneigentliche Integral

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} \, dx.$$

Wir haben

$$\int_a^b \frac{1}{1+x^2} = \arctan(x) \Big|_a^b = \arctan(b) - \arctan(a).$$

Es ist

$$\begin{aligned} \lim_{b \rightarrow \infty} \arctan(b) &= \frac{\pi}{2} \\ \lim_{a \rightarrow -\infty} \arctan(a) &= -\frac{\pi}{2}. \end{aligned}$$

Also haben wir schlussendlich

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} \, dx = \frac{1}{\pi} \left(\frac{\pi}{2} - \left(-\frac{\pi}{2} \right) \right) = 1.$$

Bei der Funktion handelt es sich um die Dichte der sogenannten Cauchy-Verteilung.

Literaturverzeichnis

Freedman, D., Pisani, R. and Purves, R. (2007) *Statistics*. W.W. Norton & Company.

Stahel, W. (2007) *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*. Vieweg + Teubner Verlag.

Index

A

Alternativhypothese, 77
Annahmebereich, 79
Approximation
 Normal-, 61
 Poisson-, 24
Arithmetisches Mittel, 37
Ausfallrate, 30
Ausreisser, 40
Axiome der Wahrscheinlichkeitsrechnung, 5

B

bayes'sch, 5
bedingt
 Dichte, 50
 Erwartungswert, 47, 50
 Verteilung, 47
 Wahrscheinlichkeit, 8
Beobachtungsstudien, 103
Bernoulliverteilung, 19
bimodal, 40
Binomialtest, 20
Binomialverteilung, 20
 Normalapproximation der B., 61
 Poissonapproximation der B., 24
Bonferroni-Korrektur, 90
Boxplot, 40

C

confounder, 103

D

De Morgan'sche Regeln, 4
Designmatrix, 118
Dichte, 25
 bedingte, 50
 einer Summe von ZV., 55
 gemeinsame, 48
 Rand-, 50
Differenz von Mengen, 4
disjunkt, 4
doppelblind, 103
Durchschnitt von Mengen, 4

E

Elementarereignis, 3

empirisch, 37
Ereignis, 3
erklärende Variable, 109
erwartungstreu, 75
Erwartungswert, 18
 bedingter, 47, 50
 bei mehreren Zufallsvariablen, 50
Exponentialverteilung, 29

F

Faltung, 55
Fehler
 1. Art, 78
 2. Art, 78
Freiheitsgrad, 82
frequentistisch, 5
F-Test, 119

G

gemeinsam
 Dichte, 48
 Verteilung, 47
 Wahrscheinlichkeitsfunktion, 47
Geometrische Verteilung, 20
gepaart, 101
Gesetz der grossen Zahlen, 60
Globaltest, 119
Grundgesamtheit, 37
Grundraum, 3

H

Histogramm, 39

I

i.i.d. Annahme, 59
Indikatorfunktion, 55
interquartile range, 38
IQR, 38

J

Jensen'sche Ungleichung, 32

K

Kleinste-Quadrate Schätzer, 110
Komplement von M., 4
Korrelation, 52

empirische, 41
 Kovarianz, 52
 empirische, 41
 kumulative Verteilungsfunktion, 16
 empirische, 40
 kurzschwänzig, 69

L
 Lageparameter, 18
 langschwänzig, 67, 82
 leere Menge, 4
 Likelihoodfunktion, 72
 log-, 72
 lineare Regression
 einfache, 109
 multiple, 117
 Lognormalverteilung, 32

M
 Macht, 84
 Mann-Whitney U-Test, 107
 Maximum-Likelihood Methode, 72
 Median, 27
 empirischer, 38
 Mengen
 De Morgan'sche Regeln, 4
 Differenz von M., 4
 disjunkte, 4
 Durchschnitt von M., 4
 Komplement von M., 4
 Operationen der Mengenlehre, 4
 Vereinigung von M., 4
 Mittel
 arithmetisches, 37
 Modell von Laplace, 7
 Momentenmethode, 71

N
 Normalverteilung, 28
 Standard-, 29
 zweidimensionale, 54
 Nullhypothese, 77

O
 Operationen der Mengenlehre, 4

P
 Partition, 12
 Poissonprozess, 33
 homogener, 33
 Intensität eines P., 33
 Poissonverteilung, 23
 Normalapproximation der P., 62
 Prognoseintervall, 115
 p-Wert, 87

Q
 QQ-Plot, 66
 Normalplot, 67
 Quantil, 26
 empirisches, 38
 Quartil, 38
 Quartilsdifferenz, 38

R
 Rand-
 Dichte, 50
 Verteilung, 47
 Randomisierung, 102
 Relevanzbereich, 93
 Residuum, 111
 robust, 38

S
 Satz der totalen Wahrscheinlichkeit, 11
 Satz von Bayes, 13
 Schätzer, 70
 allgemeine für Erw.wert und Varianz, 74
 erwartungstreuer, 75
 Genauigkeit von S., 75
 Maximum-Likelihood Methode, 72
 Momentenmethode, 71
 Schwerpunkt, 18
 Signifikanz, 79
 Signifikanzniveau, 78
 Simulation, 33
 Standardabweichung, 19
 empirische, 38
 Standardfehler, 76, 113
 Standardisierung, 31
 Stichproben, 37
 gepaarte, 101
 ungepaarte oder unabhängige, 101
 Zufalls-, 37
 Stichprobenmittel, 37
 Störparameter, 110
 Streuung, 18
 Streuungsparameter, 19

T
 Tabellen
 Standardnormalverteilung, 126
 t-Verteilung, 127
 Teststatistik, 81
 Trägheitsmoment, 19
 t-Test
 für eine Stichprobe, 82
 für unabhängige Stichproben, 105
 Tukey-Anscombe-Plot, 116
 t-Verteilung, 82

U

Unabhängigkeit

von Ereignissen, 8

von Stichproben, 101

von Zufallsvariablen, 15, 49

ungepaart, 101

Uniforme Verteilung, 27

Ursache-Wirkung Beziehung, 103

V

Varianz, 19

empirische, 38

Vereinigung, 4

Versuchsplanung, 102

Verteilung, 15

bedingte, 47

Bernoulli-, 19

bimodale, 40

diskrete, 16

Exponential-, 29

gemeinsame, 47

geometrische, 20

kurzschwänzige, 69

langschwänzige, 67

Lognormal-, 32, 62

Rand-, 47

schiefe, 69

Standardnormal-, 29

stetige, 25

t-, 82

uniforme, 27

zweidimensionale Normal-, 54

Verteilungsfamilie, 66

Vertrauensintervall, 76, 90

Dualität zu Tests, 90

Verwerfungsbereich, 79

Vorzeichen-Test, 93

W

Wahrscheinlichkeit

bayes'sche Interpretation der W., 5

bedingte, 8

frequentistische Interpretation der W., 5

Wahrscheinlichkeitsbaum, 11

Wahrscheinlichkeitsfunktion, 16

gemeinsame, 47

Wahrscheinlichkeitsmodell

diskretes, 6

Wahrscheinlichkeitsverteilung, 15

Welch-Test, 107

Wiederkehrperiode, 22

Wilcoxon-Test, 94

Z

Zentraler Grenzwertsatz, 60

Zentralwert, 38

Zielvariable, 109

Z-Test, 81

Zufallsexperiment, 3

Zufallsstichprobe, 37

Zufallsvariable, 15

arithm. Mittel von Z.n, 59

Summen von Z.n, 59

zweidimensionale Normalverteilung, 54