Einführung

Statistik und Wahrscheinlichkeitsrechnung

Lukas Meier

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

Fehler, Kommentare und Anregungen bitte melden an Lukas Meier: meier@stat.math.ethz.ch

Inhaltsverzeichnis

1 Einführung

Durch immer mehr und immer einfacher verfügbare Daten stellt sich die Frage einer adäquaten Modellierung und Auswertung häufiger denn je. Aus unseren Daten wollen wir (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 1.1 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. So 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, 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 1.1 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. 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 "günstigen Fälle" und die Anzahl "möglichen 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.

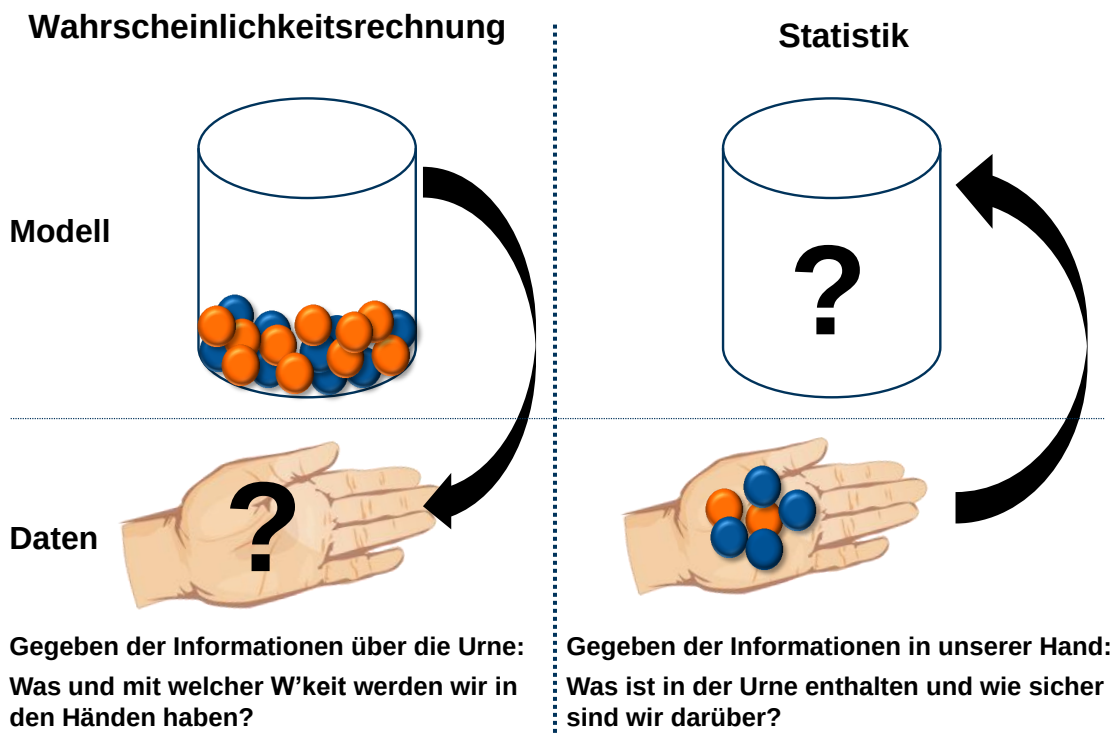


Abbildung 1.1: Darstellung der Konzepte der Wahrscheinlichkeitsrechnung und der Statistik. Das Modell wird hier durch eine Urne symbolisiert.

2 Grundlagen der Wahrscheinlichkeitsrechnung

2.1 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 $A \subset \Omega$. "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 2.1: Operationen der Mengenlehre und ihre Bedeutung.

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** (oder 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 2.1.

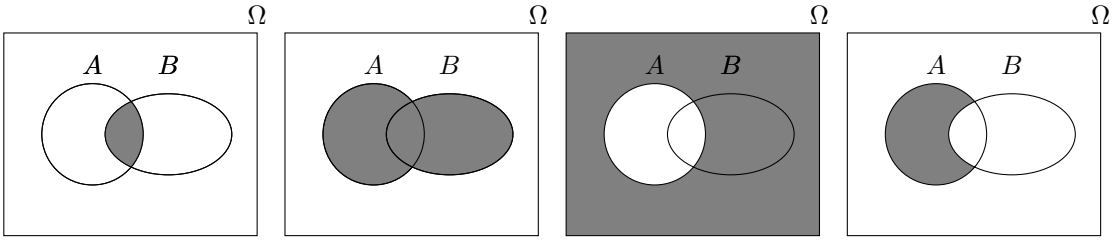


Abbildung 2.1: Illustration der Operationen der Mengenlehre an Venn-Diagrammen: $A \cap B$, $A \cup B$, A^c und $A \setminus B$ jeweils entsprechend markiert (von links nach rechts).

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). Der Durchschnitt $A \cap B$ ist das Ereignis "Beide Stahlträger haben strukturelle Mängel", A^c bedeutet, dass Stahlträger 1 keine Mängel aufweist, etc.*

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 $A, B, C, \ldots$ 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 w 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) $0 \leq \mathbb{P}(A) \leq 1$

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

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

(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.

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \quad \text{für jedes Ereignis } A \tag{2.1}$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \quad \text{für je zwei Ereignisse } A \text{ und } B \tag{2.2}$$

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq \mathbb{P}(A_1) + \ldots + \mathbb{P}(A_n) \quad \text{für je } n \text{ Ereignisse } A_1, \ldots, A_n \tag{2.3}$$

$$\mathbb{P}(B) \leq \mathbb{P}(A) \quad \text{für je zwei Ereignisse } A \text{ und } B \text{ mit } B \subseteq A \tag{2.4}$$

$$\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(B) \quad \text{für je zwei Ereignisse } A \text{ und } B \text{ mit } B \subseteq A \tag{2.5}$$

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 2.2. 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

$$f_n(A) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(A)\,.$$

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,

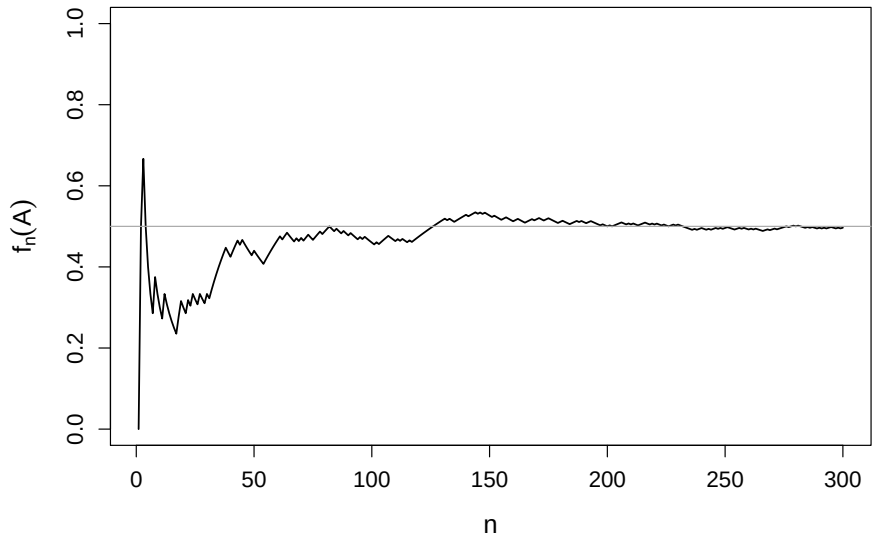


Abbildung 2.2: Relative Häufigkeiten $f_n(A)$ für das Ereignis A="Münze zeigt Kopf" beim Wurf mit einer Münze in Abhängigkeit der Anzahl Würfe n.

dass wenn sie auf ihrem Grundstück viele Bohrungen machen, dass dann 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.

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

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

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 *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 $\{w_k\}$ um zu unterstreichen, dass wir eine *Menge* (d.h. ein *Ereignis*) meinen mit *einem* Element ω_k.

Ferner gilt

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

Die Summe 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. Dies ist z.B. beim Beispiel mit dem Münzwurf 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

$$\mathbb{P}(\{\omega_k\}) = \frac{1}{|\Omega|}.$$

für alle $k \geq 1$.

Für ein Ereignis A gilt also im Laplace-Modell

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

Dies kennen sie vermutlich aus der Mittelschule. Dort bestand dann die Wahrscheinlichkeitsrechnung oft darin, durch (mühsames) Abzählen die Anzahl günstiger Fälle zu bestimmen. Wie wir aber sehen werden, geht die Wahrscheinlichkeitsrechnung weit über das Laplace-Modell hinaus. Insbesondere ist das Laplace-Modell für viele Anwendungen ungeeignet.

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

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

2.2 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.

Ein 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 wo A eintritt (ca. $1/3$) und *davon* nochmals diejenigen, wo

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 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*, indem man verifiziert, dass zwischen zwei Ereignissen A und B kein kausaler Zusammenhang besteht (d.h. es gibt *keine* gemeinsamen Ursachen oder Ausschliessungen).

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 in obigem Beispiel bei einem Ausfall des einen Bauteils das andere Bauteil auch gerade ausfällt (also ist die Unabhängigkeit nicht mehr gegeben), dann steigt die Ausfallwahrscheinlichkeit des Geräts auf 1/100 (da in diesem Fall $A_1 = A_2$ und somit $A_1 \cap A_2 = A_1 = A_2$)!

Wenn man also Ausfallwahrscheinlichkeiten unter der Annahme von Unabhängigkeit berechnet, aber diese in der Realität nicht erfüllt ist, so ist das Resultat oft um einige Grössenordnungen zu klein!

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

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

2.3 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, dann werden wir vermutlich annehmen, dass dann der zweite Träger eher auch betroffen ist (da sie aus der selben Produktion stammen und den gleichen Witterungsbedingungen ausgesetzt waren etc.). Die Wahrscheinlichkeit für den zweiten Träger (dessen Zustand wir noch nicht kennen) würden wir dann nach Erhalt der Information über den ersten Träger höher einschätzen als ursprünglich.*

Dies führt zum Konzept der bedingten Wahrscheinlichkeiten. Diese treten 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".

2.3.1 Definition der bedingten Wahrscheinlichkeit

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

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

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

Wie kann man die Formel verstehen? Da wir wissen, dass B schon eingetreten ist (wir haben also ein neues $\Omega' = B$), müssen wir nur noch Ereignisse darin anschauen wenn wir uns für A interessieren (daher $A \cap B$). Die Normierung mit $\mathbb{P}(B)$ sorgt dafür, dass $\mathbb{P}(\Omega') = \mathbb{P}(B) = 1$. Dies ist auch in Abbildung 2.3 illustriert. Wenn man wieder mit Flächen denkt, dann ist die bedingte Wahrscheinlichkeit $\mathbb{P}(A \mid B)$ die schraffierte Fläche dividiert durch die Fläche von B.

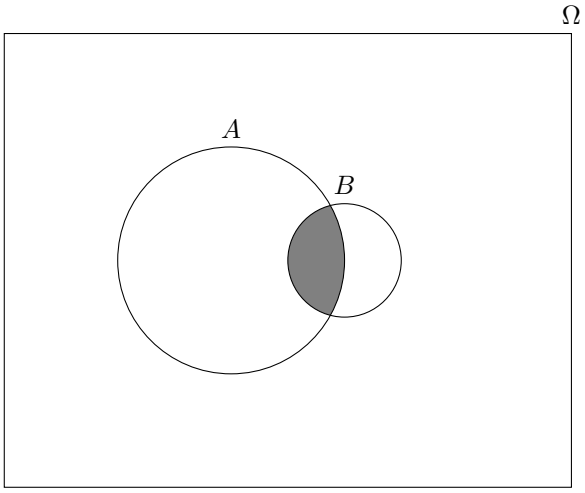


Abbildung 2.3: Hilfsillustration für bedingte Wahrscheinlichkeiten.

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?*

Es ist $\Omega = \{1, \dots, 6\}$*,* $A = \{6\}$ *und* $B = \{2, 4, 6\}$*. Also ist* $A \cap B = \{6\}$*. Weiter ist* $\mathbb{P}(B) = 3/6 = 1/2$*. Also haben wir damit*

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

Durch die zusätzliche Information (gerade Augenzahl) hat sich die Wahrscheinlichkeit also geändert.

Wenn wir B *festhalten*, dann gelten für die bedingte Wahrscheinlichkeit $\mathbb{P}(A \mid B)$ die normalen Rechenregeln von früher.

Rechenregeln

$0 \leq \mathbb{P}(A \mid B) \leq 1$	für jedes Ereignis A
$\mathbb{P}(B \mid B) = 1$	
$\mathbb{P}(A_1 \cup A_2 \mid B) = \mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B)$	für A_1, A_2 disjunkt
$\mathbb{P}(A^c \mid B) = 1 - \mathbb{P}(A \mid B)$	für jedes Ereignis A

So lange man am "bedingenden Ereignis" B nichts ändert, kann man also mit bedingten Wahrscheinlichkeiten wie gewohnt rechnen.

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

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A) \tag{2.6}$$

Deshalb können wir die Unabhängigkeit auch folgendermassen definieren:

$$A, B \text{ unabhängig} \quad \Longleftrightarrow \quad \mathbb{P}(A \mid B) = \mathbb{P}(A) \quad \Longleftrightarrow \quad \mathbb{P}(B \mid A) = \mathbb{P}(B) \tag{2.7}$$

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 z.B. 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 2.3 ersichtlich. $\mathbb{P}(A \mid B)$ ist dort viel grösser als $\mathbb{P}(B \mid A)$.

2.3.2 Satz der totalen Wahrscheinlichkeit und Satz von Bayes

Wie wir in (2.6) 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

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

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 2.4. 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 summieren.

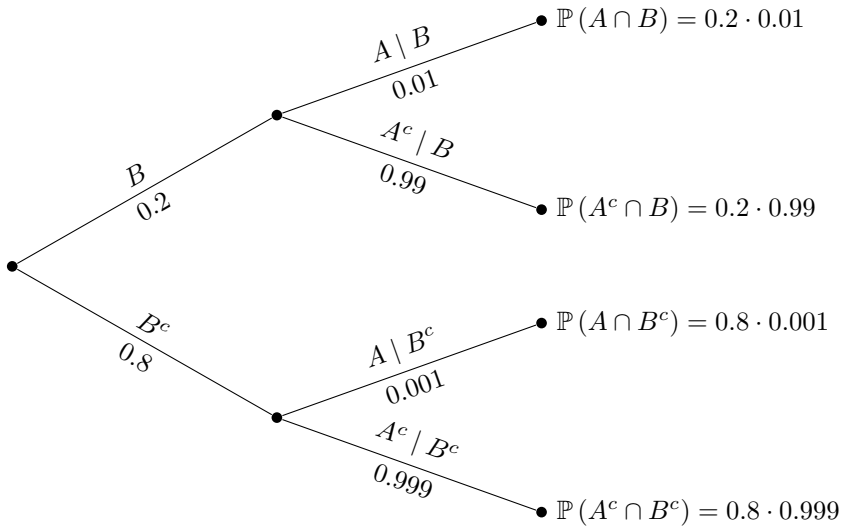


Abbildung 2.4: Wahrscheinlichkeitsbaum.

Diese Aufteilung in verschiedene Situationen (B, B^c) funktioniert ganz allgemein und führt zum Satz der totalen Wahrscheinlichkeit.

Satz der totalen Wahrscheinlichkeit

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

$$B_1 \cup \ldots \cup B_k = \Omega \qquad \text{("alle möglichen Fälle sind abgedeckt")}$$

Dann gilt

$$\mathbb{P}(A) \overset{(A3)}{=} \sum_{i=1}^{k} \mathbb{P}(A \cap B_i) \overset{(2.6)}{=} \sum_{i=1}^{k} \mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i)\,.$$

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.

$B_1, \ldots, B_k$ heisst auch **Partitionierung** von Ω. Sie deckt alle möglichen Fälle ab und zwei Ereignisse B_i und B_j können nicht zusammen eintreten. Ein Illustration einer Partitionierung findet man in Abbildung 2.5.

Manchmal will man die bedingten Wahrscheinlichkeiten auch "umkehren". Sie haben z.B. ein technisches Verfahren entwickelt, um Haarrisse in einem Flugzeugflügel zu detektieren. Wir betrachten

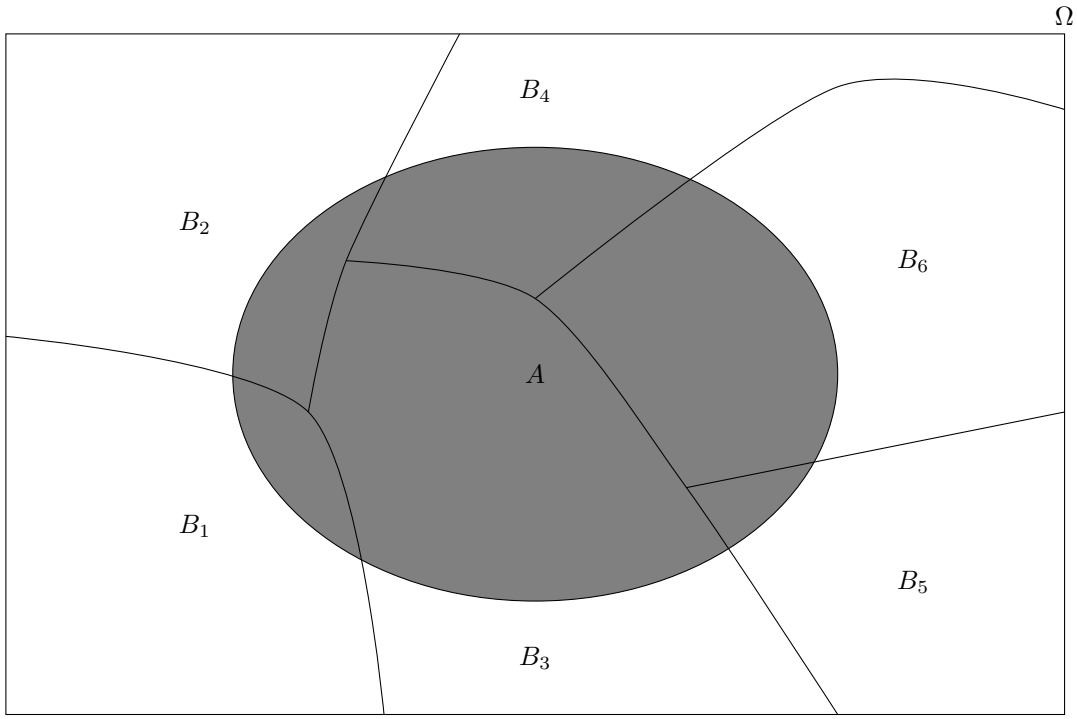


Abbildung 2.5: Illustration einer Partitionierung von Ω $(B_1, \ldots, B_6)$.

folgende Ereignisse

$$\begin{aligned} A &= \text{“Technisches Verfahren indiziert, dass Risse da sind”} \\ B_1 &= \text{“Flügel weist in der Tat Haarrisse auf”} \\ B_2 = B_1^c &= \text{“Flügel weist in der Tat } \textit{keine} \text{ Haarrisse auf”} \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}(A \mid B_1) &= 0.99 \\ \mathbb{P}(A \mid B_2) &= 0.03 \end{aligned}$$

Zudem nehmen wir an, dass gilt

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

Wenn der Flügel also tatsächlich defekt ist, 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 ein Flügel mit Wahrscheinlichkeit 0.001 überhaupt Risse aufweist (a-priori, ohne einen Test gemacht zu haben).

Die Frage ist nun: Gegeben, dass das Verfahren einen Mangel 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

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

In der Situation des Satzes der totalen Wahrscheinlichkeit haben wir

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

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

Beispiel. *In obigem Beispiel haben wir also*

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

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.

3 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), oft Zahlen. Man führt für diesen Spezialfall den Begriff der Zufallsvariable ein.

3.1 Der Begriff der Zufallsvariable

Eine **Zufallsvariable** X ist der Ausgang eines Zufallsexperiments mit möglichen Werten in $\mathbb{R}$, bzw. in einer Teilmenge von $\mathbb{R}$, z.B. $\mathbb{N}_0 = \{0, 1, \ldots\}$. 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 "Wortschwall" vorstellen (z.B. "Messung der Druckfestigkeit"). Beim Kleinbuchstaben setzt man einen konkreten Wert ein.

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.

3.1.1 Wahrscheinlichkeitsverteilungen

Von Interesse ist die Frage, mit welchen Wahrscheinlichkeiten eine Zufallsvariable in welchen Bereichen liegt. Man spricht von der sogenannten **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).

3.2 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

$W = \{0, 1, 2, \ldots, 100\}$, $W = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ oder ganz allgemein $W = \{x_1, x_2, \ldots\}$.

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

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

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

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.

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

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

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

Die ganze Funktion ist in Abbildung 3.1 dargestellt.

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.

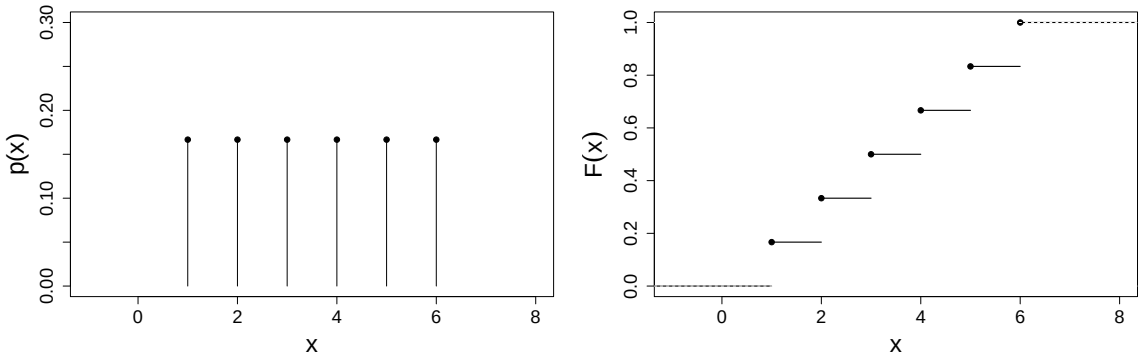


Abbildung 3.1: Wahrscheinlichkeitsfunktion (links) und kumulative Verteilungsfunktion (rechts) beim Beispiel mit dem Würfel.

Rechenregeln und Eigenschaften

Es gilt (egal ob X diskret ist oder nicht)

$$\begin{aligned}
\mathbb{P}(a < X \leq b) &= \mathbb{P}(X \in (a, b]) \\
&\overset{(2.5)}{=} \mathbb{P}(X \in (-\infty, b]) - \mathbb{P}(X \in (-\infty, a)) \\
&= F(b) - F(a) \\
\mathbb{P}(X > x) &\overset{(2.1)}{=} 1 - \mathbb{P}(X \leq x) = 1 - F(x)
\end{aligned}$$

Die kumulative Verteilungsfunktion F erfüllt zudem immer:

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

3.2.1 Erwartungswert und Varianz

Wir haben gesehen, dass die Verteilung einer diskreten Zufallsvariable durch eine (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 durch

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

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

Physikalisch 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).

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 Wiederho-

lungen". D.h. er ist eine Idealisierung des arithmetischen Mittels der Werte einer Zufallsvariablen bei vielen Wiederholungen. Also: $\mathbb{E}[X]$ ist eine Kennzahl im Modell der Wahrscheinlichkeitstheorie.

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 3.1. 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 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.

Manchmal betrachtet man statt der Zufallsvariablen X eine Transformation $f(X)$. Für den Erwartungswert einer transformierten diskreten Zufallsvariablen $Y = f(X)$ gilt

$$\mathbb{E}[Y] = \mathbb{E}[f(X)] = \sum_{k \geq 1} f(x_k) p(x_k). \tag{3.1}$$

Wieder wie vorher summiert man über "was passiert" $(f(x_k))$ × "mit welcher Wahrscheinlichkeit passiert es" $(p(x_k))$.

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

$$\mathrm{Var}(X) = \sigma_X^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \overset{(3.1)}{=} \sum_{k \geq 1} (x_k - \mu_X)^2 p(x_k).$$

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.

Die Varianz ist also ein **Mass für die Streuung der Verteilung** um die mittlere Lage, ein sogenannter **Streuungsparameter**.

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

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

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

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

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

3.2.2 Bernoulliverteilung [Bernoulli (p)]

Die **Bernoulliverteilung** 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!)

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

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

3.2.3 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 $W=\{0,1,\ldots,n\}$.

Erfolg und Experiment kann 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

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

Eine Herleitung für die Wahrscheinlichkeitsfunktion findet man in Kapitel A.5. In Abbildung 3.2 sind einige Fälle mit verschiedenen Parametern dargestellt. Für grosses n hat man schon ein ziemlich "glockenförmiges" Bild, mehr dazu später.

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; insbesondere muss man nicht mühsam die Anzahl Fälle bestimmen. Was ist z.B. die Wahrscheinlichkeit, dass von 10 Bauteilen genau 3 mangelhaft sind? Diese Wahrscheinlichkeit ist gegeben

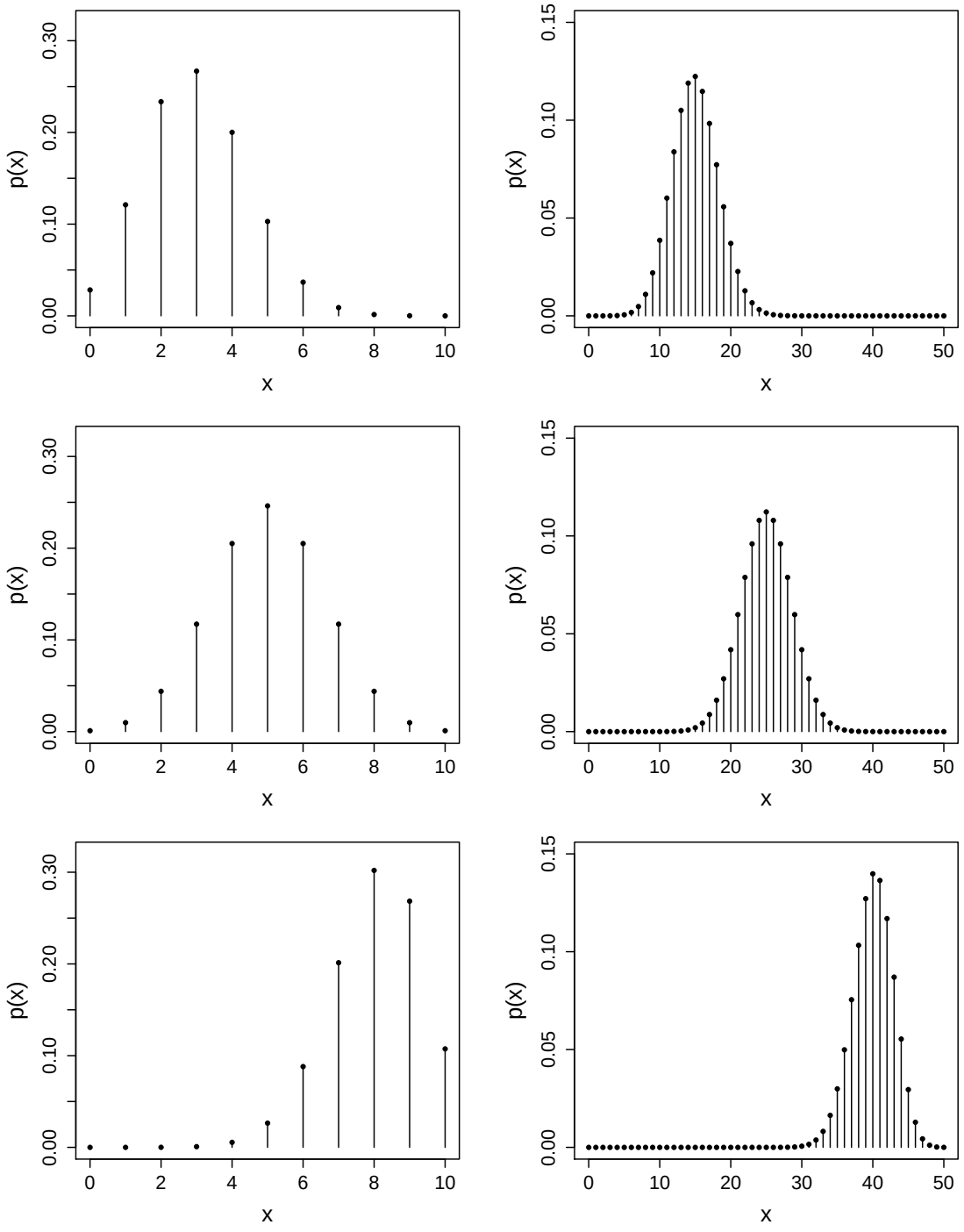


Abbildung 3.2: Wahrscheinlichkeitsfunktion der Binomialverteilung für $n = 10$ (links) und $n = 50$ (rechts) für jeweils $p = 0.3, 0.5, 0.8$ (oben nach unten).

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

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

mühsam zu bestimmen, erhalten wir direkt mit dem komplementären Ereignis

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

dass

$$\mathbb{P}(X \geq 1) \overset{(2.1)}{=} 1 - \mathbb{P}(X = 0) = 1 - p(0) = 1 - 0.95^{10} = 0.401.$$

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 $p(x)$.

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

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

$$\begin{aligned} p(x) &= p \cdot (1-p)^{x-1} \\ \mathbb{E}[X] &= \frac{1}{p} \\ \mathrm{Var}(X) &= \frac{1-p}{p^2}. \end{aligned}$$

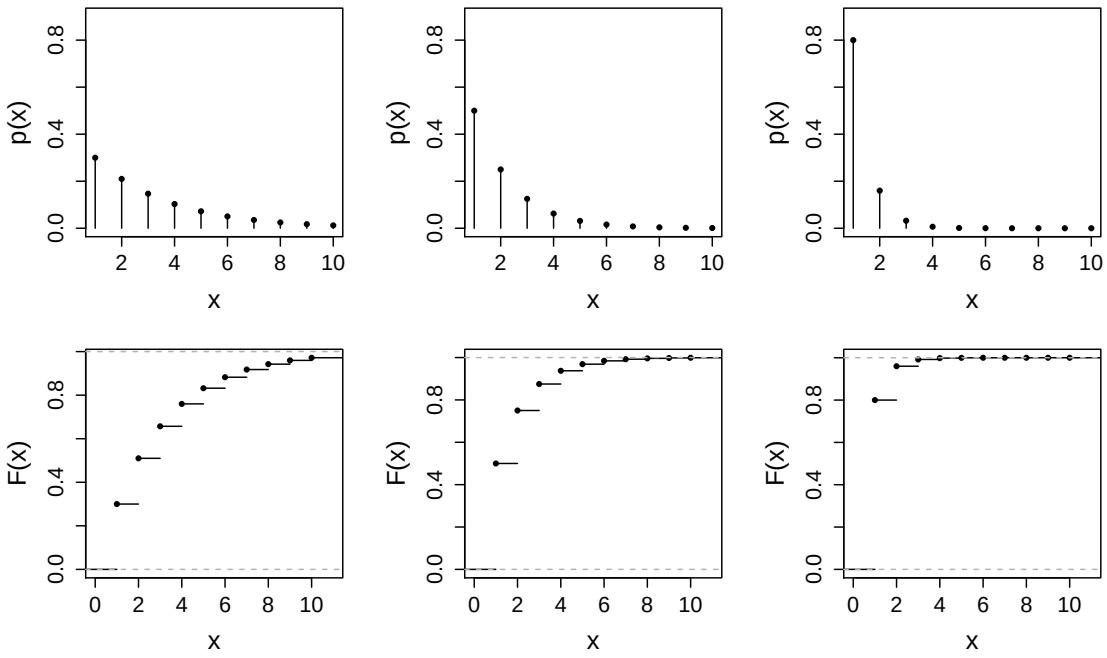


Abbildung 3.3: Wahrscheinlichkeitsfunktion (oben) und kumulative Verteilungsfunktion (unten) der geometrischen Verteilung für $p = 0.3, 0.5, 0.8$ (links nach rechts), jeweils abgeschnitten bei $x = 10$.

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

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

für $x \in W$. Dazwischen ist F *konstant*, siehe auch Abbildung 3.3.

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. das 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, so 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.

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%!

3.2.5 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 Poissonverteilung gut eignet.

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

$$\begin{aligned} p(x) &= e^{-\lambda}\frac{\lambda^x}{x!},\ x \in W \\ \mathbb{E}[X] &= \lambda \\ \mathrm{Var}(X) &= \lambda. \end{aligned}$$

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

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

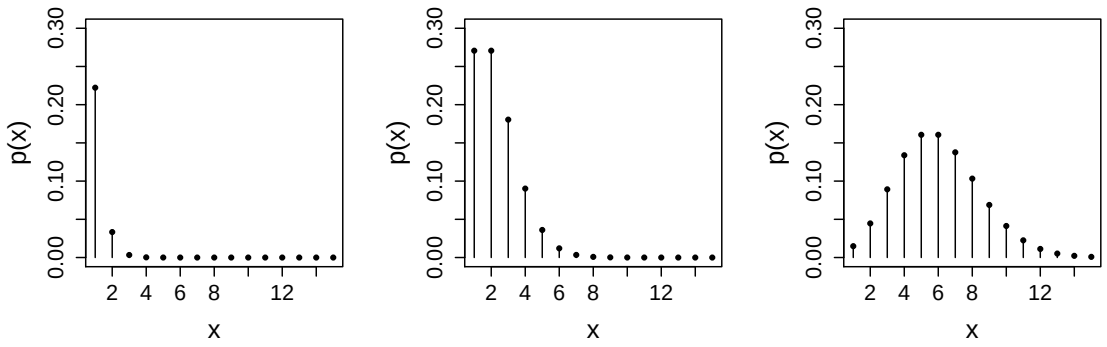


Abbildung 3.4: Wahrscheinlichkeitsfunktion der Poissonverteilung für $\lambda = 0.3, 2, 6$ (links nach rechts).

$\text{Pois}(\lambda), \lambda = 5$. *Damit können wir nun "alle" Wahrscheinlichkeiten berechnen, z.B. die Wahrscheinlichkeit, dass in einer Minute niemand anruft:*

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

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

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

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 seltene 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.

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 zwansläufig gerade addieren wegen den Rechenregeln für den Erwartungswert.

Wenn wir aber $\frac{1}{2}(X + Y)$ betrachten, so liegt nicht etwa eine 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 $\frac{1}{2}(\lambda_1 + \lambda_2)$.

3.3 Stetige Verteilungen

Eine Zufallsvariable X heisst **stetig**, falls die Menge der möglichen Werte W ein Intervall enthält, wie zum Beispiel $W = [0, 1]$ oder $W = \mathbb{R}$. Im Gegensatz zu früher haben wir hier nicht mehr einfach

eine "Liste" 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 proportional zur Intervallbreite sein, d.h. es gilt

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

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

3.3.1 Wahrscheinlichkeitsdichte

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

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

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:

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

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

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

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 3.5. 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

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

All dies folgt aus den ursprünglichen Axiomen. Man beachte insbesondere, dass es durchaus (kleine) Intervalle geben kann, wo gilt $f(x) > 1$, siehe z.B. Abbildung 3.10. Dies im Gegensatz zum diskreten Fall, wo jeweils immer gilt $0 \leq p(x_k) \leq 1$.

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.

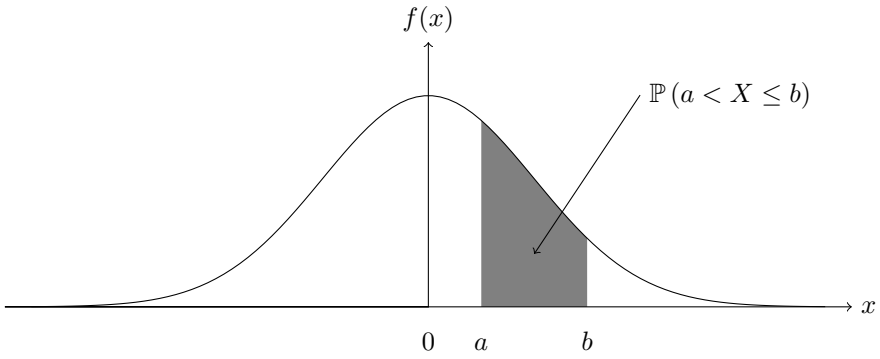


Abbildung 3.5: Illustration einer Dichte einer Zufallsvariablen und der Wahrscheinlichkeit, in das Intervall $[a, b]$ zu fallen (graue Fläche).

3.3.2 Kennzahlen von stetigen Verteilungen

Erwartungswert und Varianz

Der Erwartungswert berechnet sich im stetigen Fall als

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

Für $g(X)$ gilt analog zu früher

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

Für die Varianz haben wir entsprechend

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

Alle diese Formeln sind genau gleich wie früher: Man ersetzt die Summe durch das Integral und die Wahrscheinlichkeit $p(x)$ durch $f(x)\,\mathrm{d}x$. Es gelten insbesondere die gleichen Rechenregeln wie im diskreten Fall. Auch die Interpretation bleibt unverändert.

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

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

siehe auch Abbildung 3.6.

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.

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

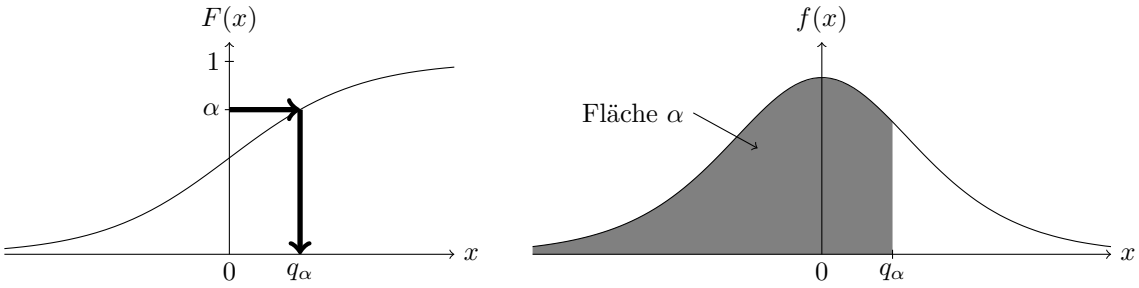


Abbildung 3.6: Illustration des Quantils q_α anhand der Verteilungsfunktion (links) und der Dichte (rechts) für $\alpha = 0.75$.

3.3.3 Uniforme Verteilung $[\mathrm{Uni}\,(a, b)]$

Die **uniforme Verteilung** mit den Parametern $a, b \in \mathbb{R}$, tritt z.B. auf bei Rundungsfehlern und als Formalisierung der völligen "Ignoranz". Hier ist $W = [a, b]$ und

$$\begin{aligned} f(x) &= \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{sonst} \end{cases} \\ F(x) &= \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b. \end{cases} \end{aligned}$$

Die Dichte ist also *konstant* und die kumulative Verteilungsfunktion eine *lineare* Funktion auf dem Definitionsbereich $[a, b]$, siehe Abbildung 3.7.

Für Erwartungswert und Varianz gilt

$$\begin{aligned} \mathbb{E}\,[X] &= \frac{a+b}{2} \\ \mathrm{Var}\,(X) &= \frac{(b-a)^2}{12}. \end{aligned}$$

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*

$$\mathbb{P}\,(2 \le X \le 4) = 2/5,$$

denn das Integral ist hier einfach eine Rechtecksfläche.

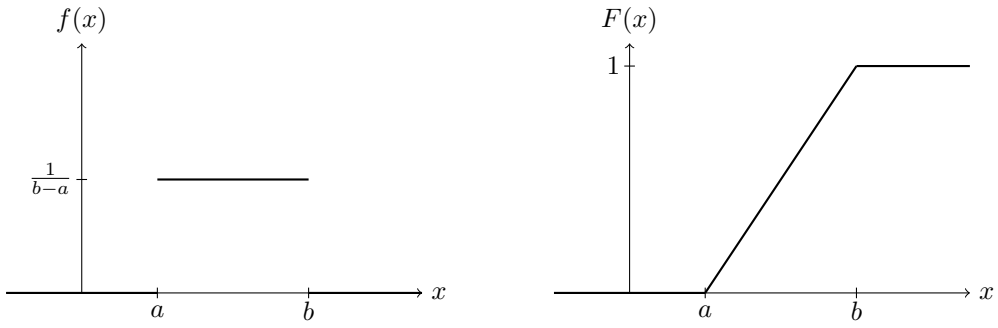


Abbildung 3.7: Dichte (links) und Verteilungsfunktion (rechts) der uniformen Verteilung.

3.3.4 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

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

mit

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

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

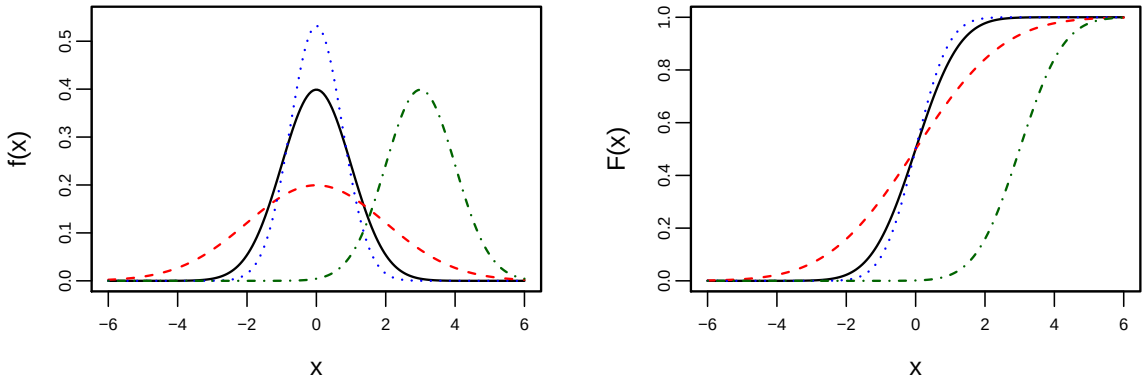


Abbildung 3.8: Dichte und Verteilungsfunktion der Normalverteilung für $\mu = 0$, $\sigma = 1$ (schwarz, durchgezogen), $\mu = 0$, $\sigma = 2$ (rot, gestrichelt), $\mu = 0$, $\sigma = 0.75$ (blau, gepunktet) und $\mu = 3$, $\sigma = 1$ (grün, strichpunktiert).

Die Dichte der Normalverteilung ist symmetrisch um den Erwartungswert μ. Je grösser σ, desto flacher/breiter wird die Dichte. Für kleine σ gibt es einen "schmalen und hohen" Gipfel. Mit μ verschieben wir einfach die Dichte nach links/rechts, siehe auch Abbildung 3.8.

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 3.9.

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%.

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 ist

$$\begin{aligned} \varphi(x) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \\ \Phi(x) &= \int_{-\infty}^{x} \varphi(u)\,\mathrm{d}u. \end{aligned}$$

Die Funktion Φ ist leider *nicht* geschlossen darstellbar. Eine Tabelle findet man in Kapitel A.2.

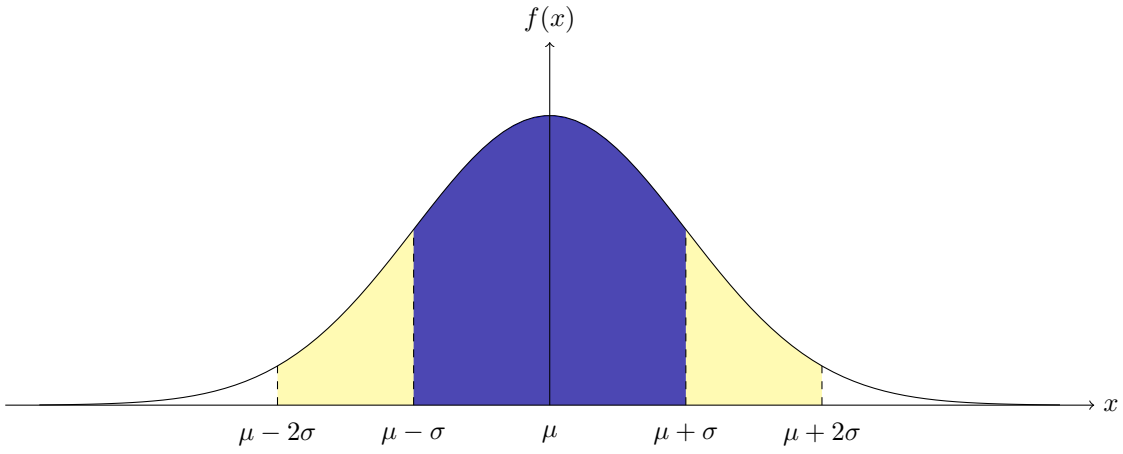


Abbildung 3.9: Dichte der Normalverteilung. Ca. 66% der Fläche befindet sich im Intervall $[\mu - \sigma, \mu + \sigma]$, ca. 95% der Fläche im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$.

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.

3.3.5 Exponentialverteilung $[\mathrm{Exp}\,(\lambda)]$

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

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

Das führt zu

$$\begin{aligned} \mathbb{E}[X] &= 1/\lambda \\ \mathrm{Var}(X) &= 1/\lambda^2. \end{aligned}$$

Beispiel. *Die Lebensdauer T eines Bauteils (in Wochen) sei exponential-verteilt mit erwarteter Lebensdauer 15 Wochen. Es ist also $T \sim \mathrm{Exp}\,(\lambda)$ 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

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

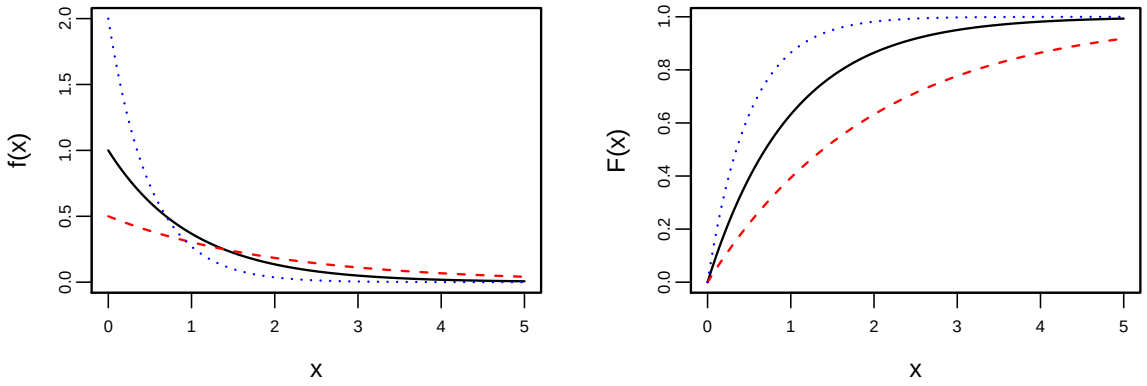


Abbildung 3.10: Dichte und Verteilungsfunktion der Exponentialverteilung für $\lambda = 1$ (schwarz, durchgezogen), $\lambda = 2$ (blau, gepunktet) und $\lambda = 1/2$ (rot, gestrichelt).

3.3.6 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 kennen wie auch die Funktion g. 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

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

Durch Ableiten erhalten wir die Dichte. Wir schreiben diese gerade für den allgemeinen Fall auf, d.h. für $b \neq 0$

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

Insbesondere gilt: Wenn $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, dann gilt für $Y = a + bX$, dass $Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right)$, denn nach obiger Transformationsformel gilt

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

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 kann man auch alle Exponentialverteilungen ineinander überführen.

Natürlich haben wir auch direkt mit den Rechenregeln von früher

$$\begin{aligned} \mathbb{E}[Y] &= a + b\mathbb{E}[X] \\ \sigma_Y &= |b|\sigma_X = b\sigma_X. \end{aligned}$$

Diese Kennzahlen müssen wir *nicht* via Umweg über die transformierten Dichten berechnen.

Standardisierung

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

$$g(x) = \frac{x - \mu}{\sigma}$$

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

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

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 $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Wie gross ist dann $\mathbb{P}(X \leq 3)$? Wir haben

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

wobei $Z \sim \mathcal{N}(0, 1)$. 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.2 lesen wir ab, dass $\Phi(0.25) = 0.5987$ (Zeile "`.2`" und Spalte "`.05`"). Wir können also mit diesem Trick alle Fälle 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 $Y = g(X)$ Dichte

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

Die Herleitung geht genau gleich wie beim linearen Fall.

Beispiel. *Wenn $X \sim \mathcal{N}(\mu, \sigma^2)$ normalverteilt ist, dann heisst $Y = e^X$* **lognormal-verteilt**. *Eine Zufallsvariable $Y > 0$ heisst also lognormal-verteilt, wenn der Logarithmus davon normalverteilt ist. Es ist hier*

$$g(x) = e^x,\ g'(x) = e^x \text{ und } g^{-1}(y) = \log(y).$$

Die Dichte ist gemäss obiger Transformationsformel gegeben durch

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

Für beliebiges g gilt zudem immer

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

Achtung: Der Erwartungswert transformiert nicht einfach mit. Falls g konvex ist, so gilt die **Jensen'sche Ungleichung**

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

Beispiel. *Ist Y lognormal-verteilt, so gilt $\mathbb{E}[Y] = e^{\mu+\sigma^2/2} > e^{\mu} = g(\mu)$.*

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.*

3.3.7 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

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

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 3.3.3.

Was bringt uns dieses Resultat? Es ist sehr nützlich, um Zufallsvariablen zu **simulieren**. So lange wir eine Implementierung der $\mathrm{Uni}(0, 1)$-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 $U \sim \mathrm{Uni}(0, 1)$. Dies wird mittels einem "Standard-Paket" gemacht.
2. Berechne $x = F^{-1}(u)$. Gemäss obigem Faktum ist dann x eine Realisation einer Zufallsvariablen X mit kumulativer Verteilungsfunktion F.

3.3.8 Vergleich der Konzepte: Diskrete vs. stetige Verteilungen

Die wichtigsten Konzepte der stetigen und diskreten Verteilungen sind in Tabelle 3.1 einander gegenüber gestellt.

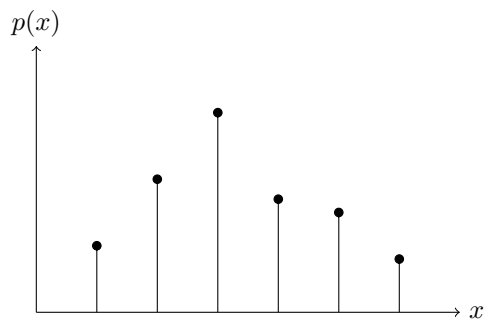
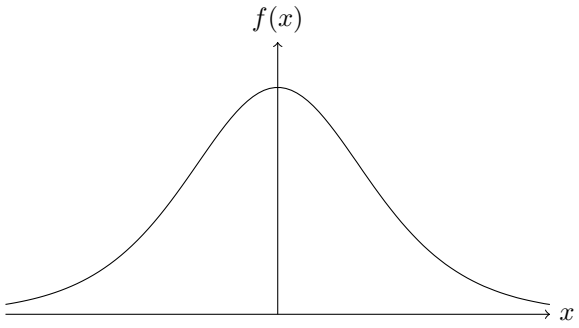
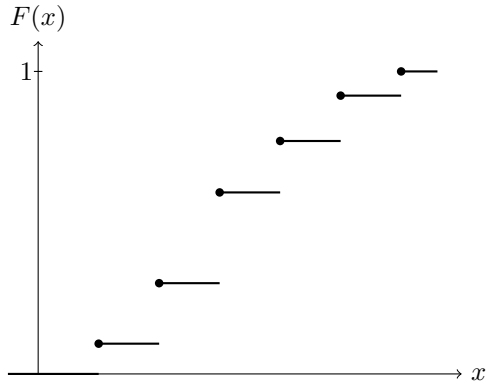
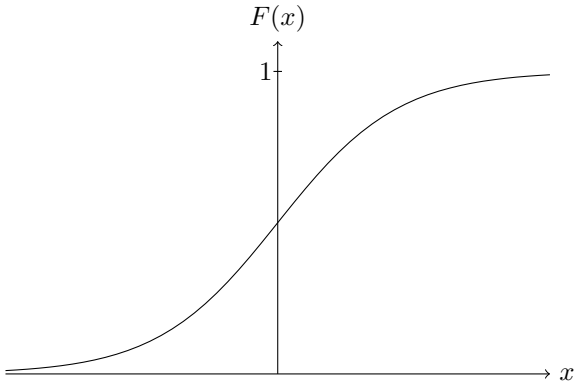
diskret	**stetig**
Wahrscheinlichkeitsfunktion	*Dichte*
	
Kumulative Verteilungsfunktion	*Kumulative Verteilungsfunktion*
	
$F(x) = \sum_{k\,:\,x_k \le x} p(x_k)$	$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u$
Ewartungswert	*Erwartungswert*
$\mathbb{E}[X] = \sum_{k\ge 1} x_k p(x_k)$	$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$
etc.	

Tabelle 3.1: Konzepte der diskreten Verteilungen (links) und der stetigen Verteilungen (rechts).

3.4 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 ferner 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.

4 Deskriptive Statistik

4.1 Einführung

In der *schliessenden* Statistik wird es später darum gehen, aus beobachteten Daten Schlüsse über den 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 geht es in einem ersten Schritt aber zunächst einmal darum, die vorhandenen Daten übersichtlich darzustellen und zusammenzufassen. Dies ist das Thema der *beschreibenden* oder *deskriptiven* Statistik.

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!

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

4.2 Kennzahlen

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

$$\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$$

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 im Sinne von experimentell beobachtet bzw. ermittelt).

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

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

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**

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

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 gerade (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

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

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) andere Definitionen des empirischen Quantils; für grosse n wird der Unterschied zwischen den Definitionen vernachlässigbar.

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

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

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** ist das obere Quartil minus das untere 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

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

Der empirische Median ist gegeben durch

$$q_{0.5} = \frac{1}{2}(3.333 + 3.6000) = 3.4665.$$

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*

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

4.3 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.

4.3.1 Histogramm

Beim **Histogramm** teilen wir den Wertebereich der Beobachtungen auf, d.h. wir bilden Klassen (Intervalle) $(c_{k-1}, c_k]$ und berechnen die Häufigkeiten h_k, d.h. die Anzahl Beobachtungen im entsprechenden Intervall.

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 ist. Wenn man überall die gleiche Klassenbreite $c_k - c_{k-1}$ wählt, so kann man auch direkt die Anzahlen auftragen.

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 nächste grössere ganze Zahl aufrundet.

4.3.2 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 4.1) 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 Sterne eingezeichnet.

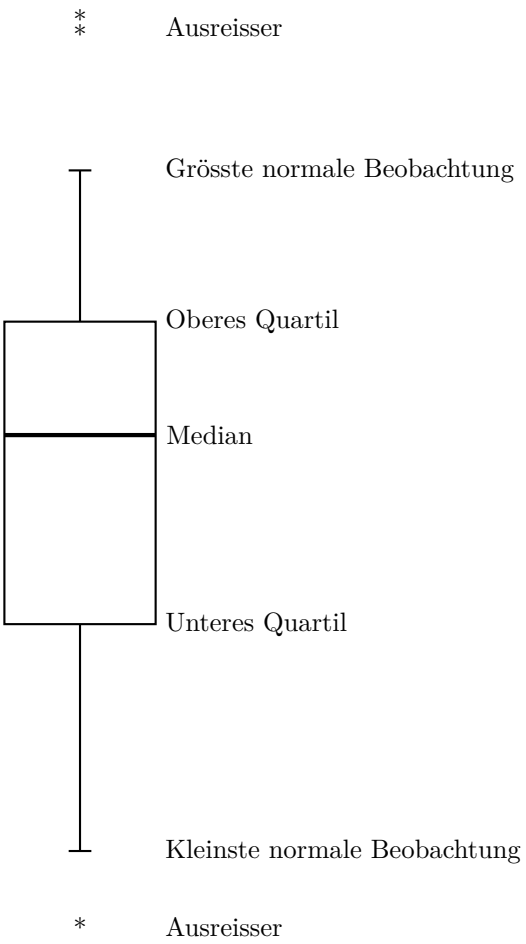


Abbildung 4.1: Schematische Darstellung eines Boxplots.

4.3.3 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 zum Beispiel 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 Abbildung 4.2 sind Histogramm, Boxplot und empirische kumulative Verteilungsfunktion 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 (eine sogenannte *bimodale* Verteilung) hat.

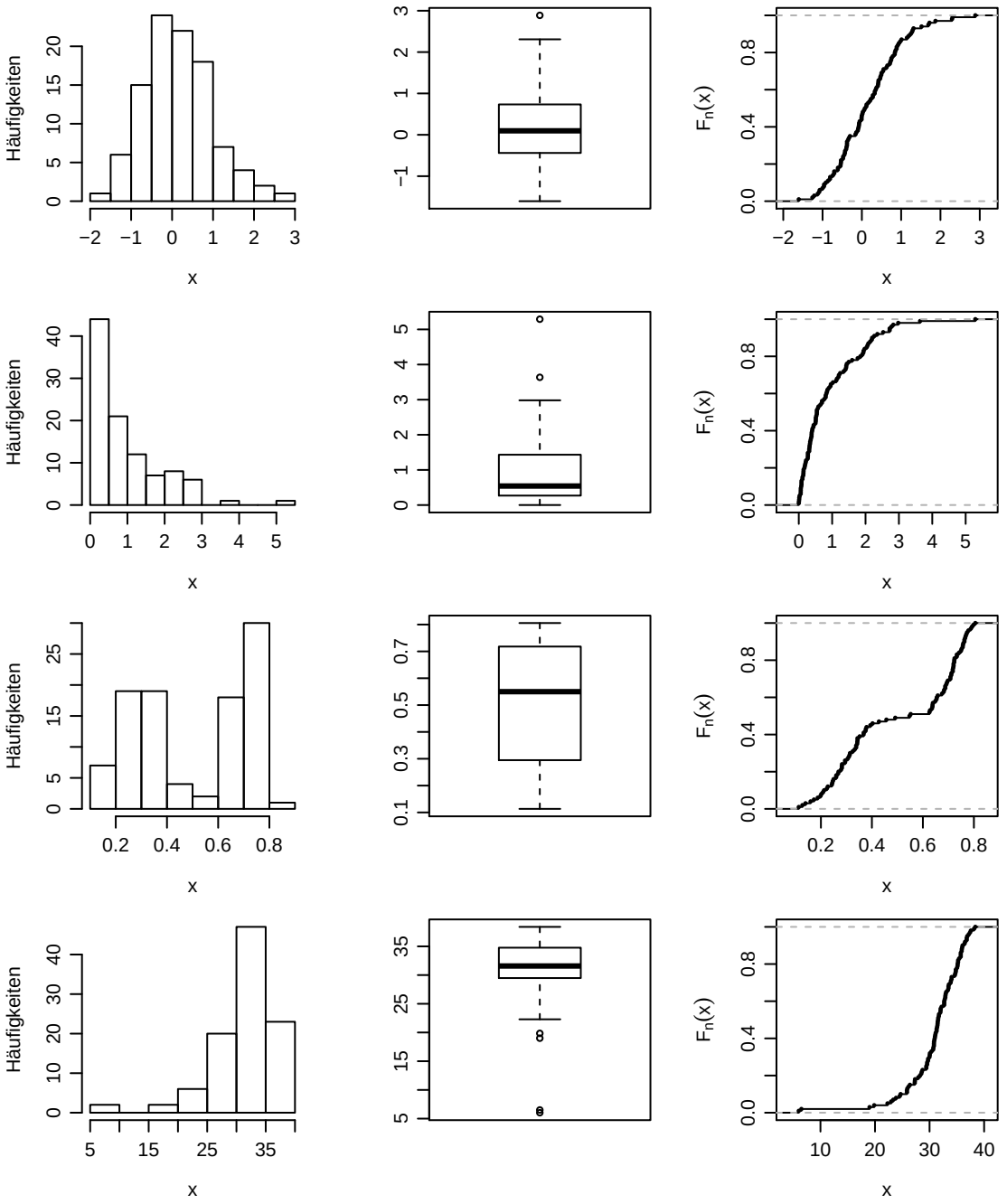


Abbildung 4.2: Histogramm (links), Boxplot (mitte) und empirische kumulative Verteilungsfunktion (rechts) von 4 Datensätzen der Grösse $n = 100$.

4.4 Mehrere Variablen

Oft misst man nicht nur eine einzelne, sondern mehrere Grössen gleichzeitig. Die einfachste Situation ist, falls die Daten paarweise vorliegen. Wir beobachten 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 an 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

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

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

$r = +1$ genau dann wenn $y_i = a + bx_i$ für ein $a \in \mathbb{R}$ und ein $b > 0$.
$r = -1$ genau dann wenn $y_i = a + bx_i$ für ein $a \in \mathbb{R}$ und ein $b < 0$.

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 4.3.

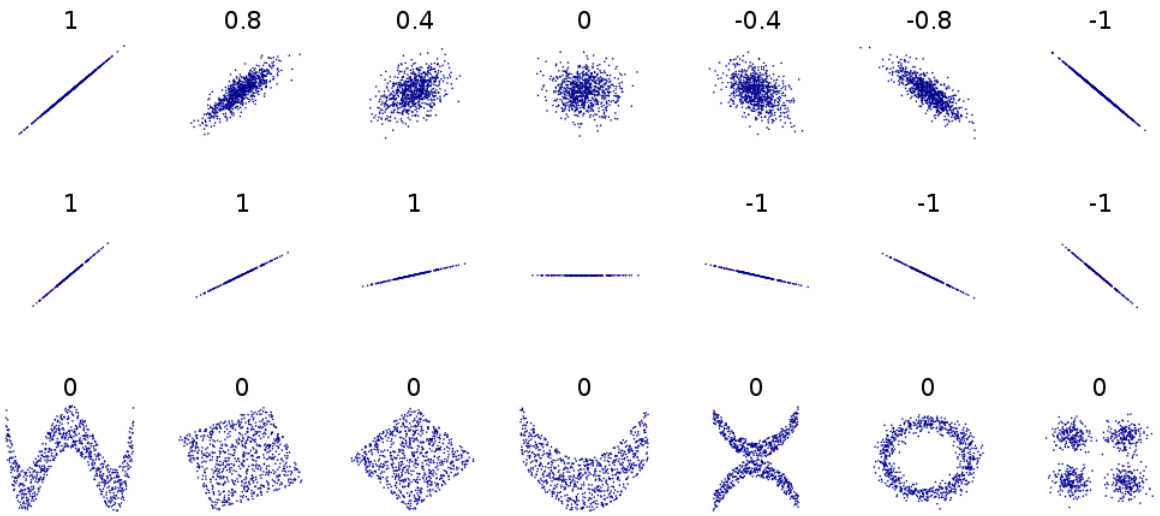


Abbildung 4.3: Empirische Korrelation bei verschiedenen Datensätzen (Quelle: Wikipedia)

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 4.4.

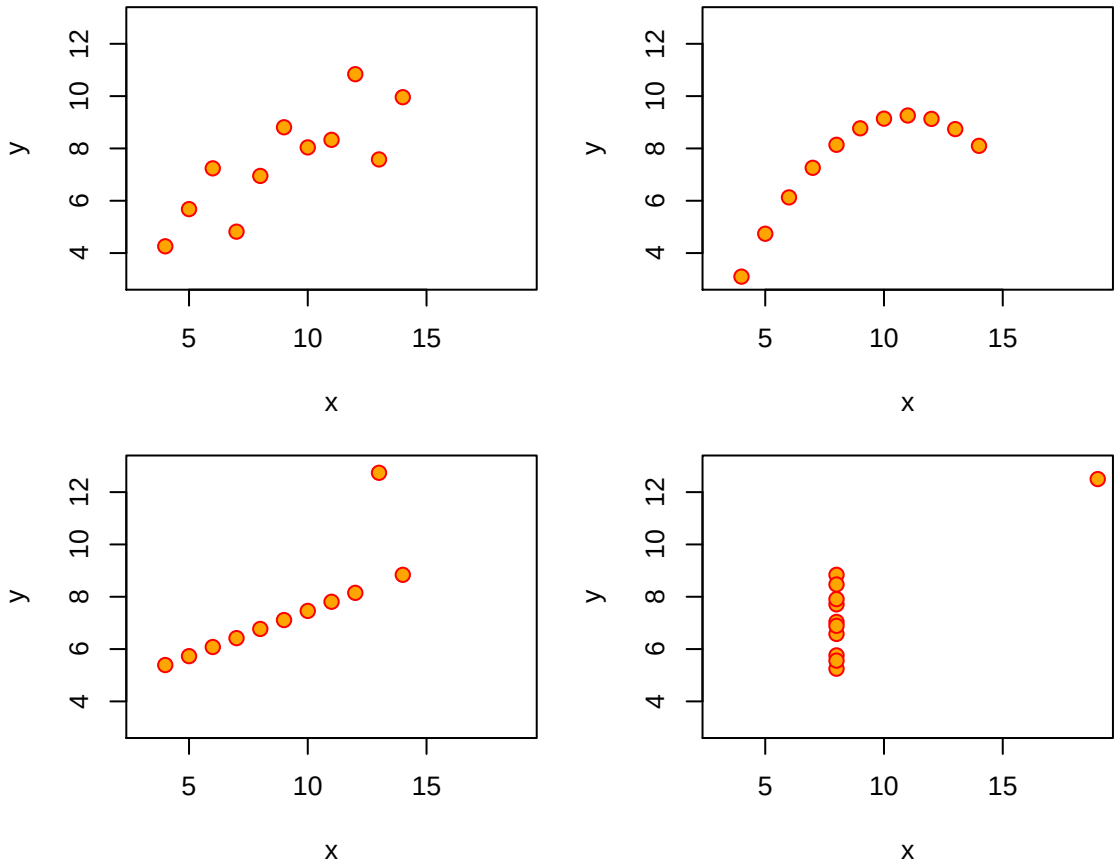


Abbildung 4.4: Vier Datensätze (von Anscombe) mit jeweils identischer empirischer Korrelation $r = 0.82$ zwischen x und y.

4.5 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.

In Tabelle 4.1 sind die entsprechenden "Gegenstücke" nochmals aufgelistet.

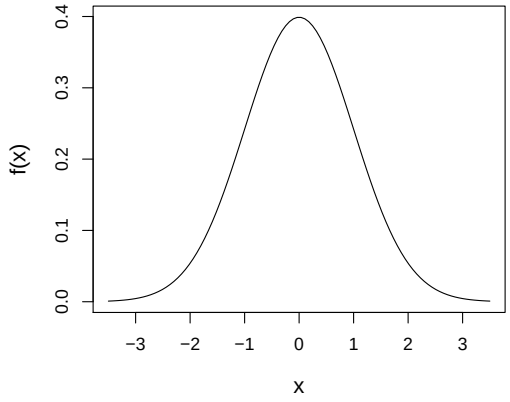
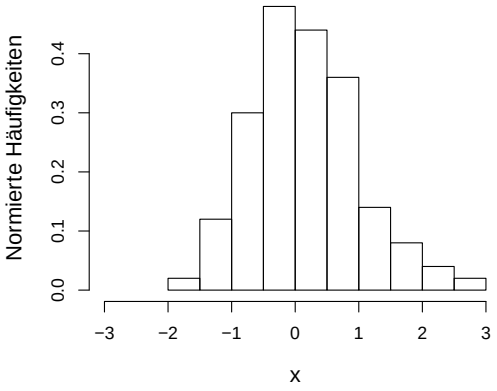
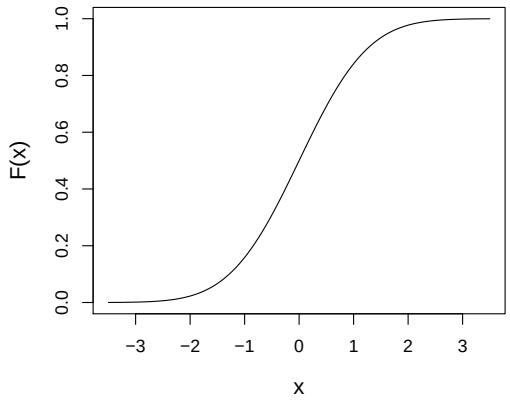
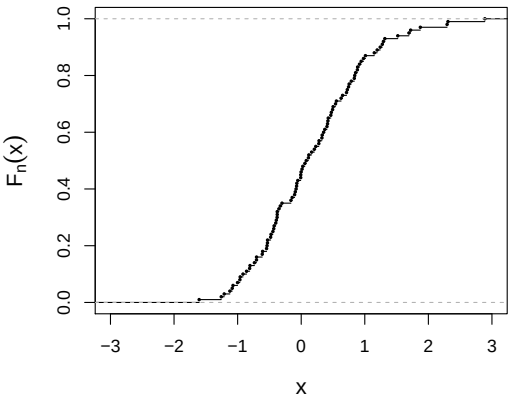
Modell	**Daten**
Dichte	*Histogramm*
	
Kumulative Verteilungsfunktion	*Empirische kumulative Verteilungsfunktion*
	
Erwartungswert $\mathbb{E}[X]$	**Arithm. Mittel** $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Varianz $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$	**Emp. Varianz** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
Quantile q_α	**Emp. Quantile** q_α
etc.	

Tabelle 4.1: Modell vs. Daten.

5 Mehrdimensionale Verteilungen

Oft misst bzw. modelliert man mehrere Grössen *gleichzeitig* (bei der deskriptiven Statistik haben wir dies schon ganz kurz gesehen), z.B. der Wasserstand an zwei verschiedenen Positionen A und B eines Flusses und will diese *gemeinsam* modellieren. In diesem Beispiel kann man wohl nicht von Unabhängigkeit ausgehen. Wenn an Position A der Wasserstand hoch ist, dann wird dies wohl mit hoher Wahrscheinlichkeit auch an Position B der Fall sein (bzw. umgekehrt). Für solche Fälle greift man auf sogenannte gemeinsame Verteilungen zurück.

5.1 Gemeinsame und bedingte Verteilungen

5.1.1 Diskreter Fall

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

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

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

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

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

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

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

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

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.

Ausser den neuen Begriffen haben wir soweit eigentlich alles schon einmal in leicht anderer Form gesehen, siehe Kapitel 2.

$X\backslash Y$	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 5.1: Gemeinsame diskrete Verteilung von (X, Y) im Beispiel mit den Wetterstationen.

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 5.1. Es ist z.B.*

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

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

y	1	2	3	4
$\mathbb{P}(Y = y \mid X = 1)$	0.8	0.15	0.03	0.02

Dies ist die erste Zeile aus Tabelle 5.1 dividiert durch $\mathbb{P}(X = 1) = 0.1$. Wir können auch die Wahrscheinlichkeit berechnen, dass die beiden Stationen das Gleiche messen. Es ist dies die Summe der Wahrscheinlichkeiten auf der Diagonalen

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

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.

5.1.2 Stetiger Fall

Bei zwei oder mehreren stetigen Zufallsvariablen muss man das Konzept der Dichten 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

$$\mathbb{P}(x \le X \le x + \mathrm{d}x,\ y \le Y \le y + \mathrm{d}y) = f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y.$$

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 durch Integration der Dichte über den entsprechenden Bereich berechnen

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

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

$$f_{X,Y}(x, y) = f_X(x) f_Y(y),\ x, y \in \mathbb{R}^2. \tag{5.1}$$

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* $X \sim \mathrm{Exp}(\lambda_1)$ *und* $Y \sim \mathrm{Exp}(\lambda_2)$*, 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 5.1. Also haben wir

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

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

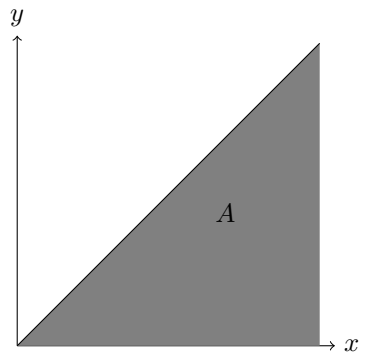


Abbildung 5.1: Integrationsbereich im Beispiel mit zwei Lebensdauern.

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

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \,\mathrm{d}y \qquad\qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \,\mathrm{d}x.$$

Dies ist genau gleich wie im diskreten Fall, dort haben wir einfach mit Hilfe des Satzes der totalen Wahrscheinlichkeit summiert statt integriert. Eine Illustration findet man in Abbildung 5.2.

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

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

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.

Aus den obigen Definitionen ist klar, dass alle wahrscheinlichkeitstheoretischen Aspekte von zwei Zufallsvariablen X und Y durch deren *gemeinsamen* Dichte $f_{X,Y}(x, y)$ vollständig bestimmt sind.

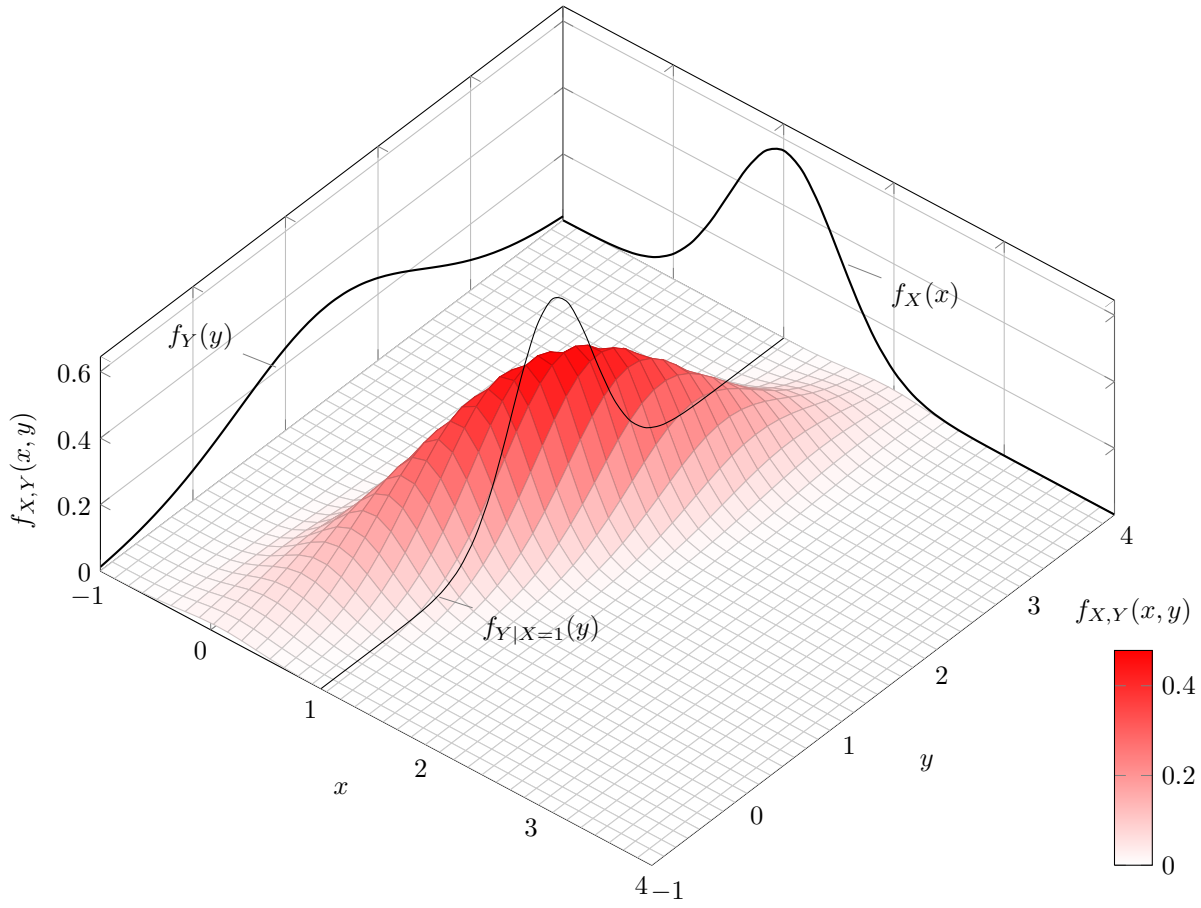


Abbildung 5.2: Illustration einer zweidimensionalen Dichte, deren Randverteilungen und der bedingten Verteilung gegeben $X = 1$. (tikZ Code von http://tex.stackexchange.com/questions/31708/draw-a-bivariate-normal-distribution-in-tikz).

5.2 Erwartungswert bei mehreren Zufallsvariablen

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

$$\mathbb{E}\left[g(X, Y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

Im diskreten Fall lautet die entsprechende Formel:

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

Der Erwartungswert der Zufallsvariablen Y gegeben $X = x$ ist

$$\mathbb{E}\left[Y \mid X=x\right] = \int_{-\infty}^{\infty} y f_{Y|X=x}(y)\,\mathrm{d}y.$$

5.3 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.

Die **Kovarianz** sowie die **Korrelation** zwischen X und Y sind wie folgt definiert:

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

Unmittelbar aus der Definition der Kovarianz folgt sofort

$$\mathrm{Var}\left(X\right) = \mathrm{Cov}\left(X,X\right),$$

sowie die wichtige Formel

$$\mathrm{Cov}\left(X,Y\right) = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

zur praktischen Berechnung der Kovarianz.

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

$$\mathrm{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \,\mathrm{Cov}\left(X_i, Y_j\right),$$

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

Für die Varianz der Summe erhalten wir also

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

Weitere Rechenregeln sind

$$\begin{aligned}
\mathrm{Cov}\left(a+bX, c+dY\right) &= bd\,\mathrm{Cov}\left(X,Y\right)\\
\mathrm{Corr}\left(a+bX, c+dY\right) &= \mathrm{sign}(b)\,\mathrm{sign}(d)\,\mathrm{Corr}\left(X,Y\right)\\
\mathrm{Var}\left(X+Y\right) &= \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right) + 2\,\mathrm{Cov}\left(X,Y\right).
\end{aligned}$$

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

$$\mathrm{Var}\left(X_1+\cdots+X_m\right) = \mathrm{Var}\left(X_1\right)+\cdots+\mathrm{Var}\left(X_m\right) \quad (X_1,\ldots,X_m \textbf{ unabhängig}).$$

Die Korrelation misst Stärke und Richtung der *linearen Abhängigkeit* zwischen X und Y. Es gilt

$$\begin{aligned}
&\mathrm{Corr}\left(X,Y\right) = +1 \text{ genau dann wenn } Y = a+bX \text{ für ein } a\in\mathbb{R} \text{ und ein } b>0\\
&\mathrm{Corr}\left(X,Y\right) = -1 \text{ genau dann wenn } Y = a+bX \text{ für ein } a\in\mathbb{R} \text{ und ein } b<0.
\end{aligned}$$

Im Gegensatz zur Kovarianz ist die Korrelation eine *dimensionslose* Grösse. Ferner gilt

$$X \text{ und } Y \text{ unabhängig} \quad\Longrightarrow\quad \mathrm{Corr}\left(X,Y\right) = 0. \tag{5.2}$$

Die Umkehrung gilt i.A. *nicht*. Ein Spezialfall, wo auch die Umkehrung gilt, wird in Kapitel 5.4 diskutiert.

5.4 Zweidimensionale Normalverteilung

Die wichtigste zweidimensionale Verteilung ist die Normalverteilung mit Erwartungswerten (μ_X, μ_Y) und Kovarianzmatrix Σ wobei $\Sigma_{11} = \mathrm{Var}(X)$, $\Sigma_{22} = \mathrm{Var}(Y)$ und $\Sigma_{12} = \Sigma_{21} = \mathrm{Cov}(X, Y)$. Sie hat die Dichte

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

Wir sehen von dieser Formel: Wenn $\mathrm{Cov}(X,Y) = 0$ wird Σ eine Diagonalmatrix und man kann nachrechnen, dass dann die Bedingung (5.1) gilt. Das heisst: im Falle der zweidimensionalen Normalverteilung gilt auch die Umkehrung von (5.2). Zudem: die Rand- sowie die bedingten Verteilungen sind wieder (eindimensionale) normalverteilte Zufallsvariablen.

5.5 Dichte einer Summe von zwei Zufallsvariablen

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

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

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

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

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 5.2 bzw. 5.3 vorgehen.

Beispiel. *Schauen wir einmal einen auf den ersten Blick "einfachen" Fall an. Wir betrachten zwei unabhängige Arbeitsprozesse. Der erste dauerst 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 $X \sim \mathrm{Uni}(3,5)$ *und* $Y \sim \mathrm{Uni}(6,10)$*, wobei* X *und* Y *unabhängig sind.*

Die Dichten der beiden uniformen Verteilungen können wir auch schreiben als

$$\begin{aligned} f_X(x) &= \frac{1}{2} 1_{[3,5]}(x), \\ f_Y(y) &= \frac{1}{4} 1_{[6,10]}(y) \end{aligned}$$

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

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

Wir können mit S *nur Werte zwischen 9 und 15 erreichen, ausserhalb ist die Dichte daher sicher 0. Wir haben daher gemäss obiger Formel*

$$\begin{aligned} f_S(s) &= \int_{-\infty}^{\infty} \frac{1}{8} \cdot 1_{[3,5]}(x) \cdot 1_{[6,10]}(s-x)\,\mathrm{d}x \\ &= \frac{1}{8} \int_3^5 1_{[6,10]}(s-x)\,\mathrm{d}x, \end{aligned}$$

wobei wir ausgenutzt haben, dass die erste Indikatorfunktion nur auf dem Intervall $[3, 5]$ *von Null verschieden ist und wir daher nur über diesen Bereich integrieren müssen.*

Jetzt müssen wir noch schauen, wann der Integrand verschieden von Null ist. Es ist dies der Fall, falls gilt $6 \leq s - x \leq 10$ *für ein fixes* $9 \leq s \leq 15$ *und* $3 \leq x \leq 5$*. Die obere Grenze des Integrals ist also gegeben durch* $\min\{5, s - 6\}$*, während die untere Grenze* $\max\{3, s - 10\}$ *ist. Dies führt zu*

$$f_S(s) = \begin{cases} \frac{1}{8}\Big(\min\{5, s-6\} - \max\{3, s-10\}\Big) & 9 \leq s \leq 15 \\ 0 & \textit{sonst} \end{cases}$$

Die entsprechende Funktion ist in Abbildung 5.3 dargestellt.

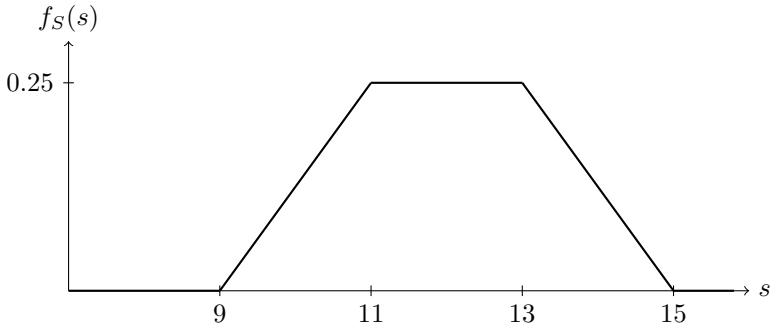


Abbildung 5.3: Dichte der Summe von zwei uniformen Verteilungen (Faltung).

5.6 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.

6 Grenzwertsätze

Wir wollen in diesem Kapitel schauen, was passiert, wenn wir viele unabhängige Wiederholungen von einer Zufallsvariablen haben und diese mitteln. Natürlich erhoffen wir uns eine grössere Genauigkeit, sonst würden wir den Mehraufwand für das mehrfache Durchführen eines Experimentes/Messung ja nicht machen. Es stellt sich natürlich die Frage, wie "schnell" die Genauigkeit zunimmt.

6.1 Die i.i.d. Annahme

Wir betrachten also n Zufallsvariablen $X_1, \ldots, X_n$, wobei X_i die i-te Wiederholung von unserem Experiment 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 $X_1, \ldots, X_n$ **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 scheint. Die Annahme bringt erhebliche Vereinfachungen, um mit mehreren Zufallsvariablen zu rechnen.

6.2 Summen und arithmetische Mittel von Zufallsvariablen

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

$$S_n = X_1 + \ldots + X_n$$

und arithmetisches Mittel

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

Wir nehmen stets an, dass $X_1, \ldots X_n$ 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 $\overline{X}_n$ nichts anderes als die relative Häufigkeit dieses Ereignisses. Die Verteilung von S_n ist im Allgemeinen schwierig exakt zu bestimmen, mit den folgenden 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 $S_n \sim \mathcal{N}(n\mu, n\sigma^2)$.

Einfacher sind die Berechnungen von Erwartungswert, Varianz und Standardabweichung.

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

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 $1/\sqrt{n}$. Um die Genauigkeit des arithmetischen Mittels zu *verdoppeln* (d.h. die Standardabweichung zu halbieren), braucht man *viermal* so viele Beobachtungen.

6.3 Das Gesetz der Grossen Zahlen und der Zentrale Grenzwertsatz

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

- $\mathbb{E}\left[\overline{X}_n\right] = \mathbb{E}\left[X_i\right]$: das heisst $\overline{X}_n$ hat denselben Erwartungswert wie die einzelnen X_i's.
- $\mathrm{Var}\left(\overline{X}_n\right) \to 0$ $(n \to \infty)$: das heisst, $\overline{X}_n$ besitzt keine Variabilität mehr im Limes.

Diese beiden Punkte implizieren den folgenden Satz.

Gesetz der Grossen Zahlen (GGZ)
Seien $X_1, \ldots, X_n$ i.i.d. mit Erwartungswert μ. Dann gilt

$$\overline{X}_n \overset{n\to\infty}{\longrightarrow} \mu.$$

Ein Spezialfall davon ist

$$f_n[A] \overset{n\to\infty}{\longrightarrow} \mathbb{P}\left(A\right).$$

Korrekterweise müsste man den Begriff der Konvergenz für Zufallsvariablen zuerst 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.

Zur Berechnung der genäherten **Verteilung** von S_n und $\overline{X}_n$ (dies ist ein bedeutend stärkeres Resultat als das GGZ) stützt man sich auf den folgenden berühmten Satz.

Zentraler Grenzwertsatz (ZGWS)
Seien $X_1, \ldots, X_n$ i.i.d. mit Erwartungswert μ und Varianz σ^2, dann ist

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

für grosse n. Wie gut diese Approximationen für ein gegebenes n sind, hängt von der Verteilung der X_i ab.

D.h. selbst wenn wir die Verteilung der X_i nicht kennen, so haben wir eine Ahnung über die approximative Verteilung von S_n und $\overline{X}_n$. Der Zentrale Grenzwertsatz ist mitunter ein Grund für die Wichtigkeit der Normalverteilung.

In Abbildung 6.1 sieht man den zentralen Grenzwertsatz an einem empirischen Beispiel. Wir betrachten $X_1, \ldots, X_8$ i.i.d. $\sim \mathrm{Uni}\left(-1/2, 1/2\right)$. Von jeder Zufallsvariablen simulieren wir 5'000 Realisierungen. Wir betrachten die Histogramme für U_1, $U_1 + U_2, \ldots$ etc. und sehen, dass schon bei wenigen Summanden einen glockenförmige Struktur vorliegt ist.

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

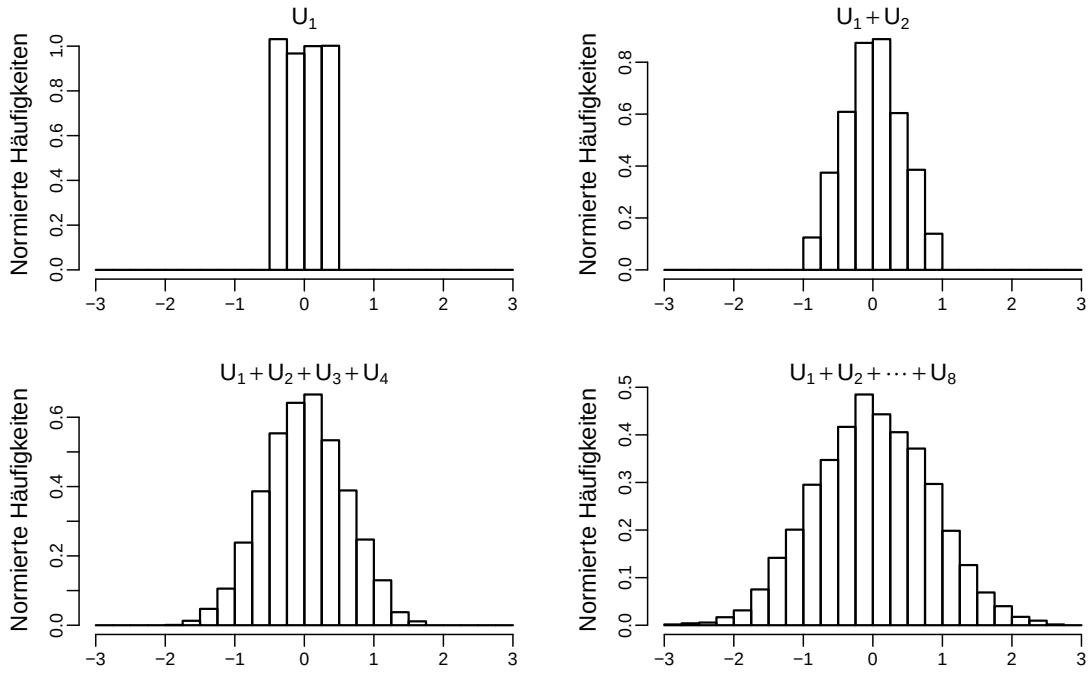


Abbildung 6.1: Histogramme der Summen von simulierten uniform-verteilten Zufallsvariablen. Die Stichprobengrösse beträgt jeweils 5'000.

Die X_i's können natürlich auch diskret sein. Wir haben schon bei der Binomialverteilung in Abbildung 3.2 gesehen, dass diese für n gross "glockenförmig" aussieht. Dasselbe gilt für die Poissonverteilung in Abbildung 3.4 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 Bernoulliverteilungen). Man spricht dann von der sogenannten **Normalapproximation** der Binomialverteilung.

Wenn $X \sim \text{Bin}(n, p)$, dann haben wir $\mathbb{E}[X] = np$ und $\text{Var}(X) = np(1-p)$. Es ist daher für n gross

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

d.h. wir approximieren X mit einer Normalverteilung mit Erwartungswert np und Varianz $np(1-p)$.

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

A Zusammenfassungen, Tabellen und Herleitungen

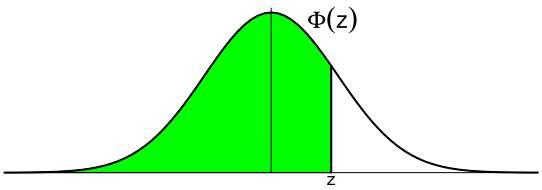
A.1 Die wichtigsten eindimensionalen Verteilungen

Beachte $\mathbb{R}_+ := \{x \in \mathbb{R} \,|\, x \geq 0\}$.

Verteilung	$p(x)$ bzw. $f(x)$	W_X	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
$\mathrm{Bin}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,\ldots,n\}$	np	$np(1-p)$
Geometrisch(p)	$p(1-p)^{x-1}$	$\{1,2,\ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\mathrm{Pois}(\lambda)$	$e^{-\lambda}\frac{\lambda^x}{x!}$	$\{0,1,\ldots\}$	λ	λ
$\mathrm{Uni}(a,b)$	$\frac{1}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(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 Tabelle der Normalverteilung

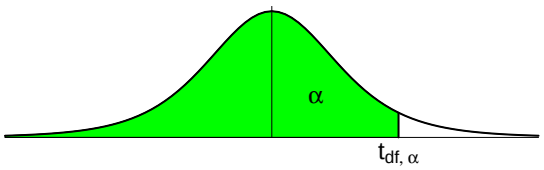
$$\Phi(z) = \mathbb{P}(Z \leq z), \; Z \sim \mathcal{N}(0,1)$$



Bsp: $\mathbb{P}(Z \leq 1.96) = 0.975$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
.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

A.3 Quantile der t-Verteilung



Bsp: $t_{9,\,0.975} = 2.262$

df	$t_{0.60}$	$t_{0.70}$	$t_{0.80}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
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
31	0.255	0.530	0.853	1.309	1.696	2.040	2.452	2.744
32	0.255	0.530	0.853	1.309	1.694	2.037	2.449	2.738
33	0.255	0.530	0.853	1.308	1.693	2.035	2.445	2.733
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724
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

A.4 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

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

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

$$\int_{0}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Wie ist dieses Integral genau zu verstehen und wie berechnet man es? Das Integral

$$\int_{0}^{\infty} f(x)\,\mathrm{d}x$$

ist ein uneigentliches Integral und ist definiert als

$$\int_{0}^{\infty} f(x)\,\mathrm{d}x = \lim_{a\to\infty} \int_{0}^{a} f(x)\,\mathrm{d}x.$$

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

$$\int_{0}^{a} \lambda e^{-\lambda x}\,\mathrm{d}x = -e^{-\lambda x}\Big|_{0}^{a} = -e^{-\lambda a} + 1.$$

Wenn wir jetzt den Limes $a \to \infty$ ziehen, so haben wir

$$\int_{0}^{\infty} f(x)\,\mathrm{d}x = 1,$$

da $\lim_{a\to\infty} e^{-\lambda a} = 0$.

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.

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \lim_{a\to-\infty} \int_{a}^{c} f(x)\,\mathrm{d}x + \lim_{b\to\infty} \int_{c}^{b} f(x)\,\mathrm{d}x.$$

Das heisst implizit, dass wir die beiden Grenzen *unabhängig voneinander* nach $\pm\infty$ gehen lassen:

$$\lim_{\substack{a\to-\infty \\ b\to\infty}} \int_{a}^{b} f(x)\,\mathrm{d}x.$$

Man darf *nicht*

$$\lim_{a\to\infty}\int_{-a}^{a} f(x)\,\mathrm{d}x$$

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

$$\int_{-\infty}^{0} x\,\mathrm{d}x \quad \text{bzw.} \quad \int_{0}^{\infty} x\,\mathrm{d}x$$

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

$$\int_{-\infty}^{\infty} \frac{1}{\pi}\frac{1}{1+x^2}\,\mathrm{d}x.$$

Wir haben

$$\int_a^b \frac{1}{1+x^2} = \arctan(x)\Big|_a^b = \arctan(b) - \arctan(a).$$

Es ist

$$\begin{aligned}\lim_{b\to\infty}\arctan(b) &= \frac{\pi}{2}\\ \lim_{a\to-\infty}\arctan(a) &= -\frac{\pi}{2}.\end{aligned}$$

Also haben wir schlussendlich

$$\int_{-\infty}^{\infty} \frac{1}{\pi}\frac{1}{1+x^2}\,\mathrm{d}x = \frac{1}{\pi}\left(\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right) = 1.$$

Bei der Funktion handelt es sich um die Dichte der sogenannten Cauchy-Verteilung.

A.5 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 • markiert wenn Erfolg eintritt und sonst mit dem Symbol ○.

1	2	3	4	5	6	$\cdots$	$\cdots$	$n-1$	n
•	○	•	○	○	•	○	○	○	○

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 $1 \to 4 \to 6,\ 1 \to 6 \to 4,\ 4 \to 1 \to 6,\ 4 \to 6 \to 1,\ 6 \to 1 \to 4$ und $6 \to 4 \to 1$ 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.