DR. JO’S DIVE IN

Ereignis. Zufall. Wahrscheinlichkeit.

Eine klare Einführung in die mathematische Sprache der Unsicherheit.

STATISTIK · BAND 1
Granaria Foundation · Ausgabe 2026

Inhaltsverzeichnis

KAPITEL 1

Ereignisse und Wahrscheinlichkeit

1.1 Einführung

Wahrscheinlichkeitstheorie ist die mathematische Sprache der Unsicherheit. Jedes Mal, wenn wir eine Münze werfen, eine Messung durchführen oder eine Prognose treffen, haben wir es mit Zufall zu tun – und genau diesen Zufall wollen wir präzise fassen.

Die Grundidee ist bestechend einfach: Wir listen alle denkbaren Ausgänge eines Experiments auf, fassen interessante Teilmengen zu **Ereignissen** zusammen und ordnen diesen Wahrscheinlichkeiten zu. Aus diesen drei Zutaten – Ergebnisraum, Ereignisse, Wahrscheinlichkeitsmaß – entsteht das gesamte Gebäude der Wahrscheinlichkeitstheorie.

1.2 Ergebnisräume und Ereignisse

Der **Ergebnisraum** Ω ist die Menge aller möglichen Ausgänge eines Experiments. Punkte $\omega \in \Omega$ heißen **Ergebnisse** oder **Realisierungen**. Ereignisse sind die Elemente einer **Ereignisklasse** $\mathcal{A}$ – einer Familie von Teilmengen von Ω, der man konsistent Wahrscheinlichkeiten zuordnen kann.

Bemerkung 1.2.1 (Vorgriff: messbare Ereignisse)

Für endliche oder abzählbare Ω darf man $\mathcal{A} = \mathcal{P}(\Omega)$ wählen, also *jede* Teilmenge als Ereignis zulassen; so halten wir es in diesem Kapitel. Zwingend ist das nicht – auch kleinere Familien sind zulässig, solange sie unter den unten eingeführten Mengenoperationen stabil bleiben. Bei überabzählbarem Ω (etwa $\Omega = \mathbb{R}$) wird die Potenzmenge dagegen unbequem: Verlangt man von $\mathbb{P}$ zusätzlich Eigenschaften wie Verschiebungsinvarianz, so lässt sie sich nicht mehr auf allen Teilmengen erklären, und man beschränkt sich auf eine echt kleinere Familie. Die zulässigen Familien heißen σ-Algebren – die Potenzmenge ist die größte von ihnen –; präzisiert wird das in §1.8. Bis dahin darf man Ereignisse gefahrlos als Teilmengen von Ω lesen.

Beispiel 1.2.1

Beim zweimaligen Münzwurf ist $\Omega = \{KK, KZ, ZK, ZZ\}$. Das Ereignis „erster Wurf ist Kopf“ ist $A = \{KK, KZ\}$.

Beispiel 1.2.2

Sei ω das Ergebnis einer physikalischen Messung, z. B. der Temperatur. Dann ist $\Omega = \mathbb{R}$. Man könnte argumentieren, dass $\Omega = \mathbb{R}$ ungenau ist, da Temperatur eine untere Grenze besitzt. In der Praxis schadet es jedoch nicht, den Ergebnisraum großzügiger zu wählen. Das Ereignis, dass die Messung größer als 10 aber höchstens 23 ist, lautet $A = (10, 23]$.

Beispiel 1.2.3

Beim unendlich oft wiederholten Münzwurf ist

$$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \ldots) : \omega_i \in \{K, Z\}\}.$$

Das Ereignis, dass der erste Kopf beim dritten Wurf erscheint, ist

$$E = \{(\omega_1, \omega_2, \omega_3, \ldots) : \omega_1 = Z,\, \omega_2 = Z,\, \omega_3 = K\}.$$

Mengenoperationen

Für ein Ereignis A bezeichnen wir mit $A^c = \{\omega \in \Omega : \omega \notin A\}$ das **Komplement** von A („nicht A“). Das Komplement von Ω ist die leere Menge $\emptyset$.

Die **Vereinigung** $A \cup B = \{\omega : \omega \in A$ oder $\omega \in B\}$ und für eine Folge $A_1, A_2, \ldots$ ist $\bigcup_{i=1}^{\infty} A_i = \{\omega : \omega \in A_i$ für mindestens ein $i\}$.

Der **Durchschnitt** $A \cap B = \{\omega : \omega \in A$ und $\omega \in B\}$. Wir schreiben auch AB statt $A \cap B$. Für eine Folge: $\bigcap_{i=1}^{\infty} A_i = \{\omega : \omega \in A_i$ für alle $i\}$.

Die **Differenz** ist $A \setminus B = \{\omega : \omega \in A,\, \omega \notin B\}$. Ist jedes Element von A in B enthalten, schreiben wir $A \subset B$. Für eine endliche Menge A bezeichnet $|A|$ die Anzahl ihrer Elemente.

Symbol	Bedeutung
Ω	Ergebnisraum (sicheres Ereignis)
ω	Ergebnis (Element von Ω)
A	Ereignis (Teilmenge von Ω)
A^c	Komplement von A
$A \cup B$	Vereinigung (A oder B)
$A \cap B$	Durchschnitt (A und B)
$A \setminus B$	Mengendifferenz
$A \subset B$	Inklusion
$\emptyset$	Leere Menge (unmögliches Ereignis)

Ereignisse $A_1, A_2, \ldots$ heißen **disjunkt** (oder **paarweise disjunkt**), falls $A_i \cap A_j = \emptyset$ für $i \neq j$.

Definition 1.2.1 (Zerlegung)

Eine **Zerlegung** (auch *Partition*) von Ω ist eine endliche oder abzählbare Familie nicht-leerer, paarweise disjunkter Ereignisse $A_1, A_2, \ldots$ mit $\bigcup_i A_i = \Omega$.

Die **Indikatorfunktion** von A ist

$$I_A(\omega) = \begin{cases} 1 & \text{falls } \omega \in A, \\ 0 & \text{falls } \omega \notin A. \end{cases}$$

Eine Folge $A_1, A_2, \ldots$ ist **monoton wachsend**, falls $A_1 \subset A_2 \subset \cdots$, und wir definieren $\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$. Sie ist **monoton fallend**, falls $A_1 \supset A_2 \supset \cdots$, und dann $\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i$. In beiden Fällen schreiben wir $A_n \to A$.

Beispiel 1.2.4

Sei $\Omega = \mathbb{R}$ und $A_i = [0, 1/i)$ für $i = 1, 2, \ldots$. Dann ist $A_1 \supset A_2 \supset \cdots$ fallend mit $\bigcap_{i=1}^{\infty} A_i = \{0\}$. Definiert man stattdessen $A_i = (0, 1/i)$, so ist $\bigcap_{i=1}^{\infty} A_i = \emptyset$ – der Punkt 0 fehlt in jedem A_i.

Ereignisfolgen und Grenzwerte

Eine beliebige Ereignisfolge muss weder wachsen noch fallen. Trotzdem lässt sich präzise beschreiben, welche Ergebnisse immer wieder auftreten und welche ab einem gewissen Index dauerhaft auftreten. Dazu betrachten wir für jedes $n \geq 1$ den Rest der Folge ab dem Index n.

Checkerbox 1.2.1 (A13: Ereignisfolgen und Grenzwerte)

Sei Ω ein Stichprobenraum und seien $A_1, A_2, \ldots$ Ereignisse. Definiere

$$B_n = \bigcup_{i=n}^{\infty} A_i \qquad \text{und} \qquad C_n = \bigcap_{i=n}^{\infty} A_i.$$

(a) Zeige, dass $B_1 \supset B_2 \supset \cdots$ und dass $C_1 \subset C_2 \subset \cdots$.

(b) Zeige, dass $\omega \in \bigcap_{n=1}^{\infty} B_n$ genau dann, wenn ω zu unendlich vielen der Ereignisse $A_1, A_2, \ldots$ gehört.

(c) Zeige, dass $\omega \in \bigcup_{n=1}^{\infty} C_n$ genau dann, wenn ω zu allen Ereignissen $A_1, A_2, \ldots$ gehört, außer möglicherweise zu endlich vielen dieser Ereignisse.

(d) Folgere aus (b) und (c), dass stets $\bigcup_{n=1}^{\infty} C_n \subset \bigcap_{n=1}^{\infty} B_n$ gilt.

Beweis. **(a)** Aus $\omega \in B_{n+1}$ folgt $\omega \in A_i$ für mindestens ein $i \geq n+1$. Dann ist auch $i \geq n$, also $\omega \in B_n$. Somit ist $B_{n+1} \subset B_n$. Liegt dagegen $\omega \in C_n$, so gehört ω zu allen A_i mit $i \geq n$ und damit insbesondere zu allen A_i mit $i \geq n+1$. Also ist $C_n \subset C_{n+1}$.
(b) Sei $\omega \in \bigcap_{n=1}^{\infty} B_n$. Dann gibt es zu jedem n einen Index $i \geq n$ mit $\omega \in A_i$. Die Menge solcher Indizes kann daher nicht endlich sein. Umgekehrt liegt unter unendlich vielen Indizes mit $\omega \in A_i$ zu jedem n mindestens einer jenseits von n. Folglich ist $\omega \in B_n$ für jedes n, also $\omega \in \bigcap_{n=1}^{\infty} B_n$.

(c) Sei $\omega \in \bigcup_{n=1}^{\infty} C_n$. Dann gibt es ein n_0 mit $\omega \in A_i$ für alle $i \geq n_0$; Ausnahmen sind also nur unter $A_1, \ldots, A_{n_0-1}$ möglich. Gibt es umgekehrt nur endlich viele Ausnahmen, so wählen wir n_0 größer als deren größten Index; gibt es keine Ausnahme, setzen wir $n_0 = 1$. Dann liegt ω in jedem A_i mit $i \geq n_0$, also in C_{n_0} und damit in $\bigcup_{n=1}^{\infty} C_n$.
(d) Wer zu allen A_i bis auf endlich viele gehört, gehört insbesondere zu unendlich vielen. Mit (b) und (c) folgt daraus $\bigcup_{n=1}^{\infty} C_n \subset \bigcap_{n=1}^{\infty} B_n$. ■

Einordnung. Die beiden Mengenfolgen aus Checkerbox 1.2.1 führen auf zwei Grenzbegriffe, die auch für nicht-monotone Ereignisfolgen sinnvoll sind.

Definition 1.2.2 (Limes superior und Limes inferior)

Seien $A_1, A_2, \ldots$ Ereignisse und

$$B_n = \bigcup_{i=n}^{\infty} A_i, \qquad C_n = \bigcap_{i=n}^{\infty} A_i.$$

Dann heißen

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} B_n = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i$$

der **Limes superior** und

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} C_n = \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i$$

der **Limes inferior** der Ereignisfolge.

Zwei Arten des wiederholten Auftretens

B_n enthält alle Ergebnisse, die ab dem Zeitpunkt n *noch mindestens einmal* auftreten. Wer in jedem B_n liegt, tritt daher immer wieder auf. Dagegen enthält C_n nur Ergebnisse, die ab dem Zeitpunkt n *ohne weitere Ausnahme* auftreten. Somit gilt:

$$\omega \in \limsup_{n\to\infty} A_n \quad \Longleftrightarrow \quad \omega \text{ gehört zu unendlich vielen } A_i,$$
$$\omega \in \liminf_{n\to\infty} A_n \quad \Longleftrightarrow \quad \omega \text{ gehört zu allen bis auf endlich viele } A_i.$$

Kurz: $\limsup A_n$ bedeutet „unendlich oft“, $\liminf A_n$ bedeutet „schließlich immer“.

In dieser Sprache ist Teil (d) von Checkerbox 1.2.1 gerade die Inklusion

$$\liminf_{n\to\infty} A_n \subset \limsup_{n\to\infty} A_n :$$

Was schließlich immer eintritt, tritt insbesondere unendlich oft ein.

1.3 Wahrscheinlichkeit

Jedem Ereignis A ordnen wir eine reelle Zahl $\mathbb{P}(A)$ zu – die **Wahrscheinlichkeit** von A.

Definition 1.3.1

Eine Funktion $\mathbb{P}$, die jedem Ereignis A eine reelle Zahl zuordnet, ist ein **Wahrscheinlichkeitsmaß**, falls:
Axiom 1: $\mathbb{P}(A) \geq 0$ für jedes A.
Axiom 2: $\mathbb{P}(\Omega) = 1$.
Axiom 3 (σ-Additivität): Sind $A_1, A_2, \ldots$ paarweise disjunkt, so gilt

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Es gibt zwei gängige Interpretationen von $\mathbb{P}(A)$, die sich darin unterscheiden, *was* eine Wahrscheinlichkeitsaussage bedeutet:

1. Häufigkeitsinterpretation (frequentistisch). $\mathbb{P}(A)$ ist der langfristige relative Anteil, mit dem A bei (hypothetisch) unendlich vielen Wiederholungen desselben Experiments eintritt. Sagen wir $\mathbb{P}(\text{Kopf}) = 1/2$, so meinen wir: Wirft man die Münze n-mal und zählt die Kopf-Ergebnisse n_K, so gilt $n_K/n \to 1/2$ für $n \to \infty$. Diese Sichtweise verlangt, dass das Experiment prinzipiell wiederholbar ist.

2. Glaubensgrad-Interpretation (subjektivistisch / Bayessch). $\mathbb{P}(A)$ misst die *Stärke der persönlichen Überzeugung* eines Beobachters, dass A wahr ist. Die Zahl $\mathbb{P}(A) = 0{,}3$ drückt aus: „Ich halte es für 30 % plausibel, dass A eintritt“ – auch wenn das Experiment nicht wiederholbar ist.

Beispiel 1.3.1

Betrachte die Aussage: „Die Wahrscheinlichkeit, dass es morgen regnet, ist $0{,}3$.“

- **Frequentistisch:** Unter allen Tagen mit vergleichbaren Wetterbedingungen regnet es in 30 % der Fälle. Die Aussage bezieht sich auf eine Klasse ähnlicher Situationen.
- **Bayessch:** Aufgrund meiner aktuellen Informationen (Wolkenbild, Jahreszeit, Vorhersage) bin ich zu 30 % überzeugt, dass es morgen regnet. Die Aussage ist subjektiv und ändert sich, sobald neue Informationen vorliegen.

Der entscheidende Punkt: Beide Interpretationen verwenden *dieselben* drei Axiome – die mathematische Struktur ist identisch. Der Unterschied liegt nicht in der Rechnung, sondern in der *Bedeutung der Zahlen*. Er wird erst bei der statistischen Inferenz praktisch relevant und führt dort zu zwei Schulen: der **frequentistischen** Statistik (Wahrscheinlichkeiten sind objektive Häufigkeiten; Parameter sind feste, unbekannte Größen) und der **Bayesschen** Statistik (Wahrscheinlichkeiten drücken Überzeugungen aus; auch Parameter erhalten Wahrscheinlichkeitsverteilungen). Wir vertiefen diesen Gegensatz in Teil 2 (Kapitel 11, Bayessche Inferenz).

Satz 1.3.2 (Elementare Eigenschaften)

Für beliebige Ereignisse A, B gilt:

1. $\mathbb{P}(\emptyset) = 0$,
2. $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$,
3. $0 \leq \mathbb{P}(A) \leq 1$,
4. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
5. $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Beweis. (1) Setze $A_1 = \Omega$, $A_i = \emptyset$ für $i \geq 2$. Axiom 3: $1 = 1 + \sum_{i=2}^{\infty} \mathbb{P}(\emptyset)$, also $\mathbb{P}(\emptyset) = 0$.
(2) $B = A \cup (B \cap A^c)$ disjunkt, also $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \geq \mathbb{P}(A)$.
(3) Aus $\emptyset \subset A \subset \Omega$ und (1), (2).
(4) $A \cup A^c = \Omega$ disjunkt, also $1 = \mathbb{P}(A) + \mathbb{P}(A^c)$.
(5) Spezialfall von Axiom 3 für $n = 2$. ■

Checkerbox 1.3.1 (Elementare Eigenschaften aus den Axiomen)

Beweise jede der fünf Aussagen des vorigen Satzes direkt aus den drei Kolmogorov-Axiomen. Gib insbesondere an, welche Mengen jeweils disjunkt sind, wenn du die Additivität verwendest.

Beweis. **(1) Leere Menge.** Setze $A_1 = \Omega$ und $A_i = \emptyset$ für alle $i \geq 2$. Diese Folge ist paarweise disjunkt und ihre Vereinigung ist Ω. Die σ-Additivität liefert daher

$$1 = \mathbb{P}(\Omega) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = 1 + \sum_{i=2}^{\infty} \mathbb{P}(\emptyset).$$

Alle Summanden sind nichtnegativ. Deshalb kann die letzte Summe nur dann 0 sein, wenn $\mathbb{P}(\emptyset) = 0$ gilt.
(2) Monotonie. Aus $A \subset B$ folgt die disjunkte Zerlegung

$$B = A \cup (B \setminus A).$$

Deshalb gilt $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$, weil Wahrscheinlichkeiten nach Axiom 1 nichtnegativ sind.
(3) Schranken. Axiom 1 liefert $\mathbb{P}(A) \geq 0$. Wegen $A \subset \Omega$ ergibt die eben bewiesene Monotonie außerdem $\mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$.
(4) Komplement. A und A^c sind disjunkt und ergeben zusammen Ω. Daher

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c),$$

also $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
(5) Endliche Additivität. Sind A und B disjunkt, wenden wir die σ-Additivität auf $A_1 = A$, $A_2 = B$ und $A_i = \emptyset$ für $i \geq 3$ an. Mit (1) folgt

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

■

Aus Punkt (5) gewinnt man sofort ein Werkzeug, das unscheinbar aussieht und uns doch durch das ganze Kapitel begleitet: die **Fallunterscheidung**. Zu einem Ereignis A und einem beliebigen zweiten Ereignis B gibt es nur zwei Möglichkeiten – entweder tritt B mit ein oder eben nicht. Die Wahrscheinlichkeit von A teilt sich entlang dieser Grenze sauber auf.

Checkerbox 1.3.2

Seien A und B beliebige Ereignisse. Zeige

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

Halte dabei fest, an welcher Stelle du die Additivität benutzt und warum sie dort anwendbar ist.

Beweis. **Schritt 1: Zerlegung von A.** Die Ereignisse B und B^c bilden eine Zerlegung von Ω: Sie sind disjunkt wegen $B \cap B^c = \emptyset$, und sie schöpfen Ω aus wegen $B \cup B^c = \Omega$. Schnitt mit A liefert

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

Schritt 2: Die beiden Teile sind disjunkt.

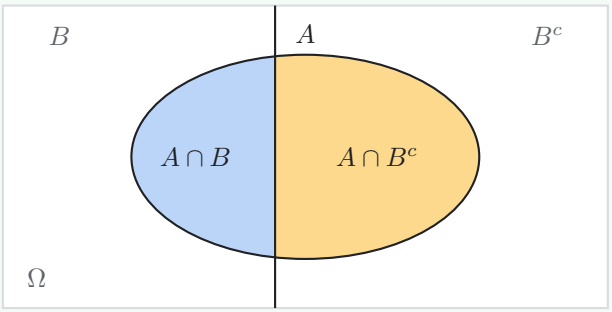
$$(A \cap B) \cap (A \cap B^c) = A \cap (B \cap B^c) = A \cap \emptyset = \emptyset.$$

Schritt 3: Additivität. Punkt (5) des vorigen Satzes erlaubt für disjunkte Ereignisse das Addieren:

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

Jedes Ereignis zerfällt in „mit B“ und „ohne B“

Das Ereignis B zerschneidet den Ergebnisraum Ω in zwei Hälften. Jedes andere Ereignis A wird von diesem Schnitt mitgeteilt – in den Anteil, der zugleich in B liegt, und den Rest:



Die blaue und die gelbe Fläche überlappen sich nirgends und ergeben zusammen genau A. Mehr steckt hinter der Aussage nicht – es ist die Additivität, angewandt auf einen besonders nützlichen Schnitt.

Beispiel 1.3.2 (Fallunterscheidung beim Doppelwurf)

Zweimaliger fairer Münzwurf, $\Omega = \{KK, KZ, ZK, ZZ\}$ mit Gleichverteilung. Sei A = „mindestens einmal Kopf“ $= \{KK, KZ, ZK\}$ und B = „erster Wurf ist Kopf“ $= \{KK, KZ\}$. Wir zerlegen A entlang B:

$$A \cap B = \{KK, KZ\}, \qquad A \cap B^c = \{ZK\},$$

also $\mathbb{P}(A \cap B) = \frac{2}{4} = \frac{1}{2}$ und $\mathbb{P}(A \cap B^c) = \frac{1}{4}$. Die Zerlegung liefert

$$\mathbb{P}(A) = \tfrac{1}{2} + \tfrac{1}{4} = \tfrac{3}{4}.$$

Statt A als Ganzes abzuzählen, haben wir zwei kleinere, klar getrennte Fälle betrachtet. Genau dieses Ereignis begegnet uns weiter unten noch einmal – dort auf dem Weg über die Einschluss-Ausschluss-Formel, mit demselben Ergebnis $3/4$.

Kurz-Check: Was bedeutet das konkret?

$A \cap B$ und $A \cap B^c$ sind beide Teilmengen von A. Warum darf man ihre Wahrscheinlichkeiten trotzdem einfach addieren – und was ginge schief, wenn man A entlang zweier *überlappender* Ereignisse B_1, B_2 zerlegte?

Antwort: Addieren ist erlaubt, weil B und B^c disjunkt sind und sich diese Disjunktheit auf $A \cap B$ und $A \cap B^c$ überträgt – genau das verlangt Punkt (5). Überlappen sich B_1 und B_2, so wird der gemeinsame Teil doppelt gezählt und die Summe fällt zu groß aus. Man muss ihn wieder abziehen – und landet bei der Einschluss-Ausschluss-Formel, dem nächsten Lemma.

Lemma 1.3.3 (Einschluss-Ausschluss)

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

Beweis. Zerlege $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$ disjunkt. Dann

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \qquad \blacksquare$$

Checkerbox 1.3.3 (Einschluss-Ausschluss für drei Ereignisse)

Verallgemeinere das vorige Lemma auf drei Ereignisse und zeige

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) = {} & \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ & - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \\ & + \mathbb{P}(A \cap B \cap C).\end{aligned}$$

Beweis. Wir wenden die Formel für zwei Ereignisse zunächst auf $A \cup B$ und C an:

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A \cup B) + \mathbb{P}(C) - \mathbb{P}\big((A \cup B) \cap C\big).$$

Der erste Term ist

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

Für den letzten Term benutzen wir das Distributivgesetz $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ und nochmals die Formel für zwei Ereignisse:

$$\begin{aligned}\mathbb{P}\big((A \cup B) \cap C\big) &= \mathbb{P}(A \cap C) + \mathbb{P}(B \cap C) \\ &\quad - \mathbb{P}(A \cap B \cap C).\end{aligned}$$

Einsetzen liefert

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) &= \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ &\quad - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \\ &\quad + \mathbb{P}(A \cap B \cap C).\end{aligned}$$

Die paarweisen Schnittmengen werden zunächst doppelt gezählt und daher abgezogen. Der Dreifachschnitt wurde dabei dreimal addiert und dreimal abgezogen; deshalb muss er einmal wieder hinzugefügt werden. ■

Checkerbox 1.3.4 (Siebformel)

Beweise das Inklusions-Exklusions-Prinzip für n Ereignisse:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) \\ &\quad - \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap \cdots \cap A_n).\end{aligned}$$

Beweis. Wir beweisen die Aussage durch Induktion über n. Übersichtlich lässt sich die behauptete Summe als

$$\sum_{\emptyset \neq J \subset \{1,\dots,n\}} (-1)^{|J|+1} \mathbb{P}\left(\bigcap_{j \in J} A_j\right) \qquad (*)$$

schreiben: Für $|J| = 1$ entstehen die Einzelwahrscheinlichkeiten, für $|J| = 2$ die paarweisen Schnitte und so weiter.

Induktionsanfang. Für $n = 1$ ist die Aussage unmittelbar richtig; für $n = 2$ ist sie genau das bereits bewiesene Einschluss-Ausschluss-Lemma.

Induktionsschritt. Die Formel gelte für n Ereignisse. Setze $U_n = \bigcup_{i=1}^{n} A_i$. Dann liefert die Zweierformel

$$\mathbb{P}(U_n \cup A_{n+1}) = \mathbb{P}(U_n) + \mathbb{P}(A_{n+1}) - \mathbb{P}(U_n \cap A_{n+1}).$$

Mit dem Distributivgesetz gilt

$$U_n \cap A_{n+1} = \bigcup_{i=1}^{n} (A_i \cap A_{n+1}).$$

Auf diese Vereinigung von n Ereignissen dürfen wir die Induktionsvoraussetzung anwenden. Nach dem Subtrahieren ihrer Inklusions-Exklusions-Summe entstehen genau alle Schnitte, die A_{n+1} enthalten: Ein Schnitt mit insgesamt r Mengen erhält das Vorzeichen $(-1)^{r+1}$. Die bereits in $\mathbb{P}(U_n)$ enthaltenen Terme sind dagegen genau die Schnitte ohne A_{n+1}. Zusammen erhält

man somit alle nichtleeren Teilmengen von $\{1, \dots, n+1\}$ mit dem Vorzeichen aus $(*)$. Das ist die Siebformel für $n+1$. ■

Korollar 1.3.4 (Bonferroni-Ungleichung)

$\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1.$

Beweis. $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$, da $\mathbb{P}(A \cup B) \leq 1$. ■

Korollar 1.3.5 (Unionsschranke)

Für beliebige Ereignisse $A_1, \dots, A_n$ gilt $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Beweis. Induktion. $n = 2$: $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2)$. Schritt:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

■

Beispiel 1.3.3

Zweimaliger Münzwurf. Sei H_j = „Kopf beim j-ten Wurf“. Bei Gleichverteilung:

$$\mathbb{P}(H_1 \cup H_2) = \tfrac{1}{2} + \tfrac{1}{2} - \tfrac{1}{4} = \tfrac{3}{4}.$$

Die Wahrscheinlichkeit, mindestens einmal Kopf zu werfen, ist also 75%.

Checkerbox 1.3.5 (Komplement einer Vereinigung)

Seien A und B Ereignisse mit

$$\mathbb{P}(A) = \frac{1}{3}, \qquad \mathbb{P}(B) = \frac{1}{4}, \qquad \mathbb{P}(A \cup B) = \frac{1}{2}.$$

Bestimme $\mathbb{P}(A^c \cap B^c)$.

Beweis. Nach dem De-Morgan-Gesetz ist

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

Gesucht ist also das Gegenereignis zu $A \cup B$. Mit der Komplementformel folgt unmittelbar

$$\mathbb{P}(A^c \cap B^c) = \mathbb{P}\big((A \cup B)^c\big) = 1 - \mathbb{P}(A \cup B) = 1 - \frac{1}{2} = \frac{1}{2}.$$

Die Einzelwerte $\mathbb{P}(A) = 1/3$ und $\mathbb{P}(B) = 1/4$ werden für diese Rechnung nicht benötigt; sie legen zusätzlich $\mathbb{P}(A \cap B) = 1/12$ fest. ■

Satz 1.3.6 (Stetigkeit von Wahrscheinlichkeiten)

Gilt $A_n \to A$, so folgt $\mathbb{P}(A_n) \to \mathbb{P}(A)$.

Beweis. Sei $A_1 \subset A_2 \subset \cdots$ und $A = \bigcup_{i=1}^{\infty} A_i$. Definiere $B_1 = A_1$, $B_n = A_n \setminus A_{n-1}$ für $n \geq 2$. Dann sind $B_1, B_2, \ldots$ disjunkt mit $A_n = \bigcup_{i=1}^{n} B_i$ und $A = \bigcup_{i=1}^{\infty} B_i$. Also

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \mathbb{P}(A).$$

Fallender Fall: A_n^c ist wachsend, also $\mathbb{P}(\bigcap A_i) = 1 - \mathbb{P}(\bigcup A_i^c) = 1 - \lim \mathbb{P}(A_n^c) = \lim \mathbb{P}(A_n)$. ■

Checkerbox 1.3.6 (Stetigkeit von Wahrscheinlichkeiten)

Vervollständige den Beweis des vorigen Satzes im monoton wachsenden Fall: Zeige sorgfältig, dass die Mengen

$$B_1 = A_1, \qquad B_n = A_n \setminus A_{n-1} \quad (n \geq 2)$$

paarweise disjunkt sind und dass $A_n = \bigcup_{i=1}^{n} B_i$ sowie $A = \bigcup_{i=1}^{\infty} B_i$ gilt. Beweise anschließend auch den monoton fallenden Fall.

Beweis. **Wachsender Fall.** Seien $A_1 \subset A_2 \subset \cdots$ und $A = \bigcup_{i=1}^{\infty} A_i$. Für $i < j$ gilt $B_i \subset A_i \subset A_{j-1}$, während $B_j = A_j \setminus A_{j-1}$ außerhalb von A_{j-1} liegt. Daher ist $B_i \cap B_j = \emptyset$; die Mengen $B_1, B_2, \ldots$ sind also paarweise disjunkt.
Außerdem gilt

$$A_1 = B_1, \qquad A_n = A_{n-1} \cup (A_n \setminus A_{n-1}) = A_{n-1} \cup B_n.$$

Induktiv folgt daraus $A_n = \bigcup_{i=1}^{n} B_i$. Vereinigen wir diese Gleichheit über alle n, erhalten wir $A = \bigcup_{i=1}^{\infty} B_i$.
Die Additivität für die disjunkten B_i liefert nun

$$\mathbb{P}(A_n) = \sum_{i=1}^{n} \mathbb{P}(B_i) \quad \text{und} \quad \mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(B_i).$$

Die endlichen Partialsummen der nichtnegativen Reihe konvergieren gegen ihre Summe. Somit gilt $\mathbb{P}(A_n) \to \mathbb{P}(A)$.
Fallender Fall. Seien $A_1 \supset A_2 \supset \cdots$ und $A = \bigcap_{i=1}^{\infty} A_i$. Dann ist $A_1^c \subset A_2^c \subset \cdots$ wachsend und nach De Morgan

$$\bigcup_{i=1}^{\infty} A_i^c = \left(\bigcap_{i=1}^{\infty} A_i \right)^c = A^c.$$

Der bereits bewiesene wachsende Fall ergibt $\mathbb{P}(A_n^c) \to \mathbb{P}(A^c)$. Mit der Komplementformel folgt schließlich

$$\mathbb{P}(A_n) = 1 - \mathbb{P}(A_n^c) \longrightarrow 1 - \mathbb{P}(A^c) = \mathbb{P}(A).$$

■

1.4 Wahrscheinlichkeit auf endlichen Ergebnisräumen

Sei $\Omega = \{\omega_1, \dots, \omega_n\}$ endlich. Sind alle Ergebnisse gleichwahrscheinlich, so gilt die **Gleichverteilung**:

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

Um $|A|$ zu bestimmen, brauchen wir **kombinatorische Methoden**. Die Anzahl der Anordnungen von n Objekten ist $n! = n(n-1)\cdots 1$ (mit $0! = 1$). Der Binomialkoeffizient

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

gibt die Anzahl der Möglichkeiten, k Objekte aus n auszuwählen (ohne Berücksichtigung der Reihenfolge).

1.5 Unabhängige Ereignisse

Wirft man zweimal eine faire Münze, ist die Wahrscheinlichkeit für zweimal Kopf gleich $\frac{1}{2} \times \frac{1}{2}$. Wir multiplizieren, weil die Würfe als **unabhängig** gelten – das Ergebnis des ersten Wurfs beeinflusst den zweiten nicht.

Definition 1.5.1

Zwei Ereignisse A und B sind **unabhängig**, geschrieben $A \perp B$, falls

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

Eine Familie $\{A_i : i \in I\}$ ist unabhängig, falls für jede endliche Teilmenge $J \subset I$ gilt

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i).$$

Bemerkung 1.5.1

Sind A und B disjunkt mit $\mathbb{P}(A), \mathbb{P}(B) > 0$, so können sie *nicht* unabhängig sein: $\mathbb{P}(A) \cdot \mathbb{P}(B) > 0$, aber $\mathbb{P}(A \cap B) = 0$. Disjunktheit und Unabhängigkeit sind gegensätzliche Konzepte!

Checkerbox 1.5.1 (Unabhängigkeit und Komplemente)

Seien A und B unabhängig. Zeige, dass auch A^c und B, A und B^c sowie A^c und B^c unabhängig sind.

Beweis. Aus der Unabhängigkeit folgt $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Zunächst zerfällt A disjunkt in $A \cap B$ und $A \cap B^c$. Daher

$$\begin{aligned}\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \mathbb{P}(A)\big(1 - \mathbb{P}(B)\big) = \mathbb{P}(A)\mathbb{P}(B^c).\end{aligned}$$

Also sind A und B^c unabhängig. Vertauschen wir die Rollen von A und B, erhalten wir ebenso

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

Somit sind auch A^c und B unabhängig.
Schließlich zerfällt A^c disjunkt in $A^c \cap B$ und $A^c \cap B^c$. Deshalb

$$\begin{aligned}\mathbb{P}(A^c \cap B^c) &= \mathbb{P}(A^c) - \mathbb{P}(A^c \cap B) \\ &= \mathbb{P}(A^c) - \mathbb{P}(A^c)\mathbb{P}(B) \\ &= \mathbb{P}(A^c)\mathbb{P}(B^c).\end{aligned}$$

Damit sind auch A^c und B^c unabhängig. ■

Checkerbox 1.5.2 (Ziehen ohne Zurücklegen)

Eine Urne enthält drei rote und zwei blaue Kugeln. Es werden zweimal ohne Zurücklegen Kugeln gezogen. Sei

$$A = \text{„die erste Kugel ist rot“} \quad \text{und} \quad B = \text{„die zweite Kugel ist rot“}.$$

Sind A und B unabhängig? Begründe deine Antwort rechnerisch.

Beweis. Für den ersten Zug sind drei der fünf Kugeln rot, also

$$\mathbb{P}(A) = \frac{3}{5}.$$

Aus Symmetrie hat jede Kugel dieselbe Chance, an zweiter Stelle gezogen zu werden. Daher ist auch $\mathbb{P}(B) = 3/5$.
Sind im ersten Zug eine rote Kugel gezogen worden, bleiben zwei rote unter vier Kugeln übrig. Somit

$$\mathbb{P}(B \mid A) = \frac{2}{4} = \frac{1}{2}$$

und folglich

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A) = \frac{3}{5} \cdot \frac{1}{2} = \frac{3}{10}.$$

Bei Unabhängigkeit müsste dagegen

$$\mathbb{P}(A)\mathbb{P}(B) = \frac{3}{5} \cdot \frac{3}{5} = \frac{9}{25}$$

gelten. Da $3/10 \neq 9/25$, sind A und B nicht unabhängig. Anschaulich senkt eine rote Kugel im ersten Zug die Rot-Chance im zweiten Zug von $3/5$ auf $1/2$. ■

Beispiel 1.5.1 (Faire Münze, $n = 10$)

Wir werfen eine faire Münze zehnmal unabhängig. Sei $A =$ „mindestens ein Kopf“.
Mit $T_j =$ „Zahl beim j-ten Wurf“ ergibt sich:

$$\mathbb{P}(A) \;=\; 1 - \mathbb{P}(T_1 \cap \cdots \cap T_{10}) \;=\; 1 - \prod_{j=1}^{10} \mathbb{P}(T_j) \;=\; 1 - \left(\frac{1}{2}\right)^{10} \;=\; 1 - \frac{1}{1024} \;\approx\; 0{,}999.$$

Interpretation: Das Gegenereignis – zehnmal Zahl in Folge – hat die Wahrscheinlichkeit

$$\mathbb{P}(A^c) = \frac{1}{1024} \approx 0{,}098\,\%.$$

Führt man dieses Experiment 1024-mal durch, erwartet man im Schnitt genau *einmal* das Ergebnis $(Z, Z, \ldots, Z)$. Bereits nach $n = 7$ Würfen überschreitet $\mathbb{P}(A_n)$ die $99\,\%$-Marke ($\mathbb{P}(A_7) = 127/128 \approx 99{,}2\,\%$).

Bemerkung 1.5.2 (Warum das Komplementärprinzip?)

Das direkte Berechnen von $\mathbb{P}(A)$ wäre mühsam, da A aus $1024 - 1 = 1023$ Elementarereignissen besteht (alle Folgen außer $(Z, Z, \ldots, Z)$). Das Komplementärprinzip reduziert das Problem auf ein einziges ungünstiges Elementarereignis.

Checkerbox 1.5.3 (Verdoppelungsstrategie: Behauptung prüfen)

Bei einem fairen Spiel wird das gesamte Kapital gesetzt: Bei Gewinn verdoppelt es sich, bei Verlust ist es vollständig verloren. Die Gewinnchance beträgt in jeder unabhängigen Runde $1/2$, das Anfangskapital ist $K_0 = 1$.
Im Übungsbuch wird behauptet, die Wahrscheinlichkeit, irgendwann beliebig reich zu werden, sei 1, während der Erwartungswert des Kapitals stets 1 bleibe. Prüfe diese Behauptung kritisch: Bestimme für jedes n die Verteilung von K_n und $\mathbb{E}[K_n]$ (Vorgriff auf Kapitel 3). Berechne anschließend

$$\mathbb{P}\left(\sup_{n \geq 0} K_n = \infty\right) \qquad \text{und} \qquad \mathbb{P}(K_n = 0 \text{ für ein } n \geq 1)\,.$$

Beweis. Nach dem ersten Verlust ist das Kapital 0 und bleibt dort. Nach n Runden gibt es daher nur zwei mögliche Kapitalstände:

$$K_n = \begin{cases} 2^n, & \text{wenn alle ersten } n \text{ Spiele gewonnen wurden,} \\ 0, & \text{wenn mindestens eines davon verloren wurde.} \end{cases}$$

Wegen der Unabhängigkeit der Spiele ist die Wahrscheinlichkeit einer Serie von n Gewinnen gleich $(1/2)^n$. Damit

$$\mathbb{P}(K_n = 2^n) = 2^{-n}, \qquad \mathbb{P}(K_n = 0) = 1 - 2^{-n}.$$

Der Erwartungswert zu jedem festen Zeitpunkt ist tatsächlich konstant:

$$\mathbb{E}[K_n] = 2^n \cdot 2^{-n} + 0 \cdot (1 - 2^{-n}) = 1.$$

Um beliebig reich zu werden, müsste der Spieler ohne einen einzigen Verlust unendlich lange gewinnen. Bezeichne mit G_n das Ereignis „die ersten n Spiele werden gewonnen“. Dann gilt $G_1 \supset G_2 \supset \cdots$ und $\mathbb{P}(G_n) = 2^{-n}$. Mit der Stetigkeit von Wahrscheinlichkeiten folgt

$$\mathbb{P}\left(\sup_{n\geq 0} K_n = \infty\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} G_n\right) = \lim_{n\to\infty} 2^{-n} = 0.$$

Das Gegenereignis ist, dass irgendwann mindestens ein Verlust eintritt und das Kapital auf 0 fällt. Daher

$$\mathbb{P}(K_n = 0 \text{ für ein } n \geq 1) = 1 - 0 = 1.$$

Die ursprüngliche Behauptung ist also in ihrem ersten Teil falsch: Fast sicher tritt der Ruin ein, obwohl $\mathbb{E}[K_n] = 1$ für jedes feste n gilt. ■

Beispiel 1.5.2 (Abwechselnde Würfe – Erstvorteil)

Zwei Personen werfen abwechselnd auf einen Basketballkorb; Person 1 beginnt. Die Trefferwahrscheinlichkeiten seien $p_1 = \frac{1}{3}$ und $p_2 = \frac{1}{4}$, alle Würfe sind unabhängig. Gesucht ist $\mathbb{P}(E)$ mit $E =$ „Person 1 trifft zuerst“.

Modell. Person 1 wirft in den Runden $1, 3, 5, \ldots$, also beim k-ten eigenen Versuch in Runde $2k-1$ ($k = 1, 2, 3, \ldots$). Zu diesem Zeitpunkt sind genau $2(k-1)$ Würfe absolviert – je $k-1$ von beiden Personen, alle ohne Treffer. Die Wahrscheinlichkeit, einen vollständigen Zyklus (je ein Fehlwurf beider Personen) zu überstehen, beträgt:

$$q \;\coloneqq\; (1-p_1)(1-p_2) \;=\; \frac{2}{3}\cdot\frac{3}{4} \;=\; \frac{1}{2}.$$

Die Wahrscheinlichkeit, dass *genau* in Runde $2k-1$ Person 1 zum ersten Mal trifft, ist daher $q^{k-1}\cdot p_1$.

Berechnung. Da die Ereignisse „Person 1 gewinnt in Runde $2k-1$“ für verschiedene k disjunkt sind, ergibt die Summe über alle Möglichkeiten:

$$\mathbb{P}(E) = \sum_{k=1}^{\infty} q^{k-1}\cdot p_1 = p_1 \sum_{j=0}^{\infty} q^j = \frac{p_1}{1-q} = \frac{\frac{1}{3}}{1-\frac{1}{2}} = \frac{2}{3}.$$

Kontrolle. Analog gewinnt Person 2 nur, wenn Person 1 vorher verfehlt; ihre Gewinnwahrscheinlichkeit ist:

$$\mathbb{P}(F) = \frac{(1-p_1)\,p_2}{1-q} = \frac{\frac{2}{3}\cdot\frac{1}{4}}{\frac{1}{2}} = \frac{1}{3}, \qquad \mathbb{P}(E)+\mathbb{P}(F) = 1. \;\checkmark$$

Allgemein und Fairness. Für beliebige $p_1, p_2 \in (0,1)$ mit $q = (1-p_1)(1-p_2) < 1$ gilt $\mathbb{P}(E) = p_1/(1-q)$. Das Spiel ist genau dann fair ($\mathbb{P}(E) = \frac{1}{2}$), wenn $p_1 = (1-p_1)\,p_2$. Für $p_1 = \frac{1}{3}$ liefert das $p_2 = \frac{1}{2}$ – Person 2 müsste ihre eigene Quote also verdoppeln (von $\frac{1}{4}$ auf $\frac{1}{2}$) und damit das $1{,}5$-Fache der Quote von Person 1 erreichen, um den Erstvorteil zu neutralisieren.

Person 1 gewinnt hier mit Wahrscheinlichkeit $\frac{2}{3}$, also doppelt so oft wie Person 2 – obwohl ihre Einzeltreffsicherheit mit $\frac{1}{3}$ gegenüber $\frac{1}{4}$ nur um den Faktor $\frac{4}{3}$ höher liegt. Der Vorteil des ersten

Wurfs verstärkt den kleinen Vorsprung in der Trefferquote. Dieses Muster – geometrische Reihe als Modell für *wer trifft zuerst?* – tritt auch in Wartezeitproblemen, Bernoulli-Ketten und der Erneuerungstheorie auf; die Konvergenz ist wegen $q < 1$ stets gesichert ($\lim_{n\to\infty} q^n = 0$, das Spiel endet fast sicher).

1.6 Bedingte Wahrscheinlichkeit

Definition 1.6.1 (Bedingte Wahrscheinlichkeit)

Falls $\mathbb{P}(B) > 0$, ist die **bedingte Wahrscheinlichkeit** von A gegeben B:

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

Intuition. $\mathbb{P}(A \mid B)$ beantwortet die Frage: *Wie wahrscheinlich ist A, wenn ich bereits weiß, dass B eingetreten ist?* Das Wissen über B verkleinert den relevanten Ergebnisraum von Ω auf B; innerhalb dieses neuen Raums wird $\mathbb{P}(A \cap B)$ auf $\mathbb{P}(B)$ normiert, damit die Axiome der Wahrscheinlichkeit erhalten bleiben.

Geometrische Lesart. Stellt man Ω als Fläche dar, so ist $\mathbb{P}(A \mid B)$ der *Anteil von $A \cap B$ an der Fläche B*:

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

Nur der Schnittbereich $A \cap B$ zählt; alles außerhalb von B ist durch die Bedingung ausgeblendet.

Äquivalente Formulierung: Multiplikationsregel. Umstellen liefert die für Berechnungen zentrale Form:

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

Sie erlaubt, Verbundwahrscheinlichkeiten schrittweise aus bedingten Wahrscheinlichkeiten aufzubauen – Grundlage von Baumdiagrammen und der totalen Wahrscheinlichkeit.

Bedingte Wahrscheinlichkeit als Wahrscheinlichkeitsmaß. $\mathbb{P}(\,\cdot \mid B)$ erfüllt selbst alle Kolmogorov-Axiome auf Ω, insbesondere:

$$\begin{aligned} \mathbb{P}(\Omega \mid B) &= 1, \\ \mathbb{P}(A_1 \cup A_2 \mid B) &= \mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B) \quad \text{für } A_1 \cap A_2 = \emptyset. \end{aligned}$$

Die gesamte Wahrscheinlichkeitsrechnung gilt also unverändert im *konditionierten Universum* B.

Stochastische Unabhängigkeit als Spezialfall. Die in Abschnitt 1.5 über $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$ definierte Unabhängigkeit liest sich hier als „das Eintreten von B liefert keinerlei Information über A“: Setzt man das Produkt in den Quotienten ein, kürzt sich $\mathbb{P}(B)$ heraus und es bleibt $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Lemma 1.6.3 führt das aus und zeigt zugleich, dass sich die Unabhängigkeit auf die Komplemente überträgt.

Beispiel 1.6.1 (Würfel)

Ein fairer Würfel. Sei $A = \{6\}$, $B = \{\text{gerade Zahl}\} = \{2, 4, 6\}$.

$$\mathbb{P}(A \mid B) \;=\; \frac{\mathbb{P}(\{6\})}{\mathbb{P}(\{2,4,6\})} \;=\; \frac{\frac{1}{6}}{\frac{3}{6}} \;=\; \frac{1}{3}.$$

Ohne Zusatzwissen: $\mathbb{P}(A) = \frac{1}{6}$. Mit dem Wissen „die Zahl ist gerade“ steigt die Wahrscheinlichkeit auf $\frac{1}{3}$ – der Ergebnisraum schrumpft von $\{1, \ldots, 6\}$ auf $\{2, 4, 6\}$.

Voraussetzung $\mathbb{P}(B) > 0$**.** Die Definition ist nur sinnvoll, wenn B tatsächlich eintreten kann. Für $\mathbb{P}(B) = 0$ ist die Bedingung „B ist eingetreten“ informationslos; der Quotient wäre undefiniert. In modernen Ansätzen (Maßtheorie, reguläre bedingte Wahrscheinlichkeiten) wird dieser Fall gesondert behandelt.

Lemma 1.6.2 (Produktregel)

Für Ereignisse $A_1, \ldots, A_n$ mit $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$ gilt:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) \;=\; \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \mid A_1) \cdot \mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Beweis. Durch vollständige Induktion über $n \geq 2$.

Induktionsanfang ($n = 2$). Aus der Definition der bedingten Wahrscheinlichkeit folgt unmittelbar:

$$\mathbb{P}(A_1 \cap A_2) \;=\; \mathbb{P}(A_2 \mid A_1) \cdot \mathbb{P}(A_1). \tag{IA}$$

Induktionsschritt ($n - 1 \to n$). Sei die Aussage für $n - 1$ Ereignisse bereits bewiesen. Setze $B := A_1 \cap \cdots \cap A_{n-1}$; nach Voraussetzung ist $\mathbb{P}(B) > 0$. Dann gilt:

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) &= \mathbb{P}(A_n \cap B) \\
&= \mathbb{P}(A_n \mid B) \cdot \mathbb{P}(B) && \text{(wie (IA), da } \mathbb{P}(B) > 0) \\
&= \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}) \cdot \mathbb{P}\left(\bigcap_{i=1}^{n-1} A_i\right) \\
&= \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}) \cdot \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \mid A_1) \cdots \mathbb{P}(A_{n-1} \mid A_1 \cap \cdots \cap A_{n-2}). && \text{(IV)}
\end{aligned}$$

Das ist genau die behauptete Formel für n. ∎

Checkerbox 1.6.1 (Wechselseitige Erhöhung)

Zeige: Gilt $\mathbb{P}(A \mid B) > \mathbb{P}(A)$, dann gilt auch $\mathbb{P}(B \mid A) > \mathbb{P}(B)$. Begründe dabei, warum beide bedingten Wahrscheinlichkeiten definiert sind.

Beweis. Da $\mathbb{P}(A \mid B)$ in der Voraussetzung vorkommt, gilt zunächst $\mathbb{P}(B) > 0$. Aus der Definition der bedingten Wahrscheinlichkeit erhalten wir

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

Multiplikation mit der positiven Zahl $\mathbb{P}(B)$ liefert

$$\mathbb{P}(A \cap B) > \mathbb{P}(A)\mathbb{P}(B). \tag{1}$$

Insbesondere ist $\mathbb{P}(A \cap B) > 0$. Wegen $A \cap B \subset A$ folgt daraus $\mathbb{P}(A) > 0$; somit ist auch $\mathbb{P}(B \mid A)$ definiert. Teilen wir (1) durch die positive Zahl $\mathbb{P}(A)$, ergibt sich

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} > \mathbb{P}(B). \qquad \blacksquare$$

Lemma 1.6.3

Sind A und B unabhängig, so gilt $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Zudem sind dann auch A und B^c, A^c und B sowie A^c und B^c unabhängig.

Beweis. **(i)** $\mathbb{P}(A \mid B) = \mathbb{P}(A)$**.** Da $A \perp B$, gilt $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Mit $\mathbb{P}(B) > 0$:

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

(ii) $A \perp B^c$**.**

$$\begin{aligned} \mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \mathbb{P}(A)\big(1 - \mathbb{P}(B)\big) = \mathbb{P}(A)\mathbb{P}(B^c). \end{aligned}$$

(iii) $A^c \perp B$**.** Analog zu (ii) mit den Rollen von A und B vertauscht:

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B)(1 - \mathbb{P}(A)) = \mathbb{P}(A^c)\mathbb{P}(B).$$

(iv) $A^c \perp B^c$**.** Mit dem Ergebnis aus (iii):

$$\mathbb{P}(A^c \cap B^c) = \mathbb{P}(A^c) - \mathbb{P}(A^c \cap B) = \mathbb{P}(A^c) - \mathbb{P}(A^c)\mathbb{P}(B) = \mathbb{P}(A^c)(1 - \mathbb{P}(B)) = \mathbb{P}(A^c)\mathbb{P}(B^c). \qquad \blacksquare$$

Checkerbox 1.6.2 (Simpson-Paradoxon)

Konstruiere ein Beispiel mit Ereignissen A, B, C_1, C_2, wobei C_1, C_2 eine Zerlegung des Ergebnisraums bilden, sodass

$$\mathbb{P}(A \mid B \cap C_1) < \mathbb{P}(A \mid B^c \cap C_1), \qquad \mathbb{P}(A \mid B \cap C_2) < \mathbb{P}(A \mid B^c \cap C_2),$$

aber nach Zusammenfassen beider Gruppen

$$\mathbb{P}(A \mid B) > \mathbb{P}(A \mid B^c)$$

gilt. Gib eine konkrete Vierfeldertafel für jede der beiden Gruppen an und prüfe alle drei Ungleichungen rechnerisch.

Beweis. Wir betrachten 200 gleichgewichtete Personen. Das Ereignis A bedeute „Erfolg“, B bedeute „Behandlung“, und C_1, C_2 seien zwei disjunkte Gruppen, die zusammen alle Personen enthalten. Die Häufigkeiten seien:

Gruppe	Status	A	A^c	Gesamt
C_1	B	81	9	90
C_1	B^c	19	1	20
C_2	B	1	9	10
C_2	B^c	16	64	80

Innerhalb der ersten Gruppe ist die Erfolgsquote unter B kleiner:

$$\mathbb{P}(A \mid B \cap C_1) = \frac{81}{90} = 0{,}90 < \frac{19}{20} = 0{,}95 = \mathbb{P}(A \mid B^c \cap C_1).$$

Dasselbe gilt innerhalb der zweiten Gruppe:

$$\mathbb{P}(A \mid B \cap C_2) = \frac{1}{10} = 0{,}10 < \frac{16}{80} = 0{,}20 = \mathbb{P}(A \mid B^c \cap C_2).$$

Nach dem Zusammenfassen kehrt sich der Vergleich jedoch um. Unter B gibt es $81 + 1 = 82$ Erfolge bei $90 + 10 = 100$ Personen; unter B^c gibt es $19 + 16 = 35$ Erfolge bei $20 + 80 = 100$ Personen. Also

$$\mathbb{P}(A \mid B) = \frac{82}{100} = 0{,}82 > \frac{35}{100} = 0{,}35 = \mathbb{P}(A \mid B^c).$$

Der Grund für die Umkehrung ist die sehr unterschiedliche Zusammensetzung: B kommt überwiegend in der erfolgsstarken Gruppe C_1 vor, B^c überwiegend in der erfolgsschwachen Gruppe C_2. Das Vermischen dieser Gruppen verdeckt daher die beiden Vergleiche innerhalb der Gruppen. ■

1.7 Satz von Bayes

Satz 1.7.1 (Gesetz der totalen Wahrscheinlichkeit)

Sei $(A_i)_{i=1}^{k}$ eine Zerlegung von Ω mit $\mathbb{P}(A_i) > 0$ für alle i. Dann gilt:

$$\mathbb{P}(B) \;=\; \sum_{i=1}^{k} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i).$$

Beweis. **Schritt 1: Zerlegung von B.** Da $A_1, \dots, A_k$ eine Zerlegung von Ω bilden, gilt $\bigcup_{i=1}^{k} A_i = \Omega$ und $A_i \cap A_j = \emptyset$ für $i \neq j$. Schnitt mit B liefert:

$$B \;=\; B \cap \Omega \;=\; B \cap \bigcup_{i=1}^{k} A_i \;=\; \bigcup_{i=1}^{k} (B \cap A_i).$$

Die Mengen $B \cap A_i$ sind paarweise disjunkt, denn für $i \neq j$:

$$(B \cap A_i) \cap (B \cap A_j) = B \cap (A_i \cap A_j) = B \cap \emptyset = \emptyset.$$

Schritt 2: σ-Additivität. Da die $B \cap A_i$ disjunkt sind, liefert das dritte Kolmogorov-Axiom:

$$\mathbb{P}(B) \;=\; \mathbb{P}\left(\bigcup_{i=1}^{k} (B \cap A_i) \right) \;=\; \sum_{i=1}^{k} \mathbb{P}(B \cap A_i).$$

Schritt 3: Multiplikationsregel. Da $\mathbb{P}(A_i) > 0$ für alle i, ist $\mathbb{P}(B \mid A_i)$ definiert und es gilt:

$$\mathbb{P}(B \cap A_i) \;=\; \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i).$$

Einsetzen in Schritt 2 ergibt:

$$\mathbb{P}(B) \;=\; \sum_{i=1}^{k} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i).$$

■

Bemerkung 1.7.1 (Derselbe Gedanke, nur feiner zerschnitten)

Die Schritte 1 und 2 dieses Beweises sind wortgleich die Schritte von Checkerbox 1.3.2: erst das interessierende Ereignis entlang einer Zerlegung von Ω aufschneiden, dann die disjunkten Stücke addieren. Dort bestand die Zerlegung aus der gröbstmöglichen Wahl – einem Ereignis und seinem Komplement, also zwei Fällen. Hier sind es k Fälle. Neu ist allein Schritt 3: Weil inzwischen die bedingte Wahrscheinlichkeit zur Verfügung steht, lässt sich jeder Summand $\mathbb{P}(B \cap A_i)$ als $\mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)$ schreiben – und erst dadurch wird aus der Zerlegungsformel ein Rechenwerkzeug.

Satz 1.7.2 (Satz von Bayes)

Unter den Voraussetzungen des vorigen Satzes gilt für $\mathbb{P}(B) > 0$

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)}{\sum_{j=1}^{k} \mathbb{P}(B \mid A_j) \cdot \mathbb{P}(A_j)}.$$

Beweis. $\mathbb{P}(A_i \mid B) = \mathbb{P}(A_i \cap B)/\mathbb{P}(B) = \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)/\mathbb{P}(B)$. Der Nenner ist das Gesetz der totalen Wahrscheinlichkeit. ■

Bemerkung 1.7.2

$\mathbb{P}(A_i)$ heißt **A-priori-Wahrscheinlichkeit** (Prior), $\mathbb{P}(A_i \mid B)$ heißt **A-posteriori-Wahrscheinlichkeit** (Posterior). Bayes beschreibt, wie neue Information B unsere Überzeugung über A_i aktualisiert.

Korollar 1.7.3 (Bayes für zwei Ereignisse)

Für $\mathbb{P}(A), \mathbb{P}(B) > 0$ gilt

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A) + \mathbb{P}(B \mid A^c) \cdot \mathbb{P}(A^c)}.$$

Beweis. Die beiden Ereignisse A und A^c bilden eine Zerlegung von Ω: Sie sind disjunkt, ihre Vereinigung ist Ω, und wegen $0 < \mathbb{P}(A) < 1$ – was aus $\mathbb{P}(A) > 0$ und $\mathbb{P}(B) > 0$ zusammen mit $\mathbb{P}(B \mid A^c)$ stillschweigend vorausgesetzt ist – haben beide positive Wahrscheinlichkeit. Der Satz von Bayes mit $k = 2$, $A_1 = A$ und $A_2 = A^c$ liefert unmittelbar

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A) + \mathbb{P}(B \mid A^c) \cdot \mathbb{P}(A^c)}.$$

Der Nenner ist dabei nichts anderes als das Gesetz der totalen Wahrscheinlichkeit für die Zerlegung $\{A, A^c\}$, also $\mathbb{P}(B)$. ■

Beispiel 1.7.1 (Medizinischer Test – Basisraten-Paradoxon)

Situation. Ein diagnostischer Test auf die Krankheit D ist technisch nahezu perfekt:

$$\underbrace{\mathbb{P}(+ \mid D)}_{\text{Sensitivität}} = 0{,}993, \qquad \underbrace{\mathbb{P}(- \mid D^c)}_{\text{Spezifität}} = 0{,}9999.$$

Die Krankheit ist jedoch extrem selten:

$$\mathbb{P}(D) = 0{,}0001 \quad \text{(1 von 10 000 Personen ist krank).}$$

Gesucht: $\mathbb{P}(D \mid +)$ – die Wahrscheinlichkeit, *tatsächlich krank zu sein*, gegeben ein positives Testergebnis.

Schritt 1: Alle vier Testausgänge. Aus den gegebenen Größen ergibt sich das vollständige Bild:

$$\begin{aligned}
\mathbb{P}(+ \mid D) &= 0{,}9930 && \text{(richtig positiv – Sensitivität)} \\
\mathbb{P}(- \mid D) &= 0{,}0070 && \text{(falsch negativ)} \\
\mathbb{P}(- \mid D^c) &= 0{,}9999 && \text{(richtig negativ – Spezifität)} \\
\mathbb{P}(+ \mid D^c) &= 0{,}0001 && \text{(falsch positiv – Rate = } 1- \text{ Spezifität)}
\end{aligned}$$

Schritt 2: Gesamtwahrscheinlichkeit eines positiven Tests. Mit dem Gesetz der totalen Wahrscheinlichkeit (Satz 1.7.1, Zerlegung $\{D, D^c\}$ von Ω):

$$\begin{aligned}\mathbb{P}(+) &= \mathbb{P}(+ \mid D)\,\mathbb{P}(D) \;+\; \mathbb{P}(+ \mid D^c)\,\mathbb{P}(D^c)\\ &= 0{,}9930 \times 0{,}0001 \;+\; 0{,}0001 \times 0{,}9999\\ &= 0{,}000\,099\,3 \;+\; 0{,}000\,099\,99\\ &= 0{,}000\,199\,29 \;\approx\; 0{,}000\,199\,3.\end{aligned}$$

Damit sind von $10\,000$ zufällig getesteten Personen im Schnitt ≈ 2 positiv getestet – je etwa *eine* davon krank, *eine* gesund aber falsch positiv.

Schritt 3: Satz von Bayes (Korollar 1.7.3)**.**

$$\begin{aligned}\mathbb{P}(D \mid +) &= \frac{\mathbb{P}(+ \mid D)\,\mathbb{P}(D)}{\mathbb{P}(+)}\\ &= \frac{0{,}9930 \times 0{,}0001}{0{,}000\,199\,3}\\ &= \frac{0{,}000\,099\,3}{0{,}000\,199\,3} \;\approx\; \mathbf{0{,}498}.\end{aligned} \tag{1.1}$$

Ergebnis

Trotz Sensitivität $99{,}3\,\%$ und Spezifität $99{,}99\,\%$ ist die Wahrscheinlichkeit, bei positivem Befund tatsächlich krank zu sein, nur $\approx 50\,\%$.

Beispiel 1.7.2 (Medizinischer Test – Fortsetzung)

Warum? – Natürliche Häufigkeiten. Betrachte eine Referenzpopulation von $1\,000\,000$ Personen:

	Test $+$	Test $-$	Gesamt
Krank (D)	99	1	100
Gesund (D^c)	100	999 800	999 900
Gesamt	**199**	999 801	1 000 000

Von den 199 positiv Getesteten sind 99 tatsächlich krank – also $99/199 \approx 49{,}7\,\%$. Die 100 *falsch positiven* Gesunden entstehen, weil die Spezifitätslücke von $0{,}01\,\%$ auf die riesige Gruppe der $999\,900$ Gesunden wirkt. **Der entscheidende Faktor: die Prävalenz.** Schreibt man den Bayes-Quotienten explizit aus, wird sichtbar, warum die Basisrate dominiert:

$$\frac{\mathbb{P}(D \mid +)}{\mathbb{P}(D^c \mid +)} = \underbrace{\frac{\mathbb{P}(+ \mid D)}{\mathbb{P}(+ \mid D^c)}}_{\text{Likelihood-Ratio} = 9930} \cdot \underbrace{\frac{\mathbb{P}(D)}{\mathbb{P}(D^c)}}_{\text{Prior-Odds} \approx 0{,}0001}.$$

Die Likelihood-Ratio $LR = 0{,}993/0{,}0001 = 9930$ ist eindrucksvoll. Aber die Prior-Odds sind $0{,}0001/0{,}9999 \approx 1{:}10\,000$. Das Produkt $9930 \times 0{,}0001 \approx 0{,}993$ liefert Posterior-Odds nahe $1:1$, also $\mathbb{P}(D \mid +) \approx 50\,\%$.

Basisraten-Paradoxon. Das *Basisraten-Paradoxon* (*base rate neglect*) beschreibt den kognitiven Fehler, die Prävalenz $\mathbb{P}(D)$ zu ignorieren und stattdessen $\mathbb{P}(D \mid +) \approx \mathbb{P}(+ \mid D)$ zu setzen – hier eine Überschätzung um den Faktor ≈ 2. In der Medizin führt das zu:

- falscher Panik nach positivem Befund bei Screening-Tests,
- Über-Diagnose und unnötigen Folgeuntersuchungen,
- systematischer Fehlkalibrierung ärztlicher Urteilskraft.

Sensitivitätsanalyse: Einfluss der Prävalenz. Die Formel

$$\mathbb{P}(D \mid +) = \frac{p_1 \cdot \pi}{p_1 \cdot \pi + p_2(1-\pi)}, \quad p_1 = \mathbb{P}(+ \mid D),\ p_2 = \mathbb{P}(+ \mid D^c),\ \pi = \mathbb{P}(D)$$

zeigt: Mit steigender Prävalenz π wächst $\mathbb{P}(D \mid +)$ rasch.

Prävalenz π	$\mathbb{P}(D \mid +)$
$0{,}01\,\%$	$\approx 50\,\%$
$0{,}1\,\%$	$\approx 91\,\%$
$1\,\%$	$\approx 99\,\%$
$10\,\%$	$\approx 99{,}9\,\%$

Bei gezielter Testung einer *Risikogruppe* (höheres π) ist der prädiktive Wert dramatisch besser – derselbe Test, deutlich verlässlichere Aussage. Screening-Programme für seltene Krankheiten erfordern deshalb zwingend eine Vorselektion oder einen Bestätigungstest.

Checkerbox 1.7.1 (Seltener Krankheitstest)

Ein Test auf eine seltene Krankheit D hat eine Sensitivität von $95\,\%$ und eine Spezifität von $90\,\%$. Die Prävalenz beträgt $\mathbb{P}(D) = 0{,}001$. Wie groß ist die Wahrscheinlichkeit $\mathbb{P}(D \mid +)$, tatsächlich krank zu sein, wenn der Test positiv ausfällt?
Übersetze Sensitivität und Spezifität zunächst in bedingte Wahrscheinlichkeiten und verwende dann den Satz von Bayes.

Beweis. Sensitivität $95\,\%$ bedeutet

$$\mathbb{P}(+ \mid D) = 0{,}95.$$

Spezifität $90\,\%$ bedeutet $\mathbb{P}(- \mid D^c) = 0{,}90$; die Rate falsch positiver Tests unter Gesunden ist daher

$$\mathbb{P}(+ \mid D^c) = 1 - 0{,}90 = 0{,}10.$$

Außerdem gilt $\mathbb{P}(D) = 0{,}001$ und $\mathbb{P}(D^c) = 0{,}999$. Mit Bayes folgt

$$\begin{aligned}\mathbb{P}(D \mid +) &= \frac{\mathbb{P}(+ \mid D)\mathbb{P}(D)}{\mathbb{P}(+ \mid D)\mathbb{P}(D) + \mathbb{P}(+ \mid D^c)\mathbb{P}(D^c)} \\ &= \frac{0{,}95 \cdot 0{,}001}{0{,}95 \cdot 0{,}001 + 0{,}10 \cdot 0{,}999} \\ &= \frac{0{,}00095}{0{,}10085} \approx 0{,}00942.\end{aligned}$$

Somit beträgt die gesuchte Wahrscheinlichkeit nur ungefähr $0{,}942\,\%$. Die hohe Zahl falsch positiver Tests entsteht, weil die Krankheit sehr selten ist und die Fehlalarmrate von $10\,\%$ auf die große Gruppe der Gesunden wirkt. ■

Beispiel 1.7.3 (E-Mail-Klassifikation – Bayes'scher Spam-Filter)

Situation. E-Mails werden in drei disjunkte Kategorien eingeteilt, die eine Zerlegung von Ω bilden:

$$A_1 = \text{Spam}, \quad A_2 = \text{niedrige Priorität}, \quad A_3 = \text{hohe Priorität},$$

mit den *Priorwahrscheinlichkeiten* (Basisraten aus historischen Daten):

$$\mathbb{P}(A_1) = 0{,}7, \quad \mathbb{P}(A_2) = 0{,}2, \quad \mathbb{P}(A_3) = 0{,}1.$$

Das Schlüsselwort B = „enthält das Wort ‚frei'" tritt kategorienabhängig auf:

$$\mathbb{P}(B \mid A_1) = 0{,}9, \qquad \mathbb{P}(B \mid A_2) = \mathbb{P}(B \mid A_3) = 0{,}01.$$

„Frei" ist also ein starkes Spam-Signal: Es erscheint in $90\,\%$ aller Spam-Mails, aber nur in $1\,\%$ der legitimen Mails.

Schritt 1: Totale Wahrscheinlichkeit von B. Mit Satz 1.7.1 und der Zerlegung $\{A_1, A_2, A_3\}$:

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \mid A_1)\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\mathbb{P}(A_2) + \mathbb{P}(B \mid A_3)\mathbb{P}(A_3) \\ &= 0{,}9 \times 0{,}7 \;+\; 0{,}01 \times 0{,}2 \;+\; 0{,}01 \times 0{,}1 \\ &= 0{,}630 \;+\; 0{,}002 \;+\; 0{,}001 \\ &= 0{,}633.\end{aligned}$$

Von allen E-Mails enthalten also $63{,}3\,\%$ das Wort „frei" – dominiert vom Spam-Anteil $0{,}630$.

Schritt 2: Satz von Bayes für alle drei Kategorien.

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\mathbb{P}(B)}, \quad i = 1, 2, 3.$$

Kategorie	$\mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)$	$\mathbb{P}(A_i \mid B)$	Interpretation
A_1 Spam	$0{,}9 \times 0{,}7 = 0{,}630$	$\frac{0{,}630}{0{,}633} \approx 0{,}9953$	$99{,}5\,\%$ Spam
A_2 niedrig	$0{,}01 \times 0{,}2 = 0{,}002$	$\frac{0{,}002}{0{,}633} \approx 0{,}0032$	$0{,}3\,\%$
A_3 hoch	$0{,}01 \times 0{,}1 = 0{,}001$	$\frac{0{,}001}{0{,}633} \approx 0{,}0016$	$0{,}2\,\%$
Summe	$0{,}633$	$1{,}0000$	

Ergebnis

$$\mathbb{P}(A_1 \mid B) = \frac{0{,}630}{0{,}633} \approx 0{,}995.$$

Eine E-Mail mit dem Wort „frei“ ist mit $99{,}5\,\%$ Wahrscheinlichkeit Spam.

Beispiel 1.7.4 (Fortsetzung)

Warum so eindeutig? – Likelihood-Ratio. Das Verhältnis der Spam-Evidenz gegenüber dem Nicht-Spam-Durchschnitt:

$$LR = \frac{\mathbb{P}(B \mid A_1)}{\mathbb{P}(B \mid A_2)} = \frac{0{,}9}{0{,}01} = 90.$$

Das Wort „frei“ ist 90-mal häufiger in Spam als in legitimen Mails – ein extrem starkes diagnostisches Signal. Kombiniert mit dem bereits hohen Prior $\mathbb{P}(A_1) = 0{,}7$ bleibt kaum Raum für Zweifel.

Bayes'sche Filterlogik. Praxissysteme verallgemeinern dieses Prinzip auf n Schlüsselwörter $B_1, \ldots, B_n$. Unter der – vereinfachenden – *Naïve-Bayes-Annahme* bedingter Unabhängigkeit der Wörter gegeben die Kategorie gilt:

$$\mathbb{P}(A_1 \mid B_1, \ldots, B_n) \propto \mathbb{P}(A_1) \prod_{j=1}^{n} \mathbb{P}(B_j \mid A_1).$$

Jedes neue Wort multipliziert einen weiteren Likelihood-Faktor hinzu – das Posterior wird iterativ durch eingehende Evidenz aktualisiert. Trifft die Unabhängigkeitsannahme nicht exakt zu, bleibt das Verfahren in der Praxis dennoch leistungsfähig; die Klassifikationsgrenzen bleiben stabil, auch wenn die absoluten Wahrscheinlichkeiten verzerrt sind.

1.8 σ-Algebren

Wir holen die zu Kapitelbeginn angekündigte Präzisierung der Ereignisklasse $\mathcal{A}$ nach. Ist Ω groß (z. B. $\Omega = \mathbb{R}$), kann man nicht jeder Teilmenge eine Wahrscheinlichkeit zuordnen – das führt auf messtheoretische Widersprüche (Vitali-Mengen). Man beschränkt sich auf eine geeignete Klasse von Mengen.

Definition 1.8.1 (σ-Algebra)

Eine Familie $\mathcal{A}$ von Teilmengen von Ω heißt **σ-Algebra**, falls:

1. $\emptyset \in \mathcal{A}$,
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ (Abgeschlossenheit unter Komplementbildung),
3. $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ (Abgeschlossenheit unter abzählbaren Vereinigungen).

Die Mengen in $\mathcal{A}$ heißen **messbar**. Das Paar $(\Omega, \mathcal{A})$ heißt **messbarer Raum**, das Tripel $(\Omega, \mathcal{A}, \mathbb{P})$ heißt **Wahrscheinlichkeitsraum**.

Unmittelbare Folgerungen. Die drei Axiome erzwingen weitere Stabilität:

- $\Omega \in \mathcal{A}$: denn $\Omega = \emptyset^c$ und $\emptyset \in \mathcal{A}$.
- Abzählbare Durchschnitte: Aus $A_i \in \mathcal{A}$ folgt $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$, denn

$$\bigcap_{i=1}^{\infty} A_i \;=\; \left(\bigcup_{i=1}^{\infty} A_i^c\right)^c.$$

- Endliche Operationen: Jede σ-Algebra ist insbesondere abgeschlossen unter endlichen Vereinigungen, Durchschnitten und Differenzen $A \setminus B = A \cap B^c$.

Beispiel 1.8.1 (Kanonische σ-Algebren)

Drei Standardbeispiele für Ω beliebig:

1. **Triviale σ-Algebra**: $\mathcal{A} = \{\emptyset, \Omega\}$. Kleinstmögliche σ-Algebra; kein Ereignis außer dem sicheren und dem unmöglichen ist messbar.
2. **Potenzmenge**: $\mathcal{A} = \mathcal{P}(\Omega)$. Größtmögliche σ-Algebra; jede Teilmenge ist messbar. Für überabzählbares Ω ist $\mathbb{P}$ auf $\mathcal{P}(\Omega)$ im Allgemeinen *nicht* konsistent definierbar.
3. **Von $\mathcal{E}$ erzeugte σ-Algebra**: $\sigma(\mathcal{E})$ ist der Durchschnitt aller σ-Algebren, die $\mathcal{E}$ enthalten – also die *kleinste* σ-Algebra mit $\mathcal{E} \subseteq \sigma(\mathcal{E})$.

Borel-σ-Algebra $\mathcal{B}(\mathbb{R})$. Für $\Omega = \mathbb{R}$ wählt man

$$\mathcal{B}(\mathbb{R}) \;:=\; \sigma\big(\{\text{offene Mengen in } \mathbb{R}\}\big),$$

die kleinste σ-Algebra, die alle offenen Mengen enthält. Äquivalente Erzeuger sind:

$$\mathcal{B}(\mathbb{R}) = \sigma\big(\{(a,b) : a < b\}\big) = \sigma\big(\{(-\infty, b] : b \in \mathbb{R}\}\big) = \sigma\big(\{[a,b] : a \le b\}\big).$$

Alle in der Analysis vorkommenden Mengen – offene, abgeschlossene und kompakte Mengen, Folgengrenzwerte, F_σ- und G_δ-Mengen – sind Borel-messbar. Nicht Borel-messbare Mengen existieren zwar (Vitali 1905), sind aber nicht explizit konstruierbar.

Warum nicht einfach $\mathcal{P}(\mathbb{R})$? Auf $\mathbb{R}$ existiert keine Funktion $\mathbb{P}\colon \mathcal{P}(\mathbb{R}) \to [0,1]$, die gleichzeitig

- σ-additiv,
- translationsinvariant ($\mathbb{P}(A + t) = \mathbb{P}(A)$),
- normiert ($\mathbb{P}([0,1]) = 1$)

erfüllt (Vitali-Konstruktion mit Auswahlaxiom). Die Einschränkung auf $\mathcal{B}(\mathbb{R})$ umgeht diesen Widerspruch: Das Lebesgue-Maß erfüllt alle drei Bedingungen auf $\mathcal{B}(\mathbb{R})$.

Übertragung auf $\mathbb{R}^n$. Für multivariate Zufallsvariablen verwendet man

$$\mathcal{B}(\mathbb{R}^n) \;:=\; \sigma\big(\{U \subseteq \mathbb{R}^n : U \text{ offen}\}\big) \;=\; \mathcal{B}(\mathbb{R})^{\otimes n},$$

wobei $\otimes$ das Produkt-σ-Algebra-Symbol bezeichnet. Die offenen Rechtecke $(a_1, b_1) \times \cdots \times (a_n, b_n)$ bilden *gemeinsam* einen Erzeuger: Ihre Gesamtheit erzeugt $\mathcal{B}(\mathbb{R}^n)$, ein einzelnes Rechteck dagegen nur die vier Mengen $\emptyset$, es selbst, sein Komplement und $\mathbb{R}^n$.

Vom messbaren Raum zum Maß

Eine σ-Algebra legt fest, *welche* Mengen man messen darf – noch nicht, *wie groß* sie sind. Diese zweite Zutat ist das Maß. Wir haben mit $\mathbb{P}$ bereits eines kennengelernt, allerdings in normierter Form. Die allgemeine Fassung brauchen wir später an mehreren Stellen: beim Carathéodoryschen Erweiterungssatz in Abschnitt 2.12 und beim Satz von Radon-Nikodým, der die Existenz des bedingten Erwartungswerts trägt.

Definition 1.8.2 (σ-Additivität)

Sei $\mathcal{A}$ eine σ-Algebra über Ω. Eine Mengenfunktion $\mu\colon \mathcal{A} \to [0, \infty]$ heißt **σ-additiv** (gleichbedeutend: *abzählbar additiv*), wenn für jede Folge paarweise disjunkter Mengen $A_1, A_2, \ldots \in \mathcal{A}$ gilt

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i).$$

Fordert man die Gleichung nur für endlich viele disjunkte Mengen, heißt μ **endlich additiv**.

Bemerkung 1.8.1 (Warum die Reihenfolge keine Rolle spielen darf)

Die linke Seite kennt keine Reihenfolge – eine Vereinigung ändert sich nicht, wenn man die A_i anders aufzählt. Die rechte Seite ist eine Reihe, und Reihen sind sehr wohl reihenfolgeabhängig. Damit die Definition überhaupt widerspruchsfrei ist, muss die Summe also umordnungsstabil sein.
Für $\mu \geq 0$ bekommt man das geschenkt: Eine Reihe mit nichtnegativen Gliedern hat in $[0, \infty]$ stets denselben Wert, gleichgültig in welcher Reihenfolge man summiert. Sobald negative Werte zugelassen sind – beim signierten Maß in Definition 3.5.2 –, gilt das nicht mehr, und man muss *absolute* Konvergenz zusätzlich fordern. Der Riemannsche Umordnungssatz zeigt, warum: Eine nur bedingt konvergente Reihe lässt sich durch Umordnen auf jeden beliebigen Wert bringen.

Definition 1.8.3 (Maß und Maßraum)

Sei $(\Omega, \mathcal{A})$ ein messbarer Raum. Eine Funktion

$$\mu\colon \mathcal{A} \longrightarrow [0, \infty]$$

heißt **Maß**, falls

1. $\mu(\emptyset) = 0$,
2. μ ist σ-additiv im Sinne von Definition 1.8.2.

Das Tripel $(\Omega, \mathcal{A}, \mu)$ heißt **Maßraum**.
Ein Maß heißt **endlich**, falls $\mu(\Omega) < \infty$, und **σ-endlich**, falls es Mengen $\Omega_1, \Omega_2, \ldots \in \mathcal{A}$ gibt mit $\Omega = \bigcup_i \Omega_i$ und $\mu(\Omega_i) < \infty$ für alle i.

Bemerkung 1.8.2 (Wahrscheinlichkeit ist der normierte Spezialfall)

Der Vergleich mit Definition 1.3.1 lohnt sich, weil er zeigt, wie wenig neu ist: Nichtnegativität und σ-Additivität stehen dort wörtlich genauso. Ein **Wahrscheinlichkeitsmaß ist genau ein Maß mit** $\mu(\Omega) = 1$. Umgekehrt ist $\mu(\emptyset) = 0$ dort eine *Folgerung* aus der Normierung, hier dagegen eine *Forderung* – ohne $\mu(\Omega) = 1$ lässt sie sich nicht mehr herleiten.
Der substanzielle Unterschied ist der Wertebereich $[0, \infty]$: Ein Maß darf unendlich sein. Das Lebesgue-Maß λ misst $\lambda(\mathbb{R}) = \infty$, ist aber σ-endlich wegen $\mathbb{R} = \bigcup_n [-n, n]$ – genau dafür ist dieser Begriff da. Weitere Standardbeispiele sind das **Zählmaß** $\mu(A) = |A|$ und das **Dirac-Maß** $\delta_{\omega_0}(A) = I_{\{\omega_0 \in A\}}$.
Aus Satz 1.3.2 überträgt sich alles, dessen Beweis ohne die Normierung auskommt: die Monotonie $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$ und die endliche Additivität. Die Komplementregel $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ dagegen nicht – für $\mu(\Omega) = \infty$ wäre sie sinnlos.

KAPITEL 2

Zufallsvariablen

2.1 Einführung

In der Wahrscheinlichkeitstheorie arbeiten wir mit einem abstrakten Ergebnisraum Ω, dessen Elemente nicht einmal Zahlen sein müssen (Münzwurf: $\Omega = \{K, Z\}$). In der Statistik dagegen haben wir konkrete Daten – Zahlen, mit denen wir rechnen wollen. Die **Zufallsvariable** ist das Bindeglied zwischen beiden Welten: Sie übersetzt abstrakte Ergebnisse in reelle Zahlen.

Definition 2.1.1

Eine **Zufallsvariable** ist eine (messbare) Abbildung $X : \Omega \to \mathbb{R}$, die jedem Ergebnis ω eine reelle Zahl $X(\omega)$ zuordnet.

Ab einem gewissen Punkt wird Ω in den meisten Darstellungen kaum noch erwähnt, und man arbeitet direkt mit Zufallsvariablen. Man sollte jedoch im Hinterkopf behalten, dass der Ergebnisraum stets vorhanden ist.

Was ist eine Zufallsvariable – anschaulich?

Eine Zufallsvariable X übersetzt ein zufälliges Ereignis $\omega \in \Omega$ in einen **beobachtbaren Wert**, mit dem wir rechnen können:

$$X : \Omega \to \mathbb{R} \quad \text{(oder in einen anderen messbaren Raum).}$$

Beispiel 2.1.1 (Würfelwurf)

$\Omega = \{1, 2, 3, 4, 5, 6\}$ (hier sind die Elemente schon Zahlen). Zufallsvariable $X(\omega) = \omega$ – X ist einfach die Augenzahl. Beim 100-fachen Würfeln erhalten wir eine Liste von beobachteten Werten $x_1, \ldots, x_{100}$ (z. B. 3, 5, 1, …). Das sind **Realisierungen** von X.

Beispiel 2.1.2 (Münzwurf)

$\Omega = \{\text{Kopf}, \text{Zahl}\}$, Zufallsvariable X: Kopf $\mapsto 1$, Zahl $\mapsto 0$. Jetzt können wir mit Zahlen rechnen (Erwartungswert, Varianz usw.). Daten: eine Sequenz von 0en und 1en.

Beispiel 2.1.3 (Körpergröße in einer Population)

Das „Experiment“ ist: *wähle zufällig eine Person aus*. Ω ist extrem komplex (alle genetischen und umweltbedingten Faktoren). Zufallsvariable $X(\omega)$ = Körpergröße dieser Person in cm. Wir beobachten nur die Werte von X (z. B. 171, 168, 182,...), den zugrunde liegenden Ergebnisraum Ω sehen wir nie direkt.

Warum ist die Zufallsvariable so wichtig für Statistik und Data Mining?

- Sie ermöglicht es, **Wahrscheinlichkeiten auf die Daten zu übertragen**: $\mathbb{P}(X \leq 170)$ statt $\mathbb{P}(\text{Ereignis „Person} \leq 170\,\text{cm“})$.
- Alle statistischen Konzepte (Erwartungswert, Varianz, Verteilung, Konfidenzintervalle, Hypothesentests, Regression, Clustering, ...) sind auf Zufallsvariablen definiert.
- Unsere Daten sind nichts anderes als **beobachtete Realisierungen** (Samples) einer oder mehrerer Zufallsvariablen – meist als unabhängig und identisch verteilt (iid) modelliert.

Zusammengefasst: Ergebnisraum und Ereignisse liefern das theoretische Fundament der Wahrscheinlichkeit. Die **Zufallsvariable** ist die Brücke, die dieses Fundament mit den tatsächlichen Daten verbindet, mit denen Statistik und Data Mining arbeiten.

Bemerkung 2.1.1 (Messbarkeit – technische Voraussetzung)

Streng genommen muss eine Zufallsvariable **messbar** sein: für jede Borel-Menge $B \subset \mathbb{R}$ muss $X^{-1}(B) = \{\omega : X(\omega) \in B\}$ ein Ereignis sein, also in der σ-Algebra $\mathcal{A}$ liegen (vgl. §1.8). In der Praxis ist diese Bedingung für alle in der Statistik vorkommenden Abbildungen erfüllt.

Beispiel 2.1.4

Zehnmaliger Münzwurf. Sei $X(\omega)$ die Anzahl der Köpfe. Für $\omega = KKZKKZKKZZ$ ist $X(\omega) = 6$.

Beispiel 2.1.5

Sei $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ die Einheitskreisscheibe und $\omega = (x, y)$ ein zufälliger Punkt. Dann sind $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$ und $W(\omega) = x^2 + y^2$ allesamt Zufallsvariablen.

Für eine Zufallsvariable X und eine Borel-Menge $A \subseteq \mathbb{R}$ definieren wir

$$\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}), \quad \mathbb{P}(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\}).$$

Beachte: X ist die Zufallsvariable, x ein konkreter Wert.

Beispiel 2.1.6

Zweimaliger Münzwurf, X = Anzahl Köpfe:

ω	$\mathbb{P}(\{\omega\})$	$X(\omega)$
ZZ	1/4	0
ZK	1/4	1
KZ	1/4	1
KK	1/4	2

x	$\mathbb{P}(X = x)$
0	1/4
1	1/2
2	1/4

2.2 Verteilungsfunktionen und Wahrscheinlichkeitsfunktionen

Definition 2.2.1

Die **kumulative Verteilungsfunktion** (cdf) ist $F_X(x) = \mathbb{P}(X \leq x)$.

Die cdf enthält alle probabilistischen Informationen über X. Oft schreiben wir kurz F statt F_X.

Beispiel 2.2.1

Zweimaliger fairer Münzwurf, X = Anzahl Köpfe:

$$F_X(x) = \begin{cases} 0 & x < 0, \\ 1/4 & 0 \leq x < 1, \\ 3/4 & 1 \leq x < 2, \\ 1 & x \geq 2. \end{cases}$$

Die Funktion ist rechtsstetig, monoton nichtfallend und für alle $x \in \mathbb{R}$ definiert.

Satz 2.2.2

Haben X die cdf F und Y die cdf G mit $F(x) = G(x)$ für alle x, so ist $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ für alle Borel-Mengen A.

Beweis. Für $A = (-\infty, x]$ ist $\mathbb{P}(X \in A) = F(x) = G(x) = \mathbb{P}(Y \in A)$. Da solche Intervalle die Borel-σ-Algebra erzeugen, folgt die Aussage durch Standardargumente der Maßtheorie. ∎

Satz 2.2.3 (Charakterisierung von Verteilungsfunktionen)

Eine Funktion $F : \mathbb{R} \to [0, 1]$ ist genau dann eine cdf, wenn sie

1. **monoton nichtfallend** ist: $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$,

2. **normiert** ist: $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$,
3. **rechtsstetig** ist: $F(x) = \lim_{y\downarrow x} F(y)$.

Beweis. ($\Rightarrow$) Sei $F(x) = \mathbb{P}((-\infty, x])$. Monotonie folgt aus $(-\infty, x_1] \subset (-\infty, x_2]$. Normierung aus $\mathbb{P}(\mathbb{R}) = 1$, $\mathbb{P}(\emptyset) = 0$. Rechtsstetigkeit: Für $x_n \downarrow x$ ist $(-\infty, x_n] \downarrow (-\infty, x]$; Stetigkeit von $\mathbb{P}$ liefert das Resultat.

($\Leftarrow$) Definiere $\mu((a, b]) := F(b) - F(a)$. Die Rechtsstetigkeit garantiert σ-Additivität. Der Erweiterungssatz von Carathéodory liefert ein eindeutiges Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $\mathcal{B}(\mathbb{R})$ mit $\mathbb{P}((-\infty, x]) = F(x)$. ■

Diskrete Zufallsvariablen

Definition 2.2.4

X ist **diskret**, falls sie abzählbar viele Werte $\{x_1, x_2, \ldots\}$ annimmt. Die **Wahrscheinlichkeitsfunktion** (pmf) ist $f_X(x) = \mathbb{P}(X = x)$.

Lemma 2.2.5

Für eine diskrete Zufallsvariable X mit Wertebereich $\{x_1, x_2, \ldots\}$ gilt:

1. $\sum_i f_X(x_i) = 1$,
2. $F_X(x) = \mathbb{P}(X \le x) = \sum_{x_i \le x} f_X(x_i)$.

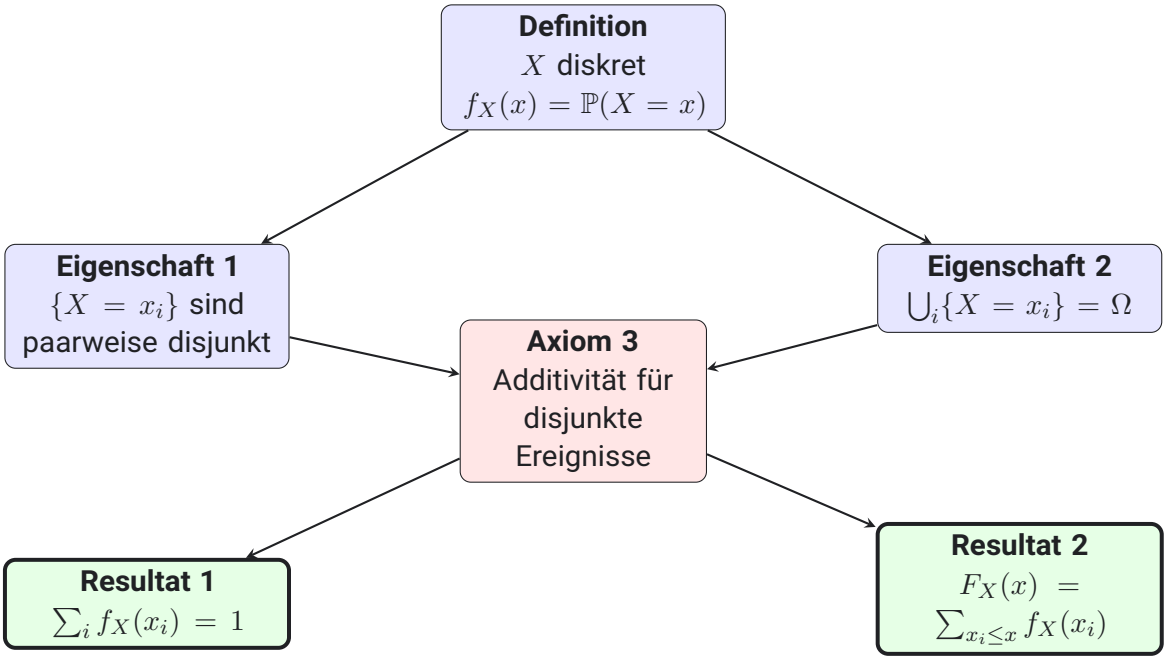
Beweis. (1) Die Ereignisse $\{X = x_i\}$ für $i = 1, 2, \ldots$ bilden eine Partition von Ω (paarweise disjunkt, Vereinigung $= \Omega$). Mit Axiom 3:

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} \{X = x_i\}\right) = \sum_{i=1}^{\infty} \mathbb{P}(X = x_i) = \sum_{i=1}^{\infty} f_X(x_i).$$

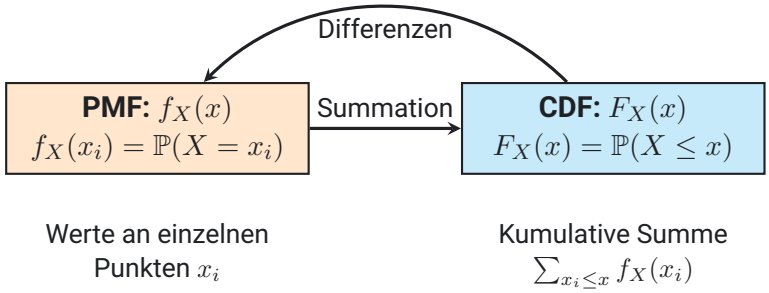
(2) Das Ereignis $\{X \le x\}$ ist die disjunkte Vereinigung aller $\{X = x_i\}$ mit $x_i \le x$:

$$F_X(x) = \mathbb{P}\left(\bigcup_{x_i \le x} \{X = x_i\}\right) = \sum_{x_i \le x} f_X(x_i).$$ ■

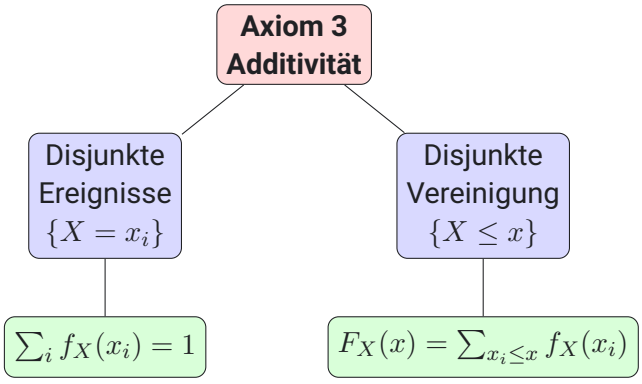
Visuelle Struktur der Beweisführung



Beziehung zwischen pmf und cdf



Axiomatische Struktur



Beispiel 2.2.2 (Zweimaliger fairer Münzwurf, X = Anzahl Köpfe)

Wertebereich: $\{0, 1, 2\}$. **PMF:**

x	0	1	2
$f_X(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

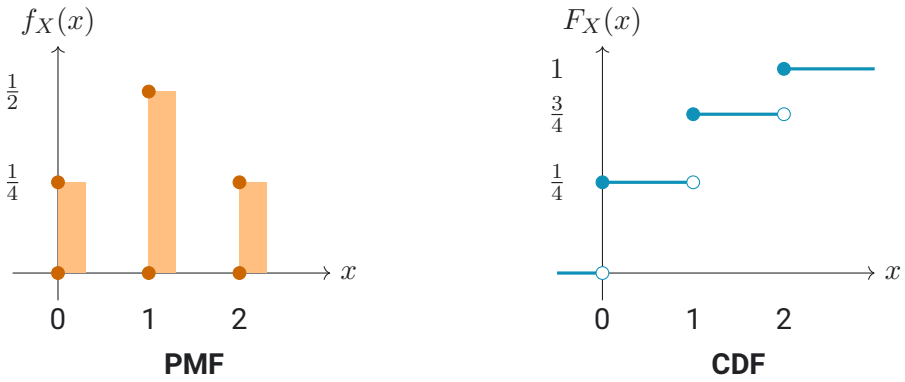
Verifikation von Lemma (1):

$$\sum_i f_X(x_i) = \tfrac{1}{4} + \tfrac{1}{2} + \tfrac{1}{4} = 1. \quad \checkmark$$

CDF aus Lemma (2):

$$\begin{aligned} F_X(0) &= f_X(0) = \tfrac{1}{4}, \\ F_X(1) &= f_X(0) + f_X(1) = \tfrac{1}{4} + \tfrac{1}{2} = \tfrac{3}{4}, \\ F_X(2) &= f_X(0) + f_X(1) + f_X(2) = 1. \end{aligned}$$

Visualisierung: PMF und CDF



Stetige Zufallsvariablen

Definition 2.2.6

X ist **stetig**, falls es eine Funktion $f_X \geq 0$ mit $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ gibt, sodass

$$\mathbb{P}(X \in A) = \int_A f_X(x)\,dx \quad \text{für jede Borel-Menge } A \subseteq \mathbb{R}.$$

f_X heißt **Dichte** (pdf).

Für stetige Zufallsvariablen gilt $\mathbb{P}(X = x) = 0$ für alle x und

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt, \quad f_X(x) = F_X'(x) \text{ (wo differenzierbar).}$$

Beispiel 2.2.3

$X \sim$ Uniform$(0,1)$ hat $f_X(x) = 1$ für $0 < x < 1$ und $F_X(x) = x$ für $0 \le x \le 1$.

Beispiel 2.2.4

Sei $f_X(x) = 3x^2$ für $0 < x < 1$. Dann $F_X(x) = x^3$ für $0 \le x \le 1$, und

$$\mathbb{P}(X \in [0{,}1,\ 0{,}5]) = \int_{0{,}1}^{0{,}5} 3x^2\,dx = 0{,}5^3 - 0{,}1^3 = 0{,}124.$$

Lemma 2.2.7

Sei F die cdf von X. Dann: (i) $\mathbb{P}(X = x) = F(x) - \lim_{y \uparrow x} F(y)$, (ii) $\mathbb{P}(x < X \le y) = F(y) - F(x)$, (iii) $\mathbb{P}(X > x) = 1 - F(x)$, (iv) ist X stetig, so $\mathbb{P}(X = x) = 0$ für alle x.

Beweis. **(i)** Da F monoton nichtfallend ist, existiert der linksseitige Grenzwert und stimmt für jede streng wachsende Folge $y_n \uparrow x$ mit $\lim_n F(y_n)$ überein. Für eine solche Folge gilt

$$\{X \le y_1\} \subset \{X \le y_2\} \subset \cdots, \qquad \bigcup_{n=1}^{\infty} \{X \le y_n\} = \{X < x\},$$

denn ω liegt genau dann in einem der Ereignisse, wenn $X(\omega) \le y_n < x$ für ein n gilt, und umgekehrt wird jedes $X(\omega) < x$ von einem y_n überholt. Die Stetigkeit von Wahrscheinlichkeiten liefert $\lim_{y \uparrow x} F(y) = \mathbb{P}(X < x)$. Wegen der disjunkten Zerlegung $\{X \le x\} = \{X < x\} \cup \{X = x\}$ folgt

$$\mathbb{P}(X = x) = \mathbb{P}(X \le x) - \mathbb{P}(X < x) = F(x) - \lim_{y \uparrow x} F(y).$$

(ii) Für $x < y$ ist $\{X \le x\} \subset \{X \le y\}$, und $\{X \le y\} = \{X \le x\} \cup \{x < X \le y\}$ ist eine disjunkte Zerlegung. Additivität liefert $F(y) = F(x) + \mathbb{P}(x < X \le y)$.

(iii) $\{X > x\}$ ist das Komplement von $\{X \le x\}$, also $\mathbb{P}(X > x) = 1 - F(x)$.

(iv) Ist X stetig mit Dichte f_X, so gilt nach Definition $\mathbb{P}(X \in A) = \int_A f_X(t)\,dt$ für jede Borel-Menge A. Für die einpunktige Menge $A = \{x\}$ verschwindet dieses Integral, also $\mathbb{P}(X = x) = 0$. Nach (i) bedeutet das zugleich, dass F in jedem Punkt linksstetig und wegen der Rechtsstetigkeit damit stetig ist. ∎

Definition 2.2.8

Die **Quantilfunktion** ist $F^{-1}(q) = \inf\{x : F(x) \ge q\}$ für $q \in [0,1]$.

Verteilungs- und Quantilfunktion übersetzen zwischen einer gegebenen Verteilung und der Gleichverteilung. Diese Beziehung ist insbesondere für die Simulation von Zufallsvariablen grundlegend.

Checkerbox 2.2.1 (Wahrscheinlichkeitsintegraltransformation)

(a) Sei X eine Zufallsvariable mit *stetiger* Verteilungsfunktion F. Zeige, dass $U = F(X)$ auf $(0,1)$ gleichverteilt ist.

(b) Sei umgekehrt $U \sim \mathrm{Uniform}(0,1)$ und $Y = F^{-1}(U)$. Zeige, dass Y die Verteilungsfunktion F besitzt; hier ist keine Stetigkeit von F erforderlich.

(c) Erkläre anhand eines Gegenbeispiels, weshalb die Aussage aus (a) ohne Stetigkeitsannahme im Allgemeinen falsch ist.

Beweis. **(a)** Sei $u \in (0,1)$. Da F stetig und monoton nichtfallend ist und an den Rändern gegen 0 bzw. 1 konvergiert, nimmt F den Wert u an. Setze

$$x_u = \sup\{x \in \mathbb{R} : F(x) \le u\}.$$

Aus Stetigkeit und Monotonie folgt $F(x_u) = u$. Nach der Definition des Supremums gilt auch dann, wenn F auf einem Intervall konstant gleich u ist,

$$\{F(X) \le u\} = \{X \le x_u\}.$$

Folglich

$$\mathbb{P}(U \le u) = \mathbb{P}(F(X) \le u) = \mathbb{P}(X \le x_u) = F(x_u) = u.$$

Dies ist die Verteilungsfunktion von $\mathrm{Uniform}(0,1)$.
(b) Nach der Definition der Quantilfunktion gilt für $u \in (0,1)$ und $x \in \mathbb{R}$ die Äquivalenz

$$F^{-1}(u) \le x \quad \Longleftrightarrow \quad u \le F(x).$$

Daher

$$\mathbb{P}(Y \le x) = \mathbb{P}(F^{-1}(U) \le x) = \mathbb{P}(U \le F(x)) = F(x).$$

Somit besitzt Y die Verteilungsfunktion F.
(c) Ohne Stetigkeit kann die Behauptung aus (a) scheitern. Ist etwa $X \sim \delta_a$, so gilt $F(X) = F(a) = 1$ fast sicher; die transformierte Variable ist dann nicht gleichverteilt. ■

2.3 Wichtige diskrete Verteilungen

Diskrete Gleichverteilung

Definition 2.3.1

Eine Zufallsvariable X ist auf den k verschiedenen Werten $\{x_1, \dots, x_k\}$ **diskret gleichverteilt**, wenn

$$\mathbb{P}(X = x_i) = \frac{1}{k}, \qquad i = 1, \dots, k.$$

Ein fairer sechsseitiger Würfel liefert das Standardbeispiel: $X \in \{1,2,3,4,5,6\}$ und $\mathbb{P}(X = i) = 1/6$ für jedes i.

Punktmasse (Dirac-Verteilung)

Definition 2.3.2

Eine Zufallsvariable X hat eine **Punktmasse-Verteilung** (auch **Dirac-Verteilung**), geschrieben $X \sim \delta_a$, falls

$$\mathbb{P}(X = a) = 1.$$

Die gesamte Wahrscheinlichkeitsmasse ist auf den einzelnen Punkt a konzentriert.

PMF:

$$f_X(x) = \begin{cases} 1 & \text{falls } x = a, \\ 0 & \text{sonst.} \end{cases}$$

CDF:

$$F_X(x) = \begin{cases} 0 & \text{falls } x < a, \\ 1 & \text{falls } x \geq a. \end{cases}$$

Beispiel 2.3.1

Eine Zufallsvariable X, die den Wert 5 mit Sicherheit annimmt, also $X \sim \delta_5$:

- die „zufällige“ Wahl einer festen Zahl,
- eine Konstante als Zufallsvariable betrachtet,
- $\mathbb{P}(X = 5) = 1$, $\mathbb{P}(X \neq 5) = 0$.

Bemerkung 2.3.1 (Shannon-Entropie – Definition zur Einordnung)

Die **Shannon-Entropie** einer diskreten Zufallsvariable X mit pmf f ist

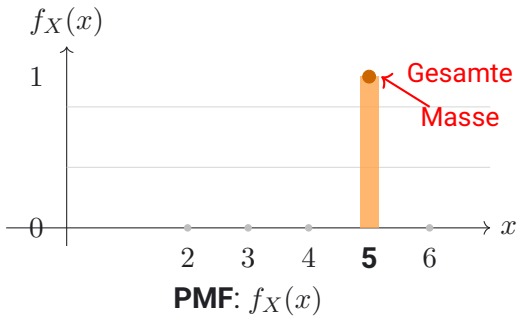
$$H(X) \;=\; -\sum_i f(x_i)\,\log f(x_i),$$

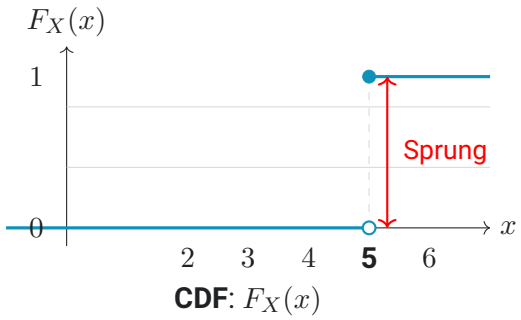
mit der Konvention $0 \log 0 := 0$. Sie misst die mittlere Unsicherheit (bzw. den mittleren Informationsgehalt pro Beobachtung) und ist genau dann minimal, $H(X) = 0$, wenn die gesamte Masse auf einem einzigen Punkt sitzt – also $X \sim \delta_a$. Das Maximum $H(X) = \log k$ wird von der diskreten Gleichverteilung auf k Punkten angenommen. Die Wahl der Logarithmus-Basis (2: Bits, e: Nats, 10: Dits) ändert nur die Einheit.

Eigenschaften

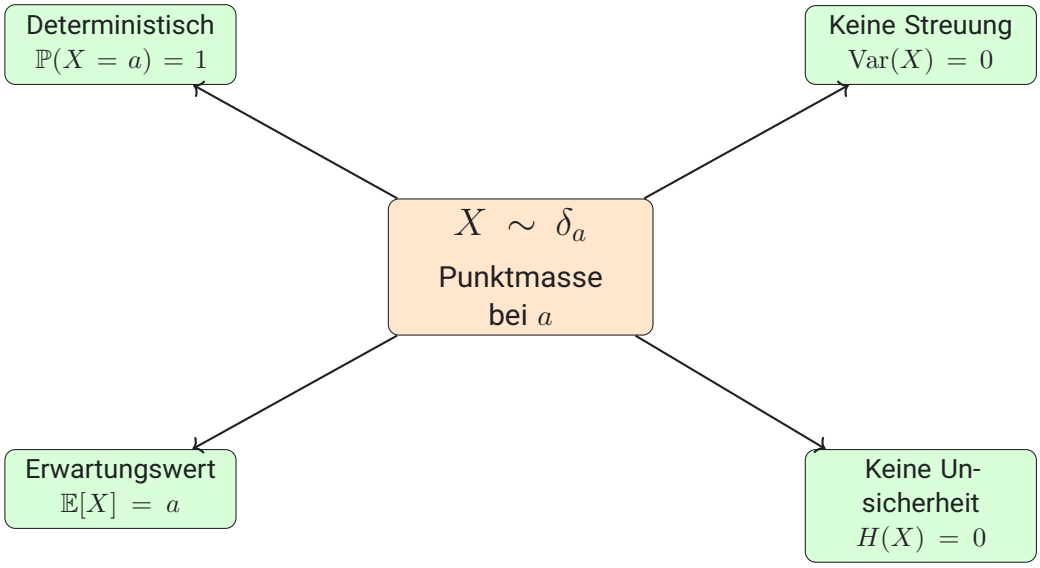
- **Erwartungswert:** $\mathbb{E}[X] = a$
- **Varianz:** $\mathrm{Var}(X) = 0$ (keine Streuung)
- **Entropie:** $H(X) = 0$ (keine Unsicherheit – minimaler Wert, vgl. Bemerkung oben)
- **Grenzfall:** deterministische Variable (keine echte Zufälligkeit)

Visualisierung: PMF und CDF für δ_5

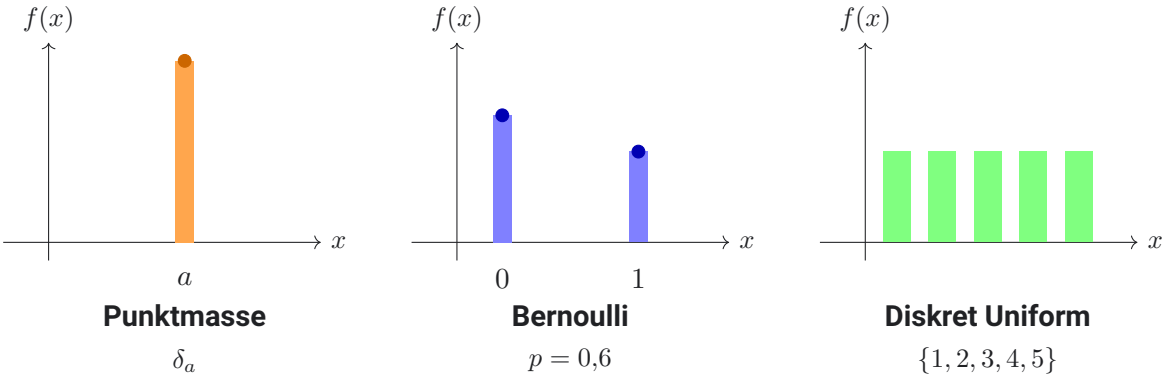




Konzeptuelle Darstellung



Vergleich: Punktmasse vs. andere diskrete Verteilungen



Mathematische Eigenschaften im Detail

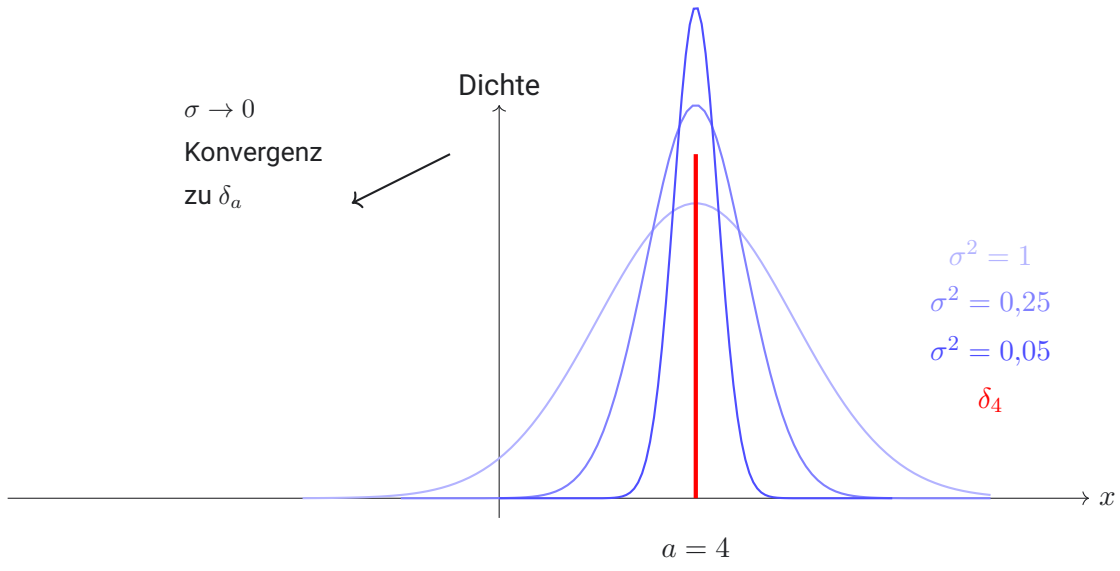
Eigenschaft	Wert für δ_a
Träger (Support)	$\{a\}$
Erwartungswert	$\mathbb{E}[X] = a$
Varianz	$\mathrm{Var}(X) = 0$
Standardabweichung	$\sigma_X = 0$
Schiefe (Skewness)	nicht definiert
Kurtosis	nicht definiert
MGF	$\psi_X(t) = e^{at}$
Charakteristische Funktion	$\varphi_X(t) = e^{iat}$

Anwendungen

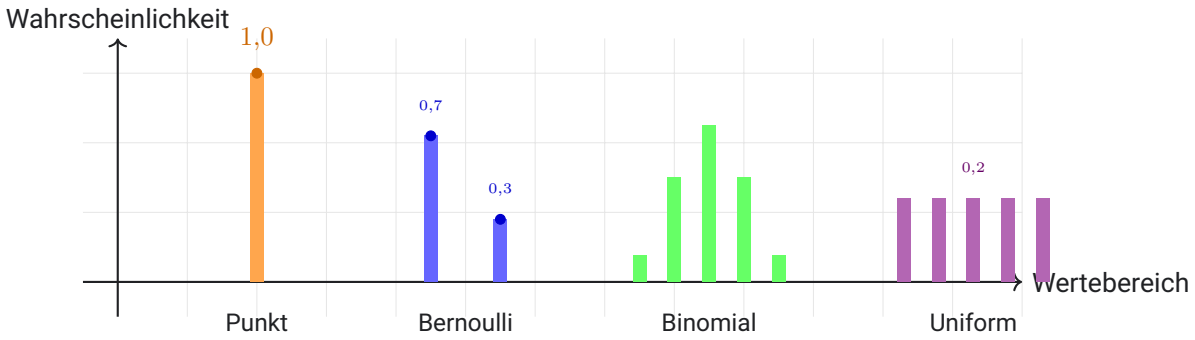
- **Modellierung von Sicherheit:** wenn ein Ereignis mit Sicherheit eintritt
- **Grenzfälle:** Approximation durch sehr konzentrierte Verteilungen
- **Dirac-Delta-Funktion:** stetige Verallgemeinerung in der Analysis
- **Bayes'sche Statistik:** Prior-Verteilung bei exaktem Vorwissen
- **Anfangsbedingungen:** Startpunkt in stochastischen Prozessen

Konvergenz zu Punktmasse

Betrachte eine Folge von Normalverteilungen mit schrumpfender Varianz:



Vergleich: PMF-Darstellung verschiedener Verteilungen



Bernoulli-Verteilung

$X \sim$ Bernoulli(p): $X \in \{0,1\}$ mit $\mathbb{P}(X=1)=p$ und $\mathbb{P}(X=0)=1-p$. Die pmf lautet

$$f(x) = p^x(1-p)^{1-x}, \qquad x \in \{0,1\}.$$

Die Verteilung modelliert einen einzelnen Versuch mit den beiden Ausgängen Erfolg und Misserfolg; ein fairer Münzwurf entspricht $p = 1/2$.

Binomialverteilung

$X \sim$ Binomial(n,p): Summe von n unabhängigen Bernoulli(p)-Variablen.

$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,\dots,n.$$

Beispielsweise ist die Anzahl der Köpfe bei zehn Würfen einer fairen Münze $\mathrm{Binomial}(10, 1/2)$-verteilt.

Geometrische Verteilung

$X \sim$ Geom(p): Anzahl der Versuche bis zum ersten Erfolg. $\mathbb{P}(X=k) = p(1-p)^{k-1}$, $k = 1,2,\dots$ Beim wiederholten Würfeln bis zur ersten Sechs gilt $p = 1/6$.

Poisson-Verteilung

$X \sim$ Poisson(λ): $\mathbb{P}(X=k) = e^{-\lambda}\lambda^k/k!$, $k = 0,1,2,\dots$ Approximiert Binomial(n,p) für großes n, kleines p mit $np \approx \lambda$. So kann X etwa die Zahl der Anrufe in einem Callcenter während einer Stunde modellieren; bei durchschnittlich vier Anrufen setzt man $\lambda = 4$.

Checkerbox 2.3.1 (Poisson-Approximation der Binomialverteilung)

Sei $\lambda > 0$ fest und für jede ganze Zahl $n > \lambda$ sei $X_n \sim \mathrm{Binomial}(n, p_n)$ mit $p_n = \lambda/n$. Zeige für jedes feste $k \in \mathbb{N}_0$:

$$\mathbb{P}(X_n = k) \longrightarrow e^{-\lambda}\frac{\lambda^k}{k!}.$$

Damit wird die Poisson-Approximation für großes n und entsprechend kleines p_n bei $np_n = \lambda$ begründet.

Beweis. Für $n > \max\{k, \lambda\}$ gilt

$$\begin{aligned}\mathbb{P}(X_n = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\ &= \frac{\lambda^k}{k!}\left(\prod_{j=0}^{k-1}\left(1-\frac{j}{n}\right)\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}.\end{aligned}$$

Da k fest ist, konvergiert das endliche Produkt gegen 1. Ferner gelten

$$\left(1-\frac{\lambda}{n}\right)^n \longrightarrow e^{-\lambda} \qquad \text{und} \qquad \left(1-\frac{\lambda}{n}\right)^{-k} \longrightarrow 1.$$

Multiplikation der Grenzwerte liefert

$$\mathbb{P}(X_n = k) \longrightarrow \frac{\lambda^k}{k!}\,e^{-\lambda}.$$

■

2.4 Wichtige stetige Verteilungen

Gleichverteilung

$X \sim \text{Uniform}(a, b)$: $f(x) = \frac{1}{b-a}$ für $a < x < b$.

Normalverteilung

$X \sim N(\mu, \sigma^2)$:

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

$N(0,1)$ ist die Standardnormalverteilung mit cdf Φ.

Exponentialverteilung

$X \sim \text{Exp}(\beta)$: $f(x) = \frac{1}{\beta}e^{-x/\beta}$ für $x > 0$. Gedächtnislos: $\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t)$.

Checkerbox 2.4.1 (Gedächtnislosigkeit der Exponentialverteilung)

Sei $X \sim \text{Exp}(\beta)$ mit $\beta > 0$. Beweise für $s, t \geq 0$:

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t).$$

Beweis. Die Überlebensfunktion der Exponentialverteilung ist für $x \geq 0$

$$\mathbb{P}(X > x) = \int_x^\infty \frac{1}{\beta}e^{-u/\beta}\,du = e^{-x/\beta}.$$

Da $\{X > s+t\} \subseteq \{X > s\}$ und $\mathbb{P}(X > s) = e^{-s/\beta} > 0$, folgt

$$\begin{aligned}\mathbb{P}(X > s+t \mid X > s) &= \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} \\ &= \frac{e^{-(s+t)/\beta}}{e^{-s/\beta}} = e^{-t/\beta} = \mathbb{P}(X > t).\end{aligned}$$

■

Gamma-Verteilung

Für $\alpha > 0$ und $\beta > 0$ gilt $X \sim \text{Gamma}(\alpha, \beta)$: $f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$, $x > 0$, wobei

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,dt$$

die Gamma-Funktion ist.

Beta-Verteilung

$X \sim \text{Beta}(\alpha, \beta)$: $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$, $0 < x < 1$.

t-Verteilung und χ^2-Verteilung

$X \sim t_\nu$ mit $\nu > 0$ Freiheitsgraden besitzt die Dichte

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \qquad x \in \mathbb{R}.$$

Die Dichte der t-Verteilung nimmt in den Randbereichen langsamer ab als die Normaldichte; je kleiner die Zahl ν der Freiheitsgrade ist, desto stärker ist dieser Effekt. Bei normalverteilten Daten entsteht sie insbesondere beim Standardisieren mit geschätzter statt bekannter Streuung und konvergiert für $\nu \to \infty$ gegen $N(0,1)$. Eine Zufallsvariable $X \sim \chi^2_\nu$ ist gamma-verteilt mit Parametern $\nu/2$ und 2: $\chi^2_\nu = \text{Gamma}(\nu/2, 2)$.

2.5 Bivariate Verteilungen

Definition 2.5.1

Die **gemeinsame cdf** ist $F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y)$.
Im diskreten Fall ist die **gemeinsame pmf** $f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$. Im stetigen Fall heißt eine nichtnegative Funktion $f_{X,Y}$ mit

$$\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dx\,dy = 1$$

gemeinsame pdf, wenn für jede Borel-Menge $A \subseteq \mathbb{R}^2$ gilt

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

Beispiel 2.5.1 (Minimum und Maximum zweier Würfel)

Wir werfen zwei faire Würfel und setzen X gleich dem Minimum und Y gleich dem Maximum der beiden Augenzahlen. Dann

$$\mathbb{P}(X = 1, Y = 6) = \mathbb{P}(\{(1,6),(6,1)\}) = \frac{2}{36} = \frac{1}{18}.$$

Beispiel 2.5.2 (Gleichverteilung auf dem Einheitsquadrat)

Ist (X, Y) gleichverteilt auf $[0,1]^2$, so lautet die gemeinsame Dichte

$$f_{X,Y}(x,y) = \begin{cases} 1, & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{sonst.} \end{cases}$$

2.6 Randverteilungen

Definition 2.6.1

Die **Randdichte/-funktion** ergibt sich durch Marginalisierung:
Diskret: $f_X(x) = \sum_y f_{X,Y}(x,y)$. Stetig: $f_X(x) = \int f_{X,Y}(x,y)\,dy$.

Beispiel 2.6.1

Sei $f_{X,Y}(x,y) = \frac{6}{5}(x + y^2)$ für $0 \le x, y \le 1$. Dann $f_X(x) = \int_0^1 \frac{6}{5}(x+y^2)\,dy = \frac{6}{5}(x + 1/3)$ für $0 < x < 1$.

Beispiel 2.6.2 (Marginalisierung einer gemeinsamen pmf)

Eine gemeinsame Wahrscheinlichkeitsfunktion sei durch die folgende Tabelle gegeben:

	$y=0$	$y=1$	$y=2$
$x=0$	$\frac{1}{9}$	$\frac{2}{9}$	0
$x=1$	$\frac{1}{3}$	0	$\frac{1}{9}$
$x=2$	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{18}$

(alle Einträge summieren sich zu 1). Durch Summation über y ergibt sich

$$f_X(0) = \frac{1}{3}, \qquad f_X(1) = \frac{4}{9}, \qquad f_X(2) = \frac{2}{9}.$$

2.7 Unabhängige Zufallsvariablen

Definition 2.7.1

$X \perp Y$ bedeutet, dass für alle Borel-Mengen $A, B \subseteq \mathbb{R}$ gilt: $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$.

Korollar 2.7.2 (Faktorisierung im diskreten Fall)

Seien X und Y diskret. Dann gilt $X \perp Y$ genau dann, wenn

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

für alle Werte x, y gilt.

Beweis. Aus der Unabhängigkeit folgt die Gleichung durch Anwendung der Definition auf die einelementigen Mengen $\{x\}$ und $\{y\}$. Gilt umgekehrt die Gleichung für alle x, y, so erhalten wir für beliebige Borel-Mengen A, B durch Summation

$$\begin{aligned}\mathbb{P}(X \in A, Y \in B) &= \sum_{x \in A} \sum_{y \in B} \mathbb{P}(X = x, Y = y) \\ &= \left(\sum_{x \in A} \mathbb{P}(X = x)\right)\left(\sum_{y \in B} \mathbb{P}(Y = y)\right) \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B).\end{aligned}$$

■

Checkerbox 2.7.1 (Summe unabhängiger Poisson-Variablen)

Seien $X \sim \text{Poisson}(\lambda_1)$ und $Y \sim \text{Poisson}(\lambda_2)$ unabhängig, wobei $\lambda_1, \lambda_2 > 0$. Zeige

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

Beweis. Für $k \in \mathbb{N}_0$ ist $\{X+Y = k\}$ die disjunkte Vereinigung der Ereignisse $\{X = j, Y = k-j\}$ für $j = 0, \ldots, k$. Mit der Unabhängigkeit folgt

$$\begin{aligned}\mathbb{P}(X + Y = k) &= \sum_{j=0}^{k} \mathbb{P}(X = j)\mathbb{P}(Y = k - j) \\ &= \sum_{j=0}^{k} e^{-\lambda_1} \frac{\lambda_1^j}{j!} e^{-\lambda_2} \frac{\lambda_2^{k-j}}{(k-j)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda_1^j \lambda_2^{k-j} \\ &= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^k}{k!}.\end{aligned}$$

Dies ist die pmf einer $\text{Poisson}(\lambda_1 + \lambda_2)$-Verteilung. ■

Satz 2.7.3

Seien X und Y stetige Zufallsvariablen mit gemeinsamer Dichte $f_{X,Y}$. Dann gilt $X \perp Y$ genau dann, wenn $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ für fast alle $(x,y) \in \mathbb{R}^2$.

Beweis. $(\Rightarrow)$ Aus der Unabhängigkeit folgt, dass die gemeinsame Verteilung das Produkt der Randverteilungen ist. Dieses Produktmaß besitzt die Dichte $(x,y) \mapsto f_X(x) f_Y(y)$. Wegen der Eindeutigkeit von Dichten bis auf Nullmengen gilt daher $f_{X,Y} = f_X f_Y$ fast überall.

$(\Leftarrow)$ $\mathbb{P}(X \in A, Y \in B) = \int_A \int_B f_X(x) f_Y(y)\, dy\, dx = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$. ■

Beispiel 2.7.1

Sind $X \sim \mathrm{Uniform}(0,1)$ und $Y \sim \mathrm{Exp}(1)$ unabhängig, so lässt sich ihre gemeinsame Dichte als Produkt schreiben:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) = e^{-y}, \qquad 0 < x < 1,\ y > 0,$$

und sie ist außerhalb dieses Rechtecks gleich 0.

Satz 2.7.4

Sei $f_{X,Y}(x,y) = g(x)h(y)$ fast überall auf einem rechteckigen Träger $\mathcal{X} \times \mathcal{Y}$, wobei g und h nichtnegative Funktionen sind. Dann sind X und Y unabhängig.

Beweis. Setze $c_g = \int_{\mathcal{X}} g(x)\, dx$ und $c_h = \int_{\mathcal{Y}} h(y)\, dy$. Aus der Normierung der gemeinsamen Dichte folgt $c_g c_h = 1$. Daher

$$f_X(x) = c_h g(x), \qquad f_Y(y) = c_g h(y),$$

und somit $f_X(x) f_Y(y) = g(x)h(y) = f_{X,Y}(x,y)$ fast überall. Nach dem vorigen Satz sind X und Y unabhängig. ■

2.8 Bedingte Verteilungen

Definition 2.8.1

Die **bedingte Dichte/pmf** von Y gegeben $X = x$ ist

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad f_X(x) > 0.$$

Beispiel 2.8.1 (Einheitsquadrat)

Für die Gleichverteilung auf $(0,1)^2$ gilt $f_{X,Y}(x,y) = 1$ und $f_X(x) = \int_0^1 1\, dy = 1$. Daher ist

$$f_{Y|X}(y \mid x) = 1, \qquad 0 < y < 1,$$

also $Y \mid X = x \sim \mathrm{Uniform}(0,1)$ für jedes $x \in (0,1)$.

Beispiel 2.8.2

Sei $X \sim \mathrm{Uniform}(0,1)$ und $Y \mid X = x \sim \mathrm{Uniform}(0,x)$. Die gemeinsame Dichte ist $f_{X,Y}(x,y) = 1/x$ für $0 < y < x < 1$. Die Randdichte von Y: $f_Y(y) = \int_y^1 \frac{1}{x}\,dx = -\ln y$ für $0 < y < 1$.

Checkerbox 2.8.1 (Gleichverteilung auf einem Dreieck)

Sei (X,Y) gleichverteilt auf dem Dreieck

$$D = \{(x,y) \in \mathbb{R}^2 : x \geq 0,\ y \geq 0,\ x + y \leq 1\}.$$

Bestimme die Randdichte $f_X(x)$ und die bedingte Dichte $f_{Y|X}(y \mid x)$.

Beweis. Das Dreieck besitzt die Fläche $1/2$. Seine konstante Dichte muss daher auf D den Wert 2 haben:

$$f_{X,Y}(x,y) = 2\, I_D(x,y).$$

Für festes $x \in (0,1)$ reicht y von 0 bis $1-x$. Somit

$$f_X(x) = \int_0^{1-x} 2\,dy = 2(1-x), \qquad 0 < x < 1,$$

und $f_X(x) = 0$ sonst. Für $0 < x < 1$ ist $f_X(x) > 0$, also

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{1-x}, \qquad 0 < y < 1-x,$$

und die bedingte Dichte ist außerhalb dieses Intervalls gleich 0. Damit gilt $Y \mid X = x \sim \mathrm{Uniform}(0, 1-x)$. ∎

2.9 Multivariate Verteilungen und iid-Stichproben

Definition 2.9.1

$X_1, \ldots, X_n$ heißen **unabhängig und identisch verteilt** (iid) mit gemeinsamer Dichte f, wenn sie unabhängig sind und jede Zufallsvariable die Dichte f besitzt. Dann gilt

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i).$$

Definition 2.9.2 (Ordnungsstatistiken)

Die durch Sortieren einer Stichprobe entstehenden Zufallsvariablen

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

heißen **Ordnungsstatistiken**. Insbesondere sind $X_{(1)} = \min_i X_i$ und $X_{(n)} = \max_i X_i$.

Beispiel 2.9.1 (Zwei Gleichverteilungen)

Für zwei unabhängige Zufallsvariablen $X_1, X_2 \sim \mathrm{Uniform}(0,1)$ können bei $0 < x_1 < x_2 < 1$ entweder (X_1, X_2) oder (X_2, X_1) in kleinen Umgebungen von (x_1, x_2) liegen. Beide Zuordnungen haben dieselbe Dichte 1; daher hat $(X_{(1)}, X_{(2)})$ auf dem Dreieck $0 < x_1 < x_2 < 1$ die gemeinsame Dichte 2.

Checkerbox 2.9.1 (Gemeinsame Dichte der Ordnungsstatistiken)

Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit einer stetigen Dichte f. Zeige, dass die gemeinsame Dichte ihrer Ordnungsstatistiken

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i)$$

für $x_1 < \cdots < x_n$ gilt und außerhalb dieses Bereichs gleich 0 ist.
Hinweis: Zähle die $n!$ möglichen Zuordnungen der Zufallsvariablen $X_1, \ldots, X_n$ zu n kleinen, geordneten Intervallen.

Beweis. Seien $x_1 < \cdots < x_n$ fest. Wähle kleine, paarweise disjunkte Intervalle $I_i = [x_i, x_i + h_i]$, die in dieser Reihenfolge angeordnet sind. Das Ereignis

$$\{X_{(i)} \in I_i \text{ für } i = 1, \ldots, n\}$$

tritt genau dann ein, wenn in jedem Intervall I_i genau eine der n Zufallsvariablen $X_1, \ldots, X_n$ liegt. Es gibt $n!$ Möglichkeiten, die Zufallsvariablen den Intervallen zuzuordnen. Diese Ereignisse sind disjunkt, und jede Zuordnung hat wegen der Unabhängigkeit die Wahrscheinlichkeit

$$\prod_{i=1}^{n} \mathbb{P}(X_i \in I_i) = \prod_{i=1}^{n} \int_{I_i} f(t)\, dt;$$

die identische Verteilung macht den Wert unabhängig von der gewählten Zuordnung. Daher

$$\mathbb{P}\big(X_{(i)} \in I_i, i = 1, \ldots, n\big) = n! \prod_{i=1}^{n} \int_{I_i} f(t)\, dt.$$

Division durch $h_1 \cdots h_n$ und der Grenzübergang $h_1, \ldots, h_n \downarrow 0$ liefern $n! \prod_i f(x_i)$, weil f stetig ist. Außerhalb von $x_1 \leq \cdots \leq x_n$ kann der sortierte Vektor keine Werte annehmen. Die Randmenge mit Gleichheiten hat Wahrscheinlichkeit 0, denn für $i \neq j$ gilt $\mathbb{P}(X_i = X_j) = 0$. ∎

2.10 Zwei wichtige multivariate Verteilungen

In diesem Abschnitt betrachten wir die beiden zentralen Standardverteilungen für *multivariate* Modellierung: die **Multinomialverteilung** als diskreten Vektor-Zähler und die **multivariate Normalverteilung** als ihr stetiges Pendant. Die Multinomialverteilung verallgemeinert die Binomialverteilung von zwei auf k Kategorien und beschreibt die Häufigkeiten kategorialer Daten; die multivariate Normalverteilung verallgemeinert die Normalverteilung auf Vektoren und ist die Grundlage praktisch jeder mehrdimensionalen Inferenz.

Multinomialverteilung

Wirft man n-mal einen k-seitigen Würfel mit Klassenwahrscheinlichkeiten $p = (p_1, \dots, p_k)$, $p_i \geq 0$, $\sum_{i=1}^{k} p_i = 1$, und zählt, wie oft jede Seite gefallen ist, erhält man einen ganzzahligen Zählvektor $X = (X_1, \dots, X_k)$. Wir schreiben

$$X \sim \text{Multinomial}(n, p).$$

Per Konstruktion gilt $X_i \in \{0, 1, \dots, n\}$ und

$$X_1 + X_2 + \cdots + X_k \;=\; n,$$

da jeder Versuch in genau einer Klasse landet. Die Massenfunktion ist

$$\mathbb{P}(X_1 = x_1, \dots, X_k = x_k) \;=\; \frac{n!}{x_1!\, x_2!\, \cdots\, x_k!}\, p_1^{x_1}\, p_2^{x_2} \cdots p_k^{x_k} \qquad \text{für } x_i \in \mathbb{N}_0 \text{ mit } \sum_{i=1}^{k} x_i = n,$$

und null außerhalb dieses Trägers. Der Multinomialkoeffizient $\frac{n!}{x_1!\cdots x_k!}$ zählt die Anordnungen der n Würfe, die zum Häufigkeitsvektor $(x_1, \dots, x_k)$ führen.

Bemerkung 2.10.1

Für $k = 2$ ist $X_2 = n - X_1$, und man erhält die Binomialverteilung zurück: $X_1 \sim \text{Binomial}(n, p_1)$.

Die Multinomialverteilung modelliert diskrete Zählvektoren. Wenden wir uns nun ihrem stetigen Gegenstück zu — der Verallgemeinerung der Normalverteilung auf Vektoren.

Multivariate Normalverteilung

Sei $X = (X_1, \dots, X_k)^\top$ ein Zufallsvektor in $\mathbb{R}^k$, $\mu \in \mathbb{R}^k$ ein Erwartungswertvektor und $\Sigma \in \mathbb{R}^{k \times k}$ eine symmetrische, positiv definite Kovarianzmatrix. Wir sagen, X ist **multivariat normalverteilt** mit Parametern μ und Σ, geschrieben

$$X \sim N(\mu, \Sigma),$$

falls X die Dichte

$$f(x) \;=\; \frac{1}{(2\pi)^{k/2}\, |\Sigma|^{1/2}} \exp\!\Big(-\tfrac{1}{2}\,(x-\mu)^\top \Sigma^{-1} (x-\mu)\Big), \qquad x \in \mathbb{R}^k,$$

besitzt, wobei $|\Sigma|$ die Determinante von Σ bezeichnet.

Bemerkung 2.10.2

Es gilt $\mathbb{E}[X] = \mu$ und $\mathrm{Var}(X) = \Sigma$, d. h. $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Für $k = 1$ reduziert sich die Dichte auf die bekannte univariate Normaldichte mit Varianz $\sigma^2 = \Sigma$. Ist $\Sigma = I$ und $\mu = 0$, so heißt X *Standard-multivariat-normalverteilt*; in diesem Fall sind $X_1, \dots, X_k$ unabhängig und jeweils $N(0,1)$-verteilt.

Für die folgende Konstruktion benötigen wir die mehrdimensionale Version der Dichtetransformation.

Satz 2.10.1 (Dichtetransformation, mehrdimensional)

Seien $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ offen und sei $g : \mathcal{X} \to \mathcal{Y}$ ein C^1-Diffeomorphismus mit Umkehrung $h = g^{-1}$. Der Zufallsvektor X besitze die Dichte f_X, die außerhalb von $\mathcal{X}$ verschwindet. Dann besitzt $Y = g(X)$ die Dichte

$$f_Y(y) = f_X(h(y)) \, |\det Dh(y)|, \qquad y \in \mathcal{Y},$$

und außerhalb von $\mathcal{Y}$ ist $f_Y(y) = 0$.

Beweis. Für jede Borel-Menge $B \subseteq \mathcal{Y}$ gilt nach dem Transformationssatz für mehrdimensionale Integrale

$$\begin{aligned} \mathbb{P}(Y \in B) &= \mathbb{P}(X \in h(B)) = \int_{h(B)} f_X(x)\, dx \\ &= \int_B f_X(h(y)) \, |\det Dh(y)| \, dy. \end{aligned}$$

Damit ist der Integrand die Dichte von Y. ∎

Eine elegante Möglichkeit, multivariate Normalverteilungen zu konstruieren oder zu simulieren, liefert der folgende Satz.

Satz 2.10.2

Ist $Z \sim N(0, I)$ in $\mathbb{R}^k$ und $X = \mu + \Sigma^{1/2} Z$, so ist $X \sim N(\mu, \Sigma)$.

Beweis. Setze $A = \Sigma^{1/2}$ (die symmetrische Wurzel der positiv definiten Matrix Σ). Die Umkehrung lautet $Z = A^{-1}(X - \mu)$, mit Jacobi-Determinante $|J| = |\Sigma|^{-1/2}$. Mit Satz 2.10.1 folgt

$$f_X(x) \;=\; f_Z\big(A^{-1}(x-\mu)\big) \cdot |\Sigma|^{-1/2} \;=\; \frac{1}{(2\pi)^{k/2}\,|\Sigma|^{1/2}} \exp\!\Big(-\tfrac{1}{2}\,(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\Big)\,. \qquad ∎$$

2.11 Transformationen von Zufallsvariablen

Sei $Y = g(X)$.

Diskreter Fall: $f_Y(y) = \sum_{x:\, g(x)=y} f_X(x)$.

Stetiger Fall: Ist X stetig und g streng monoton sowie differenzierbar, so liefert der folgende Satz eine geschlossene Formel für die Dichte von $Y = g(X)$.

Satz 2.11.1 (Dichtetransformation, eindimensional)

Sei X eine stetige Zufallsvariable mit Dichte f_X und Träger $\mathcal{X} \subseteq \mathbb{R}$. Ist $g : \mathcal{X} \to \mathcal{Y} := g(\mathcal{X})$ streng monoton und stetig differenzierbar mit $g'(x) \neq 0$ für alle $x \in \mathcal{X}$, so besitzt $Y = g(X)$ die Dichte

$$f_Y(y) \;=\; f_X\big(g^{-1}(y)\big)\,\left|\frac{d}{dy}\,g^{-1}(y)\right|, \qquad y \in \mathcal{Y},$$

und $f_Y(y) = 0$ außerhalb von $\mathcal{Y}$.

Beweis. Da g streng monoton, stetig differenzierbar und $g'(x) \neq 0$ ist, existiert die Umkehrfunktion $g^{-1} : \mathcal{Y} \to \mathcal{X}$ und ist nach dem Satz über die Umkehrfunktion ebenfalls stetig differenzierbar. Strategie: Wir berechnen zunächst die Verteilungsfunktion $F_Y(y) = \mathbb{P}(Y \leq y)$ und differenzieren anschließend nach y, um die Dichte zu erhalten. Wir unterscheiden zwei Fälle.

Fall 1: g streng monoton wachsend. Dann ist auch g^{-1} streng monoton wachsend, und es gilt die Äquivalenz

$$g(x) \leq y \iff x \leq g^{-1}(y) \qquad \text{für alle } x \in \mathcal{X},\ y \in \mathcal{Y}.$$

Damit

$$F_Y(y) \;=\; \mathbb{P}(g(X) \leq y) \;=\; \mathbb{P}\big(X \leq g^{-1}(y)\big) \;=\; F_X\big(g^{-1}(y)\big)\,.$$

Differenzieren nach y mit der Kettenregel und unter Verwendung von $F_X'(x) = f_X(x)$ liefert

$$f_Y(y) \;=\; \frac{d}{dy}\,F_X\big(g^{-1}(y)\big) \;=\; f_X\big(g^{-1}(y)\big)\cdot\frac{d}{dy}\,g^{-1}(y).$$

Da g^{-1} wachsend ist, gilt $\frac{d}{dy}\,g^{-1}(y) > 0$. Damit stimmt der Faktor mit seinem Betrag überein, und es folgt

$$f_Y(y) \;=\; f_X\big(g^{-1}(y)\big)\,\left|\frac{d}{dy}\,g^{-1}(y)\right|.$$

Fall 2: g streng monoton fallend. Dann ist g^{-1} ebenfalls streng monoton fallend, und die Richtung der Ungleichung dreht sich um:

$$g(x) \leq y \iff x \geq g^{-1}(y).$$

Da X stetig ist, gilt $\mathbb{P}\big(X = g^{-1}(y)\big) = 0$, sodass

$$F_Y(y) \;=\; \mathbb{P}\big(X \geq g^{-1}(y)\big) \;=\; 1 - F_X\big(g^{-1}(y)\big)\,.$$

Differenzieren ergibt

$$f_Y(y) \;=\; -\,f_X\big(g^{-1}(y)\big)\cdot\frac{d}{dy}\,g^{-1}(y).$$

Weil g^{-1} fallend ist, gilt $\frac{d}{dy}\,g^{-1}(y) < 0$; das Minuszeichen vor dem Term macht den Gesamtausdruck nicht-negativ. Mit $|a| = -a$ für $a < 0$ folgt

$$f_Y(y) \;=\; f_X\big(g^{-1}(y)\big)\,\left|\frac{d}{dy}\,g^{-1}(y)\right|.$$

In beiden Fällen ergibt sich dieselbe Formel. Werte außerhalb von $\mathcal{Y}$ kann $Y = g(X)$ nicht annehmen, sodass dort $f_Y(y) = 0$ gilt. ■

Bemerkung 2.11.1

Mit der Identität $\frac{d}{dy}\, g^{-1}(y) = \frac{1}{g'(g^{-1}(y))}$ aus dem Satz über die Umkehrfunktion lässt sich die Formel auch in der äquivalenten Gestalt

$$f_Y(y) \;=\; \frac{f_X\big(g^{-1}(y)\big)}{|g'(g^{-1}(y))|}$$

schreiben. Beide Versionen sind in der Literatur gebräuchlich.

Beispiel 2.11.1

$\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = 1/4$, $\mathbb{P}(X = 0) = 1/2$, $Y = X^2$. Dann $f_Y(0) = 1/2$, $f_Y(1) = 1/2$.

Checkerbox 2.11.1 (Logarithmus einer Gleichverteilung)

Sei $X \sim \mathrm{Uniform}(0,1)$ und $Y = -\ln X$. Bestimme die Dichte von Y und identifiziere ihre Verteilung.

Beweis. Da $0 < X < 1$ fast sicher, gilt $Y > 0$. Für $y \geq 0$ kehrt die streng fallende Exponentialfunktion die Ungleichung um:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(-\ln X \leq y) = \mathbb{P}(X \geq e^{-y}) \\ &= 1 - \mathbb{P}(X < e^{-y}) = 1 - e^{-y}. \end{aligned}$$

Die Unterscheidung zwischen $<$ und $\leq$ ist wegen der Stetigkeit von X unerheblich. Ableiten liefert

$$f_Y(y) = e^{-y}, \qquad y > 0,$$

und $f_Y(y) = 0$ für $y \leq 0$. Somit ist $Y \sim \mathrm{Exp}(1)$. ■

Bemerkung 2.11.2 (Endlich viele Umkehrzweige)

Zerfällt der Träger von X in endlich viele offene Teilintervalle, auf denen g jeweils ein C^1-Diffeomorphismus auf sein Bild ist, so werden für fast alle Bildwerte y die Beiträge der Umkehrzweige addiert:

$$f_Y(y) = \sum_{x:\, g(x)=y} \frac{f_X(x)}{|g'(x)|}.$$

Jeder Summand erfasst die Wahrscheinlichkeitsmasse, die über genau einen monotonen Zweig auf die Umgebung von y abgebildet wird.

Beispiel 2.11.2 (Quadrat einer Gleichverteilung auf einem asymmetrischen Intervall)

Sei $X \sim \mathrm{Uniform}(-1, 3)$ und $Y = X^2$. Für $0 < y < 1$ liegen beide Urbilder $\pm\sqrt{y}$ im Träger von X, für $1 < y < 9$ nur $\sqrt{y}$. Daher

$$f_Y(y) = \begin{cases} \dfrac{1}{4\sqrt{y}}, & 0 < y < 1, \\ \dfrac{1}{8\sqrt{y}}, & 1 < y < 9, \\ 0, & \text{sonst.} \end{cases}$$

Die Werte an den Randpunkten $0, 1, 9$ sind für eine Dichte unerheblich.

Checkerbox 2.11.2 (Quadrat einer Standardnormalvariablen)

Sei $X \sim N(0, 1)$. Zeige für $Y = X^2$:

$$Y \sim \chi^2_1.$$

Beweis. Für $y > 0$ besitzt die Gleichung $x^2 = y$ die beiden Lösungen $x = \pm\sqrt{y}$. Mit der Formel für endlich viele Umkehrzweige und der Standardnormaldichte $\varphi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ folgt

$$\begin{aligned} f_Y(y) &= \frac{\varphi(\sqrt{y})}{2\sqrt{y}} + \frac{\varphi(-\sqrt{y})}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi y}} e^{-y/2}, \qquad y > 0. \end{aligned}$$

Für $y \le 0$ ist die Dichte gleich 0. Andererseits hat $\mathrm{Gamma}(1/2, 2)$ die Dichte

$$\frac{1}{2^{1/2}\Gamma(1/2)} y^{-1/2} e^{-y/2} = \frac{1}{\sqrt{2\pi y}} e^{-y/2},$$

weil $\Gamma(1/2) = \sqrt{\pi}$. Nach der Definition der χ^2-Verteilung ist $\mathrm{Gamma}(1/2, 2) = \chi^2_1$. ∎

Transformationen mehrerer Zufallsvariabler

Satz 2.10.1 liefert die allgemeine Formel. Das folgende Beispiel zeigt die Rechnung in zwei Dimensionen.

Beispiel 2.11.3 (Summe und Differenz zweier Gleichverteilungen)

Seien $X_1, X_2 \sim \mathrm{Uniform}(0, 1)$ unabhängig und

$$Y_1 = X_1 + X_2, \qquad Y_2 = X_1 - X_2.$$

Die Umkehrung lautet $X_1 = (Y_1 + Y_2)/2$, $X_2 = (Y_1 - Y_2)/2$, und

$$\det \begin{pmatrix} \tfrac{1}{2} & \tfrac{1}{2} \\ \tfrac{1}{2} & -\tfrac{1}{2} \end{pmatrix} = -\frac{1}{2}.$$

Da $f_{X_1,X_2} = 1$ auf $(0,1)^2$, folgt

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2}$$

auf dem Bildbereich $|y_2| < y_1 < 2 - |y_2|$ und 0 außerhalb.

Checkerbox 2.11.3 (Summe zweier Exponentialvariablen)

Seien X und Y unabhängig und jeweils $\mathrm{Exp}(1)$-verteilt. Bestimme die Verteilung von $Z = X + Y$.

Beweis. Ergänze $Z = X + Y$ um $W = X$. Die Umkehrung der Transformation lautet

$$X = W, \qquad Y = Z - W,$$

und ihre Jacobi-Determinante hat den Betrag 1. Aus $X > 0$ und $Y > 0$ wird $z > 0$ sowie $0 < w < z$. Wegen der Unabhängigkeit ist $f_{X,Y}(x,y) = e^{-(x+y)}$ auf dem positiven Quadranten. Somit

$$f_{Z,W}(z,w) = e^{-z}, \qquad z > 0,\ 0 < w < z.$$

Marginalisierung über w ergibt

$$f_Z(z) = \int_0^z e^{-z}\, dw = z e^{-z}, \qquad z > 0,$$

und $f_Z(z) = 0$ für $z \le 0$. Da $\Gamma(2) = 1$, ist dies die Dichte von $\mathrm{Gamma}(2,1)$. Also

$$X + Y \sim \mathrm{Gamma}(2,1).$$ ■

Checkerbox 2.11.4 (Box-Muller-Transformation (Vertiefung))

Seien $U_1, U_2 \sim \mathrm{Uniform}(0,1)$ unabhängig. Definiere

$$X = \sqrt{-2\ln U_1}\cos(2\pi U_2), \qquad Y = \sqrt{-2\ln U_1}\sin(2\pi U_2).$$

Zeige, dass X und Y unabhängig und jeweils standardnormalverteilt sind.
Hinweis: Bestimme zuerst die Umkehrung samt Bildbereich und berechne danach die Jacobi-Determinante.

Beweis. Setze $R = \sqrt{X^2 + Y^2}$ und bezeichne den zu (X,Y) gehörenden Polarwinkel mit $\Theta \in (0, 2\pi)$. Bis auf eine Nullmenge ist die Transformation bijektiv, und ihre Umkehrung lautet

$$U_1 = \exp\left(-\frac{X^2+Y^2}{2}\right), \qquad U_2 = \frac{\Theta}{2\pi}.$$

Für $r^2 = x^2 + y^2 > 0$ gelten

$$\frac{\partial\Theta}{\partial x} = -\frac{y}{r^2}, \qquad \frac{\partial\Theta}{\partial y} = \frac{x}{r^2}.$$

Daher beträgt der Betrag der Determinante der Jacobi-Matrix der Umkehrung

$$\left|\det \frac{\partial(u_1,u_2)}{\partial(x,y)}\right| = \left|\det \begin{pmatrix} -xu_1 & -yu_1 \\ -\dfrac{y}{2\pi r^2} & \dfrac{x}{2\pi r^2} \end{pmatrix}\right| = \frac{u_1}{2\pi} = \frac{1}{2\pi} e^{-(x^2+y^2)/2}.$$

Die gemeinsame Dichte von (U_1, U_2) ist auf $(0,1)^2$ gleich 1. Die Transformationsformel liefert deshalb

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} = \left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-y^2/2}\right).$$

Die gemeinsame Dichte ist das Produkt zweier Standardnormaldichten. Somit sind X und Y unabhängig und jeweils $N(0,1)$-verteilt. ■

2.12 Maßtheoretische Vertiefung: Äußeres Maß und Carathéodory

Die Rückrichtung des Charakterisierungssatzes für Verteilungsfunktionen verwendete den **Carathéodoryschen Erweiterungssatz**. Wir holen den dahinter stehenden maßtheoretischen Apparat hier nach. Der Abschnitt kann beim ersten Lesen übersprungen werden.

Motivation

Das **äußere Maß** (outer measure) ist ein zentrales Konzept der Maßtheorie. Es ermöglicht, aus einer „vorbereiteten" Mengenfunktion (z. B. einem Prämaß auf einem Ring oder Semiring) eine Mengenfunktion auf der gesamten Potenzmenge zu konstruieren. Es dient als Zwischenschritt im Carathéodoryschen Erweiterungssatz, um ein echtes Maß auf einer σ-Algebra zu erhalten (z. B. das Lebesgue-Maß oder Wahrscheinlichkeitsmaße aus CDFs).

Definition 2.12.1 (Ring und Semiring)

Sei X eine Menge. Eine Familie $\mathcal{R} \subseteq \mathcal{P}(X)$ heißt **Ring**, falls

1. $\emptyset \in \mathcal{R}$,
2. $A, B \in \mathcal{R} \Rightarrow A \cup B \in \mathcal{R}$,
3. $A, B \in \mathcal{R} \Rightarrow A \setminus B \in \mathcal{R}$.

Die Abgeschlossenheit unter Durchschnitten folgt daraus, denn $A \cap B = A \setminus (A \setminus B)$.
Eine Familie $\mathcal{S} \subseteq \mathcal{P}(X)$ heißt **Semiring** (auch: Halbring), falls

1. $\emptyset \in \mathcal{S}$,
2. $A, B \in \mathcal{S} \Rightarrow A \cap B \in \mathcal{S}$,
3. für $A, B \in \mathcal{S}$ lässt sich $A \setminus B$ als *endliche disjunkte* Vereinigung von Mengen aus $\mathcal{S}$ schreiben.

Bemerkung 2.12.1 (Die Hierarchie – und wozu der Semiring gut ist)

Die Mengensysteme dieses Buches ordnen sich der Stärke nach:

$$\sigma\text{-Algebra} \subset \text{Algebra} \subset \text{Ring} \subset \text{Semiring},$$

zu lesen als „jede σ-Algebra ist eine Algebra, jede Algebra ein Ring, jeder Ring ein Semiring". Dabei ist eine **Algebra** ein Ring, der zusätzlich X selbst enthält, und eine σ-Algebra (Definition 1.8.1) eine Algebra, die überdies unter *abzählbaren* Vereinigungen abgeschlossen ist. Dass jeder Ring ein Semiring ist, sieht man an Punkt (3): Dort liegt $A \setminus B$ bereits selbst im System und ist damit eine disjunkte Vereinigung aus einem einzigen Stück.
Der Semiring ist der schwächste dieser Begriffe – und genau deshalb der übliche Startpunkt. Die halboffenen Intervalle

$$\mathcal{S} = \{(a, b] : a \le b\}$$

bilden einen Semiring, aber keinen Ring, denn

$$(0, 3] \setminus (1, 2] = (0, 1] \cup (2, 3]$$

ist kein Intervall mehr. Auf $\mathcal{S}$ ist das Messen dafür völlig unmittelbar: Die Länge $\mu_0((a, b]) = b - a$ – im Wahrscheinlichkeitsfall $F(b) - F(a)$ – steht ohne jede Konstruktion da. Man beginnt also bei dem System, auf dem der Messwert offensichtlich ist, und arbeitet sich von dort zur σ-Algebra hoch. Genau diesen Weg geht der Erweiterungssatz weiter unten.

Definition 2.12.2 (Äußeres Maß aus einem Prämaß)

Sei X eine Menge und $\mathcal{R}$ ein Ring oder Semiring von Teilmengen von X (Definition 2.12.1). Sei $\mu_0 : \mathcal{R} \to [0, \infty]$ ein **Prämaß**, d. h. $\mu_0(\emptyset) = 0$ und μ_0 ist σ-additiv im Sinne von Definition 1.8.2 – allerdings nur, soweit die Vereinigung den Ring nicht verlässt: für paarweise disjunkte $A_1, A_2, \ldots \in \mathcal{R}$ *mit* $\bigcup_i A_i \in \mathcal{R}$ gilt $\mu_0(\bigcup_i A_i) = \sum_i \mu_0(A_i)$. Diese Einschränkung ist nötig, weil ein Ring anders als eine σ-Algebra nicht unter abzählbaren Vereinigungen abgeschlossen ist. Das zugehörige **äußere Maß** $\mu^* : \mathcal{P}(X) \to [0, \infty]$ ist

$$\mu^*(E) := \inf\left\{ \sum_{n=1}^{\infty} \mu_0(A_n) \;\middle|\; A_n \in \mathcal{R},\ E \subset \bigcup_{n=1}^{\infty} A_n \right\},$$

wobei das Infimum über alle abzählbaren Überdeckungen von E durch Mengen aus $\mathcal{R}$ genommen wird. (Existiert keine Überdeckung, so $\mu^*(E) = \infty$; die leere Summe ist 0, also $\mu^*(\emptyset) = 0$.)

Allgemeiner lässt sich ein äußeres Maß direkt axiomatisch definieren:

Definition 2.12.3 (Äußeres Maß (axiomatisch))

Eine Funktion $\mu^* : \mathcal{P}(X) \to [0, \infty]$ heißt **äußeres Maß**, wenn

1. $\mu^*(\emptyset) = 0$,
2. **Monotonie:** $E \subset F \implies \mu^*(E) \le \mu^*(F)$,

3. **abzählbare Subadditivität:** für $E_n \subset X$ gilt

$$\mu^*\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} \mu^*(E_n).$$

Korollar 2.12.4 (Eigenschaften des aus einem Prämaß konstruierten μ^*)

Das aus μ_0 konstruierte μ^* ist tatsächlich ein äußeres Maß und erfüllt zusätzlich:

- μ^* erweitert μ_0: für alle $A \in \mathcal{R}$ gilt $\mu^*(A) = \mu_0(A)$.
- Ist μ_0 σ-endlich, so hat μ^* weitere Regularitätseigenschaften.

Ohne Beweis. Die drei Eigenschaften eines äußeren Maßes rechnet man direkt aus der Infimumsdefinition nach; der einzige nichttriviale Schritt ist $\mu^*|_{\mathcal{R}} = \mu_0$, der die σ-Additivität des Prämaßes auf dem Ring benötigt. Beides ist Standardstoff der Maßtheorie und wird hier nur benutzt, nicht entwickelt. Nachzulesen bei Elstrodt, *Maß- und Integrationstheorie*, Kap. II, §4–5, oder Billingsley, *Probability and Measure*, §11.

Bemerkungen.

- Das äußere Maß ist im Allgemeinen *nicht* additiv, sondern nur subadditiv. Additivität gilt nur für messbare Mengen (im Sinne von Carathéodory).
- Das **Lebesgue-äußere Maß** auf $\mathbb{R}^d$ entsteht aus der Längen-/Volumenfunktion auf Quadern.
- Für Wahrscheinlichkeitsmaße: aus $\mu_0((a,b]) = F(b) - F(a)$ entsteht ein äußeres Maß, das via Carathéodory zu einem Wahrscheinlichkeitsmaß auf den Borelmengen eingeschränkt wird.

Der Carathéodorysche Erweiterungssatz

Lemma 2.12.5 (Carathéodoryscher Erweiterungssatz)

Sei μ^* ein äußeres Maß auf X. Eine Menge $A \subset X$ heißt μ^***-messbar** (im Sinne von Carathéodory), wenn für alle $S \subset X$

$$\mu^*(S) = \mu^*(S \cap A) + \mu^*(S \cap A^c).$$

Dann gilt:

1. Die Menge $\mathcal{M}$ aller μ^*-messbaren Mengen ist eine σ-Algebra.
2. Die Einschränkung $\mu := \mu^*|_{\mathcal{M}}$ ist ein vollständiges Maß auf $\mathcal{M}$.
3. War μ_0 ein Prämaß auf einem Ring $\mathcal{R}$ und μ^* daraus konstruiert, so erweitert μ das Prämaß: $\mu|_{\mathcal{R}} = \mu_0$.
4. Ist zusätzlich μ_0 σ-endlich, so ist die Erweiterung μ auf der von $\mathcal{R}$ erzeugten σ-Algebra $\sigma(\mathcal{R})$ **eindeutig**.

Ohne Beweis. Der Kern des Beweises ist der Nachweis, dass $\mathcal{M}$ eine σ-Algebra und μ^* darauf σ-additiv ist; die Eindeutigkeit im σ-endlichen Fall folgt anschließend mit einem Dynkin-Argument. Beides ist eine mehrseitige maßtheoretische Konstruk-

tion, die dieser Band voraussetzt statt sie zu entwickeln. Vollständig bei Billingsley, *Probability and Measure*, §11, Bauer, *Wahrscheinlichkeitstheorie*, §5, oder Elstrodt, *Maß- und Integrationstheorie*, Kap. II, §4–5.

Bemerkungen.

- **Vollständig** heißt: Jede Teilmenge einer μ-Nullmenge ist selbst messbar und hat wieder Maß 0. Das ist eine Zusatzeigenschaft, die die Carathéodory-Konstruktion gratis mitliefert. Die Borel-σ-Algebra mit dem Lebesgue-Maß ist *nicht* vollständig; ihre Vervollständigung sind gerade die Lebesgue-messbaren Mengen. Maß, Maßraum und σ-Endlichkeit stehen in Definition 1.8.3.
- Der kritische Punkt ist der Nachweis, dass $\mathcal{M}$ eine σ-Algebra ist (insbesondere die σ-Additivität auf den messbaren Mengen). Dies erfordert Monotonie und Subadditivität des äußeren Maßes.
- Für das Lebesgue-Maß auf $\mathbb{R}^d$: Starte mit $\mu_0((a,b]) = \ell(b) - \ell(a)$ (Länge) auf dem Semiring der Halbintervalle, konstruiere μ^*, wende Carathéodory an $\to$ Lebesgue-Maß.
- Im Kontext von Wahrscheinlichkeitsmaßen auf $\mathbb{R}$ (wie im cdf-Beweis): das durch $F(b) - F(a)$ definierte μ auf Halbintervallen ist ein Prämaß (wegen Rechtsstetigkeit σ-additiv), und der Satz liefert die Erweiterung auf die Borel-σ-Algebra.
- Die σ-Endlichkeit sorgt für Eindeutigkeit; ohne sie kann es mehrere Erweiterungen geben.

KAPITEL 3

Erwartungswert

3.1 Erwartungswert einer Zufallsvariable

Der **Erwartungswert** einer Zufallsvariable X fasst eine ganze Verteilung in einer einzigen Zahl zusammen – dem langfristigen Durchschnittswert.

Definition 3.1.1

Der **Erwartungswert** von X ist

$$\mathbb{E}[X] = \int x \, dF(x) = \begin{cases} \sum_x x f(x) & \text{falls } X \text{ diskret,} \\ \int x f(x)\, dx & \text{falls } X \text{ stetig,} \end{cases}$$

falls die Summe (bzw. das Integral) wohldefiniert ist, d. h. $\int |x| \, dF_X(x) < \infty$.

Man kann $\mathbb{E}[X]$ als Durchschnitt $\frac{1}{n}\sum_{i=1}^{n} X_i$ einer großen Anzahl iid Ziehungen verstehen. Dies wird durch das **Gesetz der großen Zahlen** (Kapitel 5) rigoros gemacht.

Beispiel 3.1.1

Sei

$$X \sim \text{Bernoulli}(p), \qquad p \in (0,1).$$

Das bedeutet: X kann nur zwei Werte annehmen,

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

Die Wahrscheinlichkeitsfunktion ist also

$$p_X(x) = \mathbb{P}(X = x) = \begin{cases} 1-p & \text{für } x = 0, \\ p & \text{für } x = 1, \\ 0 & \text{sonst.} \end{cases}$$

Der Erwartungswert einer diskreten Zufallsvariable ist

$$\mathbb{E}[X] = \sum_x x \cdot \mathbb{P}(X = x).$$

Da X nur die Werte 0 und 1 annehmen kann, reicht die Summe über diese beiden Werte:

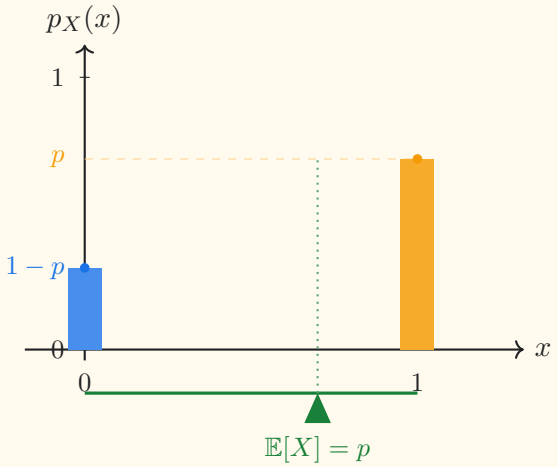
$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1).$$

Einsetzen der Wahrscheinlichkeiten liefert

$$\mathbb{E}[X] = 0 \cdot (1-p) + 1 \cdot p = p.$$

Also gilt

$$\boxed{\mathbb{E}[X] = p.}$$



Wahrscheinlichkeitsfunktion einer $\mathrm{Bernoulli}(p)$*-Variable (hier* $p = 0{,}7$*): zwei Stäbe der Höhe* $1-p$ *bei* $x = 0$ *und* p *bei* $x = 1$*. Der Erwartungswert* $\mathbb{E}[X] = p$ *ist der* Schwerpunkt *der Verteilung – der Punkt, an dem der mit den Wahrscheinlichkeiten gewichtete Balken im Gleichgewicht steht. Da* $p = 0{,}7 > \frac{1}{2}$*, liegt er näher bei* 1*.*

Interpretation: Eine Bernoulli-Zufallsvariable modelliert einen Ja/Nein-Versuch, etwa Erfolg/Misserfolg. Wenn der Erfolg mit Wahrscheinlichkeit p eintritt und als 1 gezählt wird, dann ist der langfristige Durchschnitt genau p.

Beispiel 3.1.2

Fairer Münzwurf zweimal, X = Anzahl Köpfe: $\mathbb{E}[X] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$.

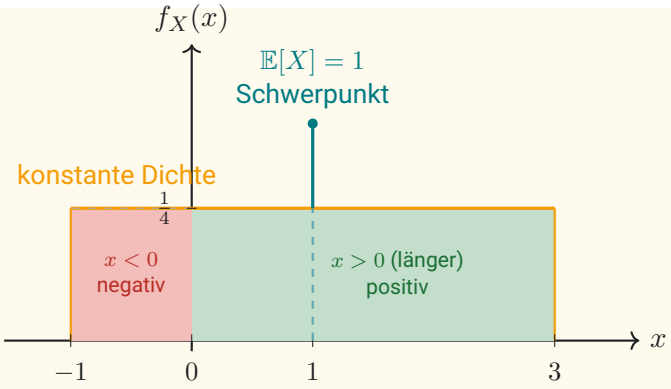
Beispiel 3.1.3

Sei $X \sim \mathrm{Uniform}(-1, 3)$. Das Intervall hat Länge $3 - (-1) = 4$; deshalb muss die konstante Dichte die Höhe $1/4$ haben:

$$f_X(x) = \begin{cases} \frac{1}{4}, & -1 \le x \le 3, \\ 0, & \text{sonst.} \end{cases}$$

Damit ist die gesamte Fläche unter der Dichte gleich $4 \cdot \frac{1}{4} = 1$. Der Erwartungswert ist

$$\mathbb{E}[X] = \int x f_X(x)\,dx = \frac{1}{4}\int_{-1}^{3} x\,dx = \frac{1}{4}\left[\frac{x^2}{2}\right]_{-1}^{3} = \frac{1}{4}\left(\frac{9}{2} - \frac{1}{2}\right) = \frac{1}{4} \cdot 4 = 1.$$



Didaktische Lesart: Eine Dichte ist keine Wahrscheinlichkeit an einem einzelnen Punkt, sondern Wahrscheinlichkeit pro Längeneinheit. Bei der Gleichverteilung ist diese Dichte überall im Intervall gleich hoch. Deshalb liegt der Schwerpunkt der Rechtecksfläche genau in der Mitte des Intervalls. Die Rechnung $\int x f_X(x)\,dx$ berechnet diesen Schwerpunkt algebraisch: Jeder Ort x wird mit derselben Dichte $\frac{1}{4}$ gewichtet. Der kleine negative Anteil von -1 bis 0 wird vom längeren positiven Anteil von 0 bis 3 übertroffen, und der Ausgleichspunkt liegt bei 1.
Kurzformel für jede Gleichverteilung:

$$X \sim \text{Uniform}(a, b) \quad \Longrightarrow \quad \mathbb{E}[X] = \frac{a+b}{2}.$$

Hier also $\mathbb{E}[X] = \frac{-1+3}{2} = 1$.

Bemerkung 3.1.1 (Tails einer Verteilung)

Die **Tails** (Randbereiche) einer Verteilung sind die Bereiche weit links und weit rechts vom Zentrum, bei einer Dichte also grob das Verhalten für

$$x \to -\infty \quad \text{und} \quad x \to +\infty.$$

Sie beschreiben, wie wahrscheinlich *sehr extreme* Werte sind. Man unterscheidet:

- **Dünne (leichte) Tails:** extreme Werte werden sehr schnell unwahrscheinlich. Beispiel: die *Normalverteilung*, deren Dichte wie $e^{-x^2/2}$ abfällt.
- **Schwere Tails:** die Dichte fällt nur langsam (etwa polynomial) ab, sodass weit außen mehr Masse liegt und Ausreißer deutlich häufiger auftreten. Beispiel: die *Cauchy-Verteilung* mit $f_X(x) \sim 1/(\pi x^2)$.

Zur Veranschaulichung: bei $x = 10$ ist die Cauchy-Dichte noch $\approx 3{,}15 \cdot 10^{-3}$, während die Normaldichte dort bereits bei $\sim 10^{-23}$ liegt. Schwere Tails sind der Grund, warum für die Cauchy-Verteilung Erwartungswert und Varianz nicht existieren (folgendes Beispiel).

Beispiel 3.1.4 (Cauchy-Verteilung)

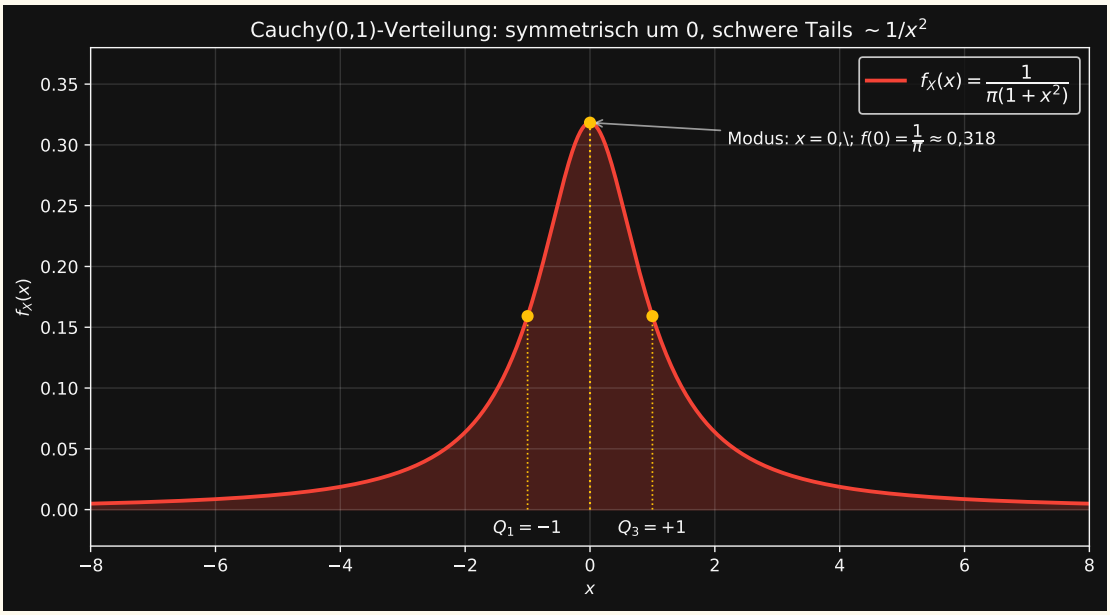
Sei X Cauchy-verteilt mit Dichte $f_X(x) = \frac{1}{\pi(1+x^2)}$ auf $\mathbb{R}$. Wir prüfen die Integrierbarkeitsbedingung $\int |x| f_X(x)\, dx < \infty$:

$$\int_{-\infty}^{\infty} |x| f_X(x)\, dx = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2}\, dx$$

(Faktor 2 wegen Symmetrie von $|x|/(1+x^2)$). Mit der Substitution $u = 1 + x^2$, $du = 2x\, dx$ folgt

$$\frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2}\, dx = \frac{1}{\pi} \left[\ln(1+x^2)\right]_0^{\infty} = \frac{1}{\pi} \cdot \infty = \infty.$$

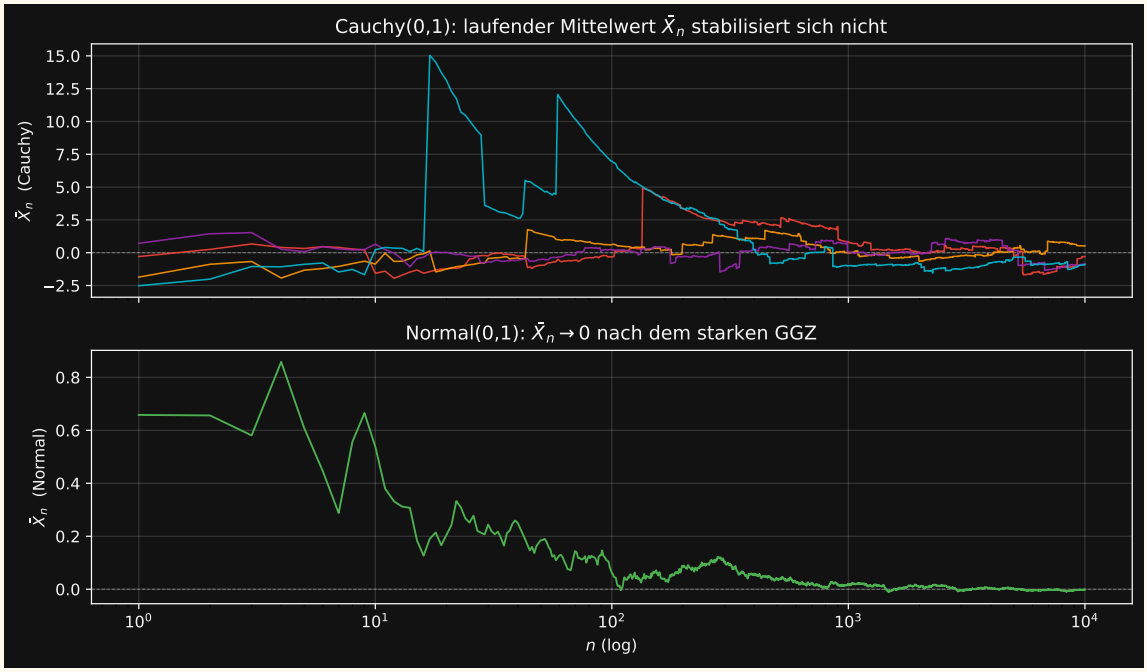
Also ist $\mathbb{E}[|X|] = \infty$, d. h. der Erwartungswert $\mathbb{E}[X]$ existiert nicht (der Hauptwert wäre formal 0 wegen Symmetrie, aber das Lebesgue-Integral ist nicht definiert). Da bereits der Mittelwert fehlt, ist für die Cauchy-Verteilung auch die Varianz nicht definiert (vgl. Bemerkung 3.3.1).



Dichte: Symmetrisch um 0 mit Modus $f(0) = 1/\pi \approx 0{,}318$ und Quartilen $Q_1 = -1$, $Q_3 = +1$. Die Tails fallen nur wie $1/(\pi x^2)$ ab – bei $x = 10$ ist $f_X(10) \approx 3{,}15 \cdot 10^{-3}$, während die Normaldichte dort schon bei $\sim 10^{-23}$ liegt.

Beispiel 3.1.5 (Cauchy-Simulation: GGZ versagt)

Praktische Konsequenz aus dem vorigen Beispiel: Simuliert man viele Cauchy-Ziehungen, stabilisiert sich der Durchschnitt $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ nie – das Gesetz der großen Zahlen versagt, weil die Tails zu schwer sind.



Simulation: Vier unabhängige Cauchy-Trajektorien (oben) wandern auch bei $n = 10{,}000$ noch in völlig verschiedene Richtungen (Endwerte $-0{,}30$, $+0{,}52$, $-0{,}85$, $-0{,}89$). Zum Vergleich konvergiert eine $N(0,1)$-Trajektorie (unten) erwartungsgemäß gegen 0 ($\bar{X}_{10000} \approx -0{,}002$). Skript: `scripts/cauchy_simulation.py`.

Satz 3.1.2 (Regel des faulen Statistikers)

Sei $Y = r(X)$. Dann $\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int r(x)\, dF_X(x)$.
Man muss nicht erst die Verteilung von Y bestimmen – eine enorme Ersparnis.

Bemerkung 3.1.2 (LOTUS – Law of the Unconscious Statistician)

Die „Regel des faulen Statistikers" heißt im Englischen **LOTUS** (*Law of the Unconscious Statistician*). Sie besagt: Will man den Erwartungswert einer transformierten Zufallsvariable $Y = g(X)$ berechnen, so muss man *nicht* zuerst die Verteilung von Y bestimmen – man darf direkt über die Verteilung von X summieren bzw. integrieren.

$$\text{diskret:}\quad \mathbb{E}[g(X)] = \sum_x g(x)\,\mathbb{P}(X = x), \qquad \text{stetig:}\quad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx.$$

Beispiel. Für $X \sim \mathrm{Uniform}(0,1)$ und $Y = X^2$ ist

$$\mathbb{E}[Y] = \mathbb{E}[X^2] = \int_0^1 x^2\, dx = \frac{1}{3}.$$

Ohne LOTUS müsste man erst die Verteilung von $Y = X^2$ herleiten. (Im Satz oben heißt die Transformation r statt g – dasselbe gemeint.)

Beweis. Die Kernidee passt in einen Satz: *statt y-Werte mit ihren Wahrscheinlichkeiten zu gewichten, gewichten wir direkt $r(x)$ mit der X-Wahrscheinlichkeit – jedes Wahrscheinlichkeitskorn wird genau einmal mitgezählt, nur die Summationsreihenfolge ist anders.*

Diskreter Fall. Sei X diskret mit Massefunktion p_X. Dann nimmt $Y = r(X)$ Werte in $r(\{x : p_X(x) > 0\})$ an, und für jedes y gilt

$$p_Y(y) = \mathbb{P}(r(X) = y) = \sum_{x : r(x) = y} p_X(x).$$

Damit

$$\mathbb{E}[Y] = \sum_y y\, p_Y(y) = \sum_y \sum_{x : r(x) = y} y\, p_X(x) = \sum_y \sum_{x : r(x) = y} r(x)\, p_X(x) = \sum_x r(x)\, p_X(x).$$

Im letzten Schritt wurde nur die Doppelsumme über die Partition $\{x : r(x) = y\}_y$ zu einer Summe über alle x zusammengezogen (jedes x liegt in genau einer Klasse). Die Umordnung ist erlaubt, weil bei Existenz von $\mathbb{E}[|r(X)|]$ alles absolut konvergiert.

Stetiger Fall (Skizze). Approximiere X durch dyadische Rundung: $X_n = 2^{-n}\lfloor 2^n X \rfloor$. X_n ist diskret, $X_n \to X$ fast sicher, und für stetiges r folgt $r(X_n) \to r(X)$ fast sicher. Der diskrete Fall liefert

$$\mathbb{E}[r(X_n)] = \sum_{k \in \mathbb{Z}} r(k/2^n)\, \mathbb{P}\big(X \in [k/2^n,\, (k+1)/2^n)\big).$$

Die rechte Seite ist eine Riemann-Stieltjes-Summe für $\int r(x)\, dF_X(x)$, deren Maschenweite gegen 0 geht. Unter $\mathbb{E}[|r(X)|] < \infty$ liefert dominierte Konvergenz auf der linken Seite $\mathbb{E}[r(X_n)] \to \mathbb{E}[r(X)]$, und Konvergenz der Riemann-Summen auf der rechten Seite das gewünschte Integral. Beide Seiten haben denselben Grenzwert.

Allgemein (messbares r ohne Stetigkeitsannahme) folgt dasselbe Resultat über das Standardargument für Bildmaße: Indikator $\to$ einfache Funktion $\to$ nichtnegativ messbar $\to$ integrierbar. ■

Beispiel 3.1.6 (LOTUS: Exponentialfunktion einer Uniformvariable)

Sei $X \sim \text{Uniform}(0, 1)$ und

$$Y = e^X.$$

Gesucht ist $\mathbb{E}[Y]$, also der mittlere Wert von e^X, wenn X gleichmäßig zwischen 0 und 1 liegt.
1. Verteilung von X. Für $X \sim \text{Uniform}(0, 1)$ ist die Dichte

$$f_X(x) = \begin{cases} 1, & 0 \le x \le 1, \\ 0, & \text{sonst.} \end{cases}$$

Die Dichte ist also auf dem Intervall konstant: jeder kleine Abschnitt gleicher Länge ist gleich wahrscheinlich.
2. LOTUS anwenden. Wir müssen nicht zuerst die Verteilung von $Y = e^X$ bestimmen. Nach LOTUS gilt direkt

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

Weil $f_X(x)$ außerhalb von $[0, 1]$ gleich 0 ist und innerhalb von $[0, 1]$ gleich 1, reduziert sich das Integral zu

$$\mathbb{E}[e^X] = \int_0^1 e^x \cdot 1\, dx.$$

3. Integral ausrechnen. Eine Stammfunktion von e^x ist wieder e^x. Daher

$$\int_0^1 e^x \, dx = [e^x]_0^1 = e^1 - e^0 = e - 1.$$

Somit

$$\boxed{\mathbb{E}[Y] = \mathbb{E}[e^X] = e - 1 \approx 1{,}718.}$$

Plausibilitätskontrolle. Da $0 \leq X \leq 1$ gilt,

$$1 = e^0 \leq e^X \leq e^1 = e.$$

Der Erwartungswert von $Y = e^X$ muss also zwischen 1 und $e \approx 2{,}718$ liegen. Der Wert $e - 1 \approx 1{,}718$ passt genau in diesen Bereich.

Beispiel 3.1.7 (Stab der Länge 1, zufällig gebrochen)

Aufgabe. Ein Stab der Länge 1 wird an einer rein zufällig gewählten Stelle $X \sim \text{Uniform}(0, 1)$ durchgebrochen. Es entstehen zwei Teilstücke der Längen X und $1 - X$. Gesucht: der erwartete Wert der *längeren* der beiden Längen.

Modellierung. Die Länge des längeren Stücks ist

$$Y = \max(X, 1 - X) = \begin{cases} 1 - X & \text{falls } X < \frac{1}{2} \quad \text{(linkes Stück kürzer, rechtes länger)}, \\ X & \text{falls } X \geq \frac{1}{2} \quad \text{(rechtes Stück kürzer, linkes länger)}. \end{cases}$$

Bei $X = 1/2$ sind beide Stücke gleich lang ($Y = 1/2$), und es ist offensichtlich $Y \in [1/2, 1]$: das längere Stück ist mindestens halb so lang wie der Stab und höchstens so lang wie der Stab selbst.

Anwendung von LOTUS. Setze $r(x) = \max(x, 1 - x)$. Da X stetig mit Dichte $f_X(x) = 1$ auf $[0, 1]$ ist, liefert die Regel des faulen Statistikers

$$\mathbb{E}[Y] = \int_0^1 r(x) \cdot \underbrace{f_X(x)}_{=1} \, dx = \int_0^1 \max(x, 1 - x) \, dx.$$

Das Integral spalten wir am Knickpunkt $x = 1/2$ auf:

$$\begin{aligned} \mathbb{E}[Y] &= \int_0^{1/2} \underbrace{(1 - x)}_{=\max(x,1-x)} \, dx + \int_{1/2}^1 \underbrace{x}_{=\max(x,1-x)} \, dx \\ &= \left[x - \frac{x^2}{2}\right]_0^{1/2} + \left[\frac{x^2}{2}\right]_{1/2}^1 \\ &= \left(\frac{1}{2} - \frac{1}{8}\right) + \left(\frac{1}{2} - \frac{1}{8}\right) \\ &= \frac{3}{8} + \frac{3}{8} = \frac{3}{4}. \end{aligned}$$

Die beiden Beiträge sind gleich groß – das musste aufgrund der Symmetrie $X \leftrightarrow 1 - X$ so sein.

Plausibilitätscheck. Das längere Stück ist immer $\geq 1/2$ und ≤ 1, also muss $\mathbb{E}[Y] \in [1/2, 1]$ liegen. Der Wert $3/4$ ist genau der Mittelpunkt dieses Intervalls.

Beispiel 3.1.8 (Stab: Verteilung des längeren Stücks)

Im vorigen Beispiel haben wir mit LOTUS direkt $\mathbb{E}[Y] = 3/4$ berechnet. Jetzt bestimmen wir die ganze Verteilung von

$$Y = \max(X, 1 - X),$$

wobei $X \sim \text{Uniform}(0, 1)$ die Bruchstelle des Stabs beschreibt. Dieses Beispiel zeigt gut, wie man eine Verteilung einer transformierten Zufallsvariable systematisch über ihre Verteilungsfunktion findet.

1. Wertebereich verstehen. Das längere der beiden Stücke kann nie kürzer als die Hälfte des Stabs sein. Es kann aber beliebig nahe an die volle Länge 1 herankommen, wenn der Bruch sehr nahe an einem Ende liegt. Also gilt

$$Y \in \left[\frac{1}{2}, 1\right].$$

Damit ist sofort klar:

$$F_Y(y) = \mathbb{P}(Y \le y) = 0 \quad \text{für } y < \frac{1}{2}, \qquad F_Y(y) = 1 \quad \text{für } y \ge 1.$$

2. Das Ereignis übersetzen. Für $1/2 \le y < 1$ fragen wir:

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(\max(X, 1 - X) \le y).$$

Das Maximum ist genau dann höchstens y, wenn *beide* Stücke höchstens y lang sind:

$$\max(X, 1 - X) \le y \quad \Longleftrightarrow \quad X \le y \text{ und } 1 - X \le y.$$

Die zweite Ungleichung formen wir um:

$$1 - X \le y \quad \Longleftrightarrow \quad X \ge 1 - y.$$

Zusammen erhalten wir

$$\max(X, 1 - X) \le y \quad \Longleftrightarrow \quad 1 - y \le X \le y.$$

3. Die Wahrscheinlichkeit berechnen. Da X gleichverteilt auf $[0, 1]$ ist, ist die Wahrscheinlichkeit eines Intervalls einfach seine Länge. Also

$$F_Y(y) = \mathbb{P}(1 - y \le X \le y) = y - (1 - y) = 2y - 1, \qquad \frac{1}{2} \le y < 1.$$

Insgesamt lautet die Verteilungsfunktion daher

$$F_Y(y) = \begin{cases} 0, & y < \frac{1}{2}, \\ 2y - 1, & \frac{1}{2} \le y < 1, \\ 1, & y \ge 1. \end{cases}$$

4. Dichte und Verteilung erkennen. Auf dem Intervall $(1/2, 1)$ gilt

$$f_Y(y) = F_Y'(y) = 2.$$

Das ist genau die Dichte einer Gleichverteilung auf $[1/2, 1]$, denn die Länge dieses Intervalls ist $1/2$ und daher ist die konstante Dichte $1/(1/2) = 2$. Somit

$$Y \sim \text{Uniform}\left(\frac{1}{2}, 1\right).$$

5. Kontrolle des Erwartungswerts. Aus der bekannten Formel für die Gleichverteilung folgt

$$\mathbb{E}[Y] = \frac{a+b}{2} = \frac{\frac{1}{2}+1}{2} = \frac{3}{4}.$$

Das stimmt mit dem LOTUS-Ergebnis aus dem vorigen Beispiel überein. LOTUS war schneller, aber die Verteilungsfunktion erklärt zusätzlich, *warum* das längere Stück gleichverteilt zwischen $1/2$ und 1 ist.
Interpretation. Im langfristigen Mittel ist das längere Stück $75\,\%$ des Stabs, das kürzere entsprechend $25\,\%$. Da die beiden Stücke zusammen immer Länge 1 haben, ist das längere Stück im Mittel dreimal so lang wie das kürzere.

Existenz niedrigerer Momente

Lemma 3.1.3 (Monotonie positiver Momente)

Seien $0 \leq j < k < \infty$. Gilt $\mathbb{E}[|X|^k] < \infty$, so gilt auch

$$\mathbb{E}[|X|^j] < \infty.$$

Ein endliches höheres absolutes Moment erzwingt also die Existenz aller niedrigeren absoluten Momente.

Beweis. Wir zerlegen den Erwartungswert in die Bereiche $\{|X| \leq 1\}$ und $\{|X| > 1\}$. Auf dem ersten Bereich gilt $|X|^j \leq 1$, auf dem zweiten wegen $j < k$ die Abschätzung $|X|^j \leq |X|^k$. Daher

$$\mathbb{E}[|X|^j] = \mathbb{E}\left[|X|^j I_{|X|\leq 1}\right] + \mathbb{E}\left[|X|^j I_{|X|>1}\right] \leq 1 + \mathbb{E}[|X|^k] < \infty.$$

■

Lemma 3.1.4 (Monotonie negativer Momente)

Seien $0 < j < k < \infty$. Gilt $\mathbb{E}[|X|^{-k}] < \infty$, so gilt auch

$$\mathbb{E}[|X|^{-j}] < \infty.$$

Dabei wird $|0|^{-r} = \infty$ verstanden; die Voraussetzung impliziert insbesondere $\mathbb{P}(X = 0) = 0$.

Beweis. Auf $\{0 < |X| \leq 1\}$ gilt $|X|^{-j} \leq |X|^{-k}$, auf $\{|X| > 1\}$ gilt $|X|^{-j} \leq 1$. Somit

$$\mathbb{E}[|X|^{-j}] \leq \mathbb{E}[|X|^{-k}] + 1 < \infty.$$

■

3.2 Eigenschaften des Erwartungswerts

Satz 3.2.1 (Linearität)

Für Zufallsvariablen $X_1, \dots, X_n$ und Konstanten $a_1, \dots, a_n$ gilt

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

Beweis. Per Induktion über n genügt es, den Fall $n = 2$ zu zeigen: $\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$ für X, Y mit $\mathbb{E}[|X|], \mathbb{E}[|Y|] < \infty$.

Diskreter Fall. Sei $p(x,y) = \mathbb{P}(X = x,\, Y = y)$ die gemeinsame Massefunktion mit Randverteilungen $p_X(x) = \sum_y p(x,y)$ und $p_Y(y) = \sum_x p(x,y)$. Mit LOTUS für $r(x,y) = ax + by$:

$$\begin{aligned}
\mathbb{E}[aX + bY] &= \sum_{x,y} (ax + by)\, p(x,y) \\
&= a \sum_x x \underbrace{\sum_y p(x,y)}_{= p_X(x)} + b \sum_y y \underbrace{\sum_x p(x,y)}_{= p_Y(y)} \\
&= a\,\mathbb{E}[X] + b\,\mathbb{E}[Y].
\end{aligned}$$

Stetiger Fall. Mit gemeinsamer Dichte $f(x,y)$ und Randdichten $f_X(x) = \int f(x,y)\,dy$, $f_Y(y) = \int f(x,y)\,dx$ läuft dieselbe Rechnung mit Integralen statt Summen – der Fubini-Satz erlaubt die Vertauschung der Integrationsreihenfolge (absolute Integrierbarkeit folgt aus $\mathbb{E}[|X|], \mathbb{E}[|Y|] < \infty$).

Induktionsschritt. Anwenden des Zweier-Falls auf $S_n = \sum_{i=1}^{n} a_i X_i$ und $a_{n+1} X_{n+1}$ liefert $\mathbb{E}[S_{n+1}] = \mathbb{E}[S_n] + a_{n+1}\,\mathbb{E}[X_{n+1}]$ und damit per Induktionsvoraussetzung die Behauptung. ∎

Bemerkenswert. Im Beweis wurde *keine* Annahme über die gemeinsame Verteilung der X_i gemacht – die Linearität gilt selbst dann, wenn die X_i stark abhängig oder identisch sind. Das macht sie zum mit Abstand mächtigsten Werkzeug für Erwartungswert-Rechnungen (vgl. das folgende Binomial-Beispiel: ohne Linearität bräuchte man Binomialkoeffizienten, mit ist es eine Zeile).

Beispiel 3.2.1 (Erwartungswert der Binomialverteilung)

Sei

$$X \sim \mathrm{Binomial}(n, p).$$

Dann zählt X die Anzahl der Erfolge in n unabhängigen Versuchen, wobei jeder einzelne Versuch mit Wahrscheinlichkeit p erfolgreich ist.
Für jeden Versuch führen wir eine Indikatorvariable ein:

$$X_i = \begin{cases} 1, & \text{falls Versuch } i \text{ ein Erfolg ist,} \\ 0, & \text{falls Versuch } i \text{ kein Erfolg ist.} \end{cases}$$

Dann gilt $X_i \sim \text{Bernoulli}(p)$ und

$$X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i.$$

X ist somit die Summe der einzelnen Erfolgsindikatoren.
Aus dem Bernoulli-Beispiel wissen wir

$$\mathbb{E}[X_i] = 0 \cdot (1-p) + 1 \cdot p = p.$$

Mit der Linearität des Erwartungswerts folgt daher

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np.$$

Also

$$\boxed{\mathbb{E}[X] = np.}$$

Wichtig: Für diese Erwartungswertrechnung braucht man die Unabhängigkeit der X_i nicht. Entscheidend ist nur, dass jeder Indikator denselben Erwartungswert p hat. Die Unabhängigkeit gehört zur üblichen Definition der Binomialverteilung, aber die Linearität selbst funktioniert auch ohne sie.

Satz 3.2.2

Sind $X_1, \ldots, X_n$ unabhängig, so ist $\mathbb{E}[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} \mathbb{E}[X_i]$.

Beweis. Wir setzen $\mathbb{E}[|X_i|] < \infty$ für alle i voraus, damit alle Erwartungswerte existieren. Es genügt, den Fall zweier unabhängiger Zufallsvariablen X, Y zu zeigen, also

$$\mathbb{E}[XY] = \mathbb{E}[X]\ \mathbb{E}[Y];$$

der allgemeine Fall folgt daraus per Induktion.

1. Zwei Faktoren. Unabhängigkeit bedeutet, dass die gemeinsame Verteilung faktorisiert. Im stetigen Fall mit gemeinsamer Dichte gilt $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$, und mit LOTUS für $g(x,y) = xy$ folgt

$$\mathbb{E}[XY] = \int_{\mathbb{R}}\int_{\mathbb{R}} xy\, f_{X,Y}(x,y)\, dx\, dy = \Big(\int_{\mathbb{R}} x\, f_X(x)\, dx\Big)\Big(\int_{\mathbb{R}} y\, f_Y(y)\, dy\Big) = \mathbb{E}[X]\ \mathbb{E}[Y].$$

Im diskreten Fall ersetzt man Dichten durch Massefunktionen und Integrale durch Summen:

$$\mathbb{E}[XY] = \sum_{x}\sum_{y} xy\, p_X(x)\, p_Y(y) = \Big(\sum_{x} x\, p_X(x)\Big)\Big(\sum_{y} y\, p_Y(y)\Big) = \mathbb{E}[X]\ \mathbb{E}[Y].$$

Das Vertauschen von Doppelsumme bzw. Doppelintegral und Produkt ist erlaubt, weil bei $\mathbb{E}[|X|], \mathbb{E}[|Y|] < \infty$ alles absolut konvergiert (Satz von Fubini–Tonelli).

2. Induktion über n. Für $n = 1$ ist nichts zu zeigen. Die Aussage gelte für $n-1$ Faktoren. Da X_n unabhängig vom Vektor $(X_1, \dots, X_{n-1})$ ist, ist X_n auch unabhängig von der Funktion $\prod_{i=1}^{n-1} X_i$ dieses Vektors. Mit dem Zwei-Faktoren-Fall und der Induktionsvoraussetzung erhalten wir

$$\mathbb{E}\Big[\prod_{i=1}^{n} X_i\Big] = \mathbb{E}\Big[\Big(\prod_{i=1}^{n-1} X_i\Big) X_n\Big] = \mathbb{E}\Big[\prod_{i=1}^{n-1} X_i\Big]\,\mathbb{E}[X_n] = \Big(\prod_{i=1}^{n-1} \mathbb{E}[X_i]\Big)\,\mathbb{E}[X_n] = \prod_{i=1}^{n} \mathbb{E}[X_i]. \qquad \blacksquare$$

3.3 Varianz und Kovarianz

Definition 3.3.1

Die **Varianz** von X mit Mittelwert μ ist $\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2]$. Die **Standardabweichung** ist $\sigma = \sqrt{\mathrm{Var}(X)}$.

Bemerkung 3.3.1 (Wann existiert die Varianz?)

Die Varianz ist *genau dann* definiert und endlich, wenn X ein endliches zweites Moment besitzt:

$$\mathbb{E}[X^2] < \infty.$$

Die Menge dieser Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ heißt $L^2(\Omega, \mathcal{A}, \mathbb{P})$ – **auf L^2, und nur dort, ist die Varianz berechenbar.** Ein einziges Kriterium genügt, denn wegen $|X| \le 1 + X^2$ gilt die Inklusionskette $L^2 \subset L^1$:

$$\mathbb{E}[X^2] < \infty \implies \mathbb{E}[|X|] < \infty \implies \mu = \mathbb{E}[X] \text{ existiert endlich.}$$

Das endliche zweite Moment liefert also den Mittelwert μ *und* die Endlichkeit von $\mathbb{E}[(X-\mu)^2]$ zugleich. Praktisch prüft man

$$\sum_x x^2 f(x) < \infty \quad \text{(diskret)} \qquad \text{bzw.} \qquad \int_{-\infty}^{\infty} x^2 f(x)\,dx < \infty \quad \text{(stetig).}$$

Wo es scheitert: Bei der **Cauchy-Verteilung** existiert schon μ nicht ($\mathbb{E}[|X|] = \infty$, siehe oben), also erst recht keine Varianz. Bei der t_ν-Verteilung ist die Varianz für $1 < \nu \le 2$ unendlich (und μ fehlt für $\nu \le 1$); bei der Pareto-Verteilung mit Index $\alpha \le 2$ ebenfalls unendlich.
Auf welcher Menge berechne ich die Varianz? Der Erwartungswert ist ein Integral über *ganz* Ω ($\mathbb{E}[\cdot] = \int_\Omega \cdot\, d\mathbb{P}$); man berechnet die Varianz nicht auf einer Teilmenge. Die Streuung eingeschränkt auf ein Ereignis A mit $\mathbb{P}(A) > 0$ ist die *bedingte* Varianz $\mathrm{Var}(X \mid A)$, gebildet bezüglich $\mathbb{P}(\cdot \mid A)$.

Die Varianz misst die *mittlere quadratische Abweichung* vom Erwartungswert und damit die **Streuung** der Verteilung um μ. Drei Beobachtungen machen die Definition $\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2]$ präzise:

- **Warum quadriert?** Die mittlere *vorzeichenbehaftete* Abweichung trägt keine Information, denn

  $$\mathbb{E}[X-\mu] = \mathbb{E}[X] - \mu = 0$$

  – positive und negative Abweichungen heben sich exakt auf. Das Quadrat beseitigt das Vorzeichen und gewichtet große Abweichungen überproportional stark.

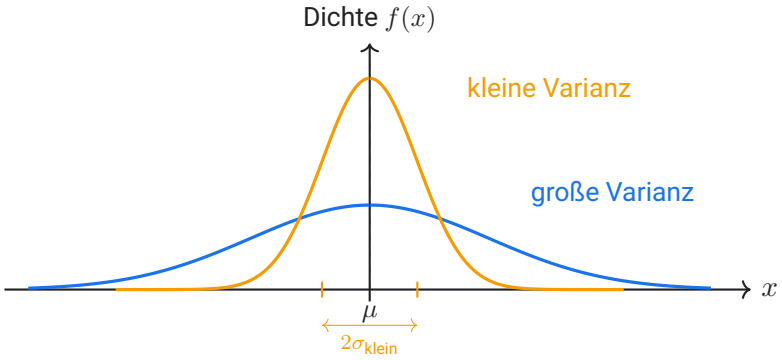
- **Minimaleigenschaft des Mittelwerts.** Unter allen Konstanten c wird die mittlere quadratische Abweichung $\mathbb{E}[(X-c)^2]$ *genau* bei $c=\mu$ minimal, und der Minimalwert ist $\operatorname{Var}(X)$. Denn

$$\mathbb{E}[(X-c)^2] = \mathbb{E}\left[((X-\mu)+(\mu-c))^2\right] = \operatorname{Var}(X) + (\mu-c)^2 \;\geq\; \operatorname{Var}(X),$$

  mit Gleichheit nur für $c=\mu$ (der gemischte Term verschwindet wegen $\mathbb{E}[X-\mu]=0$). Der Erwartungswert ist also der beste konstante Vorhersagewert im quadratischen Sinn.

- **Einheiten.** $\operatorname{Var}(X)$ trägt die *quadrierte* Einheit von X (z. B. m^2). Die Standardabweichung $\sigma = \sqrt{\operatorname{Var}(X)}$ hat dieselbe Einheit wie X und ist daher die direkt interpretierbare „typische“ Abweichung vom Mittelwert.

Zwei Verteilungen können denselben Erwartungswert besitzen und sich trotzdem stark in der Varianz unterscheiden. Die folgende Abbildung zeigt zwei Dichten mit *identischem* μ und gleicher Gesamtfläche 1: Die schmale, hohe Kurve hat kleine Varianz (die Werte konzentrieren sich nahe bei μ), die breite, flache Kurve große Varianz.



Beide Kurven haben denselben Mittelwert μ und dieselbe Fläche 1. Kleine Varianz bedeutet eine schmale, hohe Dichte (Werte nah bei μ); große Varianz eine breite, flache Dichte (Werte weit gestreut).

Satz 3.3.2

1. $\operatorname{Var}(X) = \mathbb{E}[X^2] - [\mathbb{E}[X]]^2$ (Verschiebungssatz),
2. $\operatorname{Var}(aX+b) = a^2 \operatorname{Var}(X)$,
3. $\operatorname{Var}(X) \geq 0$.

Beweis. Wir setzen voraus, dass $X \in L^2$ ist, also $\mathbb{E}[X^2] < \infty$. Dann ist auch $\mathbb{E}[|X|] < \infty$, der Erwartungswert

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

existiert endlich, und alle folgenden Ausdrücke sind wohldefiniert.

1. Verschiebungssatz. Aus der Definition der Varianz folgt

$$\operatorname{Var}(X) = \mathbb{E}[(X-\mu)^2].$$

Wir entwickeln das Quadrat:

$$(X-\mu)^2 = X^2 - 2\mu X + \mu^2.$$

Mit Linearität des Erwartungswerts erhalten wir

$$\mathrm{Var}(X) = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\,\mathbb{E}[X] + \mu^2.$$

Da $\mu = \mathbb{E}[X]$ gilt,

$$-2\mu\,\mathbb{E}[X] + \mu^2 = -2\mu^2 + \mu^2 = -\mu^2.$$

Also

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu^2 = \mathbb{E}[X^2] - [\mathbb{E}[X]]^2.$$

Das ist der Verschiebungssatz: Statt die mittlere quadratische Abweichung direkt zu berechnen, darf man das zweite Moment minus Quadrat des ersten Moments nehmen.

2. Lineare Transformation. Sei

$$Y = aX + b.$$

Dann ist

$$\mathbb{E}[Y] = \mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b = a\mu + b.$$

Die zentrierte Zufallsvariable ist daher

$$Y - \mathbb{E}[Y] = (aX + b) - (a\mu + b) = a(X - \mu).$$

Also

$$\mathrm{Var}(aX + b) = \mathrm{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}\left[a^2(X - \mu)^2\right].$$

Da a^2 konstant ist,

$$\mathrm{Var}(aX + b) = a^2\,\mathbb{E}[(X - \mu)^2] = a^2\,\mathrm{Var}(X).$$

Die Konstante b verschiebt die Verteilung nur nach links oder rechts; sie ändert die Streuung nicht. Der Faktor a skaliert Abstände, deshalb skaliert die Varianz mit a^2.

3. Nichtnegativität. Für jedes $\omega \in \Omega$ gilt

$$(X(\omega) - \mu)^2 \geq 0.$$

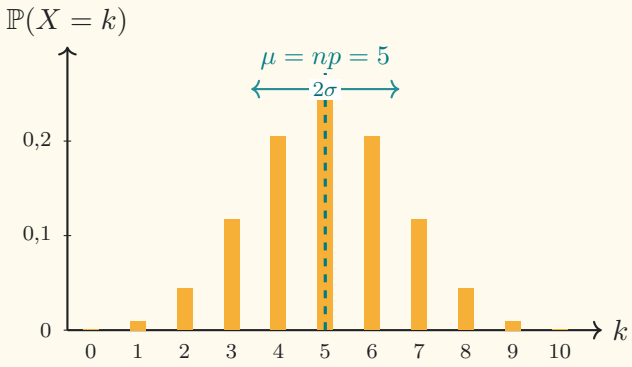
Der Erwartungswert einer nichtnegativen Zufallsvariablen ist nichtnegativ. Daher

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2] \geq 0.$$

Gleichheit gilt genau dann, wenn $(X - \mu)^2 = 0$ fast sicher, also $X = \mu$ fast sicher. In diesem Fall streut X gar nicht. ■

Beispiel 3.3.1

$X \sim$ Binomial(n, p): $\mathrm{Var}(X_i) = p(1-p)$. Da die X_i unabhängig: $\mathrm{Var}(X) = np(1-p)$.



Wahrscheinlichkeitsfunktion einer $\mathrm{Binomial}(n, p)$*-Variable (hier* $n = 10$*,* $p = 0{,}5$*). Erwartungswert* $\mu = np = 5$*, Varianz* $\mathrm{Var}(X) = np(1-p) = 2{,}5$ *und damit Standardabweichung* $\sigma = \sqrt{np(1-p)} \approx 1{,}58$*. Der eingezeichnete Bereich* $[\mu - \sigma,\ \mu + \sigma]$ *hat Breite* 2σ *und enthält den Großteil der Wahrscheinlichkeitsmasse – je kleiner* $p(1-p)$*, desto schmaler dieser Bereich.*

Satz 3.3.3

Sind $X_1, \ldots, X_n$ iid mit $\mu = \mathbb{E}[X_i]$, $\sigma^2 = \mathrm{Var}(X_i)$ und $\overline{X}_n = \frac{1}{n}\sum X_i$, so ist

$$\mathbb{E}[\overline{X}_n] = \mu, \quad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Was bedeutet iid?

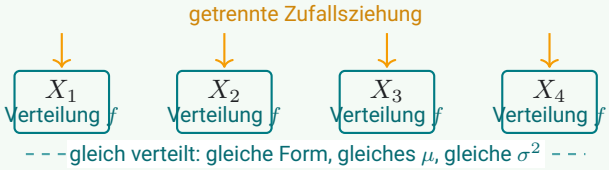
Die Abkürzung **iid** steht für *independent and identically distributed*, also:

$$X_1, \ldots, X_n \text{ iid} \iff \begin{cases} \text{alle } X_i \text{ sind unabhängig,} \\ \text{alle } X_i \text{ haben dieselbe Verteilung.} \end{cases}$$

Gleiche Verteilung bedeutet: Jede Beobachtung entsteht nach demselben Zufallsmechanismus. Deshalb haben alle X_i denselben Erwartungswert μ und dieselbe Varianz σ^2.
Unabhängigkeit bedeutet: Eine Beobachtung verrät nichts über die nächste. Deshalb faktorisiert der Erwartungswert von Produkten, und die gemischten Terme verschwinden:

$$i \neq j \implies \mathbb{E}[(X_i - \mu)(X_j - \mu)] = \mathbb{E}[X_i - \mu]\ \mathbb{E}[X_j - \mu] = 0.$$

Anschaulich: $X_1, \ldots, X_n$ sind wiederholte Messungen unter denselben Bedingungen, ohne gegenseitige Beeinflussung. Genau dann mittelt $\overline{X}_n$ denselben Zufall n-mal; die Lage bleibt μ, aber die Streuung schrumpft auf σ^2/n.



Beweis. Mit Linearität des Erwartungswerts:

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\mu = \mu.$$

Für die Varianz rechnen wir direkt mit der Definition. Wegen $\overline{X}_n - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$ ist

$$\mathrm{Var}(\overline{X}_n) = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[(X_i-\mu)(X_j-\mu)].$$

Für $i \neq j$ sind $X_i - \mu$ und $X_j - \mu$ unabhängig; nach Satz 3.2.2 faktorisiert der Erwartungswert, und es bleibt $\mathbb{E}[X_i-\mu]\,\mathbb{E}[X_j-\mu] = 0\cdot 0 = 0$. Von der Doppelsumme überleben also nur die n Diagonalterme $\mathbb{E}[(X_i-\mu)^2] = \sigma^2$:

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

■

Die Varianz des Stichprobenmittels schrumpft mit $1/n$ – das ist der mathematische Kern von „mehr Daten = mehr Präzision“.

Definition 3.3.4

Die **Kovarianz** misst, ob zwei Zufallsvariablen gemeinsam nach oben oder nach unten von ihren Mittelwerten abweichen. Sie ist definiert als

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y.$$

Bemerkung 3.3.2 (Rückblick auf das Stichprobenmittel)

Die gemischten Terme $\mathbb{E}[(X_i-\mu)(X_j-\mu)]$, die im Beweis von $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ verschwanden, sind genau die Kovarianzen $\mathrm{Cov}(X_i, X_j)$. Jene Rechnung ist damit rückblickend der Spezialfall von Satz 3.3.7 für unkorrelierte Summanden – dort wurde sie nur noch ohne den Begriff geführt, der erst hier zur Verfügung steht.

Sinn der Kovarianz

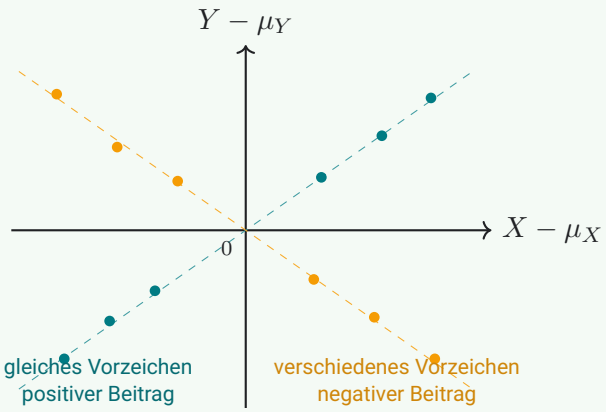
Die Varianz betrachtet eine Variable allein:

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu_X)^2].$$

Sie fragt: *Wie stark streut X um seinen Mittelwert?*
Die Kovarianz betrachtet zwei Variablen gleichzeitig:

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)].$$

Sie fragt: *Weichen X und Y typischerweise gemeinsam in dieselbe Richtung ab?*



Liegt X über seinem Mittelwert und liegt gleichzeitig auch Y über seinem Mittelwert, ist das Produkt $(X - \mu_X)(Y - \mu_Y)$ positiv. Dasselbe gilt, wenn beide unter ihren Mittelwerten liegen. Solche Beobachtungen sprechen für **positive Kovarianz**.
Liegt dagegen X über seinem Mittelwert, während Y unter seinem Mittelwert liegt, ist das Produkt negativ. Solche Beobachtungen sprechen für **negative Kovarianz**. Heben sich positive und negative Beiträge im Mittel auf, ist die Kovarianz nahe 0.

Satz 3.3.5

1. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$,
2. $X \perp Y \Rightarrow \mathrm{Cov}(X, Y) = 0$ (Umkehrung gilt i. A. nicht!),
3. Cov ist symmetrisch, bilinear, und $\mathrm{Cov}(X, a) = 0$.

Beweis. Wir verwenden durchgehend die Darstellung

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

1. Für $Y = X$ folgt sofort

$$\mathrm{Cov}(X, X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathrm{Var}(X).$$

2. Sind X und Y unabhängig, dann gilt $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. Daher

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.$$

Die Umkehrung gilt im Allgemeinen nicht: Aus $\mathrm{Cov}(X, Y) = 0$ folgt nur, dass der lineare Zusammenhang verschwindet, nicht aber Unabhängigkeit.

3. Symmetrie. Da $XY = YX$, gilt

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

Bilinearität. Für Konstanten $\alpha, \beta \in \mathbb{R}$ gilt

$$\begin{aligned}\mathrm{Cov}(\alpha X + \beta Z, Y) &= \mathbb{E}[(\alpha X + \beta Z)Y] - \mathbb{E}[\alpha X + \beta Z]\,\mathbb{E}[Y] \\ &= \alpha\,\mathbb{E}[XY] + \beta\,\mathbb{E}[ZY] - (\alpha\,\mathbb{E}[X] + \beta\,\mathbb{E}[Z])\,\mathbb{E}[Y] \\ &= \alpha\,\mathrm{Cov}(X, Y) + \beta\,\mathrm{Cov}(Z, Y).\end{aligned}$$

Wegen Symmetrie folgt dieselbe Linearität auch im zweiten Argument.

Ist a konstant, dann ist $\mathbb{E}[a] = a$ und somit

$$\operatorname{Cov}(X, a) = \mathbb{E}[Xa] - \mathbb{E}[X]\,\mathbb{E}[a] = a\,\mathbb{E}[X] - a\,\mathbb{E}[X] = 0.$$

■

Satz 3.3.6 (Varianz einer Linearkombination)

Seien $X_1, \ldots, X_n \in L^2$ und $a_1, \ldots, a_n \in \mathbb{R}$. Dann gilt

$$\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{Cov}(X_i, X_j).$$

Sind die X_i paarweise unkorreliert – insbesondere, wenn sie unabhängig sind –, reduziert sich die Formel zu

$$\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}(X_i).$$

Beweis. Setze $Z = \sum_{i=1}^{n} a_i X_i$. Mit $\operatorname{Var}(Z) = \operatorname{Cov}(Z, Z)$ und der Bilinearität der Kovarianz folgt

$$\operatorname{Var}(Z) = \operatorname{Cov}\left(\sum_i a_i X_i, \sum_j a_j X_j\right) = \sum_i \sum_j a_i a_j \operatorname{Cov}(X_i, X_j).$$

Im unkorrelierten Fall verschwinden alle Summanden mit $i \neq j$. ■

Satz 3.3.7 (Varianz einer Summe)

Sind $X, Y \in L^2$, so gilt

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

Beweis. Nach Punkt 1 des vorigen Satzes ist die Varianz die Kovarianz einer Zufallsvariable mit sich selbst, $\operatorname{Var}(Z) = \operatorname{Cov}(Z, Z)$. Angewandt auf $Z = X \pm Y$ und ausmultipliziert mit der Bilinearität (Punkt 3):

$$\begin{aligned}\operatorname{Var}(X \pm Y) &= \operatorname{Cov}(X \pm Y,\ X \pm Y)\\ &= \operatorname{Cov}(X, X) \pm \operatorname{Cov}(X, Y) \pm \operatorname{Cov}(Y, X) + \operatorname{Cov}(Y, Y)\\ &= \operatorname{Var}(X) + \operatorname{Var}(Y) \pm 2\operatorname{Cov}(X, Y).\end{aligned}$$

Im zweiten Schritt liefert das gemischte Vorzeichen zweimal denselben Beitrag $\pm\operatorname{Cov}(X, Y)$, im letzten Term dagegen $(\pm 1)\cdot(\pm 1) = +1$ – deshalb steht vor $\operatorname{Var}(Y)$ immer ein Plus. Der Schritt von der dritten zur letzten Zeile nutzt die Symmetrie $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$.

Ist $\operatorname{Cov}(X, Y) = 0$ – etwa weil X und Y unabhängig sind –, so fällt der gemischte Term weg und die Varianzen addieren sich einfach. Genau dieser Schritt steckt hinter $\operatorname{Var}(\overline{X}_n) = \sigma^2/n$. ■

Definition 3.3.8

Sind $\operatorname{Var}(X) > 0$ und $\operatorname{Var}(Y) > 0$, so heißt

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\,\operatorname{Var}(Y)}} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

die **Korrelation** von X und Y.

Lemma 3.3.9 (Korrelationsschranke)

Für $X, Y \in L^2$ mit $\sigma_X, \sigma_Y > 0$ gilt

$$-1 \le \rho(X,Y) \le 1.$$

Gleichheit tritt genau dann ein, wenn Y fast sicher affin von X abhängt: $\rho = +1$ bei $Y = aX + b$ mit $a > 0$, und $\rho = -1$ bei $a < 0$.

Beweis. **Schritt 1: Standardisieren.** Wir setzen

$$U = \frac{X}{\sigma_X}, \qquad V = \frac{Y}{\sigma_Y}.$$

Aus $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$ folgt $\operatorname{Var}(U) = \sigma_X^2/\sigma_X^2 = 1$ und ebenso $\operatorname{Var}(V) = 1$. Die Bilinearität der Kovarianz liefert

$$\operatorname{Cov}(U,V) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} = \rho.$$

Schritt 2: Nichtnegativität ausnutzen. Satz 3.3.7, angewandt auf U und V, ergibt

$$\operatorname{Var}(U \pm V) = 1 + 1 \pm 2\rho = 2(1 \pm \rho).$$

Varianzen sind nie negativ, also ist $1 + \rho \ge 0$ und $1 - \rho \ge 0$ – zusammen genau $-1 \le \rho \le 1$.

Schritt 3: Der Gleichheitsfall. $\rho = 1$ bedeutet $\operatorname{Var}(U - V) = 0$. Aus $\mathbb{E}[(Z - \mathbb{E}[Z])^2] = 0$ und $(Z - \mathbb{E}[Z])^2 \ge 0$ folgt $Z = \mathbb{E}[Z]$ fast sicher; die Differenz $U - V$ ist also f.s. eine Konstante c. Einsetzen und Auflösen nach Y:

$$\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = c \quad \Longleftrightarrow \quad Y = \frac{\sigma_Y}{\sigma_X} X - c\,\sigma_Y,$$

eine affine Funktion mit positiver Steigung. Der Fall $\rho = -1$ verläuft analog über $\operatorname{Var}(U + V) = 0$ und liefert negative Steigung.

Umgekehrt sei $Y = aX + b$ mit $a \neq 0$. Dann ist $\operatorname{Cov}(X,Y) = a \operatorname{Var}(X)$ und $\sigma_Y = |a|\sigma_X$, also

$$\rho = \frac{a\,\sigma_X^2}{\sigma_X \cdot |a| \sigma_X} = \frac{a}{|a|} = \pm 1.$$

∎

Bemerkung 3.3.3 (Zwei Wege zur selben Schranke)

Die Korrelationsschranke ist ein Spezialfall der Cauchy-Schwarz-Ungleichung (Satz 4.2.1 in Kapitel 4), die $|\operatorname{Cov}(X,Y)| \le \sigma_X \sigma_Y$ direkt liefert. Der Beweis hier kommt ohne dieses schwerere Werkzeug aus: Er braucht nur $\operatorname{Var} \ge 0$ und die Rechenregel für $\operatorname{Var}(X \pm Y)$ von der vorigen Seite.

3.4 Erwartungswert und Varianz wichtiger Verteilungen

Verteilung	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
Punktmasse an a	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometrisch(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ,σ^2)	μ	σ^2
Exp(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
t_ν	0 $(\nu>1)$	$\frac{\nu}{\nu-2}$ $(\nu>2)$
χ^2_p	p	$2p$

Lemma 3.4.1 (Multivariate Varianz)

Sei $X=(X_1,\dots,X_k)^\top$ ein Zufallsvektor mit endlichen zweiten Momenten, Erwartungswertvektor μ und Kovarianzmatrix Σ. Für jedes $a\in\mathbb{R}^k$ gilt

$$\mathbb{E}[a^\top X]=a^\top\mu, \qquad \mathrm{Var}(a^\top X)=a^\top\Sigma a.$$

Beweis. Die erste Formel ist die Linearität des Erwartungswerts. Für die zweite wenden wir Satz 3.3.6 an:

$$\mathrm{Var}(a^\top X)=\sum_{i=1}^{k}\sum_{j=1}^{k}a_i a_j\,\mathrm{Cov}(X_i,X_j)=a^\top\Sigma a.$$

■

Bemerkung 3.4.1 (Strukturelle Beziehungen wichtiger Verteilungen)

Die Tabelle wird durch einige besonders nützliche Beziehungen ergänzt:

- Eine Binomialvariable ist die Summe unabhängiger Bernoulli-Variablen; bei festem $\lambda=np$ konvergiert $\mathrm{Binomial}(n,\lambda/n)$ gegen $\mathrm{Poisson}(\lambda)$.
- Die geometrische Verteilung ist gedächtnislos: $\mathbb{P}(X>n+m\mid X>n)=\mathbb{P}(X>m)$. Dasselbe stetige Prinzip besitzt die Exponentialverteilung.
- Sind $X_i\sim N(\mu_i,\sigma_i^2)$ unabhängig, so gilt

$$\sum_i a_iX_i\sim N\left(\sum_i a_i\mu_i,\ \sum_i a_i^2\sigma_i^2\right).$$

- $\mathrm{Exp}(\beta)=\mathrm{Gamma}(1,\beta)$ und $\mathrm{Uniform}(0,1)=\mathrm{Beta}(1,1)$.

- Für unabhängige $Z_1, \dots, Z_p \sim N(0,1)$ gilt $\sum_{i=1}^{p} Z_i^2 \sim \chi_p^2$; außerdem konvergiert t_ν für $\nu \to \infty$ in Verteilung gegen $N(0,1)$.

3.5 Bedingter Erwartungswert

Definition 3.5.1

Der **bedingte Erwartungswert** von X gegeben $Y = y$ ist

$$\mathbb{E}[X \mid Y = y] = \begin{cases} \sum_x x\, f_{X|Y}(x \mid y) & \text{falls } X \text{ diskret,} \\ \int x\, f_{X|Y}(x \mid y)\, dx & \text{falls } X \text{ stetig,} \end{cases}$$

mit der bedingten Dichte bzw. pmf $f_{X|Y}(x \mid y) = f_{X,Y}(x,y)/f_Y(y)$ aus Abschnitt 2.8, vorausgesetzt $f_Y(y) > 0$.

Definiere $g(y) = \mathbb{E}[X \mid Y = y]$. Dann ist $\mathbb{E}[X \mid Y] = g(Y)$ eine Zufallsvariable.

Bemerkung 3.5.1 (Idee des bedingten Erwartungswerts)

Der bedingte Erwartungswert $\mathbb{E}[X \mid Y = y]$ ist der **mittlere Wert von X, wenn wir bereits wissen, dass Y den Wert y angenommen hat.** Es ist dieselbe Konstruktion wie beim gewöhnlichen Erwartungswert $\mathbb{E}[X] = \int x f_X(x)\, dx$ – Wert mal Gewicht, integriert –; der *einzige* Unterschied liegt in der Gewichtung:

$$\mathbb{E}[X] \text{ gewichtet mit } f_X(x), \qquad \mathbb{E}[X \mid Y = y] \text{ gewichtet mit } f_{X|Y}(x \mid y).$$

Die Information „$Y = y$“ verändert also nur die *Gewichte*: Statt über alle möglichen Welten zu mitteln, mitteln wir nur noch über die Teilwelt, in der $Y = y$ gilt.
Anschaulich: Man stelle sich die gemeinsame Dichte $f_{X,Y}$ als Gebirge über der (x,y)-Ebene vor. Ein senkrechter Schnitt bei festem y liefert das Profil $x \mapsto f_{X,Y}(x,y)$; dieses hat i. A. nicht die Fläche 1, also normiert man durch $f_Y(y)$ und erhält die bedingte Dichte $f_{X|Y}(x \mid y) = f_{X,Y}(x,y)/f_Y(y)$. Deren *Schwerpunkt* auf der x-Achse ist gerade $\mathbb{E}[X \mid Y = y]$. Der Nenner $f_Y(y)$ ist dabei nur die Normierung, die $\int f_{X|Y}(x \mid y)\, dx = 1$ sicherstellt.
Zahl vs. Zufallsvariable: Für festes y ist $\mathbb{E}[X \mid Y = y]$ eine *Zahl*. Lässt man y variieren, entsteht die Funktion $g(y) = \mathbb{E}[X \mid Y = y]$, und $\mathbb{E}[X \mid Y] = g(Y)$ ist selbst eine *Zufallsvariable*. Dies ist die Brücke zur Turm-Eigenschaft $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$: Mittelt man die bedingten Mittelwerte über alle y (gewichtet mit f_Y), erhält man den Gesamtmittelwert zurück – man berechnet $\mathbb{E}[X]$ „in zwei Etappen“.

Maßtheoretischer Einschub: Der Satz von Radon-Nikodým

Die formale Existenz von $\mathbb{E}[X \mid Y]$ steht und fällt mit einem zentralen Resultat der Maßtheorie. Wir führen ihn hier explizit ein. Maß, Maßraum und σ-Endlichkeit stehen in Definition 1.8.3; einen Begriff brauchen wir zusätzlich, weil das gleich konstruierte ν auch negative Werte annehmen kann.

Definition 3.5.2 (Signiertes Maß)

Ein **signiertes Maß** auf $(\Omega, \mathcal{F})$ ist eine Funktion

$$\nu \colon \mathcal{F} \to \mathbb{R} \quad \text{mit} \quad \nu(\emptyset) = 0,$$

für die bei paarweise disjunkten $A_1, A_2, \ldots \in \mathcal{F}$

$$\nu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \nu(A_i)$$

gilt, wobei die Reihe *absolut* konvergieren muss.
Gegenüber Definition 1.8.3 ist der Wertebereich geändert: negative Werte sind erlaubt, dafür sind $\pm\infty$ ausgeschlossen. Die absolute Konvergenz ist gegenüber Definition 1.8.2 die einzige Zusatzforderung; sie ist nötig, weil die Vereinigung links nicht von der Reihenfolge der A_i abhängt, eine nur bedingt konvergente Reihe rechts aber sehr wohl. Jedes endliche Maß ist ein signiertes Maß.

Definition 3.5.3 (Absolutstetigkeit)

Seien μ, ν Maße auf $(\Omega, \mathcal{F})$. ν heißt **absolutstetig bezüglich** μ, geschrieben $\nu \ll \mu$, falls

$$\mu(A) = 0 \;\Rightarrow\; \nu(A) = 0 \qquad \forall A \in \mathcal{F}.$$

Satz 3.5.4 (Radon-Nikodým)

Sei $(\Omega, \mathcal{F})$ ein messbarer Raum, μ ein σ-endliches Maß und ν ein σ-endliches signiertes Maß auf $\mathcal{F}$ mit $\nu \ll \mu$. Dann existiert eine $\mathcal{F}$-messbare Funktion $f : \Omega \to \mathbb{R}$ mit

$$\nu(A) = \int_A f \, d\mu \qquad \forall A \in \mathcal{F}.$$

f ist μ-fast überall eindeutig und wird als **Radon-Nikodým-Ableitung** bezeichnet:

$$f = \frac{d\nu}{d\mu}.$$

Beweisskizze. Der Beweis verwendet typischerweise einen der folgenden Ansätze:

(a) Hilbertraum-Ansatz (von Neumann): Für endliche Maße betrachte das Funktional $L(g) = \int g \, d\nu$ auf $L^2(\Omega, \mathcal{F}, \mu + \nu)$. Nach dem Rieszschen Darstellungssatz existiert $h \in L^2$ mit $L(g) = \int g h \, d(\mu + \nu)$. Aus der Beziehung $\int g(1-h)\, d\nu = \int gh \, d\mu$ folgt $f = h/(1-h)$ mit den gewünschten Eigenschaften. Die σ-Endlichkeit liefert die Verallgemeinerung durch Ausschöpfung.

(b) Hahn-Jordan-Zerlegung: Für signierte Maße zerlege $\nu = \nu^+ - \nu^-$ und wende den positiven Fall an.

Für den vollständigen Beweis siehe Billingsley, *Probability and Measure*, §32, oder Rudin, *Real and Complex Analysis*, Kap. 6. ∎

Bemerkung 3.5.2 (Anwendung auf die bedingte Erwartung)

Sei X integrierbar und $\mathcal{G} \subseteq \mathcal{F}$ eine Unter-σ-Algebra. Definiere auf $(\Omega, \mathcal{G})$ das signierte Maß

$$\nu(A) := \int_A X \, d\mathbb{P}, \qquad A \in \mathcal{G}.$$

Aus $\mathbb{P}(A) = 0$ folgt $\nu(A) = 0$, also $\nu \ll \mathbb{P}|_{\mathcal{G}}$. Nach Radon-Nikodým (Theorem 3.5.4) existiert eine $\mathcal{G}$-messbare Funktion Z mit $\int_A Z \, d\mathbb{P} = \int_A X \, d\mathbb{P}$ für alle $A \in \mathcal{G}$. Diese Funktion wird per Definition als bedingte Erwartung bezeichnet:

$$\mathbb{E}[X \mid \mathcal{G}] := \frac{d\nu}{d\mathbb{P}|_{\mathcal{G}}}.$$

Für $\mathcal{G} = \sigma(Y)$ ergibt sich $\mathbb{E}[X \mid Y]$.

Zurück zur bedingten Erwartung

Satz 3.5.5 ($\mathbb{E}[X \mid Y] = g(Y)$ ist eine Zufallsvariable)

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, X integrierbar ($\mathbb{E}[|X|] < \infty$) und Y eine Zufallsvariable. Dann existiert eine Borel-messbare Funktion $g : \mathbb{R} \to \mathbb{R}$ mit

$$\mathbb{E}[X \mid Y] = g(Y) \quad \mathbb{P}\text{-f.s.},$$

und $g(Y)$ ist eine Zufallsvariable auf $(\Omega, \mathcal{F})$.

Beweis. **(1) Definition.** $\mathbb{E}[X \mid Y] := \mathbb{E}[X \mid \sigma(Y)]$ mit $\sigma(Y) = \{Y^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$. Nach Theorem 3.5.4 existiert eine $\sigma(Y)$-messbare Funktion Z mit $\int_A Z \, d\mathbb{P} = \int_A X \, d\mathbb{P}$ für alle $A \in \sigma(Y)$, $\mathbb{P}$-f.s. eindeutig.

(2) Faktorisierungslemma (Doob-Dynkin). Jede $\sigma(Y)$-messbare Funktion Z besitzt eine Darstellung $Z = g \circ Y$ mit Borel-messbarem g. Beweis per maßtheoretischer Induktion: *Indikatoren:* $A = Y^{-1}(B) \Rightarrow I_A = I_B \circ Y$. *Einfache Funktionen:* Linearkombinationen übertragen sich. *Nichtnegative Funktionen:* Approximation $Z_n \uparrow Z$ mit $Z_n = g_n(Y)$; setze $g := \limsup_n g_n$, Borel-messbar. *Allgemeiner Fall:* Zerlegung $Z = Z^+ - Z^-$.

(3) Anwendung. Da $\mathbb{E}[X \mid Y]$ nach (1) $\sigma(Y)$-messbar ist, liefert (2) ein Borel-messbares g mit $\mathbb{E}[X \mid Y] = g(Y)$ $\mathbb{P}$-f.s.

(4) Messbarkeit. Für $B \in \mathcal{B}(\mathbb{R})$ gilt $(g \circ Y)^{-1}(B) = Y^{-1}(g^{-1}(B))$. Mit g Borel-messbar folgt $g^{-1}(B) \in \mathcal{B}(\mathbb{R})$, und mit Y Zufallsvariable folgt $Y^{-1}(g^{-1}(B)) \in \mathcal{F}$. Also ist $g(Y)$ eine Zufallsvariable. ∎

Die Funktion g heißt **Regressionsfunktion** und ist $\mathbb{P}_Y$-fast überall eindeutig.

Lemma 3.5.6 (Rechenregeln für bedingte Erwartungen)

Seien U und V integrierbar, $a, b \in \mathbb{R}$, und sei $h : \mathbb{R} \to \mathbb{R}$ Borel-messbar. Dann gelten, soweit die auftretenden Produkte integrierbar sind:

1. **Linearität:**
$$\mathbb{E}[aU + bV \mid Y] = a\,\mathbb{E}[U \mid Y] + b\,\mathbb{E}[V \mid Y].$$
2. **Unabhängigkeit:** Ist $U \perp Y$, so ist
$$\mathbb{E}[U \mid Y] = \mathbb{E}[U] \qquad \mathbb{P}\text{-f.s.}$$
3. **Herausziehen bekannter Faktoren:**
$$\mathbb{E}[h(Y)U \mid Y] = h(Y)\,\mathbb{E}[U \mid Y].$$

Beweis. Alle rechten Seiten sind $\sigma(Y)$-messbar. Die Aussagen folgen aus der definierenden Integralgleichheit der bedingten Erwartung. Für $A \in \sigma(Y)$ gilt etwa
$$\int_A \big(a\,\mathbb{E}[U \mid Y] + b\,\mathbb{E}[V \mid Y]\big)\,d\mathbb{P} = a\int_A U\,d\mathbb{P} + b\int_A V\,d\mathbb{P} = \int_A (aU + bV)\,d\mathbb{P}.$$
Ist $U \perp Y$, so erfüllt die konstante Zufallsvariable $\mathbb{E}[U]$ dieselbe Gleichheit, denn $\mathbb{E}[UI_A] = \mathbb{E}[U]\mathbb{P}(A)$ für $A \in \sigma(Y)$. Beim dritten Punkt ist $h(Y)$ bereits durch die Bedingung bekannt; die Gleichheit folgt zunächst für Indikatoren und einfache Funktionen h und danach durch Approximation für integrierbare $h(Y)U$. Die fast sichere Eindeutigkeit der bedingten Erwartung liefert jeweils die Behauptung. ∎

Satz 3.5.7 (Iterierte Erwartungswerte, Turm-Eigenschaft)

Ist Y integrierbar, also $\mathbb{E}[|Y|] < \infty$, so gilt
$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y].$$

Erst innerhalb der Gruppen mitteln, dann über die Gruppen – das Ergebnis ist der Gesamtmittelwert.

Beweis. **Diskreter Fall (die Mechanik).** Sei $g(x) = \mathbb{E}[Y \mid X = x] = \sum_y y\,\mathbb{P}(Y = y \mid X = x)$. Der äußere Erwartungswert mittelt $g(X)$ über die Verteilung von X:
$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y \mid X]] &= \sum_x g(x)\,\mathbb{P}(X = x) = \sum_x \Big(\sum_y y\,\mathbb{P}(Y = y \mid X = x)\Big)\mathbb{P}(X = x)\\ &= \sum_x \sum_y y\,\underbrace{\mathbb{P}(Y = y \mid X = x)\,\mathbb{P}(X = x)}_{=\mathbb{P}(X=x,\,Y=y)} = \sum_y y \sum_x \mathbb{P}(X = x,\,Y = y)\\ &= \sum_y y\,\mathbb{P}(Y = y) = \mathbb{E}[Y].\end{aligned}$$
Der Kern ist der Übergang von der bedingten zur gemeinsamen Wahrscheinlichkeit (Produktregel) und danach das Aufsummieren über x, das die Randverteilung von Y zurückgibt. Das Vertauschen der Summationsreihenfolge ist erlaubt, weil wegen $\mathbb{E}[|Y|] < \infty$ alles absolut konvergiert.

Stetiger Fall. Wortgleich mit Dichten statt Massefunktionen; aus $f_{Y|X}(y \mid x) f_X(x) = f_{X,Y}(x, y)$ folgt
$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} y\,f_{Y|X}(y \mid x)\,dy\Big) f_X(x)\,dx = \int_{\mathbb{R}} y \underbrace{\int_{\mathbb{R}} f_{X,Y}(x, y)\,dx}_{=f_Y(y)}\,dy = \mathbb{E}[Y],$$

wobei das Vertauschen der Integrationsreihenfolge durch Fubini–Tonelli gerechtfertigt ist.

Allgemeiner Fall (in einer Zeile). Maßtheoretisch ist $Z = \mathbb{E}[Y \mid X]$ per Definition diejenige $\sigma(X)$-messbare Zufallsvariable mit

$$\int_A Z \, d\mathbb{P} = \int_A Y \, d\mathbb{P} \qquad \text{für alle } A \in \sigma(X).$$

Da $\Omega \in \sigma(X)$, darf man $A = \Omega$ wählen – und das ist bereits die Behauptung:

$$\mathbb{E}[Z] = \int_\Omega Z \, d\mathbb{P} = \int_\Omega Y \, d\mathbb{P} = \mathbb{E}[Y].$$

Die Turm-Eigenschaft ist also kein Zusatzresultat, sondern der Spezialfall der definierenden Eigenschaft für das größtmögliche Ereignis. ■

Beispiel 3.5.1 (Gleichverteilung unter der Diagonalen)

Sei $X \sim \mathrm{Uniform}(0,1)$ und, nach Beobachtung von $X = x$,

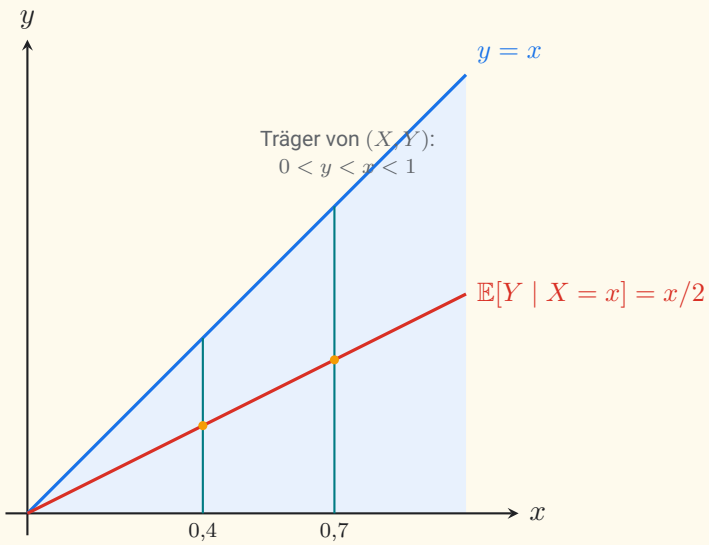
$$Y \mid X = x \sim \mathrm{Uniform}(0, x).$$

Für $0 < x < 1$ ist die bedingte Dichte $f_{Y|X}(y \mid x) = 1/x$ auf $0 < y < x$. Daher

$$\mathbb{E}[Y \mid X = x] = \int_0^x y \frac{1}{x} \, dy = \frac{x}{2}, \qquad \mathbb{E}[Y \mid X] = \frac{X}{2}.$$

Die Turm-Eigenschaft liefert sofort

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{X}{2}\right] = \frac{1}{2}\,\mathbb{E}[X] = \frac{1}{4}.$$



Die senkrechten cyanfarbenen Schnitte zeigen die bedingten Gleichverteilungen; ihre Mittelpunkte liegen auf der roten Regressionsgeraden.
Zur direkten Kontrolle ist die gemeinsame Dichte $f_{X,Y}(x,y) = 1/x$ auf $0 < y < x < 1$. Marginalisieren ergibt

$$f_Y(y) = \int_y^1 \frac{1}{x} \, dx = -\ln y, \qquad 0 < y < 1,$$

und damit

$$\mathbb{E}[Y] = \int_0^1 y(-\ln y)\, dy = \frac{1}{4}.$$

Definition 3.5.8 (Bedingte Varianz)

Für $Y \in L^2$ ist die **bedingte Varianz** gegeben X

$$\operatorname{Var}(Y \mid X) = \mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2 \mid X\big].$$

Für eine bedingte Dichte und festes x lautet dieselbe Definition

$$\operatorname{Var}(Y \mid X = x) = \int_{\mathbb{R}} \big(y - \mathbb{E}[Y \mid X = x]\big)^2 f_{Y|X}(y \mid x)\, dy.$$

Beispiel 3.5.2 (Bivariate Normalverteilung: die Regressionsgerade)

Sei (X, Y) gemeinsam normalverteilt mit Erwartungswerten μ_X, μ_Y, Varianzen σ_X^2, σ_Y^2 und Korrelation $\rho \in (-1, 1)$. Dann ist die bedingte Verteilung von X gegeben $Y = y$ wieder normal,

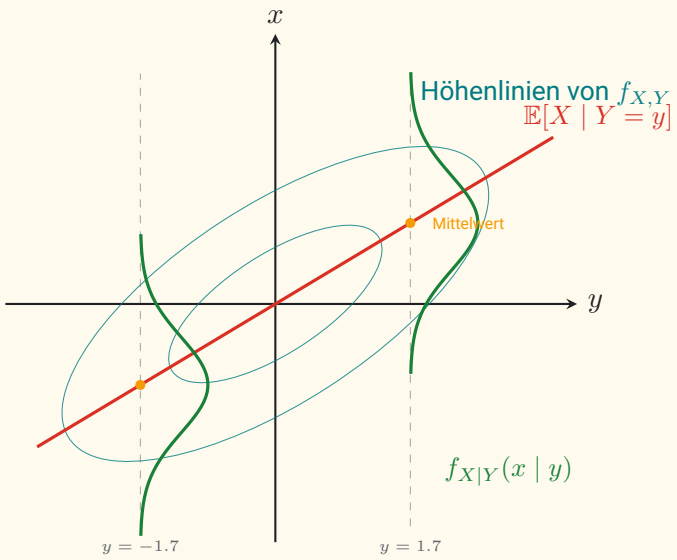
$$X \mid Y = y \;\sim\; N\!\left(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y),\; \sigma_X^2(1 - \rho^2)\right),$$

und insbesondere

$$\mathbb{E}[X \mid Y = y] = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y).$$

Der bedingte Erwartungswert ist *linear* in y – die Regressionsfunktion $g(y) = \mathbb{E}[X \mid Y = y]$ ist eine Gerade, die **Regressionsgerade**. Ihre Steigung $\rho\,\sigma_X/\sigma_Y$ wird von der Korrelation gesteuert: Für $\rho = 0$ ist $\mathbb{E}[X \mid Y = y] = \mu_X$ konstant (bei Normalverteilung bedeutet $\rho = 0$ Unabhängigkeit), für $\rho \neq 0$ verschiebt eine Abweichung $(y - \mu_Y)$ die Erwartung von X proportional dazu.
Die bedingte Varianz $\operatorname{Var}(X \mid Y = y) = \sigma_X^2(1 - \rho^2)$ hängt *nicht* von y ab (Homoskedastizität) und ist um den Faktor $(1 - \rho^2)$ kleiner als σ_X^2: Die Kenntnis von Y reduziert die Unsicherheit über X. Probe über die Turm-Eigenschaft:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}\underbrace{\mathbb{E}[Y - \mu_Y]}_{=0} = \mu_X. \;\checkmark$$



Jeder senkrechte Schnitt $y = \text{const}$ liefert eine glockenförmige bedingte Dichte (grün) mit *gleicher* Streuung $\sigma_X\sqrt{1-\rho^2}$; ihre Mittelwerte (orange) liegen exakt auf der Regressionsgeraden (rot), die durch das Zentrum (μ_Y, μ_X) der Höhenlinien (cyan) verläuft.

Satz 3.5.9 (Varianzzerlegung)

Ist $Y \in L^2$, so gilt

$$\operatorname{Var}(Y) = \mathbb{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}[\mathbb{E}[Y \mid X]].$$

Die Gesamtvarianz zerfällt in „Varianz innerhalb der Gruppen“ plus „Varianz zwischen den Gruppen“ – die mathematische Grundlage der Varianzanalyse (ANOVA).

Beweis. Wir schreiben abkürzend $m = \mathbb{E}[Y \mid X]$ für die Regressionsfunktion. Die bedingte Varianz ist definiert als die gewöhnliche Varianz unter der bedingten Verteilung,

$$\operatorname{Var}(Y \mid X) = \mathbb{E}\left[(Y-m)^2 \,\middle|\, X\right] = \mathbb{E}[Y^2 \mid X] - m^2,$$

wobei die zweite Gleichung der Verschiebungssatz ist, angewandt auf die bedingte Verteilung. Wegen $Y \in L^2$ sind alle auftretenden Erwartungswerte endlich.

Erster Summand. Erwartungswert bilden und zweimal die Turm-Eigenschaft (Satz 3.5.7) anwenden – einmal auf Y^2, einmal implizit auf Y:

$$\mathbb{E}[\operatorname{Var}(Y \mid X)] = \mathbb{E}\left[\mathbb{E}[Y^2 \mid X]\right] - \mathbb{E}[m^2] = \mathbb{E}[Y^2] - \mathbb{E}[m^2].$$

Zweiter Summand. Der Verschiebungssatz für die Zufallsvariable m, wieder mit der Turm-Eigenschaft $\mathbb{E}[m] = \mathbb{E}[Y]$:

$$\operatorname{Var}(m) = \mathbb{E}[m^2] - (\mathbb{E}[m])^2 = \mathbb{E}[m^2] - (\mathbb{E}[Y])^2.$$

Addieren. Der Term $\mathbb{E}[m^2]$ tritt einmal mit Minus und einmal mit Plus auf und hebt sich weg:

$$\mathbb{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}(m) = \mathbb{E}[Y^2] \underbrace{- \mathbb{E}[m^2] + \mathbb{E}[m^2]}_{=0} - (\mathbb{E}[Y])^2 = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \operatorname{Var}(Y). \qquad \blacksquare$$

Beispiel 3.5.3 (Varianz innerhalb und zwischen Gruppen)

Wähle zunächst zufällig eine Grafschaft in den USA und anschließend zufällig eine Person aus dieser Grafschaft. Sei X die gewählte Grafschaft und Y das Einkommen der Person. Dann zerlegt sich die gesamte Einkommensvarianz als

$$\operatorname{Var}(Y) = \underbrace{\mathbb{E}[\operatorname{Var}(Y \mid X)]}_{\text{Varianz innerhalb der Grafschaften}} + \underbrace{\operatorname{Var}(\mathbb{E}[Y \mid X])}_{\text{Varianz zwischen den Grafschaften}}.$$

Der erste Term misst die durchschnittliche Streuung der Einkommen innerhalb einer Grafschaft, der zweite die Streuung der grafschaftsspezifischen Mittelwerte.

Bemerkung 3.5.3 (Bedingen kann die Streuung nur verringern)

Beide Summanden sind als Erwartungswert einer nichtnegativen Größe bzw. als Varianz selbst nichtnegativ. Aus der Zerlegung folgt daher sofort

$$\operatorname{Var}\big(\mathbb{E}[Y \mid X]\big) \le \operatorname{Var}(Y),$$

mit Gleichheit genau dann, wenn $\mathbb{E}[\operatorname{Var}(Y \mid X)] = 0$, also Y f.s. bereits durch X festgelegt ist. Umgekehrt ist $\operatorname{Var}(\mathbb{E}[Y \mid X]) = 0$, wenn X nichts über Y verrät. Das Beispiel der bivariaten Normalverteilung oben zeigt genau diesen Effekt: Die bedingte Varianz $\sigma_X^2(1-\rho^2)$ ist um den Faktor $(1-\rho^2)$ kleiner als die unbedingte.

3.6 Momenterzeugende Funktionen

Definition 3.6.1

Die **momenterzeugende Funktion** (mgf) von X ist

$$\psi_X(t) = \mathbb{E}[e^{tX}] \in (0, \infty].$$

Da $e^{tX} > 0$ ist, ist dieser Erwartungswert stets definiert, kann aber $+\infty$ sein. Wir sagen, X **besitzt eine mgf**, wenn es ein $\delta > 0$ gibt mit

$$\psi_X(t) < \infty \qquad \text{für alle } t \in (-\delta, \delta).$$

Alle folgenden Aussagen setzen diese Bedingung voraus.

Nicht jede Verteilung hat eine mgf

Die Endlichkeit auf einer *Umgebung* von 0 ist eine echte Einschränkung, keine Formalie. Für die Cauchy-Verteilung und die Lognormalverteilung ist $\psi_X(t) = \infty$ für jedes $t \neq 0$ – beide besitzen keine mgf. Bei $X \sim \operatorname{Exp}(\beta)$ – in der Skalenkonvention dieses Buches, also $f(x) = \frac{1}{\beta} e^{-x/\beta}$ – existiert sie nur für $t < 1/\beta$, immerhin auf einer Umgebung von 0. Wer mit mgfs argumentiert, ohne ihre Existenz zu prüfen, kann daher zu falschen Schlüssen kommen. Das allgemeine

Werkzeug ohne diese Lücke ist die *charakteristische Funktion* $\varphi_X(t) = \mathbb{E}[e^{itX}]$, die wegen $|e^{itX}| = 1$ immer existiert.

Beispiel 3.6.1 (mgf der Exponentialverteilung)

Sei $X \sim \mathrm{Exp}(\beta)$ mit Dichte $f_X(x) = \beta^{-1}e^{-x/\beta}$ für $x > 0$. Dann gilt für $t < 1/\beta$

$$\begin{aligned}\psi_X(t) = \mathbb{E}[e^{tX}] &= \int_0^\infty e^{tx}\frac{1}{\beta}e^{-x/\beta}\,dx \\ &= \frac{1}{\beta}\int_0^\infty e^{-(1/\beta - t)x}\,dx = \frac{1}{1-\beta t}.\end{aligned}$$

Für $t \geq 1/\beta$ divergiert das Integral. Im Spezialfall $\beta = 1$ ist $\psi_X(t) = 1/(1-t)$ für $t < 1$.

Lemma 3.6.2 (Rechenregeln für mgfs)

Sei ψ_X endlich auf $(-\delta, \delta)$. Dann gilt:

1. X besitzt Momente jeder Ordnung, und $\mathbb{E}[X^k] = \psi_X^{(k)}(0)$,
2. $X \perp Y \Rightarrow \psi_{X+Y}(t) = \psi_X(t)\,\psi_Y(t)$, sofern beide mgfs bei t endlich sind,
3. $\psi_{aX+b}(t) = e^{bt}\psi_X(at)$ für alle t mit $at \in (-\delta, \delta)$.

Beweis. **1. Ableiten unter dem Erwartungswert.** Hier steckt die eigentliche Arbeit – $\frac{d}{dt}\,\mathbb{E}[\,\cdot\,] = \mathbb{E}[\frac{d}{dt}\,\cdot\,]$ ist nicht selbstverständlich und braucht genau die Voraussetzung der Definition. Wähle $t_0 < t_1 < \delta$. Für $|t| \leq t_0$ gilt die Abschätzung

$$|X^k e^{tX}| \leq |X|^k e^{t_0|X|} = \underbrace{\left(|X|^k e^{-(t_1-t_0)|X|}\right)}_{\leq C_k < \infty} e^{t_1|X|} \leq C_k\, e^{t_1|X|},$$

denn $u \mapsto u^k e^{-(t_1-t_0)u}$ ist auf $[0,\infty)$ beschränkt. Die dominierende Größe ist integrierbar:

$$\mathbb{E}\left[e^{t_1|X|}\right] \leq \mathbb{E}\left[e^{t_1 X}\right] + \mathbb{E}\left[e^{-t_1 X}\right] = \psi_X(t_1) + \psi_X(-t_1) < \infty.$$

Der Satz von der majorisierten Konvergenz erlaubt daher, Ableitung und Erwartungswert zu vertauschen:

$$\psi_X^{(k)}(t) = \mathbb{E}\left[X^k e^{tX}\right], \qquad |t| \leq t_0.$$

Für $t = 0$ folgt $\psi_X^{(k)}(0) = \mathbb{E}[X^k]$; die Abschätzung mit $t = 0$ zeigt zugleich $\mathbb{E}[|X|^k] < \infty$, also die Existenz aller Momente. Äquivalent liest man dasselbe an der Reihenentwicklung ab:

$$\psi_X(t) = \mathbb{E}\left[\sum_{k\geq 0}\frac{(tX)^k}{k!}\right] = \sum_{k\geq 0}\frac{\mathbb{E}[X^k]}{k!}\,t^k,$$

die Momente sind die Taylor-Koeffizienten – daher der Name *momenterzeugend*.

2. Unabhängigkeit. Sind X und Y unabhängig, so sind auch e^{tX} und e^{tY} unabhängig, denn messbare Funktionen unabhängiger Zufallsvariablen sind unabhängig. Mit dem Produktsatz für Erwartungswerte folgt

$$\psi_{X+Y}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX}e^{tY}\right] = \mathbb{E}\left[e^{tX}\right]\,\mathbb{E}\left[e^{tY}\right] = \psi_X(t)\,\psi_Y(t).$$

3. Affine Transformation. Der Faktor e^{bt} ist deterministisch und wandert aus dem Erwartungswert heraus:

$$\psi_{aX+b}(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = \mathbb{E}\left[e^{bt}e^{(at)X}\right] = e^{bt}\,\mathbb{E}\left[e^{(at)X}\right] = e^{bt}\,\psi_X(at).$$ ■

Beispiel 3.6.2 (mgf der Binomialverteilung)

Für $B \sim \text{Bernoulli}(p)$ gilt

$$\psi_B(t) = \mathbb{E}[e^{tB}] = (1-p) + pe^t.$$

Ist $X = \sum_{i=1}^{n} B_i \sim \text{Binomial}(n,p)$ mit unabhängigen Bernoulli-Variablen B_i, so liefert Lemma 3.6.2

$$\psi_X(t) = \prod_{i=1}^{n} \psi_{B_i}(t) = \left((1-p) + pe^t\right)^n.$$

Beispiel 3.6.3 (Wichtige Momente der Normalverteilung)

Für $X \sim N(\mu, \sigma^2)$ ist

$$\psi_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Ableiten an der Stelle $t = 0$ ergibt neben $\mathbb{E}[X] = \mu$ und $\text{Var}(X) = \sigma^2$ die höheren Rohmomente

$$\mathbb{E}[X^3] = \mu^3 + 3\mu\sigma^2, \qquad \mathbb{E}[X^4] = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4.$$

Für die Standardnormalverteilung folgen insbesondere $\mathbb{E}[X^3] = 0$ und $\mathbb{E}[X^4] = 3$.

Satz 3.6.3 (Eindeutigkeitssatz)

Existiert ein $\delta > 0$ mit $\psi_X(t) = \psi_Y(t) < \infty$ für alle $t \in (-\delta, \delta)$, so gilt $F_X = F_Y$.

Ohne Beweis. Der Beweis führt über die charakteristische Funktion $\varphi_X(t) = \mathbb{E}[e^{itX}]$ und deren Umkehrformel; die Brücke von ψ_X zu φ_X ist eine analytische Fortsetzung in die komplexe Ebene und benötigt Funktionentheorie, die wir hier nicht voraussetzen. Ausführlich in Billingsley, *Probability and Measure*, §30, oder Durrett, *Probability: Theory and Examples*, Kap. 3.3.

Die mgf charakterisiert die Verteilung also eindeutig – das macht sie zum idealen Werkzeug, um Faltungen zu berechnen: Statt Summenverteilungen direkt zu bestimmen, multipliziert man mgfs nach Lemma 3.6.2 (2) und erkennt das Ergebnis wieder.

Beispiel 3.6.4

$X_1 \sim \text{Binomial}(n_1, p)$, $X_2 \sim \text{Binomial}(n_2, p)$ unabhängig: $\psi_{X_1+X_2}(t) = [(1-p) + pe^t]^{n_1+n_2}$, also $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$.

Beispiel 3.6.5

$Y_1 \sim \text{Poisson}(\lambda_1)$, $Y_2 \sim \text{Poisson}(\lambda_2)$ unabhängig: $\psi_{Y_1+Y_2}(t) = e^{(\lambda_1+\lambda_2)(e^t-1)}$, also $Y_1 + Y_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

KAPITEL 4

Ungleichungen

4.1 Wahrscheinlichkeitsungleichungen

Wir werfen eine Münze 100-mal und erhalten 65-mal Kopf. Ist die Münze manipuliert, oder war es nur Zufall? Mathematisch: Wie wahrscheinlich ist es, bei einer fairen Münze so weit vom Erwartungswert abzuweichen?

Genau hier kommen Wahrscheinlichkeitsungleichungen ins Spiel – das Schweizer Taschenmesser der Statistik. Mit minimalem Wissen über eine Verteilung (manchmal nur Erwartungswert oder Varianz) können wir Aussagen über seltene Ereignisse treffen. Das Beste: Diese Ungleichungen gelten *unabhängig von der konkreten Verteilung*.

Satz 4.1.1 (Markov-Ungleichung)

Sei $X \geq 0$ mit $\mathbb{E}[X] < \infty$. Für $t > 0$ gilt

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Beweis. Sei $A = \{X \geq t\}$ und I_A die zugehörige Indikatorfunktion. Punktweise gilt

$$X \geq t\, I_A :$$

für $\omega \in A$ ist $X(\omega) \geq t = t\, I_A(\omega)$, für $\omega \notin A$ ist $t\, I_A(\omega) = 0 \leq X(\omega)$ wegen $X \geq 0$.

Die Differenz $X - t\, I_A$ ist demnach nichtnegativ, und der Erwartungswert einer nichtnegativen Zufallsvariablen ist nichtnegativ – Summe bzw. Integral laufen über lauter nichtnegative Terme. Mit der Linearität (Satz 3.2.1) folgt daraus

$$\mathbb{E}[X] - t\, \mathbb{E}[I_A] = \mathbb{E}[X - t\, I_A] \geq 0.$$

Schließlich nimmt I_A nur die Werte 1 und 0 an, ist also diskret mit $\mathbb{E}[I_A] = 1 \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = \mathbb{P}(A)$. Zusammengesetzt ergibt das $\mathbb{E}[X] \geq t\, \mathbb{P}(X \geq t)$. ■

Bemerkung 4.1.1

Der Beweis benutzt weder eine Dichte noch eine Zähldichte, sondern nur die punktweise Abschätzung $X \geq t\, I_A$ und die Linearität. Er trägt daher so weit, wie der Erwartungswert überhaupt erklärt ist – in diesem Band für diskrete und stetige Zufallsvariablen, in der maßtheoretischen Fassung $\mathbb{E}[X] = \int X \, d\mathbb{P}$ für jede nichtnegative Zufallsvariable.

Interpretation: Ist $\mathbb{E}[X] = 5$ und $t = 50$, so $\mathbb{P}(X \geq 50) \leq 1/10$. Die Masse kann nicht beliebig weit vom Erwartungswert wegwandern.

Satz 4.1.2 (Tschebyschow-Ungleichung)

Sei $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}(X)$. Dann

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Mit $Z = (X - \mu)/\sigma$: $\mathbb{P}(|Z| \geq k) \leq 1/k^2$.

Beweis. Markov auf $|X - \mu|^2$: $\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^2 \geq t^2) \leq \sigma^2/t^2$. ■

Interpretation: Bei *jeder* Verteilung liegen mindestens 75% der Masse innerhalb von 2σ und $\approx 89\%$ innerhalb von 3σ um den Mittelwert.

Beispiel 4.1.1 (Fehlerrate eines Prädiktors)

Teste ein neuronales Netz auf n Fällen. Sei $X_i = 1$ bei Fehler, p die wahre Fehlerrate. Die beobachtete Fehlerrate $\overline{X}_n$ hat $\mathrm{Var}(\overline{X}_n) = p(1-p)/n \leq 1/(4n)$. Tschebyschow:

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq \frac{1}{4n\epsilon^2}.$$

Bei $n = 1000$, $\epsilon = 0{,}05$: $\mathbb{P}(|\overline{X}_n - p| > 0{,}05) \leq 0{,}1$. Mit 90% Wahrscheinlichkeit liegt die gemessene Fehlerrate innerhalb von $\pm 5\%$ der wahren.

Satz 4.1.3 (Hoeffding-Ungleichung)

Seien $Y_1, \ldots, Y_n$ unabhängig mit $\mathbb{E}[Y_i] = 0$ und $a_i \leq Y_i \leq b_i$. Dann

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Beweis. **Hoeffding-Lemma:** Für $\mathbb{E}[Y] = 0$, $a \leq Y \leq b$ gilt $\mathbb{E}[e^{tY}] \leq e^{t^2(b-a)^2/8}$.

Beweis des Lemmas: Konvexität von e^{tx} liefert $e^{ty} \leq \frac{b-y}{b-a}e^{ta} + \frac{y-a}{b-a}e^{tb}$. Erwartungswert bilden mit $\mathbb{E}[Y] = 0$: $\mathbb{E}[e^{tY}] \leq e^{L(h)}$ wobei $h = t(b-a)$ und $L(h) = -ph + \log(1 - p + pe^h)$, $p = -a/(b-a)$. Da $L(0) = L'(0) = 0$ und $L''(h) \leq 1/4$ (weil $p(1-p) \leq 1/4$), folgt $L(h) \leq h^2/8$.

Chernoff-Methode: Für $t > 0$, mit Markov und Unabhängigkeit:

$$\mathbb{P}\left(\sum Y_i \geq \epsilon\right) = \mathbb{P}\left(e^{t\sum Y_i} \geq e^{t\epsilon}\right) \leq e^{-t\epsilon}\prod_{i=1}^{n}\mathbb{E}[e^{tY_i}] \leq \exp\left(\frac{t^2\sum(b_i - a_i)^2}{8} - t\epsilon\right).$$

Minimierung über t (optimal: $t^* = 4\epsilon/\sum(b_i - a_i)^2$) liefert die Behauptung. ■

Korollar 4.1.4

Für $X_1, \ldots, X_n \sim$ Bernoulli(p) iid gilt $\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$.

Beweis. Setze $Y_i = X_i - p$, also $b_i - a_i = 1$. Hoeffding für $\sum Y_i > n\epsilon$ und symmetrisch für $< -n\epsilon$, dann Unionsschranke. ∎

Beispiel 4.1.2 (Tschebyschow vs. Hoeffding)

$X_i \sim$ Bernoulli(p), $n = 100$, $\epsilon = 0{,}2$:

$$\text{Tschebyschow: } \leq 0{,}0625. \quad \text{Hoeffding: } \leq 2e^{-8} \approx 0{,}00067.$$

Hoeffding ist fast **100-mal schärfer**! Die exponentielle Schranke dominiert für große Abweichungen.

Bemerkung 4.1.2 (Quantitativer Vergleich)

Schranken für $\mathbb{P}(|\overline{X}_n - p| > \epsilon)$ bei $X_i \sim$ Bernoulli$(0{,}5)$, $n = 100$:

ϵ	Tschebyschow	Hoeffding	Exakt
0,10	0,250	0,271	0,035
0,15	0,111	0,022	0,002
0,20	0,063	0,0007	4×10^{-5}
0,30	0,028	3×10^{-8}	10^{-9}

Für große ϵ ist Hoeffding dramatisch schärfer – der exponentielle Abfall macht den Unterschied.

Satz 4.1.5 (Mills-Ungleichung)

Für $Z \sim N(0,1)$ und $t > 0$ gilt

$$\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t} \quad \text{und} \quad \mathbb{P}(|Z| > t) \leq 2e^{-t^2/2}.$$

Beweis. **Erste Schranke:** Für $t > 0$:

$$\mathbb{P}(Z > t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx \leq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{t} \int_t^\infty x e^{-x^2/2} dx = \frac{e^{-t^2/2}}{t\sqrt{2\pi}}.$$

Verdoppeln (Symmetrie) liefert die erste Ungleichung.

Zweite Schranke: Chernoff mit $\mathbb{E}[e^{sZ}] = e^{s^2/2}$: $\mathbb{P}(Z > t) \leq e^{-st+s^2/2}$. Optimal bei $s = t$: $\mathbb{P}(Z > t) \leq e^{-t^2/2}$. ∎

4.2 Ungleichungen für Erwartungswerte

Satz 4.2.1 (Cauchy-Schwarz-Ungleichung)

Für $X, Y \in L^2$ gilt

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}, \qquad |\operatorname{Cov}(X,Y)| \leq \sqrt{\operatorname{Var}(X)\,\operatorname{Var}(Y)}.$$

Gleichheit gilt genau bei linearer Abhängigkeit.

Beweis. Für $t \in \mathbb{R}$ ist $0 \leq \mathbb{E}[(X - tY)^2] = \mathbb{E}[X^2] - 2t\,\mathbb{E}[XY] + t^2\,\mathbb{E}[Y^2]$. Als quadratische Funktion in t ist die Diskriminante ≤ 0: $[\mathbb{E}[XY]]^2 \leq \mathbb{E}[X^2]\,\mathbb{E}[Y^2]$.

Kovarianz-Version: Ersetze X durch $X - \mathbb{E}[X]$, Y durch $Y - \mathbb{E}[Y]$. ■

Dividiert man die Kovarianz-Version durch $\sigma_X \sigma_Y$, so steht dort erneut die Korrelationsschranke $|\rho| \leq 1$ (Lemma 3.3.9) – dort elementar aus $\operatorname{Var} \geq 0$ gewonnen, hier als Spezialfall $p = q = 2$ eines allgemeinen Prinzips.

Satz 4.2.2 (Jensen-Ungleichung)

Ist g konvex, so gilt $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. Ist g konkav, so gilt $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$.

Beweis. Sei $\mu = \mathbb{E}[X]$. Da g konvex, existiert ein Subgradient a mit $g(x) \geq a(x - \mu) + g(\mu)$ für alle x. Erwartungswert bilden:

$$\mathbb{E}[g(X)] \geq a(\mathbb{E}[X] - \mu) + g(\mu) = g(\mu).$$

Der konkave Fall folgt durch Anwenden auf $-g$. ■

Geometrische Intuition: Konvexe Funktionen „biegen nach oben“. Erst g anwenden, dann mitteln liefert mehr als erst mitteln, dann g anwenden.

Beispiel 4.2.1

(1) $g(x) = x^2$ konvex: $\mathbb{E}[X^2] \geq [\mathbb{E}[X]]^2$, also $\operatorname{Var}(X) \geq 0$ – gratis!
(2) $g(x) = \log x$ konkav: $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$, d. h. das geometrische Mittel ist $\leq$ dem arithmetischen.

Bemerkung 4.2.1 (Hölder und Minkowski)

Hölder: Für $1/p + 1/q = 1$ ($p, q > 1$): $\mathbb{E}[|XY|] \leq [\mathbb{E}[|X|^p]]^{1/p} [\mathbb{E}[|Y|^q]]^{1/q}$. Für $p = q = 2$: Cauchy-Schwarz.
Minkowski: Für $p \geq 1$: $[\mathbb{E}[|X + Y|^p]]^{1/p} \leq [\mathbb{E}[|X|^p]]^{1/p} + [\mathbb{E}[|Y|^p]]^{1/p}$ (Dreiecksungleichung in L^p).

KAPITEL 5

Konvergenz von Zufallsvariablen

5.1 Einführung

Der wichtigste Aspekt der Wahrscheinlichkeitstheorie betrifft das Verhalten von Folgen von Zufallsvariablen – die **Grenzwerttheorie**.

In der Praxis haben wir oft eine Folge $X_1, X_2, \ldots$ (Messungen, Renditen, Münzwürfe). Die zentrale Frage: *Was passiert für* $n \to \infty$*?* In der Analysis ist Konvergenz einer Zahlenfolge klar. Für Zufallsvariablen – Funktionen $X : \Omega \to \mathbb{R}$ – gibt es mehrere sinnvolle Konvergenzbegriffe.

Zwei fundamentale Resultate stehen im Zentrum:

1. **Gesetz der großen Zahlen:** $\overline{X}_n \to \mu$.
2. **Zentraler Grenzwertsatz:** $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \to N(0,1)$.

Sie bilden das theoretische Fundament der gesamten Statistik.

5.2 Konvergenzarten

Definition 5.2.1 (Konvergenz in Wahrscheinlichkeit)

$X_n \xrightarrow{P} X$ falls für jedes $\epsilon > 0$: $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$.

Intuition: Mit hoher Wahrscheinlichkeit liegt X_n nahe bei X für großes n.

Definition 5.2.2 (Konvergenz in Verteilung)

$X_n \xrightarrow{d} X$ falls $\lim_{n\to\infty} F_n(t) = F(t)$ an allen Stetigkeitsstellen t von F.

Die X_n und X müssen nicht auf demselben Wahrscheinlichkeitsraum leben – nur die Verteilungen zählen. Dies ist die schwächste Konvergenzform.

Definition 5.2.3 (Konvergenz in L^p)

Für $p \geq 1$: $X_n \to X$ in L^p falls $\mathbb{E}[|X_n - X|^p] \to 0$.

Definition 5.2.4 (Fast-sichere Konvergenz)

$X_n \xrightarrow{\text{f.s.}} X$ falls $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.

Intuition: Vier Begriffe – eine Landkarte

Alle vier Begriffe sagen „X_n nähert sich X“, aber sie messen das unterschiedlich streng:

- **Fast sicher** ($\xrightarrow{\text{f.s.}}$): Verfolgt man *einen* Zufallspfad ω, läuft die Zahlenfolge $X_n(\omega)$ tatsächlich ins Ziel – für (fast) jeden Pfad.
- **In L^p**: Der *durchschnittliche* Fehler (in der p-ten Potenz) schrumpft auf 0.
- **In Wahrscheinlichkeit** ($\xrightarrow{P}$): Die *Chance* auf einen nennenswerten Fehler $|X_n - X| > \epsilon$ schrumpft auf 0 – einzelne Ausreißer bleiben erlaubt.
- **In Verteilung** ($\xrightarrow{d}$): Nur die *Form* der Verteilung (die cdf) stabilisiert sich. Über den konkreten Wert von X_n sagt das nichts.

Merksatz: $\xrightarrow{\text{f.s.}}$ und L^p sind „stark“, $\xrightarrow{P}$ liegt in der Mitte, $\xrightarrow{d}$ ist die *schwächste* Form – sie verlangt nicht einmal, dass X_n und X auf demselben Wahrscheinlichkeitsraum leben.

Beispiel 5.2.1 (Normalverteilung mit schrumpfender Varianz)

$X_n \sim N(0, 1/n)$. Dann $\mathbb{E}[X_n^2] = 1/n \to 0$ (L^2), $\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}(|Z| > \epsilon\sqrt{n}) \to 0$ (Konvergenz in Wahrscheinlichkeit), und $F_n(t) = \Phi(t\sqrt{n}) \to I_{\{t\geq 0\}}$ *für jedes* $t \neq 0$, also $X_n \xrightarrow{d} \delta_0$ (Verteilung). Bei $t = 0$ konvergiert dagegen nichts: Dort ist $F_n(0) = \Phi(0) = 1/2$ für jedes n, während die Grenz-Verteilungsfunktion den Wert 1 hat. Genau dafür steht die Stetigkeitsstellen-Klausel in der Definition der Verteilungskonvergenz – und $t = 0$ ist die einzige Unstetigkeitsstelle der Punktmasse δ_0.

Beispiel 5.2.2 (Konvergenz in Verteilung $\not\Rightarrow$ Konvergenz in Wahrscheinlichkeit)

$X \sim N(0, 1)$, $X_n = -X$ für alle n. Dann $X_n \xrightarrow{d} X$ (da $F_n = F$), aber $\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|2X| > \epsilon) > 0$ für alle n.

Typischer Fehler

„$X_n \xrightarrow{d} X$“ bedeutet *nicht*, dass X_n und X nahe beieinander liegen. Im Beispiel oben ist $X_n = -X$: die Verteilung ist sogar *identisch* ($F_n = F$), trotzdem ist X_n von X maximal entfernt. Verteilungskonvergenz vergleicht nur die Form der cdf, niemals die konkreten Werte – deshalb braucht sie nicht einmal denselben Wahrscheinlichkeitsraum. Wer „$\xrightarrow{d}$“ wie „die Werte werden gleich“ liest, zieht falsche Schlüsse.

Hierarchie der Konvergenzarten

Satz 5.2.5 (Hierarchie)

1. L^2-Konvergenz $\Rightarrow$ Konvergenz in Wahrscheinlichkeit.
2. Fast-sichere Konvergenz $\Rightarrow$ Konvergenz in Wahrscheinlichkeit.
3. Konvergenz in Wahrscheinlichkeit $\Rightarrow$ Konvergenz in Verteilung.
4. $X_n \xrightarrow{d} c$ (Konstante) $\Rightarrow X_n \xrightarrow{P} c$.

Die Umkehrungen gelten im Allgemeinen nicht.

Beweis. **(1)** Markov: $\mathbb{P}(|X_n - X| > \epsilon) \leq \mathbb{E}[(X_n - X)^2]/\epsilon^2 \to 0$.

(2) Definiere $A_n(\epsilon) = \{|X_m - X| \leq \epsilon \text{ für alle } m \geq n\}$. Dann $A_n \uparrow$ und $\mathbb{P}(\bigcup A_n) = 1$ (f.s.-Konvergenz). Also $\mathbb{P}(|X_n - X| > \epsilon) \leq 1 - \mathbb{P}(A_n) \to 0$.

(3) Sei t Stetigkeitsstelle von F. Für $\epsilon > 0$:

$$F_n(t) \leq \mathbb{P}(X \leq t + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) = F(t + \epsilon) + o(1).$$

Analog $F(t - \epsilon) - o(1) \leq F_n(t)$. Da ϵ beliebig und F stetig bei t: $F_n(t) \to F(t)$.

(4) $\mathbb{P}(|X_n - c| > \epsilon) = F_n(c - \epsilon) + 1 - F_n(c + \epsilon) \to 0 + 0$, da $c \pm \epsilon$ Stetigkeitsstellen der Grenz-cdf (Punktmasse bei c) sind. ■

Kurz-Check: Was bedeutet das konkret?

Eine Folge konvergiert in Verteilung gegen eine *Konstante* c. Folgt daraus auch Konvergenz in Wahrscheinlichkeit gegen c?

Antwort: Ja. Das ist Punkt (4) der Hierarchie. Konvergenz in Verteilung ist sonst die schwächste Form, aber bei einem konstanten Grenzwert fällt sie mit der Konvergenz in Wahrscheinlichkeit zusammen: Die Grenzverteilung ist eine Punktmasse bei c, sodass „die Form stabilisiert sich“ und „die Werte konzentrieren sich bei c“ dasselbe bedeuten. Genau dieser Spezialfall macht das Slutsky-Theorem erst nutzbar.

5.3 Das Slutsky-Theorem

Satz 5.3.1 (Slutsky)

1. $X_n \xrightarrow{P} X$ und g stetig $\Rightarrow g(X_n) \xrightarrow{P} g(X)$.
2. $X_n \xrightarrow{d} X$ und $Y_n \xrightarrow{P} c \Rightarrow X_n + Y_n \xrightarrow{d} X + c$, $X_n Y_n \xrightarrow{d} cX$, $X_n / Y_n \xrightarrow{d} X/c$ (falls $c \neq 0$).

Beweis. **(1)** Sei $\epsilon > 0$. Wähle M mit $\mathbb{P}(|X| > M) < \epsilon/2$. Auf $[-M, M]$ ist g gleichmäßig stetig: $\exists \delta$: $|x - y| < \delta \Rightarrow |g(x) - g(y)| < \epsilon$. Dann

$$\mathbb{P}(|g(X_n) - g(X)| > \epsilon) \leq \mathbb{P}(|X_n - X| > \delta) + \mathbb{P}(|X| > M) \to 0 + \epsilon/2.$$

(2) Für t Stetigkeitsstelle von F_{X+c} und $\epsilon > 0$: $\mathbb{P}(X_n + Y_n \leq t) \leq \mathbb{P}(X_n \leq t - c + \epsilon) + \mathbb{P}(|Y_n - c| > \epsilon)$. Grenzübergang und $\epsilon \to 0$ liefert $F_n(t) \to F_{X+c}(t)$. Multiplikation analog. ■

5.4 Das Gesetz der großen Zahlen

Satz 5.4.1 (Schwaches Gesetz der großen Zahlen (WLLN))

Seien $X_1, X_2, \ldots$ iid mit $\mathbb{E}[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Dann

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

Beweis. Tschebyschow: $\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \mathrm{Var}(\overline{X}_n)/\epsilon^2 = \sigma^2/(n\epsilon^2) \to 0$. ■

Satz 5.4.2 (Starkes Gesetz der großen Zahlen)

Unter denselben Voraussetzungen gilt sogar $\overline{X}_n \xrightarrow{\text{f.s.}} \mu$.

Ohne Beweis. Fast sichere Konvergenz lässt sich nicht mehr wie beim schwachen Gesetz mit einer einzigen Ungleichung erschlagen: Man muss die Ausnahmemengen über alle n hinweg kontrollieren. Die üblichen Wege sind das Lemma von Borel-Cantelli zusammen mit einer Blockzerlegung der Teilfolgen (Etemadis Beweis, der nur $\mathbb{E}\,|X_1| < \infty$ verlangt) oder das Martingalkonvergenzsatz-Argument. Beides braucht Werkzeuge, die dieser Band nicht bereitstellt. Ausführlich in Durrett, *Probability: Theory and Examples*, 5. Aufl. 2019, Kap. 2.4, oder Billingsley, *Probability and Measure*, §22.

Beispiel 5.4.1 (Münzwurf und frequentistische Interpretation)

Sei $X_i = I_{\{\text{Kopf}\}}$ mit $\mathbb{P}(X_i = 1) = p$. Dann ist $\overline{X}_n$ die relative Häufigkeit von Kopf, und $\overline{X}_n \xrightarrow{P} p$. Dies rechtfertigt die frequentistische Interpretation: Wahrscheinlichkeit als Grenzwert der relativen Häufigkeit.
Quantitativ: Für $n = 10000$, $p = 0{,}5$, $\epsilon = 0{,}01$: $\mathbb{P}(|\overline{X}_n - 0{,}5| > 0{,}01) \leq 0{,}25/(10000 \cdot 0{,}0001) = 0{,}25$.

5.5 Der zentrale Grenzwertsatz

Satz 5.5.1 (Zentraler Grenzwertsatz (CLT))

Seien $X_1, X_2, \ldots$ iid mit $\mathbb{E}[X_i] = \mu$, $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Dann

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

Äquivalent: $\overline{X}_n \approx N(\mu, \sigma^2/n)$.

Beweisskizze. Sei $Y_i = (X_i - \mu)/\sigma$ mit $\mathbb{E}[Y_i] = 0$, $\mathrm{Var}(Y_i) = 1$. Die charakteristische Funktion $\varphi(t) = \mathbb{E}[e^{itY_i}]$ hat Taylor-Entwicklung $\varphi(t) = 1 - t^2/2 + o(t^2)$. Für $Z_n = \frac{1}{\sqrt{n}} \sum Y_i$:

$$\varphi_{Z_n}(t) = [\varphi(t/\sqrt{n})]^n = \left[1 - \frac{t^2}{2n} + o(t^2/n)\right]^n \to e^{-t^2/2}.$$

Dies ist die charakteristische Funktion von $N(0,1)$. Der Stetigkeitssatz von Lévy liefert $Z_n \xrightarrow{d} N(0,1)$.

Skizze bleibt dies aus zwei Gründen: Der Grenzübergang $[1 - t^2/(2n) + o(t^2/n)]^n \to e^{-t^2/2}$ verlangt eine sorgfältige Abschätzung des komplexen Logarithmus, und der Stetigkeitssatz von Lévy wird hier nur zitiert. Für den vollständigen Beweis siehe Billingsley, *Probability and Measure*, §27, oder Durrett, *Probability: Theory and Examples*, Kap. 3.4. ∎

Beispiel 5.5.1 (Binomialapproximation)

$S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n,p)$. CLT: $\frac{S_n - np}{\sqrt{np(1-p)}} \approx N(0,1)$.
Für $n = 100$, $p = 0{,}3$ ist $np = 30$ und $np(1-p) = 21$. Exakt gilt $\mathbb{P}(S_n \geq 35) \approx 0{,}162\,86$.

$$\text{CLT: } \mathbb{P}\left(Z \geq \frac{5}{\sqrt{21}}\right) = \mathbb{P}(Z \geq 1{,}0911) \approx 0{,}1376.$$

Mit Stetigkeitskorrektur: $\mathbb{P}\left(Z \geq (34{,}5 - 30)/\sqrt{21}\right) = \mathbb{P}(Z \geq 0{,}9820) \approx 0{,}1631$. Der Fehler schrumpft damit von $0{,}025$ auf $0{,}0002$ – die Korrektur trifft den exakten Wert nahezu.

Satz 5.5.2 (Berry-Esséen-Ungleichung)

Falls $\mathbb{E}\left[|X_1 - \mu|^3\right] < \infty$, so gilt

$$\sup_z |\mathbb{P}(Z_n \leq z) - \Phi(z)| \leq \frac{C\, \mathbb{E}\left[|X_1 - \mu|^3\right]}{\sigma^3 \sqrt{n}}$$

mit $C \approx 0{,}4748$. Der Approximationsfehler ist also $O(n^{-1/2})$.

Ohne Beweis. Der Beweis beruht auf der Esséenschen Glättungsungleichung, die den Abstand zweier Verteilungsfunktionen durch ein Integral über die Differenz ihrer charakteristischen Funktionen abschätzt. Das ist ein rein analytisches Argument, das über den Rahmen dieses Bandes hinausgeht. Ausführlich in Feller, *An Introduction to Probability Theory and Its Applications*, Bd. II, Kap. XVI.5, oder Petrov, *Sums of Independent Random Variables*, Kap. V. Die zulässige Konstante C wurde seit Esséen mehrfach verbessert; der angegebene Wert ist der heute übliche.

5.6 Die Delta-Methode

Satz 5.6.1 (Delta-Methode)

Sei g differenzierbar mit $g'(\mu) \neq 0$. Falls $\sqrt{n}(\overline{X}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$, dann

$$\frac{\sqrt{n}(g(\overline{X}_n) - g(\mu))}{\sigma |g'(\mu)|} \xrightarrow{d} N(0,1).$$

Äquivalent: $g(\overline{X}_n) \approx N(g(\mu),\, \sigma^2 [g'(\mu)]^2 / n)$.

Beweis. Taylor um μ liefert

$$g(\overline{X}_n) = g(\mu) + g'(\mu)(\overline{X}_n - \mu) + R_n,$$

wobei der Restterm R_n gegenüber $|\overline{X}_n - \mu|$ verschwindet, d. h. $R_n / |\overline{X}_n - \mu| \to 0$ für $\overline{X}_n \to \mu$. Also

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) = g'(\mu)\sqrt{n}(\overline{X}_n - \mu) + \sqrt{n}\, R_n,$$

und der Restterm $\sqrt{n}\, R_n$ konvergiert in Wahrscheinlichkeit gegen 0. Slutsky liefert $\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2 [g'(\mu)]^2)$. ∎

Beispiel 5.6.1 (Konfidenzintervall für μ^2)

$X_1, \ldots, X_n$ iid mit $\mathbb{E}[X_i] = \mu \neq 0$, $\mathrm{Var}(X_i) = \sigma^2$. Schätze $\tau = \mu^2$ durch $\hat{\tau} = \overline{X}_n^2$. Die Voraussetzung $\mu \neq 0$ ist wesentlich: Nur dann ist $g'(\mu) \neq 0$ und der Satz überhaupt anwendbar.
Mit $g(x) = x^2$, $g'(\mu) = 2\mu$: $\sqrt{n}(\overline{X}_n^2 - \mu^2) \approx N(0, 4\mu^2\sigma^2)$.
Approximatives 95%-Konfidenzintervall:

$$\mu^2 \in \left[\overline{X}_n^2 \pm 1{,}96 \cdot \frac{2|\overline{X}_n| \cdot s}{\sqrt{n}} \right],$$

wobei s die Stichprobenstandardabweichung ist.

Delta-Methode bei $g'(\mu) = 0$

Der Fall $\mu = 0$ fällt aus dem obigen Beispiel heraus, und zwar nicht nur formal: Dort ist $g'(0) = 0$, der lineare Term der Taylor-Entwicklung verschwindet, und es gilt

$$\sqrt{n}\,(\overline{X}_n^2 - 0) = \frac{(\sqrt{n}\,\overline{X}_n)^2}{\sqrt{n}} \xrightarrow{P} 0,$$

die Grenzverteilung ist also die Punktmasse in 0 und nicht normal. Die richtige Rate ist dann n statt $\sqrt{n}$:

$$\frac{n\,\overline{X}_n^2}{\sigma^2} = \left(\frac{\sqrt{n}\,\overline{X}_n}{\sigma} \right)^2 \xrightarrow{d} \chi^2_1,$$

was aus dem zentralen Grenzwertsatz zusammen mit der Stetigkeit von $x \mapsto x^2$ folgt. Das ist die Delta-Methode zweiter Ordnung. Das oben angegebene Intervall verliert bei $\mu = 0$ sein Niveau – und weil die Approximation in μ nicht gleichmäßig ist, ist auch bei kleinem $|\mu|$ Vorsicht geboten.

Literatur

Die Kurzformen im Text – etwa „Billingsley, *Probability and Measure*, §22“ – verweisen auf die hier genannten Ausgaben. Die Paragraphen- und Kapitelnummern sind auflagenabhängig; sie beziehen sich auf die jeweils angegebene Auflage.

Literaturverzeichnis

[Bauer] Heinz Bauer: *Wahrscheinlichkeitstheorie.* 5. Auflage, de Gruyter, Berlin 2002.

[Billingsley] Patrick Billingsley: *Probability and Measure.* 3. Auflage, Wiley, New York 1995.

[Durrett] Rick Durrett: *Probability: Theory and Examples.* 5. Auflage, Cambridge University Press, Cambridge 2019.

[Elstrodt] Jürgen Elstrodt: *Maß- und Integrationstheorie.* 8. Auflage, Springer Spektrum, Berlin 2018.

[Feller] William Feller: *An Introduction to Probability Theory and Its Applications, Band II.* 2. Auflage, Wiley, New York 1971.

[Petrov] Valentin V. Petrov: *Sums of Independent Random Variables.* Springer, Berlin 1975.

[Rudin] Walter Rudin: *Real and Complex Analysis.* 3. Auflage, McGraw-Hill, New York 1987.

[Wasserman] Larry Wasserman: *All of Statistics. A Concise Course in Statistical Inference.* Springer, New York 2004. – Vorlage dieses Bandes.

Index