Dr. Jo's dive in:

Ereignis - Zufall - Wahrscheinlichkeit

Inhaltsverzeichnis

Kapitel 1

Ereignisse und Operationen

1.1 Einführung

Wahrscheinlichkeitstheorie (künftig Wtheorie) ist die Mathematik zur Quantifizierung von Unsicherheit. Wir zeigen hier grundlegenden Konzepte und starten mit dem Ergebnisraum als Menge aller möglichen Ausgänge.

1.2 Ergebnisräume und Ereignisse

Der **Ergebnisraum** Ω ist die Menge aller möglichen Ausgänge eines Experiments. Punkte $\omega \in \Omega$ heißen **Ergebnisse**, **Realisierungen** oder **Elemente**. Teilmengen von Ω nennen wir **Ereignisse**.

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 $\Omega = \mathbb{R} = (-\infty, \infty)$. Man könnte argumentieren, dass $\Omega = \mathbb{R}$ ungenau ist, da Temperatur eine untere Grenze besitzt. In der Praxis schadet es jedoch nicht, den Ergebnisraum größer 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

$$E = \{(\omega_1, \omega_2, \omega_3, \ldots) : \omega_1 = Z, \omega_2 = Z, \omega_3 = K, \omega_i \in \{K, Z\} \text{ für } i > 3\}\,.$$

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** von A und B ist

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ oder } \omega \in B\}$$

(„"A oder B). Für eine Folge $A_1, A_2, \ldots$ ist

$$\bigcup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ für mindestens ein } i\}\,.$$

Der **Durchschnitt** ist

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ und } \omega \in B\}$$

(„"A und B). Wir schreiben auch AB statt $A \cap B$. Für eine Folge gilt

$$\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ für alle } i\}\,.$$

Die **Differenz** ist $A - B = \{\omega : \omega \in A, \omega \notin B\}$. 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
ω	Ergebnis (Punkt, Element)
A	Ereignis (Teilmenge von Ω)
A^c	Komplement von A (nicht A)
$A \cup B$	Vereinigung (A oder B)
$A \cap B$	Durchschnitt (A und B)
$A - B$	Mengendifferenz (ω in A, aber nicht in B)
$A \subset B$	Inklusion
$\emptyset$	Leere Menge (unmögliches Ereignis)
Ω	Sicheres Ereignis

Ereignisse $A_1, A_2, \ldots$ heißen **disjunkt** oder **paarweise disjunkt**, falls $A_i \cap A_j = \emptyset$ für $i \neq j$. Eine **Zerlegung** (Partition) von Ω ist eine Folge disjunkter Mengen $A_1, A_2, \ldots$ mit $\bigcup_{i=1}^{\infty} A_i = \Omega$.

Die **Indikatorfunktion** von A ist

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

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 $\bigcup_{i=1}^{\infty} A_i = [0, 1)$ und $\bigcap_{i=1}^{\infty} A_i = \{0\}$. Definiert man stattdessen $A_i = (0, 1/i)$, so ist $\bigcup_{i=1}^{\infty} A_i = (0, 1)$ und $\bigcap_{i=1}^{\infty} A_i = \emptyset$.

Lemma 1

Sei $A_1, A_2, \ldots$ eine monoton wachsende Folge von Ereignissen mit $A = \lim_{n\to\infty} A_n$. Dann gilt

$$P(A) = \lim_{n\to\infty} P(A_n).$$

Analog gilt für eine monoton fallende Folge $P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n)$.

Beweis. Für monoton wachsende Folgen folgt dies direkt aus Theorem 1.8 (Stetigkeit von Wahrscheinlichkeiten). Für monoton fallende Folgen $A_1 \supset A_2 \supset \cdots$ ist $A_1^c \subset A_2^c \subset \cdots$ monoton wachsend mit

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

Mit dem ersten Teil folgt

$$P\left(\bigcap_{i=1}^{\infty} A_i \right) = 1 - P\left(\bigcup_{i=1}^{\infty} A_i^c \right) = 1 - \lim_{n\to\infty} P(A_n^c) = \lim_{n\to\infty} P(A_n). \qquad \square$$

1.3 Wahrscheinlichkeit

Jedem Ereignis A ordnen wir eine reelle Zahl $P(A)$ zu, die **Wahrscheinlichkeit** von A. Wir nennen P auch eine **Wahrscheinlichkeitsverteilung** oder ein **Wahrscheinlichkeitsmaß**.

Definition 1

Eine Funktion P, die jedem Ereignis A eine reelle Zahl $P(A)$ zuordnet, ist eine **Wahrscheinlichkeitsverteilung** oder ein **Wahrscheinlichkeitsmaß**, falls folgende drei Axiome gelten:
Axiom 1: $P(A) \geq 0$ für jedes A
Axiom 2: $P(\Omega) = 1$
Axiom 3: Sind $A_1, A_2, \ldots$ disjunkt, so gilt

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

Es gibt zwei gängige Interpretationen von $P(A)$: die **Häufigkeitsinterpretation** und die **Glaubensgrad-Interpretation**. In der Häufigkeitsinterpretation ist $P(A)$ der langfristige Anteil, mit dem A bei Wiederholungen eintritt. Sagen wir etwa, die Wahrscheinlichkeit für Kopf sei $1/2$, so meinen wir, dass bei vielen Würfen der Anteil von Kopf gegen $1/2$ konvergiert.

In der Glaubensgrad-Interpretation misst $P(A)$ die Stärke der Überzeugung eines Beobachters, dass A wahr ist. In beiden Interpretationen müssen die Axiome 1 bis 3 gelten. Der Unterschied wird erst bei der statistischen Inferenz relevant und führt zu zwei Schulen: der **frequentistischen** und der **Bayesschen** Statistik (siehe Kapitel 11).

Aus den Axiomen folgen viele Eigenschaften:

Satz 2

Für beliebige Ereignisse A, B gilt:

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

Beweis. (1) Wähle $A_1 = \Omega$ und $A_i = \emptyset$ für $i \geq 2$. Da $A_1, A_2, \ldots$ disjunkt sind, folgt mit Axiom 3:

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

Mit Axiom 2 ist $P(\Omega) = 1$, also $1 = 1 + \sum_{i=2}^{\infty} P(\emptyset)$, woraus $P(\emptyset) = 0$ folgt.

(2) Ist $A \subset B$, so ist $B = A \cup (B \cap A^c)$ eine disjunkte Vereinigung. Mit Axiom 3:

$$P(B) = P(A) + P(B \cap A^c) \geq P(A),$$

da $P(B \cap A^c) \geq 0$ nach Axiom 1.

(3) Aus $\emptyset \subset A \subset \Omega$ folgt mit (1), (2) und Axiom 2: $0 = P(\emptyset) \leq P(A) \leq P(\Omega) = 1$.

(4) Da A und A^c disjunkt sind mit $A \cup A^c = \Omega$, folgt:

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c).$$

(5) Folgt direkt aus Axiom 3 für $n = 2$. □

Lemma 3

Für beliebige Ereignisse A und B gilt

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

Beweis. Wir zerlegen $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$ in disjunkte Mengen. Durch wiederholte Anwendung der Additivität ergibt sich

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

Korollar 4

[Bonferroni-Ungleichung] Für beliebige Ereignisse A und B gilt

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

Beweis. Aus dem vorherigen Lemma folgt $P(A \cap B) = P(A) + P(B) - P(A \cup B)$. Da $P(A \cup B) \leq 1$, ergibt sich die Behauptung. □

Korollar 5

[Unionsschranke] Für beliebige Ereignisse $A_1, \ldots, A_n$ gilt

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

Beweis. Beweis durch Induktion. Für $n = 1$ ist die Aussage trivial. Für $n = 2$ folgt aus dem Lemma:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2),$$

da $P(A_1 \cap A_2) \geq 0$. Sei die Aussage für $n - 1$ bewiesen. Dann ist

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= P\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cup A_n\right) \leq P\left(\bigcup_{i=1}^{n-1} A_i\right) + P(A_n) \\ &\leq \sum_{i=1}^{n-1} P(A_i) + P(A_n) = \sum_{i=1}^{n} P(A_i). \end{aligned}$$ □

Beispiel 1.3.1

Zweimaliger Münzwurf. Sei H_1 das Ereignis „"Kopf beim ersten Wurf und H_2 „"Kopf beim zweiten Wurf. Sind alle Ergebnisse gleichwahrscheinlich, so ist

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

Satz 6

[Stetigkeit von Wahrscheinlichkeiten] Gilt $A_n \to A$, so folgt $P(A_n) \to P(A)$ für $n \to \infty$.

Beweis. Sei $A_1 \subset A_2 \subset \cdots$ monoton wachsend und $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$. Definiere $B_1 = A_1$, $B_2 = \{\omega \in \Omega : \omega \in A_2, \omega \notin A_1\}$, $B_3 = \{\omega : \omega \in A_3, \omega \notin A_2, \omega \notin A_1\}$, usw. Dann sind $B_1, B_2, \ldots$ disjunkt, $A_n = \bigcup_{i=1}^{n} B_i$ und $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$. Nach Axiom 3 gilt

$$P(A_n) = P\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} P(B_i)$$

und somit

$$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) = \sum_{i=1}^{\infty} P(B_i) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = P(A). \qquad \square$$

1.4 Wahrscheinlichkeit auf endlichen Ergebnisräumen

Sei $\Omega = \{\omega_1, \ldots, \omega_n\}$ endlich. Beim zweimaligen Würfelwurf hat Ω etwa 36 Elemente: $\Omega = \{(i,j) : i,j \in \{1,\ldots,6\}\}$. Sind alle Ergebnisse gleichwahrscheinlich, so ist $P(A) = |A|/36$, wobei $|A|$ die Anzahl der Elemente von A bezeichnet.

Bei endlichem Ω mit gleichwahrscheinlichen Ergebnissen gilt die **Gleichverteilung**:

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

Um Wahrscheinlichkeiten zu berechnen, müssen wir Elemente zählen – sogenannte **kombinatorische Methoden**. Einige wichtige Fakten:

Die Anzahl der Anordnungen von n Objekten ist $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$. Definitionsgemäß ist $0! = 1$. Ferner ist

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

die Anzahl der Möglichkeiten, k Objekte aus n auszuwählen („"n über k). Beispiel: In einer Klasse von 20 Personen gibt es

$$\binom{20}{3} = \frac{20!}{3! \cdot 17!} = \frac{20 \times 19 \times 18}{3 \times 2 \times 1} = 1140$$

Möglichkeiten, ein 3-Personen-Komitee zu bilden.

Eigenschaften:

$$\binom{n}{0} = \binom{n}{n} = 1, \quad \binom{n}{k} = \binom{n}{n-k}.$$

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 die Wahrscheinlichkeiten, weil wir die Würfe als **unabhängig** betrachten.

Definition 1

Zwei Ereignisse A und B sind **unabhängig**, falls

$$P(A \cap B) = P(A) \cdot P(B),$$

und wir schreiben $A \perp B$. Eine Menge von Ereignissen $\{A_i : i \in I\}$ ist unabhängig, falls für jede endliche Teilmenge $J \subset I$ gilt

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

Sind A und B nicht unabhängig, schreiben wir $A \not\perp B$.

Unabhängigkeit kann auf zwei Arten entstehen: Entweder nehmen wir explizit an, dass Ereignisse unabhängig sind (z. B. beim wiederholten Münzwurf), oder wir verifizieren $P(A \cap B) = P(A) \cdot P(B)$.

Bemerkung 1.5.1. Sind A und B disjunkt mit $P(A), P(B) > 0$, so können sie nicht unabhängig sein, denn $P(A) \cdot P(B) > 0$, aber $P(A \cap B) = P(\emptyset) = 0$.

Beispiel 1.5.1

Faire Münze, 10 Würfe. Sei A = „mindestens ein Kopf und T_j das Ereignis „SZahl beim j-ten Wurf. Dann ist

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

Beispiel 1.5.2

Zwei Personen versuchen abwechselnd, einen Basketball in den Korb zu werfen. Person 1 trifft mit Wahrscheinlichkeit $1/3$, Person 2 mit $1/4$. Wie groß ist die Wahrscheinlichkeit, dass Person 1 zuerst trifft?
Sei E das gesuchte Ereignis und A_j das Ereignis „erster Treffer im j-ten Versuch. Person 1 wirft in den Versuchen 1, 3, 5, . . . , Person 2 in 2, 4, 6, ... Es ist

$$\begin{aligned} P(E) &= P(A_1) + P(A_3) + P(A_5) + \cdots \\ &= \frac{1}{3} + \left(\frac{2}{3}\cdot\frac{3}{4}\right)\cdot\frac{1}{3} + \left(\frac{2}{3}\cdot\frac{3}{4}\right)^2\cdot\frac{1}{3} + \cdots \\ &= \frac{1}{3}\sum_{j=0}^{\infty}\left(\frac{1}{2}\right)^j = \frac{1}{3}\cdot\frac{1}{1-1/2} = \frac{2}{3}. \end{aligned}$$

1.6 Bedingte Wahrscheinlichkeit

Definition 1

Falls $P(B) > 0$, ist die **bedingte Wahrscheinlichkeit** von A gegeben B definiert als

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

Lemma 2

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

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)\cdot P(A_2 \mid A_1)\cdot P(A_3 \mid A_1 \cap A_2)\cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Beweis. Durch wiederholte Anwendung der Definition der bedingten Wahrscheinlichkeit:

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_2 \mid A_1)\cdot P(A_1), \\ P(A_1 \cap A_2 \cap A_3) &= P(A_3 \mid A_1 \cap A_2)\cdot P(A_1 \cap A_2) \\ &= P(A_3 \mid A_1 \cap A_2)\cdot P(A_2 \mid A_1)\cdot P(A_1). \end{aligned}$$

Fortsetzen dieser Argumentation liefert die Behauptung. □

Für festes B mit $P(B) > 0$ definiert $P(\cdot \mid B)$ ein neues Wahrscheinlichkeitsmaß auf Ω. Insbesondere gelten die Axiome:

1. $P(A \mid B) \geq 0$ für alle A,
2. $P(\Omega \mid B) = 1$,
3. Sind $A_1, A_2, \ldots$ disjunkt, so ist $P(\bigcup_{i=1}^{\infty} A_i \mid B) = \sum_{i=1}^{\infty} P(A_i \mid B)$.

Beispiel 1.6.1

[Medizinischer Test] Ein Test für eine Krankheit D hat Ergebnisse $+$ (positiv) und $-$ (negativ). Die Sensitivität ist $P(+ \mid D) = 0{,}993$, die Spezifität $P(- \mid D^c) = 0{,}9999$. Die Prävalenz sei $P(D) = 0{,}0001$. Wie groß ist $P(D \mid +)$?
Nach der Definition der bedingten Wahrscheinlichkeit ist

$$P(D \mid +) = \frac{P(D \cap +)}{P(+)} = \frac{P(+ \mid D) \cdot P(D)}{P(+)}.$$

Mit dem Gesetz der totalen Wahrscheinlichkeit (Satz 1.7) gilt

$$\begin{aligned} P(+) &= P(+ \mid D) \cdot P(D) + P(+ \mid D^c) \cdot P(D^c) \\ &= 0{,}993 \times 0{,}0001 + 0{,}0001 \times 0{,}9999 = 0{,}00019992. \end{aligned}$$

Somit ist

$$P(D \mid +) = \frac{0{,}993 \times 0{,}0001}{0{,}00019992} \approx 0{,}497.$$

Trotz hoher Sensitivität und Spezifität ist die Wahrscheinlichkeit, tatsächlich krank zu sein, nur etwa 50 %, da die Krankheit selten ist.

Lemma 3

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

Beweis. Aus $A \perp B$ folgt $P(A \cap B) = P(A) \cdot P(B)$, also

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

Für die Unabhängigkeit von A und B^c schreiben wir $A = (A \cap B) \cup (A \cap B^c)$ als disjunkte Vereinigung:

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

Also ist $A \perp B^c$. Analog zeigt man $A^c \perp B$.

Für $A^c \perp B^c$ wenden wir das eben Bewiesene auf A^c und B an: Da $A \perp B$, ist auch $A^c \perp B$, und daraus folgt $A^c \perp B^c$. □

Beispiel 1.6.2

Ziehen zweier Karten ohne Zurücklegen. Sei A = „erste Karte ist Herz-As und B = „zweite Karte ist Herz-2. Dann ist

$$P(B \mid A) = \frac{1}{51}, \quad P(B \mid A^c) = \frac{1}{51}.$$

Überraschenderweise ist $P(B \mid A) = P(B \mid A^c)$, also $P(B \mid A) = P(B)$. Somit sind A und B unabhängig.

1.7 Satz von Bayes

Satz 1

[Gesetz der totalen Wahrscheinlichkeit] Sei $A_1, \ldots, A_k$ eine Zerlegung von Ω mit $P(A_i) > 0$ für alle i. Dann gilt für jedes Ereignis B

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

Beweis. Definiere $C_j = B \cap A_j$. Dann sind $C_1, \ldots, C_k$ disjunkt und $B = \bigcup_{j=1}^{k} C_j$. Somit

$$P(B) = \sum_j P(C_j) = \sum_j P(B \cap A_j) = \sum_j P(B \mid A_j) \cdot P(A_j). \qquad \square$$

Satz 2

[Satz von Bayes] Sei $A_1, \ldots, A_k$ eine Zerlegung von Ω mit $P(A_i) > 0$ für alle i. Falls $P(B) > 0$, so gilt für jedes $i \in \{1, \ldots, k\}$

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

Bemerkung 1.7.1. Wir nennen $P(A_i)$ die **A-priori-Wahrscheinlichkeit** (engl. *prior probability*) und $P(A_i \mid B)$ die **A-posteriori-Wahrscheinlichkeit** (engl. *posterior probability*).

Beweis. Zweimalige Anwendung der Definition der bedingten Wahrscheinlichkeit liefert

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

Mit dem Gesetz der totalen Wahrscheinlichkeit folgt die Behauptung. $\square$

Korollar 3

[Bayes für zwei Ereignisse] Sind A und B Ereignisse mit $P(A), P(B) > 0$, so gilt

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

Beweis. Spezialfall des Satzes von Bayes mit der Zerlegung $\{A, A^c\}$. □

Beispiel 1.7.1

[E-Mail-Klassifikation] Ich teile E-Mails in drei Kategorien: A_1 = SSpam", A_2 = niedrige Priorität", A_3 = „"hohe Priorität. Aus Erfahrung weiß ich: $P(A_1) = 0{,}7$, $P(A_2) = 0{,}2$, $P(A_3) = 0{,}1$. Sei B = „Ë-Mail enthält Wort „frei“. Die bedingten Wahrscheinlichkeiten sind:

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

Erhalte ich eine E-Mail mit dem Wort „frei, wie groß ist die Wahrscheinlichkeit, dass es Spam ist? Nach Bayes gilt

$$\begin{aligned} P(A_1 \mid B) &= \frac{P(B \mid A_1) \cdot P(A_1)}{P(B \mid A_1) \cdot P(A_1) + P(B \mid A_2) \cdot P(A_2) + P(B \mid A_3) \cdot P(A_3)} \\ &= \frac{0{,}9 \times 0{,}7}{0{,}9 \times 0{,}7 + 0{,}01 \times 0{,}2 + 0{,}01 \times 0{,}1} \\ &= \frac{0{,}63}{0{,}633} \approx 0{,}995. \end{aligned}$$

Die E-Mail ist mit hoher Wahrscheinlichkeit Spam.

1.8

Ist der Ergebnisraum Ω groß (z. B. $\Omega = \mathbb{R}$), kann man nicht jedem beliebigen Ereignis $A \subset \Omega$ eine Wahrscheinlichkeit zuordnen. Stattdessen beschränkt man sich auf eine Klasse von Mengen, die eine σ**-Algebra** (oder σ-Feld) bildet.

Definition 1

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

1. $\emptyset \in \mathcal{A}$,
2. Ist $A \in \mathcal{A}$, so auch $A^c \in \mathcal{A}$,
3. Sind $A_1, A_2, \ldots \in \mathcal{A}$, so ist $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Die Mengen in $\mathcal{A}$ heißen **messbar**. Das Paar $(\Omega, \mathcal{A})$ nennt man **messbarer Raum**. Ist P ein Wahrscheinlichkeitsmaß auf $\mathcal{A}$, so heißt $(\Omega, \mathcal{A}, P)$ **Wahrscheinlichkeitsraum**.

Ist $\Omega = \mathbb{R}$, wählt man $\mathcal{A}$ als kleinste σ-Algebra, die alle offenen Teilmengen enthält – die

sogenannte **Borel-σ-Algebra**.

Kapitel 2

Zufallsvariablen

2.1 Einführung

In der **Wahrscheinlichkeitstheorie** arbeiten wir zunächst mit einem abstrakten Modell eines Zufallsexperiments:

- Der **Ergebnisraum** (auch Stichprobenraum, Sample Space) Ω ist die Menge **aller möglichen Ergebnisse** des Experiments.
  Beispiel Würfelwurf: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- **Ereignisse** sind Teilmengen von Ω, für die wir Wahrscheinlichkeiten definieren können.
  Beispiel: „gerade Augenzahl“ = $\{2, 4, 6\}$.

Das Modell ist rein mathematisch-abstrakt. Die Elemente von Ω müssen nicht einmal Zahlen sein (z. B. beim Münzwurf: $\Omega = \{\text{Kopf}, \text{Zahl}\}$ oder beim Wetter: $\Omega = \{\text{Sonne}, \text{Regen}, \text{Schnee}, \ldots$

In **Statistik und Data Mining** haben wir jedoch **konkrete Daten** – also Zahlen, Kategorien, Texte, Bilder usw., mit denen wir rechnen, visualisieren und Muster finden wollen.

Die **Zufallsvariable** ist genau das Bindeglied, das die abstrakte Wahrscheinlichkeitstheorie mit den realen Daten verbindet.

Was ist eine Zufallsvariable?

Eine **Zufallsvariable** X ist eine **messbare Funktion**, die jedem möglichen Ereignis $\omega \in \Omega$ einen **beobachtbaren Wert** zuordnet:

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

Sie übersetzt also das zufällige Ereignis in etwas Konkretes, das wir **messen oder beobachten** können.

Beispiel 2.1.1

Würfelwurf
Ergebnisraum $\Omega = \{1, 2, 3, 4, 5, 6\}$ (hier sind die Elemente schon Zahlen, aber das ist Zufall).
Zufallsvariable $X(\omega) = \omega \to X$ ist einfach die Augenzahl.
Daten: Wenn wir 100-mal würfeln, erhalten wir eine Liste von beobachteten Werten $x_1, x_2, \ldots, x_{100}$ (z. B. 3, 5, 1, . . .). Das sind **Realisationen** der Zufallsvariablen X.

Beispiel 2.1.2

Münzwurf
$\Omega = \{\text{Kopf}, \text{Zahl}\}$
Zufallsvariable X: Kopf $\to$ 1, Zahl $\to$ 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, umweltbedingten Faktoren usw.).
Zufallsvariable $X(\omega) =$ Körpergröße dieser Person in cm.
Wir beobachten nur die Werte von X (z. B. 171, 168, 182, ... cm). Den zugrunde liegenden Ergebnisraum Ω sehen wir nie direkt.

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

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

Zusammengefasst:
Der Ergebnisraum und die 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.

Statistik und Data Mining befassen sich mit Daten. Wie verbinden wir Ergebnisräume und Ereignisse mit Daten? Das Bindeglied ist die **Zufallsvariable**.

Definition 1

Eine **Zufallsvariable** ist eine Abbildung

$$X : \Omega \to \mathbb{R},$$

die jedem Ergebnis ω eine reelle Zahl $X(\omega)$ zuordnet.

Ab einem gewissen Punkt in den meisten Wahrscheinlichkeitsvorlesungen wird der Er-

gebnisraum kaum noch erwähnt und wir arbeiten direkt mit Zufallsvariablen. Man sollte jedoch im Hinterkopf behalten, dass der Ergebnisraum stets vorhanden ist.

Beispiel 2.1.4

Zehnmaliger Münzwurf. Sei $X(\omega)$ die Anzahl der Köpfe in der Sequenz ω. Beispiel: Ist $\omega = KKZKKZKKZZ$, so ist $X(\omega) = 6$.

Beispiel 2.1.5

Sei $\Omega = \{(x, y) : x^2+y^2 \leq 1\}$ die Einheitskreisscheibe. Wählen wir zufällig einen Punkt aus Ω. Ein typisches Ergebnis hat die Form $\omega = (x, y)$. Beispiele für Zufallsvariablen sind $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$ und $W(\omega) = x^2 + y^2$.

Für eine Zufallsvariable X und eine Teilmenge $A \subset \mathbb{R}$ definieren wir $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$ und setzen

$$P(X \in A) = P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\}),$$
$$P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) = x\}).$$

Beachten Sie: X bezeichnet die Zufallsvariable, x einen konkreten Wert.

Beispiel 2.1.6

Zweimaliger Münzwurf, X = Anzahl der Köpfe. Dann ist $P(X = 0) = P(\{ZZ\}) = 1/4$, $P(X = 1) = P(\{KZ, ZK\}) = 1/2$ und $P(X = 2) = P(\{KK\}) = 1/4$. Zusammenfassung:

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

x	$P(X = x)$
0	1/4
1	1/2
2	1/4

2.2 Verteilungsfunktionen und Wahrscheinlichkeitsfunktionen

Definition 1

Die **kumulative Verteilungsfunktion** (engl. *cumulative distribution function, cdf*) ist die Funktion $F_X : \mathbb{R} \to [0, 1]$ definiert durch

$$F_X(x) = P(X \leq x).$$

Die cdf enthält alle Informationen über die Zufallsvariable. Oft schreiben wir kurz F statt F_X.

Beispiel 2.2.1

Zweimaliger fairer Münzwurf, X = Anzahl Köpfe. Dann ist $P(X = 0) = P(X = 2) = 1/4$ und $P(X = 1) = 1/2$. Die Verteilungsfunktion lautet

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

Die Funktion ist rechtsseitig stetig, monoton wachsend und für alle x definiert, obwohl X nur Werte $0, 1, 2$ annimmt. Warum ist $F_X(1{,}4) = 0{,}75$?

Satz 2

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

Beweis. Es genügt zu zeigen, dass $P(X \in A) = P(Y \in A)$ für alle Intervalle A gilt, da sich alle Borel-Mengen aus Intervallen erzeugen lassen. Für ein Intervall $A = (-\infty, x]$ ist per Definition

$$P(X \in A) = F(x) = G(x) = P(Y \in A).$$

Durch Mengendifferenzen und abzählbare Vereinigungen solcher Intervalle folgt die Aussage für alle Borel-Mengen. □

Satz 3

Eine Funktion $F : \mathbb{R} \to [0, 1]$ ist genau dann die kumulative Verteilungsfunktion (CDF) eines Wahrscheinlichkeitsmaßes P auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, d. h.

$$F(x) = P((-\infty, x]),$$

wenn sie die folgenden Eigenschaften erfüllt:

1. **Monoton nichtfallend:** $x_1 < x_2 \implies F(x_1) \le F(x_2)$,
2. **Normierung:** $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$,
3. **Rechtsstetig:** $F(x) = \lim_{y\downarrow x} F(y)$ für alle $x \in \mathbb{R}$.

Beweis. **„⇒“ (CDF ⇒ Eigenschaften).**
Sei $F(x) = P((-\infty, x])$ für ein Wahrscheinlichkeitsmaß P.

- Monotonie: $x_1 < x_2 \implies (-\infty, x_1] \subset (-\infty, x_2] \implies F(x_1) \le F(x_2)$.

- Normierung: $\lim_{x\to\infty} F(x) = P(\mathbb{R}) = 1$ und $\lim_{x\to-\infty} F(x) = P(\emptyset) = 0$ (da $\bigcap_n (-\infty, x_n] = \emptyset$ für $x_n \to -\infty$).

- Rechtsstetigkeit: $\lim_{y\downarrow x} F(y) = \lim_{y\downarrow x} P((-\infty, y]) = P((-\infty, x])$ (Stetigkeit von oben für die abnehmende Folge $(-\infty, y]$, $y \downarrow x$).

„⇐“ (Eigenschaften ⇒ existiert eindeutiges P).
Definiere für $a < b$

$$\mu((a, b]) := F(b) - F(a).$$

(Die Monotonie sichert $\mu \geq 0$; für $a = -\infty$ bzw. $b = \infty$ verwenden wir die Grenzwerte.)

Die Menge $\mathcal{S} = \{(-\infty, x] : x \in \mathbb{R}\} \cup \{(a, b] : a < b\} \cup \{\emptyset, \mathbb{R}\}$ ist ein Semiring, und μ ist auf disjunkten endlichen Vereinigungen additiv (folgt aus Monotonie und Rechtsstetigkeit).

Wegen der Rechtsstetigkeit ist μ sogar σ-additiv auf dem von $\mathcal{S}$ erzeugten Ring (das ist der kritische Schritt: die Rechtsstetigkeit verhindert „Massensprünge“ und garantiert die σ-Additivität bei abzählbaren disjunkten Vereinigungen von Halbintervallen).

Nach dem **Carathéodoryschen Erweiterungssatz** existiert eine eindeutige Erweiterung von μ zu einem Wahrscheinlichkeitsmaß P auf der σ-Algebra $\mathcal{B}(\mathbb{R})$.

Schließlich gilt per Konstruktion

$$P((-\infty, x]) = \lim_{y\downarrow x} \mu((-\infty, y]) = \lim_{y\downarrow x} F(y) = F(x)$$

(Rechtsstetigkeit), also ist F die CDF von P.

Eindeutigkeit folgt daraus, dass die Halbintervalle $(-\infty, x]$ ein π-System sind, das die Borel-σ-Algebra erzeugt, und zwei Maße, die auf einem erzeugenden π-System übereinstimmen, sind gleich (Eindeutigkeitssatz für Maße). □

Motivation

Das **äußere Maß** (outer measure) ist ein zentrales Konzept der Maßtheorie, das 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 4

[Äußeres Maß aus einem Prämaß] Sei X eine Menge und $\mathcal{R}$ ein Ring (oder Semiring) von Teilmengen von X. Sei $\mu_0 : \mathcal{R} \to [0, \infty]$ ein Prämaß (d. h. $\mu_0(\emptyset) = 0$ und σ-additiv auf disjunkten Vereinigungen in $\mathcal{R}$).
Das zugehörige **äußere Maß** $\mu^* : \mathcal{P}(X) \to [0, \infty]$ ist definiert durch

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

wobei das Infimum über alle abzählbaren Überdeckungen von E durch Mengen aus $\mathcal{R}$ genommen wird. (Falls keine solche Überdeckung existiert, setzt man $\mu^*(E) = \infty$. Die leere Summe ist 0, also $\mu^*(\emptyset) = 0$.)

Allgemeiner kann man ein äußeres Maß direkt axiomatisch definieren:

Definition 5

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

1. $\mu^*(\emptyset) = 0$,
2. **Monotonie:** $E \subset F \implies \mu^*(E) \leq \mu^*(F)$,
3. **abzählbare Subadditivität:** Für beliebige $E_n \subset X$ gilt

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

Wichtige Eigenschaften

Korollar 6

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

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

Bemerkungen

- Das äußere Maß ist im Allgemeinen **nicht additiv**, sondern nur subadditiv. Additivität gilt nur für messbare Mengen (im Sinne von Carathéodory).

- Beispiel: Das **Lebesgue-äußere Maß** auf $\mathbb{R}^d$ entsteht aus der Längen-/Volumenfunktion auf Quadern/Halbintervallen.

- Im Kontext von Wahrscheinlichkeitsmaßen: Aus der durch eine CDF F definierten $\mu_0((a, b]) = F(b) - F(a)$ entsteht ein äußeres Maß, das dann via Carathéodory zu einem Wahrscheinlichkeitsmaß auf den Borelmengen eingeschränkt wird.

Der Carathéodorysche Erweiterungssatz

Lemma 7

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

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

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. Falls μ_0 ein Prämaß auf einem Ring $\mathcal{R}$ war und μ^* daraus konstruiert wurde, so ist μ eine Erweiterung von μ_0 (d. h. $\mu|_{\mathcal{R}} = \mu_0$).
4. Falls zusätzlich μ_0 σ-endlich ist (d. h. $X = \bigcup_n X_n$ mit $\mu_0(X_n) < \infty$), so ist die Erweiterung μ auf der von $\mathcal{R}$ erzeugten σ-Algebra $\sigma(\mathcal{R})$ **eindeutig**.

Bemerkungen

- Der kritische Punkt ist der Nachweis, dass $\mathcal{M}$ eine σ-Algebra ist (insbesondere σ-Additivität der messbaren Mengen). Dies erfordert die 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 $\rightarrow$ 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.

Der vollständige Beweis (insbesondere der σ-Algebra-Nachweis) findet sich in Standardwerken wie Billingsley *Probability and Measure*, Bauer *Wahrscheinlichkeitstheorie* oder Elstrodt *Maß- und Integrationstheorie*.

Diskrete Zufallsvariablen

Definition 8

X ist **diskret**, falls sie abzählbar viele Werte $\{x_1, x_2, \ldots\}$ annimmt. Die **Wahrscheinlichkeitsfunktion** (engl. *probability mass function, pmf*) ist

$$f_X(x) = P(X = x).$$

Lemma 9

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

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

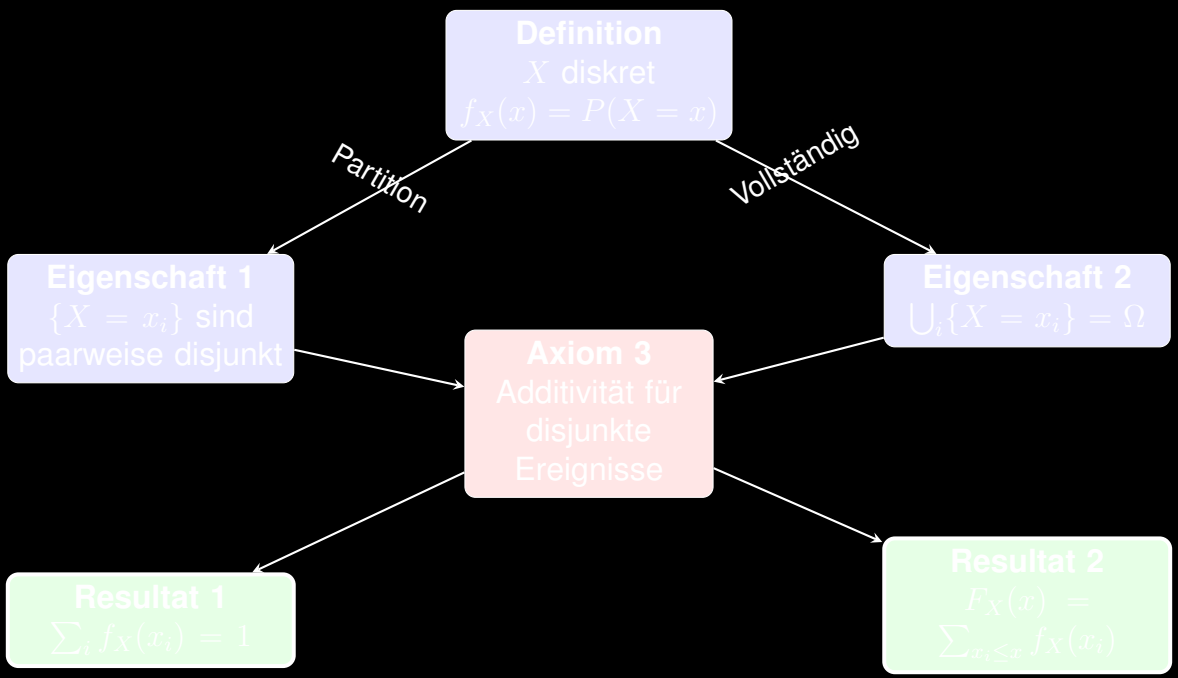
Beweis. (1) Die Ereignisse $\{X = x_i\}$ für $i = 1, 2, \ldots$ bilden eine Partition von Ω, d. h. sie sind paarweise disjunkt und ihre Vereinigung ist Ω. Mit Axiom 3 der Wahrscheinlichkeitstheorie folgt:

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

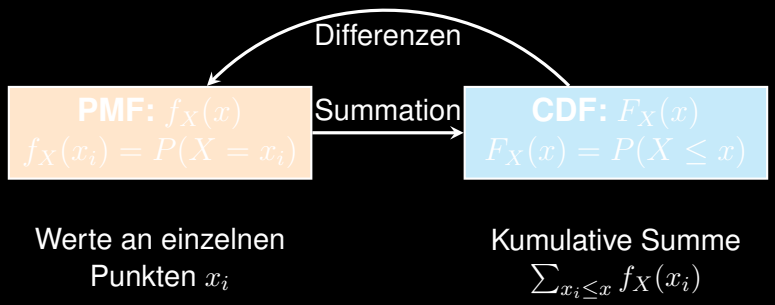
(2) Das Ereignis $\{X \leq x\}$ ist die disjunkte Vereinigung aller Ereignisse $\{X = x_i\}$ mit $x_i \leq x$. Mit Axiom 3 folgt:

$$F_X(x) = P(X \leq x) = P\left(\bigcup_{x_i \leq x} \{X = x_i\}\right) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} f_X(x_i). \qquad \square$$

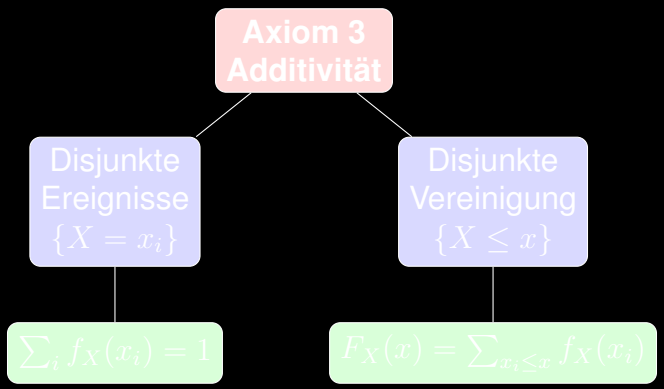
Visuelle Struktur der Beweisführung



Beziehung zwischen pmf und cdf



Axiomatische Struktur



Beispiel 2.2.2

2.2 Zweimaliger fairer Münzwurf, $X =$ Anzahl der 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):

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

CDF aus Lemma (2):

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4} & 0 \le x < 1 \\ \frac{3}{4} & 1 \le x < 2 \\ 1 & x \ge 2 \end{cases}$$

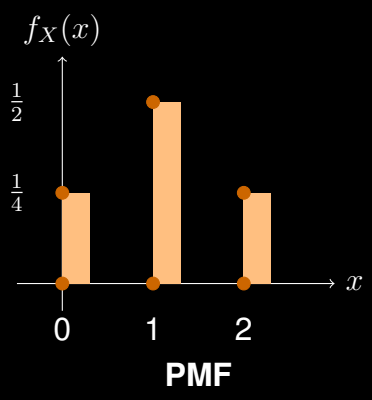
Berechnung:

$$F_X(0) = \sum_{x_i \le 0} f_X(x_i) = f_X(0) = \frac{1}{4}$$

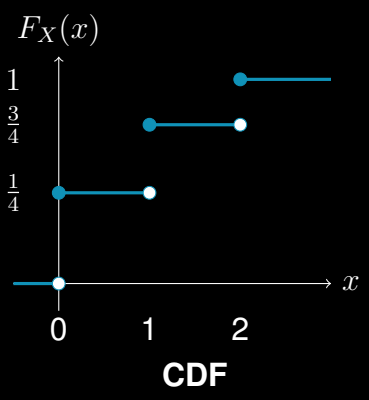
$$F_X(1) = \sum_{x_i \le 1} f_X(x_i) = f_X(0) + f_X(1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

$$F_X(2) = \sum_{x_i \le 2} f_X(x_i) = f_X(0) + f_X(1) + f_X(2) = 1$$

Visualisierung: PMF und CDF



PMF



CDF

Stetige Zufallsvariablen

Definition 10

X ist **stetig**, falls es eine Funktion f_X gibt, sodass $f_X(x) \geq 0$ für alle x, $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ und für jedes $A \subset \mathbb{R}$ gilt

$$P(X \in A) = \int_A f_X(x)\,dx.$$

Die Funktion f_X heißt **Wahrscheinlichkeitsdichtefunktion** oder **Dichte** (engl. *probability density function, pdf*).

Für stetige Zufallsvariablen gilt $P(X = x) = 0$ für alle x und

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

an allen Stellen, wo F_X differenzierbar ist.

Beispiel 2.2.3

Sei X mit pdf

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

Dies ist die Gleichverteilung auf $(0,1)$, geschrieben $X \sim \text{Uniform}(0,1)$. Die cdf ist

$$F_X(x) = \begin{cases} 0 & x < 0, \\ x & 0 \leq x \leq 1, \\ 1 & x > 1. \end{cases}$$

Beispiel 2.2.4

Sei $f_X(x) = 3x^2$ für $0 < x < 1$ und $f_X(x) = 0$ sonst. Dann ist

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \begin{cases} 0 & x < 0, \\ x^3 & 0 \leq x \leq 1, \\ 1 & x > 1. \end{cases}$$

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

Lemma 11

Sei F die cdf einer Zufallsvariable X. Dann gilt:

1. $P(X = x) = F(x) - \lim_{y \uparrow x} F(y)$,
2. $P(x < X \leq y) = F(y) - F(x)$,
3. $P(X > x) = 1 - F(x)$,
4. Falls X stetig ist, gilt $P(X = x) = 0$ für alle x.

Definition 12

Sei X eine Zufallsvariable mit cdf F. Das **inverse cdf** oder **Quantilfunktion** ist

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

für $q \in [0, 1]$. Falls F streng monoton wachsend und stetig ist, dann ist $F^{-1}(q)$ die eindeutige Lösung von $F(x) = q$.

2.3 Wichtige diskrete Verteilungen

Punktmasse (Dirac-Verteilung)

Definition 1

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

$$P(X = a) = 1.$$

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

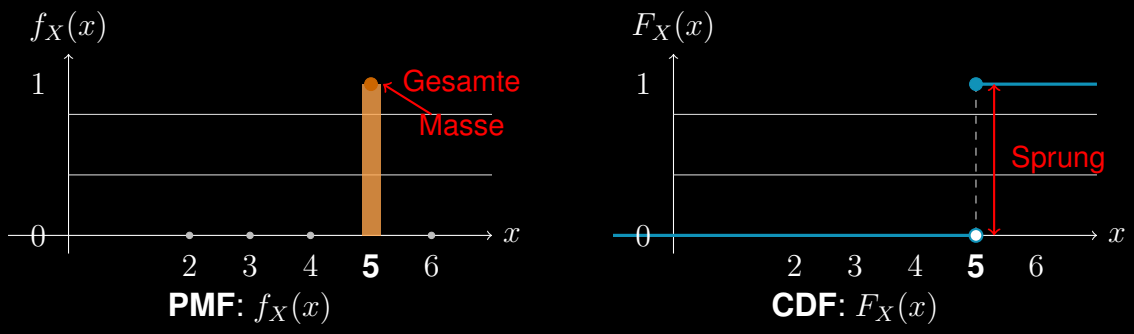
Eine Zufallsvariable X, die den Wert 5 mit Sicherheit annimmt, also $X \sim \delta_5$.

- Die „zufällige"Wahl einer festen Zahl
- Eine Konstante als Zufallsvariable betrachtet
- $P(X = 5) = 1$, $P(X \neq 5) = 0$

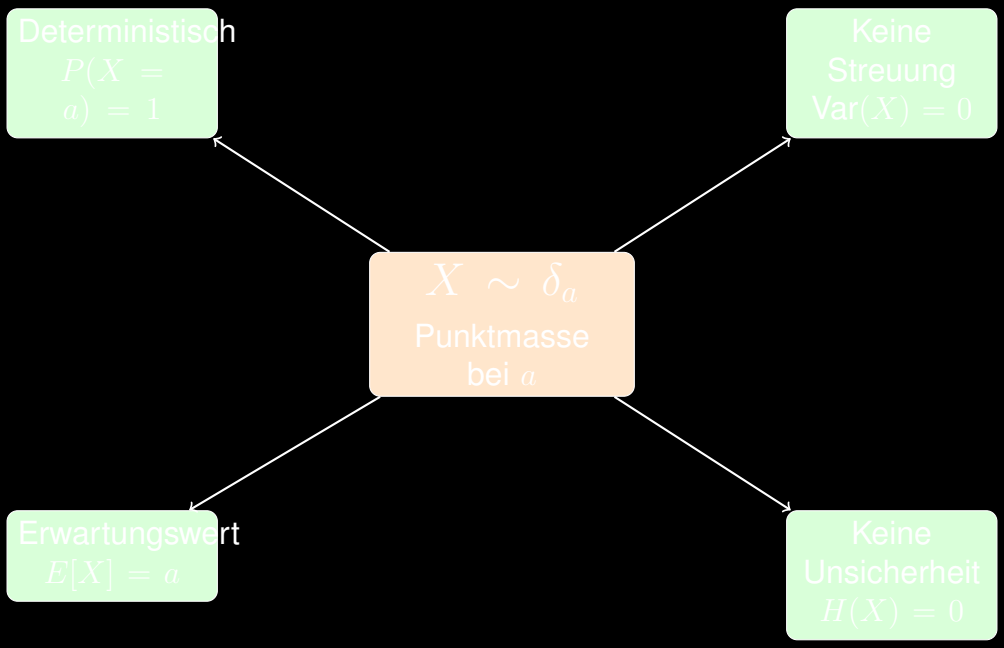
Eigenschaften

- **Erwartungswert:** $E[X] = a$
- **Varianz:** $\text{Var}(X) = 0$ (keine Streuung!)
- **Entropie:** $H(X) = 0$ (keine Unsicherheit)
- **Grenzfall:** Deterministische Variable (keine echte Zufälligkeit)

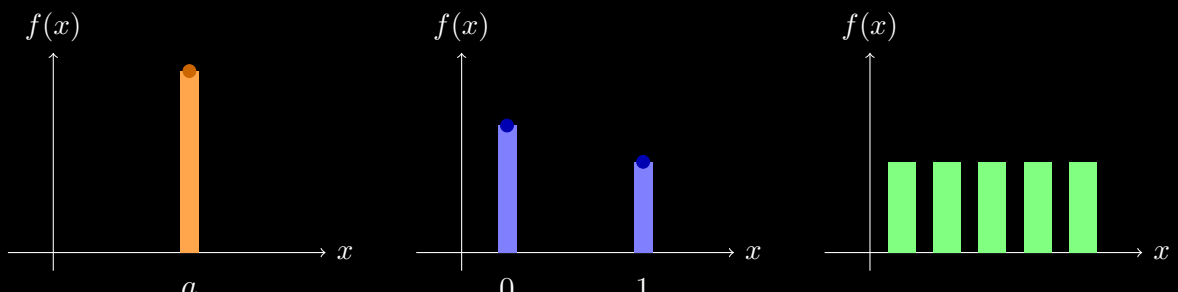
Visualisierung: PMF und CDF für δ_5



Konzeptuelle Darstellung



Vergleich: Punktmasse vs. andere diskrete Verteilungen



Mathematische Eigenschaften im Detail

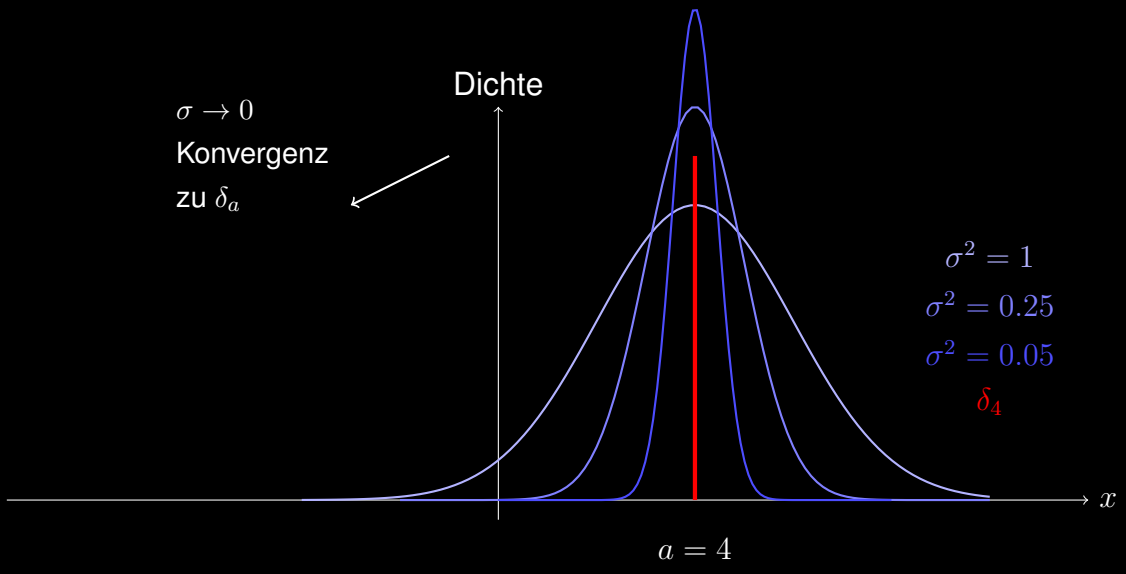
Eigenschaft	**Wert für δ_a**
Träger (Support)	$\{a\}$
Erwartungswert	$E[X] = a$
Varianz	$\text{Var}(X) = 0$
Standardabweichung	$\sigma_X = 0$
Schiefe (Skewness)	Nicht definiert
Kurtosis	Nicht definiert
MGF	$M_X(t) = e^{at}$
Charakteristische Funktion	$\phi_X(t) = e^{iat}$

Anwendungen

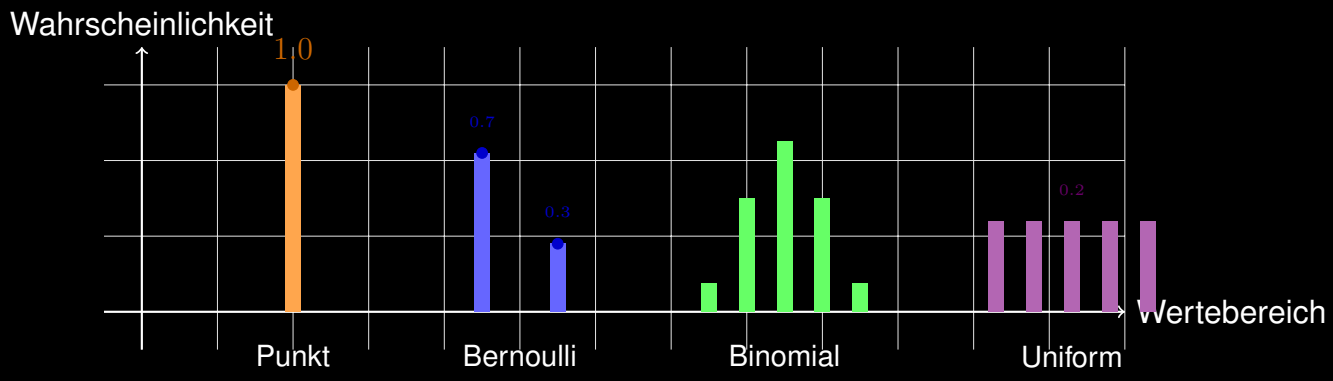
- **Modellierung von Sicherheit:** Wenn ein Ereignis mit Sicherheit eintritt
- **Grenzfälle:** Approximation durch sehr konzentrierte Verteilungen
- **Dirac-Delta-Funktion:** Kontinuierliche Verallgemeinerung in der Analysis
- **Bayessche Statistik:** Prior-Verteilung bei exaktem Vorwissen
- **Anfangsbedingungen:**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

Konvergenz zu Punktmasse

Betrachte eine Folge von Normalverteilungen mit schrumpfender Varianz:



Vergleich: PMF-Darstellung verschiedener Verteilungen



Diskrete Gleichverteilung

Definition: X nimmt endlich viele Werte $\{x_1, \ldots, x_k\}$ an mit gleicher Wahrscheinlichkeit $P(X = x_i) = \frac{1}{k}$ für alle i.

Beispiel: Würfelwurf mit einem fairen sechsseitigen Würfel: Werte $\{1, 2, 3, 4, 5, 6\}$, jeweils mit Wahrscheinlichkeit $1/6$.

Bernoulli-Verteilung

Definition: $X \in \{0, 1\}$ mit $P(X = 1) = p$ und $P(X = 0) = 1 - p$, $p \in [0, 1]$. PMF: $f(x) = p^x(1 - p)^{1-x}$ für $x \in \{0, 1\}$. Modelliert einen einzelnen Versuch mit zwei Ausgängen (Erfolg/Misserfolg).

Beispiel: Münzwurf mit einer fairen Münze: $p = 0{,}5$ für Kopf ($X = 1$) bzw. Zahl ($X = 0$).

Binomialverteilung

Definition: $X \sim \text{Binomial}(n, p)$ ist die Anzahl der Erfolge in n unabhängigen Bernoulli-Versuchen mit Erfolgswahrscheinlichkeit p. PMF:

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

Beispiel: Anzahl der Köpfe bei $n = 10$ Würfen einer fairen Münze ($p = 0{,}5$).

Geometrische Verteilung

Definition: $X \sim \text{Geometrisch}(p)$ ist die Anzahl der Versuche bis zum ersten Erfolg in einer Sequenz unabhängiger Bernoulli-Versuche mit Erfolgswahrscheinlichkeit p. PMF:

$$P(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \ldots$$

Beispiel: Anzahl der Würfe mit einem fairen Würfel ($p = 1/6$ für eine Sechs), bis die erste Sechs erscheint.

Poisson-Verteilung

Definition: $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$, modelliert die Anzahl seltener Ereignisse in einem festen Intervall. PMF:

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Approximiert $\text{Binomial}(n, p)$ für großes n und kleines p mit $np = \lambda$.

Beispiel: Anzahl der eingehenden Anrufe in einem Callcenter pro Stunde, wenn im Mittel $\lambda = 4$ Anrufe pro Stunde erwartet werden.

2.4 Wichtige stetige Verteilungen

Gleichverteilung (Uniform)

$X \sim \text{Uniform}(a, b)$ mit pdf

$$f(x) = \frac{1}{b-a}, \quad a < x < b.$$

Normalverteilung (Gauß-Verteilung)

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

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

μ ist der Erwartungswert, σ^2 die Varianz. $N(0, 1)$ ist die Standardnormalverteilung.

Exponentialverteilung

$X \sim \text{Exp}(\beta)$ mit $\beta > 0$ hat die pdf

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0.$$

Die Exponentialverteilung ist **gedächtnislos**: $P(X > s + t \mid X > s) = P(X > t)$ für alle $s, t > 0$.

Gamma-Verteilung

$X \sim$ Gamma(α, β) mit $\alpha, \beta > 0$ hat die pdf

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0,$$

wobei $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$ die Gamma-Funktion ist.

Beta-Verteilung

$X \sim$ Beta(α, β) mit $\alpha, \beta > 0$ hat die pdf

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1.$$

t-Verteilung

$X \sim t_\nu$ mit ν Freiheitsgraden hat die pdf

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

χ^2-Verteilung

$X \sim \chi^2_\nu$ mit ν Freiheitsgraden ist ein Spezialfall der Gamma-Verteilung: $\chi^2_\nu =$ Gamma$(\nu/2, 2)$.

2.5 Bivariate Verteilungen

Definition 1

Die **gemeinsame cdf** von (X, Y) ist

$$F_{X,Y}(x, y) = P(X \le x, Y \le y).$$

Definition 2

Im diskreten Fall ist die **gemeinsame pmf**

$$f_{X,Y}(x, y) = P(X = x, Y = y).$$

Im stetigen Fall ist die **gemeinsame pdf** eine Funktion $f_{X,Y}$ mit $f_{X,Y}(x, y) \ge 0$, $\iint f_{X,Y}(x, y)\, dx\, dy = 1$ und

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

Beispiel 2.5.1

Werfen zwei faire Würfel. Sei X das Minimum und Y das Maximum. Dann ist

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

Beispiel 2.5.2

Sei (X, Y) gleichverteilt auf dem Einheitsquadrat $[0,1] \times [0,1]$. Dann ist

$$f_{X,Y}(x,y) = 1, \quad 0 < x, y < 1.$$

2.6 Randverteilungen

Definition 1

Im diskreten Fall sind die Randdichten

$$f_X(x) = \sum_y f_{X,Y}(x,y), \quad f_Y(y) = \sum_x f_{X,Y}(x,y).$$

Beispiel 2.6.1

Für die Tabelle aus Beispiel 2.18 ist

$$f_X(0) = \frac{1}{9} + \frac{2}{9} + 0 = \frac{1}{3}, \quad f_X(1) = \frac{1}{3} + 0 + \frac{1}{9} = \frac{4}{9}, \quad f_X(2) = \frac{1}{18} + \frac{1}{9} + \frac{1}{18} = \frac{2}{9}.$$

Definition 2

Im stetigen Fall sind die Randdichten

$$f_X(x) = \int f_{X,Y}(x,y)\,dy, \quad f_Y(y) = \int f_{X,Y}(x,y)\,dx.$$

Beispiel 2.6.2

Sei $f_{X,Y}(x,y) = \frac{6}{5}(x + y^2)$ für $0 \le x, y \le 1$. Dann ist

$$f_X(x) = \int_0^1 \frac{6}{5}(x + y^2)\,dy = \frac{6}{5}\left[xy + \frac{y^3}{3}\right]_0^1 = \frac{6}{5}\left(x + \frac{1}{3}\right), \quad 0 < x < 1.$$

2.7 Unabhängige Zufallsvariablen

Definition 1

Zwei Zufallsvariablen X und Y sind **unabhängig**, geschrieben $X \perp Y$, falls für alle $A, B \subset \mathbb{R}$ gilt

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B).$$

Satz 2

Seien X und Y mit gemeinsamer pdf $f_{X,Y}$. Dann gilt $X \perp Y$ genau dann, wenn

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

für alle x, y.

Beweis. ($\Rightarrow$) Angenommen $X \perp Y$. Dann gilt für alle A, B:

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B).$$

Wähle $A = (-\infty, x]$ und $B = (-\infty, y]$. Differentiation nach x und y ergibt $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$.

($\Leftarrow$) Sei $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$. Dann ist für beliebige Mengen A, B:

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x, y)\, dy\, dx \\ &= \int_A \int_B f_X(x) f_Y(y)\, dy\, dx \\ &= \int_A f_X(x)\, dx \cdot \int_B f_Y(y)\, dy \\ &= P(X \in A) \cdot P(Y \in B). \end{aligned}$$

□

Beispiel 2.7.1

Seien X und Y unabhängig mit $X \sim \text{Uniform}(0, 1)$ und $Y \sim \text{Exp}(1)$. Dann ist

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) = 1 \cdot e^{-y} = e^{-y}$$

für $0 < x < 1$ und $y > 0$.

Satz 3

Sind die Wertebereiche von X und Y ein (möglicherweise unendliches) Rechteck $\mathcal{X} \times \mathcal{Y}$ und lässt sich $f_{X,Y}$ in der Form

$$f_{X,Y}(x,y) = g(x)h(y)$$

schreiben, wobei g eine Funktion nur von x und h eine Funktion nur von y ist, dann sind X und Y unabhängig.

Beweis. Aus der Faktorisierung $f_{X,Y}(x,y) = g(x)h(y)$ folgt für die Randdichten:

$$f_X(x) = \int_{\mathcal{Y}} g(x)h(y)\,dy = g(x)\int_{\mathcal{Y}} h(y)\,dy = g(x)\cdot c_1,$$
$$f_Y(y) = \int_{\mathcal{X}} g(x)h(y)\,dx = h(y)\int_{\mathcal{X}} g(x)\,dx = h(y)\cdot c_2,$$

wobei c_1, c_2 Konstanten sind. Wegen $\int_{\mathcal{X}}\int_{\mathcal{Y}} f_{X,Y}(x,y)\,dy\,dx = 1$ ist $c_1 \cdot c_2 = 1$. Somit

$$f_{X,Y}(x,y) = g(x)h(y) = \frac{g(x)}{c_2}\cdot\frac{h(y)}{c_1}\cdot c_1c_2 = f_X(x)\cdot f_Y(y).$$

Nach dem vorherigen Theorem sind X und Y unabhängig. □

2.8 Bedingte Verteilungen

Definition 1

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

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

falls $f_X(x) > 0$.

Definition 2

Im stetigen Fall ist die **bedingte pdf** von Y gegeben $X = x$

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

falls $f_X(x) > 0$.

Beispiel 2.8.1

Gleichverteilung auf dem Einheitsquadrat: $f_{X,Y}(x,y) = 1$ für $0 < x, y < 1$. Die Randdichte ist $f_X(x) = \int_0^1 1\,dy = 1$ für $0 < x < 1$. Somit ist

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

für $0 < y < 1$. Also ist $Y \mid X = x \sim \text{Uniform}(0,1)$.

Beispiel 2.8.2

Sei $X \sim \text{Uniform}(0,1)$. Nach Beobachtung von $X = x$ wählen wir $Y \sim \text{Uniform}(0,x)$. Die gemeinsame Dichte ist

$$f_{X,Y}(x,y) = f_X(x) \cdot f_{Y|X}(y \mid x) = 1 \cdot \frac{1}{x} = \frac{1}{x}$$

für $0 < y < x < 1$. Die Randdichte von Y ist

$$f_Y(y) = \int_y^1 \frac{1}{x}\,dx = [-\ln x]_y^1 = -\ln y$$

für $0 < y < 1$.

2.9 Multivariate Verteilungen und iid-Stichproben

Für n Zufallsvariablen $X_1, \ldots, X_n$ ist die gemeinsame pdf

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

Definition 1

Falls $X_1, \ldots, X_n$ unabhängig sind und jede die gleiche Randverteilung f hat, nennen wir sie **unabhängig und identisch verteilt** (engl. *independent and identically distributed, iid*) und schreiben $X_1, \ldots, X_n \sim f$. Dann ist

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

2.10 Zwei wichtige multivariate Verteilungen

Multinomialverteilung

Verallgemeinerung der Binomialverteilung. Haben n Versuche mit k möglichen Ausgängen und Wahrscheinlichkeiten $p_1, \ldots, p_k$ (mit $\sum_{i=1}^k p_i = 1$), so ist $X = (X_1, \ldots, X_k) \sim$

Multinomial(n, p) mit

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

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

Multivariate Normalverteilung

Ein Zufallsvektor $X = (X_1, \ldots, X_k)^T$ hat eine **multivariate Normalverteilung** $X \sim N(\mu, \Sigma)$, falls die pdf

$$f(x) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

ist, wobei $\mu \in \mathbb{R}^k$ der Erwartungswertvektor und Σ die $k \times k$ Kovarianzmatrix (positiv definit) ist.

Satz 1

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

Beweis. Sei $A = \Sigma^{1/2}$. Die Transformation $X = \mu + AZ$ hat die Umkehrung $Z = A^{-1}(X - \mu)$ mit Jacobi-Determinante $|J| = |\det(A^{-1})| = |\Sigma|^{-1/2}$. Die pdf von Z ist

$$f_Z(z) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2} z^T z\right).$$

Mit der Transformationsformel folgt:

$$\begin{aligned} f_X(x) &= f_Z(A^{-1}(x-\mu)) \cdot |J| \\ &= \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}(x-\mu)^T (A^{-1})^T A^{-1} (x-\mu)\right) \cdot |\Sigma|^{-1/2} \\ &= \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right), \end{aligned}$$

da $(A^{-1})^T A^{-1} = (AA^T)^{-1} = \Sigma^{-1}$. Dies ist die Dichte von $N(\mu, \Sigma)$. □

2.11 Transformationen von Zufallsvariablen

Sei $Y = g(X)$ für eine Funktion $g : \mathbb{R} \to \mathbb{R}$.

Diskreter Fall

Ist X diskret mit pmf f_X, so hat Y die pmf

$$f_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x : g(x) = y} f_X(x).$$

Beispiel 2.11.1

Sei $P(X=-1)=P(X=1)=1/4$ und $P(X=0)=1/2$. Sei $Y=X^2$. Dann ist

$$f_Y(0)=P(Y=0)=P(X=0)=1/2, \quad f_Y(1)=P(Y=1)=P(X=-1)+P(X=1)=1/2.$$

Stetiger Fall

Ist g streng monoton wachsend oder fallend, so hat Y die pdf

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

Beispiel 2.11.2

Sei $X\sim$ Uniform$(-1,3)$ mit $f_X(x)=1/4$ für $-1<x<3$. Sei $Y=X^2$. Der Wertebereich von Y ist $[0,9]$. Für $0<y<9$ gibt es zwei Lösungen: $x=\pm\sqrt{y}$.

- Für $0<y<1$: beide Werte liegen in $(-1,3)$, also

$$f_Y(y)=f_X(\sqrt{y})\cdot\frac{1}{2\sqrt{y}}+f_X(-\sqrt{y})\cdot\frac{1}{2\sqrt{y}}=\frac{1}{4}\cdot\frac{1}{2\sqrt{y}}+\frac{1}{4}\cdot\frac{1}{2\sqrt{y}}=\frac{1}{4\sqrt{y}}.$$

- Für $1\le y<9$: nur $\sqrt{y}\in(-1,3)$, also

$$f_Y(y)=\frac{1}{4}\cdot\frac{1}{2\sqrt{y}}=\frac{1}{8\sqrt{y}}.$$

2.12 Transformationen mehrerer Zufallsvariabler

Seien X_1,X_2 mit gemeinsamer pdf f_{X_1,X_2} und $Y_1=g_1(X_1,X_2)$, $Y_2=g_2(X_1,X_2)$. Falls die Transformation bijektiv ist mit Umkehrung $X_1=h_1(Y_1,Y_2)$, $X_2=h_2(Y_1,Y_2)$, so ist die pdf von (Y_1,Y_2)

$$f_{Y_1,Y_2}(y_1,y_2)=f_{X_1,X_2}(h_1(y_1,y_2),h_2(y_1,y_2))|J|,$$

wobei J die Determinante der Jacobi-Matrix ist:

$$J=\det\begin{pmatrix}\frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2}\\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2}\end{pmatrix}.$$

Beispiel 2.12.1

Seien $X_1, X_2 \sim \text{Uniform}(0,1)$ unabhängig. Sei $Y_1 = X_1 + X_2$ und $Y_2 = X_1 - X_2$. Die Umkehrung ist $X_1 = (Y_1 + Y_2)/2$, $X_2 = (Y_1 - Y_2)/2$. Die Jacobi-Determinante ist

$$J = \det \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix} = -1/2,$$

also $|J| = 1/2$. Die gemeinsame pdf ist

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

im Bildbereich, d. h. für $0 < (y_1 + y_2)/2 < 1$ und $0 < (y_1 - y_2)/2 < 1$, also $|y_2| < y_1 < 2 - |y_2|$.

2.13 Anhang

Technisch gesehen muss eine Zufallsvariable **messbar** sein, d. h. für jede Borel-Menge $B \subset \mathbb{R}$ muss $X^{-1}(B) = \{\omega : X(\omega) \in B\}$ ein Ereignis sein (d. h. in der σ-Algebra liegen). Dies ist in der Praxis meist erfüllt.

Kapitel 3

Erwartungswert

3.1 Erwartungswert einer Zufallsvariable

Der **Erwartungswert** (oder Mittelwert) einer Zufallsvariable X ist ihr durchschnittlicher Wert.

Definition 1

Der **Erwartungswert** oder **Mittelwert** 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. Notation:

$$\mathbb{E}(X) = \mathbb{E}\, X = \int x \, dF(x) = \mu = \mu_X.$$

Der Erwartungswert ist eine Ein-Zahlen-Zusammenfassung der Verteilung. Man kann $\mathbb{E}(X)$ als Durchschnitt $\frac{1}{n}\sum_{i=1}^n X_i$ einer großen Anzahl iid Ziehungen $X_1, \ldots, X_n$ verstehen. Dies ist mehr als eine Heuristik – es ist ein Satz, das **Gesetz der großen Zahlen** (Kapitel 5).

Damit $\mathbb{E}(X)$ wohldefiniert ist, fordern wir $\int |x| \, dF_X(x) < \infty$. Andernfalls existiert der Erwartungswert nicht.

Beispiel 3.1.1

Sei $X \sim$ Bernoulli(p). Dann ist $\mathbb{E}(X) = \sum_{x=0}^1 x f(x) = 0 \cdot (1-p) + 1 \cdot p = p$.

Beispiel 3.1.2

Fairer Münzwurf zweimal, X = Anzahl Köpfe. Dann ist

$$\mathbb{E}(X) = \sum_x x f_X(x) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

Beispiel 3.1.3

Sei $X \sim$ Uniform$(-1, 3)$. Dann ist

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

Beispiel 3.1.4

[Cauchy-Verteilung] Eine Zufallsvariable hat Cauchy-Verteilung mit Dichte $f_X(x) = \frac{1}{\pi(1+x^2)}$. Dann ist

$$\int |x|\, dF(x) = \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}\, dx = \infty,$$

also existiert der Erwartungswert nicht. Simuliert man viele Cauchy-Ziehungen, stabilisiert sich der Durchschnitt nie, da die Cauchy-Verteilung dicke Tails hat.

Von nun an setzen wir implizit voraus, dass Erwartungswerte existieren.

Satz 2

[Regel des faulen Statistikers] Sei $Y = r(X)$. Dann ist

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x)\, dF_X(x) = \begin{cases} \sum_x r(x) f_X(x) & \text{falls } X \text{ diskret,} \\ \int r(x) f_X(x)\, dx & \text{falls } X \text{ stetig.} \end{cases}$$

Man muss also nicht erst die Verteilung von Y bestimmen.

Beispiel 3.1.5

Sei $X \sim$ Unif$(0, 1)$ und $Y = e^X$. Dann ist

$$\mathbb{E}(Y) = \mathbb{E}(e^X) = \int_0^1 e^x \cdot 1\, dx = [e^x]_0^1 = e - 1.$$

Beispiel 3.1.6

Stab der Länge 1, zufällig gebrochen bei $X \sim$ Unif$(0, 1)$. Sei Y die Länge des längeren Stücks. Dann ist $Y = \max(X, 1 - X)$ und

$$\mathbb{E}(Y) = \int_0^{1/2} (1 - x)\, dx + \int_{1/2}^1 x\, dx = \frac{3}{4}.$$

Satz 3

Existiert das k-te Moment und ist $j < k$, so existiert auch das j-te Moment.

3.2 Eigenschaften des Erwartungswerts

Satz 1

Seien $X_1, \ldots, X_n$ Zufallsvariablen und $a_1, \ldots, a_n$ Konstanten. Dann ist

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

Der Erwartungswert ist also **linear**.

Beispiel 3.2.1

[Binomialverteilung] Sei $X \sim \text{Binomial}(n, p)$. Schreibe $X = \sum_{i=1}^{n} X_i$, wobei $X_i \sim \text{Bernoulli}(p)$. Dann ist

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

Satz 2

Seien $X_1, \ldots, X_n$ unabhängig. Dann ist

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}(X_i).$$

3.3 Varianz und Kovarianz

Definition 1

Sei X eine Zufallsvariable mit Mittelwert μ. Die **Varianz** ist

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

Die **Standardabweichung** ist $\sigma = \sqrt{\text{Var}(X)}$. Notation: $\text{Var}(X) = \sigma^2 = \sigma_X^2$.

Satz 2

Die Varianz hat folgende Eigenschaften (falls wohldefiniert):

1. $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$,
2. Falls a, b Konstanten sind, gilt $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$,
3. $\mathrm{Var}(X) \geq 0$.

Beispiel 3.3.1

Sei $X \sim$ Binomial(n, p) mit $X = \sum_{i=1}^{n} X_i$, wobei $X_i = 1$ für Erfolg beim i-ten Versuch. Da die X_i unabhängig sind, ist

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Für $X_i \sim$ Bernoulli(p) ist $\mathrm{Var}(X_i) = \mathbb{E}(X_i^2) - [\mathbb{E}(X_i)]^2 = p - p^2 = p(1-p)$. Also

$$\mathrm{Var}(X) = np(1-p).$$

Satz 3

Seien $X_1, \ldots, X_n$ iid mit $\mu = \mathbb{E}(X_i)$ und $\sigma^2 = \mathrm{Var}(X_i)$. Sei $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ der Stichprobenmittelwert. Dann ist

$$\mathbb{E}(\overline{X}_n) = \mu, \quad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Definition 4

Seien X und Y Zufallsvariablen mit Erwartungswerten μ_X und μ_Y. Die **Kovarianz** ist

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mu_X \mu_Y.$$

Satz 5

Die Kovarianz erfüllt:

1. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$,
2. Falls X und Y unabhängig sind, ist $\mathrm{Cov}(X, Y) = 0$,
3. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$,
4. $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)$,
5. $\mathrm{Cov}(X, a) = 0$ für eine Konstante a,
6. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$.

Satz 6

Es gilt

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

und

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

Definition 7

Die **Korrelation** zwischen X und Y ist

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Es gilt $-1 \leq \rho(X, Y) \leq 1$.

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)	$\alpha/(\alpha+\beta)$	$\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$
t_ν	0 $(\nu > 1)$	$\nu/(\nu-2)$ $(\nu > 2)$
χ^2_p	p	$2p$

Lemma 1

Sei a ein Vektor und X ein Zufallsvektor mit Erwartungswert μ und Kovarianzmatrix Σ. Dann ist

$$\mathbb{E}(a^T X) = a^T \mu, \quad \mathrm{Var}(a^T X) = a^T \Sigma a.$$

3.5 Bedingter Erwartungswert

Definition 1

Der **bedingte Erwartungswert** von X gegeben $Y = y$ ist

$$\mathbb{E}(X \mid Y = y) = \int x f_{X|Y}(x \mid y)\, dx.$$

Definiere $g(y) = \mathbb{E}(X \mid Y = y)$. Dann ist $\mathbb{E}(X \mid Y)$ die Zufallsvariable $g(Y)$.

Beispiel 3.5.1

Ziehe $X \sim \mathrm{Unif}(0, 1)$. Nach Beobachtung von $X = x$ ziehe $Y \sim \mathrm{Unif}(0, x)$. Dann ist

$$\mathbb{E}(Y \mid X = x) = \frac{x}{2}, \quad \mathbb{E}(Y \mid X) = \frac{X}{2}.$$

Satz 2

[Regel der iterierten Erwartungswerte] Für Zufallsvariablen X und Y gilt

$$\mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}(Y).$$

Beispiel 3.5.2

Im vorigen Beispiel ist

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}\left(\frac{X}{2}\right) = \frac{1}{2}\mathbb{E}(X) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Man kann auch direkt rechnen: $\mathbb{E}(Y) = \int_0^1 y(-\ln y)\, dy = 1/4$.

Definition 3

Die **bedingte Varianz** ist

$$\mathrm{Var}(Y \mid X = x) = \int (y - \mu(x))^2 f_{Y|X}(y \mid x)\, dy,$$

wobei $\mu(x) = \mathbb{E}(Y \mid X = x)$.

Satz 4

Für Zufallsvariablen X und Y gilt

$$\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(\mathbb{E}(Y \mid X)).$$

Beispiel 3.5.3

Ziehe zufällig eine Grafschaft in den USA. Ziehe dann zufällig eine Person aus dieser Grafschaft. Sei Y das Einkommen. Wir haben

$$\mathrm{Var}(Y) = \underbrace{\mathbb{E}(\mathrm{Var}(Y \mid X))}_{\text{Varianz innerhalb Grafschaften}} + \underbrace{\mathrm{Var}(\mathbb{E}(Y \mid X))}_{\text{Varianz zwischen Grafschaften}}.$$

3.6 Momenterzeugende Funktionen

Definition 1

Die **momenterzeugende Funktion** (engl. *moment generating function, mgf*) ist

$$\psi_X(t) = \mathbb{E}(e^{tX}) = \int e^{tx}\, dF_X(x).$$

Beispiel 3.6.1

Sei $X \sim \mathrm{Exp}(1)$. Für $t < 1$ ist

$$\psi_X(t) = \int_0^\infty e^{tx} e^{-x}\, dx = \int_0^\infty e^{(t-1)x}\, dx = \frac{1}{1-t}.$$

Lemma 2

Eigenschaften der mgf:

1. Das k-te Moment ist $\mathbb{E}(X^k) = \psi_X^{(k)}(0)$, wobei $\psi_X^{(k)}$ die k-te Ableitung ist.
2. Sind X und Y unabhängig, so ist $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t)$.
3. $\psi_{aX+b}(t) = e^{bt}\psi_X(at)$.

Beispiel 3.6.2

Sei $X \sim$ Binomial(n,p) mit $X = \sum_{i=1}^n X_i$, wobei $X_i \sim$ Bernoulli(p). Dann ist $\psi_{X_i}(t) = (1-p) + pe^t$ und

$$\psi_X(t) = \prod_{i=1}^n \psi_{X_i}(t) = [(1-p) + pe^t]^n.$$

Satz 3

Seien X und Y Zufallsvariablen. Falls $\psi_X(t) = \psi_Y(t)$ für alle t in einer Umgebung von 0, dann ist $F_X(x) = F_Y(x)$ für alle x, d. h. X und Y haben die gleiche Verteilung.

Beispiel 3.6.3

Seien $X_1 \sim$ Binomial(n_1,p) und $X_2 \sim$ Binomial(n_2,p) unabhängig. Sei $Y = X_1 + X_2$. Dann ist

$$\psi_Y(t) = \psi_{X_1}(t)\psi_{X_2}(t) = [(1-p) + pe^t]^{n_1}[(1-p) + pe^t]^{n_2} = [(1-p) + pe^t]^{n_1+n_2}.$$

Also ist $Y \sim$ Binomial$(n_1 + n_2, p)$.

Beispiel 3.6.4

Seien $Y_1 \sim$ Poisson(λ_1) und $Y_2 \sim$ Poisson(λ_2) unabhängig. Dann ist $\psi_{Y_1}(t) = e^{\lambda_1(e^t-1)}$ und

$$\psi_{Y_1+Y_2}(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}.$$

Also ist $Y_1 + Y_2 \sim$ Poisson$(\lambda_1 + \lambda_2)$.

3.7 Anhang

Wichtige Momente

Für $X \sim N(\mu, \sigma^2)$ gilt:

- $\mathbb{E}(X) = \mu$,
- $\mathrm{Var}(X) = \sigma^2$,
- $\mathbb{E}(X^3) = \mu^3 + 3\mu\sigma^2$,
- $\mathbb{E}(X^4) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$.

Ungleichungen

Markov-Ungleichung: Für $X \geq 0$ und $t > 0$ gilt

$$P(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Tschebyschow-Ungleichung: Für beliebige X mit Erwartungswert μ und Varianz σ^2 gilt für $t > 0$

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Kapitel 4

Ungleichungen

4.1 Wahrscheinlichkeitsungleichungen

Stellen Sie sich vor, Sie haben eine Münze, bei der Sie nicht sicher sind, ob sie fair ist. Sie werfen sie 100-mal und erhalten 65-mal Kopf. Ist die Münze manipuliert? Oder hatten Sie nur Pech? Oder – mathematisch präziser gefragt: Wie wahrscheinlich ist es, bei einer fairen Münze so weit vom Erwartungswert abzuweichen?

Genau hier kommen Wahrscheinlichkeitsungleichungen ins Spiel. Sie sind das Schweizer Taschenmesser der Statistik: Mit minimalem Wissen über eine Verteilung – manchmal nur ihrem Erwartungswert oder ihrer Varianz – können wir mächtige Aussagen über die Wahrscheinlichkeit seltener Ereignisse treffen.

Das Schöne ist: Diese Ungleichungen gelten *unabhängig von der konkreten Verteilung*. Ob normal, binomial, exponentiell oder etwas ganz Exotisches – die Gesetze gelten universal. Das macht sie besonders wertvoll in der Praxis, wo wir die wahre Verteilung oft nicht kennen.

Aber Ungleichungen sind mehr als nur ein praktisches Werkzeug. Sie sind das theoretische Fundament der gesamten Konvergenztheorie (Kapitel 5), sie garantieren die Zuverlässigkeit statistischer Verfahren und sie erklären, warum maschinelles Lernen überhaupt funktioniert. Wenn Sie verstehen wollen, wie aus Daten Wissen wird, führt kein Weg an diesen Ungleichungen vorbei.

Lassen Sie uns mit der einfachsten beginnen – und einer der mächtigsten.

Satz 1

[Markov-Ungleichung] Sei X eine nichtnegative Zufallsvariable mit existierendem Erwartungswert $\mathbb{E}(X)$. Für jedes $t > 0$ gilt

$$P(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Beweis. Da $X \geq 0$, können wir den Erwartungswert aufspalten:

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty x f(x)\,dx \\ &= \int_0^t x f(x)\,dx + \int_t^\infty x f(x)\,dx \\ &\geq \int_t^\infty x f(x)\,dx \geq t \int_t^\infty f(x)\,dx = t\,P(X \geq t).\end{aligned}$$

Division durch $t > 0$ liefert die Behauptung. □

Interpretation: Die Markov-Ungleichung sagt uns: Wenn der Erwartungswert klein ist, kann X nicht zu oft große Werte annehmen. Ist etwa $\mathbb{E}(X) = 5$ und $t = 50$, so ist $P(X \geq 50) \leq 5/50 = 0{,}1$. Die Masse der Verteilung kann nicht beliebig weit vom Erwartungswert wegwandern.

Satz 2

[Tschebyschow-Ungleichung] Sei $\mu = \mathbb{E}(X)$ und $\sigma^2 = \mathrm{Var}(X)$. Dann gilt

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

und mit $Z = (X - \mu)/\sigma$ (standardisierte Variable)

$$P(|Z| \geq k) \leq \frac{1}{k^2}.$$

Insbesondere ist $P(|Z| > 2) \leq 1/4$ und $P(|Z| > 3) \leq 1/9$.

Beweis. Wende Markov auf die nichtnegative Zufallsvariable $|X - \mu|^2$ an:

$$P(|X - \mu| \geq t) = P(|X - \mu|^2 \geq t^2) \leq \frac{\mathbb{E}(|X - \mu|^2)}{t^2} = \frac{\sigma^2}{t^2}.$$

Setze $t = k\sigma$ für die zweite Aussage. □

Interpretation: Tschebyschow ist stärker als Markov, weil wir zusätzliche Information (die Varianz) nutzen. Die Ungleichung garantiert, dass bei *jeder* Verteilung mindestens 75% der Masse innerhalb von 2σ um den Mittelwert liegt. Bei 3σ sind es mindestens $\approx 89\%$. Das ist die mathematische Basis für die „68-95-$99{,}7$-Regel"bei Normalverteilungen – nur allgemeiner!

Beispiel 4.1.1

[Fehlerrate eines Prädiktors] Stellen Sie sich vor, Sie testen ein neuronales Netz auf n neuen Testfällen. Sei $X_i = 1$ falls die i-te Vorhersage falsch ist, sonst $X_i = 0$. Die beobachtete Fehlerrate ist dann

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

Jedes X_i ist Bernoulli-verteilt mit unbekanntem Parameter p (der wahren Fehlerrate). Die Frage ist: Wie weit kann $\overline{X}_n$ von p abweichen?
Da $\mathrm{Var}(X_i) = p(1-p) \leq 1/4$ und die X_i unabhängig sind, ist

$$\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X_1)}{n} = \frac{p(1-p)}{n} \leq \frac{1}{4n}.$$

Mit Tschebyschow erhalten wir

$$P(|\overline{X}_n - p| > \epsilon) \leq \frac{1}{4n\epsilon^2}.$$

Konkretes Beispiel: Bei $n = 1000$ Testfällen und $\epsilon = 0{,}05$ (5 % Abweichung):

$$P(|\overline{X}_n - p| > 0{,}05) \leq \frac{1}{4 \cdot 1000 \cdot 0{,}0025} = \frac{1}{10} = 0{,}1.$$

Mit mindestens 90% Wahrscheinlichkeit liegt unsere gemessene Fehlerrate innerhalb von $\pm 5\%$ der wahren Fehlerrate. Nicht schlecht für eine verteilungsfreie Garantie!

Satz 3

[Hoeffding-Ungleichung] Seien $Y_1, \ldots, Y_n$ unabhängig mit $\mathbb{E}(Y_i) = 0$ und $a_i \leq Y_i \leq b_i$. Sei $\epsilon > 0$. Dann gilt

$$P\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Beweis. Der Beweis nutzt zwei Ideen: die **Chernoff-Methode** und das **Hoeffding-Lemma**.
Schritt 1 (Hoeffding-Lemma): Für $\mathbb{E}(Y) = 0$ und $a \leq Y \leq b$ gilt

$$\mathbb{E}(e^{tY}) \leq \exp\left(\frac{t^2(b-a)^2}{8}\right) \quad \text{für alle } t \in \mathbb{R}.$$

Beweis des Lemmas: Da Y beschränkt ist, können wir Y als Konvexkombination der Randpunkte schreiben. Für $y \in [a, b]$ ist

$$y = \frac{b-y}{b-a} \cdot a + \frac{y-a}{b-a} \cdot b.$$

Wegen Konvexität von e^{tx} gilt

$$e^{ty} \leq \frac{b-y}{b-a}e^{ta} + \frac{y-a}{b-a}e^{tb}.$$

Mit $\mathbb{E}(Y) = 0$ folgt $\mathbb{E}(Y) = \lambda a + (1-\lambda)b = 0$ für $\lambda = b/(b-a)$, also

$$\begin{aligned}\mathbb{E}(e^{tY}) &\leq \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} \\ &= e^{ta}\left[\frac{b - b\cdot 1 + a\cdot e^{t(b-a)}}{b-a}\right] \\ &= e^{-ta/(b-a)}\cdot\frac{-a + be^{t(b-a)}}{b-a}.\end{aligned}$$

Setze $h = t(b-a)$ und $p = -a/(b-a) \in [0,1]$. Definiere

$$L(h) := -ph + \log(1 - p + pe^h).$$

Es ist $L(0) = L'(0) = 0$ und $L''(h) = \frac{p(1-p)e^h}{(1-p+pe^h)^2} \leq \frac{1}{4}$ (da $p(1-p) \leq 1/4$). Mit Taylor folgt

$$L(h) \leq \frac{h^2}{8} \quad \Longrightarrow \quad \mathbb{E}(e^{tY}) \leq e^{t^2(b-a)^2/8}.$$

Schritt 2 (Chernoff-Methode): Für $t > 0$ gilt mit Markov:

$$\begin{aligned}P\left(\sum_{i=1}^n Y_i \geq \epsilon\right) &= P\left(e^{t\sum Y_i} \geq e^{t\epsilon}\right) \\ &\leq \frac{\mathbb{E}(e^{t\sum Y_i})}{e^{t\epsilon}} \\ &= e^{-t\epsilon}\prod_{i=1}^n \mathbb{E}(e^{tY_i}) \quad \text{(Unabhängigkeit)} \\ &\leq e^{-t\epsilon}\prod_{i=1}^n \exp\left(\frac{t^2(b_i-a_i)^2}{8}\right) \quad \text{(Hoeffding-Lemma)} \\ &= \exp\left(t^2\sum_{i=1}^n \frac{(b_i-a_i)^2}{8} - t\epsilon\right).\end{aligned}$$

Schritt 3 (Optimierung): Minimiere die rechte Seite über $t > 0$. Ableitung nach t:

$$\frac{d}{dt}\left[\frac{t^2}{8}\sum(b_i-a_i)^2 - t\epsilon\right] = \frac{t}{4}\sum(b_i-a_i)^2 - \epsilon = 0.$$

Optimal ist $t^* = 4\epsilon/\sum(b_i-a_i)^2$. Einsetzen liefert

$$\begin{aligned}&\exp\left(\frac{16\epsilon^2}{8\sum(b_i-a_i)^2}\cdot\frac{\sum(b_i-a_i)^2}{16} - \frac{4\epsilon^2}{\sum(b_i-a_i)^2}\right) \\ &= \exp\left(\frac{2\epsilon^2}{\sum(b_i-a_i)^2} - \frac{4\epsilon^2}{\sum(b_i-a_i)^2}\right) = \exp\left(-\frac{2\epsilon^2}{\sum(b_i-a_i)^2}\right).\end{aligned}$$

□

Bemerkung: Hoeffding ist eine der mächtigsten Konzentrationsungleichungen überhaupt. Sie sagt: Summen von beschränkten Zufallsvariablen konzentrieren sich exponentiell schnell um ihren Erwartungswert. Die Wahrscheinlichkeit großer Abweichungen fällt *exponentiell* mit ϵ^2 – viel schneller als die polynomiellen Schranken von Markov oder Tschebyschow!

Alternativbeweis (kompakte Version)

Der obige Beweis ist vollständig und zeigt alle Details des Hoeffding-Lemmas. Hier eine **kompakte Alternative**, die die Kernidee auf das Wesentliche reduziert:

Beweis (Chernoff-Methode): Für jedes $t > 0$ gilt

$$\begin{aligned} P\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) &= P\left(\exp\left(t\sum_{i=1}^{n} Y_i\right) \geq e^{t\epsilon}\right) \\ &\leq \frac{\mathbb{E}\left(\exp\left(t\sum Y_i\right)\right)}{e^{t\epsilon}} \quad \text{(Markov)} \\ &= \frac{\prod_{i=1}^{n} \mathbb{E}(e^{tY_i})}{e^{t\epsilon}} \quad \text{(Unabhängigkeit)}. \end{aligned}$$

Das **Hoeffding-Lemma** (siehe Beweis oben) besagt: Für $\mathbb{E}(Y) = 0$ und $a \leq Y \leq b$ gilt

$$\mathbb{E}(e^{tY}) \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

Einsetzen liefert

$$P\left(\sum Y_i \geq \epsilon\right) \leq \exp\left(\frac{t^2\sum(b_i-a_i)^2}{8} - t\epsilon\right).$$

Minimierung über t (optimal: $t^* = 4\epsilon/\sum(b_i-a_i)^2$) ergibt

$$P\left(\sum Y_i \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum(b_i-a_i)^2}\right).$$

Kernidee: Durch Exponenzieren wird die Summe zum Produkt (Unabhängigkeit!), dann Hoeffding-Lemma, dann Optimierung über den freien Parameter t.

Korollar 4

Seien $X_1, \ldots, X_n \sim$ Bernoulli(p) unabhängig. Dann gilt für jedes $\epsilon > 0$

$$P(|\overline{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Beweis. Setze $Y_i = X_i - p$. Dann ist $\mathbb{E}(Y_i) = 0$, und wegen $0 \leq X_i \leq 1$ gilt $-p \leq Y_i \leq 1-p$, also $b_i - a_i = 1$. Mit Hoeffding folgt

$$\begin{aligned} P(\overline{X}_n - p > \epsilon) &= P\left(\sum_{i=1}^{n} Y_i > n\epsilon\right) \\ &\leq \exp\left(-\frac{2(n\epsilon)^2}{n\cdot 1^2}\right) = e^{-2n\epsilon^2}. \end{aligned}$$

Symmetrisch für $P(\overline{X}_n - p < -\epsilon)$. Die Unionsschranke $P(A \cup B) \leq P(A) + P(B)$ liefert die Behauptung. □

Beispiel 4.1.2

[Tschebyschow vs. Hoeffding] Betrachte $X_1, \ldots, X_n \sim$ Bernoulli(p) mit $n = 100$ und $\epsilon = 0{,}2$.

Tschebyschow:

$$P(|\overline{X}_n - p| > 0{,}2) \leq \frac{p(1-p)}{100 \cdot 0{,}04} \leq \frac{0{,}25}{4} = 0{,}0625.$$

Hoeffding:

$$P(|\overline{X}_n - p| > 0{,}2) \leq 2e^{-2 \cdot 100 \cdot 0{,}04} = 2e^{-8} \approx 0{,}00067.$$

Hoeffding ist hier fast **100-mal schärfer**! Für große Abweichungen ist die exponentielle Schranke unschlagbar.

Bemerkung 4.1.1 (Quantitativer Vergleich). Die folgende Tabelle zeigt die Schranken für $P(|\overline{X}_n - p| > \epsilon)$ bei $X_i \sim$ Bernoulli$(0{,}5)$, $n = 100$:

ϵ	Tschebyschow	Hoeffding	Exakt (Binomial)
0,05	0,100	0,606	0,729
0,10	0,025	0,135	0,157
0,20	0,006	0,001	0,0003
0,30	0,003	10^{-8}	10^{-11}

Beobachtungen:

- Für **kleine** ϵ (0,05–0,10): Beide Schranken sind konservativ, aber Hoeffding ist näher am wahren Wert.
- Für **große** ϵ (0,20–0,30): Hoeffding ist dramatisch schärfer – der exponentielle Abfall macht den Unterschied!
- Die Hoeffding-Schranke ist *universell*: Sie gilt für **alle** beschränkten Verteilungen, nicht nur Bernoulli.

Satz 5

[Mill-Ungleichung] Sei $Z \sim N(0,1)$. Dann gilt

$$P(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t} \quad \text{und} \quad P(|Z| > t) \leq 2e^{-t^2/2}$$

für $t > 0$.

Beweis. Wegen Symmetrie ist $P(|Z| > t) = 2P(Z > t)$. Die Dichte von Z ist $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Erste Schranke: Für $t > 0$ gilt mit partieller Integration

$$\begin{aligned} P(Z > t) &= \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \int_t^\infty \frac{1}{\sqrt{2\pi}} \cdot \frac{x}{x} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{1}{x} \cdot x e^{-x^2/2} dx \\ &\leq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{t} \int_t^\infty x e^{-x^2/2} dx \quad \text{(da } 1/x \text{ fallend)} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{t} \left[-e^{-x^2/2} \right]_t^\infty = \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t}. \end{aligned}$$

Also $P(|Z| > t) = 2P(Z > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}$.

Zweite Schranke: Mit Markov für e^{tZ} und $s > 0$:

$$P(Z > t) = P(e^{sZ} > e^{st}) \leq \frac{\mathbb{E}(e^{sZ})}{e^{st}}.$$

Für $Z \sim N(0,1)$ ist $\mathbb{E}(e^{sZ}) = e^{s^2/2}$ (MGF der Normalverteilung). Somit

$$P(Z > t) \leq e^{s^2/2 - st} = e^{-st + s^2/2}.$$

Minimierung über s (optimal: $s = t$) liefert $P(Z > t) \leq e^{-t^2/2}$, also $P(|Z| > t) \leq 2e^{-t^2/2}$. □

Bemerkung 4.1.2. Für große t ist Mill deutlich besser als Tschebyschow:

t	Tschebyschow	Mill
2	$\leq 0{,}25$	$\leq 0{,}27$
3	$\leq 0{,}11$	$\leq 0{,}02$
4	$\leq 0{,}0625$	$\leq 0{,}0027$

Für $t \geq 3$ ist Mill um Größenordnungen schärfer. Das liegt daran, dass Mill die spezielle Struktur der Normalverteilung ausnutzt, während Tschebyschow nur Erwartungswert und Varianz kennt.

4.2 Ungleichungen für Erwartungswerte

Während die vorherigen Ungleichungen Wahrscheinlichkeiten abschätzen, beschäftigen wir uns jetzt mit Ungleichungen zwischen Erwartungswerten. Diese sind fundamental für viele Beweise in der Wahrscheinlichkeitstheorie und ermöglichen elegante Argumente.

Satz 1

[Cauchy-Schwarz-Ungleichung] Haben X und Y endliche zweite Momente, so gilt

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\,\mathbb{E}(Y^2)}$$

und

$$|\operatorname{Cov}(X,Y)| \leq \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}.$$

Gleichheit gilt genau dann, wenn X und Y linear abhängig sind.

Beweis. Für $t \in \mathbb{R}$ ist

$$\begin{aligned} 0 &\leq \mathbb{E}[(X - tY)^2] \\ &= \mathbb{E}(X^2) - 2t\,\mathbb{E}(XY) + t^2\,\mathbb{E}(Y^2). \end{aligned}$$

Dies ist eine quadratische Funktion in t mit Minimum bei $t^* = \mathbb{E}(XY)/\mathbb{E}(Y^2)$. Einsetzen liefert

$$0 \leq \mathbb{E}(X^2) - \frac{[\mathbb{E}(XY)]^2}{\mathbb{E}(Y^2)},$$

woraus $[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2)\,\mathbb{E}(Y^2)$ folgt. Ersetze X durch $|X|$ und Y durch $|Y|$ für die erste Aussage.

Für die zweite Aussage wende die erste auf $X - \mathbb{E}(X)$ und $Y - \mathbb{E}(Y)$ an. □

Interpretation: Cauchy-Schwarz sagt, dass der Erwartungswert eines Produkts nie größer sein kann als das geometrische Mittel der zweiten Momente. Die Kovarianz-Version zeigt: $|\rho(X,Y)| = |\operatorname{Cov}(X,Y)|/\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)} \leq 1$, was wir bereits wussten, aber hier als Spezialfall erhalten.

Satz 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. Wir beweisen die Aussage für konvexes g. Sei $\mu = \mathbb{E}(X)$.

Schritt 1: Existenz der Stützgeraden. Da g konvex ist, existiert zu jedem Punkt μ ein Subgradient $a \in \mathbb{R}$, sodass die affin-lineare Funktion

$$h(x) = a(x - \mu) + g(\mu)$$

eine Stützgerade an g im Punkt μ ist. Dies bedeutet per Definition der Konvexität:

$$g(x) \geq h(x) = a(x - \mu) + g(\mu) \quad \text{für alle } x \in \mathbb{R}.$$

Begründung: Für konvexe Funktionen gilt die Ungleichung

$$g(y) \geq g(x) + a(y - x)$$

für einen geeigneten Subgradienten a (der bei differenzierbaren Funktionen gleich $g'(x)$ ist). Setzen wir $x = \mu$ und y beliebig, folgt die Stützgeraden-Eigenschaft.

Schritt 2: Anwendung der Linearität des Erwartungswerts. Da $h(x) = a(x - \mu) + g(\mu) = ax - a\mu + g(\mu)$ affin-linear ist, gilt

$$\begin{aligned} \mathbb{E}(h(X)) &= \mathbb{E}(aX - a\mu + g(\mu)) \\ &= a\,\mathbb{E}(X) - a\mu + g(\mu) \\ &= a\mu - a\mu + g(\mu) \\ &= g(\mu) = g(\mathbb{E}(X)). \end{aligned}$$

Schritt 3: Anwendung der Stützgeraden-Eigenschaft. Aus $g(x) \geq h(x)$ für alle x folgt durch Anwendung des Erwartungswerts (unter Nutzung der Monotonie: aus $Y \geq Z$ folgt $\mathbb{E}(Y) \geq \mathbb{E}(Z)$):

$$\mathbb{E}(g(X)) \geq \mathbb{E}(h(X)) = g(\mathbb{E}(X)).$$

Konkaver Fall: Ist g konkav, so ist $-g$ konvex. Aus dem konvexen Fall folgt

$$\mathbb{E}(-g(X)) \geq -g(\mathbb{E}(X)) \quad \Longleftrightarrow \quad \mathbb{E}(g(X)) \leq g(\mathbb{E}(X)).$$

Dies zeigt die Jensen-Ungleichung für konkave Funktionen. □

Geometrische Intuition: Konvexe Funktionen „biegen nach oben". Wenn wir X durch seinen Mittelwert ersetzen und dann g anwenden, erhalten wir einen kleineren Wert als wenn wir erst g anwenden und dann mitteln. Die Funktion „verstärkt"die Variabilität von X.

Beispiel 4.2.1

[Anwendungen der Jensen-Ungleichung] **(1)** Da $g(x) = x^2$ konvex ist, folgt

$$\mathbb{E}(X^2) \geq [\mathbb{E}(X)]^2 \quad \Longleftrightarrow \quad \mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \geq 0.$$

Die Varianz ist also automatisch nichtnegativ – keine separate Rechnung nötig!
(2) Da $g(x) = \log x$ konkav ist (für $x > 0$), folgt

$$\mathbb{E}(\log X) \leq \log \mathbb{E}(X).$$

Äquivalent: Das *geometrische Mittel* ist höchstens so groß wie das *arithmetische Mittel*:

$$\exp[\mathbb{E}(\log X)] \leq \mathbb{E}(X).$$

Bemerkung 4.2.1 (Weitere wichtige Ungleichungen). Für Vollständigkeit erwähnen wir zwei weitere Klassiker:

Minkowski-Ungleichung: Für $p \geq 1$ gilt

$$[\mathbb{E}(|X + Y|^p)]^{1/p} \leq [\mathbb{E}(|X|^p)]^{1/p} + [\mathbb{E}(|Y|^p)]^{1/p}.$$

Dies ist die L^p-Dreiecksungleichung: Der L^p-Abstand zwischen X und $-Y$ ist höchstens die Summe der Abstände.

Hölder-Ungleichung: Für $p, q > 1$ mit $1/p + 1/q = 1$ gilt

$$\mathbb{E}(|XY|) \leq [\mathbb{E}(|X|^p)]^{1/p} [\mathbb{E}(|Y|^q)]^{1/q}.$$

Für $p = q = 2$ erhalten wir gerade Cauchy-Schwarz. Hölder ist die allgemeine Version für konjugierte Exponenten.

Kapitel 5

Konvergenz von Zufallsvariablen

5.1 Einführung und Motivation

Der wichtigste Aspekt der Wahrscheinlichkeitstheorie betrifft das Verhalten von Folgen von Zufallsvariablen. Dieser Teil wird **Große-Stichproben-Theorie**, **Grenzwerttheorie** oder **asymptotische Theorie** genannt.

Motivation: In der Praxis haben wir oft eine Folge $X_1, X_2, X_3, \ldots$ von Zufallsvariablen. Beispiele:

- X_i = Ergebnis des i-ten Münzwurfs
- X_i = Messung einer physikalischen Größe zum Zeitpunkt i
- X_i = Rendite einer Aktie am Tag i

Die zentrale Frage lautet: *Was können wir über das Grenzverhalten für $n \to \infty$ aussagen?*

In der Analysis konvergiert eine Zahlenfolge x_n gegen x, falls für jedes $\epsilon > 0$ gilt: $|x_n - x| < \epsilon$ für alle hinreichend großen n. In der Wahrscheinlichkeitstheorie sind Zufallsvariablen Funktionen $X : \Omega \to \mathbb{R}$, und es gibt verschiedene sinnvolle Konzepte von Konvergenz.

Zwei fundamentale Resultate:

1. Das **Gesetz der großen Zahlen**: Der Stichprobenmittelwert $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ konvergiert gegen $\mu = \mathbb{E}(X_i)$.

2. Der **zentrale Grenzwertsatz**: Die normierte Summe $\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}}$ konvergiert in Verteilung gegen $N(0,1)$.

Diese Sätze bilden das theoretische Fundament der gesamten Statistik.

5.2 Konvergenzarten

Wir beginnen mit präzisen Definitionen der verschiedenen Konvergenzarten.

Definition 1

[Konvergenz in Wahrscheinlichkeit] Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen und X eine Zufallsvariable, alle definiert auf demselben Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$.
Wir sagen, X_n **konvergiert in Wahrscheinlichkeit gegen** X, geschrieben $X_n \xrightarrow{P} X$, falls für jedes $\epsilon > 0$ gilt:

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

Ist X konstant gleich c, schreiben wir $X_n \xrightarrow{P} c$.

Intuition: Die Wahrscheinlichkeit, dass X_n weit von X entfernt ist, wird beliebig klein. Mit hoher Wahrscheinlichkeit liegt X_n nahe bei X für großes n.

Definition 2

[Konvergenz in Verteilung] Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen mit Verteilungsfunktionen F_n und X eine Zufallsvariable mit Verteilungsfunktion F.
Wir sagen, X_n **konvergiert in Verteilung gegen** X, geschrieben $X_n \xrightarrow{d} X$, falls

$$\lim_{n\to\infty} F_n(t) = F(t)$$

an allen Stetigkeitsstellen t von F.

Bemerkung: Die X_n und X müssen nicht auf demselben Wahrscheinlichkeitsraum definiert sein – nur ihre Verteilungen zählen. Dies ist die schwächste Form von Konvergenz.

Definition 3

[Konvergenz in L^p] Für $p \geq 1$ sagen wir, X_n **konvergiert in** L^p **gegen** X, falls

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

Für $p = 2$ spricht man von **Konvergenz in quadratischem Mittel**.

Definition 4

[Fast-sichere Konvergenz] Wir sagen, X_n **konvergiert fast sicher gegen** X, geschrieben $X_n \xrightarrow{\text{a.s.}} X$, falls

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1,$$

d.h.

$$P\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1.$$

Intuition: Mit Wahrscheinlichkeit 1 konvergiert die Folge punktweise gegen X.

Beispiel 5.2.1

[Normalverteilung mit schrumpfender Varianz] Sei $X_n \sim N(0, 1/n)$. Intuitiv konzentriert sich X_n bei 0.
(a) L^2-Konvergenz:

$$\mathbb{E}(X_n^2) = \mathrm{Var}(X_n) = \frac{1}{n} \to 0.$$

Also $X_n \to 0$ in L^2.
(b) Konvergenz in Wahrscheinlichkeit: Für $\epsilon > 0$ und $Z \sim N(0, 1)$:

$$P(|X_n| > \epsilon) = P\left(\left|\frac{Z}{\sqrt{n}}\right| > \epsilon\right) = P(|Z| > \epsilon\sqrt{n}) \to 0.$$

Also $X_n \xrightarrow{P} 0$.
(c) Konvergenz in Verteilung:

$$F_n(t) = P(X_n \leq t) = \Phi(t\sqrt{n}),$$

wobei Φ die cdf von $N(0, 1)$ ist. Für $t < 0$ ist $\Phi(t\sqrt{n}) \to 0$, und für $t > 0$ ist $\Phi(t\sqrt{n}) \to 1$. Bei $t = 0$ haben wir $\Phi(0) = 1/2$ für alle n, aber 0 ist eine Unstetigkeitsstelle der Grenzverteilung (Punktmasse bei 0). An allen anderen Punkten gilt:

$$F_n(t) \to F(t) = \begin{cases} 0 & t < 0, \\ 1 & t \geq 0. \end{cases}$$

Also $X_n \xrightarrow{d} 0$.

Beispiel 5.2.2

[Gegenbeispiel: Konvergenz in Verteilung impliziert nicht Konvergenz in Wahrscheinlichkeit] Sei $X \sim N(0, 1)$ und definiere $X_n = -X$ für alle n. Dann hat jedes X_n dieselbe Verteilung wie X, also $X_n \xrightarrow{d} X$ (sogar $F_n = F$ für alle n).
Aber $X_n - X = -2X$ hat Verteilung $N(0, 4)$, also

$$P(|X_n - X| > \epsilon) = P(|2X| > \epsilon) > 0$$

für alle n und $\epsilon < \infty$. Somit $X_n \not\xrightarrow{P} X$.
Fazit: Konvergenz in Verteilung ist echt schwächer als Konvergenz in Wahrscheinlichkeit.

Hierarchie der Konvergenzarten

Der folgende Satz ordnet die verschiedenen Konvergenzarten.

Satz 5

[Hierarchie der Konvergenzarten] Für Zufallsvariablen X_n, X gelten folgende Implikationen:

1. $X_n \to X$ in $L^2 \Rightarrow X_n \xrightarrow{P} X$.
2. $X_n \xrightarrow{\text{a.s.}} X \Rightarrow X_n \xrightarrow{P} X$.
3. $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$.
4. Falls $X_n \xrightarrow{d} c$ für eine Konstante c, dann $X_n \xrightarrow{P} c$.

Die Umkehrungen gelten im Allgemeinen nicht.

Beweis. **(1)** $L^2 \Rightarrow P$**:** Sei $\epsilon > 0$. Mit der Markov-Ungleichung:

$$P(|X_n - X| > \epsilon) = P((X_n - X)^2 > \epsilon^2) \leq \frac{\mathbb{E}[(X_n - X)^2]}{\epsilon^2} \to 0.$$

(2) a.s. $\Rightarrow P$**:** Sei $\epsilon > 0$ und definiere

$$A_n(\epsilon) = \{\omega : |X_m(\omega) - X(\omega)| \leq \epsilon \text{ für alle } m \geq n\}.$$

Dann ist $(A_n(\epsilon))$ eine aufsteigende Folge und

$$\bigcup_{n=1}^{\infty} A_n(\epsilon) = \{\omega : \lim_{m\to\infty} X_m(\omega) = X(\omega)\}.$$

Da $X_n \xrightarrow{\text{a.s.}} X$, ist $P(\bigcup_{n=1}^{\infty} A_n(\epsilon)) = 1$. Mit Stetigkeit von unten:

$$P(|X_n - X| \leq \epsilon) \geq P(A_n(\epsilon)) \to 1.$$

Also $P(|X_n - X| > \epsilon) \to 0$.

(3) $P \Rightarrow d$**:** Sei t eine Stetigkeitsstelle von F und $\epsilon > 0$. Dann:

$$\begin{aligned} F_n(t) &= P(X_n \leq t) \\ &= P(X_n \leq t, X \leq t+\epsilon) + P(X_n \leq t, X > t+\epsilon) \\ &\leq P(X \leq t+\epsilon) + P(|X_n - X| > \epsilon) \\ &= F(t+\epsilon) + P(|X_n - X| > \epsilon). \end{aligned}$$

Grenzübergang $n \to \infty$: $\limsup_{n\to\infty} F_n(t) \leq F(t+\epsilon)$.

Analog: $P(X \leq t-\epsilon) \leq P(X_n \leq t) + P(|X_n - X| > \epsilon)$, also

$$F(t-\epsilon) \leq \liminf_{n\to\infty} F_n(t) + 0.$$

Somit $F(t-\epsilon) \leq \liminf F_n(t) \leq \limsup F_n(t) \leq F(t+\epsilon)$.

Da ϵ beliebig und F stetig bei t, folgt $F_n(t) \to F(t)$.

(4) $d(c) \Rightarrow P$**:** Sei $F(t) = \mathbb{1}_{\{t \geq c\}}$ (Punktmasse bei c). Für $\epsilon > 0$:

$$P(|X_n - c| > \epsilon) = F_n(c-\epsilon) + (1 - F_n(c+\epsilon)).$$

Die Punkte $c \pm \epsilon$ (für $\epsilon > 0$) sind Stetigkeitsstellen von F, daher:

$$F_n(c-\epsilon) \to F(c-\epsilon) = 0, \quad F_n(c+\epsilon) \to F(c+\epsilon) = 1.$$

Also $P(|X_n - c| > \epsilon) \to 0$. □

5.3 Das Slutsky-Theorem

Das Slutsky-Theorem ist fundamental für Anwendungen, da es erlaubt, Konvergenz unter stetigen Transformationen zu erhalten.

Satz 1

[Slutsky] Seien X_n, X, Y_n Zufallsvariablen und $g : \mathbb{R} \to \mathbb{R}$ eine stetige Funktion. Dann gilt:

1. Falls $X_n \xrightarrow{P} X$, dann $g(X_n) \xrightarrow{P} g(X)$.
2. Falls $X_n \xrightarrow{d} X$ und $Y_n \xrightarrow{P} c$ (Konstante), dann:
   - $X_n + Y_n \xrightarrow{d} X + c$,
   - $X_n \cdot Y_n \xrightarrow{d} c \cdot X$,
   - $X_n / Y_n \xrightarrow{d} X/c$ (falls $c \neq 0$).

Beweis. **(1)** Sei $\epsilon > 0$. Da g stetig auf $\mathbb{R}$ ist, existiert für jedes kompakte $K \subset \mathbb{R}$ ein $\delta > 0$ mit: Falls $x, y \in K$ und $|x - y| < \delta$, dann $|g(x) - g(y)| < \epsilon$.

Wähle $K = [-M, M]$ groß genug, sodass $P(|X| \leq M) > 1 - \epsilon/2$. Dann:

$$\begin{aligned} P(|g(X_n) - g(X)| > \epsilon) &\leq P(|g(X_n) - g(X)| > \epsilon, |X| \leq M, |X_n| \leq M) \\ &\quad + P(|X| > M) + P(|X_n| > M) \\ &\leq P(|X_n - X| > \delta) + \epsilon/2 + P(|X_n - X| > M - |X|) \\ &\to 0 + \epsilon/2 < \epsilon. \end{aligned}$$

(2a) Addition: Sei F_n, F die cdf von X_n, X und t eine Stetigkeitsstelle von F. Für $\epsilon > 0$:

$$\begin{aligned} P(X_n + Y_n \leq t) &= P(X_n + Y_n \leq t, |Y_n - c| \leq \epsilon) + P(X_n + Y_n \leq t, |Y_n - c| > \epsilon) \\ &\leq P(X_n \leq t - c + \epsilon) + P(|Y_n - c| > \epsilon) \\ &= F_n(t - c + \epsilon) + P(|Y_n - c| > \epsilon). \end{aligned}$$

Grenzübergang: $\limsup P(X_n + Y_n \leq t) \leq F(t - c + \epsilon)$. Analog für liminf. Da ϵ beliebig: $P(X_n + Y_n \leq t) \to F(t - c)$.

(2b) Multiplikation: Analog mit $P(X_n Y_n \leq t) = P(X_n \leq t/Y_n)$ und Approximation $Y_n \approx c$. □

5.4 Das Gesetz der großen Zahlen

Satz 1

[Schwaches Gesetz der großen Zahlen] Seien $X_1, X_2, \ldots$ iid mit $\mathbb{E}(X_i) = \mu$ und $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Dann gilt

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

Beweis. Mit der Tschebyschow-Ungleichung: Für $\epsilon > 0$

$$\begin{aligned} P(|\overline{X}_n - \mu| > \epsilon) &\leq \frac{\mathrm{Var}(\overline{X}_n)}{\epsilon^2} \\ &= \frac{1}{\epsilon^2} \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2\epsilon^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) \quad \text{(Unabhängigkeit)} \\ &= \frac{n\sigma^2}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0. \end{aligned}$$

□

Satz 2

[Starkes Gesetz der großen Zahlen (ohne Beweis)] Unter denselben Voraussetzungen gilt sogar

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu.$$

Bemerkung: Der Beweis des starken Gesetzes ist technisch anspruchsvoll und verwendet entweder Martingaltheorie oder die Borel-Cantelli-Lemmata. Siehe Durrett (2019) für Details.

Beispiel 5.4.1

[Münzwurf und frequentistische Interpretation] Sei $p \in (0,1)$ die Wahrscheinlichkeit für Kopf. Definiere

$$X_i = \begin{cases} 1 & \text{Kopf beim } i\text{-ten Wurf}, \\ 0 & \text{Zahl beim } i\text{-ten Wurf}. \end{cases}$$

Dann ist $\overline{X}_n$ die relative Häufigkeit von Kopf in n Würfen.
Nach dem Gesetz der großen Zahlen: $\overline{X}_n \xrightarrow{P} p$.
Interpretation: Die relative Häufigkeit konvergiert in Wahrscheinlichkeit gegen die Wahrscheinlichkeit. Dies rechtfertigt die frequentistische Interpretation: Wahrscheinlichkeit ist der Grenzwert der relativen Häufigkeit.
Quantitativ (mit Tschebyschow): Für $n = 10000$, $p = 0{,}5$, $\epsilon = 0{,}01$:

$$P(|\overline{X}_n - 0{,}5| > 0{,}01) \leq \frac{\mathrm{Var}(X_1)}{n\epsilon^2} = \frac{0{,}25}{10000 \cdot 0{,}0001} = 0{,}25.$$

Mit Wahrscheinlichkeit mindestens 75% liegt die relative Häufigkeit innerhalb von 1% um 50%.

5.5 Der zentrale Grenzwertsatz

Satz 1

[Zentraler Grenzwertsatz (CLT)] Seien $X_1, X_2, \ldots$ iid mit $\mathbb{E}(X_i) = \mu$ und $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Sei

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}}.$$

Dann gilt

$$Z_n \xrightarrow{d} N(0,1),$$

d.h. für alle $z \in \mathbb{R}$:

$$\lim_{n\to\infty} P(Z_n \le z) = \Phi(z),$$

wobei Φ die cdf der Standardnormalverteilung ist.
Äquivalent: $\overline{X}_n$ ist approximativ $N(\mu, \sigma^2/n)$ verteilt.

Bemerkung: Der Beweis verwendet charakteristische Funktionen und ist technisch. Eine Beweisskizze:

Beweisskizze via charakteristische Funktionen. Sei $Y_i = (X_i - \mu)/\sigma$, sodass $\mathbb{E}(Y_i) = 0$, $\mathrm{Var}(Y_i) = 1$. Die charakteristische Funktion von Y_i ist $\varphi(t) = \mathbb{E}(e^{itY_i})$.

Taylor-Entwicklung (falls $\mathbb{E}(|Y_i|^3) < \infty$):

$$\varphi(t) = 1 + it \cdot 0 - \frac{t^2}{2} \cdot 1 + o(t^2) = 1 - \frac{t^2}{2} + o(t^2).$$

Die charakteristische Funktion von $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$ ist

$$\begin{aligned}
\varphi_{Z_n}(t) &= \prod_{i=1}^n \varphi(t/\sqrt{n}) = [\varphi(t/\sqrt{n})]^n \\
&= \left[1 - \frac{t^2}{2n} + o(t^2/n)\right]^n \\
&\to e^{-t^2/2} \quad \text{(mit } (1 + x/n)^n \to e^x\text{)}.
\end{aligned}$$

Dies ist die charakteristische Funktion von $N(0,1)$. Nach dem Stetigkeitssatz von Lévy folgt $Z_n \xrightarrow{d} N(0,1)$. $\square$

Beispiel 5.5.1

[Binomialverteilung] Sei $X_i \sim$ Bernoulli(p) iid. Dann $\mathbb{E}(X_i) = p$, $\mathrm{Var}(X_i) = p(1-p)$. Die Summe $S_n = \sum_{i=1}^n X_i \sim$ Binomial(n,p) hat exakte Verteilung

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Nach dem CLT:

$$\frac{S_n - np}{\sqrt{np(1-p)}} \approx N(0,1).$$

Numerisch: Für $n = 100$, $p = 0{,}3$ schätze $P(S_n \geq 35)$:
Exakt (mit Computer):

$$P(S_n \geq 35) = \sum_{k=35}^{100} \binom{100}{k} 0{,}3^k 0{,}7^{100-k} \approx 0{,}1762.$$

CLT-Approximation:

$$\begin{aligned} P(S_n \geq 35) &\approx P\left(Z \geq \frac{35 - 30}{\sqrt{100 \cdot 0{,}3 \cdot 0{,}7}}\right) \\ &= P\left(Z \geq \frac{5}{\sqrt{21}}\right) = P(Z \geq 1{,}091) \approx 0{,}1377. \end{aligned}$$

Mit Stetigkeitskorrektur (besser für diskrete Verteilungen):

$$P(S_n \geq 35) \approx P\left(Z \geq \frac{34{,}5 - 30}{\sqrt{21}}\right) = P(Z \geq 0{,}982) \approx 0{,}1631.$$

Näher am exakten Wert!

Satz 2

[Berry-Esséen-Ungleichung] Falls $\mathbb{E}\,|X_1 - \mu|^3 < \infty$, dann existiert eine universelle Konstante C mit

$$\sup_z |P(Z_n \leq z) - \Phi(z)| \leq \frac{C\,\mathbb{E}\,|X_1 - \mu|^3}{\sigma^3 \sqrt{n}}.$$

Die beste bekannte Konstante ist $C \approx 0{,}4748$.

Interpretation: Der Approximationsfehler ist $O(n^{-1/2})$. Für Genauigkeit 0,01 benötigen wir $n \approx 10000$ (abhängig vom dritten Moment).

5.6 Die Delta-Methode

Satz 1

[Delta-Methode] Sei $g : \mathbb{R} \to \mathbb{R}$ differenzierbar mit $g'(\mu) \neq 0$. Falls

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1),$$

dann gilt

$$\frac{\sqrt{n}(g(\overline{X}_n) - g(\mu))}{\sigma|g'(\mu)|} \xrightarrow{d} N(0,1).$$

Äquivalent: $g(\overline{X}_n)$ ist approximativ $N(g(\mu), \sigma^2[g'(\mu)]^2/n)$ verteilt.

Beweis. Taylor-Entwicklung von g um μ:

$$g(\overline{X}_n) = g(\mu) + g'(\mu)(\overline{X}_n - \mu) + R_n,$$

wobei $R_n = o(|\overline{X}_n - \mu|)$ der Rest ist.

Multipliziere mit $\sqrt{n}$:

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) = g'(\mu)\sqrt{n}(\overline{X}_n - \mu) + \sqrt{n}R_n.$$

Da $\overline{X}_n \xrightarrow{P} \mu$ (WLLN) und $R_n = o(|\overline{X}_n - \mu|)$, gilt $\sqrt{n}R_n = o_P(1)$ (verschwindet in Wahrscheinlichkeit).

Mit Slutsky:

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) = g'(\mu)\sqrt{n}(\overline{X}_n - \mu) + o_P(1) \xrightarrow{d} g'(\mu) \cdot N(0,\sigma^2) = N(0, \sigma^2[g'(\mu)]^2). \quad \square$$

Beispiel 5.6.1

[Konfidenzintervall für μ^2] Seien $X_1, \ldots, X_n$ iid mit $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$. Schätze $\tau = \mu^2$ durch $\hat{\tau} = \overline{X}_n^2$.
Mit $g(x) = x^2$ ist $g'(x) = 2x$, also $g'(\mu) = 2\mu$. Nach der Delta-Methode:

$$\sqrt{n}(\overline{X}_n^2 - \mu^2) \approx N(0, 4\mu^2\sigma^2).$$

Ein approximatives 95%-Konfidenzintervall für μ^2:

$$\mu^2 \in \left[\overline{X}_n^2 \pm 1{,}96\frac{2|\overline{X}_n| \cdot s}{\sqrt{n}}\right],$$

wobei s die Stichprobenstandardabweichung ist.