Vorlesung aus dem Wintersemester 2010/11

Stochastik

Prof. Dr. Franz Merkl

geTEXt von Viktor Kleen & Florian Stecker

Inhaltsverzeichnis

1 Wahrscheinlichkeitstheorie

1.1 Wahrscheinlichkeitsmodelle

Definition. Ein *Wahrscheinlichkeitsmodell* ist ein Tripel $(\Omega, \mathcal{A}, P)$ mit einer informalen Interpretationsregel, was die Komponenten bedeuten sollen.

1.1.1 Der Ergebnisraum Ω

Ω ist eine nichtleere Menge, der Ergebnisraum. Die Elemente $\omega \in \Omega$ heißen Ergebnisse und werden als mögliche Ausgänge des Zufallsexperiments interpretiert.

Beispiel (Einmaliger Wurf eines Spielwürfels).

Modell 1 $\Omega_1 = \{1, 2, \ldots, 5, 6\}$, Interpretation von $\omega \in \Omega_1$ ist die oben liegende Augenzahl
Modell 2 $\Omega_2 = \{1, \ldots, 6, \text{ungültig}\}$, Modell 2 ist feiner als Modell 1, durch Ignorieren ungültiger Ergebnisse erhält man Modell 1
Modell 3 $\Omega_3 = \mathbb{R}^3 \times SO(3)$, Lage des Schwerpunkts und Orientierung des Würfels im Raum

Beispiel (n-facher Münzwurf).

Modell 1 $\Omega = \{0, 1\}^n$ (Kopf und Zahl für jeden Wurf)
Modell 2 $\Omega' = \{0, \ldots, n\}$, Interpretation von $\omega' \in \Omega'$ als Anzahl von Würfen, bei denen Zahl aufgetreten ist

Modell 1 enthält mehr Informationen als Modell 2. Der Zusammenhang wird durch die Abbildung $S\colon \Omega \to \Omega', (\omega_1, \ldots, \omega_n) = \sum_{i=1}^n \omega_i$ vermittelt.

Beispiel (Ziehen von n Kugeln aus eine Urne mit $m \geq n$ unterscheidbaren Kugeln). $\Omega = \{\omega\colon \{1, \ldots, n\} \to \{1, \ldots, m\}\colon \omega \text{ ist injektiv}\}$

Beispiel (Glücksrad).

Modell 1 $\Omega = S^1$, Interpretation von $\omega \in \Omega$ als Koordinatenvektor der Zeigerspitze
Modell 2 $\Omega' = [0, 1)$, Interpretation von $t \in \Omega$, als Winkel $\alpha = 2\pi t$ zwischen pos. x-Achse und Zeiger

Abbildung zwischen den Modellen: $\Omega' \to \Omega, t \mapsto \exp(2\pi i t) \in \mathbb{R}^2$

Beispiel (Pegelstand im Ammersee in einem Zeitintervall). $\Omega = C([t_0, t_1])$

1.1.2 Die Ereignis-σ-Algebra $\mathcal{A}$

Ja/Nein-Fragen an das zufällige Ergebnis $\omega \in \Omega$ werden durch Teilmengen $A \subseteq \Omega$ modelliert. $\omega \in A$ wird als "Ja" interpretiert, $\omega \notin A$ als "Nein". Dabei muss man nicht alle Teilmengen von Ω als zulässige Fragen erlauben, sondern nur manche, die *beobachtbar* oder *messbare* Teilmengen heißen. Eine messbare Teilmenge von Ω heißt auch *Ereignis*. Die Menge $\mathcal{A}$ aller messbaren Teilmengen von Ω soll Abschlusseigenschaften erfüllen, die im Begriff der σ-Algebra zusammengefasst werden.

Definition. Sei Ω ein Ergebnisraum. Eine Menge $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ heißt *σ-Algebra* über Ω, wenn gilt:

1. $\Omega \in \mathcal{A}$
2. Für alle $A \in \mathcal{A}$ gilt $A^c = \Omega \setminus A \in \mathcal{A}$.
3. Für alle Folgen $(A_n)_{n\in\mathbb{N}}$ mit Werten in $\mathcal{A}$ gilt $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

Definition. Ein Paar $(\Omega, \mathcal{A})$, bestehend aus einem Ergebnisraum Ω und einer σ-Algebra $\mathcal{A}$ über Ω, heißt *Ereignisraum* oder *messbarer Raum*. Die Elemente von $\mathcal{A}$ heißen *Ereignisse*, $\Omega \in \mathcal{A}$ heißt *sicheres Ereignis* und $\emptyset = \Omega^c \in \mathcal{A}$ heißt *unmögliches Ereignis*.

Beispiel. Die Potenzmenge $\mathcal{P}(\Omega)$ von Ω ist eine σ-Algebra über Ω. Meist wählt man diese, wenn Ω endlich oder abzählbar unendlich ist.

Beispiel. $\mathcal{A} = \{\Omega, \emptyset\}$ ist eine σ-Algebra über Ω, die *triviale σ-Algebra*.

Beispiel (Einfacher Würfelwurf). $\Omega = \{1, \dots, 6\}$. Das Ereignis "gerade Augenzahl" wird durch $\{2, 4, 6\}$ beschrieben.

Lemma 1. *Sei $\mathcal{A}$ eine σ-Algebra über Ω. Dann gilt:*

1. $\emptyset \in \mathcal{A}$
2. *Aus* $A, B \in \mathcal{A}$ *folgt* $A \cup B \in \mathcal{A}$
3. *Aus* $A, B \in \mathcal{A}$ *folgt* $A \cap B \in \mathcal{A}$
4. *Aus* $A, B \in \mathcal{A}$ *folgt* $A \setminus B \in \mathcal{A}$
5. *Aus* $A, B \in \mathcal{A}$ *folgt* $A \Delta B = (A \setminus B) \cup (B \setminus A) \in \mathcal{A}$

Für jede Menge $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ von Teilmengen gibt es eine kleinste σ-Algebra über Ω, die $\mathcal{M}$ umfasst, nämlich:

$$\begin{aligned}\sigma(\mathcal{M}) = \sigma(\mathcal{M}, \Omega) &= \bigcap \{\mathcal{A} \subseteq \mathcal{P}(\Omega) \colon \mathcal{A} \text{ ist } \sigma\text{-Algebra mit } \mathcal{M} \subseteq \mathcal{A}\} \\ &= \{A \subseteq \Omega \colon \forall \mathcal{A} \subseteq \mathcal{P}(\Omega)\ \sigma\text{-Algebra. } \mathcal{M} \subseteq \mathcal{A} \Rightarrow A \in \mathcal{A}\}\end{aligned}$$

Es ist leicht zu sehen, dass $\sigma(\mathcal{M})$ eine σ-Algebra über Ω ist, die $\mathcal{M}$ umfasst, und dass jede σ-Algebra, die $\mathcal{M}$ umfasst, eine Obermenge von $\sigma(\mathcal{M})$ ist. $\sigma(\mathcal{M}, \Omega)$ heißt die *von $\mathcal{M}$ erzeugte σ-Algebra*.

Beispiel. $\Omega = \{1, \dots, 6\}$. Wir betrachten die Ereignisse $A = \{2, 4, 6\}$, $B = \{6\}$. Dann gilt $\sigma(\{A\}) = \{\emptyset, \Omega, A, A^c\}$ und

$$\sigma(\{A, B\}) = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4\}, \{6\}, \{1, 2, 3, 4, 5\}, \{1, 3, 5, 6\}, \{2, 4, 6\}\}$$

Bemerkung. Ist Ω endlich (oder abzählbar unendlich), so wird jede σ-Algebra $\mathcal{A}$ über Ω von einer eindeutig bestimmten Partition von Ω erzeugt.

Beispiel (Fort.). $\sigma(\{A, B\})$ wird auch von der Partition $\{\{1, 3, 5\}, \{2, 4\}, \{6\}\}$ erzeugt.

Definition. Als die Standard-σ-Algebra über $\mathbb{R}$ verwendet man

$$\mathcal{B}(\mathbb{R}) = \sigma(\{(a, b) \colon a, b \in \mathbb{R}, a < b\})$$

Sie heißt *Borelsche σ-Algebra* über $\mathbb{R}$ und ihre Elemente heißen *Borelmengen* oder auch *Borel-messbar*.

Bemerkung. Die Borelsche σ-Algebra $\mathcal{B}(\mathbb{R})$ wird *nicht* von einer Partition von $\mathbb{R}$ erzeugt, sie ist echt kleiner als $\mathcal{P}(\mathbb{R})$.

Definition. Allgemeiner definiert man für jeden metrischen Raum (M, d) (oder topologischen Raum (M, τ)) die Borelsche σ-Algebra $\mathcal{B}(M) = \sigma(\{A \subseteq M \colon A \text{ ist offen}\})$.

1.1.3 Warum arbeiten wir mit σ-Algebren über Ω statt stets mit $\mathcal{P}(\Omega)$?

σ-Algebren erlauben die Modellierung unvollständiger, wechselnder Beobachtungsmöglichkeiten.

Beispiel (mehrfacher Münzwurf). $\Omega_n = \{0,1\}^n$. Beobachten wir nur die Würfe bis zum m-ten, $m \leq n$, so sind i.A. nicht alle Teilmengen von Ω_n beobachtbar, sondern nur die in

$$\mathcal{F}_m = \{\Pi_{n,m}^{-1}(A) \colon A \subseteq \{0,1\}^m\},$$

wobei $\Pi_{n,m} \colon \{0,1\}^n \to \{0,1\}^m, (\omega_1, \ldots, \omega_n) \mapsto (\omega_1, \ldots, \omega_m)$. $\mathcal{F}_m$ ist eine σ-Algebra, aber $\mathcal{F}_m \neq \mathcal{P}(\Omega_n)$ für $m < n$.

Bei kontinuierlichen Modellen, z.B. dem Glücksrad $\Omega = S^1$, gibt es kein sinnvolles Modell für die "Gleichverteilung", aber sehr wohl auf $\mathcal{B}(\Omega)$.

1.1.4 Wahrscheinlichkeitsmaß P

Definition. Sei $(\Omega, \mathcal{A})$ eine Ereignisraum. Eine Abbildung $P \colon \mathcal{A} \to [0,1]$ heißt *Wahrscheinlichkeitsmaß* auf $(\Omega, \mathcal{A})$, wenn gilt:

1. $P(\Omega) = 1$
2. Für jede Folge $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{A}$ von paarweise disjunkten Ereignisse gilt:

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$$

Die Eigenschaft (2) heißt *σ-Additivität*.

Definition. Fúr Ereignisse A heißt $P(A)$ die *Wahrscheinlichkeit von A* (im Modell $(\Omega, \mathcal{A}, P)$), dass das Ereignis A eintritt.

Definition. Ein Tripel $(\Omega, \mathcal{A}, P)$ heißt *Wahrscheinlichkeitsraum*, wenn Ω ein Ergebnisraum, $\mathcal{A}$ eine σ-Algebra über Ω und P ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$ ist.

Definition. Ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ zusammen mit einer Interpretation, was die $\omega \in \Omega$ und die Wahrscheinlichkeiten $P(A)$, $A \in \mathcal{A}$, in der Anwendung bedeuten sollen, heißt *Wahrscheinlichkeitsmodell*.

Definition. Sei $(\Omega, \mathcal{A})$ ein messbarer Raum. Eine Abbildung $\mu \colon \mathcal{A} \to [0, \infty]$ heißt *Maß* auf $(\Omega, \mathcal{A})$, wenn gilt:

1. $\mu(\emptyset) = 0$
2. Für jede Folge $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{A}$ von paarweise disjunkten Mengen gilt:

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

$(\Omega, \mathcal{A}, \mu)$ heißt dann *Maßraum*.

Bemerkung. In der Definition des Wahrscheinlichkeitsraums wäre ein Fordefung $P(\emptyset) = 0$ überflüssig, da sie automatisch folgt:

$$P(\emptyset) = P\left(\bigcup_{n\in\mathbb{N}} \emptyset\right) = \sum_{n\in\mathbb{N}} P(\emptyset)$$

Also folgt wegen $P(\emptyset) \in [0,1]$ bereits $P(\emptyset) = 0$.

1.1.5 Einfache Eigenschaften von (Wahrscheinlichkeits-)Maßen

Lemma. *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Dann gilt:*

1. *Endliche Additivität: Sind $A, B \in \mathcal{A}$ disjunkt, so ist $P(A \cup B) = P(A) + P(B)$.*
2. *Für beliebige Ereignisse $A, B \in \mathcal{A}$ gilt $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.*
3. *Für $A, B \in \mathcal{A}$ gilt $P(A^c) = 1 - P(A)$.*
4. *Monotonie: Für alle $A, B \in \mathcal{A}$ mit $A \subseteq B$ gilt $P(A) \leq P(B)$.*
5. *σ-Stetigkeit von unten: Ist $(A_n)_{n\in\mathbb{N}}$ eine aufsteigende Folge in $\mathcal{A}$, also $A_1 \subseteq A_2 \subseteq \dots$, so gilt:*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{n\in\mathbb{N}} A_n\right)$$

6. *σ-Stetigkeit von oben: Ist $(A_n)_{n\in\mathbb{N}}$ eine absteigende Folge in $\mathcal{A}$, also $A_1 \supseteq A_2 \supseteq \dots$, so gilt:*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcap_{n\in\mathbb{N}} A_n\right)$$

Beweis. 1. Seien $A_1, \dots, A_N \in \mathcal{A}$ paarweise disjunkt. Für die Folge $(A_1, \dots, A_n, \emptyset, \dots)$ erhalten wir mit der σ-Additivität

$$\sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{n} P(A_i) + \sum_{i=n+1}^{\infty} P(\emptyset) = P(A_1 \cup \dots \cup A_n \cup \emptyset \cup \dots) = P(A_1 \cup \dots \cup A_n)$$

2. Es gilt $A \cup B = A \sqcup (B \setminus A)$ und $B = (A \cap B) \sqcup (B \setminus A)$, also folgt:

$$P(A \cup B) + P(A \cap B) = P(A) + P(B \setminus A) + P(A \cap B) = P(A) + P(B)$$

3. Aus $A \sqcup A^c = \Omega$ folgt $P(A) + P(A^c) = P(\Omega) = 1$.
4. Aus $A \subseteq B$ folgt $B = A \sqcup (B \setminus A)$, also $P(B) = P(A) + P(B \setminus A) \geq P(A)$.
5. Setzen wir formal $A_0 = \emptyset$, so folgt

$$\bigsqcup_{m\in\mathbb{N}} A_m \setminus A_{m-1} = \bigcup_{m\in\mathbb{N}} A_m$$

Es folgt:

$$P\left(\bigcup_{m\in\mathbb{N}} A_m\right) = \sum_{m\in\mathbb{N}} P(A_m \setminus A_{m-1}) = \lim_{n\to\infty} \sum_{m=1}^{n} P(A_m \setminus A_{m-1}) =$$
$$= \lim_{n\to\infty} P\left(\bigcup_{m=1}^{n} A_m \setminus A_{m-1}\right) = \lim_{n\to\infty} P(A_n)$$

6. Die Folge $(A_n^c)_{n\in\mathbb{N}}$ ist aufsteigend, also mit 5:

$$1 - P(A_n) = P(A_n^c) \underset{n\to\infty}{\longrightarrow} P\left(\bigcup_{n\in\mathbb{N}} A_n^c\right) = 1 - P\left(\bigcap_{n\in\mathbb{N}} A_n\right) \qquad \square$$

Bemerkung. Die Eigenschaften 1, 4 und 5 gelten immernoch für Maße, ebenso 2 in der Version $P(A \cup B) + P(A \cap B) = P(A) + P(B)$. Die σ-Stetigkeit von oben kann für Maße verletzt werden, aber nur falls $\mu(A_n) = \infty$ für alle $n \in \mathbb{N}$.

Beispiel.

1. ist Ω ein endlicher oder abzählbarer unendlicher Ergebnisraum und ist $\rho = (\rho_\omega)_{\omega\in\Omega}$ eine Familie nichtnegativer Zahlen und $\sum_{\omega\in\Omega} \rho_\omega = 1$, so wird durch
$$P\colon \mathcal{P}(\Omega) \to [0,1], A \to \sum_{\omega\in A} \rho_\omega$$
ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{P}(\Omega))$ definiert. Umgekehrt ist jedes Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{P}(\Omega))$ auf endlichen oder abzählbar unendlichem Ω von dieser Gestalt und ρ ist eindeutig bestimmt. ρ heißt *Zähldichte* (oder *Wahrscheinlichkeitsfunktion*) von P.
2. Ist Ω wieder eine Ergebnisraum, so definiert
$$\mu\colon \mathcal{P}(\Omega) \to \mathbb{N}_0 \cup \{\infty\}, \mu(A) = |A|$$
eine Maß auf $(\Omega, \mathcal{P}(\Omega))$, das *Zählmaß* auf Ω. Außer im Fall $|\Omega| = 1$ ist μ kein Wahrscheinlichkeitsmaß. Ist Ω endlich, so wird durch $P\colon \mathcal{P}(\Omega) \to [0,1], A \mapsto \mu(A)/\mu(\Omega)$ ein Wahrscheinlichkeitsmaß auf $\mathcal{P}(\Omega)$ definiert. Es heißt (diskrete) *Gleichverteilung* auf Ω. Es besitzt die Zähldichte $(1/\mu(\Omega))_{\omega\in\Omega}$.
3. Ist $(\Omega, \mathcal{A})$ ein Ereignisraum und $b \in \Omega$, so wird durch
$$\delta_b\colon \mathcal{A} \to [0,1], A \mapsto \begin{cases} 1 & \text{falls } b \in A \\ 0 & \text{sonst} \end{cases}$$
ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$ definiert. Es heißt *Diracmaß*. Wir können damit ein Maß mit Zähldichte $(\rho_\omega)_{\omega\in\Omega}$ wie folgt schreiben:
$$P = \sum_{\omega\in\Omega} \rho_\omega \delta_\omega$$
4. In der Analysis 3 wird gezeigt, dass es ein (eindeutiges) Maß $\lambda\colon \mathcal{B}(\mathbb{R}) \to [0,\infty]$ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gibt, für das $\lambda((a,b]) = b - a$ für alle $a, b \in \mathbb{R}$,$b \geq a$ gilt. Es heißt *Lebesguemaß* (oder Borel-Lebesgue-Maß) auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Allgemeiner gibt es für jedes $n \in \mathbb{N}$ ein eindeutiges Maß $\lambda_n\colon \mathcal{B}(\mathbb{R}^n) \to [0,\infty]$ auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ für das gilt
$$\lambda_n\left(\prod_{i=1}^n (a_i, b_i]\right) = \prod_{i=1}^n (b_i - a_i)$$
für alle $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$ mit $a_i \leq b_i$ für alle $1 \leq i \leq n$. λ_n heißt *n-dimensionales Lebesguemaß* (oder Volumenmaß). $\lambda_n(A)$ wird als Volumen von $A \in \mathcal{B}(\mathbb{R}^n)$ interpretiert.

Beispiel (Kontinuierliche Gleichverteilung auf einem Intervall)*.* Es seien $a, b \in \mathbb{R}$, $a < b$. Dann wird durch
$$P\colon \mathcal{B}(\mathbb{R}) \to [0,1], A \mapsto \frac{\lambda(A \cap [a,b])}{\lambda([a,b])}$$
ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R})$ definiert. Es heißt *(kontinuierliche) Gleichverteilung auf* $[a,b]$ (oder auch uniforme Verteilung auf $[a,b]$), kurz: unif$[a,b]$. Mit ihrer Hilfe können wir ein sinnvolles Wahrscheinlichkeitsmodell für das "Glücksrad" definieren: Ist $f\colon [0,1) \to S^1, t \mapsto \exp(2\pi i t)$ die Darstellung in Polarkoordinaten, so gilt $f^{-1}(A) \in \mathcal{B}([0,1))$ für alle $A \in \mathcal{B}(S^1)$ (Beweis in viel größerer Allgemeinheit später). Wir setzen
$$P\colon \mathcal{B}(S^1) \to [0,1], A \mapsto \text{unif}[0,1)(f^{-1}(A))$$
Dann ist $(S^1, \mathcal{B}(S^1), P)$ ein Wahrscheinlichkeitsraum. P heißt die Gleichverteilung auf S^1.

Beispiel (Gleichverteilung in höheren Dimensionen). Für $B \in \mathcal{B}(\mathbb{R}^n)$ mit $0 < \lambda_n(B) < \infty$ definieren wir

$$P\colon \mathcal{B}(\mathbb{R}^n) \to [0,1], A \mapsto \frac{\lambda_n(A \cap B)}{\lambda_n(B)}$$

P ist ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. P heißt Gleichverteilung auf B. *Achtung*: Die Gleichverteilung auf S^1 ist kein Spezialfall, da $\lambda_2(S^1) = 0$.

Bemerkung. In den Beispielen kann man sehen, dass aus $P(A) = 0$ nicht $A = \emptyset$ folgt. Z.B. ist $\mathrm{unif}[a,b](\{x\}) = 0$ für alle $x \in \mathbb{R}$ oder $P(S^1) = 0$ für die Gleichverteilung P auf $[-1,1]^2$, obwohl S^1 sogar überabzählbar ist.

Definition. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine Menge $N \subseteq \Omega$ heißt *Nullmenge* bzgl. P, wenn ein $A \in \mathcal{A}$ mit $N \subseteq A$ und $P(A) = 0$ existiert. Im Fall $N \in \mathcal{A}$ ist dies äquivalent zu $P(N) = 0$. Eine Aussage $\Phi(\omega)$ über ein Ergebnis $\omega \in \Omega$ heißt *P-fast sicher gültig* (oder *P-fast überall gültig*), wenn $\{\omega \in \Omega\colon \Phi(\omega)\}^c$ eine Nullmenge ist.

1.1.6 Interpretation von Wahrscheinlichkeiten

Je nach philosophischem Standpunkt sind verschiedene Interpretationen sinnvoll:

Objektivistische Interpretation durch relative Häufigkeiten Führt man ein Zufallsexperiment mit Werten in Ω wiederholt aus, sagen wir n-mal, so erhält man Daten $\omega_1, \dots, \omega_n \in \Omega$. Die *relative Häufigkeit* eines Ereignisses $A \subseteq \Omega$ ist definiert durch

$$r_{\omega_1,\dots,\omega_n}(A) = \frac{|\{i \in \{1,\dots,n\}\colon \omega_i \in A\}|}{n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\omega_i}(A)$$

$r_{\omega_1,\dots,\omega_n}$ heißt die *empirische Verteilung* gegeben die Beobachtungen $\omega_1, \dots, \omega_n$.
Objektivistische Interpretation: Führt man ein Zufallsexperiment immer wieder aus, so liegt nach vielen Versuchen die relative Häufigkeit von A typischerweise nahe bei $P(A)$.
Später stellen wir dieser Interpretation "innermathematische" Theoreme gegenüber, die "Gesetze der großen Zahlen".

von Mises-Interpretation Versuch einer Verschärfung der objektivistischen Interpretation. Bei unendlicher Wiederholung gilt

$$\lim_{n\to\infty} r_{\omega_1,\dots,\omega_n}(A) = P(A)$$

Für die praktische Anwendung ist dies wenig nützlich, da unendlich viele Wiederholungen in der Realität unmöglich sind. Weitere Schwäche: bei unendlichem Würfeln eines fairen Spielwürfels ist die Konstante Folge $1, 1, \dots$ zwar "extrem untypisch", aber nicht unmöglich.

Subjektivistische Interpretation $P(A)$ bedeutet den Grad meiner Überzeugheit vom Eintreten von A. Die definierenden Bedingungen an P (die "Kolmogorov-Axiome") spielen dann die Rolle von Konsistenzbedingungen an das System meiner subjektiven Überzeugungen.

Glücksspiel-Interpretation Versuch die subjektivistische Interpretation schärfer zu fassen. Das Ereignis A hat die subjektive Wahrscheinlichkeit $P(A)$, wenn ich bereit bin, die folgenden beiden Wetten einzugehen:

1. Wenn das Ereignis A eintritt, *bekomme* ich α€, wenn A^c eintritt, *zahle* ich β€, wobei $\alpha/\beta = P(A^c)/P(A)$.
2. Wenn A eintritt, *zahle* ich α€, wenn A^c eintritt *bekomme* ich β€.

Reale Glückspiele (oder reales Anlegerverhalten) sind viel komplexer als diese Interpretation: Ich bin vielleicht bereit 5ct gegen 1ct zu wetten, aber nicht 5M€ gegen 1M€.

Die Probleme bei der Quantifizierung von Wahrscheinlichkeiten motivieren dazu eine möglichst voraussetzungsarme Interpretation zu versuchen, die *Minimal-Interpretation von Wahrscheinlichkeiten*:

1. Wahrscheinlichkeiten $P(A)$ nahe bei 1 bedeuten: A tritt "praktisch sicher" ein.
2. Wahrscheinlichkeiten $P(A)$ nahe bei 0 bedeuten: A ist "praktisch unmöglich".
3. Wahrscheinlichkeiten $P(A)$, die weder nahe bei 0 noch nahe bei 1 liegen, bedeuten Unsicherheit. Es sind Rechengrößen.

1.2 Verteilungsfunktionen und Eindeutigkeitssatz für Wahrscheinlichkeitsmaße

Als Funktionen auf $\mathcal{B}(\mathbb{R})$ sind Wahrscheinlichkeitsmaße über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ sehr komplexe Gebilde. Man kann sie jedoch in einer viel einfacheren Funktion $\mathbb{R} \to \mathbb{R}$ kodieren:

Definition. Sei P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Die Funktion $F\colon \mathbb{R} \to [0,1], x \mapsto P((-\infty, x])$ heißt die *Verteilungsfunktion* von P.

Beispiel. Die Verteilungsfunktion von unif$[0,1]$ ist gegeben durch

$$F(x) = \begin{cases} 0 & \text{für } x < 0 \\ x & \text{für } 0 \leq x \leq 1 \\ 1 & \text{für } x > 1 \end{cases}$$

Beispiel. Die Verteilungsfunktion von δ_a, $a \in \mathbb{R}$, lautet:

$$F(x) = \begin{cases} 0 & \text{für } x < a \\ 1 & \text{für } x \geq a \end{cases}$$

Beispiel. Modell für den fairen Münzwurf: $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1)$. Die Verteilungsfunktion von P lautet:

$$F(x) = \frac{1}{2}\left(1_{(-\infty,x]}(0) + 1_{(-\infty,x]}(1)\right) = \begin{cases} 0 & \text{für } x < 0 \\ \frac{1}{2} & \text{für } 0 \leq x < 1 \\ 1 & \text{für } x \geq 1 \end{cases}$$

Hier haben wir die *Indikatorfunktion* verwendet. Für $A \subseteq \Omega$:

$$1_A\colon \Omega \to \{0,1\}, \omega \mapsto \begin{cases} 1 & \text{für } \omega \in A \\ 0 & \text{für } \omega \notin A \end{cases}$$

Lemma (charakteristische Eigenschaften von Verteilungsfunktionen)**.** *Für jede Verteilungsfunktion F eines Wahrscheinlichkeitsmaßes P über $\mathbb{R}$ gilt:*

1. *F ist monoton steigend.*
2. *F ist rechtsseitig stetig.*
3. $\lim\limits_{x\to\infty} F(x) = 1$
4. $\lim\limits_{x\to-\infty} F(x) = 0$

Beweis.

1. Seien $x, y \in \mathbb{R}$ mit $x \leq y$. Dann gilt $(-\infty, x] \subseteq (-\infty, y]$, also $F(x) = P((-\infty, x]) \leq P((-\infty, y]) = F(y)$.
2. Es sei $(x_n)_{n\in\mathbb{N}}$ eine monoton fallende Folge in $\mathbb{R}$ mit $x_n \xrightarrow{n\to\infty} x \in \mathbb{R}$. Dann ist $((-\infty, x_n])_{n\in\mathbb{N}}$ eine monton fallende Folge aus $\mathcal{B}(\mathbb{R})$ mit

$$\bigcap_{n\in\mathbb{N}} (-\infty, x_n] = (-\infty, x]$$

   Mit der σ-Stetigkeit von oben von P folgt

$$F(x_n) = P((-\infty, x_n]) \xrightarrow{n\to\infty} P\Big(\bigcap_{n\in\mathbb{N}} (-\infty, x_n]\Big) = P((-\infty, x]) = F(x)$$

3. Es sei $(x_n)_{n\in\mathbb{N}}$ eine monoton steigende Folge in $\mathbb{R}$ mit $x_n \xrightarrow{n\to\infty} \infty$. Die Folge $((-\infty, x_n])_{n\in\mathbb{N}}$ ist monoton steigend mit

$$\bigcup_{n\in\mathbb{N}} (-\infty, x_n] = \mathbb{R}$$

   Mit der σ-Stetigkeit von unten von P folgt.

$$F(x_n) = P((-\infty, x_n]) \xrightarrow{n\to\infty} P\Big(\bigcup_{n\in\mathbb{N}} (-\infty, x_n]\Big) = P(\mathbb{R}) = 1$$

4. Es sei $(x_n)_{n\in\mathbb{N}}$ eine monoton fallende Folge in $\mathbb{R}$ mit $x_n \xrightarrow{n\to\infty} -\infty$. Die Folge $((-\infty, x_n])_{n\in\mathbb{N}}$ ist monoton fallend mit

$$\bigcap_{n\in\mathbb{N}} (-\infty, x_n] = \emptyset$$

   Mit der σ-Stetigkeit von oben von P folgt.

$$F(x_n) = P((-\infty, x_n]) \xrightarrow{n\to\infty} P\Big(\bigcap_{n\in\mathbb{N}} (-\infty, x_n]\Big) = P(\emptyset) = 0$$ □

Bemerkung. Es gilt $\lim_{x\nearrow a} F(x) = P((-\infty, a))$.

Bemerkung. Wir werden später sehen, dass jede Funktion $F\colon \mathbb{R} \to [0, 1]$, die die Eigenschaften 1 – 4 des Lemmas erfüllt, die Verteilungsfunktion einers Wahrscheinlichkeitsmaßes P über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ist.

Verteilungsfunktionen charakterisieren das zugehörige Wahrscheinlichkeitsmaß eindeutig:

Satz. *Sind P,Q zwei Wahrscheinlichkeitsmaße über* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *mit der gleichen Verteilungsfunktion von F. Dann gilt* $P = Q$.

Bemerkung. Zwei *verschiedene* Wahrscheinlichkeitsmaße P und Q auf $(\Omega, \mathcal{A})$ können auf einem Erzeuger von $\mathcal{A}$ übereinstimmen.

Definition. Sei Ω ein Ergebnisraum und $\mathcal{M} \subseteq \mathcal{P}(\Omega)$. $\mathcal{M}$ heißt *durchschnittstabil* (kurz σ-stabil oder Π-System), wenn für alle $A, B \in \mathcal{M}$ $A \cap B \in \mathcal{M}$ gilt. Ist $\mathcal{M} \subseteq \mathcal{A}$ mit einer σ-Algebra $\mathcal{A}$ mit $\sigma(\mathcal{M}) = \mathcal{A}$, so heißt $\mathcal{M}$ ein *durschnittstabiler Erzeuger* von $\mathcal{A}$.

Zum Beweis des Eindeutigkeitssatzes ist folgende Abschwächung des Begriffs der σ-Algebra nützlich:

Definition. Sei Ω eine Ergebnisraum. Ein Mengensystem $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ heißt *Dynkin-System* über Ω, wenn gilt:

1. $\emptyset \in \Omega$.
2. Aus $A \in \mathcal{D}$ folgt $A^c \in \mathcal{D}$.
3. Ist $(A_n)_{n\in\mathbb{N}}$ eine Folge von paarweise disjunkten Mengen aus $\mathcal{D}$, so gilt $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{D}$.

Es ist oft leichter zu sehen, dass ein System ein Dynkin-System ist, als zu zeigen, dass es eine σ-Algebra ist. Ein Beispiel:

Lemma. *Sind P, Q zwei Wahrscheinlichkeitsmaße über $(\Omega, \mathcal{A})$, so ist $\mathcal{D} = \{A \in \mathcal{A} \colon P(A) = Q(A)\}$ ein Dynkin-System.*

Beweis. Offensichtlich ist $P(\emptyset) = Q(\emptyset) = 0$, also $\emptyset \in \mathcal{D}$. Aus $A \in \mathcal{D}$, also $P(A) = Q(A)$, folgt $P(A^c) = 1 - P(A) = 1 - Q(A) = Q(A^c)$, also $A^c \in \mathcal{D}$. Nun sei $(A_n)_{n\in\mathbb{N}}$ eine Folge paarweise disjunkter Mengen in $\mathcal{D}$. Dann gilt

$$P\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} P(A_n) = \sum_{n\in\mathbb{N}} Q(A_n) = Q\Big(\bigcup_{n\in\mathbb{N}} A_n\Big),$$

also $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{D}$. □

Beispiel. $(\Omega, \mathcal{A}) = (\{1,2,3,4\}, \mathcal{P}(\{1,2,3,4\}))$. $P = \frac{1}{4}(\delta_1 + \delta_2 + \delta_3 + \delta_4)$, $Q = \frac{1}{2}(\delta_1 + \delta_4)$. Hier ist $\mathcal{D} = \{A \in \mathcal{A} \colon P(A) = Q(A)\}$ *keine* σ-Algebra, aber ein Dynkin-System mit $\sigma(\mathcal{D}) = \mathcal{P}(\Omega)$. Außerdem ist $\mathcal{D}$ nicht durchschnittstabil.

Der Kernpunkt des Beweises des Eindeutigkeitssatzes steckt in folgendem mengentheoretischen Lemma.

Lemma (Dynkin-Lemma od. Π-Λ-Theorem)**.** *Sei Ω eine Ergebnisraum, $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ ein durchschnittstabiles System und $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ ein Dynkin-System über Ω. Dann gilt: Ist $\mathcal{M} \subseteq \mathcal{D}$, so auch $\sigma(\mathcal{M}) \subseteq \mathcal{D}$.*

Beweis. `http://www.mathematik.uni-muenchen.de/~merkl/ws10/dynkin.pdf`

Satz (Eindeutigkeitssatz für Wahrscheinlichkeitsmaße)**.** *Sind P, Q zwei Wahrscheinlichkeitsmaße über dem gleichen Ereignisraum $(\Omega, \mathcal{A})$ und ist $\mathcal{M}$ ein durchschnittstabiler Erzeuger von $\mathcal{A}$, auf dem P und Q übereinstimmen, d.h. $\forall A \in \mathcal{M}.\ P(A) = Q(A)$, so gilt $P = Q$.*

Beweis. Seien P, Q zwei Wahrscheinlichkeitsmaße über $(\Omega, \mathcal{A})$, $\mathcal{M}$ ein durchschnittstabiler Erzeuger von $\mathcal{A}$ mit $P|\mathcal{M} = Q|\mathcal{M}$. Dann folgt:

$$\mathcal{M} \subseteq \mathcal{D} := \{A \in \mathcal{A} \colon P(A) = Q(A)\}$$

Weil $\mathcal{D}$ ein Dynkin-System ist folgt $\mathcal{A} = \sigma(\mathcal{M}) \subseteq \mathcal{D}$, also $P = Q$. □

Bemerkung. Im Spezialfall $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ bildet $\mathcal{M} = \{(-\infty, x] \colon x \in \mathbb{R}\}$ einen durschnittstabilen Erzeuger von $\mathcal{B}(\mathbb{R})$. In der Tat ist $\sigma(\mathcal{M}) = \mathcal{B}(\mathbb{R})$ und für alle $x, y \in \mathbb{R}$ gilt:

$$(-\infty, x] \cap (-\infty, y] = (-\infty, \min(x, y)] \in \mathcal{M}$$

Der obige Eindeutigkeitssatz für Verteilungsfunktionen ist also eine Konsequenz des allgemeinen Eindeutigkeitssatzes für Wahrscheinlichkeitsmaße.

Bemerkung. Für allgemeine (nicht endliche) Maße gilt der Eindeutigkeitssatz im Allgemeinen nicht. Z.B. ist $\lambda \neq 2\lambda$, aber $\lambda((-\infty, x]) = \infty = 2\lambda((-\infty, x])$.

Definition. . Für $t \in \mathbb{R}$ sei

$$D_t = \begin{pmatrix} \cos(2\pi t) & -\sin(2\pi t) \\ \sin(2\pi t) & \cos(2\pi t) \end{pmatrix}$$

Für $A \subseteq S^1$ sei

$$D_t A = \left\{ D_t \begin{pmatrix} x \\ y \end{pmatrix} : \begin{pmatrix} x \\ y \end{pmatrix} \right\}$$

das um $2\pi t$ Gedrehte von A. Man kann zeigen, dass $D_t A \in \mathcal{B}(S^1) \iff A \in \mathcal{B}(S^1)$.

Lemma (Rotationsinvarianz der Gleichverteilung auf S^1)**.** *Für alle $A \in \mathcal{B}(S^1)$ und alle $t \in \mathbb{R}$ gilt:*

$$P(A) = P(D_t A)$$

Beweisskizze. Die Aussage ist offensichtlich richtig für alle A der Gestalt $\{\exp(2\pi x) \colon x \in I\}$, wenn I ein Intervall in $[0, 1)$ ist. Sei $\mathcal{M}$ alles A dieser Gestalt. $\mathcal{M}$ ist ein durchschnittstabiler Erzeuger von $\mathcal{B}(S^1)$. Weiter ist $\mathcal{D} = \{A \in \mathcal{B}(S^1) \colon P(D_t A) = P(A)\}$ ein Dynkin-System, das offensichtlich $\mathcal{M}$ umfasst. Aus dem Eindeutigkeitssatz folgt $\mathcal{D} = \mathcal{B}(S^1)$ also die Behauptung. □

Satz. *Es gibt* kein *rotationsinvariantes Wahrscheinlichkeitsmaß $Q \colon \mathcal{P}(S^1) \to [0, 1]$, d.h. kein Maß Q, für das gilt*

$$\forall A \subseteq S^1 \ \forall t \in \mathbb{R}.\ Q(A) = Q(D_t A)$$

Beweis. wir definieren folgende Relation $\sim$ auf S^1:

$$x \sim y \iff \exists t \in \mathbb{Q}.\ D_t x = y$$

Es ist leicht zu sehen, dass $\sim$ eine Äquivalenzrelation ist. Es sei $[x] = \{y \in S^1 \colon y \sim x\}$ die Äquivalenzklasse von $x \in S^1$ und $\sim$ und $S^1/\sim = \{[x] \colon x \in S^1\}$ die Menge der Äquivalenzklassen. Es sei $f \colon S^1/\sim \to S^1$ eine *Auswahlfunktion*, also eine Abbildung, die jeder Äquivalenzklasse $[x]$ ein Element $f([x]) \in [x]$ zuordnet. (f existiert nach dem Auswahlaxiom der Mengenlehre, aber kann nicht konstruktiv angegeben werden.) Es gilt also $f([x]) \sim x$ für alle $x \in S^1$. Sei $A = \operatorname{im} f = \{f([x]) \colon x \in S^1\}$. Dann ist $(D_t A)_{t \in \mathbb{Q} \cap [0,1)}$ eine Zerlegung von S^1 mit abzählbar vielen Mengen (insbesondere sind die $D_t A$ paarweise disjunkt). Wäre nun Q ein rotationsinvariantes Wahrscheinlichkeitsmaß auf $(S^1, \mathcal{P}(S^1))$, so folgte

$$1 = Q(S^1) = Q\Big(\bigsqcup\nolimits_{t \in \mathbb{Q} \cap [0,1)} D_t A\Big) = \sum_{t \in \mathbb{Q} \cap [0,1)} Q(D_t A) = \sum_{t \in \mathbb{Q} \cap [0,1)} Q(A)$$

Das ist weder verträglich mit $Q(A) = 0$ noch mit $Q(A) > 0$, ein Widerspruch. □

Korollar. $\mathcal{P}(S^1) \neq \mathcal{B}(S^1)$

1.3 Borel-messbare Funktionen und Maße mit Dichten

Definition. Es sei $(\Omega, \mathcal{A})$ ein Ereignisraum. Eine Funktion $f \colon \Omega \to \overline{\mathbb{R}}$, heißt *Borel-messbar* bezüglich $\mathcal{A}$ (bzw. *messbar*), wenn für alle $a \in \mathbb{R}$ gilt:

$$f^{-1}([-\infty, a]) = \{\omega \in \Omega \colon f(\omega) \leq a\} \in \mathcal{A}$$

Dies ist ein Spezialfall des Begriffs messbarer Funktionen, den wird später besprechen.

Beispiel. Alle stetigen Funktionen $f\colon \mathbb{R} \to \overline{\mathbb{R}}$ sind Borel-messbar bezüglich $\mathcal{B}(\mathbb{R})$, denn Urbilder abgeschlossener Mengen unter stetigen Abbildungen sind abgeschlossen und $[-\infty, a]$ ist abgeschlossen; also ist $f^{-1}([-\infty, a])$ abgeschlossen und damit eine Borelmenge.

Beispiel. Ist $A \in \mathcal{A}$, dann ist $1_A\colon \Omega \to \mathbb{R}$ messbar.

Beispiel. Sind $f, g\colon \Omega \to \mathbb{R}$ messbar, sind auch $\alpha f + \beta g$, $\alpha.\beta \in \mathbb{R}$ messbar.

Beispiel. Ist $(f_n)_{n\in\mathbb{N}}$ eine Folge messbarer Funktionen, so sind $\liminf_{n\to\infty} f_n$ und $\limsup_{n\to\infty} f_n$ (punktweise zu lesen) wieder messbar. Existiert $\lim_{n\to\infty} f_n$ puntweise, so ist auch $\lim_{n\to\infty} f_n$ wieder messbar.

Für nichtnegative messbare Funktionen $f\colon \Omega \to [0, \infty]$ wird in der Maßtheorie ein Integral

$$\int_\Omega f \,\mathrm{d}\mu \in [0, \infty]$$

bezüglich eines Maßes $\mu\colon \mathcal{A} \to [0, \infty]$ definiert, und zwar so:

$$\int_\Omega f \,\mathrm{d}\mu = \sup\Big\{\sum_{i=1}^n \alpha_i \mu(A_i)\colon n \in \mathbb{N} \wedge \alpha_i \geq 0 \wedge A_i \in \mathcal{A} \wedge \sum_{i=1}^n \alpha_i 1_{A_i} \leq f\Big\}$$

Das Integral wird durch folgende Eigenschaften charakterisiert:

1. Für alle $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \geq 0$, $A_1, \ldots, A_n \in \mathcal{A}$ gilt:
$$\int_\Omega \sum_{i=1}^n \alpha_i 1_{A_i} \,\mathrm{d}\mu = \sum_{i=1}^n \alpha_i \mu(A_i)$$
2. "Satz von der monotonen Konvergenz" Ist $0 \leq f_1 \leq f_2 \leq \ldots$ eine aufsteigende Folge messbarer Funktionen $\Omega \to [0, \infty]$ und setzen wir $f(\omega) = \lim_{n\to\infty} f_n(\omega)$ für $\omega \in \Omega$, so ist auch f messbar und es gilt
$$\int_\Omega f \,\mathrm{d}\mu = \lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}\mu$$

Bemerkung. Ist $f\colon \mathbb{R}^n \to [0, \infty]$ stückweise stetig, so existiert das (uneigentliche) Riemannintegral $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$, f ist messbar und es gilt

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

Wir schreiben auch:

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x \text{ statt } \int_{\mathbb{R}} f \,\mathrm{d}\lambda$$

Andere Notation:

$$\int_\Omega f \,\mathrm{d}\mu = \int_\Omega f(\omega)\,\mu(\mathrm{d}\omega)$$

Bemerkung. Sind $f, g\colon \Omega \to [0, \infty]$ messbar und μ-fast überall gleich, d.h. $\mu(\{\omega \in \Omega\colon f(\omega) \neq g(\omega)\}) = 0$, so gilt

$$\int_\Omega f \,\mathrm{d}\mu = \int_\Omega g \,\mathrm{d}\mu$$

Beim Integral kommt es also auf Nullmengen nicht an.

Bemerkung. Linearität: Sind $f, g \geq 0$ messbare Funktionen und $\alpha, \beta \in \mathbb{R}^+$, so gilt

$$\int_\Omega (\alpha f + \beta g) \,\mathrm{d}\mu = \alpha \int_\Omega f \,\mathrm{d}\mu + \beta \int_\Omega g \,\mathrm{d}\mu$$

Bemerkung. Monotonie: Sind $f, g\colon \Omega \to [0,\infty]$ messbar mit $f \le g$, so gilt

$$\int_\Omega f \,\mathrm{d}\mu \le \int_\Omega g \,\mathrm{d}\mu$$

Beispiel. Ist Ω abzählbar, $\mathcal{A} = \mathcal{P}(\Omega)$, μ ein Maß auf $(\Omega, \mathcal{A})$ mit Zähldichte $\rho = (\rho_\omega)_{\omega\in\Omega}$, so gilt für alle $f\colon \Omega \to [0,\infty]$

$$\int_\Omega f \,\mathrm{d}\mu = \sum_{\omega\in\Omega} f(\omega)\rho_\omega$$

Satz. *Ist μ ein Maß auf dem Ereignisraum $(\Omega, \mathcal{A})$ und ist $f\colon \Omega \to [0,\infty]$ messbar, so wird durch*

$$\nu\colon \mathcal{A} \to [0,\infty], A \mapsto \int_\Omega f \cdot 1_A \,\mathrm{d}\mu =: \int_A f \,\mathrm{d}\mu$$

ein Maß auf $(\Omega, \mathcal{A})$ definiert. ν ist ein Wahrscheinlichkeitsmaß genau dann, wenn

$$\int_\Omega f \,\mathrm{d}\mu = 1$$

Sprechweise: Wir sagen, dass ν bezüglich μ eine Dichte *f besitzt, wenn ν wie eben gegeben. Ist sogar $\int_\Omega f \,\mathrm{d}\mu = 1$, so heißt f eine* Wahrscheinlichkeitsdichte *bezüglich μ. (Notation: $f = \frac{d\nu}{d\mu}$)*

Beweis. Es gilt:

- $\nu(\emptyset) = \int_\emptyset f \,\mathrm{d}\mu = \int_\Omega f 1_\emptyset \,\mathrm{d}\mu = 0$
- $\nu(\Omega) = \int_\Omega f \,\mathrm{d}\mu$
- Sind $A_1, A_2, \ldots \in \mathcal{A}$ paarweise disjunkt, so gilt:

$$\begin{aligned}
\nu\Big(\bigsqcup_{n\in\mathbb{N}} A_n\Big) &= \int_\Omega f 1_{\bigsqcup_{n\in\mathbb{N}} A_n} \,\mathrm{d}\mu = \int_\Omega \sum_{n\in\mathbb{N}} f 1_{A_n} \,\mathrm{d}\mu = \\
&= \int_\Omega \lim_{m\to\infty} \sum_{n=1}^m f 1_{A_n} \,\mathrm{d}\mu \overset{\text{mon. Konv.}}{=} \lim_{m\to\infty} \int_\Omega \sum_{n=1}^m f 1_{A_n} \,\mathrm{d}\mu = \\
&= \lim_{m\to\infty} \sum_{n=1}^m \int_\Omega f 1_{A_n} \,\mathrm{d}\mu = \lim_{m\to\infty} \sum_{n=1}^m \nu(A_n) = \sum_{n\in\mathbb{N}} \nu(A_n)
\end{aligned}$$

□

Beispiel. Ist Ω abzählbar, $\mathcal{A} = \mathcal{P}(\Omega)$, μ das Zählmaß, und besitzt ν die Dichte f bezüglich μ, so ist f die Zähldichte von ν.

Beispiel. Im Fall $\mu = \lambda$ hat man folgende Veranschaulichung: Ist f eine Wahrscheinlichkeitsdichte eines Wahrscheinlichkeitsmaßes P über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ bezüglich λ, so bedeutet $P(A)$ die Fläche unterhalb des Graphen von f über A. Anschaulich bedeutet f die "Wahrscheinlichkeit pro Längeneinheit", daher der Name "Dichte".

Beispiel. Die Gleichverteilung $\mathrm{unif}[a,b]$ besitzt die Dichte $\frac{1}{b-a}1_{[a,b]}$ (bezüglich des Lebesguemaßes λ). Ebenso sind z.B. $\frac{1}{b-a}1_{(a,b)}$ oder $\frac{1}{b-a}1_{(a,b]}$ auch Dichten der gleichen Verteilung $\mathrm{unif}[a,b]$.

Beispiel. Es sei $a > 0$. Das Wahrscheinlichkeitsmaß P auf $\mathcal{B}(\mathbb{R})$ mit der Dichte

$$f(x) = 1_{[0,\infty)}(x) a e^{-ax} \ge 0, \quad x \in \mathbb{R}$$

heißt Exponentialverteilung zum Parameter a. Notation: $P := \mathrm{Exp}(a)$. In der Tat ist P ein Wahrscheinlichkeitsmaß, denn

$$\int_\mathbb{R} f \,\mathrm{d}\lambda = \int_0^\infty a e^{-ax} \,\mathrm{d}x = -e^{-ax}\big|_{x=0}^\infty = 1$$

Beispiel. In der Analysis lernen Sie das "Gaußsche Integral" kennen:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,\mathrm{d}x = \sqrt{2\pi}$$

Also ist $\mathbb{R} \to [0,1], x \mapsto (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2)$ eine Wahrscheinlichkeitsdichte (bezüglich λ). Das Wahrscheinlichkeitsmaß mit dieser Dichte (bezüglich λ) heißt *Standardnormalverteilung* und wird mit normal$(0,1)$ oder $N(0,1)$ abgekürzt.

Beispiel. Seien $a > 0$ ("Skalenparameter") und $s > 0$ ("Formparameter"). Das Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der Dichte

$$f(x) = 1_{(0,\infty)}(x)\frac{a^s}{\Gamma(s)}x^{s-1}e^{-ax} \geq 0, \quad x \in \mathbb{R}$$

wobei

$$\Gamma(s) = \int_0^{\infty} x^{s-1}e^{-x}\,\mathrm{d}x$$

die Gammafunktion bei s bezeichnet, heißt *Gammaverteilung* mit Parametern a und s und wird mit Gamma(a,s) bezeichnet. Es gilt $\mathrm{Exp}(a) = \mathrm{Gamma}(a,1)$.

Bemerkung. Nicht jedes Wahrscheinlichkeitsmaß über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ hat eine Dichte bezüglich λ, z.B. hat δ_0 keine Dichte bezüglich λ. Ein Maß ν heißt *absolut stetig* bezüglich μ, wenn ν eine Dichte bezüglich μ besitzt.

1.3.1 Zusammenhang zwischen Dichten und Verteilungsfunktionen

Sei P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Dichte f, so wird die Verteilungsfunktion F von P wie folgt gegeben:

$$F(a) = P((-\infty, a]) = \int_{(-\infty,a]} f\,\mathrm{d}\lambda = \int_{-\infty}^{a} f(x)\,\mathrm{d}x, \quad a \in \mathbb{R}$$

Aus dem Hauptsatz der Differential- und Integralrechnung folgt: Ist P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit stetig differenzierbarer Verteilungsfunktion F, so ist $F' =: f$ eine Dichte von P. Das Gleiche gilt, wenn F nur stetig und stückweise stetig differenzierbar ist.

Beispiel. Die Exponentialverteilung $\mathrm{Exp}(a)$ hat die Verteilungsfunktion

$$F(t) = \int_{(-\infty,t]} 1_{[0,\infty)}(x)ae^{-ax}\,\mathrm{d}x = 1_{[0,\infty)}(t)\int_0^t ae^{-ax}\,\mathrm{d}x = 1_{[0,\infty)}(t)(1-e^{-at}), \quad t \in \mathbb{R}$$

1.4 Allgemeine messbare Funktionen und Zufallsvariablen

Definition. Seien $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$ zwei Ereignisräume. Eine Abbildung $f\colon \Omega \to \Omega'$ heißt $\mathcal{A}$-$\mathcal{A}'$-*messbar* (oder *messbar*, wenn klar ist, welches $\mathcal{A}$ und welches $\mathcal{A}'$ gemeint ist), wenn für alle $A' \in \mathcal{A}'$ gilt:

$$f^{-1}(A') = \{\omega \in \Omega\colon f(\omega) \in A'\} \in \mathcal{A}$$

Folgendes Kriterium ist nützlich zum Nachweisen der Messbarkeit:

Lemma. *Es seien $(\Omega, \mathcal{A})$, $(\Omega', \mathcal{A}')$ Ereignisräume, $\mathcal{M}' \subseteq \mathcal{A}'$ ein Erzeugendensystem von $\mathcal{A}'$. Dann sind folgende Aussagen über eine Funktion $f\colon \Omega \to \Omega'$ äquivalent:*

1. *f ist $\mathcal{A}$-$\mathcal{A}'$-messbar.*
2. *Für alle $A' \in \mathcal{M}'$ gilt $f^{-1}(A') \in \mathcal{A}$.*

Beweis.

$1 \Rightarrow 2$ trivial
$2 \Rightarrow 1$ Es sei

$$\mathcal{B} = \{A' \in \mathcal{A}' \colon f^{-1}(A') \in \mathcal{A}\}$$

Nach Voraussetzung ist $\mathcal{M}' \subseteq \mathcal{B}$. Zudem ist $\mathcal{B}$ eine σ-Algebra über Ω', denn es gilt:

- $\Omega' \in \mathcal{B}$, denn $f^{-1}(\Omega') = \Omega \in \mathcal{A}$.
- Für $A' \in \mathcal{B}$ folgt $\Omega' \setminus A' \in \mathcal{B}$ wegen $f^{-1}(\Omega' \setminus A') = \Omega \setminus f^{-1}(A') \in \mathcal{A}$.
- Für $A'_1, A'_2, \ldots \in \mathcal{B}$ folgt $\bigcup_{n\in\mathbb{N}} A'_n \in \mathcal{B}$, denn

$$f^{-1}\Big(\bigcup_{n\in\mathbb{N}} A'_n\Big) = \bigcup_{n\in\mathbb{N}} f^{-1}(A'_n) \in \mathcal{A}$$

Es folgt $\mathcal{A}' = \sigma(\mathcal{M}') \subseteq \mathcal{B} \subseteq \mathcal{A}'$, also $\mathcal{B} = \mathcal{A}'$. □

Beispiel. Weil $\{(-\infty, a] \colon a \in \mathbb{R}\}$ ein Erzeugendensystem von $\mathcal{B}(\mathbb{R})$ ist, ist eine Abbildung $f \colon \Omega \to \mathbb{R}$ genau dann $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-messbar, wenn sie Borel-messbar im früheren Sinn ist.

Beispiel. Seien (M, d_M) und (N, d_N) metrische Räume und $f \colon M \to N$ stetig. Dann ist f $\mathcal{B}(M)$-$\mathcal{B}(N)$-messbar, denn das System der offenen Mengen in N ist ein Erzeugendensystem von $\mathcal{B}(N)$. Urbilder offener Mengen in N unter f sind offen, also in $\mathcal{B}(M)$ enthalten.

Beispiel. Jede Abbildung $(\Omega, \mathcal{P}(\Omega)) \to (\Omega', \mathcal{P}(\Omega'))$ ist messbar.

Satz. *Sei $(\Omega, \mathcal{A}, \mu)$ eine Maßraum und $f \colon (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ eine messbare Abbildung. Dann wird durch*

$$\nu \colon \mathcal{A}' \to [0, \infty], A' \mapsto \mu(f^{-1}(A'))$$

ein Maß auf $(\Omega', \mathcal{A}')$ definiert. Es heißt Bildmaß *von μ unter f und wird mit $f[\mu]$ oder μf^{-1} bezeichnet.*

Beweis. ν ist wohldefiniert, weil f messbar ist.

- $\nu(\emptyset) = \mu(f^{-1}(\emptyset)) = \mu(\emptyset) = 0$
- Ist $A'_1, A'_2, \ldots$ eine Folge von paarweise disjunkten Ereignissen in $\mathcal{A}'$, so sind auch die Urbilder $f^{-1}(A'_n)$, $n \in \mathbb{N}$, paarweise disjunkt und messbar. Es folgt

$$\begin{aligned}\nu\Big(\bigsqcup_{n\in\mathbb{N}} A'_n\Big) = \mu\Big(f^{-1}\Big(\bigsqcup_{n\in\mathbb{N}} A'_n\Big)\Big) = \mu\Big(\bigsqcup_{n\in\mathbb{N}} f^{-1}(A'_n)\Big) = \\ = \sum_{n\in\mathbb{N}} \mu(f^{-1}(A'_n)) = \sum_{n\in\mathbb{N}} \nu(A'_n)\end{aligned}$$

□

Bemerkung. Ist μ ein Wahrscheinlichkeitsmaß, so ist auch $\nu = f[\mu]$ ein Wahrscheinlichkeitsmaß, da $\nu(\Omega') = \mu(f^{-1}(\Omega')) = \mu(\Omega) = 1$.

1.4.1 Sprechweisen in der Stochastik

Definition. Sei $(\Omega, \mathcal{A}, P')$ ein Wahrscheinlichkeitsraum, $(\Omega', \mathcal{A}')$ ein Ereignisraum. Eine $\mathcal{A}$-$\mathcal{A}'$-messbare Abbildung $X\colon \Omega \to \Omega'$ heißt auch *Zufallsvariable* mit Werten in $(\Omega', \mathcal{A}')$. Das Bildmaß $X[P]$ heißt auch *Verteilung* von X (unter P) (engl. "law", Notation $\mathcal{L}_P(X) = \mathcal{L}(X)$). Im Fall $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ heißt X eine reelle (oder reellwertige) Zufallsvariable.

Konvention. Für eine Zufallsvariable X schreibt man statt

$$\{\omega \in \Omega\colon X(\omega) \text{ hat die Eigenschaft } \Phi\}$$

kurz $\{X \text{ hat die Eigenschaft } \Phi\}$. Zum Beispiel steht $\{X \in A'\}$ für $X^{-1}(A')$. Für reelle Zufallsvariablen und $a \in \mathbb{R}$ bedeutet zum Beispiel $\{X \leq a\} = X^{-1}((-\infty, a])$. Die Notation wird analog für mehrere Zufallsvariablen verwendet. Sind zum Beispiel X, Y reelle Zufallsvariablen auf dem gleichen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ so steht $\{X < Y\}$ kurz für $\{\omega \in \Omega\colon X(\omega) < Y(\omega)\}$. Eine analoge Notation wird für Wahrscheinlichkeiten verwendet:

$$P(\{\omega \in \Omega\colon X(\omega) \text{ hat die Eigenschaft } \Phi\}) =: P[X \text{ hat die Eigenschaft } \Phi]$$

Zum Beispiel steht $P[X < 2]$ für $P(\{\omega \in \Omega\colon X(\omega) < 2\}) = P(X^{-1}((-\infty, 2])) = \mathcal{L}_P(X)((-\infty, 2])$. Die Verteilungsfunktion F von X, also die Verteilungsfunktion der Verteilung von $\mathcal{L}_P(X)$, lässt sich damit für $a \in \mathbb{R}$ so schreiben:

$$F(a) = P(\{\omega \in \Omega\colon X(\omega) \leq a\}) = P[X \leq a] = PX^{-1}((-\infty, a]) = \mathcal{L}_P(X)((-\infty, a])$$

Beispiel. Es sei $\Omega = \{0,1\}^n$, $\mathcal{A} = \mathcal{P}(\Omega)$, $P = \frac{1}{2^n}\sum_{\omega\in\Omega}\delta_\omega$. Es sei $X_i\colon \Omega \to \mathbb{R}$ die i-te kanonische Projektion. Dann gilt für $a \in \{0,1\}$

$$P[X_i = a] = P(\{(\omega_1, \ldots, \omega_n) \in \Omega\colon \omega_i = a\}) = \frac{2^{n-1}}{2^n} = \frac{1}{2}$$

Also ist $\mathcal{L}_P(X_i) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Die Anzahl der "1"en im Ergebnis des Münzwurfs wird modelliert durch $S = \sum_{i=1}^n X_i$. S ist eine Zufallsvariable. Für sie gilt für $k \in \{0, \ldots, n\}$:

$$P[S = k] = P(\{\omega \in \Omega\colon S(\omega) = k\}) = \binom{n}{k} 2^{-n}$$

Es gilt also:

$$\mathcal{L}_P(S) = \sum_{k=0}^n P[S = k]\delta_k = \sum_{k=0}^n \binom{n}{k} 2^{-n}\delta_k$$

Beispiel. Ist allgemeiner $X\colon (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ eine Zufallsvariable mit endlich vielen Werten $x_1, \ldots, x_n \in \Omega'$ (paarweise verschieden), so gilt

$$\mathcal{L}_P(X) = \sum_{i=1}^n P[X = x_i]\delta_{x_i} = \sum_{i=1}^n P(X^{-1}(\{x_i\}))\delta_{x_i}$$

Beispiel. ist P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ und bezeichnen $X_1, \ldots, X_n\colon \mathbb{R}^n \to \mathbb{R}$ die kanonischen Projektionen, so sind alle X_i stetig, also Zufallsvariablen. Die Verteilung $\mathcal{L}_P(X_i)$ wird die *i-te Randverteilung* von P genannt. Ist zum Beispiel P die uniforme Verteilung auf einem Rechteck $(a, b] \times (c, d] \subseteq \mathbb{R}^2$, so ist die uniforme Verteilung auf $(a, b]$ die 1. Randverteilung und die uniforme Verteilung auf $(c, d]$ die 2. Randverteilung von P.

Als eine Anwendung von Verteilungen von Bildmaßen zeigen wir jetzt die Existenz von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit vorgegebener Verteilungsfunktion.

Satz. *Sei $F\colon \mathbb{R} \to [0,1]$. Dann sind äquivalent:*

1) Es gibt ein Wahrscheinlichkeitsmaß μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der Verteilungsfunktion F.
2) F ist monoton steigend, rechtsseitig stetig und es gilt

$$\lim_{x\to-\infty} F(x) = 0 \text{ und } \lim_{x\to\infty} F(x) = 1$$

Beweis.
1) $\Rightarrow$ 2) Früher gezeigt.
2) $\Rightarrow$ 1) Wir definieren eine "Quasi-Inverse" zu F,

$$G\colon (0,1) \to \mathbb{R}, q \mapsto \sup\{s \in \mathbb{R}\colon F(s) \le q\} = \sup F^{-1}([0,q])$$

G nimmt in der Tat Werte in $\mathbb{R}$ (und nicht etwa $\pm\infty$) an:

- $G(q) > -\infty$ für $q \in (0,1)$ folgt aus $\{s \in \mathbb{R}\colon F(s) \le q\} \neq \emptyset$, denn $F(s) \xrightarrow{s\to-\infty} 0 < q$.
- $G(q) < \infty$ für $q \in (0,1)$, denn für alle genügend großen s gilt $F(s) > q$, denn $F(s) \xrightarrow{s\to\infty} 1 > q$.

Nun sei P die Gleichverteilung $\mathrm{unif}(0,1)$ auf $((0,1), \mathcal{B}((0,1)))$. G ist $\mathcal{B}((0,1))$-$\mathcal{B}(\mathbb{R})$-messbar, da monoton steigend. Also ist das Bildmaß $\mu := \mathcal{L}_P(G)$ definiert. Wir zeigen jetzt, dass μ die Verteilungsfunktion F besitzt: Sei hierzu $s \in \mathbb{R}$ und $q \in (0,1)$. Wir zeigen:

1) Falls $q < F(s)$, gilt $\forall t \in \mathbb{R}(F(t) \le q \Rightarrow t \le s)$.
2) Falls $q > F(s)$, gilt *nicht* $\forall t \in \mathbb{R}(F(t) \le q \Rightarrow t \le s)$.

zu 1) Sei $q < F(s)$ und $t \in \mathbb{R}$ mit $F(t) \le q$. Dann folgt $F(t) \le q < F(s)$, also $t \le s$ wegen der Monotonie von F.
zu 2) Sei $q > F(s)$. Weil F rechtsseitig stetig in s ist, gibt es ein $t > s$ mit $q \ge F(t)$. Das bedeutet $\exists t \in \mathbb{R}(F(t) \le q \wedge t > s)$.

Damit haben wir gezeigt

1') $q < F(s) \Longrightarrow G(q) \le s$
2') $q > F(s) \Longrightarrow G(q) > s$

Es folgt $(0, F(s)) \subseteq \{q \in (0,1)\colon G(q) \le s\} \subseteq (0, F(s)]$ und daher

$$F(s) = P((0,F(s))) \le P[G \le s] \le P((0,F(s)] \cap (0,1)) = F(s)$$

also $P[G \le s] = F(s)$. □

Bemerkung. Der Satz liefert uns ein praktisches Verfahren zur Simulation von Zufallszahlen mit einer vorgegebenen Verteilungsfunktion F, wenn $\mathrm{unif}(0,1)$-verteilte Zufallszahlen S gegeben sind: $G(S)$ leistet das gewünschte.

Beispiel. Verfahren zur Simulation Exp(1)-verteilter Zufallszahlen: Ist ω eine $\mathrm{unif}(0,1)$-verteilte Zufallszahl, so ist $-\log(1-\omega)$ eine Exp(1)-verteilte Zufallszahl. In der Tat: Die Verteilungsfunktion $F(t) = (1-e^{-t})1_{[0,\infty)}(t)$ von Exp(1) besitzt die Quasiinverse bzw. Quantilsfunktion $G(q) = -\log(1-q)$, $q \in (0,1)$. Natürlich ist auch $-\log\omega$ eine Exp(1)-verteilte Zufallszahl, da mit ω auch $1-\omega$ $\mathrm{unif}(0,1)$-verteilt ist.

Definition. Sei μ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Verteilungsfunktion F. Für jedes $q \in (0,1)$ heißt jedes t mit $F(t) = q$ ein *q-Quantil* von F. Etwas allgemeiner heißt jedes $t \in \mathbb{R}$ mit

$$\lim_{s \nearrow t} F(s) \leq q \leq F(t) = \lim_{s \searrow t} F(s)$$

ein *q-Quantil* von F.
Jede Funktion G, die jedem $q \in (0,1)$ ein q-Quantil zuordnet, heißt *Quantilsfunktion*. Insbesondere ist die Quasiinverse G von F eine Quantilsfunktion.

1.5 Berechnung von Dichten und Verteilungen

Wir besprechen zwei Fälle, in denen das Bildmaß unter einer Abbildung eine Dichte hat, wenn das Ausgangsmaß eine Dichte hat.

1.5.1 Dichten von Randverteilungen

Satz 2. *Ist μ ein Maß über $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ mit der Dichte f, d.h. $\mu(A) = \int_A f \, \mathrm{d}\lambda_n$ für alle $A \in \mathcal{B}(\mathbb{R}^n)$, so sei $m < n$ und $\rho \colon \mathbb{R}^n \to \mathbb{R}^m, (x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_m)$ die kanonische Projektion. Dann besitzt $\rho(\mu)$ eine Dichte $g \colon \mathbb{R}^m \to [0, \infty]$, die durch*

$$g(x) = \int_{\mathbb{R}^{n-m}} f(x,y) \lambda_{n-m}(\mathrm{d}y)$$

für $x \in \mathbb{R}^m$ definiert ist.

Dieser Satz beruht auf dem Satz von Fubini für das Lebesgue-Maß für nichtnegative Funktionen.

Satz 3 (Fubini)**.** *Sei $f \colon \mathbb{R}^n \to [0, \infty]$ messbar und $m < n$. Dann ist auch*

$$g \colon \mathbb{R}^m \to [0, \infty], x \mapsto \int_{\mathbb{R}^{n-m}} f(x,y) \lambda_{n-m}(\mathrm{d}y)$$

wohldefiniert und messbar und es gilt

$$\int_{\mathbb{R}^n} f \, \mathrm{d}\lambda_n = \int_{\mathbb{R}^m} g \, \mathrm{d}\lambda_m$$

Bemerkung.

$$\begin{aligned} \int_{\mathbb{R}^n} f(z) \lambda_n(\mathrm{d}z) &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^{n-m}} f(x,y) \lambda_{n-m}(\mathrm{d}y) \lambda_m(\mathrm{d}x) = \\ &= \int_{\mathbb{R}^{n-m}} \int_{\mathbb{R}^m} f(x,y) \lambda_m(\mathrm{d}x) \lambda_{n-m}(\mathrm{d}y) \end{aligned}$$

Beweis. Sei $A \in \mathcal{B}(\mathbb{R}^m)$. Dann ist $\rho^{-1}(A) = A \times \mathbb{R}^{n-m} \in \mathcal{B}(\mathbb{R}^n)$. Dann gilt:

$$\begin{aligned} \rho(\mu)(A) = \mu(\rho^{-1}(A)) = \int_{\rho^{-1}(A)} f \, \mathrm{d}\lambda_n &= \int_{\mathbb{R}^n} \underbrace{1_{\rho^{-1}(A)}}_{1_{A \times \mathbb{R}^{n-m}}} f \, \mathrm{d}\lambda_n = \\ &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^{n-m}} \underbrace{1_{A \times \mathbb{R}^{n-m}}(x,y)}_{1_A(x)} f(x,y) \lambda_{n-m}(\mathrm{d}y) \lambda_m(\mathrm{d}x) = \\ &= \int_A \int_{\mathbb{R}^{n-m}} f(x,y) \lambda_{n-m}(\mathrm{d}y) \lambda_m(\mathrm{d}x) = \\ &= \int_A g(x) \lambda_m(\mathrm{d}x) \end{aligned}$$

Das bedeutet $\rho(\mu)$ hat die Dichte g. Diese wird auch *Randdichte* genannt. □

Bemerkung. Es besteht eine Analogie zum diskreten Fall. Sind Ω_1, Ω_2 endliche Ergebnisräume, ist $\rho\colon \Omega_1 \times \Omega_2 \to \Omega_1, (x,y) \mapsto x$ die erste kanonische Projektion und μ ein Maß auf $(\Omega_1 \times \Omega_2, \mathcal{P}(\Omega_1 \times \Omega_2))$ mit Zähldichte f, also $\mu(A) = \sum_{\omega \in A} f(\omega)$ für $A \subseteq \Omega_1 \times \Omega_2$, so hat $\rho(\mu)$ die Zähldichte $g\colon \Omega_1 \to [0,\infty], x \mapsto \sum_{y \in \Omega_2} f(x,y)$. In der Tat: Für alle $A \subseteq \Omega_1 \times \Omega_2$ gilt:

$$\rho(\mu)(A) = \mu(A \times \Omega_2) = \sum_{(x,y) \in A \times \Omega_2} f(x,y) = \sum_{x \in A} \sum_{y \in \Omega_2} f(x,y) = \sum_{x \in A} g(x)$$

Die Analogie wird noch deutlicher, wenn man die Summen als Integrale über Zählmaße schreibt. Der diskrete und der kontinuierliche Fall sind Spezialfälle des allgemeinen Satzes von Fubini.

Beispiel. Sei P die Gleichverteilung auf der Einheitskreisscheibe $B = \{z \in \mathbb{R}^2\colon \|z\|_2 < 1\}$ und $X\colon \mathbb{R}^2 \to \mathbb{R}$ die Projektion auf die erste Koordinate. Dann besitzt $\mathcal{L}_P(X)$ die Dichte

$$g\colon \mathbb{R} \to [0,\infty], x \mapsto \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{für } |x| < 1 \\ 0 & \text{für } |x| \geq 1 \end{cases}$$

denn due Gleichverteilung P besitzt die Dichte $f\colon \mathbb{R}^2 \to [0,\infty], x \mapsto \frac{1}{\pi} 1_B(x,y)$. Dann gilt

$$f(x) = \begin{cases} \frac{1}{\pi} 1_{(-\sqrt{1-x^2}, \sqrt{1-x^2})}(y) & \text{für } |x| < 1 \\ 0 & \text{für } |x| \geq 1 \end{cases}$$

Es folgt: $\mathcal{L}_P(X)$ hat die Dichte

$$g(x) = \int_{\mathbb{R}} \frac{1}{\pi} 1_B(x,y)\,\mathrm{d}y = \begin{cases} \frac{1}{\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \mathrm{d}y = \frac{2}{\pi}\sqrt{1-x^2} & \text{für } |x| < 1 \\ 0 & \text{für } |x| \geq 1 \end{cases}$$

Beispiel. Sei P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ mit einer Dichte der Gestalt $f\colon \mathbb{R}^2 \to [0,\infty], (x,y) \mapsto g(x)h(y)$ mit zwei Wahrscheinlichkeitsdichten $g, h\colon \mathbb{R} \to [0,\infty]$. Dann haben die beiden Randverteilungen von f die Dichten g bzw. h, denn die erste Randverteilung hat die Dichte

$$x \mapsto \int_{\mathbb{R}} f(x,y)\,\mathrm{d}y = \int_{\mathbb{R}} g(x)h(y)\,\mathrm{d}y = g(x) \int_{\mathbb{R}} h(y)\,\mathrm{d}y = g(x)$$

und analog für $h(y)$. Ist umgekehrt $f(x,y) = g(x)h(y)$ gegeben, so gilt

$$\begin{aligned} \int_{\mathbb{R}^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_{\mathbb{R}} \int_{\mathbb{R}} g(x)h(y)\,\mathrm{d}y\,\mathrm{d}x = \\ &= \int_{\mathbb{R}} g(x) \int_{\mathbb{R}} h(y)\,\mathrm{d}y\,\mathrm{d}x = \int_{\mathbb{R}} h(y)\,\mathrm{d}y \int_{\mathbb{R}} g(x)\,\mathrm{d}x = 1 \end{aligned}$$

Also ist f eine Wahrscheinlichkeitsdichte.

1.5.2 Bildmaße unter Diffeomorphismen

Satz. *Seien $U, V \subseteq \mathbb{R}^n$ offen und $f\colon U \to V$ ein $\mathcal{C}^1$-Diffeomorphismus, d.h. f ist stetig differenzierbar und bijektiv mit stetig differenzierbarer Inversen. Dann gilt für alle messbaren $g\colon V \to [0,\infty]$*

$$\int_V g(y)\lambda_n(\mathrm{d}y) = \int_U g(f(x)) |\det Df(x)| \lambda_m(\mathrm{d}x)$$

Beweis. In der Analysis 3.

Für unsere Zwecke impliziert das

Satz. *Seien $U, V \subseteq \mathbb{R}^n$ offen und $f\colon U \to V$ ein $\mathcal{C}^1$-Diffeomorphismus. Sei weiter P ein Wahrscheinlichkeitsmaß auf $(V, \mathcal{B}(V))$ mit der Dichte g bezüglich λ_n auf $\mathcal{B}(V)$. Dann besitzt $\mathcal{L}_P(f^{-1})$ die Dichte $(g \circ f) \cdot |\det Df|$ bezüglich λ_n auf $\mathcal{B}(U)$.*

Beweis. Für alle $A \in \mathcal{B}(U)$ gilt

$$\begin{aligned}\mathcal{L}_P(f^{-1})(A) = P(f(A)) &= \int_V 1_{f(A)}(y) g(y) \lambda_n(\mathrm{d}y) = \\ &= \int_U 1_{f(A)}(f(x) g(f(x)) |\det Df(x)| \lambda_n(\mathrm{d}x) = \\ &= \int_A g(f(x)) |\det Df(x)| \lambda_n(\mathrm{d}x)\end{aligned}$$

□

Beispiel. Sei P das Wahrscheinlichkeitsmaß auf $((0,\infty)^2, \mathcal{B}((0,\infty)^2)) = (\Omega, \mathcal{A})$ mit der Dichte $g(x,y) = e^{-x}e^{-y}$ für $x, y > 0$. Insbesondere sind die Randverteilungen von P jeweils Exp(1)-Verteilungen. Sei

$$h\colon (0,\infty)^2 \to (0,\infty) \times (0,1), (x,y) \mapsto (x+y, \frac{y}{x+y})$$

ein $\mathcal{C}^1$-Diffeomorphismus mit der Umkehrung

$$f\colon (0,\infty) \times (0,1) \to (0,\infty)^2, (s,t) \mapsto (s - st, st)$$

Die Umkehrabbildung f besitzt besitzt die Jacobimatrix

$$Df(s,t) = \begin{pmatrix} 1-t & -s \\ t & s \end{pmatrix}$$

Also ist $\det Df(s,t) = s$. Es folgt $\mathcal{L}_P(h)$ besitzt die Dichte $(0,\infty)\times(0,1) \ni (s,t) \mapsto g(f(s,t))|\det Df(s,t)| = se^{-(s-st)}e^{-st} = se^{-s}$ bezuglich λ_2 auf $\mathcal{B}((0,\infty)\times(0,1))$. Wir können das auch so formulieren: Bezeichen $X, Y\colon (0,\infty)^2 \to \mathbb{R}$ die Projektionen auf die 1. bzw. 2. Komponente, so besitzt der Zufallsvektor $(X+Y, \frac{Y}{X+Y})$ die Dichte

$$(s,t) \mapsto 1_{(0,\infty)}(s) s e^{-s} 1_{(0,1)}(t), \quad (s,t) \in \mathbb{R}^2$$

Insbesondere ist $X+Y$ Γ-verteilt mit Skalenparameter 1 und Formparameter 2 und $\frac{Y}{X+Y}$ unif$(0,1)$-verteilt.

Beispiel (Simulation standardnormalverteilter Zufallszahlen). Die 2-dimensionale Standardnormalverteilung ist das Wahrscheinlichkeitsmaß P über $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ mit der Dichte

$$f(x,y) = \varphi(x)\varphi(y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$$

wobei

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

die Dichte der Standardnormalverteilung bezeichnet. Insbesondere sind beide Randverteilungen von P standardnormalverteilt. Offensichtlich ist f und damit auch P rotationsinvariant, weil $f(x,y)$ nur vom Radiusquadrat x^2+y^2 abhängt. Dies motiviert folgendes Simulationsverfahren:

Es sei $Z = (U, V)$ ein Zufallsvektor, gleichverteilt auf $(0,1)^2$. Wir bilden: $\phi := 2\pi V$, $R := \sqrt{-2\log U}$, $X := R\cos\phi$ und $Y := R\sin\phi$. Dann ist der Zufallsvektor (X, Y) 2-dimensional standardnormalverteilt, insbesondere sind X und Y (einzeln) standardnormalverteilt. Man beachte, dass $\frac{1}{2}R^2 = -\log U$ Exp(1)-verteilt ist. Begründung des Verfahrens:

Die Abbildung

$$g\colon (0,1)^2 \to \mathbb{R}^2 \setminus ([0,\infty)\times\{0\}), (u,v) \mapsto \Big(\sqrt{-2\log u}\cos(2\pi v), \sqrt{-2\log u}\sin(2\pi v)\Big)$$

ist ein Diffeomorphismus mit der Jacobimatrix

$$Dg(u,v) = \begin{pmatrix} -\frac{2}{u}\frac{1}{2\sqrt{-2\log u}}\cos(2\pi v) & -2\pi\sqrt{-2\log u}\sin(2\pi v) \\ -\frac{2}{u}\frac{1}{2\sqrt{-2\log u}}\sin(2\pi v) & 2\pi\sqrt{-2\log u}\cos(2\pi v) \end{pmatrix}$$

und der Jacobideterminante

$$\det Dg(u,v) = -\frac{2\pi}{u}$$

Für die Umkehrabbildung

$$g^{-1}\colon \mathbb{R}^2 \setminus ([0,\infty)\times\{0\}) \to (0,1)^2, (x,y) \mapsto (u,v)$$

gilt also

$$\det D(g^{-1})(x,y) = \big(\det Dg(u,v)\big)^{-1} = -\frac{u}{2\pi} = -\frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}$$

Nun besitzt die Gleichverteilung auf $(0,1)^2$ die Dichte 1 auf $(0,1)^2$. Nach der Transformationsformel für Dichten folgt $\mathcal{L}_{\mathrm{unif}(0,1)^2}(g)$ besitzt die Dichte

$$f(x,y) = 1 \cdot |\det D(g^{-1})(x,y)| = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}$$

auf $\mathbb{R}^2 \setminus ([0,\infty)\times\{0\})$ und daher $\mathcal{L}_{\mathrm{unif}(0,1)^2}(g\colon (0,1)^2 \to \mathbb{R}^2)$ eine Dichte

$$f(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}1_{\mathbb{R}^2\setminus([0,\infty)\times\{0\})}(x,y), \quad (x,y) \in \mathbb{R}^2$$

Dies ist jedoch auch eine Dichte der 2-dimensionalen Standardnormalverteilung, da $[0,\infty)\times\{0\}$ eine λ_2-Nullmenge ist.

1.6 Die von Zufallsvariablen erzeugte σ-Algebra

Definition. Sei Ω eine Ergebnisraum, $(\Omega', \mathcal{A}')$ ein Ereignisraum und $X\colon \Omega \to \Omega'$ eine Abbildung. Dann ist $\sigma(X) = \{X^{-1}(A')\colon A' \in \mathcal{A}'\}$ eine σ-Algebra. Sie heißt die von X *erzeugte σ-Algebra*. $\sigma(X)$ wird interpretiert als das System der beobachtbaren Ereignisse, wenn nur X beobachtet wird. Ist allgemeiner $(\Omega_i, \mathcal{A}_i)_{i\in I}$ eine Familie von Ereignisräumen und $(X_i\colon \Omega \to \Omega_i)_{i\in I}$ eine Familie von Abbildungen, so heißt

$$\sigma(X_i\colon i \in I) := \sigma(\{X_i^{-1}(A_i)\colon i \in I, A_i \in \mathcal{A}_i\})$$

die von den X_i, $i \in I$, erzeugte σ-Algebra.

Bemerkung. $\sigma(X)$ ist die kleinste σ-Algebra $\mathcal{A}$ über Ω, bezüglich der $X\colon (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ messbar ist. Ebenso ist $\sigma(X_i\colon i \in I)$ die kleinste σ-Algebra über Ω, bezüglich der alle $X_i\colon (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ messbar sind. Eine Abbildung $X\colon (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ ist genau dann messbar, wenn $\sigma(X) \subseteq \mathcal{A}$ ist.

Beispiel. Sind $X_1, \dots, X_n \colon \mathbb{R}^n \to \mathbb{R}$ die kanonischen Projektionen, so ist $\sigma(X_i \colon i = 1, \dots, n) = \sigma(X_1, \dots, X_n) = \mathcal{B}(\mathbb{R}^n)$, wobei $\mathbb{R}$ mit $\mathcal{B}(\mathbb{R})$ versehen wird. Allgemeiner: Sind $(\Omega_1, \mathcal{A}_1), \dots, (\Omega_n, \mathcal{A}_n)$ Ereignisräume, $\Omega = \Omega_1 \times \cdots \times \Omega_n$ und $X_i \colon \Omega \to \Omega_i$, $i = 1, \dots, n$ die kanonischen Projektionen, so heißt $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n := \sigma(X_i \colon i = 1, \dots, n)$ die *Produkt-σ-Algebra* der $\mathcal{A}_1, \dots, \mathcal{A}_n$. Sie wird von den "Quadern" $A_1 \times \cdots \times A_n$, $A_i \in \mathcal{A}_i$, $i = 1, \dots, n$ erzeugt. Ist $(\Omega_i, \mathcal{A}_i)_{i \in I}$ eine Familie von Ereignisräumen, $\Omega = \prod_{i \in I} \Omega_i$, $X_i \colon \Omega \to \Omega_i$ für $i \in I$ die kanonische Projektion, so heißt $\bigotimes_{i \in I} \mathcal{A}_i = \sigma(X_i \colon i \in I)$ die *Produkt-σ-Algebra* der $\mathcal{A}_i$, $i \in I$. Sie enthält *im Allgemeinen nicht* beliebige Quader $\prod_{i \in I} A_i$, $A_i \in \mathcal{A}_i$, $i \in I$ falls I überabzählbar ist. $\bigotimes \mathcal{A}_i$ wird jedoch von den Zylindermengen $\prod_{i \in I} A_i$, $A_i \in \mathcal{A}_i$, $i \in I$ aber $A_i \neq \Omega_i$ nur für höchstens abzählbar viele $i \in I$ erzeugt.

Beispiel. Ist $X \colon \mathbb{R}^2 \to \mathbb{R}, (x, y) \to x$, so ist $\sigma(X)$ die Menge der "Streifen" $A \times \mathbb{R}$, $A \in \mathcal{B}(\mathbb{R})$.

1.7 Elementare bedingte Wahrscheinlichkeit

Definition. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $B \in \mathcal{A}$ ein Ereignis mit $P(B) > 0$. Für jedes $A \in \mathcal{A}$ heißt

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

die *bedingte Wahrscheinlichkeit von A gegeben B.*

Bemerkung. $P(\cdot|B) \colon \mathcal{A} \to [0, 1], A \mapsto P(A|B)$ ist ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$. Es heißt *bedingtes Maß zu P gegeben B.*

Interpretation. Beobachtet man bei einem Zufallsexperiment die Teilinformation, dass B eingetreten ist, so interpretiert man $P(A|B)$ als die neue Wahrscheinlichkeit von A, gegeben diese Teilinformation. Der Nenner $P(B)$ normiert die bedingte Wahrscheinlichkeit, sodass B die bedingte Wahrscheinlichkeit 1 bekommt.

Beispiel. Modellieren wir ein Spielwürfel-Experiment mit $\Omega = \{1, 2, 3, 4, 5, 6, \text{ungültig}\}$, $\mathcal{A} = \mathcal{P}(\Omega)$

$$P = q \frac{1}{6} \sum_{i=1}^{6} \delta_i + (1 - q) \delta_{\text{ungültig}}$$

so dass $q \in (0, 1)$ die Wahrscheinlichkeit eines gültigen Ergebnisses beschreibt. Dann modelliert

$$P(\cdot|\omega \neq \text{ungültig}) = \frac{\frac{q}{6} \sum_{i=1}^{6} \delta_i}{P[\omega \neq \text{ungültig}]} = \frac{1}{6} \sum_{i=1}^{6} \delta_i$$

das Würfelexperiment, bei dem ungültige Ergebnisse ignoriert werden.

Beispiel. Sind $A, B \in \mathcal{B}(\mathbb{R}^n)$, $A \subseteq B$ mit $0 < \lambda_n(A) \leq \lambda_n(B) < \infty$, und ist P die Gleichverteilung auf B über $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, so ist $P(\cdot|A)$ die Gleichverteilung auf A. Praktische Anwendung:

Simulation von Zufallszahlen mit gegebener Dichte: Ist P die Gleichverteilung auf dem Quadrat $((0, 1)^2, \mathcal{B}((0, 1)^2))$ und ist $f \colon (0, 1) \to \mathbb{R}$ eine *beschränkte* Wahrscheinlichkeitsdichte, sagen wir $f \leq c \in (0, \infty)$, so setzen wir $g := f/c \leq 1$, $B = \{(x, y) \in (0, 1)^2 \colon y < g(x)\}$. Dann ist $P(\cdot|B)$ die Gleichverteilung auf B. Bezeichnet $X \colon (0, 1)^2 \to (0, 1), (x, y) \mapsto x$ die 1. Projektion, so hat X bezüglich $P(\cdot|B)$ die Dichte f. Zur praktischen Anwendung wählt man zuerst einen Punkt $\omega = (x, y) \in (0, 1)^2$ gleichverteilt. Ist dann $y < g(x)$, so gibt man das Ergebnis x aus, ansonsten verwirft man ω und startet unabhängig neu. Der Iterationsschritt ist jetzt noch nicht modelliert. (Wir holen das nach.)

Beweis. Für $A \in \mathcal{B}((0,1))$ gilt:

$$\begin{aligned}\mathcal{L}_{P(\cdot|B)}(X \in A) &= P[X \in A|B] = P(X^{-1}(A)|B) = P(A \times (0,1)|B) = \\ &= \frac{P((A \times (0,1)) \cap B)}{P(B)} = \frac{\int_A \int_{(0,g(x))} 1 \,\mathrm{d}y \,\mathrm{d}x}{\int_{(0,1)} \int_{(0,g(x))} 1 \,\mathrm{d}y \,\mathrm{d}x} = \\ &= \frac{\int_A g(x) \,\mathrm{d}x}{\int_{(0,1)} g(x) \,\mathrm{d}x} = \int_A f(x) \,\mathrm{d}x\end{aligned}$$

□

Beispiel ("Stochastische Fallunterscheidung"). "Fälle" werden durch eine endliche oder abzählbare Partition $A_1, \ldots, A_n \in \mathcal{A}$ (bzw. $(A_k)_{k\in\mathbb{N}}$) modelliert. Formel für die totale Wahrscheinlichkeit: Falls $P(A_k) > 0$ für alle k, so gilt für alle $B \in \mathcal{A}$:

$$P(B) = \sum_k P(B|A_k)P(A_k)$$

Beweis. Wegen $B = \bigsqcup_k (B \cap A_k)$ folgt:

$$P(B) = \sum_k P(B \cap A_k) = \sum_k P(B|A_k)P(A_k)$$

□

Beispiel. Ein Spielwürfel wird geworfen, und dann nochmal so oft, wie die Augenzahl des ersten Wurfs anzeigt. Man berechne die Wahrscheinlichkeit, dass ab dem 2. Wurf keine "6" auftritt. Modell $\Omega = \bigsqcup_{k=1}^{k} A_k$, wobei $A_k = \{k\} \times \{1,2,3,4,5,6\}^k$, $\mathcal{A} = \mathcal{P}(\Omega)$ mit dem Modell für P:

- $P(A_k) = \frac{1}{6}$ für alle k
- $P(\{(k, \omega_1, \ldots, \omega_k)\}|A_k) = \frac{1}{6^k}$ für $k, \omega_1, \ldots, \omega_k \in \{1, \ldots, 6\}$, so dass $P(\cdot|A_k)$ die Gleichverteilung auf A_k.

Das bedeutet:

$$P = \sum_{k=1}^{6} \sum_{\omega \in \{1,\ldots,6\}^k} \frac{1}{6} \cdot \frac{1}{6^k} \delta_{(k,\omega)}$$

Es folgt:

$$\begin{aligned}P(\text{ab 2. Wurf keine "6"}) &= \sum_{k=1}^{6} P(A_k) P(\text{ab 2.W. keine "6"}|A_k) \\ &= \sum_{k=1}^{6} \frac{1}{6} \cdot \frac{5^k}{6^k} = 0.554\ldots\end{aligned}$$

Ausblick. Alle $P(B|A_k)$, $k = 1, \ldots, n$, kann man in der einen Zufallsvariable

$$\sum_{k=1}^{n} P(B|A_k) 1_{A_k}$$

zusammenfassen, die auf A_k den Wert $P(B|A_k)$ annimmt. Interpretation: "Prognose für B gegeben die Information aus der σ-Algebra $\mathcal{F} := \sigma(\{A_k \colon k = 1, \ldots, n\})$. Notation: $P(B|\mathcal{F}) := \sum_{k=1}^{n} P(B|A_k) 1_{A_k}$ heißt die bedingte Wahrscheinlichkeit von B gegeben $\mathcal{F}$. Ausblick auf Spezialfall: Ist P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ mit Dichte f, und bezeichnen $X, Y \colon \mathbb{R}^2 \to \mathbb{R}$ die beiden kanonischen Projektionen, so nennen wir für $B \in \mathcal{B}(\mathbb{R}^2)$, $x \in \mathbb{R}$

$$P(B|X = x) = \frac{\int_{\mathbb{R}} 1_B(x,y) f(x,y) \,\mathrm{d}y}{\int_{\mathbb{R}} f(x,y) \,\mathrm{d}y}$$

eine bedingte Wahrscheinlichkeit von P gegeben $X = x$. Wegen der Mehrdeutigkeit von f ist dies nur P-fast überall eindeutig. $y \mapsto f(x,y)(\int_{\mathbb{R}} f(x,t) \,\mathrm{d}t)^{-1}$ heißt die *bedingte Dichte* von Y gegeben $X = x$.

1.8 Die Formel von Bayes

Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, A_k, $k = 1, \ldots, n$ (oder $k \in \mathbb{N}$) eine endliche (oder abzählbare) Partition von Ω mit $P(A_k) > 0$ für alle k. Dann gilt für alle $B \in \mathcal{A}$ mit $P(B) > 0$ unda alle $k = 1, \ldots, n$:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

Beweis.

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$ □

Interpretation. Die Formel von Bayes dient in 2-stufigen Zufallsexperimenten zum "Rückschluss auf die Ursachen".

Beispiel. $n+1$ Urnen, beschriftet mit "0" bis "n" enthalten je n Kugeln, und zwar die Urne Nr. k k rote und $n-k$ blaue Kugeln. Man wählt zufällig eine Urne (1. Stufe) nach der Gleichverteilung und dann aus dieser Urne l Kugeln mit Zurücklegen (2. Stufe). Bedingt darauf, dass r dieser l Kugeln rot sind, mit welcher Wahrscheinlichkeit stammen sie aus der Urne k?

Wir beschreiben die Angaben formal, ohne volles Modell:

- Das Ereignis A_k bedeutet "Urne Nr. k gewählt", $k = 0, \ldots, n$
- Das Ereignis B bedeutet "r rote Kugeln gezogen"

Gegeben sind: $P(A_k) = \frac{1}{n+1}$, $k = 0, \ldots, n$, und

$$P(B|A_k) = \frac{\binom{l}{r} k^r (n-k)^{l-r}}{n^l}$$

Mit der Formel von Bayes folgt

$$P(A_k|B) = \frac{\frac{\binom{l}{r} k^r (n-k)^{l-r}}{n^l} \frac{1}{n+1}}{\sum_{j=0}^{n} \frac{\binom{l}{r} j^r (n-j)^{l-r}}{n^l} \frac{1}{n+1}} = \frac{k^r (n-k)^{l-r}}{\sum_{j=0}^{n} j^r (n-j)^{l-r}}$$

Beispiel (med. Test). 0.01% der Bevölkerung leide an einer Krankheit. Ein medizinischer Test zur Diagnose dieser Krankheit erkenne mit 99% die Krankheit korrekt, wenn der Proband tatsächlich die Krankheit hat. Der Test erkenne mit 98% Wahrscheinlichkeit das Nichtvorliegen der Krankheit korrekt, wenn der Patient die Krankheit nicht hat. Falls der Test das Vorliegen der Krankheit anzeigt, wir groß ist dann die Wahrscheinlichkeit, dass der Proband wirklich die Krankheit hat?

Modell K heiße "der Proband hat die Krankheit", T "der Test zeigt die Krankheit an".
gegeben $P(K) = 10^{-4}$, $P(T \mid K) = 0.99$, $P(T^c \mid K^c) = 0.98$.
gesucht $P(K \mid T)$

Mit der Formel von Bayes folgt:

$$P(K \mid T) = \frac{P(T \mid K)P(K)}{P(T \mid K)P(K) + P(T \mid K^c)P(K^c)} = \frac{0.99 \cdot 10^{-4}}{0.99 \cdot 10^{-4} + 0.02 \cdot (1 - 10^{-4})}$$
$$= 0.0049\ldots$$

Ausblick. Die Formel von Bayes ist das Fundament eines Zweigs der mathematischen Statistik, der "Bayesschen Statistik". Hier wird die 1. Stufe (im Beispiel: die Wahl der Urne) als *zufällige* Wahl eines Wahrscheinlichkeitsmodells interpretiert. Der Satz von Bayes erlaubt dann den Rückschluss von Beobachtungsdaten (2. Stufe) auf das zugrundeliegende Wahrscheinlichkeitsmodell. Mehr dazu später.

1.9 Stochastische Unabhängigkeit

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ mit $P(B) > 0$. Informal gesprochen nennen wir A und B unabhängig, wenn die Kenntnis des Eintretens von B die Prognose für A nicht verändert. Formal: $P(A \mid B) = P(A)$. Schreiben wir das in der Form $P(A \cap B) = P(A)P(B)$, so gibt das Anlass zu folgender Definition:

Definition. Zwei Ereignisse $A, B \in \mathcal{A}$ heißen *stochastisch unabhängig bezüglich* P, wenn gilt:

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

Allgemeiner heißt eine Familie $(A_i)_{i \in I}$ von Ereignissen *stochastisch unabhängig bezüglich* P, wenn für jedes endliche Teilfamilie $(A_i)_{i \in E}$, $\emptyset \neq E \subseteq I$, gilt:

$$P\Big(\bigcap_{i \in E} A_i\Big) = \prod_{i \in E} P(A_i)$$

Beispiel (Zweifacher Wurf eines fairen Würfels). $\Omega = \{1, \dots, 6\}^2$, $\mathcal{A} = \mathcal{P}(\Omega)$, P sei die Gleichverteilung auf Ω. Die Zufallsvariablen $X, Y \colon \Omega \to \{1, \dots, 6\}$ seien die Projektionen auf die erste bzw. zweite Koordinate. Für alle $k, l \in \{1, \dots, 6\}$ sind die Ereignisse $\{X = k\}$ und $\{Y = l\}$ unabhängig. In der Tat gilt

$$P[X = k] = \frac{|\{(k, i) \colon i = 1, \dots, 6\}|}{|\Omega|} = \frac{1}{6}$$

und analog $P[Y = l] = \frac{1}{6}$. Außerdem gilt

$$P[X = k, Y = l] := P(X^{-1}(\{k\}) \cap Y^{-1}(\{l\})) = \frac{|\{(k, l)\}|}{|\Omega|} = \frac{1}{36} = P[X = k]P[Y = l]$$

Beispiel (n-facher Wurf einer unfairen Münze). Sei $\Omega = \{0, 1\}^n$, $\mathcal{A} = \mathcal{P}(\Omega)$, $0 \leq p \leq 1$ mit der Interpretation: "1" an i-ter Stelle bedeutet der i-te Wurf liefert "Kopf", "0" an i-ter Stelle bedeutet der i-te Wurf liefert "Zahl". Wir definieren P durch seine Zähldichte $(p_\omega)_{\omega \in \Omega}$. Für $\omega = (\omega_1, \dots, \omega_n) \in \Omega$ setzen wir:

$$p_\omega = p^{S(\omega)}(1 - p)^{n - S(\omega)}, \quad \text{mit } S(\omega) = \sum_{i=1}^{n} \omega_i$$

Also ist $P = \sum_{\omega \in \Omega} p_\omega \delta_\omega$. In der Tat ist

$$\sum_{\omega \in \Omega} p_\omega = (p + (1 - p))^n = 1^n = 1$$

Also ist P ein Wahrscheinlichkeitsmaß. Es sei $X_i \colon \Omega \to \{0, 1\}, (\omega_1, \dots, \omega_n) = \omega_i$, $i = 1, \dots, n$. Dann sind die Ereignisse $\{X_1 = 1\}, \dots, \{X_n = 1\}$ unabhängig, denn sei $E \subseteq \{1, \dots, n\}$. Dann gilt

$$P[\forall i \in E.\ X_i = 1] = \sum_{\substack{\omega \in \Omega \\ \forall i \in E. \omega_i = 1}} p_\omega = \sum_{\substack{\omega \in \Omega \\ \forall i \in E. \omega_i = 1}} p^{|E|} \prod_{\substack{i=1 \\ i \notin E}}^{n} p^{\omega_i}(1 - p)^{1 - \omega_i} = p^{|E|}(p + (1 - p))^{n - |E|} = p^{|E|}$$

Als Spezialfall $E = \{i\}$ erhalten wir $P[X_i = 1] = p$ für alle $i = 1, \dots, n$, und daher

$$P[\forall i \in E.\ X_i = 1] = p^{|E|} = \prod_{i \in E} P[X_i = 1]$$

Definition. Die Verteilung der Summe S in dem eben besprochenen Modell heißt *Binomialverteilung* zu dem Parametern n und p, kurz $\text{binomial}(n,p) := \mathcal{L}_P(S)$. Die Binomialverteilung beschreibt also die Anzahl des Ergebnisses "Kopf" bei n-fachem unabhängigen Münzwurf mit Wahrscheinlichkeit p von "Kopf" in einem Wurf.

Bemerkung. Es gilt:

$$\text{binomial}(n,p) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k$$

d.h.

$$\text{binomial}(n,p)(A) = \sum_{k \in A} \binom{n}{k} p^k (1-p)^{n-k}$$

denn

$$\text{binomial}(n,k)(\{k\}) = P[S=k] = \sum_{\substack{\omega \in \Omega \\ S(\omega)=k}} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^k$$

Beispiel. Unabhängigkeit ist nicht dasselbe wie paarweise Unabhängigkeit! Seien $\Omega = \{0,1\}^2$, $\mathcal{A} = \mathcal{P}(\Omega)$, P die Gleichverteilung auf Ω und X, Y die kanonischen Projektionen. Sei $Z = X + Y \bmod 2$. Dann sind die Ereignisse $\{X=1\}$, $\{Y=1\}$, $\{Z=1\}$ *paarweise* unabhängig, aber dennoch nicht unabhängig, denn

$$P[X=1, Y=1, Z=1] = P(\emptyset) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P[X=1]P[Y=1]P[Z=1]$$

Definition. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine Familie von Zufallsvariablen $\left(X_i \colon (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i)\right)_{i \in I}$ heißt stochastisch unabhängig bezüglich P, wenn für alle Familien $(A_i \in \mathcal{A}_i)_{i \in I}$ gilt:

$$\left(\{X_i \in A_i\}\right)_{i \in I} = \left(X_i^{-1}(A_i)\right)_{i \in I} \text{ ist unabhängig bezüglich } P$$

Eine Familie von $\cap$-stabilen Ereignissystemen $(\mathcal{M}_i)_{i \in I}$, $\emptyset \neq \mathcal{M}_i \in \mathcal{A}_i$ heißt unabhängig bezüglich P, wenn alle Familien $(A_i \in \mathcal{M}_i)_{i \in I}$ unabhängig sind.

Bemerkung. Nach Definition gilt also für Zufallsvariablen X_i, $i \in I$

$$(X_i)_{i \in I} \text{ unabhängig} \iff \left(\sigma(X_i)\right)_{i \in I} \text{ unabhängig}$$

Beispiel (n-facher Münzwurf). $\Omega = \{0,1\}^n$, $X_i \colon \Omega \to \{0,1\}$ die i-te Projektion. Oben wurde gezeigt, dass $\{X_i = 1\}$, $i = 1, \ldots, n$, unabhängig sind. Es gilt sogar: X_i, $i = 1, \ldots, n$, sind unabhängig (Übung).

Abschlusseigenschaften der Unabhängigkeit:

Lemma. *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $B \in \mathcal{A}$. Dann gilt:*

0. $\emptyset$ ist unabhängig von B.
1. Ist $A \in \mathcal{A}$ unabhängig von B, so ist auch A^c unabhängig von B.
2. Sind $A_n \in \mathcal{A}$, $n \in \mathbb{N}$, paarweise disjunkt und unabhängig von B, so ist auch $\bigcup_{n \in \mathbb{N}} A_n$ unabhängig von B.

Anders gesagt: $\{A \in \mathcal{A} \colon A, B \text{ unabhängig}\}$ ist ein Dynkin-System.

Beweis.

0. $P(\emptyset \cap B) = 0 = P(\emptyset)P(B)$.
1. $P(A^c \cap B) = P(B \setminus (A \cap B)) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = (1 - P(A))P(B) = P(A^c)P(B)$.
2. $P\Big(\bigcup_{n\in\mathbb{N}} A_n \cap B\Big) = \sum_{n\in\mathbb{N}} P(A_n \cap B) = \sum_{n\in\mathbb{N}} P(A_n)P(B) = P\Big(\bigcup_{n\in\mathbb{N}} A_n\Big)P(B)$. □

Korollar. *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $\emptyset \neq \mathcal{B} \subseteq \mathcal{A}$. Dann ist*

$$\mathcal{D} = \{A \in \mathcal{A} \colon \forall B \in \mathcal{B}.\ A, B \text{ unabhängig}\}$$

ein Dynkin-System.

Beweis. $\mathcal{D} = \bigcap_{B\in\mathcal{B}}\{A \in \mathcal{A} \colon A, B \text{ unabhängig}\}$, also ist $\mathcal{D}$ ein Dynkin-System. □

Satz. *Es seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $\mathcal{F}, \mathcal{G} \subseteq \mathcal{A}$ zwei $\cap$-stabile, nichtleere Ereignissysteme. $\mathcal{F}$ und $\mathcal{G}$ seien unabhängig, d.h. $\forall A \in \mathcal{F}\ \forall B \in \mathcal{G}.\ P(A \cap B) = P(A)P(B)$. Dann sind $\sigma(\mathcal{F})$, $\sigma(\mathcal{G})$ unabhängig.*

Beweis. $\mathcal{D} = \{A \in \mathcal{A} \colon \forall B \in \mathcal{G}.\ A, B \text{ unabhängig}\}$ ist ein Dynkin-System mit $\mathcal{F} \subseteq \mathcal{D}$. Da $\mathcal{F}$ $\cap$-stabil ist, folgt aus dem Dynkin-Lemma, dass $\sigma(\mathcal{F}) \subseteq \mathcal{D}$, also dass $\sigma(\mathcal{F})$ und $\mathcal{G}$ unabhängig sind. Nun sei $\mathcal{D}' = \{B \in \mathcal{A} \colon \forall A \in \sigma(\mathcal{F}).\ A, B \text{ unabhängig}\}$. $\mathcal{D}'$ ist ebenfalls ein Dynkin-System, und aus $\sigma(\mathcal{F})$, $\mathcal{G}$ unabhängig folgt $\mathcal{G} \subseteq \mathcal{D}'$. Da $\mathcal{G}$ $\cap$-stabil ist, folgt aus dem Dynkin-Lemma, dass $\sigma(\mathcal{G}) \subseteq \mathcal{D}'$, bzw. dass $\sigma(\mathcal{F})$ und $\sigma(\mathcal{G})$ unabhängig sind. □

Verallgemeinerung.

a) Seien $(\mathcal{F}_i)_{i\in I}$ nichtleere, $\cap$-stabile Ereignissysteme. Ist $(\mathcal{F}_i)_{i\in I}$ unabhängig, so ist auch $\big(\sigma(\mathcal{F}_i)\big)_{i\in I}$ unabhängig.

b) Unter den Voraussetzungen von a) sei $(E_j)_{j\in J}$ eine Familie von paarweise disjunkten Teilmengen von I. Wenn $(\mathcal{F}_i)_{i\in I}$ unabhängig ist, so ist auch

$$\Big(\sigma\Big(\bigcup_{i\in E_j} \mathcal{F}_i\Big)\Big)_{j\in J}$$

unabhängig.

Diese Sätze werden sehr häufig — oft implizit — angewandt:

Beispiel. Sein $X, Y, Z \colon \Omega \to \mathbb{R}$ unabhängige Zufallsvariablen und ist $f \colon \mathbb{R}^2 \to \mathbb{R}$ Borel-messbar, so sind auch $f(X, Y), Z$ unabhängig, wobei $f(X, Y) \colon \Omega \to \mathbb{R}$ definiert ist durch $f(\omega) = f(X(\omega), Y(\omega))$, denn X, Y, Z sind unabhängig genau dann, wenn $\sigma(X)$, $\sigma(Y)$, $\sigma(Z)$ unabhängig sind. Also sind $\mathcal{M} = \{A \cap B \colon A \in \sigma(X), B \in \sigma(Y)\}$ und $\sigma(Z)$ unabhängig. Aber $\mathcal{M}$ ist $\cap$-stabil, also sind $\sigma(\mathcal{M})$ und $\sigma(Z)$ unabhängig. Nun gilt $\sigma(\mathcal{M}) = \sigma(\sigma(X) \cup \sigma(Y)) =: \sigma(X, Y)$. $f(X, Y)$ ist $\sigma(X, Y)$-$\mathcal{B}(\mathbb{R})$-messbar, denn $(X, Y) \colon \Omega \to \mathbb{R}^2$ ist $\mathcal{A}$-$\mathcal{B}(\mathbb{R}^2)$-messbar und $f \colon \mathbb{R}^2 \to \mathbb{R}$ ist $\mathcal{B}(\mathbb{R}^2)$-$\mathcal{B}(\mathbb{R})$-messbar. Also sind $f(X, Y)$, Z unabhängig.

1.10 Unabhängiges Zusammensetzen von zwei Zufallsexperimenten

Satz. *Seien $(\Omega, \mathcal{A}, P)$, $(\Sigma, \mathcal{B}, Q)$ zwei Wahrscheinlichkeitsräume. Dann gibt es genau ein Wahrscheinlichkeitsmaß $P \times Q$ auf $(\Omega \times \Sigma, \mathcal{A} \otimes \mathcal{B})$, so dass die beiden Projektionen $X \colon \Omega \times \Sigma \to \Omega$ und*

$Y\colon \Omega \times \Sigma \to \Sigma$ *bezüglich* $P \times Q$ *unabhängig sind mit* $\mathcal{L}_{P\times Q}(X) = P$ *und* $\mathcal{L}_{P\times Q}(Y) = Q$. $P \times Q$ *heißt Produktmaß von* P *und* Q. $P \times Q$ *wird gegeben durch*

$$\begin{aligned} P \times Q(A) &= \int_\Omega Q((\omega, \mathrm{id}_\Sigma) \in A)\, P(\mathrm{d}\omega) \\ &= \int_\Omega Q(\{\sigma \in \Sigma\colon (\omega, \sigma) \in A\})\, P(\mathrm{d}\omega) \\ &= \int_\Omega \int_\Sigma 1_A(\omega, \sigma)\, Q(\mathrm{d}\sigma)\, P(\mathrm{d}\omega) \qquad (*) \end{aligned}$$

Beweis. In der Maßtheorie.

Definition. Ein Maß μ auf $(\Omega, \mathcal{A})$ heißt σ-endlich, wenn es eine Folge $(A_n)_{n\in\mathbb{N}}$ in $\mathcal{A}$ mit $\mu(A_n) < \infty$ für $n \in \mathbb{N}$ und $\bigcup_{n\in\mathbb{N}} A_n = \Omega$ gibt.

Bemerkung. Auch für allgemeine Maße μ, ν statt P und Q liefert $(*)$ ein Maß $\mu \times \nu$. Es hat gute Eigenschaften unter folgender Zusatzvoraussetzung: μ, ν sind σ-endlich.

Bemerkung. Das Produktmaß σ-endlicher Maße wird durch $\mu \times \nu(A \times B) = \mu(A)\mu(B)$ für $A \in \mathcal{A}, B \in \mathcal{B}$ charakterisiert.

Beispiel. $\lambda_2 = \lambda_1 \times \lambda_1$.

Beispiel. Die Gleichverteilung auf $(a, b] \times (c, d] \subseteq \mathbb{R}^2$ $(a < b,\ c < d)$ ist das Produktmaß der Gleichverteilungen auf $(a, b]$ und $(c, d]$.

Die Integration bezüglich des Produktmaßes kann man auf sukzessive Integration bezüglich der Faktoren zurückführen:

Satz (Fubini)**.** *Seien* $(\Omega, \mathcal{A}, \mu)$ *und* $(\Sigma, \mathcal{B}, \nu)$ *zwei* σ*-endliche Maßräume und* $f\colon \Omega \times \Sigma \to [0, \infty]$ $\mathcal{A} \otimes \mathcal{B}$-$\mathcal{B}([0, \infty])$*-messbar. Dann ist*

$$\Omega \ni \omega \mapsto \int_\Sigma f(\omega, \sigma)\, \nu(\mathrm{d}\sigma)$$

$\mathcal{A}$-$\mathcal{B}([0, \infty])$*-messbar und*

$$\Sigma \ni \sigma \mapsto \int_\Omega f(\omega, \sigma)\, \mu(\mathrm{d}\omega)$$

$\mathcal{B}$-$\mathcal{B}([0, \infty])$*-messbar und es gilt:*

$$\int_{\Omega\times\Sigma} f\, \mathrm{d}(\mu \times \nu) = \int_\Omega \int_\Sigma f(\omega, \sigma)\, \nu(\mathrm{d}\sigma)\, \mu(\mathrm{d}\omega) = \int_\Sigma \int_\Omega f(\omega, \sigma)\, \mu(\mathrm{d}\omega)\, \nu(\mathrm{d}\sigma)$$

Beweis. In der Maßtheorie.

Bemerkung. Der Spezialfall $(\Omega, \mathcal{A}, \mu) = (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m), \lambda_m)$ und $(\Sigma, \mathcal{B}, \nu) = (\mathbb{R}^{n-m}, \mathcal{B}(\mathbb{R}^{n-m}), \lambda_{n-m})$ liefert den früher besprochenen Satz von Fubini.

Bemerkung. Der Spezialfall $(\Omega, \mathcal{A}, \mu) = (\Sigma, \mathcal{B}, \nu) = (\mathbb{N}, \mathcal{P}(\mathbb{N}), |\cdot|)$ kann man als den großen Umordnungssatz für Reihen mit nichtnegativen Koeffizienten auffassen.

Satz. *Sei* $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum,* $X, Y\colon \Omega \to \mathbb{R}$ *Zufallsvariablen mit Dichten* f *bzw.* g*. Dann sind* X *und* Y *unabhängig genau dann, wenn* $h\colon \mathbb{R}^2 \to [0, \infty], (x, y) \mapsto f(x)g(y)$ *eine Dichte für* $Z = (X, Y)\colon \Omega \to \mathbb{R}^2$ *ist.*

Beweis. Übungen.

Verallgemeinerung. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $(\Sigma_1, \mathcal{B}_1, \mu_1)$, $(\Sigma_2, \mathcal{B}_2, \mu_2)$ σ-endliche Maßräume und $X_i \colon \Omega \to \Sigma_i$, $i = 1, 2$, Zufallsvariablen mit Dichten f_i bezüglich m_i. Dann sind X_1 und X_2 genau dann unabhängig, wenn $h \colon \Sigma_1 \times \Sigma_2 \to [0, \infty], (x_1, x_2) \mapsto f_1(x_1) f_2(x_2)$ eine Dichte für $Z = (X_1, X_2) \colon \Omega \to \Sigma_1 \times \Sigma_2$ bezüglich $\mu_1 \times \mu_2$ ist.

Beweis. Übungen.

Bemerkung. Produktmaßbildung ist assoziativ. Die besprochenen Sätze für das Produktmaß lassen sich alle auf endlich viele Faktoren verallgemeinern.

1.11 Die Faltung (engl. convolution)

Definition. Es seien μ_1, μ_2 σ-endliche Maße über $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Die *Faltung* von μ_1 und μ_2, in Zeichen $\mu_1 * \mu_2$, ist definiert als das Bildmaß von $\mu_1 \times \mu_2$ unter der Addition $+ \colon \mathbb{R}^{2n} \to \mathbb{R}^n$. Für Wahrscheinlichkeitsmaße P und Q können wir das auch so formulieren: Sind X und Y unabhängige Zufallsvariablen mit $\mathcal{L}(X) = P$ und $\mathcal{L}(Y) = Q$, so ist $\mathcal{L}(X + Y) = P * Q$.

Beispiel. Ist $P = p\delta_1 + (1-p)\delta_0$ mit $0 \le p \le 1$, so ist

$$\begin{aligned} P * P &= p^2\delta_2 + 2p(1-p)\delta_1 + (1-p)^2\delta_0 \\ P * P * P &= p^3\delta_3 + 3p^2(1-p)\delta_2 + 3p(1-p)^2\delta_1 + (1-p)^3\delta_0 \\ P^{*n} &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k = \text{binomial}(n, p) \end{aligned}$$

Satz. *Seien μ, ν σ-endliche Maße auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ mit Dichten f bzw. g. Dann besitzt $\mu * \nu$ die Dichte $f * g$, definiert durch*

$$f * g(z) = \int_{\mathbb{R}^n} f(x) g(z - x) \, \mathrm{d}x = \int_{\mathbb{R}^n} f(z - y) g(y) \, \mathrm{d}y, \quad \text{für } z \in \mathbb{R}^n$$

$f * g$ heißt ebenfalls Faltung von f mit g.

Beweis. Das Produktmaß $\mu \times \nu$ besitzt die Dichte $h \colon \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto f(x)g(y)$. Nun betrachten wir den Diffeomorphismus

$$k \colon \mathbb{R}^2 \to \mathbb{R}^2, (x, y) \mapsto (x + y, x)$$

mit der Inversen

$$k^{-1} \colon \mathbb{R}^2 \to \mathbb{R}^2, (z, x) \mapsto (x, z - x)$$

sowie die 1. Projektion $\pi \colon \mathbb{R}^2 \to \mathbb{R}, (z, x) \mapsto z$. Dann ist $\pi \circ k = + \colon \mathbb{R}^2 \to \mathbb{R}$. Nun gilt

$$\left| \det Dk^{-1}(z, x) \right| = \left| \det \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \right| = 1$$

Aus der Transformationsformel folgt, dass das Bildmaß $k[\mu \times \nu]$ die Dichte $j \colon \mathbb{R}^2 \to \mathbb{R}$ besitzt mit

$$j(z, x) = h(k^{-1}(z, x)) \left| \det Dk^{-1}(z, x) \right| = h(x, z - x) = f(x)g(z - x), \quad \text{für } (z, x) \in \mathbb{R}^2$$

Also besitzt die 1. Randverteilung $\mu \times \nu = \pi[k[\mu \times \nu]]$ von $k[\mu \times \nu]$ die Dichte

$$\mathbb{R} \ni z \mapsto \int_{\mathbb{R}} j(z, x) \, \mathrm{d}x = \int_{\mathbb{R}} f(x) g(z - x) \, \mathrm{d}x = f * g(z)$$

Analog für $\mathbb{R}^n$. Die zweite Darstellung von $f * g$ folgt mittels Substitution $y = z - x$ oder auch mittels $\mu * \nu = \nu * \mu$. □

Beispiel. Sei $P = \text{unif}[0,1]$. Dann besitzt $P \times P = \text{unif}[0,1]^2$ die Dichte $1_{[0,1]^2}$ und daher ist $P * P$ die Dichte

$$1_{[0,1]} * 1_{[0,1]}(z) = z \cdot 1_{[0,1]}(z) + (2-z) \cdot 1_{(1,2]}(z), \quad \text{für } z \in \mathbb{R}$$

Beispiel. Es seien X, Y unabhängige, Gamma-verteilte Zufallsvariablen:

$$\mathcal{L}_P(X) = \text{Gamma}(a,s) \text{ und } \mathcal{L}_P(Y) = \text{Gamma}(a,t), \quad \text{für } a,s,t > 0$$

Wei zeigen, dass $X + Y$ Gamma$(a, s+t)$-verteil ist:

$$\text{Gamma}(a,s) * \text{Gamma}(a,t) = \text{Gamma}(a,s+t)$$

Beweis. X bzw. Y besitzen die Dichte

$$f(x) = 1_{(0,\infty)}(x) \frac{a^s}{\Gamma(s)} x^{s-1} e^{-ax}, \quad x \in \mathbb{R}$$

bzw.

$$g(y) = 1_{(0,\infty)}(y) \frac{a^t}{\Gamma(t)} y^{t-1} e^{-ay}, \quad y \in \mathbb{R}$$

Es folgt:

$$\begin{aligned} f * g(z) &= \frac{a^{s+t}}{\Gamma(s)\Gamma(t)} \int_{\mathbb{R}} 1_{(0,\infty)}(x) 1_{(0,\infty)}\,(z-x) x^{s-1} (z-x)^{t-1} e^{-ax} e^{-a(z-x)}\,\mathrm{d}x \\ &= \frac{a^{s+t} e^{-az}}{\Gamma(s)\Gamma(t)} 1_{(0,\infty)}(z) \int_0^z x^{s-1}(z-x)^{t-1}\,\mathrm{d}x = \frac{a^{s+t}e^{-az}}{\Gamma(s)\Gamma(t)} 1_{(0,\infty)}(z) \int_0^1 (zu)^{s-1}(z-zu)^{t-1} z\,\mathrm{d}u \\ &= 1_{(0,\infty)}(z) \frac{a^{s+t} B(s,t)}{\Gamma(s)\Gamma(t)} z^{s+t-1} e^{-az}, \quad \text{für } z \in \mathbb{R} \end{aligned}$$

wobei die Betafunktion $B(s,t)$ definiert ist durch

$$B(s,t) = \int_0^1 u^{s-1}(1-u)^{t-1}\,\mathrm{d}u$$

Dies ist bis auf die Konstante $\frac{B(s,t)}{\Gamma(s)\Gamma(t)}$ statt $\frac{1}{\Gamma(s+t)}$ die Dichte von Gamma$(a, s+t)$. Weil sowohl Gamma$(a,s) \times \Gamma(a,t)$ als auch Gamma$(a, s+t)$ Wahrscheinlichkeitsmaße sind, müssen die Konstanten übereinstimmen:

$$\frac{B(s,t)}{\Gamma(s)\Gamma(t)} = \frac{1}{\Gamma(s+t)} \text{ bzw. } B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$$

Also folgt die Behauptung. □

Beispiel. Seien $X_1, \ldots, X_n$ unabhängige, standardnormalverteilte Zufallsvariablen. Sei $\chi_n^2 = \sum_{k=1}^n X_k^2$. Aus einer Hausaufgabe wissen wir $\mathcal{L}(X_k^2) = \text{Gamma}(\frac{1}{2}, \frac{1}{2})$ für alle $k = 1, \ldots, n$. Es folgt:

$$\mathcal{L}(\chi_n^2) = \text{Gamma}\left(\tfrac{1}{2}, \tfrac{1}{2}\right)^{*n} = \text{Gamma}\left(\tfrac{1}{2}, \tfrac{n}{2}\right)$$

Diese Verteilung heißt χ^2-*Verteilung mit n Freiheitsgraden.* Die gemeinsame Verteilung der $X_1, \ldots, X_n$, also die Verteilung des Zufallsvektors $X = (X_1, \ldots, X_n)$, heißt die *n-dimensionale Standardnormalverteilung*, sie besitzt die Dichte

$$\mathbb{R}^n \ni x \mapsto f(x) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_k^2} = (2\pi)^{\frac{n}{2}} e^{-\frac{1}{2}\|x\|_2^2}, \text{ wobei } x = (x_1, \ldots, x_n)$$

bezüglich λ_n. Die χ^2-Verteilung mit n Freiheitsgraden ist also die Verteilung von $\|X\|_2^2$, wenn X n-dimensional standardnormalverteilt ist.

Beispiel. Die Normalverteilung mit den Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ ist die Verteilung von $X = \sigma Z + \mu$, wenn Z standardnormalverteilt ist. Bezeichnung: $N(\mu, \sigma^2)$. Sie besitzt die Dichte

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

denn sei $g\colon \mathbb{R} \to \mathbb{R}, x \mapsto \sigma z + \mu$. Dann gilt für alle $A \in \mathcal{B}(\mathbb{R})$:

$$N(\mu,\sigma^2)(A) = P[g(Z) \in A] = \frac{1}{\sqrt{2\pi}} \int_{g^{-1}(A)} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_A \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \frac{1}{\sigma}\,\mathrm{d}x =$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_A \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x$$

Definition. Die Normalverteilung mit Parametern μ und $\sigma^2 > 0$ ist das Wahrscheinlichkeitsmaß $N(\mu, \sigma^2)$ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Dichte

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

Satz. *Für $\mu_1, \mu_2 \in \mathbb{R}$ und $\sigma_1^2, \sigma_2^2 > 0$ gilt:*

$$N(\mu_1, \sigma_1^2) * N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Anders gesagt: Sind X und Y zweei unabhängige Zufallsvariablen mit den Verteilungen $\mathcal{L}(X) = N(\mu_1, \sigma_1^2)$ und $\mathcal{L}(Y) = N(\mu_2, \sigma_2^2)$, so gilt:

$$\mathcal{L}(X + Y) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Beweis. Wir zeigen:

$$f_{\mu_1,\sigma_1^2} * f_{\mu_2,\sigma_2^2}(x) = f_{\mu_1+\mu_2,\sigma_1^2+\sigma^2}(x)$$

durch direkte Rechnung. (Alternativ auch für n Dimensionen: Übungen). Für $x \in \mathbb{R}$ gilt:

$$f_{\mu_1,\sigma_1} * f_{\mu_2,\sigma_2}(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}} \underbrace{\int \exp\left(-\frac{(x-y-\mu_1)^2}{2\sigma_1^2}\right) \exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right) \mathrm{d}y}_{\mathrm{I}}$$

Wir substituieren $\tilde{x} = x - \mu_1 - \mu_2$ und $\tilde{y} = y - \mu_2$. Damit:

$$\mathrm{I} = \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\underbrace{\frac{(\tilde{x}-\tilde{y})^2}{\sigma_1^2} + \frac{\tilde{y}^2}{\sigma_2^2}}_{\mathrm{II}}\right)\right) \mathrm{d}\tilde{y}$$

Wir schreiben II mit quadratischer Ergänzung um:

$$\mathrm{II} = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)\tilde{y}^2 - \frac{2}{\sigma_1^2}\tilde{x}\tilde{y} + \frac{\tilde{x}^2}{\sigma_1^2} =$$
$$= \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)\left(\tilde{y} - \frac{\sigma_1^{-2}\tilde{x}}{\sigma_1^{-2} + \sigma_2^{-2}}\right)^2 \underbrace{- \frac{\sigma_1^{-4}}{\sigma_1^{-2} + \sigma_2^{-2}}\tilde{x}^2 + \frac{\tilde{x}^2}{\sigma_1^2}}_{\mathrm{III}}$$

Abkürzung: $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. Damit schreiben wir III wie folgt:

$$\mathrm{III} = \frac{-\sigma_1^{-4} + \sigma_1^{-4} + \sigma_1^{-2}\sigma_2^{-2}}{\sigma_1^{-2}\sigma_2^{-2}}\tilde{x}^2 = \frac{\sigma_1^{-2}\sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}}\tilde{x}^2 = \frac{\tilde{x}^2}{\sigma^2}$$

Man beachte: $\sigma_1^{-2} + \sigma_2^{-2} = \left(\frac{\sigma}{\sigma_1\sigma_2}\right)^2$. Eingesetzt erhalten wir:

$$\mathrm{I} = \exp\left(-\frac{\tilde{x}^2}{2\sigma^2}\right) \underbrace{\int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left(\sigma_1^{-2} + \sigma_2^{-2}\right)\left(\tilde{y} - \frac{\sigma_1^{-2}\tilde{x}}{\sigma_1^{-2} + \sigma_2^{-2}}\right)^2\right) \mathrm{d}\tilde{y}}_{\mathrm{IV}}$$

Wir substituieren

$$z = \frac{\sigma}{\sigma_1\sigma_2}\left(\tilde{y} - \frac{\sigma_1^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}}\tilde{x}\right), \quad \frac{dz}{d\tilde{y}} = \frac{\sigma}{\sigma_1\sigma_2}$$

Wir erhalten

$$\mathrm{IV} = \int_{\mathbb{R}} e^{-\frac{z^2}{2}} \left(\frac{dz}{d\tilde{y}}\right)^{-1} \mathrm{d}z = \frac{\sigma_1\sigma_2}{\sigma}\int_{\mathbb{R}} e^{-\frac{z^2}{2}}\,\mathrm{d}z = \frac{\sigma_1\sigma_2}{\sigma}\sqrt{2\pi}$$

Es folgt:

$$\mathrm{I} = \frac{\sigma_1\sigma_2}{\sigma}\sqrt{2\pi}\exp\left(-\frac{\tilde{x}^2}{2\sigma^2}\right)$$

Oben eingesetzt:

$$\begin{aligned} f_{\mu_1,\sigma_1^2} * f_{\mu_2,\sigma_2^2} &= \frac{1}{\sqrt{2\pi\sigma_2^2}}\frac{1}{\sqrt{2\pi\sigma_2^2}}\frac{\sigma_1\sigma_2}{\sigma}\sqrt{2\pi}\exp\left(-\frac{1}{2}\frac{\tilde{x}^2}{\sigma^2}\right) = \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{\tilde{x}^2}{\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \end{aligned}$$ □

Beispiel. Die Poissonverteilung Poisson(λ) zum Parameter $\lambda > 0$ ist die Verteilung auf $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$ mit

$$\mathrm{Poisson}(\lambda) = \sum_{n\in\mathbb{N}_0} e^{-\lambda}\frac{\lambda^n}{n!}\delta_n$$

Es gilt:

$$\mathrm{Poisson}(\lambda_1) * \mathrm{Poisson}(\lambda_2) = \mathrm{Poisson}(\lambda_1 + \lambda_2)$$

1.12 Folgen unabhängiger Zufallsvariablen

Satz. *Es sei $(\Omega_i, \mathcal{A}_i, P_i)_{i\in I}$ eine beliebige Familie von Wahrscheinlichkeitsräumen und $\Omega = \prod_{i\in I}\Omega_i$ das kartesische Produkt der Ω_i, $X_i\colon \Omega \to \Omega_i, (x_j)_{j\in I} \mapsto x_i$, $i \in I$ sei die i-te kanonische Projektion und $\mathcal{A} = \bigotimes_{i\in I}\mathcal{A}_i = \sigma(X_i\colon i \in I)$ die Produkt-σ-Algebra. Dann gibt es genau ein Wahrscheinlichkeitsmaß P auf $(\Omega, \mathcal{A})$, so dass die X_i, $i \in I$, unabhängig mit den Verteilungen $\mathcal{L}_P(X_i) = P_i$ sind. P heißt (allg.)* Produktmaß, *$P = \prod_{i\in I} P_i$.*

Beweis. Maßtheorie.

Alternative. Direkte Konstruktion einer Folge unabhängiger Zufallsvariablen auf dem Wahrscheinlichkeitsraum $\Omega = ([0,1), \mathcal{B}([0,1)), \mathrm{unif}([0,1)))$.

Definition (Unabhängige, identisch verteilte Zufallsvariablen)**.** Eine Familie $(X_i)_{i\in I}$ von Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ heißt *i.i.d.* (engl.: *i*ndepent, *i*dentically *d*istributed), wenn die X_i, $i \in I$ unabhängig mit der gleichen Verteilung $\mathcal{L}_P(X_i) = \mathcal{L}_p(X_j)$, $i, j \in I$, sind.

Alternative (Fort.)**.** Für $\omega \in \Omega$, $(\Omega, \mathcal{A}, P) = ([0,1), \mathcal{B}([0,1)), \mathrm{unif}([0,1)))$, sei $X_n(\omega)$, $n \in \mathbb{N}$, die n-te Nachkommaziffer in der Binärdarstellung von ω, also $X_n(\omega) = [2^n\omega] - 2 \cdot [2^n\omega]$, wobei $[x] = \max\{z \in \mathbb{Z}\colon z \le x\}$.

Satz. *Die Binärziffern $(X_n)_{n\in\mathbb{N}}$ sind i.i.d. $\frac{1}{2}(\delta_0 + \delta_1)$-verteilte Zufallsvariablen über $(\Omega, \mathcal{A}, P)$.*

Beweis. Wir müssen für alle endlichen $E \subset \mathbb{N}$ zeigen: Die (gemeinsame) Verteilung von $(X_n)_{n\in E}$ ist gleich:

$$\mathcal{L}_P(X_n\colon n \in E) = \prod_{n\in E} \frac{1}{2}(\delta_0 + \delta_1)$$

also gleich der Gleichverteilung auf $\{0,1\}^E$. Es genügt, die für den Spezialfall $E = \{1, \dots, n\}$ zu zeigen. Nun sei $X = (X_1, \dots, X_n) \in \{0,1\}^n$ und

$$a = \sum_{k=1}^{n} 2^{-k} x_k = (0.x_1 x_2 \dots x_n)_2 \in [0,1)$$

Dann gilt $\{X_k = x_k, k = 1, \dots, n\} = [a, a+2^n)$. Also ist $P[X_k = x_k, k = 1, \dots, n] = P([a, a+2^{-n})) = 2^{-n} = |\{0,1\}^n|^{-1}$. Die Verteilung $\mathcal{L}_P(X_1, \dots, X_n)$ hat also die Zähldichte $(|\{0,1\}^n|^{-1})_{x\in\{0,1\}^n}$, ist also die Gleichverteilung auf $\{0,1\}^n$. □

Lemma. *Es seien $(X_n)_{n\in\mathbb{N}}$ i.i.d. $\frac{1}{2}(\delta_0+\delta_1)$-verteilte Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann ist $Z := \sum_{n\in\mathbb{N}} 2^{-n} X_n\colon \Omega \to [0,1]$ uniform auf $[0,1]$ verteilt.*

Beweis. Wir betrachten das Mengensystem

$$\mathfrak{I} = \{[a, a+2^{-n})\colon n \in \mathbb{N}, (x_1, \dots, x_n) \in \{0,1\}^n, a = \sum_{k=1}^{n} 2^{-k} x_k\} \cup \{\emptyset\}$$

Ein Intervall $I \in \mathfrak{I}$, $I \neq \emptyset$, besteht also aus allen Zahlen $\omega \in [0,1)$, die ein vorgegebenes Anfangsstück $(0, x_1 \dots x_n)_2$ in ihrer Binärdarstellung besitzen. $\mathfrak{I}$ ist ein Erzeugendensystem von $\mathcal{B}([0,1])$. Zudem ist $\mathfrak{I}$ $\cap$-stabil. Um zu zeigen, dass $\mathcal{L}_P(Z) = \mathrm{unif}[0,1]$ ist, reicht es, dass gilt

$$P[Z \in I] = \mathrm{unif}[0,1](I), \quad I \in \mathfrak{I}$$

Das ist klar für $I = \emptyset$. Für $I = [a, a+2^{-n})$ mit $a = (0, x_1 \dots x_n)_2$ gilt: Die Ereignisse $A = \{X_k = x_k, k = 1, \dots, n\}$ und $B = \{Z \in I\}$ unterscheiden sich nur um eine Nullmenge, denn (für $a = 0$ ist $B \setminus A = \emptyset$)

$$A \setminus B = \{X_k = x_k, k = 1, \dots, n\} \cap \{X_k = 1, k > n\}$$
$$B \setminus A = \{X_k = y_k, k = 1, \dots, n\} \cap \{X_k = 1, k > n\}, \text{ wobei } (0, y_1 \dots y_n)_2 = a - 2^{-n}$$

Es folgt

$$A \triangle B \subseteq \{X_k = 1, k > n\}$$

und damit

$$P(A \triangle B) \leq P[X_k = 1, k > n] = \lim_{m\to\infty} P[X_k = 1, n < k \leq m] \underset{X_k \text{ i.i.d.}}{=} \lim_{m\to\infty} \prod_{k=n+1}^{m} P[X_k = 1] =$$

$$= \lim_{m\to\infty} \frac{1}{2^m} = 0$$

Daher ist $P[Z \in I] = P(B) = P(A) = \frac{1}{2^n} = \text{unif}[0,1](I)$. Damit ist $\mathcal{L}_P(Z) = \text{unif}[0,1]$. □

Lemma. *Sei $\iota\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ eine Injektion und $(X_n)_{n\in\mathbb{N}}$ eine i.i.d Folge von $\frac{1}{2}(\delta_0 + \delta_1)$-verteilten Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann ist die Folge $(Z_n)_{n\in\mathbb{N}}$ von Zufallsvariablen,*

$$Z_m := \sum_{k\in\mathbb{N}} 2^{-k} X_{\iota(k,m)}\colon \Omega \to [0,1]$$

i.i.d. mit der Verteilung unif$[0,1]$.

Beweis. Für jedes $m \in \mathbb{N}$ ist die Folge $(X_{\iota(k,m)})_{k\in\mathbb{N}}$ i.i.d $\frac{1}{2}(\delta_0+\delta_1)$-verteilt. Außerdem ist Z_m messbar bezüglich der σ-Algebra $\mathcal{F}_m = \sigma(X_{i(k,m)}\colon k \in \mathbb{N})$. Nun ist die Familie $(\mathcal{F}_m)_{m\in\mathbb{N}}$ unabhängig, da ι injektiv ist und da $(X_n)_{n\in\mathbb{N}}$ i.i.d. ist. Also sind auch die $(Z_m)_{m\in\mathbb{N}}$ unabhängig, da $\sigma(Z_m) \subseteq \mathcal{F}_m$. □

Lemma. *Es sei $(Z_m)_{m\in\mathbb{N}}$ eine i.i.d.* unif$[0,1]$*-verteilte Zufallsvariable auf $(\Omega, \mathcal{A}, P)$. Weiter sei $(P_n)_{n\in\mathbb{N}}$ eine Folge von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Es seien $q_m\colon (0,1) \to \mathbb{R}$, $m \in \mathbb{N}$, Quantilsfunktionen der P_n. Wir setzen $\tilde{q}_m\colon \mathbb{R} \to \mathbb{R}$, $\tilde{q}_m(x) = q_m(x)$ für $x \in (0,1)$ und $\tilde{q}_m(x)$ beliebig messbar sonst. Damit ist $(\tilde{q}_m(Z_m))_{m\in\mathbb{N}}$ eine Folge unabhängiger Zufallsvariablen mit den Verteilungen $\mathcal{L}_P(\tilde{q}_m(Z_m)) = P_m$, $m \in \mathbb{N}$.*

Bemerkung. Zu jeder Folge $(P_m)_{m\in\mathbb{N}}$ von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gibt es also eine Folge $(X_m)_{m\in\mathbb{N}}$ unabhängiger Zufallsvariablen auf $(\Omega, \mathcal{A}, P) = ([0,1), \mathcal{B}([0,1)), \text{unif}[0,1))$ mit diesen Verteilungen: $\mathcal{L}_P(X_m) = P_m$.

Bemerkung. Unser Interesse verschiebt sich damit von den Ergebnisräumen Ω zu Zufallsvariablen und ihren gemeinsamen Verteilungen. Wenn nötig, können wir für fast alle Anwendungen auf $(\Omega, \mathcal{A}, P) = ([0,1), \mathcal{B}([0,1)), \text{unif}[0,1))$ arbeiten

Folgendes Kriterium zum Nachweis der Unabhängigkeit diskreter Zufallsvariablen ist nützlich:

Lemma. *Seien $X_1, \ldots, X_n$ Zufallsvariablen über $(\Omega, \mathcal{A}, P)$ mit Werten in der abzählbaren Menge $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Dann sind genau dann $X_1, \ldots, X_n$ unabhängig, wenn für alle $k_1, \ldots, k_n \in \mathbb{N}$ gilt:*

$$P[X_i = k_i, i = 1, \ldots, n] = \prod_{i=1}^{n} P[X_i = k_i]$$

Beweis. $\xi_i = \{\{X_i = k\}\colon k \in \mathbb{N}\} \cup \{\emptyset\}$ ist ein $\cap$-stabiler Erzeuger von $\sigma(X_i)$, $i = 1, \ldots, n$. Nach Voraussetzung sind die ξ_i, $i = 1, \ldots, n$, unabhängig. Also sind auch $\sigma(\xi_i) = \sigma(X_i)$, $i = 1, \ldots, n$, unabhängig. Die andere Richtung ist trivial. □

1.13 Beispiele und Standardverteilungen

1.13.1 Die geometrische Verteilung

Eine (möglicherweise unfaire) Münze wird bis zum 1. Auftreten von "1" geworfen. Wir bestimmen die Verteilung der Anzahl der Würfe.

Modell Sei $0 < p < 1$, und $(X_t)_{t\in\mathbb{N}}$ i.i.d. $p\delta_1 + (1-p)\delta_0$-verteilte Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Wir setzen

$$T = \inf\{t \in \mathbb{N} \colon X_t = 1\}, \text{ wobei } \inf\emptyset := +\infty$$

Für $s \in \mathbb{N}$ gilt:

$$P[T = s] = P[X_t \neq 1 \text{ für } t < s, X_s = 1] = \Big(\prod_{t=1}^{s-1} P[X_t \neq 1]\Big) P[X_s = 1] = (1-p)^{s-1}p$$

$$P[T = \infty] = P[X_t \neq 1 \text{ für alle } t \in \mathbb{N}] = \lim_{s\to\infty} P[X_t \neq 1 \text{ für } t < s] = \lim_{s\to\infty} \prod_{t=1}^{s} P[X_t \neq 1] =$$

$$= \lim_{s\to\infty} (1-p)^s = 0$$

Es folgt

$$\mathcal{L}_P(T) = \sum_{s\in\mathbb{N}} (1-p)^{s-1} p\ \delta_s$$

Diese Verteilung (oder auch die Verteilung $\mathcal{L}_P(T-1) = \sum_{s\in\mathbb{N}_0}(1-p)^s p\ \delta_s$) heißt *geometrische Verteilung* zum Parameter p. Sie tritt typischerweise als Wartezeit bis zum ersten Auftreten eines Ereignisses in diskreter Zeit auf.

1.13.2 Die negative Binomialverteilung

Im Modell von eben sei T_n, $n \in \mathbb{N}$, die Anzahl der Würfe bis zur n-ten "1". Formal sei

$$T_0 = 0$$
$$T_n = \inf\{t > T_{n-1} \colon X_t = 1\}, \text{ für } n > 0$$

Insbesondere ist T_1 geometrisch verteilt. P-fast sicher sind alle T_n endlich, denn

$$P[\exists n \in \mathbb{N}.\ T_n = \infty] = P[X_t \neq 1 \text{ schließlich für } t \to \infty] = P[\exists s \in \mathbb{N} \underbrace{\forall t \geq s.\ X_t \neq 1}_{\text{monoton steigend in } s}]$$

$$= \lim_{s\to\infty} P[\forall t \geq s.\ X_t \neq 1] = 0$$

Insbesondere ist $T_n - T_{n-1}$ (Wartezeit zwischen n-ter und $(n-1)$-ter "1") P-fast sicher wohldefiniert. Wir zeigen, dass $T_n - T_{n-1}$, $n \in \mathbb{N}$, i.i.d. geometrisch verteilt sind.

Beweis. Seien $s_1, \ldots, s_n \in \mathbb{N}$, $t_k = \sum_{i=1}^k s_i$ für $k = 1, \ldots, n$. Dann gilt für $A = P[T_k - T_{k-1} = s_k \text{ für } k = 1, \ldots, n]$:

$$\begin{aligned} A &= P[T_k = t_k \text{ für } k = 1, \ldots, n] = \\ &= P[X_{t_k} = 1 \text{ für } k = 1, \ldots, n, X_t \neq 1 \text{ für alle } t \in \{1, \ldots, t_n\} \setminus \{t_1, \ldots, t_n\}] = \\ &= p^n (1-p)^{t_n - n} = \prod_{k=1}^{n} \Big[(1-p)^{s_k - 1} p\Big] = \prod_{k=1}^{n} P[T = s_k] \end{aligned}$$

□

Anschaulich interpretiert: "Gedächtnislosigkeit" des Münzwurfs. Die Verteilung der Wartezeit auf die "nächste 1" ist immer die gleiche, gleichgültig, welche Wartezeiten vorher auftraten. Wegen $T_k = \sum_{k=1}^n (T_k - T_{k-1})$, $n \in \mathbb{N}$, bedeutet das: T_n ist eine Summe von n i.i.d. geom(p)-verteilten Zufallsvariablen und damit

$$\mathcal{L}_P(T_n) = \text{geom}(p)^{*n}$$

Diese Verteilung heißt *negative Binomialverteilung* mit den Parametern n und p. Wir berechnen nun die Zähldichte von $\mathcal{L}_P(T_n)$. Sei hierzu $t \in \mathbb{N}$:

$$\begin{aligned} P[T_n = t] &= p[X_t = 1, |\{s < t \colon X_s = 1\}| = n-1] = \\ &= \sum_{\substack{E \subseteq \{1,\dots,t-1\} \\ |E| = n-1}} P[X_t = 1, \forall s \in E.\ X_s = 1, X_s \neq 1 \text{ für alle } s \in \{1,\dots,t-1\} \setminus E] = \\ &= \binom{t-1}{n-1} p^n (1-p)^{t-n} \end{aligned}$$

Die negative Binomialverteilung zu den Parametern n und p ist also gleich:

$$\mathcal{L}_P(T_n) = \sum_{t=n}^{\infty} \binom{t-1}{n-1} p^n (1-p)^{t-n} \delta_t$$

1.13.3 Seltene Ereignisse: Die Poissonverteilung

Seltene Ereignisse treten auf bei häufiger Wiederholung eines Experiments mit kleiner Erfolgswahrscheinlichkeit.

Beispiel. Die Anzahl der Haftpflichtschäden in einem Monat: Es gibt viele voneinander unabhängige Versicherte, aber die Wahrscheinlichkeit, dass ein bestimmer Versicherter einen Unfall hat, ist sehr klein.

Beispiel. Die Anzahl an Regentropen pro Sekunde, die auf einen Regenschirm fallen: Es gibt viele Regentropfen, die voneinander unabhängig sind, aber die Wahrscheinlichkeit, dass ein fester Regentropen den Regenschirm trifft, ist sehr klein.

Beispiel. Anzahl der radioaktiven Zerfälle in einer Probe Uran pro Sekunde: Es gibt viele Urankerne, aber die Wahrscheinlichkeit, dass ein bestimmer Kern in einer Sekunde zerfällt, ist extrem gering.

Modell Binomialverteilung mit Parametern n und p im Limes $n \to \infty$ und $p \to 0$ aber so, dass $np \to \lambda \in (0, \infty)$.

Satz. *Sei $(p_n)_{n\in\mathbb{N}}$ eine Folge in $(0,1)$ mit $np_n \xrightarrow{n\to\infty} \lambda \in (0,\infty)$. Dann gilt für alle $k \in \mathbb{N}_0$:*

$$\text{binomial}(n, p_n)(\{k\}) \xrightarrow{n\to\infty} e^{-\lambda} \frac{\lambda^k}{k!} = \text{Poisson}(\lambda)(\{k\})$$

Zur Erinnereung: Die Poisson-Verteilung mit Parameter λ ist die Verteilung

$$\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \delta_k, \quad \textit{für } \lambda > 0$$

Beweis. Es gilt $\log(1-p) = -p(1+o(1))$ fur $p \to 0$, also:

$$\begin{aligned} \text{binomial}(n, p_n)(\{k\}) = \binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{1}{k!} \underbrace{\left(\prod_{l=0}^{k-1} \frac{n-l}{n} \right)}_{\to 1} \underbrace{(np_n)^k}_{\to \lambda^k} \exp\left(\underbrace{\frac{n-k}{n}}_{\to 1} \underbrace{n \log(1-p_n)}_{\substack{= -np_n(1+o(1)) \\ \to -\lambda}} \right) \to \\ &\xrightarrow{n\to\infty} \frac{1}{k!} \lambda^k e^{-\lambda} \end{aligned}$$

□

Korollar. *Für alle $A \subseteq \mathbb{N}_0$ gilt unter Voraussetzungen wie eben:*

$$\text{binomial}(n, p_n)(A) \xrightarrow{n\to\infty} \text{Poisson}(\lambda)(A)$$

Beweis. Wir verwenden das Lemma von Fatou aus der Maßtheorie: Ist X_n, $n \in \mathbb{N}$, eine Folge messbarer Funktionen über einen Maßraum $(\Omega, \mathcal{A}, \mu)$ mit $X_n \geq 0$, $n \in \mathbb{N}$, so gilt für alle $A \in \mathcal{A}$:

$$\int_A \liminf_{n\to\infty} X_n \, \mathrm{d}\mu \leq \liminf_{n\to\infty} \int_A X_n \, \mathrm{d}\mu$$

Wir wenden dieses Lemma auf $\Omega = \mathbb{N}_0$, μ das Zählmaß und $X_n(k) = \binom{n}{k} p_n^{k-1}(1-p_n)^{n-k}$. Wir verwenden also die "Reihenversion" des Lemmas von Fatou. Für jedes $A \subseteq \mathbb{N}_0$ gilt:

$$\begin{aligned}\text{Poisson}(\lambda)(A) = \sum_{k\in A} e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k\in A} \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} \leq \\ \leq \liminf_{n\to\infty} \sum_{k\in A} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \liminf_{n\to\infty} \text{binomial}(n, p_n)(A)\end{aligned}$$

und ebenso

$$\text{Poisson}(\lambda)(A^c) \leq \liminf_{n\to\infty} \text{binomial}(n, p_n)(A^c)$$

Damit folgt

$$\text{Poisson}(\lambda)(A) = 1 - \text{Poisson}(\lambda)(A^c) \geq 1 - \liminf_{n\to\infty} \text{binomial}(n, p_n)(A^c) = \limsup_{n\to\infty} \text{binomial}(n, p_n)(A)$$

Zusammen gilt also:

$$\limsup_{n\to\infty} \text{binomial}(n, p_n)(A) \leq \text{Poisson}(\lambda)(A) \leq \liminf_{n\to\infty} \text{binomial}(n, p_n)(A)$$

und es folgt $\lim_{n\to\infty} \text{binomial}(n, p_n)(A) = \text{Poisson}(\lambda)(A)$. □

1.13.4 Ordnungsstatistiken und Betaverteilungen

Definition. Für $x_1, \ldots, x_n \in \mathbb{R}$ sei $x_{[1]}, \ldots, x_{[n]}$ diejenige Permutation von $x_1, \ldots, x_n$, die diese Zahlen der Größe nach anordnet; d.h. $x_{[1]} \leq \cdots \leq x_{[n]}$. $x_{[1]}, \ldots, x_{[n]}$ heißt *Ordnungsstatistik* von $x_1, \ldots, x_n$.

Definition. Die *Betaverteilung* mit den Parametern $s > 0$ und $t > 0$ ist die Verteilung mit der Dichte

$$\beta_{s,t}(a) = 1_{(0,1)}(a)\frac{1}{B(s,t)} a^{s-1}(1-a)^{t-1}$$

mit der Betafunktion

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$$

Satz. *Es seien $U_1, \ldots, U_n$ i.i.d. $\text{unif}(0,1)$-verteilte Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und $U_{[1]}, \ldots, U_{[n]}$ ihre Ordnungsstatistik. Dann ist $U_{[k]}$ $B(k, n-k+1)$-verteilt für alle $k = 1, \ldots, n$.*

Beweis. Wir beweisen den Satz durch Induktion über k "rückwarts". Sei $a \in [0,1]$. Dann gilt:

$$P[U_{[n]} \leq a] = P[\forall i = 1, \ldots, n \colon U_i \leq a] = \prod_{i=1}^{n} P[U_i \leq a] = a^n$$

Es gilt:

$$\frac{1}{B(k, n-k+1)} = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} = \frac{n!}{(k-1)!(n-k)!} = k\binom{n}{k}$$

also

$$P[U_{[n]} \leq a] = \int_0^a n x^{n-1} \,\mathrm{d}x = \int_0^a n\binom{n}{n} x^{n-1}(1-x)^{n-n} \,\mathrm{d}x = B(n,1)([0,a])$$

Zum Induktionsschritt, sei also $k \in \{1, \ldots, n-1\}$. Für $a \in [0,1]$ gilt $\{U_{[k]} \leq a\} = \{U_{[k+1]} \leq a\} \sqcup \{U_{[k]} \leq a < U_{[k+1]}\}$, also

$$P[U_{[k]} \leq a] = P[U_{[k+1]} \leq a] + P[U_{[k]} \leq a < U_{[k+1]}] = \int_0^a \beta_{k+1,n-k} \,\mathrm{d}x + P[U_{[k]} \leq a < U_{[k+1]}]$$

Weiterhin gilt

$$\begin{aligned}
P[U_{[k]} \leq a < U_{[k+1]}] &= P[|\{i = 1, \ldots, n \colon U_i \leq a\}| = k] = \\
&= \sum_{\substack{E \subseteq \{1,\ldots,n\} \\ |E|=k}} P[U_i \leq a \text{ für } i \in E, U_i > a \text{ für } i \notin E, i = 1, \ldots, n] = \\
&= \binom{n}{k} a^k (1-a)^{n-k} = \int_0^a \frac{d}{dx}\left[\binom{n}{k} x^k (1-x)^{n-k}\right] \mathrm{d}x = \\
&= \int_0^a k\binom{n}{k} x^{k-1}(1-x)^{n-k} \,\mathrm{d}x - \int_0^a \underbrace{(n-k)\binom{n}{k}}_{(k+1)\binom{n}{k+1}} x^k (1-x)^{n-k-1} \,\mathrm{d}x = \\
&= \int_0^a \beta_{k,n-k+1} \,\mathrm{d}x - \int_0^a \beta_{k+1,n-k} \,\mathrm{d}x
\end{aligned}$$

Zusammen gilt also:

$$P[U_{[k]} \leq a] = \int_0^a \beta_{k,n-k+1} \,\mathrm{d}x = B(k, n-k+1)([0,a]) \qquad \square$$

1.14 Erwartungswert und Varianz

Definition (Integrale von Funktionen mit beliebigem Vorzeichen)**.** Sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum und $X \colon (\Omega, \mathcal{A}) \to (\mathbb{R} \cup \{\pm\infty\}, \mathcal{B}(\mathbb{R} \cup \{\pm\infty\}))$ messbar. Der Positivteil von X wird durch $X_+ := \max(X, 0)$, der Negativteil von X durch $X_- := \max(-X, 0)$ definiert. Insbesondere gilt $X = X_+ - X_-$. Im Fall, dass $\int_\Omega X_+ \,\mathrm{d}\mu$ oder $\int_\Omega X_- \,\mathrm{d}\mu$ endlich ist, definieren wir

$$\int_\Omega X \,\mathrm{d}\mu = \int_\Omega X_+ \,\mathrm{d}\mu - \int_\Omega X_- \,\mathrm{d}\mu$$

Wir setzen

$$\mathcal{L}^1(\Omega, \mathcal{A}, \mu) := \{X \colon (\Omega, A) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ messbar} \colon \int_\Omega X_+ \,\mathrm{d}\mu < \infty \text{ und } \int_\Omega X_- \,\mathrm{d}\mu < \infty\}$$

Die Integralabbildung

$$\int : \mathcal{L}^1(\Omega, \mathcal{A}, \mu) \to \mathbb{R}, X \mapsto \int_\Omega X \, \mathrm{d}\mu$$

ist wohldefiniert und linear. Die Elemente des Vektorraums $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ heißen *integrierbare Funktionen* bezüglich μ.

Im Fall, dass $\mu = P$ ein Wahrscheinlichkeitsmaß ist, führen wir folgende Sprechweise ein:

Definition. Es sei $X \colon (\Omega, \mathcal{A}, P) \to (\mathbb{R} \cup \{\pm\infty\}, \mathcal{B}(\mathbb{R} \cup \{\pm\infty\}))$ eine Zufallsvariable. Falls $\int_\Omega X \, \mathrm{d}P$ existiert, sagen wir X *besitzt einen Erwartungswert* bezüglich P und nennen

$$E_P[X] := \int_\Omega X \, \mathrm{d}P$$

den *Erwartungswert von* X bezüglich P. Kurznotation $E[X] = E_P[X]$, wenn klar ist, welches P gemeint ist.

Ist $A \in \mathcal{A}$ ein Ereignis, so nennen wir

$$E_p[X, A] = E[X, A] := E_P[X \cdot 1_A] = \int_\Omega X \cdot 1_A \, \mathrm{d}P = \int_A X \, \mathrm{d}P$$

den *Erwartungswert von* X *auf dem Ereignis* A.

Bemerkung. Der Erwartungswert ist linear, d.h. für alle $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ gilt

$$E_P[X + Y] = E_P[X] + E_P[Y]$$

und für alle $\alpha \in \mathbb{R}$ und $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ gilt

$$E_P[\alpha X] = \alpha E_P[X]$$

Bemerkung. Der Erwartungswert ist monoton, d.h. sind X und Y Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit existierender Erwartung und ist $X \leq Y$, so ist auch $E_P[X] \leq E_P[Y]$.

Bemerkung. Für alle $A \in \mathcal{A}$ gilt $E_P[1_A] = P(A)$, insbesondere ist $E_P[1] = 1$.

Bemerkung. Die Kombination der letzen Bemerkung mir der Monotonie ist ein wichtiger Trick Wahrscheinlichkeiten nach oben abzuschätzen.

Beispiel. Ist Ω abzählbar, $\mathcal{A} = \mathcal{P}(\Omega)$ und besitzt P die Zähldichte $(p_\omega)_{\omega \in \Omega}$, so gilt für $X \colon \Omega \to \mathbb{R}$:

$$X \in \mathcal{L}^{-1}(\Omega, \mathcal{A}, P) \iff \sum_{\omega \in \Omega} |X(\omega)| p_\omega < \infty$$

und falls X einen Erwartungswert besitzt, gilt

$$E_P[X] = \sum_{\omega \in \Omega} X(\omega) p_\omega$$

Beispiel. Mit $(\Omega, \mathcal{A}, P) = (\{1, \ldots, 6\}, \mathcal{P}(\Omega), \mathrm{unif})$ und $X = \mathrm{id}_\Omega$ gilt

$$E_P[X] = \frac{1}{6} \cdot 1 + \cdots + \frac{1}{6} \cdot 6 = 3.5$$

Das Beispiel zeigt, dass $E_P[X]$ nicht notwendigerweise ein möglicher Wert von X zu sein braucht.

Der folgende Satz zeigt unter anderem, dass $E_P[X]$ nur von der Verteilung $\mathcal{L}_P(X)$ von X abhängt.

Satz (Integration bezüglich des Bildmaßes)**.** *Sei* $(\Omega, \mathcal{A}, \mu)$ *ein Maßraum,* $X\colon (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ *messbar und* $f\colon (\Omega', \mathcal{A}') \to (\mathbb{R} \cup \{\pm\infty\}, \mathcal{B}(\mathbb{R} \cup \{\pm\infty\}))$ *messbar. Dann ist*

$$\int_\Omega f \circ X \,\mathrm{d}\mu \text{ existiert} \iff \int_{\Omega'} f \,\mathrm{d}(X[\mu]) \text{ existiert}$$

In diesem Fall gilt

$$\int_\Omega f \circ X \,\mathrm{d}\mu = \int_{\Omega'} f \,\mathrm{d}(X[\mu])$$

Für reelle Zufallsvariablen $X\colon (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *bedeutet das*

$$E_P[X] = \int_{\mathbb{R}} x \; L_P(X)(\mathrm{d}x)$$

falls eine der beiden Seiten existiert. Etwas allgemeiner: Besitzt $X\colon (\Omega, \mathcal{A}, P) \to (\Omega', \mathcal{A}')$ *die Verteilung* $Q = \mathcal{L}_P(X)$*, so gilt*

$$E_P[f(X)] = \int_{\Omega'} f \,\mathrm{d}Q$$

Beweisidee. Maßtheoretische Induktion: Zuerst betrachtet man den Fall $f = 1_A$, anschließend den Fall, dass f eine nichtnegative Linearkombination von Indikatorfunktionen ist. Dann betrachtet man den Fall eine messbaren Abbildung $f \geq 0$, indem man eine Approximation von f von unten durch Treppenfunktionen betrachtet. Zuletzt betrachtet man für allgemeine f die Zerlegung $f = f_+ + f_-$.

Bemerkung. Besitzt ein Maß μ auf $(\Omega, \mathcal{A})$ eine Dichte g bezüglich eines weiteren Maßes ν, so gilt für alle messbaren $f\colon (\Omega, \mathcal{A}) \to (\mathbb{R} \cup \{\pm\infty\}, \mathcal{B}(\mathbb{R} \cup \{\pm\infty\}))$

$$\int_\Omega f \,\mathrm{d}\mu = \int_\Omega fg \,\mathrm{d}\nu$$

Symbolische Notation: $\mathrm{d}\mu = g \,\mathrm{d}\nu$ bedeutet μ besitzt eine Dichte g bezüglich ν. Im Fall, dass eine Zufallvariable X eine Dichte $g\colon \mathbb{R} \to [0, \infty]$ bezüglich λ besitzt, bedeutet das

$$E_P[f(X)] = \int_{\mathbb{R}} f(x)g(x) \,\mathrm{d}x$$

Beispiel. Ist X normalverteilt mit Parametern μ und σ^2, so gilt

$$\begin{aligned} E[X] &= \int_{\mathbb{R}} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} (t-\mu) \exp\left(-\frac{t^2}{2\sigma^2}\right) \mathrm{d}t = \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} t \exp\left(-\frac{t^2}{2\sigma^2}\right) \mathrm{d}t}_{=0} + \frac{\mu}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{t^2}{2\sigma^2}\right) \mathrm{d}t = \mu \end{aligned}$$

Der Parameter μ ist also gleich dem Erwartungswert einer $N(\mu, \sigma^2)$-verteilten Zufallsvariablen.

Beispiel. Nimmt die Zufallsvariable $X\colon (\Omega, \mathcal{A}, P) \to \mathbb{R}$ nur endlich viele Werte $x_1, \ldots, x_n$ an, so gilt

$$Q = \mathcal{L}_P(X) = \sum_{k=1}^{n} P[X = x_k] \delta_{x_k}$$

also ist

$$E_P[f(X)] = \int_{\mathbb{R}} f \,\mathrm{d}Q = \sum_{k=1}^{n} f(x_k) P[X = x_k]$$

Analoges gilt, wenn X abzählbar unendlich viele Werte $x_1, \dots$ annimmt:

$$E_P[f(X)] = \sum_{k \in \mathbb{N}} f(x_k) P[X = x_k] \text{ falls } f \geq 0 \text{ oder die Reihe absolute summierbar ist}$$

Definition. Sei $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$. Wir definieren die *Varianz* von X bezüglich P durch:

$$\mathrm{Var}_P(X) = \mathrm{Var}(X) = E_P[(X - E_P[X])^2]$$

Die Quadratwurzel der Varianz heißt *Standardabweichung* von X bezüglich P:

$$\sigma_P(X) = \sigma(X) = \sqrt{\mathrm{Var}_P(X)}$$

Wir definieren

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) = \{X \in \mathcal{L}^1(\Omega, \mathcal{A}, P) \colon E_P[X^2] < \infty\} = \{X \in \mathcal{L}^1(\Omega, \mathcal{A}, P) \colon \mathrm{Var}_P(X) < \infty\}$$

Die letzte Gleichheit folgt aus folgender Formel:

Satz. *Für alle $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ gilt*

$$\mathrm{Var}_P(X) = E_P[X^2] - E_P[X]^2$$

Beweis. Es gilt

$$\begin{aligned} \mathrm{Var}_P(X) &= E_P[(X - E_P[X])^2] = E_P[X^2 - 2X E_P[X] + E_P[X]^2] = \\ &= E_P[X^2] - 2E_P[X]E_P[X] + E_P[X]^2 = E_P[X^2] - E_P[X]^2 \end{aligned}$$ □

Korollar. *Für $x \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ gilt stets*

$$E_P[X^2] \geq E_P[X]^2$$

da $\mathrm{Var}_P(X) \geq 0$.

Beispiel. Die Varianz einer $p\delta_1 + (1-p)\delta_0$-verteilten Zufallsvariable X mit Werten in $\{0, 1\}$ beträgt:

$$\mathrm{Var}_P[X] = E_P[X^2] - E_P[X]^2 = E_P[X] - E_P[X]^2 = p - p^2 = p(1-p)$$

Beispiel. Wir berechnen die Varianz der Gleichverteilung unif$[0, 1]$: Sei X eine unif$[0, 1]$-verteilte Zufallsvariable.

$$\begin{aligned} E_P[X] &= \int_{\mathbb{R}} t \cdot 1_{[0,1]}(t) \,\mathrm{d}t = \int_0^1 t \,\mathrm{d}t = \frac{1}{2} \\ E_P[X^2] &= \int_{\mathbb{R}} t^2 \cdot 1_{[0,1]}(t) \,\mathrm{d}t = \int_0^1 t^2 \,\mathrm{d}t = \frac{1}{3} \end{aligned}$$

und daher

$$\begin{aligned} \mathrm{Var}_P(X) &= E_P[X^2] - E_P[X]^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12} \\ \sigma_P(X) &= \sqrt{\mathrm{Var}_P(X)} = \sqrt{\frac{1}{12}} \end{aligned}$$

Beispiel. Sei X normalverteilt: $\mathcal{L}_P(X) = N(\mu, \sigma^2)$. Dann gilt:

$$\begin{aligned}\operatorname{Var}_P(X) &= \int_{\mathbb{R}} (t-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \mathrm{d}t = \frac{1}{\sqrt{2\pi\sigma^2}} \sigma^3 \int_{\mathbb{R}} z^2 e^{-\frac{z^2}{2}} \,\mathrm{d}z = \\ &= \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} z \frac{d}{dz}\left(-e^{-\frac{z^2}{2}}\right) \mathrm{d}z = \frac{\sigma^2}{\sqrt{2\pi}} \left[-ze^{-\frac{z^2}{2}}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \,\mathrm{d}z\right] = \sigma^2\end{aligned}$$

Es ist also μ der Erwartungswert, σ^2 die Varianz und σ die Standardabweichung der Normalverteilung $N(\mu, \sigma^2)$.

Bemerkung. Für $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ und $a \in \mathbb{R}$ gilt:

1) $\operatorname{Var}_P(aX) = a^2 \operatorname{Var}_P(X)$ (Skalierungseigenschaft)
2) $\operatorname{Var}_P(X + a) = \operatorname{Var}_P(X)$ (Verschiebungseigenschaft)

Beweis.

1) $\operatorname{Var}_P(aX) = E_P[(aX - E_P[aX])^2] = E_P[(aX - aE_P[X])^2] = a^2 E_P[(X - E_P[X])^2] = a^2 \operatorname{Var}_P(X)$.
2) $\operatorname{Var}_P(X + a) = E_P[(X + a - E_P[X + a])^2] = E_P[(X - E_P[X] + a - a)^2] = E_P[(X - E_P[X])^2] = \operatorname{Var}_P(X)$. □

Bemerkung. Es folgt $\sigma_P(aX) = |a|\sigma_P(X)$.

Die Varianz ist eine quadratische Form. Die zugehörige symmetrische Bilinearform heißt *Covarianz*:

Definition. Seien $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$. Wir definieren die *Covarianz* von X und Y bezüglich P durch

$$\operatorname{Cov}_P(X, Y) = E_P[(X - E_P[X])(Y - E_P[Y])]$$

Insbesondere gilt $\operatorname{Cov}_P(X, X) = \operatorname{Var}_P(X)$.

Bemerkung. Analog zur Varianz gilt $\operatorname{Cov}_P(X, Y) = E_P[XY] - E_P[X]E_P[Y]$.

Lemma. *Es gilt für alle $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$*

$$|\operatorname{Cov}_P(X, Y)| \leq \sigma_P(X)\sigma_P(Y)$$

Das ist ein Spezialfall der allgemeine Cauchy-Schwarz-Ungleichung für positiv semidefinite quadratische Formen.

Lemma. *Für alle $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, \mu)$ gilt:*

$$\left(\int_\Omega XY \,\mathrm{d}\mu\right) \leq \left(\int_\Omega X^2 \,\mathrm{d}\mu\right)^{\frac{1}{2}} \left(\int_\Omega Y^2 \,\mathrm{d}\mu\right)^{\frac{1}{2}}$$

Anders geschrieben für Wahrscheinlichkeitsmaße

$$E_P[XY] \leq E_P[X^2]^{\frac{1}{2}} E_P[Y^2]^{\frac{1}{2}}$$

Setzt man hier $X - E_P[X]$ statt X und $Y - E_P[Y]$ statt Y, so folgt das vorige Lemma.

Beweis. Wir betrachten die quadratische Form $q\colon \mathbb{R}^2 \to [0, \infty)$, definiert durch

$$q(\alpha, \beta) = \int_\Omega (\alpha X + \beta Y)^2 \,\mathrm{d}\mu = \alpha^2 \int_\Omega X^2 \,\mathrm{d}\mu + 2\alpha\beta \int_\Omega XY \,\mathrm{d}\mu + \beta^2 \int_\Omega Y^2 \,\mathrm{d}\mu$$

Man beachte, dass $q(\alpha, \beta) < \infty$, denn es gilt $(\alpha X + \beta Y)^2 \leq 2\alpha^2 X^2 + 2\beta^2 Y^2$ und

$$q(\alpha, \beta) \leq 2\alpha^2 \int X^2 \, \mathrm{d}\mu + 2\beta^2 \int_\Omega Y^2 \, \mathrm{d}\mu < \infty$$

Setzen wir speziell

$$\beta = \sqrt{\int_\Omega X^2 \, \mathrm{d}\mu}, \quad \alpha = \pm\sqrt{\int_\Omega Y^2 \, \mathrm{d}\mu}$$

so folgt

$$0 \leq q(\alpha, \beta) = 2 \int_\Omega X^2 \, \mathrm{d}\mu \int_\Omega Y^2 \, \mathrm{d}\mu \pm 2\sqrt{\int_\Omega X^2 \, \mathrm{d}\mu}\sqrt{\int_\Omega Y^2 \, \mathrm{d}\mu} \int_\Omega XY \, \mathrm{d}\mu$$

Im Fall $\alpha\beta \neq 0$ folgt die Behauptung. Im Fall $\alpha = 0$ folgt $Y^2 = 0$ μ-fast sicher, also $Y = 0$ μ-fast sicher, also $XY = 0$ μ-fast-sicher und daher $\int_\Omega XY \, \mathrm{d}\mu = 0$. Der Fall $\beta = 0$ ist analog. □

Bemerkung. Gleichheit in der Cauchy-Schwarz-Ungleichung gilt genau dann, wenn X und Y bis auf einer Nullmenge linear abhängig sind.

Definition. Seien $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ und $\sigma_P(X)\sigma_P(Y) \neq 0$, d.h. X und Y sind nicht P-fast sicher konstant). Wir definieren den *Korrelationskoeffizienten*

$$r_P(X, Y) = \mathrm{Cor}_P(X, Y) = \frac{\mathrm{Cov}_P(X, Y)}{\sigma_P(X)\sigma_P(Y)}$$

Es gilt also $-1 \leq r_P(X, Y) \leq 1$ mit $r_P(X, Y) = 1$ für $X - E_P[X] = \beta(Y - E_P[Y])$ P-fast sicher mit $\beta > 0$ und $r_P(X, Y) = -1$ für $\beta < 0$.

Lemma. *Für* unabhängige $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ *gilt* $\mathrm{Cor}_P(X, Y) = 0$*, also* $r_p(X, Y) = 0$*, falls* X *und* Y *nicht* P*-fast sicher konstant sind. Anders gesagt* $E_P[XY] = E_P[X]E_P[Y]$ *für unabhängige* X, Y*.*

Satz (Satz von Fubini für integrierbare Funktionen)**.** *Seien* $(\Omega, \mathcal{A}, \mu)$ *und* $(\Sigma, \mathcal{B}, \nu)$ *zwei* σ*-endliche Maßräume und* $f \colon \Omega \times \Sigma \to \mathbb{R}$ $\mathcal{A} \otimes \mathcal{B}$*-messbar. Es gelte außerdem* $|f| \leq g$ *für ein* $g \in \mathcal{L}^1(\Omega \times \Sigma, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$*. Dann gilt auch* $f \in \mathcal{L}^1(\Omega \times \Sigma, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ *und*

$$\int_{\Omega \times \Sigma} f \, \mathrm{d}(\mu \times \nu) = \int_\Omega \int_\Sigma f(x, y) \, \nu(\mathrm{d}y) \, \mu(\mathrm{d}x) = \int_\Sigma \int_\Omega f(x, y) \, \mu(\mathrm{d}x) \, \nu(\mathrm{d}y)$$

Beweis. In der Maßtheorie.

Korollar. *Sind* $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ *unabhängig, so gilt auch* $XY \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ *und* $E_P[XY] = E_P[X]E_P[Y]$*.*

Beweis. Wir zeigen dies zuerst für $X, Y \geq 0$. Wir setzen $\mu = \mathcal{L}_P(X)$ und $\nu = \mathcal{L}_P(Y)$. Also ist wegen der Unabhängigkeit $\mathcal{L}_P(X, Y) = \mu \times \nu$. Es folgt

$$\begin{aligned} E_P[XY] &= \int_{\mathbb{R}^2} x_1 x_2 \, \mu \times \nu(\mathrm{d}x) = \int_{[0,\infty)} \int_{[0,\infty)} x_1 x_2 \, \nu(\mathrm{d}x_2) \, \mu(\mathrm{d}x_1) = \\ &= \int_{[0,\infty)} x_1 \, \mu(\mathrm{d}x_1) \int_{[0,\infty)} x_2 \, \nu(\mathrm{d}x_2) = E_P[X]E_P[Y] \end{aligned}$$

Nun seien X, Y von beliebigem Vorzeichen. Die Rechnung von eben zeigt, $E_P[|XY|] = E_P[|X|]E_P[|Y|] < \infty$, also $|XY| \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ und daher $XY \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$. Die Rechung von eben mit dem Satz von Fubini für integrierbare Funktionen zeigt die Behauptung. □

Beispiel. Sind X, Y unabhängig $N(0,1)$-verteilt, so gilt für $\alpha, \beta \in \mathbb{R}$:

$$\mathrm{Cov}_P(X, \alpha X + \beta Y) = \alpha \,\mathrm{Cov}_P(X, X) + \beta \,\mathrm{Cov}_P(X, Y) = \alpha$$

und

$$\begin{aligned}\mathrm{Var}_P(\alpha X + \beta Y) &= \mathrm{Cov}_P(\alpha X + \beta Y, \alpha X + \beta Y) = \alpha^2 \,\mathrm{Var}_P(X) + 2\alpha\beta \,\mathrm{Cov}_P(X, Y) + \beta^2 \,\mathrm{Var}_P(Y) = \\ &= \alpha^2 + \beta^2\end{aligned}$$

Damit gilt für den Korrelationskoeffizienten

$$r_P(X, \alpha X + \beta Y) = \frac{\alpha}{\sqrt{\alpha^2 + \beta^2}}$$

Bemerkung. Für $X_1, \ldots, X_n \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ wissen wir

$$E_P[X_1 + \cdots + X_n] = E_P[X_1] + \cdots + E_P[X_n]$$

Satz. *Seien $X_1, \ldots, X_n \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ unabhängige Zufallsvariablen. Dann gilt*

$$\mathrm{Var}_P\Big(\sum_{k=1}^n X_k\Big) = \sum_{k=1}^n \mathrm{Var}_P(X_k)$$

Beweis. Es gilt

$$\mathrm{Var}_P\Big(\sum_{k=1}^n X_k\Big) = \mathrm{Cov}_P\Big(\sum_{k=1}^n X_k, \sum_{k=1}^n X_k\Big) = \sum_{k,l=1}^n \mathrm{Cov}_P(X_k, X_l)$$

Nun gilt $\mathrm{Cov}_P(X_k, X_k) = \mathrm{Var}_P(X_k)$ und $\mathrm{Cov}_P(X_k, X_l) = 0$ für $k \neq l$ wegen der Unabhängigkeit, also

$$\mathrm{Var}_P\Big(\sum_{k=1}^n X_k\Big) = \sum_{k=1}^n \mathrm{Var}_P(X_k)$$ □

Bemerkung. Die Aussage gilt auch, wenn man statt Unabhängigkeit nur die Unkorreliertheit von $X_1, \ldots, X_n$, d.h. $\mathrm{Cov}_P(X_i, X_j) = 0$ für $i \neq j$, fordert.

Beispiel. Seien $X_1, \ldots, X_n$ i.i.d. $p\delta_1 + (1-p)\delta_0$-verteilt, $0 \leq p \leq 1$. Dann ist $S = \sum_{k=1}^n X_k$ binomial(n,p)-verteilt. Es folgt

$$E[S] = \sum_{k=1}^n \underbrace{E[X_k]}_{p} = np$$

und

$$\mathrm{Var}[S] = \sum_{k=1}^n \underbrace{\mathrm{Var}[X_k]}_{p(1-p)} = np(1-p), \quad \sigma(S) = \sqrt{np(1-p)}$$

Also ist $E[S] = O(n)$ und $\sigma(S) = O(\sqrt{n})$ für $n \to \infty$. Die binomial(n,p)-Verteilung hat also den Erwartungswert np, die Varianz $np(1-p)$ und die Standardabweichung $\sqrt{np(1-p)}$.

1.15 Momente und momentenerzeugende Funktionen

Definition. Es sei X eine Zufallsvariable auf $(\Omega, \mathcal{A}, P)$, $m \in \mathbb{N}$. Falls $E_p[X^m]$ existiert, heißt er das *m-te Moment* von X und $E_P[(X - E_P[X])^m]$ das *m-te zentrierte Moment*. Sei außerdem

$$\mathcal{L}^m(\Omega, \mathcal{A}, P) = \{X \colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ messbar} \colon E_P[|X|^m] < \infty\}$$

Die *Laplacetransformierte* oder *momentenerzeugende Funktion* von X wird definiert durch

$$L_X \colon \mathbb{R} \to [0, \infty], L_X(s) = E_P\Big[e^{sX}\Big]$$

Ausblick: für komplexe s wird die sogenannte Fourier-Laplacetransformierte analog definiert.

Satz (Lebesgue)**.** *Sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum, $V \subseteq \mathbb{R}$ offen, $f \colon \Omega \times V \to \mathbb{R}$ messbar im 1. Argument und differenzierbar im 2. Argument. Es gelte $f(\cdot, s) \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ für alle $s \in V$ und es existiere ein $g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ mit*

$$\left|\frac{\partial}{\partial s} f(\cdot, s)\right| \leq g, \quad s \in V$$

Dann ist

$$V \ni s \mapsto \int_\Omega f(\omega, s)\, \mu(\mathrm{d}\omega)$$

differenzierbar und es gilt

$$\frac{d}{ds} \int_\Omega f(\omega, s)\, \mu(\mathrm{d}\omega) = \int_\Omega \frac{\partial}{\partial s} f(\omega, s)\, \mu(\mathrm{d}\omega)$$

Beweis. In der Maßtheorie.

Satz. *Sei X eine Zufallsvariable über $(\Omega, \mathcal{A}, P)$ und $U \subseteq \mathbb{R}$ offen, so dass $L_X(s) < \infty$ für alle $s \in U$. Dann ist L_X auf U beliebig oft differenzierbar und es gilt für alle $m \in \mathbb{N}$ und $s \in U$:*

$$L_X^{(m)}(s) = E_P\Big[X^m e^{sX}\Big]$$

Besonders interessant ist das für $s = 0 \in U$. Dann gilt $L_X^{(m)}(0) = E_P[X^m]$.

Beweis. Der Beweis besteht in der Begründung der Vertauschbarkeit von E_P und $\frac{\partial}{\partial s}$:

$$\frac{d^m}{ds^m} E_P\Big[e^{sX}\Big] = E_P\left[\frac{\partial^m}{\partial s^m} e^{sX}\right] = E_P\Big[X^m e^{sX}\Big]$$

Um dies zu beweisen, brauchen wir für jedes $s \in U$, $m \in \mathbb{N}$ eine offene Umgebung $V \subseteq U$ von s und eine Zufallsvariable $g \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ mit

$$\Big|X^m e^{tX}\Big| \leq g, \quad t \in V$$

Sei hierzu $\varepsilon > 0$ so klein, dass $[s - 2\varepsilon, s + 2\varepsilon] \subseteq U$. Wir setzen $V = (s - \varepsilon, s + \varepsilon)$. Dann folgt für $t \in V$:

$$\Big|X^m e^{tX}\Big| = |X^m|\, e^{(t-s)X} e^{sX} \leq \underbrace{\left(\frac{m!}{\varepsilon^m} \sum_{k=0}^{\infty} \frac{|\varepsilon X|^k}{k!}\right)}_{m\text{-ter Summand ist } |X^m|} e^{\varepsilon|X|} e^{sX} = \frac{m!}{\varepsilon^m} e^{2\varepsilon|X|} e^{sX}$$

$$\leq \frac{m!}{\varepsilon^m} \left(e^{(s-2\varepsilon)X} + e^{(s+2\varepsilon)X}\right) \in \mathcal{L}^1(\Omega, \mathcal{A}, P) \qquad \square$$

1.15.1 Wichtige Eigenschaften der Laplacetransformierten

Satz. *Sind X und Y unabhängige Zufallsvariablen, so gilt*

$$L_{X+Y}(s) = L_X(s)L_Y(s)$$

Beweis.

$$L_{X+Y}(s) = E_P\Big[e^{s(X+Y)}\Big] = E_P\Big[e^{sX}e^{sY}\Big] = E_P\Big[e^{sX}\Big]E_P\Big[e^{sY}\Big] = L_X(s)L_Y(s) \qquad \square$$

Bemerkung. Analog gilt für jedes $s \in \mathbb{R}$

$$L_{\sum_{k=1}^n X_k}(s) = \prod_{k=1}^n L_{X_k}(s)$$

falls $X_1, \dots, X_n$ unabhängige Zufallsvariablen sind.

Beispiel. Seien $X_1, \dots, X_n$ i.i.d. $p\delta_1 + (1-p)\delta_0$-verteilt, $0 \le p \le 1$, also $S_n = X_1 + \cdots + X_n$ binomial(n,p)-verteilt, so gilt

$$L_{X_k}(t) = pe^{t1} + (1-p)e^{t0} = pe^t + 1 - p, \quad t \in \mathbb{R}$$

also

$$L_{S_n}(t) = \prod_{k=1}^n L_{X_k}(t) = \Big(pe^t + 1 - p\Big)^n$$

Insbesondere

$$L'_{S_n}(t) = npe^t\Big(pe^t + 1 - p\Big)^{n-1}$$
$$L''_{S_n}(t) = npe^t\Big(pe^t + 1 - p\Big)^{n-1} + n(n-1)\Big(pe^t\Big)^2\Big(pe^t + 1 - p\Big)^{n-2}$$

also folgt $E_P[S_n] = L'_{S_n}(0) = np$ und $E_P[S_n^2] = np + n(n-1)p^2$ und daher $\mathrm{Var}_P(S_n) = E_P[S_n^2] - E_P[S_n]^2 = np(1-p)$.

1.15.2 Allgemeine Tschebyscheffungleichung

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $A \in \mathcal{A}$. Die einfache Gleichung

$$P(A) = E_P[1_A]$$

hat viele Konsequenzen, z.B. die Siebformel: Für $A_1, \dots, A_n$ gilt

$$P(A_1 \cup \cdots \cup A_n) = \sum_{\substack{E \subseteq \{1,\dots,n\} \\ E \neq \emptyset}} (-1)^{|E|+1} P\Big(\bigcap_{i \in E} A_i\Big)$$

Lemma (Allgemeine Tschebyscheffungleichung)**.** *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $A \in \mathcal{A}$, $X \ge 0$ eine Zufallsvariable und $c \ge 0$. Es gelte $X(\omega) \ge c$ für alle $\omega \in A$. Dann folgt*

$$E_P[X] \ge cP(A)$$

Beweis. Nach Voraussetzung gilt $X \ge c1_A$, also

$$E_P[X] \ge E_P[c1_A] = cE_P[1_A] = cP(A) \qquad \square$$

Beispiel. Für alle $s \geq 0$, $a \in \mathbb{R}$ gilt

$$E_P\left[e^{sX}\right] \geq e^{sa} P[X \geq a]$$

denn es gilt $e^{sX} \geq e^{sa} 1_{\{X \geq a\}}$, also folgt dies aus der allgemeinen Tschebyscheffungleichung. Optimierung über s liefert:

$$P[X \geq a] \leq \inf_{s \geq 0} e^{-sa} E_P\left[e^{sX}\right]$$

Analog:

$$P[X \leq a] \leq \inf_{s \leq 0} e^{-sa} E_P\left[e^{sX}\right]$$

Beispiel. Seien Y_k, $k \in \mathbb{N}$m i.i.d. $p\delta_1 + (1-p)\delta_0$-verteilte Zufallsvariablen, $0 < p < 1$. Dann ist $X_n = \sum_{k=1}^n Y_k$ binomial(n,p)-verteilt, und es folgt:

$$\begin{aligned}\text{binomial}(n,p)([na,\infty)) = P[X_n \geq na] \leq \inf_{s \geq 0} e^{-sna} E_P\left[e^{sX_n}\right] = \inf_{s \geq 0} e^{-sna} \left(pe^s + 1 - p\right)^n = \\ = \inf_{s \geq 0} \exp\left(n \underbrace{\left(-sa + \log\left(pe^s + 1 - p\right)\right)}_{H(s)}\right)\end{aligned}$$

Wir optimieren über $s \geq 0$:

$$\begin{aligned}H'(s) &= -a + \frac{pe^s}{pe^s + 1 - p} = -a + \frac{1}{1 + \frac{1-p}{p} e^{-s}} \\ H''(s) &= \frac{\frac{1-p}{p} e^{-s}}{\left(1 + \frac{1-p}{p} e^{-s}\right)^2} > 0\end{aligned}$$

Um das Minimum von H zu finden, lösen wir die Gleichung $H'(s) = 0$.

$$\begin{aligned}H'(s) &\iff 1 + \frac{1-p}{p} e^{-s} = \frac{1}{a} \\ &\iff \frac{1-p}{p} e^{-s} = \frac{1-a}{a} \\ &\iff s = \log\left(\frac{1-p}{p} \frac{a}{1-a}\right)\end{aligned}$$

Wenn $1 > a \geq p > 0$ ist $\frac{a}{p} \geq 1$ und $\frac{1-p}{1-a} \geq 1$, also $s \geq 0$. Eingesetzt erhalten wir für das Optimum:

$$pe^s + 1 - p = (1-p)\frac{a}{1-a} + 1 - p = \frac{1-p}{1-a}$$

also

$$\begin{aligned}H(s) = -sa + \log(pe^s + 1 - p) = a \log\left(\frac{p}{1-p} \frac{1-a}{a}\right) + \log \frac{1-p}{1-a} = a \log \frac{p}{a} + (1-a) \log \frac{1-p}{1-a} \leq \\ \leq a\left(\frac{p}{a} - 1\right) + (1-a)\left(\frac{1-p}{1-a} - 1\right) = 0\end{aligned}$$

Es ist also $H(s) < 0$ für $1 > a > p > 0$.

Satz. *Ist* X_n binomial(n,p)*-verteilt,* $0 < p < a < 1$*, so gilt:*

$$P[X_n \geq na] \leq \exp\left(n \underbrace{\left(a \log \frac{p}{a} + (1-a) \log \frac{1-p}{1-a}\right)}_{<0}\right)$$

Interpretation. $\frac{X_n}{n}$ bedeutet die relative Häufigkeit von "1" in einem Münzwurfexperiment.

$$P\left[\frac{X_n}{n} \geq a\right] \xrightarrow{n\to\infty} 0 \quad \text{exponentiell schnell}$$

Numerisches Beispiel: fairer Münzwurf $p = \frac{1}{2}$, $a = 0.6 = 60\%$.

$$H = a \log \frac{p}{a} + (1-a) \log \frac{1-p}{1-a} = -0.020\ldots$$

Wir erhalten für $n = 1000$

$$P[X_{1000} \geq 600] \leq e^{1000H} = 1.79\ldots \cdot 10^{-9}$$

und für $n = 10000$

$$P[X_{10000} \geq 6000] \leq e^{10000H} = 3.5\ldots \cdot 10^{-88}$$

Eine weitere Anwendung der allgemeinen Tschebyscheff-Ungleichung ist die Markov-Ungleichung:

Satz (Markov-Ungleichung)**.** *Sei* X *eine Zufallsvariable über* $(\Omega, \mathcal{A}, P)$*,* $m > 0$ *und* $a > 0$*. Dann gilt*

$$a^m P[|X| \geq a] \leq E_P[|X|^m]$$

Beweis. Es gilt

$$a^m 1_{\{|X| \geq a\}} \leq |X|^m$$

also folgt

$$a^m P[|X| \geq a] = E_P[a^m 1_{\{|X| \geq a\}}] \leq E_P[|X|^m] \qquad \square$$

Bemerkung. Im Fall $m = 1$ gilt $P[|X| \geq a] \leq \frac{E_P[|X|]}{a}$.

Bemerkung. Im Fall $m = 2$ ist dies die *quadratische Tschebyscheff-Ungleichung*: Ist $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$, so gilt für alle $a > 0$

$$a^2 P\big[|X - E_P[X]| \geq a\big] \leq \mathrm{Var}_P(X)$$

denn mit $Y = X - E_P[X]$ und $m = 2$ folgt aus der Markovungleichung

$$a^2 P[|Y| \geq a] \leq E_P[Y^2] = \mathrm{Var}_P(X)$$

1.16 Gesetze der großen Zahlen

1.16.1 Das schwache Gesetz der großen Zahlen

Es seien $X_1, X_2, \ldots$ i.i.d. Zufallsvariablen über $(\Omega, \mathcal{A}, P)$ mit endlicher Erwartung $E_P[X_1]$.

Intuition. Für "große" n ist das Mittel $\overline{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ "typischerweise" nahe bei $E_P[X]$.

Definition. Eine Folge Y_n, $n \in \mathbb{N}$, von Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ *konvergiert in Wahrscheinlichkeit* oder *konvergiert stochastisch* gegen $a \in \mathbb{R}$, in Zeichen $Y_n \xrightarrow[P]{n\to\infty} a$, wenn gilt

$$\forall \varepsilon > 0.\ P[|Y_n - a| \geq \varepsilon] \xrightarrow{n\to\infty} 0$$

Satz (Schwaches Gesetz der großen Zahlen)**.** *Es seien X_n, $n \in \mathbb{N}$, i.i.d. in $\mathcal{L}^2(\Omega, \mathcal{A}, P)$. Dann gilt für den Mittelwert $\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$:*

$$\overline{X}_n \xrightarrow[P]{n\to\infty} E_P[X_1]$$

Beweis. Sei $\varepsilon > 0$. Wir kürzen ab: $a = E_P[X_1] = E_P[X_k]$ für alle $k \in \mathbb{N}$, also gilt auch $a = \frac{1}{n}\sum_{k=1}^n E_P[X_k] = E_P[\overline{X}_k]$. Weiter gilt

$$\mathrm{Var}_P(\overline{X}_n) = \mathrm{Var}_P\left(\frac{1}{n}\sum_{k=1}^n X_k\right) = \frac{1}{n^2}\mathrm{Var}_P\left(\sum_{k=1}^n X_k\right) = \frac{1}{n^2}\sum_{k=1}^n \mathrm{Var}_P(X_k) = \frac{1}{n}\mathrm{Var}_P(X_1)$$

Es folgt

$$P\Big[|\overline{X}_n - a| \geq \varepsilon\Big] \leq \frac{1}{\varepsilon^2} E_P\Big[(\overline{X}_n - a)^2\Big] = \frac{1}{\varepsilon^2}\mathrm{Var}_P(\overline{X}_n) = \frac{1}{\varepsilon^2 n}\mathrm{Var}_P(X_1) \xrightarrow{n\to\infty} 0 \qquad \square$$

Bemerkung. Die X_n, $n \in \mathbb{N}$, müssen nicht unabhängig sein; der Beweis funktioniert genauso, wenn sie unkorreliert sind.

Im Spezialfall $X_k = 1_{A_k}$ mit $P(A_k) = p$ für alle k erhalten wir das schwache Gesetz der großen Zahlen für relative Häufigkeiten:

Korollar (Schwaches Gesetz der großen Zahlen für relative Häufigkeiten)**.** *Sind A_n, $n \in \mathbb{N}$, unabhängige Ereignisse mit gleicher Wahrscheinlichkeit $P(A_k) = p$, so gilt*

$$\frac{1}{n}\sum_{k=1}^n 1_{A_k} \xrightarrow[P]{n\to\infty} p$$

Das schwache Gesetz der großen Zahlen kann als innermathematisches Analogon der objektivistischen Interpretation von Wahrscheinlichkeiten aufgefasst werden. Es lierfert auch ein Fundament für die Minimalinterpretation von Wahrscheinlichkeiten: Eine beliebige Wahrscheinlichkeit $P[A_k] = p$ wird durch unabhängige Wiederholung des Experiments mit dem Gesetz der großen Zahlen zu der Aussage $P\big[\big|\frac{1}{n}\sum_{k=1}^n 1_{A_k} - p\big| \geq \varepsilon\big] \xrightarrow{n\to\infty} 0$.

1.16.2 Das starke Gesetz der großen Zahlen

Seien $(A_n)_{n\in\mathbb{N}}$ unabhängige Ereignisse mit gleicher Wahrscheinlichkeit p. Aus der quadratischen Tschebyscheff-Ungleichung folgt

$$P\Big[\Big|\frac{1}{n}\sum_{k=1}^n 1_{A_k} - p\Big| \geq \varepsilon\Big] \leq \frac{1}{n\varepsilon^2}\mathrm{Var}_P(1_{A_k}) = \frac{p(1-p)}{\varepsilon^2 n}$$

Für große n ist diese Abschätzung extrem unscharf, denn aus der exponentiellen Tschebyscheff-Ungleichung folgt:

$$P\Big[\frac{1}{n}\sum_{k=1}^n 1_{A_k} - p \geq \varepsilon\Big] \leq e^{-H_+ n}$$

$$P\Big[\frac{1}{n}\sum_{k=1}^n 1_{A_k} - p \leq -\varepsilon\Big] \leq e^{-H_- n}$$

mit Konstanten $H_+, H_- > 0$. Insbesondere folgt mit $\alpha = \min\{H_+, H_-\}$

$$\sum_{n\in\mathbb{N}} P\Big[\Big|\frac{1}{n}\sum_{k=1}^{n} - p\Big| \geq \varepsilon\Big] \leq \sum_{n\in\mathbb{N}} 2e^{\alpha n} = 2\frac{e^{-\alpha}}{1-e^{-\alpha}} < \infty$$

Die quadratische Tschebyscheff-Ungleichung reicht hierfür nicht, weil die harmonische Reihe divergiert. Diese Summierbarkeit wird mit folgendem Lemma bedeutsam:

Lemma (1. Lemma von Borel-Cantelli)**.** *Ist $(B_n)_{n\in\mathbb{N}}$ eine Folge von Ereignissen mit $\sum_{n\in\mathbb{N}} P(B_n) < \infty$, so gilt*

$$P[B_n \text{für unendlich viele } n] = 0$$

d.h. P-fast sicher tritt B_n nur für endlich viele n ein.

Beweis. Es ist

$$\{B_n \text{ für unendlich viele } n\} = \{\omega \in \Omega\colon \forall m \in \mathbb{N}\ \exists n \geq m.\ \omega \in B_n\} = \bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} B_n$$

Nun gilt für alle $m \in \mathbb{N}$ wegen $\sum_{k\in\mathbb{N}} P(B_n) < \infty$:

$$P\Big(\bigcup_{n\geq m} B_n\Big) = \lim_{k\to\infty} P\Big(\bigcup_{n=m}^{k} B_n\Big) \leq \lim_{k\to\infty} \sum_{n=m}^{k} P(B_n) = \sum_{n=m}^{\infty} P(B_n) \xrightarrow{m\to\infty} 0$$

Nun fällt die Folge $\big(\bigcup_{n\geq m} B_n\big)_{m\in\mathbb{N}}$ monoton. Mit der σ-Stetigkeit von oben folgt

$$P\Big(\bigcup_{n\geq m} B_n\Big) \xrightarrow{m\to\infty} P\Big(\bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} B_n\Big)$$

Es folgt also

$$P[B_n \text{ für unendlich viele } n] = 0$$ □

Satz (Starkes Gesetz der großen Zahlen für relative Häufigkeiten)**.** *Es sei $(A_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger Ereignisse über $(\Omega, \mathcal{A}, P)$ mit gleicher Wahrscheinlichkeit p. Dann gilt P-fast sicher*

$$\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} \xrightarrow{n\to\infty} p$$

d.h.

$$P\Big(\Big\{\omega \in \Omega\colon \frac{1}{n}\sum_{k=1}^{n} 1_{A_k}(\omega) \xrightarrow{n\to\infty} p\Big\}\Big) = 1$$

Beweis. Die Fälle $p = 0$ und $p = 1$ sind trivial. Wir nehmen also $0 < p < 1$ an. Wir wissen für alle $\varepsilon > 0$:

$$\sum_{n\in\mathbb{N}} P\Big[\underbrace{\Big|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\Big| \geq \varepsilon}_{\text{Ereignis } B_n(\varepsilon)}\Big] < \infty$$

Mit dem 1. Borel-Cantelli-Lemma folgt

$$P[\underbrace{B_n(\varepsilon) \text{ für unendlich viele } n}_{\text{Ereignis } C(\varepsilon)}] = 0$$

Wegen $|\mathbb{Q}^+| = |\mathbb{N}|$, folgt

$$P\Big(\bigcup_{\varepsilon\in\mathbb{Q}^+} C(\varepsilon)\Big) = 0$$

Also gilt P-fast sicher

$$\forall\varepsilon > 0, \varepsilon \in \mathbb{Q}^+ \ \exists m \in \mathbb{N} \ \forall n \geq m. \ \Big|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\Big| < \varepsilon$$

d.h. es gilt P-fast sicher

$$\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} \xrightarrow{n\to\infty} p \qquad \square$$

Das starke Gesetz der großen Zahlen (st. G. d. gr. Z.) für relative Häufigkeit kann als innermathematisches Analogon zu von Mises-Interpretation von Wahrscheinlichkeiten aufgefasst werden: Die relative Häufigkeit bei n Versuchen konvergiert mit Wahrscheinlichkeit 1 für $n \to \infty$ gegen die Wahrscheinlichkeit p.

Verallgemeinerung (Starkes Gesetz der großen Zahlen für i.i.d. Zufallsvariablen mit exponentiellem Abfall)**.** Es seien X_n, $n \in \mathbb{N}$, i.i.d. Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Es existiere ein $\alpha > 0$ mit $E_P\big[e^{\alpha|X_1|}\big] < \infty$. Dann gilt P-fast sicher

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{n\to\infty} E_P[X_1]$$

Anders gesagt:

$$P\Big(\Big\{\omega \in \Omega\colon \frac{1}{n}\sum_{k=1}^{n} X_k(\omega) \xrightarrow{n\to\infty} E_P[X_1]\Big\}\Big) = 1$$

Beweis. Aus der Voraussetzung $E_P\big[e^{\alpha|X_1|}\big] < \infty$ folgt für alle $s \in [-\alpha, \alpha]$:

$$L_{X_1}(s) = E_P\big[e^{sX_1}\big] \leq E_P\big[e^{\alpha|X_1|}\big] < \infty$$

Also ist L_{X_1} in einer Umgebung von 0 beliebig oft differenzierbar mit $L'_{X_1}(0) = E_P[X_1] =: \mu$. Es sei $s \in (0, \alpha]$. Dann folgt für $\varepsilon > 0$:

$$\begin{aligned} P\Big[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \mu + \varepsilon\Big] &= P\big[\sum_{k=1}^{n} X_k \geq n(\mu+\varepsilon)\big] \leq e^{-n(\mu+\varepsilon)s} E_P\Big[\exp\Big(s\sum_{k=1}^{n} X_k\Big)\Big] = \\ &= e^{-n(\mu+\varepsilon)s} E_P\Big[\prod_{k=1}^{n} e^{sX_k}\Big] = e^{-n(\mu+\varepsilon)s} \prod_{k=1}^{n} E_P\big[e^{sX_k}\big] = \\ &= e^{-n(\mu+\varepsilon)s} L_{X_1}(s)^n = \Big(\underbrace{e^{-(\mu+\varepsilon)s} L_{X_1}(s)}_{=1 \text{ für } s=0}\Big)^n \end{aligned}$$

Nun gilt

$$\frac{d}{ds}\Big|_{s=0} \Big[e^{-(\mu+\varepsilon)s} L_{X_1}(s)\Big] = \Big[-(\mu+\varepsilon)e^{-(\mu+\varepsilon)s} L_{X_1}(s) + e^{-(\mu+\varepsilon)s} L'_{X_1}(s)\Big]_{s=0} = -\varepsilon < 0$$

Zusammen mit $e^{-(\mu+\varepsilon)s} L_{X_1}(s)\Big|_{s=0} = 1$ folgt, dass es ein $s \in (0, a)$ gibt, mit

$$0 \leq \xi := e^{-(\mu+\varepsilon)s} L_{X_1}(s) < e^{-(\mu+\varepsilon)0} L_{X_1}(0) = 1$$

also $\sum_{n\in\mathbb{N}} \xi^n < \infty$. Damit ist gezeigt:

$$\sum_{n\in\mathbb{N}} P\Big[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \mu + \varepsilon\Big] \leq \sum_{n\in\mathbb{N}} \xi^n < \infty$$

Mit dem 1. Borel-Cantelli-Lemma folgt, dass P-fast sicher

$$\frac{1}{n}\sum_{k=1}^{n} X_k < \mu + \varepsilon \text{ für alle bis auf endlich viele } n$$

Analog folgt P-fast sicher

$$\frac{1}{n}\sum_{k=1}^{n} X_k > \mu - \varepsilon \text{ für alle bis auf endlich viele } n$$

Weil das für alle $\varepsilon > 0$, insbesondere für alle rationalen $\varepsilon > 0$, gilt folgt P-fast sicher

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{n\to\infty} \mu \qquad \square$$

Bemerkung. Die Voraussetzung $E_P\big[e^{\alpha|X_1|}\big] < \infty$ für geeignetes α kann zu $X_1 \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ abgeschwächt werden.

1.16.3 Der zentrale Grenzwertsatz

In Anwendungen nimmt man oft an, dass Fehler normalverteilt sind. Der zentrale Grenzwertsatz liefert ein Motiv für diese Annahme. Gegeben sei eine Folge $(X_n)_{n\in\mathbb{N}}$ binomial(n,p)-verteilter Zufallsvariablen, $0 < p < 1$. Insbesondere ist $E[X_n] = np$ und $\sigma(X_n) = \sqrt{np(1-p)}$. Wir betrachten die Dichte einer Normalverteilung mit genau den gleichen Parametern:

$$f_{n,p}\colon \mathbb{R} \to \mathbb{R},\ f_{n,p}(x) = \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{1}{2}\frac{(x-np)^2}{np(1-p)}\right)$$

Dann gilt

Satz (Satz von de Moivre-Laplace)**.** *Für alle $M > 0$ gilt*

$$\max\left\{\left|\frac{P[X_n = k]}{f_{n,p}(k)} - 1\right| : k \in \mathbb{N}, \left|\frac{k-np}{\sqrt{np(1-p)}}\right| \leq M\right\} \xrightarrow{n\to\infty} 0$$

Anschaulich, vergröbert gesagt: $P[X_n = k]$ liegt nahe bei $f_{n,p}(k)$, wenn nur $|k - E_P[X_n]|$ höchstens ein vorgegebenes Vielfaches von $\sigma_P(X_n)$ ist (für $n \to \infty$).

Beweis. Der Beweis beruht auf einer Näherungsformel für $n!$, der Stirlingformel:

$$\frac{m!}{\sqrt{2\pi}m^{m+\frac{1}{2}}e^{-m}} \xrightarrow{m\to\infty} 1$$

Wir benötigen hier eine quantitative Verstärkung davon:

$$\forall m \in \mathbb{N}\ \exists \vartheta_m \in \left(0, \frac{1}{12m}\right).\ m! = \sqrt{2\pi}m^{m+\frac{1}{2}}e^{-m}e^{\vartheta_m}$$

d.h.

$$0 < \log m! - \left(\log\sqrt{2\pi} + \left(m+\frac{1}{2}\right)\log m - m\right) < \frac{1}{12m}$$

Wir erhalten damit folgende Näherungsformel für $\binom{n}{k}$:

$$\begin{aligned}\log\binom{n}{k} &= \log\frac{n!}{k!(n-k)!} = \log n! - \log k! - \log(n-k)! = \\ &= -\log\sqrt{2\pi} + \left(n+\frac{1}{2}\right)\log n - \left(k+\frac{1}{2}\right)\log k - \left(n-k+\frac{1}{2}\right)\log(n-k) + \vartheta_n - \vartheta_k - \vartheta_{n-k} \\ &= -\log\sqrt{2\pi} + \frac{1}{2}\log\frac{n}{k(n-k)} + n\log n - k\log k - (n-k)\log(n-k) + \vartheta_n - \vartheta_k - \vartheta_{n-k}\end{aligned}$$

Es folgt

$$\begin{aligned}\log P[X_n = k] &= \log\left(\binom{n}{k}p^k(1-p)^{n-k}\right) = \log\binom{n}{k} + k\log p + (n-k)\log(1-p) = \\ &= -\log\sqrt{2\pi} + \frac{1}{2}\log\frac{n}{k(n-k)} - k\log\frac{k}{np} - (n-k)\log\frac{n-k}{n(1-p)} + \vartheta_n - \vartheta_k - \vartheta_{n-k}\end{aligned}$$

Wir analysieren die Terme einzeln. Sei

$$\mathcal{M}_n = \left\{k \in \mathbb{N}\colon \left|\frac{k-np}{\sqrt{np(1-p)}}\right| \le M\right\}$$

Es gilt

$$\begin{aligned}\max_{k\in\mathcal{M}_n}\left|\frac{k}{np} - 1\right| &= \max_{k\in\mathcal{M}_n}\left|\frac{k-np}{np}\right| = \\ &= \frac{\sqrt{np(1-p)}}{np}\max_{k\in\mathcal{M}_n}\left|\frac{k-np}{\sqrt{np(1-p)}}\right| \le \frac{\sqrt{np(1-p)}}{np}M = \\ &= M\sqrt{\frac{1-p}{p}}\frac{1}{\sqrt{n}} \xrightarrow{n\to\infty} 0\end{aligned}$$

Analog erhält man

$$\max_{k\in\mathcal{M}_n}\left|\frac{n-k}{n(1-p)} - 1\right| = \max_{k\in\mathcal{M}_n}\left|\frac{k-np}{n(1-p)}\right| \le M\sqrt{\frac{p}{1-p}}\frac{1}{\sqrt{n}} \xrightarrow{n\to\infty} 0$$

Wir schreiben

$$\begin{aligned}k\log\frac{k}{np} &= np\frac{k}{np}\log\frac{k}{np} \\ (n-k)\log\frac{n-k}{n(1-p)} &= n(1-p)\frac{n-k}{n(1-p)}\log\frac{n-k}{n(1-p)}\end{aligned}$$

Für eine Näherung hiervon entwickeln wir $f(x) = x\log x$ um $x_0 = 1$:

$$\begin{aligned}f'(x) &= 1 + \log x & f''(x) &= \frac{1}{x} & f'''(x) &= -\frac{1}{x^2} \\ f'(1) &= 1 & f''(1) &= 1\end{aligned}$$

Also gilt

$$x \log x = (x-1) + \frac{1}{2}(x-1)^2 + r(x)$$

mit $|r(x)| \leq \text{const} \cdot |x-1|^3$ für x nahe bei 1. Es folgt:

$$\frac{k}{np} \log \frac{k}{np} = \frac{k}{np} - 1 + \frac{1}{2}\left(\frac{k}{np} - 1\right)^2 + r\left(\frac{k}{np}\right)$$
$$\frac{k}{\log} \frac{k}{np} = k - np + \frac{1}{2}(1-p)\frac{(k-np)^2}{np(1-p)} + npr\left(\frac{k}{np}\right)$$

mit der Schranke für den Restterm:

$$\max_{k \in \mathcal{M}_n} \left| npr\left(\frac{k}{np}\right)\right| \leq \max_{k \in \mathcal{M}_n} \left(np \left| \frac{k}{np} - 1 \right|^3 \right) \leq M^3 (1-p)^{\frac{3}{2}} p^{-\frac{1}{2}} n \frac{1}{\sqrt{n}^3} \xrightarrow{n \to \infty} 0$$

Analog folgt

$$(n-k) \log \frac{n-p}{n(1-p)} = np - k + \frac{1}{2} p \frac{(k-np)^2}{np(1-p)} + n(1-p) r\left(\frac{n-k}{n(1-p)}\right)$$

wobei

$$\max_{k \in \mathcal{M}_n} \left| n(1-p) r\left(\frac{n-k}{n(1-p)}\right)\right| \leq \max_{k \in \mathcal{M}_n} n(1-p) \left| \frac{n-k}{n(1-p)} - 1 \right|^3 \leq M^3 p^{\frac{3}{2}} (1-p)^{-\frac{1}{2}} \frac{1}{\sqrt{n}} \xrightarrow{n \to \infty} 0$$

Zusammen folgt:

$$k \log \frac{k}{np} + (n-k) \log \frac{n-k}{n(1-p)} = \frac{1}{2} \frac{(k-np)^2}{np(1-p)} + r_2(n,p,k)$$

wobei

$$r_2(n,p,k) = npr\left(\frac{k}{np}\right) + n(1-p) r\left(\frac{n-k}{n(1-p)}\right)$$

folgende Schranke erfüllt:

$$\max_{k \in \mathcal{M}_n} |r_2(n,p,k)| \xrightarrow{n \to \infty} 0$$

Weiter gilt

$$\log \frac{n}{k(n-k)} = \log \frac{1}{np(1-p)} - \log \frac{k}{np} - \log \frac{n-k}{n(1-p)}$$
$$= \log \frac{1}{np(1-p)} + r_3(n,k,p)$$

wobei auch $\max_{k \in \mathcal{M}_n} |r_3(n,k,p)| \xrightarrow{n \to \infty} 0$. Schließlich gilt

$$\min_{k \in \mathcal{M}_n} k \geq np - M\sqrt{np(1-p)} \xrightarrow{n \to \infty} \infty$$
$$\min_{k \in \mathcal{M}_n} (n-k) \geq n(1-p) - M\sqrt{np(1-p)} \xrightarrow{n \to \infty} \infty$$

Also folgt

$$\max_{k \in \mathcal{M}n} \vartheta_k \leq \max_{k \in \mathcal{M}_n} \frac{1}{12k} \xrightarrow{n \to \infty} 0$$
$$\max_{k \in \mathcal{M}_n} \vartheta_{n-k} \leq \max_{k \in \mathcal{M}_n} \frac{1}{12(n-k)} \xrightarrow{n \to \infty} 0$$

Fassen wir zusammen:

$$\log P[X_n = k] = -\log\sqrt{2\pi} + \frac{1}{2}\log\frac{1}{np(1-p)} - \frac{1}{2}\frac{k-np}{np(1-p)} + r_4(n,k,p)$$

wobei

$$r_4(n,k,p) = \frac{1}{2}r_3(n,k,p) - r_2(n,k,p) + \vartheta_n - \vartheta_k - \vartheta_{n-k}$$

folgende Fehlerschranke erfüllt:

$$\max_{k\in\mathcal{M}_n} |r_4(n,k,p)| \xrightarrow{n\to\infty} 0$$

Es folgt

$$\max_{k\in\mathcal{M}_n} \left|\log\frac{P[X_n=k]}{f_{n,p}(k)}\right| = \max_{k\in\mathcal{M}_n} |r_4(n,k,p)| \xrightarrow{n\to\infty} 0$$

also auch

$$\max_{k\in\mathcal{M}_n} \left|\frac{P[X_n=k]}{f_{n,p}(k)} - 1\right| \xrightarrow{n\to\infty} 0$$ □

Bemerkung. Der Beweis funktioniert auch, wenn $M = M_n$ von n abhängig gemacht wird, solange $\frac{M_n}{\sqrt{n}} \xrightarrow{n\to\infty} 0$.

Korollar. *Es seien X_n, $n \in \mathbb{N}$, Zufallsvariablen mit $\mathcal{L}(X_n) = \text{binomial}(n,p)$ und*

$$Z_n = \frac{X_n - E_P[X_n]}{\sigma_P(X_n)} = \frac{X_n - np}{\sqrt{np(1-p)}}$$

Weiter sei Z standardnormalverteilt. Dann gilt für alle $a, b \in \mathbb{R}$ mit $a < b$:

$$P[a \le Z_n \le b] \xrightarrow{n\to\infty} P[a \le Z \le b]$$

1.16.4 Der zentrale Grenzwertsatz für i.i.d. Zufallsvariablen aus $\mathcal{L}^2$

Satz (Zentraler Grenzwertsatz)**.** *Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von i.i.d. Zufallsvariablen mit endlicher positiver Varianz,*

$$S_n = \sum_{k=1}^{n} X_k \qquad\qquad Z_n = \frac{S_n - E_P[S_n]}{\sigma_P(S_n)} = \frac{S_n - nE_P[X_1]}{\sqrt{n}\sigma_P(X_1)}$$

Weiter sei Z standardnormalverteilt. Dann gilt für alle Intervalle $I \subseteq \mathbb{R}$:

$$P[Z_n \in I] \xrightarrow{n\to\infty} P[Z \in I]$$

Wir führen den Satz auf folgende Variante zurück:

Satz. *Sei $f\colon \mathbb{R} \to \mathbb{R}$ dreimal stetig differenzierbar und beschränkt mit beschränkten Ableitungen bis zur 3. Stufe, d.h. $f \in \mathcal{C}_b^3(\mathbb{R})$. Dann gilt mit den Voraussetzungen von eben*

$$E_P[f(Z_n)] \xrightarrow{n\to\infty} E_P[f(Z)]$$

Beweis. Ohne Einschränkung gelte $E_P[X_n] = 0$ und $\mathrm{Var}_P(X_n) = 1$, sonst ersetze wir X_n durch $\tilde{X}_n = \frac{X_n - E_P[X_n]}{\sigma_P(X_n)}$. Weiter existiere ohne Einschränkung auf $(\Omega, \mathcal{A}, P)$ eine i.i.d. Folge $(Y_n)_{n\in\mathbb{N}}$ von standardnormalverteilten Zufallsvariablen, unabhängig von $(X_n)_{n\in\mathbb{N}}$. (Wenn nötig ersetzen wir die X_n durch eine andere i.i.d. Folge mit der gleichen Verteilung auf einem anderen Wahrscheinlichkeitsraum.) Dann gilt:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k \qquad\qquad \mathcal{L}_P\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k\right) = N(0,1)$$

Wir müssen also zeigen:

$$\Delta := E_P\left[f\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k\right)\right] - E_P\left[f\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k\right)\right] \xrightarrow{n\to\infty} 0$$

Wir zerlegen diese Differenz in eine Teleskopsumme, sei $T_{n,l} = \frac{1}{\sqrt{n}} \sum_{k=1}^{l-1} Y_k + \frac{1}{\sqrt{n}} \sum_{k=l+1}^{n} X_k$.

$$\Delta = \sum_{l=1}^{n} \underbrace{E\left[f\left(T_{n,l} + \frac{1}{\sqrt{n}} X_l\right) - f\left(T_{n,l} + \frac{1}{\sqrt{n}} Y_l\right)\right]}_{=:\Delta'}$$

Wir entwickeln f nach Taylor um $t \in \mathbb{R}$. Für $x \in \mathbb{R}$ existieren $\vartheta_2, \vartheta_3 \in [0,1]$ mit

$$\begin{aligned} f(t+x) &= f(t) + xf'(t) + \frac{x^2}{2} f''(t) + \frac{x^3}{6} f''(t + \theta_3 x) = \\ &= f(t) + xf'(t) + \frac{x^2}{2} f''(t) + \frac{x^2}{2}(f''(t + \vartheta_2 x) - f''(t)) \end{aligned}$$

also

$$f(t+x) = f(t) + xf'(t) + \frac{x^2}{2} f''(t) + r(t,x) \text{ mit } |r(t,x)| \le \min\left\{x^2 \sup|f''|, \frac{|x|^3}{6} \sup|f'''|\right\}$$

Es gilt also $|r(t,x)| \le c_f \min\{x^2, |x|^3\}$ mit $c_f = \max\left\{\sup|f''|, \frac{1}{6}\sup|f'''|\right\} < \infty$. Es folgt:

$$\begin{aligned} \Delta' = &E\left[\frac{X_l}{\sqrt{n}} f'(T_{n,l})\right] - E\left[\frac{Y_l}{\sqrt{n}} f'(T_{n,l})\right] + \frac{1}{2} E\left[\frac{X_l^2}{n} f''(T_{n,l})\right] - \frac{1}{2} E\left[\frac{Y_l^2}{n} f''(T_{n,l})\right] + \\ &+ E\left[r\left(T_{n,l}, \frac{X_l}{\sqrt{n}}\right) - r\left(T_{n,l}, \frac{Y_l}{\sqrt{n}}\right)\right] \end{aligned}$$

Nun gilt $E\left[\frac{X_l}{\sqrt{n}} f'(T_{n,l})\right] = \frac{1}{\sqrt{n}} E[X_l] E[f'(T_{n,l})] = 0$ wegen der Unabhängigkeit von X_l und $f'(T_{n,l})$ und analog $E\left[\frac{Y_l}{\sqrt{n}} f'(T_{n,l})\right] = 0$. Weiterhin gilt:

$$E\left[\frac{X_l^2}{n} f''(T_{n,l})\right] = \frac{1}{n} E[X_l^2] E[f''(T_{n,l})] = \frac{1}{n} E[f''(T_{n,l})] = E\left[\frac{Y_l^2}{n} f''(T_{n,l})\right]$$

und daher

$$\begin{aligned} |\Delta'| &= \left|E\left[r\left(T_{n,l}, \frac{X_l}{\sqrt{n}}\right) - r\left(T_{n,l}, \frac{Y_l}{\sqrt{n}}\right)\right]\right| \le E\left[\left|r\left(T_{n,l}, \frac{X_l}{\sqrt{n}}\right)\right|\right] + E\left[\left|r\left(T_{n,l}, \frac{Y_l}{\sqrt{n}}\right)\right|\right] \le \\ &\le \frac{c_f}{n}\left(E\left[\min\left\{X_l^2, \frac{|X_l|^3}{\sqrt{n}}\right\}\right] + E\left[\min\left\{Y_l^2, \frac{|Y_l|^3}{\sqrt{n}}\right\}\right]\right) = \\ &= \frac{c_f}{n}\left(E\left[\min\left\{X_1^2, \frac{|X_1|^3}{\sqrt{n}}\right\}\right] + E\left[\min\left\{Y_1^2, \frac{|Y_1|^3}{\sqrt{n}}\right\}\right]\right) \end{aligned}$$

Eingesetzt in die Teleskopsumme erhalten wir:

$$|\Delta| \leq n\frac{c_f}{n}\left(E\left[\min\left\{X_1^2, \frac{|X_1|^3}{\sqrt{n}}\right\}\right] + E\left[\min\left\{Y_1^2, \frac{|Y_1|^3}{\sqrt{n}}\right\}\right]\right) \xrightarrow{n\to\infty} 0$$

denn es gilt für alle $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$:

$$E\left[\min\left\{X^2, \frac{|X|^3}{\sqrt{n}}\right\}\right] \xrightarrow{n\to\infty} 0$$

Dies gilt, denn $0 \leq X^2 - \min\left\{X^2, \frac{|X|^3}{\sqrt{n}}\right\} \xrightarrow{n\to\infty} X^2$ monoton steigend. Aus dem Satz von der montonen Konvergenz folgt

$$E\left[X^2 - \min\left\{X^2, \frac{|X|^3}{\sqrt{n}}\right\}\right] \xrightarrow{n\to\infty} E[X^2] \qquad \square$$

Um den zentralen Grenzwersatz aus obigem Satz herzuleiten, verwenden wir:

Satz. *Seien $(Z_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen und Z eine weiter Zufallsvariable mit Verteilungsfunktion F. Dann sind äquivalent:*
1) Für jede beschränkte stetige Funktion $f \in \mathcal{C}_b(\mathbb{R})$ gilt:

$$E[f(Z_n)] \xrightarrow{n\to\infty} E[f(Z)]$$

2) Für alle $f \in \mathcal{C}_b^3(\mathbb{R})$ gilt

$$E[f(Z_n)] \xrightarrow{n\to\infty} E[f(Z)]$$

3) Für jedes Intervall $I = [a,b], (a,b], [a,b)$ oder (a,b), so dass F in a und b stetig ist, gilt:

$$P[Z_n \in I] \xrightarrow{n\to\infty} P[Z \in I]$$

Beweis. Wir beweisen hier nur die Implikation 1) $\Rightarrow$ 2). Es sei I ein Intervall wie in 3) und $\varepsilon > 0$. Nach der Voraussetzung über die Grenzen von I gibt es ein offenes Intervall $I_1 \supseteq \overline{I}$ und ein abgeschlossenes Intervall $I_2 \subseteq \mathrm{int}(I)$ mit

$$P[Z \in I_1] - P[Z \in I] < \varepsilon \quad \text{und} \quad P[Z \in I] - P[Z \in I_2] < \varepsilon$$

Wir wählen $f_1, f_2 \colon \mathbb{R} \to [0,1]$ mit $f_1, f_2 \in \mathcal{C}_b^3(\mathbb{R})$ mit

$$1_{I_1} \geq f_1 \geq 1_I \geq f_2 \geq 1_{I_2}$$

Dann gilt:

$$P[Z_n \in I] = E[1_I(Z_n)] \leq E[f_1(Z_n)] \xrightarrow{n\to\infty} E[f_1(Z)] \leq E[1_{I_1}(Z)] = P[Z \in I_1] \leq P[Z \in I] + \varepsilon$$

und

$$P[Z_n \in I] = E[1_I(Z_n)] \geq E[f_2(Z_n)] \xrightarrow{n\to\infty} E[f_2(Z)] \geq E[1_{I_2}(Z)] = P[Z \in I_2] \geq P[Z \in I] - \varepsilon$$

Weil $\varepsilon > 0$ beliebig war, folgt die Behauptung $P[Z_n \in I] \xrightarrow{n\to\infty} P[Z \in I]$. $\square$

Bemerkung. Ist Z $N(0,1)$-verteilt (oder allgemeiner: hat Z eine Dichte), so ist F stetig. Die Einschränkung in 3) auf Stetigkeitspunkte liefert dann keine Einschränkung.

Definition. Sind die äquivalenten Bedingungen 1) bis 3) des Satzes erfüllt, so *konvergiert Z_n in Verteilung gegen Z* oder auch *Z_n konvergiert schwach gegen Z*.

2 Mathematische Statistik

	Wahrscheinlichkeitstheorie	mathematische Statistik
Wahrscheinlichkeitsmaß P	bekannt	unbekannt bis auf einige allgemeine "Rahmenannahmen"
Ergebnis ω des Zufallsexperiments	unbekannt	bekannt (beobachtete Daten)
typische Aufgaben	Berechnung oder Abschätzung von Wahrscheinlichkeiten interessanter Ereignisse	Schätzen von Parametern über die unbekannte Verteilung P oder testen von Hypothesen über die unbekannte Verteilung P

Sowohl Wahrscheinlichkeitstheorie als auch mathematische Statistik beschäftigen sich mit zufälligen Phänomenen. Die Statistik beschäftigt sich mit "inversen Problemen" zur Wahrscheinlichkeitstheorie: Man will Informationen über unbekannte Wahrscheinlichkeitsverteilungen aus Beobachtungsdaten gewinnen. Die Daten werden als beobachtete Werte einer Zufallsvariablen interpretiert.

Definition. Ein *statistisches Modell* (oder auch statistisches Rahmenmodell) ist ein Tripel $(\Omega, \mathcal{A}, \mathcal{P})$, bestehend aus einem Ergebnisraum Ω, einer Ereignis-σ-Algebra $\mathcal{A}$ über Ω und einer *Menge* $\mathcal{P}$ von Wahrscheinlichkeitsmaßen über $(\Omega, \mathcal{A})$, zusammen mit einer Interpretation dieser Komponenten. Die Beobachtungsdaten werden in einem Ergebnis $\omega \in \Omega$ kodiert.

Beispiel. Eine möglicherweise unfaire Münze wird n-mal geworfen. Wir erhalten Beobachtungsdaten $\omega = (\omega_1, \dots, \omega_n) \in \Omega = \{0,1\}^n$. Sei $\mathcal{A} = \mathcal{P}(\Omega)$. Ohne die Beobachtungsdaten ω schon zu kennen, ist es plausibel folgende Rahmenannahmen zu treffne: Unter der unbekannten Wahrscheinlichkeitsverteilung P sind die $\omega_1, \dots, \omega_n$ *unabhängig* voneinander und *identisch verteilt*. Formalisiert bedeutet das: Unser Rahmenmodell verwendet die Klasse von Wahrscheinlichkeitsmaßen:

$$\mathcal{P} = \left\{ \left(p\delta_1 + (1-p)\delta_0\right)^n : 0 \le p \le 1 \right\}$$

Definition. Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell. Wird $\mathcal{P}$ als eine Klasse von Verteilungen P_ϑ mit endlich vielen Parametern $\vartheta = (\vartheta_1, \dots, \vartheta_d) \in \mathbb{R}^d$ gegeben, so heißt $\left(\Omega, \mathcal{A}, (P_\vartheta)_{\vartheta \in \Theta}\right)$, $\Theta \subseteq \mathbb{R}^d$, ein *parametrisches Modell*. Andernfalls — typischerweise für unendlichdimensionale $\mathcal{P}$ — heißt $(\Omega, \mathcal{A}, \mathcal{P})$ ein *nichtparametrisches Modell*.

Beispiel. Sei $\Omega = \{0,1\}^n$, $\mathcal{A} = \mathcal{P}(\Omega)$ und $\mathcal{P} = \{P_p^n : 0 \le p \le 1\}$ mit $P_p = p\delta_1 + (1-p)\delta_0$ ist ein parametrisches Modell mit Parameter $p \in [0,1]$.

Beispiel. Sei $\Omega = \mathbb{R}^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ und

$$\mathcal{P} = \{P^n : P \text{ ist ein Wahrscheinlichkeitsmaß über } \mathbb{R}\}$$

Dann ist $(\Omega, \mathcal{A}, \mathcal{P})$ ein Modell für n i.i.d. Beobachtungen mit Werten in $\mathbb{R}$, über deren Verteilung weiter nichts bekannt ist. Es ist ein nichparametrisches Modell.

2.1 Frequentistische und Bayessche Sicht

Es gibt zwei grundsätzlich verschiedene Herangehensweisen an die Statistik, die *frequentistische Sicht* und die *Bayessche Sicht*.

Frequentistische Sicht Die beobachteten Daten $\omega \in \Omega$ werden als *zufälliges* Ergebnis eines Zufallsexperiments interpretiert. Das zugrundeliegende Wahrscheinlichkeitsmaß P wird als fest, *nicht zufällig*, aber *unbekannt* aufgefasst.

Bayessche Sicht Die Klasse $\mathcal{P}$ der plausiblen Wahrscheinlichkeitsmaße wird selbst mit einer σ-Algebra $\mathbb{A}$ und einem Wahrscheinlichkeitsmaß $\mathbb{P}$ versehen. $\mathbb{P}$ heißt *a priori Verteilung* (engl. "prior distribution", kurz "prior"). Die Beobachtungsdaten $\omega \in \Omega$ werden als Ergebnis eines *zweistufigen* Zufallsexperiments interpretiert: In der 1. Stufe wählt "die Natur" ein Wahrscheinlichkeitsmaß $P \in \mathcal{P}$ im Modell $(\mathcal{P}, \mathbb{A}, \mathbb{P})$, in der 2. Stufe wird das Beobachtungsergebnis $\omega \in \Omega$ zufällig im Modell $(\Omega, \mathcal{A}, P)$ gezogen. Der Statistiker studiert die Verteilung von $P \in \mathcal{P}$, *bedingt auf die Beobachtung* $\omega \in \Omega$. Sie heißt *a posteriori Verteilung* (engl. "posterior distribution").

Erinnerung. Ein Maß μ auf $(\Omega, \mathcal{A})$ heißt *σ-endlich*, wenn es eine Folge A_n, $n \in \mathbb{N}$, in $\mathcal{A}$ mit $A_n \nearrow \Omega$ und $\mu(A_n) < \infty$, $n \in \mathbb{N}$, gibt.

Definition. Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein Rahmenmodell. Ein *dominierendes Maß* auf $\mathcal{P}$ ist ein σ-endliches Maß μ auf $(\Omega, \mathcal{A})$, bezüglich dem alle $P \in \mathcal{P}$ eine Dichte $\frac{dP}{d\mu}$ besitzen. Existiert so ein dominierendes Maß μ, so heißt $(\Omega, \mathcal{A}, P)$ *dominiert*.

Bemerkung. Ist P ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$ und μ ein Maß auf $(\Omega, \mathcal{A})$, so dass P eine Dichte f bezüglich μ besitzt, so ist f bis auf Abänderung auf einer μ-Nullmenge eindeutig bestimmt. Wir schreiben:

$$f = \frac{dP}{d\mu} \ \mu\text{-fast überall}$$

Definition. Sei $(\Omega, \mathcal{A}, (P_\vartheta)_{\vartheta \in \Theta})$ ein parametrisches Modell mit dominierendem Maß μ. Die Abbildung

$$f \colon \Omega \times \Theta \to \mathbb{R}, f(\omega, \vartheta) := \frac{dP_\vartheta}{d\mu}(\omega)$$

heißt *Likelihood-Funktion*. Für jedes $\vartheta \in \Theta$ ist $f(\cdot, \vartheta)$ μ-fast überall eindeutig.

Beispiel. Ist Ω endlich oder abzählbar unendlich, $\mathcal{A} = \mathcal{P}(\Omega)$ und μ das Zählmaß, so ist die Likelihood-Funktion gegeben durch

$$f \colon \Omega \times \Theta \to \mathbb{R}, f(\omega, \vartheta) = P_\vartheta(\{\omega\})$$

Im Münzwurfmodell von vorhin bedeutet das:

$$f \colon \{0,1\}^n \times [0,1] \to \mathbb{R}, (\omega, p) \mapsto p^{S(\omega)}(1-p)^{n-S(\omega)}, \quad \text{wobei } S \colon \{0,1\}^n \to \mathbb{N}_0, (\omega_1, \ldots, \omega_n) \mapsto \sum_{i=1}^n \omega_i$$

Die Likelihood-Funktion kodiert also alle $(P_\vartheta)_{\vartheta \in \Theta}$ in einer einzigen Funktion.

Definition. Für $P, Q \in \mathcal{P}$, so dass P eine Dichte $\frac{dP}{dQ}$ bezüglich Q besitzt, heißt diese auch der *Likelihood-Quotient*. In der Tat ist der Likelihood-Quotient der Quotient der Likelihood-Funktionen:

$$\frac{dP_{\vartheta_1}}{dP_{\vartheta_2}}(\omega) = \frac{\frac{dP_{\vartheta_1}}{d\mu}(\omega)}{\frac{dP_{\vartheta_2}}{d\mu}(\omega)} = \frac{f(\omega, \vartheta_1)}{f(\omega, \vartheta_2)} \quad P_{\vartheta_2}\text{-fast überall}$$

wann immer dies definiert ist.

2.2 Grundbegriffe der Schätztheorie

Definition. Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell. Ein *Parameter* ist eine Abbildung $\vartheta\colon \mathcal{P} \to \mathbb{R}^n$. Ein Schätzer für einen Parameter ϑ ist eine $\mathcal{A}$-$\mathcal{B}(\mathbb{R}^n)$-messbare Abbildung $\hat{\vartheta}\colon \Omega \to \mathbb{R}^n$.

Bemerkung. Die Definition des Schätzers sagt nichts darüber, ob ein Schätzer "gut" oder "schlecht" ist. Für eine Beobachtung $\omega \in \Omega$ bei zugrundeliegender Verteilung $P \in \mathcal{P}$ heißt $\hat{\vartheta}(\omega) - \vartheta(P)$ der *Schätzfehler*. Ein mögliches Kriterium zur Beurteilung von Schätzern ist:

Definition. Es sei $\vartheta\colon \mathcal{P} \to \mathbb{R}$ ein Parameter. Ein Schätzer $\hat{\vartheta}\colon \Omega \to \mathbb{R}$ heißt *erwartungstreu* (oder *unverfälscht*, engl. *unbiased*), wenn für *alle* $P \in \mathcal{P}$ gilt: $E_P[\hat{\vartheta}]$ existiert und es gilt

$$E_P[\hat{\vartheta}] = \vartheta(P)$$

Anders gesagt:

$$\forall P \in \mathcal{P}.\ E_P[\hat{\vartheta} - \vartheta(P)] = 0$$

Beispiel. Es seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit unbekannter Verteilung P und existierender, aber unbekannter Erwartung $\mu(P) = E_P[X_i]$. Wir beschreiben das mit dem nichtparametrischen Modell $(\Omega, \mathcal{A}, P)$, wobei $\Omega = \mathbb{R}^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ und

$$\mathcal{P} = \{P^n\colon P \text{ ist Wahrscheinlichkeitsmaß über } (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ mit endlicher Erwartung } \mu(P)\}$$

Weiter sei $X_i(\omega_1, \ldots, \omega_n) = \omega_i$. Das Stichprobenmittel

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\colon \Omega \to \mathbb{R}$$

ist ein erwartungstreuer Schätzer für μ, denn für alle $P^n \in \mathcal{P}$ gilt

$$E_{P^n}[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n} E_{P^n}[X_i] = \mu(P)$$

Jedes X_i ist ebenfalls ein erwartungstreuer Schätzer.

Beispiel. Nehmen wir noch zusätzlich an, dass die $X_1, \ldots, X_n$ bezüglich P^n eine endliche Varianz $\sigma^2(P)$ besitzen:

$$\mathcal{P} = \{P^n\colon P \text{ ist Wahrscheinlichkeitsmaß über } (\mathbb{R}, \mathcal{B}(R)) \text{ mit endlicher Varianz } \sigma^2(P)\}$$

Die *empirische Varianz* der Stichprobe $X_1, \ldots, X_n$ wird definiert durch

$$s_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Sie ist ein erwartungstreuer Schätzer für $\sigma^2(P)$, aber der (vielleicht naheliegendere) Schätzer $\frac{n-1}{n}s_X^2$ ist *nicht* erwartungstreu, denn sei $P^n \in \mathcal{P}$. Wegen $E_{P^n}[X_i - \overline{X}] = 0$ ist $E_{P^n}[(X_i - \overline{X})^2] = \mathrm{Var}_{P^n}(X_i - \overline{X})$. Nun gilt

$$\mathrm{Var}_{P^n}(X_i - \overline{X}) = \mathrm{Var}_{P^n}\left(\left(1 - \frac{1}{n}\right)X_i - \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_j\right) = \left(1 - \frac{1}{n}\right)^2 \mathrm{Var}_{P^n}(X_i) + \frac{1}{n^2}\sum_{\substack{j=1\\ j\neq i}}^{n} \mathrm{Var}_{P^n}(X_j) =$$

$$= \left(\left(1 - \frac{1}{n}\right)^2 + \frac{n-1}{n^2}\right)\sigma^2(P) = \frac{n-1}{n}\sigma^2(P)$$

Es folgt:

$$E_{P^n}[s_X^2] = \frac{1}{n-1}\sum_{i=1}^{n} E_P[(X_i - \overline{X})^2] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2(P) = \sigma^2(P)$$

Bemerkung. $s_X = \sqrt{s_X^2}$ heißt *empirische Standardabweichung*. Sie ist *kein* erwartungstreuer Schätzer für die echte Standardabweichung $\sigma(P) = \sqrt{\sigma^2(P)}$. Es gibt *keinen* erwartungstreuen Schätzer für die Standardabweichung $\sigma(P)$.

In der Praxis will man oft keine erwartungstreuen Schätzer (z.B. Sicherheitsabstand bei der Schätzung eines Bremswegs oder bei der Länge eines Transatlantikkabels). Ein weiteres (asymptotisches) Kriterium zur Beurteilung von Schätzern: Wir betrachten eine i.i.d. Stichprobe $X_1, \dots, X_n$ mit Werten in Ω und unbekannter Verteilung P, formal betrachten wir ein Produkt-Rahmenmodell $(\Omega^n, \mathcal{A}^{\otimes n}, \mathcal{P}_n)$, $n \in \mathbb{N}$, mit einem statistischen Modell der Einzelbeobachtungen $(\Omega, \mathcal{A}, \mathcal{P})$ und $\mathcal{P}_n = \{P^n \colon P \in \mathcal{P}\}$.

Definition. Eine Folge $\hat{\vartheta}_n \colon \Omega^n \to \mathbb{R}$, $n \in \mathbb{N}$, von Schätzern für einen Parameter $\vartheta \colon \mathcal{P} \to \mathbb{R}$ heißt *konsistent*, wenn für alle $P \in \mathcal{P}$ und alle $\varepsilon > 0$ gilt

$$\lim_{n\to\infty} P^n\big[|\hat{\vartheta}_n - \vartheta(P)| > \varepsilon\big] = 0$$

bzw.

$$\hat{\vartheta}_n \xrightarrow{n\to\infty} \vartheta(P) \quad \text{in Wahrscheinlichkeit bezüglich } P^n.$$

Beispiel. Das Stichprobenmittel $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $n \in \mathbb{N}$, ist eine konsistente Folge von Schätzern für die Erwartung $E_{P^n}[X_i] = \mu(P)$ (nicht jedoch ein Stichprobenwert, z.B. X_1). Das folgt aus dem schwachen Gesetz der großen Zahlen.

2.2.1 Maximum-Likelihood-Schätzer

Definition. Sei $(\Omega, \mathcal{A}, (P_\vartheta)_{\vartheta\in\Theta})$ ein parametrisches statistisches Modell mit dominierendem Maß μ und Likelihood-Funktion $f \colon \Omega \times \Theta \to \mathbb{R}$. Der *Maximum-Likelihood-Schätzer* für den Parameter $\vartheta \in \Theta$ wird wie folgt definiert:

$$\hat{\vartheta} = \hat{\vartheta}_{ML} \colon \Omega \to \Theta, \omega \mapsto \arg\max_{\vartheta\in\Theta} f(\omega, \vartheta)$$

Beispiel. Sei $\Omega = \{0, 1, \dots, n\}$, $\mathcal{A} = \mathcal{P}(\Omega)$, $\mathcal{P} = \big\{\text{binomial}(n, p) \colon p \in [0, 1]\big\}$, $\Theta = [0, 1]$ und $P_p = \text{binomial}(n, p)$. Das dominierende Maß μ sei das Zählmaß auf Ω. Es gilt also für die Likelihood-Funktion:

$$f \colon \Omega \times [0, 1] \to \mathbb{R}, (\omega, p) \mapsto \binom{n}{\omega} p^\omega (1 - p)^{n-\omega}$$

Gegeben eine Beobachtung $\omega \in \Omega$ maximieren wir $f(\omega, p)$ in p. Wir rechnen

$$\frac{\partial}{\partial p} \log f(\omega, p) = \frac{\partial}{\partial p}\big(\omega \log p + (n - \omega) \log(1 - p)\big) = \frac{\omega}{p} - \frac{n - \omega}{1 - p},$$

was für $0 < p < 1$ monoton fällt. Die Ableitung wird 0 an der Stelle $\hat{p}$ mit $\frac{\omega}{\hat{p}} = \frac{n-\omega}{1-\hat{p}}$, also bei $\hat{p}(\omega) = \frac{\omega}{n}$, falls $\omega \in \{1, \dots, n-1\}$, und auch für alle $\omega \in \{0, \dots, n\}$ ist $\hat{p}(\omega) = \frac{\omega}{n}$ die Stelle, an der die Likelihood-Funktion maximal wird. Der Maximum-Likelihood-Schätzer für p im Münzwurfmodell ist also gleich der relativen Häufigkeit von "1" in den Beobachtungen.

Bemerkung. In Produktmodellen $\mathcal{P}_n = \{P_\vartheta^n \colon \vartheta \in \Theta\}$ mit einer Likelihood-Funktion

$$f_n(\omega_1, \dots, \omega_n; \vartheta) = \prod_{i=1}^n f(\omega_i, \vartheta)$$

ist es rechnerisch meist einfacher statt f direkt den Logarithmus von f_n zu maximieren, denn

$$\log f_n(\omega_1, \dots, \omega_n; \vartheta) = \sum_{i=1}^{n} \log f(\omega_i, \vartheta)$$

$\log f_n$ heißt *Log-Likelihood-Funktion.*

2.2.2 Momentenschätzer

Es sei $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (P_\vartheta^n)_{\vartheta\in\Theta})$ ein parametrisches Modell mit k Parametern aus $\Theta \subseteq \mathbb{R}^k$. Ein einfache Methode zur Gewinnung von Schätzern für ϑ ist das Schätzen der ersten k Momente (oder zentrierten Momente) von P_ϑ, z.B. für $k = 1$ das Mittel der Stichprobe als Schätzer für die Erwartung, für $k = 2$ das Mittel und die empirische Varianz als Schätzer für die Erwartung und Varianz, mit anschließender Wahl des $\hat{\vartheta} \in \Theta$, für das die ersten k Momente von $P_{\hat{\vartheta}}$ mit den Schätzwerten übereinstimmen.

Beispiel. Seien $X_1, \dots, X_n$ i.i.d. normal(μ, σ^2)-verteilt, μ und σ^2 unbekannt. Der Momentenschätzer für die Parameter (μ, σ^2) ist $(\overline{X}, s_X^2)$.

2.3 Regression

2.3.1 Lineare Gleichungssysteme mit zufällig gestörter rechter Seite

Wir betrachten ein lineares Gleichungssystem $Ax = b$ mit $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$ und $b \in \mathbb{R}^m$. Das Gleichungssystem sei überbestimmt, d.h. $m > n$. Wir nehmen an, dass A bekannt ist und $\ker A = \{x \in \mathbb{R}^n \colon Ax = 0\} = 0$ gilt. Das System $Ax = b$ habe also höchstens eine Lösung x. Die rechte Seite b sei nicht bekannt, sondern nur eine zufällige Störung β davon. Wir machen die Modellannahmen, dass die Komponenten $\beta_1, \dots, \beta_m$ von β unabhängig voneinander sind, und dass die β_i normalverteilt sind mit Erwartung b_i, $b = (b_1, \dots, b_m)$, und unbekannter Varianz σ^2 (für alle i gleich). Gesucht ist eine Schätzung $\hat{x}$ der unbekannten Lösung des Gleichungssystems $Ax = b$ mit nicht genau bekannter rechten Seite b. Weiter suchen wir eine Schätzung $\hat{\sigma}^2$ für die Varianz σ^2.

Wir betrachten folgendes statistisches Modell: Sei $\Omega = \mathbb{R}^m \ni \beta$, $\mathcal{A} = \mathcal{B}(\Omega)$, $\Theta = \mathbb{R}^n \times (0, \infty) \ni (x, \sigma^2)$ und $\mathcal{P} = \{P_{x,\sigma^2} \colon (x, \sigma^2) \in \Theta\}$, wobei $P_{x,\sigma^2} = \text{normal}(Ax, \sigma^2 I_m)$. Das Modell besitzt folgende Likelihood-Funktion:

$$L \colon \Omega \times \Theta \to \mathbb{R}, L(\beta; x, \sigma^2) = (2\pi)^{-\frac{m}{2}} \sigma^{-m} \exp\left(-\frac{1}{2\sigma^2} \|Ax - \beta\|_2^2\right)$$

Gegeben β, suchen wir eine Schätzung $(\hat{x}, \hat{\sigma}^2)$ für den unbekannten Parameter (x, σ^2). Hierzu verwenden wir einen Maximum-Likelihood-Schätzer. Wir maximieren zunächst $L(\beta; x, \sigma^2)$ über x bei festgehaltenem β und σ^2. Hierzu muss $\|Ax - \beta\|_2^2$ möglichst klein sein ("Methode der kleinsten Quadrate" von Gauß).

Behauptung. Ist $\hat{b} = A\hat{x}$ die orthogonale Projektion von β auf den Raum $\operatorname{im} A = \{Ax \colon x \in \mathbb{R}^n\}$, so erfüllt $\hat{x}$ das Ziel

$$\|A\hat{x} - \beta\|_2^2 \le \|Ax - \beta\|_2^2, \quad \text{für alle } x \in \mathbb{R}^n.$$

Beweis. Ist $\langle Ax, A\hat{x} - \beta\rangle = 0$ für alle $x \in \mathbb{R}^n$, so folgt

$$\|Ax - \beta\|_2^2 = \|Ax - A\hat{x} + A\hat{x} - \beta\|_2^2 = \|Ax - A\hat{x}\|_2^2 + 2\langle \underbrace{Ax - A\hat{x}}_{A(x-\hat{x})\in\operatorname{im} A}, A\hat{x} - \beta\rangle + \|A\hat{x} - \beta\|_2^2 =$$

$$= \|Ax - A\hat{x}\|_2^2 + \|A\hat{x} - \beta\|_2^2 \ge \|A\hat{x} - \beta\|_2^2 \qquad \square$$

Die Gleichung $\langle Ax, A\hat{x} - \beta\rangle = 0$ ist äquivalent zu $\langle x, A^T(A\hat{x} - \beta)\rangle = 0$ für alle $x \in \mathbb{R}^n$, also zu $A^T A\hat{x} = A^T\beta$. Wegen $\ker A = 0$ ist $A^T A \in \mathrm{GL}_n(\mathbb{R})$ und es folgt

$$\hat{x} = (A^T A)^{-1} A^T \beta.$$

Damit ist der "Residuenvektor" $A\hat{x} - \beta$ und das "Residuum" $\|A\hat{x} - \beta\|_2^2 =: r^2$ bekannt. Der Wert des Schätzers $\hat{x}$ hängt also nicht von dem Wert von σ^2 ab. Nun maximieren wir $L(\beta; \hat{x}, \sigma^2)$ über σ^2, gegeben β und $\hat{x}$. Es gilt

$$\log L(\beta; \hat{x}, \sigma^2) = -\frac{m}{2}\log(2\pi) - m\log\sigma - \frac{1}{2\sigma^2}r^2.$$

Für $\sigma^2 \to 0$ und $\sigma^2 \to \infty$ erhalten wir $\log L(\beta; \hat{x}, \sigma^2) \to -\infty$, so dass wir die Ränder beim Maximieren von $\log L$ nicht berücksichtigen müssen. Es gilt:

$$\frac{\partial}{\partial\sigma}\log L(\beta; \hat{x}, \sigma^2) = -\frac{m}{\sigma} + \frac{r^2}{\sigma^3},$$

was eine Nullstelle bei $\hat{\sigma}^2 = \frac{r^2}{m}$ besitzt. $(\hat{x}, \hat{\sigma}^2)$ ist der gesuchte Maximum-Likelihood-Schätzer.

Anwendung (Regressionsgeraden)**.** Wir wenden diese Theorie an, um gegebene Messpunkte (x_i, y_i), $i = 1, \ldots, n$, "möglichst gut" durch eine Gerade anzunähern. Hierzu stellen wir uns $x_1, \ldots, x_n$ als fest und bekannt vor, die $y_1, \ldots, y_n$ jedoch auch als bekannt, aber zufällig, und zwar unabhängig voneinander und y_i normal$(ax_i + b, \sigma^2)$-verteilt, $i = 1, \ldots, n$, mit unbekannten Parametern a, b und σ^2. Wir betrachten folgendes statistische Modell: Es seien $\Omega = \mathbb{R}^n \ni (y_1, \ldots, y_n)$, $\mathcal{A} = \mathcal{B}(\Omega)$, $\Theta = \mathbb{R}^2 \times \mathbb{R}^+ \ni (a, b, \sigma^2)$ und $\mathcal{P} = \{P_{a,b,\sigma^2} \colon (a, b, \sigma^2) \in \Theta\}$, wobei $P_{a,b,\sigma^2} = \prod_{i=1}^n \mathrm{normal}(ax_i + b, \sigma^2)$. Setze zur Abkürzung $\tilde{y}_i = ax_i + b$. Das Gleichungssystem

$$ax_i + b = \tilde{y}_i, \quad i = 1, \ldots, n$$

mit der Störung y_i von $\tilde{y}_i$ führt uns auf die Theorie von vorher zurück:

$$\underbrace{\begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}}_{A} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{pmatrix}$$

mit der Störung $(y_1, \ldots, y_n)^T$ der unbekannten $(\tilde{y}_1, \ldots, \tilde{y}_n)^T$. Nach der Regressionstheorie erhalten wir die Schätzung $(\hat{a}, \hat{b})^T$ von $(a, b)^T$ wie folgt:

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (A^T A)^{-1} A^T \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Nun ist

$$A^T A = \begin{pmatrix} x_1 & \ldots & x_n \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} = n \begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}$$

und

$$A^T \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & \ldots & x_n \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix} = n \begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix}.$$

Es folgt

$$\begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix} = \begin{pmatrix}\overline{x^2} & \overline{x}\\ \overline{x} & 1\end{pmatrix}^{-1}\begin{pmatrix}\overline{xy}\\ \overline{y}\end{pmatrix} = \frac{1}{\overline{x^2}-\overline{x}^2}\begin{pmatrix}\overline{xy}-\overline{x}\ \overline{y}\\ \overline{x^2}\ \overline{y}-\overline{x}\ \overline{xy}\end{pmatrix}$$

Die Gerade $\{(x, \hat{a}x+\hat{b})\colon x\in\mathbb{R}\}$ heißt *Regressionsgerade* zu den Datenpunkten (x_i, y_i), $i=1,\dots,n$. Wir erhalten das Residuum

$$r^2 = \sum_{i=1}^{n}(\hat{a}x_i+\hat{b}-y_i)^2$$

und den Maximum-Likelihood-Schätzer für σ^2

$$\hat{\sigma}^2 = \frac{r^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(\hat{a}x_i+\hat{b}-y_i)^2.$$

2.4 Einführung in die Testtheorie

Beispiel. Im Umkreis von 5 km von Kernkraftwerken wohnten in den letzten Daten n_1 Kinder, in einer Kontrollgruppe n_2 Kinder. In der 1. Gruppe erkrankten ω_1 Kinder an Leukämie, in der Kontrollgruppe ω_2 Kinder. Wann "belegen" diese Daten, dass Kinder im Umkreis von Kernkraftwerken mit höhere Wahrscheinlichkeit an Leukämie erkranken?

Wir betrachten folgendes — stark vereinfachtes — statistisches Modell: ω_1 sei eine binomial(n_1,p_1)-verteilte Zufallsvariable, wobei p_1 unbekannt sei. ω_2 sei eine binomial(n_2,p_2)-verteilte Zufallsvariable, auch p_2 sei unbekannt. Weiterhin seien ω_1 und ω_2 unabhängig voneinander. Formaler betrachten wir ein Modell $(\Omega,\mathcal{A},\mathcal{P})$ mit

$$\begin{aligned}&\Omega = \{0,\dots,n_1\}\times\{0,\dots,n_2\} \ni (\omega_1,\omega_2)\\ &\mathcal{A} = \mathcal{P}(\Omega)\\ &\mathcal{P} = \{P_{p_1,p_2}\colon p_1,p_2\in[0,1]\} \quad \text{mit } P_{p_1,p_2} = \text{binomial}(n_1,p_1)\times\text{binomial}(n_2,p_2).\end{aligned}$$

Wir fragen uns, ob wir die Hypothese $p_1 = p_2$ aufgrund der beobachteten Daten $(\omega_1,\omega_2)\in\Omega$ verwerfen können, und zwar in "Richtung" der Alternativhypothese $p_1 > p_2$.

Definition. Sei $(\Omega,\mathcal{A},\mathcal{P})$ ein statistisches Modell. Eine *Hypothese* ist eine Teilmenge $\emptyset\neq H\subseteq\mathcal{P}$. Beim Testen treten zwei Hypothesen auf: eine Nullhypothese $H_0\subseteq\mathcal{P}$ und eine davon disjunkte Alternativhypothese $H_1\subseteq\mathcal{P}$ (kurz: "Alternative").

Beispiel. Im obigen Beispiel ist die Nullhypothese

$$H_0 = \{P_{p,p}\colon 0\le p\le 1\}$$

und die Alternative

$$H_1 = \{P_{p_1,p_2}\colon p_1 > p_2\}.$$

Nullhypothese und Alternative haben verschiedene Rollen. Die Nullhypothese beschreibt ein "einfaches Erklärungsmodell" oder "*Abwesenheit* eines Effekts". Das Vorliegen eines Effekts statistisch zu belegen bedeutet also, die Nullhypothese zu verwerfen. Die Alternative dient oft nur dazu, die Typen von Effekten, für die man sich interessiert, zu spezifizieren und die Qualität eines Tests zu messen.

Definition. Ein (nichtrandomisierter) *statistischer Test* für die Nullhypothese H_0 und die Alternative H_1 wird durch einen *Verwerfungsbereich* $V\in\mathcal{A}$ gegeben. Liegen die Beobachtungsdaten ω in V, dann *verwerfen* wir die Nullhypothese, andernfalls, $\omega\notin V$, *verwerfen* sie *nicht*.

Bemerkung. Die Situation zwischen "verwerfen" und "nicht verwerfen" ist unsymmetrisch: Wenn wir nicht verwerfen, bedeutet das *nicht*, dass H_0 richtig ist, sondern nur eine Art "Stimmenthaltung": die Daten reichen nicht aus, um die "einfache Erklärung" H_0 zu widerlegen, oder die "Anwesenheit eines Effekts" zu belegen.

Ein randomisierter statistischer Test ist ein Test, dessen Entscheidung nicht nur von den Beobachtungsdaten $\omega \in \Omega$, sondern zusätzlich noch von einem Hilfs-Zufallsexperiment, z.B. einer unif$[0,1]$-verteilten Zufallszahl, abhängt. Formaler:

Definition. Ein *randomisierter statistischer Test* zum Modell $(\Omega, \mathcal{A}, \mathcal{P})$ besteht aus einem nicht-randomisierten Test im Modell $(\Omega', \mathcal{A}', \mathcal{P}')$ mit $\Omega' = \Omega \times [0,1]$, $\mathcal{A}' = \mathcal{A} \otimes \mathcal{B}[0,1]$ und $\mathcal{P}' = \{P \otimes \text{unif}[0,1] \colon P \in \mathcal{P}\}$.

2.4.1 Typen von Fehlern

	H_0 wahr	H_0 falsch
H_0 nicht verwerfen	richtige Entscheidung	Fehler 2. Art
H_0 verwerfen	Fehler 1. Art	richtige Entscheidung

Beispiel. Ein Feuermelder soll die Hypothese "es brennt nicht" testen. Ein Fehler 1. Art liegt bei einem Fehlalarm vor. Ein Fehler 2. Art liegt vor, wenn der Brandmelder trotz Feuer nicht Alarm schlägt.

2.4.2 Ziele für gute Tests

Beim Entwurf eines guten Tests steht man vor den konträren Zielen, beide Fehlertypen möglichst zu vermeiden.

- Einerseits soll für alle $P_0 \in H_0$ die Wahrscheinlichkeit $P_0(V)$ möglichst *klein* sein.
- Andererseits soll für alle $P_1 \in H_1$ die Wahrscheinlichkeit $P_1(V)$ möglichst *groß* sein.

Definition. Das *Risiko 1. Art* $\underline{\alpha}$ ist die Wahrscheinlichkeit für den *Fehler 1. Art* unter $P_0 \in H_0$: $\underline{\alpha} = P_0(V)$ (kann von P_0 abhängen, falls H_0 mehr als nur ein Wahrscheinlichkeitsmaß enthält). Das *Risiko 2. Art* β ist die Wahrscheinlichkeit für den *Fehler 2. Art* unter der Alternative: $\beta = P_1(V^c) = 1 - P_1(V)$ mit $P_1 \in H_1$. Die *Macht* eines Tests ist $1 - \beta = P_1(V)$. Das *Signifikanzniveau* α ist das Supremum der Risiken 1. Art:

$$\alpha = \sup_{P_0 \in H_0} P_0(V)$$

Bemerkung. Dies Begriffe werden besonders einfach, wenn H_0 und H_1 einelementig sind.

Definition. Eine Hypothese $H \subseteq \mathcal{P}$ heißt *einfach*, wenn sie genau ein Element $P \in \mathcal{P}$ enthält, andernfalls *zusammengesetzt*. Für einfache $H_0 = \{P_0\} \subseteq \mathcal{P}$ ist also das Signifikanzniveau gleich dem Risiko erster Art $P_0(V)$.

Beispiel. Im obigen Beispiel ist $H_0 = \{P_{p,p} \colon 0 \le p \le 1\}$ zusammengesetzt und $H_0' = \{P_{10^{-6},10^{-6}}\}$ einfach.

Interpretation (Philosophische Interpretation des Testproblems)**.** Die Minimalinterpration von Wahrscheinlichkeiten — Wahrscheinlichkeit nahe 1 bedeutet, das Ereignis ist "praktisch sicher", Wahrscheinlichkeit nahe 0 bedeutet, das Ereignis ist "praktisch unmöglich", andere Wahrscheinlichkeiten bedeuten keine Aussage — passt genau zum Testproblem: Man wählt das Signifikanzniveau α so klein, dass ein Fehler 1. Art "praktisch unmöglich" wird. (In der Praxis heißt oft "$\alpha = 5\%$" "praktisch unmöglich"). Interpretation des Testentscheids: "H_0 verwerfen" bedeutet "Es ist praktisch unmöglich, dass H_0 die Daten beschreibt", "H_0 nicht verwerfen" bedeutet "Stimmenthaltung".

2.4.3 Optimale Tests bei einfachen Hypothesen

Wir betrachten ein Modell $(\Omega, \mathcal{A}, \mathcal{P})$ mit der Nullhypothese $H_0 = \{P_0\} \subseteq \mathcal{P}$ und der Alternative $H_1 = \{P_1\} \subseteq \mathcal{P}$ mit einem Likelihoodquotienen $\frac{dP_1}{dP_0}$. Wir geben uns eine Schranke α_{crit} für das Signifikanzniveau vor und stellen folgendes Optimierungsproblem: Unter allen Tests mit Signifikanzniveau $\alpha = P_0(V) \leq \alpha_{\text{crit}}$ finde man den/die Tests mit möglichst großer Macht $1 - \beta = P_1(V)$.

Bemerkung. Man hat eine Analogie zum "Rucksackproblem": Wir sollen Gegenstände $\omega_1, \ldots, \omega_n$ in einen Rucksack packen. Jeder Gegenstand ω_i hat ein Gewicht $P_{0,i}$ und einen Wert $P_{1,i}$. Wir können maximal das Gewicht α_{crit} tragen. Wie sollen wir den Rucksack bepacken, damit er möglichst großen Wert trägt, aber das Gewicht α_{crit} nicht überschreitet.

Testproblem	Rucksackproblem
Menge der möglichen Ergebnisse Ω	Menge der Güter
Verwerfungsbereich $V \subseteq \Omega$	Menge der Güter, die wir einpacken
Signifikanzniveau $\alpha = P_0(V)$	Gewicht der eingepackten Güter
$P_{0,i} = P_0(\{\omega_i\})$	Gewicht des Guts ω_i
$P_{1,i} = P_1(\{\omega_i\})$	Wert des Guts ω_i
Macht $1 - \beta = P_1(V)$	Wert der eingepackten Güter
$\frac{dP_1}{dP_0}(\omega)$	spezifischer Wert des Guts ω

Beispiel. Nehmen Sie an, Sie können 11.1 kg tragen und Sie haben folgende Güter zur Auswahl:

Gut	Wert	Gewicht	spezifischer Wert
Gold	1000 €	0.1 kg	10000 €/kg
Silber	500 €	1 kg	500 €/kg
Eisen	100 €	10 kg	10 €/kg
Steine	200 €	10000 kg	0.02 €/kg

Jedes Kind weiß: Man nimmt zuerst das Gold, kann man dann noch mehr tragen, dann das Silber, kann man dann noch mehr tragen, das Eisen und Steine nur dann, wenn der Rucksack dann immer noch nicht voll gepackt ist.

Lemma (Neyman-Pearson-Lemma im diskreten Fall)**.** *Sei $\Omega = \{\omega_1, \ldots, \omega_n\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ mit der Nullhypothese $H_0 = \{P_0\}$ und der Alternative $H_1 = \{P_1\}$, wobei*

$$P_0 = \sum_{i=1}^{n} p_{0,i}\delta_{\omega_i} \qquad\qquad P_1 = \sum_{i=1}^{n} p_{1,i}\delta_{\omega_i}$$

Die ω_i seien nach absteigendem Likelihoodquotienten $\frac{dP_1}{dP_0}(\omega_i) = \frac{p_{1,i}}{p_{0,i}}$ angeordnet:

$$\frac{p_{1,1}}{p_{0,1}} \geq \frac{p_{1,2}}{p_{0,2}} \geq \cdots \geq \frac{p_{1,n}}{p_{2,n}}$$

Es sei $1 \leq k \leq n$ und T der Test mit Verwerfungsbereich $V = \{\omega_1, \ldots, \omega_k\}$. Dann gilt für jeden Test T' mit Verwerfungsbereich $V' \subseteq \Omega$: Aus $P_0(V') \leq P_0(V)$ folgt $P_1(V') \leq P_1(V)$. Anders gesagt ist T optimal im folgenden Sinn: Jeder Test T' mit dem gleichen oder höchstens kleinerem Signifikanzniveau wie T hat eine kleinere oder höchstens die gleiche Macht.

Lemma (Neyman-Pearson-Lemma)**.** *Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell, $H_0 = \{P_0\} \subseteq \mathcal{P}$, $H_1 = \{P_1\} \subseteq \mathcal{P}$ zwei einfache Hypothesen mit Likelihoodquotienten $\frac{dP_1}{dP_0}$. Für den Verwerfungsbereich V eines Tests T gelte*

$$\left\{\frac{dP_1}{dP_0} > c\right\} \subseteq V \subseteq \left\{\frac{dP_1}{dP_0} \geq c\right\}$$

für ein $c \geq 0$. Dann ist T im folgenden Sinn optimal: Jeder weitere Test T' mit Verwerfungsbereich V' mit dem gleichen oder höchstens kleinerem Signifikanzniveau $P_0(V') \leq P_0(V)$ hat kleinere oder höchstens gleiche Macht $P_1(V') \leq P_1(V)$.

Beweis. Aus $P_0(V') \leq P_0(V)$ schließen wir

$$P_0(V \setminus V') - P_0(V' \setminus V) = P_0(V) - P_0(V') \geq 0.$$

Nun gilt

$$P_1(V \setminus V') = \int_{V \setminus V'} \mathrm{d}P_1 = \int_{V \setminus V'} \frac{dP_1}{dP_0}\,\mathrm{d}P_0 \geq \int_{V \setminus V'} c\,\mathrm{d}P_0 = cP_0(V \setminus V')$$

und analog

$$P_1(V' \setminus V) = \int_{V' \setminus V} \mathrm{d}P_1 = \int_{V' \setminus V} \frac{dP_1}{dP_0}\,\mathrm{d}P_0 \leq \int_{V' \setminus V} c\,\mathrm{d}P_0 = cP_0(V' \setminus V).$$

Es folgt

$$P_1(V) - P_1(V') = P_1(V \setminus V') - P_1(V' \setminus V) \geq cP_0(V \setminus V') - cP_0(V' \setminus V) \geq 0 \qquad \square$$

Bemerkung. Hat der Likelihoodquotient eine kontinuierliche Verteilung unter P_0, so gilt $P_0\left[\frac{dP_1}{dP_0} = c\right] = 0$ und damit auch $P_1\left[\frac{dP_1}{dP_0} = c\right] = 0$. Dann spielt es keine Rolle, ob man $V = \left\{\frac{dP_1}{dP_0} > c\right\}$ oder $V = \left\{\frac{dP_1}{dP_0} \geq c\right\}$ oder etwas dazwischen wählt. Bei diskreten Modellen kann aber $P_0\left[\frac{dP_1}{dP_0} = c\right] > 0$ sein. Um ein gegebenes Signifikanzniveau zu treffen und maximale Macht zu erreichen, kann es sinnvoll sein, V echt zwischen $V = \left\{\frac{dP_1}{dP_0} > c\right\}$ und $V = \left\{\frac{dP_1}{dP_0} \geq c\right\}$ zu wählen und notfalls zu randomisieren.

Bemerkung. In der Praxis verwendet man oft den Kehrwert $\frac{dP_0}{dP_1}$ statt $\frac{dP_1}{dP_0}$. Das dreht nur die "Richtung" um.

Beispiel. Seien $X_1, \ldots, X_n$ i.i.d. normalverteilte Daten mit unbekannter Erwartung μ und bekannter Varianz σ^2. Wir haben also das Rahmenmodell $\Omega = \mathbb{R}^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ und $\mathcal{P} = \{\mathrm{normal}(\mu, 1)^n \colon \mu \in \mathbb{R}\}$ mit den kanonischen Projektionen $X_1, \ldots, X_n \colon \mathbb{R}^n \to \mathbb{R}$. Wir entwerfen einen Test zum Signifikanzniveau α für die Nullhypothese H_0: "$\mu = 0$", also $H_0 = \{\mathrm{normal}(0, 1)^n\} = \{P_0\}$, bei der Alternative H_1: "$\mu = \mu_1$", also $H_1 = \{\mathrm{normal}(\mu_1, 1)^n\} = \{P_1\}$. P_0 hat die Dichte

$$f_0(x) = \frac{dP_0}{d\lambda_n}(x) = \prod_{j=1}^{n} \frac{e^{-\frac{1}{2}x_j^2}}{\sqrt{2\pi}} = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\|x\|_2^2}$$

und ebenso P_1 die Dichte

$$f_1(x) = \frac{dP_1}{d\lambda_n}(x) = \prod_{j=1}^{n} \frac{e^{-\frac{1}{2}(x_j - \mu_1)^2}}{\sqrt{2\pi}} = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}(x_j - \mu_1)^2\right)$$

Damit erhalten wir den Likelihoodquotienten

$$\frac{dP_1}{dP_0}(x) = \frac{f_1(x)}{f_2(x)} = \exp\left(\frac{1}{2}\sum_{j=1}^{n}\left(x_j^2 - (x_j - \mu_1)^2\right)\right) = e^{-\frac{n\mu_1^2}{2}} \exp\left(\mu_1 \sum_{j=1}^{n} x_j\right)$$

Man beachte, dass man nicht alle Datenpunkte $X_1, \ldots, X_n$ kennen muss, um $\frac{dP_1}{dP_0}(x)$ zu berechnen; hier genügt die Summe $S = \sum_{j=1}^{n} X_j$. Wir bestimmen jetzt die Niveaumenge $V = \left\{\frac{dP_1}{dP_0} > c\right\}$, die als

Verwerfungsbereich in einem Neyman-Pearson-Test auftritt. Für den Fall $\mu_1 > 0$ ist $s \mapsto e^{-n\mu_1^2/2} e^{\mu_1 s}$ monoton steigend, also

$$V = \left\{ \frac{dP_1}{dP_0} > c \right\} = \left\{ \sum_{j=1}^{n} x_j > s \right\}$$

für $c = \exp\left(-n\frac{\mu_1^2}{2} + \mu s\right)$. Um ein bestimmtes Signifikanzniveau α zu realisieren, wählen wir s so, dass $P_0[S > s] = \alpha$. Nun folgt aus $\mathcal{L}_{P_0}(X_1, \ldots, X_n) = \text{normal}(0,1)^n$, dass $\mathcal{L}_{P_0}(S) = \text{normal}(0, n)$, also $\mathcal{L}_{P_0}(S/\sqrt{n}) = \text{normal}(0,1)$. Bezeichnen wir mit Φ die Verteilungsfunktion der Standardnormalverteilung, also

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{\frac{x^2}{2}} \, \mathrm{d}x,$$

so folgt

$$P_0[S > s] = P_0[Z > s/\sqrt{n}] = 1 - \Phi(s/\sqrt{n}) \stackrel{!}{=} \alpha$$

mit $Z = \frac{1}{\sqrt{n}} S$. Damit erhalten wir den Verwerfungsbereich

$$V = \left\{ Z > \Phi^{-1}(1-\alpha) \right\}.$$

Im Fall $\mu_1 < 0$ ist $s \mapsto e^{-n\mu_1^2/2 + \mu_1 s}$ monoton fallend, also erhalten wir

$$V = \left\{ \frac{dP_1}{dP_0} > c \right\} = \left\{ \sum_{j=1}^{n} x_j < s \right\} = \left\{ Z < \Phi^{-1}(\alpha) \right\}.$$

Man beachte, dass der genaue Wert von μ_1 irrelevant für den Testentscheid ist, nur das Vorzeichen von μ_1 geht in die Konstruktion ein. Der erhaltene Test ist also *gleichmäßig* ein optimaler Test für *alle* $\mu_1 > 0$ bzw. alle $\mu_1 < 0$.

Definition. Eine Zufallsvariable $T \colon \Omega \to \mathbb{R}$ die den Verwerfungsbereich V_α für jede Wahl des Signifikanzniveaus α bestimmt, heißt *Teststatistik*. Allgemeiner heißt eine von Statistikern gewählte messbare Abbildung $X \colon \Omega \to \mathbb{R}$, die den Daten einen Zahlenwert zuordnet, eine *Statistik*. Im Beispiel ist $Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ unsere Teststatistik, sie ist unter H_0 standardnormalverteilt.

Definition. Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell, so dass alle $P_0, P_1 \in \mathcal{P}$ eine Dichte $\frac{dP_1}{dP_0}$ zueinander haben, z.B. ein dominiertes Modell mit positiver Likelihood-Funktion. Eine Statistik $X \colon \Omega \to \mathbb{R}^d$ heißt *suffizient* für $(\Omega, \mathcal{A}, \mathcal{P})$, wenn es für alle $P_0, P_1 \in \mathcal{P}$ eine messbare Abbildung $f \colon \mathbb{R}^d \to \mathbb{R}$ gibt mit $\frac{dP_1}{dP_0} = f(X)$ P_0-fast sicher, also der Likelihood-Quotient $\frac{dP_1}{dP_0}$ nur von X abhängt.

Bemerkung. Offensichtlich genügt der Wert $X(\omega)$ einer suffizienten Statistik (evtl. zusammen mit einer Randomisierung) zur Ausführung des Neyman-Pearson-Tests.

Lemma. *Es sei $(\Omega, \mathcal{A}, (P_\vartheta)_{\vartheta \in \Theta}$ ein parametrisches statistisches Modell mit dominierendem Maß μ und Likelihood-Funktion $f > 0$. Ist $X \colon \Omega \to \mathbb{R}^d$ eine Statistik und $g \colon \mathbb{R}^d \times \Theta \to \mathbb{R}$, $h \colon \Omega \to \mathbb{R}^+$ mit $f(\omega, \vartheta) = g(X(\omega), \vartheta) h(\omega)$ für alle $\omega \in \Omega$ und $\vartheta \in \Theta$, so ist X suffizient.*

Beweis. Seien $\vartheta_1, \vartheta_2 \in \Theta$. Dann gilt für alle $\omega \in \Omega$:

$$\frac{P_{\vartheta_1}}{P_{\vartheta_2}} = \frac{f(\omega, \vartheta_1)}{f(\omega, \vartheta_2)} = \frac{g(X(\omega), \vartheta_1)}{g(X(\omega), \vartheta_2)}.$$

□

Beispiel. Sei $\mathcal{P} = \{\text{normal}(\mu, \sigma^2)^n \colon \mu \in \mathbb{R}, \sigma^2 > 0\}$. Wir erhalten die Likelihoodfunktion

$$f(x_1, \dots, x_n; \mu, \sigma^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x_j - \mu)^2\Big) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2\Big).$$

Nun ist

$$\sum_{j=1}^{n}(x_j - \mu)^2 = \sum_{j=1}^{n}(x_j - \overline{x})^2 + 2\sum_{j=1}^{n}(x_j - \overline{x})(\overline{x} - \mu) + n(\overline{x} - \mu)^2 = (n-1)s_x^2 + n(\overline{x} - \mu)^2.$$

Es folgt

$$f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2\sigma^2}\Big((n-1)s_x^2 + n(\overline{x} - \mu)^2\Big)\Big).$$

Hier tauchen die Daten $x_1, \dots, x_n$ nur in der Kombination $\overline{X}$ und s_x^2 auf, $(\overline{X}, \sigma_X^2)$ ist also eine suffiziente Statistik.

Bemerkung. Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell und $T \colon \Omega \to \mathbb{R}^d$ eine suffiziente Statistik. Gegeben eine Nullhypothese $H_0 = \{P_0\} \subseteq \mathcal{P}$ und eine Alternative $H_1 = \{P_1\} \subseteq \mathcal{P}$ schreiben wir $\frac{dP_1}{dP_0} = f(T)$. Damit werden alle Tests mit Verwerfungsbereich $V_c = \{T \in f^{-1}([0, c])\}$ oder $V_c = \{T \in f^{-1}([0, c))\}$ optimal in dem Sinn, dass sie maximale Macht bei gegebenem Signifikanzniveau haben.

2.4.4 Variable Signifikanzniveaus und p-Wert

Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell, $H_0 \subseteq \mathcal{P}$ eine Nullhypothese und V_c, $c \in I \subseteq \mathbb{R}$, eine Familie von Verwerfungsbereichen, monoton steigend in c, d.h. $V_c \subseteq V_{c'}$ für $c \leq c'$, z.B. $V_c = \{T \leq c\}$ mit einer Teststatistik T.

Definition. Gegeben Beobachtungsdaten $\omega \in \Omega$ definieren wir den *p-Wert* $p = p(\omega)$ als das kleinste Niveau, auf dem die Hypothese H_0 noch verworfen werden kann. Genauer ist

$$p(\omega) = \inf\{\alpha_c \colon \omega \in V_c\}, \text{ mit } \alpha_c = \sup_{P_0 \in H_0} P_0(V_c).$$

Bemerkung. Gilt $H_0 = \{P_0\}$ und $V_c = \bigcap_{c' > c} V_{c'}$, so wird das Infimum sogar angenommen. Das ist z.B. für $V_c = \{T \leq c\}$ der Fall. In diesem Fall ist $p(\omega) = P_0(V_c)$ mit $c = T(\omega)$.

Beispiel. Seien $X_1, \dots, X_n$ i.i.d. normalverteilt und testen wir $H_0 = \{P_0\}$: $\mathcal{L}_{P_0}(X_1, \dots, X_n) = \text{normal}(\mu_0, \sigma^2)$ gegen die Alternative $H_1 = \{P_1\}$: $\mathcal{L}_{P_1}(X_1, \dots, X_n) = \text{normal}(\mu_1, \sigma^2)$ mit $\mu_1 < \mu_0$, so haben die Tests mit Verwerfungsbereich $V_c = \{Z \leq c\}$, $c \in \mathbb{R}$, mit der Teststatistik $Z = \sqrt{n}\frac{\overline{X} - \mu_0}{\sigma}$. maximale Macht bei gegebenem Niveau. Es gilt $\mathcal{L}_{P_0}(Z) = \text{normal}(0, 1)$. Gegeben Daten $X_1(\omega), \dots, X_n(\omega)$ erhalten wir den p-Wert:

$$p(\omega) = \inf\{P_0[Z \leq c] \colon Z(\omega) \leq c\} = P_0[Z \leq Z(\omega)].$$

Allgemeiner: Gegeben eine Teststatistik $T \colon \Omega \to \mathbb{R}$ mit $Q = \mathcal{L}_{P_0}(T)$, $P_0 \in H_0$, und $V_c = \{T \leq c\}$, $c \in \mathbb{R}$, so wird der p-Wert durch $p(\omega) = P_0[T \leq T(\omega)] = Q((-\infty, T(\omega)])$, $\omega \in \Omega$ und $P_0 \in H_0$, gegeben. Der p-Wert kodiert den Testentscheid bei variablem Niveau: Wenn $p(\omega) < \alpha$, so *verwerfen* wir H_0 zum Niveau α, wenn $p(\omega) > \alpha$, so *verwerfen* wir H_0 *nicht* zum Niveau α.

2.4.5 Konfidenzbereiche und Dualität

Konfidenzbereiche sind eine Art "Parameterschätzung mit Toleranzangabe". Gegeben Beobachtungsdaten $\omega \in \Omega$ möchte man die Menge $C(\omega)$ von "plausiblen" Parametern auszeichnen.

Definition. Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell und $\vartheta\colon \mathcal{P} \to \mathbb{R}^d$ ein Parameter (z.B. im parametrischen Fall $\vartheta(P_q) = q$). Weiter sei $\alpha \in (0,1)$. Eine Familie $(C(\omega))_{\omega\in\Omega}$ von Mengen $C(\omega) \subseteq \mathbb{R}^d$, $\omega \in \Omega$, heißt *Konfidenzbereich* oder *Vertrauensbereich* zum Vertrauensniveau $1-\alpha$, kurz $(1-\alpha)$-Vertrauensbereich, wenn gilt: Für alle $P \in \mathcal{P}$ ist $\{\omega \in \Omega\colon \vartheta(P) \in C(\omega)\} \in \mathcal{A}$ und es gilt

$$P(\{\omega \in \Omega\colon \vartheta(P) \in C(\omega)\}) \geq 1-\alpha.$$

Bemerkung. Man beachte: Hier ist $C(\omega)$ zufällig, d.h. von ω abhängig, aber $P \in \mathcal{P}$ *nicht zufällig*, aber allquantifiziert.

Bemerkung. Vertrauensbereiche zum Vertrauensniveau $1-\alpha$ sind am interessantesten, wenn sie möglichst klein sind. Die triviale Wahl $C(\omega) = \mathbb{R}^d$ ist zwar möglich, aber nutzlos.

Bemerkung. Im Fall $d = 1$ ist $C(\omega)$ oft ein Intervall. Es heißt dann "Konfidenzintervall".

Lemma 1. *Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell, $\vartheta\colon \mathcal{P} \to \Theta \subseteq \mathbb{R}^d$, und $K \subseteq \Omega \times \Theta$, so dass $\{\omega \in \Omega\colon (\omega, \vartheta(P)) \in K\} \in \mathcal{A}$ für alle $P \in \mathcal{P}$. Weiter sei $0 < \alpha < 1$. Dann sind äquivalent:*

1) Durch $C(\omega) = \{q \in \Theta\colon (\omega, q) \in K\}$, $\omega \in \Omega$, wir ein $(1-\alpha)$-Konfidenzbereich gegeben.
2) Für jedes $q \in \Theta$ mit $H_0(q) = \{P \in \mathcal{P}\colon \vartheta(P) = q\}$ ist $V(q) = \{\omega \in \Omega\colon (\omega, q) \notin K\}$ der Verwerfungsbereich eines Tests der Hypothese $H_0(q)$ zu einem Niveau $\leq \alpha$.

Beweis. Es gilt:

$$\begin{aligned}
1) &\iff \forall P \in \mathcal{P}.\ P[\vartheta(P) \in C] \geq 1-\alpha \iff \forall P \in \mathcal{P}.\ P(\{\omega \in \Omega\colon (\omega, \vartheta(P)) \in K\}) \geq 1-\alpha \\
&\iff \forall P \in \mathcal{P}.\ P(\{\omega \in \Omega\colon (\omega, \vartheta(P)) \notin K\}) \leq \alpha \\
&\iff \forall P \in \mathcal{P}.\ P(V(\vartheta(P))) \leq \alpha \iff \forall q \in \Theta\ \forall P_0 \in H_0(q).\ P(V(q)) \leq \alpha \\
&\iff 2)
\end{aligned}$$

□

Beispiel. Seien $X_1, \ldots, X_n$ i.i.d. normalverteilt mit Erwartung μ (unbekannt) und Varianz σ^2 (bekannt), $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Sei $0 < \alpha < 1$. Dann ist für jedes $\mu \in \mathbb{R}$

$$V(\mu) = \left\{ \sqrt{n}\frac{\overline{X} - \mu}{\sigma} \leq \Phi^{-1}(\alpha) \right\}$$

der Verwerfungsbereich zum Niveau α eines Tests der Hypothese $H_0 = \{P \in \mathcal{P}\colon E_P[X_1] = \mu\}$. Also ist

$$C(\omega) = \left\{ \mu \in \mathbb{R}\colon \sqrt{n}\frac{\overline{X}(\omega) - \mu}{\sigma} > \Phi^{-1}(\alpha) \right\} = (-\infty, \overline{X}(\omega) - \tfrac{\sigma}{\sqrt{n}}\Phi^{-1}(\alpha))$$

ein $(1-\alpha)$-Vertrauensbereich für das unbekannte μ.

Beispiel. Seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit einer unbekannten, atomlosen Verteilung P, d.h. $P(\{a\}) = 0$ für alle $a \in \mathbb{R}$. Wir verwenden folgendes Modell: $(\Omega, \mathcal{A}, \mathcal{P})$ mit $\Omega = \mathbb{R}^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ und

$$\mathcal{P} = \{P^n\colon P \text{ ist Wahrscheinlichkeitsmaß auf } \mathcal{B}(\mathbb{R}) \text{ mit stetiger Verteilungsfunktion } F_P\}.$$

Für $P^n \in \mathcal{P}$ sei $Q_P = (0,1) \to \mathbb{R}$ "die" Quantilsfunktion, $Q_P(q) = \sup\{s \in \mathbb{R} \colon F_P(s) \leq q\}$. Insbesondere gilt

$$\forall s \in \mathbb{R}\ \forall q \in (0,1)(F_P(s) \leq q \iff s \leq Q_P(q)).$$

Gegeben $0 < q < 1$ suche wir Vertrauensintervalle für den unbekannten Parameter $Q_P(q)$. Es sei $X_{[1]} \leq X_{[2]} \leq \cdots \leq X_{[n]}$ die Ordnungsstatistik zu $X_1, \ldots, X_n$.

Satz. *Es sei $1 \leq k \leq n$, $0 < q < 1$. Dann ist $C = [X_{[k]}, \infty)$ ein Vertrauensintervall für das q-Quantil $Q_P(q)$, $P^n \in \mathcal{P}$, zum Vertrauensniveau*

$$1 - \alpha = \int_0^q \beta_{k,n-k+1}(x)\,\mathrm{d}x = \mathrm{Beta}(k, n-k+1)((0,q))$$

wobei β_k die Dichte der Betaverteilung mit Parameter k und $n-k+1$. Anders gesagt: q ist das $(1-\alpha)$-Quantil dieser Betaverteilung.

Beweis. Wir zeigen

$$\forall P^n \in \mathcal{P}.\ P^n[Q_P(q) \in C] = 1 - \alpha$$

Sei hierzu $P^n \in \mathcal{P}$. Dann gilt

$$P^n[Q_P(q) \in C] = P^n[X_{[k]} \leq Q_P(q)]$$

Nun ist $U_{[k]} = F_P(X_{[k]})$ die k-te Ordnungsstatistik von $(U_i = F_P(X_i))_{i=1,\ldots,n}$. Nun sind die $U_i = F_P(X_i)$ i.i.d. unif$[0,1]$-verteilt bezüglich P^n, denn

$$\forall a \in (0,1).\ P^n[U_i \leq a] = P^n[F_P(X_i) \leq a] = P^n[X_i \leq Q_P(a)] = F_P(Q_P(a)) = a.$$

Früher haben wir gezeigt, dass für $U_1, \ldots, U_n$ i.i.d. unif$[0,1]$-verteilt gilt

$$\mathcal{L}_{P^n}(U_{[k]}) = \mathrm{Beta}(k, n-k+1).$$

Es folgt

$$P^n[Q_P(q) \in C] = P^n[U_{[k]} \leq q] = \mathrm{Beta}(k, n-k+1,)([0,1]).$$ □

Bemerkung. Nur Vertrauensniveaus $1-\alpha$ nahe bei 1 sind interessant. Hierfür muss q deutlich oberhalb des "Hauptteils der Masse" der Betaverteilung liegen.

Satz. *Für $1 \leq k \leq n$ und $0 < q < 1$ ist $C = (-\infty, X_{[k]}]$ ein Vertrauensintervall für $Q_P(q)$, $P^n \in \mathcal{P}$, zum Vertrauensniveau*

$$1 - \alpha = \int_q^1 \beta_{k,n-k+1}(x)\,\mathrm{d}x.$$

Die beiden Sätze kann man kombinieren, um ein Vertrauensintervall der Gestalt $C(\omega) = [a(\omega), b(\omega)]$ zu erhalten. Abstraktes Prinzip:

Satz. *Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell, $C_1, C_2 \colon \Omega \to \mathcal{P}(\Theta)$ zwei Vertrauensbereiche zu einem Parameter $\vartheta \colon \mathcal{P} \to \Theta$ zu den Vertrauensniveaus $1-\alpha_1$ bzw. $1-\alpha_2$. Dann ist $C_1 \cap C_2$ ein Vertrauensbereich zum Vertrauensbereich $1 - (\alpha_1 + \alpha_2)$.*

Beweis. Nach Voraussetzung ist für $i = 1, 2$

$$\forall P \in \mathcal{P}.\ P[\vartheta(P) \in C_i] \geq 1 - \alpha_i.$$

Es folgt

$$\begin{aligned} P[\vartheta(P) \in C_1 \cap C_2] &= 1 - P[\vartheta(P) \notin C_1 \vee \vartheta(P) \notin C_2] \geq 1 - P[\vartheta(P) \notin C_1] - P[\vartheta(P) \notin C_2] \\ &\geq 1 - \alpha_1 - \alpha_2 \end{aligned}$$ □

Beispiel (Fort.). Im Beispiel bedeutet das: Für $1 \le k \le l \le n$ und $0 < q < 1$ ist $c = [X_{[k]}, X_{[l]}]$ ein Vertrauensintervall zum Vertrauensniveau

$$1 - \alpha = 1 - \int_q^1 \beta_{k,n-k+1}(x)\,\mathrm{d}x - \int_0^q \beta_{k,n-l+1}(x)\,\mathrm{d}x$$

falls $\alpha < 1$. Das ist natürlich nur für $\frac{k-1}{n-1} < q < \frac{l-1}{n-1}$ interessant.

Beispiel (Einseitige Konfidenzintervalle im Binomialmodell). Sei $\Omega = \{0, \dots, n\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ und $P_\vartheta = \text{binomial}(n, \vartheta)$, $0 \le \vartheta \le 1$. Für die Nullhypothese $H_0 = \{P_{\vartheta_0}\}$ und die Alternative $H_0 = \{P_{\vartheta_1}\}$ mit $\vartheta_1 > \vartheta_0$ haben die Verwerfungsbereiche der Likelihood-Quotienten-Tests alle die Gestalt

$$V_k = \{k, k+1, \dots, n\}.$$

Die Niveaus $P_{\vartheta_0}(V_k)$ dieser Tests hängen monoton steigend von $\vartheta_0 \in [0,1]$ ab. Wir verwenden folgenden "Kopplungstrick": Es seien $U_1, \dots, U_n$ i.i.d. unif$[0,1]$-verteilt auf einem "Hilfsraum" $(\Omega', \mathcal{A}', P')$. Dann sind für $\vartheta \in [0,1]$ die Zufallsvariablen $1_{\{U_i \le \vartheta\}}$, $1 \le i \le n$, i.i.d. $\vartheta\delta_1 + (1-\vartheta)\delta_0$-verteilt, also $S_\vartheta = \sum_{i=1}^n 1_{\{U_i \le \vartheta\}}$ binomial(n, ϑ)-verteilt. Nun gilt für $0 \le \vartheta \le \vartheta' \le 1$

$$P_\vartheta(V_k) = P'[S_\vartheta \ge k] \le P'[S_{\vartheta'} \ge k] = P_{\vartheta'}(V_k).$$

Weiter gilt $P_\vartheta(V_k = P'[S_\vartheta \ge k] = P'[U_{[k]} \le \vartheta] = \text{beta}(k, n-k+1)[0, \vartheta]$. Wir setzen für $0 < \alpha < 1$ und $k \in \{0, \dots, n\}$

$$q(\alpha, k) = \begin{cases} \sup\{\vartheta \colon P_\vartheta \le \alpha\} & \text{für } k > 0 \\ 0 & \text{für } k = 0 \end{cases}$$

und

$$C_\alpha(\omega) = [q(\alpha, \omega), 1].$$

Dann ist $C_\alpha(\omega)$ ein $(1-\alpha)$-Vertrauensintervall für den Parameter ϑ, denn für alle $\vartheta \in [0,1]$ gilt mit der Abkürzung $k(\alpha, \vartheta) = \max\{k \in \{0, \dots, n\} \colon P_\vartheta(V_k) > \alpha\}$ für $\vartheta \ne 0$

$$\begin{aligned} P_\vartheta[\vartheta \in C_\alpha] &= P_\vartheta(\{\omega \in \Omega \colon \vartheta \ge q(\alpha, \omega)\}) \ge P_\vartheta(\{\omega \in \Omega \colon P_\vartheta(V_\omega) > \alpha\}) = \\ &= P_\vartheta(\{0, \dots, k(\alpha, \vartheta)\}) = 1 - P_\vartheta(V_{k(\alpha,\vartheta)+1}) \ge 1 - \alpha \end{aligned}$$

und für $\vartheta = 0$

$$P_\vartheta(\{\omega \in \Omega \colon \vartheta \ge q(\alpha, \omega)\}) = P_0(\{0\}) = 1 \ge 1 - \alpha.$$

2.4.6 t-Test

Seien $X_1, \dots, X_n$ i.i.d. normalverteilte Daten. Anders als früher nehmen wir an, dass sowohl die Erwartung μ als auch die Varianz σ^2 unbekannt sind. Wir betrachten das Modell $\Omega = \mathbb{R}^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ und $\mathcal{P} = \{P_{\mu,\sigma^2} \colon \mu \in \mathbb{R}, \sigma^2 > 0\}$ mit $P_{\mu,\sigma^2} = \text{normal}(\mu, \sigma^2)^n$ und den kanonischen Projektionen $X_1, \dots, X_n$. Wir suchen einen Test zur Überprüfung von Hypothesen über das unbekannte μ. Früher, bei bekanntem σ^2, haben wir für $H_0 = \{P_{\mu_0,\sigma^2}\}$ die Teststatistik

$$Z = \sqrt{n} \cdot \frac{\overline{X} - \mu_0}{\sigma}$$

mit $\mathcal{L}_{P_{\mu_0,\sigma^2}}(Z) = \text{normal}(0,1)$ verwendet. Bei unbekanntem σ^2 liegt es nahe, statt σ^2 die Schätzung

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$$

zu verwenden.

Definition. $T = \sqrt{n} \cdot \frac{\overline{X}-\mu_0}{s_X}$ heißt die *t-Statistik*.

Wir untersuchen die Verteilung von T unter der Hypothese $H_0 = \{P_{\mu_0,\sigma^2} : \sigma^2 > 0\}$ bei gegebenem μ_0.

Lemma. *Seien $(X_1, \ldots, X_n)$ P_{μ_0,σ^2}-verteilt. Dann hat*

$$\left(\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0), \frac{n-1}{\sigma^2}s_X^2\right)$$

die Verteilung $\text{normal}(0,1) \times \chi^2_{n-1}$*, wobei χ^2_{n-1} die Chi-Quadrat-Verteilung mit $n-1$ Freiheitsgeraden, also die Verteilung von $\|X\|_2^2$ mit $(n-1)$-dimensional standardnormalverteiltem X, sei.*

Beweis. Sei $Z_j = \frac{X_j-\mu_0}{\sigma}$, $j = 1, \ldots, n$. Dann sind $Z_1, \ldots, Z_n$ i.i.d. standardnormalverteilt, also $Z = (Z_1, \ldots, Z_n) \sim \text{normal}(0, I_n)$. Wir setzen $v = \frac{1}{\sqrt{n}}(1, \ldots, 1)^T \in \mathbb{R}^n$. Offenbar ist $\|v\|_2 = 1$. Dann gilt

$$\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} = \sqrt{n} \cdot \overline{Z} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_j = \langle Z, v \rangle = v^T Z$$

und

$$\frac{(n-1)s_X^2}{\sigma^2} = (n-1)s_Z^2 = \sum_{j=1}^{n}(Z_j - \overline{Z})^2 = \|Z - vv^T Z\|_2^2 = \|\underbrace{(I_n - vv^T)}_{\text{orthogonale Projektion auf } v^\perp} Z\|_2^2.$$

$v^T Z$ ist die Komponente von Z in Richtung v und $\|(I_n - vv^T)Z\|_2^2$ ist das Normquadrat der Komponente von Z senkrecht zu V. Nun ist die Dichte von $\text{normal}(0, I_n)$, also $x \mapsto (2\pi)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\|x\|_2^2\right)$ rotationsinvariant, also $\mathcal{L}(Z)$ rotationsinvariant. Folglich ist die Verteilung von $(v^T Z, \|(I_n - vv^T)Z\|_2^2)$ die Gleiche für *alle* Einheitsvektoren $v \in \mathbb{R}^n$. Insbesondere können wir v durch $e = (0, \ldots, 0, 1)$ ersetzen:

$$\begin{aligned}\mathcal{L}(v^T z, \|(I_n - vv^T)Z\|_2^2) = \mathcal{L}(e^T Z, \|(I_n - ee^T)Z\|_2^2) = \mathcal{L}(Z_n, \|(Z_1, \ldots, Z_{n-1})\|_2^2) = \\ = \text{normal}(0,1) \times \chi^2_{n-1}\end{aligned}$$

□

Korollar. *Die Verteilung der t-Statistik und P_{μ_0,σ^2} ist für alle $\sigma^2 > 0$ die Gleiche, nämlich diejenige von*

$$T_{n-1} = \sqrt{n-1}\frac{X}{\sqrt{Y_{n-1}}}$$

wobei $\mathcal{L}(X, Y_{n-1}) = \text{normal}(0,1) \times \chi^2_{n-1}$.

Beweis.

$$T = \sqrt{n-1}\frac{\sqrt{n}(\overline{X} - \mu_0)/\sigma}{\sqrt{(n-1)s_X^2/\sigma^2}}$$

□

Definition. Die *Student t-Verteilung* t_n mit n Freiheitsgraden ist die Verteilung von $T_n = \sqrt{n}\frac{X}{\sqrt{Y_n}}$, wenn X und Y_n unabhängig sind mit $\mathcal{L}(X) = \text{normal}(0,1)$ und $\mathcal{L}(Y_n) = \chi^2_n$.

Lemma. *Für $n \to \infty$ konvergiert t_n schwach gegen die Standardnormalverteilung.*

Lemma. *Die Dichte der t_n-Verteilung hat die Gestalt*

$$f_n(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\ \Gamma(\frac{n}{2})}\left(\frac{t^2}{n} + 1\right)^{-\frac{n+1}{2}}$$

Zusammenfassend: Wir betrachten die Nullhypothese $H_0 = \{\text{normal}(\mu_0, \sigma^2)^n \colon \sigma^2 > 0\}$ für festes $\mu_0 \in \mathbb{R}$ mit der Alternative $H_1 = \{\text{normal}(\mu_1, \sigma^2) \colon \mu_1 > \mu_0, \sigma^2 > 0\}$. Dann hat für alle $P_0 \in H_0$ hat die Teststatistik $T = \sqrt{n}\, \frac{\overline{X} - \mu_0}{s_X}$ die Verteilung $L_{P_0}(T) = t_{n-1}$. Verwerfungsbereich zum Niveau α:

$$V_\alpha = \{T > t_{n-1,\alpha-1}\}$$

mit dem $(1-\alpha)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden $t_{n-1,\alpha-1}$.