Vorlesungsnotizen zur Stochastik

Franz Merkl

(sehr vorläufige Version [1], 10. Februar 2020)

Inhaltsverzeichnis

[1] Dies ist nur der Anfang eines Entwurfs eines Stochastikskripts. Für Hinweise auf Fehler aller Art ist der Autor dankbar.

1 Einführung

Stochastik ist die Lehre von der mathematischen Analyse zufälliger Vorgänge. Sie gliedert sich in die folgenden Teilgebiete:

- **Wahrscheinlichkeitstheorie:** Typische Aufgaben sind hier die Berechnung oder die Abschätzung von Wahrscheinlichkeiten bei *gegebenem* Wahrscheinlichkeitsmodell.
- **Mathematische Statistik:** Hier geht es auf den Rückschluß aus Beobachtungsdaten auf ein *unbekanntes* Wahrscheinlichkeitsmodell oder auf Eigenschaften dieses Wahrscheinlichkeitsmodells. Typische Aufgaben sind das *Schätzen* von Parametern des Modells oder das *Testen* von Hypothesen über das nur unvollständig bekannte Wahrscheinlichkeitsmodell.

Diese Vorlesung gliedert sich auch gemäß dieser Gebietseinteilung.
Stochastik ist ein Gebiet der angewandten Mathematik. Schematisch kann man sich den Bezug zwischen dem Anwendungsgebiet und der Mathematik so vorstellen:

Anwendungsgebiet		**Mathematik**
Problemstellung	$\overset{\text{Modellierung}}{\longrightarrow}$	Modell
		$\downarrow$mathemat. Analyse
Anwendung	$\overset{\text{Interpretation}}{\longleftarrow}$	Theoreme, Verfahren

Die mathematische Modellierung und die Interpretation der Theoreme bilden die Schnittstelle zwischen Mathematik und Anwendungsgebiet, sind also strenggenommen kein Bestandteil der innermathematischen Theorie.

2 Wahrscheinlichkeitstheorie

2.1 Wahrscheinlichkeitsmodelle

Ein Wahrscheinlichkeitsmodell besteht aus drei Komponenten $(\Omega, \mathcal{A}, P)$ sowie informalen Interpretationsregeln, was diese Komponenten bedeuten sollen. Wir besprechen jetzt diese drei Komponenten und ihre Interpretation einzeln.

2.1.1 Der Ergebnisraum Ω

Ω ist eine nichtleere Menge, der *Ergebnisraum*. Die Elemente von Ω heißen *Ergebnisse* und werden als mögliche Ausgänge eines Zufallsexperiments interpretiert. Sie werden oft mit ω bezeichnet.
Die Wahl des Ergebnisraums Ω in einer Anwendung hängt stark davon ab, welche Aspekte des Experiments im Modell widergespiegelt werden sollen, und welche ignoriert werden

sollen.

Beispiel 1: Einmaliger Wurf eines Würfels.

- *Modell 1:* Das naheliegendste Modell hat den Ergebnisraum $\Omega_1 = \{1, 2, 3, 4, 5, 6\}$. Interpretation von Ergebnissen $\omega \in \Omega_1$: obenliegende Augenzahl.
- *Modell 2:* $\Omega_2 = \{1, 2, 3, 4, 5, 6, \text{ungültig}\}$. Dieses Modell ist feiner als Modell 1, denn im ersten Modell werden "ungültige" Ergebnisse ignoriert.
- *Modell 3:* $\Omega_3 = \mathbb{R}^3 \times \mathrm{SO}(3)$, wobei

  $$\mathrm{SO}(3) = \{A \in \mathbb{R}^{3\times 3} : \ A \cdot A^t = \mathrm{Id}, \det A = 1\}$$

  die Menge aller 3×3-Drehmatrizen bezeichnet. Als Interpretation eines Ergebnisses $(x, A) \in \Omega_3$ wählen wir:

  - Der Vektor x soll den Ort des Würfelschwerpunkts in einem kartesischen Koordinatensystem bedeuten.
  - Die Drehmatrix A soll die Drehung des Würfels in diesem Koordinatensystem bezüglich einer Referenzlage beschreiben.

Je nach Interesse oder Aufgabenstellung kann jedes der drei Modelle angemessen sein. Für ein Würfelspiel würde ich Modell 1 bevorzugen, für die mechanische Analyse der Dynamik des Würfelwurfs jedoch Modell 3.

Beispiel 2: n-facher Münzwurf:

- *Modell 1:* Ergebnisraum

  $$\Omega = \{0, 1\}^n = \{(\omega_1, \ldots, \omega_n) : \ \omega_i \in \{0, 1\} \text{ für alle } i = 1, \ldots, n\}.$$

  Interpretation von $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$:

  - Die Komponente $\omega_i = 1$ bedeutet "Kopf" beim i-ten Wurf.
  - Die Komponente $\omega_i = 0$ bedeutet "Zahl" beim i-ten Wurf.
- *Modell 2:* Ergebnisraum $\Omega' = \{0, 1, \ldots, n\}$. Interpretation von $\omega' \in \Omega'$: Bei den n Würfen ist ω-mal "Kopf" aufgetreten.

Das Modell 1 enthält mehr Informationen als Modell 2, weil die Reihenfolge des Auftretens von "Kopf" und "Zahl" im Modell 1 beachtet wird, im Modell 2 jedoch ignoriert wird. Wir können diesen Informationsverlust durch die Abbildung

$$S : \Omega \to \Omega', \quad S(\omega_1, \ldots, \omega_n) = \sum_{i=1}^{n} \omega_i$$

beschreiben, die Modell 1 in Modell 2 überführt.

Beispiel 3a: Ziehen von n Kugeln aus einer Urne mit $m \geq n$ Kugeln unterschiedlicher Farbe, mit Zurücklegen:
Modell: Wir verwenden die Standardnotation $[n] := \{1, \ldots, n\}$. Ein möglicher Ergebnisraum lautet

$$\Omega = [m]^{[n]} = \{\omega |\ \omega : [n] \to [m]\},$$

wobei jedes Ergebnis $\omega \in \Omega$ so interpretiert wird: Wir nummerieren die Farben mit $1, \ldots, m$. Der Wert $\omega(i) = j$ bedeutet, dass beim i-ten Zug die j-te Farbe gezogen wird.

Beispiel 3b: Ziehen von n Kugeln aus einer Urne mit $m \geq n$ Kugeln unterschiedlicher Farbe, ohne Zurücklegen:
Modell: Möglicher Ergebnisraum:

$$\Omega = \{\omega \in [m]^{[n]} |\ \omega \text{ ist injektiv}\}.$$

Wir untersuchen dieses Modell in den Übungen genauer.

Die bisher betrachteten Modelle sind *diskret*, d.h. sie besitzen einen abzählbaren Ergebnisraum. Genauer gesagt sogar *endlich*.

Beispiel 4: Glücksrad. Der Zeiger eines Glücksrads wird gedreht; er bleibt an einer zufälligen Stelle stehen.
Modell 1: Ergebnisraum

$$\Omega = S^1 = \{(x, y) \in \mathbb{R}^2 |\ x^2 + y^2 = 1\},$$

wobei ein Ergebnis (x, y) die kartesischen Koordinaten der Zeigerspitze bedeuten soll.
Modell 2: Ergebnisraum $\Omega' = [0, 1[$, wobei $\omega' \in \Omega'$ bedeuten soll, dass der Winkel zwischen der positiven x-Achse und dem Zeiger $\alpha = 2\pi\omega'$ beträgt.

Die Modelle 1 und 2 sind gleichwertig vermittels der Bijektion

$$X : \Omega' \to \Omega, \quad X(\omega') = (\cos(2\pi\omega'), \sin(2\pi\omega')).$$

Die beiden Modelle in diesem Beispiel sind *kontinuierliche Modelle*; sie besitzen einen *überabzählbaren* Ergebnisraum, der gleichmächtig zur Menge $\mathbb{R}$ der reellen Zahlen ist.

Zum Abschluß dieser Beispielliste betrachten wir noch ein Modell mit einem *funktionalen* Ergebnisraum:

Beispiel 5: Pegelstand des Ammersees in einem Zeitintervall $[t_0, t_1]$. Wir modellieren den Ereignisraum so:

$$\Omega = C[t_0, t_1] := \{\omega : [t_0, t_1] |\ \omega \text{ ist stetig}\}.$$

Dabei soll $\omega \in \Omega$ so interpretiert werden: Der Wert $\omega(t)$ bezeichnet den Pegelstand in Meter zur Zeit $t \in [t_0, t_1]$.
Ähnliche funktionale Ergebnisräume werden in der Finanzmathematik in kontinuierlicher Zeit oft verwendet, zum Beispiel zur Modellierung des Kursverläufs eines Anlageguts.

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

Ja/Nein-Fragen an das zufällige Ergebnis $\omega \in \Omega$ werden durch Teilmengen $A \subseteq \Omega$ des Ergebnisraums modelliert. Die Aussage $\omega \in A$ entspricht dabei der Antwort "ja", $\omega \notin A$ der Antwort "nein". Dabei muss man nicht beliebige Fragen zulassen, sondern kann nur manchen Teilmengen $A \subseteq \Omega$ als "*beobachtbar*", "*messbar*" auszeichnen. Diese messbaren Teilmengen heißen *Ereignisse*.

Bitte nicht verwechseln:

Ergebnisse sind *Elemente* von Ω,
Ereignisse sind *messbare Teilmengen* von Ω.

Die Menge $\mathcal{A}$ aller Ereignisse soll einige Abschlusseigenschaften unter endlichen und abzählbaren booleschen Operationen besitzen, die im Begriff der *σ-Algebra* (engl.: σ-field) zusammengefasst werden:[2]

Definition 2.1 (σ-Algebra) *Es sei Ω ein Ergebnisraum. Eine Menge $\mathcal{A}$ von Teilmengen von Ω heißt eine 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:*

$$\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}.$$

Die Menge A^c wird das Komplement von A genannt. Ein Paar (Ω, A), bestehend aus einem Ergebnisraum Ω und einer σ-Algebra $\mathcal{A}$ über Ω, heißt *Ereignisraum* (synonym: *messbarer Raum*). Die Elemente $A \in \mathcal{A}$ heißen *Ereignisse* (synonym: *messbare Menge*).
Das Ereignis $\Omega \in \mathcal{A}$ heißt das *sichere Ereignis*.
Das Ereignis $\emptyset = \Omega^c \in \mathcal{A}$ heißt das *unmögliche Ereignis*.

Beispiel 1: Die Potenzmenge von Ω, also $\mathcal{P}(\Omega) = \{A : A \subseteq \Omega\}$, ist eine σ-Algebra über Ω. Meist wählt man diese σ-Algebra, wenn Ω endlich oder abzählbar unendlich ist, oft sogar, ohne dies explizit zu erwähnen.

Beispiel 2: Das Mengensystem $\mathcal{A} = \{\emptyset, \Omega\}$ ist eine σ-Algebra über Ω. Sie heißt die *triviale σ-Algebra* über Ω.

Beispiel 3: Einfacher Wurf eines Würfels: $\Omega = [6]$, $\mathcal{A} = \mathcal{P}(\Omega)$. Das Ereignis "gerade Augenzahl" wird durch

$$A = \{2, 4, 6\}$$

[2] Das Symbol σ steht für *abzählbar unendliche* Mengenoperationen, insbesondere aufsteigende Operationen, wie abzählbar unendliche Vereinigungen.

beschrieben.

In der Definition von σ-Algebren ist nur ein minimaler Satz von Abschlusseigenschaften aufgenommen. Weitere Abschlusseigenschaften unter (endlichen oder abzählbar unendlichen) booleschen Kombinationen sind Konsequenzen daraus:

Lemma 2.2 (Abschlusseigenschaften unter booleschen Operationen) *Es sei $\mathcal{A}$ eine σ-Algebra über Ω. Dann gilt:*

1. *$\emptyset \in \mathcal{A}$,*
2. *Für alle $A, B \in \mathcal{A}$ folgen $A \cup B, A \cap B, A \setminus B, A \triangle B := (A \setminus B) \cup (B \setminus A) \in \mathcal{A}$.*
3. *Für alle Folgen $(A_n)_{n\in\mathbb{N}}$ mit Werten in $\mathcal{A}$[3] gilt: $\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{A}$.*

Beweis:

1. Mit den definierenden Eigenschaften 1. und 2. von σ-Algebren folgt: $\emptyset = \Omega^c \in \mathcal{A}$.
2. Die Folge $A, B, \emptyset, \emptyset, \emptyset, \ldots$ nimmt Werte in $\mathcal{A}$ an. Die Vereinigung ihrer Folgenglieder ist $A \cup B$. Mit Bedingung 3. in Definition 2.1 folgt: $A \cup B \in \mathcal{A}$. Die übrigen Fälle $A \cap B, A \setminus B, A \triangle B$ sollen Sie als Übung behandeln.
3. Es gilt $A_n^c \in \mathcal{A}$ für alle $n \in \mathbb{N}$ wegen Bed. 2. in Def. 2.1, also $\bigcup_{n\in\mathbb{N}} A_n^c \in \mathcal{A}$ wegen Bed. 3. in Def. 2.1. Nochmal mit Bed. 2. in Def. 2.1 folgt:

$$\bigcap_{n\in\mathbb{N}} A_n = \left(\bigcup_{n\in\mathbb{N}} A_n^c\right)^c \in \mathcal{A}.$$

□

Für jede Teilmenge $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ gibt es eine *kleinste* σ-Algebra über Ω, die $\mathcal{M}$ enthält, nämlich den Durchschnitt aller σ-Algebren über Ω, die $\mathcal{M}$ enthalten:

$$\begin{aligned}
&\sigma(\mathcal{M}) = \sigma(\mathcal{M}, \Omega) \\
&:= \bigcap \{\mathcal{A} \subseteq \mathcal{P}(\Omega) |\ \mathcal{A} \text{ ist eine } \sigma\text{-Algebra über } \Omega \text{ mit } \mathcal{M} \subseteq \mathcal{A}\} \\
&= \{A \subseteq \Omega :\ \text{Für jede } \sigma\text{-Algebra } \mathcal{A} \text{ über } \Omega \text{ mit } \mathcal{M} \subseteq \mathcal{A} \text{ gilt } A \in \mathcal{A}\}. \qquad (1)
\end{aligned}$$

Übung 2.3 (Erzeugte σ-Algebra) Beweisen Sie, dass $\sigma(\mathcal{M})$ in der Tat eine σ-Algebra über Ω ist, die $\mathcal{M}$ umfasst.

[3] Kurznotation dafür: $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$

$\sigma(\mathcal{M})$ wird die von $\mathcal{M}$ *erzeugte σ-Algebra* genannt. $\mathcal{M}$ wird ein *Erzeugendensystem* der σ-Algebra $\sigma(M)$ genannt.

Beispiel Würfeln, $\Omega = [6]$: Wir betrachten die Ereignisse

$$A = \{2,4,6\} = \text{“gerade Augenzahl”}, \qquad B = \{6\} = \text{“Augenzahl 6”}.$$

Dann gilt mit den Abkürzungen $\sigma(A) := \sigma(\{A\})$ und $\sigma(A,B) := \sigma(\{A,B\})$:

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

Beobachtet man also nur, ob die Ereignisse A und B eintreten, so weiß man auch, ob das Ereignis $\{1,3,5,6\}$ eingetreten ist, aber zum Beispiel nicht, ob das Ereignis $\{1,5\}$ eingetreten ist: Gegeben nur die in A und B enthaltene Information, ist also das Ereignis $\{1,5\}$ nicht messbar.

Ausblick/Bemerkung: Man kann $\sigma(\mathcal{M})$ auch “von unten” statt “von oben” konstruieren:

- *“von oben”:* Durchschnitt über größere σ-Algebren
- *“von unten”:* Rekursives Hinzunehmen von Elementen von $\mathcal{M}$, von Ω, von Komplementen, von abzählbaren Vereinigungen. Allerdings reichen abzählbar viele Rekursionsschritte i.a. nicht aus: Hat man in jedem Rekursionsschritt $n \in \mathbb{N}$ eine Menge A_n neu hinzugenommen, so muss man schließlich auch die Diagonal-Vereinigung $\bigcup_{n\in\mathbb{N}} A_n$ hinzunehmen, und rekursiv so fort...
  Formal kann man diese rekursive Konstruktion “von unten” mit Hilfe der sogenannten “*transfiniten Rekursion*” ausführen. Wir brauchen diese mengentheoretische Konstruktion in dieser Vorlesung jedoch nicht, obwohl sie eine bessere Anschauung des Begriffs der erzeugten σ-Algebra liefert.

Bemerkung: Ist der Ergebnisraum Ω endlich oder abzählbar unendlich, so wird jede σ-Algebra $\mathcal{A}$ über Ω von einer eindeutig bestimmten *Partition* $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ von Ω erzeugt. Eine Partition von Ω ist eine Menge von paarweise disjunkten nichtleeren Teilmengen von Ω, deren Vereinigung Ω ist. Die Elemente von $\mathcal{A}$ sind dann genau die Vereinigungen von Teilmengen von $\mathcal{M}$.

Beispiel Würfeln (Fortsetzung): Im obigen Beispiel wird $\sigma(\{A,B\})$ auch von der Partition $\{\{1,3,5\},\{2,4\},\{6\}\}$ von Ω erzeugt.

Als Standard-σ-Algebra über der Menge $\mathbb{R}$ der reellen Zahlen verwendet man die von den offenen Intervallen erzeugte σ-Algebra:

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

Sie wird die *Borelsche σ-Algebra* über $\mathbb{R}$ genannt. Ihre Elemente heißen *Borelmengen* oder auch *Borel-messbar*. Diese Borelsche σ-Algebra wird zum Beispiel auch von den folgenden Mengensystemen erzeugt:

$$\begin{aligned}
\mathcal{B}(\mathbb{R}) &= \sigma(]-\infty, a] :\ a \in \mathbb{R}) && (2)\\
&= \sigma(]a, b] :\ a, b \in \mathbb{R}, a \leq b) && (3)\\
&= \sigma([a, b] :\ a, b \in \mathbb{R}, a \leq b) && (4)\\
&= \sigma(]-\infty, a] :\ a \in \mathbb{Q}) && (5)\\
&= \sigma(A :\ A \subseteq \mathbb{R} \text{ ist offen}). && (6)
\end{aligned}$$

Die Borelsche σ-Algebra $\mathcal{B}(\mathbb{R})$ ist echt kleiner als die Potenzmenge $\mathcal{P}(\mathbb{R})$. Sie wird nicht von einer Partition von $\mathbb{R}$ erzeugt.

Allgemeiner definiert man für jeden metrischen Raum (M, d) oder auch für jeden topologischen Raum $(M, \mathcal{T})$ die Borelsche σ-Algebra $\mathcal{B}(M)$ als die von dem Mengensystem der offenen Mengen erzeugte σ-Algebra. Insbesondere verwenden wir diese Notation für Teilmengen $M \subseteq \mathbb{R}^n$ mit der euklidischen Metrik.

Warum arbeitet man mit σ-Algebren über Ω statt nur mit der Potenzmenge $\mathcal{P}(\Omega)$? Es gibt eine positiven und einen negativen Grund dafür:

- *positiv:* σ-Algebren erlauben die Modellierung unvollständiger, wechselnder Beobachtungsmöglichkeiten, zum Beispiel beim Informationszuwachs im Verlauf der Zeit.

  Beispiel n-facher Münzwurf: Betrachten wir das Modell $\Omega_n = \{0, 1\}^n$ mit ω_i in $(\omega_1, \ldots, \omega_n) \in \Omega_n$ als Ergebnis des i-ten Wurfs. Beobachten wir zunächst nicht alle Würfe, sondern nur die erste $m < n$ davon, so sind nicht alle Teilmengen von Ω_n beobachtbar, sondern nur jene in

  $$\mathcal{F}_m := \{\Pi_{n,m}^{-1}[A] :\ A \subseteq \{0, 1\}^m\},$$

  wobei $\Pi_{n,m} : \{0, 1\}^n \to \{0, 1\}^m$, $\Pi_{n,m}(\omega_1, \ldots, \omega_n) = (\omega_1, \ldots, \omega_m)$ die Projektion auf die ersten m Koordinaten bezeichnet. Zur Beschreibung des zunehmenden Informationsgewinns im Verlauf der Zeit kann man also die "Filtration" von σ-Algebren

  $$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_n$$

  über Ω_n verwenden, wobei größere σ-Algebren größere Beobachtungsmöglichkeiten bedeuten.

- *negativ:* Bei kontinuierlichen Modellen, zum Beispiel dem Glücksrad $\Omega = S^1$, gibt es sinnvolle Wahrscheinlichkeitsmodelle über der Borelschen σ-Algebra $\mathcal{B}(S^1)$, zum Beispiel die später zu besprechende "Gleichverteilung", die sich jedoch nicht sinnvoll auf ganz $\mathcal{P}(S^1)$ erweitern lassen.

2.1.3 Wahrscheinlichkeitsmaße P

Bis in die erste Hälfte des 20. Jahrhunderts, lange nach den Anfängen der Stochastik, gab es keine präzise Theorie für Wahrscheinlichkeiten in kontinuierlichen Modellen oder z.B. für unendlich häufige Wiederholungen eines Münzwurfs. Erst dem russische Mathematiker Andrei Nikolajewitsch Kolmogoroff (1903-1987) gelang es, dem Wahrscheinlichkeitsbegriff auch für nicht abzählbare Ergebnisräume einen mathematisch rigorosen Sinn zu geben. Er benutzte hierzu die moderne Maßtheorie von Lebesgue und Borel, mit der auch dem Volumenbegriff ein rigoroser Sinn gegeben wurde. Zu Ehren Kolmogoroffs werden die definierenden Eigenschaften von Wahrscheinlichkeitsmaßen auch *Kolmogoroff-Axiome* genannt. Heute ist die Maßtheorie gleichzeitig das Fundament und die Sprache der modernen Wahrscheinlichkeitstheorie, ähnlich wie die Mengenlehre gleichzeitig das Fundament und die Sprache der modernen Mathematik ist.

Definition 2.4 (Wahrscheinlichkeitsmaß: Kolmogoroff-Axiome) *Es sei $(\Omega, \mathcal{A})$ ein Ereignisraum. Eine Abbildung $P : \mathcal{A} \to [0, 1]$ heißt Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$, wenn gilt:*

1. *$P(\Omega) = 1$.*
2. *Für jede Folge $(A_n)_{n\in\mathbb{N}}$ von paarweise disjunkten Ereignissen $A_n \in \mathcal{A}$ (d.h. $A_i \cap A_j \neq \emptyset$ für alle $i, j \in \mathbb{N}$ mit $i \neq j$) gilt*

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n). \tag{7}$$

Die Eigenschaft 2. in dieser Definition heißt auch σ-Additivität. Der Funktionswert $P(A)$ eines Ereignisses A wird die *Wahrscheinlichkeit* von A genannt. Ein Tripel $(\Omega, \mathcal{A}, P)$, bestehend aus einem Ergebnisraum Ω, einer σ-Algebra $\mathcal{A}$ über Ω und einem Wahrscheinlichkeitsmaß P über $(\Omega, \mathcal{A})$ wird *Wahrscheinlichkeitsraum* genannt. Ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, zusammen mit einer Interpretation, was die Ergebnisse $\omega \in \Omega$, die Ereignisse $A \in \mathcal{A}$ und ihre Wahrscheinlichkeiten $P(A)$ in einer Anwendung bedeuten sollen, wird ein *Wahrscheinlichkeitsmodell* genannt.
Verzichtet man auf die Normierungsbedingung $P(\Omega) = 1$, so gelangt man zum allgemeineren Maßbegriff:

Definition 2.5 (Maß) *Es sei $(\Omega, \mathcal{A})$ ein Ereignisraum. Eine Abbildung $P : \mathcal{A} \to [0, \infty]$ (Wert $+\infty$ erlaubt!) heißt ein Maß auf $(\Omega, \mathcal{A})$, wenn gilt:*

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

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n). \tag{8}$$

In diesem Fall nennt man das Tripel $(\Omega, \mathcal{A}, \mu)$ einen Maßraum.

Bemerkung: In der Definition von Wahrscheinlichkeitsmaßen wäre die Forderung $P(\emptyset) = 0$ überflüssig, weil sie automatisch folgt:

$$P(\emptyset) = P\left(\bigcup_{n\in\mathbb{N}} \emptyset\right) = \sum_{n\in\mathbb{N}} P(\emptyset). \tag{9}$$

Wegen $P(\emptyset) \in [0,1]$ folgt $P(\emptyset) \neq \infty$, und hiermit $P(\emptyset) = 0$. Weil beim allgemeinen Maßbegriff auch der Wert $+\infty$ zugelassen ist, ist dort die Forderung $\mu(\emptyset) = 0$ jedoch nötig.

Einfache Eigenschaften von Wahrscheinlichkeitsmaßen

Lemma 2.6 (Eigenschaften von Wahrscheinlichkeitsmaßen) *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Dann gilt:*

1. *Endliche Additivität: Sind A und B* disjunkte *Ereignisse, d.h. $A \cap B = \emptyset$, so folgt*

$$P(A \cup B) = P(A) + P(B) \tag{10}$$

   Allgemeiner gilt für paarweise disjunkte *Ereignisse $A_1, \ldots, A_n$ die Formel*

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k). \tag{11}$$

   Gleiches gilt für allgemeine Maße.

2. *Sind A und B* beliebige *Ereignisse, so folgt*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{12}$$

   Die Variante

$$P(A \cup B) + P(A \cap B) = P(A) + P(B) \tag{13}$$

   dieser Formel, die undefinierte Ausdrücke $\infty - \infty$ vermeidet, gilt auch für allgemeine Maße.

3. *Komplementwahrscheinlichkeit: Für alle $A \in A$ gilt*

$$P(A^c) = 1 - P(A) \tag{14}$$

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 monoton aufsteigende Folge von Ereignissen, $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$, so folgt*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{n\in\mathbb{N}} A_n\right). \tag{15}$$

Diese Aussage gilt auch für allgemeine Maße.

6. *σ-Stetigkeit von oben: Ist $(A_n)_{n\in\mathbb{N}}$ eine monoton fallende Folge von Ereignissen, $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$, so folgt*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcap_{n\in\mathbb{N}} A_n\right). \tag{16}$$

Beweis:

1. Wir beweisen gleich den allgemeinen Fall. Die gegebene Liste $A_1, \ldots, A_n$ paarweise disjunkter Ereignisse setzen wir mit leeren Mengen zur Folge $A_1, \ldots, A_n, \emptyset, \emptyset, \emptyset, \ldots$ fort. Mit $P(\emptyset) = 0$ und der σ-Additivität von P folgt

$$\sum_{k=1}^{n} P(A_k) = \sum_{k=1}^{n} P(A_k) + \sum_{k=n+1}^{\infty} P(\emptyset) = P\left(\bigcup_{k=1}^{n} A_k \cup \bigcup_{k=n+1}^{\infty} \emptyset\right) = P\left(\bigcup_{k=1}^{n} A_k\right).$$

2. Wir schreiben $A \cup B$ als Vereinigung der disjunkten Ereignisse A und $B \setminus A$, und B als Vereinigung der disjunkten Ereignisse $A \cap B$ und $B \setminus A$. Mit der endlichen Additivität von P folgt die Variante der Behauptung so:

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

3. Weil Ω die Vereinigung der disjunkten Ereignisse A und A^c ist, folgt

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

und hieraus die Behauptung.

4. Wir schreiben die Obermenge B von A als Vereinigung der disjunkten Ereignisse A und $B \setminus A$. Mit der endlichen Additivität und mit $P(B \setminus A) \geq 0$ folgt

$$P(B) = P(A) + P(B \setminus A) \geq P(A).$$

5. Setzen wir $A_0 = \emptyset$ und $B_n = A_n \setminus A_{n-1}$ für $n \in \mathbb{N}$, so sind die B_n, $n \in \mathbb{N}$, paarweise disjunkte Ereignisse mit

$$\bigcup_{k=1}^{n} B_k = A_n, \tag{17}$$

$$\bigcup_{k\in\mathbb{N}} B_k = \bigcup_{k\in\mathbb{N}} A_k \tag{18}$$

für alle $n \in \mathbb{N}$. Mit der endlichen Additivität und der σ-Additivität von P folgt die Behauptung so:

$$P(A_n) = P\left(\bigcup_{k=1}^{n} B_k\right) = \sum_{k=1}^{n} P(B_k)$$
$$\stackrel{n\to\infty}{\longrightarrow} \sum_{k=1}^{\infty} P(B_k) = P\left(\bigcup_{k\in\mathbb{N}} B_k\right) = P\left(\bigcup_{k\in\mathbb{N}} A_k\right). \tag{19}$$

6. Die Folge $(A_n^c)_{n\in\mathbb{N}}$ ist aufsteigend, weil $(A_n)_{n\in\mathbb{N}}$ absteigend ist. Mit den Teilen 4. (Komplemente) und 5. (σ-Stetigkeit von unten) des Lemmas folgt:

$$1 - P(A_n) = P(A_n^c) \stackrel{n\to\infty}{\longrightarrow} P\left(\bigcup_{n\in\mathbb{N}} A_n^c\right) = P\left(\left(\bigcap_{n\in\mathbb{N}} A_n\right)^c\right) = 1 - P\left(\bigcap_{n\in\mathbb{N}} A_n\right) \tag{20}$$

und daher

$$P(A_n) \stackrel{n\to\infty}{\longrightarrow} P\left(\bigcap_{n\in\mathbb{N}} A_n\right). \tag{21}$$

□

Beispiele:

1. **Wahrscheinlichkeitsmaße auf diskreten Räumen:** Ist Ω ein endlicher oder abzählbar unendlicher Ergebnisraum und ist $(p_\omega)_{\omega\in\Omega}$ eine Familie von Zahlen mit $p_\omega \geq 0$ für alle $\omega \in \Omega$ und

$$\sum_{\omega\in\Omega} p_\omega = 1, \tag{22}$$

so wird durch

$$P : \mathcal{P}(\Omega) \to [0,1],$$
$$P(A) = \sum_{\omega\in A} p_\omega \tag{23}$$

ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{P}(\Omega))$ definiert. Umgekehrt wird jedes Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{P}(\Omega))$ durch eine eindeutig bestimmte solche Familie $(p_\omega)_{\omega\in\Omega}$ gegeben, nämlich

$$p_\omega = P(\{\omega\}). \tag{24}$$

Man nennt $(p_\omega)_{\omega\in\Omega}$ die *Zähldichte* oder auch die *Wahrscheinlichkeitsfunktion* des Wahrscheinlichkeitsmaßes P.

2. **Zählmaß:** Ist Ω wieder ein Ergebnisraum, so definiert

$$\mu : \mathcal{P}(\Omega) \to \mathbb{N}_0 \cup \{+\infty\}, \quad , \mu(A) = |A| \tag{25}$$

ein Maß auf $(\Omega, \mathcal{P}(\Omega))$. Hierbei bezeichnet $|A|$ die Mächtigkeit von A, wobei $|A| = +\infty$ für unendliche A gelten soll. Das Maß μ wird das *Zählmaß* auf Ω genannt. Außer im Fall $|\Omega| = 1$ ist es kein Wahrscheinlichkeitsmaß.

3. **Diskrete Gleichverteilung:** Ist Ω sogar ein endlicher Ergebnisraum, so wird durch

$$P : \mathcal{P}(\Omega) \to [0,1] \quad P(A) = \frac{|A|}{|\Omega|} \tag{26}$$

ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{P}(\Omega))$ definiert. Es heißt *diskrete Gleichverteilung* kurz *Gleichverteilung*[4] auf Ω, manchmal auch *normiertes Zählmaß* oder *(diskrete) uniforme Verteilung* auf $(\Omega, \mathcal{P}(\Omega))$. Es besitzt die Zähldichte

$$\left(\frac{1}{|\Omega|}\right)_{\omega \in \Omega}. \tag{27}$$

4. **Diracmaß:** Ist $(\Omega, \mathcal{A})$ ein Ereignisraum und $b \in \Omega$ fest, so wird durch

$$\delta_b : \mathcal{A} \to [0,1], \quad \delta_b(A) = \begin{cases} 1, & \text{falls } b \in A \\ 0, & \text{falls } b \notin A \end{cases} \tag{28}$$

ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$ definiert. Es heißt *Diracmaß* in b. Ein "Zufalls"experiment mit Ergebnisraum Ω, bei dem sicher das Ergebnis b eintritt, kann durch das Modell $(\Omega, \mathcal{A}, \delta_b)$ beschrieben werden. Das Diracmaß erlaubt uns die folgende schöne Notation für ein Wahrscheinlichkeitsmaß P auf einer abzählbaren Menge Ω mit Zähldichte $(p_\omega)_{\omega \in \Omega}$:

$$P = \sum_{\omega \in \Omega} p_\omega \delta_\omega. \tag{29}$$

5. **Lebesguemaß:** In der Maßtheorievorlesung lernen Sie, dass es ein eindeutig bestimmtes Maß

$$\lambda : \mathcal{B}(\mathbb{R}) \to [0, +\infty] \tag{30}$$

auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gibt, das jedem Intervall $]a, b]$ mit reellen Grenzen $a \leq b$ seine Intervalllänge

$$\lambda(]a, b]) = b - a \tag{31}$$

[4]Bitte nicht mit der unten beprochenen kontinuierlichen Gleichverteilung verwechseln!

zuordnet. Es heißt *Lebesgue-Maß* oder auch *Borel-Lebesgue-Maß* auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Allgemeiner gibt es für jedes $n \in \mathbb{N}$ rein eindeutig bestimmtes Maß

$$\lambda_n : \mathcal{B}(\mathbb{R}^n) \to [0, \infty], \tag{32}$$

das jedem n-dimensionalen Quader

$$Q = \prod_{i=1}^{n}]a_i, b_i] \tag{33}$$

mit reellen Koordinaten $a_i \leq b_i$ sein Volumen

$$\lambda_n(Q) = \prod_{i=1}^{n} (b_i - a_i) \tag{34}$$

zuordnet. Dieses Maß heißt das *n-dimensionale (Borel-)Lebesguemaß* oder auch *Volumenmaß*. Der Funktionswert $\lambda_n(A)$ wird als Volumen einer Borelmenge $A \in \mathcal{B}(\mathbb{R}^n)$ interpretiert.

6. **Gleichverteilung auf einem Intervall:** Für jedes Intervall $[a, b] \subseteq \mathbb{R}$ positiver Länge $b - a > 0$ wird durch

$$P : \mathcal{B}(\mathbb{R}) \to [0, 1], \quad P(A) = \frac{\lambda(A \cap [a, b])}{b - a} \tag{35}$$

ein Wahrscheinlichkeitsmaß auf $\mathcal{B}([a, b])$ definiert. Es (oder auch seine Einschränkung auf $\mathcal{B}([a, b])$) wird *kontinuierliche Gleichverteilung* oder auch *uniforme Verteilung* auf $[a, b]$ genannt und oft mit $\text{unif}[a, b]$ bezeichnet.
Mit ihrer Hilfe können wir auch ein sinnvolles Wahrscheinlichkeitsmodell für das "Glücksrad" definieren:
Ist

$$f : [0, 1[\to S^1, \quad f(t) = (\cos(2\pi t), \sin(2\pi t)) \tag{36}$$

die Beschreibung des Glücksrads in Polarkoordinaten, so gilt, wie wir später viel allgemeiner sehen werden, für alle $A \in \mathcal{B}(S^1)$:[5]

$$f^{-1}[A] \in \mathcal{B}([0, 1[). \tag{37}$$

Wir setzen

$$P : \mathcal{B}(S^1) \to [0, 1], \quad P(A) = \text{unif}[0, 1](f^{-1}[A]). \tag{38}$$

Dann ist $(S^1, \mathcal{B}(S^1), P)$ ein Wahrscheinlichkeitsraum, der als ein Modell für das "Glücksrad" dient. Das Wahrscheinlichkeitsmaß P wird auch *Gleichverteilung auf* S^1 genannt.

[5]Erinnern Sie sich an die Definition des Urbilds: $f^{-1}[A] = \{t \in [0, 1[: \ f(t) \in A\}$

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

$$P : \mathcal{B}(\mathbb{R}^n) \to [0,1], \quad P(A) = \frac{\lambda_n(A \cap B)}{\lambda_n(B)}. \tag{39}$$

P ist ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Es (oder auch seine Einschränkung auf $\mathcal{B}(B)$ heißt *Gleichverteilung auf B.*

Achtung: Die Gleichverteilung auf S^1 ist kein Spezialfall hiervon, da $\lambda_2(S^1) = 0$.

Bemerkung: In den letzten Beispielen kann man schön sehen, dass i.a. aus $P(A) = 0$ *nicht* folgt: $A = \emptyset$. Zum Beispiel ist $\mathrm{unif}[a,b](\{x\}) = 0$ für alle $x \in [a,b]$, und $P(S^1) = 0$ für die Gleichverteilung P auf $[-1,1]^2 \subseteq \mathbb{R}^2$.

"Wahrscheinlichkeit 0" und "unmöglich" sind verschiedene Begriffe!

Diese Beobachtung gibt Anlaß zu folgender

Definition 2.7 (Nullmengen; fast sicher) *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum (oder auch ein beliebiger Maßraum). Eine Menge $N \subseteq \Omega$ heißt eine* Nullmenge *bezüglich P, wenn es ein $A \in \mathcal{A}$ mit $N \subseteq A$ und $P(A) = 0$ gibt.*
Im Fall $N \in \mathcal{A}$ ist dies gleichwertig mit $P(N) = 0$.
*Eine Aussage $\Phi(\omega)$ über das zufällige Ergebnis $\omega \in \Omega$ eines Zufallsexperiments heißt P-*fast sicher *gültig (synonym, vor allem im Fall allgemeiner Maßräume verwendet: P-fast überall gültig), wenn gilt: $\{\omega \in \Omega |\ \Phi(\omega)$ gilt nicht$\}$ ist eine Nullmenge bezüglich P.*

Beispiel: Ist $\omega \in [0,1]$ ein nach der Gleichverteilung P auf $[0,1]$ gezogenes zufälliges Ergebnis, so gilt P-fast sicher: $\omega > 0$. Ebenso gilt für einen nach der Gleichverteilung Q auf dem Quadrat $[-1,1]^2$ zufällig gewählten Punkt (x,y) Q-fast sicher die Aussage $x^2 + y^2 \neq 1$.

Interpretation von Wahrscheinlichkeiten. Als Verbindungsglied zwischen der Mathematik und den Anwendungen brauchen wir *Interpretationen*, was Wahrscheinlichkeiten bedeuten sollen. Diese sind naturgemäß *kein* Bestandteil des mathematischen Formalismus und daher informal formuliert.
Je nach philosophischem Standpunkt kann man verschiedene Interpretationen vertreten.

1. **Objektivistische Interpretation mit relativen Häufigkeiten:** Führt man ein Zufallsexperiment mit Werten in Ω wiederholt aus, sagen wir n-mal, so erhält man Beobachtungsdaten $\omega_1, \omega_2, \ldots, \omega_n \in \Omega$. Die *relative Häufigkeit* eines Ereignisses A ist definiert durch

$$r(\omega_1, \ldots, \omega_n; A) := \frac{|\{i \in [n] |\ \omega_i \in A\}|}{n} = \frac{\text{Anzahl der günstigen Ergebnisse}}{\text{Anzahl aller Ergebnisse}} \tag{40}$$

Objektivistische Interpretation von Wahrscheinlichkeiten:
Führt man ein Zufallsexperiment immer wieder aus, so liegt nach vielen Wiederholungen die relative Häufigkeit eines Ereignisses A typischerweise nahe bei der Wahrscheinlichkeit $P(A)$.

Kritik: Durch die unpräzisen Begriffe "viel", "typischerweise" und "nahe" bleibt diese Interpretation notwendigerweise unscharf.

Ein innermathematisches Analogon dieser Interpretation ist das schwache Gesetz der großen Zahlen, das wir später besprechen.

Ein Versuch, die objektivistische Interpretation schärfer zu fassen, geht auf von Mises zurück:

von-Mises-Interpretation von Wahrscheinlichkeiten:
Bei unendlicher Wiederholung des Zufallsexperiments konvergieren die beobachteten relativen Häufigkeiten gegen die Wahrscheinlichkeit:

$$\lim_{n\to\infty} r(\omega_1, \ldots, \omega_n; A) = P(A).$$

Kritik: Für praktische Zwecke ist diese Interpretation nur wenig nützlich, da man das Zufallsexperiment nie wirklich unendlich oft ausführen kann. Die Interpretation beschreibt also nur idealisierte, nicht real ausführbare unendliche Beobachtungsreihen. Außerdem kann die Interpretation nicht völlig richtig sein, da es ja möglich ist, beim Würfeln stets nur die "6" zu würfeln, auch wenn die Wahrscheinlichkeit für die "6" nur $\frac{1}{6}$ betragen sollte.

Ein innermathematisches Analogon der von-Mises-Interpretation ist das starke Gesetz der großen Zahlen, das wir auch später besprechen.

2. **Subjektivistische Interpretation:** Bei dieser Interpretation bekommen Wahrscheinlichkeiten nur eine subjektive Bedeutung:

   Subjektivistische Interpretation von Wahrscheinlichkeiten:
   Die Wahrscheinlichkeit $P(A)$ quantifiziert meinen Grad von Überzeugtheit vom Eintreten von A.

   Die definierenden Bedingungen für das Wahrscheinlichkeitsmaß P, also die Kolmogoroff-Axiome, werden dann als Konsistenzbedingungen für das System meiner subjektiven Überzeugungen interpretiert.

   Kritik: Naturwissenschaftliche Theorien, die Wahrscheinlichkeiten vorhersagen, z.B. die Quantentheorie, haben den Anspruch, nicht nur subjektive Zustände von Menschen zu beschreiben, sondern objektive Tatsachen, unabhängig von psychischen

Gegebenheiten. Hierfür erscheint die subjektivistische Interpretation zu vage, zu unscharf.

Die folgende *Glückspiel-Interpretation* versucht, die subjektivistische Interpretation quantitativ genauer zu fassen:

Glückspiel-Interpretation von Wahrscheinlichkeiten:
Das Ereignis A hat für mich die subjektive Wahrscheinlichkeit $P(A)$, wenn ich bereit bin, jede der folgenden beiden Wetten einzugehen:

(a)
- wenn das Ereignis A eintritt, *bekomme* ich α Euro,
- wenn das Ereignis A^c eintritt, *zahle* ich β Euro,

wobei

$$\frac{\alpha}{\beta} = \frac{1 - P(A)}{P(A)}.$$

(b) Wie eben, nur mit vertauschten Rollen:

- wenn das Ereignis A eintritt, *zahle* ich α Euro,
- wenn das Ereignis A^c eintritt, *bekomme* ich β Euro.

Kritik: Die Glückspielinterpretation passt gut zur stochastischen Beschreibung von Finanzmärkten, bei denen Kauf- und Verkaufentscheidungen von Anlegern auch nach subjektiven Erwartungen getroffen werden. Subjektive Wahrscheinlichkeiten kann man also mit Preisverhältnissen für Wetten auf zukünftige Ereignisse in Wettbüros quantifizieren.

Allerdings sind reale Glücksspiele viel komplexer als die vereinfachte Sicht der Glückspielinterpretation: Vielleicht bin ich zwar bereit $\alpha = 1$ Euro gegen $\beta = 0,50$ Euro auf ein Ereignis zu wetten, aber nicht $\alpha = 1000000$ Euro gegen $\beta = 500000$ Euro.

In der Realität wollen Anleger auch einen Risikozuschlag: Vielleicht bin ich bereit $\alpha = 1$ Euro gegen $\beta = 0,50$ Euro auf ein Ereignis zu wetten, aber nur $\alpha = 0,90$ Euro zu $\beta = 0,60$ Euro bei vertauschten Rollen.

3. **Minimalinterpretation von Wahrscheinlichkeiten:** Die Probleme bei der Quantifizierung von Wahrscheinlichkeiten motivieren dazu, eine möglichst voraussetzungsarme Interpretation zu suchen:

Minimalinterpretation von Wahrscheinlichkeiten:

- Ereignisse A mit Wahrscheinlichkeiten $P(A)$ nahe bei 1 treten *praktisch sicher* ein, also "*mit an Sicherheit grenzender Wahrscheinlichkeit*".
- Ereignisse A mit Wahrscheinlichkeiten $P(A)$ nahe bei 0 treten *praktisch sicher nicht* ein.
- Wahrscheinlichkeiten $P(A)$, die weder nahe bei 0 noch bei 1 liegen, bedeuten *Unsicherheit*. Ihre genauen Werte sind nur Rechengrößen.

Kritik: Vielleicht erscheint Ihnen diese Interpretation zu wenig quantitativ und zu schwach, um nützlich zu sein. Außerdem verwendet sie wieder unscharfe Begriffe wie "nahe bei" und "praktisch sicher". Wir werden später, zum Beispiel beim schwachen Gesetz der großen Zahlen und bei der statistischen Testtheorie, sehen, dass die Minimalinterpretation trotzdem sehr wohl sinnvoll und nützlich sein kann.

In der praktischen Arbeit, insbesondere in der Statistik, muss man den unscharfen Begriff "nahe bei" natürlich quantifizieren: Hierzu dient die Wahl eines *Signifikanzniveaus* α nahe bei 0: Ein Ereignis A mit einer Wahrscheinlichkeit $P(A) \leq \alpha$ gilt dann als "praktisch unmöglich"; wenn es doch eintritt, kann man die *Hypothese*, dass das verwendete Wahrscheinlichkeitsmodell das Zufallsexperiment beschreibt, *verwerfen*. In der statistischen Praxis gewählte Signifikanzniveaus sind allerdings vielfach gar nicht so klein, z.B. $\alpha = 5\%$.

In jedem Fall kann man sagen, dass der Begriff der Wahrscheinlichkeit ein fundamentaler Grundbegriff ist, der eine ganz andere Modellierung als nur die Beschreibung deterministischer Vorgänge erlaubt. Wie viele andere Grundbegriffe auch kann er philosophisch verschieden interpretiert werden.

Die mathematische Wahrscheinlichkeitstheorie kann zum Glück unabhängig von der Interpretation und vom philosophischen Standpunkt betrieben werden, obwohl der philosophische Standpunkt manchmal sehr wohl die Art, mathematische Fragen zu stellen, beeinflusst.

2.2 Verteilungsfunktion und Eindeutigkeitssatz für Maße

Als Funktionen auf der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R})$ sind Wahrscheinlichkeitsmaße auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ recht komplizierte Gebilde. Zum Glück lässt sich ihre Information auch in einer einfacheren Funktion kodieren:

Definition 2.8 (Verteilungsfunktion) *Es sei P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Die Funktion*

$$F : \mathbb{R} \to [0,1] \quad F(x) = P(]-\infty, x]) \tag{41}$$

wird Verteilungsfunktion *von P genannt.*

Der Wert $F(x)$ der Verteilungsfunktion an einer Stelle x gibt also die Wahrscheinlichkeit an, ein zufälliges Ergebnis kleiner oder gleich x zu erhalten.
Beispiele:

1. Die Verteilungsfunktion der Gleichverteilung auf $[0,1]$ wird wie folgt gegeben

$$F(x) = \begin{cases} 0 & \text{für } x < 0, \\ x & \text{für } 0 \leq x \leq 1, \\ 1 & \text{für } x > 1. \end{cases} \tag{42}$$

2. Die Verteilungsfunktion des Diracmaßes $\delta_a : \mathcal{B}(\mathbb{R}) \to \{0,1\}$ in $a \in \mathbb{R}$ lautet

$$F(x) = \begin{cases} 0 & \text{für } x < a, \\ 1 & \text{für } x \geq a. \end{cases} \tag{43}$$

3. Modellieren wir das Ergebnis eines fairen Münzwurfs durch $(\Omega = \mathbb{R}, \mathcal{A} = \mathcal{B}(\mathbb{R}), P = \frac{1}{2}(\delta_0 + \delta_1))$, so lautet die Verteilungsfunktion für P wie folgt:

$$F(x) = \begin{cases} 0 & \text{für } x < 0, \\ \frac{1}{2} & \text{für } 0 \leq x < 1, \\ 1 & \text{für } x \geq 1. \end{cases} \tag{44}$$

   Mit Hilfe der *Indikatorfunktion*

$$1_A : \Omega \to \{0,1\}, \quad 1_A(\omega) = \begin{cases} 0 & \text{für } \omega \in A, \\ 1 & \text{für } \omega \in A^c \end{cases} \tag{45}$$

   einer Teilmenge $A \subseteq \Omega$ können wir das auch so abkürzen:

$$F = \frac{1}{2}1_{[0,1[} + 1_{[1,\infty[} = \frac{1}{2}1_{[0,\infty[} + \frac{1}{2}1_{[1,\infty[}. \tag{46}$$

Lemma 2.9 (charakteristische Eigenschaften von Verteilungsfunktionen) *Es sei F die Verteilungsfunktion eines Wahrscheinlichkeitsmaßes P über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Dann gilt:*

1. *F ist monoton steigend:*

$$\forall x, y \in \mathbb{R} : \ x \leq y \Rightarrow F(x) \leq F(y). \tag{47}$$

2. *F ist rechtsseitig stetig: Für alle monoton fallenden Folgen $x_n \downarrow_{n\to\infty} x \in \mathbb{R}$ mit reellem Grenzwert x gilt*

$$F(x_n) \overset{n\to\infty}{\longrightarrow} F(x). \tag{48}$$

3. $\lim_{x\to+\infty} F(x) = 1$.

4. $\lim_{x\to-\infty} F(x) = 0$.

Beweis:

1. Aus $x \leq y$ folgt $]-\infty, x] \leq]-\infty, y]$, und hieraus
$$F(x) = P(]-\infty, x]) \leq P(]-\infty, y]) \leq F(y). \tag{49}$$

2. Es sei $(x_n)_{n\in\mathbb{N}}$ eine monoton fallende Folge in $\mathbb{R}$ mit $x_n \overset{n\to\infty}{\longrightarrow} x \in \mathbb{R}$. Dann ist $(]-\infty, x_n])_{n\in\mathbb{N}}$ eine monoton fallende Folge in $\mathcal{B}(\mathbb{R})$ mit dem Durchschnitt
$$\bigcap_{n\in\mathbb{N}}]-\infty, x_n] =]-\infty, x]. \tag{50}$$
Mit der σ-Stetigkeit von P von oben folgt:
$$F(x) = P(]-\infty, x]) = \lim_{n\to\infty} P(]-\infty, x_n]) = \lim_{n\to\infty} F(x_n). \tag{51}$$

3. Es sei $(x_n)_{n\in\mathbb{N}}$ eine monoton steigende Folge in $\mathbb{R}$ mit $x_n \overset{n\to\infty}{\longrightarrow} +\infty$. Dann ist die Intervallfolge $(]-\infty, x_n])_{n\in\mathbb{N}}$ bezüglich der Inklusionsrelation $\subseteq$ aufsteigend mit der Vereinigung
$$\bigcup_{n\in\mathbb{N}}]-\infty, x_n] = \mathbb{R}. \tag{52}$$
Aus der σ-Stetigkeit von P von unten folgt
$$1 = P(\mathbb{R}) = \lim_{n\to\infty} P(]-\infty, x_n]) = \lim_{n\to\infty} F(x_n). \tag{53}$$
Wir schließen
$$F(x) \overset{x\to\infty}{\longrightarrow} F(x). \tag{54}$$

4. Diese Aussage wird analog mit Hilfe der σ-Stetigkeit von P von oben bewiesen; sie wird Ihnen als Übungsaufgabe überlassen.

□

Übung 2.10 *Beweisen Sie in der Situation des Lemmas für alle $x \in \mathbb{R}$:*
$$P(]-\infty, x[) = \lim_{y\uparrow x} F(y).$$

Bemerkung: Wir werden später sehen, dass es zu jeder Funktion $F : \mathbb{R} \to [0,1]$ mit den Eigenschaften 1.–4. des Lemmas ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der Verteilungsfunktion F gibt.
Verteilungsfunktionen charakterisieren das zugehörige Wahrscheinlichkeitsmaß P über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ eindeutig:

Satz 2.11 (Eindeutigkeitssatz für Verteilungsfunktionen) *Es seien P, Q zwei Wahrscheinlichkeitsmaße über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der gleichen Verteilungsfunktion F. Dann gilt $P = Q$.*

Dieser Satz ist ein Spezialfall des allgemeinen Eindeutigkeitssatzes für Wahrscheinlichkeitsmaße. Um diesen formulieren zu können, brauchen wir noch einige Vorbereitungen:

Definition 2.12 (Durchschnittsstabilität) *Es sei Ω ein Ergebnisraum und $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ ein Mengensystem darüber. Das System $\mathcal{M}$ heißt durchschnittstabil, (kurz $\cap$-stabil, synonym Π-System), wenn für alle $A, B \in \mathcal{M}$ gilt: $A \cap B \in \mathcal{M}$. Ist $(\Omega, \mathcal{A})$ ein Ereignisraum und $\mathcal{M} \subseteq \mathcal{A}$ ein $\cap$-stabiles Mengensystem mit $\sigma(\mathcal{M}) = \mathcal{A}$, so wird $\mathcal{M}$ ein $\cap$-stabiler Erzeuger von $\mathcal{A}$ genannt.*

Zum Beispiel bildet $\mathcal{M} = \{]-\infty, x] |\ x \in \mathbb{R}\}$ einen $\cap$-stabilen Erzeuger der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R})$, denn in der Tat ist $\sigma(\mathcal{M}) = \mathcal{B}(\mathbb{R})$, und für alle $x, y \in \mathbb{R}$ gilt

$$]-\infty, x] \cap]-\infty, y] =]-\infty, x \wedge y] \in \mathcal{M}, \tag{55}$$

wobei wir die folgende Notation verwendet haben:

$$x \wedge y := \min\{x, y\}. \tag{56}$$

Damit folgt der Eindeutigkeitssatz für Verteilungsfunktionen aus dem folgenden allgemeinen Eindeutigkeitssatz für Wahrscheinlichkeitsmaße:

Satz 2.13 (Eindeutigkeitssatz für Wahrscheinlichkeitsmaße) *Es seien P, Q zwei Wahrscheinlichkeitsmaße über dem gleichen Ereignisraum $(\Omega, \mathcal{A})$. Weiter sei $\mathcal{M}$ ein $\cap$-stabiler Erzeuger von $\mathcal{A}$, auf dem P und Q übereinstimmen: Dann gilt $P = Q$.*

Bemerkung: Für allgemeine Maße gilt der Eindeutigkeitssatz nicht in dieser Form: Zum Beispiel stimmen das Lebesguemaß λ und sein Doppeltes 2λ nicht überein, obwohl sie auf dem $\cap$-stabilen Erzeuger $\{]-\infty, x] |\ x \in \mathbb{R}\}$ der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R})$ übereinstimmen, nämlich mit dem konstanten Wert $+\infty$.
Zum Beweis des Eindeutigkeitssatzes führen wir die folgende Abschwächung des Begriffs der σ-Algebra ein:

Definition 2.14 (Dynkin-System) Es sei Ω eine Menge. Ein Mengensystem $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ heißt *Dynkin-System* über Ω, wenn gilt:

1. $\emptyset \in \mathcal{D}$
2. Für alle $A \in \mathcal{D}$ folgt $A^c \in \mathcal{D}$.
3. Für jede Folge $(A_n)_{n\in\mathbb{N}}$ *paarweise disjunkter* Mengen in $\mathcal{D}$ gilt $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{D}$.

Man beachte: Anders als in der Definition von σ-Algebren wird hier in der dritten Bedingung die paarweise Disjunktheit der Mengen A_n, $n \in \mathbb{N}$, vorausgesetzt!
Weil die Voraussetzung der paarweisen Disjunktheit 3. in der Definition von Dynkin-Systemen gut zur Voraussetzung der paarweisen Disjunktheit bei der Formulierung der σ-Additivität passt, ist es im Zusammenhang mit Maßen oft leichter, zu überprüfen, dass ein Mengensystem ein Dynkin-System ist, als zu zeigen, dass es eine σ-Algebra ist.
Hierzu ein Beispiel:

Lemma 2.15 (Übereinstimmungsbereich von Wahrscheinlichkeitsmaßen) *Sind P und Q Wahrscheinlichkeitsmaße über einem Ereignisraum $(\Omega, \mathcal{A})$, so ist*

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

ein Dynkin-System.

Beweis: 1. $\emptyset \in \mathcal{D}$ folgt wegen $P(\emptyset) = 0 = Q(\emptyset)$.
2. Gegeben $A \in \mathcal{D}$ folgt $P(A^c) = P(\Omega) - P(A) = 1 - P(A) = 1 - Q(A) = Q(\Omega) - Q(A) = Q(A^c)$, also auch $A^c \in \mathcal{D}$.
3. Ist eine Folge $(A_n)_{n\in\mathbb{N}}$ paarweise disjunkter Mengen in $\mathcal{D}$ gegeben und setzen wir $A = \bigcup_{n\in\mathbb{N}} A_n$, so folgt wegen der σ-Additiviät von P und Q:

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

also auch $A \in \mathcal{D}$.

□

Lemma 2.16 (Dynkin-Lemma, auch Π-Λ-Theorem genannt) *Es sei Ω eine nichtleere Menge, $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ ein durchschnittstabiles Mengensystem und $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ ein Dynkin-System über Ω. Dann gilt:*
Aus $\mathcal{M} \subseteq \mathcal{D}$ folgt $\sigma(\mathcal{M}) \subseteq \mathcal{D}$.

Das Dynkin-Lemma wird zwar in der Stochastik oft benutzt, obwohl sein Beweis eher zur Maßtheorie gehört. Anders als die *Aussage* des Lemmas gehört sein *Beweis* nicht zum Kernbestand dieser Vorlesung. Er befindet sich im Anhang.

Übung 2.17 *Es seien $\Omega = [4]$ und $\mathcal{A} = \mathcal{P}(\Omega)$. Der Übereinstimmungsbereich*

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

der beiden Wahrscheinlichkeitsmaße

$$P = \frac{1}{4}(\delta_1 + \delta_2 + \delta_3 + \delta_4) \quad \textit{und} \quad Q = \frac{1}{2}(\delta_1 + \delta_2) \tag{57}$$

ist zwar ein Dynkin-System, aber keine σ-Algebra. Außerdem ist $\mathcal{D}$ nicht $\cap$-stabil.

Wir zeigen jetzt den Eindeutigkeitssatz für Wahrscheinlichkeitsmaße mit Hilfe des Dynkin Lemmas:
Beweis von Satz 2.13: Sine P und Q zwei Wahrschinlichkeitsmaße über $(\Omega, \mathcal{A})$ und $\mathcal{M}$ ein $\cap$-stabiler Erzeuger von $\mathcal{A}$ mit $P|_{\mathcal{M}} = Q|_{\mathcal{M}}$, so folgt

$$\mathcal{M} \subseteq \mathcal{D} = \{A \in \mathcal{A} |\ P(A) = Q(A)\}.$$

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

□

Als eine weitere Illustration, wie das Dynkin-Lemma typischerweise eingesetzt wird, untersuchen wir nun die Rotationsinvarianz der Gleichverteilung auf S^1.
Für $t \in \mathbb{R}$ sei

$$D_t = \begin{pmatrix} \cos(2\pi t) & -\sin(2\pi t) \\ \sin(2\pi t) & \cos(2\pi t) \end{pmatrix} \tag{58}$$

die Drehmatrix um den Winkel $2\pi t$, und für $A \subseteq S^1$ sei

$$D_t A = \left\{ D_t \begin{pmatrix} x \\ y \end{pmatrix} :\ \begin{pmatrix} x \\ y \end{pmatrix} \in A \right\} \tag{59}$$

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

Lemma 2.18 (Rotationsinvarianz der Gleichverteilung auf der Kreislinie) *Es sei P die Gleichverteilung auf $(S^1, \mathcal{B}(S^1))$. Für alle $A \in \mathcal{B}(S^1)$ und alle $t \in \mathbb{R}$ gilt dann*

$$P(A) = P(D_t A). \tag{60}$$

Beweisskizze: Die Aussage ist offensichtlich richtig für alle A der Gestalt

$$A = \left\{ \begin{pmatrix} \cos(2\pi x) \\ \sin(2\pi x) \end{pmatrix} :\ x \in I \right\}, \tag{61}$$

wenn I ein Intervall in $[0, 1[$ (inklusive der leeren Menge) ist. Nun sei $\mathcal{M}$ die Menge aller A dieser Gestalt. Dann ist $\mathcal{M}$ ein $\cap$-stabiler Erzeuger von $\mathcal{B}(S^1)$. Weiter kann man zeigen, dass für gegebenes $t \in \mathbb{R}$ die Menge

$$\mathcal{D} := \{A \in \mathcal{B}(S^1) :\ P(D_t) = P(A)\} \tag{62}$$

ein Dynkin-System ist, das $\mathcal{M}$ umfaßt. Es folgt aus dem Dynkin-Lemma:

$$\mathcal{B}(S^1) = \sigma(\mathcal{M}) \subseteq \mathcal{D}, \tag{63}$$

also die Behauptung.

□

Wir kommen jetzt zu dem "negativen" Grund zurück, σ-Algebren $\mathcal{A}$ über Ω statt nur der Potenzmenge $\mathcal{P}(\Omega)$ zu betrachten:
Die Gleichverteilung auf $(S^1, \mathcal{B}(S^1))$ kann nicht zu einem rotationsinvarianten Maß auf $(S^1, \mathcal{P}(S^1))$ fortgesetzt werden. Es gilt nämlich:

Satz 2.19 (Nichtexistenz der Gleichverteilung auf der Potenzmenge der Kreislinie) *Es gibt kein rotationsinvariantes Wahrscheinlichkeitsmaß Q auf $(S^1, \mathcal{P}(S^1))$, also kein Wahrscheinlichkeitsmass $Q : \mathcal{P}(S^1) \to [0,1]$, für das gilt:*

$$\forall A \subseteq S^1 \; \forall t \in \mathbb{R} : \; Q(A) = Q(D_t A). \tag{64}$$

Beweis: Wir definieren die folgende Relation $\sim$ auf S^1: Für $x, y \in S^1$ soll $x \sim y$ genau dann gelten, wenn es ein $t \in \mathbb{Q}$ mit der Eigenschaft $D_t x = y$ gibt, d.h. wenn der Winkel $\measuredangle x0y$ ein rationales Vielfaches von 2π ist. Dann ist $\sim$ eine Äquivalenzrelation. Es sei $[x] = \{y \in S^1 |\; x \sim y\}$ die Äquivalenzklasse von $x \in S^1$, und $S^1/\sim := \{[x] |\; x \in S^1\}$ die Menge aller Äquivalenzklassen. Wir wählen eine Auswahlfunktion $f : S^1/\sim \to S^1$, also eine Abbildung, die aus jeder Äquivalenzklasse ein Element auswählt: $f([x]) \sim x$ für alle $x \in S^1$. Die Existenz einer solchen Auswahlfunktion wird durch das Auswahlaxiom der Mengenlehre garantiert. Nun sei

$$A = \operatorname{Bild} f = \{f([x]) |\; x \in S^2\}. \tag{65}$$

Dann ist $(D_t A)_{t \in \mathbb{Q} \cap [0,1[}$ eine Partition der Kreislinie S^1 mit abzählbar vielen Mengen. Wäre nun Q ein rotationsinvariantes Wahrscheinlichkeitsmaß auf $(S^1, \mathcal{P}(S^1))$, so folgte:

$$1 = Q(S^1) = Q\left(\bigcup_{t \in \mathbb{Q} \cap [0,1[} D_t A\right) = \sum_{t \in \mathbb{Q} \cap [0,1[} Q(D_t A) = \sum_{t \in \mathbb{Q} \cap [0,1[} Q(A). \tag{66}$$

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

□

2.2.1 Anhang: Beweis des Dynkin-Lemmas

Beweis des Lemmas 2.16:

1. Schritt. Ist $\mathcal{D}_1$ ein Dynkin-System über Ω, so gilt:
Für alle $A, B \in \mathcal{D}_1$ mit $A \subseteq B$ folgt $B \setminus A \in \mathcal{D}_1$.
Beweis hierzu: Die Mengen B^c und A sind disjunkt. Nun gilt $B^c \in \mathcal{D}_1$ wegen $B \in \mathcal{D}_1$. Zusammen mit $A \in \mathcal{D}_1$ folgt

$$(B \setminus A)^c = A \cup B^c \cup \emptyset \cup \emptyset \cup \emptyset \cup \ldots \in \mathcal{D}_1$$

und damit $B \setminus A \in \mathcal{D}_1$.

2. Schritt. Es gibt ein kleinstes Dynkin-System $\mathcal{G}$ über Ω, das $\mathcal{M}$ umfasst, nämlich den Durchschnitt aller Dynkin-Systeme über Ω, die $\mathcal{M}$ umfassen:

$$\mathcal{G} = \{A \subseteq \Omega \mid \text{Für jedes Dynkin-System } \mathcal{E} \text{ über } \Omega \text{ mit } \mathcal{M} \subseteq \mathcal{E} \text{ gilt } A \in \mathcal{E}.\}$$

3. Schritt. Wir setzen:

$$\mathcal{G}_1 = \{A \in \mathcal{G} \mid \forall B \in \mathcal{M} : \ A \cap B \in \mathcal{G}\}$$

Dann ist $\mathcal{G}_1$ ein Dynkin-System über Ω, das $\mathcal{M}$ umfasst: $\mathcal{M} \subseteq \mathcal{G}_1$.
Beweis hierzu:

- Es gilt $\emptyset \in \mathcal{G}_1$, da für alle $B \in \mathcal{M}$ gilt: $\emptyset \cap B = \emptyset \in \mathcal{G}$.
- Ist $A \in \mathcal{G}_1$, so folgt für alle $B \in \mathcal{M}$ mit Hilfe des 1. Schritts:

  $$A^c \cap B = B \setminus (A \cap B) \in \mathcal{G},$$

  denn es gelten $B \in \mathcal{G}$ wegen $B \in \mathcal{M}$, $A \cap B \in \mathcal{G}$ wegen $A \in \mathcal{G}_1$, und $A \cap B \subseteq B$. Wir schließen: $A^c \in \mathcal{G}_1$.
- Ist $(A_n)_{n\in\mathbb{N}}$ eine Folge paarweise disjunkter Elemente von $\mathcal{G}_1$, so folgt für alle $B \in \mathcal{M}$:

  $$\left(\bigcup_{n\in\mathbb{N}} A_n\right) \cap B = \bigcup_{n\in\mathbb{N}} (A_n \cap B) \in \mathcal{G},$$

  denn die $A_n \cap B$, $n \in \mathbb{N}$, sind paarweise disjunkte Elemente von $\mathcal{G}$. Wir schließen: $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{G}_1$. item Es gilt $\mathcal{M} \subseteq \mathcal{G}_1$. Zum Beweis hiervon sei $A \in \mathcal{M}$ gegeben. Dann folgt für alle $B \in \mathcal{M}$ die Aussage $A \cap B \in \mathcal{M} \subseteq \mathcal{G}$, weil $\mathcal{M}$ durchschnittstabil ist. Das bedeutet $A \in \mathcal{G}_1$.

4. Schritt. Weil $\mathcal{G}_1 \subseteq \mathcal{G}$ ein Dynkin-System über Ω mit $\mathcal{M} \subseteq \mathcal{G}_1$ ist, folgt auch $\mathcal{G}_1 \supseteq \mathcal{G}$, weil $\mathcal{G}$ das *kleinste* Dynkin-System über Ω ist, das $\mathcal{M}$ als Teilmenge hat. Das bedeutet: $\mathcal{G}_1 = \mathcal{G}$. Nach der Definition von $\mathcal{G}_1$ impliziert das für alle $A \in \mathcal{G}$ und alle $B \in \mathcal{M}$ die Aussage $A \cap B \in \mathcal{G}$.

5. Schritt. Wir setzen:

$$\mathcal{G}_2 := \{B \in \mathcal{G} \mid \forall A \in \mathcal{G} : \ A \cap B \in \mathcal{G}\}$$

Dann ist $\mathcal{G}_2$ ein Dynkin-System über Ω mit $\mathcal{M} \subseteq \mathcal{G}_2$.
Der *Beweis hierzu* ähnelt ein wenig dem Beweis im 3. Schritt:

- Es gilt $\emptyset \in \mathcal{G}_2$, da für alle $A \in \mathcal{G}$ gilt: $A \cap \emptyset = \emptyset \in \mathcal{G}$.

- Es sei $B \in \mathcal{G}_2$ gegeben. Dann folgt für alle $A \in \mathcal{G}$ mit Hilfe des 1. Schritts:

$$A \cap B^c = A \setminus (A \cap B) \in \mathcal{G},$$

da $A \cap B \in \mathcal{G}$ wegen $B \in \mathcal{G}_2$. Dies impliziert $B^c \in \mathcal{G}_2$.

- Ist $(B_n)_{n\in\mathbb{N}}$ eine Folge paarweise disjunkter Elemente von $\mathcal{G}_2$, so folgt für alle $A \in \mathcal{G}$:

$$A \cap \bigcup_{n\in\mathbb{N}} B_n = \bigcup_{n\in\mathbb{N}} (A \cap B_n) \in \mathcal{G},$$

denn die Mengen $A \cap B_n$, $n \in \mathbb{N}$, sind paarweise disjunkte Elemente von $\mathcal{G}$. Wir schließen: $\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{G}_2$.

- $\mathcal{M} \subseteq \mathcal{G}_2$ ist klar nach dem vierten Schritt.

6. Schritt. Wie im 4. Schritt folgt $\mathcal{G}_2 = \mathcal{G}$, also ist das Mengensystem $\mathcal{G}$ durchschnittstabil.

7. Schritt. Jedes durchschnittstabile Dynkin-System ist eine σ-Algebra. Insbesondere ist $\mathcal{G}$ eine σ-Algebra.
Beweis hierzu: Es sei $\mathcal{G}'$ ein durchschnittstabiles Dynkin-System. Weil $\mathcal{G}'$ als Dynkin-System die leere Menge als Element enthält und abgeschlossen unter Komplementbildung ist, bleibt nur zu zeigen, dass $\mathcal{G}'$ auch abgeschlossen unter abzählbarer Vereinigungsbildung ist. Hierzu sei $(A_n)_{n\in\mathbb{N}}$ eine Folge von Elementen von $\mathcal{G}'$. Wir setzen für $n \in \mathbb{N}$:

$$B_n = A_n \cap \bigcap_{m<n} A_m^c.$$

Dann gilt $B_n \in \mathcal{G}'$, weil $\mathcal{G}'$ durchschnittstabil und abgeschlossen unter Komplementbildung ist. Nun sind die B_n, $n \in \mathbb{N}$, paarweise disjunkt, und es gilt

$$\bigcup_{n\in\mathbb{N}} A_n = \bigcup_{n\in\mathbb{N}} B_n \in \mathcal{G}'.$$

Das Mengensystem $\mathcal{G}'$ ist also in der Tat abgeschlossen unter abzählbarer Vereinigungsbildung.

8. Schritt. Aus $\mathcal{M} \subseteq \mathcal{G}$ folgt $\sigma(\mathcal{M}) \subseteq \mathcal{G}$, weil $\mathcal{G}$ nach dem 7. Schritt eine σ-Algebra ist und $\sigma(\mathcal{M})$ die *kleinste* σ-Algebra ist, die $\mathcal{M}$ umfasst. Weiter wissen wir $\mathcal{G} \subseteq \mathcal{D}$, weil $\mathcal{G}$ das *kleinste* Dynkin-System ist, das $\mathcal{M}$ umfasst. Es folgt die Behauptung: $\sigma(\mathcal{M}) \subseteq \mathcal{D}$.

□

2.3 Borel-messbare Funktionen und Maße mit Dichten

Zur Erweiterung unseres Beispielvorrats für kontinuierliche Wahrscheinlichkeitsmaße führen wir den Begriff der Wahrscheinlichkeitsdichte ein. Dichten sind sogenannte *Borel-messbare* Funktionen mit Werten in $[0, \infty]$. Definieren wir hierzu:

Definition 2.20 (Borel-Messbarkeit) *Es sei $(\Omega, \mathcal{A})$ ein Ereignisraum. Eine Funktion $f : \Omega \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ heißt* Borel-messbar *beüglich $\mathcal{A}$ (kurz auch nur:* messbar*), wenn für alle $a \in \mathbb{R}$ gilt:*

$$\{\omega \in \Omega : f(\omega) \leq a\} \in \mathcal{A}. \tag{67}$$

Dies ist ein Spezialfall des allgemeinen Begriffs messbarer Funktionen, den wir später behandeln.

Beispiele: In der Maßtheorie wird gezeigt:

1. Alle stetigen oder auch stückweise stetigen Funktionen $f : \mathbb{R} \to \overline{\mathbb{R}}$ sind Borel-messbar über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

2. Ist $A \in \mathcal{A}$, so ist die Indikatorfunktion $1_A : \Omega \to \mathbb{R}$ messbar.

3. Sind $f, g : \Omega \to \mathbb{R}$ messbare Funktionen, so ist auch jede Linearkombination $\alpha f + \beta g$ mit Koeffizienten $\alpha, \beta \in \mathbb{R}$ sowie die Produktfunktion $f \cdot g$ messbar.

4. Ist $(f_n)_{n\in\mathbb{N}}$ eine punktweise konvergente Folge messbarer Funktionen, so ist auch der punktweise Grenzwert $\lim_{n\to\infty} f_n$ messbar.

In der Masstheorie wird auch für jede nichtnegative messbare Funktion $f : \Omega \to [0, \infty]$ und jedes Maß μ auf $(\Omega, \mathcal{A})$ ein *Integral* so definiert:

$$\int_\Omega f \, d\mu := \tag{68}$$

$$\sup \left\{ \sum_{i=1}^n \alpha_i \mu(A_i) \,\middle|\, n \in \mathbb{N}, \alpha_i \geq 0, A_i \in \mathcal{A} \text{ für alle } i \in [n], \sum_{i=1}^n \alpha_i 1_{A_i} \leq f \right\} \tag{69}$$

Es wird durch folgende Eigenschaften charakterisiert:

1. **Integral von nichtnegativen Treppenfunktionen mit messbaren Stufen:** Für alle $n \in \mathbb{N}$, alle $\alpha_1, \ldots, \alpha_n \geq 0$ und alle $A_1, \ldots, A_n \in \mathcal{A}$ gilt:

$$\int_\Omega \left(\sum_{i=1}^n \alpha_i 1_{A_i} \right) d\mu = \sum_{i=1}^n \alpha_i \mu(A_i) \tag{70}$$

2. **Satz von der monotonen Konvergenz:** Für jede aufsteigende Folge $0 \leq f_1 \leq f_2 \leq f_3 \leq \ldots$ messbarer Funktionen $f_n : \Omega \to [0, \infty]$ mit punktweisem Grenzwert f, d.h.

$$f(\omega) = \lim_{n\to\infty} f_n(\omega) \text{ für alle } \omega \in \Omega \tag{71}$$

gilt:

$$\int_\Omega f\,d\mu = \lim_{n\to\infty} \int_\Omega f_n\,d\mu. \tag{72}$$

Wir beweisen das hier nicht, sondern verweisen auf die Maßtheorievorlesung.

Beispiele:

1. **Integrale bzgl. Zählmaß = Summen:** Ist μ das Zählmaß auf einer abzählbaren Menge Ω, so gilt für jede Funktion $f : \Omega \to [0, \infty]$:

$$\int_\Omega f\,d\mu = \sum_{\omega\in\Omega} f(\omega). \tag{73}$$

   Integrale bezüglich des Zählmaßes sind also das Gleiche wie Summen. In diesem Sinne ist der Integralbegriff eine Verallgemeinerung des Reihenbegriffs.

2. **Integrale auf abzählbaren Mengen:** Etwas allgemeiner: Ist Ω abzählbar (versehen mit der Potenzmenge als σ-Algebra) und ist μ ein Maß über Ω mit Zähldichte $(p_\omega)_{\omega\in\Omega}$, so gilt für jede Funktion $f : \Omega \to [0, \infty]$ mit nichtnegativen Werten:

$$\int_\Omega f\,d\mu = \sum_{\omega\in\Omega} f(\omega)p_\omega. \tag{74}$$

   Integrale auf abzählbaren Mengen sind also mit der Zähldichte gewichtete Summen.

3. **Riemann-Integrale als spezielle Lebesgueintegrale:** Ist $f : \mathbb{R} \to \mathbb{R}_0^+$ eine sückweise stetige nichtnegative Funktion, so existiert das (uneigentliche) Riemannintegral

$$\int_{-\infty}^{\infty} f(x)\,dx \in [0, \infty], \tag{75}$$

   und es gilt:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{\mathbb{R}} f\,d\lambda, \tag{76}$$

   wobei λ das Lebesguemaß bezeichnet. Das Integral bezüglich des Lebesguemaßes λ wird auch *Lebesgueintegral* genannt. In diesem Sinne verallgemeinert das Lebesgueintegral das Riemannintegral.

Notation: Für $A \in \mathcal{A}$ verwendet man auch die Schreibweisen für das Integral:

$$\int_A f(x)\,\mu(dx) = \int_A f\,d\mu := \int_\Omega f1_A\,d\mu. \tag{77}$$

Bemerkungen: Das Integral besitzt auch die folgenden Eigenschaften:

1. **Integrale sehen Nullmengen nicht:** Sind $f, g : \Omega \to [0, \infty]$ messbar und μ-fast überall gleich, d.h. ist

$$\mu(\{\omega \in \Omega : \ f(\omega) = g(\omega)) = 0, \tag{78}$$

so stimmen die Integrale über f und g überein:

$$\int_\Omega f \, d\mu = \int_\Omega g \, d\mu. \tag{79}$$

2. **Monotonie:** Sind $f, g : \Omega \to [0, \infty]$ nichtnegative messbare Funktionen mit $f \leq g$, so gilt auch

$$\int_\Omega f \, d\mu \leq \int_\Omega g \, d\mu. \tag{80}$$

Mit Hilfe von Integralen können wir unseren Beispielvorrat für Wahrscheinlichkeitsmaße insbesondere im Kontinuum enorm erweitern:

Satz 2.21 (Wahrscheinlichkeitsdichten) *Es sei μ ein Maß auf einem Ereignisraum $(\Omega, \mathcal{A})$ und $f : \Omega \to [0, \infty]$ messbar mit*

$$\int_\Omega f \, d\mu = 1. \tag{81}$$

Dann wird durch

$$P : \mathcal{A} \to [0, 1], \tag{82}$$

$$P(A) = \int_A f \, d\mu \tag{83}$$

ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$ definiert. Wir sagen dann, das Wahrscheinlichkeitsmaß P besitze eine Wahrscheinlichkeitsdichte *f bezüglich μ Verzichtet man auf die Normierungsbedingung* (81)*, so wird die Abbildung $\mathcal{A} \ni A \mapsto \int_A f \, d\mu \in [0, \infty]$ nur ein Maß. In diesem Fall heißt f eine* Dichte *dieses Maßes.*

Im Spezialfall $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, versehen mit dem Lebesguemaß $\mu = \lambda$, erwähnt man meist nicht explizit, dass das Lebesguemaß als Referenzmaß gemeint ist. Das gleiche gilt in n Dimensionen $(\Omega, \mathcal{A}, \mu) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \lambda_n)$.
Anschaulich gesprochen beschreibt die Wahrscheinlichkeitsdichte f die "*Wahrscheinlichkeit pro Längeneinheit*". Anders als das Wahrscheinlichkeitsmaß können Wahrscheinlichkeitsdichten auch Werte größer als 1 annehmen.

Der **Beweis des Satzes 2.21** folgt unmittelbar aus den Eigenschaften des Integrals:

1. **Normierung:** $P(\Omega) = \int_\Omega f \, d\mu = 1.$

2. **σ-Additivität:** Die σ-Additivität von P ist eine Folge des Satzes von der monotonen Konvergenz. Das sieht man so: Ist $(A_n)_{n\in\mathbb{N}}$ eine Folge paarweise disjunkter Ereignisse mit Vereinigung

$$A = \bigcup_{n\in\mathbb{N}} A_n, \tag{84}$$

so folgt:

$$1_A = \sum_{n\in\mathbb{N}} 1_{A_n} \tag{85}$$

wegen der paarweisen Disjunktheit, und daher

$$\begin{aligned}
P(A) &= \int_\Omega 1_A f\, d\mu = \\
&= \int_\Omega \sum_{n\in\mathbb{N}} 1_{A_n} f\, d\mu \\
&= \int_\Omega \lim_{m\to\infty} \sum_{n=1}^{m} 1_{A_n} f\, d\mu \qquad \text{(Limes punktweise gemeint)} \\
&= \lim_{m\to\infty} \int_\Omega \sum_{n=1}^{m} 1_{A_n} f\, d\mu \qquad \text{(mit dem Satz v.d. monontonen Konvergenz)} \\
&= \lim_{m\to\infty} \sum_{n=1}^{m} \int_\Omega 1_{A_n} f\, d\mu \qquad \text{(mit der Linearität des Integrals)} \\
&= \sum_{n\in\mathbb{N}} \int_{A_n} f\, d\mu = \sum_{n\in\mathbb{N}} P(A_n).
\end{aligned} \tag{86}$$

□

Beispiele

1. **Dichte der Gleichverteilung:** Die Gleichverteilung auf einem Intervall $[a, b]$ der Länge $b - a > 0$ besitzt die Wahrscheinlichkeitsdichte

$$\frac{1_{[a,b]}}{b-a} \tag{87}$$

(bezüglich des Lebesguemaßes λ). Anschaulich gesprochen bekommen Borelmengen im Intervall $[a, b]$ die "*Wahrscheinlichkeit* $1/(b-a)$ *pro Längeneinheit*" und im Komplement $[a, b]^c$ die Wahrscheinlichkeit 0. Ebenso sind $1_{]a,b]}/(b-a)$ und $1_{]a,b[}/(b-a)$ Dichten dieser Gleichverteilung; sie unterscheiden sich ja nur auf Lebesgue-Nullmengen.

2. **Zähldichte = Dichte bzgl. Zählmaß:** Ist Ω abzählbar und ist

$$P = \sum_{\omega \in \Omega} p_\omega \delta_\omega : \mathcal{P}(\Omega) \to [0,1] \tag{88}$$

ein Wahrscheinlichkeitsmaß mit Zähldichte $(p_\omega)_{\omega\in\Omega}$, so ist die Zähldichte eine Wahrscheinlichkeitsdichte von P bezüglich des Zählmaßes. Das erklärt den Namen "Zähldichte".

3. **Diracmaß.** Das Diracmaß δ_a in einem Punkt a besitzt *keine* Dichte bezüglich des Lebesguemaßes λ.[6] Wäre nämlich f so eine Dichte, so folgte der folgende Widerspruch, weil $\{a\}$ eine Lebesgue-Nullmenge ist:

$$1 = \delta_a(\mathbb{R}) = \int_{\mathbb{R}} f \, d\lambda = \int_{\mathbb{R}} 1_{\{a\}^c} f \, d\lambda = \delta_a(\{a\}^c) = 0. \tag{89}$$

4. **Die Exponentialverteilung.** Für $a > 0$ sei P_a das Maß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der Dichte

$$f_a(x) = 1_{[0,\infty[}(x) a e^{-ax}, \quad (x \in \mathbb{R}). \tag{90}$$

Es heißt die *Exponentialverteilung* mit dem Parameter a.

5. **Die Standardnormalverteilung.** Aus der Analysis kennen Sie das "*Gaußsche Integral*"

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \, dx = \sqrt{2\pi}. \tag{91}$$

Das Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der Dichte

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad (x \in \mathbb{R}) \tag{92}$$

spielt in der Stochastik eine äußerst wichtige Rolle. Es heißt die *Standardnormalverteilung*. Es wird uns noch viel begegnen.

6. **Gammaverteilungen.** Es seien zwei Parameter $a > 0$ und $s > 0$ gegeben. Weiter bezeichne

$$\Gamma(s) = \int_0^\infty y^{s-1} e^{-y} \, dy \tag{93}$$

[6]Dennoch ist es in der Physik üblich, trotzdem symbolisch mit einer "Dichte" $\delta(x)$ von δ_0 zu arbeiten; diese eigentlich nicht existierende Dichte heißt dort "Diracsche Deltafunktion". Indem man solche Schreibweisen in Maßnotation oder verwandte Konzepte (z.B. Distributionen, also Linearformen auf Funktionenräumen) übersetzt, kann man ihnen in den meisten Fällen eine mathematische Interpretation geben.

die Gammafunktion an der Stelle s. Das Wahrscheinlichkeitsmass Gamma(a, s) auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der Dichte

$$f_{a,s}(x) = 1_{]0,\infty[}(x)\frac{a^s}{\Gamma(s)}x^{s-1}e^{-ax}, \qquad (x \in \mathbb{R}) \tag{94}$$

wird die *Gammaverteilung* mit dem Skalenparameter a und dem Formparameter s bezeichnet. Insbesondere ist $\Gamma(a, 1)$ die Exponentialverteilung mit dem Parameter a.

Zusammenhang zwischen Dichten und Verteilungsfunktionen. Ist P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit der Dichte f, so wird die Verteilungsfunktion F von P wie folgt gegeben:

$$F(a) = P(]-\infty, a]) = \int_{]-\infty,a]} f(x)\,dx = \int_{-\infty}^{a} f(x)\,dx, \qquad (a \in \mathbb{R}), \tag{95}$$

wobei das letzte Gleichheitszeichen nur ein Notationswechsel ist. Aus dem Hauptsatz der Differential- und Integralrechnung folgt: Ist P ein Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R})$ mit stetig differenzierbarer Verteilungsfunktion F, so ist $f = F'$ eine Dichte von P. Das Gleiche gilt, wenn F stetig und stückweise stetig differenzierbar ist; auf die Werte der Dichte bei den Definitionslücken der Ableitung kommt es hierbei nicht an, weil sie eine Lebesgue-Nullmenge bilden.

Beispiel: Die Exponentialverteilung mit dem Parameter $a > 0$ besitzt die Verteilungsfunktion

$$F_a(t) = \int_{]-\infty,t]} 1_{[0,\infty[}(x)ae^{-ax}\,dx = 1_{[0,\infty[}(t)(1 - e^{-at}), \qquad (t \in \mathbb{R}). \tag{96}$$

In der Tat ist F_a stetig und stückweise stetig differenzierbar mit der Ableitung

$$f_a(t) = F_a'(t) = \begin{cases} 0 & \text{für } t < 0, \\ ae^{-at} & \text{für } t > 0, \\ \text{undefiniert} & \text{für } t = 0. \end{cases} \tag{97}$$

Diese Funktion, genauer gesagt jede beliebige Fortsetzung davon auf ganz $\mathbb{R}$, ist eine Dichte der Exponentialverteilung mit dem Parameter $a > 0$. Man beachte, dass die Menge $\{0\}$, auf der der Wert der Dichte beliebig fortgesetzt werden durfte, eine Lebesgue-Nullmenge bildet.

2.4 Allgemeine messbare Funktionen und Zufallsvariablen

Wir verallgemeinern nun den Begriff der Borel-messbaren Funktionen wesentlich:

Definition 2.22 (messbare Abbildungen) *Es seien $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$ zwei Ereignisräume. Eine Abbildung $f : \Omega \to \Omega'$) heißt $\mathcal{A}$-$\mathcal{A}'$-*messbar *(oder kurz* messbar*, wenn klar ist, welche σ-Algebren gemeint sind), wenn für alle $A' \in \mathcal{A}'$ gilt:*[7]

$$f^{-1}[A'] := \{\omega \in \Omega : \ f(\omega) \in A'\} \in \mathcal{A}. \tag{98}$$

Bemerkung: Beachten Sie die Analogie zwischen messbaren und stetigen Abbildungen:

Eine Abbildung $f : M \to N$ zwischen zwei metrischen (oder auch topologischen) Räumen (M, d_M) und (N, d_M) ist *stetig*, wenn für alle *offenen* Mengen $A' \subseteq N$ gilt: Das Urbild $f^{-1}[A'] \subseteq M$ ist *offen*.

Merke:
Stetigkeit: Urbilder *offener* Mengen sind *offen*.
Messbarkeit: Urbilder *messbarer* Mengen sind *messbar*.

Das folgende Kriterium ist oft nützlich zum Nachweis der Messbarkeit:

Lemma 2.23 (Kriterium für Messbarkeit von Abbildungen) *Es sei $f : \Omega \to \Omega'$ eine Abbildung zwischen zwei Ereignisräumen $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$. Weiter sei $\mathcal{M}' \subseteq \mathcal{A}'$ ein Erzeugendensystem für $\mathcal{A}'$, in Formeln: $\sigma(\mathcal{M}') = \mathcal{A}'$. Dann sind äquivalent:*

1. *Die Abbildung 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." ist trivial.
Zu "2. $\Rightarrow$ 1.": Wir setzen

$$\mathcal{B}' = \{A' \in \mathcal{A}' : f^{-1}[A'] \in \mathcal{A}\}. \tag{99}$$

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

- $\Omega' \in \mathcal{B}'$ wegen $f^{-1}[\Omega'] = \Omega \in \mathcal{A}$.
- Für alle $A' \in \mathcal{B}'$ folgt $\Omega' \setminus A' \in \mathcal{B}'$, weil $f^{-1}[A'] \in \mathcal{A}$ impliziert:

$$f^{-1}[\Omega' \setminus A'] = \Omega \setminus f^{-1}[A'] \in \mathcal{A}. \tag{100}$$

- Für jede Folge $(A'_n)_{n\in\mathbb{N}}$ mit Werten in $\mathcal{B}'$ folgt $\bigcup_{n\in\mathbb{N}} A'_n \in \mathcal{B}'$ wegen

$$f^{-1}\left[\bigcup_{n\in\mathbb{N}} A'_n\right] = \bigcup_{n\in\mathbb{N}} \underbrace{f^{-1}[A'_n]}_{\in\mathcal{A}} \in \mathcal{A}. \tag{101}$$

[7] Man unterscheide sorgfältig zwischen dem Urbild $f^{-1}[A]$ (eckige Klammern!) einer *Menge* $A \subseteq \Omega'$ und der Umkehrfunktion $f^{-1}(y)$ (runde Klammern!) bei einem *Wert* $y \in \Omega'$. Während die Urbildabbildung $f^{-1}[\cdot] : \mathcal{P}(\Omega') \to \mathcal{P}(\Omega)$ für *jede* Abbildung $f : \Omega \to \Omega'$ definiert ist, ist die Umkehrfunktion $f^{-1} : \Omega' \to \Omega$ nur für *Bijektionen* $f : \Omega \to \Omega'$ definiert.

Es folgt $\sigma(\mathcal{M}') \subseteq \mathcal{B}'$ und hieraus die Behauptung $\mathcal{B}' = \mathcal{A}'$ wegen

$$\mathcal{A}' = \sigma(\mathcal{M}') \subseteq \mathcal{B}' \subseteq \mathcal{A}'. \tag{102}$$

□

Beispiele:

1. Weil das Mengensystem $\{]-\infty, a] : \ a \in \mathbb{R}\}$ ein Erzeugendensystem der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R})$ ist, ist eine Abbildung $f : \Omega \to \mathbb{R}$ genau dann $\mathcal{A}$-$\mathcal{B}(\mathbb{R})$-messbar, wenn für alle $a \in \mathbb{R}$ gilt:

   $$f^{-1}\big[]-\infty, a]\big] = \{\omega \in \Omega |\ f(\omega) \leq a\} \in \mathcal{A}. \tag{103}$$

   Der allgemeine Messbarkeitsbegriff verallgemeinert also in der Tat den Begriff der Borel-Messbarkeit.

2. Sine (M, d_M) und (N, d_N) zwei metrische (oder topologische) Räume und ist $f : M \to N$ eine stetige Abbildung, so ist f auch $\mathcal{B}(M)$-$\mathcal{B}(N)$-messbar, wobei $\mathcal{B}(M)$ bzw. $\mathcal{B}(N)$ die Borelsche σ-Algebra über M bzw. N bezeichnet. In der Tat erzeugt das System der offenen Mengen über N die σ-Algebra $\mathcal{B}(N)$, und Urbilder offener Mengen in N sind offen in M, also messbar.

3. Jede Abbildung $f : (\Omega, \mathcal{P}(\Omega)) \to (\Omega', \mathcal{A}')$ ist trivialerweise $\mathcal{P}(\Omega)$-$\mathcal{A}$-messbar. Für diskrete Ergebnisräume, versehen mit ihrer Potenzmenge, liefert der Messbarkeitsbegriff also nichts Neues.

Der folgende wichtige Begriff erlaubt es, Maße und insbesondere auch Wahrscheinlichkeitsmaße mit einer messbaren Abbildung vom Ausgangsraum auf den Zielraum zu transportieren:

Satz/Definition 2.24 (Bildmaß) *Es sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum und $f : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ eine messbare Abbildung. Dann wird durch*

$$\begin{aligned} &\nu : \mathcal{A}' \to [0, \infty], \\ &\nu(A') = \mu(f^{-1}[A']) \end{aligned} \tag{104}$$

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

Zusatz: *Ist μ sogar ein Wahrscheinlichkeitsmaß, so ist auch das Bildmaß ein Wahrscheinlichkeitsmaß.*

Beweis:

- Die Messbarkeit von f impliziert die Wohldefiniertheit von ν: Für jedes $A' \in \mathcal{A}'$ ist nämlich $f^{-1}[A']$ ein Element des Definitionsbereichs $\mathcal{A}$ von μ.

- $\nu(\emptyset) = \mu(f^{-1}[\emptyset] = \mu(\emptyset) = 0$.
- Ist $(A'_n)_{n\in\mathbb{N}}$ eine Folge paarweise disjunkter Mengen in $\mathcal{A}'$, so sind auch die Folgenglieder in $(f^{-1}[A'_n])_{n\in\mathbb{N}}$ paarweise disjunkt, und wir erhalten

$$\nu\left(\bigcup_{n\in\mathbb{N}} A'_n\right) = \mu\left(f^{-1}\left[\bigcup_{n\in\mathbb{N}} A'_n\right]\right) = \mu\left(\bigcup_{n\in\mathbb{N}} f^{-1}[A'_n]\right) = \sum_{n\in\mathbb{N}} \mu(f^{-1}[A'_n]) = \sum_{n\in\mathbb{N}} \nu(A'_n). \tag{105}$$

Zum Beweis des Zusatzes bemerken wir für Wahrscheinlichkeitsmaße μ:

$$\nu(\Omega') = \mu(f^{-1}[\Omega']) = \mu(\Omega) = 1. \tag{106}$$

□

Im Fall von Wahrscheinlichkeitsmaßen verwendet man fast immer eine andere Sprechweise:

Definition 2.25 (Zufallsvariablen und Verteilungen) *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $(\Omega', \mathcal{A}')$ ein Ereignisraum. Eine $\mathcal{A}$-$\mathcal{A}'$-messbare Abbildung*

$$X : \Omega \to \Omega' \tag{107}$$

wird auch Zufallsvariable *(engl.: "*random variable*") genannt. Das Bildmaß $X[P] = PX^{-1}$ von P unter X heißt auch* Verteilung *von X (unter P), englisch "*law *of X with respect to P" oder auch "*distribution*", und wird auch mit $\mathcal{L}_P(X)$ oder, wenn klar ist, welches Wahrscheinlichkeitsmaß P gemeint ist, mit $\mathcal{L}(X)$ bezeichnet.*

sehr wichtig!

Zufallsvariablen und ihre Verteilung sind zentrale Begriffe in der Stochastik. Zu Beginn sind diese Begriffe natürlich etwas gewöhnungsbedürftig: Das "Zufällige" an Zufallsvariablen $X : \Omega \to \Omega'$ ist nicht die Abbildung X, sondern ihr Argument $\omega \in \Omega$, also das Ergebnis eines Zufallsexperiments. Die Zufallsvariable X ordnet dem zufälligen Ergebnis $\omega \in \Omega$ also nur eine "Maßgröße" $X(\omega) \in \Omega'$ zu. Der Wert $Q(A')$ der Verteilung $Q = \mathcal{L}_P(X)$ von X bei einer messbare Menge A' gibt an, mit welcher Wahrscheinlichket X einen Wert $X(\omega)$ in A' annimmt, in Formeln:[8]

$$Q(A') = P(\{\omega \in \Omega |\ X(\omega) \in A'\}). \tag{108}$$

[8]Bitte verwechseln Sie nie Zufallsvariablen mit ihrer Verteilung! Beachten Sie, dass verschiedene Zufallsvariablen die gleiche Verteilung besitzen können: Nehmen wir zum Beispiel das zweifache faire Münzwurfexperiment

$$(\Omega, \mathcal{A}, \mathcal{P}) = (\{0,1\}^2, \mathcal{P}(\Omega), \frac{1}{4}(\delta_{(0,0)} + \delta_{(0,1)} + \delta_{(1,0)} + \delta_{(1,1)}))$$

und die beiden Projektionen $X, Y : \Omega \to \{0,1\}$, $X(\omega_1, \omega_2) = \omega_1$, $Y(\omega_1, \omega_2) = \omega_2$ als Modelle für das Ergebnis des ersten bzw. zweiten Wurfs. Hier sind X und Y *verschiedene* Zufallsvariablen mit der gleichen Verteilung $\frac{1}{2}(\delta_0 + \delta_1)$.

Konventionen und Sprechweisen:

- Der häufigste Spezialfall ist $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$: Solche Zufallsvariablen $X : \Omega \to \mathbb{R}$ heißen auch *reelle* oder genauer *reellwertige* Zufallsvariablen. Wenn der Zielraum $(\Omega', \mathcal{A}')$ nicht spezifiziert wird, sind fast immer reelle Zufallsvariablen gemeint.
- Eine Zufallsvariable X mit Verteilung Q heißt auch Q-*verteilte* Zufallsvariable.
- Zufallsvariablen werden oft mit Großbuchstaben $X, Y, Z, \ldots$ bezeichnet.
- Sehr häufig wird die folgende abkürzende Notation für Ereignisse verwendet: Ein Ereignis

$$\{\omega \in \Omega |\ X(\omega) \text{ hat die Eigenschaft } \Phi\} \tag{109}$$

wird abgekürzt so geschrieben:

$$\{X \text{ hat die Eigenschaft } \Phi\}. \tag{110}$$

Das Argument ω wird in dieser stenographischen Schreibweise also weggelassen. Zum Beispiel steht

$$\{X \in A'\} \tag{111}$$

(wobei $A' \in \mathcal{A}'$) für

$$\{\omega \in \Omega |\ X(\omega) \in A'\} = X^{-1}[A'], \tag{112}$$

und für reelle Zufallsvariablen X und Zahlen $a \in \mathbb{R}$ steht

$$\{X \leq a\} \tag{113}$$

für

$$\{\omega \in \Omega |\ X(\omega) \leq a\}. \tag{114}$$

Diese Notation wird ebenso für mehrere Zufallsvariablen verwendet. Sind zum Beispiel X und Y reelle Zufallsvariablen auf dem gleichen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, so steht $\{X < Y\}$ kurz für $\{\omega \in \Omega |\ X(\omega) < Y(\omega)\}$. Die abkürzende Notation wird auch für Wahrscheinlichkeiten verwendet:

$$P(\{X \text{ hat die Eigenschaft } \Phi\}) \tag{115}$$

oder noch kürzer

$$P[X \text{ hat die Eigenschaft } \Phi] \quad \text{oder auch} \quad P(X \text{ hat die Eigenschaft } \Phi) \tag{116}$$

steht kurz für

$$P(\{\omega \in \Omega |\ X(\omega) \text{ hat die Eigenschaft } \Phi\}) \tag{117}$$

Zum Beispiel steht $P[X < 2]$ für

$$P(\{\omega \in \Omega |\ X(\omega) < 2\}) = P(X^{-1}\big[]-\infty, 2[\big]) = \mathcal{L}_P(X)(]-\infty, 2[). \tag{118}$$

Zugegeben ist diese Notation zunächst sehr gewöhnungsbedürftig und ziemlich verschieden von der sonst in der Mathematik üblichen Schreibweise, doch andererseits ist sie sehr praktisch und intuitiv.

Beispiele:

1. Es sei

$$\Omega = \{0,1\}^n, \quad \mathcal{A} = \mathcal{P}(\Omega), \quad P = \frac{1}{2^n}\sum_{\omega\in\Omega}\delta_\omega = \text{ Gleichverteilung auf } \Omega \tag{119}$$

ein Modell für den n-fachen fairen Münzwurf, wobei die i-te Komponente ω_i im Ergebnis $\omega = (\omega_1, \ldots, \omega_n)$ das Ergebnis des i-ten Wurfs beschreiben soll. Die kanonische Projektion

$$X_i : \Omega \to \mathbb{R}, \quad X_i(\omega) = \omega_i, \quad (i \in [n]) \tag{120}$$

ist dann eine Zufallsvariable, die den i-ten Wurf modelliert. Es gilt für $a \in \{0,1\}$:

$$P[X_i = a] = P(\{(\omega_1, \ldots, \omega_n) \in \Omega |\ \omega_i = a\}) = \frac{2^{n-1}}{2^n} = \frac{1}{2}. \tag{121}$$

Daher besitzt X_i die Verteilung

$$\mathcal{L}_P(X_i) = \frac{1}{2}(\delta_0 + \delta_1). \tag{122}$$

Ebenso ist die Anzahl der Ergebnisse "1" in der Münzwurffolge,

$$S = \sum_{i=1}^n X_i, \qquad \text{also} \qquad S(\omega) = \sum_{i=1}^n \omega_i, \tag{123}$$

eine Zufallsvariable. Für sie gilt mit $k \in \{0, 1 \ldots, n\}$:

$$P[S = k] = P[\{\omega \in \Omega |\ S(\omega) = k\}] = \frac{1}{2^n}\binom{n}{k}, \tag{124}$$

denn es gibt genau $\binom{n}{k}$ Elemente von $\{0,1\}^n$ mit genau k Einträgen 1. Also besitzt die Zufallsvariable S die Verteilung

$$\mathcal{L}_P(S) = \sum_{k=0}^n P[S = k]\delta_k = \frac{1}{2^n}\sum_{k=0}^n \binom{n}{k}\delta_k. \tag{125}$$

2. Ist allgemeiner $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ eine Zufallsvariable[9] mit nur endlich vielen Werten $x_1, \ldots, x_n \in \Omega'$ (paarweise verschieden), so besitzt X die Verteilung[10]

$$\mathcal{L}_P(X) = \sum_{i=1}^n P[X = x_i]\delta_{x_i} = \sum_{i=1}^n P(X^{-1}[\{x_i\}])\delta_{x_i}. \tag{126}$$

[9]Diese Schreibweise soll bedeuten, dass $X : \Omega \to \Omega'$ eine Zufallsvariable vom Ereignisraum $(\Omega, \mathcal{A})$ in den Ereignisraum $(\Omega', \mathcal{A}')$ ist, also eine $\mathcal{A}$=$\mathcal{A}'$-messbare Abbildung.

[10]Erinnerung: $\delta_{x_i} : \mathcal{A}' \to \{0,1\}$ bezeichnet das Diracmaß in x_i.

3. Ist P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ und bezeichnen $X_1, \ldots, X_n : \mathbb{R}^n \to \mathbb{R}$, die kanonischen Projektionen $X_i(\omega_1, \ldots, \omega_n) = \omega_i$, so sind alle X_i stetig, also Zufallsvariablen. Die Verteilung $\mathcal{L}_P(X_i) = X_i[P]$ von X_i unter P ist ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$; es wird die i-te *Randverteilung* von P genannt.
   Ist zum Beispiel P die uniforme Verteilung auf einem Rechteck

$$]a, b] \times]c, d] \subseteq \mathbb{R}^2, \tag{127}$$

   so sind die uniformen Verteilungen auf $]a, b]$ bzw. auf $]c, d]$ die beiden Randverteilungen von P.

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

Satz 2.26 (Charakterisierung von Verteilungsfunktionen) *Es sei $F : \mathbb{R} \to [0, 1]$ eine Abbildung. Dann sind äquivalent:*

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

$$\lim_{x \to +\infty} F(x) = 1 \qquad \textit{und} \qquad \lim_{x \to -\infty} F(x) = 0. \tag{128}$$

Beweis:
"1.⇒ 2." haben wir bereits früher in Lemma 2.9 gezeigt.
Zu "2.⇒ 1.": Wir definieren eine "Quasi-Inverse" G von F so:[11]

$$G :]0, 1[\to \mathbb{R}, \qquad G(q) = \sup\{s \in \mathbb{R} |\ F(s) \leq q\}. \tag{129}$$

Die Funktion G nimmt in der Tat nur Werte in $\mathbb{R}$ an:

- Die Aussage $G(q) > -\infty$ für $q \in]0, 1[$ folgt aus $\{s \in \mathbb{R} |\ F(s) \leq q\} \neq 0$, denn $F(s) \overset{s \to -\infty}{\longrightarrow} 0$.
- Ebenso gilt $G(q) < +\infty$ für $q \in]0, 1[$, denn für alle genügend großen $s \in \mathbb{R}$ gilt $F(s) > q$ wegen $F(s) \overset{s \to +\infty}{\longrightarrow} 1$.

Nun sei P die Gleichverteilung auf $(]0, 1[, \mathcal{B}(]0, 1[))$. Die Funktion G ist $\mathcal{B}(]0, 1[)$–$\mathcal{B}(\mathbb{R})$–messbar, da monoton steigend. Also ist das Bildmaß

$$\mu := G[P] = \mathcal{L}_P(G) \tag{130}$$

definiert.
Wir zeigen jetzt, dass das Wahrscheinlichkeitsmaß μ die Verteilungsfunktion F besitzt. Hierzu seien $s \in \mathbb{R}$ und $a \in]0, 1[$ gegeben. Wir zeigen:

[11] Anschaulich gesprochen kann man sich den Graphen von G so vorstellen: Man entfernt horizontale Stücke im Graphen von F bis auf ihren rechten Endpunkt, und fügt vertikale Stücke bei allen Sprungstellen im Graphen von F ein. Anschließend vertauscht man Abszisse und Ordinate. Horizontale Stücke im Graphen von F entsprechen also Sprungstellen im Graphen von G, und Sprungstellen im Graphen von F entsprechen horizontalen Stücken im Graphen von G.

1. Falls $q < F(s)$, so gilt für alle $t \in \mathbb{R}$ die Implikation $F(t) \leq q \;\Rightarrow\; t \leq s$.
2. Falls $q > F(s)$, so gilt *nicht* für alle $t \in \mathbb{R}$ die Implikation $F(t) \leq q \;\Rightarrow\; t \leq s$.

- *Beweis zu 1.:* Gegeben $q < F(s)$ und $t \in \mathbb{R}$ mit $F(t) \leq q$ folgt $F(t) \leq q < F(s)$, also $t \leq s$ wegen der Monotonie von F.
- *Beweis zu 2.:* Es sei $q > F(s)$ gegeben. Weil F rechtsseitig stetig in s ist, gibt es ein $t > s$ mit $q \geq F(t)$. Das bedeutet

$$\exists t \in \mathbb{R} : \; (F(t) \leq q \;\text{ und }\; t > s), \tag{131}$$

also

$$\text{nicht } \forall t \in \mathbb{R} : \; (F(t) \leq q \;\Rightarrow\; t \leq s). \tag{132}$$

Nun ist die Aussage $\forall t \in \mathbb{R} : \; (F(t) \leq q \;\Rightarrow\; t \leq s)$ gleichwertig mit

$$\sup\{t \in \mathbb{R} |\; F(t) \leq q\} \leq s, \tag{133}$$

also mit $G(q) \leq s$. Damit haben wir gezeigt:

1'. $q < F(s) \;\Rightarrow\; G(q) \leq s$,

2'. $q > F(s) \;\Rightarrow\; G(q) > s$.

Es folgt:

$$]0, F(s)[\subseteq \underbrace{\{q \in]0,1[|\; G(q) \leq s\}}_{\text{kurz: } \{G \leq s\}} \subseteq]0, F(s)] \tag{134}$$

und daher

$$F(s) = P(]0, F(s)[) \leq P[G \leq s] \leq P(]0, F(s) \cap]0,1[) = F(s), \tag{135}$$

also die Behauptung:

$$\mu(]-\infty, s]) = P[G \leq s] = F(s). \tag{136}$$

□

Der letzte Satz liefert uns auch ein praktisches Verfahren zur Simulation von Zufallszahlen mit vorgegebener Verteilungsfunktion F, wenn auf dem Einheitsintervall $]0,1[$ gleichverteilte Zufallszahlen gegeben sind.

Beispiel: Ist ω eine uniform auf $]0,1[$ verteilte Zufallszahl, so ist $-\log(1-\omega)$ eine exponentialverteilte Zufallszahl zum Parameter 1.
In der Tat: Die Verteilungsfunktion

$$F(t) = (1 - e^{-t}) 1_{[0,\infty[}(t), \qquad (t \in \mathbb{R}) \tag{137}$$

der Exponentialverteilung zum Parameter 1 besitzt die Quasiinverse

$$G(q) = -\log(1-q), \qquad (q \in]0,1[). \tag{138}$$

Natürlich ist auch $-\log\omega$ eine exponentialverteilte Zufallszahl, denn mit ω ist auch $1-\omega$ uniform auf $]0,1[$ verteilt.

Definition 2.27 (Quantil) *Es sei μ ein Wahrscheinlichkeitsmaßa uf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Verteilungsfunktion F. Für $q \in]0,1[$ heißt jedes $t \in \mathbb{R}$ mit $F(t) = q$ ein* q-Quantil *von F. Ist G die in Formel* (129) *definierte Quasiinverse zu F, so ist $G(q)$ ein q-Quantil zu F. Die Funktion G heißt daher manchmal auch "die"* Quantilsfunktion *zu F.*

Man beachte, dass Quantile nicht eindeutig bestimmt sind, falls der Graph von F horizontale Stücke besitzt, also falls F nicht injektiv ist. Unsere Variante G der Quantilsfunktion ist die rechtsstetige Version.

2.5 Berechnung der Dichte von Verteilungen

Wir besprechen nun zwei Fälle, in denen das Bildmaß unter einer Abbildung eine Dichte hat, wenn das Ausgangsmaß eine Dichte besitzt. Für die praktische Arbeit mit kontinuierlichen Modellen sind diese Fälle sehr wichtig! Die Beweise beruhen auf Sätzen der Maßtheorie.

Satz 2.28 (Dichten von Randverteilungen) *Es sei μ ein Maß über $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ mit der Dichte $f : \mathbb{R}^n \to [0,\infty]$, also*

$$\mu(A) = \int_A f \, d\lambda_n \qquad \text{für alle } A \in \mathcal{B}(\mathbb{R}^n). \tag{139}$$

Gegeben $m < n$, sei

$$\rho : \mathbb{R}^n \to \mathbb{R}^m, \qquad q(x_1, \ldots, x_n) = (x_1, \ldots, x_m) \tag{140}$$

die Projektion auf die ersten m Komponenten. Dann besizt das Bildmaß $\rho[\mu]$ eine Dichte

$$g : \mathbb{R}^m \to [0, +\infty], \qquad g(x) = \int_{\mathbb{R}^{n-m}} f(x,y)\, \lambda_{n-m}(dy). \tag{141}$$

Dieser Satz beruht auf dem Satz von Fubini für das Lebesguemaß für nichtnegative Funktionen aus der Maßtheorie:

Satz 2.29 (Satz von Fubini – Version für das Lebesguemaß und nichtnegative Funktionen) *Es sei $f : \mathbb{R}^n \to [0,\infty]$ messbar und $m < n$. Dann ist auch die Abbildung*

$$g : \mathbb{R}^m \to [0,\infty], \quad g(x) = \int_{\mathbb{R}^{n-m}} f(x,y)\, \lambda_{n-m}(dy) \tag{142}$$

messbar, und es gilt

$$\int_{\mathbb{R}^n} f \, d\lambda_n = \int_{\mathbb{R}^m} g \, d\lambda_m. \tag{143}$$

Merke:

Für messbare $f \geq 0$ gilt:

$$\begin{aligned}\int_{\mathbb{R}^n} f(z)\,\lambda_n(dz) &= \int_{\mathbb{R}^m}\int_{\mathbb{R}^{n-m}} f(x,y)\,\lambda_{n-m}(dy)\,\lambda_m(dx)\\ &= \int_{\mathbb{R}^{n-m}}\int_{\mathbb{R}^m} f(x,y)\,\lambda_m(dx)\,\lambda_{n-m}(dy) \qquad (144)\end{aligned}$$

Wir beweisen jetzt den Satz 2.28 über Dichten von Randverteilungen mit dem Satz von Fubini:
Es sei $A \in \mathcal{B}(\mathbb{R}^m)$. Dann ist

$$\rho^{-1}[A] = A \times \mathbb{R}^{n-m} \in \mathcal{B}(\mathbb{R}^n), \qquad (145)$$

und es gilt

$$\begin{aligned}\rho[\mu](A) &= \int_{\rho^{-1}[A]} f\,d\lambda_n = \int_{\mathbb{R}^n} 1_{A\times\mathbb{R}^{n-m}} f\,d\lambda_n\\ &= \int_{\mathbb{R}^m}\int_{\mathbb{R}^{n-m}} \underbrace{1_{A\times\mathbb{R}^{n-m}}(x,y)}_{=1_A(x)} f(x,y)\,\lambda_{n-m}(dy)\,\lambda_m(dx)\\ &= \int_{\mathbb{R}^m} 1_A(x)\int_{\mathbb{R}^{n-m}} f(x,y)\,\lambda_{n-m}(dy)\,\lambda_m(dx)\\ &= \int_A\int_{\mathbb{R}^{n-m}} f(x,y)\,\lambda_{n-m}(dy)\,\lambda_m(dx)\\ &= \int_A g\,\lambda_m(dx). \qquad (146)\end{aligned}$$

Das bedeutet: Das Bildmaß $\rho[\mu]$ besitzt die Dichte g.

□

Man beachte die folgende Analogie zum diskreten Fall: Sind Ω_1 und Ω_2 endliche Ergebnisräume,

$$\rho : \Omega_1 \times \Omega_2 \to \Omega_1, \quad \rho(x,y) = x \qquad (147)$$

die erste kanonische Projektion, und μ ein Maß auf dem kartesischen Produkt $(\Omega_1 \times \Omega_2, \mathcal{P}(\Omega_1 \times \Omega_2))$ mit der Zähldichte f, also

$$\mu(A) = \sum_{\omega\in A} f(\omega) \qquad \text{für alle } A \subseteq \Omega_1 \times \Omega_2, \qquad (148)$$

so besitzt das Bildmaß $\rho[\mu]$ die Zähldichte

$$g : \Omega_1 \to [0,\infty], \qquad g(x) = \sum_{y\in\Omega_2} f(x,y). \qquad (149)$$

In der Tat: Für alle $A \subseteq \Omega_1$ gilt:

$$\rho[\mu](A) = \mu(A \times \Omega_2) = \sum_{(x,y)\in A\times\Omega_2} f(x,y) = \sum_{x\in A}\sum_{y\in\Omega_2} f(x,y) = \sum_{x\in A} g(x). \tag{150}$$

Die Analogie wird noch deutlicher, wenn man die Summen als Integrale bezüglich der Zählmaße ν_1 auf Ω_1, ν_2 auf Ω_2 bzw. ν auf $\Omega_1 \times \Omega_2$ auffaßt.

Sowohl der diskrete als auch der kontinuierliche Fall sind Spezialfälle des allgemeinen Satzes von Fubini, den wir später besprechen.

Beispiele

1. Es seien P die Gleichverteilung auf der Einheitskreisscheibe

$$B = \{z \in \mathbb{R}^2 |\ \|z\|_2 < 1\} \tag{151}$$

und $X : \mathbb{R}^2 \to \mathbb{R}$ die Projektion auf die erste Koordinate. Dann besitzt die Verteilung $\mathcal{L}_P(X)$ die Dichte

$$g : \mathbb{R} \to [0,\infty], \quad g(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{für } |x| < 1, \\ 0 & \text{für } |x| \geq 1. \end{cases} \tag{152}$$

Beweis: Die Gleichverteilung P besitzt die Dichte

$$f : \mathbb{R}^2 \to [0,\infty],$$
$$f(x,y) = \frac{1}{\pi} 1_B(x,y) = \begin{cases} \frac{1}{\pi} 1_{]-\sqrt{1-x^2},\sqrt{1-x^2}[}(y) & \text{für } |x| < 1, \\ 0 & \text{für } |x| \geq 1. \end{cases} \tag{153}$$

Man beachte dabei, dass die Kreisscheibe B den Flächeninhalt $\lambda_2(B) = \pi$ besitzt. Es folgt: Die Verteilung $\mathcal{L}_P(X)$ hat die Dichte

$$g(x) = \int_{\mathbb{R}} \frac{1}{\pi} 1_B(x,y)\, dy = \begin{cases} \frac{1}{\pi}\int_{\mathbb{R}} 1_{]-\sqrt{1-x^2},\sqrt{1-x^2}[}(y)\, dy = \frac{2}{\pi}\sqrt{1-x^2} & \text{für } |x| < 1, \\ 0 & \text{für } |x| \geq 1. \end{cases} \tag{154}$$

2. Es sei P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ mit einer Dichte der Gestalt

$$f : \mathbb{R}^2 \to [0,\infty], \qquad f(x,y) = g(x) \cdot h(y) \tag{155}$$

mit zwei Wahrscheinlichkeitsdichten $g, h : \mathbb{R} \to [0,\infty]$. Dann besitzen die beiden Randverteilungen von P die Dichten g bzw. h.

Beweis: Wir zeigen das hier nur für die erste Randverteilung, für die zweite Randverteilung folgt es analog. Die erste Randverteilung besitzt die Dichte

$$\mathbb{R} \ni x \mapsto \int_{\mathbb{R}} f(x,y)\, dy = \int_{\mathbb{R}} g(x) \cdot h(y)\, dy = g(x) \int_{\mathbb{R}} h(y)\, dy = g(x), \tag{156}$$

denn es gilt $\int_{\mathbb{R}} h(y)\, dy = 1$, weil h eine Wahrscheinlichkeitsdichte ist.

□

Satz 2.30 (Bildmaße unter Diffeomorphismen) *Es seien $U, V \subseteq \mathbb{R}^n$ offen und $f : U \to V$ ein C^1-Diffeomorphismus, also stetig differenzierbar und bijektiv mit stetig differenzierbarer Inverser $f^{-1} : V \to U$. Dann gilt für alle messbaren $g : V \to [0, \infty]$ die folgende* Transformationsformel: [12]

$$\int_V g(y)\, \lambda_n(dy) = \int_U g(f(x)) |\det Df(x)|\, \lambda_n(dx) \quad (157)$$

Hierbei bezeichnet $\det Df$ die Jacobideterminante von f, also die Determinante der Jacobimatrix Df von f.

Dieser Satz ist ein zentrales Resultat der Integralrechnung mehrerer Variablen.
Für unsere Zwecke impliziert das:

Satz 2.31 (Transformationssatz für Dichten) *Mit den Bezeichnungen von oben sei P ein (Wahrscheinlichkeits-)Maß auf $(V, \mathcal{B}(V))$ mit der Dichte g bezüglich des Lebesguemaßes λ_n auf $\mathcal{B}(V)$. Dann besitzt die Verteilung $\mathcal{L}_P(f^{-1})$ der Zufallsvariablen $f^{-1} : V \to U$ die Dichte*

$$(g \circ f) \cdot |\det Df| \quad (158)$$

bezüglich des Lebesguemaßes λ_n auf $\mathcal{B}(U)$.

In der Tat: Für alle $A \in \mathcal{B}(U)$ erhalten wir mit der Transformationsformel:

$$\begin{aligned} \mathcal{L}_P(f^{-1})(A) = P(f[A]) &= \int_V 1_{f[A]}(y) g(y)\, \lambda_n(dy) \\ &= \int_U \underbrace{1_{f[A]}(f(x))}_{=1_A(x)} g(f(x)) |\det Df(x)|\, \lambda_n(dx) \\ &= \int_A g(f(x)) |\det Df(x)|\, \lambda_n(dx). \end{aligned} \quad (159)$$

□

[12] Ein häufiger Fehler ist es, die Jacobideterminante auf die falsche Seite der Transformationsformel zu schreiben, also die Jacobideterminante mit ihrem Kehrwert zu verwechseln. Vielleicht hilft Ihnen die folgende nicht ganz präzise, aber intuitive mnemotechnische Notation, diesen Fehler zu vermeiden:
Mit $y = f(x)$ merke man sich:

$$g(y)\, dy = g(f(x)) \left| \frac{dy}{dx} \right| dx.$$

Hierbei stehen dy bzw. dx symbolisch für das Lebesguemaß auf V bzw. auf U, und $\frac{dy}{dx}$ symbolisch für die Jacobideterminante.

Beispiel: Simulation standardnormalverteilter Zufallszahlen: Die zweidimensionale Standardnormalverteilung ist das Wahrscheinlichkeitsmaß P auf $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ mit der Dichte

$$f(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} = \varphi(x)\varphi(y), \qquad (x,y) \in \mathbb{R}^2, \tag{160}$$

wobei

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \qquad x \in \mathbb{R}, \tag{161}$$

die Dichte der Standardnormalverteilung bezeichnet. Insbesondere sind beide Randverteilungen von P Standardnormalverteilungen. Offensichtlich ist f und damit auch P invariant unter Drehungen um $(0,0)$, weil $f(x,y)$ nur vom Radiusquadrat $r^2 = x^2+y^2$ abhängt. Diese Drehinvarianz liegt letztlich dem folgenden Simulationsverfahren zu Grunde:

Es sei $Z = (U,V)$ ein Vektor zweier Zufallszahlen, gleichverteilt auf $]0,1[^2$. Wir bilden:

$$\phi := 2\pi V, \tag{162}$$

$$R := \sqrt{-2\log U}, \tag{163}$$

und rechnen von Polarkoordinaten in kartesische Koordinaten um:

$$X := R\cos\phi, \tag{164}$$

$$Y := R\sin\phi. \tag{165}$$

Dann ist der Vektor (X,Y) zweidimensional standardnormalverteilt, insbesondere sind X und Y beide standardnormalverteilt. Man beachte auch, dass

$$\frac{1}{2}R^2 = -\log U \tag{166}$$

exponentialverteilt mit dem Parameter 1 ist.

Begründung des Verfahrens: Die Abbildung

$$g :]0,1[^2 \to \mathbb{R}^2 \setminus ([0,\infty[\times\{0\}), \tag{167}$$

$$g(u,v) = (\sqrt{-2\log u}\cos(2\pi v), \sqrt{-2\log u}\sin(2\pi v)) = (x,y) \tag{168}$$

ist ein Diffeomorphismus mit der Jacobimatrix

$$Dg(u,v) = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} -\frac{2}{u}\cdot\frac{1}{2\sqrt{-2\log u}}\cos(2\pi v) & -2\pi\sqrt{-2\log u}\sin(2\pi v) \\ -\frac{2}{u}\cdot\frac{1}{2\sqrt{-2\log u}}\sin(2\pi v) & 2\pi\sqrt{-2\log u}\cos(2\pi v) \end{pmatrix} \tag{169}$$

und der Jacobi-Determinante

$$\det Dg(u,v) = -\frac{2\pi}{u}[\cos^2(2\pi v) + \sin^2(2\pi v)] = -\frac{2\pi}{u}. \tag{170}$$

Für die Umkehrabbildung

$$g^{-1} : \mathbb{R}^2 \setminus ([0,\infty[\times\{0\}) \to]0,1[^2, \quad (x,y) \mapsto (u,v) \tag{171}$$

gilt also:

$$\det D(g^{-1})(x,y) = [\det Dg(u,v)]^{-1} = -\frac{u}{2\pi} = -\frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}, \tag{172}$$

wobei wir $-\frac{1}{2}(x^2+y^2) = \log u$ verwendet haben. Nun besitzt die Gleichverteilung auf $]0,1[^2$ die Dichte 1 auf $]0,1[^2$. Nach der Transformationsformel folgt: Die Verteilung $\mathcal{L}_{\text{unif}]0,1[^2}(g)$ von g besitzt die Dichte

$$f(x,y) = 1 \cdot |\det D(g^{-1})(x,y)| = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \tag{173}$$

für $(x,y) \in \mathbb{R}^2 \setminus ([0,\infty[\times\{0\})$. Dies ist jedoch eine Dichte der zweidimensionalen Standardnormalverteilung, da die positive x-Achse $[0,\infty[\times\{0\}$ eine λ_2-Nullmenge ist. [13]

□

Beispiel: Gebrochen-rationale Transformation exponentialverteilter Zufallsvariablen: Es sei P das Wahrscheinlichkeitsmaß auf $(]0,\infty[^2, \mathcal{B}(]0,\infty[^2) = (\Omega, \mathcal{A})$ mit der Dichte

$$g(x,y) = e^{-x}e^{-y}, \qquad (x,y > 0). \tag{175}$$

Insbesondere sind beide Randverteilungen von P Exponentialverteilungen. Wir betrachten die Transformation

$$\begin{aligned} h :&]0,\infty[^2 \to]0,\infty[\times]0,1[, \\ &h(x,y) = \left(x+y, \frac{y}{x+y}\right). \end{aligned} \tag{176}$$

Sie ist ein C^1-Diffeomorphismus mit der Umkehrabbildung

$$\begin{aligned} f :&]0,\infty[\times]0,1[\to]0,\infty[^2, \\ &f(s,t) = (s - st, st). \end{aligned} \tag{177}$$

[13] Das Argument des Beweises kann man auch verwenden, um das Gaußsche Integral

$$\int_{\mathbb{R}} e^{-\frac{1}{2}x^2}\, dx = \sqrt{2\pi} \tag{174}$$

zu berechnen: Weil die Dichte der Gleichverteilung $]0,1[^2$ eine Wahrscheinlichkeitsdichte ist, bildet auch die mit g transformierte Dichte, also die Dichte $\varphi(x)\varphi(y) = \exp(-(x^2+y^2)/2)/(2\pi)$ der zweidimensionalen Standardnormalverteilung, eine Wahrscheinlichkeitsdichte. Damit wird auch die Dichte $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ der (eindimensionalen) Standardnormalverteilung eine Wahrscheinlichkeitsdichte.

Diese Umkehrabbildung besitzt die Jacobimatrix

$$Df(s,t) = \begin{pmatrix} \frac{\partial}{\partial s}(s-st) & \frac{\partial}{\partial t}(s-st) \\ \frac{\partial}{\partial s}(st) & \frac{\partial}{\partial t}(st) \end{pmatrix} = \begin{pmatrix} 1-t & -s \\ t & s \end{pmatrix} \tag{178}$$

mit der Jacobideterminante

$$\det Df(s,t) = \begin{vmatrix} 1-t & -s \\ t & s \end{vmatrix} = s. \tag{179}$$

Es folgt: Die Verteilung $\mathcal{L}_P(h)$ besitzt die Dichte

$$]0,\infty[\times]0,1[\ni (s,t) \mapsto g(f(s,t)) \cdot |\det Df(s,t)| = e^{-(s-st)}e^{-st} \cdot s = de^{-s} \tag{180}$$

bezüglich des Lebesguemaßes λ_2 auf der Borelschen σ-Algebra $\mathcal{B}(]0,\infty[\times]0,1[)$.

Wir können das auch so formulieren: Bezeichnen

$$X, Y :]0,\infty[^2 \to \mathbb{R} \tag{181}$$

die Projektionen auf die erste bzw. die zweite Koordinate, so besitzt die Zufallsvariable

$$\left(X+Y, \frac{Y}{X+Y}\right) : (\Omega, \mathcal{A}) \to (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)) \tag{182}$$

(genauer gesagt ihre Verteilung) die Dichte

$$\mathbb{R}^2 \ni (s,t) \mapsto 1_{]0,\infty[}(s) s e^{-s} \cdot 1_{]0,1[}(t). \tag{183}$$

Insbesondere ist $X+Y$ Gamma(1,2)-verteilt, und $Y/(X+Y)$ ist uniform auf $]0,1[$ verteilt.

2.6 Die von Zufallsvariablen erzeugte σ-Algebra

Satz/Definition 2.32 (erzeugte σ-Algebra) *Es sei Ω ein Ergebnisraum, $(\Omega', \mathcal{A}')$ ein Ereignisraum und $X : \Omega \to \Omega'$ eine Abbildung. Dann ist*

$$\sigma(X) := \{X^{-1}[A'] :\ A' \in \mathcal{A}'\} = \{\{X \in A'\} :\ A' \in \mathcal{A}'\} \tag{184}$$

eine σ-Algebra.[14] *Sie heißt die von X* erzeugte σ-Algebra.
Ist allgemeiner $(\Omega_i, \mathcal{A}_i)_{i\in I}$ eine Familie von Ereignisräumen und $(X_i : \Omega \to \Omega_i)_{i\in I}$ eine Familie von Abbildungen, so heißt

$$\sigma(X_i :\ i \in I) := \sigma(X_i^{-1}[A_i] :\ i \in I,\ A_i \in \mathcal{A}_i) \tag{185}$$

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

[14] Natürlich hängt $\sigma(X)$ nicht nur von X, sondern auch von der σ-Algebra $\mathcal{A}'$ ab, doch das wird in der Notation üblicherweise unterdrückt.

Übung 2.33 *Beweisen Sie, dass $\sigma(X)$ in der Tat eine σ-Algebra ist.*

Bemerkungen:

1. $\sigma(X)$ ist die kleinste σ-Algebra $\mathcal{A}$ über Ω, bezüglich der
$$X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}') \tag{186}$$
messbar ist. Ebenso ist $\sigma(X_i : \ i \in I)$ die kleinste σ-Algebra $\mathcal{A}$ über Ω, bezüglich der alle Abbildungen
$$X_i : (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i), \qquad (i \in I), \tag{187}$$
messbar sind.

2. Die σ-Algebra $\sigma(X)$ bzw. $\sigma(X_i : \ i \in I)$ wird als die Menge aller beobachtbaren Ereignisse interpretiert, wenn man nicht das Ergebnis $\omega \in \Omega$ des Zufallsexperiments beobachtet, sondern nur den Wert $X(\omega)$ bzw. die Werte $X_i(\omega)$, $i \in I$.

3. Eine Abbildung $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ ist also genau dann eine Zufallsvariable, wenn $\sigma(X) \subseteq \mathcal{A}$ gilt.

Beispiele:

1. Sind $X_1, \ldots, X_n : \mathbb{R}^n \to \mathbb{R}$ die kanonischen Projektionen, so ist $\sigma(X_i : \ i \in [n]) = \mathcal{B}(\mathbb{R}^n)$ die Borelsche σ-Algebra, wenn $\mathbb{R}$ mit der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R})$ versehen wird.

2. Sind allgemeiner $(\Omega_1, \mathcal{A}_1), \ldots, (\Omega_n, \mathcal{A}_n)$ Ereignisräume, $\Omega = \Omega_1 \times \ldots \times \Omega_n$ ihr kartesisches Produkt, und $X_i : \Omega \to \Omega_i$, $(i \in [n])$, die kanonischen Projektionen, so heißt
$$\mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n := \sigma(X_i : \ i \in [n]) \tag{188}$$
die *Produkt-σ-Algebra* der $\mathcal{A}_1, \ldots, \mathcal{A}_n$. Sie wird von den "Quadern" $A_1 \times \ldots \times A_n$ mit $A_i \in \mathcal{A}_i$ für alle $i \in [n]$ erzeugt.

   Der Begriff läßt sich auch auf unendlich viele Faktoren erweitern: Ist $(\Omega_i, \mathcal{A}_i)_{i \in I}$ eine Familie von Ereignisräumen, $\Omega = \prod_{i \in I} \Omega_i$ ihr kartesisches Produkt, und $X_i : \Omega \to \Omega_i$, $(i \in I)$, die kanonischen Projektionen, so heißt
$$\bigotimes_{i \in I} \mathcal{A}_i := \sigma(X_i : \ i \in I) \tag{189}$$
die *Produkt-σ-Algebra* aller $\mathcal{A}_i$. Sie enthält i.a. *nicht* beliebige Quader $\prod_{i \in I} A_i$ mit $A_i \in \mathcal{A}_i$, $(i \in I)$, falls überabzählbar viele A_i von Ω verschieden sind. Die Produkt-σ-Algebra wird jedoch von den "Zylindermengen" $\prod_{i \in I} A_i$ mit $A_i \in \mathcal{A}_i$, aber $A_i \neq \Omega$ für höchstens endlich viele $i \in I$ erzeugt.

3. Ist $X : \mathbb{R}^2 \to \mathbb{R}$ die erste kanonische Projektion mit dem Zielraum $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, so ist $\sigma(X)$ die Menge aller "Streifen" $A \times \mathbb{R}$ mit $A \in \mathcal{B}(\mathbb{R})$.

2.7 Elementare bedingte Wahrscheinlichkeiten

Sehr häufig hat man Teilinformationen über den Ausgang eines Zufallsexperiments und möchte Wahrscheinlichkeitsvorhersagen über das zufällige Ergebnis, gegeben diese Teilinformation, machen. In dieser Vorlesung besprechen wir erst die einfache Situation, dass wir den Antwort "ja" auf eine ja-nein-Frage über das zufällige Ergebnis kennen. Das führt uns auf den Begriff der *bedingten Wahrscheinlichkeit* gegeben das Eintreten eines Ereignisses. In der Vorlesung Wahrscheinlichkeitstheorie wird dieses Konzept dann wesentlich abstrahiert und verallgemeinert zu bedingten Wahrscheinlichkeiten gegeben die Information, die in einer σ-Algebra codiert ist.

Definition 2.34 (bedingte Wahrscheinlichkeit) *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $B \in \mathcal{A}$ ein Ereignis mit positiver Wahrscheinlichkeit: $P(B) > 0$. Für jedes Ereignis $A \in \mathcal{A}$ heißt*

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

die bedingte Wahrscheinlichkeit *des Ereignisses A gegeben B.*

Übung 2.35 (Bedingtes Maß) *Zeigen Sie, dass die Abbildung*

$$P(\cdot|B) : \mathcal{A} \to [0,1], \qquad A \mapsto P(A|B) \tag{190}$$

ein Wahrscheinlichkeitsmaß über $(\Omega, \mathcal{A})$ ist. Es wird das bedingte Maß zu P gegeben B genannt.

Interpretation: Beobachtet man bei dem Zufallsexperiment nur die Teilinformation, dass das Ereignis B eingetreten ist, so interpretiert man die bedingte Wahrscheinlichkeit $P(A|B)$ als die Wahrscheinlichkeit für das Eintreten von A gegeben diese Teilinformation. Der Nenner $P(B)$ normiert die bedingte Wahrscheinlichkeit so, dass nicht nur das sichere Ereignis Ω, sondern auch die Bedingung B die bedingte Wahrscheinlichkeit $P(B|B) = 1$ erhält. Das Ereignis B kann man sich hier als einen neuen, verkleinerten Ergebnisraum vorstellen.
Beispiele:

1. **Ignorieren ungültiger Ergebnisse beim Würfeln:** Modellieren wir ein Spielwürfel-Experiment mit ungültigen Ergebnissen durch

   $$(\Omega = \{1,2,3,4,5,6, \text{ungültig}\}, \mathcal{A} = \mathcal{P}(\Omega), P) \tag{191}$$

   mit einem Wahrscheinlichkeitsmaß

   $$P = \frac{q}{6} \sum_{i=1}^{6} \delta_i + (1-q)\delta_{\text{ungültig}}, \tag{192}$$

wobei $q \in]0,1[$ die Wahrscheinlichkeit beschreibt, ein gültiges Ergebnis zu erhalten, so modelliert

$$P(\cdot|\{\text{ungültig}\}^c) = \frac{1}{6}\sum_{i=1}^{6}\delta_i \tag{193}$$

das Würfelexperiment, bei dem ungültige Ergebnisse ignoriert werden.

2. **Bedingen mit der Gleichverteilung:** Sind $A, B \in \mathcal{B}(\mathbb{R}^n)$ zwei Borelmengen mit $A \subseteq B$ mit positivem, aber nicht unendlichem Volumen,

$$0 < \lambda_n(A) \leq \lambda_n(B) < \infty, \tag{194}$$

und bezeichnet P die Gleichverteilung auf B über $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, so ist das bedingte Maß $P(\cdot|A)$ die Gleichverteilung auf A.

3. **Simulation von Zufallszahlen mit vorgegebener Dichte:** Man kann die Beobachtung von eben so praktisch zur Simulation von Zufallsvariablen mit gegebener Dichte verwenden:

   Es sei P die Gleichverteilung auf dem Quadrat $(]0,1[^2, \mathcal{B}(]0,1[^2))$ und $f :]0,1[\to \mathbb{R}$ eine *beschränkte* Wahrscheinlichkeitsdichte (bezüglich des Lebesguemaßes auf $\mathcal{B}(]0,1[)$), sagen wir

$$f \leq c \in]0,\infty[. \tag{195}$$

   Wir reskalieren f so, dass es nur mehr Werte in $[0,1]$ annimmt:

$$g := \frac{f}{c}. \tag{196}$$

   Betrachten wir nun die Fläche unterhalb des Graphen von g,

$$B := \{(x,y) \in]0,1[^2 :\ y < g(x)\} \tag{197}$$

   so ist $P(\cdot|B)$ die Gleichverteilung auf B. Bezeichnet

$$X :]0,1[^2 \to]0,1[, \quad X(x,y) = x, \tag{198}$$

   die erste kanonische Projektion, so hat die Zufallsvariable X bezüglich des bedingten Maßes $P(\cdot|B)$ die Dichte f.
   Beweis: Für alle Ereignisse $A \in \mathcal{B}(]0,1[)$ gilt:

$$\mathcal{L}_{P(\cdot|B)}(X)(A) = P(X \in A|\ B) = P(\underbrace{X^{-1}[A]}_{=A\times]0,1[}|\ B) \tag{199}$$

$$= \frac{P((A\times]0,1[) \cap B)}{P(B)} = \frac{\int_A \int_{]0,g(x)[} 1\,dy\,dx}{\int_{]0,1[}\int_{]0,g(x)[} 1\,dy\,dx} \qquad \text{(mit Fubini)} \tag{200}$$

$$= \frac{\int_A g(x)\,dx}{\int_{]0,1[} g(x)\,dx} \tag{201}$$

$$= \frac{\frac{1}{c}\int_A f(x)\,dx}{\frac{1}{c}\int_{]0,1[} f(x)\,dx} = \int_A f(x)\,dx \tag{202}$$

wegen $\int_{]0,1[} f(x)\,dx = 1$.

□

Praktisch wendet man das so an:

Verfahren 2.36 (Simulation von Zufallszahlen mit vorgegebener beschränkter Dichte auf dem Einheitsintervall)

1. *Man wählt eine zufälligen Punkt $\omega = (x, y) \in]0, 1[^2$ gleichverteilt.*
2. *Ist $y < g(x)$, so gibt man das Ergebnis x aus.*
3. *Ist $y \geq g(x)$, so verwirft man ω und startet unabhängig neu bei Schritt 1.*

Den letzten Schritt 3. in diesem Verfahren, den "unabhängigen Neustart", haben wir noch nicht mathematisch modelliert. Wir holen das später nach.

Bedingte Wahrscheinlichkeiten erlauben eine stochastische Version von "Fallunterscheidungen": Die Fälle werden durch eine endliche oder abzählbar unendliche Partition von Ω in Ereignisse $A_1, \ldots, A_n$ oder $(A_k)_{k\in\mathbb{N}}$ modelliert:

Satz 2.37 (totale Wahrscheinlichkeit) *Es sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Ist $(A_k)_{k\in I}$ eine endliche oder abzählbar unendliche Partition von Ω in Ereignisse $A_k \in \mathcal{A}$ mit positiver Wahrscheinlichkeit $P(A_k) > 0$ für alle $k \in I$, so gilt für alle Ereignisse $B \in \mathcal{A}$:*

Formel von der totalen Wahrscheinlichkeit:

$$P(B) = \sum_{k\in I} P(B|A_k)P(A_k)$$

Beweis: Weil die Mengen $B \cap A_k$, $(k \in I)$, paarweise disjunkt mit der Vereinigung

$$B = \bigcup_{k\in I} (B \cap A_k) \tag{203}$$

sind, folgt:

$$P(B) = \sum_{k\in I} P(B \cap A_k) = \sum_{k\in I} P(B|A_k)P(A_k). \tag{204}$$

□

Beispiel: Ein Spielwürfel wird geworfen, und dann ein zweiter Spielwürfel so oft, wie der erste Augen zeigt. Man berechne die Wahrscheinlichkeit, dass der zweite Spielwürfel nie

eine “6” zeigt.
Wir modellieren das so: Wir nehmen das Modell

$$\Omega = \bigcup_{k=1}^{6} A_k \qquad \text{mit} \qquad A_k = \{k\} \times [6]^k = \text{“erster Würfel zeigt Augenzahl } k\text{”}, \tag{205}$$

$$\mathcal{A} = \mathcal{P}(\Omega) \tag{206}$$

wobei das Ergebnis $\omega = (k, (\omega_i)_{i\in[k]}) \in \Omega$ bedeuten soll, dass der erste Würfel die Augenzahl k zeigte, und der i-te Wurf des zweiten Würfels die Augenzahl ω_i, $(i \in [k])$. Wir nehmen das Wahrscheinlichkeitsmaß P, das durch die folgenden Annahmen charakterisiert wird:

$$P(A_k) = \frac{1}{6}, \tag{207}$$

$$P(\{(k, (\omega_i)_{i\in[k]})\}|\ A_k) = \frac{1}{6^k} \text{ für } k \in [6], (k, (\omega_i)_{i\in[k]}) \in A_k. \tag{208}$$

Das bedeutet

$$P = \sum_{k=1}^{6} \sum_{\omega\in[6]^k} \frac{1}{6} \cdot \frac{1}{6^k} \delta_{(k,\omega)}. \tag{209}$$

Mit der Formel für die totale Wahrscheinlichkeit folgt:

$$P(\text{“2. Würfel zeigt nie eine 6”}) = \sum_{k=1}^{6} P(\text{“2. Würfel zeigt nie eine 6”}|A_k)P(A_k) \tag{210}$$

$$= \sum_{k=1}^{6} \frac{5^k}{6^k} \frac{1}{6} \tag{211}$$

$$= \frac{155155}{279936} = 0.554\ldots \tag{212}$$

Ausblick: Bedingte Wahrscheinlichkeiten gegeben eine σ-Algebra. Für eine Partition $(A_k)_{k\in I}$ des Ergebnisraums Ω wie im Satz von der totalen Wahrscheinlichkeit kann man die Gesamtheit aller bedingten Wahrscheinlichkeiten $P(B|A_k)$, $k \in I$, auch zu einer Zufallsvariablen

$$\sum_{k\in I} P(B|A_k) 1_{A_k} \tag{213}$$

zusammenfassen, die den Wert $P(B|A_k)$ auf dem Ereignis A_K annimmt. Man kann diese Zufallsvariable als “Prognose für das Eintreffen von B” auffassen, wenn nur die Information aus der Unter-σ-Algebra

$$\mathcal{F} = \sigma(A_k :\ k \in I) \tag{214}$$

von $\mathcal{A}$ zur Verfügung steht. Man schreibt auch

$$P(B|\mathcal{F}) := \sum_{k\in I} P(B|A_k)1_{A_k} \tag{215}$$

und nennt die Zufallsvariable $P(B|\mathcal{F})$ die bedingte Wahrscheinlichkeit des Ereignisses B gegeben die σ-Algebra $\mathcal{F}$.
Eine kleine Zusatzkomplikation entsteht, wenn man auch P-Nullmengen unter den A_k zuläßt, weil dann $P(B|A_k)$ wegen einer Division durch 0 undefiniert bleibt. In diesem Fall bleibt $P(B|\mathcal{F}) = \sum_{k\in I} P(B|A_k)1_{A_k}$ auf einer Nullmenge undefiniert, ist also nur mehr P-fast überall definiert.
In der Vorlesung Wahrscheinlichkeitstheorie wird der Begriff der bedingten Wahrscheinlichkeit gegeben $\mathcal{F}$ auf beliebige Unter-σ-Algebren $\mathcal{F} \subseteq \mathcal{A}$ verallgemeinert, auch auf solche, die nicht von einer Partition erzeugt werden.
Hier als Ausblick nur ein Spezialfall: Ist P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ mit Dichte f und bezeichnet $X : \mathbb{R}^2 \to \mathbb{R}$ die Projektion auf die x-Koordinate, so nennen wir für $B \in \mathcal{B}(|R^2)$ und $x \in R$ den Quotienten

$$P(B|\ X = x) = \frac{\int_{\mathbb{R}} 1_B(x,y)f(x,y)\,dy}{\int_{\mathbb{R}} f(x,y)\,dy}, \tag{216}$$

falls er wohldefiniert ist, also falls weder 0 noch ∞ im Nenner steht, eine *bedingte Wahrscheinlichkeit von B gegeben $X = x$.* Die Mehrdeutigkeit der Dichte f, die ja auf einer Nullmenge abgeändert werden kann, impliziert, dass auch die Funktion $P(B|\ X = \cdot)$ nicht eindeutig bestimmt ist, sondern auf Nullmengen abgeändert werden kann.
Setzt man noch die Zufallsvariable X in das Argument der Funktion $P(B|\ X = \cdot)$ ein, so erhält man eine $\sigma(X)$-messbare Funktion

$$P(B|X) = P(B|\sigma(X)) := P(B|\ X = \cdot) \circ X \tag{217}$$

Auch sie ist nur bis auf Abänderung auf P-Nullmengen in $\sigma(X)$ eindeutig bestimmt. Sie wird *bedingte Wahrscheinlichkeit von B gegeben die σ-Algebra $\sigma(X)$* genannt.

Der Satz von Bayes. Eng verwandt mit dem Satz von der totalen Wahrscheinlichkeit ist die folgende Formel von Bayes, die es erlaubt, die beiden Ereignisse im Argument einer bedingten Wahrscheinlichkeit zu vertauschen:

Satz 2.38 (Satz von Bayes) *Es seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $(A_k)_{k\in I}$ eine endliche oder abzählbar unendliche Partition von Ω in Ereignisse A_k positiver Wahrscheinlichkeit: $P(A_k) > 0$. Dann gilt für alle Ereignisse $B \in \mathcal{A}$ mit $P(B) > 0$ und alle $k \in I$:*

Formel von Bayes:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j\in I} P(B|A_j)P(A_j)}$$

Beweis: Wir rechnen mit dem Satz von der totalen Wahrscheinlichkeit:

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(B|A_k)P(A_k)}{\sum_{j\in I} P(B|A_j)P(A_j)} \tag{218}$$

□

Die Formel von Bayes dient in mehrstufigen Zufallsexperimenten zum "Rückschluß auf die Ursachen":

Beispiele:

1. $0,01\%$ der Bevölkerung leide an einer seltenen Krankheit X. Ein medizinischer Test zur Diagnose dieser Krankheit erkenne mit einer Wahrscheinlichkeit 99% die Krankheit X korrekt, wenn der Proband tatsächlich an X leidet, und erkenne mit der Wahrscheinlichkeit 98% das Nichtvorliegen der Krankheit X korrekt, wenn der Proband nicht an X leidet.
   "Rückschluss auf die Ursache:" Falls der Test das Vorliegen von X anzeigt, wie groß ist die Wahrscheinlichkeit, dass der Proband tatsächlich an X leidet?
   Wir modellieren das so: Es sollen bedeuten:

   - Ereignis K: "Proband leidet an X."
   - Ereignis T: "Test zeigt X an."

   Gegeben sind:

$$P(K) = 10^{-4}, \tag{219}$$
$$P(T|K) = 0,99, \tag{220}$$
$$P(T^c|K^c) = 0,98. \tag{221}$$

   Mit der Formel von Bayes folgt:

$$\begin{aligned} P(K|T) &= \frac{P(T|K)P(K)}{P(T|K)P(K) + P(T|K^c)P(K^c)} \\ &= \frac{P(T|K)P(K)}{P(T|K)P(K) + (1 - P(T^c|K^c))(1 - P(K))} \\ &= \frac{0,99 \cdot 10^{-4}}{0,99 \cdot 10^{-4} + (1 - 0,98) \cdot (1 - 10^{-4})} = 0,00492\ldots \end{aligned} \tag{222}$$

   Dieser geringe Wert von nur ungefähr $\frac{1}{2}\%$ für ein korrektes Testergebnis erscheint vielleicht zunächst kontraintuitiv, gegeben die hohe Qualität des Tests von 98% oder höher. Es liegt natürlich an der geringen "a priori" Wahrscheinlichkeit für den Probanden, die Krankheit X zu haben.

2. $n+1$ Urnen, beschriftet mit den Nummern 0 bis n, enthalten je n Kugeln, und zwar enthalte die Urne mit der Nummer k genau k rote und $n-k$ blaue Kugeln.

- **1. Stufe des Zufallsexperiments:** Man wählt zufällig eine Urne nach der Gleichverteilung.
- **2. Stufe des Zufallsexperiments:** Aus der in der 1. Stufe gewählten Urne entnimmt man zufällig l Kugeln mit Zurücklegen.

"Rückschluß auf die Ursache:" Bedingt darauf, dass r dieser l Kugeln rot sind, mit welcher Wahrscheinlichkeit stammen sie aus der Urne Nr. k

Wir verzichten auf eine vollständige Modellierung, sondern beschreiben nur die Angaben formal:

$$\text{Ereignis } A_k := \text{“Urne Nr. } k \text{ wird gewählt”}, \qquad (k = 0, \ldots, n) \tag{223}$$

$$\text{Ereignis } B := \text{“}r \text{ rote Kugeln werden gezogen”}. \tag{224}$$

Gegeben sind:

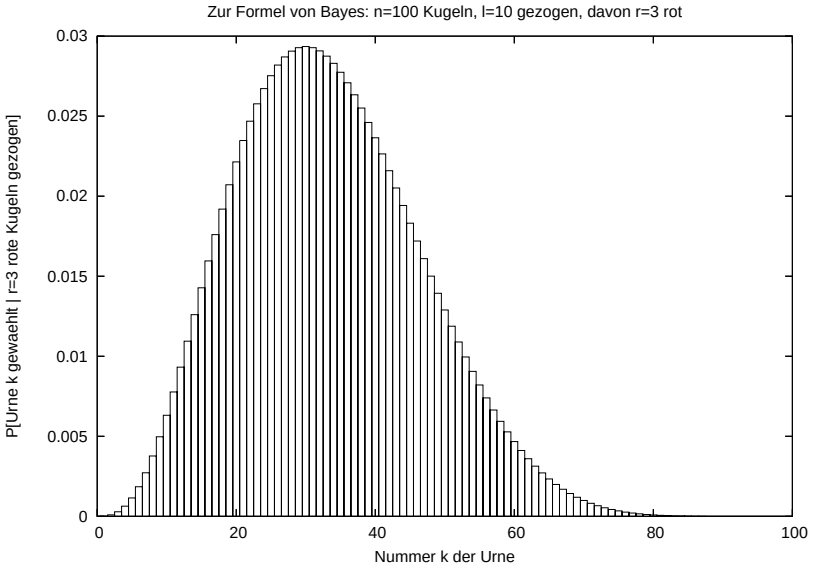
$$P(A_k) = \frac{1}{n+1}, \tag{225}$$

$$\begin{aligned} P(B|A_k) &= \frac{\text{Anzahl Möglichkeiten, } r \text{ Kugeln zu ziehen, gegeben } k}{\text{Anzahl Möglichkeiten, Kugeln zu ziehen}} \\ &= \frac{\binom{l}{r} k^r (n-k)^{l-r}}{n^l}. \end{aligned} \tag{226}$$

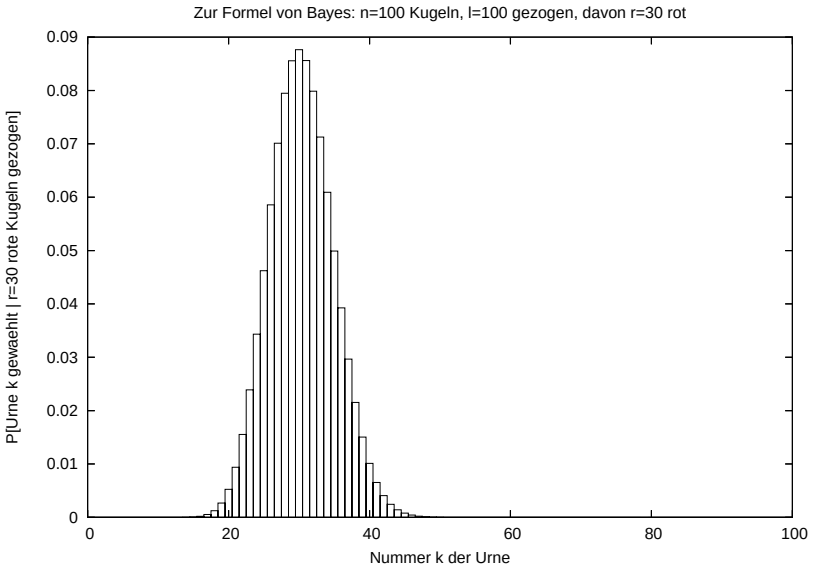
Es folgt nach der Formel von Bayes:

$$P(A_k|B) = \frac{\frac{\binom{l}{r} k^r (n-k)^{l-r}}{n^l} \cdot \frac{1}{n+1}}{\sum_{j=0}^{n} \frac{\binom{l}{r} j^r (n-j)^{l-r}}{n^l} \cdot \frac{1}{n+1}} = \frac{\binom{l}{r} k^r (n-k)^{l-r}}{\sum_{j=0}^{n} \binom{l}{r} j^r (n-j)^{l-r}}. \tag{227}$$

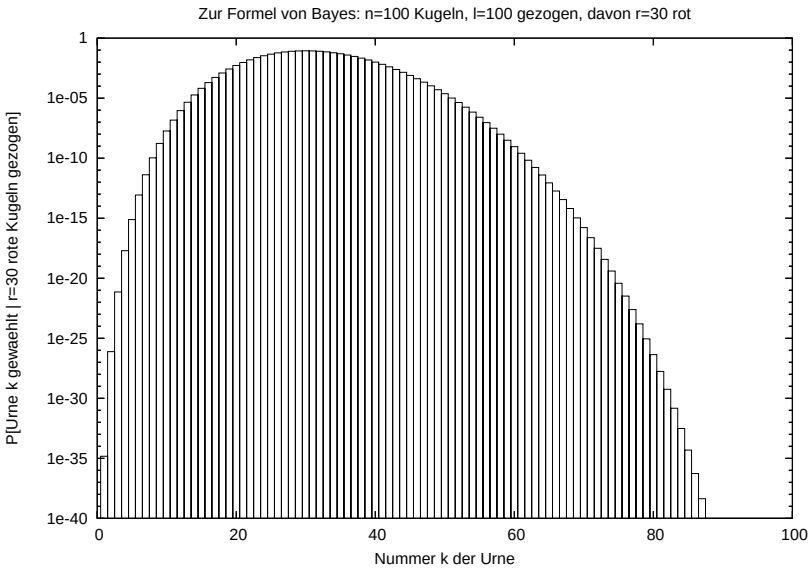
Hier ein numerisches Beispiel zu $n = 100$, $l = 10$, $r = 3$::



Man beachte das Maximum des Graphen bei $k = 30 = \frac{nr}{l}$ deutlich schärfer wird es ausgeprägt, wenn wir mehr Kugeln ziehen ($n = 100$, $l = 100$, $r = 30$):



Eine logarithmische Skalierung der Ordinate bei den gleichen Daten zeigt noch deutlicher, dass die bedingten Wahrscheinlichkeiten $P(B|A_k)$ um viele Größenordnungen abfallen, sobald man sich etwas vom Maximum entfernt:



Ausblick: Bayessche Statistik. Die Formel von Bayes ist das Fundament eines Zweigs der mathematischen Statistik, der *Bayesschen Statistik*. Statt das Zufallsexperiment (im Beispiel: Ziehen mit Zurücklegen von Kugeln unbekannter Farben) nur durch ein *einziges* Wahrscheinlichkeitsmodell zu beschreiben, betrachtet man hier viele Wahrscheinlichkeitsmodelle gleichzeitig (im Beispiel: alle $n+1$ Wahrscheinlichkeitsmodelle mit $k = 0, 1, \ldots n$ roten Kugeln in der Urne gleichzeitig). Nun versieht man die Klasse dieser Wahrscheinlichkeitsmodelle selbst mit einem Wahrscheinlichkeitsmaß, *a priori Verteilung*, englisch *prior* genannt, im Beispiel die Gleichverteilung auf $\{0, \ldots, n\}$. Man stellt sich dabei vor, "die Natur" habe in einer ersten Stufe zufällig eines der Wahrscheinlichkeitsmodelle ausgewählt. Das tatsächlich durchgeführte Zufallsexperiment stellt man sich dann als zweite Stufe vor: Mit dem in der ersten Stufe gewonnenen Wahrscheinlichkeitsmodell werden die beobachteten Daten zufällig gewonnen, im Beispiel beobachtete Zahl r gezogener roter Kugeln. Der Bayessche Statistiker schließt dann *bedingt auf die beobachteten Daten*, im Beispiel bedingt auf r, auf die Verteilung des Wahrscheinlichkeitsmodells aus der 1. Stufe zurück. Diese bedingte Verteilung wird *a posteriori Verteilung*, englisch *posterior*, genannt. Die Wahl der a priori Verteilung gibt dem Bayesschen Statistiker die Möglichkeit, Vorwissen in die statistische Modellierung einzubauen, gibt aber andererseits den Bayesschen statistischen Schlüssen eine gewisse Willkür.

2.8 Stochastische Unabhängigkeit

Es seien A und B zwei Ereignisse in einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Informal gesprochen nennen wir A und B unabhängig, wenn die Kenntnis des Eintretens von B die

Prognose für A nicht ändert. Im Fall $P(B) > 0$ können wir das mathematisch so fassen:

$$P(A|B) = P(A). \tag{228}$$

Schreiben wir das in der Form

$$P(A \cap B) = P(A)P(B), \tag{229}$$

so können wir auch den Fall $P(B) = 0$ zulassen. Das gibt Anlass zu der folgenden Definition:

Definition 2.39 (Stochastische Unabhängigkeit) *Zwei Ereignisse $A, B \in \mathcal{A}$ heißen stochastisch unabhängig bezüglich P, kurz: unabhängig, wenn gilt:* **sehr wichtig!**

$$P(A \cap B) = P(A)P(B). \tag{230}$$

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

$$P\left(\bigcap_{i\in E} A_i\right) = \prod_{i\in E} P(A_i). \tag{231}$$

Beispiele:

1. **Zweimaliger Wurf eines fairen Würfels**, modelliert durch

   $$\Omega = [6]^2, \quad \mathcal{A} = \mathcal{P}(\Omega), \quad P = \text{ Gleichverteilung auf } \Omega. \tag{232}$$

   Die Zufallsvariablen

   $$X, Y : \Omega \to [6] \tag{233}$$

   seien die Projektionen auf die erste bzw. zweite Komponente. Dann sind für alle $k, l \in [6]$ die Ereignisse $\{X = k\}$ und $\{Y = l\}$ unabhängig. In der Tat:

   $$P(X = k) = \frac{|\{k\} \times [6]|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}, \tag{234}$$

   $$P(Y = l) = \frac{|[6] \times \{l\}|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}, \tag{235}$$

   $$P(X = k,\ Y = l) = \frac{|\{(k, l)\}|}{|\Omega|} = \frac{1}{36} = P(X = k)P(Y = l). \tag{236}$$

2. **n-facher Wurf einer unfairen Münze:** Wir betrachten den Ergebnisraum $\Omega = \{0, 1\}^n$, $\mathcal{A} = \mathcal{P}(\Omega)$ mit der Interpretation "Kopf" für 1 und "Zahl" für 0. Gegeben einen Parameter $p \in [0, 1]$, wollen wir ein Wahrscheinlichkeitsmaß P auf $(\Omega, \mathcal{A})$ definieren, das einen n-fachen Münzwurf beschreibt, wenn "Kopf" bei einem Wurf

mit der Wahrscheinlichkeit p auftritt.
Wir definieren P durch seine Zähldichte $(p_\omega)_{\omega\in\Omega}$ so: Für $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ sei

$$p_\omega = \prod_{i=1}^{n} \underbrace{p^{\omega_i}(1-p)^{1-\omega_i}}_{=\begin{cases} p & \text{für } \omega_i = 1, \\ 1-p & \text{für } \omega_i = 0 \end{cases}} = p^{S(\omega)}(1-p)^{n-S(\omega)}, \tag{237}$$

wobei die Zufallsvariable

$$\begin{aligned} S &: \Omega \to \{0, \ldots, n\} \\ S(\omega) &= \sum_{i=1}^{n} \omega_i \end{aligned} \tag{238}$$

die Anzahl der Einsen bezeichnet. In der Tat ist

$$\sum_{\omega\in\Omega} p_\omega = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1, \tag{239}$$

so dass P ein Wahrscheinlichkeitsmaß ist. Nun seien

$$X_i : \Omega \to \{0, 1\}, \qquad X_i(\omega) = \omega_i, \qquad (i \in [n]), \tag{240}$$

die kanonischen Projektionen. Dann sind die Ereignisse

$$\{X_1 = 1\}, \{X_2 = 1\}, \ldots, \{X_n = 1\} \tag{241}$$

stochastisch unabhängig.
Begründung: Es sei eine Teilmenge $E \subseteq [n]$ gegeben. Dann gilt

$$\begin{aligned} P[\forall i \in E :\ X_i = 1] &= \sum_{\substack{\omega\in\Omega \\ \forall i\in E:\omega_i=1}} p_\omega = \sum_{\substack{\omega\in\Omega \\ \forall i\in E:\omega_i=1}} p^{|E|} \prod_{i\in[n]\setminus E} p^{\omega_i}(1-p)^{1-\omega_i} \\ &= p^{|E|} \sum_{k=0}^{n-|E|} \binom{n}{k} p^k (1-p)^{n-|E|-k} = p^{|E|}[p + (1-p)]^{n-|E|} = p^{|E|}. \end{aligned} \tag{242}$$

Als einen Spezialfall erhalten wir

$$P[X_i = 1] = p \text{ für alle } i \in [n] \tag{243}$$

und daher

$$P[\forall i \in E :\ X_i = 1] = p^{|E|} = \prod_{i\in E} P[X_i = 1]. \tag{244}$$

□

Definition 2.40 (Binomialverteilung) *Die Verteilung der Anzahl S der Einsen im Modell von eben heißt Binomialverteilung zu den Parametern $n \in \mathbb{N}$ und $p \in [0,1]$. Sie wird mit*

$$\text{binomial}(n,p) := \mathcal{L}_P(S) \tag{245}$$

bezeichnet.

Die Binomialverteilung beschreibt also die Verteilung der Anzahl des Ergebnisses "Kopf" bei n-facher unabhängiger Wiederholung eines Münzwurfexperiments, wenn für jeden der Würfe die Wahrscheinlichkeit, "Kopf" zu erhalten, den Wert p hat.
Es gilt:

Binomialverteilung:

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

wichtig!

Begründung: Für $k = 0, 1, \ldots, n$ gilt:

$$\begin{aligned} \text{binomial}(n,p)(\{k\}) &= P[S=k] = \sum_{\omega \in \{S=k\}} p^k (1-p)^{n-k} \\ &= |\{S=k\}| p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \tag{246}$$

□

Unabhängigkeit ist nicht das Gleiche wie paarweise Unabhängigkeit!

Achtung!

Beispiel hierzu: Es sei P die Gleichverteilung auf $(\Omega, \mathcal{A}) = (\{0,1\}^2, \mathcal{P}(\Omega))$. Weiter bezeichnen $X, Y : \Omega \to \{0,1\}$ die beiden kanonischen Projektionen, und

$$Z := (X+Y) \bmod 2 = 1_{\{X+Y \text{ ungerade }\}}. \tag{247}$$

Dann sind die Ereignisse $\{X=1\}$, $\{Y=1\}$ und $\{Z=1\}$ zwar paarweise unabhängig, aber dennoch nicht unabhängig, denn

$$\begin{aligned} P[X=1, Y=1] &= P(\{(1,1)\}) = \frac{1}{4} = P(\{(1,0),(1,1)\}) P(\{(0,1),(1,1)\}) \\ &= P[X=1] P[Y=1], \end{aligned} \tag{248}$$

$$\begin{aligned} P[X=1, Z=1] &= P(\{(1,0)\}) = \frac{1}{4} = P(\{(1,0),(1,1)\}) P(\{(0,1),(1,0)\}) \\ &= P[X=1] P[Z=1], \end{aligned} \tag{249}$$

$$\begin{aligned} P[Y=1, Z=1] &= P(\{(0,1)\}) = \frac{1}{4} = P(\{(0,1),(1,1)\}) P(\{(0,1),(1,0)\}) \\ &= P[Y=1] P[Z=1], \end{aligned} \tag{250}$$

aber

$$P[X=1,Y=1,Z=1]=P(\emptyset)=0\neq\frac{1}{8}=P[X=1]P[Y=1]P[Z=1]. \tag{251}$$

Wir führen nun eine praktische, sehr häufig verwendete Sprechweise ein:

Definition 2.41 (Unabhängigkeit von Zufallsvariablen) *Eine Familie* $(X_i:(\Omega,\mathcal{A})\to(\Omega_i,\mathcal{A}_i))_{i\in I}$ *von Zufallsvariablen auf einem Wahrscheinlichkeitsraum* $(\Omega,\mathcal{A},P)$ *heißt* unabhängig*, wenn für alle Familien* $(A_i\in\mathcal{A}_i)_{i\in I}$ *von Ereignissen in den Zielräumen gilt: Die Familie*

$$(\{X_i\in A_i\})_{i\in I}=(X_i^{-1}[A_i])_{i\in I} \tag{252}$$

ist unabhängig.
Eine Familie von $\cap$*-stabilen Mengensystemen* $(\mathcal{F}_i)_{i\in I}$ *mit* $\emptyset\neq\mathcal{F}_i\subseteq\mathcal{A}$ *heißt* unabhängig*, wenn alle Familien* $(A_i\in\mathcal{F}_i)_{i\in I}$ *unabhängig sind.*

Nach Definition gilt also für eine Familie $(X_i)_{i\in I}$ von Zufallsvariablen auf $(\Omega,\mathcal{A},P)$:

$$(X_i)_{i\in I}\text{ ist unabhängig}\quad\Leftrightarrow\quad(\sigma(X_i))_{i\in I}\text{ ist unabhängig.} \tag{253}$$

Beispiel: Im obigen Beispiel 2. (n-facher Münzwurf) sind die Zufallsvariablen $X_1,\ldots,X_n$, die die Ergebnisse der einzelnen Würfe beschreiben, unabhängig.

Wir zeigen jetzt einige Abschlusseigenschaften der Unabhängigkeit:

Lemma 2.42 (Dynkin-System aus unabhängigen Ereignissen) *Es sei* $(\Omega,\mathcal{A},P)$ *ein Wahrscheinlichkeitsraum und* $B\in\mathcal{A}$ *ein Ereignis. Dann gilt:*

1. *Die leere Menge* $\emptyset$ *ist unabhängig von* B*.*
2. *Ist ein Ereignis* $A\in\mathcal{A}$ *unabhängig von* B*, so ist auch* A^c *unabhängig von* B*.*
3. *Sind* $A_n\in\mathcal{A}$*,* $(n\in\mathbb{N})$*, paarweise disjunkte Ereignisse, die alle unabhängig von* B *sind, so ist auch deren Vereinigung* $\bigcup_{n\in\mathbb{N}}A_n$ *unabhängig von* B*.*

Anders gesagt: Das Mengensystem

$$\{A\in\mathcal{A}|\ A\text{ und }B\text{ sind unabhängig bzgl. }P\} \tag{254}$$

bildet ein Dynkin-System.

Beweis:

1. $P(\emptyset\cap B)=0=P(\emptyset)P(B)$.

2. Für unabhängige Ereignisse A, B gilt

$$\begin{aligned} P(A^c \cap B) &= P(B \setminus (A \cap B)) = P(B) - P(A \cap B) \qquad \text{(wegen } A \cap B \subseteq B) \\ &= P(B) - P(A)P(B) \qquad \text{(weil } A, B \text{ unabhängig)} \\ &= (1 - P(A))P(B) \\ &= P(A^c)P(B) \end{aligned} \tag{255}$$

3. Für die Vereinigung paarweise disjunkter, von B unabhängigen Ereignissen erhalten wir:

$$\begin{aligned} P\left(\left(\bigcup_{n\in\mathbb{N}} A_n\right) \cap B\right) &= P\left(\bigcup_{n\in\mathbb{N}} \underbrace{(A_n \cap B)}_{\text{paarweise disjunkt}}\right) \\ &= \sum_{n\in\mathbb{N}} P(A_n \cap B) \qquad \text{(wegen } \sigma\text{-Additivität)} \\ &= \sum_{n\in\mathbb{N}} P(A_n)P(B) \qquad \text{(weil } A_n, B \text{ unabhängig)} \\ &= P\left(\bigcup_{n\in\mathbb{N}} A_n\right) P(B) \qquad \text{(weil } A_n, (n \in \mathbb{N}), \text{ paarweise disjunkt)}, \end{aligned} \tag{256}$$

also die Behauptung.

□

Korollar 2.43 (Dynkin-System aus unabhängigen Ereignissen – Version für Mengensysteme) *Es seien ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und ein nichtleeres Ereignissystem $\emptyset \neq \mathcal{B} \subseteq \mathcal{A}$ gegeben. Dann ist*

$$\mathcal{D} = \{A \in \mathcal{A} |\ \forall B \in \mathcal{B} :\ A, B \textit{ sind unabhängig}\} \tag{257}$$

ein Dynkin-System.

Beweis: Das Mengensystem

$$\mathcal{D} = \bigcap_{B\in\mathcal{B}} \{A \in \mathcal{A} |\ A, B \text{ sind unabhängig}\} \tag{258}$$

ist ein Durchschnitt von Dynkin-Systemen, also auch selbst ein Dynkin-System.

□

Satz 2.44 (Hochheben der Unabhängigkeit) *Es seien ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und zwei $\cap$-stabile, nichtleere unabhängige Ereignissysteme $\mathcal{F}, \mathcal{G} \subseteq \mathcal{A}$ gegeben, d.h. es gelte*

$$\forall A \in \mathcal{F} \forall B \in \mathcal{G} :\ A, B \textit{ sind unabhängig.} \tag{259}$$

Dann sind auch $\sigma(\mathcal{F})$ und $\sigma(\mathcal{G})$ unabhängig.

Beweis: Das Mengensystem

$$\mathcal{D} := \{A \in \mathcal{A} |\ \forall B \in \mathcal{G} :\ A, B \text{ sind unabhängig}\} \tag{260}$$

ist ein Dynkin-System, das das $\cap$-stabile System $\mathcal{F}$ enthält: $\mathcal{F} \subseteq \mathcal{D}$. Aus dem Dynkin-Lemma folgt $\sigma(\mathcal{F}) \subseteq \mathcal{D}$, also sind $\sigma(\mathcal{F})$ und $\mathcal{G}$ voneinander unabhängig. Nun sei

$$\mathcal{D}' := \{B \in \mathcal{A} |\ \forall A \in \sigma(\mathcal{F}) :\ A, B \text{ sind unabhängig}\}. \tag{261}$$

Das Mengensystem $\mathcal{D}'$ ist ebenfalls ein Dynkin-System, und es gilt $\mathcal{G} \subseteq \mathcal{D}'$. Weil das Mengensystem $\mathcal{G}$ $\cap$-stabil ist, folgt aus dem Dynkin-Lemma: $\sigma(\mathcal{G}) \subseteq \mathcal{D}'$. Das bedeutet, dass $\sigma(F)$ und $\sigma(G)$ ebenfalls unabhängig voneinander sind.

□

Hier eine Verallgemeinerung davon:

Satz 2.45 (Vererbung der stochastischen Unabhängigkeit)

a) Es sei $(\mathcal{M}_i)_{i\in I}$ eine stochastisch unabhängige Familie von nichtleeren, durchschnittsstabilen Ereignissystemen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann ist auch $(\sigma(\mathcal{M}_i))_{i\in I}$ eine stochastisch unabhängige Familie von σ-Algebren.

b) Unter der Voraussetzung von Teil a) sei $(I_j)_{j\in J}$ eine Familie von paarweise disjunkten Teilmengen von I. Dann ist auch $(\sigma(\bigcup_{i\in I_j} \mathcal{M}_i))_{j\in J}$ eine stochastisch unabhängige Familie von σ-Algebren.

Beweis: Ohne Beschränkung der Allgemeinheit dürfen wir $\Omega \in \mathcal{M}_i$ für alle $i \in I$ voraussetzen, sonst nehmen wir Ω zu $\mathcal{M}_i$ hinzu. Man beachte, dass diese eventuelle Hinzunahme die Unabhängigkeit der $(\mathcal{M}_i)_{i\in I}$ nicht zerstört.
Teil a) des Satzes ist der Spezialfall $J = I$, $I_j = \{j\}$ des Teils b). Zur Übersichtlichkeit beweisen wir dennoch zuerst den Teil a).
Behauptung: Es seien $i_1, i_2, \ldots, i_m \in I$ paarweise verschiedene Indizes, $m \in \mathbb{N}$. Dann gilt für alle $n = 0, 1, \ldots, m$ und alle Familien von Ereignissen $(A_j)_{j=1,\ldots,m}$ mit $A_j \in \sigma(\mathcal{M}_{i_j})$ für $1 \leq j \leq n$ und $A_j \in \mathcal{M}_{i_j}$ für $n < j \leq m$ die folgende Aussage:

$$P\left(\bigcap_{i=1}^{m} A_i\right) = \prod_{i=1}^{m} P(A_i) \tag{262}$$

Wir zeigen dies durch vollständige Induktion über n.
Der *Induktionsanfang*, $n = 0$, ist klar nach der vorausgesetzten Unabhängigkeit der Familie $(\mathcal{M}_i)_{i\in I}$.
Induktionsvoraussetzung: Es sei die Behauptung schon für gegebenes $n \in \{0, 1, \ldots, m-1\}$ gezeigt.

Induktionsschluß: Gegeben sei eine Familie von Ereignissen $(A_j)_{\substack{j=1,\ldots,m \\ j\neq n+1}}$ mit $A_j \in \sigma(\mathcal{M}_{i_j})$ für $1 \leq j \leq n$ und $A_j \in \mathcal{M}_{i_j}$ für $n+1 < j \leq m$. Wir bilden das folgende Mengensystem:

$$\mathcal{D} := \left\{ A \in \mathcal{A} \;\middle|\; P\left(A \cap \bigcap_{\substack{i=1 \\ i\neq n+1}}^{m} A_i \right) = P(A) \prod_{\substack{i=1 \\ i\neq n+1}}^{m} P(A_i) \right\}$$

Hiermit können wir die Induktionsbehauptung in der Form $\sigma(\mathcal{M}_{i_{n+1}}) \subseteq \mathcal{D}$ schreiben. Die Induktionsvoraussetzung impliziert $\mathcal{M}_{i_{n+1}} \subseteq \mathcal{D}$. Wir zeigen jetzt, dass $\mathcal{D}$ ein Dynkin-System ist:

- Es ist $\Omega \in \mathcal{D}$ nach dem eben Bemerkten wegen $\Omega \in \mathcal{M}_{i_{n+1}}$.
- Aus $A \in \mathcal{D}$ folgt $A^c \in \mathcal{D}$ wie folgt:

$$\begin{aligned} P\left(A^c \cap \bigcap_{\substack{i=1 \\ i\neq n+1}}^{m} A_i \right) &= P\left(\Omega \cap \bigcap_{\substack{i=1 \\ i\neq n+1}}^{m} A_i \right) - P\left(A \cap \bigcap_{\substack{i=1 \\ i\neq n+1}}^{m} A_i \right) \\ &= P(\Omega) \prod_{\substack{i=1 \\ i\neq n+1}}^{m} P(A_i) - P(A) \prod_{\substack{i=1 \\ i\neq n+1}}^{m} P(A_i) \\ &= [P(\Omega) - P(A)] \prod_{\substack{i=1 \\ i\neq n+1}}^{m} P(A_i) = P(A^c) \prod_{\substack{i=1 \\ i\neq n+1}}^{m} P(A_i) \end{aligned}$$

  wobei wir $\Omega, A \in \mathcal{D}$ beim zweiten Gleichheitszeichen angewandt haben.

- Es seien $B_j \in \mathcal{D}$ für $j \in \mathbb{N}$ paarweise disjunkt. Wir setzen $B = \bigcup_{j\in\mathbb{N}} B_j$. Mit Hilfe der paarweisen Disjunktheit schließen wir:

$$\begin{aligned} P\left(B \cap \bigcap_{\substack{i=1 \\ i\neq n+1}}^{m} A_i \right) &= P\left(\bigcup_{j\in\mathbb{N}} \left(B_j \cap \bigcap_{\substack{i=1 \\ i\neq n+1}}^{m} A_i \right) \right) \\ &= \sum_{j\in\mathbb{N}} P\left(B_j \cap \bigcap_{\substack{i=1 \\ i\neq n+1}}^{m} A_i \right) = \sum_{j\in\mathbb{N}} P(B_j) \prod_{\substack{i=1 \\ i\neq n+1}}^{m} P(A_i) \\ &= P(B) \prod_{\substack{i=1 \\ i\neq n+1}}^{m} P(A_i) \end{aligned}$$

  wobei wir $B_j \in \mathcal{D}$ beim dritten Gleichheitszeichen angewandt haben.

Aus dem Dynkin-Lemma folgt $\sigma(\mathcal{M}_{i_{n+1}}) \subseteq \mathcal{D}$, also die Induktionsbehauptung.
Die Aussage a) des Satzes ist der Fall $n = m$ des eben Gezeigten.
Zum Beweis der Behauptung b) des Satzes definieren wir die Mengensysteme

$$\mathcal{N}_j = \left\{ \bigcap_{i\in I_j} A_i \,\middle|\, A_i \in \mathcal{M}_i \text{ für } i \in I_j,\ A_i \neq \Omega \text{ für höchstens endlich viele } i \in I_j \right\}, \quad j \in J.$$

Da die Mengensysteme $\mathcal{M}_i$, $i \in I$, durchschnittstabil sind und Ω enthalten, sind auch auch die Mengensysteme $\mathcal{N}_j$, $j \in J$, durchschnittstabil und enthalten Ω.
Es sei $E \subseteq J$ endlich. Wir setzen $F = \bigcup_{j\in J} I_j$. Es sei weiter eine Familie $(A_i \in \mathcal{M}_i)_{i\in F}$ mit der Eigenschaft $A_i \neq \Omega$ für höchstens endlich viele $i \in F$ gegeben. Aus der Unabhängigkeit der $(\mathcal{M}_i)_{i\in I}$ und der paarweisen Disjunktheit der I_j, $j \in E$, folgt

$$P\left(\bigcap_{j\in E}\bigcap_{i\in I_j} A_i\right) = \prod_{j\in E}\prod_{i\in I_j} P(A_i) = \prod_{j\in E} P\left(\bigcap_{i\in I_j} A_i\right);$$

man beachte, dass höchstens endlich viele Faktoren von 1 verschieden sein können. Dies bedeutet genau die Unabhängigkeit der Familie $(\mathcal{N}_j)_{j\in J}$. Teil a) des Satzes, angewandt auf diese Familie, zeigt, dass auch $(\sigma(\mathcal{N}_j))_{j\in J}$ unabhängig ist. Unter Verwendung von $\Omega \in \mathcal{M}_i$ für $i \in I$, wissen wir $\sigma(\mathcal{N}_j)) = \sigma(\bigcup_{i\in I_j} \mathcal{M}_i)$ für $j \in J$. Zusammen folgt die Behauptung b).

□

Diese Sätze werden vielfach – oft implizit, ohne sie zu erwähnen – gebraucht.

Hierzu ein **Beispiel:**
Sind $X, Y, Z : \Omega \to \mathbb{R}$ drei voneinander unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und ist $f : \mathbb{R}^2 \to \mathbb{R}$ Borel-messbar, so sind auch $f(X, Y)$[15] und Z voneinander unabhängige Zufallsvariablen.

Begründung: Wir schließen

$$\begin{aligned}
& X, Y, Z \text{ sind unabhängig} \\
\Leftrightarrow\ & \sigma(X), \sigma(Y), \sigma(Z) \text{ sind unabhängig} \\
\Rightarrow\ & \underbrace{\{A \cap B |\ A \in \sigma(X), B \in \sigma(Y)\}}_{\cap\text{-stabil}}, \sigma(Z) \text{ sind unabhängig} \\
\Rightarrow\ & \sigma(X, Y) = \sigma(\sigma(X) \cup \sigma(Y)), \sigma(Z) \text{ sind unabhängig.} \qquad (263)
\end{aligned}$$

Nun ist $f(X, Y)$ $\sigma(X, Y)$-$\mathcal{B}(\mathbb{R})$-messbar, denn $(X, Y) : \Omega \to \mathbb{R}$ ist $\sigma(X, Y)$-$\mathcal{B}(\mathbb{R})$-messbar und $f : \mathbb{R}^2 \to \mathbb{R}$ ist $\mathcal{B}(\mathbb{R}^2)$-$\mathcal{B}(\mathbb{R})$-messbar. Also sind $f(X, Y)$ und Z voneinander unabhängig.

□

[15] Für diese in der Stochastik gebräuchliche Notation für $f(X, Y) : \Omega \ni \omega \mapsto f(X(\omega), Y(\omega)) \in \mathbb{R}$, in der das Argument $\omega \in \Omega$ wieder weggelassen wird, würde man in anderen Gebieten der Mathematik eher explizit die Komposition schreiben: $f(X, Y) = f \circ (X, Y)$ mit $(X, Y) : \Omega \to \mathbb{R}^2$.

Existenz zweier unabhängiger Zufallsvariablen mit vorgegebener Verteilung: In der Maßtheorie wird der folgende Satz bewiesen:

Satz 2.46 (Produktmaß) *Es seien $(\Omega, \mathcal{A}, P)$ und $(\Sigma, \mathcal{B}, Q)$ zwei Wahrscheinlichkeitsräume. Dann gibt es genau ein Wahrscheinlichkeitsmaß*[16] *$P \times Q$ auf $(\Omega \times \Sigma, \mathcal{A} \times \mathcal{B})$, so dass die beiden Projektionen*

$$X : \Omega \times \Sigma \to \Omega, \qquad X(\omega, \sigma) = \omega \qquad \text{und} \qquad Y : \Omega \times \Sigma \to \Sigma, \qquad Y(\omega, \sigma) = \sigma \quad (264)$$

bezüglich $P \times Q$ voneinander unabhängig mit den Verteilungen $\mathcal{L}_{P\times Q}(X) = P$ und $\mathcal{L}_{P\times Q}(Y) = Q$ sind. Das Maß $P \times Q$ heißt Produktmaß *von P und Q. Das Wahrscheinlichkeitsmaß $P \times Q$ wird für $A \in \mathcal{A} \otimes \mathcal{B}$ so gegeben:*

$$\begin{aligned} P \times Q(A) &= \int_\Omega Q[(\omega, \mathrm{id}_\Sigma) \in A]\, P(d\omega) \\ &= \int_\Omega Q(\{\sigma \in \Sigma |\ (\omega, \sigma) \in A\})\, P(d\omega) \\ &= \int_\Omega \int_\Sigma 1_A(\omega, \sigma)\, Q(d\sigma)\, P(d\omega). \end{aligned} \quad (265)$$

Bemerkung: Auch im Fall, dass P und Q nur Maße sind, liefert die Formel (265) ein Maß $P \times Q$; es heißt ebenfalls Produktmaß. Falls P und Q endliche Maße sind, wird es durch

$$P \times Q(A \times B) = P(A)Q(B) \text{ für } A \in \mathcal{A}, B \in \mathcal{B} \quad (266)$$

charakterisiert.[17]
Beispiele:

1. Das zweidimensionale Lebesguemaß λ_2 ist das Produktmaß $\lambda_1 \times \lambda_1$.
2. Die Gleichverteilung auf dem Rechteck $]a, b] \times]c, d] \subseteq \mathbb{R}^2$, wobei $a < b$ und $c < d$, ist das Produktmaß der Gleichverteilung auf dem Intervall $]a, b]$ mit der Gleichverteilung auf dem Intervall $]c, d]$.

Integration bezüglich des Produktmaßes Für die Formulierung des folgenden Satzes brauchen wir noch eine Verallgemeinerung des Begriffs endlicher Maße:

Definition 2.47 (σ-Endlichkeit) *Ein Maß μ auf einem Ereignisraum $(\Omega, \mathcal{A})$ heißt* σ-endlich*, wenn es eine Folge $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$ mit $\mu(A_n) < \infty$ für alle $n \in \mathbb{N}$ gibt, die den ganzen Raum ausschöpft: $\bigcup_{n\in\mathbb{N}} A_n = \Omega$.*

[16] Statt $P \times Q$ schreibt man manchmal auch $P \otimes Q$.

[17] Gleiches gilt für die nachfolgend definierten σ-endlichen Maße, wenn man die Konventionen $\infty \cdot a = \infty = a \cdot \infty$ für $a \in]0, \infty]$ und $\infty \cdot 0 = 0 = 0 \cdot \infty$ verwendet.

Zum Beispiel ist das Lebesguemaß λ zwar kein endlichens Maß, doch es ist noch σ-endlich, wie man mit der Ausschöpfung $A_n = [-n, n]$, $(n \in \mathbb{N})$ von $\mathbb{R}$ sieht. Fast alle praktisch relevanten Maße sind σ-endlich; nicht σ-endliche Maße haben eine sehr viel geringere Bedeutung.
Die Integration bezüglich des Produktmaßes kann man im σ-endlichen Fall auf sukzessive Integration bezüglich der Faktoren zurückführen, wobei es auf die Integrationsreihenfolge nicht ankommt:[18]

Satz 2.48 (Satz von Fubini für nichtnegative Funktionen) *Es seien $(\Omega, \mathcal{A}, \mu)$ und $(\Sigma, \mathcal{B}, \nu)$ zwei σ-endliche Maßräume und $f : \Omega \times \Sigma \to [0, \infty[$ eine $\mathcal{A} \otimes \mathcal{B}$-$\mathcal{B}([0, \infty])$-messbare Abbildung. Dann sind auch die Abbildungen*

$$\Omega \ni \omega \mapsto \int_\Sigma f(\omega, \sigma)\, \nu(d\sigma) \quad \text{und} \tag{267}$$

$$\Sigma \ni \sigma \mapsto \int_\Omega f(\omega, \sigma)\, \mu(d\omega) \tag{268}$$

$\mathcal{A}$–$\mathcal{B}([0, \infty])$–messbar bzw. $\mathcal{B}$–$\mathcal{B}([0, \infty])$–messbar, und es gilt: **Wichtig!**

$$\begin{aligned} \int_{\Omega \times \Sigma} f\, d(\mu \times \nu) &= \int_\Omega \int_\Sigma f(\omega, \sigma)\, \nu(d\sigma)\, \mu(d\omega) \\ &= \int_\Sigma \int_\Omega f(\omega, \sigma)\, \mu(d\omega)\, \nu(d\sigma). \end{aligned}$$

Dieser Satz wird in der Maßtheorie bewiesen.

Bemerkungen:

1. Im Spezialfall

$$(\Omega, \mathcal{A}, \mu) = (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m), \lambda_m), \tag{269}$$

$$(\Sigma, \mathcal{B}, \nu) = (\mathbb{R}^{n-m}, \mathcal{B}(\mathbb{R}^{n-m}), \lambda_m), \tag{270}$$

$$\lambda_m \times \lambda_{n-m} = \lambda_n \tag{271}$$

erhalten wir daraus die früher besprochene Variante 2.29 des Satzes von Fubini für das Lebesguemaß.

[18] Für nicht σ-endliche Maße kann das falsch werden: Bezeichnen zum Beispiel λ das Lebesguemaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ und ν das Zählmaß auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, so gilt für die Diagonale $\Delta = \{(x, x) |\ x \in \mathbb{R}\}$:

$$\int_{\mathbb{R}} \int_{\mathbb{R}} 1_\Delta(x, y)\, \lambda(dx)\, \nu(dy) = 0 \neq \infty = \int_{\mathbb{R}} \int_{\mathbb{R}} 1_\Delta(x, y)\, \nu(dy)\, \lambda(dx).$$

2. Den Spezialfall

$$(\Omega, \mathcal{A}, \mu) = (\Sigma, \mathcal{B}, \nu) = (\mathbb{N}, \mathcal{P}(\mathbb{N}), \text{Zählmaß}) \tag{272}$$

kann man als den großen Umordnungssatz für Reihen mit nichtnegativen Summanden auffassen.

Übung 2.49 (Dichten unabhängiger Zufallsvariablen) *1. Es seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X, Y : \Omega \to \mathbb{R}$ Zufallsvariablen mit Dichten f bzw. g. Dann sind X und Y genau dann unabhängig, wenn*

$$h : \mathbb{R}^2 \to [0, \infty], \quad h(x, y) = f(x)g(y)$$

eine Dichte für $Z = (X, Y) : \Omega \to \mathbb{R}^2$ ist.

2. Verallgemeinerung: Es seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $(\Sigma_1, \mathcal{B}_1, \mu_1)$ und $(\Sigma_2, \mathcal{B}_2, \mu_2)$ σ-endliche Maßräume und $X_i : \Omega \to \Sigma_i$, $(i = 1, 2)$, Zufallsvariablen mit Dichten f_i bezüglich μ_i. Dann sind X_1 und X_2 genau dann unabhängig voneinander, wenn

$$h : \Sigma_1 \times \Sigma_2 \to [0, \infty], \quad h(x_1, x_2) = f_1(x_1)f_2(x_2) \tag{273}$$

eine Dichte der zusammengesetzten Zufallsvariablen

$$Z = (X_1, X_2) : \Omega \to \Sigma_1 \times \Sigma_2 \tag{274}$$

bezüglich des Produktmaßes $\mu_1 \times \mu_2$ ist.

Diese Aussagen haben auch unmittelbare Verallgemeinerungen auf n Faktoren, die wir hier weder formulieren noch beweisen.

Bemerkung: Produktmaßbildung ist assoziativ:

$$(\mu_1 \times \mu_2) \times \mu_3 = \mu_1 \times (\mu_2 \times \mu_3) \tag{275}$$

2.9 Die Faltung

Die Faltung beschreibt die Verteilung der Summe zweier unabhängiger Zufallsvariablen mit gegebenen Verteilungen:

Definition 2.50 (Faltung von Maßen) *Es seien μ_1 und μ_2 zwei σ-endliche Maße über $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Die* Faltung *von μ_1 mit μ_2, in Zeichen $\mu_1 * \mu_2$, wird als das Bildmaß des Produktmaßes $\mu_1 \times \mu_2$ unter der Additionsabbildung $+ : \mathbb{R}^2 \to \mathbb{R}$ definiert.*

Für Wahrscheinlichkeitsmaße P_1 und P_2 können wir das auch so formulieren:
Sind X und Y zwei unabhängige Zufallsvariablen mit den Verteilungen $\mathcal{L}(X) = P_1$ und $\mathcal{L}(Y) = P_2$, so hat die Summe $X + Y$ die Verteilung $\mathcal{L}(X + Y) = P_1 * P_2$.

Die Faltung zweier σ-endlicher Maße auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ wird analog definiert.

Beispiel: Ist $P = p\delta_1 + (1-p)\delta_0$ mit $0 \leq p \leq 1$ die Verteilung für einen Münzwurf, so folgt:

$$P * P = p^2\delta_2 + 2p(1-p)\delta_1 + (1-p)^2\delta_0, \tag{276}$$

$$P * P * P = p^3\delta_3 + 3p^2(1-p)\delta_2 + 3p(1-p)^2\delta_2 + (1-p)^3\delta_0, \tag{277}$$

$$\vdots$$

$$P^{*n} = \underbrace{P * \ldots * p}_{n \text{ Faktoren}} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k = \text{binomial}(n,p) \tag{278}$$

Übung 2.51 (Faltung über $\mathbb{Z}$) *Es seien μ und ν zwei Verteilungen auf $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}))$ mit den Zähldichten f bzw. g. Zeigen Sie, dass die Faltung $\mu * \nu$ die Zähldichte*

$$f * g(z) := \sum_{x \in \mathbb{Z}} f(x) g(z-x) = \sum_{y \in \mathbb{Z}} f(z-y) g(y) \tag{279}$$

besitzt.

Im kontinuierlichen Fall gilt das folgende Analogon:

Satz 2.52 (Faltung von Dichten) *Es seien μ und ν σ-endliche Maße auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit den Dichten f bzw. g. Dann besitzt die Faltung $\mu * \nu$ die Dichte $f * g$, definitert durch*

$$\boxed{f * g(z) = \int_{\mathbb{R}} f(x) g(z-x)\, dx = \int_{\mathbb{R}} f(z-y) g(y)\, dy \quad (z \in \mathbb{R})} \tag{280}$$

Die Funktion $f * g$ wird ebenfalls die Faltung *der Dichten f und g genannt.*

Beweis: Das Produktmaß $\mu \times \nu$ besitzt die Dichte

$$h : \mathbb{R}^2 \to \mathbb{R}, \quad h(x,y) = f(x)g(y). \tag{281}$$

Nun betrachten wir den Diffeomorphismus

$$k : \mathbb{R}^2 \to \mathbb{R}^2, \quad k(x,y) = (x+y, x) \tag{282}$$

mit der Inversen

$$k^{-1} : \mathbb{R}^2 \to \mathbb{R}^2, \quad k^{-1}(z,x) = (x, z-x) \tag{283}$$

sowie die erste Projektion

$$\pi : \mathbb{R}^2 \to \mathbb{R}^2, \quad \pi(z,x) = z. \tag{284}$$

Durch Komposition erhalten wir

$$\pi \circ k = + : (x,y) \mapsto x + y \tag{285}$$

Wir berechnen die Jacobideterminante

$$|\det D(k^{-1}(z,x)| = \left|\det \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}\right| = 1. \tag{286}$$

Mit der Transformationsformel folgt: Das Bildmaß $k[\mu \times \nu]$ besitzt die Dichte

$$\begin{aligned} &j : \mathbb{R}^2 \to \mathbb{R}, \\ &j(z,x) = h(k^{-1}(z,x) \underbrace{|\det D(k^{-1})(z,x)|}_{=1} = h(x, z-x) = f(x)g(z-x). \end{aligned} \tag{287}$$

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

$$\mathbb{R} \ni z \mapsto \int_{\mathbb{R}} j(z,x)\, dx = \int_{\mathbb{R}} f(x)g(z-x)\, dx = f * g(z). \tag{288}$$

Die Gleichung

$$\int_{\mathbb{R}} f(x)g(z-x)\, dx = \int_{\mathbb{R}} f(z-y)g(y)\, dy \tag{289}$$

folgt mit der Substitution $y = z - x$ oder auch aus der Beobachtung $\mu * \nu = \nu * \mu$.

□

Beispiele:

1. **"Rechteck∗Rechteck=Dreieck":** Es sei $P = \text{unif}[0,1]$. Das Wahrscheinlichkeitsmaß P hat die Dichte $1_{[0,1]}$. Dann besitzt das Produktmaß $P \times P = \text{unif}([0,1]^2)$ die Dichte $1_{[0,1]^2}$, und $P * P$ die Dichte

$$1_{[0,1]} * 1_{[0,1]}(z) = \int_{\mathbb{R}} 1_{[0,1]}(x) 1_{[0,1]}(z-x)\, dx = z 1_{[0,1]}(z) + (2-z) 1_{]1,2]}(z). \tag{290}$$

Übung 2.53 *Verifizieren Sie die Rechnung* (290).

2. **Gammaverteilungen als Faltungshalbgruppe:** Es seien X und Y zwei voneinander unabhänige, gammaverteilte Zufallsvariablen:

$$\mathcal{L}_P(X) = \text{Gamma}(a,s), \quad \mathcal{L}_P(Y) = \text{Gamma}(a,t), \tag{291}$$

mit dem gleichen Skalenparameter $a > 0$ und den Formparametern $s, t > 0$. Wir zeigen: $X + Y$ ist $\text{Gamma}(a, s+t)$-verteilt, d.h.

$$\boxed{\text{Gamma}(a,s) * \text{Gamma}(a,t) = \text{Gamma}(a, s+t)} \tag{292}$$

Beweis: X bzw. Y besitzen die Dichte

$$f(x) = 1_{]0,\infty[}(x)\frac{a^s}{\Gamma(s)}x^{s-1}e^{-ax}, \qquad (x \in \mathbb{R}) \qquad bzw. \tag{293}$$

$$g(y) = 1_{]0,\infty[}(y)\frac{a^t}{\Gamma(t)}y^{t-1}e^{-ay}, \qquad (y \in \mathbb{R}). \tag{294}$$

Es folgt:

$$\begin{aligned} f * g(z) &= \frac{a^{s+t}}{\Gamma(s)\Gamma(t)} \int_{\mathbb{R}} 1_{]0,\infty[}(x)1_{]0,\infty[}(z-x)x^{s-1}(z-x)^{t-1}e^{-ax}e^{-a(z-x)}\,dx \\ &= \frac{a^{s+t}e^{-az}}{\Gamma(s)\Gamma(t)}1_{]0,\infty[}(z)\int_0^z x^{s-1}(z-x)^{t-1}\,dx. \end{aligned} \tag{295}$$

Das letzte Integral berechnen wir mit der Substitution $x = zu$:

$$\begin{aligned} &\int_0^z x^{s-1}(z-x)^{t-1}\,dx = \int_0^1 (zu)^{s-1}(z-zu)^{t-1}z\,du \\ &= z^{s+t-1}\int_0^1 u^{s-1}(1-u)^{t-1}\,du = z^{s+t-1}B(s,t), \end{aligned} \tag{296}$$

wobei wir die *Betafunktion*

$$B :]0,\infty[^2 \to \mathbb{R}, \qquad B(s,t) = \int_0^1 u^{s-1}(1-u)^{t-1}\,du \tag{297}$$

verwenden. Damit ist gezeigt:

$$f * g(z) = 1_{]0,\infty[}(z)a^{s+t}\frac{B(s,t)}{\Gamma(s)\Gamma(t)}z^{s+t-1}e^{-az}, \qquad (z \in \mathbb{R}). \tag{298}$$

Nun ist dies bis auf die Konstante $\frac{B(s,t)}{\Gamma(s)\Gamma(t)}$ statt $\frac{1}{\Gamma(s+t)}$ die Dichte der Verteilung Gamma$(a, s+t)$. Weil sowohl die Faltung Gamma$(a,s) *$ Gamma(a,t) als auch die Gammaverteilung Gamma$(a, s+t)$ Wahrscheinlichkeitsmaße sind, müssen die Normierungskonstanten übereinstimmen:

$$\frac{B(s,t)}{\Gamma(s)\Gamma(t)} = \frac{1}{\Gamma(s+t)}, \tag{299}$$

und wir erhalten die Behauptung.

□

Bemerkung: Als Nebenprodukt lieferte der Beweis die folgende Beziehung zwischen der Betafunktion und der Gammafunktion:

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}, \qquad (s,t > 0) \tag{300}$$

3. **Die Chi-Quadrat-Verteilung.**

Definition 2.54 (χ^2-Verteilung) *Es seien $X_1, \ldots, X_n$ unabhängige standardnormalverteilte Zufallsvariablen, wobei $n \in \mathbb{N}$. Die Verteilung der Quadratsumme*

$$\chi_n^2 := \sum_{k=1}^{n} X_k^2 \tag{301}$$

wird die Chi-Quadrat-Verteilung (synonym: χ^2-Verteilung) mit n Freiheitsgraden genannt. Sie wird ebenfalls mit χ_n^2 bezeichnet.

Wir überlegen uns nun, dass Chi-Quadrat-Verteilungen spezielle Gammaverteilungen sind.

Übung 2.55 *Beweisen Sie, dass $\mathcal{L}(X_1^2) = \text{Gamma}(\frac{1}{2}, \frac{1}{2})$ gilt.*

Es folgt mit der Faltungseigenschaft der Gammaverteilungen:

$$\mathcal{L}(\chi_n^2) = \text{Gamma}(\frac{1}{2}, \frac{1}{2})^{*n} = \text{Gamma}(\frac{1}{2}, \frac{n}{2}). \tag{302}$$

Damit ist gezeigt:

Lemma 2.56 (Zusammenhang zwischen χ^2-Verteilung und Gammaverteilung) *Für alle $n \in \mathbb{N}$ ist die χ^2-Verteilung mit n Freiheitsgraden das Gleiche wie die Gammaverteiung $\Gamma(\frac{1}{2}, \frac{n}{2})$.*

Die gemeinsame Verteilung von $X = (X_1, \ldots, X_n)$ unserer n unabhängigen, standardnormalverteilten Zufallsvariablen $X_1, \ldots, X_n$, also die Verteilung auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ mit der Dichte

$$\begin{aligned} f(x) &= \prod_{k=1}^{n} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_k^2} \right) \\ &= (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \sum_{k=1}^{n} x_k^2 \right) \\ &= (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\|x\|_2^2}, \qquad (x = (x_1, \ldots, x_n) \in \mathbb{R}^n), \end{aligned} \tag{303}$$

heißt auch die *n-dimensionale Standardnormalverteilung*. Die χ^2-Verteilung mit n Freiheitsgraden ist also die Verteilung von $\|X\|_2^2$, wenn X n-dimensional standardnormalverteilt ist.

4. **Faltung von Normalverteilungen:**

Definition 2.57 (Normalverteilung) *Die (1-dimensionale)* Normalverteilung *mit den Parametern $\mu \in \mathbb{R}$ (dem "Erwartungswert") und $\sigma^2 > 0$ (der "Varianz") ist die Verteilung der Zufallsvariablen*

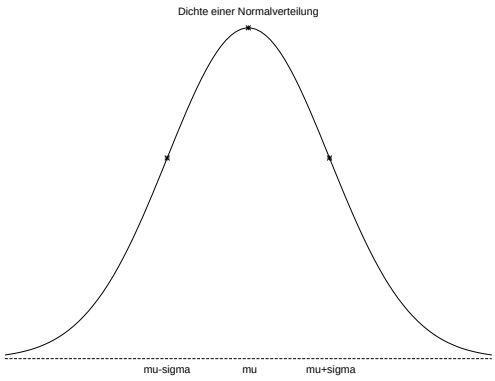
$$X = \sigma Z + \mu, \tag{304}$$

wenn Z standardnormalverteilt ist und $\sigma = \sqrt{\sigma^2}$. Sie wird mit $\mathrm{N}(\mu, \sigma^2)$ bezeichnet.

Die Normalverteilung $N(\mu, \sigma^2)$ besitzt die folgende Dichte: **sehr wichtig!**

$$\boxed{f_{\mu,\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \tag{305}$$

Der Graph dieser Dichte ist die berühmte Gaußsche Glockenkurve:



Begründung der Formel (305)**:** Es sei $g : \mathbb{R} \to \mathbb{R}$, $x = g(z) = \sigma z + \mu$ die Transformation in der Definition der Normalverteilung $\mathrm{N}(\mu, \sigma^2)$. Dann gilt für alle $A \in \mathcal{B}(\mathbb{R})$:

$$\begin{aligned}
\mathrm{N}(\mu, \sigma^2)(A) &= \frac{1}{\sqrt{2\pi}} \int_{g^{-1}[A]} e^{-\frac{z^2}{2}}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \int_A e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \underbrace{\left|\frac{dz}{dx}\right|}_{=\frac{1}{\sigma}} dx \qquad \text{(mit der Transformation } z = \frac{x-\mu}{\sigma}\text{)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_A e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx. \qquad (306)
\end{aligned}$$

□

Die Faltung zweier Normalverteilungen ergibt wieder eine Normalverteilung. Genauer gesagt gilt:

Satz 2.58 (Faltung von Normalverteilungen) *Für $\mu_1, \mu_2 \in \mathbb{R}$ und $\sigma_1^2, \sigma_2^2 > 0$ gilt:*

$$\boxed{\mathrm{N}(\mu_1, \sigma_1^2) * \mathrm{N}(\mu_2, \sigma_2^2) = \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)} \tag{307}$$

Anders gesagt: Sind X und Y voneinander unabängige Zufallsvariablen mit den Verteilungen

$$\mathcal{L}(X) = \mathrm{N}(\mu_1, \sigma_1^2) \quad \text{und} \quad \mathcal{L}(Y) = \mathrm{N}(\mu_2, \sigma_2^2), \tag{308}$$

so ist auch die Summe $X + Y$ wieder normalverteilt:

$$\mathcal{L}(X + Y) = \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \tag{309}$$

Beweis: Wir zeigen

$$f_{\mu_1,\sigma_1^2} * f_{\mu_2,\sigma_2^2} = f_{\mu_1+\mu_2,\sigma_1^2+\sigma_2^2} \tag{310}$$

durch direkte Rechnung. Es gilt für $x \in \mathbb{R}$:

$$f_{\mu_1,\sigma_1^2} * f_{\mu_2,\sigma_2^2}(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_1^2}} \underbrace{\int_{\mathbb{R}} e^{-\frac{(x-y-\mu_1)^2}{2\sigma_1^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}} \, dy}_{=:I_1}. \tag{311}$$

Wir substituieren

$$\tilde{x} := x - \mu_1 - \mu_2, \tag{312}$$

$$\tilde{y} := y - \mu_2. \tag{313}$$

Damit erhalten wir für das letzte Integral:

$$I_1 = \int_{\mathbb{R}} \exp\left[-\frac{1}{2} \underbrace{\left(\frac{(\tilde{x} - \tilde{y})^2}{\sigma_1^2} + \frac{\tilde{y}^2}{\sigma_2^2} \right)}_{=:I_2} \right] d\tilde{y} \tag{314}$$

Wir schreiben den Term I_2 mit einer quadratischen Ergänzung um:

$$\begin{aligned} I_2 &= (\sigma_1^{-2} + \sigma_2^{-1})\tilde{y}^2 - 2\sigma_1^{-2}\tilde{x}\tilde{y} + \sigma_1^{-2}\tilde{x}^2 \\ &= (\sigma_1^{-2} + \sigma_2^{-1}) \left(\tilde{y} - \frac{\sigma_1^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \tilde{x} \right)^2 \underbrace{- \frac{\sigma_1^{-4}}{\sigma_1^{-2} + \sigma_2^{-2}} \tilde{x}^2 + \sigma_1^{-2}\tilde{x}^2}_{=:I_3}. \end{aligned} \tag{315}$$

Den letzten Term I_3 vereinfachen wir:

$$I_3 = \frac{-\sigma_1^{-4} + \sigma_1^{-4} + \sigma_1^{-2}\sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \tilde{x}^2 = \frac{1}{\sigma_1^{-2} + \sigma_2^{-2}} \tilde{x}^2. \tag{316}$$

Eingesetzt erhalten wir:

$$I_1 = \exp\left(-\frac{1}{2}\frac{\tilde{x}^2}{\sigma_1^2+\sigma_2^2}\right) \underbrace{\exp\left[-\frac{1}{2}(\sigma_1^{-2}+\sigma_2^{-2})\left(\tilde{y}-\frac{\sigma_2^{-2}}{\sigma_1^{-2}+\sigma_2^{-2}}\right)^2\right] d\tilde{y}}_{=:I_4} \tag{317}$$

Wir setzen

$$\sigma := \sqrt{\sigma_1^2+\sigma_2^2}, \tag{318}$$

also

$$\sigma_1^{-2}+\sigma_2^{-2} = \frac{\sigma^2}{\sigma_1^2\sigma_2^2} \tag{319}$$

und substituieren die Integrationsvariable:

$$z := \frac{\sigma}{\sigma_1\sigma_2}\left(\tilde{y}-\frac{\sigma_1^{-2}}{\sigma_1^{-2}\sigma_2^{-2}}\tilde{x}\right), \tag{320}$$

$$\frac{dz}{d\tilde{y}} = \frac{\sigma}{\sigma_1\sigma_2}. \tag{321}$$

Wir erhalten für das letzte Integral:

$$I_4 = \int_{\mathbb{R}} e^{-\frac{1}{2}z^2}\left(\frac{dz}{d\tilde{y}}\right)^{-1} dz = \frac{\sigma_1\sigma_2}{\sigma}\int_{\mathbb{R}} e^{-\frac{1}{2}z^2}\,dz = \frac{\sigma_1\sigma_2}{\sigma}\sqrt{2\pi}. \tag{322}$$

Es folgt

$$I_1 = \frac{\sigma_1\sigma_2}{\sigma}\sqrt{2\pi}\exp\left(-\frac{1}{2}\frac{\tilde{x}^2}{\sigma^2}\right). \tag{323}$$

Eingesetzt in Formel (311) ergibt das:

$$f_{\mu_1,\sigma_1^2} * f_{\mu_2,\sigma_2^2}(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{1}{2}\frac{\tilde{x}^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{\tilde{x}^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \tag{324}$$

wobei wir $\mu := \mu_1+\mu_2$ abgekürzt haben.

□

5. **Die Poissonhalbgruppe:** Die *Poissonverteilung* zum Parameter $\lambda > 0$ ist die Verteilung auf $\mathbb{N}_0$ mit der Zähldichte

$$p_\lambda(n) = e^{-\lambda}\frac{\lambda^n}{n!} \qquad (n \in \mathbb{N}_0). \tag{325}$$

Sie wird mit Poisson(λ) bezeichnet. Dann gilt für alle $\lambda_1, \lambda_2 > 0$:

$$\boxed{\text{Poisson}(\lambda_1) * \text{Poisson}(\lambda_2) = \text{Poisson}(\lambda_1+\lambda_2)} \tag{326}$$

Übung 2.59 *Beweisen Sie die Formel* (326).

2.10 Folgen unabhängiger Zufallsvariablen

Wir zitieren zunächst einen allgemeinen Existenzsatz für Produkte beliebig vieler Wahrscheinlichkeitsräume:

Satz 2.60 (Produktmaß von beliebig vielen Wahrscheinlichkeitsmaßen) *Es sei $(\Omega_i, \mathcal{A}_i, P_i)_{i\in I}$ eine beliebige Familie von Wahrscheinlichkeitsräumen, $\Omega = \prod_{i\in I} \Omega_i$ das kartesische Produkt der Ω_i, $X_i : \Omega \to \Omega_i$, $X(\omega) = \omega_i$ für $i \in I$, $\omega = (\omega_j)_{j\in I} \in \Omega$ die kartesische Projektion und $\mathcal{A} = \bigotimes_{i\in I} \mathcal{A}_i = \sigma(X_i : \ i \in I)$ die von den X_i erzeugte σ-Algebra. Dann gibt es genau ein Wahrscheinlichkeitsmaß P auf $(\Omega, \mathcal{A})$, so dass alle X_i, $i \in I$, unabhängig mit den Verteilungen P_i sind. Wir schreiben*

$$\prod_{i\in I} P_i := P \tag{327}$$

für dieses Maß; es wird das Produktmaß *der P_i genannt.*

Wir beweisen diesen Satz nicht in dieser Vorlesung, weil der Beweis eher in die Maßtheorie als in die Stochastik gehört. Der Beweis benutzt den Fortsetzungssatz von Carathéodory.

Als eine Alternative zum Satz konstruieren wir statt dessen direkt Folgen unabhängiger Zufallsvariablen mit Hilfe der kontinuierlichen Gleichverteilung.

Unabhängige, identisch verteilte Zufallsvariablen

Definition 2.61 (i.i.d.) *Eine Familie $(X_i)_{i\in I}$ von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Werten in stets dem gleichen Ereignisraum $(\Omega', \mathcal{A}')$ wird* i.i.d. *genannt (Abkürzung aus dem Englischen* independent identically distributed*), wenn die X_i, $i \in I$ unabhängig voneinander mit der gleichen Verteilung $\mathcal{L}_P(X_i) = \mathcal{L}_P(X_j)$, $(i, j \in I)$ sind.*

Wir zeigen jetzt die Existenz von i.i.d. "fairen Münzwürfen". Es sei dazu

$$(\Omega, \mathcal{A}, P) = ([0, 1[, \mathcal{B}([0, 1[), P). \tag{328}$$

Für $\omega \in \Omega$ sei $X_n(\omega)$, $n \in \mathbb{N}$, die n-te Nachkommaziffer in der Binärdarstellung von ω, also

$$X_n(\omega) = \lfloor 2^n\omega \rfloor - 2 \cdot \lfloor 2^{n-1}\omega \rfloor, \tag{329}$$

wobei

$$\lfloor x \rfloor = \max\{z \in \mathbb{Z} |\ z \le x\} \tag{330}$$

den ganzzahligen Anteil von $x \in \mathbb{R}$ bezeichnet.

Satz 2.62 (Binärziffern als faire Münzwürfe) *Die Binärziffern $(X_n)_{n\in\mathbb{N}}$ sind i.i.d. auf $\{0, 1\}$ gleichverteilte Zufallsvariablen über $(\Omega, \mathcal{A}, P)$.*

Die Folge $(X_n)_{n\in\mathbb{N}}$ ist also ein Modell für die abzählbar unendliche Wiederholung eines fairen Münzwurfs.

Beweis: Wir müssen für alle endlichen $n \in \mathbb{N}$ zeigen: Die (gemeinsame) Verteilung des Zufallsvektors $(X_k)_{k\in[n]}$ ist gleich $\prod_{k\in[n]} \frac{1}{2}(\delta_0 + \delta_2)$, also gleich der Gleichverteilung auf $\{0,1\}^n$. Hierzu sei $x = (x_k)_{k\in[n]} \in \{0,1\}^n$ gegeben. Wir setzen

$$a := \sum_{k=1}^{n} 2^{-k} x_k \in [0,1[. \tag{331}$$

Dannn gilt

$$\begin{aligned}
&\{X_k = x_k \text{ für alle } k \in [n]\} && (332)\\
&= \{\omega \in [0,1[|\ 0, x_1 \dots x_n \text{ ist der Beginn der Binärdarstellung von } \omega\} && (333)\\
&= [a, a + 2^{-n}[, && (334)
\end{aligned}$$

also

$$P[X_k = x_k \text{ für alle } k \in [n]] = P([a, a + 2^{-n}[) = 2^{-n} = \frac{1}{|\{0,1\}|^n}. \tag{335}$$

Die Verteilung $\mathcal{L}(X_1, \dots, X_n)$ hat also in der Tat die Zähldichte $(|\{0,1\}|^{-1})_{x\in\{0,1\}^n}$.

□

Aus i.i.d. fairen Münzwürfen kann man auch umgekehrt die Gleichverteilung auf $[0,1]$ rekonstruieren:

Übung 2.63 (i.i.d. Folgen mit beliebiger Verteilung aus i.i.d. fairen Münzwürfen)
Es seien $(X_n)_{n\in\mathbb{N}}$ i.i.d. $\frac{1}{2}(\delta_0 + \delta_1)$-verteilte Zufallsvariablen auf einem beliebigen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$.

1. *Zeigen Sie, dass*

$$Z = \sum_{n\in\mathbb{N}} 2^{-n} X_n : \Omega \to [0,1] \tag{336}$$

   uniform auf $[0,1]$ verteilt ist.

2. *Konstruieren Sie eine i.i.d. Folge $(Z_n)_{n\in\mathbb{N}}$ von uniform auf $[0,1]$ verteilten Zufallsvariablen über dem gegebenen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$.*
   Hinweis: Verwenden Sie eine Bijektion $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$.

3. *Gegeben ein Wahrscheinlichkeitsmaß Q auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, konstruieren Sie eine i.i.d. Folge $(Y_n)_{n\in\mathbb{N}}$ auf $(\Omega, \mathcal{A}, P)$ mit Verteilung $\mathcal{L}_P(Y_n) = Q$.*
   Hinweis: Quantilstransformation. Es genügt auch, wenn Sie die Zufallsvariablen Y_n nur P-fast sicher definieren.

Das folgende Kriterium ist zum Nachweis der Unabhängigkeit diskreter Zufallsvariablen nützlich:

Lemma 2.64 (Unabhängigkeit im diskreten Fall) *Es seien $X_1, \ldots, X_n$ Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Werten in einer abzählbaren Menge $(N, \mathcal{P}(N))$. Dann sind die $X_1, \ldots, X_n$ genau dann unabhängig, wenn für alle $k_1, \ldots, k_n \in N$ gilt:*

$$P[X_i = k_i \text{ für alle } i \in [n]] = \prod_{i=1}^{n} P[X_i = k_i]. \tag{337}$$

Beweis: Der Beweisteil "$\Rightarrow$" folgt unmittelbar aus der Definition der Unabhängigkeit von Zufallsvariablen.
Zu "$\Leftarrow$": Wir beobachten, dass das Mengensystem

$$\{\{X_i = k_i \text{ für alle } i \in [n]\} |\ k_1, \ldots, k_n \in N\} \cup \{\emptyset\} \tag{338}$$

ein $\cap$-stabiles Erzeugendensystem der σ-Algebra $\sigma(X_1, \ldots, X_n)$ ist. Die Behauptung folgt dann unmittelbar aus dem Satz 2.45 (Vererbung der stochastischen Unabhängigkeit).

□

2.11 Einige Standardverteilungen

2.11.1 Die geometrische Verteilung

Eine unfaire Münze wird bis zum ersten Auftreten der "1" geworfen. Wir bestimmen die Verteilung der Anzahl der Würfe.

Modell: Es sei $0 < p < 1$ und X_t, $t \in \mathbb{N}$, i.i.d. $p\delta_1 + (1-p)\delta_0$-verteilt über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Wir setzen

$$T := \inf\{t \in \mathbb{N} |\ X_t\}. \tag{339}$$

Insbesondere gilt für alle $s \in \mathbb{N}$:

$$\{T = s\} = \{X_t = 0 \text{ für alle } t < s,\ X_s = 1\}, \tag{340}$$

also

$$\begin{aligned} P[T=s] &= P[X_t = 0 \text{ für alle } t < s,\ X_s = 1] \\ &= \left(\prod_{t=1}^{s-1} P[X_t = 0]\right) P[X_s = 1] \quad (\text{da } (X_t)_t \text{ unabhängig}) \qquad (341) \\ &= (1-p)^{s-1}p, \\ P[T=\infty] &= P[X_t = 0 \text{ für alle } t \in \mathbb{N}] \\ &= \lim_{s\to\infty} P[X_t = 0 \text{ für alle } t \in [s]] \quad (\text{da } P \ \sigma\text{-stetig von oben}) \\ &= \lim_{s\to\infty} \prod_{t=1}^{s} P[X_t = 0] \quad (\text{da } (X_t)_t \text{ unabhängig}) \\ &= \lim_{s\to\infty} (1-p)^s = 0.^{19} \qquad (342) \end{aligned}$$

Es folgt:

$$\boxed{\mathcal{L}(T) = \sum_{s=1}^{\infty} (1-p)^{s-1} p \delta_s} \qquad (344)$$

Diese Verteilung heißt *geometrische Verteilung* mit dem Parameter p.

Bemerkung: Hier und in vielen anderen Anwendungen ist es nicht nötig, den Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ zu spezifizieren; nur die (gemeinsame) Verteilung der relevanten Zufallsvariablen wird fixiert. Das Zentrum unserer Untersuchungen hat sich also nun von den Wahrscheinlichkeitsräumen zu den Zufallsvariablen und ihren Verteilungen verschoben.

2.11.2 Die negative Binomialverteilung

Im Modell von eben sei T_n für $n \in \mathbb{N}$ die Anzahl der Würfe bis zum n-ten Auftreten einer "1". Formal definieren wir das so:

$$T_0 := 0, \qquad (345)$$

$$T_n := \inf\{t > T_{n-1} |\ X_t = 1\} \quad \text{für } n \in \mathbb{N}. \qquad (346)$$

[19] Das gleiche Argument liefert für alle $s \in \mathbb{N}$:

$$P[X_t = 0 \text{ für alle } t \geq s] = 0 \qquad (343)$$

Insbesondere ist $T_1 = T$ geometrisch verteilt. P-fast sicher sind alle T_n endlich. In der Tat:

$$\begin{aligned}
&P[\exists n:\ T_n = \infty] = P[X_t = 0 \text{ schließlich für } t \to \infty] \\
&= P[\exists s \in \mathbb{N} \underbrace{\forall t \geq s:\ X_t = 0}_{\substack{\text{monoton steigendes} \\ \text{Ereignis in } s}}] \\
&= \lim_{s\to\infty} P[\forall t \geq s:\ X_t = 0] \quad \text{(mit } \sigma\text{-Stetigkeit von unten)} \\
&= \lim_{s\to\infty} 0 = 0 \quad \text{(mit Formel (343)).}
\end{aligned} \tag{347}$$

Insbesondere ist die Wartezeit[20] $T_n - T_{n-1}$ zwischen dem $(n-1)$-ten und dem n-ten Auftreten der "1" P-fast sicher wohldefiniert.
Wir zeigen jetzt:

Lemma 2.65 (i.i.d. geometrisch verteilte Wartezeiten) *Die Zufallsvariablen $T_n - T_{n-1}$, $n \in \mathbb{N}$, sind i.i.d. geometrisch verteilt mit dem Parameter p.*

Beweis: Es seien $s_1, \ldots, s_n \in \mathbb{N}$ gegeben. Wir setzen

$$t_k := \sum_{i=1}^{k} s_i \tag{348}$$

für $k \in [n]$. Dann gilt:

$$\begin{aligned}
&P[T_k - T_{k-1} = s_k \text{ für alle } k \in [n]] = P[T_k = t_k \text{ für alle } k \in [n]] \\
&= P[X_{t_k} = 1 \text{ für alle } k \in [n] \text{ und } X_t = 0 \text{ für alle } t \in [t_n] \setminus \{t_1, t_2, \ldots, t_n\}] \\
&= p^n (1-p)^{t_n - n} \quad \text{(da } (X_t)_t \text{ i.i.d. } p\delta_1 + (1-p)\delta_0\text{-verteilt)} \\
&= \prod_{k=1}^{n} [(1-p)^{s_k - 1} p] \\
&= \prod_{k=1}^{n} P[T = s_k] \quad \text{(da } T \text{ geometrisch}(p) \text{ verteilt).}
\end{aligned} \tag{349}$$

□

Anschaulich gesprochen reflektiert das die "Gedächtnislosigkeit" des Münzwurfs: Die Verteilung der Wartezeit auf die nächste "1" ist immer die gleiche, gleichgültig welche Wartezeiten früher aufgetreten sind.
Wegen

$$T_n = \sum_{k=1}^{n} (T_k - T_{k-1}) \tag{350}$$

[20] Diese Interpretation gilt für $n > 1$.

können wir das auch so formulieren:
Die Wartezeit T_n ist eine Summe von n i.i.d. geometrisch(p)-verteilten Zufallsvariablen. Insbesondere ist $\mathcal{L}_P(T_n)$ die Faltung von n gleichen geometrischen Verteilungen.
Wir berechnen jetzt die Verteilung $\mathcal{L}_P(T_n)$ explizit. Hierzu sei $t \in \mathbb{N}$ mit $t \geq n$ gegeben. (Man beachte, dass der Fall $T_n < n$ nicht auftreten kann.) Wir erhalten:

$$P[T_n = t] = P[X_t = 1,\ |\{s < t|\ X_s = 1\}| = n-1] \tag{351}$$

$$= \sum_{\substack{E \subseteq [t-1] \\ |E| = n-1}} P[X_t = 1,\ \forall s \in E : X_s = 1,\ \forall s \in [t-1] \setminus E : X_s = 0] \tag{352}$$

$$= |\{E \subseteq [t-1] :\ |E| = n-1\}| p^n (1-p)^{t-n} = \binom{t-1}{n-1} p^n (1-p)^{t-n}. \tag{353}$$

Definition 2.66 (negative Binomialverteilung) *Die Verteilung*

$$\mathcal{L}_P(T_n) = \sum_{t=n}^{\infty} \binom{t-1}{n-1} p^n (1-p)^{t-n} \delta_t \tag{354}$$

auf $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ *wird* negative Binomialverteilung *zu den Parametern* $n \in \mathbb{N}$ *und* $p \in]0,1[$ *genannt.*

2.11.3 Seltene Ereignisse: Die Poissonverteilung

"Sehr häufige unabhängige Wiederholungen "seltener" Ereignisse treten in Anwendungen oft auf:
Beispiele:

- Anzahl der Regentropfen, die einen Regenschirm während eines Regenschauers während einer Sekunde treffen:
  - Viele Tropfen sind in der Regenwolken, die unabhängig voneinander den Schirm treffen könnten.
  - Die Wahrscheinlichkeit, dass ein bestimmter Tropfen den Schirm trifft, ist sehr klein.
- Anzahl der Kernzerfälle in einem radioaktiven Uranpräparat in einer Sekunde:
  - Im Präparat sind viele radioaktive Atomkerne, die alle zerfallen könnten.
  - Die Wahrscheinlichkeit, dass ein vorgegebener Atomkern in einer Sekunde tatsächlich zerfällt, ist sehr klein.
- Anzahl der Haftpflichtfälle bei einer Versicherung in einem Monat
  - Die Versicherung hat viele Kunden.

– Die Wahrscheinlichkeit, dass ein bestimmter Kunde tatsächlich ein Haftpflichtereignis hat, ist sehr klein.

Modell: Wir modellieren die Verteilung der Anzahl der "seltenen Ereignisse" mit Binomialverteilungen mit Parametern n und p in der Asymptotik $n \to \infty$ und $p \to 0$ mit der Nebenbedingung $np = \lambda$ mit einer Konstanten $\lambda > 0$.
Es bezeichne

$$\text{Poisson}(\lambda) = \sum_{k \in \mathbb{N}_0} e^{-\lambda} \frac{\lambda^k}{k!} \tag{355}$$

die Poissonverteilung zum Parameter $\lambda > 0$.

Satz 2.67 (Poisson-Limes der Binomialverteilung) *Es sei $(p_n)_{n\in\mathbb{N}}$ eine Folge mit Werten in $]0,1[$ mit*

$$np_n \overset{n\to\infty}{\longrightarrow} \lambda > 0. \tag{356}$$

Dann gilt für alle $k \in \mathbb{N}_0$:

$$\text{binomial}(n, p_n)(\{k\}) = \binom{n}{k} p_n^k (1-p_n)^{n-k} \overset{n\to\infty}{\longrightarrow} e^{-\lambda} \frac{\lambda^k}{k!} = \text{Poisson}(\lambda)(\{k\}). \tag{357}$$

Beweis: Wir rechnen[21]

$$\begin{aligned}
&\text{binomial}(n, p_n)(\{k\}) = \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \underbrace{\left(\prod_{l=0}^{k-1} \frac{n-l}{n} \right)}_{\overset{n\to\infty}{\longrightarrow} 1} \cdot \frac{1}{k!} \cdot \underbrace{(np_n)^k}_{\overset{n\to\infty}{\longrightarrow} \lambda^k} \cdot \exp\Big[\underbrace{\frac{n-k}{n}}_{\overset{n\to\infty}{\longrightarrow} 1} \cdot \underbrace{n \log(1-p_n)}_{\substack{= -np_n(1+o(1)) \to -\lambda \\ \text{für } n\to\infty}} \Big] \\
&\overset{n\to\infty}{\longrightarrow} \frac{1}{k!} \lambda^k e^{-\lambda} = \text{Poisson}(\lambda)(\{k\}).
\end{aligned} \tag{358}$$

□

Korollar 2.68 (Poisson-Limes der Binomialverteilung – Version für Ereignisse) *Für alle Ereignisse $A \subseteq \mathbb{N}_0$ gilt unter den Voraussetzungen des Satzes 2.67:*

$$\text{binomial}(n, p_n)(A) = \sum_{k \in A} \binom{n}{k} p_n^k (1-p_n)^{n-k} \overset{n\to\infty}{\longrightarrow} \sum_{k \in A} e^{-\lambda} \frac{\lambda^k}{k!} = \text{Poisson}(\lambda)(A). \tag{359}$$

Der **Beweis** des Korollars verwendet das Lemma von Fatou aus der Maßtheorie. Es lautet:

[21] *Zur Notation:* Das Landausymbol $o(f(n))$ steht für $n \to \infty$ für irgendein $g(n)$ mit $g(n)/f(n) \overset{n\to\infty}{\longrightarrow} 0$. Zum Beispiel steht $o(1)$ für irgendeine Nullfolge.

Lemma 2.69 (Fatou) *Ist $(X_n)_{n\in\mathbb{N}}$ eine Folge messbarer Funktionen über einem Maßraum $(\Omega, \mathcal{A}, \mu)$ mit $X_n \geq 0$ für alle $n \in \mathbb{N}$, so gilt für alle messbaren Mengen $A \in \mathcal{A}$:*

$$\int_A \liminf_{n\to\infty} X_n \, d\mu \leq \liminf_{n\to\infty} \int_A X_n \, d\mu. \tag{360}$$

Gegeben $A \subseteq \mathbb{N}_0$, wenden wir das im Fall $\Omega = \mathbb{N}_0$ mit dem Zählmaß μ und

$$X_n(k) = \text{binomial}(n, p_n)(\{k\}) = \binom{n}{k} p_n^k (1-p_n)^{n-k} \tag{361}$$

an, verwenden also die "Reihen-Version" des Lemmas von Fatou. Wir erhalten:

$$\begin{aligned}
&\text{Poisson}(\lambda)(A) = \sum_{k\in A} e^{-\lambda} \frac{\lambda^k}{k!} \\
&= \sum_{k\in A} \lim_{n\to\infty} \text{binomial}(n, p_n)(\{k\}) \\
&\leq \liminf_{n\to\infty} \sum_{k\in A} \text{binomial}(n, p_n)(\{k\}) \\
&\leq \liminf_{n\to\infty} \text{binomial}(n, p_n)(A)
\end{aligned} \tag{362}$$

und ebenso

$$\text{Poisson}(\lambda)(A^c) \leq \liminf_{n\to\infty} \text{binomial}(n, p_n)(A^c). \tag{363}$$

Weil sowohl die Poissonverteilung als auch die Binomialverteilung Wahrscheinlichkeitsmaße sind, können wir schließen:

$$\begin{aligned}
&\text{Poisson}(\lambda)(A) = 1 - \text{Poisson}(\lambda)(A^c) \\
&\geq 1 - \liminf_{n\to\infty} \text{binomial}(n, p_n)(A^c) = \limsup_{n\to\infty} (1 - \text{binomial}(n, p_n)(A^c)) \\
&= \limsup_{n\to\infty} \text{binomial}(n, p_n)(A).
\end{aligned} \tag{364}$$

Die Ungleichungen (364) und (364) zusammen zeigen die Behauptung (359).

□

2.11.4 Ordnungsstatistik und Betaverteilung

Definition 2.70 (Ordnungsstatistik) *Für $x_1, x_2, \ldots, x_n \in \mathbb{R}$ sei $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ die monoton aufsteigende Permutation von $x_1, x_2, \ldots, x_n$:*

$$x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}. \tag{365}$$

Das Tupel $(x_{[1]}, x_{[2]}, \ldots, x_{[n]})$ wird Ordnungsstatistik *von $(x_1, x_2, \ldots, x_n)$ genannt.*

Die $x_{[1]}, \ldots, x_{[n]}$ sind also die $x_1, \ldots, x_n$ in aufsteigender Reihenfolge angeordnet.

Definition 2.71 (Betaverteilung) *Die Betaverteilung* $\mathrm{Beta}(a, b)$ *mit den Parametern* $a > 0$ *und* $b > 0$ *ist die Verteilung auf* $\mathbb{R}$ *mit der Dichte*

$$\beta_{a,b}(x) = 1_{]0,1[}(x) \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \tag{366}$$

mit der Betafunktion als Normierungskonstanten:

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{367}$$

Weil $\Gamma(n+1) = n!$ für alle $n \in \mathbb{N}_0$, können wir das für $k, n \in \mathbb{N}$ mit $k \leq n$ auch schreiben:

$$\frac{1}{B(k, n-k+1)} = \frac{n!}{(k-1)!(n-k)!} = k\binom{n}{k}. \tag{368}$$

Wir zeigen:

Satz 2.72 (Verteilung von Komponenten der Ordnungsstatistik) *Gegeben* $n \in \mathbb{N}$, *seien* $U_1, \ldots, U_n$ *i.i.d.* $\mathrm{uniform}(]0,1[)$*-verteilte Zufallsvariablen über einem Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, P)$ *und* $U_{[1]}, \ldots, U_{[n]}$ *ihre Ordnungsstatistik, also* $U_{[k]}(\omega)$ *der* k*-t kleinste unter den* $U_1(\omega), \ldots, U_n(\omega)$*, wobei* $\omega \in \Omega$*. Dann ist für alle* $k \in [n]$ *die* k*-te Komponente* $U_{[k]}$ *der Ordnungsstatistik betaverteilt mit den Parametern* k *und* $n - k + 1$*.*

Beweis: Gegeben $a \in [0,1]$, beweisen wir für alle $k \in [n]$ die Behauptung

$$P[U_{[k]} \leq a] = \int_0^a \beta_{k,n-k+1}(x)\, dx = \mathrm{Beta}(k, n-k+1)([0,a]), \tag{369}$$

und zwar durch Induktion über k "rückwärts":
Induktionsanfang, $k = n$*:* Es gilt:

$$\begin{aligned}
&P[U_{[n]} \leq a] = P[\forall i \in [n] :\ U_i \leq a] \\
&= \prod i = 1^n P[U_i \leq a] = a^n \quad \text{(da } U_i \text{ i.i.d. uniform]0,1[-vert.)} \\
&= \int_0^a n x^{n-1}\, dx = \int_0^a n\binom{n}{n} x^{n-1}(1-x)^{n-n}\, dx \\
&= \int_0^a \beta_{n,1}(x)\, dx = \mathrm{Beta}(n,1)([0,a]).
\end{aligned} \tag{370}$$

Induktionsvoraussetzung für $k+1$*:* Gegeben $k \in [n]$ mit $k \geq 2$, nehmen wir an:

$$P[U_{[k+1]} \leq a] = \mathrm{Beta}(k+1, n-k)([0,a]). \tag{371}$$

Induktionsschluss $k+1 \rightsquigarrow k$*:* Wir rechnen:

$$\begin{aligned} P[U_{[k]} \le a] &= P[U_{[k+1]} \le a] + P[U_{[k]} \le a < U_{[k+1]}] \quad \text{(wegen } U_{[k]} \le U_{[k+1]}\text{)} \\ &= \int_0^a \beta_{k+1,n-k}(x)\,dx + P[U_{[k]} \le a < U_{[k+1]}] \quad \text{(mit der I.V. (371))} \end{aligned} \tag{372}$$

Wir berechnen den zweiten Summanden in der letzten Summe:

$$\begin{aligned} &P[U_{[k]} \le a < U_{[k+1]}] = P[|\{i \in [n]|\ U_i \le a\}| = k] \\ &= \sum_{\substack{E\subseteq[n]\\|E|=k}} P[\forall i \in E:\ U_i \le a,\ \forall i \in [n]\setminus E:\ U_i > a] \\ &= \sum_{\substack{E\subseteq[n]\\|E|=k}} \prod_{i\in E} P[U_i \le a] \cdot \prod_{i\in[n]\setminus E} P[U_i > a] = \binom{n}{k} a^k (1-a)^{n-k} \quad \text{(da } (U_i) \text{ i.i.d. uniform}]0,1[) \\ &= \int_0^a \frac{d}{dx}\left[\binom{n}{k} x^k (1-x)^{n-k}\right] dx \quad \text{(da } \tbinom{n}{k} x^k (1-x)^{n-k} = 0 \text{ für } x = 0) \\ &= \int_0^a \underbrace{k\binom{n}{k}}_{=B(k,n-k+1)^{-1}} x^{k-1}(1-x)^{n-k}\,dx - \int_0^a \underbrace{(n-k)\binom{n}{k}}_{\substack{=(k+1)\binom{n}{k+1}\\=B(k+1,n-k)^{-1}}} x^k (1-x)^{n-k-1}\,dx \\ &= \int_0^a \beta_{k,n-k+1}(x)\,dx - \int_0^a \beta_{k+1,n-k}(x)\,dx \end{aligned} \tag{373}$$

Eingesetzt in (372) folgt die Induktionsbehauptung:

$$P[U_{[k]} \le a] = \int_0^a \beta_{k,n-k+1}(x)\,dx = \mathrm{Beta}(k, n-k+1)([0,a]). \tag{374}$$

Weil die Verteilungsfunktion die Verteilung eindeutig bestimmt, bedeutet das:

$$\mathcal{L}_P(U_{[k]}) = \mathrm{Beta}(k, n-k+1)([0,a]), \tag{375}$$

wie zu zeigen war.

□

2.12 Erwartungswert und Varianz

Bis jetzt haben wir nur Integrale nichtnegativer Funktionen im Zusammenhang mit Dichten betrachtet. Die Verallgemeinerung des Integralbegriffs auf Funktionen beliebigen Vorzeichens erfolgt so:

Definition 2.73 (Integral von Funktionen mit beliebigem Vorzeichen, Integrierbarkeit) *Es sei* $(\Omega, \mathcal{A}, \mu)$ *ein Maßraum und*

$$X : (\Omega, \mathcal{A}) \to (\mathbb{R} \cup \{\pm\infty\}, \mathcal{B}(\mathbb{R} \cup \{\pm\infty\})) \tag{376}$$

eine messbare Abbildung.[22] *Der* Positivteil *von X wird durch*

$$X_+ := \max\{X, 0\} \tag{377}$$

definiert, der Negativteil *durch*

$$X_- := \max\{-X, 0\}. \tag{378}$$

Im Fall, dass $\int_\Omega X_+ \, d\mu$ oder(!) $\int_\Omega X_- \, d\mu$ endlich sind, ist die Differenz

$$\int_\Omega X \, d\mu := \int_\Omega X_+ \, d\mu - \int_\Omega X_- \, d\mu \in \mathbb{R} \cup \{\pm\infty\} \tag{379}$$

definiert. Wir nennen sie das Integral *von X. Sind sogar $\int_\Omega X_+ \, d\mu$ und(!) $\int_\Omega X_- \, d\mu$ endlich, so ist das Integral $\int_\Omega X \, d\mu \in \mathbb{R}$ endlich; in diesem Fall nennen wir X* integrierbar *bezüglich μ. Wir setzen*

$$\mathcal{L}^1(\Omega, \mathcal{A}, \mu) := \{X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ } \textit{messbar}| \text{ } X \text{ } \textit{ist } \mu\textit{-integrierbar}\} \tag{380}$$

Die Integralabbildung

$$\int : \mathcal{L}^1(\Omega, \mathcal{A}, \mu) \to \mathbb{R}, \quad X \mapsto \int_\Omega X \, d\mu \tag{381}$$

ist linear.[23] Im Fall, dass der Integrator $\mu = P$ ein Wahrscheinlichkeitsmaß ist, verwendet man in der Stochastik meist eine andere Sprechweise für das Integral:

Definition 2.74 (Erwartungswert) *Es sei*

$$X : (\Omega, \mathcal{A}) \to (\mathbb{R} \cup \{\pm\infty\}, \mathcal{B}(\mathbb{R} \cup \{\pm\infty\})) \tag{384}$$

eine Zufallsvariable über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Falls das Integral $\int_\Omega X \, dP$ existiert, nennen wir

$$\boxed{E_P[X] := \int_\Omega X \, dP} \tag{385}$$

den Erwartungswert *(synonym:* Erwartung*, engl.* expectation *oder* expected value*) von X bezüglich P.* **Wichtig!**

[22] Eigentlich genügt es, wenn X nur μ-fast überall definiert ist. Wir führen das nicht genauer aus.

[23] Das Integral $\int_\Omega X \, d\mu$ ist jedoch nicht nur linear im Integranden X, sondern auch im Integrator μ. Genauer gesagt gilt für zwei Maße μ und ν auf $(\Omega, \mathcal{A})$ und alle $a, b \geq 0$:

$$\mathcal{L}^1(\Omega, \mathcal{A}, a\mu + b\nu) \subseteq \mathcal{L}^1(\Omega, \mathcal{A}, \mu) \cap \mathcal{L}^1(\Omega, \mathcal{A}, \nu), \tag{382}$$

und für alle $X \in \mathcal{L}^1(\Omega, \mathcal{A}, a\mu + b\nu)$ oder auch für alle messbaren $X \geq 0$ gilt:

$$\int_\Omega X \, d(a\mu + b\nu) = a \int_\Omega X \, d\mu + b \int_\Omega X \, d\nu. \tag{383}$$

Notationen: Falls klar ist, welches Wahrscheinlichkeitsmaß P gemeint ist, schreiben wir auch $E[X]$ statt $E_P[X]$. Ist $A \in \mathcal{A}$ ein Ereignis, so heißt

$$E_P[X, A] := E[X, A] := E_P[X \cdot 1_A] = \int_A X \, dP \tag{386}$$

die Erwartung von X auf dem Ereignis A.[24] Hier ist eine Liste von Eigenschaften des Erwartungswerts, die man bei der Arbeit damit sehr oft verwendet:

Lemma 2.75 (Grundlegende Eigenschaften des Erwartungswerts)

1. **Linearität:**
   - *Für alle* $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ *gilt:* $E_P[X + Y] = E_P[X] + E_P[Y]$.
   - *Für alle* $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ *und* $\alpha \in \mathbb{R}$ *gilt:* $E_P[\alpha X] = \alpha E_P[X]$.
2. **Monotonie:** *Für alle Zufallsvariablen* X, Y *auf* $(\Omega, \mathcal{A}, P)$ *mit existierender Erwartung gilt die Implikation*

$$X \leq Y \quad \Longrightarrow \quad E_P[X] \leq E_P[Y]. \tag{387}$$

3. **Erwartung von Konstanten:** *Für konstante Zufallsvariablen mit einem Wert* $c \in \mathbb{R}$ *gilt:*

$$E_P[c] = c. \tag{388}$$

4. **Erwartung von Indikatorfunktionen:** *Für jedes Ereignis* $A \in \mathcal{A}$ *gilt:* $E_P[1_A] = P(A)$.

Diese Eigenschaften folgen unmittelbar aus den entsprechenden Eigenschaften des Integrals.

Beispiele:

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

$$X \in \mathcal{L}^1(\Omega, \mathcal{A}, P) \quad \Leftrightarrow \quad \sum_{\omega \in \Omega} |X(\omega)| p_\omega < \infty, \tag{389}$$

   und falls X einen Erwartungswert besitzt, wird dieser durch

$$E_P[X] = \sum_{\omega \in \Omega} X(\omega) p_\omega \tag{390}$$

   gegeben.

[24] Ein häufiger Notationsfehler von Anfängern ist es, die Typen von Wahrscheinlichkeiten und Erwartungen zu verwechseln. Beachten Sie daher: Die Erwartung $E[X]$ ist höchstens für *Zufallsvariablen* X definiert, die Wahrscheinlichkeit $P(A)$ dagegen nur für *Ereignisse* A. Bitte nie verwechseln, vor allem nicht in Klausuren!

2. Für das Modell

$$(\Omega, \mathcal{A}, P) = ([6], \mathcal{P}([6]), \text{Gleichverteilung}) \tag{391}$$

des fairen Würfelns und

$$X = \mathrm{id} : \Omega \to \mathbb{R}, \quad X(\omega) = \omega \tag{392}$$

gilt:

$$E_P[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3,5. \tag{393}$$

Dieses Beispiel illustriert, dass der Erwartungswert kein möglicher Wert der Zufallsvariable zu sein braucht.

Der folgende Satz zeigt unter anderem, dass der Erwartungswert einer Zufallsvariablen X nur von der Verteilung von X abhängt. Deshalb spricht man manchmal auch von der Erwartung einer Verteilung statt von der Erwartung einer Zufallsvariablen.

Satz 2.76 (Integral bezüglich des Bildmaßes) *Es seien $(\Omega, \mathcal{A}, \mu)$ ein Maßraum. Weiter seien $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ und $f : (\Omega', \mathcal{A}') \to (\mathbb{R} \cup \{\pm\infty\}, \mathcal{B}(\mathbb{R} \cup \{\pm\infty\}))$ messbare Abbildungen. Dann existiert das Integral $\int_\Omega f \circ X \, d\mu$ genau dann, wenn das Integral $\int_{\Omega'} f \, d(X[\mu])$ existiert. In diesem Fall gilt:*

$$\boxed{\int_\Omega f \circ X \, d\mu = \int_{\Omega'} f \, d(X[\mu])} \tag{394}$$

Der Beweis wird in der Maßtheorie erst für Treppenfunktionen f, dann für messbare $f \geq 0$ und schließlich für allgemeine f geführt; wir verzichten hier darauf.

Im Fall von Wahrscheinlichkeitsmaßen $\mu = P$ können wir die Formel (394) auch so schreiben:

$$\boxed{Q = \mathcal{L}_P(X) \quad \Longrightarrow \quad E_P[f(X)] = \int_{\Omega'} f(x) \, Q(dx)} \tag{395}$$

Im Fall, dass die Verteilung Q von $X : \Omega \to \Omega' = \mathbb{R}^n$ eine Dichte $g : \mathbb{R}^n \to [0, \infty]$ besitzt, bedeutet das:

$$\boxed{E_P[f(X)] = \textstyle\int_{\mathbb{R}^n} f(x) g(x) \, dx} \tag{396}$$

wobei hier "dx" kurz für "$\lambda_n(dx)$" steht. Wir verwenden hierbei das folgende Lemma aus der Maßtheorie:

Lemma 2.77 (Integral bezüglich eines Maßes mit Dichte) *Besitzt ein Maß μ auf $(\Omega, \mathcal{A})$ bezüglich eines Maßes ν eine Dichte g, so gilt für alle messbaren Abbildungen $f : \Omega \to \mathbb{R} \cup \{\pm\infty\}$:*

$$\int_\Omega f \, d\mu = \int_\Omega f g \, d\nu, \tag{397}$$

wobei hierin die linke Seite genau dann definiert ist, wenn auch die rechte Seite definiert ist.[25]

Beispiele:

1. Ist X eine normalverteilte Zufallsvariable mit den Parametern μ und σ, so gilt

$$\begin{aligned} E[X] &= \int_{\mathbb{R}} x \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} (t+\mu) e^{-\frac{t^2}{2\sigma^2}} \, dt \quad \text{(mit der Substitution } t = x - \mu) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \underbrace{\int_{\mathbb{R}} t e^{-\frac{t^2}{2\sigma^2}} \, dt}_{=0 \text{ wg. Symmetrie}} + \mu \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-\frac{t^2}{2\sigma^2}} \, dt}_{=1} \\ &= \mu. \end{aligned} \tag{400}$$

   Der Parameter μ der Normalverteilung ist also in der Tat der Erwartungswert von X.

2. Nimmt eine Zufallsvariable X auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ nur endlich viele verschiedene Werte $x_1, \ldots, x_n \in \mathbb{R}$ an, besitzt sie also die Verteilung

$$Q := \mathcal{L}_P(X) = \sum_{k=1}^{n} P[X = x_k] \delta_{x_k}, \tag{401}$$

   so erhalten wir die Erwartung

$$\boxed{E_P[X] = \sum_{k=1}^{n} x_k P[X = x_k]} \tag{402}$$

[25] **Notation:** Im Hinblick auf dieses Lemma schreibt man für die Dichte auch symbolisch

$$g = \frac{d\mu}{d\nu} \quad \nu\text{-f.ü.} \tag{398}$$

und nennt $d\mu/d\nu$ auch die *Radon-Nikodym-Ableitung* von μ nach ν. Die Gleichung (397) bekommt damit die intuitive Form

$$\int_\Omega f \, d\mu = \int_\Omega f \frac{d\mu}{d\nu} \, d\nu. \tag{399}$$

Etwas allgemeiner erhalten für eine reellwertige Zufallsvariable $f(X)$ die folgende Erwartung, wenn X nur endlich viele Werte $x_1, \ldots, x_n$ annimmt:

$$\boxed{E_P[f(X)] = \sum_{k=1}^{n} f(x_k) P[X = x_k]} \tag{403}$$

Analoges gilt, wenn X abzählbar viele verschiedene Werte x_k, $k \in \mathbb{N}$, annimmt:

$$\boxed{E_P[f(X)] = \sum_{k \in \mathbb{N}} f(x_k) P[X = x_k]} \tag{404}$$

falls die Reihe absolut konvergiert.

3. Manchmal hat die Verteilung einer Zufallsvariablen sowohl einen Anteil mit Dichte, "absolutstetiger Teil" genannt, also auch einen Teil, der auf einer Lebesgue-Nullmenge lebt, "singulärer Teil" genannt. Hierzu ein Beispiel: Ist

$$(\Omega, \mathcal{A}, P) = ([0,1], \mathcal{B}([0,1]), \text{uniform}([0,1])) \tag{405}$$

und

$$X : \Omega \to \mathbb{R}, \quad X(\omega) = \tfrac{1}{3} \vee (\omega \wedge \tfrac{2}{3}) = \max\left\{\tfrac{1}{3}, \min\left\{\omega, \tfrac{2}{3}\right\}\right\}, \tag{406}$$

so gilt

$$\mathcal{L}_P(X) = \underbrace{\frac{1}{3} \text{uniform}\left(\left[\tfrac{1}{3}, \tfrac{2}{3}\right]\right)}_{\text{absolutstetig}} + \underbrace{\frac{1}{3}(\delta_{\frac{1}{3}} + \delta_{\frac{2}{3}})}_{\text{singulär}}. \tag{407}$$

Hier gilt für jede Borel-messbare beschränkte Funktion $f : \mathbb{R} \to \mathbb{R}$:

$$E_P[f(X)] = \int_{\mathbb{R}} f \, d\mathcal{L}_P(X) = \int_{1/3}^{2/3} f(x)\, dx + \frac{1}{3}(f(\tfrac{1}{3}) + f(\tfrac{2}{3})), \tag{408}$$

weil die uniforme Verteilung auf $[1/3, 2/3]$ die Dichte $3 \cdot 1_{[1/3,2/3]}$ besitzt. Natürlich kann man ebensogut die gleiche Erwartung auch so berechnen:

$$\begin{aligned} E_P[f(X)] &= \int_0^1 f(X(\omega))\, d\omega \\ &= \int_0^{1/3} f\left(\frac{1}{3}\right)\, d\omega + \int_{1/3}^{2/3} f\left(\omega\right)\, d\omega + \int_{2/3}^{1} f\left(\frac{2}{3}\right)\, d\omega \\ &= \int_{1/3}^{2/3} f(x)\, dx + \frac{1}{3}(f(\tfrac{1}{3}) + f(\tfrac{2}{3})). \end{aligned} \tag{409}$$

Die Erwartung einer Zufallsvariablen ist ein Lageparameter ihrer Verteilung. Dagegen misst die Varianz einer Zufallsvariablen, die wir nun einführen, die Fluktuationen der Werte der Zufallsvariablen um ihre Erwartung:

Definition 2.78 (Varianz, Standardabweichung) *Es sei $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$. Wir definieren die* Varianz *von X bezüglich P durch* **Wichtig!**

$$\boxed{\mathrm{Var}(X) = \mathrm{Var}_P(X) := E_P[(X - E_P[X])^2] \in [0, \infty]} \tag{410}$$

Ihre Quadratwurzel[26] *heißt* Standardabweichung *von X bezüglich P:*

$$\sigma_P(X) := \sqrt{\mathrm{Var}_P(X)} \tag{411}$$

Den Raum der Zufallsvariablen mit endlicher Varianz bezeichnet man so:

$$\begin{aligned} \mathcal{L}^2(\Omega, \mathcal{A}, P) &:= \{X \in \mathcal{L}^1(\Omega, \mathcal{A}, P) |\ \mathrm{Var}_P(X) < \infty\} \\ &= \{X \in \mathcal{L}^1(\Omega, \mathcal{A}, P) |\ E_P[X^2] < \infty\}. \end{aligned} \tag{412}$$

Die letzte Gleichheit wird durch die folgende Formel gerechtfertigt:

Satz 2.79 (Eine Berechnungsformel für die Varianz) *Für alle $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ gilt:* **Wichtig!**

$$\boxed{\mathrm{Var}_P(X) := E_P[X^2] - E_P[X]^2} \tag{413}$$

Beweis:

$$\begin{aligned} \mathrm{Var}_P(X) &= E_P[(X - E_P[X])^2] = E_P[X^2 - 2E_P[X]X + E_P[X]^2] \\ &= E_P[X^2] - 2E_P[X]E_P[X] + E_P[X]^2 \quad \text{(wg. Linearität von } E_P \text{ und } E_P[1] = 1) \\ &= E_P[X^2] - E_P[X]^2. \end{aligned} \tag{414}$$

□

Übung 2.80 (Eigenschaften der Varianz) *Zeigen Sie für $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ und $a \in \mathbb{R}$:*[27]

$$\mathrm{Var}_P(X) \geq 0, \tag{415}$$

$$\mathrm{Var}_P(aX) = a^2 \,\mathrm{Var}_P(X), \tag{416}$$

$$\sigma_P(aX) = |a|\sigma_P(X), \tag{417}$$

$$\mathrm{Var}_P(X + a) = \mathrm{Var}_P(X). \tag{418}$$

[26] Konvention: $\sqrt{\infty} := \infty$

[27] *Warnung vor einem Standardrechenfehler:* Vergisst man z.B. das letzte Quadrat in Formel (413), so kann es vorkommen, dass man irrtümlich ein negatives Ergebnis für eine Varianz bekommt. Diesen Fehler sollte man ebenso sofort bemerken, wie man es bemerken sollte, wenn eine Wahrscheinlichkeit irrtümlich als negativ oder größer als 1 berechnet wird.

Beispiele:

1. Die Varianz einer $p\delta_1 + (1-p)\delta_0$-verteilten Zufallsvariablen X mit Werten 0 oder 1 beträgt

$$\begin{aligned} \mathrm{Var}_P(X) &= E_P[X^2] - E_P[X]^2 = E_P[X] - E_P[X]^2 \quad \text{(wegen } X^2 = X\text{)} \\ & E_P[X](1 - E_P[X]) = p(1-p), \end{aligned} \tag{419}$$

   denn

$$E[X] = 1 \cdot \underbrace{P[X=1]}_{=p} + 0 \cdot P[X=0] = p. \tag{420}$$

2. Wir berechnen die Erwartung und die Varianz der Gleichverteilung auf dem Einheitsintervall. Es sei also X eine Zufallsvariable mit $\mathcal{L}_P(X) = \mathrm{uniform}[0,1]$. Dann folgt:

$$E_P[X] = \int_{\mathbb{R}} t 1_{[0,1]}(t)\, dt = \int_0^1 t\, dt = \frac{1}{2}, \tag{421}$$

$$E_P[X^2] = \int_{\mathbb{R}} t^2 1_{[0,1]}(t)\, dt = \int_0^1 t^2\, dt = \frac{1}{3} \tag{422}$$

   und daher

$$\mathrm{Var}_P(X) = E_P[X^2] - E_P[X]^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}. \tag{423}$$

   Ist nun $[a,b]$ irgendein Intervall ($a < b$ reelle Zahlen),so ist $(b-a)X + a$ uniform auf $[a,b]$ verteilt. Für die Erwartung und die Varianz dieser Verteilung folgt mit Skalierung:

$$E_P[(b-a)X + a] = (b-a)E_P[X] + a = (b-a) \cdot \frac{1}{2} + a = \frac{1}{2}(a+b), \tag{424}$$

$$\mathrm{Var}_P((b-a)X + a) = (b-a)^2 \,\mathrm{Var}_P(X) = \frac{1}{12}(b-a)^2. \tag{425}$$

   Durch Wurzelziehen erhalten wir die Standardabweichung der Gleichverteilung auf $[a,b]$:

$$\sigma_P((b-a)X + a) = \frac{1}{\sqrt{12}}(b-a). \tag{426}$$

3. Wir berechnen nun die Varianz normalverteilter Zufallsvariablen. Hierzu sei Z eine

standardnormalverteilte Zufallsvariable. Dann folgt wegen $E_P[Z] = 0$:

$$\begin{aligned}
\mathrm{Var}_P(Z) = E_P[Z^2] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} z^2 e^{-\frac{z^2}{2}}\, dz \\
&= -\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} z \frac{d}{dz} e^{-\frac{z^2}{2}}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \left[\underbrace{\left[-z e^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty}}_{=0} + \underbrace{\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz}_{=\sqrt{2\pi}} \right] \qquad \text{(mit partieller Integration)} \\
&= 1. \qquad (427)
\end{aligned}$$

Gegeben $\mu \in \mathbb{R}$ und $\sigma > 0$, ist die Zufallsvariable $X := \sigma Z + \mu$ normalverteilt: $\mathcal{L}_P(X) = \mathrm{N}(\mu, \sigma^2)$. Für ihre Varianz folgt mit Skalierung:

$$\mathrm{Var}_P(X) = \mathrm{Var}_P(\sigma Z + \mu) = \sigma^2 \,\mathrm{Var}_P(Z) = \sigma^2. \qquad (428)$$

Der Parameter σ^2 der Normalverteilung trägt also zu Recht den Namen "Varianz".

Die Varianz ist eine quadratische Form auf $\mathcal{L}^2(\Omega, \mathcal{A}, P)$. Die zugehörige symmetrische Bilinearform heißt *Covarianz*.

Definition 2.81 (Covarianz) *Es seien $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$. Die Covarianz von X und Y bezüglich P wird durch*

$$\boxed{\mathrm{Cov}_P(X, Y) := E_P[(X - E_P[X])(Y - E_P[Y])]} \qquad (429)$$

definiert.

Bemerkungen:

1. Insbesondere gilt

$$\mathrm{Cov}_P(X, X) = \mathrm{Var}_P(X). \qquad (430)$$

2. Die Covarianz ist bilinear und symmetrisch, d.h. für alle $X, Y, Z \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ und alle $\alpha \in \mathbb{R}$ gilt:

$$\begin{aligned}
\mathrm{Cov}_P(X, Y) &= \mathrm{Cov}_P(Y, X), && (431)\\
\mathrm{Cov}_P(X + Y, Z) &= \mathrm{Cov}_P(X, Z) + \mathrm{Cov}_P(Y, Z), && (432)\\
\mathrm{Cov}_P(\alpha X, Y) &= \alpha \,\mathrm{Cov}_P(X, Y). && (433)
\end{aligned}$$

3. Analog zur Formel (413) für die Varianz gilt für alle $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$:

$$\boxed{\mathrm{Cov}_P(X, Y) = E_P[XY] - E_P[X] E_P[Y]} \qquad (434)$$

Übung 2.82 *Beweisen Sie die Formel* (434).

Lemma 2.83 (Cauchy-Schwarz-Ungleichung für die Covarianz) *Für alle* $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ *gilt:*

$$\boxed{|\operatorname{Cov}_P(X,Y)| \leq \sigma_P(X)\sigma_P(Y)} \tag{435}$$

Dies ist ein Spezialfall der folgenden Cauchy-Schwarz-Ungleichung:

Lemma 2.84 (L^2-Version der Cauchy-Schwarz-Ungleichung) *Es seien* $X, Y : \Omega \to \mathbb{R}$ *messbare Funktionen auf einem Maßraum* $(\Omega, \mathcal{A}, \mu)$ *mit*

$$\int_\Omega X^2 \, d\mu < \infty \text{ und } \int_\Omega Y^2 \, d\mu < \infty. \tag{436}$$

Dann ist XY *bezüglich* μ *integrierbar, und es gilt*

$$\boxed{\left(\int_\Omega XY \, d\mu\right)^2 \leq \int_\Omega X^2 \, d\mu \cdot \int_\Omega Y^2 \, d\mu.} \tag{437}$$

Dasselbe in anderer Notation für Wahrscheinlichkeitsmaße P*:*

$$\boxed{E_P[XY]^2 \leq E_P[X^2]E_P[Y^2]} \tag{438}$$

Setzen wir hierin $X - E_P[X]$ statt X und $Y - E_P[Y]$ statt Y, erhalten wir die Cauchy-Schwarz-Ungleichung (435) für die Covarianz.

Beweis des Lemmas 2.84: Wir betrachten die quadratische Form

$$q : \mathbb{R}^2 \to [0, \infty[, \tag{439}$$

$$q(\alpha, \beta) = \int_\Omega (\alpha X + \beta Y)^2 \, d\mu = \alpha^2 \int_\Omega X^2 \, d\mu + 2\alpha\beta \int_\Omega XY \, d\mu + \beta^2 \int_\Omega Y^2 \, d\mu. \tag{440}$$

Zur Wohldefiniertheit von q: Wegen

$$2\alpha^2 X^2 + 2\beta^2 Y^2 - (\alpha X + \beta Y)^2 = (\alpha X - \beta Y)^2 \geq 0 \tag{441}$$

gilt

$$(\alpha X + \beta Y)^2 \leq 2\alpha^2 X^2 + 2\beta^2 Y^2 \tag{442}$$

und daher

$$q(\alpha, \beta) \leq 2\alpha^2 \int_\Omega X^2 \, d\mu + 2\beta^2 \int_\Omega Y^2 \, d\mu < \infty. \tag{443}$$

Setzen wir speziell

$$\alpha = \sqrt{\int_\Omega Y^2 \, d\mu} \quad \text{und } \beta = \pm\sqrt{\int_\Omega X^2 \, d\mu} \tag{444}$$

in die quadratische Form q ein, so folgt:

$$0 \leq q(\alpha, \beta) = 2\int_\Omega X^2 \, d\mu \int_\Omega Y^2 \, d\mu \pm 2\sqrt{\int_\Omega X^2 \, d\mu}\sqrt{\int_\Omega Y^2 \, d\mu} \int_\Omega XY \, d\mu. \tag{445}$$

- Im Fall $\alpha\beta \neq 0$ folgt hieraus 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 die Behauptung folgt auch hier.
- Im Fall $\beta = 0$ folgt analog $X^2 = 0$ μ-fast sicher, also $X = 0$ μ-fast sicher, also $XY = 0$ μ-fast sicher, und die Behauptung folgt ebenso.

□

Bemerkung: Gleichheit in der Cauchy-Schwarz-Ungleichung gilt nur im Fall $\alpha X = \beta Y$ μ-fast sicher, also falls X und Y bis auf Abänderung auf Nullmengen linear abhängig voneinander sind.

Definition 2.85 (Korrelation) *Es seien $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ zwei Zufallsvariablen mit positiver Standardabweichung. Wir definieren den* Korrelationskoeffizienten *(synonym: die* Korrelation*) zwischen X und Y:*

$$\boxed{\mathrm{r}_P(X, Y) = \mathrm{Corr}_P(X, Y) := \frac{\mathrm{Cov}_P(X, Y)}{\sigma_P(X)\sigma_P(Y)}} \tag{446}$$

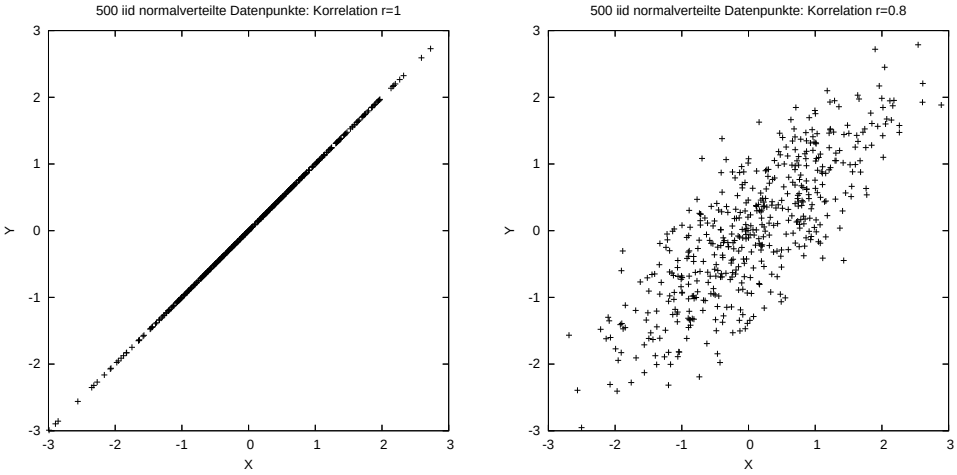
Es gilt also:

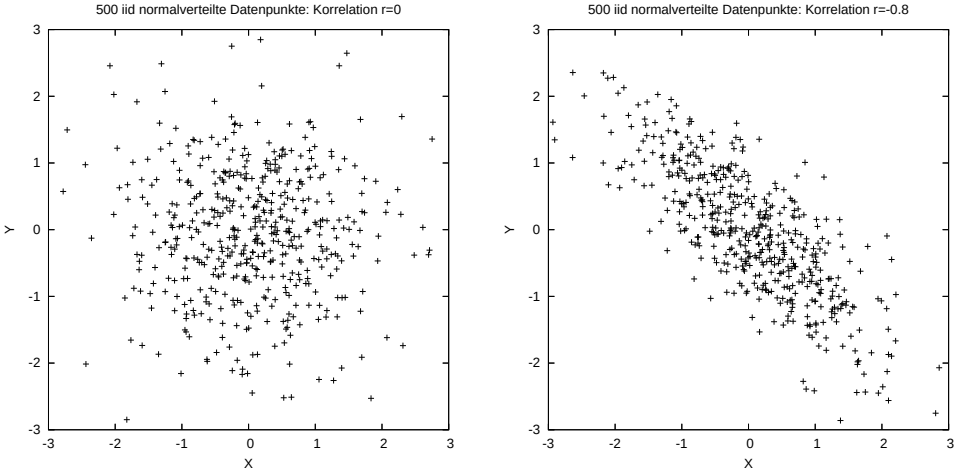
$$\boxed{-1 \leq \mathrm{r}_P(X, Y) \leq 1} \tag{447}$$

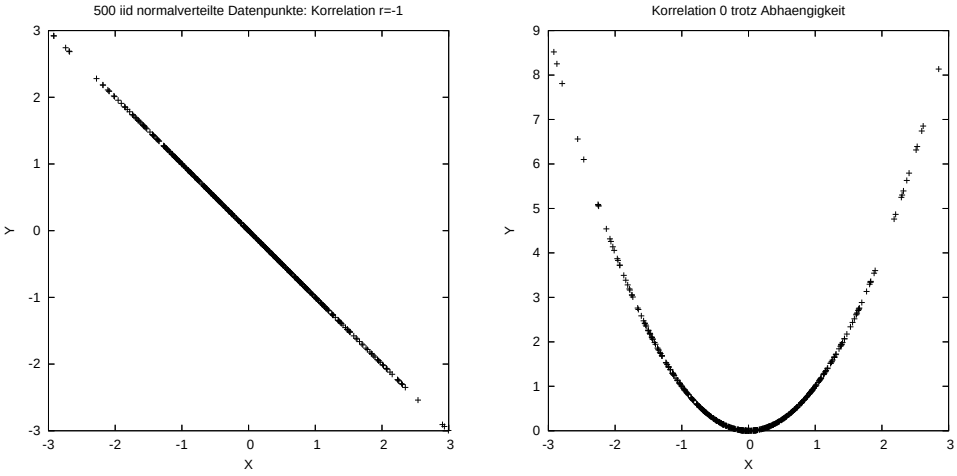
wobei

$$\mathrm{r}_P(X, Y) = +1 \text{ für } X - E_P[X] = \beta(Y - E_P[Y]) \text{ mit } \beta > 0, \tag{448}$$
$$\mathrm{r}_P(X, Y) = -1 \text{ für } X - E_P[X] = \beta(Y - E_P[Y]) \text{ mit } \beta < 0. \tag{449}$$

Die Korrelation misst, wie stark X und Y linear voneinander abhängen. Die folgenden Graphiken mit simulierten Datensätzen verschiedener Korrelation sollen das illustrieren.


500 iid normalverteilte Datenpunkte: Korrelation r=1
500 iid normalverteilte Datenpunkte: Korrelation r=0.8
X
Y


500 iid normalverteilte Datenpunkte: Korrelation r=0
500 iid normalverteilte Datenpunkte: Korrelation r=-0.8
X
Y


500 iid normalverteilte Datenpunkte: Korrelation r=-1
Korrelation 0 trotz Abhaengigkeit
X
Y

Insbesondere sind *unabhängige* Zufallsvariablen $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ unkorreliert, d.h. es gilt

$$\mathrm{Cov}_P(X,Y) = 0, \qquad \mathrm{r}_P(X,Y) = 0, \tag{450}$$

wobei natürlich X und Y nicht P-fast sicher konstant sein dürfen, damit die Korrelation $\mathrm{r}_P(X,Y)$ definiert ist. Die Gleichungen (450) sind Konsequenzen der nachfolgenden Version des Satzes von Fubini aus der Maßtheorie:

Satz 2.86 (Satz von Fubini für integrierbare Funktionen) *Es seien $(\Omega, \mathcal{A}, \mu)$ und $(\Sigma, \mathcal{B}, \nu)$ zwei σ-endliche Maßräume und $f : \Omega \times \Sigma \to \mathbb{R}$ eine $\mathcal{A} \otimes \mathcal{B}$–$\mathcal{B}(\mathbb{R})$–messbare Abbildung. Es gelte $|f| \leq g$ 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*[28] **Wichtig!**

$$\begin{aligned} \int_{\Omega\times\Sigma} f \, d(\mu \times \nu) &= \int_\Omega \int_\Sigma f(x,y)\, \nu(dy)\, \mu(dx) \\ &= \int_\Sigma \int_\Omega f(x,y)\, \mu(dx)\, \nu(dy). \end{aligned} \tag{452}$$

Wir erhalten:

Korollar 2.87 (Faktorisierung der Erwartung unabhängiger Zufallsvariablen) *Sind $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ bezüglich P unabhängig, so gilt $XY \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ und*

$$E_P[XY] = E_P[X] E_P[Y]. \tag{453}$$

Beweis: Wir zeigen das zuerst im Fall $X \geq 0$, $Y \geq 0$. Wir setzen

$$\mu := \mathcal{L}_P(X), \qquad \nu := \mathcal{L}_P(Y), \tag{454}$$

also

$$\mathcal{L}_P(X,Y) = \mu \times \nu \tag{455}$$

[28] Genauer: Unter den Voraussetzungen des Satzes ist $f(x,\cdot)$ für μ-fast alle $x \in \Omega$ bzgl. ν integrierbar, und die Abbildung

$$\Omega \ni x \mapsto \begin{cases} \int_\Sigma f(x,y)\, \nu(dy) & \text{falls das Integral existiert,} \\ 0 & \text{sonst} \end{cases} \tag{451}$$

ist bezüglich μ integrierbar mit dem in Formel (452) angegebenem Integral. Analoges gilt für vertauschte Rollen von x und y.

wegen der vorausgesetzten Unabhängigkeit. Es folgt:

$$\begin{aligned}
E_P[XY] &= \int_{\mathbb{R}^2} x_1 x_2 \, \mu \times \nu(dx) \quad \text{(Notation: } x = (x_1, x_2)\text{)} \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x_1 x_2 \, \nu(dx_2) \, \mu(dx_1) \quad \text{(mit Fubini für nichtnegative Integranden)} \\
&= \int_{\mathbb{R}} x_1 \underbrace{\int_{\mathbb{R}} x_2 \, \nu(dx_2)}_{=E_P[Y]} \, \mu(dx_1) \\
&= \int_{\mathbb{R}} x_1 \, \mu(dx_1) \cdot E_P[Y] \\
&= E_P[X] E_P[Y]. \qquad (456)
\end{aligned}$$

Nun seien beliebige Vorzeichen für Werte von X und Y zugelassen. Die Rechnung von eben zeigt

$$E_P[|XY|] = E_P[|X|] E_P[|Y|] < \infty, \qquad (457)$$

also $|XY| \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ und daher $XY \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$. Damit gilt die Rechnung (456) auch für diese X und Y, wobei die Anwendung des Satzes von Fubini für nichtnegative Integranden (Satz 2.48) durch eine Anwendung des Satzes von Fubini für integrierbare Funktionen (Satz 2.86) ersetzt wird.

□

Korollar 2.88 (Unabhängigkeit impliziert Unkorreliertheit) *Sind $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ bezüglich P unabhängig, so sind X und Y unkorreliert:*

$$\mathrm{Cov}_P(X, Y) = 0. \qquad (458)$$

Beweis:

$$\mathrm{Cov}_P(X, Y) = E_P[XY] - E_P[X] E_P[Y] = 0. \qquad (459)$$

□

Beispiel: Es seien X und Y unabhängige, standardnormalverteilte Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$. Es seien weiter $\alpha, \beta \in \mathbb{R}$ und

$$Z := \alpha X + \beta Y. \qquad (460)$$

Dann folgt

$$\mathrm{Cov}_P(X, Z) = \alpha \underbrace{\mathrm{Cov}_P(X, X)}_{=\mathrm{Var}_P(X)=1} + \beta \underbrace{\mathrm{Cov}_P(X, Y)}_{=0} = \alpha, \qquad (461)$$

$$\begin{aligned}
\mathrm{Var}_P(Z) &= \mathrm{Cov}_P(\alpha X + \beta Y, \alpha X + \beta Y) \\
&= \alpha^2 \underbrace{\mathrm{Var}_P(X)}_{=1} + 2\alpha\beta \underbrace{\mathrm{Cov}_P(X, Y)}_{=0} + \beta^2 \underbrace{\mathrm{Var}_P(Y)}_{=1} \\
&= \alpha^2 + \beta^2. \qquad (462)
\end{aligned}$$

Im Fall $(\alpha, \beta) \neq (0,0)$ bedeutet das:

$$\mathrm{r}_P(X,Z) = \frac{\alpha}{\sqrt{1 \cdot (\alpha^2+\beta^2)}} = \frac{\alpha}{\sqrt{\alpha^2+\beta^2}}. \tag{463}$$

Für beliebige Zufallsvariablen $X_1, \ldots, X_n \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$, ohne jede Annahme über Abhängigkeiten, wissen wir:

$$E_P[X_1 + \ldots + X_n] = E_P[X_1] + \ldots + E_P[X_n]. \tag{464}$$

Falls die $X_1, \ldots, X_n$ jedoch unabhängig oder wenigstens unkorreliert sind und endliche Varianzen besitzen, gilt mehr:

Satz 2.89 (Additivität der Varianz unabhängiger Zufallsvariablen)
Es seien $X_1, \ldots, X_n \in \mathcal{L}(\Omega, \mathcal{A}, P)$ unabhängige Zufallsvariablen, oder wenigstens unkorrelierte Zufallsvariablen, d.h. $\mathrm{Cov}_P(X_i, X_j) = 0$ für $i, j \in [n]$ mit $i \neq j$. Dann gilt **Wichtig!**

$$\mathrm{Var}_P\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n \mathrm{Var}_P(X_k). \tag{465}$$

Beweis: Die vorausgesetzte Unabhängigkeit impliziert die Unkorreliertheit. Wir rechnen:

$$\mathrm{Var}_P\left(\sum_{k=1}^n X_k\right) = \mathrm{Cov}_P\left(\sum_{k=1}^n X_k, \sum_{l=1}^n X_l\right) = \sum_{k=1}^n \sum_{l=1}^n \mathrm{Cov}_P\left(X_k, X_l\right). \tag{466}$$

Nun gilt

$$\mathrm{Cov}_P(X_k, X_l) = \begin{cases} \mathrm{Var}_P(X_k) & \text{für } k = l, \\ 0 & \text{für } k \neq l. \end{cases} \tag{467}$$

In der Doppelsumme bleiben also nur die Diagonalterme übrig, und es folgt die Behauptung (465).

□

Beispiel: Es seien $X_1, \ldots, X_n$ i.i.d. $p\delta_1 + (1-p)\delta_0$-verteilte Zufallsvariablen, wobei $0 \leq p \leq 1$. Dann ist

$$S := \sum_{k=1}^n X_k \tag{468}$$

binomial(n,p)-verteilt. Es folgt:

$$E[S] = \sum_{k=1}^n \underbrace{E[X_k]}_{=p} = np, \tag{469}$$

$$\mathrm{Var}(S) = \sum_{k=1}^n \underbrace{\mathrm{Var}(X_k)}_{=p(1-p)} = np(1-p). \tag{470}$$

Die Binomialverteilung mit den Parametern n und p besitzt den Erwartungswert np, die Varianz $np(1-p)$ und die Standardabweichung $\sqrt{np(1-p)}$.

Man beachte, dass der Erwartungswert np für $n \to \infty$ proportional zu n wächst, die Standardabweichung $\sqrt{np(1-p)}$ jedoch nur proportional zu $\sqrt{n}$, also viel langsamer.

2.12.1 Momente und momentenerzeugende Funktion

Definition 2.90 (Momente und momentenerzeugende Funktion) *Es sei X eine Zufallsvariable über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und $m \in \mathbb{N}$. Wir nennen die Erwartung $E_P[X^m]$, falls sie existiert, das* m-te Moment *von X, und die Erwartung $E_P[(X - E_P[X])^m]$, falls sie existiert, das* m-te zentrierte Moment. *Weiter bezeichne*

$$\mathcal{L}^m(\Omega, \mathcal{A}, P) := \{X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ Zufallsvariable} |\ E_P[|X|^m] < \infty\} \tag{471}$$

den Raum aller Zufallsvariablen über $(\Omega, \mathcal{A}, P)$ mit endlichem m-ten Moment.[29] *Wir nennen die Abbildung*

$$L_X : \mathbb{R} \to [0, +\infty], \quad L_X(s) := E_P[e^{sX}] \tag{472}$$

die momentenerzeugende Funktion *oder auch die* Laplacetransformierte *von X.*

Ausblick. Obwohl wir in dieser Vorlesung nur mit der Laplacetransformierten im Reellen arbeiten, spielt ihre Erweiterung auf komplexe Argumente $s \in \mathbb{C}$ ebenfalls eine wichtige Rolle: Für komplexwertige Zufallsvariablen $Z : \Omega \to \mathbb{C}$ definiert man die Erwartung so:[30]

$$E_P[Z] := E_P[\operatorname{Re} Z] + iE_P[\operatorname{Im} Z], \tag{473}$$

wann immer diese Erwartungen endlich sind. Damit definiert man für $s \in \mathbb{C}$:

$$L_X(s) := E_P[e^{sX}] = E_P[e^{(\operatorname{Re} s)X}[\cos((\operatorname{Im} s)X) + i \sin((\operatorname{Im} s)X)]], \tag{474}$$

wann immer diese Erwartung definiert ist. Man nennt $L_X(s)$ die *Fourier-Laplace-Transformierte* der Zufallsvariablen X an der Stelle s. Die Einschränkung der Fourier-Laplace-Transformierten auf die reelle Achse ist die Laplacetransformierte; die Einschränkung auf die imaginäre Achse

$$L_X(it) = E_P[e^{itX}], \quad t \in \mathbb{R}, \tag{475}$$

heißt die *Fouriertransformierte* von X an der Stelle $t \in \mathbb{R}$. Sie ist stets für alle $t \in \mathbb{R}$ definiert.

Der Name "momentenerzeugende Funktion" findet seine Rechtfertigung im folgenden Satz:

[29] Dieser Raum wird auch analog für $m \in \mathbb{R}$ mit $m \geq 1$ definiert. Für $m = 2$ ist es der uns schon bekannte Raum $\mathcal{L}^2$ der Zufallsvariablen mit endlicher Varianz.

[30] Hier bezeichnet $i \in \mathbb{C}$ die imaginäre Einheit: $i^2 = -1$.

Satz 2.91 (Wie die momentenerzeugende Funktion die Momente erzeugt) *Es sei X eine Zufallsvariable über $(\Omega, \mathcal{A}, P)$ und $U \subseteq \mathbb{R}$ offen. Es gelte $L_X(s) < \infty$ für alle $s \in U$. Dann ist die Laplacetransformierte L_X auf U beliebig oft differenzierbar, und es gilt für alle $m \in \mathbb{N}$, $s \in U$:*

$$\frac{d^m}{ds^m} E_P[e^{sX}] = E_P[X^m e^{sX}]. \tag{476}$$

Besonders interessant ist das im Fall $s = 0 \in U$: Hier gilt:

$$L_X^{(m)}(0) = E_P[X^m]. \tag{477}$$

Der **Beweis** besteht in einer Rechtfertigung der folgenden Vertauschbarkeit von Ableitung und Erwartung:

$$\frac{d^m}{ds^m} E_P[e^{sX}] = E_P\left[\underbrace{\frac{\partial^m}{\partial s^m} e^{sX}}_{=X^m e^{sX}}\right] \tag{478}$$

Hierzu verwenden wir den folgenden Satz aus der Maßtheorie:

Satz 2.92 (Satz von Lebesgue von der Vertauschbarkeit von Integral und Ableitung) *Es sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum, $V \subseteq \mathbb{R}$ eine offene Menge, $f : \Omega \times V \to \mathbb{R}$ messbar im ersten Argument und differenzierbar im zweiten Argument, d.h.*

$$\forall s \in V : \ f(\cdot, s) : \Omega \to \mathbb{R} \text{ sei messbar,} \tag{479}$$

$$\forall \omega \in \Omega : \ f(\omega, \cdot) : V \to \mathbb{R} \text{ sei differenzierbar.} \tag{480}$$

Es gelte

$$\forall s \in V : \ f(\cdot, s) \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu), \tag{481}$$

und es existiere eine "integrierbare Majorante" $g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$:

$$\forall s \in V : \ \left|\frac{\partial}{\partial s} f(\cdot, s)\right| \leq g. \tag{482}$$

Dann ist die Abbildung

$$V \ni s \mapsto \int_\Omega f(\omega, s)\, \mu(d\omega) \tag{483}$$

differenzierbar, und es gilt

$$\frac{d}{ds} \int_\Omega f(\omega, s)\, \mu(d\omega) = \int_\Omega \frac{\partial}{\partial s} f(\omega, s)\, \mu(d\omega). \tag{484}$$

Um die Formel (478) zu beweisen, brauchen wir für jedes $s \in U$ und $m \in \mathbb{N}$ also nur eine offene Umgebung $V \subseteq U$ von s und eine Zufallsvariable $g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ zu finden, so daß gilt:

$$\forall t \in V : \left| \underbrace{X^m e^{tX}}_{=\frac{\partial^m}{\partial t^m} e^{tX}} \right| \leq g. \tag{485}$$

Das folgt so: Es sei $\epsilon > 0$ so klein, dass $[s-2\epsilon, s+2\epsilon] \subseteq U$ gilt. Wir setzen $V :=]s-\epsilon, s+\epsilon[$. Dann folgt für alle $t \in V$:

$$\begin{aligned}
|X^m e^{tX}| &= |X^m| e^{(t-s)X} e^{sX} \\
&\leq \frac{m!}{\epsilon^m} \frac{|\epsilon X|^m}{m!} \cdot e^{\epsilon|X|} e^{sX} \qquad \text{(wegen } |(t-s)X| \leq |\epsilon X|) \\
&\leq \frac{m!}{\epsilon^m} \underbrace{\sum_{k=0}^{\infty} \frac{|\epsilon X|^k}{k!}}_{=e^{\epsilon|X|}} \cdot e^{\epsilon|X|} e^{sX} \\
&= \frac{m!}{\epsilon^m} e^{2\epsilon|X|} e^{sX} \\
&\leq \frac{m!}{\epsilon^m} [e^{(s-2\epsilon)X} + e^{(s-2\epsilon)X}] \\
&=: g \in \mathcal{L}^1(\Omega, \mathcal{A}, P) \qquad \text{(wg. } s \pm 2\epsilon \in U)
\end{aligned} \tag{486}$$

Man beachte, dass die Majorante g nicht von der Wahl von $t \in V$ abhängt.

□

Übung 2.93 (Varianz und log-Laplacetransformierte) *Es sei X eine reellwertige Zufallsvariable, deren Laplacetransformierte L_X in einer Umgebung von 0 endlich sei. Weiter sei $F = \log L_X$ der Logarithmus davon; er wird auch log-Laplacetransformierte genannt. Zeigen Sie:*

$$F''(0) = \mathrm{Var}(X). \tag{487}$$

Beispiele:

1. Die Laplacetransformierte einer Poisson(λ)-verteilten Zufallsvariablen X lautet:

$$\begin{aligned}
L_X(s) &= \sum_{k=0}^{\infty} e^{sk} P[X = k] = \sum_{k=0}^{\infty} e^{sk} e^{-\lambda} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^s \lambda)^k}{k!} = \exp(-\lambda + e^s \lambda), \quad s \in \mathbb{R}.
\end{aligned} \tag{488}$$

Es folgt

$$L'_X(s) = e^s\lambda \exp(-\lambda + e^s\lambda), \tag{489}$$
$$L''_X(s) = [e^s\lambda + (e^s\lambda)^2] \exp(-\lambda + e^s\lambda), \tag{490}$$

und daher

$$E_P[X] = L'_X(0) = \lambda, \tag{491}$$
$$E_P[X^2] = L''_X(0) = \lambda + \lambda^2, \tag{492}$$
$$\mathrm{Var}_P(X) = E_P[X^2] - E_P[X]^2 = \lambda. \tag{493}$$

Alternativ erhält man die gleiche Varianz aus der log-Lapacetransformierten:

$$\log L_X(s) = -\lambda + e^s\lambda, \tag{494}$$
$$(\log L_X)''(s) = e^s\lambda, \tag{495}$$
$$\mathrm{Var}_P(X) = (\log L_X)''(0) = \lambda. \tag{496}$$

2. Sind X und Y *unabhängige* Zufallsvariablen, so gilt für alle $s \in \mathbb{R}$:

$$\boxed{L_{X+Y}(s) = L_X(s) \cdot L_Y(s)} \tag{497}$$

Beweis:

$$\begin{aligned} L_{X+Y}(s) &= E_P[e^{s(X+Y)}] \\ &= E_P[e^{sX}e^{sY}] \\ &= E_P[e^{sX}]E_P[e^{sY}] \quad (\text{da } e^{sX},\ e^{sY} \text{ unabhängig}) \\ &= L_X(s) \cdot L_Y(s). \end{aligned} \tag{498}$$

Analog gilt:

$$\boxed{L_{\sum_{k=1}^n X_k}(s) = \prod_{k=1}^n L_{X_k}(s) \text{ für unabhängige } X_1, \dots X_n} \tag{499}$$

3. Sind $X_1, \dots, X_n$ i.i.d. $p\delta_1+(1-p)\delta_0$–verteilte Zufallsvariablen, also $S_n = X_1+\ldots+X_n$ binomial(n,p)-verteilt, so gilt für $k \in [n]$:

$$L_{X_k}(t) = pe^{t\cdot 1} + (1-p)e^{t\cdot 0} = pe^t + 1 - p, \tag{500}$$

also

$$\begin{aligned} L_{S_n}(t) &= E\left[\prod_{k=1}^n e^{tX_k}\right] \\ &= \prod_{k=1}^n E\left[e^{tX_k}\right] \quad (\text{wg. Unabhängigkeit der Faktoren}) \\ &= (pe^t + 1 - p)^n. \end{aligned} \tag{501}$$

Insbesondere gilt für $n \geq 2$:

$$L'_{S_n}(t) = npe^t(pe^t + 1 - p)^{n-1}, \tag{502}$$
$$E[S_n] = L'_{S_n}(0) = np, \tag{503}$$
$$L''_{S_n}(t) = npe^t(pe^t + 1 - p)^{n-1} + n(n-1)p^2e^{2t}(pe^t + 1 - p)^{n-2}, \tag{504}$$
$$E[S_n^2] = np + n(n-1)p^2, \tag{505}$$

und daher

$$\mathrm{Var}(S_n) = E[S_n^2] - E[S_n]^2 = np + n(n-1)p^2 - (np)^2 = np(1-p), \tag{506}$$

wie uns aus Formel (470) auch schon bekannt ist.

Alternativ erhalten wir mit der log-Laplacetransformierten das gleiche Ergebnis:

$$\log L_{S_n}(t) = n \log(pe^t + 1 - p), \tag{507}$$
$$(\log L_{S_n})'(t) = n\frac{pe^t}{pe^t + 1 - p}, \tag{508}$$
$$(\log L_{S_n})''(t) = n\frac{pe^t(pe^t + 1 - p) - pe^t \cdot pe^t}{(pe^t + 1 - p)^2} = n\frac{pe^t(1-p)}{(pe^t + 1 - p)^2}, \tag{509}$$
$$\mathrm{Var}(S_n) = (\log L_{S_n})'(0) = np(1-p). \tag{510}$$

2.12.2 Erwartung von Indikatorfunktionen und allgemeine Tschebyscheff-Ungleichung

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

$$\boxed{P(A) = E_P[1_A]} \tag{511}$$

hat erstaunliche Konsequenzen, von denen wir zwei jetzt besprechen:

Das Inklusions-Exklusions-Prinzip

Satz 2.94 (Inklusions-Exklusions-Prinzip) *Es seien $A_1, \ldots, A_n \in \mathcal{A}$ Ereignisse, wobei $n \in \mathbb{N}$. Dann gilt:*

$$\boxed{P(A_1 \cup \ldots \cup A_n) = \sum_{\substack{E \subseteq [n] \\ E \neq \emptyset}} (-1)^{|E|-1} P\left(\bigcap_{i \in E} A_i\right)} \tag{512}$$

Das ist eine Verallgemeinerung der bekannten Formeln

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2), \tag{513}$$
$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) =& P(A_1) + P(A_2) + P(A_3) \\ & - P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_1 \cap A_3) \\ & + P(A_1 \cap A_2 \cap A_3). \end{aligned} \tag{514}$$

Beweis des Inklusions-Exklusions-Prinzips: Wir rechnen:

$$\begin{aligned} P\left(\bigcup_{i\in[n]} A_i\right) &= 1 - P\left(\left(\bigcup_{i\in[n]} A_i\right)^c\right) \\ &= 1 - P\left(\bigcap_{i\in[n]} A_i^c\right) \\ &= 1 - E_P\left[1_{\bigcap_{i\in[n]} A_i^c}\right] \\ &= 1 - E_P\left[\prod_{i\in[n]} 1_{A_i^c}\right]. \end{aligned} \tag{515}$$

Das Argument der letzten Erwartung formen wir so um:

$$\begin{aligned} \prod_{i\in[n]} 1_{A_i^c} &= \prod_{i\in[n]} (1 - 1_{A_i}) \\ &= \sum_{E\subseteq[n]} \prod_{i\in E} (-1_{A_i}) \quad \text{(mit dem Distributivgesetz)} \\ &= 1 - \sum_{\substack{E\subseteq[n] \\ E\neq\emptyset}} (-1)^{|E|-1} \prod_{i\in E} 1_{A_i}. \\ &= 1 - \sum_{\substack{E\subseteq[n] \\ E\neq\emptyset}} (-1)^{|E|-1} 1_{\bigcap_{i\in E} A_i}. \end{aligned} \tag{516}$$

Eingesetzt in (515) folgt:

$$\begin{aligned} P\left(\bigcup_{i\in[n]} A_i\right) &= 1 - E_P\left[1 - \sum_{\substack{E\subseteq[n] \\ E\neq\emptyset}} (-1)^{|E|-1} 1_{\bigcap_{i\in E} A_i}\right] \\ &= \sum_{\substack{E\subseteq[n] \\ E\neq\emptyset}} (-1)^{|E|-1} E_P\left[1_{\bigcap_{i\in E} A_i}\right] \\ &= \sum_{\substack{E\subseteq[n] \\ E\neq\emptyset}} (-1)^{|E|-1} P\left(\bigcap_{i\in E} A_i\right), \end{aligned} \tag{517}$$

wie zu beweisen war.

□

Die allgemeine Tschebyscheff-Ungleichung

Lemma 2.95 (allgemeine Tschebyscheff-Ungleichung) *Es seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $A \in \mathcal{A}$ ein Ereignis, $X \geq 0$ eine nichtnegative Zufallsvariable, und $c \geq 0$ eine nichtnegative reelle Zahl. Es gelte*

$$X(\omega) \geq c \text{ für alle } \omega \in A. \tag{518}$$

Dann folgt: **Wichtig!**

$$E_P[X] \geq cP(A). \tag{519}$$

Beweis: Nach Voraussetzung gilt

$$X \geq c1_A. \tag{520}$$

Nehmen wir die Erwartung hiervon, folgt die Behauptung:

$$E_P[X] \geq E_P[c1_A] = cE_P[1_A] = cP(A). \tag{521}$$

□

Trotz ihrer Einfachheit hat diese Ungleichung erstaunliche Anwendungen:

Satz 2.96 (Exponentielle Tschebyscheff-Ungleichung) *Es sei X eine reellwertige Zufallsvariable über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann gilt für alle $s \geq 0$ und $a \in \mathbb{R}$:*

$$\boxed{E_P[e^{sX}] \geq e^{sa}P[X \geq a]} \tag{522}$$

und daher: **Wichtig!**

$$\boxed{P[X \geq a] \leq \inf_{s \geq 0}(e^{-sa}E_P[e^{sX}])} \tag{523}$$

Beweis: Die erste Version (522) der exponentiellen Tschebyscheff-Ungleichung folgt aus der allgemeinen Tschebyscheff-Ungleichung angewandt auf die Ungleichung

$$e^{sX} \geq e^{sa}1_{\{X \geq a\}}. \tag{524}$$

Diese Ungleichung sieht man so: Gegeben $\omega \in \Omega$, unterscheiden wir zwei Fälle:

- Ist $X(\omega) < a$, so gilt

$$e^{sX(\omega)} > 0 = e^{sa}1_{\{X \geq a\}}(\omega). \tag{525}$$

- Ist $X(\omega) \geq a$, so gilt wegen $s \geq 0$ die Ungleichung $sX(\omega) \geq sa$ und daher

$$e^{sX(\omega)} \geq e^{sa} = e^{sa}1_{\{X \geq a\}}(\omega). \tag{526}$$

Dividieren wir Ungleichung (522) durch $e^{sa} > 0$ und bilden wir das Infimum über $s \geq 0$, folgt auch die zweite Version (523) der Ungleichung.

□

Korollar 2.97 (Exponentielle Tschebyscheff-Ungleichung – Version für Abweichungen nach unten) *Für alle reellwertigen Zufallsvariablen X auf $(\Omega, \mathcal{A}, P)$ und alle $a \in \mathbb{R}$ gilt:*

$$\boxed{P[X \leq a] \leq \inf_{s \leq 0}(e^{-sa} E_P[e^{sX}])} \tag{527}$$

Übung 2.98 *Beweisen Sie die Ungleichung* (527)*.*

Beispiel: Große Abweichungen für binomialverteilte Zufallsvariablen. Sind $(Y_k)_{i \in \mathbb{N}}$ i.i.d. $p\delta_1 + (1-p)\delta_0$-verteilte Zufallsvariablen mit einem Parameter $0 < p < 1$, so ist für $n \in \mathbb{N}$ die Zufallsvariable

$$X_n := \sum_{k=1}^{n} Y_k \tag{528}$$

binomial(n,p)-verteilt. Nun sei eine Zahl $a \in \mathbb{R}$ mit $p < a < 1$ gegeben. Dann gilt:

$$\begin{aligned}
\text{binomial}(n,p)([na, \infty[) &= P[X_n \geq na] \\
&\leq \inf_{s \geq 0} e^{-sna} E_P[e^{sX_n}] \quad \text{(mit der exp. Tschebyscheff-Ungl.)} \\
&\leq \inf_{s \geq 0} e^{-sna}(pe^s + 1 - p)^n \quad \text{(mit Formel (501))} \\
&= \inf_{s \geq 0} \exp\{n[\underbrace{-sa + \log(pe^s + 1 - p)}_{=:H(s)}]\}
\end{aligned} \tag{529}$$

Wir optimieren nun über $s \geq 0$. Hierzu bilden wir die Ableitung:

$$\begin{aligned}
H'(s) &= \frac{\partial}{\partial s}[-sa + \log(pe^s + 1 - p)] = -a + \frac{pe^s}{pe^s + 1 - p} \\
&= -a + \left(1 + \frac{1-p}{p} e^{-s}\right)^{-1}, \qquad (530) \\
H''(s) &= \frac{\frac{1-p}{p} e^{-s}}{\left(1 + \frac{1-p}{p} e^{-s}\right)^2} > 0. \qquad (531)
\end{aligned}$$

Insbesondere ist H eine konvexe Funktion. Um das Minimum zu finden, lösen wir die Gleichung $H'(s) = 0$: Es gelten die folgenden Äquivalenzen:

$$H'(s) = 0 \quad \Leftrightarrow \quad -a + \left(1 + \frac{1-p}{p} e^{-s}\right)^{-1} = 0 \tag{532}$$

$$\Leftrightarrow \quad 1 + \frac{1-p}{p} e^{-s} = \frac{1}{a} \tag{533}$$

$$\Leftrightarrow \quad \frac{1-p}{p} e^{-s} = \frac{1}{a} - 1 = \frac{1-a}{a} \tag{534}$$

$$\Leftrightarrow \quad s = \log\left(\frac{1-p}{p}\frac{a}{1-a}\right). \tag{535}$$

Wegen $1 > a \geq p > 0$ ist $\frac{a}{p} \geq 1$ und $\frac{1-p}{1-a} \geq 1$, also $s \geq 0$ für den optimalen Wert s aus Gleichung (535). Eingesetzt erhalten wir an diesem Optimum:

$$pe^s + 1 - p = (1-p)\frac{a}{1-a} + 1 - p = \frac{1-p}{1-a}, \tag{536}$$

und daher den Wert des Optimums

$$\begin{aligned} h(a,p) := H(s) &= -sa + \log(pe^s + (1-p)) \\ &= a \log\left(\frac{p}{1-p}\frac{1-a}{a}\right) + \log\left(\frac{1-p}{1-a}\right) \qquad (537) \\ &= a \log\frac{p}{a} + (1-a)\log\frac{1-p}{1-a}. \qquad (538) \end{aligned}$$

Es gilt wegen

$$\log x < x - 1 \text{ für } x \in \mathbb{R}^+ \setminus \{1\} \tag{539}$$

die Ungleichung[31]

$$\begin{aligned} h(a,p) &= a \underbrace{\log \frac{p}{a}}_{<\frac{p}{a}-1} + (1-a) \underbrace{\log \frac{1-p}{1-a}}_{<\frac{1-p}{1-a}-1} \\ &< a\left(\frac{p}{a}-1\right) + (1-a)\left(\frac{1-p}{1-a}-1\right) \\ &= (p-a) + ((1-p)-(1-a)) \\ &= 0. \end{aligned} \tag{544}$$

Damit ist gezeigt:

Satz 2.99 (Große Abweichungen für die Binomialverteilung – obere Schranke) *Gegeben $0 < p < a < 1$ und $n \in \mathbb{N}$, sei X_n eine binomial(n,p)-verteilte Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Dann gilt:*

$$\boxed{P[X_n \geq na] \leq e^{nh(a,p)}} \tag{545}$$

*mit der "*relativen Entropie*":* **Wichtig!**

$$\boxed{h(a,p) = a\log\frac{p}{a} + (1-a)\log\frac{1-p}{1-a} < 0} \tag{546}$$

Insbesondere gilt

$$\boxed{P\left[\frac{X_n}{n} \geq a\right] \overset{n\to\infty}{\longrightarrow} 0 \text{ exponentiell schnell.}} \tag{547}$$

[31] Man nennt die Größe

$$h(a,p) = a\log\frac{p}{a} + (1-a)\log\frac{1-p}{1-a} \tag{540}$$

auch die *relative Entropie* der Verteilung $\mu = a\delta_1 + (1-a)\delta_0$ bezüglich der Verteilung $\nu = p\delta_1 + (1-p)\delta_0$. Sie wird allgemeiner durch

$$h(\mu,\nu) := E_\mu\left[\log\frac{d\nu}{d\mu}\right] \tag{541}$$

definiert. Relative Entropien spielen nicht nur in der Theorie großer Abweichungen eine wichtige Rolle, von der wir hier einen ersten Eindruck bekommen, sondern auch in der mathematischen Statistik, in der Informationstheorie und in der statistischen Physik. Auch allgemeiner gilt

$$h(\mu,\nu) \leq 0 \tag{542}$$

wegen

$$h(\mu,\nu) = E_\mu\left[\log\frac{d\nu}{d\mu}\right] \leq E_\mu\left[\frac{d\nu}{d\mu} - 1\right] = E_\nu[1] - E_\nu[1] = 0. \tag{543}$$

Gleichheit $h(\mu,\nu) = 0$ tritt dabei nur im Fall $\mu = \nu$ auf.

Interpretation: Interpretieren wir X_n/n als die relative Häufigkeit der "1" in einer i.i.d. Münzwurfsequenz mit Wahrscheinlichkeit p für "1", so gilt $p = E_P[X_n/n]$. Die Wahrscheinlichkeit, eine "große Abweichung" $X_n/n \geq a$ der relativen Häufigkeit zu sehen, wobei $a > p$, ist also exponentiell klein in der Anzahl n der Münzwürfe.

Übung 2.100 *Beweisen Sie die folgende Variante des Satzes 2.99: Gegeben* $0 < a < p < 1$ *und* $n \in \mathbb{N}$*, sei* X_n *eine binomial*(n,p)*-verteilte Zufallsvariable auf einem Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, P)$*. Dann gilt*

$$P\left[X_n \leq na\right] \leq e^{nh(a,p)} \qquad mit \qquad h(a,p) = a \log \frac{p}{a} + (1-a) \log \frac{1-p}{1-a} < 0. \tag{548}$$

Sie können dazu zum Beispiel "0" und "1" in Münzwurfsequenzen vertauschen.

Numerisches Beispiel für den fairen Münzwurf, $p = \frac{1}{2}$, $a = 0,6 = 60\%$:

$$h := h(a,p) = -0.020\ldots \tag{549}$$

Wir erhalten:

$$P[X_{100} \geq 60] \leq e^{100h} = 13,35\ldots\,\%, \tag{550}$$

$$P[X_{1000} \geq 600] \leq e^{1000h} = 1,79\ldots \cdot 10^{-9}, \tag{551}$$

$$P[X_{10000} \geq 6000] \leq e^{10000h} = 3,5\ldots \cdot 10^{-88}. \tag{552}$$

Es ist also "praktisch unmöglich", bei 10000 fairen Münzwürfen eine relative Häufigkeit der "1" von mindestens 60% zu beobachten.[32]

Die Markoff-Ungleichung Andere Anwendungen der allgemeinen Tschebyscheff-Ungleichung erhält man, wenn man statt der Exponentialfunktion Potenzfunktionen verwendet:

Satz 2.101 (Markoff-Ungleichung) *Es sei* X *eine Zufallsvariable über* $(\Omega, \mathcal{A}, P)$*. Weiter seien* $m > 0$ *und* $a > 0$ *zwei positive Zahlen. Dann gilt:* **Wichtig!**

$$\boxed{P[|X| \geq a] \leq a^{-m} E_P[|X|^m]} \tag{553}$$

[32] Um die extreme Kleinheit der Wahrscheinlichkeit $P[X_{10000} \geq 6000]$ zu veranschaulichen, stelle man sich vor, dass alle Supercomputer, die es heute gibt oder es bis heute gegeben hat, in jeder Mikrosekunde ihrer gesamten Betriebszeit 10000 faire Münzwürfe simulieren und die relative Häufigkeit der "1" darin berechnen. Auch dann würde man mit an Sicherheit grenzender Wahrscheinlichkeit in keiner der Simulationen eine relative Häufigkeit der 1 von mindestens 60% gesehen haben. Das illustriert, dass die Stochastik auch schon bei moderat großem Stichprobenumfang manchmal sehr scharfe Prognosen machen kann, die praktisch nicht mehr von "sicheren" Prognosen zu unterscheiden sind. In der Philosophie der Minimalinterpretation von Wahrscheinlichkeiten (Seite 19 in Abschnitt 2.1.3) gesprochen, wird die "Rechengröße" $p = \frac{1}{2}$, die zunächst nur "Unsicherheit" bedeutet, durch 10000-fache i.i.d. Wiederholung des Zufallsexperiments in eine "an Sicherheit grenzende Wahrscheinlichkeit" $\geq 1 - 3,5\ldots \cdot 10^{-88}$ transformiert.

Beweis: Es gilt

$$a^m 1_{\{|X|\geq a\}} \leq |X|^m. \tag{554}$$

In der Tat: Für $\omega \in \Omega$ unterscheiden wir zwei Fälle:

- Falls $|X(\omega)| \geq a$ gilt

$$a^m 1_{\{|X|\geq a\}}(\omega) = a^m \leq |X|^m. \tag{555}$$

- Falls $|X(\omega)| < a$ gilt

$$a^m 1_{\{|X|\geq a\}}(\omega) = 0 \leq |X|^m. \tag{556}$$

Erwartungswertbildung liefert

$$a^m P[|X| \geq a] = E_P[a^m 1_{\{|X|\geq a\}}] \leq E_P[|X|^m]. \tag{557}$$

Dividieren wir diese Ungleichung durch a^m, folgt die behauptete Markoff-Ungleichung (553).

□

Spezialfälle:

1. $m = 1$. In diesem Fall lautet die Markov-Ungleichung:

$$P[|X| \geq a] \leq \frac{E_P[|X|]}{a}. \tag{558}$$

2. $m = 2$:

Korollar 2.102 (quadratische Tschebyscheff-Ungleichung) *Für alle $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ und $a \geq 0$ gilt:* **Wichtig!**

$$\boxed{P\left[|X - E_P[X]| \geq a\right] \leq \frac{\mathrm{Var}_P(X)}{a^2}} \tag{559}$$

Beweis: Das ist der Spezialfall $m = 2$ der Markoff-Ungleichung, angewandt auf $Y = X - E_P[X]$. Wegen seiner hohen Bedeutung schreiben wir hier noch einmal den Beweis für diesen Spezialfall auf, obwohl es redundant ist:

$$a^2 1_{\{|Y|\geq a\}} \leq Y^2 \tag{560}$$

$$\Rightarrow \quad a^2 E_P[|Y| \geq a] \leq E_P[Y^2] = \mathrm{Var}_P(X). \tag{561}$$

□

Numerischer Vergleich beim fairen Münzwurf:
Schranke nach quadratischer Tschebyscheff-Ungleichung:

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \epsilon\right] \leq \frac{p(1-p)}{n\epsilon^2}$$

Schranke nach exponentieller Tschebyscheff-Ungleichung:

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \epsilon\right] \leq \exp(nh(p+\epsilon, p)) + \exp(nh(p-\epsilon, p)),$$

wobei

$$h(q,p) = q\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q}$$

Exakt:

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \epsilon\right] = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} 1_{\{|\frac{k}{n}-p|\geq\epsilon\}}$$

Numerisches Beispiel: $p = \frac{1}{2}$, $\epsilon = \frac{1}{10}$:

n	quadratisch	exponentiell	exakt
10	2.5	$1.63524\ldots$	$0.753906\ldots$
100	0.25	$0.267027\ldots$	$0.0568879\ldots$
1000	0.025	$3.59988\ldots \cdot 10^{-9}$	$2.72846\ldots \cdot 10^{-10}$
10000	0.0025	$7.13848\ldots \cdot 10^{-88}$	$1.74043\ldots \cdot 10^{-89}$

Es fällt auf, dass hier die Schranke nach der exponentiellen Tschebyscheff-Ungleichung die richtige Größenordnung liefert, während die quadratische Tschebyscheff-Ungleichung hier oft um viele Größenordnungen gröber ist.

2.13 Gesetze der großen Zahlen

2.13.1 Das schwache Gesetz der großen Zahlen

Es seien $X_1, X_2, X_3, \ldots$ i.i.d. Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit endlicher Erwartung $E_P[X_1]$. Informal gesprochen erwartet man, dass der Mittelwert

$$\overline{X}_n := \frac{1}{n}\sum_{k=1}^{n} X_k \tag{562}$$

für "*große*" n "*typischerweise*" "*nahe*" bei dem Erwartungswert $E_P[X_1]$ liegt.
Beispiel: Würfeln wir n-mal mit einem fairen Spielwürfel, so liegt die Augensumme für "große" n "typischerweise" "nahe" bei 3,5.
Wir fassen das formaler:

Definition 2.103 (Konvergenz in Wahrscheinlichkeit) *Eine Folge Y_n, $n \in \mathbb{N}$, von Zufallsvariablen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$* konvergiert in Wahrscheinlichkeit, *synonym* konvergiert stochastisch *gegen $a \in \mathbb{R}$, in Zeichen $Y_n \xrightarrow[P]{n\to\infty} a$, wenn gilt:* **Wichtig!**

$$\boxed{\forall \epsilon > 0 : \; P[|Y_n - a| \geq \epsilon] \overset{n\to\infty}{\longrightarrow} 0} \tag{563}$$

Analog wird die Konvergenz in Wahrscheinlichkeit definiert, wenn die Zufallsvariablen Y_n auf verschiedenen Wahrscheinlichkeiträumen $(\Omega_n, \mathcal{A}_n, P_n)$ definiert sind; es wird dann nur P_n statt P in Formel (563) verwendet.
Mit dieser Definition zeigen wir:

Satz 2.104 (Schwaches Gesetz der großen Zahlen)
Es seien $(X_n)_{n\in\mathbb{N}}$ i.i.d. Zufallsvariablen in $\mathcal{L}^2(\Omega, \mathcal{A}, P)$. Dann gilt für den Mittelwert $\overline{X}_n$, definiert in Formel (562):

$$\overline{X}_n \xrightarrow[P]{n\to\infty} E_P[X_1] \tag{564}$$

Wichtig!

Beweis: Es sei $\epsilon > 0$ gegeben. Wir kürzen ab:

$$a := E_P[X_1] = E_P[X_k] \quad \text{für alle } k \in \mathbb{N}, \tag{565}$$

also auch

$$a := \frac{1}{n} \sum_{k=1}^{n} E_P[X_k] \quad \text{für alle } n \in \mathbb{N}. \tag{566}$$

Es folgt für alle $n \in \mathbb{N}$:

$$\begin{aligned}
\mathrm{Var}_P(\overline{X}_n) &= \mathrm{Var}_P \left(\frac{1}{n} \sum_{k=1}^{n} X_k \right) \\
&= \frac{1}{n^2} \mathrm{Var}_P \left(\sum_{k=1}^{n} X_k \right) \\
&= \frac{1}{n^2} \underbrace{\sum_{k=1}^{n} \mathrm{Var}_P (X_k)}_{= n \, \mathrm{Var}_P(X_1)} \quad (\text{da } (X_k) \text{ i.i.d.}) \\
&= \frac{1}{n} \mathrm{Var}_P (X_1)
\end{aligned} \tag{567}$$

und daher mit der quadratischen Tschebyscheffungleichung (Korollar 2.102):

$$\begin{aligned} P[|\overline{X}_n - a| \geq \epsilon] &\leq E_P\left[\frac{(\overline{X}_n - a)^2}{\epsilon^2}\right] \\ &= \frac{1}{\epsilon^2}\operatorname{Var}_P(\overline{X}_n) \\ &= \frac{1}{\epsilon^2 n}\operatorname{Var}_P(\overline{X}_1) \overset{n\to\infty}{\longrightarrow} 0. \end{aligned} \tag{568}$$

Das bedeutet $\overline{X}_n \overset{n\to\infty}{\longrightarrow}_P a$.

□

Bemerkung: Wir können die Voraussetzungen des Satzes so abschwächen: Man kann die Voraussetzung, dass die X_k i.i.d. sind, ersetzen durch

$$E_P[X_k] = a, \quad (k \in \mathbb{N}) \tag{569}$$

$$\operatorname{Var}_P(X_k) = \sigma^2, \quad (k \in \mathbb{N}) \tag{570}$$

$$\operatorname{Cov}_P(X_k, X_l) = 0 \text{ für } k, l \in \mathbb{N} \text{ mit } k \neq l \tag{571}$$

mit einer Zahl $\sigma^2 \geq 0$. Auch in diesem Fall gilt nämlich

$$E_P\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = a, \tag{572}$$

$$\operatorname{Var}_P\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}\operatorname{Cov}_P(X_k, X_l) = \frac{1}{n^2}\sum_{k=1}^{n}\operatorname{Var}_P(X_k) = \frac{\sigma^2}{n}, \tag{573}$$

und der Beweis funktioniert ebenso.

Korollar 2.105 (Schwaches Gesetz der großen Zahlen für relative Häufigkeiten) *Es seien $(A_n)_{n\in\mathbb{N}}$ unabhängige Ereignisse mit gleicher Wahrscheinlichkeit $P(A_n) = p$ für alle $n \in \mathbb{N}$, so gilt:*

$$\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} \overset{n\to\infty}{\underset{P}{\longrightarrow}} P(A_1), \tag{574}$$

$$\textit{relative Häufigkeit} \underset{P}{\longrightarrow} \textit{Wahrscheinlichkeit} \tag{575}$$

Wichtig!

Das ist der Spezialfall $X_n = 1_{A_n}$ des schwachen Gesetzes der großen Zahlen.

Vergleich des schwachen Gesetzes der großen Zahlen mit Interpretationen des Wahrscheinlichkeitsbegriffs:

- Das schwache Gesetz der großen Zahlen kann als innermathematisches Analogon der objektivistischen Interpretation von Wahrscheinlichkeiten mit relativen Häufigkeiten aufgefasst werden, siehe Abschnitt 1 auf Seite 16f. Anders als die objektivistische Interpretation, die ein Bindeglied zwischen der Theorie und den Anwendungen liefert, ein Bestandteil der stochastischen Modellbildung ist, und auch nicht beweisen wird, ist das schwache Gesetz der großen Zahlen ein mathematisches Theorem, das wie jedes Theorem einen Beweis besitzt.

- **Transformation beliebiger Wahrscheinlichkeiten in "an Sicherheit grenzende Wahrscheinlichkeiten":** Das schwache Gesetz der großen Zahlen bildet auch ein Fundament für die Minimalinterpretation von Wahrscheinlichkeiten (siehe Seite 19 in Abschnitt 2.1.3): Es erlaubt, eine *beliebige* Wahrscheinlichkeit

$$P(A_k) = p, \tag{576}$$

die nicht notwendig nahe bei 0 oder nahe bei 1 liegen muss und daher im Sinne der Minimalinterpretation nur Unsicherheit oder nur eine Rechengrösse bedeutet, in Wahrscheinlichkeiten nahe bei 1 zu transformieren:

$$P\left[\left|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\right| < \epsilon\right] \overset{n\to\infty}{\longrightarrow} 1 \qquad \text{für alle } \epsilon > 0. \tag{577}$$

Im Sinne der Minimalinterpretation bedeutet das:
Mit an Sicherheit grenzender Wahrscheinlichkeit wird bei unabhängiger n-facher Wiederholung des Zufallsexperiments die beobachtete relative Häufigkeit eines Ereignisses nahe (genauer gesagt: näher als ϵ) an der Wahrscheinlichkeit des Ereignisses liegen, wenn die Anzahl n der Versuche groß ist.

2.13.2 Das starke Gesetz der großen Zahlen

Es seien $(A_n)_{n\in\mathbb{N}}$ unabhängige Ereignisse über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit gleicher Wahrscheinlichkeit $P(A_n) = p \in]0,1[$ für alle $n \in \mathbb{N}$. Aus der quadratischen Tschebyscheff-Ungleichung im Beweis des schwachen Gesetzes der gorßen Zahl wissen wir für alle $\epsilon > 0$:

$$P\left[\left|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\right| \geq \epsilon\right] \leq \frac{\mathrm{Var}_P(1_{A_1})}{n\epsilon^2} = \frac{p(1-p)}{\epsilon^2}\cdot\frac{1}{n} \tag{578}$$

Diese obere Schranke konvergiert zwar für $n \to \infty$ gegen 0, aber nur langsam, proportional zu $1/n$.

Aus der exponentiellen Tschebyscheff-Ungleichung wissen wir nämlich für $\epsilon > 0$ mit $0 < p \pm \epsilon < 1$:

$$P\left[\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p \geq \epsilon\right] \leq e^{nh(p+\epsilon,p)}, \tag{579}$$

$$P\left[\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p \leq -\epsilon\right] \leq e^{nh(p-\epsilon,p)} \text{ mit} \tag{580}$$

$$h(p \pm \epsilon, p) = (p \pm \epsilon)\log\frac{p}{p \pm \epsilon} + (1 - (p \pm \epsilon))\log\frac{1-p}{1-(p \pm \epsilon)} < 0, \tag{581}$$

siehe Formeln (545) und (546). Setzen wir

$$\alpha := \min\{-h(p+\epsilon, p), -h(p-\epsilon, p)\} > 0, \tag{582}$$

so folgt

$$P\left[\left|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\right| \geq \epsilon\right] \leq 2e^{-\alpha n}, \tag{583}$$

was exponentiell schnell für $n \to \infty$ abfällt, also viel schneller also die nur proportional zu $1/n$ abfallende Schranke (578) aus der quadratischen Tschebyscheff-Ungleichung. Insbesondere folgt aus der Schranke (583) mit der geometrischen Reihe:

$$\sum_{n\in\mathbb{N}} P\left[\left|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\right| \geq \epsilon\right] \leq 2\sum_{n\in\mathbb{N}} e^{-\alpha n} = \frac{2e^{-\alpha}}{1-e^{-\alpha}} < \infty, \tag{584}$$

während die Schranke (578) aus der quadratischen Tschebyscheff-Ungleichung wegen der Divergenz der harmonischen Reihe, $\sum_{n\in\mathbb{N}} n^{-1} = \infty$, nicht ausreicht, dies zu zeigen. Mit dem nächsten Lemma bekommt die Konvergenz der Reihe (584) eine besondere Bedeutung. Zu seiner Formulierung definieren wir:

Definition 2.106 (lim inf und lim sup für Ereignisse) *Es sei $(B_n)_{n\in\mathbb{N}}$ eine Folge von Mengen. Wir definieren*

$$\begin{aligned}\limsup_{n\to\infty} B_n &:= \{B_n \textit{ tritt für unendlich viele } n \textit{ ein}\} \\ &= \{\omega \in \Omega :\ \forall m \in \mathbb{N}\ \exists n \geq m :\ \omega \in B_n\} \\ &= \bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} B_n\end{aligned} \tag{585}$$

und

$$\begin{aligned}\liminf_{n\to\infty} B_n &:= \{B_n \textit{ tritt für } n \to \infty \textit{ schließlich ein}\} \\ &= \{\omega \in \Omega :\ \exists m \in \mathbb{N}\ \forall n \geq m :\ \omega \in B_n\} \\ &= \bigcup_{m\in\mathbb{N}} \bigcap_{n\geq m} B_n\end{aligned} \tag{586}$$

Die Bezeichnungen lim inf, lim sup kommen von den folgenden Gleichungen:

$$1_{\limsup_{n\to\infty} B_n} = \limsup_{n\to\infty} 1_{B_n}, \tag{587}$$

$$1_{\liminf_{n\to\infty} B_n} = \liminf_{n\to\infty} 1_{B_n}. \tag{588}$$

Es gilt

$$\left(\limsup_{n\to\infty} B_n\right)^c = \left(\bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} B_n\right)^c = \bigcup_{m\in\mathbb{N}} \left(\bigcup_{n\geq m} B_n\right)^c = \bigcup_{m\in\mathbb{N}} \bigcap_{n\geq m} B_n^c = \liminf_{n\to\infty} B_n^c. \tag{589}$$

Lemma 2.107 (erstes Lemma von Borel-Cantelli)
Es sei $(B_n)_{n\in\mathbb{N}}$ eine Folge von Ereignissen in einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit

$$\sum_{n\in\mathbb{N}} P(B_n) < \infty. \tag{590}$$

Dann gilt

$$P\left(\limsup_{n\to\infty} B_n\right) = 0. \tag{591}$$

Mit anderen Worten: P-fast sicher tritt das Ereignis B_n nur für endlich viele n ein.

Wichtig!

Beweis: Für alle $m \in \mathbb{N}$ gilt:

$$P\left(\bigcup_{n\geq m} B_n\right) = \lim_{k\to\infty} P\left(\bigcup_{n=m}^{k} B_n\right) \text{ (mit der } \sigma\text{-Stetigkeit von } P \text{ von unten)}$$

$$\leq \lim_{k\to\infty} \sum_{n=m}^{k} P(B_n) = \sum_{n=m}^{\infty} P(B_n) \overset{m\to\infty}{\longrightarrow} 0 \tag{592}$$

wegen der Voraussetzung (590). Damit folgt auch

$$P\left(\bigcup_{n\geq m} B_n\right) \overset{m\to\infty}{\longrightarrow} 0. \tag{593}$$

Weil die Folge $(\bigcup_{n\geq m} B_n)_{m\in\mathbb{N}}$ monoton fällt, folgt mit der σ-Stetigkeit von P von oben"

$$P\left(\limsup_{n\to\infty} B_n\right) = P\left(\bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} B_n\right) = \lim_{m\to\infty} P\left(\bigcup_{n\geq m} B_n\right) = 0. \tag{594}$$

□

Bemerkung: Es gibt eine "Umkehrung" des 1. Borel-Cantelli-Lemmas für unabhängige Ereignisse, das "2. Borel-Cantelli-Lemma. Wir behandeln es erst in der Vorlesung "Wahrscheinlichkeitstheorie".
Als eine Folgerung des erten Borel-Cantelli-Lemmas erhalten wir:

Satz 2.108 (Starkes Gesetz der großen Zahlen für relative Häufigkeiten)
Es sei $(A_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger Ereignisse in einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit gleicher Wahrscheinlichkeit $P(A_n) = p$ für alle $n \in \mathbb{N}$. Dann gilt:

$$\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} \overset{n\to\infty}{\longrightarrow} p \quad P\text{-fast sicher,} \tag{595}$$

Wichtig!

d.h.

$$P\left(\left\{\omega \in \Omega : \frac{1}{n}\sum_{k=1}^{n} 1_{A_k}(\omega) \overset{n\to\infty}{\longrightarrow} p\right\}\right) = 1. \tag{596}$$

Man kann diese Version des Gesetzes der großen Zahlen als ein innermathematisches Analogon der von-Mises-Interpretation von Wahrscheinlichkeiten, siehe Seite 17, auffassen:

relative Häufigkeit bei n unabhängigen Wiederholungen $\underset{P\text{-fast sicher}}{\overset{n\to\infty}{\longrightarrow}}$ Wahrscheinlichkeit p.

Beweis des Satzes: Die Fälle $p = 0$ und $p = 1$ sind trivial. Wir nehmen also $0 < p < 1$ an. Wir definieren für alle $\epsilon > 0$ das Ereignis

$$\begin{aligned} B_\epsilon &:= \limsup_{n\to\infty} \left\{\left|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\right| \geq \epsilon\right\} \\ &= \left\{\left|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\right| \geq \epsilon \text{ für unendlich viele } n \in \mathbb{N}\right\} \end{aligned} \tag{597}$$

Aus Formel (584) wissen wir:

$$\sum_{n\in\mathbb{N}} P\left[\left|\frac{1}{n}\sum_{k=1}^{n} 1_{A_k} - p\right| \geq \epsilon\right] < \infty. \tag{598}$$

Das 1. Borel-Cantelli-Lemma, Lemma 2.107, liefert hiermit:

$$P(B_\epsilon) = 0 \tag{599}$$

für alle $\epsilon > 0$. Da die Menge $\mathbb{Q}^+$ aller positiven rationalen Zahlen abzählbar ist, folgt:

$$P\left(\bigcup_{\epsilon\in\mathbb{Q}^+} B_\epsilon\right) = 0, \tag{600}$$

also

$$P\text{-fast sicher: } \forall \epsilon > 0, \epsilon \in \mathbb{Q}\ \exists m \in \mathbb{N}\ \forall n \geq m : \left| \frac{1}{n} \sum_{k=1}^{n} 1_{A_k} - p \right| < \epsilon, \tag{601}$$

anders gesagt:

$$\frac{1}{n} \sum_{k=1}^{n} 1_{A_k} \overset{n\to\infty}{\longrightarrow} p \ P\text{-fast sicher.} \tag{602}$$

□

Wir besprechen jetzt eine Verallgemeinerung des starken Gesetzes der großen Zahlen auf Zufallsvariablen:

Satz 2.109 (Starkes Gesetz der großen Zahlen für Zufallsvariablen mit exponentiellem Abfall)
Es seien $(X_n)_{n\in\mathbb{N}}$ i.i.d. Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Es existiere $\alpha > 0$ mit

$$E_P[e^{\alpha|X_1|}] < \infty. \tag{603}$$

Dann gilt

Wichtig!

$$\frac{1}{n} \sum_{k=1}^{n} X_k \overset{n\to\infty}{\longrightarrow} E_P[X_1] \quad P\text{-fast sicher,} \tag{604}$$

anders gesagt

$$P\left(\left\{ \omega \in \Omega : \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) \overset{n\to\infty}{\longrightarrow} E_P[X_1] \right\} \right) = 1. \tag{605}$$

Bemerkung: Die Voraussetzung (603) kann zu

$$E_P[X_1] < \infty, \tag{606}$$

also zu $X_1 \in \mathcal{L}_1(\Omega, \mathcal{A}, P)$ abgeschwächt werden. Diese Variante des starken Gesetzes der großen Zahlen wird allerdings erst in der Vorlesung "Wahrscheinlichkeitstheorie" diskutiert.

Beweis des Satzes 2.109: Die Voraussetzung (603) impliziert für alle $s \in [-\alpha, \alpha]$:

$$L_{X_1}(s) = E_P[e^{sX_1}] \leq E_P[e^{\alpha|X_1|}] < \infty \quad \text{wegen } e^{sX_1} \leq e^{\alpha|X_1|}. \tag{607}$$

Also ist die Laplacetransformierte L_{X_1} in einer offenen Umgebung von 0 endlich und daher dort beliebig oft differenzierbar mit

$$L'_{X_1}(0) = E_P[X_1] =: \mu. \tag{608}$$

Für $0 \leq s \leq \alpha$ und $\epsilon > 0$ schätzen wir ab:

$$\begin{aligned}
&P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \mu + \epsilon\right] \\
=&P\left[\sum_{k=1}^{n} X_k \geq n(\mu+\epsilon)\right] \\
\leq& e^{-n(\mu+\epsilon)s} E_P\left[\exp\left(s\sum_{k=1}^{n} X_k\right)\right] \qquad \text{(mit der exp. Tschebyscheff-Ungl. (522))} \\
=& e^{-n(\mu+\epsilon)s}\prod_{k=1}^{n} E_P\left[e^{sX_k}\right] = e^{-n(\mu+\epsilon)s} L_{X_1}(s)^n \qquad \text{(da } (X_k)_k \text{ i.i.d.)} \\
=& \left(e^{-(\mu+\epsilon)s} L_{X_1}(s)\right)^n, \qquad (609)
\end{aligned}$$

wobei wir die Faltungseigenschaft (499) der Laplacetransformierten verwendet haben. Nun gilt:

$$\begin{aligned}
&\frac{d}{ds} e^{-(\mu+\epsilon)s} L_{X_1}(s)\Big|_{s=0} \\
=& \left(-(\mu+\epsilon)e^{-(\mu+\epsilon)s} L_{X_1}(s) + e^{-(\mu+\epsilon)s} L'_{X_1}(s)\right)\Big|_{s=0} \\
=& -(\mu+\epsilon) + \mu \qquad \text{(wegen } L_{X_1}(0) = 1,\ L'_{X_1}(0) = \mu) \\
=& -\epsilon \qquad (610)
\end{aligned}$$

und

$$e^{-(\mu+\epsilon)s} L_{X_1}(s)\big|_{s=0} = L_{X_1}(0) = 1. \qquad (611)$$

Unter Verwendung der Definition der Ableitung

$$0 > \frac{d}{ds} e^{-(\mu+\epsilon)s} L_{X_1}(s)\Big|_{s=0} = \lim_{s\to 0} \frac{e^{-(\mu+\epsilon)s} L_{X_1}(s) - 1}{s} \qquad (612)$$

schließen wir für alle $s \in]0, \alpha[$, die nahe genug bei 0 sind, die Ungleichungskette

$$0 \leq e^{-(\mu+\epsilon)s} L_{X_1}(s) < 1. \qquad (613)$$

Wir fixieren solch ein s und setzen hierfür

$$\xi := e^{-(\mu+\epsilon)s} L_{X_1}(s) \in]0, 1[. \qquad (614)$$

Damit konvergiert die folgende geometrische Reihe:

$$\sum_{n\in\mathbb{N}} \xi^n < \infty. \qquad (615)$$

Damit ist insgesamt gezeigt:

$$\sum_{n\in\mathbb{N}} P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \mu+\epsilon\right] \leq \sum_{n\in\mathbb{N}} \xi^n < \infty. \tag{616}$$

Mit dem ersten Borel-Cantelli-Lemma, Lemma 2.107, schließen wir

$$P\text{-fast sicher: } \frac{1}{n}\sum_{k=1}^{n} X_k < \mu+\epsilon \text{ für alle bis auf endlich viele } n\in\mathbb{N}. \tag{617}$$

Ganz analog, z.B. indem wir $-X_k$ statt X_k betrachten, folgt

$$P\text{-fast sicher: } \frac{1}{n}\sum_{k=1}^{n} X_k > \mu-\epsilon \text{ für alle bis auf endlich viele } n\in\mathbb{N}. \tag{618}$$

Weil das für alle $\epsilon > 0$ und daher insbesondere für alle (abzählbar vielen) $\epsilon \in \mathbb{Q}^+$ gilt, folgt:

$$P\text{-fast sicher: } \frac{1}{n}\sum_{k=1}^{n} X_k \overset{n\to\infty}{\longrightarrow} \mu. \tag{619}$$

□

2.14 Der zentrale Grenzwertsatz

In Anwendungen sind Messfehler oder zufällige Schwankungen von Messgrößen oft ungefähr normalverteilt. In diesem Abschnitt besprechen wir den mathematischen Grund für diese "Universalität" der Normalverteilung.
Wir betrachten zunächst eine Folge $(X_n)_{n\in\mathbb{N}}$ von binomialverteilten Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, also $\mathcal{L}_P(X_n) = \text{binomial}(n,p)$ mit einem festen Parameter $0 < p < 1$. Es gilt also

$$E_P[X_n] = np, \tag{620}$$
$$\text{Var}_P(X_n) = np(1-p). \tag{621}$$

Wir betrachten die Dichte

$$f_{n,p} : \mathbb{R} \to \mathbb{R}, \quad f_{n,p}(x) = \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{1}{2}\frac{(x-np)^2}{np(1-p)}\right) \tag{622}$$

der Normalverteilung mit dem gleichen Erwartungswert np und der gleichen Varianz $np(1-p)$. Für große n liegt dann die Zähldichte der Binomialverteilung im folgenden Sinn nahe an der Dichte der Normalverteilung:

Satz 2.110 (Satz von de Moivre-Laplace) *Für alle $M > 0$ gilt mit der Abkürzung*

$$\mathcal{M}_n(M) := \left\{ k \in \mathbb{N} : \left| \frac{k - np}{\sqrt{np(1-p)}} \right| \leq M \right\} \tag{623}$$

Wichtig!

der folgende Vergleich der Binomial-Zähldichte mit der Dichte einer Normalverteilung:

$$\max_{k \in \mathcal{M}_n(M)} \left| \frac{P[X_n = k]}{f_{n,p}(k)} - 1 \right| \overset{n\to\infty}{\longrightarrow} 0 \tag{624}$$

Anschaulich gesprochen bedeutet das: Für große n liegt $P[X_n = k]$ nahe bei $f_{n,p}(k)$, solange die "standardisierte Größe"

$$\frac{k - E_P[X_n]}{\sigma_P(X_n)} \tag{625}$$

im kompakten Intervall $[-M, M]$ liegt.

Der **Beweis** des Satzes 2.110 beruht auf der Stirlingformel, also der folgenden Näherungsformel für $m!$ für große $m \in \mathbb{N}$:

$$\frac{m!}{\sqrt{2\pi} m^{m+\frac{1}{2}} e^{-m}} \overset{m\to\infty}{\longrightarrow} 1. \tag{626}$$

Wir verwenden die folgende quantitative Version davon; der Beweis davon gehört zur Analysis:
Für alle $m \in \mathbb{N}$ existiert ein $\theta_m \in \mathbb{R}$ mit

$$0 < \theta_m < \frac{1}{12m} \tag{627}$$

und

$$m! = \sqrt{2\pi} m^{m+\frac{1}{2}} e^{-m} e^{\theta_m}, \tag{628}$$

anders gesagt:

$$0 < \theta_m = \log(m!) - \left[\log\sqrt{2\pi} + (m + \frac{1}{2}) \log m - m \right] < \frac{1}{12m} \tag{629}$$

Wir erhalten die folgende Näherungsformel für den Binomialkoeffizienten $\binom{n}{k}$ für $n \in \mathbb{N}$ und $k \in [n-1]$:

$$\begin{aligned}
&\log \binom{n}{k} = \log(n!) - \log(k!) - \log((n-k)!) \\
&= -\log\sqrt{2\pi} + \left(n + \frac{1}{2}\right) \log n - \left(k + \frac{1}{2}\right) \log k - \left(n - k + \frac{1}{2}\right) \log(n-k) + \theta_n - \theta_k - \theta_{n-k} \\
&= -\log\sqrt{2\pi} + \frac{1}{2} \log \frac{n}{k(n-k)} + n \log n - k \log k - (n-k) \log(n-k) + \theta_n - \theta_k - \theta_{n-k}.
\end{aligned} \tag{630}$$

Wir kürzen ab:

$$q := 1 - p. \tag{631}$$

Es folgt:

$$\begin{aligned}
&\log P[X_n = k] = \log \left[\binom{n}{k} p^k q^{n-k} \right] \\
&= \log \binom{n}{k} + k \log p + (n-k) \log q \\
&= -\log \sqrt{2\pi} + \frac{1}{2} \log \frac{n}{k(n-k)} - k \log \frac{k}{np} - (n-k) \log \frac{n-k}{nq} + \theta_n - \theta_k - \theta_{n-k}
\end{aligned} \tag{632}$$

Wir analysieren die Terme einzeln: Mit der Definition (623) von $\mathcal{M}_n(M)$ schließen wir

$$\max_{k \in \mathcal{M}_n(M)} \left| \frac{k}{np} - 1 \right| = \max_{k \in \mathcal{M}_n(M)} \left| \frac{k - np}{np} \right| \leq M \frac{\sqrt{npq}}{np} = M \sqrt{\frac{q}{p}} \frac{1}{\sqrt{n}} \overset{n \to \infty}{\longrightarrow} 0, \tag{633}$$

$$\max_{k \in \mathcal{M}_n(M)} \left| \frac{n-k}{nq} - 1 \right| = \max_{k \in \mathcal{M}_n(M)} \left| \frac{k - np}{nq} \right| \leq M \sqrt{\frac{p}{q}} \frac{1}{\sqrt{n}} \overset{n \to \infty}{\longrightarrow} 0. \tag{634}$$

Insbesondere gilt

$$\exists n_0 = n_0(M, p) \; \forall n \geq n_0 \; \forall k \in \mathcal{M}_n(M) : \; \frac{k}{np} \geq \frac{1}{2}, \; \frac{n-k}{nq} \geq \frac{1}{2}. \tag{635}$$

Für $x > 0$ setzen wir

$$f(x) := x \log x \tag{636}$$

und schreiben

$$k \log \frac{k}{np} = npf\left(\frac{k}{np} \right), \tag{637}$$

$$(n-k) \log \frac{n-k}{nq} = nqf\left(\frac{n-k}{nq} \right). \tag{638}$$

Um eine Näherung dafür zu gewinnen, entwickeln wir die Funktion f um $x_0 = 1$ mit der Taylorformel: Für $x > 0$ gilt

$$f'(x) = 1 + \log x, \quad f''(x) = \frac{1}{x}, \quad f'''(x) = -\frac{1}{x^2}, \tag{639}$$

$$f(1) = 0, \quad f'(1) = 1, \quad f''(1) = 1, \tag{640}$$

also

$$f(x) = f(1) + f'(1)(x-1) + \frac{f''(1)}{2}(x-1)^2 + r(x) = (x-1) + \frac{1}{2}(x-1)^2 + r(x) \tag{641}$$

mit einem Restterm $r(x)$, der die Abschätzung

$$|r(x)| \leq \frac{|x-1|^3}{3!} \max\{|f'''(\xi)| \mid \xi \text{ zwischen 1 und } x\} = \frac{|x-1|^3}{6} \max\{1, x^{-2}\} \tag{642}$$

erfüllt. Insbesondere gilt

$$|r(x)| \leq |x-1|^3 \text{ für } x \geq \frac{1}{2}. \tag{643}$$

Es folgt

$$f\left(\frac{k}{np}\right) = \left(\frac{k}{np} - 1\right) + \frac{1}{2}\left(\frac{k}{np} - 1\right)^2 + r\left(\frac{k}{np}\right), \tag{644}$$

$$npf\left(\frac{k}{np}\right) = (k-np) + \frac{1}{2}q\frac{(k-np)^2}{npq} + npr\left(\frac{k}{np}\right), \tag{645}$$

wobei für alle großen $n \in \mathbb{N}$, genauer gesagt für $n \geq n_0(M,p)$ mit der Schranke aus Formel (635), die folgende Resttermabschätzung gilt:

$$\begin{aligned}
\max_{k \in \mathcal{M}_n(M)} \left| npr\left(\frac{k}{np}\right)\right| &\leq np \max_{k \in \mathcal{M}_n(M)} \left|\frac{k}{np} - 1\right|^3 \\
&\leq np\left(M\sqrt{\frac{q}{p}}\frac{1}{\sqrt{n}}\right)^3 \quad \text{(mit Schranke (633))} \\
&= M^3 q^{\frac{3}{2}} p^{-\frac{1}{2}} n^{-\frac{1}{2}} \overset{n\to\infty}{\longrightarrow} 0
\end{aligned} \tag{646}$$

Ebenso, mit vertauschten Rollen $k \leftrightarrow n-k$ und $p \leftrightarrow q$, erhalten wir die Taylorentwicklung

$$f\left(\frac{n-k}{nq}\right) = \left(\frac{n-k}{nq} - 1\right) + \frac{1}{2}\left(\frac{n-k}{nq} - 1\right)^2 + r\left(\frac{n-k}{nq}\right), \tag{647}$$

$$nqf\left(\frac{n-k}{nq}\right) = (np-k) + \frac{1}{2}p\frac{(k-np)^2}{npq} + nqr\left(\frac{n-k}{nq}\right), \tag{648}$$

wobei wir $(n-k)-nq = np-k$ verwendet haben, mit der folgenden Resttermabschätzung für alle großen $n \in \mathbb{N}$:

$$\begin{aligned}
\max_{k \in \mathcal{M}_n(M)} \left| nqr\left(\frac{n-k}{nq}\right)\right| &\leq nq \max_{k \in \mathcal{M}_n(M)} \left|\frac{n-k}{nq} - 1\right|^3 \\
&\leq nq\left(M\sqrt{\frac{p}{q}}\frac{1}{\sqrt{n}}\right)^3 \quad \text{(mit Schranke (634))} \\
&= M^3 p^{\frac{3}{2}} q^{-\frac{1}{2}} n^{-\frac{1}{2}} \overset{n\to\infty}{\longrightarrow} 0.
\end{aligned} \tag{649}$$

Zusammen erhalten wir

$$\begin{aligned}
npf\left(\frac{k}{np}\right) + nqf\left(\frac{n-k}{nq}\right) &= \frac{1}{2}(p+q)\frac{(k-np)^2}{npq} + r_2(n,k,p) \\
&= \frac{1}{2}\frac{(k-np)^2}{npq} + r_2(n,k,p)
\end{aligned} \tag{650}$$

mit dem Restterm

$$r_2(n,k,p) := npr\left(\frac{k}{np}\right) + nqr\left(\frac{n-k}{nq}\right), \tag{651}$$

der die Schranke

$$\max_{k\in\mathcal{M}_n(M)} |r_2(n,k,p)| \leq M^3(q^{\frac{3}{2}}p^{-\frac{1}{2}} + p^{\frac{3}{2}}q^{-\frac{1}{2}})n^{-\frac{1}{2}} \overset{n\to\infty}{\longrightarrow} 0 \tag{652}$$

erfüllt.
Betrachten wir nun die übrigen Terme in der Formel (632):

$$\begin{aligned}\log\frac{n}{k(n-k)} &= \log\frac{1}{npq} - \log\frac{k}{np} - \log\frac{n-k}{nq} \\ &=: \log\frac{1}{npq} + r_3(n,k,p)\end{aligned} \tag{653}$$

wobei der Restterm $r_3(n,k,p)$ die folgende Schranke erfüllt:

$$\max_{k\in\mathcal{M}_n(M)} |r_3(n,k,p)| \leq \max_{k\in\mathcal{M}_n(M)} \left[\left|\log\frac{k}{np}\right| + \left|\log\frac{n-k}{nq}\right|\right] \overset{n\to\infty}{\longrightarrow} 0. \tag{654}$$

Im letzten Schritt haben wir nochmal die Schranken (633) und (634) verwendet.
Schließlich gilt

$$\min_{k\in\mathcal{M}_n(M)} k \geq np - M\sqrt{npq} \overset{n\to\infty}{\longrightarrow} \infty, \tag{655}$$

$$\min_{k\in\mathcal{M}_n(M)} (n-k) \geq nq - M\sqrt{npq} \overset{n\to\infty}{\longrightarrow} \infty, \tag{656}$$

also mit der Fehlerschätzung (627) zur Stirlingformel:

$$\max_{k\in\mathcal{M}_n(M)} \theta_k \leq \max_{k\in\mathcal{M}_n(M)} \frac{1}{12k} \overset{n\to\infty}{\longrightarrow} 0, \tag{657}$$

$$\max_{k\in\mathcal{M}_n(M)} \theta_{n-k} \leq \max_{k\in\mathcal{M}_n(M)} \frac{1}{12(n-k)} \overset{n\to\infty}{\longrightarrow} 0. \tag{658}$$

Setzen wir diese Abschätzungen zusammen mit (650) und (653) in Formel (632) ein:

$$\begin{aligned}\log P[X_n = k] &= -\log\sqrt{2\pi} + \frac{1}{2}\log\frac{1}{npq} - \frac{1}{2}\frac{(k-np)^2}{npq} + r_4(n,k,p) \\ &= \log f_{n,p}(k) + r_4(n,k,p)\end{aligned} \tag{659}$$

mit dem Restterm

$$r_4(n,k,p) = \frac{1}{2}r_3(n,k,p) - r_2(n,k,p) + \theta_n - \theta_k - \theta_{n-k}, \tag{660}$$

der die Fehlerschranke

$$\max_{k\in\mathcal{M}_n(M)} |r_4(n,k,p)| \stackrel{n\to\infty}{\longrightarrow} 0 \tag{661}$$

erfüllt. Es folgt:

$$\max_{k\in\mathcal{M}_n(M)} \left|\log \frac{P[X_n=k]}{f_{n,p}(k)}\right| = \max_{k\in\mathcal{M}_n(M)} |r_4(n,k,p)| \stackrel{n\to\infty}{\longrightarrow} 0, \tag{662}$$

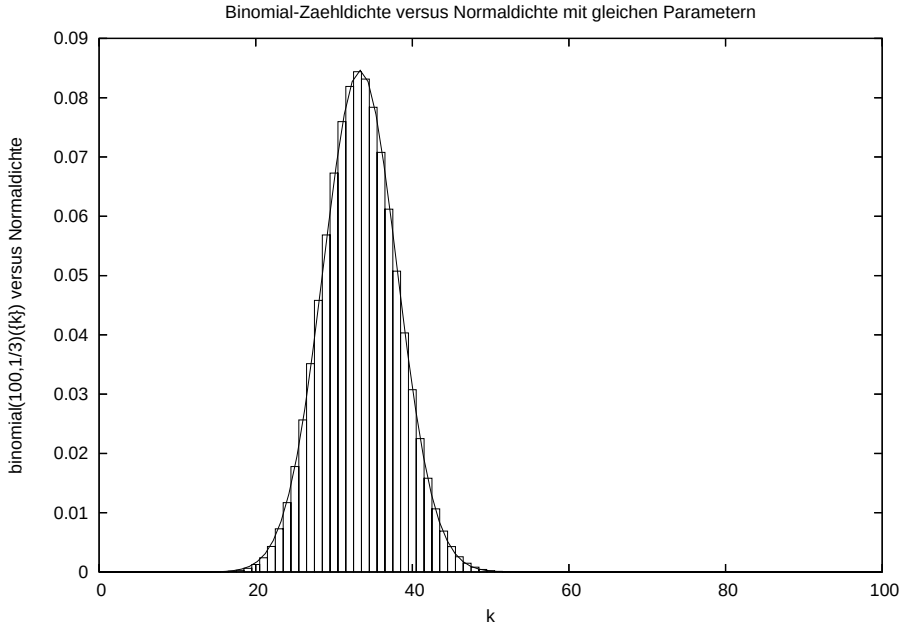
was wir wegen der Stetigkeit der Exponentialfunktion und $e^0 = 1$ auch in der Form

$$\max_{k\in\mathcal{M}_n(M)} \left|\frac{P[X_n=k]}{f_{n,p}(k)} - 1\right| = \max_{k\in\mathcal{M}_n(M)} |r_4(n,k,p)| \stackrel{n\to\infty}{\longrightarrow} 0 \tag{663}$$

schreiben können.[33]

□

Die folgende Graphik illustriert, dass die Binomialdichte zu den Parametern $n = 100$ und $p = 0,3$ in der Tat nahe an der Dichte der Normalverteilung mit der gleichen Erwartung np und der gleichen Varianz $np(1-p)$ liegt:

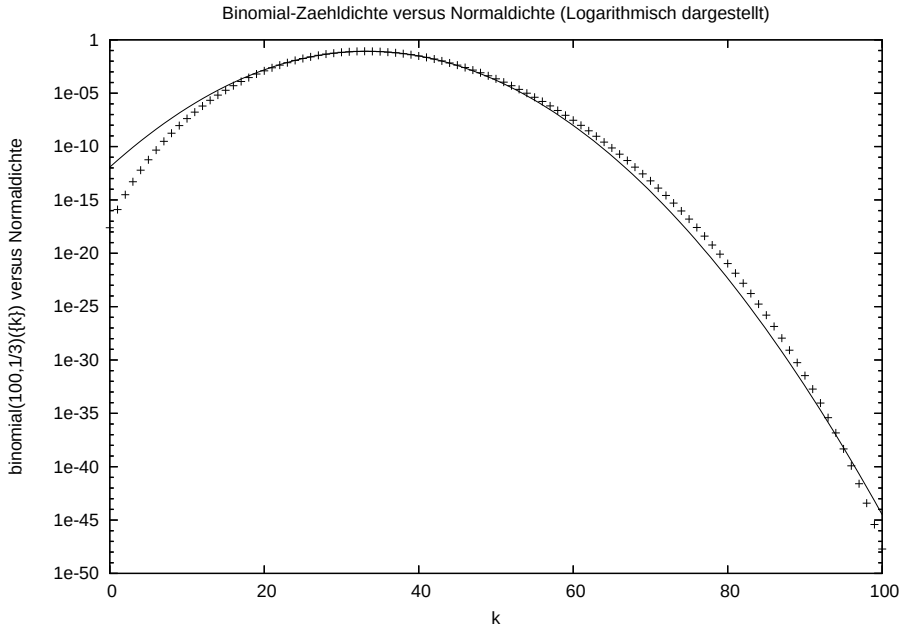


[33]Ein Rückblick auf den Beweis zeigt, dass wir den Satz von de Moivre-Laplace noch etwas verschärfen können, indem wir die feste Schranke M durch eine n-abhängige Schranke M_n ersetzen, die für $n \to \infty$ leicht ansteigt: Genauer gesagt brauchen wir nur die Voraussetzung

$$\frac{M_n^3}{\sqrt{n}} \stackrel{n\to\infty}{\longrightarrow} 0, \tag{664}$$

mit anderen Worten $M_n = o(n^{-\frac{2}{3}})$ für $n \to \infty$, siehe Formeln (633), (634) und (652).

Wozu die Beschränkung im Satz von de Moivre-Laplace auf einen kompakten Bereich $[-M, M]$ nötig ist, kann man jedoch erst graphisch sehen, wenn die gleichen Daten logarithmisch dargestellt werden, weil "weg vom Zentrum" np die Zahlenwerte für $P[X_n = k]$ und $f_{n,p}(k)$ beide schon sehr nahe bei 0 liegen:



Weg vom Zentrum sieht man hier deutliche relative Abweichungen, die aus den dort größern Fehlern der Taylorapproximation stammen, aber in absoluten Wahrscheinlichkeiten gesehen dennoch klein sind.[34]

Korollar 2.111 (Zentraler Grenzwertsatz für die Binomialverteilung) *Gegeben* $0 < p < 1$*, seien* X_n*,* $n \in \mathbb{N}$*, binomialverteilte Zufallsvariablen mit den Parametern* n *und* p *und*

$$Z_n := \frac{X_n - E_P[X_n]}{\sigma_P(X_n)} = \frac{X_n - np}{\sqrt{np(1-p)}} \tag{665}$$

ihre Standardisierung. Weiter sei Z *eine standardnormalverteilte Zufallsvariable. Dann gilt für alle Zahlen* $a, b \in \mathbb{R}$ *mit* $a < b$*:*

$$P[a \leq Z_n \leq b] \overset{n\to\infty}{\longrightarrow} P[a \leq Z \leq b] \tag{666}$$

[34]Vergröbert gesagt: Der "zentrale Grenzwertsatz" gilt nur "im Zentrum" (Fluktuationen auf der Skala $k - np \approx O(\sqrt{n})$), nicht für "große Abweichungen" (Fluktuationen auf der Skala $k - np \approx O(n)$). Der zentrale Grenzwertsatz hat also seinen Namen von seinem Gültigkeitsbereich "im Zentrum der Verteilung", im Gegensatz zu "großen Abweichungen". Natürlich kann man das Wort "zentral" auch ganz anders lesen: Der zentrale Grenzwertsatz ist einer der zentralen Sätze der Stochastik!

In der letzten Formel kann man das $\leq$-Symbol auch durch das $<$-Symbol ersetzen.
Beweis: Es sei

$$M = \max\{|a|, |b|\}, \tag{667}$$

$$K_n = \left\{ k \in \mathbb{N} : \ a \leq \frac{k - np}{\sqrt{np(1-p)}} \leq b \right\} \subseteq \mathcal{M}_n(M), \tag{668}$$

wobei die Menge $\mathcal{M}_n$ in Formel (623) definiert wurde. Wir setzen

$$\underline{c}_n := \min_{k \in K_n} \frac{P[X_n = k]}{f_{n,p}(k)}, \tag{669}$$

$$\overline{c}_n := \max_{k \in K_n} \frac{P[X_n = k]}{f_{n,p}(k)}. \tag{670}$$

Insbesondere folgt aus dem Satz 2.110 von de Moivre-Laplace:

$$\underline{c}_n \overset{n\to\infty}{\longrightarrow} 1, \tag{671}$$

$$\overline{c}_n \overset{n\to\infty}{\longrightarrow} 1. \tag{672}$$

Nun gilt:

$$P[a \leq Z_n \leq b] = \sum_{k \in K_n} P[X_n = k] \left\{ \begin{array}{ll} \leq & \overline{c}_n \sum_{k \in K_n} f_{n,p}(k), \\ \geq & \underline{c}_n \sum_{k \in K_n} f_{n,p}(k) \end{array} \right. . \tag{673}$$

Wir setzen

$$G_n = \left\{ \frac{k - np}{\sqrt{np(1-p)}} : \ k \in K_n \right\} \subseteq [a, b]. \tag{674}$$

Für alle n groß genug ist G_n eine Zerlegung des Intervalls $[a, b]$ in Stücke der Länge $(np(1-p))^{-\frac{1}{2}}$ mit zwei eventuell kürzeren Randstücken. Es sei

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{675}$$

die Dichte der Standardnormalverteilung. Dann gilt

$$f_{n,p}(k) = \frac{1}{\sqrt{np(1-p)}} \phi\left(\frac{k - np}{\sqrt{np(1-p)}} \right), \quad (k \in K_n) \tag{676}$$

und daher wegen der Konvergenz von Riemannsummen gegen ein Integral:

$$\sum_{k \in K_n} f_{n,p}(k) = \frac{1}{\sqrt{np(1-p)}} \sum_{x \in G_n} \phi(x) \overset{n\to\infty}{\longrightarrow} \int_a^b \phi(x)\, dx = P[a \leq Z \leq b]. \tag{677}$$

Zusammen mit den Limiten (671), (672) und der Abschätzung (673) folgt die Behauptung.

□

Wir verallgemeinern nun dieses Korollar wesentlich auf allgemeinere Verteilungen:

Satz 2.112 (Zentraler Grenzwertsatz für i.i.d. Zufallsvariablen in $\mathcal{L}^2$) *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von i.i.d. Zufallsvariablen in $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ mit positiver Varianz $\mathrm{Var}_P(X_1) > 0$,*

$$S_n := \sum_{k=1}^{n} X_k, \tag{678}$$

$$Z_n := \frac{S_n - E_P[S_n]}{\sigma_P(S_n)} = \frac{S_n - nE_P[X_1]}{\sqrt{n}\sigma_P(X_1)} \tag{679}$$

Wichtig!

Weiter sei Z eine standardnormalverteilte Zufallsvariable. Dann gilt für alle Intervalle $I \subseteq \mathbb{R}$ (gleichgültig, ob endlich oder unendlich, offen, abgeschlossen oder halboffen):

$$P[Z_n \in I] \overset{n\to\infty}{\longrightarrow} P[Z \in I]. \tag{680}$$

Der Zentrale Grenzwertsatz macht das häufige Auftreten der Normalverteilung in Anwendungen für zufällige Schwankungen verständlich, wenn sich eine solche Schwankung aus einer Summe von sehr vielen unabhängigen gleichartigen Beiträgen zusammensetzt.
Beweisidee: Die Idee des hier vorgestellten[35] Beweises des Zentralen Grenzwertsatzes beruht darauf, dass die Summanden X_k in der Summe S_n sukzessive durch unabhängige normalverteilte Summanden Y_k mit der gleichen Erwartung und der gleichen Varianz ersetzt werden. Die dabei entstehenden Fehler werden mit einer Taylorentwicklung kontrolliert, wobei die ersten Terme in der Taylorentwicklung sich wegen $E[X_k] = E[Y_k]$ und $\mathrm{Var}(X_k) = \mathrm{Var}(Y_k)$ nicht ändern. Sind schließlich alle Summanden X_k durch normalverteilte Summanden Y_k ersetzt so verwenden wir, dass die Summe der Y_k wegen der Faltungseigenschaft der Normalverteilung, Satz 2.58, wieder normalverteilt ist.

Nun zu den Einzelheiten des Beweises: Wir führen den Satz 2.112 auf die folgende Variante zurück:

Satz 2.113 (Zentraler Grenzwertsatz – Version für C_b^3-Testfunktionen) *Es sei $f : \mathbb{R} \to \mathbb{R}$ eine dreimal stetig differenzierbare beschränkte Abbildung mit beschränkten Ableitungen bis zur 3. Stufe, in Zeichen: $f \in C_b^3(\mathbb{R})$. Dann gilt mit den Bezeichnungen und Voraussetzungen von Satz 2.112:*

$$E_P[f(Z_n)] \overset{n\to\infty}{\longrightarrow} E_P[f(Z)]. \tag{681}$$

Beweis: O.B.d.A. gelte $E_P[X_n] = 0$ und $\mathrm{Var}_P(X_n) = 1$; andernfalls ersetzen wir nämlich

[35] Es gibt auch mehrere andere Beweise des Zentralen Grenzwertsatzes, die auf ganz anderen Ideen beruhen, zum Beispiel der Fouriertransformation oder der sogenannten Steinschen Methode. Diese alternativen Beweise besprechen wir hier nicht.

X_n durch

$$\tilde{X}_n := \frac{X_n - E_P[X_n]}{\sigma_P(X_n)} \tag{682}$$

und damit S_n durch

$$\tilde{S}_n := \sum_{k=1}^{n} X_k, \tag{683}$$

bei ungeänderter Standardisierung:

$$Z_n = \frac{S_n - E_P[S_n]}{\sigma_P(S_n)} = \frac{\tilde{S}_n - E_P[\tilde{S}_n]}{\sigma_P(\tilde{S}_n)}. \tag{684}$$

Wir dürfen auch o.B.d.A. annehmen, dass auf $(\Omega, \mathcal{A}, P)$ eine i.i.d. Folge $(Y_n)_{n\in\mathbb{N}}$ von standardnormalverteilten Zufallsvariablen existiert, unabhängig von den $(X_n)_{n\in\mathbb{N}}$.[36] Dann gilt:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k. \tag{685}$$

Außerdem ist

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k \tag{686}$$

standardnormalverteilt, denn die Summe $\sum_{k=1}^{n} Y_k$ der i.i.d. standardnormalverteilten Zufallsvariable $Y_1, \ldots, Y_n$ ist $\mathrm{N}(0, n)$-verteilt aufgrund der Faltungseigenschaft der Normalverteilung, Satz 2.58, und die Reskalierung mit $\frac{1}{\sqrt{n}}$ reskaliert die Varianz um den Faktor $\frac{1}{n}$.
Wir müssen also zeigen:

$$E_P\left[f\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k\right)\right] - E_P\left[f\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k\right)\right] \overset{n\to\infty}{\longrightarrow} 0 \tag{687}$$

Hierzu zerlegen wir diese Differenz in eine Teleskopsumme, indem wir sukzessive Summanden X_k durch Y_k ersetzen:

$$\begin{aligned} &E_P\left[f\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k\right)\right] - E_P\left[f\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k\right)\right] \\ &= \sum_{l=1}^{n} E_P\left[f\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{l-1} Y_k + \frac{1}{\sqrt{n}} \sum_{k=l}^{n} X_k\right) - f\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{l} Y_k + \frac{1}{\sqrt{n}} \sum_{k=l+1}^{n} X_k\right)\right]. \end{aligned} \tag{688}$$

[36] Wenn nötig, ersetzen wir $(\Omega, \mathcal{A}, P)$ durch einen Produktraum, wobei die X_n nur von der ersten Komponente abhängen und die Y_n nur von der zweiten Komponente.

Mit der Abkürzung

$$T_{n,l} := \frac{1}{\sqrt{n}} \sum_{k=1}^{l-1} Y_k + \frac{1}{\sqrt{n}} \sum_{k=l+1}^{n} X_k \tag{689}$$

für die Summe aller Summanden, die in der l-ten Stufe nicht ausgetauscht werden, können wir den l-ten Summanden auf der rechten Seite in Formel (688) auch so schreiben:

$$E_P \left[f\left(T_{n,l} + \frac{X_l}{\sqrt{n}} \right) - f\left(T_{n,l} + \frac{Y_l}{\sqrt{n}} \right) \right]. \tag{690}$$

Wir entwickeln die Funktion f mit der Taylorformel: Für alle $t, x \in \mathbb{R}$ existierten $\theta_2, \theta_3 \in [0,1]$, so dass gilt:

$$f(t+x) = f(t) + xf'(t) + \frac{x^2}{2} f''(t) + \frac{x^3}{6} f'''(t + \theta_3 x), \tag{691}$$

$$f(t+x) = f(t) + xf'(t) + \frac{x^2}{2} f''(t) + \frac{x^2}{2} (f''(t + \theta_2 x) - f''(t)), \tag{692}$$

also

$$f(t+x) = f(t) + xf'(t) + \frac{x^2}{2} f''(t) + r(t,x) \tag{693}$$

mit einem Restterm $r(t,x)$, der die folgende Schranke erfüllt:

$$|r(t,x)| \leq \min \left\{ x^2 \sup_{\mathbb{R}} |f''|, \frac{|x|^3}{6} \sup_{\mathbb{R}} |f'''| \right\} \leq c_f \min\{x^2, |x|^3\}, \tag{694}$$

wobei

$$c_f := \sup_{\mathbb{R}} |f''| + \frac{1}{6} \sup_{\mathbb{R}} |f'''| < \infty. \tag{695}$$

Es folgt:

$$\begin{aligned}
& E_P \left[f\left(T_{n,l} + \frac{X_l}{\sqrt{n}} \right) - f\left(T_{n,l} + \frac{X_l}{\sqrt{n}} \right) \right] \\
=& E_P \left[\frac{X_l}{\sqrt{n}} f'(T_{n,l}) \right] - E_P \left[\frac{Y_l}{\sqrt{n}} f'(T_{n,l}) \right] \\
& + \frac{1}{2} E_P \left[\frac{X_l^2}{n} f''(T_{n,l}) \right] - \frac{1}{2} E_P \left[\frac{Y_l^2}{n} f''(T_{n,l}) \right] \\
& + E_P \left[r\left(T_{n,l}, \frac{X_l}{\sqrt{n}} \right) - r\left(T_{n,l}, \frac{Y_l}{\sqrt{n}} \right) \right].
\end{aligned} \tag{696}$$

Weil $T_{n,l}$, X_l und Y_l unabhängig sind, schließen wir:

$$E_P\left[\frac{X_l}{\sqrt{n}}f'(T_{n,l})\right] = \underbrace{E_P\left[\frac{X_l}{\sqrt{n}}\right]}_{=0} E_P\left[f'(T_{n,l})\right] = 0, \tag{697}$$

$$E_P\left[\frac{Y_l}{\sqrt{n}}f'(T_{n,l})\right] = \underbrace{E_P\left[\frac{Y_l}{\sqrt{n}}\right]}_{=0} E_P\left[f'(T_{n,l})\right] = 0,$$

$$\begin{aligned} E_P\left[\frac{X_l^2}{n}f''(T_{n,l})\right] &= \frac{1}{n}\underbrace{E_P\left[X_l^2\right]}_{=1} E_P\left[f''(T_{n,l})\right] \\ &= \frac{1}{n}\underbrace{E_P\left[Y_l^2\right]}_{=1} E_P\left[f''(T_{n,l})\right] \\ &= E_P\left[\frac{Y_l^2}{n}f''(T_{n,l})\right] \end{aligned} \tag{698}$$

und daher

$$\begin{aligned} &\left|E_P\left[f\left(T_{n,l}+\frac{X_l}{\sqrt{n}}\right)-f\left(T_{n,l}+\frac{X_l}{\sqrt{n}}\right)\right]\right| \\ &= \left|E_P\left[r\left(T_{n,l},\frac{X_l}{\sqrt{n}}\right)-r\left(T_{n,l},\frac{Y_l}{\sqrt{n}}\right)\right]\right| \\ &\leq E_P\left[\left|r\left(T_{n,l},\frac{X_l}{\sqrt{n}}\right)\right|\right]+E_P\left[\left|r\left(T_{n,l},\frac{Y_l}{\sqrt{n}}\right)\right|\right] \\ &\leq c_f E_P\left[\min\left\{\left(\frac{X_l}{\sqrt{n}}\right)^2,\left|\frac{X_l}{\sqrt{n}}\right|^3\right\}\right]+c_f E_P\left[\min\left\{\left(\frac{Y_l}{\sqrt{n}}\right)^2,\left|\frac{Y_l}{\sqrt{n}}\right|^3\right\}\right] \quad \text{(mit Formel (694))} \\ &= \frac{c_f}{n}E_P\left[\min\left\{X_1^2,\frac{|X_1|^3}{\sqrt{n}}\right\}\right]+\frac{c_f}{n}E_P\left[\min\left\{Y_1^2,\frac{|Y_1|^3}{\sqrt{n}}\right\}\right], \end{aligned} \tag{699}$$

wobei wir im letzten Schritt $\mathcal{L}_P(X_l) = \mathcal{L}_P(X_1)$ und $\mathcal{L}_P(Y_l) = \mathcal{L}_P(Y_1)$ verwendet haben. Eingesetzt in die Teleskopsumme (688):

$$\begin{aligned} &\left|E_P\left[f\left(\frac{1}{\sqrt{n}}\sum_{k=1}^n X_k\right)\right]-E_P\left[f\left(\frac{1}{\sqrt{n}}\sum_{k=1}^n Y_k\right)\right]\right| \\ &\leq \underbrace{n\cdot\frac{c_f}{n}}_{=c_f}\left(E_P\left[\min\left\{X_1^2,\frac{|X_1|^3}{\sqrt{n}}\right\}\right]+E_P\left[\min\left\{Y_1^2,\frac{|Y_1|^3}{\sqrt{n}}\right\}\right],\right) \overset{n\to\infty}{\longrightarrow} 0, \end{aligned} \tag{700}$$

denn es gilt für alle $X \in \mathcal{L}^2(\Omega,\mathcal{A},P)$:

$$E_P\left[\min\left\{X^2,\frac{|X|^3}{\sqrt{n}}\right\}\right] \overset{n\to\infty}{\longrightarrow} 0. \tag{701}$$

Die Konvergenz (701) sieht man so: Das Argument $\min\{X^2, |X|^3 n^{-\frac{1}{2}}\}$ der Erwartung besitzt die integrierbare, von n nicht abhängende Majorante X^2, und es konvergiert punktweise für $n \to \infty$ gegen 0. Damit folgt die Behauptung (701) aus dem Satz von der dominierten Konvergenz.

□

Der Zentrale Grenzwertsatz 2.112 folgt nun aus der Implikation im folgenden Satz:

Satz 2.114 (Äquivalente Charakterisierungen der Konvergenz in Verteilung) *Es seien $(Z_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen und Z eine weitere Zufallsvariable. Es sei $F : \mathbb{R} \to [0,1]$ die Verteilungsfunktion von Z. Dann sind äquivalent:*

1. *Für jede beschränkte stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ gilt*
$$E[f(Z_n)] \overset{n\to\infty}{\longrightarrow} E[f(Z)]. \tag{702}$$
2. *Für jede Funktion $f \in C_b^3(\mathbb{R})$ gilt*
$$E[f(Z_n)] \overset{n\to\infty}{\longrightarrow} E[f(Z)]. \tag{703}$$
3. *Für jedes Intervall $I = [a,b],]a,b], [a,b[$ oder $]a,b[$, so dass a und b Stetigkeitspunkte[37] der Verteilungsfunktion F oder $\pm\infty$ sind, gilt:*
$$P[Z_n \in I] \overset{n\to\infty}{\longrightarrow} P[Z \in I]. \tag{704}$$

Bemerkung: In unserer Anwendung auf den Zentralen Grenzwertsatz ist die Zufallsvariable Z standardnormalverteilt. Ihre Verteilungsfunktion F ist stetig. Daher liefert die Einschränkung in dritten der drei äquivalenten Aussagen des Satzes 2.114 hier keine Bedingung.

Definition 2.115 (Konvergenz in Verteilung) *Gelten die drei äquivalenten Bedingungen des Satzes 2.114, so heißt die Folge Z_n* konvergent in Verteilung *gegen Z (oder auch gegen die Verteilung von Z), englisch "*convergent in distribution*", synonym: Z_n* konvergiert schwach *gegen Z, in Zeichen* **Wichtig!**

$$Z_n \xrightarrow[n\to\infty]{d} Z. \tag{705}$$

Wir beweisen hier nur den Teil 2.⇒3. des Satzes 2.114, den wir brauchen:

Es seien I ein Intervall wie in Aussage 3. und $\epsilon > 0$. Es bezeichne $\overline{I}$ den topologischen Abschluss von I und I° den offenen Kern von I. Nach der Voraussetzung an das Intervall I gibt es ein offenes Intervall $I_1 \supseteq \overline{I}$ und ein abgeschlossenes Intervall $I_2 \subseteq I^\circ$ mit

$$P[Z \in I_1] \le P[Z \in I] + \epsilon \quad \text{und} \quad P[Z \in I_2] \ge P[Z \in I] - \epsilon. \tag{706}$$

[37] Ein Punkt x heißt Stetigkeitspunkt einer Funktion f, wenn f bei x stetig ist.

Wir wählen zwei Funktionen $f_1, f_2 \in C_b^3(\mathbb{R})$ mit Werten im Einheitsintervall $[0,1]$ mit

$$1_{I_2} \leq f_2 \leq 1_I \leq f_1 \leq 1_{I_1}; \tag{707}$$

solche Funktionen existieren. Dann gilt wegen Voraussetzung 1.:

$$P[Z_n \in I] \leq E_P[f_1(Z_n)] \stackrel{n\to\infty}{\longrightarrow} E_P[f_1(Z)] \leq E_P[Z \in I_1] \leq P[Z \in I] + \epsilon, \tag{708}$$

$$P[Z_n \in I] \geq E_P[f_2(Z_n)] \stackrel{n\to\infty}{\longrightarrow} E_P[f_2(Z)] \geq E_P[Z \in I_2] \geq P[Z \in I] - \epsilon. \tag{709}$$

Weil $\epsilon > 0$ beliebig war, folgt die Behauptung

$$P[Z_n \in I] \stackrel{n\to\infty}{\longrightarrow} P[Z \in I]. \tag{704}$$

□

2.14.1 Anhang: Beweis der Stirlingformel

In diesem Anhang wird die Stirlingformel (627)–(628) bewiesen. Der Beweis ist rein analytisch; er verwendet keine Methoden aus der Stochastik und gehört daher nicht zum prüfungsrelevanten Stoff der Stochastik.

Lemma 2.116 (Asymptotik des mittleren Binomialkoeffizienten) *Es gilt:*

$$\boxed{\lim_{n\to\infty} \frac{\sqrt{\pi n}}{2^{2n}} \left(\begin{array}{c} 2n \\ n \end{array} \right) = 1} \tag{710}$$

Beweis: Für $n \in \mathbb{N}$ und $x \in \mathbb{R}$ gilt nach der binomischen Formel

$$\begin{aligned} 2^{2n} \cos^{2n} x = (2\cos x)^{2n} &= (e^{ix} + e^{-ix})^{2n} \\ &= \sum_{k=0}^{2n} \left(\begin{array}{c} 2n \\ k \end{array} \right) e^{ikx} e^{-i(2n-k)x} \\ &= \sum_{k=0}^{2n} \left(\begin{array}{c} 2n \\ k \end{array} \right) e^{2i(k-n)x}. \end{aligned}$$

Integrieren wir von $-\pi/2$ bis $\pi/2$:

$$2^n \int_{-\pi/2}^{\pi/2} \cos^{2n} x \, dx = \sum_{k=0}^{2n} \left(\begin{array}{c} 2n \\ k \end{array} \right) \int_{-\pi/2}^{\pi/2} e^{2i(k-n)x} \, dx$$

Das Integral im mittleren Summanden, $k = n$, lautet

$$\int_{-\pi/2}^{\pi/2} e^{2i(n-n)x} \, dx = \int_{-\pi/2}^{\pi/2} 1 \, dx = \pi.$$

Alle anderen Summanden, $k \neq n$, verschwinden, denn hier gilt

$$\int_{-\pi/2}^{\pi/2} e^{2i(k-n)x}\,dx = \left[\frac{e^{2i(k-n)x}}{2i(k-n)}\right]_{x=-\pi/2}^{\pi/2} = 0$$

wegen der π-Periodizität von $x \mapsto e^{2i(k-n)x}$. Damit ist gezeigt:

$$\int_{-\pi/2}^{\pi/2} \cos^{2n} x\,dx = 2^{-2n}\pi \left(\begin{array}{c} 2n \\ n \end{array}\right). \tag{711}$$

Wir werten das letzte Integral asymptotisch auch anders aus, indem wir es mit einem Gaußschen Integral vergleichen: Für $-\pi/2 < x < \pi/2$ und $n \in \mathbb{N}$ gilt:

$$\cos^{2n} x \leq e^{-nx^2}$$

Das sieht man so: Wegen der Symmetrie des Kosinus reicht es, $0 \leq x < \pi/2$ zu betrachten. Für diese x gilt $\cos x > 0$ und

$$\log\cos x = -\int_0^x \tan u\,du \leq -\int_0^x u\,du = -\frac{x^2}{2}$$

wegen

$$\frac{d}{dx}\log\cos x = -\frac{\sin x}{\cos x} = -\tan x$$

und $\tan u \geq u$ für $0 \leq u < \pi/2$. Es folgt

$$\cos^{2n} x = e^{2n\log\cos x} \leq e^{-2n\frac{x^2}{2}} = e^{-nx^2}. \tag{712}$$

Wir erhalten:

$$\int_{-\pi/2}^{\pi/2} \cos^{2n} x\,dx \leq \int_{-\pi/2}^{\pi/2} e^{-nx^2}\,dx \leq \int_{\mathbb{R}} e^{-nx^2}\,dx = \sqrt{\frac{\pi}{n}}$$

Zusammen ergibt das

$$2^{-2n}\pi \left(\begin{array}{c} 2n \\ n \end{array}\right) \leq \sqrt{\frac{\pi}{n}},$$

anders geschrieben:

$$\left(\begin{array}{c} 2n \\ n \end{array}\right) \leq \frac{2^{2n}}{\sqrt{\pi n}}.$$

Um zu sehen, dass diese Formel asymptotisch für $n \to \infty$ sogar scharf ist, betrachten wir die Substitution $u = \sqrt{n}x$ und erhalten

$$\sqrt{n}\int_{-\pi/2}^{\pi/2} \cos^{2n} x\,dx = \int_{\mathbb{R}} 1_{]-\sqrt{n}\pi/2,\sqrt{n}\pi/2[}(u)\cos^{2n}\frac{u}{\sqrt{n}}\,du, \tag{713}$$

wobei wir die Integralgrenzen nun mit einer Indikatorfunktion

$$1_A(u) := \begin{cases} 1 & \text{für } u \in A \\ 0 & \text{für } u \notin A \end{cases}$$

geschrieben haben. Der Integrand besitzt wegen der Abschätzung (712) eine integrierbare Majorante:

$$0 \leq 1_{]-\sqrt{n}\pi/2,\sqrt{n}\pi/2[}(u) \cos^{2n} \frac{u}{\sqrt{n}} \leq \exp\left(-n\left(\frac{u}{\sqrt{n}}\right)^2\right) = e^{-u^2}.$$

Er konvergiert auch punktweise gegen diese obere Schranke, also

$$\lim_{n\to\infty} 1_{]-\sqrt{n}\pi/2,\sqrt{n}\pi/2[}(u) \cos^{2n} \frac{u}{\sqrt{n}} = \exp\left(-n\left(\frac{u}{\sqrt{n}}\right)^2\right) = e^{-u^2}$$

für alle $u \in \mathbb{R}$. Um das zu sehen, verwenden wir die Taylorentwicklung

$$\log\cos x = -\frac{x^2}{2} + r(x)$$

mit einem Restterm $r(x)$ mit $\lim_{x\to 0} x^{-2} r(x) = 0$. Es folgt für $n > (2|u|/\pi)^2$:

$$\begin{aligned} 1_{]-\sqrt{n}\pi/2,\sqrt{n}\pi/2[}(u) \cos^{2n} \frac{u}{\sqrt{n}} &= e^{2n \log\cos \frac{u}{\sqrt{n}}} \\ = \exp\left(-2n \cdot \frac{1}{2}\left(\frac{u}{\sqrt{n}}\right)^2 + 2nr\left(\frac{u}{\sqrt{n}}\right)\right) &\overset{n\to\infty}{\longrightarrow} e^{-u^2}. \end{aligned}$$

Aus dem Satz von der dominierten Konvergenz folgt

$$\lim_{n\to\infty} \int_{\mathbb{R}} 1_{]-\sqrt{n}\pi/2,\sqrt{n}\pi/2[}(u) \cos^{2n} \frac{u}{\sqrt{n}}\, du = \int_{\mathbb{R}} e^{-u^2}\, du = \sqrt{\pi}$$

Fassen wir zusammen: Mit den Formeln (711) und (713) erhalten wir:

$$\begin{aligned} \frac{\sqrt{\pi n}}{2^{2n}} \binom{2n}{n} &= \sqrt{\frac{n}{\pi}} \int_{-\pi/2}^{\pi/2} \cos^{2n} x\, dx \\ &= \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} 1_{]-\sqrt{n}\pi/2,\sqrt{n}\pi/2[}(u) \cos^{2n} \frac{u}{\sqrt{n}}\, du \overset{n\to\infty}{\longrightarrow} 1. \end{aligned}$$

Die folgende Stirling-Formel liefert eine Näherungsformel für die Fakultätsfunktion:

Lemma 2.117 (Stirling-Formel) *Für alle $n \in \mathbb{N}$ gilt:*

$$\boxed{\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \leq n! \leq \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}}} \tag{714}$$

Insbesondere folgt:

$$\boxed{\frac{n!}{\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}} \overset{n\to\infty}{\longrightarrow} 1} \tag{715}$$

Beweis: Vereinfacht gesagt beruht der Beweis auf der Approximation der Summe

$$\log n! = \sum_{m=1}^{n} \log m$$

durch das Integral

$$\int_1^n \log x \, dx.$$

Dabei wird

$$\sum_{m=1}^{n} \log m - \frac{1}{2}(\log 1 + \log n) = \sum_{m=1}^{n-1} \frac{1}{2}[\log m + \log(m+1)]$$

als eine Summe von Trapezflächen interpretiert.
Daher verwenden wir als ein Hilfmittel die Trapezregel zur numerischen Integration mit verschiedenen Darstellungen des Restglieds.[38]
Ist $f : [0,1] \to \mathbb{R}$ glatt, so erhalten wir mit partieller Integration:

$$\begin{aligned} \int_0^1 f(x)\,dx &= [(x-\tfrac{1}{2})f(x)]_{x=0}^1 - \int_0^1 (x-\tfrac{1}{2})f'(x)\,dx \\ &= \tfrac{1}{2}[f(0)+f(1)] - \int_0^1 (x-\tfrac{1}{2})f'(x)\,dx. \end{aligned} \tag{716}$$

Nun gilt einerseits, nochmals mit partieller Integration:

$$\begin{aligned} \int_0^1 (x-\tfrac{1}{2})f'(x)\,dx &= [\tfrac{1}{2}(x^2-x)f'(x)]_{x=0}^1 - \int_0^1 \tfrac{1}{2}(x^2-x)f''(x)\,dx \\ &= \tfrac{1}{2}\int_0^1 x(1-x)f''(x)\,dx, \end{aligned}$$

also in (716) eingesetzt:

$$\int_0^1 f(x)\,dx = \tfrac{1}{2}[f(0)+f(1)] - \tfrac{1}{2}\int_0^1 x(1-x)f''(x)\,dx. \tag{717}$$

Ist f konkav, also $f'' \leq 0$, so eignet sich diese Restglieddarstellung gut für eine untere Schranke des Integrals, da die Gewichtsfunktion $x(1-x)$ nichtnegative Werte annimmt. Für obere Schranken ist sie in unserem Fall weniger gut geeignet. Wir können aber die Integrationskonstante[39] bei der zweiten partiellen Integration auch alternativ so wählen,

[38] Man kann die folgende Rechnung auch als eine Herleitung der ersten Instanzen der Euler-McLaurinschen Summenformel auffassen.

[39] Die Integrationskonstante bei den in den Stammfunktionen verwendeten Polynomen

$$B_1(x) = x - \tfrac{1}{2}, \quad B_2(x) = x^2 - x + \tfrac{1}{6}, \quad B_3(x) = x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x$$

ist dabei jeweils so gewählt, daß $\int_0^1 B_j(x)\,dx = 0$ für $j = 1,2,3$ gilt. Die Polynome $B_j(x)$ heißen auch *Bernoulli-Polynome*.

dass ein zu $f'(1) - f'(0)$ proportionaler Randterm entsteht.[40]

$$\begin{aligned}\int_0^1 (x - \tfrac{1}{2}) f'(x)\,dx &= [\tfrac{1}{2}(x^2 - x + \tfrac{1}{6}) f'(x)]_{x=0}^1 - \int_0^1 \tfrac{1}{2}(x^2 - x + \tfrac{1}{6}) f''(x)\,dx \\ &= \tfrac{1}{12}[f'(1) - f'(0)] - \int_0^1 \tfrac{1}{2}(x^2 - x + \tfrac{1}{6}) f''(x)\,dx. \end{aligned} \tag{718}$$

Für unsere gewünschten Quadraturfehlerschranken ist die Gewichtsfunktion $x^2 - x + \frac{1}{6}$ im letzten Integranden immer noch nicht gut geeignet, da sie kein einheitliches Vorzeichen besitzt. Deshalb integrieren wir das letzte Integral zwei weitere Male partiell:

$$\begin{aligned}&\int_0^1 \tfrac{1}{2}(x^2 - x + \tfrac{1}{6}) f''(x)\,dx = [\tfrac{1}{6}(x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x) f''(x)]_{x=0}^1 - \int_0^1 \tfrac{1}{6}(x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x) f'''(x)\,dx \\ &= -\int_0^1 \tfrac{1}{6}(x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x) f'''(x)\,dx \\ &= -[\tfrac{1}{24}(x^4 - 2x^3 + x^2) f'''(x)]_{x=0}^1 + \int_0^1 \tfrac{1}{24}(x^4 - 2x^3 + x^2) f''''(x)\,dx \\ &= -\int_0^1 \tfrac{1}{24} x^2 (1-x)^2 f''''(x)\,dx. \end{aligned}$$

Man beachte, dass die Gewichtsfunktion $x^2(1-x)^2$ im Integranden nun nichtnegativ ist. In (718) eingesetzt erhalten wir:

$$\int_0^1 f(x)\,dx = \tfrac{1}{2}[f(0) + f(1)] + \frac{1}{12}[f'(0) - f'(1)] + \int_0^1 \tfrac{1}{24} x^2 (1-x)^2 f''''(x)\,dx. \tag{719}$$

Die beiden Formeln (717) und (719) kann man als Trapezregel mit zwei verschiedenen Darstellungen des Restglieds auffassen. Wir verwenden diese beiden Formeln für eine obere bzw. untere Schranke des Trapezregel-Quadraturfehlers

$$R_n := \int_n^{n+1} \log t\,dt - \tfrac{1}{2}[\log n + \log(n+1)].$$

Wenden wir also die beiden Formeln (717) und (719) auf $f(x) = \log(x+n)$, $f'(x) = 1/(x+n)$, $f''(x) = -(x+n)^{-2} < 0$ und $f''''(x) = -6(x+n)^{-4} < 0$ mit $n \in \mathbb{N}$ an. Zunächst mit Formel (717):

$$\begin{aligned}\int_n^{n+1} \log t\,dt &= \int_0^1 \log(x+n)\,dx \\ &= \tfrac{1}{2}[\log n + \log(n+1)] + \tfrac{1}{2}\int_0^1 x(1-x)(x+n)^{-2}\,dx \\ &\geq \tfrac{1}{2}[\log n + \log(n+1)], \end{aligned}$$

[40] Beim Aufsummieren zur iterierten Trapezregel liefert dieser Randterm eine Teleskopsumme. Davon machen wir später Gebrauch.

wobei wir verwendet haben, dass der Integrand $x(1-x)(x+n)^{-2}$ im letzten Integral nichtnegativ ist. Andererseits mit Formel (719):

$$\begin{aligned}\int_n^{n+1} \log t\,dt &= \int_0^1 \log(x+n)\,dx \\ &= \tfrac{1}{2}[\log n + \log(n+1)] + \tfrac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right) - \int_0^1 \tfrac{1}{4}x^2(1-x)^2(x+n)^{-2}\,dx \\ &\leq \tfrac{1}{2}[\log n + \log(n+1)] + \tfrac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right),\end{aligned}$$

wobei wir auch hier die Nichtnegativität des Integranden $\frac{1}{4}x^2(1-x)^2(x+n)^{-2}$ verwendet haben. Zusammen ist damit gezeigt:

$$0 \leq R_n \leq \tfrac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right). \tag{720}$$

Wir setzen für $n \in \mathbb{N}$:

$$L_n := \log n! - (n+\tfrac{1}{2})\log n + n = \log \frac{n!}{n^{n+\frac{1}{2}}e^{-n}}.$$

Um die Quadraturfehler R_n in einer Teleskopsumme aufzusummieren, schreiben wir sie wie folgt als Differenzen:

$$\begin{aligned}R_n &= \int_n^{n+1} \log t\,dt - \tfrac{1}{2}[\log n + \log(n+1)] \\ &= [(n+1)\log(n+1) - (n+1)] - [n\log n - n] - \tfrac{1}{2}[\log n + \log(n+1)] \\ &= [\log n! - (n+\tfrac{1}{2})\log n + n] - [\log(n+1)! - (n+1+\tfrac{1}{2})\log(n+1) + (n+1)] \\ &= L_n - L_{n+1},\end{aligned}$$

wobei wir $\log(n+1)! - \log n! = \log(n+1)$ verwendet haben. Summieren wir R_n für $n = n_1, \ldots, n_2 - 1$ mit natürlichen Zahlen $n_1 < n_2$ als Teleskopsumme auf verwenden die Quadraturfehlerschranken (720):

$$\begin{aligned}0 \leq \sum_{n=n_1}^{n_2-1} R_n &= \sum_{n=n_1}^{n_2-1} (L_n - L_{n+1}) = L_{n_1} - L_{n_2} \\ &\leq \sum_{n=n_1}^{n_2-1} \tfrac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right) = \tfrac{1}{12}\left(\frac{1}{n_1} - \frac{1}{n_2}\right)\end{aligned} \tag{721}$$

Insbesondere ist $(L_n)_{n\in\mathbb{N}}$ eine monoton fallende Cauchyfolge, also konvergent:

$$L := \lim_{n\to\infty} L_n \in \mathbb{R}.$$

Aus (721) erhalten wir im Limes $n_2 \to \infty$ für alle $n = n_1 \in \mathbb{N}$:

$$0 \leq L_n - L \leq \frac{1}{12n}. \tag{722}$$

Um die Konstante L zu identifizieren, verwenden wir die Asymptotik des mittleren Binomialkoeffizienten aus Lemma 2.116 wie folgt:

$$\exp(L_{2n} - 2L_n) = \frac{(2n)!}{(2n)^{2n+\frac{1}{2}}e^{-2n}} \frac{(n^{n+\frac{1}{2}}e^{-n})^2}{(n!)^2} = \sqrt{\frac{n}{2}} 2^{-2n} \binom{2n}{n} \overset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}},$$

also

$$L = \lim_{n\to\infty} (2L_n - L_{2n}) = \log\sqrt{2\pi}.$$

Die Schranken (722) lauten damit

$$0 \leq L_n - \log\sqrt{2\pi} = \log \frac{n!}{\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}} \leq \frac{1}{12n},$$

woraus die Stirlingformel (714) folgt.

3 Mathematische Statistik

3.1 Grundlagen

Wahrscheinlichkeitstheorie und Mathematische Statistik beschäftigen sich beide mit zufälligen Vorgängen. Die Sichtweisen unterscheiden sich jedoch:

	Wahrscheinlichkeitstheorie	Statistik
Wahrscheinlichkeitsmaß P	*bekannt*	*unbekannt* (bis auf einige allgemeine Annahmen)
Ergebnis ω des Zufallsexperiments	*unbekannt*	*bekannt* (Beobachtungsdaten)
Typische Aufgaben	Berechnung oder Abschätzung von Wahrscheinlichkeiten interessanter Ereignisse	Testen von Hypothesen über die unbekannte Verteilung P, Schätzen von Parametern der unbekannten Verteilung P

Wichtig!

Statistik beschäftigt sich mit "inversen Problemen" zur Wahrscheinlichkeitstheorie.

Man möchte Informationen über unbekannte Wahrscheinlichkeitsverteilungen aus Beobachtungsdaten gewinnen. Die Daten werden als beobachtete Werte einer Zufallsvariablen interpretiert.

Definition 3.1 (statistisches Modell) *Ein* statistisches Modell*, synonym* Rahmenmodell*, ist ein Tripel* $(\Omega, \mathcal{A}, \mathcal{P})$*, bestehend aus einem Ergebnisraum* Ω*, einer Ereignis-*σ*-Algebra* $\mathcal{A}$ *darüber und einer* Menge $\mathcal{P}$ *von Wahrscheinlichkeitsmaßen über* Ω*. Die Beobachtungsdaten werden in einem Ergebnis* $\omega \in \Omega$*, synonym: in einer* Stichprobe $\omega \in \Omega$*, englisch:* sample*, codiert.*

fundamental!

Beispiel: Eine möglicherweise unfaire Münze wird n-mal geworfen. Wir erhalten Beobachtungsdaten

$$\omega = (\omega_1, \ldots, \omega_n) \in \Omega := \{0,1\}^n. \tag{723}$$

Es sei $\mathcal{A} = \mathcal{P}(\Omega)$. Ohne die Beobachtungsdaten ω zu kennen, ist es plausibel, die folgenden Modellannahmen zu treffen:

Modellannahmen für das Rahmenmodell: Unter der unbekannten Wahrscheinlichkeitsverteilung P auf $(\Omega, \mathcal{A})$ sind die $\omega_1, \ldots, \omega_n$ *unabhängig* und *identisch verteilt.*

Formalisierung der Modellannahmen: Unser Rahmenmodell verwendet die folgende Klasse von Wahrscheinlichkeitsmaßen:

$$\mathcal{P} = \{(p\delta_1 + (1-p)\delta_0)^n |\ 0 \le p \le 1\}. \tag{724}$$

Grundsätzlich sollte man das Rahmenmodell entwerfen, bevor *man die Beobachtungsdaten $\omega \in \Omega$ kennt.*

Damit sollen nämlich Verfälschungen der auf der Zufälligkeit der Daten beruhenden statistischen Vorhersagen vermieden werden.

Definition 3.2 (parametrische und nichtparametrische Modelle) *Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell. Wird die Menge $\mathcal{P}$ der Wahrscheinlichkeitsmaße im Modell als eine Klasse von Verteilungen P_θ mit endlich vielen reellen Parametern $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ gegeben, also*

$$\mathcal{P} = \{P_\theta |\ \theta \in \Theta\} \tag{725}$$

mit $\Theta \subseteq \mathbb{R}^d$, so heißt $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ ein parametrisches Modell. *Andernfalls – typischerweise für unendlichdimensionale $\mathcal{P}$ – heißt das Rahmenmodell $(\Omega, \mathcal{A}, \mathcal{P})$ ein* nichtparametrisches Modell.

Beispiele:

1. Das Modell für den n-fachen Münzwurf

$$\Omega = \{0,1\}^n, \quad \mathcal{A} = \mathcal{P}(\Omega), \quad \mathcal{P} = \{P_p^n |\ 0 \le p \le 1\} \tag{726}$$

   mit

$$P_p = p\delta_1 + (1-p)\delta_0 \tag{727}$$

   ist ein parametrisches Modell mit einem Parameter p.

2. Das Tripel

$$(\Omega = \mathbb{R}^n, \mathcal{A} = \mathcal{B}(\mathbb{R}^n), \mathcal{P} = \{P^n |\ P \text{ ist ein Wahrscheinlichkeitsmaß über } (\mathbb{R}, \mathcal{B}(\mathbb{R}))\}) \tag{728}$$

   ist ein statistisches Modell für n i.i.d. Beobachtungen mit Werten in $\mathbb{R}$, über deren Verteilung nichts bekannt ist. Es ist ein nichtparametrisches Modell.

Frequentistische versus Bayessche Sicht. Es gibt zwei grundsätzliche verschiedene Herangehensweisen an die Statistik:

Frequentistische Sicht:

- Die *beobachteten, bekannten* Daten $\omega \in \Omega$ werden als *zufälliges* Ergebnis eines Zufallsexperiments interpretiert.
- Das zugrundeliegende Wahrscheinlichkeitsmaß P wird als fest, *nicht zufällig*, aber *unbekannt* aufgefasst.

Wichtig!

Bayessche Sicht:

- Die Klasse $\mathcal{P}$ der plausiblen Wahrscheinlichkeitsverteilungen P wird selbst mit einer σ-Algebra $\mathbb{A}$ und einem Wahrscheinlichkeitsmaß $\mathbb{P}$ versehen. Die Verteilung $\mathbb{P}$ wird *a priori Verteilung* (englisch: *prior distribution* oder kurz *prior*) genannt; sie erlaubt es, Vorwissen des Statistikers über die Plausibilität der verschiedenen möglichen Wahrscheinlichkeitsmaße $P \in \mathcal{P}$ zu modellieren.
- Die Beobachtungsdaten $\omega \in \Omega$ werden als Ergebnis eines *zweistufigen* Zufallsexperiments interpretiert:
  1. In der ersten Stufe wählt "die Natur" *zufällig* ein Wahrscheinlichkeitsmaß $P \in \mathcal{P}$ nach dem Wahrscheinlichkeitsmodell $(\mathcal{P}, \mathbb{A}, \mathbb{P})$ aus.
  2. In der zweiten Stufe wird das Beobachtungsergebnis $\omega \in \Omega$ zufällig im Modell $(\Omega, \mathcal{A}, P)$ gezogen.
- Der Statistiker studiert die Verteilung des Wahrscheinlichkeitsmaßes $P \in \mathcal{P}$ *bedingt auf die Beobachtung* $\omega \in \Omega$. Sie heißt *a posteriori Verteilung*, englisch *posterior distribution*.

Wichtig!

Beispiel: Für eine Bayessche Modellierung des n-fachen Münzwurfs nehmen wir an, dass der unbekannte Parameter $p \in [0,1]$ des dem Münzwurfexperiment zugrundeliegenden Wahrscheinlichkeitsmaßes P_p^n unter dem a priori Maß $\mathbb{P}$ uniform auf $[0,1]$ verteilt sei. Das gesamte, zweistufige Münzwurf-Zufallsexperiment wird dann durch das folgende Wahrscheinlichkeitsmaß Q über $[0,1] \times \{0,1\}^n$ modelliert:

$$\begin{aligned} Q(A) &= \int_{[0,1]} P_p^n(\{\omega \in \{0,1\}^n |\ (p,\omega) \in A\})\, dp \\ &= \int_{[0,1]} \sum_{\omega \in \{0,1\}^n} 1_A(p,\omega) \prod_{i=1}^n p^{\omega_i}(1-p)^{1-\omega_i}\, dp, \qquad (A \in \mathcal{B}([0,1] \times \{0,1\}^n)) \end{aligned} \tag{729}$$

Gegeben Beobachtungsdaten $\omega \in \{0,1\}^n$ mit $k = \sum_{i=1}^n \omega_i$ Einsen, wird die a posteriori Verteilung des Parameters $p \in [0,1]$ dann durch

$$\begin{aligned} Q(A \times \{\omega\} |\ [0,1] \times \{\omega\}) &= \frac{\int_A \prod_{i=1}^n p^{\omega_i}(1-p)^{1-\omega_i}\, dp}{\int_{[0,1]} \prod_{i=1}^n p^{\omega_i}(1-p)^{1-\omega_i}\, dp} \\ &= \frac{\int_A p^k(1-p)^{n-k}\, dp}{\int_0^1 p^k(1-p)^{n-k}\, dp} \qquad \text{für } A \in \mathcal{B}([0,1]) \end{aligned} \tag{730}$$

gegeben, also durch die Betaverteilung mit den Parametern $k+1$ und $n-k+1$.

Wir beschäftigen uns im Rest der Vorlesung allerdings nur mehr mit der frequentistischen Sicht.

Definition 3.3 (dominierendes Maß) *Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell. Ein*

dominierendes Maß *für $\mathcal{P}$ ist ein σ-endliches* [41] *Maß μ auf $(\Omega, \mathcal{A})$, bezüglich dem alle $P \in \mathcal{P}$ eine Dichte $\frac{dP}{d\mu}$ besitzen. Existiert ein dominierendes Maß μ, so heißt das Modell* $(\Omega, \mathcal{A}, \mathcal{P})$ dominiert.

Bemerkung: Ist P ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$, so dass P eine Dichte f bezüglich μ besitzt, so ist diese Dichte f bis auf Abänderung auf einer μ-Nullmenge eindeutig bestimmt. Wir schreiben dafür

$$f = \frac{dP}{d\mu} \quad \mu\text{-fast überall.} \tag{731}$$

Definition 3.4 (Likelihood-Funktion) *Es sei $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ ein parametrisches statistisches Modell mit dominierendem Maß μ. Die Abbildung*

$$L : \Omega \to \mathbb{R}, \qquad L(\omega, \theta) = \frac{dP_\theta}{d\mu}(\omega) \tag{732}$$

heißt Likelihood-Funktion *des Modells. Für jeden Parameter $\theta \in \Theta$ ist $L(\cdot, \theta)$ μ-fast überall eindeutig bestimmt.*

Beispiel: Ist Ω endlich oder abzählbar unendlich, $\mathcal{A} = \mathcal{P}(\Omega)$, μ =Zählmaß, so ist die Likelihood-Funktion gegeben durch

$$L : \Omega \times \Theta \to \mathbb{R}, \qquad L(\omega, \theta) = P_\theta(\{\omega\}). \tag{733}$$

Im Münzwurfmodell (726) lautet die Likelihoodfunktion also

$$\begin{aligned} L : \{0,1\}^n \times [0,1] &\to \mathbb{R}, \\ L(\omega, p) = p^{S(\omega)}(1-p)^{n-S(\omega)}&, \end{aligned} \tag{734}$$

wobei

$$S : \{0,1\}^n \to \mathbb{N}_0, \qquad S(\omega_1, \ldots, \omega_n) = \sum_{k=1}^{n} \omega_k \tag{735}$$

die Anzahl der Einsen in einer Stichprobe bezeichnet.

Die Likelihood-Funktion codiert also alle Wahrscheinlichkeitsmaße P_θ, $\theta \in \Theta$, des Modells in einer einzigen Abbildung.

Definition 3.5 (Likelihood-Quotient) *Für zwei Wahrscheinlichkeitsmaße $P, Q \in \mathcal{P}$, 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, wann immer dieser definiert ist:

$$\frac{dP_{\theta_1}}{dP_{\theta_2}}(\omega) = \frac{\frac{dP_{\theta_1}}{d\mu}(\omega)}{\frac{dP_{\theta_2}}{d\mu}(\omega)} = \frac{L(\omega, \theta_1)}{L(\omega, \theta_2)} \quad P_{\theta_2}\text{-f.ü.} \qquad (\omega \in \Omega,\ \theta_1, \theta_2 \in \Theta) \tag{736}$$

[41] Erinnerung: Ein Maß μ auf $(\Omega, \mathcal{A})$ heißt σ-endlich, wenn es eine Ereignisfolge $A_n \uparrow \Omega$ mit $\mu(A_n) < \infty$ für alle $n \in \mathbb{N}$ gibt, siehe Definition 2.47.

3.2 Elemente der Schätztheorie

Definition 3.6 (Schätzer eines Parameters) *Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell. Ein* Parameter *ist eine Abbildung $\theta : \mathcal{P} \to \mathbb{R}$ (oder d-dimensional: $\theta : \mathcal{P} \to \mathbb{R}^d$).*
Ein Schätzer *für einen Parameter θ ist eine $\mathcal{A}$–$\mathcal{B}(\mathbb{R})$–messbare Abbildung $\hat{\theta} : \Omega \to \mathbb{R}$ (oder d-dimensional: $\hat{\theta} : \mathcal{P} \to \mathbb{R}^d$).*

Bemerkungen:

1. Schätzer werden traditionell mit einem Symbol " ˆ " gekennzeichnet.

2. Die Definition des Schätzers sagt nichts darüber aus, wie "gut" der Schätzer ist. Für eine Beobachtung $\omega \in \Omega$ bei zugrundeliegender Verteilung $P \in \mathcal{P}$ heißt $\hat{\theta}(\omega) - \theta(P)$ der *Schätzfehler*.

Hier ist ein mögliches Kriterium für die "Güte" von Schätzern:

Definition 3.7 (erwartungstreue Schätzer) *Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell und $\theta : \mathcal{P} \to \mathbb{R}$ ein Parameter. Ein Schätzer $\hat{\theta} : \Omega \to \mathbb{R}$ heißt* erwartungstreu *(synonym:* unverfälscht*, englisch* unbiased*), wenn für alle $P \in \mathcal{P}$ gilt:*

$$E_P[\hat{\theta}] = \theta(P). \tag{737}$$

Anders gesagt:

$$\forall P \in \mathcal{P} : \; E_P[\hat{\theta} - \theta(P)] = 0. \tag{738}$$

Beispiele:

1. Es seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit unbekannter Verteilung P mit einer unbekannten Erwartung $\mu(P) = E_P[X_i]$. Wir beschreiben diese Situation mit dem nichtparametrischen Modell

$$\Omega = \mathbb{R}^n, \tag{739}$$
$$\mathcal{A} = \mathcal{B}(\mathbb{R}^n), \tag{740}$$
$$\mathcal{P} = \left\{ P^n \,\middle|\, \begin{array}{l} P \text{ ist ein Wahrscheinlichkeitsmaß} \\ \text{über } \mathcal{B}(\mathbb{R}) \text{ mit endlicher Erwartung } \mu(P) \end{array} \right\}, \tag{741}$$

   und

$$X_i(\omega_1, \ldots, \omega_n) = \omega_i. \tag{742}$$

   Das Stichprobenmittel

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i : \Omega \to \mathbb{R} \tag{743}$$

ist ein erwartungstreuer Schätzer für die Erwartung $\mu(P)$, denn für alle $P^n \in \mathcal{P}$ gilt:

$$E_{P^n}[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n} E_{P^n}[X_i] = \mu(P). \tag{744}$$

Allerdings ist jedes X_i, $i \in [n]$, ebenfalls ein erwartungstreuer Schätzer für $\mu(P)$.

2. Nehmen wir nun in Beispiel 1 zusätzlich an, dass die Zufallsvariablen $X_1, \ldots, X_n$ für alle $P \in \mathcal{P}$ eine endliche Varianz $\sigma^2(P)$ besitzen:

$$\Omega = \mathbb{R}^n, \tag{745}$$

$$\mathcal{A} = \mathcal{B}(\mathbb{R}^n), \tag{746}$$

$$\mathcal{P} = \left\{ P^n \,\middle|\, \begin{array}{l} P \text{ ist ein Wahrscheinlichkeitsmaß} \\ \text{über } \mathcal{B}(\mathbb{R}) \text{ mit endlicher Varianz } \sigma^2(P) \end{array} \right\}. \tag{747}$$

Die *empirische Varianz* der Stichprobe $X_1, \ldots, X_n$ wird durch

$$s_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{748}$$

definiert, wobei natürlich $n > 1$ gelten muss. Die empirische Varianz ist ein erwartungstreuer Schätzer für die Varianz $\sigma^2(P)$, anders als der vielleicht naheliegendere Schätzer

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}s_X^2. \tag{749}$$

Beweis: Es sei $P^n \in \mathcal{P}$. Wegen

$$E_{P^n}[X_i - \overline{X}] = \mu(P) - \mu(P) = 0 \tag{750}$$

erhalten wir

$$\begin{aligned} E_{P^n}[(X_i - \overline{X})^2] &= \mathrm{Var}_{P^n}(X_i - \overline{X}) \\ &= \mathrm{Var}_{P^n} \underbrace{\left(\left(1 - \frac{1}{n}\right) X_i - \sum_{j \in [n]\setminus\{i\}} \frac{X_j}{n} \right)}_{\text{alle Summanden unabhängig}} \\ &= \left(1 - \frac{1}{n}\right)^2 \underbrace{\mathrm{Var}_{P^n}(X_i)}_{=\sigma^2(P)} + \sum_{j \in [n]\setminus\{i\}} \frac{1}{n^2} \mathrm{Var}_{P^n} X_j \\ &= \left(1 - \frac{1}{n}\right)^2 \sigma^2(P) + (n-1) \cdot \frac{1}{n^2}\sigma^2(P) \\ &= \frac{n-1}{n}\sigma^2(P). \end{aligned} \tag{751}$$

Es folgt:

$$E_{P^n}[s_X^2] = \sum_{i=1}^{n} E_{P^n}[(X_i - \overline{X})^2] = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2(P) = \sigma^2(P), \tag{752}$$

aber für $\sigma^2(P) \neq 0$:

$$E_{P^n}\left[\frac{n-1}{n}s_X^2\right] = \frac{n-1}{n}\sigma^2(P) \neq \sigma^2(P). \tag{753}$$

□

Bemerkung: $s_X = \sqrt{s_X^2}$ ist *kein* erwartungstreuer Schätzer für die Standardabweichung $\sigma(P) = \sqrt{\sigma^2(P)}$. Es gibt nicht einmal einen erwartungstreuen Schätzer für $\sigma(P)$:

Übung 3.8 Nichtexistenz erwartungstreuer Schätzer für die Standardabweichung. *Eine möglicherweise unfaire Münze (beschriftet mit "0" und "1" bei unbekannter Wahrscheinlichkeit $p \in [0,1]$ für "1") wird n-mal unabhängig geworfen; man beobachtet eine Folge $\omega \in \{0,1\}^n$ mit $S(\omega)$ Einsen. Zeigen Sie: Es gibt keinen erwartungstreuen Schätzer für die unbekannte Standardabweichung $\sqrt{np(1-p)}$ von S.*

Obwohl es praktisch ist, erwartungstreue Schätzer zu haben, möchte man in einigen Fällen einen Bias:
Beispiel: (nach R. Gill) Ein Transatlantikkabel sollte verlegt werden. Die nötige Länge wurde geschätzt. Doch es war zu kurz: Einige hundert Meter vor dem Ziel fiel das Ende ins Wasser, mit großem Schaden. Hier wäre es besser gewesen, keinen erwartungstreuen Schätzer zu nehmen, sondern einen Sicherheitsabstand zu verwenden. Solch ein "Bias" ist immer dann sinnvoll, wenn Abweichungen in die eine Richtung viel schwerwiegendere Konsequenzen als Abweichungen in die entgegengesetzte Richtung haben.

Hier ist ein weiteres Kriterium für die Güte von Schätzern: Wir betrachten eine i.i.d. Stichprobe $X_1, \ldots, X_n$ mit Werten in Ω und unbekannter Verteilung P. Anders gesagt: Wir betrachten das Produkt-Rahmenmodell

$$(\Omega^n, \mathcal{A}^{\otimes n}, \mathcal{P}_n), \qquad (n \in \mathbb{N}) \tag{754}$$

mit einem statistischen Modell $(\Omega, \mathcal{A}, \mathcal{P})$ für Einzelbeobachtungen und

$$\mathcal{P}_n = \{P^n |\ P \in \mathcal{P}\}. \tag{755}$$

Definition 3.9 (Konsistenz von Schätzern) *Eine Folge $(\hat{\theta}_n : \Omega^n \to \mathbb{R})_{n\in\mathbb{N}}$ für einen Parameter $\theta : \mathcal{P} \to \mathbb{R}$ heißt* konsistent*, wenn für alle $P \in \mathcal{P}$ und alle $\epsilon > 0$ gilt:*

$$P^n[|\hat{\theta}_n - \theta(P)| > \epsilon] \overset{n\to\infty}{\longrightarrow} 0, \tag{756}$$

mit anderen Worten $\hat{\theta}_n \overset{n\to\infty}{\longrightarrow} \theta(P)$ in Wahrscheinlichkeit bezüglich P^n.

Beispiel: Das Stichprobenmittel

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (n \in \mathbb{N}) \tag{757}$$

ist eine konsistente Folge von Schätzern für die Erwartung

$$E_{P^n}[X_i] = \mu(P). \tag{758}$$

Das folgt aus dem schwachen Gesetz der großen Zahlen.

Maximum-Likelihood-Schätzer Wir besprechen nun einen recht allgemeinen Ansatz, sinnvolle Schätzer zu gewinnen:

Definition 3.10 (Maximum-Likelihood-Schätzer) *Es sei $(\Omega, \mathcal{A}, (\mathcal{P}_\theta)_{\theta\in\Theta})$ ein parametrisches statistisches Modell mit dominierendem Maß μ und Likelihood-Funktion $L : \Omega \times \Theta \to \mathbb{R}$. Der Maximum-Likelihood-Schätzer für den Parameter $\theta \in \Theta$ wird wie folgt definiert:*

$$\begin{aligned}
\hat{\theta}_{\mathrm{ML}} : \Omega \to & \Theta, \\
\hat{\theta}_{\mathrm{ML}}(\omega) = & \arg\max_{\theta\in\Theta} L(\omega, \theta) \\
= & \textit{Die Stelle } \theta \in \Theta, \textit{ an der} && (759)\\
& L(\omega, \cdot) \textit{ sein Maximum annimmt,} && (760)
\end{aligned}$$

falls es ein eindeutig bestimmtes solches Maximum gibt.[42]

Beispiel: *unfairer Münzwurf:*

$$\begin{aligned}
\Omega &= \{0, 1, \ldots, n\}, && (761)\\
\mathcal{A} &= \mathcal{P}(\Omega), && (762)\\
\mathcal{P} &= \{P_p |\ p \in \Theta\}, \quad \text{wobei} && (763)\\
P_p &= \text{binomial}(n, p), && (764)\\
\Theta &= [0, 1], && (765)\\
\mu &= \text{Zählmaß}. && (766)
\end{aligned}$$

Die Likelihood-Funktion lautet

$$\begin{aligned}
L : \Omega \times [0, 1] &\to \mathbb{R}, \\
L(\omega, p) &= \binom{n}{\omega} p^\omega (1-p)^{n-\omega}. && (767)
\end{aligned}$$

[42]Manchmal existiert die geforderte Maximumstelle nicht für *alle* $\omega \in \Omega$, sondern z.B. auf einer μ-Nullmenge nicht. In diesem Fall ist der Maximum-Likelihood-Schätzer nur partiell definiert, in Zeichen $\hat{\theta}_{\mathrm{ML}} : \Omega \dashrightarrow \Theta$, bleibt also z.B. auf einer μ-Nullmenge undefiniert.

Gegeben $\omega \in \Omega$, maximieren wir $L(\omega, p)$. Wir rechnen:

$$\begin{aligned}\frac{\partial}{\partial p} \log L(\omega, p) &= \frac{\partial}{\partial p}[\omega \log p + (n-\omega)\log(1-p)] \\ &= \frac{\omega}{p} - \frac{n-\omega}{1-p},\end{aligned} \tag{768}$$

was für $0 < p < 1$ streng monoton fällt. Die Ableitung wird 0 an der Stelle

$$\hat{p}_{\mathrm{ML}}(\omega) = \frac{\omega}{n}, \tag{769}$$

falls $\omega \in \{1, \ldots, n-1\}$, und auch für *alle* $\omega \in \{0, \ldots, n\}$ ist $\hat{p}_{\mathrm{ML}}(\omega)$ die Stelle, an der $L(\omega, p)$ maximal wird. Der Maximum-Likelihood-Schätzer für den Parameter p im Münzwurfmodell ist also die relative Häufigkeit der "1" in den Beobachtungen.

Bemerkung: In Produktmodellen

$$\mathcal{P}_n = \{P_\theta^n |\ \theta \in \Theta\} \tag{770}$$

mit einer Likelihood-Funktion

$$L_n(\omega_1, \ldots, \omega_n; \theta) = \prod_{i=1}^{n} L(\omega_i, \theta) \tag{771}$$

ist es rechnerisch oft einfacher, den Logarithmus von L_n zu minimieren, denn

$$\log L_n(\omega_1, \ldots, \omega_n; \theta) = \sum_{i=1}^{n} \log L(\omega_i, \theta), \tag{772}$$

statt L_n direkt zu betrachten. Die Funktion $\log L_n$ heißt die *Log-Likelihood-Funktion.*

Momentenschätzer. Es sei $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (P_\theta^n)_{\theta\in\Theta})$ ein k-parametrisches Modell. Eine sehr einfache Methode zur Gewinnung von Schätzern erhält man so:
Man schätzt die ersten k Momente (oder auch zentrierten Momente) von P_θ, z.B. für $k = 1$ mit dem Mittel der Stichprobe ω und für $k = 2$ mit dem Mittel und der empirischen Varianz der Stichprobe. Dann wählt man dasjenige $\hat{\theta}(\omega) \in \Theta$, dessen erste k Momente zu $P_{\hat{\theta}(\omega)}$ (bzw. zentrierte Momente) mit den Schätzern übereinstimmt.

Beispiel: Es seien $X_1, \ldots, X_n$ i.i.d. $\mathrm{N}(\mu, \sigma^2)$-verteilt mit unbekanntem μ und σ^2, also $P_{\mu,\sigma}^n = \mathrm{N}(\mu, \sigma^2)^n$. Der Momentenschätzer für die Parameter (μ, σ) ist $(\overline{X}, s_X^2)$.

3.2.1 Ausgleichsrechnung: die Methode der kleinsten Quadrate

Als ein Beispiel betrachten wir lineare Gleichungssysteme mit zufällig gestörter rechter Seite:

Wir betrachten ein lineares Gleichungssystem

$$Ax = b \tag{773}$$

mit einer Matrix $A \in \mathbb{R}^{m\times n}$ und rechter Seite $b \in \mathbb{R}^m$ und Lösungsvektor $x \in \mathbb{R}^n$. Wir stellen uns die Matrix A als bekannt vor und nehmen an, dass ihr Nullraum verschwindet,

$$N(A) = \{x \in \mathbb{R}^n |\ Ax = 0\} = \{0\}, \tag{774}$$

so dass das System $Ax = b$ nur höchstens eine Lösung besitzt. Die rechte Seite b sei jedoch nicht bekannt, sondern nur eine zufällige Störung $\beta \in \mathbb{R}^m$ davon sei bekannt. Wir machen die Modellannahme, dass die Komponenten $\beta_1, \ldots, \beta_m$ von β unabhängig voneinander sind und β_i normalverteilt mit der Erwartung b_i und der Varianz σ^2 sind, wobei weder der Erwartungswertvektor

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}, \tag{775}$$

also die ungestörte rechte Seite des Gleichungssystems, noch die Varianz σ^2 bekannt seien. Gesucht ist eine Schätzung $\hat{x}$ der Lösung x des Gleichungssystems mit nicht genau bekannter rechter Seite b, sowie eine Schätzung $\widehat{\sigma^2}$ der unbekannten Varianz σ^2.

Formalisieren wir die Angaben mit einem statistischen Modell:

$$\text{Stichprobenraum } \Omega = \mathbb{R}^m \ni \beta, \quad \mathcal{A} = \mathcal{B}(\Omega), \tag{776}$$

$$\text{Parameterraum } \Theta = \mathbb{R}^n \times \mathbb{R}^+ \ni (x, \sigma^2), \tag{777}$$

$$\mathcal{P} = \{P_{x,\sigma^2} |\ (x, \sigma^2) \in \Theta\}, \tag{778}$$

wobei

$$P_{x,\sigma^2} = \mathrm{N}(Ax, \sigma^2 1_m) = \prod_{i=1}^{m} \mathrm{N}((Ax)_i, \sigma^2) \tag{779}$$

die multidimensionale Normalverteilung mit Erwartungswertvektor Ax und dem σ^2-fachen der $m \times m$ Einheitsmatrix 1_m als Covarianzmatrix bezeichnet.
Das Modell besitzt also die Likelihoodfunktion[43]

$$L : \Omega \times \Theta \to \mathbb{R},$$

$$L(\beta, x, \sigma^2) = (2\pi)^{-\frac{m}{2}} \sigma^{-m} \exp\left(-\frac{1}{2\sigma^2} \|Ax - \beta\|_2^2\right). \tag{781}$$

[43]Die euklidische Norm von $y \in \mathbb{R}^m$ wird mit

$$\|y\|_2 = \sqrt{\sum_{i=1}^{m} y_i^2} \tag{780}$$

bezeichnet.

Gegeben β, suchen wir den Maximum-Likelihood-Schätzer $(\hat{x}, \widehat{\sigma^2})$ für die unbekannten Parameter (x, σ^2). 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:

Wir suchen

$$\hat{x} = \arg\min_{x \in \mathbb{R}^m} \|Ax - \beta\|_2^2, \tag{782}$$

also dasjenige $\hat{x} \in \mathbb{R}^m$, für das gilt:

$$\|A\hat{x} - \beta\|_2^2 \leq \|Ax - \beta\|_2^2 \text{ für alle } x \in \mathbb{R}^m. \tag{783}$$

Behauptung: Ist $\hat{b} = A\hat{x}$ die orthogonale Projektion von β auf den Raum

$$R(A) = \{Ax |\ x \in \mathbb{R}^n\}, \tag{784}$$

so erfüllt $\hat{x}$ das Ziel (782).

In der Tat:[44] Ist

$$\langle Ax, A\hat{x} - \beta \rangle = 0 \text{ für alle } x \in \mathbb{R}^n, \tag{785}$$

so folgt:

$$\begin{aligned}
&\|Ax - \beta\|_2^2 \\
&= \|(Ax - A\hat{x}) + (A\hat{x} - \beta)\|_2^2 \\
&= \|Ax - A\hat{x}\|_2^2 + \underbrace{2 \left\langle \underbrace{Ax - A\hat{x}}_{\in R(A)}, A\hat{x} - \beta \right\rangle}_{=0} + \|A\hat{x} - \beta\|_2^2 \\
&= \|Ax - A\hat{x}\|_2^2 + \|A\hat{x} - \beta\|_2^2 \\
&\geq \|A\hat{x} - \beta\|_2^2.
\end{aligned} \tag{786}$$

Die Gleichung (785) ist äquivalent zu[45]

$$\left\langle x, A^t(A\hat{x} - \beta) \right\rangle = 0 \text{ für alle } x \in \mathbb{R}^n, \tag{787}$$

also zu

$$A^t(A\hat{x} - \beta) = 0, \tag{788}$$

d.h.

$$A^t A\hat{x} = A^t\beta. \tag{789}$$

[44] $\langle u, v \rangle = \sum_{i=1}^m u_i v_i$ bezeichnet das euklidische Skalarprodukt von $u, v \in \mathbb{R}^m$.

[45] A^t bezeichnet die transponierte Matrix zu A.

Nun ist die Matrix A^tA wegen der Voraussetzung $N(A) = \{0\}$ in Formel (774) invertierbar, so dass folgt:

Schätzer nach der Methode der kleinsten Quadrate:

$$\hat{x} = (A^tA)^{-1}A^t\beta \tag{790}$$

Damit ist $\hat{x}$ und damit auch das sogenannte *Residuum*

$$r^2 := \|A\hat{x} - \beta\|_2^2 \tag{791}$$

bekannt. Insbesondere hängt $\hat{x}$ nicht von der unbekannten Varianz σ^2 ab. Nun bestimmen wir den Maximum-Likelihood-Schätzer $\widehat{\sigma^2}$ für σ^2. Wir maximieren also $L(\beta, \hat{x}, \sigma^2)$ über σ^2, gegeben β und damit $\hat{x}$:
Es gilt

$$\log L(\beta, \hat{x}, \sigma^2) = -\frac{m}{2}\log(2\pi) - m\log\sigma - \frac{1}{2\sigma^2}r^2. \tag{792}$$

Für $\sigma^2 \to \infty$ oder $\sigma^2 \to 0$ 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:

$$\frac{\partial}{\partial\sigma}\log L(\beta, \hat{x}, \sigma^2) = -\frac{m}{\sigma} + \frac{r^2}{\sigma^3}, \tag{793}$$

was eine Nullstelle bei

$$\boxed{\widehat{\sigma^2} = \frac{r^2}{m}} \tag{794}$$

besitzt. Durch die Gleichungen (790) und (794) wird also der gesuchte Maximum-Likelihood-Schätzer gegeben.
Wir wenden diese Theorie an, um gegebene Messpunkte (x_i, y_i), $i \in [n]$ "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 als bekannt, aber zufällig, und zwar unabhängig voneinander und normalverteilt mit den Erwartungen $ax_i + b$ und der Varianz σ^2 (Varianz), $i \in [n]$, wobei $a, b \in \mathbb{R}$ und $\sigma^2 > 0$ die unbekannten Parameter des Modells bezeichnen.

Beschreiben wir das formaler mit einem statistischen Modell:

$$\Omega = \mathbb{R}^n \ni (y_1, \ldots, y_n), \quad \mathcal{A} = \mathcal{B}(\Omega), \tag{795}$$

$$\Theta = \mathbb{R}^2 \times \mathbb{R}^+ \ni (a, b, \sigma^2), \tag{796}$$

$$\mathcal{P} = \{P_{a,b,\sigma^2} |\, (a, b, \sigma^2) \in \Theta\} \text{ mit} \tag{797}$$

$$P_{a,b,\sigma^2} = \prod_{i=1}^{n} \mathrm{N}(\underbrace{ax_i + b}_{=:\tilde{y}_i}, \sigma^2). \tag{798}$$

Das Gleichungssystem

$$ax_i + b = \tilde{y}_i, \quad (i \in [n]) \tag{799}$$

mit der Störung y_i von $\tilde{y}_i$, anders geschrieben

$$\underbrace{\begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}}_{=:A} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{pmatrix}, \tag{800}$$

mit der Störung

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ von } \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{pmatrix}, \tag{801}$$

führt uns nach der Methode der kleinsten Quadrate, Formel (790), zur Schätzung

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (A^t A)^{-1} A^t \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}. \tag{802}$$

Nun ist

$$A^t A = \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} = \begin{pmatrix} \sum_{i\in[n]} x_i^2 & \sum_{i\in[n]} x_i \\ \sum_{i\in[n]} x_i & n \end{pmatrix} = n \begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}, \tag{803}$$

wobei $\overline{x}$ bzw. $\overline{x^2}$ das Mittel der x_i bzw. der x_i^2 bezeichnet, und

$$A^t \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i\in[n]} x_i y_i \\ \sum_{i\in[n]} y_i \end{pmatrix} = n \begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix} \tag{804}$$

Es folgt mit der Cramerschen Regel:

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix} = \frac{1}{\begin{vmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{vmatrix}} \begin{pmatrix} \begin{vmatrix} \overline{xy} & \overline{x} \\ \overline{y} & 1 \end{vmatrix} \\ \begin{vmatrix} \overline{x^2} & \overline{xy} \\ \overline{x} & \overline{y} \end{vmatrix} \end{pmatrix}, \tag{805}$$

also

Koeffizienten der Regressionsgeraden:

$$\hat{a} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}, \tag{806}$$

$$\hat{b} = \frac{\overline{x^2}\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} \tag{807}$$

mit dem Residuum

$$r^2 = \sum_{i\in[n]} (\hat{a}x_i - \hat{b} - y_i)^2 \tag{808}$$

und dem Maximum-Likelihood-Schätzer für σ^2:

$$\boxed{\widehat{\sigma^2} = \frac{r^2}{n} = \frac{1}{n}\sum_{i\in[n]} (\hat{a}x_i - \hat{b} - y_i)^2} \tag{809}$$

3.3 Einführung in die Testtheorie

Beispiel: Im Umkreis von 5 Kilometern von Kernkraftwerken wuchsen in den letzten Jahren n_1 Kinder auf; ω_1 davon erkrankten an Leukämie. In einer Kontrollgruppe von n_2 Kindern erkrankten ω_2 an Leukämie. Wann "belegen" diese Daten, dass Kinder im Umkreis von Kernkraftwerken eine höhere Wahrscheinlichkeit haben, an Leukämie zu erkranken?

Wir verwenden dazu folgendes (zugegeben stark vereinfachtes[46]) **Modell:**

- ω_1 ist eine binomial(n_1, p_1)-verteilte Zufallsvariable mit unbekanntem p_1.
- ω_2 ist eine binomial(n_2, p_2)-verteilte Zufallsvariable mit unbekanntem p_2.
- ω_1 und ω_2 sind unabhängig.

Formaler beschreiben wir das mit dem statistischen Modell

$$\begin{aligned}&(\Omega, \mathcal{A}, \mathcal{P})\\ &= (\{0,\dots,n_1\}\times\{0,\dots,n_2\}, \mathcal{P}(\Omega), \{\text{binomial}(n_1,p_1)\times\text{binomial}(n_2,p_2) |\ p_1,p_2\in[0,1]\})\end{aligned} \tag{810}$$

Wir fragen uns, ob wir die Hypothese $p_1 = p_2$ aufgrund der beobachteten Daten $(\omega_1, \omega_2) \in \Omega$ zu Gunsten der Alternativhypothese $p_1 > p_2$ verwerfen können.

[46] In der statistischen Praxis treten zahlreiche zusätzliche Probleme auf. Zum Beispiel gibt es meist weitere Variablen, *Kovariaten* genannt, die eine Rolle spielen können, wie Sozialstruktur der Bevölkerung, Inhomogenitäten der natürlichen Radioaktivität, unterschiedliches Ernährungsverhalten und viele mehr. Auch die Frage nach der *Kausalität* ist viel schwieriger zu beantworten als nur die Frage nach *Korrelationen*: Sind Emissionen von Kernkraftwerken die *Ursache* für beobachtete Krankheitsraten, oder gibt es vielleicht eine ganz andere Ursache, die nicht bedacht wurde?

Definition 3.11 (Hypothese) *Es sei* $(\Omega, \mathcal{A}, \mathcal{P})$ *ein statistisches Modell. Eine* Hypothese *ist eine Teilmenge* $H \subseteq \mathcal{P}$*,* $H \neq \emptyset$*. Beim Testproblem treten zwei Hypothesen auf: eine* Nullhypothese $H_0 \subseteq \mathcal{P}$ *und eine davon disjunkte* Alternativhypothese*, kurz* Alternative $H_1 \subseteq \mathcal{P}$*.*

Im obigen Beispiel lautet die Nullhypothese:

$$H_0 = \{\text{binomial}(n_1, p) \times \text{binomial}(n_2, p) |\ p \in [0, 1]\} \tag{811}$$

oder kurz, etwas informaler, geschrieben: "$p_1 = p_2$". Sie soll das *Nichtvorliegen*[47] eines Effekts von Kernkraftwerken auf die unbekannten Leukämiewahrscheinlichkeiten modellieren. Als Alternativhypothese wählen wir

$$H_1 = \{\text{binomial}(n_1, p_1) \times \text{binomial}(n_2, p_2) |\ 0 \leq p_2 < p_1 \leq 1\} \tag{812}$$

Sie spiegelt im Modell die Ausgangsfrage wider, ob die Erkrankungswahrscheinlichkeit in der Nähe von Kernkraftwerken *erhöht* ist, und nicht etwa, ob sie *erniedrigt* ist.

Obwohl sie in der Definition scheinbar eine symmetrische Rolle haben, spielen die Nullhypothese und die Alternative ganz verschiedene Rollen:

Interpretation der Nullhypothese:
Die Nullhypothese beschreibt ein *"einfaches Erklärungsmodell"* oder die *Abwesenheit* des Effekts, dessen Existenz man statistisch untersuchen möchte. Das Vorliegen eines Effekts zu belegen bedeutet also, die Nullhypothese zu *verwerfen.*

Die Alternative dient oft nur dazu, die Typen von Effekten zu spezifizieren, für die man sich interessiert, und auch dazu, die Qualität eines Tests zu messen.

Definition 3.12 (Test) *Ein (nichtrandomisierter)* statistischer Test *für die Nullhypothese* H_0 *wird durch einen Verwerfungsbereich* $V \in \mathcal{A}$ *gegeben. Liegen die Beobachtungsdaten* ω *im Verwerfungsbereich* V*, so* verwerfen *wir die Nullhypothese; andernfalls verwerfen wir sie nicht.*

Man beachte die Sprechweise: "*nicht verwerfen*" der Nullhypothese, nicht etwa "bestätigen"! Die Situation zwischen "verwerfen" und "nicht verwerfen" ist *nicht symmetrisch.* Wenn wir die Nullhypothese, z.B. $p_1 = p_2$, verwerfen, wollen wir "praktisch sicher" sein, dass sie falsch ist. Wenn wir sie nicht verwerfen, heißt das *nicht*, dass die Nullhypothese richtig ist, sondern es ist nur eine Art "Stimmenthaltung", "Unsicherheit": Die Daten reichen nicht aus, die "einfache Erklärung" H_0 zu widerlegen, oder die "Anwesenheit eines Effekts" zu belegen.[48]

[47] Daher der Name *Null*hypothese.

[48] Diese Interpretation ist konsistent mit den philosophischen Ideen von Karl Popper, dass wissenschaftliche Theorien *nie verifiziert* werden können, sondern höchstens *falsifiziert.* Man ist an Theorien interessiert, die viele Versuche, sie zu falsifizieren, überstehen.

Randomisierung. In Pattsituationen wirft man im täglichen Leben manchmal eine Münze, um eine Entscheidung herbeizuführen, führt also ein Zufallsexperiment aus, das zunächst nichts mit der eigentlichen Frage zu tun hat. Auch dieses Vorgehen hat in der Testtheorie ein Analogon:

Definition 3.13 (Randomisierung) *Ein* randomisierter *statistischer Test für die Nullhypothese H_0 wird durch eine $\mathcal{A}$-messbare Funktion $\varphi : \Omega \to [0,1]$ gegeben, Verwerfungsfunktion genannt. Gegeben Daten $\omega \in \Omega$, wählt man in einem unabhängigen Hilfszufallsexperiment eine uniform auf $[0,1]$ verteilte Zufallszahl r. Ist $r \leq \varphi(\omega)$, so verwirft man die Nullhypothese; andernfalls verwirft man sie nicht.*

Den Spezialfall, dass φ nur die Werte 0 und 1 annimmt, kann man als einen nichtrandomisierten Test mit Verwerfungsbereich $V = \{\omega \in \Omega |\ \varphi(\omega) = 1\}$ auffassen. In diesem Fall ist also $\varphi = 1_V$.
Man kann randomisierte Tests auch auf nichtrandomisierte Tests zurückführen, indem man das Ergebnis $r \in [0,1]$ des Hilfszufallsexperiments als eine weitere Komponente zu den Daten hinzunimmt. Formal gesagt: Aus einem randomisierten statistischen Test

$$\varphi : \Omega \to [0,1] \tag{813}$$

über einem Rahmenmodell $(\Omega, \mathcal{A}, \mathcal{P})$ wird ein nichtrandomisierter statistischer Test über dem Rahmenmodell $(\Omega \times [0,1], \mathcal{A} \otimes \mathcal{B}([0,1]), \mathcal{P}')$, wobei

$$\mathcal{P}' = \{P \times \text{uniform}[0,1] |\ P \in \mathcal{P}\} \tag{814}$$

mit dem Verwerfungsbereich

$$V = \{(\omega, r) \in \Omega \times [0,1] |\ r \leq \varphi(\omega)\}. \tag{815}$$

Fehlertypen bei Tests. Beim Testentscheid können zwei verschiedene Arten von Fehlern auftreten:

	H_0 wahr	H_0 falsch
H_0 nicht verwerfen	richtige Entscheidung	Fehler 2. Art
H_0 verwerfen	Fehler 1. Art	richtige Entscheidung

Wichtig!

Beispiel: Ein Feuermelder soll die Nullhypothese "Es brennt nicht" testen. Ein Fehler erster Art liegt vor, wenn der Feuermelder grundlos Alarm schlägt. Ein Fehler zweiter Art liegt vor, wenn das Haus brennt, aber der Feuermelder nicht alarmiert.

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 Verwerfungswahrscheinlichkeit $P_0(V)$[49] möglichst klein sein.
- Andererseits soll für alle $P_1 \in H_1$ die Nichtverwerfungswahrscheinlichkeit $P_1(V^c)$[50]

[49] bzw. $E_{P_0}[\varphi]$ im randomisierten Fall
[50] bzw. $1 - E_{P_1}[\varphi]$ im randomisierten Fall

möglichst klein sein.

Definition 3.14 (Risiken, Signifikanzniveau und Macht von Tests) *Das* Risiko erster Art *eines Tests ist die Wahrscheinlichkeit für den* Fehler erster Art *unter der* Nullhypothese*:*

$$P_0(V),^{51} \qquad (P_0 \in H_0) \tag{816}$$

Es kann von der Wahl von $P_0 \in H_0$ abhängen, wenn H_0 mehr als ein Element enthält. Das Risiko zweiter Art *eines Tests ist die Wahrscheinlichkeit β für den* Fehler zweiter Art *unter der* Alternative*:*

$$\beta = P_1(V^c) = 1 - P_1(V)^{52} \qquad (P_1 \in H_1). \tag{817}$$

Das Signifikanzniveau α*, abgekürzt "*Niveau*", englisch "*significance level*" oder kurz "*level*", ist das Supremum der Risiken 1. Art:* **Wichtig!**

$$\alpha = \sup_{P_0 \in H_0} P_0(V).^{53} \tag{818}$$

Die Macht $1 - \beta$ *ist die Wahrscheinlichkeit für die richtige Entscheidung unter der Alternative:*

$$\textit{Macht} = 1 - \beta = P_1(V) = 1 - \textit{Risiko 2. Art} \qquad (P_1 \in H_1) \tag{819}$$

Diese Begriffe werden besonders einfach, wenn H_0 und H_1 einelementig sind.

Definition 3.15 (einfache Hypothese) *Eine Hypothese $H \subseteq \mathcal{P}$ heißt* einfach*, wenn sie einelementig ist: $H = \{P\}$ für ein $P \in \mathcal{P}$. Andernfalls heißt sie* zusammengesetzt*.*

Für einfache Nullhypothesen $H_0 = \{P_0\}$ gilt also:

Signifikanzniveau α = Risiko 1. Art $= P_0(V)$

Beispiel: Im obigen Beispiel ist

$$H_0 = \{\text{binomial}(n_1, p) \times \text{binomial}(n_2, p) |\ 0 \le p \le 1\} \tag{820}$$

eine zusammengesetzte Hypothese, und

$$H_0 = \{\text{binomial}(n_1, 10^{-6}) \times \text{binomial}(n_2, 10^{-6})\} \tag{821}$$

eine einfache Hypothese.

[51] bzw. $E_{P_0}[\varphi]$ im randomisierten Fall
[52] bzw. $1 - E_{P_1}[\varphi]$ im randomisierten Fall
[53] $\sup_{P_0 \in H_0} E_{P_0}[\varphi]$ im randomisierten Fall

Interpretation des Testproblems. Die Minimalinterpretation von Wahrscheinlichkeiten

- Wahrscheinlichkeit nahe bei 1 bedeutet praktisch sicheres Eintreten,
- Wahrscheinlichkeit nahe bei 0 bedeutet praktisch unmögliches Eintreten,
- Wahrscheinlichkeit weder nahe bei 0 noch bei 1 bedeutet Unsicherheit

vgl. Seite 19, ist genau an die Interpretation von statistischen Tests angepasst; sie ist sogar dadurch motiviert:
Man wählt das Signifikanzniveau α so nahe bei 0, dass man Fehler erster Art als "praktisch unmöglich" ansieht.[54] Damit wird der Testentscheid "H_0 verwerfen" interpretiert als: "Es ist praktisch unmöglich, dass H_0 den Daten zugrundeliegt." Dementsprechend wird Testentscheid "H_0 nicht verwerfen" so interpretiert: "Die Nullhypothese H_0 kann nicht mit der nötigen Sicherheit ausgeschlossen werden."

3.3.1 Optimale Tests bei einfachen Hypothesen

Wir betrachten ein statistisches Modell $(\Omega, \mathcal{A}, \mathcal{P})$ und zwei einfache Hypothesen $H_0 = \{P_0\} \subseteq \mathcal{P}$ und $H_1 = \{P_1\} \subseteq \mathcal{P}$ mit Likelihoodquotienten dP_1/dP_0. Wir geben uns eine Schranke α_{crit} für Signifikanzniveaus vor und stellen folgendes Optimierungsproblem:

Unter allen Tests mit Signifikanzniveaus $\alpha = P_0(V) \leq \alpha_{\text{crit}}$ finde man denjenigen oder diejenigen mit größter Macht $1 - \beta = P_1(V)$.

Zur Veranschaulichung betrachten wir folgendes mathematisch analoges "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 mit einem möglichst großem Gesamtwert bepacken, ohne uns zu überladen?

Übersetzungstabelle:

Testproblem	**Rucksackproblem**
Stichprobenraum Ω	Menge der Güter
Verwerfungsbereich $V \subseteq \Omega$	Menge der Güter, die wir einpacken
Signifikanzniveau $\alpha = P_0(V)$	Gewicht der eingepackten Güter
Wahrscheinlichkeit unter H_0 von ω_i: $p_{0,i} = P_0(\{\omega_i\})$	Gewicht des Guts ω_i
Wahrscheinlichkeit unter H_1 von ω_i: $p_{1,i} = P_1(\{\omega_i\})$	Wert des Guts ω_i
Macht $1 - \beta = P_1(V)$	Wert der eingepackten Güter
Likelihoodquotient $\frac{dP_1}{dP_0}(\omega)$	Wert pro Gewicht des Guts ω

[54] In der statistischen Praxis nimmt man allerdings oft "unwahrscheinlicher als 5%" schon als "praktisch unmöglich", d.h. man toleriert noch vergleichsweise hohe Fehlerwahrscheinlichkeiten.

Beispiel: Nemen wir an, wir könnten 11,1 Kilogramm tragen, und wir haben die folgenden Güter zur Auswahl:[55]

Gut	Wert	Gewicht	Wert pro Gewicht
Gold	1000€	0,1kg	10000€/kg
Silber	500€	1kg	500€/kg
Eisen	100€	10kg	10€/kg
Steine	200€	10000kg	0,02€/kg

Intuitiv ist völlig klar, wie man den Rucksack bepacken soll:

- Man nimmt zuerst das Gold (10000€/kg).
- Kann man noch mehr tragen, dann nimmt man das Silber (500€/kg).
- Kann man immer noch mehr tragen, dann nimmt man das Eisen (10€/kg),
- und Steine nur dann, wenn man nicht vorher schon voll bepackt ist.

Anders gesagt: Wir nehmen nur Güter mit hohem spezifischen Wert, aber keine mit niedrigem spezifischen Wert.

Übersetzt ins Testproblem:

Satz 3.16 (Neyman-Pearson-Lemma – diskreter Fall) *Gegeben seien* $\Omega = \{\omega_1, \dots, \omega_n\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ *und zwei einfachen Hypothesen* $H_0 = \{P_0\}$ *und* $H_1 = \{P_1\}$, *wobei*

$$P_0 = \sum_{i=1}^{n} p_{0,i}\delta_{\omega_i}, \tag{822}$$

$$P_1 = \sum_{i=1}^{n} p_{1,i}\delta_{\omega_i} \tag{823}$$

mit positiven Wahrscheinlichkeiten $p_{0,i}, p_{1,i} > 0$. *Die* ω_i *seien nach absteigenden Likelihoodquotienten*

$$\frac{dP_1}{dP_0}(\omega_i) = \frac{p_{1,i}}{p_{0,i}} \tag{824}$$

angeordnet:

$$\frac{p_{1,1}}{p_{0,1}} \geq \frac{p_{1,2}}{p_{0,2}} \geq \dots \geq \frac{p_{1,n}}{p_{0,n}}. \tag{825}$$

Gegeben $k \in [n]$, *sei* T *der Test mit dem Verwerfungsbereich*

$$V = \{\omega_1, \dots, \omega_k\}. \tag{826}$$

Dann gilt für jeden Test T' mit einem Verwerfungsbereich $V' \subseteq \Omega$:

$$\text{Wenn } P_0(V') \leq P_0(V), \text{ so folgt } P_1(V') \leq P_1(V). \tag{827}$$

Anders gesagt ist T optimal im folgenden Sinn: **Wichtig!**

Jeder Test T' mit dem gleichen oder höchstens kleinerem Signifikanzniveau wie T hat eine kleinere oder höchstens gleiche Macht wie T.

Wir verallgemeinern diesen Satz sofort in mehrfacher Hinsicht:

- Einerseits lassen wir beliebige Stichprobenräume Ω statt nur endlicher Ω zu.
- Andererseits betrachten wir auch randomisierte Tests.

Zuerst der nichtrandomisierte Fall:

Satz 3.17 (Neyman-Pearson-Lemma – nichtrandomisierter Fall) *Es seien $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell und $H_0 = \{P_0\}$, $H_1 = \{P_1\}$ zwei einfache Hypothesen mit Likelihoodquotienten dP_1/dP_0. Weiter sei ein Test T mit einem Verwerfungsbereich V gegeben. gegeben. Es existiere ein $c \geq 0$, "*kritischer Wert*" genannt mit der Eigenschaft*

$$\left\{\frac{dP_1}{dP_0} > c\right\} \subseteq V \subseteq \left\{\frac{dP_1}{dP_0} \geq c\right\}. \tag{828}$$

Dann ist T im folgenden Sinn optimal: Jeder weitere Test T' mit Verwerfungsbereich V' **Wichtig!**
mit dem gleichen oder kleineren Signifikanzniveau

$$P_0(V') \leq P_0(V) \tag{829}$$

hat kleinere oder höchstens die gleiche Macht:

$$P_1(V') \leq P_1(V) \tag{830}$$

Nun zum allgemeinen Fall:

Satz 3.18 (Neyman-Pearson-Lemma – randomisierter Fall) *Es seien $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell und $H_0 = \{P_0\}$, $H_1 = \{P_1\}$ zwei einfache Hypothesen mit Likelihoodquotienten dP_1/dP_0. Weiter sei ein randomisierter Test T mit einer Verwerfungsfunktion $\varphi : \Omega \to [0,1]$ gegeben. Es existiere ein kritischer Wert $c \geq 0$ mit der Eigenschaft*

$$1_{\left\{\frac{dP_1}{dP_0} > c\right\}} \leq \varphi \leq 1_{\left\{\frac{dP_1}{dP_0} \geq c\right\}}. \tag{831}$$

Dann ist T im folgenden Sinn optimal: Jeder weitere randomisierter Test T' mit einer **Wichtig!**
Verwerfungsfunktion φ' mit dem gleichen oder kleinerem Signifikanzniveau

$$E_{P_0}[\varphi'] \leq E_{P_0}[\varphi] \tag{832}$$

hat kleinere oder höchstens die gleiche Macht:

$$E_{P_1}[\varphi'] \leq E_{P_1}[\varphi] \tag{833}$$

[55] Die angegebenen Preise dienen nur zur Veranschaulichung; sie sind nicht realistisch.

Der nichtrandomisierte Fall ist im randomisierten Fall als Spezialfall $\varphi = 1_V$, $\varphi' = 1_{V'}$ enthalten. Es genügt daher, den randomisierten Fall zu beweisen:
Beweis: Wir zeigen zuerst für alle $\omega \in \Omega$:

$$(\varphi(\omega) - \varphi'(\omega)) \left(\frac{dP_1}{dP_0}(\omega) - c \right) \geq 0. \tag{834}$$

Gegeben $\omega \in \Omega$, unterscheiden wir hierzu drei Fälle:

- Ist

$$\frac{dP_1}{dP_0}(\omega) - c = 0, \tag{835}$$

  so ist die linke Seite in (834) gleich 0.

- Ist

$$\frac{dP_1}{dP_0}(\omega) - c < 0, \tag{836}$$

  so gilt $\varphi(\omega) \leq 0$[56] nach der Voraussetzung (831) und daher

$$\varphi(\omega) - \varphi'(\omega) \leq -\varphi'(\omega) \leq 0, \tag{837}$$

  so dass beide Faktoren auf der linken Seite in (834) ≤ 0 und damit ihr Produkt ≥ 0 ist.

- Ist

$$\frac{dP_1}{dP_0}(\omega) - c > 0, \tag{838}$$

  so gilt $\varphi(\omega) \geq 1$[57] nach der Voraussetzung (831) und daher

$$\varphi(\omega) - \varphi'(\omega) \geq 1 - \varphi'(\omega) \geq 0, \tag{839}$$

  so dass beide Faktoren auf der linken Seite in (834) ≥ 0 und damit ihr Produkt wieder ≥ 0 ist.

[56] Genauer gesagt gilt hier sogar Gleichheit, doch das brauchen wir hier nicht.
[57] Genauer gesagt gilt auch hier Gleichheit.

Wir bilden die Erwartung bezüglich P_0 in (834):[58]

$$\begin{aligned}
0 &\leq E_{P_0}\left[(\varphi-\varphi')\left(\frac{dP_1}{dP_0}-c\right)\right] \\
&= E_{P_0}\left[(\varphi-\varphi')\frac{dP_1}{dP_0}\right] - cE_{P_0}\left[\varphi-\varphi'\right] \\
&= E_{P_1}\left[\varphi-\varphi'\right] - cE_{P_0}\left[\varphi-\varphi'\right] \\
&\leq E_{P_1}\left[\varphi-\varphi'\right] \text{ (wegen } c\geq 0 \text{ und Voraussetzung (832))} \\
&= E_{P_1}[\varphi] - E_{P_1}[\varphi'] \qquad (841)
\end{aligned}$$

Es folgt die Behauptung (833).

□

Definition 3.19 (Likelihood-Quotienten-Tests) *Tests mit einem Verwerfungsbereich V mit der Eigenschaft* (828)[59] *heißen* Likelihood-Quotienten-Tests*, synonym* Neyman-Pearson-Tests.

Bemerkung: Hat der Likelihood-Quotient dP_1/dP_0 eine kontinuierliche Verteilung unter P_0, so gilt

$$P_0\left[\frac{dP_1}{dP_0}=c\right]=0 \qquad (842)$$

und damit auch

$$P_1\left[\frac{dP_1}{dP_0}=c\right]=0 \qquad (843)$$

In diesem Fall spielt es keine Rolle, ob man

$$V=\left\{\frac{dP_1}{dP_0}>c\right\} \text{ oder } \quad V=\left\{\frac{dP_1}{dP_0}\geq c\right\} \qquad (844)$$

oder etwas dazwischen wählt. Insbesondere in diskreten Modellen kann jedoch

$$P_0\left[\frac{dP_1}{dP_0}=c\right]>0 \qquad (845)$$

[58] Bei der Rechnung verwenden wir die Rechenregel

$$E_{P_0}\left[X\frac{dP_1}{dP_0}\right]=\int_\Omega X\frac{dP_1}{dP_0}\,dP_0=\int_\Omega X\,dP_1=E_{P_1}[X], \qquad (840)$$

die für jede Zufallsvariable X gilt, für die diese Erwartung existiert.

[59] bzw. im randomisierten Fall mit einer Verwerfungsfunktion mit der Eigenschaft (831)

gelten. Um ein vorgegebenes Signifikanzniveau genau zu treffen, kann es sinnvoll sein, V echt zwischen $\left\{\frac{dP_1}{dP_0} > c\right\}$ und $\left\{\frac{dP_1}{dP_0} \geq c\right\}$ und wenn nötig eine Randomisierung vorzunehmen.

Beispiel (Analogie): Im Rucksackproblem von oben nehmen wir an, dass wir 5kg tragen können. Es ist dann optimal, dass wir alles Gold (0.1kg), alles Silber (1kg), aber nur einen Teil des Eisens (3,9kg von 10 kg) einzupacken. Dem "Zerteilen" des Eisens in der Analogie entspricht dann dem "Zerteilen" des Ereignisses $\left\{\frac{dP_1}{dP_0} = c\right\}$ mittels Randomisierung. In der Praxis beim Testen wird es zwar nur selten angewandt, in anderen Zusammenhängen dagegen schon.
In der Praxis verwendet man meist den Kehrwert $\frac{dP_0}{dP_1}$ von $\frac{dP_1}{dP_0}$. Das macht aber keinen Unterschied:

$$\left\{\frac{dP_0}{dP_1} < c^{-1}\right\} = \left\{\frac{dP_1}{dP_0} > c\right\}. \tag{846}$$

Beispiel: Es seien $X_1, \ldots, X_n$ i.i.d. normalverteilte Daten mit unbekannter Erwartung μ und bekannter Varianz $\sigma^2 = 1$. Wir verwenden also das Rahmenmodell

$$\Omega = \mathbb{R}^n, \tag{847}$$
$$\mathcal{A} = \mathcal{B}(\mathbb{R}^n), \tag{848}$$
$$\mathcal{P} = \{\mathrm{N}(\mu, 1)^n |\ \mu \in \mathbb{R}\} \tag{849}$$

mit den kanonischen Projektionen $X_1, \ldots, X_n : \mathbb{R}^n \to \mathbb{R}$. Wir entwerfen jetzt einen Likelihood-Quotienten-Test zum Signifikanzniveau α für die Nullhypothese H_0: $\mu = 0$, also

$$H_0 = \{\mathrm{N}(0, 1)^n\} = \{P_0\}, \tag{850}$$

bei der Alternative H_1: $\mu = \mu_1$, also

$$H_1 = \{\mathrm{N}(\mu_1, 1)^n\} = \{P_1\} \tag{851}$$

mit gegebenem $\mu_1 \neq 0$.
Das Maß P_0 hat die Dichte

$$f_0(x) = \frac{dP_0}{d\lambda_n}(x) = \prod_{j=1}^{n} \frac{e^{-\frac{1}{2}x_j^2}}{\sqrt{2\pi}} = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\|x\|_2^2}, \quad (x = (x_1, \ldots, x_n) \in \mathbb{R}^n), \tag{852}$$

und ebenso P_1 die Dichte

$$f_1(x) = \frac{dP_1}{d\lambda_n}(x) = \prod_{j=1}^{n} \frac{e^{-\frac{1}{2}(x_j-\mu_1)^2}}{\sqrt{2\pi}} = (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(x_j - \mu_1)^2\right\}. \tag{853}$$

Wir erhalten den Likelihood-Quotienten

$$\frac{dP_1}{dP_0}(x) = \frac{f_1(x)}{f_0(x)} = \exp\left\{\frac{1}{2}\sum_{j=1}^{n}[x_j^2 - (x_j-\mu_1)^2]\right\} = e^{-\frac{n\mu_1^2}{2}}\exp\left\{\mu_1\sum_{j=1}^{n}x_j\right\}. \tag{854}$$

Man beachte, dass man nicht alle Datenpunkte $X_1, \ldots, X_n$ einzeln kennen muss, um den Likelihood-Quotienten $\frac{dP_1}{dP_0}$ zu berechnen; in diesem Beispiel genügt die Summe

$$S := \sum_{j=1}^{n} X_j. \tag{855}$$

Wir bestimmen jetzt die Niveaumengen

$$V = \left\{\frac{dP_1}{dP_0} > c\right\}, \tag{856}$$

die als Verwerfungsbereich im Neyman-Pearson-Test auftreten. Hierzu unterscheiden wir zwei Fälle:

1. Fall: Für $\mu_1 > 0$ ist die Abbildung $s \mapsto e^{-\frac{n\mu_1^2}{2}}\exp\{\mu_1 s\}$ monoton steigend, also erhalten wir für jedes $s \in \mathbb{R}$ mit dem zugehörigen kritischen Wert

$$c = \exp\{-\frac{n\mu_1^2}{2} + \mu s\} \tag{857}$$

den Verwerfungsbereich

$$V = \left\{\frac{dP_1}{dP_0} > c\right\} = \{S > s\}. \tag{858}$$

Um ein bestimmtes Signifikanzniveau α zu realisieren, wählen wir s so, dass

$$P_0[S > s] = \alpha \tag{859}$$

gilt. Nun folgt aus $\mathcal{L}_{P_0}(X_1, \ldots, x_n) = \mathrm{N}(0,1)^n$ mit der Faltungseigenschaft der Normalverteilung, Satz 2.58, die Aussage

$$\mathcal{L}_{P_0}(S) = \mathrm{N}(0, n), \tag{860}$$

also

$$\mathcal{L}_{P_0}\left(\frac{S}{\sqrt{n}}\right) = \mathrm{N}(0,1). \tag{861}$$

Bezeichnen wir mit $\Phi : \mathbb{R} \to [0,1]$ die Verteilungsfunktion der Standardnormalverteilung $\mathrm{N}(0,1)$, also

$$\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{x^2}{2}}\,dx, \tag{862}$$

so folgt

$$P_0[S > s] = P_0\left[\frac{S}{\sqrt{n}} > \frac{s}{\sqrt{n}}\right] = 1 - P_0\left[\frac{S}{\sqrt{n}} \leq \frac{s}{\sqrt{n}}\right] = 1 - \Phi\left(\frac{s}{\sqrt{n}}\right) \tag{863}$$

und damit die Bedingung

$$\alpha = 1 - \Phi\left(\frac{s}{\sqrt{n}}\right) \tag{864}$$

oder äquivalent

$$s = \sqrt{n}\Phi^{-1}(1-\alpha). \tag{865}$$

Mit der Abkürzung

$$Z = \frac{S}{\sqrt{n}} \tag{866}$$

erhalten wir daher den Verwerfungsbereich

$$V = \{S > \sqrt{n}\Phi^{-1}(1-\alpha)\} = \{Z > \Phi^{-1}(1-\alpha)\}. \tag{867}$$

2. Fall: Für $\mu_1 < 0$ ist die Abbildung $s \mapsto e^{-\frac{n\mu_1^2}{2}} \exp\{\mu_1 s\}$ monoton fallend, also diesmal

$$V = \left\{\frac{dP_1}{dP_0} > c\right\} = \{S < s\} \tag{868}$$

statt (858). Analog zu oben gilt

$$\alpha = P_0(V) = P_0[S < s] = \Phi\left(\frac{s}{\sqrt{n}}\right), \tag{869}$$

und wir erhalten diesmal den Verwerfungsbereich

$$V = \{S < \sqrt{n}\Phi^{-1}(\alpha)\} = \{Z < \Phi^{-1}(\alpha)\}. \tag{870}$$

Für den Testentscheid ist also in beiden Fällen nur der Wert von Z nötig.

Definition 3.20 (Teststatistik, Statistik) *Eine Zufallsvariable $T : \Omega \to \mathbb{R}$ (oder allgemeiner ein Zufallsvektor T), der den Verwerfungsbereich V_α für jede Wahl des Signifikanzniveaus α bestimmt, wird* Teststatistik *genannt. Allgemeiner heißt eine vom Statistiker gewählte messbare Abbildung $X : \Omega \to \mathbb{R}^d$, die den Daten Zahlenwerte zuordnet, eine* Statistik.

Im Synonym "*Statistik*" für "*Zufallsvariable*" schwingt also stets zusätzlich die Konnotation mit, dass die Abbildung *vom Statistiker zu einem bestimmten Zweck* gewählt wurde, z.B. um einen Testentscheid darauf aufzubauen.

Beispiel (Fortsetzung): Die Teststatistik

$$Z = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j, \tag{871}$$

vgl. Formel (866), im obigen Beispiel ist unter der Nullhypothese, Maß P_0, standardnormalverteilt und unter der Alternativhypothese, Maß P_1, $\mathrm{N}(\sqrt{n}\mu_1, 1)$-verteilt. Ihr Wert genügt, um den Likelihood-Quotienten-Test im Beispiel durchzuführen.

Zur Wahl von Teststatistiken ist der folgende Begriff der *suffizienten Statistik* sehr nützlich:

Definition 3.21 (suffiziente Statistik) *Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell, so dass alle $P_0, P_1 \in \mathcal{P}$ eine Dichte dP_1/dP_0 zueinander besitzen, z.B. ein dominiertes Modell mit positiver Likelihoodfunktion. Eine Statistik $X : \Omega \to \mathbb{R}^d$ heißt* suffizient *für $(\Omega, \mathcal{A}, \mathcal{P})$, wenn es für alle P_0, P_1 eine Borel-messbare Funktion $f : \mathbb{R}^d \to \mathbb{R}$ mit der Eigenschaft*

$$\frac{dP_1}{dP_0} = f(X) \quad P_0\text{-fast sicher} \tag{872}$$

gibt, also der Likelihood-Quotient $\frac{dP_1}{dP_0}$ nur von X abhängt.

Anschaulich gesprochen enthält eine suffiziente Statistik alle für statistische Schlüsse im Modell *relevanten* Informationen. Zum Beispiel genügt offensichtlich der Wert $X(\omega)$ einer suffizienten Statistik X (zusammen mit einer von den Daten unabhängigen, uniform auf $[0, 1]$ verteilten Zufallszahl r, falls randomisiert werden soll) zur Ausführung eines Likelihood-Quotienten-Tests.

Lemma 3.22 (Kriterium für Suffizienz) *Es sei $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ ein parametrisches statistisches Modell mit dominierendem Maß μ und positiver Likelihoodfunktion $L > 0$, $L : \Omega \times \Theta \to \mathbb{R}^+$. Weiter sei $X : \Omega \to \mathbb{R}^d$ eine Statistik und*

$$g : \mathbb{R}^d \times \Theta \to \mathbb{R}, \tag{873}$$

$$h : \Omega \to \mathbb{R}^+ \tag{874}$$

messbare Abbildungen mit

$$L(\omega, \theta) = g(X(\omega), \theta)h(\omega) \tag{875}$$

für alle $\omega \in \Omega$ und $\theta \in \Theta$. Dann ist X eine suffiziente Statistik.

Beweis: Für alle $\theta_1, \theta_2 \in \Theta$ und $\omega \in \Omega$ gilt:

$$\frac{dP_{\theta_1}}{dP_{\theta_2}}(\omega) = \frac{L(\omega, \theta_1)}{L(\omega, \theta_2)} = \frac{g(X(\omega), \theta_1)}{g(X(\omega), \theta_2)}. \tag{876}$$

Hier tritt ω nur in der Form $X(\omega)$ auf.

□

Beispiel: Wir betrachten n i.i.d. normalverteilte Datenpunkte mit unbekannter Erwartung μ und unbekannter Varianz σ^2. Formaler gesagt betrachten wir das folgende statistische Modell:

$$\Omega = \mathbb{R}^n \tag{877}$$
$$\mathcal{A} = \mathcal{B}(\Omega) \tag{878}$$
$$\mathcal{P} = \{\mathrm{N}(\mu, \sigma^2)^n |\ \mu \in \mathbb{R},\ \sigma^2 > 0\}. \tag{879}$$

Wir erhalten die Likelihood-Funktion

$$\begin{aligned} &L : \mathbb{R}^n \times (\mathbb{R} \times \mathbb{R}^+) \to \mathbb{R}, \\ &L(x_1, \ldots, x_n;\, \mu, \sigma^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_j - \mu)^2\right) \\ &\qquad = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2\right). \end{aligned} \tag{880}$$

Nun gilt, wenn $\overline{x}$ das Mittel der Datenpunkte $x_1, \ldots x_n$ und s_x^2 ihre empirische Varianz bezeichnet:

$$\begin{aligned} \sum_{j=1}^{n}(x_j - \mu)^2 &= \sum_{j=1}^{n}[(x_j - \overline{x}) + (\overline{x} - \mu)]^2 \\ &= \sum_{j=1}^{n}(x_j - \overline{x})^2 + 2\underbrace{\sum_{j=1}^{n}(x_j - \overline{x})(\overline{x} - \mu)}_{=0} + n(\overline{x} - \mu)^2 \\ &= (n-1)s_x^2 + n(\overline{x} - \mu)^2, \end{aligned} \tag{881}$$

wobei wir

$$\sum_{j=1}^{n}(x_j - \overline{x})(\overline{x} - \mu) = (\overline{x} - \mu)\left(\sum_{j=1}^{n} x_j - n\overline{x}\right) = 0 \tag{882}$$

verwendet haben. Es folgt:

$$L(x;\, \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s_x^2 + n(\overline{x} - \mu)^2\right]\right\}. \tag{883}$$

Hier treten die Datenpunkte $x_1, \ldots, x_n$ nur in der Kombination s_x^2 und $\overline{x}$ auf. Die zweidimensionale Statistik $(x_1, \ldots, x_n) \mapsto (\overline{x}, s_x^2)$ ist also suffizient.

Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell und $T : \Omega \to \mathbb{R}^d$ 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

$$\frac{dP_1}{dP_0} = f(T). \tag{884}$$

Damit werden alle Tests mit Verwerfungsbereichen der Gestalt

$$V_c = \{f(T) \leq c\} \text{ oder auch } V_c = \{f(T) < c\} \quad (c \in \mathbb{R}^+) \tag{885}$$

optimal in dem Sinn, dass sie maximale Macht bei gegebenem Signifikanzniveau besitzen.

3.3.2 Variable Signifikanzniveaus und p-Wert

Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell, $H_0 \subseteq \mathcal{P}$ eine Nullhypothese und $(V_c)_{c \in I}$ eine Familie von Verwerfungsbereichen von Tests, indiziert mit einer Menge $I \subseteq \mathbb{R}$ und monoton steigend im Index c:

$$V_c \subseteq V_{c'} \text{ für } c \leq c'. \tag{886}$$

Eine typische Situation ist

$$V_c = \{f(T) \leq c\} \tag{887}$$

mit einer Teststatistik T, z.B. $f(T) = \frac{dP_0}{dP_1}$. Die Gesamtheit aller Testentscheidungen zu einer Stichprobe wird mit dem folgenden Begriff des p-Werts zusammengefasst:

Definition 3.23 (p-Wert) *Gegeben Beobachtungsdaten $\omega \in \Omega$, definieren wir den* p-Wert *zu der Familie $(V_c)_{c \in I}$ von Verwerfungsbereichen als das Infimum aller Signifikanzniveaus, auf dem die Hypothese H_0 noch verworfen werden kann. In Formeln:*

$$p(\omega) := \inf\{\alpha_c |\ c \in I,\ \omega \in V_c\},\ \textit{wobei}\ \alpha_c := \sup_{P_0 \in H_0} P_0(V_c) \tag{888}$$

wichtig!

das Signifikanzniveau des Tests mit Verwerfungsbereich V_c bezeichnet.

Bemerkung: Gilt $H_0 = \{P_0\}$ und $V_c = \bigcap_{c' > c} V_{c'}$, so ist das Infimum sogar ein Minimum. Das ist zum Beispiel für $V_c = \{T \leq c\}$ der Fall. In diesem Fall kann man den p-Wert einfacher so charakterisieren:

Der p-Wert ist das kleinste Signifikanzniveau, auf dem die Nullhypothese bei den gegebenen Daten noch verworfen werden kann.

Beispiel: Sind $X_1, \ldots, X_n$ i.i.d. normalverteilt und testen wir die Nullhypothese

$$H_0 = \{P_0\} = \{\mathrm{N}(\mu_0, \sigma^2)^n\} \tag{889}$$

gegen die Alternative

$$H_1 = \{P_1\} = \{\mathrm{N}(\mu_1, \sigma^2)^n\} \tag{890}$$

mit gegebenen Parametern $\mu_1 < \mu_0$ und $\sigma^2 > 0$, so haben die Tests mit den Verwerfungsbereichen

$$V_c = \{Z \leq c\}, \quad c \in \mathbb{R} \tag{891}$$

mit der Teststatistik

$$Z = \sqrt{n}\frac{\overline{X} - \mu_0}{\sigma} \tag{892}$$

maximale Macht bei gegebenem Signifikanzniveau. Nun gilt

$$\mathcal{L}_{P_0}(Z) = \mathrm{N}(0, 1), \tag{893}$$

d.h. die Teststatistik Z ist unter der Nullhypothese standardnormalverteilt. Gegeben Daten $X_1(\omega), \ldots, X_n(\omega)$, erhalten wir den p-Wert

$$p(\omega) = \inf\{P_0[Z \leq c] |\ Z(\omega) \leq c\} = P_0[Z \leq Z(\omega)]. \tag{894}$$

Allgemeiner: Gegeben seien eine Teststatistik $T : \Omega \to \mathbb{R}$ mit der gleichen Verteilung $Q = \mathcal{L}_{P_0}(T)$ für alle $P_0 \in H_0$, sowie Verwerfungsbereiche

$$V_c = \{T \leq c\}, \quad c \in \mathbb{R}. \tag{895}$$

Dann wird der p-Wert durch

$$p(\omega) = P_0[T \leq T(\omega)] = Q(-\infty, T(\omega)]), \quad (\omega \in \Omega, P_0 \in H_0) \tag{896}$$

gegeben.
Der p-Wert codiert also die Testentscheidung bei variablem Signifikanzniveau α:

Codierung des Testentscheids im p-Wert:

- Wenn $p(\omega) \leq \alpha$, so *verwerfen* wir H_0 zum Signifikanzniveau α.
- Wenn $p(\omega) > \alpha$, so *verwerfen* wir H_0 zum Signifikanzniveau α *nicht*.

Wichtig!

Der p-Wert $p : \Omega \to [0, 1]$ hier also selbst die Teststatistik einer Familie von Tests mit Verwerfungsbereichen $\{p \leq \alpha\}$, $0 < \alpha < 1$, In diesem Fall kann man daher als kritische Werte α die Signifikanzniveaus α selbst nehmen.

Zusammenfassend anschaulich gesagt:

Interpretation von p-Werten:

Der p-Wert ist die Wahrscheinlichkeit unter der Nullhypothese, dass das Experiment die tatsächlich beobachteten Daten ω oder ein noch extremeres Ergebnis liefert. Für zusammengesetzte Hypothesen H_0 gilt das nur, falls diese Wahrscheinlichkeit nicht von der Wahl von $P_0 \in H_0$ abhängt. Dabei wird durch die Alternativhypothese vorgegeben, was mit "noch extremer" gemeint sein soll.

wichtig!

Übung 3.24 (Uniforme Verteilung des p-Werts unter der Nullhypothese) *Nehmen Sie in der Situation von eben an, dass die Verteilung Q der Teststatistik T unter H_0 atomlos sei, d.h. $Q(\{a\}) = 0$ für alle $a \in \mathbb{R}$. Zeigen Sie, dass dann der p-Wert unter allen $P_0 \in H_0$ uniform auf $[0,1]$ verteilt ist.*

Beispiel: unfairer Münzwurf. Gegeben sei das statistische Modell für den unfairen Münzwurf

$$\Omega = \{0,1\}^n, \quad \mathcal{A} = \mathcal{P}(\Omega), \quad \mathcal{P} = \{P_p = (p\delta_1 + (1-p)\delta_0)^p |\ 0 < p < 1\}. \tag{897}$$

Die Statistik

$$S : \Omega \to \{0, \dots, n\}, \quad S(\omega) = \sum_{k=1}^{n} \omega_k \tag{898}$$

ist suffizient, und wir haben die Likelihoodquotienten

$$\frac{dP_{p_1}}{dP_{p_0}} = \left(\frac{p_1}{p_0}\right)^S \left(\frac{1-p_1}{1-p_0}\right)^{n-S}, \tag{899}$$

die für $p_1 < p_0$ monoton in S fallen. Gegeben eine Nullhypothese $H_0 = \{P_{p_0}\}$ und eine Alternative $H_1 = \{P_{p_1}\}$ mit $p_1 < p_0$, gehören also alle Verwerfungsbereiche der Gestalt

$$V_c := \{S \le c\}, \quad (c \in \{0, \dots, n\}) \tag{900}$$

zu Neyman-Pearson-Tests. Bezüglich dieser Testfamilie lautet der p-Wert zu einer Stichprobe $\omega \in \Omega$ also

$$\begin{aligned} p(\omega) &= \min\{P_{p_0}(V_c) |\ c \in \{0, \dots, n\},\ S(\omega) \le c\} \\ &= P_{p_0}(V_{S(\omega)}) \\ &= \text{binomial}(n, p_0)(\{0, \dots, S(\omega)\}) \\ &= \sum_{k=0}^{S(\omega)} \binom{n}{k} p_0^k (1-p_0)^{n-k}. \end{aligned} \tag{901}$$

Seine Verteilung unter der Nullhypothese lautet

$$\mathcal{L}_{P_{p_0}}(p) = \sum_{s=0}^{n} \text{binomial}(n, p_0)(\{s\}) \delta_{\text{binomial}(n,p_0)(\{0,\dots,s\})}. \tag{902}$$

Die zugehörige Verteilungsfunktion ist eine Treppenfunktion, die die Verteilungsfunkton der Gleichverteilung auf $[0,1]$ im Limes $n \to \infty$ approximiert.

Warnung vor falschen Interpretationen mit p-Werten! Obwohl er so praktisch zur Codierung vieler Testentscheide gleichzeitig ist, ist es nicht leicht, eine korrekte anschauliche Vorstellung vom Begriff des p-Werts zu bekommen und ihn richtig zu interpretieren. Seine so häufige falsche Verwendung und fehlerhafte Interpretation in den Anwendungswissenschaften und sogar in Lehrbüchern[60] war der Anlass für eine kritische Stellungnahme der *American Statistical Association (ASA)*[61] aus der hier einige Prinzipien zitiert werden sollen:

- *P-values can indicate how incompatible the data are with a specified statistical model.*
  (P-Werte können anzeigen, wie inkompatibel die Daten mit einem spezifischen statistischen Modell sind.)

- *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
  (P-Werte messen nicht die Wahrscheinlichkeit, dass die betrachtete Hypothese wahr ist[62], oder die Wahrscheinlichkeit, dass die Daten nur durch reinen Zufall entstanden sind.)

- *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
  (Wissenschaftliche Schlüsse oder Geschäftsentscheidungen oder politische Entscheidungen sollten nicht nur darauf gründen, ob ein p-Wert einen spezifischen Wert unter- oder überschreitet.)

- *Proper inference requires full reporting and transparency.*
  (Richtige Schlüsse erfordern vollständige Dokumentation und Transparenz.)

- *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
  (Ein p-Wert, oder statistische Signifikanz, misst nicht die Größe eines Effekts oder die Bedeutung eines Resultats.)

3.3.3 Konfidenzbereiche und Dualität

Konfidenzbereiche sind eine Art "Parameterschätzer mit Toleranzangabe". Gegeben Beobachtungen $\omega \in \Omega$, möchte man eine Menge $C(\omega)$ von "plausiblen" Parametern auszeichnen.

[60] dokumentiert in S. Cassidy, R. Dimova, B. Gigure, J. Spence, D. Stanley: *Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly.* In: Advances in methods and practices in psychological science. Juni 2019, doi:10.1177/2515245919858072.

[61] *Referenz:* Ronald L. Wasserstein & Nicole A. Lazar (2016) "*The ASA Statement on p-Values: Context, Process, and Purpose*", The American Statistician, 70:2, 129-133, DOI:10.1080/00031305.2016.1154108

[62] In der Tat: In einer frequentistischen Modellierung ergibt es gar keinen Sinn, von der Wahrscheinlichkeit, dass die Hypothese wahr ist, zu sprechen. Schon eine solche Wahrscheinlichkeit hinzuschreiben wäre ein Typfehler.

Definition 3.25 (Konfidenzbereich) *Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell und $\theta : \mathcal{P} \to \mathbb{R}^d$ ein Parameter.*[63] *Weiter sei $\alpha \in]0,1[$ ein gegebenes Signifikanzniveau. Eine Familie $(C(\omega))_{\omega\in\Omega}$ von Mengen $C(\omega) \subseteq \mathbb{R}^d$ heißt* Konfidenzbereich*, synonym* Vertrauensbereich*, zum Vertrauensniveau $1-\alpha$, kurz $(1-\alpha)$-Konfidenzbereich, wenn gilt:* **wichtig!**

- *Für alle $P \in \mathcal{P}$ ist $\{\omega \in \Omega |\ \theta(P) \in C(\omega)\} \in \mathcal{A}$.*
- *Für alle $P \in \mathcal{P}$ gilt*

$$P(\{\omega \in \Omega |\ \theta(P) \in C(\omega)\}) \geq 1-\alpha. \tag{903}$$

Dasselbe in Kurznotation:

$$\forall P \in \mathcal{P} :\ P[\theta(P) \in C] \geq 1-\alpha. \tag{904}$$

Bemerkungen:

- Man beachte: Hier ist der Wert $C(\omega)$ des Vertrauensbereichs zufällig, da von der bekannten, aber zufälligen Stichprobe ω abhängig, während das unbekannte Wahrscheinlichkeitsmass P allquantifiziert wird und *nicht zufällig* ist.
- Konfidenzbereiche zu gegebenem Vertrauensniveau $1-\alpha$ sind dann am interessantesten, wenn sie möglichst klein sind, ähnlich wie Fehlerschranken in der Numerik dann am interessantesten sind, wenn sie möglichst klein sind.[64] Man fordert daher meist, dass die Konfidenzbereiche möglichst klein sein sollen.
- Im eindimensionalen Fall $\theta : \mathcal{P} \to \mathbb{R}$ nimmt man oft ein Intervall $C(\omega)$ als Konfidenzbereich. Es heißt dann *Konfidenzintervall* oder *Vertrauensintervall*.

Das folgend Lemma stellt die Verbindung zwischen Konfidenzbereichen und Tests her:

Lemma 3.26 (Dualität zwischen Tests und Konfidenzbereichen) *Es sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell, $\theta : \mathcal{P} \to \Theta \subseteq \mathbb{R}^d$ ein Parameter und $K \subseteq \Omega \times \Theta$ so, dass für alle $P \in \mathcal{P}$ gilt:*

$$\{\omega \in \Omega |\ (\omega, \theta(P)) \in K\} \in \mathcal{A}. \tag{905}$$

Weiter sei eine Zahl $0 < \alpha < 1$ gegeben. Dann sind äquivalent: **wichtig!**

1. *Durch*

$$C(\omega) := \{q \in \Theta |\ (\omega, q) \in K\}, \quad (\omega \in \Omega) \tag{906}$$

wird ein $(1-\alpha)$-Konfidenzbereich gegeben.

[63] Im parametrischen Fall $\mathcal{P} = \{P_q |\ q \in \Theta\}$ kann man zum Beispiel $\theta(P_q) = q$ wählen.

[64] Die triviale Wahl $C(\omega) = \mathbb{R}^d$ für alle $\omega \in \Omega$ ist zwar zulässig, aber völlig nutzlos, ähnlich wie eine Fehlerschranke der Art "der Fehler ist $\leq \infty$" in der Numerik zulässig, aber völlig nutzlos ist.

2. Für jedes $q \in \Theta$ mit

$$H_0(q) := \{P \in \mathcal{P} |\ \theta(P) = q\} \neq \emptyset \tag{907}$$

ist die Menge

$$V(q) := \{\omega \in \Omega |\ (\omega, q) \notin K\} \tag{908}$$

der Verwerfungsbereich eines Tests der Hypothese zu einem Signifikanzniveaus $\leq \alpha$.

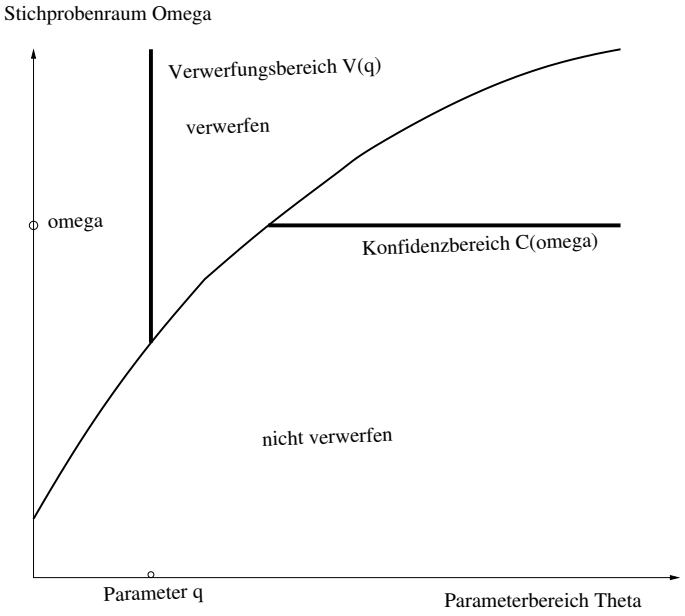
Beweis:

$$\begin{aligned}
\text{Aussage 1.} \quad &\Leftrightarrow \quad \forall P \in \mathcal{P}:\ P[\theta(P) \in C] \geq 1 - \alpha \\
&\Leftrightarrow \quad \forall P \in \mathcal{P}:\ P(\{\omega \in \Omega |\ (\omega, \theta(P)) \in K\}) \geq 1 - \alpha \qquad \text{(wegen (906))} \\
&\Leftrightarrow \quad \forall P \in \mathcal{P}:\ P(\{\omega \in \Omega |\ (\omega, \theta(P)) \notin K\}) \leq \alpha \\
&\Leftrightarrow \quad \forall P \in \mathcal{P}:\ P(V(\theta(P))) \leq \alpha \qquad \text{(wegen (908))} \\
&\Leftrightarrow \quad \forall q \in \Theta\ \forall P_0 \in H_0(q):\ P_0(V(q))) \leq \alpha \qquad \text{(wegen (907))} \\
&\Leftrightarrow \quad \text{Aussage 2.}
\end{aligned} \tag{909}$$

□

Man kann also die Angabe eines Vertrauensbereichs als die Durchführung vieler Tests gleichzeitig auffassen, wobei die Nullhypothese H_0 variiert, aber eine obere Schranke α für das Signifikanzniveau fixiert ist.
Äquivalent sind also:

- Der Parameter q liegt nicht im Vertrauensbereich $C(\omega)$ bei beobachteter Stichprobe ω;
- Der Test mit Verwerfungsberich $V(q)$ verwirft die Hypothese H_0: "$\theta(P) = q$" bei der beobachteten Stichprobe ω.



Beispiel (i.i.d. normalverteilte Daten): Es seien $X_1, \ldots, X_n$ i.i.d. normalverteilt mit unbekannter Erwartung μ und bekannter Varianz σ^2 und $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$ ihr Mittelwert. Weiter sei $0 < \alpha < 1$ gegeben. Dann ist für jedes $\mu \in \mathbb{R}$ die Menge[65]

$$V(\mu) := \left\{ \sqrt{n} \frac{\overline{X} - \mu}{\sigma} < \Phi^{-1}(\alpha) \right\} \tag{910}$$

der Verwerfungsbereich zum Signifikanzniveau α eines Tests der Hypothese H_0: "Erwartung μ", formaler gesagt

$$H_0 = \{P \in \mathcal{P} : \ E_P[X_1] = \mu\}. \tag{911}$$

Mit dem Dualitätslemma 3.26 schließen wir, dass für zufällige Stichproben $\omega \in \Omega$ die Menge

$$\begin{aligned} C(\omega) &= \{\mu \in \mathbb{R} |\ \omega \notin V(\mu)\} \\ &= \left\{ \mu \in \mathbb{R} \,|\ \sqrt{n} \frac{\overline{X}(\omega) - \mu}{\sigma} \geq \Phi^{-1}(\alpha) \right\} \end{aligned} \tag{912}$$

ein $(1-\alpha)$-Vertrauensbereich für den unbekannten Parameter μ ist. Man beachte, dass für kleine α gilt: $\Phi^{-1}(\alpha) < 0$, also $\overline{X}(\omega) \in C(\omega)$. Der Schätzwert $\overline{X}(\omega)$ liegt also für kleine Signifikanzniveaus α, also große Vertrauensniveaus $1-\alpha$, im Vertrauensintervall $C(\omega)$.

Beispiel (Vertrauensintervall für ein Quantil bei nichtparametrische Modellierung): Es seien nun $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit einer unbekannten, atomlosen Verteilung P. Der Ausdruck "*atomlos*" bedeutet $P(\{a\}) = 0$ für alle $a \in \mathbb{R}$, oder das Gleiche mit anderen Worten gesagt: Die Verteilungsfunktion F_P von P sei stetig. Formaler gesagt verwenden wir also das nichtparametrische Modell $(\Omega, \mathcal{A}, \mathcal{P})$ mit

$$\Omega = \mathbb{R}^n, \tag{913}$$

$$\mathcal{A} = \mathcal{B}(\mathbb{R}^n), \tag{914}$$

$$\mathcal{P} = \left\{ P^n \,\middle|\, \begin{matrix} P \text{ ist ein Wahrscheinlichkeitsmaß auf } \mathcal{B}(\mathbb{R}) \\ \text{mit stetiger Verteilungsfunktion } F_P \end{matrix} \right\}. \tag{915}$$

Wir realisieren $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ als kanonische Projektionen. Für $P^n \in \mathcal{P}$ sei $Q_P :]0,1[\to \mathbb{R}$ die Quantilsfunktion zu P (Quasiinverse von F_P):

$$Q_P(q) = \sup\{s \in \mathbb{R} |\ F_P(s) \leq q\}. \tag{916}$$

Insbesondere gilt

$$\forall s \in \mathbb{R}\ \forall q \in]0,1[:\ F_P(s) \leq q \quad \Leftrightarrow \quad s \leq Q_P(q), \tag{917}$$

[65] Erinnerung: Φ bezeichnet die Verteilungsfunktion der Standardnormalverteilung. Also ist $\Phi^{-1}(\alpha)$ das α-Quantil der Standardnormalverteilung.

weil P keine Atome hat, sowie

$$\forall P^n \in \mathcal{P}\ \forall q \in]0,1[:\ P(]-\infty, Q_P(q)]) = F_P(Q_P(q) = q. \tag{918}$$

Gegeben $0 < q < 1$, finden wir jetzt Vertrauensintervalle für den unbekannten Parameter $Q_P(q)$:

Es sei

$$X_{[1]} \leq \ldots \leq X_{[n]} \tag{919}$$

die Ordnungsstatistik zu $X_1, \ldots, X_n$, also ist $X_{[1]}(\omega) \leq \ldots \leq X_{[n]}(\omega)$ die monoton steigende Anordnung von $X_1(\omega), \ldots, X_n(\omega)$ für $\omega \in \Omega$.

Satz 3.27 (Vertrauensintervall für Quantile – untere Schranke) *Es seien* $1 \leq k \leq n$ *und* $0 < q < 1$. *Dann ist*

$$C = [X_{[k]}, \infty[\tag{920}$$

ein Vertrauensintervall für das q-Quantil $Q_P(q)$, *(*$P^n \in \mathcal{P}$*), zum Vertrauensniveau*

$$1 - \alpha = \int_0^q \beta_{k,n-k+1}(x)\, dx = \mathrm{Beta}(k, n-k+1)([0, q[), \tag{921}$$

also der Verteilungsfunktion der Betaverteilung zu den Parametern k, $n-k+1$ *an der Stelle* q.[66]

Bemerkung: Nur Vertrauensniveaus $1-\alpha$ nahe bei 1 sind von Interesse für die Statistik. Hierfür muss q deutlich oberhalb des "Hauptteils der Masse" der Betaverteilung liegen. Das Maximum der Beta-Dichte $\beta_{k,n-k+1}(x)$ liegt bei $x = \frac{k-1}{n-1}$. Es muss also $\frac{k-1}{n-1}$ deutlich unter q liegen, um im für die Statistik relevanten Bereich zu sein.
Beweis des Satzes: Wir zeigen:

$$\forall P^n \in \mathcal{P}:\ P^n[Q_P(q) \in C] = 1 - \alpha. \tag{923}$$

Hierzu sei $P^n \in \mathcal{P}$ gegeben. Dann gilt

$$P^n[Q_P(q) \in C] = P^n[X_{[k]} \leq Q_P(q)] = P^n[F_P(X_{[k]}) \leq q]. \tag{924}$$

Nun ist $U_{[k]} = F_P(X_{[k]})$ die k-te Ordnungsstatistik der mit der Verteilungsfunktion transformierten Zufallsvariablen

$$(U_i := F_P(X_i))_{i \in [n]}. \tag{925}$$

[66] Erinnerung an Definition 2.71: Die Betaverteilung $\mathrm{Beta}(k, n-k+1)$ ist die Vertelung auf $\mathcal{B}(\mathbb{R})$ mit der Dichte

$$x \mapsto \frac{1}{\mathrm{B}(k, n-k+1)} x^{k-1}(1-x)^{n-k+1} 1_{]0,1[}(x) = \beta_{k,n-k+1}(x). \tag{922}$$

Unter dem Wahrscheinlichkeitsmaß P^n sind die $U_1, \ldots, U_n$ i.i.d. uniform auf dem Einheitsintervall $[0,1]$ verteilt, denn

$$\forall a \in]0,1[:\ P^n[U_i \leq a] = P^n[F_P(X_i) \leq a] = P^n[X_i \leq Q_P(a)] = F_P(Q_P(a)) = a. \quad (926)$$

In Satz 2.72 (Verteilung von Komponenten der Ordnungsstatistik) haben wir gezeigt, dass hieraus folgt:

$$\mathcal{L}_{P^n}(U_{[k]}) = \text{Beta}(k, n-k+1). \quad (927)$$

Es folgt:

$$P^n[Q_P(q) \in C] = P^n[U_{[k]} \leq q] = \text{Beta}(k, n-k+1)([0,q]). \quad (928)$$

□

Hier ist eine Version des Satzes für Vertrauensintervalle "in umgekehrter Richtung":

Satz 3.28 (Vertrauensintervall für Quantile – obere Schranke) *Für $1 \leq k \leq n$, $0 < q < 1$ ist $C =]-\infty, X_{[k]}]$ ein Vertrauensintervall für $Q_P(q)$, $P^n \in \mathcal{P}$, zum Vertrauensniveau*

$$1 - \alpha - \int_q^1 \beta_{k,n-k+1}(x)\, dx \quad (929)$$

Dieser Satz folgt aus dem vorhergehenden Satz 3.27, indem man "links" und "rechts" vertauscht, formaler gesagt die X_i durch $-X_i$ ersetzt. Wir verzichten auf die Ausführung im Detail.

Man kann die beiden Sätze 3.27 und 3.27 auch kombinieren, um "zweiseitige" Vertrauensintervalle der Gestalt $C(\omega) = [a(\omega), b(\omega)]$ zu bekommen. Wir untersuchen das erst abstrakt:

Satz 3.29 (Kombination von Vertrauensbereichen – Abstraktion zu zweiseitigen Vertrauensbereichen) *Es seien $(\Omega, \mathcal{A}, \mathcal{P})$ ein statistisches Modell und $C_1, C_2 : \Omega \to \mathcal{P}(\Theta)$ zwei Vertrauensbereiche für einen Parameter $\theta : \mathcal{P} \to \Theta$ zu den Vertrauensniveaus $1-\alpha_1$ bzw. $1-\alpha_2$, wobei $\alpha_1 + \alpha_2 < 1$. Dann ist der Durchschnitt $C_1 \cap C_2$ ein Vertrauensbereich zum Vertrauensniveau $1-(\alpha_1+\alpha_2)$.*

Beweis: Nach Voraussetzung gilt

$$\forall P \in \mathcal{P} :\ P[\theta(P) \in C_i] \geq 1 - \alpha_i \quad (930)$$

für $i = 1, 2$. Es folgt für jedes $P \in \mathcal{P}$:

$$\begin{aligned} &P[\theta(P) \in C_1 \cap C_2] = 1 - P[\theta(P) \notin C_1 \text{ oder } \theta(P) \notin C_2] \\ &\geq 1 - \underbrace{P[\theta(P) \notin C_1]}_{\leq \alpha_1} - \underbrace{P[\theta(P) \notin C_2]}_{\leq \alpha_2} \geq 1 - \alpha_1 - \alpha_2. \end{aligned} \quad (931)$$

□

Korollar 3.30 (Zweiseitige Vertrauensintervalle für Quantile) *In der Situation des Modells* (913)–(913) *ist für* $1 \leq k \leq l \leq n$ *und* $0 < q < 1$ *das zufällige Intervall*

$$C = [X_{[k]}, X_{[l]}] \tag{932}$$

ein Vertrauensintervall zum Vertrauensniveau

$$1 - \alpha = 1 - \int_q^1 \beta_{k,n-k+1}(x)\,dx - \int_0^q \beta_{l,n-l+1}(x)\,dx, \tag{933}$$

falls $\alpha < 1$.

Das ist natürlich nur im Fall

$$\frac{k-1}{n-1} < q < \frac{l-1}{n-1} \tag{934}$$

interessant.

Beispiel: Einseitige Konfidenzintervalle im Binomialmodell. Wir betrachten das 1-parametrische Modell $(\Omega, \mathcal{A}, (P_\theta)_{0\leq\theta\leq1})$ mit

$$\Omega = \{0, \ldots, n\}, \tag{935}$$
$$\mathcal{A} = \mathcal{P}(\Omega), \tag{936}$$
$$P_\theta = \text{binomial}(n, \theta), \quad (0 \leq \theta \leq 1). \tag{937}$$

Für Nullhypothesen $H_0 = \{P_{\theta_0}\}$ und Alternativen $H_1 = \{P_{\theta_1}\}$ mit $\theta_1 > \theta_0$ haben die Verwerfungsbereiche der Neyman-Pearson-Tests alle die Gestalt

$$V_k = \{k, \ldots, n\} \text{ mit } k \in \{0, \ldots, n\}. \tag{938}$$

Die Niveaus $P_{\theta_0}(V_k)$ dieser Tests hängen monoton steigen von $\theta_0 \in [0, 1]$ ab. Um das zu sehen, betrachten wir "Hilfszufallsvariablen": Es seien $U_1, \ldots, U_n$ uniform auf $[0, 1]$ verteilte Zufallsvariablen auf einem "Hilfs-Wahrscheinlichkeitsraum" $(\Omega', \mathcal{A}', P')$. Dann sind für jedes $\theta \in [0, 1]$ die Zufallsvariablen $1_{\{U_i \leq \theta\}}$, $i \in [n]$, i.i.d. $\theta\delta_1 + (1-\theta)\delta_0$-verteilt, also ihre Summe

$$S_\theta := \sum_{i=1}^n 1_{\{U_i \leq \theta\}} \tag{939}$$

binomial(n, θ)-verteilt:

$$\mathcal{L}_{P'}(S_\theta) = P_\theta. \tag{940}$$

Nun gilt für $0 \leq \theta \leq \theta' \leq 1$:

$$\begin{aligned} & \forall i \in [n] : \ \{U_i \leq \theta\} \subseteq \{U_i \leq \theta'\} \\ \Rightarrow \quad & \forall i \in [n] : \ 1_{\{U_i \leq \theta\}} \leq 1_{\{U_i \leq \theta'\}} \\ \Rightarrow \quad & S_\theta \leq S_{\theta'} \\ \Rightarrow \quad & P_\theta(V_k) = P'[S_\theta \geq k] \leq P'[S_{\theta'} \geq k] = P_{\theta'}(V_k). \end{aligned} \tag{941}$$

Mit anderen Worten gesagt: Das Signifikanzniveau $P_\theta(V_k)$ des Tests mit Verwerfungsbereichs V_k zur Nullhypothese $H_0(\theta) = \{P_\theta\}$ steigt monoton in θ.

Weiter gilt für die k-te Ordnungsstatistik $U_{[k]}$ der $U_1, \ldots, U_n$ die folgende Äquivalenz:

$$U_{[k]} \leq \theta \quad \Leftrightarrow \quad S_\theta \geq k. \tag{942}$$

Wir schließen mit Satz 2.72 (Verteilung von Komponenten der Ordnungsstatistik):

$$P_\theta(V_k) = P'[U_{[k]} \leq \theta] = \text{Beta}(k, n-k+1)[0, \theta]. \tag{943}$$

Wir setzen für $0 < \alpha < 1$ und $k \in \{0, \ldots, n\}$:

$$q(\alpha, k) = \begin{cases} \sup\{\theta : \ P_\theta(V_k) \leq \alpha\} = \alpha\text{-Quantil von Beta}(k, n-k+1) & \text{für } k > 0, \\ 0 & \text{für } k = 0 \end{cases} \tag{944}$$

und

$$C_\alpha(\omega) := [q(\alpha, \omega), 1] \text{ für } \omega \in \Omega. \tag{945}$$

Wir zeigen nun, dass das zufällige Intervall C_α ein Konfidenzintervall zum Vertrauensniveau $1 - \alpha$ für den Parameter θ ist. Zu zeigen ist also:

$$\forall \theta \in [0, 1] : \ P_\theta[\theta \in C_\alpha] \geq 1 - \alpha. \tag{946}$$

Gegeben $\theta \in [0, 1]$, schließen wir mit der Abkürzung

$$k(\alpha, \theta) := \max\{k \in \{0, \ldots, n\} |\ P_\theta(V_k) > \alpha\} \tag{947}$$

Folgendes:

- Im Fall $\theta = 0$:

$$P_\theta[\theta \in C_\alpha] = P_\theta[\{\omega \in \Omega |\ \theta \geq q(\alpha, \omega)] = P_0(\{0\}) = 1 \geq 1 - \alpha \tag{948}$$

- Im Fall $\theta > 0$ gilt wegen der Monotonie (941) von $P_\theta(V_\omega)$ in θ und der Definition (944) von $q(\alpha, \omega)$ die Implikationskette

$$P_\theta(V_\omega) > \alpha \quad \Rightarrow \quad \theta \geq q(\alpha, \omega) \quad \Rightarrow \quad \theta \in C_\alpha(\omega) \tag{949}$$

und daher (mit der Konvention $V_{n+1} := \emptyset$):

$$\begin{aligned} P_\theta[\theta \in C_\alpha] &\geq P_\theta(\{\omega \in \Omega |\ P_\theta(V_\omega) > \alpha\}) \\ &= P_\theta(\underbrace{\{0, \ldots, k(\alpha,\theta)\}}_{=V^c_{k(\alpha,\theta)+1}}) = 1 - P_\theta(V_{k(\alpha,\theta)+1}) \\ &\geq 1 - \alpha, \end{aligned} \tag{950}$$

weil $P_\theta(V_{k(\alpha,\theta)+1}) \leq \alpha$ nach der Definition (947) von $k(\alpha, \theta)$.

Damit ist die Behauptung (946) in jedem Fall gezeigt.

3.3.4 Einige Standardtests

Der Einstichproben-t-Test Es seien $X_1, \ldots, X_n$ i.i.d. normalverteilte Daten. Anders als früher nehmen wir an, dass sowohl der Erwartungswert als auch die Varianz der X_i unbekannt seien. Wir verwenden also das 2-parametrische Modell

$$\Omega = \mathbb{R}^n, \tag{951}$$

$$\mathcal{A} = \mathcal{B}(\mathbb{R}^n), \tag{952}$$

$$\mathcal{P} = \{P_{\mu,\sigma^2} |\ \mu \in \mathbb{R},\ \sigma^2 > 0\},\ \text{wobei } P_{\mu,\sigma^2} = \mathrm{N}(\mu, \sigma^2)^n, \tag{953}$$

und beschreiben die Daten durch die kanonischen Projektionen $X_i : \mathbb{R}^n \to \mathbb{R}$, $i \in [n]$. Wir entwickeln nun Tests (oder dual ausgedrückt: Konfidenzintervalle) für den unbekannten Erwartungswert μ.

Bei bekannter Varianz σ^2 hatten wir die folgende Teststatistik zur Hypothese

$$H_0(\mu_0, \sigma^2) = \{P_{\mu_0,\sigma^2}\} \tag{954}$$

verwendet, siehe Formel (892):

$$Z = \sqrt{n}\frac{\overline{X} - \mu_0}{\sigma}, \qquad \mathcal{L}_{P_{\mu_0,\sigma^2}}(Z) = \mathrm{N}(0, 1). \tag{955}$$

Bei nun unbekannter Varianz liegt es nahe, statt einer *im Modell angenommenen* Varianz σ^2 die *Schätzung*

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2, \tag{956}$$

also die *empirische* Varianz zu verwenden:

Definition 3.31 (t-Statistik) *Die Statistik*[67]

$$T := \sqrt{n}\frac{\overline{X} - \mu_0}{s_X} \tag{957}$$

heißt die t-Statistik.

[67] $s_X := \sqrt{s_X^2}$

Wir untersuchen nun die Verteilung der t-Statistik unter der *zusammengesetzten* Hypothese

$$H_0(\mu_0) = \{P_{\mu_0,\sigma^2} |\ \sigma^2 > 0\} \tag{958}$$

bei gegebenen μ_0. Insbesondere zeigen wir, dass diese Verteilung für alle $P_{\mu_0,\sigma^2} \in H_0(\mu_0)$ die Gleiche ist, genauer gesagt weder von μ_0 noch von σ^2 abhängt.

Lemma 3.32 (Verteilung von Termen in der t-Statistik) *Es sei $(X_1, \ldots, X_n)$ P_{μ_0,σ^2}-verteilt. Dann hat der zweidimensionale Zufallsvektor*

$$\left(\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0), \frac{n-1}{\sigma^2} s_X^2\right) \tag{959}$$

die Verteilung $\mathrm{N}(0,1) \times \chi^2_{n-1}$, wobei χ^2_{n-1} die χ^2-Verteilung mit $n-1$ Freiheitsgraden bezeichnet.[68]

Beweis: Es sei

$$Z_j = \frac{X_j - \mu_0}{\sigma}, \quad (j \in [n]). \tag{960}$$

Dann sind $Z_1, \ldots, Z_n$ i.i.d. standardnormalverteilt, also

$$Z := (Z_1, \ldots, Z_n) \tag{961}$$

$N(0, 1_n)$-verteilt. Wir setzen

$$v := \frac{1}{\sqrt{n}} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n. \tag{962}$$

v ist ein Einheitsvektor: $\|v\|_1 = 1$. Dann gilt:[69]

$$\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} = \sqrt{n}\,\overline{Z} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_j = v^t Z \tag{963}$$

und

$$\frac{(n-1)s_X^2}{\sigma^2} = (n-1)s_Z^2 = \sum_{j=1}^{n} (Z_j - \overline{Z})^2 = \|Z - vv^t Z\|_2^2 = \|(1_m - vv^t)Z\|_2^2. \tag{964}$$

[68] *Erinnerung:* Die χ^2-Verteilung χ^2_{n-1} mit $n-1$ Freiheitsgraden ist die Verteilung von $\|X\|_2^2$, wenn X $(n-1)$-dimensional standardnormalverteilt ist, siehe Definition 2.54.

[69] $\overline{Z} = \frac{1}{n}(Z_1 + \ldots + Z_n)$ bezeichnet das Mitel der Z_j.

Wir interpretieren das geometrisch: $v^t Z$ ist die Komponente von Z in Richtung von v, und $1_n - vv^t$ ist die orthogonale Projektion auf den Orthogonalraum von v. Nun ist die Dichte

$$x \mapsto (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\|x\|_2^2\right) \tag{965}$$

der n-dimensioanlen Standardnormalverteilung $\mathcal{L}(Z) = \mathrm{N}(0, 1_n)$ rotationsinvariant um den Nullpunkt, also ist auch $\mathcal{L}(Z)$ rotationsinvariant. Folglich ist die Verteilung von

$$(v^t Z, \|(1_n - vv^t)Z\|_2^2) \tag{966}$$

die Gleiche für alle Einheitsvektoren$v \in \mathbb{R}^n$. Insbesondere können wir v durch $e = (0, \ldots, 0, 1)^t \in \mathbb{R}^n$ ersetzen:

$$\begin{aligned} &\mathcal{L}(v^t Z, \|(1_n - vv^t)Z\|_2^2) = \mathcal{L}(e^t Z, \|(1_n - ee^t)Z\|_2^2) \\ =&\mathcal{L}(Z_n, \|(Z_1, \ldots, Z_{n-1})\|_2^2) = \mathrm{N}(0,1) \times \chi^2_{n-1}. \end{aligned} \tag{967}$$

□

Korollar 3.33 (Verteilung der t-Statistik unter der Nullhypothese) *Die Verteilung der t-Statistik*

$$T = \sqrt{n}\frac{\overline{X} - \mu_0}{s_X} \tag{968}$$

unter P_{μ_0,σ^2} ist für alle $\sigma^2 > 0$ die Gleiche, nämlich diejenige von

$$T_{n-1} = \sqrt{n-1}\frac{X}{\sqrt{Y_{n-1}}}, \tag{969}$$

wenn

$$\mathcal{L}(X, Y_{n-1}) = \mathrm{N}(0,1) \times \chi^2_{n-1}. \tag{970}$$

Beweis: Wir schreiben

$$T = \sqrt{n-1}\frac{\sqrt{n}(\overline{X} - \mu_0)/\sigma}{\sqrt{(n-1)s_X^2/\sigma^2}} \tag{971}$$

Hier sind Zähler und Nenner unabhängig voneinander; der Zähler ist standardnormalverteilt, und das Argument der Wurzel im Nenner ist χ^2_{n-1}-verteilt. Es folgt die Behauptung $\mathcal{L}(T) = \mathcal{L}(T_{n-1})$.

Definition 3.34 (Student-t-Verteilung) *Die Verteilung von*

$$T_n = \sqrt{n}\frac{X}{\sqrt{Y_n}}, \tag{972}$$

wenn X und Y_n uabhängig sind mit $\mathcal{L}(X) = \mathrm{N}(0,1)$ und $\mathcal{L}(Y_n) = \chi^2_n$, wird Student-t-Verteilung *oder kurz* t-Verteilung *mit n Freiheitsgraden genannt und mit t_n bezeichnet.*[70]

[70] "Student" war das Pseudonym, unter dem *William Sealy Gosset* seine Arbeiten zur Statistik veröffentlichte. Sein Arbeitgeber, eine Bierbrauerei, hatte ihm nämlich die Publikation von Arbeiten verboten, um den Verrat von Betriebsgeheimnissen zu verhindern.

Lemma 3.35 (Asymptotik der t-Verteilung für große Freiheitsgradzahlen) *Für $t \to \infty$ konvergiert t_n schwach*[71] *gegen die Standardnormalverteilung.*

Beweis: Wir realisieren X, Y_n als $Z = Z_0$ und $Y_n = \|(Z_1, \ldots, Z_n)\|_2^2$ mit einer i.i.d. Folge $(Z_i)_{i\in\mathbb{N}}$ standardnormalverteilter Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Nach dem starken Gesetz der großen Zahlen, Satz 2.109, gilt

$$\frac{Y_n}{n} = \frac{1}{n}\sum_{j=1}^{n} Z_j^1 \overset{n\to\infty}{\longrightarrow} 1 \quad P\text{-fast sicher} \tag{973}$$

Es folgt

$$T_n = \sqrt{n}\frac{X}{Y_n} = \frac{X}{\sqrt{\frac{Y_n}{n}}} \overset{n\to\infty}{\longrightarrow} X \quad P\text{-fast sicher,} \tag{974}$$

also für jede stetige, beschränkte Testfunktion $f : \mathbb{R} \to \mathbb{R}$:

$$E[f(T_n)] \overset{n\to\infty}{\longrightarrow} E[f(X)] \tag{975}$$

nach dem Satz von der dominierten Konvergenz. Das Gleiche anders ausgedrückt:

$$T_n \underset{n\to\infty}{\overset{d}{\longrightarrow}} X. \tag{976}$$

□

Lemma 3.36 (Dichte der Student-t-Verteilung) *Die Student-t-Verteilung t_n besitzt die Dichte*

$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(\frac{t^2}{n}+1\right)^{-\frac{n+1}{2}}. \tag{977}$$

Bemerkung: Diese Dichte $f_n(t)$ fällt für $|t| \to \infty$ nur mit einem Potenzgesetz, genauer gesagt nur wie $\text{const}\cdot|t|^{-(n+1)}$ ab, also sehr viel langsamer als die Dichte $(2\pi)^{-1/2}e^{-t^2/2}$, die superexponentiell schnell abfällt. Daher sind die für das Testproblem wichtigen q-Quantile der Student-t-Verteilung t_n für q nahe bei 0 oder nahe bei 1 viel Betragsgrößer als jene der Standardnormalverteilung. Dieser Effekt ist für kleine Freiheitsgradzahlen, also kleine Stichprobengrößen, besonders spürbar. Hier ist es daher besonders wichtig, die t-Verteilung statt der Standardnormalverteilung zu verwenden, während bei großen Stichprobenanzahlen der Unterschied nicht mehr so ausgeprägt ist.

Beweis des Lemmas: Es sei

$$\mathcal{L}(X, Y_n) = \mathrm{N}(0,1) \times \chi_n^2. \tag{978}$$

[71] *Erinnerung:* Die schwache Konvergenz von Zufallsvariablen oder auch von Verteilungen wurde in Definition 2.115 eingeführt.

Dann hat der Zufallsvektor (X, Y_n) die gemeinsame Dichte[72]

$$f(x, y_n) = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}_{\text{Dichte von N}(0,1)} \cdot \underbrace{\frac{2^{-n/2}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}}_{\substack{\text{Dichte von } \chi^2_n \\ =\text{Gamma}(a=1/2, s=n/2)}} . \tag{979}$$

Wir setzen

$$Q := \frac{X}{\sqrt{Y_n}}. \tag{980}$$

Wir berechnen die Dichte $f_{(Q,Y_n)}$ des Zufallsvektors (Q, Y_n):
Der Diffeomorphismus

$$G : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R} \times \mathbb{R}^+, \quad G(x, y) = \left(\frac{x}{\sqrt{y}}, y \right) \tag{981}$$

hat die Inverse

$$G^{-1}(q, y) = (q\sqrt{y}, y) \tag{982}$$

mit der Jacobideterminante

$$\det DG^{-1}(q, y) = \begin{vmatrix} \sqrt{y} & \frac{q}{2\sqrt{y}} \\ 0 & 1 \end{vmatrix} = \sqrt{y}. \tag{983}$$

Nach der Transformationsregel, Satz 2.30, hat also der Zufallsvektor (Q, Y_n) die Dichte

$$\begin{aligned} f_{(Q,Y_n)} &= f_{(X,Y_n)}(G^{-1}(q, y)) |\det DG^{-1}(q, y)| \\ &= \frac{1}{2\pi} \frac{2^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} e^{-\frac{q^2 y}{2}} y^{\frac{n-1}{2}} e^{-\frac{y}{2}}. \end{aligned} \tag{984}$$

Integration über y liefert die Dichte von Q:

$$f_Q(q) = \frac{1}{\sqrt{2\pi}} \frac{2^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \underbrace{\int_0^\infty e^{-(q^2+1)\frac{y}{2}} y^{\frac{n-1}{2}} \, dy}_{=\Gamma\left(\frac{n+1}{2}\right)\left(\frac{q^2+1}{2}\right)^{-\frac{n+1}{2}}} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} (q^2+1)^{-\frac{n+1}{2}}. \tag{985}$$

Damit hat

$$T_n = \sqrt{n} Q \tag{986}$$

die Dichte

$$f_{T_n}(t) = \frac{1}{\sqrt{n}} f_Q\left(\frac{t}{\sqrt{n}}\right), \tag{987}$$

wie behauptet.

[72] *Erinnerung:* Für die Dichte der Gammaverteilung Gamma(a, s) siehe Formel (94). Zur Beziehung χ^2_n = Gamma$(a = 1/2, s = n/2)$ zwischen der χ^2-Verteilung und der Gammaverteilung siehe Formel (302).

□

Fassen wir zusammen:

Ein-Stichproben t-Test für die Erwartung:

Modell:

$$\Omega = \mathbb{R}^n, \tag{988}$$
$$\mathcal{A} = \mathcal{B}(\mathbb{R}^n), \tag{989}$$
$$\mathcal{P} = \{N(\mu, \sigma^2)^n |\ \mu \in \mathbb{R},\ \sigma^2 > 0\}, \tag{990}$$
$$X_1, \ldots, X_n : \Omega \to \mathbb{R} \text{ Koordinatenabbildungen.} \tag{991}$$

Nullhypothese:

$$H_0 = \{\mathrm{N}(\mu_0, \sigma^2)^n |\ \sigma^2 > 0\}, \tag{992}$$

also H_0: "$\mu = \mu_0$" mit gegebenem $\mu_0 \in \mathbb{R}$.

Alternativhypothese:

$$H_1 = \{\mathrm{N}(\mu, \sigma^2)^n |\ \mu > \mu_0,\ \sigma^2 > 0\}, \tag{993}$$

lso H_0: "$\mu > \mu_0$".

Teststatistik:

$$T = \sqrt{n}\frac{\overline{X} - \mu_0}{s_X}. \tag{994}$$

Für alle $P_0 \in H_0$ hat die Teststatistik T die gleiche Verteilung

$$\mathcal{L}_{P_0}(T) = t_{n-1}, \tag{995}$$

also die Student-t-Verteilung mit $n-1$ Freiheitsgraden.

Verwerfungsbereich zum Signifikanzniveau α:

$$V_\alpha = \{T \geq t_{n-1,1-\alpha}\}, \tag{996}$$

wobei $t_{n-1,1-\alpha}$ das $1-\alpha$-Quantil von t_{n-1} bezeichnet. Dann gilt:

$$\forall P_0 \in H_0 :\ P_0(V_\alpha) = \alpha. \tag{997}$$

Bemerkungen:

1. Der t-Test ist *unverfälscht* im folgenden Sinn:

$$\forall P_1 \in H_1\ \forall P_0 \in H_0 :\ P_0(V_\alpha) \leq \alpha \leq P_1(V_\alpha). \tag{998}$$

Es ist sogar ein *bester* unverfälschter Test im folgenden Sinn: Für jeden anderen Test im gleichen Rahmenmodell mit einem Verwerfungsbereich $\tilde{V}_\alpha$ mit

$$\forall P_1 \in H_1 \, \forall P_0 \in H_0 : \; P_0(\tilde{V}_\alpha) \leq \alpha \leq P_1(\tilde{V}_\alpha) \tag{999}$$

gilt

$$\forall P_1 \in H_1 : \; P_1(V_\alpha) \geq P_1(\tilde{V}_\alpha). \tag{1000}$$

Der t-Test hat also unter allen unverfälschten Tests bei gegebenem Niveau gleichmäßig die größte Macht. Ein Beweis dieser Aussage wird in Satz (10.19) im Buch "Stochastik" von Hans-Otto Georgii (De Gruyter Lehrbuch 2009) gegeben.

2. Analog erhält man einen ein-Stichproben-t-Test für die Alternative H_1': "$\mu < \mu_0$" statt H_1: "$\mu > \mu_0$" mit dem Verwerfungsbereich

$$V_\alpha' = \{T \leq t_{n-1,\alpha}\} \quad \text{statt} \quad V_\alpha = \{T \geq t_{n-1,1-\alpha}\}. \tag{1001}$$

Hier gelten analoge Optimalitätsaussagen.

Anwendung: Gekoppelte normalverteilte Stichproben In einer klinischen Studie eines Cholesterinsenkers werden n Probanden untersucht. Man erhält $2n$ Datenpunkte

$$X_1, \dots X_n \quad \text{(Cholesterinspiegel vor Gabe des Medikaments)} \tag{1002}$$

$$Y_1, \dots Y_n \quad \text{(Cholesterinspiegel bei den gleichen Probanden nach Gabe des Medikaments)} \tag{1003}$$

Hier wäre es unvernünftig, anzunehmen, dass die $X_1, \dots, X_n$ unabhängig von den $Y_1, \dots, Y_n$ sind, weil X_i und Y_i auf Daten des gleichen Probanden beruht. Plausibel ist dagegen die folgende Rahmenannahme:

$(X_i, Y_i)_{i \in [n]}$ sind i.i.d. mit einer unbekannten Verteilung P über $\mathbb{R}^2$.

Wir verwenden nun eine parametrische Modellierung mit einer zusätzlichen Normalverteilungsannahme:

$$\Omega = (\mathbb{R}^2)^n, \tag{1004}$$

$$\mathcal{A} = \mathcal{B}(\Omega), \tag{1005}$$

$$\mathcal{P} = \{\mathrm{N}((\mu_X, \mu_Y); \Sigma)^n | \; (\mu_X, \mu_Y) \in \mathbb{R}^2, \; \Sigma \in \mathbb{R}^{2\times 2} \text{ positiv definit}\}, \tag{1006}$$

wobei $(X_i, Y_i) : \Omega \to \mathbb{R}^2$ als i-te kanonische Projektion realisiert wird.

Nullhypothese: (X_i, Y_i) sind *austauschbar*, d.h. $\mathcal{L}_P(X_i, Y_i) = \mathcal{L}_P(Y_i, X_i)$. Anders gesagt:

$$H_0 = \{\mathrm{N}((\mu, \mu); \Sigma)^n | \; \mu \in \mathbb{R}, \; \Sigma \in \mathbb{R}^{2\times 2} \text{ positiv definit}, \; \Sigma_{XX} = \Sigma_{YY}\} \tag{1007}$$

wobei

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \tag{1008}$$

mit $\Sigma_{XY} = \Sigma_{YX}$.

Alternativhypothese:

$$H_1 = \{\mathrm{N}((\mu_X, \mu_Y); \Sigma)^n |\ \mu_X > \mu_Y,\ \Sigma \in \mathbb{R}^{2\times 2} \text{ positiv definit}, \Sigma_{XX} = \Sigma_{YY}\} \tag{1009}$$

Testverfahren: Man wendet einen Einstichproben t-Test auf die Differenzdaten $(X_i - Y_i)_{i\in[n]}$ an.

Nichtparametrische Alternativen zum t-Test für gekoppelte Stichproben. Im unserem Beispiel (Cholesterinsenker) mag die parametrische Modellierung $\mathcal{L}_P(X, Y) = \mathrm{N}(\mu, \Sigma)$ fragwürdig erscheinen. Strenggenommen kann sie gar nicht richtig sein, denn wir wissen $X_i, Y_i \geq 0$, weil negative Cholesterinspiegel keinen Sinn ergeben, was einer Normalverteilungsannahme widerspricht.

"Robuster", weil auf viel schwächeren Annahmen beruhend, ist das folgende nichtparametrische Rahmenmodell:

$$\Omega = (\mathbb{R}^2)^n, \tag{1010}$$

$$\mathcal{A} = \mathcal{B}(\Omega), \tag{1011}$$

$$\mathcal{P} = \left\{ P^n |\ P \text{ ist ein Wahrscheinlichkeitsmaß auf } (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)) \text{ mit einer Dichte } \frac{dP}{d\lambda_2} \right\} \tag{1012}$$

mit $(X_i, Y_i) : \Omega \to \mathbb{R}^2$, $i \in [n]$, realisiert als kanonische Projektionen.

Nullhypothese: H_0: "Die X_i, Y_i seien austauschbar". Das Gleiche einer Formel ausgedrückt:

$$H_0 = \{P^n \in \mathcal{P} |\ \forall i \in [n] :\ \mathcal{L}_{P^n}(X_i, Y_i) = \mathcal{L}_{P^n}(Y_i, X_i)\} \tag{1013}$$

Der Vorzeichentest. Hier ist ein sehr einfacher Test bei dieser nichtparametrischen Modellierung:

Teststatistik:

$$N = \sum_{i=1}^{n} 1_{\{X_i < Y_i\}} \tag{1014}$$

Verteilung der Teststatistik unter der Nullhypothese H_0:

$$\forall P_0^n \in H_0 : \ \mathcal{L}_{P_0^n}(N) = \text{binomial}(n, \frac{1}{2}) \tag{1015}$$

Unter beliebigen $P^n \in \mathcal{P}$ hat man dagegen die folgende Verteilung der Teststatistik:

$$\mathcal{L}_{P^n}(N) = \text{binomial}(n, p) \text{ mit } p = P^n[X_1 < Y_1]. \tag{1016}$$

Damit kann man Tests auf die Hypothese "$p = \frac{1}{2}$" für den unbekannten Parameters p der Binomialverteilung anwenden.

Ein Rangtest. Eine i.a. höhere Macht als der Vorzeichentest – bei gleicher Modellannahme und Nullhypothese – hat der folgende *Rangtest*: Es seien $Z_j = X_j - Y_j$, $j \in [n]$ die Differenzen der Datenpunkten; unter der Nullhypothese sind sie i.i.d. und symmetrisch um den Nullpunkt verteilt:

$$\forall P_0^n \in H_0 : \ \mathcal{L}_{P_0^n}(Z_j) = \mathcal{L}_{P_0^n}(-Z_j) \tag{1017}$$

Der Rangtest berücksichtigt nicht nur das Vorzeichen der Z_j, sondern auch ihre absolute Größe: Wir ordnen die Beträge $|Z_j|$, $j \in [n]$, der Größe nach an: Es sei

$$\sigma : [n] \to [n] \tag{1018}$$

die (fast sicher eindeutig bestimmte) Permutation mit

$$|Z_{\sigma(1)}| \leq |Z_{\sigma(2)}| \leq \ldots \leq |Z_{\sigma(n)}|; \tag{1019}$$

unter der Nullhypothese ist sie auf der Menge S_n aller Permutationen von $[n]$ gleichverteilt und unabhängig von der Familie $(1_{\{Z_i<0\}})_{i\in[n]}$ der Vorzeichendaten. Dann bilden wir die "Rangsummen-Teststatistik"

$$S = \sum_{k=1}^{n} 1_{\{Z_{\sigma(k)}<0\}}. \tag{1020}$$

Wir bilden also die Summe über die "Ränge" der j mit $Z_j < 0$.

Zahlenbeispiel:

i	1	2	3
X_i	10	4	2
Y_i	8	11	5
Z_i	2	-7	-3
$\lvert Z_i \rvert$	2	7	3
Rang	1	$\boxed{3}$	$\boxed{2}$

wobei die Ränge in den zwei Boxen, die zu negativen Z_i gehören, bei der Summe berücksichtigt werden. Die Rangsumme ist also in diesem Beispiel gleich 5. Man verwirft die Nullhypothese – je nach Alternative – für große oder für kleine Werte der Teststatistik S.

Übung 3.37 (Verteilung der Rangsummen-Teststatistik) *Zeigen Sie, dass unter allen $P_0 \in H_0$ die Teststatistik S die gleiche Verteilung hat, nämlich die Verteilung von $\sum_{j=1}^n jB_j$, wenn $B_1, \dots B_n$ i.i.d. $\frac{1}{2}(\delta_0 + \delta_1)$-verteilte Zufallsvariablen sind.*

Der χ^2-Test für einfache Hypothesen. Wir starten mit einem
Anwendungsbeispiel: In Quasiland gibt es vier politische Parteien: A, B, C,D. Bei den Wahlen vor einem Jahr erhielten sie die folgenden Stimmanteile:

Partei	A	B	C	D
Stimmanteil	39,6%	31,5%	21,7%	7,2%

Ein jahr später ergibt eine Stichprobe unter 1000 zufällig gewählten wahlberechtigten Quasiländern das folgende Ergebnis:

Partei	A	B	C	D
Anzahl Stimmen	375	326	209	50

Belegt das einen Stimmungswandel, oder kann die Abweichung "rein zufällig" sein?

Wir formalisieren das Problem mit dem folgenden
Modell: Es sei $\Omega = \{A, B, C, D\}$ oder allgemeiner $\Omega = \{A_1, \dots, A_d\}$ die Menge der Wahlmöglichkeiten. Wir verwenden das Rahmenmodell $(\Omega^n, \mathcal{P}(\Omega^n), \mathcal{P}_n)$ mit

$$\mathcal{P}_n = \{P^n | \ P \text{ ist ein Wahrscheinlichkeitsmaß auf } (\Omega, \mathcal{P}(\Omega))\}. \tag{1021}$$

Im Beispiel ist $d = 4$ und $n = 1000$. Es sei

$$Y_j : \Omega^n \to \Omega, \quad (j \in [n]) \tag{1022}$$

die j-te Projektion; sie beschreibt die Antwort der j-ten befragten Person. Wir parametrisieren P mit seiner Zähldichte $(p_i)_{i\in[d]}$, wobei $\sum_{i=1}^d p_i = 1$ und $p_i \geq 0$ für alle $i \in [d]$:

$$P = \sum_{i=1}^d p_i \delta_{A_i}. \tag{1023}$$

Unter $P^n \in \mathcal{P}_n$ sind also die Y_j, $j \in [n]$, i.i.d. mit der unbekannten Verteilung P.
Die Anzahl der Antworten "A_i" lautet:

$$N_i := \sum_{j=1}^n 1_{\{Y_j = A_i\}}, \quad (i \in [d]). \tag{1024}$$

Definition 3.38 (Multinomialverteilung) *Die gemeinsame Verteilung $\mathcal{L}_{P^n}(N_1, \dots, N_d)$ auf $\{(n_1, \dots, n_d) \in \mathbb{N}_0^d | \ n_1 + \dots + n_d = n\}$ wird* Multinomialverteilung *mit den Parametern $(n; p_1, \dots, p_d)$ genannt.*

Bemerkung: Die Multinomialverteilung hat die folgende Zähldichte:

$$P^{\lfloor} N_i = n_i \text{ für alle } i \in [d]] = \frac{n!}{\prod_{i=1}^d n_i!} \prod_{i=1}^d p_i^{n_i} \tag{1025}$$

für alle $(n_1, \ldots, n_d) \in \mathbb{N}_0^d$ mit $n_1 + \ldots + n_d = n$. Die Randverteilungen der Multinomialverteilung sind Binomialverteilungen:

$$\mathcal{L}_{P^n}(N_i) = \text{binomial}(n, p_i). \tag{1026}$$

Wir verwenden die folgende Vektornotation:

$$p := \begin{pmatrix} p_1 \\ \vdots \\ p_d \end{pmatrix}, \quad X_j := \begin{pmatrix} 1_{\{Y_j = A_1\}} \\ \vdots \\ 1_{\{Y_j = A_d\}} \end{pmatrix}, \quad N = \begin{pmatrix} N_1 \\ \vdots \\ N_d \end{pmatrix} = \sum_{j=1}^n X_j. \tag{1027}$$

Wir erhalten (komponentenweise zu lesen):

$$E_{P^n}[N] = np, \tag{1028}$$

$$\begin{aligned} \text{Cov}_{P^n}(N^t, N) := \text{Cov}_{P^n}(N_i, N_{i'}))_{i,i' \in [d]} &= \sum_{j=1}^n \text{Cov}_{P^n}(X_j^t, X_j) \\ &\quad \text{(weil } X_1, \ldots, X_n \text{ unabhängig)} \\ &= n \, \text{Cov}_{P^n}(X_1^t, X_1) = n(E_{P^n}[X_1^t X_1] - E_{P^n}[X_1^t] E_{P^n}[X_1]) \\ &= n(\text{diag}(p) - p^t p), \end{aligned} \tag{1029}$$

wobei $\text{diag}(p) = \text{diag}(p_1, \ldots, p_d)$ die $d \times d$-Diagonalmatrix mit den Diagonaleinträgen $p_1, \ldots, p_d$ bezeichnet. Im letzten Schritt haben wir

$$E_{P^n}[1_{\{Y_1 = A_i\}} 1_{\{Y_1 = A_{i'}\}}] = p_i \delta_{ii'} \tag{1030}$$

verwendet.[73] Man beachte, dass die Kovarianzmatrix $n(\text{diag}(p) - p^t p)$ singulär ist, denn

$$(1, \ldots, 1)(\text{diag}(p) - p^t p) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 0. \tag{1031}$$

Wir verwenden jetzt ohne Beweis[74] die multidimensionale Version des Zentralen Grenzwertsatzes:[75]

---

[73] $\delta_{ii'} = \begin{cases} 1 & \text{falls } i = i' \\ 0 & \text{sonst} \end{cases}$ bezeichnet das Kronecker-Delta.

[74] Man kann diese multidimensionale Version des zentralen Grenzwertsatzes entweder analog zur eindimensionalen Version beweisen, oder auch auf die eindimensionale Version zurückführen.

[75] Erinnerung an die Übungen: Für jede positiv semidefinite, aber nicht positiv definite Matrix $\Sigma \in \mathbb{R}^{d \times d}$ vom Rang $k < d$ ist die singuläre Normalverteilung $\text{N}(0, \Sigma)$ definiert als die Verteilung von LY, wobei $L \in \mathbb{R}^{d \times k}$ irgendeine Matrix mit $LL^t = \Sigma$ und $\mathcal{L}(Y) = \text{N}(0, 1_k)$ ist. Sie ist wohldefiniert, hängt also nicht von der Wahl von L ab. Weiter gilt hiermit: Ist X ein Zufallsvektor mit Werten in $\mathbb{R}^d$ mit $\mathcal{L}(X) = \text{N}(0, \Sigma)$ und ist $A \in \mathbb{R}^{l \times d}$ mit $l \in \mathbb{N}$, so ist $\mathcal{L}(AX) = \text{N}(0, A\Sigma A^t)$.

Satz 3.39 (Multidimensionaler Zentraler Grenzwertsatz) *Sind $(X_j)_{j\in\mathbb{N}}$ i.i.d. Zufallsvektoren mit Werten in $\mathbb{R}^d$, Erwartungswertvektor $E[X_j] = \mu \in \mathbb{R}^d$ und endlicher Kovarianzmatrix $\mathrm{Cov}(X_j^t, X_j) = \Sigma \in \mathbb{R}^{d\times s}$, so konvergiert*

$$Z_n := \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (X_j - \mu) \tag{1032}$$

für $n \to \infty$ schwach gegen die multidimensionale Normalverteilung $\mathrm{N}(0, \Sigma)$, d.h. für alle stetigen beschränkten Funktionen $f : \mathbb{R}^d \to \mathbb{R}$ gilt

$$E[f(Z_n)] \overset{n\to\infty}{\longrightarrow} E[f(Z)], \tag{1033}$$

wobei Z einen $\mathrm{N}(0, \Sigma)$-verteilten Zufallsvektor bezeichnet.

Wir erhalten: Die Verteilung bezüglich P^n von

$$\frac{1}{\sqrt{n}}(N - np) = \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{n} (X_j - E_{P^n}[X_j]) \right) \tag{1034}$$

konvergiert für $n \to \infty$ gegen die multidimensionale Normalverteilung $\mathrm{N}(0, \mathrm{diag}(p) - p^t p)$. Um eine geometrisch einfachere Struktur zu sehen, skalieren wir die Einträge der Vektoren noch mit $\frac{1}{\sqrt{p_i}}$, wobei wir annehmen, dass alle p_i positiv sind, um nicht durch 0 zu dividieren. Es sei

$$D = \mathrm{diag} \left(\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_d}} \right) \tag{1035}$$

die Diagonalmatrix mit den Diagonaleinträgen $1/\sqrt{p_i}$, und

$$v = \begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_d} \end{pmatrix} = Dp. \tag{1036}$$

Insbesondere ist v ein Einheitsvektor:

$$\|v\|_2^2 = \sum_{i=1}^{d} p_i = 1. \tag{1037}$$

Definition 3.40 (χ^2-Statistik) *Mit der Abkürzung*

$$X_{(n)} := \frac{1}{\sqrt{n}} D \cdot (N - np) = \left(\frac{N_i - np_i}{\sqrt{np_i}} \right)_{i\in[d]} \tag{1038}$$

nennen wir

$$\boxed{\chi^2_{(n)} := \|X_{(n)}\|_2^2 = \sum_{i=1}^{n} \frac{N - np_i}{np_i}} \tag{1039}$$

die χ^2-Statistik zur Nullhypothese $H_0 = \{P^n\}$, wobei $P = \sum_{i=1}^{d} p_i \delta_i$.

Die Asymptotik im Limes $n \to \infty$ der Verteilung $\mathcal{L}_{P^n}(\chi^2_{(n)})$ der χ^2-Statistik unter der Nullhypothese P^n wird durch den folgenden Satz von Pearson beschrieben:

Satz 3.41 (Satz von Pearson) *Für $n \to \infty$ konvergiert die χ^2-Statistik $\chi^2_{(n)}$ bezüglich P^n in Verteilung gegen die χ^2-Verteilung mit $n-1$ Freiheitsgraden.*

Beweisskizze: Aus

$$\frac{1}{\sqrt{n}}(N - np) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \mathrm{diag}(p) - p^t p) \tag{1040}$$

schließen wir

$$X_{(n)} = \frac{1}{\sqrt{n}} D \cdot (N - np) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, D \cdot (\mathrm{diag}(p) - p^t p) D^t) = \mathrm{N}(0, 1_n - v^t v). \tag{1041}$$

Nun ist $1_n - v^t v$ die orthogonale Projektionsmatrix auf den Orthogonalraum $v^\perp = \{x \in \mathbb{R}^d |\ v^t x = 0\}$ von v, also auf einen $d-1$-dimensionalen Unterraum von $\mathbb{R}^d$. Wir schließen

$$\|X_{(n)}\|_2^2 = \chi^2_{(n)} \xrightarrow[n\to\infty]{d} \chi^2_{d-1}. \tag{1042}$$

□

Wir erhalten

χ^2-Test für einfache Hypothesen

Für ein Niveau $\alpha \in]0,1[$ (nahe bei 0) sei c_α das $1-\alpha$-Quantil der χ^2-Verteilung χ^2_{d-1} mit $d-1$ Freiheitsgraden. Dann hat der Test mit dem Verwerfungsbereich

$$V_\alpha := \{\chi^2_{(n)} \geq c_\alpha\} \tag{1043}$$

der Nullhypothese $H_0(p) = \{P^n\}$ bei der Alternative $H_1(p) = \mathcal{P} \setminus H_0(p)$ im Limes $n \to \infty$ asymptotisch das Niveau α.

Bemerkungen/Ausblick:

1. In der Praxis muss man nicht n groß wählen. Für praktisch tolerable Genauigkeiten reichen erstaunlich kleine n: Zum Beispiel gibt es die folgende Daumenregel nach Prof. Hampel (emeritierter Professor an der ETH Zürich): Es soll $np_i \geq 1$ für alle $i \in [d]$ gelten, und $np_i \geq 4$ für mindestens $\frac{4}{5}$ aller $i \in [d]$.

2. Es gibt auch eine Variante des χ^2-Tests für zusammengesetzte Hypothesen, die auf Fisher zurückgeht: Hierzu nimmt man als Nullhypothese an, dass der Parameter p von P in einer k-dimensionalen differenzierbaren Mannigfaltigkeit liegt. In diesem Fall wird der Parameter p durch seinen Maximum-Likelihood-Schätzer $\hat{p}$ ersetzt. Die mit diesem Schätzer $\hat{p}$ statt p berechnete χ^2-Teststatistik $\chi^2_{(n)}$ ist dann im Limes $n \to \infty$ auch wieder χ^2-verteilt, allerdings reduziert sich die Anzahl der Freiheitsgrade um die Dimension k der Nullhypothese:

$$\chi^2_{(n)} \xrightarrow[n\to\infty]{d} \chi^2_{d-1-k}. \tag{1044}$$

Index