Stochastik I
Skript zur Vorlesung
im Sommersemester 2023

Prof. Dr. Markus Reiß
Humboldt-Universität zu Berlin

Version vom 21. Juli 2023

Inhaltsverzeichnis

Ein paar Literaturempfehlungen

Fast alle Bücher sind über den Katalog *Primus* der Universitätsbibliothek mit VPN als Ebook verfügbar. Die ersten drei Bücher werden auf jeden Fall im Skript Verwendung finden.

- **Hans-Otto Georgii**, *Stochastik*, de Gruyter: exzellentes Lehrbuch inkl. Maßtheorie
- **Achim Klenke**, *Wahrscheinlichkeitstheorie*, Springer: Lehrbuch für Stochastik I und II, aus Vorlesungen entstanden
- **Ulrich Krengel**, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Vieweg: Klassiker mit vielen Beispielen und Diskussionen, ohne Maßtheorie
- Herold Dehling, Beate Haupt, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Springer: Lehrbuch mit vielen erklärenden Skizzen und Diagrammen, ohne Maßtheorie
- William Feller, *An introduction to probability theory and its applications I*, Wiley: das alte Testament, eine Fundgrube, immer noch Standardreferenz
- Kai Lai Chung, *A Course in Probability Theory*, Academic Press: Englisch-sprachiges Standardwerk, besonders empfehlenswert für charakteristische Funktionen und Konvergenzresultate
- Richard Dudley, *Real Analysis and Probability*, Cambridge University Press: ausgezeichnetes und recht anspruchsvolles Lehrbuch zu Maßtheorie, Analysis und W-Theorie, insbesondere für Konvergenzarten und charakteristische Funktionen
- Jürgen Elstrodt, *Maß- und Integrationstheorie*, Springer: mit viel Liebe und historischen Anmerkungen verfasstes, ausführliches Maßtheoriebuch
- Heinz Bauer, *Wahrscheinlichkeitstheorie*, de Gruyter: umfassendes deutsches Standardwerk, auf dem Maßtheoriebuch des Autors aufbauend
- Albert N. Shiryaev, *Probability*, Springer: umfassendes Lehrbuch, gut als Nachschlagewerk für Stochastik I und II
- Jean Jacod, Philip Protter, *Probability Essentials*, Springer: alle wichtigen Ergebnisse auf hohem Niveau, kurz und knapp
- John A. Rice, *Mathematical Statistics and Data Analysis*, Thomson: gutes einführendes Lehrbuch in die mathematische Statistik, viele Beispiele
- Jun Shao, *Mathematical Statistics*, Springer: deckt weite Themen der math. Statistik ab, gut für den Überblick und zum Nachschlagen

1 Wahrscheinlichkeitsräume

1.1 Ereignisse, Wahrscheinlichkeiten und Zufallsvariablen

1.1 Beispiele.

(a) Würfeln mit zwei unterscheidbaren Würfeln wird durch die Grundmenge $\Omega = \{1,2,3,4,5,6\}^2$ beschrieben. Interpretation ist, dass Versuchsausgang $(n_1, n_2) \in \Omega$ bedeutet Augenzahl des 1. Würfels $= n_1$, des 2. Würfels $= n_2$. Ereignisse sind z.B. $A_1 =$*es wird ein Pasch gewürfelt*, $A_2 =$*die Augensumme ist 7*, $A_3 =$*es tritt keine 1 auf*. Ereignisse werden als Teilmengen von Ω modelliert: $A_1 = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$, $A_2 = \{(n_1,n_2) \in \Omega \,|\, n_1 + n_2 = 7\}$, $A_3 = \{(n_1,n_2) \in \Omega \,|\, n_1 \neq 1 \text{ und } n_2 \neq 1\}$. Sind alle Versuchsausgänge gleichwahrscheinlich, so ist die Wahrscheinlichkeit eines Ereignisses $A \subseteq \Omega$ gleich der Anzahl aller für A günstiger Versuchsausgänge geteilt durch die Anzahl aller möglichen Versuchsausgänge, d.h. $P(A) = \frac{|A|}{|\Omega|}$, sprich „die Wahrscheinlichkeit von A ist gleich dem Verhältnis der Anzahl Elemente von A zu der von Ω“. Wir erhalten $P(A_1) = \frac{1}{6}$, $P(A_2) = \frac{1}{6}$, $P(A_3) = \frac{25}{36}$.

(b) Beim n-fachen Wurf einer Münze setzen wir $\Omega = \{0,1\}^n$ mit der Interpretation, dass $(\omega_1, \ldots, \omega_n) \in \Omega$ das Ergebnis in den Würfen 1 bis n kodiert via $\omega_i = 1$ für Kopf im i-ten Wurf, $\omega_i = 0$ für Zahl im i-ten Wurf. Wir betrachten die Ereignisse $A_1 = \{(1,\ldots,1)\}$ (n-mal Kopf), $A_2 = \{\omega \in \Omega \,|\, \omega_1 + \cdots + \omega_n = n-1\}$ ($(n-1)$-mal Kopf), $A_3 = \{\omega \in \Omega \,|\, \omega_1 + \cdots + \omega_n \leqslant n-1\}$ (mindestens einmal Zahl). Sind alle Versuchsausgänge gleichwahrscheinlich, so erhalten wir $P(A_1) = 2^{-n}$, $P(A_2) = n2^{-n}$, $P(A_3) = 1 - 2^{-n}$. Beachte, dass $A_3 = A_1^c := \Omega \setminus A_1$ und $P(A_1) + P(A_3) = 1$ gilt. A_1 und A_3 heißen auch komplementäre Ereignisse oder Gegenereignisse.

(c) Wir nehmen nun an, dass die Münze beliebig oft hintereinander geworfen wird. Die Versuchsausgänge modellieren wir dann durch 0-1-Folgen $(\omega_n)_{n\in\mathbb{N}} = (\omega_1, \omega_2, \ldots)$, d.h. $\Omega = \{0,1\}^{\mathbb{N}}$. Ereignisse sind dann $A_1 = \{(1,1,\ldots)\}$ (immer nur Kopf), $A_2 = \{\omega \in \Omega \,|\, \forall M \in \mathbb{N}\, \exists k \in \mathbb{N} : \omega_k = \omega_{k+1} = \cdots = \omega_{k+M-1} = 1\}$ (für jedes $M \in \mathbb{N}$ gibt es einen Kopf-run der Länge M). Wegen $|\Omega| = \infty$ gibt es keine Gleichverteilung auf den Versuchsausgängen. Intuitiv würden wir aus (b) schließen $P(A_1) = \lim_{n\to\infty} 2^{-n} = 0$, während $P(A_2)$ nicht klar ist. Es fehlt eine mathematisch formale Einführung des Wahrscheinlichkeitsmaßes P und seiner Eigenschaften.

1.2 Bemerkung. Im letzten Beispiel (c) ist nicht sofort klar, welche Wahrscheinlichkeiten wir den Ereignissen A_1, A_2 zuordnen sollten. Für A_1 haben wir intuitiv eine Grenzwertüberlegung gemacht durch Ereignisse, die nur von endlich vielen Würfen abhängen, was auch für A_2 möglich ist. Könnten wir so jeder Teilmenge $A \subseteq \{0,1\}^{\mathbb{N}}$ eine Wahrscheinlichkeit zuordnen? Einfache Überlegungen zur Kardinalität der Mengen (die Potenzmenge $\mathscr{P}(\{0,1\}^{\mathbb{N}})$ hat eine größere Kardinalität als die reellen Zahlen) zeigen, dass dies nicht der Fall ist. In diesem

Fall wollen wir also vielleicht gar nicht jeder Teilmenge von Ω eine Wahrscheinlichkeit zuordnen. Unten werden wir im Satz von Vitali dann auch noch sehen, dass ein fairer Münzwurf auf $\mathscr{P}(\{0,1\}^{\mathbb{N}})$ nie definiert werden kann. Dies führt zu der Einsicht, dass wir Wahrscheinlichkeiten nur auf interessierenden Mengen definieren sollten.

1.3 Definition. Mit Ω werde die nichtleere Menge der möglichen Versuchsausgänge oder Ergebnismenge, Grundmenge bezeichnet. Ein Teilmengensystem $\mathscr{F} \subseteq \mathscr{P}(\Omega)$ heißt Menge der interessierenden Ereignisse oder mathematisch σ-Algebra, falls gilt:

(a) $\Omega \in \mathscr{F}$;

(b) $A \in \mathscr{F} \Rightarrow A^c \in \mathscr{F}$;

(c) $A_n \in \mathscr{F}, n \in \mathbb{N} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathscr{F}$.

Die Elemente von $\mathscr{F}$ heißen Ereignisse. Ein Wahrscheinlichkeitsmaß P (auch Wahrscheinlichkeitsverteilung genannt) auf $\mathscr{F}$ ist eine Abbildung $P : \mathscr{F} \to [0,1]$, die den *Kolmogorovschen Axiomen* (1933) genügt:

(a) $P(\Omega) = 1$ (Normierung);

(b) für $A_n \in \mathscr{F}$, $n \in \mathbb{N}$, paarweise disjunkt gilt

$$P\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} P(A_n) \text{ (σ-Additivität).}$$

Ein Wahrscheinlichkeitsraum ist ein Tripel $(\Omega, \mathscr{F}, P)$, bestehend aus einer Ergebnismenge Ω, einer σ-Algebra $\mathscr{F}$ über Ω sowie einem Wahrscheinlichkeitsmaß P auf $\mathscr{F}$.

1.4 Beispiele. Auf jeder nichtleeren Ergebnismenge Ω existieren die triviale σ-Algebra $\{\varnothing, \Omega\}$ sowie die Potenzmenge $\mathscr{P}(\Omega)$ als σ-Algebren. Für $\omega_0 \in \Omega$ ist das Einpunkt- oder Diracmaß $\delta_{\omega_0}(A) = \mathbf{1}(\omega_0 \in A)$, $A \in \mathscr{F}$, ein Wahrscheinlichkeitsmaß auf jeder σ-Algebra $\mathscr{F}$ über Ω. Sind $(P_n)_{n\geqslant 1}$ Wahrscheinlichkeitsmaße auf einer σ-Algebra $\mathscr{F}$, so auch jede Konvexkombination $\sum_{n\geqslant 1} w_n P_n$ mit $w_n \geqslant 0$ und $\sum_{n\geqslant 1} w_n = 1$. Die Einpunktmaße bilden Extremalpunkte der konvexen Menge aller Wahrscheinlichkeitsmaße auf $\mathscr{F}$ (recherchiere ggf. konvex, Extremalpunkt).

1.5 Lemma. *Für jede σ-Algebra $\mathscr{F}$ gilt:*

(a) $\varnothing \in \mathscr{F}$;

(b) $A_1, A_2 \in \mathscr{F} \Rightarrow A_1 \cup A_2 \in \mathscr{F}$;

(c) $A_n \in \mathscr{F}, n \in \mathbb{N} \Rightarrow \bigcap_{n\in\mathbb{N}} A_n, A_1 \cap A_2 \in \mathscr{F}$.

Beweis. Zu (a): aus Axiomen (a) und (b) folgt $\Omega \in \mathscr{F} \Rightarrow \varnothing = \Omega^c \in \mathscr{F}$. Für (b) setze einfach $A_n = \varnothing$ für $n \geqslant 3$ und wende Axiom (c) an. Behauptung (c) folgt durch Komplementbildung (Axiom (b)) aus Axiom (c) bzw. Behauptung (b). □

1.6 Lemma. *Für jedes Wahrscheinlichkeitsmaß $P : \mathscr{F} \to [0,1]$ gilt:*

(a) $P(\varnothing) = 0$*;*

(b) $A, B \in \mathscr{F},\ A \subseteq B \Rightarrow P(A) \leqslant P(B)$ *(Monotonie);*

(c) $\forall\, A, B \in \mathscr{F} :\ P(A \cup B) = P(A) + P(B) - P(A \cap B)$*;*

(d) $\forall\, A_n \in \mathscr{F},\, n \geqslant 1 :\ P(\bigcup_{n\geqslant 1} A_n) \leqslant \sum_{n\geqslant 1} P(A_n)$ *(Subadditivität, Bonferroni-Ungleichung).*

Beweis.

(a) Mit $A_n = \varnothing$ gilt wegen σ-Additivität $P(\varnothing) = \sum_{n\geqslant 1} P(\varnothing)$, also $P(\varnothing) = 0$.

(b) Mit $A_1 = A, A_2 = B \setminus A$ (und $A_n = \varnothing$, $n \geqslant 3$) gilt (Wahrscheinlichkeiten sind nicht-negativ)

$$P(B) = P(A_1 \cup A_2) = P(A_1) + P(A_2) = P(A) + P(B \setminus A) \geqslant P(A).$$

(c) Zerlege disjunkt: $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$, $A = (A \setminus B) \cup (A \cap B)$, $B = (B \setminus A) \cup (A \cap B)$ und verwende die Additivität von P:

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

(d) Setze $B_1 = A_1$, $B_n = A_n \setminus \bigcup_{i<n} A_i$, $n \geqslant 2$. Dann gilt $B_n \subseteq A_n$, $\bigcup_{n\geqslant 1} B_n = \bigcup_{n\geqslant 1} A_n$, aber die $(B_n)_{n\geqslant 1}$ sind nach Konstruktion paarweise disjunkt. Also folgt mittels σ-Additivität und Monotonie

$$P\Big(\bigcup_{n\geqslant 1} A_n\Big) = P\Big(\bigcup_{n\geqslant 1} B_n\Big) = \sum_{n\geqslant 1} P(B_n) \leqslant \sum_{n\geqslant 1} P(A_n).$$

□

1.7 Satz. *Ist P ein Wahrscheinlichkeitsmaß auf $\mathscr{F}$, so gilt für $A_n \in \mathscr{F}$, $n \geqslant 1$, mit $A_n \uparrow A$ (d.h. $A_n \subseteq A_{n+1}$, $\bigcup_{n\geqslant 1} A_n = A$) oder $A_n \downarrow A$ (d.h. $A_n \supseteq A_{n+1}$, $\bigcap_{n\geqslant 1} A_n = A$)*

$$P(A) = \lim_{n\to\infty} P(A_n) \text{ (}\sigma\text{-Stetigkeit).}$$

Andererseits ist jede normierte, additive Mengenfunktion $Q : \mathscr{F} \to [0,1]$ (d.h. $Q(\Omega) = 1$, $Q(A \cup B) = Q(A) + Q(B)$ für alle disjunkten $A, B \in \mathscr{F}$), die σ-stetig ist, auch σ-additiv und damit ein Wahrscheinlichkeitsmaß.

Beweis. Übung! □

1.8 Beispiel. Betrachte den unendlich häufigen Münzwurf aus Beispiel 1.1(c). Wir beschreiben das Wahrscheinlichkeitsmaß P auf allen Ereignissen, die nur von endlich vielen Münzwürfen abhängen. Daher fordern wir, dass für jedes $n \geqslant 1$ die Mengen $B_n \times \{0,1\} \times \{0,1\} \times \cdots = \{(\omega_i)_{i\geqslant 1} \,|\, (\omega_1, \ldots, \omega_n) \in B_n, \omega_i \in$

$\{0,1\}, i > n\}$ für $B_n \subseteq \{0,1\}^n$ in einer σ-Algebra $\mathscr{F}$ enthalten sind. Dann muss auch $A_1 = \bigcap_{n\geqslant 1}\{(\omega_i)_{i\geqslant 1} \,|\, \omega_1 = \ldots = \omega_n = 1, \omega_i \in \{0,1\}, i > n\}$ in der σ-Algebra $\mathscr{F}$ liegen. Ordnet ein Wahrscheinlichkeitsmaß P dem Ereignis $\{(\omega_i)_{i\geqslant 1} \,|\, \omega_1 = \ldots = \omega_n = 1, \omega_i \in \{0,1\}, i > n\}$ die faire Wahrscheinlichkeit 2^{-n} zu, so muss wegen σ-Stetigkeit $P(A_1) = 0$ gelten. Dies ist also eine formale Begründung unserer intuitiven Herleitung. Beachte allerdings, dass wir bislang nicht die Existenz von $\mathscr{F}$ und insbesondere von P bewiesen haben.

Ende der 1. Vorlesung

1.9 Bemerkung. Häufig interessieren uns nicht die Versuchsausgänge selbst, sondern abgeleitete Größen wie zum Beispiel die Augensumme beim Würfeln. Dies wird mit Zufallsvariablen modelliert, die als Funktionen auf Wahrscheinlichkeitsräumen Wahrscheinlichkeitsmaße „transportieren“, ähnlich den Homomorphismen der Algebra.

1.10 Definition. Es sei $(\Omega, \mathscr{F}, P)$ ein Wahrscheinlichkeitsraum und $(S, \mathscr{S})$ ein Messraum. Dann heißt eine Funktion $g : \Omega \to S$ messbar (bzgl. $(\mathscr{F}, \mathscr{S})$), falls

$$\forall\, A \in \mathscr{S} :\ g^{-1}(A) \in \mathscr{F}$$

gilt. Jede solche messbare Funktion heißt $(S, \mathscr{S})$-wertige Zufallsvariable. Für $S = \mathbb{R}^d$ wird kanonisch $\mathscr{S} = \mathfrak{B}_{\mathbb{R}^d}$ (Borel-σ-Algebra, siehe unten) gewählt, und man spricht bloß von einer Zufallsvariablen ($d = 1$) bzw. einem Zufallsvektor ($d \geqslant 2$).
Die Verteilung einer $(S, \mathscr{S})$-wertigen Zufallsvariablen X ist das Wahrscheinlichkeitsmaß (!)

$$P^X(A) := P(X \in A) = P(X^{-1}(A)), \quad A \in \mathscr{S}.$$

Die Verteilung P^X von X ist also das Bildmaß von P unter X. Später werden wir mit der Verteilungsfunktion (Dichte, Zähldichte) von X stets die zu P^X gehörige Größe meinen.

Wir schreiben kurz $\{X \in A\} := \{\omega \in \Omega \,|\, X(\omega) \in A\}$, $\{X = x\} := \{\omega \in \Omega \,|\, X(\omega) = x\}$, $P(X \in A) := P(\{X \in A\})$, $P(X = x) := P(\{X = x\})$ etc.

1.11 Beispiele.

(a) Beim Würfeln mit 2 Würfeln interessiert uns die Augensumme X als abgeleiteter Parameter. Formal betrachten wir $\Omega = \{1,2,3,4,5,6\}^2$, $X : \Omega \to \mathbb{R}$ mit $X((n_1, n_2)) = n_1 + n_2$. Das Ereignis $\{X = 7\}$ ist dann kurz für $\{\omega \in \Omega \,|\, X(\omega) = 7\} = X^{-1}(\{7\}) = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \subseteq \Omega$ und $\{X \text{ ist gerade}\} = \{\omega \in \Omega \,|\, X(\omega) \in \{2,4,6,8,10,12\}\}$.

X transportiert das Wahrscheinlichkeitsmaß P auf Ω nach $\mathbb{R}$: $P^X(A) := P(X^{-1}(A)) = P(\{\omega \in \Omega \,|\, X(\omega) \in A\}) = P(X \in A)$ für $A \subseteq \mathbb{R}$. Beachte, dass hier jede Funktion $X : \Omega \to \mathbb{R}$ messbar und damit Zufallsvariable ist, da Ω mit der Potenzmenge $\mathscr{P}(\Omega)$ als σ-Algebra versehen ist, so dass P^X sogar auf $\mathscr{P}(\mathbb{R})$ wohldefiniert ist.

(b) Auf einem Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, P)$ ist die Indikatorfunktion $X(\omega) = \mathbf{1}_A(\omega) = \mathbf{1}(\omega \in A), \omega \in \Omega$, genau dann eine reellwertige Zufallsvariable, wenn $A \in \mathscr{F}$ gilt. Für ihre Verteilung gilt $P^X = P(A^c)\delta_0 + P(A)\delta_1$.

1.2 Diskrete Verteilungen

1.12 Definition. Ist Ω eine abzählbare (d.h. endliche oder abzählbar unendliche) Menge und P ein Wahrscheinlichkeitsmaß auf $\mathscr{F} = \mathscr{P}(\Omega)$, so heißt $(\Omega, \mathscr{F}, P)$ diskreter Wahrscheinlichkeitsraum. Man nennt eine S-wertige Zufallsvariable X diskret verteilt, falls sie bezüglich $\mathscr{P}(S)$ messbar ist und einen diskreten Wahrscheinlichkeitsraum $(S, \mathscr{P}(S), P^X)$ generiert. Ist X eine diskrete S-wertige Zufallsvariable und $S \subseteq \mathbb{R}^d$ (S abzählbar und somit $\mathscr{P}(S)$ Unter-σ-Algebra der Borel-σ-Algebra), so bezeichnet man X auch als diskrete $\mathbb{R}^d$-wertige Zufallsvariable.

1.13 Beispiel. Das Modell des Würfel- und des n-fachen Münzwurfs bildet einen diskreten Wahrscheinlichkeitsraum. Die Augensumme beim Münzwurf ist eine diskret verteilte Zufallsvariable.

1.14 Lemma.

(a) Ist $(\Omega, \mathscr{F}, P)$ ein diskreter Wahrscheinlichkeitsraum, so ist P eindeutig durch seine Zähldichte $p : \Omega \to [0,1]$ mit $p(\omega) := P(\{\omega\})$ festgelegt.

Ebenso legt bei einer diskret verteilten S-wertigen Zufallsvariablen X die zugehörige Zähldichte $p^X(s) = P(X = s)$, $s \in S$, die Verteilung P^X eindeutig fest.

(b) Ist andererseits Ω eine abzählbare Menge und besitzt $p : \Omega \to [0,1]$ die Eigenschaft $\sum_{\omega\in\Omega} p(\omega) = 1$, so wird durch

$$P(A) := \sum_{\omega\in A} p(\omega), \quad A \subseteq \Omega,$$

ein Wahrscheinlichkeitsmaß P auf $\mathscr{F} = \mathscr{P}(\Omega)$ definiert, dessen Zähldichte p ist.

Beweis.

(a) Wegen $\mathscr{F} = \mathscr{P}(\Omega)$ und Ω abzählbar gilt

$$P(A) = P\Big(\bigcup_{\omega\in A} \{\omega\}\Big) = \sum_{\omega\in A} P(\{\omega\}) = \sum_{\omega\in A} p(\omega), \quad A \in \mathscr{F}.$$

Dies gilt entsprechend auch für die Verteilung P^X einer diskreten Zufallsvariablen.

(b) Offensichtlich gilt $P(\Omega) = \sum_{\omega\in\Omega} p(\omega) = 1$. Für paarweise disjunkte $A_n \subseteq \Omega$ gilt

$$P\Big(\bigcup_{n\geqslant 1} A_n\Big) = \sum_{\omega\in\bigcup_{n\geqslant 1} A_n} p(\omega) = \sum_{n\geqslant 1} \sum_{\omega\in A_n} p(\omega) = \sum_{n\geqslant 1} P(A_n),$$

also σ-Additivität. Beachte dazu, dass Reihen mit nicht-negativen Gliedern beliebig umsortiert werden dürfen.

□

1.15 Lemma (Urnenmodelle)**.** *In einer Urne liegen N Kugeln mit den Aufschriften $1, 2, \ldots, N$. Es werden n Kugeln gezogen. Dann gilt für die Anzahl verschiedener Versuchsausgänge:*

Mit Zurücklegen, mit Betrachtung der Reihenfolge:
$\Omega_1 = \{1, \ldots, N\}^n$, $|\Omega_1| = N^n$.

Ohne Zurücklegen, mit Betrachtung der Reihenfolge:
$\Omega_2 = \{(k_1, \ldots, k_n) \,|\, k_1, \ldots, k_n \in \{1, \ldots, N\}$ *paarweise verschieden*$\}$, $|\Omega_2| = \frac{N!}{(N-n)!}$ *für* $n \leqslant N$.

Ohne Zurücklegen, ohne Betrachtung der Reihenfolge:
$\Omega_3 = \{A \subseteq \{1, \ldots, N\} \,|\, |A| = n\}$, $|\Omega_3| = \binom{N}{n}$ *für* $n \leqslant N$.

Mit Zurücklegen, ohne Betrachtung der Reihenfolge:
$\Omega_4 = \{(k_1, \ldots, k_n) \,|\, 1 \leqslant k_1 \leqslant k_2 \leqslant \cdots \leqslant k_n \leqslant N\}$, $|\Omega_4| = \binom{N+n-1}{n}$.

Beweis. Siehe Krengel, Abschnitt 1.2. □

1.16 Beispiel. Wie groß ist die Wahrscheinlichkeit, dass in einem Raum mit n Personen keine zwei Personen am selben Tag Geburtstag haben? Geht man von 365 Tagen im Jahr aus, so ist die Menge aller Geburtstagskombinationen gerade Ω_1 mit $N = 365$. Das Ereignis, das keine zwei Personen am selben Tag Geburtstag haben, entspricht dann gerade Ω_2. Unter der Annahme einer Gleichverteilung ergibt sich daher für die gesuchte Wahrscheinlichkeit $\frac{N!}{N^n(N-n)!}$. Approximativ ergibt sich $\exp(-n(n-1)/(2N))$ (wie?) und konkret $0,432$ für $n = 25$; $4,4 \times 10^{-4}$ für $n = 50$; $2,2 \times 10^{-9}$ für $n = 80$; $2,7 \times 10^{-14}$ für $n = 100$.

1.17 Definition. Die Laplace-/Gleich-Verteilung ist gegeben durch die Zähldichte $p_{Lap(\Omega)}(\omega) = \frac{1}{|\Omega|}$, $\omega \in \Omega$, auf einer endlichen Grundmenge Ω. Gilt $p^X = p_{Lap(\Omega)}$ für eine diskrete Zufallsvariable X, so sagen wir, dass X Laplace- oder gleichverteilt auf Ω ist, Notation $X \sim \mathrm{Lap}(\Omega)$.

1.18 Beispiel. Der Wurf zweier fairer Würfel kann mit $\Omega = \{1, 2, 3, 4, 5, 6\}^2$ und Zähldichte $p(\omega) = \frac{1}{36}$ der Gleichverteilung auf Ω modelliert werden.

1.19 Definition. Die hypergeometrische Verteilung mit Parametern $0 \leqslant n \leqslant N$, $0 \leqslant W \leqslant N$ ist auf $\Omega = \{0, \ldots, W\}$ gegeben durch die Zähldichte

$$p_{Hyp(N,W,n)}(w) = \frac{\binom{N-W}{n-w}\binom{W}{w}}{\binom{N}{n}}, \quad w \in \{0, \ldots, W\}.$$

Gilt $p^X = p_{Hyp(N,W,n)}$ für eine diskrete Zufallsvariable X, so sagen wir, dass X hypergeometrisch verteilt ist, Notation $X \sim \mathrm{Hyp}(N, W, n)$.

1.20 Beispiel. Lotto, siehe Übung.

1.21 Definition. Das Bernoulli-Schema (die Bernoulli-Kette) der Länge $n \in \mathbb{N}$ mit Erfolgswahrscheinlichkeit $p \in [0,1]$ auf $\Omega = \{0,1\}^n$ ist gegeben durch die Zähldichte

$$p_{Bern(n,p)}(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{\sum_{i=1}^n (1-\omega_i)}, \quad \omega = (\omega_1, \ldots, \omega_n) \in \{0,1\}^n.$$

1.22 Beispiel. n-facher Münzwurf mit einer Münze, die mit Wahrscheinlichkeit p 'Kopf' (also '1') zeigt.

1.23 Definition. Die Binomialverteilung mit Anzahl $n \in \mathbb{N}$ und Erfolgswahrscheinlichkeit $p \in [0,1]$ auf $\Omega = \{0, \ldots, n\}$ ist gegeben durch die Zähldichte

$$p_{Bin(n,p)}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}.$$

Gilt $p^X = p_{Bin(n,p)}$ für eine diskrete Zufallsvariable X, so sagen wir, dass X Binomial-verteilt ist, Notation $X \sim \mathrm{Bin}(n,p)$.

1.24 Beispiel. Die Binomialverteilung zählt die Anzahl der 'Erfolge' in einem Bernoulli-Schema. Betrachte dazu auf $\Omega = \{0,1\}^n$ die $\{0, \ldots, n\}$-wertige Zufallsvariable $X(\omega_1, \ldots, \omega_n) = \omega_1 + \cdots + \omega_n$. Dann gilt für $k \in \{0, \ldots, n\}$

$$\begin{aligned} P(X = k) &= \sum_{\omega : X(\omega) = k} p_{Bern(n,p)}(\omega) \\ &= \sum_{\omega : \omega_1 + \cdots + \omega_n = k} p^k (1-p)^{n-k} \\ &= \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned}$$

Also ist die Verteilung P^X von X gerade die Binomialverteilung mit Parametern n und p, kurz $X \sim \mathrm{Bin}(n,p)$.

Ende der 2. Vorlesung

1.25 Definition. Die Multinomialverteilung mit Anzahl $n \in \mathbb{N}$, Klassenzahl $r \in \mathbb{N}$ und Erfolgswahrscheinlichkeiten $p_1, \ldots, p_r \in [0,1]$ mit $\sum_{i=1}^r p_i = 1$ ist gegeben auf $\Omega = \{k \in \{0, \ldots, n\}^r \mid k_1 + \cdots + k_r = n\}$ durch die Zähldichte

$$p_{Mult(n,r,p_1,\ldots,p_r)}(k) = \frac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r}, \quad k \in \Omega.$$

Gilt $p^X = p_{Mult(n,r,p_1,\ldots,p_r)}$ für eine diskrete Zufallsvariable X, so sagen wir, dass X Multinomial-verteilt ist, Notation $X \sim \mathrm{Mult}(n, r, p_1, \ldots, p_r)$.

1.26 Beispiel. Die Multinomialverteilung $\mathrm{Mult}(n, r, p_1, \ldots, p_r)$ zählt die Anzahl der Versuchsausgänge in r Klassen bei n Versuchen und Klassenwahrscheinlichkeiten $p_1, \ldots, p_r$. Werden die Ziffern '0' bis '9' rein zufällig n-mal erzeugt, so kann man die erhaltenen Häufigkeiten der Ziffern mit der $\mathrm{Mult}(n, 10, \frac{1}{10}, \ldots, \frac{1}{10})$-Verteilung beschreiben. Im Fall $r = 2$ erhalten wir $k_2 = n - k_1$ für $k \in \Omega$ und somit $p_{Mult(n,2,p,1-p)}(k) = p_{Bin(n,p)}(k_1)$. Mehr zur Multinomialverteilung in Abschnitt 2.2.2 bei Georgii.

1.27 Definition. Die geometrische Verteilung mit Erfolgswahrscheinlichkeit $p \in (0,1]$ ist auf $\Omega = \mathbb{N}$ gegeben durch die Zähldichte

$$p_{Geo(p)}(k) = (1-p)^{k-1}p, \quad k \in \mathbb{N}\,.$$

Gilt $p^X = p_{Geo(p)}$ für eine diskrete Zufallsvariable X, so sagen wir, dass X geometrisch verteilt ist, Notation $X \sim \mathrm{Geo}(p)$.

1.28 Beispiel. Die geometrische Verteilung beschreibt bei einem beliebig langen Bernoulli-Schema (intuitiv, da noch nicht formal konstruiert) die Anzahl der Versuche bis zum ersten Erfolg. Betrachte dazu $X : \{0,1\}^{\mathbb{N}} \to \mathbb{N}$ mit

$$X((\omega_i)_{i\geqslant 1}) = \begin{cases} 1, & \text{falls } \forall i : \omega_i = 0, \\ \min\{i \in \mathbb{N} \mid \omega_i = 1\}, & \text{sonst.} \end{cases}$$

Dann gilt (weiter intuitiv)

$$P(X=k) = P(\forall i < k : \omega_i = 0, \omega_k = 1) = p_{Bern(k,p)}(0,\dots,0,1) = (1-p)^{k-1}p,$$

und X ist geometrisch verteilt. Beachte dazu, dass das Ereignis $\{\forall i \in \mathbb{N} : \omega_i = 0\}$ wegen $p > 0$ Wahrscheinlichkeit $\lim_{n\to\infty}(1-p)^n = 0$ besitzt und der Wert 1 für X auf diesem Ereignis völlig beliebig war. Für eine rigorose Herleitung muss der Wahrscheinlichkeitsraum des beliebig langen Bernoulli-Schemas konstruiert und die Messbarkeit von X nachgewiesen werden, was weiter unten geschehen wird. Natürlich ist die geometrische Verteilung auch ohne diese Interpretation stets wohldefiniert.

1.29 Definition. Die Poissonverteilung mit Parameter $\lambda > 0$ ist auf $\Omega = \mathbb{N}_0$ gegeben durch die Zähldichte

$$p_{Pois(\lambda)}(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{N}_0\,.$$

Gilt $p^X = p_{Pois(\lambda)}$ für eine diskrete Zufallsvariable X, so sagen wir, dass X Poisson-verteilt ist, Notation $X \sim \mathrm{Pois}(\lambda)$.

1.30 Satz (Poissonscher Grenzwertsatz)**.** *Es seien $p_n \in [0,1]$ gegeben mit $\lim_{n\to\infty} np_n = \lambda > 0$. Dann gilt für alle $k \in \mathbb{N}_0$*

$$\lim_{n\to\infty} p_{Bin(n,p_n)}(k) = p_{Pois(\lambda)}(k).$$

Beweis. Wir schreiben $A_n \asymp B_n$, falls $A_n/B_n \to 1$ für $n \to \infty$. Dann ergibt sich direkt aus $p_n \to 0$ und $np_n \to \lambda$ (beachte $\log((1-p_n)^n) = n\log(1-p_n) = -np_n + O(np_n^2)$) für festes $k \in \mathbb{N}_0$:

$$\begin{aligned} p_{Bin(n,p_n)}(k) &= \frac{n^k}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}p_n^k(1-p_n)^n(1-p_n)^{-k} \\ &\asymp \frac{n^k}{k!}p_n^k e^{-np_n} \asymp \frac{\lambda^k}{k!}e^{-\lambda}. \end{aligned}$$

Wir erhalten also $p_{Pois(\lambda)}(k)$ als Grenzwert, wie behauptet. □

1.31 Bemerkung. Natürlich ist auch eine nicht-asymptotische Fehlerabschätzung wichtig. Es gilt

$$\sum_{k\geqslant 0}|p_{Bin(n,p)}(k) - p_{Pois(np)}(k)| \leqslant 2np^2,$$

wobei $p_{Bin(n,p)}(k) = 0$ für $k > n$ gesetzt wird. Dies kann man rein probabilistisch mit einem sogenannten Kopplungsargument beweisen, vergleiche Satz 5.34 in Georgii oder Satz 5.9 in Krengel.

1.32 Beispiel. Die Poissonverteilung heißt auch „Verteilung seltener Ereignisse“ wegen $p_n \to 0$ im Poissonschen Grenzwertsatz. Typische Beispiele sind die Anzahl von Geburten an einem Tag in einer Kleinstadt oder die Anzahl der radioaktiven Zerfälle in einem Präparat in einer Zeiteinheit.

1.3 Maßtheorie: allgemein und im $\mathbb{R}^d$

1.33 Satz (Vitali, 1903)**.** *Sei $\Omega = \{0,1\}^{\mathbb{N}}$. Dann gibt es kein Wahrscheinlichkeitsmaß P auf der Potenzmenge $\mathscr{P}(\Omega)$, das folgender Invarianzeigenschaft genügt:*

$$\forall A \subseteq \Omega,\, n \in \mathbb{N}:\ P(T_n(A)) = P(A),$$

wobei $T_n(\omega) = T_n(\omega_1, \omega_2, \ldots) = (\omega_1, \ldots, \omega_{n-1}, 1-\omega_n, \omega_{n+1}, \ldots)$ das Ergebnis des n-ten Wurfs umkehrt.

1.34 Bemerkung. Der Satz zeigt, dass wir das beliebig lange Münzwurfexperiment oder Bernoulli-Schema mit $p = 1/2$ nicht auf der Potenzmenge definieren können. Der Beweis beruht auf dem Auswahlaxiom. Ein anderer Satz von Vitali besagt, dass auch das Lebesguemaß nicht auf der Potenzmenge von $\mathbb{R}$ oder von $[0,1]$ definiert werden kann. Dieser kann aus dem hier angegebenen Satz per Widerspruch hergeleitet werden. Idee: Betrachte $X_k : [0,1] \to \{0,1\}$, $k \geqslant 1$, mit $X_k(u) = \mathbf{1}_{[1/2,1)}(2^{k-1}u \mod 1)$, so dass $u = \sum_{k\geqslant 1} X_k 2^{-k}$ (Binärdarstellung). Dann ist $X(u) := (X_k(u))_{k\geqslant 1}$ eine Abbildung von $[0,1]$ nach $\{0,1\}^{\mathbb{N}}$. Wäre λ das (eingeschränkte) Lebesguemaß auf $\mathscr{P}([0,1])$, so würde die Verteilung $P = \lambda^X$ von X die hier geforderten Eigenschaften erfüllen. Widerspruch!

Beweis. Wir definieren eine Äquivalenzrelation $\sim$ auf Ω:

$$\omega \sim \omega' \iff \exists N \geqslant 1 \forall n \geqslant N:\ \omega_n = \omega'_n.$$

Also sind zwei Versuchsausgänge äquivalent, wenn sie sich nur an endlich vielen Stellen unterscheiden. Nach dem Auswahlaxiom existiert eine Menge $A \subseteq \Omega$, die aus jeder Äquivalenzklasse genau einen Repräsentanten enthält. Es sei $\mathscr{S} := \{S \subseteq \mathbb{N} \mid S \text{ endlich}\}$. Wegen $\mathscr{S} = \bigcup_{m\geqslant 1}\{S \subseteq \mathbb{N} \mid \max S = m\}$ ist $\mathscr{S}$ abzählbar unendlich. Für $S = \{n_1, \ldots, n_k\} \in \mathscr{S}$ setze $T_S := T_{n_1} \circ T_{n_2} \circ \cdots \circ T_{n_k}$ ('flip' bei den Stellen in S).

Dann gilt $\Omega = \bigcup_{S\in\mathscr{S}} T_S(A)$, weil zu jedem $\omega \in \Omega$ ein $\omega' \in A$ existiert mit $\omega' \sim \omega$, also auch ein $S \in \mathscr{S}$ mit $\omega = T_S(\omega')$. Außerdem sind die Mengen $T_S(A)$, $S \in \mathscr{S}$, paarweise disjunkt; denn $T_S(\omega) = T_{S'}(\omega')$ impliziert $\omega \sim T_S(\omega) = T_{S'}(\omega') \sim \omega'$ und somit folgt für $\omega, \omega' \in A$, dass $\omega = \omega'$ gilt (A enthält genau

einen Repräsentanten jeder Äquivalenzklasse) und folglich auch $S = S'$ (nach Definition von T_S, $T_{S'}$). Wir schließen aus den Voraussetzungen an P

$$1 = P(\Omega) = P\Big(\bigcup_{S\in\mathscr{S}} T_S(A)\Big) = \sum_{S\in\mathscr{S}} P(T_S(A)) = \sum_{S\in\mathscr{S}} P(A).$$

Weil die Reihe rechts entweder null oder unendlich ist, ist dies ein Widerspruch, und ein solches P kann es nicht geben. □

1.35 Bemerkung. Im folgenden tragen wir ohne Beweis Grundlagen der Maßtheorie zusammen, wie sie in Analysis III gelehrt werden, vergleiche die Übungen. Eine ausführliche Referenz ist Elstrodt.

1.36 Lemma. *Es sei $\mathscr{E} \subseteq \mathscr{P}(\Omega)$ ein System von Teilmengen von Ω. Dann gibt es eine kleinste σ-Algebra $\mathscr{F}$, die $\mathscr{E}$ enthält.*

1.37 Definition. In der Situation des vorigen Lemmas sagt man, dass die σ-Algebra $\mathscr{F}$ von $\mathscr{E}$ erzeugt wird. $\mathscr{E}$ heißt Erzeuger von $\mathscr{F}$ und man schreibt $\mathscr{F} = \sigma(\mathscr{E})$.

1.38 Definition. Es sei (S, d) ein metrischer Raum. Dann heißt $\mathfrak{B}_S := \sigma(\{O \subseteq S \mid O \text{ offen}\})$ Borel-σ-Algebra über S.

1.39 Satz.

(a) Die Borel-σ-Algebra $\mathfrak{B}_\mathbb{R}$ über $\mathbb{R}$ wird auch erzeugt von folgenden Mengensystemen:

(i) $\mathscr{E}_1 := \{(a, b) \mid a, b \in \mathbb{R}\}$;
(ii) $\mathscr{E}_2 := \{[a, b] \mid a, b \in \mathbb{R}\}$;
(iii) $\mathscr{E}_3 := \{(a, b] \mid a, b \in \mathbb{R}\}$;
(iv) $\mathscr{E}_4 := \{(-\infty, b] \mid b \in \mathbb{R}\}$;
(v) $\mathscr{E}_5 := \{(-\infty, b) \mid b \in \mathbb{R}\}$.

(b) Die Borel-σ-Algebra $\mathfrak{B}_{\mathbb{R}^d}$ über $\mathbb{R}^d$ wird auch erzeugt von folgenden Mengensystemen:

(i) $\mathscr{E}_1^d := \{(a_1, b_1) \times \cdots \times (a_d, b_d) \mid a_k, b_k \in \mathbb{R}, k = 1, \ldots, d\}$;
(ii) $\mathscr{E}_2^d := \{[a_1, b_1] \times \cdots \times [a_d, b_d] \mid a_k, b_k \in \mathbb{R}, k = 1, \ldots, d\}$;
(iii) $\mathscr{E}_3^d := \{(a_1, b_1] \times \cdots \times (a_d, b_d] \mid a_k, b_k \in \mathbb{R}, k = 1, \ldots, d\}$;
(iv) $\mathscr{E}_4^d := \{(-\infty, b_1] \times \cdots \times (-\infty, b_d] \mid b_k \in \mathbb{R}, k = 1, \ldots, d\}$;
(v) $\mathscr{E}_5^d := \{(-\infty, b_1) \times \cdots \times (-\infty, b_d) \mid b_k \in \mathbb{R}, k = 1, \ldots, d\}$.

1.40 Lemma. *Eine Funktion $g : \Omega \to S$ ist bereits $(\mathscr{F}, \mathscr{S})$-messbar, falls für einen Erzeuger $\mathscr{E}$ von $\mathscr{S}$ gilt*

$$\forall\, A \in \mathscr{E} :\ g^{-1}(A) \in \mathscr{F}.$$

1.41 Korollar.

(a) *Jede stetige Funktion $g : S \to T$ zwischen metrischen Räumen (S, d_S) und (T, d_T) ist Borel-messbar, d.h. $(\mathfrak{B}_S, \mathfrak{B}_T)$-messbar.*

(b) *Jede Funktion $g : \Omega \to \mathbb{R}$ mit $\{g \leqslant y\} \in \mathscr{F}$ für alle $y \in \mathbb{R}$ ist $(\mathscr{F}, \mathfrak{B}_{\mathbb{R}})$-messbar.*

(c) *Falls $g_n : \Omega \to \mathbb{R}$ $(\mathscr{F}, \mathfrak{B}_{\mathbb{R}})$-messbar sind für alle $n \geqslant 1$, so auch $\inf_n g_n$, $\sup_n g_n$, $\limsup_n g_n$, $\liminf_n g_n$, sofern diese Funktionen endlich sind. Falls der punktweise Grenzwert $\lim_n g_n$ überall existiert, so ist auch dieser $(\mathscr{F}, \mathfrak{B}_{\mathbb{R}})$-messbar.*

(d) *Sind $g_1, \dots, g_d : \Omega \to \mathbb{R}$ $(\mathscr{F}, \mathfrak{B}_{\mathbb{R}})$-messbar und ist $h : \mathbb{R}^d \to \mathbb{R}^k$ Borel-messbar, so ist $\omega \mapsto h(g_1(\omega), \dots, g_d(\omega))$ $(\mathscr{F}, \mathfrak{B}_{\mathbb{R}^k})$-messbar; insbesondere sind also messbar: $(g_1, \dots, g_d)$, $g_1 + g_2$, $g_1 - g_2$, $g_1 \bullet g_2$, g_1/g_2 (falls überall wohldefiniert), $\max(g_1, g_2)$, $\min(g_1, g_2)$.*

(e) *Ist $g : \Omega \to S$ $(\mathscr{F}, \mathscr{S})$-messbar und $h : S \to T$ $(\mathscr{S}, \mathscr{T})$-messbar, so ist die Komposition $h \circ g$ $(\mathscr{F}, \mathscr{T})$-messbar.*

1.42 Definition. Es sei Ω eine nichtleere Menge. Dann heißt $\mathscr{A} \subseteq \mathscr{P}(\Omega)$ Algebra über Ω, falls gilt:

(a) $\Omega \in \mathscr{A}$;

(b) $A \in \mathscr{A} \Rightarrow A^c \in \mathscr{A}$;

(c) $A, B \in \mathscr{A} \Rightarrow A \cup B \in \mathscr{A}$.

Eine Abbildung $\mu : \mathscr{A} \to [0, \infty]$ heißt Prämaß über $\mathscr{A}$, falls

(a) $\mu(\varnothing) = 0$;

(b) für $A_n \in \mathscr{A}$, $n \in \mathbb{N}$, paarweise disjunkt mit $\bigcup_n A_n \in \mathscr{A}$ gilt

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n) \text{ (σ-Additivität).}$$

μ heißt Maß, falls $\mathscr{A}$ bereits eine σ-Algebra ist. Ein Maß μ heißt σ-endlich, falls es $A_n \in \mathscr{A}$, $n \in \mathbb{N}$, gibt mit $\mu(A_n) < \infty$ und $\Omega = \bigcup_n A_n$. Konsistent mit obiger Definition heißt ein Maß μ Wahrscheinlichkeitsmaß, falls $\mu(\Omega) = 1$ gilt.

1.43 Satz (Maßerweiterungssatz von Carathéodory, 1917)**.** *Jedes Prämaß μ auf einer Algebra $\mathscr{A}$ kann zu einem Maß $\tilde{\mu}$ auf der von $\mathscr{A}$ erzeugten σ-Algebra $\mathscr{F} = \sigma(\mathscr{A})$ fortgesetzt werden, d.h. $\tilde{\mu}$ ist ein Maß auf $\mathscr{F}$ mit $\tilde{\mu}(A) = \mu(A)$ für alle $A \in \mathscr{A}$.*

1.44 Satz (Eindeutigkeitssatz)**.** *Es seien μ und ν (σ-endliche) Maße auf $(\Omega, \mathscr{F})$ und es gebe $A_n \in \mathscr{F}$, $n \in \mathbb{N}$, mit $\mu(A_n) = \nu(A_n) < \infty$ und $\bigcup_n A_n = \Omega$. Stimmen μ und ν auf einem Erzeuger $\mathscr{E}$ von $\mathscr{F}$ überein, der in dem Sinne $\cap$-stabil ist, dass $A, B \in \mathscr{E} \Rightarrow A \cap B \in \mathscr{E}$ gilt, so stimmen μ und ν auf der ganzen σ-Algebra $\mathscr{F}$ überein. Insbesondere ist ein Wahrscheinlichkeitsmaß durch seine Werte auf einem $\cap$-stabilen Erzeuger eindeutig festgelegt (wähle $A_n = \Omega$).*

1.45 Bemerkung. Wir wollen jetzt Wahrscheinlichkeitsmaße auf $\mathbb{R}$ beschreiben, wozu folgender Begriff grundlegend ist.

1.46 Definition. Für ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ ist die zugehörige Verteilungsfunktion gegeben durch $F(x) := P((-\infty, x])$, $x \in \mathbb{R}$; für $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$-wertige Zufallsvariablen X wird durch $F^X(x) := P^X((-\infty, x]) = P(X \leqslant x)$, $x \in \mathbb{R}$, die zugehörige Verteilungsfunktion definiert.

1.47 Beispiel. Ist $P = \delta_{x_0}$ eine Einpunktverteilung, so ist $F(x) = \mathbf{1}_{[x_0,\infty)}(x)$.

Ende der 3. Vorlesung

1.48 Lemma. *Jede Verteilungsfunktion F ist monoton wachsend, rechtsstetig und erfüllt $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$.*

Beweis. Aus der Monotonie von P folgt die Monotonie von F; denn für $x < y$ gilt

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

Die σ-Stetigkeit von P impliziert die Rechtsstetigkeit von F wegen $(-\infty, x_n] \downarrow (-\infty, x]$ für jede Folge $x_n \downarrow x$ und somit $P((-\infty, x]) = \lim_{n\to\infty} P((-\infty, x_n])$. Ebenso folgt $F(x_n) \to 0$ für $x_n \downarrow -\infty$ aus $(-\infty, x_n] \downarrow \varnothing$ und $F(x_n) \to 1$ für $x_n \uparrow \infty$ aus $(-\infty, x_n] \uparrow \mathbb{R}$. □

1.49 Satz. *Es sei $F : \mathbb{R} \to \mathbb{R}$ eine monoton wachsende, rechtsstetige Funktion. Dann existiert ein Maß μ auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ mit*

$$\mu((a, b]) = F(b) - F(a), \quad a < b \in \mathbb{R}\,.$$

μ ist eindeutig durch F definiert und heißt Lebesgue-Stieltjes-Maß zu F.

Beweis. Die Eindeutigkeit folgt mit dem Eindeutigkeitssatz, weil $\{(a, b] \mid a, b \in \mathbb{R}, a < b\} \cup \{\varnothing\}$ ein $\cap$-stabiler Erzeuger von $\mathfrak{B}_\mathbb{R}$ ist und $\mu((a, b]) = F(b) - F(a)$, $\mu(\varnothing) = 0$ eindeutig festgelegt ist.

Betrachte

$$\mathscr{A} := \Big\{ \bigcup_{k=1}^{K} (a_k, b_k] \,\Big|\, K \geqslant 1, -\infty \leqslant a_1 < b_1 < \cdots < a_K < b_K \leqslant \infty \Big\} \cup \{\varnothing\},$$

wobei wir $(a_K, \infty] := (a_K, \infty)$ setzen. Dann ist $\mathbb{R} \in \mathscr{A}$ (setze $K = 1$, $a_1 = -\infty$, $b_1 = \infty$), $\mathscr{A}$ ist additiv (Vereinigungen von links-offenen, rechts-abgeschlossenen Intervallen sind $\cup$-stabil) und es gilt $A \in \mathscr{A} \Rightarrow A^c \in \mathscr{A}$ (wegen $(a, b]^c = (-\infty, a] \cup (b, \infty]$ und $\cap$-Stabilität). Damit ist $\mathscr{A}$ eine Algebra. Auf $\mathscr{A}$ definiere

$$\mu\Big(\bigcup_{k=1}^{K} (a_k, b_k] \Big) := \sum_{k=1}^{K} (F(b_k) - F(a_k)) \in [0, \infty] \text{ mit } F(\pm\infty) := \lim_{x\to\pm\infty} F(x),$$

$\mu(\varnothing) := 0$. Da die Intervalle disjunkt sind, ist μ offensichtlich wohldefiniert und additiv auf $\mathscr{A}$, d.h. $\mu(A \cup B) = \mu(A) + \mu(B)$ für disjunkte $A, B \in \mathscr{A}$.

Der Nachweis, dass μ ein Prämaß, also σ-additiv ist, ist nicht-trivial. Seien dazu $A_n = \bigcup_{k=1}^{K^n}(a_k^n, b_k^n] \in \mathscr{A}$, $n \geqslant 1$, paarweise disjunkt und $A_\infty := \bigcup_{n\geqslant 1} A_n \in \mathscr{A}$. Schreibe $A_\infty = \bigcup_{k=1}^{K^\infty}(a_k^\infty, b_k^\infty]$ (mit $K^\infty < \infty$ nach Voraussetzung!). Dann ist zu zeigen:

$$\sum_{k=1}^{K^\infty} \mu((a_k^\infty, b_k^\infty]) = \sum_{n\geqslant 1}\sum_{k=1}^{K^n} \mu((a_k^n, b_k^n]).$$

Es reicht, dies für ein Intervall links nachzuweisen und dann endliche Additivität zu benutzen. Wir müssen also zeigen ($\dot{\bigcup} B_n$ bezeichne die Vereinigung von paarweise disjunkten B_n):

$$(a^\infty, b^\infty] = \dot{\bigcup}_{n\geqslant 1}(a^n, b^n] \Rightarrow \mu((a^\infty, b^\infty]) = \sum_{n\geqslant 1}\mu((a^n, b^n]).$$

Wegen Additivität von μ gilt die Monotonie $\mu(\dot{\bigcup}_{n=1}^{N}(a^n, b^n]) \leqslant \mu(\dot{\bigcup}_{n=1}^{\infty}(a^n, b^n])$ für alle N und daher $\mu((a^\infty, b^\infty]) \geqslant \sum_{n\geqslant 1}\mu((a^n, b^n])$. Es reicht also, zu zeigen:

$$(a^\infty, b^\infty] = \dot{\bigcup}_{n\geqslant 1}(a^n, b^n] \Rightarrow \mu((a^\infty, b^\infty]) \leqslant \sum_{n\geqslant 1}\mu((a^n, b^n]).$$

Hierzu benötigen wir ein Kompaktheitsargument. Dazu sei zunächst $a^\infty > -\infty$ und $b^\infty < \infty$. Betrachte offene Intervalle $(a^n, b^n + \delta^n) \supseteq (a^n, b^n]$ mit $\delta^n > 0$, so dass $\mu((b^n, b^n + \delta^n]) \leqslant \varepsilon 2^{-n}$ für ein $\varepsilon > 0$. Die Wahl von δ^n ist möglich, weil F rechtsstetig ist. Wähle noch $\delta^\infty > 0$ mit $\mu((a^\infty, a^\infty + \delta^\infty]) \leqslant \varepsilon$. Dann erhalten wir die offene Überdeckung

$$[a^\infty + \delta^\infty, b^\infty] \subseteq (a^\infty, b^\infty] \subseteq \bigcup_{n\geqslant 1}(a^n, b^n + \delta^n).$$

Nun ist $[a^\infty + \delta^\infty, b^\infty]$ kompakt, und es gilt bereits $[a^\infty + \delta^\infty, b^\infty] \subseteq \bigcup_{n=1}^{N}(a^n, b^n + \delta^n)$ für ein endliches $N \in \mathbb{N}$. Wir können also Additivität von μ (und Monotonie) verwenden und erhalten

$$\begin{aligned}
\mu((a^\infty, b^\infty]) &= \mu((a^\infty + \delta^\infty, b^\infty]) + \mu((a^\infty, a^\infty + \delta^\infty]) \\
&\leqslant \sum_{n=1}^{N} \mu((a^n, b^n + \delta^n]) + \varepsilon \\
&\leqslant \sum_{n=1}^{N} (\mu((a^n, b^n]) + \varepsilon 2^{-n}) + \varepsilon \\
&\leqslant \sum_{n\geqslant 1} \mu((a^n, b^n]) + 2\varepsilon.
\end{aligned}$$

Mit $\varepsilon \downarrow 0$ folgt die gewünschte Ungleichung. Erlauben wir auch die Werte $a^\infty = -\infty$ und $b^\infty = \infty$, so haben wir jedenfalls $\mu(((-R) \vee a^\infty, R \wedge b^\infty]) \leqslant \sum_{n\geqslant 1}\mu((a^n, b^n])$ (mit $A \vee B := \max(A, B)$, $A \wedge B := \min(A, B)$) für alle $R > 0$ gezeigt. Wegen Monotonie von F gilt $\lim_{R\to\infty} F(R \wedge b^\infty) = F(b^\infty) \in \mathbb{R} \cup \{\infty\}$ auch für $b^\infty = \infty$ und analog $\lim_{R\to\infty} F((-R) \vee a^\infty) = F(a^\infty)$. So können wir schließen $\mu((a^\infty, b^\infty]) = \lim_{R\to\infty}\mu(((-R) \vee a^\infty, R \wedge b^\infty]) \leqslant \sum_{n\geqslant 1}\mu((a^n, b^n])$.

μ ist also ein Prämaß auf $\mathscr{A}$ und lässt sich mit dem Satz 1.43 von Caratheodory auf $\sigma(\mathscr{A}) = \mathfrak{B}_{\mathbb{R}}$ fortsetzen. □

1.50 Korollar. *Es gibt genau ein Maß λ auf $(\mathbb{R}, \mathfrak{B}_{\mathbb{R}})$ mit $\lambda((a,b]) = b-a$, das Lebesguemaß.*

Beweis. Wähle $F(x) = x$. □

1.51 Korollar. *Ist $F : \mathbb{R} \to [0,1]$ monoton wachsend und rechtsstetig mit $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$, so existiert genau ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathfrak{B}_{\mathbb{R}})$ mit $P((a,b]) = F(b)-F(a)$ für alle $a < b$. Insbesondere ist F die Verteilungsfunktion von P.*

1.52 Bemerkung. Es gibt also eine 1-1-Beziehung zwischen Verteilungsfunktionen und Wahrscheinlichkeitsmaßen auf $\mathbb{R}$. Dies lässt sich geeignet auch auf den $\mathbb{R}^d$ verallgemeinern. Auf allgemeinen Messräumen lassen sich Maße allerdings nicht mehr einfach durch Funktionen beschreiben.

Beweis. Nach Satz 1.49 ist P zunächst ein Maß mit diesen Eigenschaften. Wegen (σ-Stetigkeit!)

$$P(\mathbb{R}) = \lim_{R\to\infty} P((-R,R]) = \lim_{R\to\infty} (F(R) - F(-R)) = 1 - 0 = 1$$

folgt, dass P ein Wahrscheinlichkeitsmaß ist. Analog folgt $P((-\infty,x]) = \lim_{R\to\infty}(F(x) - F(-R)) = F(x)$ und F ist Verteilungsfunktion von P. □

1.53 Definition. Für eine Borelmenge $A \subseteq \mathbb{R}$ mit Lebesguemaß $\lambda(A) \in (0,\infty)$ ist die gleichmäßige Verteilung auf A gegeben durch das Wahrscheinlichkeitsmaß (!)

$$P(B) = \frac{\lambda(A \cap B)}{\lambda(A)}, \quad B \in \mathfrak{B}_{\mathbb{R}}.$$

Ist P die Verteilung einer reellwertigen Zufallsvariablen X, so schreiben wir $X \sim U(A)$.

1.54 Beispiel. Für eine reellwertige Zufallsvariable X mit Verteilungsfunktion F^X betrachte $U = F^X(X)$. F^X ist Borel-messbar (Beweis?) und somit U eine Zufallvariable mit Werten in $[0,1]$. Ist F^X stetig, so gilt mit der Rechtsinversen $(F^X)^{-1}(p) = \inf\{x \in \mathbb{R} \mid F^X(x) \geqslant p\}$, dass U die Verteilungsfunktion $F^U(u) = P(F^X(X) \leqslant u) = F^X((F^X)^{-1}(u)) = u$ für $u \in (0,1)$ besitzt und somit $U((0,1))$-verteilt ist.

Ist andererseits U eine $U((0,1))$-verteilte Zufallsvariable (oder eine Pseudozufallszahl in Anwendungen), so gilt für jede Verteilungsfunktion F, dass die Zufallsvariable $X = F^{-1}(U)$ die Verteilungsfunktion $F^X = F$ besitzt. Die Rechtsinverse F^{-1} heißt auch Quantilsfunktion und die Simulationsmethode Quantilstransformation. Eine einfache Anwendung ist die Erzeugung einer $\mathrm{Bin}(1,p)$-verteilten Zufallsvariablen X: es gilt $F^X(x) = (1-p)\mathbf{1}(x \geqslant 0) + p\mathbf{1}(x \geqslant 1)$ und $(F^X)^{-1}(y) = \mathbf{1}(y > 1-p)$, $y \in (0,1)$, so dass $\mathbf{1}(U > 1-p)$ $\mathrm{Bin}(1,p)$-verteilt ist für eine $U((0,1))$-verteilte Zufallsvariable U.

Ende der 4. Vorlesung

1.55 Definition. Ist $f : \mathbb{R}^d \to [0, \infty)$ eine Lebesgue-integrierbare Funktion mit $\int_{\mathbb{R}^d} f(x)\, dx = 1$, so heißt f Wahrscheinlichkeitsdichte oder kurz Dichte auf $\mathbb{R}^d$.

1.56 Satz. *Jede Wahrscheinlichkeitsdichte f auf $\mathbb{R}$ erzeugt mittels*

$$P_f((a, b]) = \int_a^b f(x)\, dx, \quad a, b \in \mathbb{R},\ a < b,$$

ein eindeutiges Wahrscheinlichkeitsmaß P_f auf $\mathfrak{B}_\mathbb{R}$. Es gilt dann

$$P_f(B) = \int_B f(x)\, dx, \quad B \in \mathfrak{B}_\mathbb{R}.$$

Aus $\lambda(B) = 0$ für ein $B \in \mathfrak{B}_\mathbb{R}$ (B ist Lebesgue-Nullmenge) folgt $P_f(B) = 0$.

1.57 Bemerkungen.

(a) Es gilt $P_f = P_g$ genau dann, wenn $f = g$ Lebesgue-fast überall. Wir können also eine Wahrscheinlichkeitsdichte auf einer Lebesgue-Nullmenge abändern, ohne das induzierte Wahrscheinlichkeitsmaß zu verändern.

(b) Der Satz spiegelt wider, was wir in Anwendungen oft wollen: wir geben eine Wahrscheinlichkeit für Intervalle durch ein Integral (oft als Riemann-Integral über eine stetige Dichte f) vor und erhalten so ein eindeutiges Wahrscheinlichkeitsmaß auf allen Borelmengen (mittels Lebesgueintegral). So können wir also auch komplizierteren Ereignissen Wahrscheinlichkeiten zuordnen und auf alle Werkzeuge der Wahrscheinlichkeitstheorie zurückgreifen.

Beweis. Durch $Q(B) = \int_B f(x)\, dx$, $B \in \mathfrak{B}_\mathbb{R}$, wird ein Wahrscheinlichkeitsmaß beschrieben (σ-Additivität folgt aus monotoner Konvergenz) mit $Q((a, b]) = \int_a^b f(x)dx$. Da die Intervalle $(a, b]$ mit $a < b$ einen $\cap$-stabilen Erzeuger von $\mathfrak{B}_\mathbb{R}$ bilden, liefert der Eindeutigkeitssatz $P_f = Q$ und somit Existenz und Eindeutigkeit von P_f sowie die Formel für $P_f(B)$. Für das Lebesgueintegral gilt $\int_B f(x)dx = 0$, wann immer $\lambda(B) = 0$ und f integrierbar ist. Also erhalten wir $\lambda(B) = 0 \Rightarrow P_f(B) = 0$. □

1.58 Beispiel. Es sei $f(x) = \min(x_+, (2 - x)_+)$, $x \in \mathbb{R}$ (mit $a_+ = \max(a, 0)$). Dann ist f stetig, also Borel-messbar, und es gilt $f \geqslant 0$ sowie $\int_\mathbb{R} f(x)\, dx = 1$ (Fläche des Dreiecks mit Grundseite $[0, 2]$ und Höhe 1). f ist also eine Wahrscheinlichkeitsdichte, und die zugehörige Verteilung P_f heißt Dreiecksverteilung.

1.59 Bemerkung. Sind alle μ-Nullmengen (Ereignisse A mit $\mu(A) = 0$) auch ν-Nullmengen ($\nu(A) = 0$) für Maße μ, ν, so heißt ν absolutstetig bezüglich μ, Notation $\nu \ll \mu$. P_f ist also absolutstetig bezüglich dem Lebesguemaß. Der Satz von Radon-Nikodym (Stochastik II, Funktionalanalysis) besagt in diesem Fall gerade, dass alle bezüglich Lebesguemaß absolutstetigen Maße von der Form P_f sind. Oft wird einfach nur von stetigen Verteilungen gesprochen, was

nicht ganz korrekt ist; denn nicht jede stetige Verteilungsfunktion definiert eine absolutstetige Verteilung. Vergleiche das sogenannte *Cantormaß*, das weder eine diskrete Verteilung ist noch absolutstetig bezüglich dem Lebesguemaß, aber eine stetige Verteilungsfunktion (die *Cantorfunktion*) besitzt.

1.60 Lemma.

(a) Ist f die Dichte eines Wahrscheinlichkeitsmaßes P auf $\mathfrak{B}_\mathbb{R}$ mit Verteilungsfunktion F, so gilt $F(x) = \int_{-\infty}^{x} f(y)\,dy$ für alle $x \in \mathbb{R}$.

(b) Ist die Verteilungsfunktion F eines Wahrscheinlichkeitsmaßes P auf $\mathfrak{B}_\mathbb{R}$ schwach differenzierbar, so ist $f := F'$ die zugehörige Wahrscheinlichkeitsdichte.

1.61 Bemerkung. Eine Funktion $g : \mathbb{R} \to \mathbb{R}$ heißt schwach differenzierbar, falls es eine Lebesgue-integrierbare Funktion $h : \mathbb{R} \to \mathbb{R}$ gibt mit $g(x) = g(0) + \int_0^x h(y)\,dy$ für alle $x \in \mathbb{R}$. Man nennt h schwache Ableitung von g und setzt $g' = h$. Nach dem Hauptsatz ist jede differenzierbare Funktion schwach differenzierbar. Beispielsweise ist die Verteilungsfunktion $F(x) = (1 - e^{-x})_+$ schwach differenzierbar mit Dichte $f(x) = F'(x) = e^{-x}\mathbf{1}(x > 0)$ (Exponentialverteilung; siehe oben und unten).

Genauso wie eine Wahrscheinlichkeitsdichte, vergleiche Bemerkung 1.57(a), ist die schwache Ableitung nur Lebesgue-fast überall eindeutig bestimmt. Teil (b) des Lemmas ist daher so zu verstehen, dass für jede schwache Ableitungsfunktion F' durch $f(x) = F'(x)\mathbf{1}(F'(x) \geqslant 0)$ eine Wahrscheinlichkeitsdichte von P definiert wird.

Beweis. Teil (a) folgt sofort aus dem Satz: $F(x) = P((-\infty, x]) = \int_{-\infty}^{x} f(y)\,dy$. Für Teil (b) schreibe $F(b) - F(a) = \int_a^b F'(x)\,dx$. Dann gilt mit montoner Konvergenz

$$\int_\mathbb{R} F'(x)\,dx = \lim_{R\to\infty} \int_{-R}^{R} F'(x)\,dx = \lim_{R\to\infty} (F(R) - F(-R)) = 1$$

wegen $F' \geqslant 0$ Lebesgue-fast überall, was wir jetzt sehen werden. Ist F' stetig, so folgt sofort $F' \geqslant 0$: sonst wäre $F' < 0$ auf einem Intervall $(a, b]$ und daher $F(b) < F(a)$. Allgemein zerlegt man $F' = F'_+ - F'_-$ in Positiv- und Negativteil ($F'_+ := \max(F', 0)$, $F'_- := \max(-F', 0)$), erhält die Maße $\mu_+(B) = \int_B F'_+(x)\,dx$, $\mu_-(B) = \int_B F'_-(x)\,dx$ mit $P = \mu_+ - \mu_-$ und folgert für die Borelmenge $B = \{x \in \mathbb{R} \mid F'(x) < 0\}$

$$\begin{aligned} 0 \leqslant P(B) &= \mu_+(B) - \mu_-(B) \\ &= \int_\mathbb{R} F'_+(x)\mathbf{1}(F'(x) < 0)\,dx - \int_\mathbb{R} F'_-(x)\mathbf{1}(F'(x) < 0)\,dx \\ &= 0 + \int_\mathbb{R} F'(x)\mathbf{1}(F'(x) < 0)\,dx. \end{aligned}$$

Das letzte Integral ist nicht-positiv, so dass es null sein und $F'(x) \geqslant 0$ Lebesgue-fast überall gelten muss.

Wir haben gezeigt, dass F' eine Wahrscheinlichkeitsdichte ist. Wegen $P((a, b]) = F(b) - F(a) = \int_a^b F'(x)\,dx$ ist $P = P_{F'}$. □

1.62 Definition. Die Exponentialverteilung mit Parameter $\lambda > 0$ ist gegeben durch die Wahrscheinlichkeitsdichte

$$f_{\mathrm{Exp}(\lambda)}(x) = \lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}^+}(x), \quad x \in \mathbb{R}\,.$$

Gilt $f^X = f_{\mathrm{Exp}(\lambda)}$ für eine reellwertige Zufallsvariable X, so schreiben wir $X \sim \mathrm{Exp}(\lambda)$.

1.63 Bemerkung. Die Exponentialverteilung ist *gedächtnislos* in dem Sinne, dass für $X \sim \mathrm{Exp}(\lambda)$ gilt

$$\forall x, t \geqslant 0 : \; P(X \geqslant x + t \,|\, X \geqslant t) = P(X \geqslant x),$$

wo wir bereits bedingte Wahrscheinlichkeiten benutzen, vergleiche Definition 2.2. Da die Familie der Exponentialverteilungen die einzigen Verteilungen mit dieser Eigenschaft sind, benutzt man sie kanonisch, um Wartezeiten zu modellieren. Ein Beispiel ist die Zeit zum nächsten Atomzerfall bei einer radioaktiven Probe („die Zeit bis zum Zerfall ist unabhängig davon, wie lange schon gewartet wurde“).

1.64 Definition. Die (eindimensionale) Normalverteilung mit Parametern $\mu \in \mathbb{R}$ und $\sigma > 0$ ist gegeben durch die Wahrscheinlichkeitsdichte

$$\varphi_{\mu,\sigma^2}(x) := f_{N(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big), \quad x \in \mathbb{R}\,.$$

Gilt $f^X = \varphi_{\mu,\sigma^2}$ für eine reellwertige Zufallsvariable X, so schreiben wir $X \sim N(\mu, \sigma^2)$ und sagen, dass X eine Gaußsche Zufallsvariable ist.

1.65 Lemma. *Die Funktion φ_{μ,σ^2} ist in der Tat eine Wahrscheinlichkeitsdichte.*

Beweis. Da φ_{μ,σ^2} offensichtlich positiv und stetig, also Borel-messbar ist, bleibt $\int \varphi_{\mu,\sigma^2} = 1$ nachzuweisen. Mit der Substitution $y = (x-\mu)/\sigma$ gilt

$$\int_{-\infty}^{\infty} \varphi_{\mu,\sigma^2}(x)\, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{y^2}{2}\Big)\, dy = \int_{-\infty}^{\infty} \varphi_{0,1}(y)\, dy.$$

Das letzte Integral berechnen wir mit einem Trick. Nach dem Satz von Fubini und mit Transformation auf Polarkoordinaten (r, φ) gilt

$$\begin{aligned}
\Big(\int_{-\infty}^{\infty} \exp\Big(- \frac{y^2}{2}\Big)\, dy\Big)^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\Big(- \frac{y^2+z^2}{2}\Big)\, dydz \\
&= \int_0^{2\pi}\int_0^{\infty} \exp\Big(- \frac{r^2}{2}\Big)\, r\, drd\varphi \\
&= -\int_0^{2\pi} \exp\Big(- \frac{r^2}{2}\Big)\Big|_{r=0}^{\infty} d\varphi \\
&= \int_0^{2\pi} 1\, d\varphi = 2\pi.
\end{aligned}$$

Teilen wir alles durch 2π, so erhalten wir $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = 1$. □

1.66 Bemerkung. Die Bedeutung der Normalverteilung rührt vom zentralen Grenzwertsatz her, siehe unten. Sie findet überall dort Verwendung, wo viele kleinere Fehlerquellen zusammenkommen, zum Beispiel bei Messfehlern physikalischer Geräte. Die Verteilungsfunktion von $N(\mu, \sigma^2)$ lässt sich nicht explizit angeben (e^{-x^2} hat keine 'einfache' Stammfunktion), umso wichtiger sind jedoch explizite Abschätzungen. Diese zeigen, dass die Überlebensfunktion (survival function) $1 - F_{N(\mu,\sigma^2)}(x)$ sogar etwas schneller für $x \to \infty$ abfällt als die Normalverteilungsdichte selbst.

1.67 Satz (Eindimensionaler Dichtetransformationssatz)**.** *Es sei X eine I-wertige Zufallsvariable für ein offenes (endliches oder unendliches) Intervall $I \subseteq \mathbb{R}$ mit Dichte f^X. Setze $Y := \varphi(X)$ mit einer stetig differenzierbaren Funktion $\varphi : I \to \mathbb{R}$, die $\varphi'(x) \neq 0$ für alle $x \in I$ erfüllt. Dann besitzt die Zufallsvariable Y die Dichte*

$$f^Y(y) = f^X(\varphi^{-1}(y))|(\varphi^{-1})'(y)|\mathbf{1}_{\varphi(I)}(y), \quad y \in \mathbb{R},$$

mit der Inversen φ^{-1} von φ und $(\varphi^{-1})'(y) = \frac{1}{\varphi'(\varphi^{-1}(y))}$.
Konvention: setze $A \bullet 0 := 0$ auch für Ausdrücke A, die nicht wohldefiniert sind.

Beweis. Sei zunächst $\varphi'(x) > 0$ für alle $x \in I$. Dann ist φ streng monoton wachsend und besitzt eine Inverse $\varphi^{-1} : \varphi(I) \to I$. Damit gilt für $y \in \varphi(I)$

$$F^Y(y) = P(Y \leqslant y) = P(\varphi(X) \leqslant y) = P(X \leqslant \varphi^{-1}(y)) = F^X(\varphi^{-1}(y))$$

und die Kettenregel (gilt auch bei schwacher Ableitung; argumentiere via Integration) zeigt

$$f^Y(y) = (F^Y)'(y) = (F^X)'(\varphi^{-1}(y))(\varphi^{-1})'(y) = f^X(\varphi^{-1}(y))(\varphi^{-1})'(y),\ y \in \varphi(I).$$

Für $y \geqslant \sup_{x\in I} \varphi(x)$ gilt $F^Y(y) = 1$ konstant, also $f^Y(y) = 0$. Für $y \leqslant \inf_{x\in I} \varphi(x)$ gilt $F^Y(y) = 0$ und ebenso $f^Y(y) = 0$. Da $\varphi(I)$ ein Intervall ist und nach Analysis I $(\varphi^{-1})'(y) = \frac{1}{\varphi'(\varphi^{-1}(y))} > 0$ für $y \in \varphi(I)$, erhalten wir obige Formel für f^Y.

Da φ' stetig ist und nicht verschwindet, ist der einzige andere Fall $\varphi' < 0$ auf I, wo wir analog argumentieren, aber sich Ungleichheitszeichen umkehren:

$$F^Y(y) = P(Y \leqslant y) = P(\varphi(X) \leqslant y) = P(X \geqslant \varphi^{-1}(y)) = 1 - F^X(\varphi^{-1}(y)).$$

Ableiten zeigt also $f^Y(y) = -f^X(\varphi^{-1}(y))(\varphi^{-1})'(y) = f^X(\varphi^{-1}(y))|(\varphi^{-1})'(y)|$. Für $y \notin \varphi(I)$ ist $F^Y(y)$ wiederum konstant, und die Formel folgt. □

1.68 Korollar. *Ist X eine reellwertige Zufallsvariable mit Dichte f^X, so besitzt $Y = \sigma X + \mu$ für $\sigma \in \mathbb{R}\setminus\{0\}$, $\mu \in \mathbb{R}$ die Dichte $f^Y(y) = |\sigma|^{-1} f^X(\frac{Y-\mu}{\sigma})$. Insbesondere ist $Y = \sigma X + \mu$ $N(\mu, \sigma^2)$-verteilt für $X \sim N(0, 1)$.*

Beweis. Setze $I = \mathbb{R}$ und $\varphi(x) = \sigma x + \mu$ im Satz und beachte $\varphi^{-1}(y) = \frac{y-\mu}{\sigma}$ sowie $(\varphi^{-1})'(y) = \sigma^{-1}$. Im Fall $X \sim N(0, 1)$ überprüft man sofort $f^Y = f_{N(\mu,\sigma^2)}$. □

1.69 Bemerkungen.

(a) Insbesondere ist also $Y = -X$ für $X \sim N(0,1)$ wiederum $N(0,1)$-verteilt. Das heißt, dass die Verteilungen von X und $-X$ gleich sind („X und $-X$ sind identisch verteilt“), aber natürlich gilt sogar $P(X = -X) = 0$ (die Zufallsvariablen sind verschieden).

(b) Falls $\varphi : I \to \mathbb{R}$ nur lokal ein Diffeomorphismus ist, also z.B. $\varphi \in C^1(I)$ gilt mit $\varphi'(x) = 0$ an endlich vielen Stellen $x_1, \dots, x_m \in I$, so betrachte die Intervalle $I_j = (x_{j-1}, x_j)$ (mit $x_0 := \inf I$, $x_{m+1} := \sup I$), die lokalen Inversen $\varphi_j^{-1} : \varphi(I_j) \to I_j$ von $\varphi_j = \varphi|_{I_j}$ auf I_j und erhalte ($\{x_1, \dots, x_m\}$ ist P^X-Nullmenge!)

$$F^Y(y) = \sum_{j=1}^{m+1} P(\varphi_j(X) \leqslant y, X \in I_j) = \sum_{j=1}^{m+1} \int_{\varphi_j^{-1}((-\infty,y])} f^X(x)\,dx.$$

Unterscheidet man die Fälle $\varphi_j^{-1}((-\infty, y]) = (-\infty, \varphi_j^{-1}(y)] \cap I_j$ und $\varphi_j^{-1}((-\infty, y]) = [\varphi_j^{-1}(y), \infty) \cap I_j$, so erhält man wieder durch Ableiten

$$f^Y(y) = (F^Y)'(y) = \sum_{j=1}^{m+1} f^X(\varphi_j^{-1}(y))|(\varphi_j^{-1})'(y)|\mathbf{1}_{\varphi(I_j)}(y).$$

1.70 Beispiel. Es sei X eine reellwertige Zufallsvariable mit Dichte f^X. Die Dichte von $Y = X^2$ ist nach Bemerkung 1.69(b) mit $\varphi(x) = x^2$ und $I_1 = (-\infty, 0)$, $I_2 = (0, \infty)$ gegeben durch

$$f^Y(y) = \big(f^X(-\sqrt{y}) + f^X(\sqrt{y})\big)(2\sqrt{y})^{-1}\mathbf{1}_{(0,\infty)}(y), \quad y \in \mathbb{R}\,.$$

Im Fall $X = N(0,1)$ heißt die Verteilung von $Y = X^2$ χ^2-Verteilung (mit einem Freiheitsgrad). Sie besitzt die Dichte

$$f_{\chi^2(1)}(y) := f^Y(y) = 2\tfrac{1}{\sqrt{2\pi}}e^{-y/2}(2\sqrt{y})^{-1}\mathbf{1}(y > 0) = \tfrac{1}{\sqrt{2\pi y}}e^{-y/2}\mathbf{1}(y > 0).$$

Beachte, dass die $\chi^2(1)$-Dichte bei Null unbeschränkt ist, aber integrierbar bleibt.

1.71 Satz. *Jede Wahrscheinlichkeitsdichte f auf $\mathbb{R}^d$ erzeugt mittels*

$$P_f((a_1, b_1] \times \cdots \times (a_d, b_d]) = \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} f(x_1, \dots, x_d)\,dx_d \cdots dx_1$$

für $a_k, b_k \in \mathbb{R}$ mit $a_k < b_k$ ein eindeutiges Wahrscheinlichkeitsmaß P_f auf $\mathfrak{B}_{\mathbb{R}^d}$, und es gilt $P_f(B) = \int_B f(x)dx$.

Beweis. Dies folgt analog zu Satz 1.56. Das Wahrscheinlichkeitsmaß $Q(B) = \int_B f(x)\,dx$, $B \in \mathfrak{B}_{\mathbb{R}^d}$, stimmt auf den Quadern $(a_1, b_1] \times \cdots \times (a_d, b_d]$ mit P_f überein, und diese Quader bilden einen $\cap$-stabilen Erzeuger von $\mathfrak{B}_{\mathbb{R}^d}$, was $P_f = Q$ impliziert. □

1.72 Bemerkung. Ganz allgemein und vollkommen analog kann man auf einem beliebigen Maßraum $(\Omega, \mathscr{F}, \mu)$ durch eine messbare Funktion $f : \Omega \to [0, \infty)$ mit $\int_\Omega f(\omega)\,\mu(d\omega) = 1$ ein Wahrscheinlichkeitsmaß P_f definieren über

$$P_f(B) = \int_B f(\omega)\,\mu(d\omega), \quad B \in \mathscr{F}.$$

f heißt dann auch μ-Dichte von P_f.

1.73 Definition. Sind $f_1, \ldots, f_d$ Wahrscheinlichkeitsdichten auf $\mathbb{R}$, so heißt

$$f(x_1, \ldots, x_d) = \prod_{k=1}^{d} f_k(x_k), \quad x_1, \ldots, x_d \in \mathbb{R},$$

Produktdichte der $(f_k)_{k=1,\ldots,d}$ im $\mathbb{R}^d$. Insbesondere ist die d-dimensionale Standard-Normalverteilung $N(0, E_d)$ im $\mathbb{R}^d$ definiert über die Produktdichte von d $N(0,1)$-Dichten:

$$f_{N(0,E_d)}(x) = (2\pi)^{-d/2} e^{-|x|^2/2},\ x \in \mathbb{R}^d, \quad \text{mit } |x|^2 = \sum_{i=1}^{d} x_i^2.$$

1.74 Bemerkung. Nach dem Satz von Fubini ist jede Produktdichte eine Wahrscheinlichkeitsdichte auf $\mathbb{R}^d$. $E_d \in \mathbb{R}^{d\times d}$ bezeichnet stets die $d \times d$-Einheitsmatrix.

Ende der 5. Vorlesung

1.75 Satz (Allgemeiner Dichtetransformationssatz)**.** *Es seien X ein d-dimensionaler Zufallsvektor mit Dichte f^X sowie $Y = \varphi(X)\mathbf{1}(X \in U)$ für einen C^1-Diffeomorphismus $\varphi : U \to V$ mit $U, V \subseteq \mathbb{R}^d$ offen und $P(X \in U) = 1$. Dann ist Y ein Zufallsvektor mit Dichte*

$$f^Y(y) = f^X(\varphi^{-1}(y))|\det(D(\varphi^{-1})(y))|\mathbf{1}(y \in V), \quad y \in \mathbb{R}^d\,.$$

1.76 Bemerkung. $Dh(y)$ bezeichnet die Ableitungs- oder Jacobimatrix einer C^1-Funktion $h : U \to \mathbb{R}^d$, $U \subseteq \mathbb{R}^d$ offen, und $\det(Dh(y))$ die Funktional- oder Jacobi-Determinante von h bei y.

Beweis. Wir greifen auf die Transformationsformel für mehrdimensionale Lebesgue-Integrale aus Analysis III zurück, nach der

$$\int_{\varphi^{-1}(O)} f^X(x)\,dx = \int_O f^X(\varphi^{-1}(y))|\det(D(\varphi^{-1})(y))|\,dy$$

für alle offenen Mengen $O \subseteq V$ gilt. Setzt man $O = V$ und beachtet $\int_{\varphi^{-1}(V)} f^X = P(X \in U) = 1$, so ist die im Satz angegebene Funktion f^Y also eine Wahrscheinlichkeitsdichte mit

$$\int_O f^Y(y)dy = \int_{\varphi^{-1}(O\cap V)} f^X(x)dx = P(X \in \varphi^{-1}(O \cap V)) = P(Y \in O)$$

für alle offenen Mengen $O \subseteq \mathbb{R}^d$. Da die offenen Mengen einen $\cap$-stabilen Erzeuger von $\mathfrak{B}_{\mathbb{R}^d}$ bilden, muss also $P_{f^Y} = P^Y$ gelten. Mit anderen Worten ist f^Y die Dichte von Y.

□

1.77 Korollar. *Ist X ein d-dimensionaler Zufallsvektor mit Dichte f^X, so besitzt $Y = AX + b$ für $A \in \mathbb{R}^{d\times d}$ invertierbar und $b \in \mathbb{R}^d$ die Dichte $f^Y(y) = f^X(A^{-1}(y-b))|\det(A)|^{-1}$, $y \in \mathbb{R}^d$.*

Beweis. Setze im Satz $\varphi(x) = Ax+b$ mit $U = V = \mathbb{R}^d$ und $\varphi^{-1}(y) = A^{-1}(y-b)$ mit $\det(D(\varphi^{-1}(y))) = \det(A^{-1}) = \det(A)^{-1}$. □

1.78 Beispiel. Sind X ein d-dimensionaler standard-normalverteilter Zufallsvektor sowie $\mu \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$ invertierbar, so ist $Y = \mu + AX$ ein d-dimensionaler Zufallsvektor mit Dichte

$$\varphi_{\mu,\Sigma}(x) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\Big(-\frac{1}{2}\langle \Sigma^{-1}(x-\mu), x-\mu\rangle\Big), \quad x \in \mathbb{R}^d,$$

wobei $\Sigma = AA^\top$ eine symmetrische positiv-definite Matrix ist. P^Y ist die Normalverteilung mit Mittelwertvektor μ und Kovarianzmatrix Σ, kurz $Y \sim N(\mu, \Sigma)$. Insbesondere ist $Y = OX$ für eine orthogonale Matrix O (d.h. $O^\top O = E_d$) wieder $N(0, E_d)$-verteilt: die Standardnormalverteilung ist invariant unter orthogonalen Transformationen wie Drehungen und Spiegelungen.

2 Bedingte Wahrscheinlichkeiten und Unabhängigkeit

2.1 Bedingte Wahrscheinlichkeiten und Bayes-Formel

2.1 Beispiel. Ein einfacher Test auf eine Krankheit liefert bei 1000 Versuchspersonen, davon 900 gesunden und 100 kranken, folgendes Resultat (dargestellt in Form einer *Kontingenztafel*):

Person	Test positiv	Test negativ	Summe
gesund	15	885	900
krank	92	8	100
Summe	107	893	1000

Was kann eine Ärztin einer Person sagen, deren Test positiv ausfällt? Mögliche Antwort: „Bei einem positiven Testergebnis liegt im Schnitt bei $\frac{92}{107}100\% \approx 86\%$ der Fälle wirklich eine Krankheit vor, in immerhin ca. 14% der Fälle ist das Testergebnis falsch.“ Andererseits könnte sie einem Patienten mit negativem Testergebnis sagen: „Nur in $\frac{8}{893}100\% \approx 0,9\%$ der Fälle liegt bei negativem Test eine Krankheit vor, es ist also sehr unwahrscheinlich, dass eine Krankheit nicht erkannt wurde.“

Für diese Aussagen beschränken wir uns auf eine Teilmenge aller Ergebnisse, z.B. nur auf die positiven Testergebnisse, und betrachten die relativen Häufigkeiten der gesunden bzw. kranken Patienten nur in dieser Teilmenge. Dies lässt sich analog auf allgemeine Wahrscheinlichkeitsmaße übertragen.

2.2 Definition. Es seien A und B Ereignisse mit $P(B) > 0$. Dann wird mit

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

die bedingte Wahrscheinlichkeit von A gegeben (oder: unter) B bezeichnet.

2.3 Beispiel. Beim Wurf von zwei fairen Würfeln mit den Ereignissen $A =$„Pasch“ und $B =$„beide Augenzahlen ungerade“ gilt $P(A \mid B) = \frac{3/36}{9/36} = \frac{1}{3}$. Intuitives Argument: wenn wir bereits wissen, dass beide Augenzahlen ungerade sind, gibt es für den zweiten Würfel nur drei Möglichkeiten, nämlich '1', '3', '5', und eine davon führt zum Pasch. Beachte, dass hier $P(A \mid B) > P(A)$ gilt und das Eintreten des Ereignisses B die Wahrscheinlichkeit des Ereignisses A beeinflusst (später: A und B sind nicht unabhängig).

Ende 6. Vorlesung

2.4 Satz. *Auf dem Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, P)$ sei B ein Ereignis mit $P(B) > 0$. Dann gilt:*

(a) Durch $Q(A) := P(A \mid B)$ wird ein Wahrscheinlichkeitsmaß Q auf $\mathscr{F}$ definiert.

(b) (Formel von der totalen Wahrscheinlichkeit) Es sei $B = \bigcup_{i=1}^{N} B_i$ Vereinigung paarweise disjunkter Ereignisse B_i mit $P(B_i) > 0$. Dann folgt für jedes Ereignis A

$$P(A \cap B) = \sum_{i=1}^{N} P(B_i) P(A \mid B_i).$$

(c) (Bayesformel) Für jedes Ereignis A mit $P(A) > 0$ und jede Zerlegung $\Omega = \bigcup_{i=1}^{N} B_i$ von Ω in paarweise disjunkte Ereignisse B_i mit $P(B_i) > 0$ gilt

$$P(B_i \mid A) = \frac{P(B_i) P(A \mid B_i)}{\sum_{j=1}^{N} P(B_j) P(A \mid B_j)}.$$

In (b) und (c) kann auch $N = \infty$ gesetzt werden.

Beweis. Für (a) beachte $Q(\Omega) = P(\Omega \cap B)/P(B) = 1$ und für paarweise disjunkte A_n

$$Q\Big(\bigcup_{n \geqslant 1} A_n\Big) = P(B)^{-1} P\Big(\bigcup_{n \geqslant 1} A_n \cap B\Big) = P(B)^{-1} \sum_{n \geqslant 1} P(A_n \cap B) = \sum_{n \geqslant 1} Q(A_n).$$

Einsetzen ergibt (b):

$$\sum_{i=1}^{N} P(B_i) P(A \mid B_i) = \sum_{i=1}^{N} P(A \cap B_i) = P\Big(\bigcup_{i=1}^{n} A \cap B_i\Big) = P(A \cap B).$$

Die Bayesformel folgt aus der Definition $P(B_i \mid A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(B_i) P(A \mid B_i)}{P(A \cap \Omega)}$ und Teil (b) mit $B = \Omega$. Nirgendwo wurde N endlich in der Herleitung benutzt. □

2.5 Bemerkungen.

(a) Bei Anwendern werden oft $P(A \mid B)$ und $P(A \cap B)$ vermischt. Mathematisch sind beide Wahrscheinlichkeiten nur gleich, wenn $P(B) = 1$ gilt. Während $P(A \mid B)$ die Wahrscheinlichkeit von A (oder äquivalent $A \cap B$!) angibt, wenn ich weiß, dass B eintritt, so gibt $P(A \cap B)$ die Wahrscheinlichkeit dafür an, dass A und B gemeinsam eintreten, die Versuchsausgänge aber nicht eingeschränkt werden. Man mache sich den Unterschied bei den relativen Häufigkeiten in Beispiel 2.1 für A='Person krank' und B='Test positiv' klar, wo $A \cap B$ die relative Häufigkeit $\frac{92}{1000}$ besitzt!

(b) Die Bayesformel führt manchmal zu philosophischen Betrachtungen zur Umkehr von Kausalitäten, weil die Reihenfolge der Ereignisse in den bedingten Wahrscheinlichkeiten ausgetauscht wird. Wie das Beispiel eines Tests auf eine Krankheit zeigt, geben bedingte Wahrscheinlichkeiten keine Kausalitäten an, sondern können entweder *frequentistisch* mit relativen Häufigkeiten wie in Beispiel 2.1 oder *subjektiv* bzw. *Bayesianisch* als Änderung ('Updating') gegebener Wahrscheinlichkeiten durch Eintreten oder Beobachtung gewisser Ereignisse (vgl. Beispiel **??**) gedeutet werden.

2.6 Beispiel (Scheinkorrelationen)**.** An einer Universität werden von 825/560/325 männlichen Bewerbern für Fach 1/2/3 jeweils 62%/63%/34% zugelassen, von 108/25/593 weiblichen Bewerberinnen hingegen 82%/68%/37%. Obwohl die Zulassungsquote in jedem Fach für Frauen höher war, ergibt sich nach der Formel von der totalen Wahrscheinlichkeit (mit A ='zugelassen', B_i ='Bewerbung für Fach i und weiblich', B ='Bewerber*in weiblich') insgesamt eine Zulassungsquote von ca. 57% für Männer und von ca. 45% für Frauen, weil Letztere sich stärker für Fach 3 mit schwierigerer Zulassung beworben haben.

2.7 Lemma (Multiplikationsformel/Pfadregel)**.** *Für Ereignisse* $A_1, \ldots, A_n$, $n \geqslant 2$, *mit* $P(A_1 \cap \cdots \cap A_{n-1}) > 0$ *gilt*

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

Beweis. Wir verwenden vollständige Induktion. Für $n = 2$ gilt $P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1)$ nach Definition. Nehmen wir an, dass die Aussage für $n-1$ mit $n \geqslant 3$ gilt, so schließen wir (beachte dazu $P(A_1 \cap \cdots \cap A_{n-1}) > 0 \Rightarrow P(A_1 \cap \cdots \cap A_{n-2}) > 0$)

$$\begin{aligned} &P(A_1 \cap \cdots \cap A_n) = P(A_1 \cap \cdots \cap A_{n-1})P(A_n \mid A_1 \cap \cdots \cap A_{n-1}) \\ &= P(A_1)P(A_2 \mid A_1) \cdots P(A_{n-1} \mid A_1 \cap \cdots \cap A_{n-2})P(A_n \mid A_1 \cap \cdots \cap A_{n-1}), \end{aligned}$$

wie für $n \geqslant 3$ zu zeigen war. □

2.8 Beispiel (Polya-Urnen-Modell)**.** Beim Ziehen aus einer Urne mit W weißen und S schwarzen Kugeln, $W \geqslant 1$, $S \geqslant 2$, ohne Zurücklegen und mit Beachtung der Reihenfolge ergibt sich mit $N = S + W$ für das Ergebnis 'SSW' (d.h. 1. und 2. Kugel schwarz, 3. Kugel weiß) nach der Pfadregel die Wahrscheinlichkeit

$\frac{S}{N} \bullet \frac{S-1}{N-1} \bullet \frac{W}{N-2}$: die Wahrscheinlichkeit für 1. Kugel schwarz ist $\frac{S}{N}$, für 2. Kugel schwarz bei $S-1$ schwarzen unter $N-1$ Kugeln (also gegeben die 1. Kugel war schwarz) ist $\frac{S-1}{N-1}$ und für 3. Kugel weiß bei W weißen unter $N-2$ Kugeln (also gegeben die 1. und 2. Kugel waren schwarz) ist $\frac{W}{N-2}$. Dieselbe Wahrscheinlichkeit besitzen die Versuchsausgänge 'SWS' und 'WSS' (man nennt die Verteilung *austauschbar*).

2.2 Unabhängige Ereignisse und Lemma von Borel-Cantelli

2.9 Bemerkung. In Beispiel 2.3 hatten wir gesehen, dass die Wahrscheinlichkeit eines Paschs steigt, wenn man weiß, dass beide Augenzahlen ungerade sind. Man sagt, dass die beiden Ereignisse (stochastisch) abhängig sind. Falls hingegen $P(A \,|\, B) = P(A)$ gilt, heißt das Ereignis A (stochastisch) unabhängig von Ereignis B. Allerdings setzt dies $P(B) > 0$ voraus. Mathematisch definiert man daher Unabhängigkeit etwas allgemeiner (multipliziere mit $P(B)$).

2.10 Definition.

(a) Zwei Ereignisse A und B heißen (stochastisch) unabhängig (unter P), falls $P(A \cap B) = P(A)P(B)$ gilt.

(b) Eine Familie $(A_i)_{i \in I}$ von Ereignissen, $I \neq \varnothing$ beliebige Indexmenge, heißt (stochastisch) unabhängig, falls für jede endliche Teilmenge $J \subseteq I$ gilt

$$P\Big(\bigcap_{j \in J} A_j\Big) = \prod_{j \in J} P(A_j).$$

2.11 Bemerkung. Die Definition der Unabhängigkeit von Familien von Ereignissen ist analog zur linearen Unabhängigkeit einer Familie von Vektoren.

2.12 Beispiel. Beim Würfeln mit zwei fairen Würfeln haben die Ereignisse „Augensumme ist 7“ und “erste Augenzahl ist 6“ jeweils Wahrscheinlichkeit 1/6. Der Schnitt der beiden Ereignisse ist „erste Augenzahl ist 6, zweite Augenzahl ist 1“ und hat Wahrscheinlichkeit 1/36, so dass die beiden Ereignisse (unter Gleichverteilung) unabhängig sind.

2.13 Definition. Für eine Folge $(A_n)_{n \geqslant 1}$ von Ereignissen setze

$$\limsup_{n \to \infty} A_n := \bigcap_{m \geqslant 1} \bigcup_{n \geqslant m} A_n = \{\omega \in \Omega \,|\, \omega \in A_n \text{ für unendlich viele } n\}.$$

2.14 Bemerkungen.

(a) Beachte die Formel $\mathbf{1}_{\limsup_{n\to\infty} A_n}(\omega) = \limsup_{n\to\infty} \mathbf{1}_{A_n}(\omega)$.

(b) Wir lernen jetzt eines der wichtigsten Resultate der Wahrscheinlichkeitstheorie kennen, das $P(\limsup_{n\to\infty} A_n)$ beschreibt. Eine wichtige Anwendung sind zufällige Folgen $(X_n)_{n\geqslant 1}$, das heißt, jedes X_n ist Zufallsvariable. Dann ist das Ereignis $\{\omega \in \Omega \,|\, (X_n(\omega))_{n\geqslant 1} \text{ ist keine Nullfolge}\}$ gleich $\bigcup_{\varepsilon>0} \limsup_{n\to\infty} \{|X_n| > \varepsilon\}$. Dabei darf man eine abzählbare Vereinigung

nur über rationale Zahlen $\varepsilon > 0$ nehmen. Gilt also $P(\limsup_{n\to\infty}\{|X_n| > \varepsilon\}) = 0$ für alle $\varepsilon > 0$, so ist (X_n) P-fast sicher (das heißt mit Wahrscheinlichkeit 1) eine Nullfolge. Dies wird beim Beweis des starken Gesetzes der großen Zahlen essentiell werden.

2.15 Satz (Lemma von Borel-Cantelli)**.** *Für eine Folge $(A_n)_{n\geqslant 1}$ von Ereignissen gilt:*

(a) Aus $\sum_{n\geqslant 1} P(A_n) < \infty$ folgt $P(\limsup_{n\to\infty} A_n) = 0$.

(b) Gilt $\sum_{n\geqslant 1} P(A_n) = \infty$ und ist $(A_n)_{n\geqslant 1}$ überdies unabhängig, so folgt $P(\limsup_{n\to\infty} A_n) = 1$.

Beweis. Es gilt $\limsup_{n\to\infty} A_n \subseteq \bigcup_{n\geqslant m} A_n$ für alle $m \geqslant 1$ und somit folgt in (a) :

$$P\Big(\limsup_{n\to\infty} A_n\Big) \leqslant \inf_{m\geqslant 1} \sum_{n\geqslant m} P(A_n) = 0;$$

denn $\sum_{n=1}^{\infty} P(A_n) < \infty$ bedeutet gerade $\lim_{m\to\infty} \sum_{n=m}^{\infty} P(A_n) = 0$.

Für (b) betrachte $(\limsup_{n\to\infty} A_n)^c = \bigcup_{m\geqslant 1} \bigcap_{n\geqslant m} A_n^c$. Dann folgt mit σ-Stetigkeit, Unabhängigkeit der $(A_n^c)_{n\geqslant 1}$ (vergleiche Beispiel 2.21 unten) und der Abschätzung $P(A_n^c) = 1 - P(A_n) \leqslant e^{-P(A_n)}$

$$\begin{aligned} P\Big(\big(\limsup_{n\to\infty} A_n\big)^c\Big) &\leqslant \sum_{m\geqslant 1} P\Big(\bigcap_{n\geqslant m} A_n^c\Big) = \sum_{m\geqslant 1} \lim_{M\to\infty} P\Big(\bigcap_{n=m}^{M} A_n^c\Big) \\ &= \sum_{m\geqslant 1} \lim_{M\to\infty} \prod_{n=m}^{M} P(A_n^c) \leqslant \sum_{m\geqslant 1} \lim_{M\to\infty} \exp\Big(-\sum_{n=m}^{M} P(A_n)\Big). \end{aligned}$$

Da für die Exponentialfunktion $\lim_{x\to\infty} e^{-x} = 0$ gilt, impliziert $\sum_{n=m}^{\infty} P(A_n) = \infty$, dass der Grenzwert Null ist. Damit ist auch die Summe Null und das Gegenereignis $\limsup A_n$ hat Wahrscheinlichkeit Eins. □

2.16 Beispiele.

(a) Ist A ein Ereignis mit $P(A) \in (0,1)$, so gilt $\sum_{n\geqslant 1} P(A_n) = \infty$ für $A_n := A$, jedoch $P(\limsup_{n\to\infty} A_n) = P(A) < 1$. Auf die Unabhängigkeit in Teil (b) des Lemmas von Borel-Cantelli kann also nicht verzichtet werden.

(b) Betrachte das Ereignis A_3, dass für alle $M \geqslant 1$ unendlich viele M-runs vorkommen im unendlich langen Würfelexperiment, vgl. mit A_2 in Beispiel 1.1. Dann gilt mit dem Ereignis $B_{k,M} = \{b \in \Omega \,|\, b_k = \cdots = b_{k+M-1} = 1\}$ eines M-runs ab dem k-ten Wurf

$$A_3 = \bigcap_{M\geqslant 1} \limsup_{k\to\infty} B_{k,M} \supseteq \bigcap_{M\geqslant 1} \limsup_{j\to\infty} B_{jM,M}.$$

Die Familie $(B_{jM,M})_{j\geqslant 1}$ ist unabhängig (formale Herleitung?) mit $P(B_{jM,M}) = 2^{-M}$. Teil (b) des Lemmas von Borel-Cantelli zeigt daher $P(\limsup_{j\to\infty} B_{jM,M}) = 1$. Damit folgt $P(A_3) = 1$ („Abzählbarer Schnitt von Einsmengen ist wiederum Einsmenge“).

2.3 Unabhängige Zufallsvariablen und σ-Algebren

2.17 Definition. Es seien $\mathscr{M}_i \subseteq \mathscr{F}$, $i \in I$, Mengen von Ereignissen. Dann heißt $(\mathscr{M}_i)_{i\in I}$ unabhängig, falls für jede beliebige Auswahl von Ereignissen $A_i \in \mathscr{M}_i$ die Familie $(A_i)_{i\in I}$ unabhängig ist.

2.18 Definition. Eine Familie $(X_i)_{i\in I}$ von $(S_i, \mathscr{S}_i)$-wertigen Zufallsvariablen heißt unabhängig, falls für jede beliebige Wahl von $A_i \in \mathscr{S}_i$ die Familie von Ereignissen $(\{X_i \in A_i\})_{i\in I}$ unabhängig ist. Äquivalent ist die Familie $(X_i)_{i\in I}$ unabhängig, falls die von X_i erzeugten σ-Algebren $\mathscr{F}^{X_i} = \{X_i^{-1}(A) \,|\, A \in \mathscr{S}_i\}$, $i \in I$, unabhängig sind.

2.19 Lemma. *Sind $(X_i)_{i\in I}$ eine Familie unabhängiger $(S_i, \mathscr{S}_i)$-wertiger Zufallsvariablen und $g_i : S_i \to T_i$ $(\mathscr{S}_i, \mathscr{T}_i)$-messbare Funktionen, so besteht auch die Familie $(g_i(X_i))_{i\in I}$ aus unabhängigen Zufallsvariablen.*

Beweis. Da die Komposition messbarer Funktionen wieder messbar ist, sind die $g_i(X_i)$ wieder Zufallsvariablen. Nun gilt

$$\mathscr{F}^{g_i(X_i)} = \{X_i^{-1}(g_i^{-1}(A)) \,|\, A \in \mathscr{T}_i\} \subseteq \{X_i^{-1}(B) \,|\, B \in \mathscr{S}_i\} = \mathscr{F}^{X_i}.$$

Aus der Definition folgt also direkt, dass $(\mathscr{F}^{X_i})_{i\in I}$ unabhängig die Unabhängigkeit von $(\mathscr{F}^{g_i(X_i)})_{i\in I}$ impliziert. □

Ende 7. Vorlesung

2.20 Satz. *Es seien $(X_i)_{i\in I}$ eine Familie von Zufallsvariablen auf $(\Omega, \mathscr{F}, P)$ mit Werten in $(S_i, \mathscr{S}_i)$ und $\mathscr{E}_i$ $\cap$-stabile Erzeuger von $\mathscr{S}_i$, $i \in I$. Dann ist $(X_i)_{i\in I}$ bereits unabhängig, falls $(\{X_i \in A_i\})_{i\in I}$ unabhängig ist für beliebige $A_i \in \mathscr{E}_i$.*

Beweis. Es reicht, dies für endliche Indexmengen $J \subseteq I$ mit $|J| \geqslant 2$ zu beweisen. Mit vollständiger Induktion über n zeigen wir die Behauptung, dass $(\{X_i \in A_i\})_{i\in J}$ unabhängig ist für alle $A_i \in \mathscr{S}_i$, die $|\{i \in J \,|\, A_i \notin \mathscr{E}_i\}| \leqslant n$ erfüllen. Für $n = 0$ ist die Induktionsbehauptung gerade die Annahme im Satz. Für den Induktionsschluss von n auf $n+1$ betrachte $A_i \in \mathscr{S}_i$ mit $|\{i \in J \,|\, A_i \notin \mathscr{E}_i\}| = n+1$ (wenn solche A_i nicht existieren, so gilt es bereits für alle $A_i \in \mathscr{S}_i$) und setze $J' = J \setminus \{j\}$ für ein $j \in J$ mit $A_j \notin \mathscr{E}_j$. Nach Induktionsannahme gilt

$$P\Big(\bigcap_{i\in J'} \{X_i \in A_i\}\Big) = \prod_{i\in J'} P(X_i \in A_i).$$

Falls dies Null ist, so gilt direkt $P(\bigcap_{i\in J}\{X_i \in A_i\}) = 0 = \prod_{i\in J} P(X_i \in A_i)$. Gelte also $P(\bigcap_{i\in J'}\{X_i \in A_i\}) > 0$. Dann sind

$$Q_1(A) := P\Big(\{X_j \in A\} \,\Big|\, \bigcap_{i\in J'} \{X_i \in A_i\}\Big),\ Q_2(A) := P(X_j \in A) \text{ für } A \in \mathscr{S}_j$$

zwei Wahrscheinlichkeitsmaße. Für $E_j \in \mathscr{E}_j$ gilt nach Induktionsannahme

$$Q_1(E_j) = \frac{P(\{X_j \in E_j\} \bigcap_{i\in J'}\{X_i \in A_i\})}{P(\bigcap_{i\in J'}\{X_i \in A_i\})} = P(X_j \in E_j) = Q_2(E_j).$$

Nach dem Eindeutigkeitssatz folgt $Q_1 = Q_2$ und somit die Induktionsbehauptung für $n+1$.

Für $n \geqslant |J|$ ergibt sich die Unabhängigkeit von $(\{X_i \in A_i\})_{i\in J}$ für alle $A_i \in \mathscr{S}_i$. □

2.21 Beispiel. Ist $(A_i)_{i\in I}$ eine Familie unabhängiger Ereignisse, so sind $X_i = \mathbf{1}_{A_i}$, $i \in I$, unabhängige $\{0,1\}$-wertige Zufallsvariablen. Zum Nachweis reicht es, jeweils den $\cap$-stabilen Erzeuger $\mathscr{E} = \{\{1\}\}$ der Potenzmenge $\mathscr{P}(\{0,1\})$ zu betrachten. Die Ereignisse $\{X_i = 1\} = A_i$, $i \in I$, sind unabhängig. Folglich sind die erzeugten σ-Algebren $\mathscr{F}^{X_i} = \{\varnothing, \Omega, A_i, A_i^c\}$, $i \in I$, unabhängig. Insbesondere sind mit $(A_i)_{i\in I}$ auch $(A_i^c)_{i\in I}$ unabhängig.

2.22 Korollar. *Es seien $X_1, \dots, X_n$ Zufallsvariablen auf $(\Omega, \mathscr{F}, P)$.*

(a) Sind X_k diskret-verteilte S_k-wertige Zufallsvariablen, so sind $X_1, \dots, X_n$ genau dann unabhängig, wenn gilt

$$p^{(X_1,\dots,X_n)}(s_1, \dots, s_n) = \prod_{k=1}^{n} p^{X_k}(s_k) \text{ für alle } s_k \in S_k.$$

(b) Reellwertige Zufallsvariablen $X_1, \dots, X_n$ sind genau dann unabhängig, wenn gilt

$$P(X_1 \leqslant b_1, \dots, X_n \leqslant b_n) = \prod_{k=1}^{n} P(X_k \leqslant b_k) \text{ für alle } b_k \in \mathbb{R}\,.$$

Beweis. In beiden Fällen (a), (b) folgt die Gleichung direkt aus der Unabhängigkeit. Es muss also nur die Unabhängigkeit jeweils aus der Gleichung gefolgert werden.

Für (a) betrachte die Menge der maximal einelementigen Mengen $\mathscr{E}_k = \{\{s_k\} \mid s_k \in S_k\} \cup \{\varnothing\}$. Da S_k abzählbar ist, gilt $\sigma(\mathscr{E}_k) = \mathscr{P}(S_k)$. Außerdem ist $\mathscr{E}_k$ $\cap$-stabil. Nach Annahme gilt $P(X_1 = s_1, \dots, X_n = s_n) = \prod_{k=1}^n P(X_k = s_k)$. Wegen $\{X_k \in \varnothing\} = \varnothing$ und $P(\varnothing) = 0$ gilt also $P(X_1 \in E_1, \dots, X_n \in E_n) = \prod_{k=1}^n P(X_k \in E_k)$ für alle $E_k \in \mathscr{E}_k$. Nach Satz 2.20 sind $X_1, \dots, X_n$ unabhängig.

Für (b) folgt auf dem $\cap$-stabilen Erzeuger $\mathscr{E} = \{(-\infty, b] \mid b \in \mathbb{R}\}$ von $\mathscr{B}_\mathbb{R}$ aus der Annahme direkt $P(X_1 \in E_1, \dots, X_n \in E_n) = \prod_{k=1}^n P(X_k \in E_k)$ für alle $E_1, \dots, E_n \in \mathscr{E}$. Satz 2.20 zeigt daher die Unabhängigkeit von $X_1, \dots, X_n$. □

2.23 Beispiel. Beim Würfelwurf mit zwei Würfeln sind $X_i : \Omega \to \{1, \dots, 6\}$ mit $X_i(\omega_1, \omega_2) = \omega_i$ für $i \in \{1, 2\}$ unabhängige Zufallsvariablen (unter Gleichverteilung). Dazu reicht es, für $k_1, k_2 \in \{1, \dots, 6\}$ für die entsprechenden Zähldichten $p^{X_1}(k_1) = p^{X_2}(k_2) = 1/6$ sowie $p^{(X_1,X_2)}(k_1, k_2) = 1/36$ nachzuprüfen und Teil (a) des Korollars anzuwenden.

2.24 Satz. *Es sei $X = (X_1, \dots, X_n)$ ein n-dimensionaler Zufallsvektor auf $(\Omega, \mathscr{F}, P)$ mit Dichte $f^X : \mathbb{R}^n \to [0, \infty)$. Dann gilt:*

(a) Jedes X_k besitzt eine Dichte, die sogenannte <u>Randdichte</u>

$$f^{X_k}(x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f^X(x_1, \dots, x_n) dx_1 \dots dx_{k-1} dx_{k+1} \dots dx_n,\ x_k \in \mathbb{R}\,.$$

(b) Die Zufallsvariablen $X_1, \dots, X_n$ sind genau dann unabhängig, wenn gilt

$$f^X(x_1, \dots, x_n) = \prod_{k=1}^{n} f^{X_k}(x_k) \text{ für Lebesgue-fast alle } x_1, \dots, x_n \in \mathbb{R}\,.$$

Die Unabhängigkeit ist also äquivalent damit, dass f^X Produktdichte der Randdichten ist.

Beweis. Für die Verteilungsfunktion von X_k gilt mit dem Satz von Fubini

$$F^{X_k}(x_k) = P(X_k \leqslant x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{1}(y_k \leqslant x_k) f^X(y_1, \dots, y_n) dy_1 \dots dy_n.$$

Also ist F^{X_k} schwach differenzierbar mit Ableitung $(F^{X_k})' = f^{X_k}$ und die Aussage (a) folgt aus Lemma 1.60.

Aus $f^X(x_1, \dots, x_n) = \prod_{k=1}^n f^{X_k}(x_k)$ in (b) folgt durch Integration (Satz von Fubini!) für alle $x_1, \dots, x_n \in \mathbb{R}$

$$P(X_1 \leqslant x_1, \dots, X_n \leqslant x_n) = \prod_{k=1}^{n} \int_{-\infty}^{x_k} f^{X_k}(y_k)\, dy_k = \prod_{k=1}^{n} P(X_k \leqslant x_k).$$

Mit Korollar 2.22(b) folgt die Unabhängigkeit von $X_1, \dots, X_n$. Umgekehrt folgt aus der Unabhängigkeit von $X_1 \dots, X_n$, dass für alle $a_k \leqslant b_k$

$$\int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f^X(x_1, \dots, x_n) dx_n \cdots dx_1 = \prod_{k=1}^{n} \int_{a_k}^{b_k} f^{X_k}(x_k) dx_k.$$

Damit stimmen die durch die entsprechenden Dichten bestimmten Wahrscheinlichkeitsmaße P_{f^X} und P_f mit f gleich der Produktdichte $f(x) := \prod_{k=1}^n f^{X_k}(x_k)$ auf dem $\cap$-stabilen Erzeuger $\{[a_1, b_1] \times \cdots \times [a_n, b_n] \,|\, a_k < b_k\}$ von $\mathfrak{B}_{\mathbb{R}^d}$ überein. Nach dem Eindeutigkeitssatz gilt $P_{f^X}(B) = P_f(B)$ für alle Borelmengen B. Daher muss $B_> = \{x \in \mathbb{R}^n \mid f^X(x) > f(x)\}$ eine Lebesgue-Nullmenge sein ($\int_{B_>}(f^X - f) = 0 \Rightarrow \lambda(B_>) = 0$) ebenso wie das Analogon $B_<$, so dass $f^X = f$ Lebesgue-fast überall. □

2.25 Beispiele.

(a) Besitzen die Zufallsvariablen $X_1, \dots, X_n$ Dichten, so hat der Zufallsvektor $(X_1, \dots, X_n)$ nicht immer eine Dichte. Einfachstes Beispiel ist der Fall $X_1 = X_2$, die Zufallsvariablen (nicht nur ihre Verteilungen) sind gleich. Dann gilt $P^{(X_1, X_2)}(D) = 1$ für die Diagonale $D := \{(x, x) \,|\, x \in \mathbb{R}\}$, aber D besitzt zweidimensionales Lebesguemaß Null. Damit kann $P^{(X_1, X_2)}$ keine Dichte (bezüglich dem Lebesguemaß) besitzen.

(b) Ist $X = (X_1, \dots, X_n) \sim N(0, E_n)$, also $f^X(x) = (2\pi)^{-n/2} e^{-|x|^2/2}$, so gilt $f^{X_k}(x_k) = (2\pi)^{-1/2} e^{-x_k^2/2}$ gemäß (a). Damit ist jede Koordinate X_k $N(0,1)$-verteilt und X ein Vektor von n unabhängigen standardnormalverteilten Zufallsvariablen. Analog ist für $X \sim N(\mu, \Sigma)$ mit $\mu \in \mathbb{R}^n$ und der Diagonalmatrix $\Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2)$ jede Koordinate X_k $N(\mu_k, \sigma_k^2)$-verteilt, und alle Koordinaten $X_1, \dots, X_n$ sind unabhängig. Ist Σ keine

Diagonalmatrix, so sind für $X \sim N(\mu, \Sigma)$ die Koordinaten nicht mehr unabhängig (rechne nach, dass gemeinsame Dichte nicht Produkt der Randdichten ist).

(c) Sind $f_1, \dots, f_n$ Dichten auf $\mathbb{R}$, so definiert die Produktdichte $f(x) = \prod_{i=1}^n f_i(x_i)$ ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n})$. Die Koordinatenprojektionen $X_k : \mathbb{R}^n \to \mathbb{R}$ mit $X_k(x_1, \dots, x_n) = x_k$, $k = 1, \dots, n$, sind Borel-messbar, also Zufallsvariablen. Gemäß Satz 2.24(a) ist ihre Verteilung P^{X_k} gegeben durch die Dichte (Satz von Fubini!)

$$f^{X_k}(x_k) = f_k(x_k) \prod_{j \neq k} \int_{-\infty}^{\infty} f_j(x_j) dx_j = f_k(x_k),$$

ihre gemeinsame Verteilung $P^{(X_1,\dots,X_n)}$ durch die Dichte f. Wir haben damit einen Wahrscheinlichkeitsraum konstruiert mit unabhängigen Zufallsvariablen $(X_k)_{1 \leqslant k \leqslant n}$, deren Verteilung P^{X_k} jeweils durch f_k bestimmt ist.

Ende 8. Vorlesung

2.26 Bemerkung. Gemäß Beispiel 2.25(a) können wir für n gegebene Dichten auf $\mathbb{R}$ einen Wahrscheinlichkeitsraum und darauf unabhängige Zufallsvariablen $X_1, \dots, X_n$ konstruieren, die gemäß den Dichten verteilt sind. Derselbe Ansatz zeigt auch die Existenz von endlich vielen unabhängigen Zufallsvariablen, die gemäß gegebener Zähldichten diskret verteilt sind.

Allgemeiner kann ich zu zwei Wahrscheinlichkeitsräumen $(\Omega_i, \mathscr{F}_i, P_i)$, $i = 1, 2$, den Produktraum $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \otimes \mathscr{F}_2, P_1 \otimes P_2)$ definieren, wobei die Produkt-σ-Algebra $\mathscr{F}_1 \otimes \mathscr{F}_2$ als die kleinste σ-Algebra definiert ist, bezüglich der die Koordinatenprojektionen $\pi_i(\omega_1, \omega_2) = \omega_i$ für $i = 1, 2$ messbar sind und das Produktmaß $P_1 \otimes P_2$ durch $(P_1 \otimes P_2)(A_1 \times A_2) = P_1(A_1)P_2(A_2)$, $A_i \in \mathscr{F}_i$, auf 'Rechteckmengen' und damit auf $\mathscr{F}_1 \otimes \mathscr{F}_2$ eindeutig festgelegt ist, vgl. Maßtheorie/Analysis III.

2.27 Definition. Es seien $(\Omega_i, \mathscr{F}_i, P_i)_{i \in I}$, $I \neq \varnothing$ beliebige Indexmenge, Wahrscheinlichkeitsräume. Setze $\Omega := \prod_{i \in I} \Omega_i$ (kartesisches Produkt) und definiere mittels der Koordinatenprojektionen $\pi_i : \Omega \to \Omega_i$, $\pi_i((\omega_j)_{j \in I}) = \omega_i$, über Ω die Produkt-σ-Algebra

$$\bigotimes_{i \in I} \mathscr{F}_i := \sigma\Big(\bigcup_{i \in I} \{\pi_i^{-1}(A_i) \mid A_i \in \mathscr{F}_i\} \Big).$$

Die Produkt-σ-Algebra $\bigotimes_{i \in I} \mathscr{F}_i$ ist also die kleinste σ-Algebra, so dass alle π_i messbar sind. Gilt für ein Wahrscheinlichkeitsmaß Q auf $\bigotimes_{i \in I} \mathscr{F}_i$

$$\forall\, J \subseteq I \text{ endlich}, A_j \in \mathscr{F}_j : Q\Big(\bigcap_{j \in J} \pi_j^{-1}(A_j) \Big) = \prod_{j \in J} P_j(A_j),$$

so heißt Q Produktmaß der P_i, Schreibweise $Q = \bigotimes_{i \in I} P_i$.

2.28 Bemerkung. Man kann sich überlegen, dass ein Produktmaß über eine unendliche Indexmenge I nur für Wahrscheinlichkeitsmaße definiert werden kann. Es gibt z.B. kein „unendlich-dimensionales Lebesguemaß“ auf $\mathbb{R}^{\mathbb{N}}$.

2.29 Lemma. *Ist $(X_i)_{i\in I}$ eine Familie unabhängiger $(\Omega_i, \mathscr{F}_i)$-wertiger Zufallsvariablen, definiert auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, P)$, so ist $X = (X_i)_{i\in I}$ eine $(\prod_{i\in I}\Omega_i)$-wertige Zufallsvariable mit Verteilung $P^X = \bigotimes_{i\in I} P^{X_i}$ auf $\bigotimes_{i\in I}\mathscr{F}_i$.*

Beweis. Übung! □

2.30 Bemerkung. Während der vorige Satz die Existenz unabhängiger Zufallsvariablen voraussetzt und zeigt, dass ihre gemeinsame Verteilung durch das Produktmaß gegeben ist, zeigt das folgende nicht-triviale Resultat, dass ein Produktmaß stets existiert. Sein Korollar, dass Familien unabhängiger Zufallsvariablen mit gegebenen Randverteilungen existieren, ist die formale Grundlage für viele Aussagen der Stochastik, die oft mit „Es sei $(X_n)_{n\geqslant 1}$ eine Folge unabhängiger Zufallsvariablen...“ starten.

2.31 Satz. *Das Produktmaß $\bigotimes_{i\in I} P_i$ existiert für alle Wahrscheinlichkeitsräume $(\Omega_i, \mathscr{F}_i, P_i)$ und Indexmengen I. Es ist eindeutig.*

Beweis. Wir führen den Beweis nur für den Fall einer abzählbaren Indexmenge, setzen also $I = \mathbb{N}$ (für endliche Indexmengen siehe Bemerkung 2.26). Betrachte die Projektionen $\Pi_n : \prod_{i\in\mathbb{N}}\Omega_i \to \prod_{i=1}^n \Omega_i$, $\Pi_n((\omega_i)_{i\geqslant 1}) = (\omega_1, \dots, \omega_n)$ auf die ersten n Koordinaten und definiere

$$P(\Pi_n^{-1}(A_n)) := (P_1 \otimes \cdots \otimes P_n)(A_n), \quad A_n \in \mathscr{F}_1 \otimes \cdots \otimes \mathscr{F}_n.$$

Dann ist P wohldefiniert, normiert und additiv auf der Algebra

$$\mathscr{A} := \Big\{ \Pi_n^{-1}(A_n) \,\Big|\, n \in \mathbb{N},\ A_n \in \mathscr{F}_1 \otimes \cdots \otimes \mathscr{F}_n \Big\}.$$

Beachte beispielsweise für disjunkte $\Pi_n^{-1}(A_n), \Pi_m^{-1}(B_m) \in \mathscr{A}$ mit $A_n \in \mathscr{F}_n$, $B_m \in \mathscr{F}_m$ und o.B.d.A. $n \leqslant m$, dass $\Pi_n^{-1}(A_n) = \Pi_m^{-1}(A_n \times \Omega_{n+1} \cdots \times \Omega_m)$ und auch $A_n \times \Omega_{n+1} \cdots \times \Omega_m$ und B_m disjunkt sind, so dass

$$\begin{aligned}
&P\Big(\Pi_n^{-1}(A_n) \cup \Pi_m^{-1}(B_m)\Big) = (P_1 \otimes \cdots \otimes P_m)\Big((A_n \times \Omega_{n+1} \cdots \times \Omega_m) \cup B_m\Big) \\
&= (P_1 \otimes \cdots \otimes P_m)(A_n \times \Omega_{n+1} \cdots \times \Omega_m) + (P_1 \otimes \cdots \otimes P_m)(B_m) \\
&= (P_1 \otimes \cdots \otimes P_n)(A_n) + (P_1 \otimes \cdots \otimes P_m)(B_m) \\
&= P\Big(\Pi_n^{-1}(A_n)\Big) + P\Big(\Pi_m^{-1}(B_m)\Big).
\end{aligned}$$

Es bleibt zu zeigen, dass P sogar ein Prämaß auf $\mathscr{A}$ ist. Dafür reicht es, σ-Stetigkeit bei $\varnothing$ zu zeigen (Übung!), also: $A_n \in \mathscr{A}$, $A_n \downarrow \varnothing \Rightarrow P(A_n) \downarrow 0$. Wir nehmen an $P(A_n) \to \delta > 0$ für eine fallende Folge $A_{n+1} \subseteq A_n$, $n \in \mathbb{N}$, und zeigen $\bigcap_{n\geqslant 1} A_n \neq \varnothing$.

O.B.d.A. schreibe $A_n = \Pi_n^{-1}(A'_n)$ mit $A'_n \in \mathscr{F}_1 \otimes \cdots \otimes \mathscr{F}_n$ (fülle ggf. die Folge auf, betrachte also $A_1, \ldots, A_1, A_2, \ldots, A_2, \ldots$). Aus $A_{n+1} \subseteq A_n$ folgt dann $A'_{n+1} \subseteq A'_n \times \Omega_{n+1}$. Für $m > n$ setze

$$h_{m,n}(\omega_1, \ldots, \omega_n) := \int_{\Omega_{n+1}} \cdots \int_{\Omega_m} \mathbf{1}_{A'_m}(\omega_1, \ldots \omega_m)\, P_m(d\omega_m) \cdots P_{n+1}(d\omega_{n+1}).$$

Dann gilt $0 \leqslant h_{m+1,n} \leqslant h_{m,n} \leqslant \cdots \leqslant h_{n+1,n} \leqslant \mathbf{1}_{A'_n}$ wegen $A'_{m+1} \subseteq A'_m \times \Omega_{m+1}$. Mit monotoner Konvergenz folgt für $h_n := \lim_{m\to\infty} h_{m,n}$:

$$\begin{aligned}\int_{\Omega_1} h_1(\omega_1)\, P_1(d\omega_1) &= \int_{\Omega_1} \lim_{m\to\infty} h_{m,1}(\omega_1)\, P_1(d\omega_1) = \lim_{m\to\infty} \int_{\Omega_1} h_{m,1}(\omega_1)\, P_1(d\omega_1) \\ &= \lim_{m\to\infty} (P_1 \otimes \cdots \otimes P_m)(A'_m) = \lim_{m\to\infty} P(A_m) = \delta > 0.\end{aligned}$$

Es existiert also ein $\bar\omega_1 \in \Omega_1$ mit $h_1(\bar\omega_1) \geqslant \delta$. Wir nehmen jetzt induktiv an, dass es $\bar\omega_1 \in \Omega_1, \ldots, \bar\omega_n \in \Omega_n$ gibt mit $h_n(\bar\omega_1, \ldots, \bar\omega_n) \geqslant \delta$. Dann folgt wiederum mit monotoner Konvergenz

$$\begin{aligned}&\int_{\Omega_{n+1}} h_{n+1}(\bar\omega_1, \ldots, \bar\omega_n, \omega_{n+1}) P_{n+1}(d\omega_{n+1}) \\ &\quad= \lim_{m\to\infty} \int_{\Omega_{n+1}} h_{m,n+1}(\bar\omega_1, \ldots, \bar\omega_n, \omega_{n+1}) P_{n+1}(d\omega_{n+1}) \\ &\quad= \lim_{m\to\infty} h_{m,n}(\bar\omega_1, \ldots, \bar\omega_n) = h_n(\bar\omega_1, \ldots, \bar\omega_n) \geqslant \delta > 0.\end{aligned}$$

Also existiert ein $\bar\omega_{n+1} \in \Omega_{n+1}$ mit $h_{n+1}(\bar\omega_1, \ldots, \bar\omega_{n+1}) \geqslant \delta$. Mit vollständiger Induktion haben wir gezeigt, dass $\mathbf{1}_{A'_n}(\bar\omega_1, \ldots, \bar\omega_n) \geqslant h_n(\bar\omega_1, \ldots, \bar\omega_n) > 0$ für alle $n \geqslant 1$ gilt. Dies impliziert $(\bar\omega_1, \ldots, \bar\omega_n) \in A'_n$ und für die Folge $(\bar\omega_i)_{i\geqslant 1} \in A_n$ für alle $n \geqslant 1$ und somit $(\bar\omega_i)_{i\geqslant 1} \in \bigcap_{n\geqslant 1} A_n$. Der Schnitt ist also nicht leer, was die Prämaß-Eigenschaft von P nachweist.

Nach dem Satz von Caratheodory lässt sich P auf $\bigotimes_{i\in I} \mathscr{F}_i$ fortsetzen. Da jede Algebra $\cap$-stabil ist und P auf $\mathscr{A}$ eindeutig festgelegt ist, ist P als Produktmaß eindeutig. □

2.32 Bemerkung. Der Beweis hier ist ein Spezialfall des Satzes von Ionescu-Tulcea für die Konstruktion von Markovketten, vgl. Satz 14.32 in Klenke. Der Fall beliebiger Indexmengen ist im Buch *Wahrscheinlichkeitstheorie* von P. Gänssler, W. Stute, Springer (1977) behandelt (kein HU-Ebook-Zugriff). Für Wahrscheinlichkeitsräume mit gewissen Borel-σ-Algebren (auf sogenannten *polnischen* metrischen Räumen, z.B. $\mathbb{R}^d$) wird dies in Stochastik II bei der Konstruktion stochastischer Prozesse mitgezeigt.

Ende 9. Vorlesung

2.33 Korollar. *Zu vorgegebenen Wahrscheinlichkeitsmaßen P_i auf $(\Omega_i, \mathscr{F}_i)$, $i \in I$, existiert ein Wahrscheinlichkeitsraum mit einer Familie unabhängiger $(\Omega_i, \mathscr{F}_i)$-wertiger Zufallsvariablen $(X_i)_{i\in I}$, deren Verteilung P_i ist.*

Beweis. Betrachte auf dem Produktraum $\Omega = \prod_{i\in I}\Omega_i$, $\mathscr{F} = \bigotimes_{i\in I}\mathscr{F}_i$, $P = \bigotimes_{i\in I}P_i$ die Koordinatenprojektionen $X_i : \Omega \to \Omega_i$, $X_i((\omega_j)_{j\in I}) = \omega_i$, die nach Definition der Produkt-σ-Algebra messbar sind und die für endliche $J \subseteq I$

$$P\Big(\bigcap_{j\in J}\{X_j \in A_j\}\Big) = P\Big(\bigcap_{j\in J}\pi_j^{-1}(A_j)\Big) = \prod_{j\in J}P_j(A_j) = \prod_{j\in J}P(X_j \in A_j)$$

für alle $A_j \in \mathscr{F}_j$ erfüllen. Damit ist $(X_i)_{i\in I}$ eine Familie unabhängiger Zufallsvariablen und es gilt $P^{X_i} = P_i$ für jedes $i \in I$. □

2.34 Bemerkung. Sind die Ereignisse $(A_n)_{n\geqslant 1}$ unabhängig, so gilt nach dem Lemma von Borel-Cantelli stets

$$P\Big(\limsup_{n\to\infty} A_n\Big) \in \{0,1\}.$$

Es kann also nie eine Wahrscheinlichkeit zwischen 0 und 1 vorkommen! Eine solche Aussage nennt man 0-1-Gesetz (das Ereignis tritt entweder fast sicher ein oder nicht ein). Wie wir sehen werden, genügen viele „Grenzwerte“ einem solchen 0-1-Gesetz. Grundlage des Beweises und der Intuition ist, dass ein Ereignis A von sich selbst unabhängig ist genau dann, wenn

$$P(A\cap A) = P(A)P(A) \iff P(A) \in \{0,1\}.$$

2.35 Definition. Es sei $(X_n)_{n\geqslant 1}$ eine Folge von Zufallsvariablen auf $(\Omega, \mathscr{F}, P)$ mit Werten in $(S_n, \mathscr{S}_n)$. Ein Ereignis $A \in \mathscr{F}$ heißt asymptotisch bezüglich (X_n), falls es für alle $k \geqslant 1$ nur von $(X_n, n \geqslant k)$ abhängt in dem Sinne, dass $A \in \mathscr{A}_X$ gilt. Hierbei ist die asymptotische σ-Algebra $\mathscr{A}_X$ definiert als

$$\mathscr{A}_X := \bigcap_{k\geqslant 1}\sigma\Big(\bigcup_{n\geqslant k}\mathscr{F}^{X_n}\Big).$$

2.36 Beispiele.

(a) Sind $(X_n)_{n\geqslant 1}$ reellwertige Zufallsvariablen und $B \in \mathfrak{B}_{\mathbb{R}}$, so ist $\limsup_{n\to\infty}\{X_n \in B\} = \{\text{unendlich viele } X_n \text{ liegen in } B\}$ ein asymptotisches Ereignis; denn $\{X_n \in B\} \in \mathscr{F}^{X_n}$ impliziert $\bigcup_{n\geqslant m}\{X_n \in B\} \in \sigma(\bigcup_{n\geqslant k}\mathscr{F}^{X_n})$ für alle $m \geqslant k$ und somit

$$\forall k \geqslant 1 : \bigcap_{m\geqslant k}\bigcup_{n\geqslant m}\{X_n \in B\} \in \sigma\Big(\bigcup_{n\geqslant k}\mathscr{F}^{X_n}\Big).$$

Nun ist aber $\bigcap_{m\geqslant 1}\bigcup_{n\geqslant m}\{X_n \in B\} = \bigcap_{m\geqslant k}\bigcup_{n\geqslant m}\{X_n \in B\}$ für alle $k \geqslant 1$ (wieso?) und daher $\bigcap_{m\geqslant 1}\bigcup_{n\geqslant m}\{X_n \in B\} \in \bigcap_{k\geqslant 1}\sigma(\bigcup_{n\geqslant k}\mathscr{F}^{X_n})$, was gerade $\limsup_{n\to\infty}\{X_n \in B\} \in \mathscr{A}_X$ bedeutet. Insbesondere ist $\mathscr{A}_X$ im Allgemeinen nicht trivial ($\mathscr{A}_X \neq \{\varnothing, \Omega\}$), was auf den ersten Blick vielleicht der Intuition widerspricht.

(b) Es seien X_k, $k \in \mathbb{N}$, reellwertige Zufallsvariablen sowie A das Ereignis, dass $\lim_{n\to\infty}\sum_{k=1}^n X_k$ existiert. Dann gilt nach dem Cauchy-Kriterium

$$A = \bigcap_{\varepsilon>0,\,\varepsilon\in\mathbb{Q}}\bigcup_{N\geqslant 1} B_N \text{ mit } B_N = \bigcap_{m\geqslant n\geqslant N}\Big\{\sum_{k=n}^m X_k \in (-\varepsilon,\varepsilon)\Big\}.$$

Nun ist $B_N \subseteq B_{N+1}$ sowie $B_N \in \sigma(\bigcup_{k\geqslant N} \mathscr{F}^{X_k})$. Daher gilt $\bigcup_{N=1}^{N_{max}} B_N = B_{N_{max}} \in \sigma(\bigcup_{k\geqslant K} \mathscr{F}^{X_k})$ für alle $N_{max} \geqslant K$. Dies impliziert $\bigcup_{N=1}^{\infty} B_N \in \sigma(\bigcup_{k\geqslant K} \mathscr{F}^{X_k})$ für alle $K \in \mathbb{N}$ und somit, dass A ein asymptotisches Ereignis ist.

2.37 Satz (0-1-Gesetz von Kolmogorov)**.** *Es seien $(X_i)_{i\geqslant 1}$ unabhängige Zufallsvariablen auf $(\Omega, \mathscr{F}, P)$ mit Werten in $(S_i, \mathscr{S}_i)$. Dann gilt für jedes bezüglich (X_i) asymptotische Ereignis $A \in \mathscr{A}_X$: $P(A) = 0$ oder $P(A) = 1$.*

2.38 Bemerkung. Für den Beweis des 0-1-Gesetzes benötigen wir zunächst ein Lemma, das sehr intuitiv, aber etwas technisch zu beweisen ist.

2.39 Lemma. *Es seien $(X_i)_{i\in I}$ eine Familie unabhängiger Zufallsvariablen mit Werten in $(S_i, \mathscr{S}_i)$ und $I = I_1 \cup I_2$ eine disjunkte Zerlegung von I. Dann sind die σ-Algebren $\mathscr{F}_1 := \sigma(\bigcup_{i\in I_1} \mathscr{F}^{X_i})$ und $\mathscr{F}_2 := \sigma(\bigcup_{i\in I_2} \mathscr{F}^{X_i})$ unabhängig.*

Beweis. Nach Lemma 2.29 sind $Y_j = (X_i)_{i\in I_j}$ mit $j = 1, 2$ jeweils $(\prod_{i\in I_j} S_i)$-wertige Zufallsvariablen. Für jedes $i \in I_j$ gilt $X_i = \pi_i^j \circ Y_j$ mit der Koordinatenprojektion $\pi_i^j : \prod_{i'\in I_j} S_{i'} \to S_i$, so dass $\mathscr{F}^{X_i} \subseteq \mathscr{F}^{Y_j}$ für alle $i \in I_j$, also $\mathscr{F}_j \subseteq \mathscr{F}^{Y_j}$ gilt. Andererseits wird $\bigotimes_{i\in I_j} \mathscr{S}_i$ erzeugt von $\bigcap_{i\in I_j'} (\pi_i^j)^{-1}(A_i)$ mit $I_j' \subseteq I_j$ endlich, $A_i \in \mathscr{S}_i$, und es gilt

$$Y_j^{-1}\Big(\bigcap_{i\in I_j'} (\pi_i^j)^{-1}(A_i)\Big) = \bigcap_{i\in I_j'} Y_j^{-1}\Big((\pi_i^j)^{-1}(A_i)\Big) = \bigcap_{i\in I_j'} X_i^{-1}(A_i) \in \mathscr{F}_j.$$

Wir schließen mit Lemma 1.40 also sogar $\mathscr{F}_j = \mathscr{F}^{Y_j}$, so dass es genügt, die Unabhängigkeit von Y_1 und Y_2 nachzuweisen.

Nach Satz 2.20 genügt es, die Unabhängigkeit auf $\cap$-stabilen Erzeugern zu zeigen. Wir betrachten dazu wieder endliche Schnitte $\bigcap_{i\in I_j'} (\pi_i^j)^{-1}(A_i)$ und erhalten wegen Unabhängigkeit der X_i

$$\begin{aligned}
&P\Big(\Big\{Y_1 \in \bigcap_{i\in I_1'} (\pi_i^1)^{-1}(A_i)\Big\} \cap \Big\{Y_2 \in \bigcap_{i\in I_2'} (\pi_i^2)^{-1}(A_i)\Big\}\Big) = P\Big(\bigcap_{i\in I_1'\cup I_2'} \{X_i \in A_i\}\Big) \\
&= \prod_{i\in I_1'} P(X_i \in A_i) \prod_{i\in I_2'} P(X_i \in A_i) \\
&= P\Big(Y_1 \in \bigcap_{i\in I_1'} (\pi_i^1)^{-1}(A_i)\Big) P\Big(\Big\{Y_2 \in \bigcap_{i\in I_2'} (\pi_i^2)^{-1}(A_i)\Big).
\end{aligned}$$

Also sind Y_1 und Y_2 sowie $\mathscr{F}_1$ und $\mathscr{F}_2$ unabhängig. □

Beweis des 0-1-Gesetzes. Betrachte die σ-Algebren $\mathscr{F}_{\leqslant n} := \sigma(\bigcup_{i=1}^{n} \mathscr{F}^{X_i})$, $\mathscr{F}_{>n} := \sigma(\bigcup_{i=n+1}^{\infty} \mathscr{F}^{X_i})$, die von den ersten n bzw. allen außer den ersten n X_i erzeugt werden. Nach dem Lemma sind $\mathscr{F}_{\leqslant n}$ und $\mathscr{F}_{>n}$ unabhängig. Die asymptotische σ-Algebra erfüllt $\mathscr{A}_X \subseteq \mathscr{F}_{>n}$ und ist daher unabhängig von $\bigcup_{n\geqslant 1} \mathscr{F}_{\leqslant n}$ (das ist eine Algebra, aber keine σ-Algebra).

$X = (X_i)_{i\geqslant 1}$ ist eine $(\prod_{i\geqslant 1} S_i)$-wertige Zufallsvariable mit $\mathscr{F}^X = \sigma(\bigcup_{n\geqslant 1} \mathscr{F}_{\leqslant n})$ (wegen Messbarkeit der $X_i = \pi_i \circ X$). Da $\bigcup_{n\geqslant 1} \mathscr{F}_{\leqslant n}$ ein $\cap$-stabiler

Erzeuger von $\mathscr{F}^X$ und unabhängig von $\mathscr{A}_X$ ist, sind also X und $\mathbf{1}_A$ für jedes $A \in \mathscr{A}_X$ gemäß Satz 2.20 unabhängige Zufallsvariablen. Es folgt, dass $\mathscr{A}_X$ und $\mathscr{F}^X$ unabhängige σ-Algebren sind.

Nun gilt $\mathscr{A}_X \subseteq \mathscr{F}^X$ wegen $\mathscr{F}^{X_i} \subseteq \mathscr{F}^X$, so dass $\mathscr{A}_X$ insbesondere von sich selbst unabhängig ist. Für jedes $A \in \mathscr{A}_X$ gilt also $P(A \cap A) = P(A)^2$ und daher $P(A) \in \{0, 1\}$. □

2.40 Beispiel (Harmonische Reihe mit zufälligen Vorzeichen)**.** Sind $(\varepsilon_k)_{k \geqslant 1}$ unabhängige $\{-1, 1\}$-wertige Zufallsvariablen, so konvergiert nach Beispiel 2.36(b) mit $X_k = \varepsilon_k / k$ und dem 0-1-Gesetz die Reihe $\sum_{k=1}^{\infty} \varepsilon_k \frac{1}{k}$ entweder mit Wahrscheinlichkeit 1 oder sie divergiert mit Wahrscheinlichkeit 1. Diese Aussage gilt für jede beliebige Randverteilung P^{ε_k} der Zufallsvariablen ε_k. Aus der Analysis sind $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$ (harmonische Reihe) und $\sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k} = 1 - \frac{1}{2} + \frac{1}{3} \mp \cdots = \log(2)$ (alternierende harmonische Reihe) wohlbekannt. Insbesondere für den symmetrischen Fall $P(\varepsilon_k = +1) = P(\varepsilon_k = -1) = 1/2$ scheint es auf Anhieb vollkommen unklar, mit welcher Wahrscheinlichkeit Konvergenz vorliegt. Das 0-1-Gesetz liefert immerhin eine klare Dichotomie; welcher der beiden Fälle vorliegt, werden wir bei den Gesetzen der großen Zahlen lernen.

3 Erwartungswert, Varianz und Kovarianz

3.1 Erwartungswert und Momente

3.1 Bemerkung. Die stochastischen Eigenschaften einer Zufallsvariablen werden über ihre Verteilung festgelegt. Weil Wahrscheinlichkeitsmaße sehr komplex sind, wollen wir $\mathbb{R}^d$-wertige Zufallsvariablen durch einfache Kenngrößen beschreiben. Die wichtigste ist ihr Erwartungswert, so er existiert, der einen Mittel- oder Schwerpunkt der Verteilung angibt. Wir gehen schrittweise vor und wiederholen dabei die Konstruktion des Maßintegrals aus stochastischer Sicht.

3.2 Definition. Eine reellwertige Zufallsvariable X auf $(\Omega, \mathscr{F}, P)$ heißt einfach, falls sie nur endlich viele Werte annimmt, d.h. es folgende Darstellung gibt:

$$X = \sum_{i=1}^{m} \alpha_i \mathbf{1}_{A_i} \text{ mit } m \in \mathbb{N},\ \alpha_i \in \mathbb{R},\ A_i \in \mathscr{F}.$$

Für eine solche Zufallsvariable definieren wir ihren Erwartungswert als

$$\mathbb{E}[X] := \sum_{i=1}^{m} \alpha_i P(A_i).$$

3.3 Beispiel. Beim Würfeln mit zwei fairen Würfeln ergibt sich für die Augensumme S

$$\mathbb{E}[S] = 2P(S = 2) + 3P(S = 3) + \cdots + 12P(S = 12) = 7.$$

3.4 Lemma. *Für eine einfache Zufallsvariable X auf $(\Omega, \mathscr{F}, P)$ gilt:*

(a) $\mathbb{E}[X] = \sum_{x \in X(\Omega)} xP(X = x)$*; insbesondere hängt der Erwartungswert nur von der Verteilung* P^X *von* X *ab.*

(b) Der Erwartungswert ist linear und monoton: ist Y *eine weitere einfache Zufallsvariable und sind* $\alpha, \beta \in \mathbb{R}$*, so gilt*

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\,\mathbb{E}[X] + \beta\,\mathbb{E}[Y];$$

aus $X \leqslant Y$ *(d.h.* $\forall \omega \in \Omega :\ X(\omega) \leqslant Y(\omega)$*) folgt* $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$*.*

(c) Falls X *und* Y *unabhängige einfache Zufallsvariablen sind, so gilt* $\mathbb{E}[X \bullet Y] = \mathbb{E}[X] \bullet \mathbb{E}[Y]$*.*

(d) Für jedes $A \in \mathscr{F}$ *gilt* $\mathbb{E}[\mathbf{1}_A] = P(A)$*.*

Beweis. Ist $X = \sum_{i=1}^m \alpha_i \mathbf{1}_{A_i}$ und sind o.B.d.A. alle α_i paarweise verschieden, so gilt

$$\mathbb{E}[X] = \sum_{i=1}^m \alpha_i P(A_i) = \sum_{i=1}^m \alpha_i P(X = \alpha_i) = \sum_{x \in X(\Omega)} xP(X = x),$$

wobei $X(\Omega) = \{\alpha_i \,|\, i = 1, \ldots, m\}$ oder $X(\Omega) = \{\alpha_i \,|\, i = 1, \ldots, m\} \cup \{0\}$ gilt, was im Erwartungswert keinen Unterschied macht. Wegen $P(X = x) = P^X(\{x\})$ hängt der Erwartungswert nur von P^X ab, so dass (a) bewiesen ist.

Gilt $X = \sum_{i=1}^m \alpha_i \mathbf{1}_{A_i}$, $Y = \sum_{j=1}^n \beta_j \mathbf{1}_{B_j}$ in (b), so können wir mit Mengen C_k der Form $A_i \cap B_j$ auch eine gemeinsame Darstellung $X = \sum_{k=1}^K \alpha'_k \mathbf{1}_{C_k}$, $Y = \sum_{k=1}^K \beta'_k \mathbf{1}_{C_k}$ finden. Dann folgt die Linearität des Erwartungswerts einfach aus

$$\sum_{k=1}^K \Big(\alpha\alpha'_k + \beta\beta'_k\Big)\mathbf{1}_{C_k} = \alpha \sum_{k=1}^K \alpha'_k \mathbf{1}_{C_k} + \beta \sum_{k=1}^K \beta'_k \mathbf{1}_{C_k}.$$

Aus $X \leqslant Y$ schließen wir $\alpha'_k \leqslant \beta'_k$ für alle k (sofern $C_k \neq \varnothing$) und jeder Summand in der Erwartungswertdarstellung von X ist nicht größer als der entsprechende Summand von Y, was $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$ impliziert.

Sind X und Y unabhängig in (c), so gilt gemäß (a) für ihr Produkt $X \bullet Y$

$$\begin{aligned}
\mathbb{E}[X \bullet Y] &= \sum_{z \in (X \bullet Y)(\Omega)} zP(X \bullet Y = z) = \sum_{z \in (X \bullet Y)(\Omega)} \sum_{x \in X(\Omega)} zP(X \bullet Y = z, X = x) \\
&= \sum_{z \in (X \bullet Y)(\Omega), z \neq 0} \sum_{x \in X(\Omega), x \neq 0} zP(X = x)P(Y = z/x) \\
&= \sum_{y \in Y(\Omega)} \sum_{x \in X(\Omega)} xyP(X = x)P(Y = y) = \mathbb{E}[X] \bullet \mathbb{E}[Y].
\end{aligned}$$

Teil (d) folgt direkt aus der Definition des Erwartungswerts. □

Ende 10. Vorlesung

3.5 Beispiel. Eine $\mathrm{Bin}(n,p)$-verteilte Zufallsvariable X ist einfach. Es gilt

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{n} kP(X=k) = \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\ &= np\sum_{l=0}^{n-1}\binom{n-1}{l}p^l(1-p)^{n-1-l} = np.\end{aligned}$$

Die $\mathrm{Bin}(n,p)$-Verteilung besitzt also den Erwartungswert np. Beachte, dass man wegen Lemma 3.4(a) auch vom *Erwartungswert einer Verteilung* spricht.

3.6 Definition. Es sei $X \geqslant 0$ eine nichtnegative Zufallsvariable. Sind dann X_n einfache nichtnegative Zufallsvariablen mit $X_n(\omega) \uparrow X(\omega)$ für $n \to \infty$ und alle $\omega \in \Omega$, so definiere den Erwartungswert

$$\mathbb{E}[X] := \lim_{n\to\infty} \mathbb{E}[X_n] \ \in [0,+\infty]$$

(man kann zeigen, dass eine solche Folge (X_n) stets existiert und $\mathbb{E}[X]$ nicht von der Auswahl der X_n abhängt). Betrachte nun auf dem Wahrscheinlichkeitsraum $(\Omega,\mathscr{F},P)$ die Menge der Zufallsvariablen

$$\mathscr{L}^1 := \mathscr{L}^1(P) := \mathscr{L}^1(\Omega,\mathscr{F},P) := \{X:\Omega\to\mathbb{R} \quad \text{messbar} \mid \mathbb{E}[|X|]<\infty\}.$$

Dann definiere für $X \in \mathscr{L}^1$ mit $X_+ := \max(X,0)$, $X_- := \max(-X,0)$ den Erwartungswert als

$$\mathbb{E}[X] := \mathbb{E}[X_+] - \mathbb{E}[X_-] \in \mathbb{R}\,.$$

Der Erwartungswert $\mathbb{E}[X]$ ist also das Lebesgueintegral von X bezüglich P, und man schreibt $\mathbb{E}[X] = \int X\,dP = \int_\Omega X(\omega)\,P(d\omega)$ sowie $\int_A X\,dP = \int_\Omega X(\omega)\mathbf{1}_A(\omega)\,P(d\omega)$ für $A \in \mathscr{F}$.

3.7 Satz. *Es seien X ein Zufallsvariable mit Werten in $(S,\mathscr{S})$ und $h: S\to\mathbb{R}$ messbar (bzgl. $\mathfrak{B}_\mathbb{R}$). Dann gilt:*

(a) $h(X)\in\mathscr{L}^1 \iff \int_S |h(x)|\,P^X(dx)<\infty$. In dem Fall gilt

$$\mathbb{E}[h(X)] = \int_S h(x)\,P^X(dx).$$

(b) Ist X ein Zufallsvektor im $\mathbb{R}^d$ mit Dichte f^X, so gilt $h(X)\in\mathscr{L}^1 \iff \int_{\mathbb{R}^d}|h(x)|f^X(x)\,dx<\infty$. In dem Fall ist

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}^d} h(x)f^X(x)\,dx.$$

(c) Ist X diskret verteilt auf $\mathbb{Z}$ mit Zähldichte p^X, so gilt $h(X)\in\mathscr{L}^1 \iff \sum_{k\in\mathbb{Z}}|h(k)|p^X(k)<\infty$. In dem Fall ist

$$\mathbb{E}[h(X)] = \sum_{k\in\mathbb{Z}} h(k)p^X(k).$$

3.8 Bemerkung. In der Maßtheorie ist (a) genau die Eigenschaft des Bildmaßes P^X von P unter X.

3.9 Beispiel. Ist X eine reellwertige Zufallsvariable mit Dichte f^X, so gilt $X \in \mathscr{L}^1$, falls $\int_{\mathbb{R}} |x| f^X(x)\,dx < \infty$, und dann $\mathbb{E}[X] = \int_{\mathbb{R}} x f^X(x)\,dx$ (setze $h(x) = x$). Stets gilt $\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f^X(x)\,dx \in [0,\infty]$ (setze $h(x) = x^2$).

Beweis. Wir beweisen alle Teile mittels *maßtheoretischer Induktion.* Für Indikatorfunktionen $h = \mathbf{1}_A$, $A \in \mathscr{S}$, sind $h(X)$ und h einfache Zufallsvariablen auf $(\Omega, \mathscr{F}, P)$ bzw. $(S, \mathscr{S}, P^X)$, so dass stets $h(X) \in \mathscr{L}^1$ und $\mathbb{E}[h(X)] = P(X \in A) = \int h dP^X$ gilt. Ist h einfach, also eine Linearkombination von Indikatorfunktionen, so folgt die Identität in (a) durch Linearität des Integrals (über einfache Funktionen). Ist $h \geqslant 0$ messbar, so existieren einfache $h_n \uparrow h$ (so dass auch $h_n \circ X \uparrow h \circ X$), und es folgt nach Definition der Integrale bezüglich P, P^X und mit der Identität für einfache Funktionen h_n

$$\mathbb{E}[h(X)] = \lim_{n\to\infty} \mathbb{E}[h_n(X)] = \lim_{n\to\infty} \int_S h_n(x)\, P^X(dx) = \int_S h(x)\, P^X(dx).$$

Schließlich sehen wir wegen $|h| \geqslant 0$ für beliebige messbare h, dass gilt

$$h(X) \in \mathscr{L}^1 \iff \int |h| dP^X = \mathbb{E}[|h|(X)] < \infty,$$

und dann mit $h = h_+ - h_-$ und $h_+, h_- \geqslant 0$, dass

$$\begin{aligned}\mathbb{E}[h(X)] &= \mathbb{E}[h_+(X)] - \mathbb{E}[h_-(X)] \\ &= \int_S h_+(x)\, P^X(dx) - \int_S h_-(x)\, P^X(dx) = \int_S h(x)\, P^X(dx).\end{aligned}$$

Für (b) müssen wir wegen (a) nur zeigen, dass $\int h dP^X = \int h f^X$ für alle messbaren $h \geqslant 0$ gilt (zerlege allgemein $h = h_+ - h_-$). Die Dichteeigenschaft zeigt gerade

$$\int_{\mathbb{R}^d} \mathbf{1}_B(x)\, P^X(dx) = P^X(B) = \int_B f^X(x)\,dx = \int_{\mathbb{R}^d} \mathbf{1}_B(x) f^X(x)\,dx, \quad B \in \mathfrak{B}_{\mathbb{R}^d}.$$

Also gilt die Identität für Indikatorfunktionen $h = \mathbf{1}_B$. Mit Linearität gilt sie auch für einfache Funktionen h und mit Approximation damit für alle messbaren $h \geqslant 0$.

Teil (c) folgt vollkommen analog zu (b). □

3.10 Satz. *Für $X, Y \in \mathscr{L}^1(\Omega, \mathscr{F}, P)$ gilt:*

(a) Der Erwartungswert ist linear: Sind $\alpha, \beta \in \mathbb{R}$, so gilt

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\, \mathbb{E}[X] + \beta\, \mathbb{E}[Y].$$

(b) Der Erwartungswert ist monoton: Ist $X \leqslant Y$, so gilt $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$. Aus $X \leqslant Y$ und $\mathbb{E}[X] = \mathbb{E}[Y]$ folgt $P(X = Y) = 1$ (man sagt: es gilt $X = Y$ P-fast sicher).

(c) *Falls* $X, Y \in \mathscr{L}^1$ *unabhängig sind, so gilt* $X \bullet Y \in \mathscr{L}^1$ *und* $\mathbb{E}[X \bullet Y] = \mathbb{E}[X] \bullet \mathbb{E}[Y]$.

Beweis. Die Linearität in (a) folgt aus der Linearität für einfache Zufallsvariablen und Approximation. Für die Monotonie in (b) genügt es, $\mathbb{E}[Z] \geqslant 0$ für $Z \geqslant 0$ zu zeigen (setze $Z = Y - X$ und nutze Linearität). Das folgt wiederum mittels Approximation durch einfache Zufallsvariablen. Ist $Z \geqslant 0$, so gibt es einfache Zufallsvariablen $Z_n \geqslant 0$ mit $Z_n \uparrow Z$ und $\mathbb{E}[Z_n] \uparrow \mathbb{E}[Z]$. Aus $\mathbb{E}[Z] = 0$ folgt daher $\mathbb{E}[Z_n] = 0$ und somit nach Definition $P(Z_n > 0) = 0$. Also gilt auch $P(Z > 0) = P(\bigcup_{n \geqslant 1} \{Z_n > 0\}) = 0$. Mit $Z = Y - X$ beweist das $P(X = Y) = 1$. Für (c) kann man wieder über Approximation mit einfachen Zufallsvariablen argumentieren. Eleganter ist die Argumentation mit dem Satz von Fubini, da bei Unabhängigkeit $P^{(X,Y)} = P^X \otimes P^Y$ gilt:

$$\begin{aligned}\mathbb{E}[|X \bullet Y|] &= \int_{\mathbb{R}^2} |xy| P^{(X,Y)}(dx, dy) = \int_{\mathbb{R}^2} |xy| (P^X \otimes P^Y)(dx, dy) \\ &= \int_{\mathbb{R}} |x| P^X(x) \int_{\mathbb{R}} |y| P^Y(y) = \mathbb{E}[|X|]\, \mathbb{E}[|Y|] < \infty.\end{aligned}$$

Also ist $X \bullet Y \in \mathscr{L}^1$. Wenn wir in der Rechnung nun die Beträge weglassen, so ergibt sich die behauptete Identität mit dem Satz von Fubini. □

3.11 Definition. Wir sagen, dass eine Zufallsvariable X in $\mathscr{L}^p = \mathscr{L}^p(P)$ liegt für $p > 0$, falls $|X|^p \in \mathscr{L}^1$, also $\mathbb{E}[|X|^p] < \infty$ gilt. Für $X \in \mathscr{L}^p$ und $p \in \mathbb{N}$ heißt $\mathbb{E}[X^p]$ das p-te Moment von X.

3.12 Satz. *Für* $X \in \mathscr{L}^p$ *und* $Y \in \mathscr{L}^q$ *mit* $p, q \geqslant 1$ *und* $\frac{1}{p} + \frac{1}{q} = 1$ *gelten* $XY \in \mathscr{L}^1$ *und die Hölder-Ungleichung*

$$|\mathbb{E}[XY]| \leqslant \mathbb{E}[|XY|] \leqslant \mathbb{E}[|X|^p]^{1/p}\, \mathbb{E}[|Y|^q]^{1/q}.$$

Insbesondere gelten $XY \in \mathscr{L}^1(P)$ *für* $X, Y \in \mathscr{L}^2(P)$ *und die Cauchy-Schwarz-Ungleichung*

$$|\mathbb{E}[XY]| \leqslant \mathbb{E}[X^2]^{1/2}\, \mathbb{E}[Y^2]^{1/2}.$$

Beweis. Siehe Analysis 3 bzw. Maßtheorie. □

3.13 Korollar. *Für* $0 < p \leqslant q$ *gelten* $\mathscr{L}^q \subseteq \mathscr{L}^p$ *und* $\mathbb{E}[|X|^p] \leqslant \mathbb{E}[|X|^q]^{p/q}$ *für* $X \in \mathscr{L}^q$. *Insbesondere ist* $\mathscr{L}^2 \subseteq \mathscr{L}^1$.

Beweis. Für $X \in \mathscr{L}^q$ gilt nach der Hölder-Ungleichung mit Exponenten q/p und $q/(q-p)$

$$\mathbb{E}[|X|^p] = \mathbb{E}[(|X|^q)^{p/q}] \leqslant \mathbb{E}[|X|^q]^{p/q}\, \mathbb{E}[1^{q/(q-p)}]^{(q-p)/q} = \mathbb{E}[|X|^q]^{p/q} < \infty,$$

was gerade die Behauptung ist. □

3.14 Bemerkung. Die Inklusion allgemeiner $\mathscr{L}^p(\mu)$-Räume ist nur bei endlichen Maßen μ wie im Korollar. Beim Lebesguemaß im $\mathbb{R}^d$ oder bei Folgenräumen ℓ^p gilt sie beispielsweise nicht.

Ende 11. Vorlesung

3.15 Satz. *Für eine Zufallsvariable $X \in \mathscr{L}^2$ gilt die Bias-Varianz-Zerlegung*

$$\forall x \in \mathbb{R}:\ \mathbb{E}[(X-x)^2] = \underbrace{(\mathbb{E}[X]-x)^2}_{Bias^2} + \underbrace{\mathbb{E}[(X-\mathbb{E}[X])^2]}_{Varianz}.$$

Die Funktion $d(x) = \mathbb{E}[(X-x)^2]$ nimmt ihr Minimum auf $\mathbb{R}$ genau bei $x = \mathbb{E}[X]$ an.

Beweis. Beachte zunächst, dass auch $X \in \mathscr{L}^1$ gilt und somit $\mathbb{E}[X]$ wohldefiniert ist. Der Nachweis erfolgt durch Einfügen einer *nahrhaften Null*:

$$\begin{aligned}\mathbb{E}[(X-x)^2] &= \mathbb{E}[(X-\mathbb{E}[X]+\mathbb{E}[X]-x)^2] \\ &= \mathbb{E}[(X-\mathbb{E}[X])^2] + 2\,\mathbb{E}[X-\mathbb{E}[X]](\mathbb{E}[X]-x) + (\mathbb{E}[X]-x)^2\end{aligned}$$

und die behauptete Identität folgt aus $\mathbb{E}[X-\mathbb{E}[X]] = 0$. Wegen der Bias-Varianz-Zerlegung ist $d(x)$ die Summe aus dem quadrierten Bias, der genau für $x = \mathbb{E}[X]$ minimal ist, und der Varianz, die nicht von x abhängt. □

3.16 Bemerkung. Die Bias-Varianz-Zerlegung ist in der Statistik von entscheidender Bedeutung, um den quadratischen Fehler eines Schätzers X (in der Statistik $\hat{\vartheta}$) eines Parameters x (in der Statistik ϑ) zu optimieren. Die Eigenschaft $\mathbb{E}[X] = \operatorname{argmin}_x d(x)$ kann auch als Motivation für den Erwartungswert herangezogen werden: $\mathbb{E}[X]$ ist diejenige deterministische Zahl, die die Zufallsvariable X am besten beschreibt (den quadratischen Fehler minimiert).

Wir kommen jetzt noch zu einer weiteren wichtigen Ungleichung, die in der Form nur für Integrale bezüglich normierten Maßen, also Erwartungswerte gilt.

3.17 Definition. Es sei $I \subseteq \mathbb{R}$ ein (ggf. auch unendliches) Intervall. Eine Funktion $\varphi : I \to \mathbb{R}$ heißt konvex, falls

$$\forall x, y \in I,\ \alpha \in [0,1]:\ \varphi(\alpha x + (1-\alpha)y) \leqslant \alpha\varphi(x) + (1-\alpha)\varphi(y).$$

3.18 Beispiel. Ist $\varphi : I \to \mathbb{R}$ stetig differenzierbar mit monoton wachsender Ableitung φ' (z.B. wenn $\varphi'' \geqslant 0$ auf I), so ist φ konvex. Insbesondere sind $\varphi(x) = e^{\alpha x}$ für $\alpha \in \mathbb{R}$ beliebig und $\varphi(x) = |x|^p$ für $p > 1$ konvex auf $I = \mathbb{R}$. Auch $\varphi(x) = |x|$ ist konvex auf $\mathbb{R}$.

3.19 Lemma. *Ist $\varphi : I \to \mathbb{R}$ konvex, so existieren für alle $x \in int(I)$ (Inneres von I) die links- und rechtsseitigen Ableitungen $\varphi'(x-)$, $\varphi'(x+) \in \mathbb{R}$ mit $\varphi'(x-) \leqslant \varphi'(x+)$. Es gilt*

$$\forall x \in int(I),\ y \in I:\ \varphi(y) \geqslant \varphi(x) + \varphi'(x+)(y-x)$$

sowie

$$\varphi(y) = \sup_{x \in int(I)} \big(\varphi(x) + \varphi'(x+)(y-x)\big), \quad y \in int(I), \tag{3.1}$$

Beweis. Für die rechtsseitige Ableitung betrachte $y > x$ und den Differenzenquotienten $D(y,x) := \frac{\varphi(y)-\varphi(x)}{y-x}$. Für $t = \alpha x + (1-\alpha)y \in (x,y)$, $\alpha \in (0,1)$, gilt dann

$$D(t,x) = \frac{\varphi(\alpha x + (1-\alpha)y) - \varphi(x)}{(1-\alpha)(y-x)} \leqslant \frac{(1-\alpha)\varphi(y) + (\alpha-1)\varphi(x)}{(1-\alpha)(y-x)} = D(y,x).$$

Zerlegt man $D(y,x)$ in $D(t,x)$ und $D(y,t)$, so zeigt dies auch $D(t,x) \leqslant D(y,t)$.

Für $x_n \downarrow x$ ist $D(x_n,x)$ also monoton fallend mit einem Grenzwert $\varphi'(x+) \in [-\infty,\infty)$. Weiterhin folgt $D(x_n,x) \geqslant D(x_n',x)$ für $x_n' < x < x_n$, so dass $\varphi'(x+) \geqslant D(x_n',x) > -\infty$ gilt. Für $x_n' \uparrow x$ ist analog $D(x_n',x)$ monoton wachsend mit Grenzwert $\varphi'(x-) \leqslant \inf_{n\geqslant 1} D(x_n,x) = \varphi'(x+)$.

Für $y > x$ impliziert $D(y,x) \geqslant \varphi'(x+)$ gerade $\varphi(y) \geqslant \varphi(x) + \varphi'(x+)(y-x)$. Für $y < x$ folgt dieselbe Ungleichung aus $D(y,x) \leqslant \varphi'(x-) \leqslant \varphi'(x+)$.

Gleichung (3.1) folgt aus der Ungleichung sowie Einsetzen von $x = y \in \mathrm{int}(I)$. □

3.20 Bemerkung. Das Lemma zeigt auch, dass jede konvexe Funktion stetig und damit Borel-messbar ist. Die reichhaltige Analysis konvexer Funktionen beruht gerade auf der Eigenschaft (3.1), dass φ als Supremum über lineare Funktionen (sogenannte Subdifferentiale) dargestellt werden kann.

3.21 Satz. *Es sei $X \in \mathscr{L}^1$ mit Werten in einem offenen Intervall I. Dann gilt $\mathbb{E}[X] \in I$. Ist $\varphi : I \to \mathbb{R}$ konvex und $\varphi(X) \in \mathscr{L}^1$, so gilt die Jensensche Ungleichung*

$$\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)].$$

Beweis. Es sei $I = (a,b)$ mit $a \in \mathbb{R}\cup\{-\infty\}$, $b \in \mathbb{R}\cup\{\infty\}$. für $a \in \mathbb{R}$ gilt nach Voraussetzung $X - a > 0$ und wegen der Monotonie des Erwartungswerts $\mathbb{E}[X-a] \geqslant 0$. Wäre $\mathbb{E}[X] = a$, so hätten wir nach Satz 3.10(b) $X = a$ P-fast sicher. Also folgt $\mathbb{E}[X] > a$. Aus $X < b$ mit $b \in \mathbb{R}$ folgt $\mathbb{E}[X] < b$ analog (oder betrachte $-X$). Daher gilt stets $\mathbb{E}[X] \in I$.

Die Darstellung (3.1) und die Monotonie des Erwartungswerts zeigen gerade

$$\forall\, x \in I:\ \mathbb{E}[\varphi(X)] \geqslant \mathbb{E}[\varphi(x) + \varphi'(x+)(X-x)] = \varphi(x) + \varphi'(x+)(\mathbb{E}[X] - x).$$

Weil I offen ist, erhalten wir unter erneuter Anwendung von (3.1)

$$\mathbb{E}[\varphi(X)] \geqslant \sup_{x\in I} \big(\varphi(x) + \varphi'(x+)(\mathbb{E}[X]-x)\big) = \varphi(\mathbb{E}[X]),$$

also unmittelbar die Jensensche Ungleichung. □

3.22 Beispiel. Für $q > p > 0$ ist $\varphi(x) = |x|^{q/p}$ konvex, und die Jensensche Ungleichung liefert ebenfalls $\mathbb{E}[|X|^p] \leqslant \mathbb{E}[|X|^q]^{p/q}$ für $X \in \mathscr{L}^p \cap \mathscr{L}^q$. Ebenso gilt stets $\exp(\alpha\,\mathbb{E}[X]) \leqslant \mathbb{E}[\exp(\alpha X)]$ für $\alpha \in \mathbb{R}$ mit $\mathbb{E}[\exp(\alpha X)] < \infty$.

3.2 Varianz, Kovarianz und Korrelation

3.23 Definition. Für eine Zufallsvariable $X \in \mathscr{L}^2$ bezeichnet

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

die Varianz von X. $\sigma(X) := \sqrt{\operatorname{Var}(X)}$ heißt Standardabweichung von X.

3.24 Bemerkung. Die Varianz von X gibt gemäß Satz 3.15 den kleinsten *mittleren quadratischen Fehler* um einen deterministischen Wert an. Sie wird daher auch manchmal als Streuung der Zufallsvariablen X bezeichnet. Wir werden dies insbesondere bei der Tschebyschew-Ungleichung später noch klarer sehen.

3.25 Satz (Eigenschaften der Varianz)**.** *Für $X, Y \in \mathscr{L}^2$ gilt:*

(a) $\operatorname{Var}(X) = 0 \iff P(X = \mathbb{E}[X]) = 1$*;*

(b) $\forall a, b \in \mathbb{R}: \operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$*;*

(c) $\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$*;*

(d) $\operatorname{Var}(X + Y) \leqslant 2\operatorname{Var}(X) + 2\operatorname{Var}(Y)$*;*

(e) falls X, Y unabhängig sind, so gilt $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.

Beweis. Nach Satz 3.10(b) folgt aus $0 = \operatorname{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$, dass $P((X - \mathbb{E}[X])^2 = 0) = 1$ und somit (a) gilt. Für (b) berechne mittels Linearität des Erwartungswerts

$$\operatorname{Var}(aX + b) = \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2] = \mathbb{E}[a^2(X - \mathbb{E}[X])^2] = a^2 \operatorname{Var}(X).$$

(c) folgt aus der Bias-Varianz-Zerlegung mit $x = 0$. Mit der Ungleichung $(A + B)^2 \leqslant 2A^2 + 2B^2$ für $A, B \in \mathbb{R}$ und der Monotonie des Erwartungswerts erhalten wir (d):

$$\begin{aligned}\operatorname{Var}(X + Y) &= \mathbb{E}[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^2] \\ &\leqslant \mathbb{E}\left[2(X - \mathbb{E}[X])^2 + 2(Y - \mathbb{E}[Y])^2\right] = 2\operatorname{Var}(X) + 2\operatorname{Var}(Y).\end{aligned}$$

Beachte dazu: $X, Y \in \mathscr{L}^2 \Rightarrow X + Y \in \mathscr{L}^2$ folgt mit demselben Argument, nämlich $\mathbb{E}[(X + Y)^2] \leqslant 2\,\mathbb{E}[X^2] + 2\,\mathbb{E}[Y^2] < \infty$. Schließlich folgt (e) wegen

$$\begin{aligned}\operatorname{Var}(X + Y) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2 + 2(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) + (Y - \mathbb{E}[Y])^2\right] \\ &= \operatorname{Var}(X) + 2\big(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\big) + \operatorname{Var}(Y) \\ &= \operatorname{Var}(X) + \operatorname{Var}(Y),\end{aligned}$$

wobei wir die Multiplikativität des Erwartungswerts $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ unter Unabhängigkeit benutzt haben. □

3.26 Bemerkung. Eine der wichtigsten Aufgaben der Statistik ist es, auf Grund der Beobachtung gewisser Einflussgrößen (*Kovariablen, Regressoren*) eine damit zusammenhängende Zielgröße vorherzusagen. Ein Anwendungsbeispiel

ist die Vorhersage des Ozongehalts in der Luft (Sommersmog) auf Grund meteorologischer Messwerte und weiterer Einflussgrößen wie Verkehrsfluss und Industrieproduktion. In erster Näherung wird häufig ein linearer Zusammenhang angenommen, und das folgende Resultat ist Grundlage der gesamten *linearen Regressionsanalyse*. Die Aussage lässt sich am einfachsten mit den Begriffen von Kovarianz und Korrelation formulieren.

3.27 Definition. Für Zufallsvariablen $X, Y \in \mathscr{L}^2$ definiert

$$\operatorname{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

die Kovarianz zwischen X und Y. Falls $\sigma(X) > 0$ und $\sigma(Y) > 0$ gilt, heißt

$$\rho(X,Y) := \frac{\operatorname{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

die Korrelation zwischen X und Y. Falls $\operatorname{Cov}(X,Y) = 0$ gilt, heißen X und Y unkorreliert.

3.28 Satz (Beste lineare Vorhersage)**.** *Es seien X, Y Zufallsvariablen in $\mathscr{L}^2$ sowie*

$$L_X := \{aX + b \,|\, a, b \in \mathbb{R}\} \subseteq \mathscr{L}^2$$

die Menge der auf linear-affinen Funktionen von X basierenden Zufallsvariablen. Dann nimmt der mittlere quadratische Fehler

$$\varphi : L_X \to [0,\infty), \quad \varphi(Z) := \mathbb{E}[(Y - Z)^2]$$

sein Minimum bei $Z = a^*X + b^*$ an mit

$$a^* = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}, \quad b^* = \mathbb{E}[Y] - a^* \,\mathbb{E}[X]$$

(a^* beliebig falls $\operatorname{Var}(X) = 0$). Für $\operatorname{Var}(X), \operatorname{Var}(Y) > 0$ gilt

$$\varphi(a^*X + b^*) = \operatorname{Var}(Y)(1 - \rho^2(X,Y)).$$

Ende 12. Vorlesung

Beweis. Minimiert man das quadratische Funktional (gemäß Bias-Varianz-Zerlegung)

$$\mathbb{E}[(Y - aX - b)^2] = \operatorname{Var}(Y - aX) + (\mathbb{E}[Y - aX] - b)^2$$

zunächst in b, so ergibt sich direkt $b^* = \mathbb{E}[Y] - a^* \,\mathbb{E}[X]$. Andererseits ist

$$\operatorname{Var}(Y - aX) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] - 2a\,\mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])] + a^2\,\mathbb{E}[(X - \mathbb{E}[X])^2],$$

was durch

$$a^* = \frac{\mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])]}{\mathbb{E}[(X - \mathbb{E}[X])^2]} = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}$$

minimiert wird, wenn der Nenner nicht null ist. Ist hingegen $\mathrm{Var}(X) = 0$, so ist $X = \mathbb{E}[X]$ P-fast sicher und somit $\mathrm{Var}(Y - aX) = \mathrm{Var}(Y)$, so dass a^* beliebig gewählt werden kann. Wegen (setze a^* in obige Formel ein)

$$\varphi(a^*X + b^*) = \mathrm{Var}(Y - a^*X) = \mathrm{Var}(Y) - \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X)} = \mathrm{Var}(Y)(1 - \rho^2(X,Y))$$

folgt auch die Darstellung des Minimalwerts durch die Korrelation $\rho(X,Y)$. □

3.29 Bemerkung. Sind X und Y unkorreliert im Satz, so kann Y nicht besser als durch seinen Erwartungswert $b^* = \mathbb{E}[Y]$ vorhergesagt werden. Je stärker X und Y korrelieren, also je größer $|\rho(X,Y)|$ ist, desto besser ist die Vorhersage. Im Extremfall $\rho(X,Y) = \pm 1$ gilt $Y = a^*X + b^*$ P-fast sicher. Aus dem Satz kann man viele Eigenschaften von Kovarianz und Korrelation direkt folgern.

3.30 Satz (Eigenschaften von Kovarianz und Korrelation)**.** *Für $X, Y, Z \in \mathscr{L}^2$ gilt:*

(a) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$, $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$*;*

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

(c) $\forall a, b \in \mathbb{R}:\ \mathrm{Cov}(aX + b, Y) = a\,\mathrm{Cov}(X,Y)$*;*

(d) $\mathrm{Cov}(X+Y, Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$*;*

(e) X, Y *unabhängig* $\Rightarrow$ X, Y *unkorreliert;*

(f) $|\mathrm{Cov}(X,Y)| \leqslant \sigma(X)\sigma(Y)$ *und* $\rho(X,Y) \in [-1,+1]$.

Beweis. Übung! □

3.31 Beispiele.

(a) Eine $\mathrm{Bin}(n,p)$-verteilte Zufallsvariable X ergibt sich als $X = \sum_{i=1}^n X_i$ mit einem Bernoullischema (X_i), d.h. $(X_i)_{i=1,\dots,n}$ unabhängig mit $P(X_i = 1) = p$, $P(X_i = 0) = 1-p$. Damit folgt $X \in \mathscr{L}^2$ und $\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = np$, $\mathrm{Var}(X) = \sum_{i=1}^n \mathrm{Var}(X_i) = np(1-p)$. Da Erwartungswert und Varianz nur von der Verteilung abhängen, sagt man, dass die $\mathrm{Bin}(n,p)$-Verteilung Erwartungswert np und Varianz $np(1-p)$ besitzt. Im Fall $p \in \{0,1\}$ gilt also $\mathrm{Var}(X) = 0$ sowie stets $\mathrm{Var}(X) \leqslant n/4$. Die *relative Häufigkeit* von Erfolgen $A = X/n$ erfüllt $\mathbb{E}[A] = p$ (*erwartungstreu* für die Erfolgswahrscheinlichkeit) und $\mathrm{Var}(A) = \frac{p(1-p)}{n}$. Die Streuung von A nimmt also mit wachsendem n ab.

(b) Für eine $N(\mu,\sigma^2)$-verteilte Zufallsvariable X mit $\mu \in \mathbb{R}$, $\sigma > 0$ gilt nach dem Dichtetransformationssatz, dass $Z = (X-\mu)/\sigma$ $N(0,1)$-verteilt ist. Nun sind $X, Z \in \mathscr{L}^2$ und wir schließen $\mathrm{Var}(X) = \mathrm{Var}(\sigma Z + \mu) = \sigma^2\,\mathrm{Var}(Z) = \sigma^2$. Beachte dazu, dass mit partieller Integration folgt

$$\mathrm{Var}(Z) = \mathbb{E}[Z^2] = \int_{-\infty}^{\infty} z \tfrac{1}{\sqrt{2\pi}} z e^{-z^2/2} dz = -\int_{-\infty}^{\infty} 1(-\tfrac{1}{\sqrt{2\pi}} e^{-z^2/2}) dz = 1.$$

(c) Bezeichnen X_1 und X_2 die Augenzahlen beim Wurf zweier Würfel, so ist $S = X_1 + X_2$ die Augensumme und $D = X_1 - X_2$ die Augendifferenz. Es gilt

$$\mathrm{Cov}(S, D) = \mathrm{Var}(X_1) + \mathrm{Cov}(X_2, X_1) - \mathrm{Cov}(X_1, X_2) - \mathrm{Var}(X_2) = 0,$$

und S und D sind unkorreliert (gilt allgemein für $X_1, X_2 \in \mathscr{L}^2$ mit $\mathrm{Var}(X_1) = \mathrm{Var}(X_2)$). Allerdings sind S und D *nicht* unabhängig; denn es gilt beispielsweise $P(S = 2, D = 5) = 0$, aber $P(S = 2)P(D = 5) > 0$. Die beste (nicht nur lineare) Vorhersage von D gegeben S ist gerade konstant gleich null, da D unter jeder Bedingung $\{S = k\}$ symmetrisch um 0 verteilt ist: $P(D = m \,|\, S = k) = P(D = -m \,|\, S = k)$.

4 Grenzwertsätze

4.1 Gesetze der großen Zahlen

4.1 Bemerkung. Wir wollen der Intuition, dass sich relative Häufigkeiten und Mittelwerte bei großen Stichprobenumfängen entsprechenden Wahrscheinlichkeiten und Erwartungswerten annähern, eine mathematische Grundlage geben. Wichtig sind zunächst einfache Ungleichungen für Abweichungswahrscheinlichkeiten.

4.2 Satz (Allgemeine Markov-Ungleichung)**.** *Es sei X eine reellwertige Zufallsvariable und $\varphi : \mathbb{R} \to [0, \infty)$ monoton wachsend. Dann gilt für jedes $K \in \mathbb{R}$ mit $\varphi(K) > 0$:*

$$P(X \geqslant K) \leqslant \frac{\mathbb{E}[\varphi(X)]}{\varphi(K)}.$$

Beweis. Da φ monoton wächst, gilt $X(\omega) \geqslant K \Rightarrow \varphi(X(\omega)) \geqslant \varphi(K)$ und daher $\varphi(X) \geqslant \varphi(K)\mathbf{1}(X \geqslant K)$. Aus der Monotonie des Erwartungswerts folgt somit

$$\mathbb{E}[\varphi(X)] \geqslant \mathbb{E}[\varphi(K)\mathbf{1}(X \geqslant K)] = \varphi(K)P(X \geqslant K).$$

Division durch $\varphi(K)$ liefert die Behauptung. Beachte dabei: jede monotone Funktion φ ist Borel-messbar und $\mathbb{E}[\varphi(X)] \in [0, \infty]$ existiert wegen $\varphi(X) \geqslant 0$. □

4.3 Beispiel. Für $X \in \mathscr{L}^p$ gilt $P(|X| \geqslant K) \leqslant \mathbb{E}[|X|^p]K^{-p}$ (betrachte $Y = |X|$ und $\varphi(y) = (y_+)^p$). Der Fall $p = 1$ ist gerade die spezielle Form der Markov-Ungleichung.

4.4 Korollar (Tschebyschev-Ungleichung)**.** *Ist X eine Zufallsvariable in $\mathscr{L}^2$, so gilt für jedes $K > 0$*

$$P(|X - \mathbb{E}[X]| \geqslant K) \leqslant \frac{\mathrm{Var}(X)}{K^2}.$$

Beweis. Wende die Markov-Ungleichung auf $Y = |X - \mathbb{E}[X]|$ und $\varphi(y) = (y_+)^2$ an:

$$P(|X - \mathbb{E}[X]| \geqslant K) \leqslant \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{K^2} = \frac{\mathrm{Var}(X)}{K^2}.$$

□

4.5 Beispiel. Eine faire Münze werde n-mal geworfen und $\hat{p}$ bezeichne die relative Häufigkeit der Würfe mit 'Kopf'. Also ist $\hat{p} = S_n/n$ mit $S_n \sim \mathrm{Bin}(n, 1/2)$. Wir erhalten $\mathbb{E}[\hat{p}] = \mathbb{E}[S_n]/n = 1/2$, $\mathrm{Var}(\hat{p}) = \mathrm{Var}(S_n)/n^2 = 1/(4n)$. Die Tschebyschev-Ungleichung gibt die Abschätzung

$$P(|\hat{p} - 1/2| > \varepsilon) \leqslant \frac{1}{4n\varepsilon^2}.$$

Die rechte Seite ist viel kleiner als Eins nur im Fall $\varepsilon \gg \frac{1}{\sqrt{n}}$. Im Fall $n = 1000$ und $\varepsilon = 0,05$ ergibt sich so $P(|\hat{p} - 1/2| \geqslant 0,05) \leqslant 0,1$. Allgemein gilt: je größer n ist, desto kleiner sind die Abweichungswahrscheinlichkeiten. Allerdings ist die Tschebyschev-Ungleichung in diesem Fall sehr pessimistisch (nicht scharf), wie wir noch sehen werden.

Ende 13. Vorlesung

4.6 Satz (schwaches Gesetz der großen Zahlen)**.** *Es sei $(X_i)_{i\geqslant 1}$ eine Folge unkorrelierter Zufallsvariablen in $\mathscr{L}^2$ mit demselben Erwartungswert $\mu \in \mathbb{R}$ und $\sup_i \mathrm{Var}(X_i) < \infty$. Dann erfüllt das arithmetische Mittel*

$$A_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

für jedes $\varepsilon > 0$

$$\lim_{n\to\infty} P(|A_n - \mu| > \varepsilon) = 0.$$

Beweis. Wegen Linearität gilt $\mathbb{E}[A_n] = \mu$ und wegen der Unkorreliertheit

$$\begin{aligned}
\mathrm{Var}(A_n) &= \frac{1}{n^2} \mathrm{Var}(X_1 + \cdots + X_n) \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) \\
&\leqslant \frac{1}{n} \sup_{i\geqslant 1} \mathrm{Var}(X_i).
\end{aligned}$$

Mit der Tschebyschev-Ungleichung folgt also

$$P(|A_n - \mu| > \varepsilon) \leqslant \frac{\sup_i \mathrm{Var}(X_i)}{n\varepsilon^2} \to 0$$

für $n \to \infty$ und jedes feste $\varepsilon > 0$. □

4.7 Bemerkung. Das schwache Gesetz wird oft für unabhängige Zufallsvariablen formuliert, aber der Beweis zeigt, dass es ausreicht, (paarweise) Unkorreliertheit zu fordern. Es gibt auch ein schwaches Gesetz der großen Zahlen unter der schwächeren Annahme $X_i \in \mathscr{L}^1$, vergleiche Satz 5.7 in Georgii.

4.8 Korollar. *(Weierstraßscher Approximationssatz) Zur stetigen Funktion* $f : [0,1] \to \mathbb{R}$ *definiere das zugehörige* Bernstein-Polynom *n-ten Grades*

$$f_n(x) := \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} x^k (1-x)^{n-k}, \quad x \in [0,1].$$

Dann gilt $\lim_{n\to\infty} \|f - f_n\|_\infty = 0$ *mit* $\|g\|_\infty := \max_{x\in[0,1]} |g(x)|$. *Insbesondere liegen also die Polynome dicht im Raum der stetigen Funktionen auf* $[0,1]$ *bezüglich der Maximumsnorm.*

Beweis. Es seien X eine $\mathrm{Bin}(n,p)$-verteilte Zufallsvariablen mit $p \in [0,1]$. Dann gilt

$$\mathbb{E}\Big[f\Big(\frac{X}{n}\Big)\Big] = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} p^k (1-p)^{n-k} = f_n(p).$$

Da f auf dem Kompaktum $[0,1]$ gleichmäßig stetig ist, existiert zu jedem $\varepsilon > 0$ ein $\delta > 0$, so dass $|x-y| \leqslant \delta \Rightarrow |f(x) - f(y)| \leqslant \varepsilon$. Wir erhalten

$$\begin{aligned}
|f_n(p) - f(p)| &\leqslant \mathbb{E}\Big[\Big|f\Big(\frac{X}{n}\Big) - f(p)\Big|\Big] \\
&\leqslant \mathbb{E}\Big[\varepsilon \mathbf{1}\Big(\Big|\frac{X}{n} - p\Big| \leqslant \delta\Big) + 2\|f\|_\infty \mathbf{1}\Big(\Big|\frac{X}{n} - p\Big| > \delta\Big)\Big] \\
&\leqslant \varepsilon + 2\|f\|_\infty P\Big(\Big|\frac{X}{n} - p\Big| > \delta\Big) \\
&\leqslant \varepsilon + \frac{2p(1-p)\|f\|_\infty}{n\delta^2},
\end{aligned}$$

wobei wir zuletzt die Tschebyschev-Ungleichung verwendet haben. Wegen $2p(1-p) \leqslant 1/2$ für $p \in [0,1]$ folgt also

$$\limsup_{n\to\infty} \sup_{p\in[0,1]} |f_n(p) - f(p)| \leqslant \varepsilon.$$

Da $\varepsilon > 0$ beliebig war, folgt die Behauptung. □

4.9 Definition. Es seien $(X_n)_{n\geqslant 1}$ und X reellwertige Zufallsvariablen auf demselben Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, P)$. Man sagt, dass X_n stochastisch (oder auch in P-Wahrscheinlichkeit) gegen X konvergiert für $n \to \infty$, falls für alle $\varepsilon > 0$ gilt

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

Notation: $X_n \xrightarrow{P} X$.

Man sagt, dass X_n P-fast sicher gegen X konvergiert, kurz $X_n \to X$ P-f.s., falls

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

4.10 Beispiel. Im schwachen Gesetz der großen Zahlen gilt $A_n \xrightarrow{P} \mu$.

4.11 Satz. *Fast sichere Konvergenz impliziert stochastische Konvergenz, aber nicht umgekehrt.*

Beweis. Übung! □

4.12 Bemerkung. Man kann zeigen, dass $d_0(X,Y) := \mathbb{E}[|X-Y| \wedge 1]$ für Zufallsvariablen X, Y auf einem Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, P)$ eine Metrik definiert, wenn man Zufallsvariablen identifiziert, die P-f.s. gleich sind. Es gilt $d_0(X_n, X) \to 0 \iff X_n \xrightarrow{P} X$. Fast sichere Konvergenz kann hingegen im Allgemeinen nicht metrisiert werden (analog zu punktweiser Konvergenz von Funktionenfolgen).

4.13 Satz. *(starkes Gesetz der großen Zahlen) Es sei $(X_i)_{i\geqslant 1}$ eine Folge unkorrelierter Zufallsvariablen in $\mathscr{L}^2$ mit demselben Erwartungswert $\mu \in \mathbb{R}$ und $\sup_i \operatorname{Var}(X_i) < \infty$. Dann konvergiert das arithmetische Mittel $A_n = \frac{1}{n}\sum_{i=1}^n X_i$ fast sicher gegen μ.*

Beweis. Wir verwenden eine allgemeine Strategie, um fast sichere Konvergenz zu zeigen: zunächst wird dies entlang einer Teilfolge mittels Lemma von Borel-Cantelli nachgewiesen. Dann wird der Abstand eines allgemeinen Folgenglieds zur Teilfolge abgeschätzt. O.B.d.A. sei $\mu = 0$, sonst betrachte $\tilde{X}_i = X_i - \mu$, so dass $\tilde{A}_n = A_n - \mu$ gegen null konvergiert.

Nach der Tschebyschev-Ungleichung gilt für alle $\varepsilon > 0$ (beachte $\mu = 0$)

$$\sum_{n\geqslant 1} P(|A_{n^2}| \geqslant \varepsilon) \leqslant \sum_{n\geqslant 1} \frac{\sup_i \operatorname{Var}(X_i)}{n^2\varepsilon^2} < \infty.$$

Das Lemma von Borel-Cantelli zeigt daher

$$P(|A_{n^2}| \geqslant \varepsilon \text{ für unendlich viele } n) = 0.$$

Betrachtet man die Vereinigung der Ereignisse für alle rationalen $\varepsilon > 0$, so folgt

$$P\Big(\lim_{n\to\infty} A_{n^2} \neq 0\Big) = P\Big(\bigcup_{\varepsilon>0,\varepsilon\in\mathbb{Q}} \{|A_{n^2}| \geqslant \varepsilon \text{ für unendlich viele } n\}\Big) = 0.$$

Es gilt also $A_{n^2} \to 0$ P-fast sicher.

Zu jedem $m \in \mathbb{N}$ wähle $n(m) \in \mathbb{N}$ mit $n(m)^2 \leqslant m < (n(m)+1)^2$. Wiederum mit der Tschebyschev-Ungleichung folgt

$$\begin{aligned} P\Big(\Big|\sum_{i=1}^m X_i - \sum_{i=1}^{n(m)^2} X_i\Big| \geqslant \varepsilon n(m)^2\Big) &\leqslant \frac{\operatorname{Var}(\sum_{i=n(m)^2+1}^m X_i)}{n(m)^4\varepsilon^2} \\ &\leqslant \frac{(m-n(m)^2)\sup_i \operatorname{Var}(X_i)}{n(m)^4\varepsilon^2}. \end{aligned}$$

Diese Wahrscheinlichkeiten sind nun summierbar:

$$\sum_{m\geqslant 1} P\Big(\Big|\sum_{i=1}^m X_i - \sum_{i=1}^{n(m)^2} X_i\Big| \geqslant \varepsilon n(m)^2\Big) \leqslant \frac{\sup_i \operatorname{Var}(X_i)}{\varepsilon^2} \sum_{n\geqslant 1} \sum_{m=n^2}^{(n+1)^2-1} \frac{m-n^2}{n^4} < \infty$$

wegen $\sum_{m=n^2}^{(n+1)^2-1}(m-n^2) \leqslant (2n+1)2n$. Dasselbe Borel-Cantelli-Argument zeigt daher

$$\lim_{m\to\infty} \frac{1}{n(m)^2}\Big|\sum_{i=1}^{m} X_i - \sum_{i=1}^{n(m)^2} X_i\Big| = 0 \quad P\text{-f.s.}$$

Mit dem ersten Teilfolgenresultat ergibt sich außerhalb einer Nullmenge (Vereinigung der beiden Nullmengen, wo keine Konvergenz vorliegt) $\lim_{m\to\infty} \frac{1}{n(m)^2}\sum_{i=1}^{m} X_i = 0$. Auf Grund von $m \geqslant n(m)^2$ impliziert dies $\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} X_i = 0$ P-fast sicher, also die Behauptung. □

4.14 Bemerkung. Auch für das starke Gesetz der großen Zahlen reicht es, $X_i \in \mathscr{L}^1$ zu fordern, wenn man paarweise Unabhängigkeit statt Unkorreliertheit annimmt, siehe Satz 5.16 in Georgii. Man kann auch zeigen, dass für $X_i \geqslant 0$ mit $\mathbb{E}[X_i] = \infty$ die Konvergenz $A_n \to \infty$ P-f.s. gilt.

4.15 Definition. Identifiziert man $X, Y \in \mathscr{L}^p(\Omega, \mathscr{F}, P)$ (Zusammenfassung in einer Äquivalenzklasse), wenn $X = Y$ P-fast sicher, d.h. $P(X = Y) = 1$, gilt, so erhält man den Vektorraum $L^p(\Omega, \mathscr{F}, P)$. Für $p \geqslant 1$ wird $L^p(\Omega, \mathscr{F}, P)$ mit der Norm $\|X\|_{L^p} = \mathbb{E}[|X|^p]^{1/p}$ zum Banachraum und mit dem Skalarprodukt $\langle X, Y\rangle = \mathbb{E}[XY]$ wird $L^2(\Omega, \mathscr{F}, P)$ zum Hilbertraum (Beweis in Analysis!). Für eine Folge (X_n) in $\mathscr{L}^p(\Omega, \mathscr{F}, P)$, $p > 0$, und ein $X \in \mathscr{L}^p(\Omega, \mathscr{F}, P)$ sagen wir, dass X_n gegen X in L^p konvergiert, falls $\mathbb{E}[|X_n - X|^p] \to 0$ für $n \to \infty$ gilt. Notation: $X_n \xrightarrow{L^p} X$.

4.16 Beispiel. Im schwachen Gesetz der großen Zahlen gilt $A_n \xrightarrow{L^2} \mu$ wegen

$$\mathbb{E}[(A_n - \mu)^2] = \mathrm{Var}(A_n) = \frac{1}{n^2}\sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j) \leqslant n^{-1} \sup_i \mathrm{Var}(X_i) \to 0.$$

Ende 14. Vorlesung

4.17 Satz. *Für reellwertige Zufallsvariablen X_n, X auf einem gemeinsamen Wahrscheinlichkeitsraum gelten folgende Implikationen:*

(a) Konvergert (X_n) gegen X in L^p für ein $p > 0$, so auch stochastisch.

(b) Konvergiert (X_n) gegen X stochastisch, so existiert eine Teilfolge (n_k), so dass $X_{n_k} \to X$ P-fast sicher für $k \to \infty$.

(c) Konvergiert (X_n) gegen X P-fast sicher und ist $\mathbb{E}[\sup_n |X_n|^p] < \infty$ für ein $p > 0$, so konvergiert X_n gegen X in L^p.

(d) In (c) reicht es, stochastische Konvergenz $X_n \xrightarrow{P} X$ statt fast sicherer Konvergenz zu fordern.

Beweis.

(a) Nach der Markovungleichung gilt für alle $\varepsilon > 0$

$$P(|X_n - X| \geqslant \varepsilon) \leqslant \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p}$$

und nach Voraussetzung konvergiert die rechte Seite gegen Null.

(b) Sei $\varepsilon_k \downarrow 0$ eine Nullfolge. Nach Voraussetzung existieren dann $n_k \in \mathbb{N}$ für $k \geqslant 1$ mit $P(|X_n - X| > \varepsilon_k) \leqslant 2^{-k}$ für alle $n \geqslant n_k$. Wegen $\sum_{k\geqslant 1} P(|X_{n_k} - X| > \varepsilon_k) < \infty$ gilt nach dem Lemma von Borel-Cantelli $P(|X_{n_k} - X| > \varepsilon_k \text{ für unendlich viele } k \geqslant 1) = 0$. Mit Wahrscheinlichkeit 1 existiert daher ein (zufälliges) $k_0 \in \mathbb{N}$ mit $\forall k \geqslant k_0 : |X_{n_k} - X| \leqslant \varepsilon_k$, was wegen $\varepsilon_k \to 0$ gerade $X_{n_k} \to X$ P-fast sicher zeigt.

(c) Wegen $X_n \to X$ P-fast sicher gilt $P(|X| \leqslant \sup_n |X_n|) = 1$. Daher gilt auch

$$\mathbb{E}\left[\sup_n |X_n - X|^p\right] \leqslant \mathbb{E}\left[\sup_n (|X_n| + |X|)^p\right] \leqslant 2^p\, \mathbb{E}\left[\sup_n |X_n|^p\right] < \infty.$$

Mit dominierter Konvergenz (Satz von Lebesgue) folgt daher $\mathbb{E}[|X_n - X|^p] \to 0$, also $X_n \xrightarrow{L^p} X$.

(d) Angenommen, X_n konvergiert nicht gegen X in L^p. Dann existiert eine Teilfolge $(n_k)_{k\geqslant 0}$ mit $\liminf_{k\to\infty} \mathbb{E}[|X_{n_k} - X|^p] > 0$. Da auch $X_{n_k} \xrightarrow{P} X$ gilt, existiert nach (b) eine Teilteilfolge $(n_{k_\ell})_{\ell\geqslant 0}$ mit $X_{n_{k_\ell}} \xrightarrow{P\text{-f.s.}} X$, so dass nach (c) $X_{n_{k_\ell}} \xrightarrow{L^p} X$, also $\mathbb{E}[|X_{n_{k_\ell}} - X|^p] \to 0$. Dies zeigt $\liminf_{k\to\infty} \mathbb{E}[|X_{n_k} - X|^p] = 0$. Widerspruch!

□

4.18 Bemerkung. Wir erhalten insbesondere folgende Implikationsreihe:

$$X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{P} X \overset{\exists (n_k)}{\Rightarrow} X_{n_k} \xrightarrow{P\text{-f.s.}} X \overset{\mathbb{E}[\sup_k |X_{n_k}|^p] < \infty}{\Rightarrow} X_{n_k} \xrightarrow{L^p} X.$$

Wir sehen, dass stochastische Konvergenz die schwächste der betrachteten Konvergenzarten ist. L^p-Konvergenz und fast sichere Konvergenzen implizieren einander nicht ohne Weiteres, sondern nur mittels Aussagen (b) und (c). Es gibt einfache Beispiele mit $X_n \to 0$ P-fast sicher und $\mathbb{E}[X_n] = 1$ für alle n (Übung!). In (c) und (d) kann $\mathbb{E}[\sup_n |X_n|^p] < \infty$ mit Hilfe der sogenannten gleichgradigen Integrierbarkeit abgeschwächt werden (Stochastik II).

4.19 Beispiel. (Harmonische Reihe mit zufälligen Vorzeichen, Beispiel 2.40) Betrachte $S_n = \sum_{k=1}^n \varepsilon_k \frac{1}{k}$ mit (ε_k) unabhängig und $P(\varepsilon_k = 1) = P(\varepsilon_k = -1) = 1/2$. Dann gilt $\mathbb{E}[S_n] = 0$, $\mathrm{Var}(S_n) = \sum_{k=1}^n \frac{1}{k^2}$. Wegen $\mathrm{Var}(S_n - S_m) \leqslant \sum_{k\geqslant m} \frac{1}{k^2}$ für $n > m$ und $\lim_{m\to\infty} \sum_{k\geqslant m} \frac{1}{k^2} = 0$ bildet (S_n) eine Cauchyfolge in L^2. Es gilt also $S_n \to S_\infty$ in L^2 und damit auch stochastisch für ein $S_\infty \in L^2$.

Wir können sogar mit der vom starken Gesetz bekannten Strategie fast sichere Konvergenz zeigen. Dazu bedeute $A_n \lesssim B_n$, dass $A_n \leqslant CB_n$ für eine Konstante $C > 0$. Nach der Tschebyshev-Ungleichung gilt

$$P(|S_n - S_\infty| > \varepsilon) \leqslant \varepsilon^{-2} \mathrm{Var}(S_\infty - S_n) = \varepsilon^{-2} \sum_{k>n} k^{-2} \lesssim n^{-1}.$$

Dies zeigt, dass $\sum_{n\geqslant 1} P(|S_{n^2} - S_\infty| > \varepsilon) < \infty$ und mit dem üblichen Borel-Cantelli-Argument $S_{n^2} \to S_\infty$ P-fast sicher. Wähle $n(m) \in \mathbb{N}$ mit $n(m)^2 \leqslant m < (n(m)+1)^2$. Dann gilt wiederum mit Tschebyschev-Ungleichung

$$P(|S_m - S_{n(m)^2}| > \varepsilon) \leqslant \varepsilon^{-2} \sum_{k=n(m)^2+1}^{(n(m)+1)^2} k^{-2} \lesssim n(m)n(m)^{-4} = n(m)^{-3},$$

wobei wir $(n(m)+1)^2 - n(m)^2 - 1 = 2n(m)$ verwendet haben. Wir schließen

$$\sum_{m\geqslant 1} P(|S_m - S_{n(m)^2}| > \varepsilon) \lesssim \sum_{n\geqslant 1} nn^{-3} < \infty.$$

Mit dem Borel-Cantelli-Argument folgt dann $S_m - S_{n(m)^2} \to 0$ P-fast sicher und daher auch $S_m \to S_\infty$ P-fast sicher.

4.2 Konvergenz in Verteilung

4.20 Bemerkung. Wir benötigen noch einen weiteren Konvergenzbegriff, der aber grundverschieden von den bisherigen ist, weil er nur die Verteilungen von Zufallsvariablen betrachtet. Diese müssen nicht einmal auf demselben Wahrscheinlichkeitsraum definiert sein. Die im folgenden definierte *schwache Konvergenz* entspricht auf kompakten Mengen in $\mathbb{R}^d$ der sogenannten *schwach*-Konvergenz* der Funktionalanalysis.

4.21 Definition. Die $\mathbb{R}^d$-wertigen Zufallsvektoren $(X_n)_{n\geqslant 1}$ konvergieren in Verteilung gegen den $\mathbb{R}^d$-wertigen Zufallsvektor X, Notation $X_n \xrightarrow{d} X$, falls für jede stetige beschränkte Funktion $\varphi : \mathbb{R}^d \to \mathbb{R}$ gilt

$$\lim_{n\to\infty} \mathbb{E}[\varphi(X_n)] = \mathbb{E}[\varphi(X)].$$

Wahrscheinlichkeitsmaße $(P_n)_{n\geqslant 1}$ auf $(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d})$ konvergieren schwach gegen ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d})$, Notation $P_n \xrightarrow{w} P$, falls für jede stetige beschränkte Funktion $\varphi : \mathbb{R}^d \to \mathbb{R}$ gilt

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} \varphi(x)\, P_n(dx) = \int_{\mathbb{R}^d} \varphi(x)\, P(dx).$$

Man definiert Konvergenz in Verteilung mit einem Wahrscheinlichkeitsmaß P als Grenzwert allgemein durch $X_n \xrightarrow{d} P :\iff P^{X_n} \xrightarrow{w} P$.

4.22 Beispiel. Sind X ein Zufallsvektor und $a_n \in \mathbb{R}$, $b_n \in \mathbb{R}^d$ mit $a_n \to a$, $b_n \to b$, so gilt $a_nX + b_n \xrightarrow{d} aX + b$: Stetigkeit von φ impliziert $\varphi(a_nX(\omega) + b_n) \to \varphi(aX(\omega) + b)$ für alle $\omega \in \Omega$. Nun ist $\|\varphi\|_\infty$ eine integrierbare Majorante und dominierte Konvergenz zeigt $\mathbb{E}[\varphi(a_nX + b_n)] \to \mathbb{E}[\varphi(aX + b)]$. Beachte, dass $P_n \xrightarrow{w} P$ nicht impliziert $P_n(B) \to P(B)$ für alle $B \in \mathfrak{B}_{\mathbb{R}^d}$: setze $a_n = a = 0$ oben und schließe $\delta_{b_n} \xrightarrow{w} \delta_b$, während für $b_n \neq b$ gilt $\delta_{b_n}(\{b\}) = 0 \neq 1 = \delta_b(\{b\})$.

4.23 Satz. *Konvergiert X_n gegen X stochastisch, so auch in Verteilung, das heißt $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$.*

Beweis. Wir verwenden ein Teilteilfolgenargument wie im Beweis von Satz 4.17(d). Falls $X_n \xrightarrow{d} X$ nicht gilt, so existiert eine stetige beschränkte Funktion φ sowie eine Teilfolge (n_k) mit $\liminf_{k\to\infty}|\mathbb{E}[\varphi(X_{n_k})] - \mathbb{E}[\varphi(X)]| > 0$. Für eine Teilteilfolge (n_{k_ℓ}) gilt aber $X_{n_{k_l}} \xrightarrow{P\text{-f.s.}} X$ und wegen Stetigkeit auch $\varphi(X_{n_{k_l}}) \xrightarrow{P\text{-f.s.}} \varphi(X)$. Wegen $\|\varphi\|_\infty < \infty$ folgt mit dominierter Konvergenz $\mathbb{E}[\varphi(X_{n_{k_l}})] \to \mathbb{E}[\varphi(X)]$ im Widerspruch zu $\liminf_{k\to\infty}|\mathbb{E}[\varphi(X_{n_k})] - \mathbb{E}[\varphi(X)]| > 0$. Also muss $X_n \xrightarrow{d} X$ gelten. □

4.24 Beispiel. Die Umkehrung gilt nicht: für X $N(0,1)$-verteilt ist auch $Y = -X$ $N(0,1)$-verteilt, so dass $X_n \xrightarrow{d} Y$ für $X_n := X$ gilt, während $P(|X_n - Y| > \varepsilon) = P(2|X| > \varepsilon)$ für $\varepsilon > 0$ nicht gegen Null konvergiert.

4.25 Satz. *Für reellwertige Zufallsvariablen sind äquivalent:*

(a) $X_n \xrightarrow{d} X$

(b) Die Verteilungsfunktionen erfüllen $F^{X_n}(x) \to F^X(x)$ *für alle* $x \in \mathbb{R}$, *an denen* F^X *stetig ist (Stetigkeitspunkte von* F^X*).*

Ende 15. Vorlesung

Beweis. (a)⇒(b): Zu $x \in \mathbb{R}, \delta > 0$ wähle eine stetige Funktion φ mit $\mathbf{1}_{(-\infty,x]} \leqslant \varphi \leqslant \mathbf{1}_{(-\infty,x+\delta]}$ punktweise (z.B. linear zwischen den Punkten $(x,1)$ und $(x+\delta,0)$). Dann gilt

$$\limsup_{n\to\infty} F^{X_n}(x) \leqslant \limsup_{n\to\infty} \mathbb{E}[\varphi(X_n)] = \mathbb{E}[\varphi(X)] \leqslant F^X(x+\delta),$$
$$\liminf_{n\to\infty} F^{X_n}(x) \geqslant \liminf_{n\to\infty} \mathbb{E}[\varphi(X_n+\delta)] = \mathbb{E}[\varphi(X+\delta)] \geqslant F^X(x-\delta).$$

Mit $\delta \to 0$ und Rechtsstetigkeit von F^X folgt also $\limsup_{n\to\infty} F^{X_n}(x) \leqslant F^X(x)$ sowie an Stetigkeitsstellen x von F^X auch $\liminf_{n\to\infty} F^{X_n}(x) \geqslant F^X(x)$. Dies zeigt (b).

(b)⇒(a): Wir müssen für jede stetige beschränkte Funktion φ zeigen $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$. Wähle dazu für $\delta > 0$ Stetigkeitsstellen $x_0 < \cdots < x_K$ von F^X mit $F^X(x_0) < \delta$, $F^X(x_K) > 1-\delta$ und $\forall x \in [x_{k-1}, x_k] : |\varphi(x) - \varphi(x_k)| < \delta$, $k = 1, \ldots, K$. Dies ist möglich, weil die monotone Funktion F^X nur abzählbar viele Unstetigkeitsstellen besitzt sowie φ auf dem Kompaktum $[x_0, x_K]$ gleichmäßig stetig ist. Daher können wir abschätzen:

$$\mathbb{E}[\varphi(X_n)] = \mathbb{E}\left[\varphi(X_n)(\mathbf{1}_{(-\infty,x_0]}(X_n) + \mathbf{1}_{(x_K,\infty)}(X_n))\right] + \sum_{k=1}^{K} \mathbb{E}[\varphi(X_n)\mathbf{1}_{(x_{k-1},x_k]}(X_n)]$$
$$\leqslant \|\varphi\|_\infty \big(F^{X_n}(x_0) + 1 - F^{X_n}(x_K)\big) + \sum_{k=1}^{K} (\varphi(x_k) + \delta)\big(F^{X_n}(x_k) - F^{X_n}(x_{k-1})\big).$$

Wegen (b) und $F^X(x_0) + 1 - F^X(x_k) < 2\delta$ folgt

$$\limsup_{n\to\infty} \mathbb{E}[\varphi(X_n)] \leqslant 2\delta\|\varphi\|_\infty + \sum_{k=1}^{K} (\varphi(x_k) + \delta)\big(F^X(x_k) - F^X(x_{k-1})\big).$$

Analog erhalten wir

$$\mathbb{E}[\varphi(X)] \geqslant -2\delta\|\varphi\|_\infty + \sum_{k=1}^{K}(\varphi(x_k)-\delta)\big(F^X(x_k)-F^X(x_{k-1})\big),$$

so dass $\limsup_{n\to\infty}\mathbb{E}[\varphi(X_n)] \leqslant \mathbb{E}[\varphi(X)] + 4\delta\|\varphi\|_\infty + 2\delta$. Mit $\delta \downarrow 0$ folgt $\limsup_{n\to\infty}\mathbb{E}[\varphi(X_n)] \leqslant \mathbb{E}[\varphi(X)]$. Dasselbe Argument, angewendet auf $-\varphi$, zeigt $\liminf_{n\to\infty}\mathbb{E}[\varphi(X_n)] \geqslant \mathbb{E}[\varphi(X)]$. Also gilt $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$. □

4.26 Beispiele.

(a) Es gilt also $X_n \xrightarrow{d} N(0,1)$ genau dann, wenn $P(X_n \leqslant x) \to \Phi(x)$ für alle $x \in \mathbb{R}$ gilt mit $\Phi(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy$. Dann folgt auch $P(X_n \in [a,b]) \to \Phi(b)-\Phi(a)$ für alle $a<b$, was für $(a,b]$ sofort klar ist und sich mit $P(X_n = a) \leqslant P(X_n \in (a-\varepsilon, a]) \to \Phi(a)-\Phi(a-\varepsilon)$ und $\varepsilon \downarrow 0$ auch auf $[a,b]$ überträgt.

(b) Sind X_n, X diskrete Zufallsvariablen mit Werten in $\mathbb{Z}$, so gilt $X_n \xrightarrow{d} X$ genau dann, wenn für die Zähldichten $p^{X_n}(k) \to p^X(k)$ für alle $k \in \mathbb{Z}$ gilt (wähle Testfunktionen φ_k mit $\varphi_k(k)=1$ und $\varphi_k(\ell)=0$ für $\ell \in \mathbb{Z}\setminus\{k\}$ oder betrachte die Verteilungsfunktionen an den Stetigkeitsstellen $x_k = k+1/2$, $k\in\mathbb{Z}$). Der Poissonsche Grenzwertsatz besagt insbesondere $\mathrm{Bin}(n,p_n) \xrightarrow{w} \mathrm{Poiss}(\lambda)$ für $np_n \to \lambda > 0$.

(c) Sind $U_1,\dots,U_n$ unabhängige $U([0,1])$-verteilte Zufallsvariablen, so besitzt $X_n = n\min(U_1,\dots,U_n)$ die Verteilungsfunktion $F^{X_n}(x) = 1-(1-x/n)_+^n$ für $x \geqslant 0$. Es folgt $n\min(U_1,\dots,U_n) \xrightarrow{d} \mathrm{Exp}(1)$.

4.27 Bemerkung. Wir werden ein Kompaktheitskriterium bezüglich schwacher Konvergenz benötigen, wofür der folgende Satz fundamental ist.

4.28 Satz. *(Auswahlsatz von Helly) Ist (P_n) eine Folge von Wahrscheinlichkeitsmaßen auf $(\mathbb{R},\mathfrak{B}_\mathbb{R})$ mit Verteilungsfunktionen (F_n), so existiert eine Teilfolge (n_k) und eine monoton wachsende rechtsstetige Funktion $F:\mathbb{R}\to[0,1]$ mit $\lim_{k\to\infty}F_{n_k}(x) = F(x)$ für alle Stetigkeitspunkte x von F.*

Beweis. Wir verwenden ein Diagonalfolgenargument. sei dazu $(q_n)_{n\geqslant 1}$ eine Abzählung von $\mathbb{Q}$, d.h. $\mathbb{Q} = \{q_n \,|\, n\in\mathbb{N}\}$. Da $(F_n(q_1))_{n\geqslant1}$ eine beschränkte Folge ist, existiert eine Teilfolge $(n_1(k))_{k\geqslant1}$ und ein $H(q_1)\in[0,1]$ mit $F_{n_1(k)}(q_1)\to H(q_1)$ für $k\to\infty$. Ist eine Teilfolge $(n_\ell(k))_{k\geqslant1}$ konstruiert mit $F_{n_\ell(k)}(q_i)\to H(q_i)$, $i=1,\dots,\ell$ und Werten $H(q_i)\in[0,1]$, so können wir eine Teilfolge $(n_{\ell+1}(k))$ von $(n_\ell(k))$ auswählen mit $F_{n_{\ell+1}(k)}(q_{\ell+1})\to H(q_{\ell+1})$ für ein $H(q_{\ell+1})\in[0,1]$. Induktiv erhalten wir so $\lim_{k\to\infty}F_{n_k(k)}(q_l) = H(q_l)$ für alle $\ell\geqslant1$ entlang der Diagonalfolge $(n_k(k))$. Da alle F_n monoton wachsend sind, ist es auch $H:\mathbb{Q}\to[0,1]$. Setze nun (für $q\in\mathbb{Q}$)

$$F(x) := \lim_{q\downarrow x} H(q) = \inf_{q>x} H(q), \quad x\in\mathbb{R}\,.$$

Dann ist $F : \mathbb{R} \to [0,1]$ auch monoton wachsend. Außerdem ist F rechtsstetig an jedem Punkt x:

$$\lim_{x_n \downarrow x} F(x_n) = \lim_{x_n \downarrow x} \inf_{q > x_n} H(q) = \inf_{q > x} H(q) = F(x).$$

Es bleibt zu zeigen, dass $F_n(x) \to F(x)$ an allen Stetigkeitspunkten x von F gilt. Wähle dazu $r_1, r_2, s \in \mathbb{Q}$ mit $r_1 < r_2 < x < s$ und $F(x) - \varepsilon \leqslant F(r_1) \leqslant F(s) \leqslant F(x) + \varepsilon$ für vorgegebenes $\varepsilon > 0$, so dass

$$\begin{aligned} \limsup_{k\to\infty} F_{n_k(k)}(x) &\leqslant \limsup_{k\to\infty} F_{n_k(k)}(s) = H(s) \leqslant F(s) \leqslant F(x) + \varepsilon, \\ \liminf_{k\to\infty} F_{n_k(k)}(x) &\geqslant \liminf_{k\to\infty} F_{n_k(k)}(r_2) = H(r_2) \geqslant F(r_1) \geqslant F(x) - \varepsilon. \end{aligned}$$

Beachte bei der Rechnung, dass $H(r_2) < F(r_2)$ durchaus vorkommen kann. Da $\varepsilon > 0$ beliebig war, folgt $\lim_{k\to\infty} F_{n_k(k)}(x) = F(x)$, wie behauptet. □

4.29 Beispiele.

(a) Sind P_n Wahrscheinlichkeitsmaße auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ mit Verteilungsfunktionen F_n, so folgt aus $P_n \xrightarrow{w} P$, dass $F_n(x) \to F(x)$ an den Stetigkeitspunkten x von F gerade für die Verteilungsfunktion F von P gilt.

(b) Ist $P_n = U([n, n+1])$ mit Verteilungsfunktionen F_n, so gilt $\lim_{n\to\infty} F_n(x) = 0$ sowie $\lim_{n\to-\infty} F_n(x) = 1$ für alle $x \in \mathbb{R}$. Für $P_n = N(0,n)$ gilt $\lim_{n\to\infty} F_n(x) = 1/2$ für alle $x \in \mathbb{R}$. Die Funktion F im Satz von Helly ist hier jeweils keine Verteilungsfunktion. Intuitiv liegt dies daran, dass die Wahrscheinlichkeitsmaße P_n Masse nach $\pm\infty$ verlieren, was in der folgenden Definition 'verboten' wird.

4.30 Definition. Eine Folge von Wahrscheinlichkeitsmaßen (P_n) auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ heißt (gleichgradig) straff, falls für jedes $\varepsilon > 0$ ein $K_\varepsilon > 0$ existiert mit $\sup_{n\geqslant 1} P_n([-K_\varepsilon, K_\varepsilon]^c) \leqslant \varepsilon$.

Ende 16. Vorlesung

4.31 Lemma. *Es gelte $P_n \xrightarrow{w} P$ für Wahrscheinlichkeitsmaße P_n, P auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$. Dann ist $(P_n)_{n\geqslant 1}$ straff.*

Beweis. F bezeichne die Verteilungsfunktion von P. Dann gibt es für $\varepsilon > 0$ ein $x > 0$ mit $F(-x) \leqslant \varepsilon/4$, $F(x) \geqslant 1 - \varepsilon/4$ und F ist stetig bei x und $-x$. Dann folgt $P_n((-x, x]) \to F(x) - F(-x) \geqslant 1 - \varepsilon/2$. Wähle nun $N \in \mathbb{N}$, so dass $P_n([-x, x]) \geqslant 1 - \varepsilon$ für alle $n \geqslant N$ sowie $y \geqslant x$ mit $P_n([-y, y]) \geqslant 1 - \varepsilon$ für $n = 1, \ldots, N-1$ (möglich wegen σ-Stetigkeit). Dann gilt $P_n([-y, y]^c) \leqslant \varepsilon$ für alle n. $(P_n)_{n\geqslant 1}$ ist also straff. □

4.32 Korollar (Satz von Prokhorov)**.** *Ist (P_n) eine straffe Folge von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$, so gibt es eine Teilfolge (n_k) und ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$, so dass $P_{n_k} \xrightarrow{w} P$ gilt.*

Beweis. Nach dem Auswahlsatz von Helly reicht es, zu zeigen, dass die Grenzfunktion F dort eine Verteilungsfunktion ist. Ist dann nämlich P das zugehörige Lebesgue-Stieltjes-Maß, so erhalten wir gerade schwache Konvergenz $P_{n_k} \xrightarrow{w} P$ gemäß Satz 4.25. Damit F Verteilungsfunktion ist, fehlen noch die Eigenschaften $\lim_{x\to-\infty} F(x) = 0$ und $\lim_{x\to\infty} F(x) = 1$. Wähle zu $\varepsilon > 0$ ein $K_\varepsilon > 0$ mit $P_n([-K_\varepsilon, K_\varepsilon]^c) \leqslant \varepsilon$ für alle n. Für Stetigkeitspunkte x, y von F mit $x < -K_\varepsilon$, $y > K_\varepsilon$ folgt dann $F(x) = \lim_k F_{n_k}(x) \leqslant \varepsilon$, $F(y) = \lim_k F_{n_k}(y) \geqslant 1-\varepsilon$. Mittels Monotonie von F folgen damit die Grenzwertaussagen. □

4.33 Bemerkung. Nach Lemma 4.31 und dem Satz von Prokhorov ist die Straffheit der Folge (P_n) äquivalent dazu, dass jede Folge (n_k) eine Teilteilfolge (n_{k_ℓ}) besitzt mit $P_{n_{k_\ell}} \xrightarrow{w} P$ für ein Wahrscheinlichkeitsmaß P (die Folge (P_n) *ist schwach relativ kompakt*). Dieses Kompaktheitskriterium gilt nicht nur in $\mathbb{R}$, sondern allgemeiner auf vollständigen und separablen metrischen Räumen S mit Borel-σ-Algebra $\mathfrak{B}_S$. Dabei heißt (P_n) straff, falls es für jedes $\varepsilon > 0$ eine kompakte Menge $K_\varepsilon \subseteq S$ gibt mit $\sup_{n\geqslant 1} P_n(K_\varepsilon^c) \leqslant \varepsilon$, vergleiche Abschnitt 13.3 in Klenke. Durch Übergang zu Verteilungsfunktionen ist der Beweis für $\mathbb{R}$ bedeutend einfacher.

4.3 Charakteristische Funktionen und Zentrale Grenzwertsätze

4.34 Bemerkung. Ähnlich wie die Fouriertransformation in der Analysis gestatten es charakteristische Funktionen viele Eigenschaften von Verteilungen einfacher abzulesen. Insbesondere wird mit ihnen Konvergenz in Verteilung besonders transparent analysiert werden können.

4.35 Definition. Für eine reellwertige Zufallsvariable X bezeichnet

$$\varphi^X(u) := \mathbb{E}[e^{iuX}] = \mathbb{E}[\cos(uX)] + i\,\mathbb{E}[\sin(uX)], \quad u \in \mathbb{R},$$

die charakteristische Funktion von X. Entsprechend ist für ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$

$$\varphi^P(u) := \int_\mathbb{R} e^{iux} P(dx) = \int_\mathbb{R} \cos(ux)\, P(dx) + i \int_\mathbb{R} \sin(ux)\, P(dx), \quad u \in \mathbb{R},$$

die charakteristische Funktion von P. Für $u \in \mathbb{R}^d$ und einen Zufallsvektor X im $\mathbb{R}^d$ ist $\varphi^X(u) := \mathbb{E}[e^{i\langle u, X\rangle}]$ und für ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d})$ ist $\varphi^P(u) := \int_{\mathbb{R}^d} e^{i\langle u,x\rangle}\, P(dx)$ die entsprechende charakteristische Funktion.

4.36 Beispiele.

(a) $\varphi^{\delta_0}(u) = 1$; denn $\int e^{iux}\delta_0(dx) = e^{iu0} = 1$.

(b) $\varphi^{\mathrm{Bin}(n,p)}(u) = (pe^{iu} + 1 - p)^n$; denn:

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} e^{iuk} = \sum_{k=0}^n \binom{n}{k} (pe^{iu})^k (1-p)^{n-k} = (pe^{iu} + 1 - p)^n.$$

(c) $\varphi^{\mathrm{Poiss}(\lambda)}(u) = \exp(\lambda(e^{iu}-1))$; denn:

$$\sum_{k\geqslant 0}\frac{\lambda^k}{k!}e^{-\lambda}e^{iuk} = e^{-\lambda}\sum_{k\geqslant 0}\frac{(\lambda e^{iu})^k}{k!} = e^{-\lambda+\lambda e^{iu}}.$$

(d) $\varphi^{N(0,1)}(u) = e^{-u^2/2}$; denn:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{iux}e^{-x^2/2}dx = e^{-u^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x-iu)^2/2}dx = e^{-u^2/2}$$

folgt, weil $f(z) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x-z)^2/2}dx$ eine holomorphe Funktion auf ganz $\mathbb{C}$ ist mit $f(z) = 1$ für alle $z \in \mathbb{R}$, so dass nach Eindeutigkeitssatz $f(z) = 1$ für alle $z \in \mathbb{C}$, insbesondere $z = iu$, gilt.

(e) $\varphi^{N(0,E_d)}(u) = e^{-|u|^2/2}$ für $u \in \mathbb{R}^d$; denn für $X \sim N(0, E_d)$ folgt wegen Unabhängigkeit der Koordinaten X_j (der Satz von Fubini gilt auch für komplexwertige Integranden, betrachte Real- und Imaginärteil getrennt)

$$\mathbb{E}\left[e^{i\langle u,X\rangle}\right] = \mathbb{E}\Big[\prod_{j=1}^{d} e^{iu_jX_j}\Big] = \prod_{j=1}^{d}\mathbb{E}\left[e^{iuX_j}\right] = \prod_{j=1}^{d} e^{-u_j^2/2} = e^{-|u|^2/2}.$$

4.37 Lemma. *Für einen Zufallsvektor X im $\mathbb{R}^d$ sowie $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$ gilt*

$$\varphi^{AX+b}(u) = \varphi^X(A^\top u)e^{i\langle u,b\rangle}.$$

Insbesondere gilt für alle $\mu \in \mathbb{R}^d$ und jede Kovarianzmatrix $\Sigma \in \mathbb{R}^{d\times d}$

$$\varphi^{N(\mu,\Sigma)}(u) = e^{-\langle \Sigma u,u\rangle/2+i\langle u,\mu\rangle}.$$

Beweis. Dies folgt sofort aus $\mathbb{E}[e^{i\langle u,AX+b\rangle}] = \mathbb{E}[e^{i\langle A^\top u,X\rangle}]e^{i\langle u,b\rangle}$ und $\Sigma^{1/2}X+\mu \sim N(\mu,\Sigma)$ für $X \sim N(0,E_d)$ in Verbindung mit Beispiel 4.36(e) und $|\Sigma^{1/2}u|^2 = \langle \Sigma u,u\rangle$. □

4.38 Lemma. *Eine charakteristische Funktion φ erfüllt $\varphi(0) = 1$, $|\varphi(u)| \leqslant 1$, $u \in \mathbb{R}^d$, und sie ist gleichmäßig stetig auf $\mathbb{R}^d$.*

Beweis. Es ist $\varphi^P(0) = \int 1\,P(dx) = 1$, $|\varphi^P(u)| \leqslant \int|e^{i\langle u,x\rangle}|\,P(dx) = 1$ sowie

$$|\varphi^P(u)-\varphi^P(v)| \leqslant \int|e^{i\langle u,x\rangle}-e^{i\langle v,x\rangle}|\,P(dx) = \int|e^{i\langle u-v,x\rangle}-1||e^{ivx}|\,P(dx)$$
$$= \int|e^{i\langle u-v,x\rangle}-1|\,P(dx), \quad u,v\in\mathbb{R}\,.$$

Wegen $e^{i\langle w,x\rangle} \to 1$ für $w \to 0$ und dominierter Konvergenz ($|e^{i\langle w,x\rangle}-1| \leqslant 2$) folgt $\int|e^{i\langle w,x\rangle}-1|\,P(dx) \to 0$. Zu $\varepsilon > 0$ existiert also ein $\delta > 0$ mit $\int|e^{i\langle w,x\rangle}-1|\,P(dx) < \varepsilon$ für $|w| < \delta$. Damit folgt $|\varphi^P(u)-\varphi^P(v)| \leqslant \varepsilon$ für alle $u,v \in \mathbb{R}^d$ mit $|u-v| < \delta$, also die gleichmäßige Stetigkeit. □

4.39 Lemma. *Es gilt $\varphi^{X_1+X_2}(u) = \varphi^{X_1}(u)\varphi^{X_2}(u)$ für unabhängige Zufallsvektoren X_1, X_2 im $\mathbb{R}^d$.*

Beweis. Dies folgt direkt aus

$$\mathbb{E}[e^{i\langle u,X_1+X_2\rangle}] = \mathbb{E}[e^{i\langle u,X_1\rangle}e^{i\langle u,X_2\rangle}] = \mathbb{E}[e^{i\langle u,X_1\rangle}]\,\mathbb{E}[e^{i\langle u,X_2\rangle}]$$

für unabhängige Zufallsvariablen X_1, X_2. □

4.40 Bemerkung. Die Faltung von Wahrscheinlichkeitsmaßen P_1, P_2 auf $\mathfrak{B}_\mathbb{R}$ wird also zum einfachen Produkt bei charakteristischen Funktionen (wie bei der Fouriertransformation in der Analysis): $\varphi^{P_1*P_2}(u) = \varphi^{P_1}(u)\varphi^{P_2}(u)$.

4.41 Lemma. *Für eine Zufallsvariable $X \in \mathscr{L}^m(\Omega,\mathcal{F},P)$ mit $m \in \mathbb{N}$ gilt $\varphi^X \in C^m(\mathbb{R})$ mit Ableitungen*

$$(\varphi^X)^{(k)}(0) = i^k\,\mathbb{E}[X^k], \quad k = 0,\dots,m.$$

Weiterhin existiert eine Funktion $r_m : \mathbb{R} \to \mathbb{C}$ mit $\sup_{|u|\leqslant 1}|r_m(u)| \leqslant 2\,\mathbb{E}[|X|^m]$ und $r_m(u) \to 0$ für $u \to 0$, so dass gilt

$$\varphi^X(u) = \sum_{k=0}^{m} \frac{(iu)^k}{k!}\,\mathbb{E}[X^k] + \frac{(iu)^m}{m!} r_m(u).$$

Beweis. Übung! □

4.42 Beispiel. Wegen $\frac{d}{du}\varphi^{\mathrm{Bin}(n,p)}(u) = n(pe^{iu}+1-p)^{n-1}ipe^{iu}$ mit Wert inp bei $u = 0$ besitzt die $\mathrm{Bin}(n,p)$-Verteilung Erwartungswert np. Wegen $\frac{d^2}{du^2}\varphi^{\mathrm{Bin}(n,p)}(u) = -n(n-1)(pe^{iu}+1-p)^{n-2}p^2e^{2iu} - n(pe^{iu}+1-p)^{n-1}pe^{iu}$ mit Wert $-n(n-1)p^2 - np$ bei $u = 0$ erhalten wir $n(n-1)p^2 + np$ als zweites Moment von $\mathrm{Bin}(n,p)$ und somit $n(n-1)p^2+np-n^2p^2 = np(1-p)$ als Varianz.

4.43 Satz. *(Eindeutigkeitssatz in $\mathbb{R}$) Für ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R},\mathfrak{B}_\mathbb{R})$ und $a < b$ gilt die Inversionsformel*

$$P((a,b)) + \tfrac{1}{2}P(\{a,b\}) = \frac{1}{2\pi}\lim_{U\to\infty}\int_{-U}^{U}\frac{e^{-iua}-e^{-iub}}{iu}\varphi^P(u)\,du.$$

Insbesondere sind zwei Wahrscheinlichkeitsmaße auf $\mathbb{R}$ mit derselben charakteristischen Funktion identisch.

Beweis. Wir verwenden aus der Analysis die Formel $\lim_{U\to\infty}\int_{-U}^{U}\frac{\sin(x)}{x}\,dx = \pi$. Setze

$$I(U) := \int_{-U}^{U}\int_{-\infty}^{\infty}\frac{e^{-iua}-e^{-iub}}{iu}e^{iux}\,P(dx)\,du, \quad U > 0,$$

mit stetiger Ergänzung bei $u = 0$. Wir erhalten mit dem Satz von Fubini (überprüfe Bedingungen!), $\int_{-U}^{U}\frac{\cos(uy_1)-\cos(uy_2)}{u}du = 0$ wegen Antisymmetrie und mit der Substitutionsregel der Integration

$$\begin{aligned}
I(U) &= \int_{-\infty}^{\infty}\int_{-U}^{U}\frac{e^{-iua}-e^{-iub}}{iu}e^{iux}du\,P(dx)\\
&= \int_{-\infty}^{\infty}\Big(\int_{-U}^{U}\frac{\sin(u(x-a))}{u}\,du - \int_{-U}^{U}\frac{\sin(u(x-b))}{u}\,du\Big)P(dx)\\
&= \int_{-\infty}^{\infty}\Big(\int_{-U(x-a)}^{U(x-a)}\frac{\sin(v)}{v}\,dv - \int_{-U(x-b)}^{U(x-b)}\frac{\sin(v)}{v}\,du\Big)P(dx).
\end{aligned}$$

Für $x > b > a$ und $U \to \infty$ konvergiert der Integrand bezüglich x gegen $\pi - \pi = 0$, für $x < a < b$ gegen $-\pi - (-\pi) = 0$ sowie für $a < x < b$ gegen $\pi - (-\pi) = 2\pi$. In den Fällen $x = a$ und $x = b$ ergibt sich der Grenzwert π. Mit dominierter Konvergenz ($\sup_U |\int_{-U}^{U} \frac{\sin(x)}{x}\, dx| < \infty$) folgt daher

$$\lim_{U\to\infty} I(U) = \int_{(a,b)} 2\pi\, P(dx) + \int_{\{a,b\}} \pi\, P(dx) = 2\pi P((a,b)) + \pi P(\{a,b\}).$$

Division durch 2π ergibt die behauptete Identität.

Die Eindeutigkeit folgt daraus, wenn man beachtet, dass es höchstens abzählbar viele $x_n \in \mathbb{R}$ gibt mit $P(\{x_n\}) > 0$. Für $a < b$ mit $P(\{a\}) = P(\{b\}) = 0$ erfüllt die Verteilungsfunktion F dann nämlich $F(b) - F(a) = \frac{1}{2\pi} \lim_{U\to\infty} \int_{-U}^{U} \frac{e^{-iua} - e^{-iub}}{iu} \varphi^P(u)\, du$, ist dort also durch φ^P eindeutig bestimmt. Wegen Rechtsstetigkeit ist F auch bei x_n eindeutig festgelegt. Mit F ist dann auch P eindeutig bestimmt. □

4.44 Satz. *(Eindeutigkeitssatz in $\mathbb{R}^d$) Es sei P ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d})$. Für $a_j < b_j$, $j = 1, \dots, d$, mit $P(\prod_{j=1}^d [a_j, b_j] \setminus \prod_{j=1}^d (a_j, b_j)) = 0$ gilt die Inversionsformel*

$$P\Big(\prod_{j=1}^d [a_j, b_j]\Big) = (2\pi)^{-d} \lim_{U\to\infty} \int_{[-U,U]^d} \prod_{j=1}^d \Big(\frac{e^{-iu_j a_j} - e^{-iu_j b_j}}{iu_j}\Big) \varphi^P(u)\, du.$$

Insbesondere sind zwei Wahrscheinlichkeitsmaße auf dem $\mathbb{R}^d$ mit derselben charakteristischen Funktion identisch.

Beweisskizze (analog zum Fall $d = 1$). Setze

$$I(U) := \int_{[-U,U]^d} \int_{\mathbb{R}^d} \prod_{j=1}^d \Big(\frac{e^{-iu_j a_j} - e^{-iu_j b_j}}{iu_j}\Big) e^{i\langle u,x\rangle}\, P(dx)\, du, \quad U > 0,$$

mit stetiger Ergänzung für $u_j = 0$. Wir erhalten für $U \to \infty$

$$\begin{aligned} I(U) &= \int_{\mathbb{R}^d} \prod_{j=1}^d \Big(\int_{-U(x_j-a_j)}^{U(x_j-a_j)} \frac{\sin(v)}{v}\, dv - \int_{-U(x_j-b_j)}^{U(x_j-b_j)} \frac{\sin(v)}{v}\, du \Big)\, P(dx) \\ &\to \int_{-\infty}^{\infty} \prod_{j=1}^d \Big(2\pi \mathbf{1}_{(a_j,b_j)}(x_j) + \pi \mathbf{1}_{\{a_j,b_j\}}(x_j) \Big)\, P(dx). \end{aligned}$$

Da P keine Masse auf dem Rand des Quaders $\prod_{j=1}^d [a_j, b_j]$ besitzt, folgt die behauptete Formel. Da P überhaupt nur höchstens abzählbar vielen Hyperebenen $H(c, j) = \{x \in \mathbb{R}^d \mid x_j = c\}$, $c \in \mathbb{R}$, $j = 1, \dots, d$, positive Wahrscheinlichkeiten zuordnen kann und die Familie aller Quader ohne solche mit Randpunkten auf abzählbar vielen Hyperebenen weiter einen $\cap$-stabilen Erzeuger von $\mathfrak{B}_{\mathbb{R}^d}$ bildet, ist P durch obige Formel eindeutig über die charakteristische Funktion festgelegt. □

4.45 Bemerkung. Es folgt aus der Definition der schwachen Konvergenz, dass $P_n \xrightarrow{w} P \Rightarrow \varphi^{P_n}(u) \to \varphi^P(u)$ punktweise. Überraschenderweise gilt auch die Umkehrung. Mehr noch, unter minimalen Annahmen ist der punktweise Grenzwert von charakteristischen Funktionen wiederum eine charakteristische Funktion, wofür wir ein entsprechendes Wahrscheinlichkeitsmaß finden müssen. Der folgende Stetigkeitsatz von Lévy ist also sehr mächtig.

4.46 Satz. *(Stetigkeitssatz von Lévy) Sind P_n Wahrscheinlichkeitsmaße auf $(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d})$ mit charakteristischen Funktionen (φ_n) und gilt $\lim_{n\to\infty} \varphi_n(u) = \psi(u)$ für alle $u \in \mathbb{R}^d$ und eine bei $u = 0$ stetige Funktion ψ, so ist $\psi = \varphi^P$, die charakteristische Funktion eines Wahrscheinlichkeitsmaßes P auf $(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d})$, und es gilt $P_n \xrightarrow{w} P$.*

Beweis. Zeige zunächst $P_n(\mathbb{R}^d \setminus [-\frac{2}{u}, \frac{2}{u}]^d) \leqslant \frac{2}{(2u)^d} \int_{[-u,u]^d} (1 - \varphi_n(v))dv$ für beliebige $u > 0$. Dazu beachte

$$(2u)^{-d} \int_{[-u,u]^d} (1 - e^{i\langle v,x\rangle})dv = 1 - (2u)^{-d} \prod_{j=1}^{d} \int_{-u}^{u} e^{iv_j x_j} dv_j, = 1 - \prod_{j=1}^{d} \frac{\sin(x_j u)}{x_j u}$$

so dass mit dem Satz von Fubini

$$\begin{aligned}
\frac{1}{(2u)^d} \int_{[-u,u]^d} (1 - \varphi_n(v))dv &= \int_{\mathbb{R}^d} \Big(1 - \prod_{j=1}^{d} \frac{\sin(x_j u)}{x_j u}\Big) P_n(dx) \\
&\geqslant \int_{\mathbb{R}^d} \Big(1 - \prod_{j=1}^{d} \frac{1 \wedge |x_j u|}{|x_j u|}\Big) P_n(dx) \geqslant \int_{\{\max_j |x_j| > 2/u\}} \Big(1 - \frac{1}{2}\Big) P_n(dx) \\
&= \frac{1}{2} P_n\Big(\mathbb{R}^d \setminus \big[-\tfrac{2}{u}, \tfrac{2}{u}\big]^d\Big).
\end{aligned}$$

Mit Hilfe obiger Ungleichung zeigen wir nun, dass (P_n) straff ist. Da ψ stetig in 0 ist und $\psi(0) = 1$ gilt, existiert für $\varepsilon > 0$ ein $u > 0$ mit $\frac{1}{(2u)^d} \int_{[-u,u]^d} (1 - \psi(v))dv < \varepsilon/2$. Wegen $\varphi_n(v) \to \psi(v)$ zeigt dominierte Konvergenz

$$\limsup_{n\to\infty} P_n\big(\mathbb{R}^d \setminus [-\tfrac{2}{u}, \tfrac{2}{u}]^d\big) \leqslant \limsup_{n\to\infty} \frac{2}{(2u)^d} \int_{[-u,u]^d} (1 - \varphi_n(v))\, dv \leqslant 2\frac{\varepsilon}{2} = \varepsilon.$$

Daher ist die Folge (P_n) straff (wieso reicht der limes superior?).

Nach dem Satz von Prokhorov (Korollar 4.32 bzw. Bemerkung 4.33 für $d > 1$) existiert eine Teilfolge (n_k) mit $P_{n_k} \xrightarrow{w} P$ für ein Wahrscheinlichkeitsmaß P. Wie oben gesehen, impliziert dies $\varphi^{P_{n_k}}(u) \to \varphi^P(u)$ punktweise und daher $\varphi^P = \psi$. Mit einem Teilteilfolgenargument weisen wir nun nach, dass dies auch für die gesamte Folge (P_n) gilt.

Angenommen es gilt nicht $P_n \xrightarrow{w} P$, so gibt es $f \in C_b(\mathbb{R}^d)$ und eine Teilfolge (n_k) mit $\liminf_{k\to\infty} |\int f\, dP_{n_k} - \int f\, dP| > 0$. Wegen Straffheit gibt es eine Teilteilfolge (n_{k_l}) und ein Wahrscheinlichkeitsmaß Q mit $P_{n_{k_l}} \xrightarrow{w} Q$, so dass

$\varphi^{P_{n_{k_l}}}(u) \to \varphi^Q(u)$. Also gilt $\varphi^Q = \psi = \varphi^P$ und nach dem Eindeutigkeitssatz $P = Q$, was $\liminf_{k\to\infty} |\int f\, dP_{n_k} - \int f\, dP| > 0$ zum Widerspruch führt. Es folgt $P_n \xrightarrow{w} P$. □

4.47 Beispiele.

(a) Für $p_n \in [0,1]$ mit $\lim_{n\to\infty} np_n = \lambda > 0$ folgt

$$\varphi^{\mathrm{Bin}(n,p_n)}(u) = (1 + p_n(e^{iu} - 1))^n \to e^{\lambda(e^{iu}-1)} = \varphi^{\mathrm{Poiss}(\lambda)}(u)$$

punktweise. Der Stetigkeitssatz von Lévy in Verbindung mit dem Eindeutigkeitssatz impliziert daher den Poissonschen Grenzwertsatz $\mathrm{Bin}(n,p_n) \xrightarrow{w} \mathrm{Poiss}(\lambda)$.

(b) Ist S_n $\mathrm{Bin}(n,p)$-verteilt mit $p \in (0,1)$, $\sigma_n^2 = np(1-p)$, so ist $S_n^* = \frac{S_n - np}{\sigma_n}$ zentriert mit Varianz 1 (*standardisiert*) und

$$\varphi^{S_n^*}(u) = \varphi^{S_n}(u/\sigma_n) e^{-iunp/\sigma_n} = \big(1 + p(e^{iu/\sigma_n} - 1)\big)^n e^{-iunp/\sigma_n}.$$

Für $n \to \infty$ zeigt also eine Taylorentwicklung um $u/\sigma_n = 0$

$$\begin{aligned}\log(\varphi^{S_n^*}(u)) &= n\big(\log(1 + p(e^{iu/\sigma_n} - 1)) - iup/\sigma_n\big) \\ &= n\big(\tfrac{iup}{\sigma_n} - \tfrac{u^2 p}{2\sigma_n^2} + \tfrac{u^2 p^2}{2\sigma_n^2} + O(\sigma_n^{-3}) - \tfrac{iup}{\sigma_n}\big) \to -\tfrac{u^2}{2}.\end{aligned}$$

$\varphi^{S_n^*}(u)$ konvergiert also punktweise gegen $\varphi^{N(0,1)}(u) = e^{-u^2/2}$. Es folgt $S_n^* \xrightarrow{d} N(0,1)$, der aus der Schule bekannte Satz von de Moivre-Laplace, der jetzt bedeutend verallgemeinert wird.

4.48 Satz. *(Zentraler Grenzwertsatz, Standardversion) Ist $(X_i)_{i\geqslant 1}$ eine Folge unabhängiger und identisch verteilter Zufallsvariablen (i.i.d.=independent and identically distributed) in $\mathscr{L}^2$ mit $\mu = \mathbb{E}[X_i]$, $\sigma^2 = \mathrm{Var}(X_i) > 0$, so erfüllt ihre standardisierte Summe*

$$S_n^* := \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \xrightarrow{d} N(0,1).$$

Beweis. Nach dem Stetigkeitssatz von Lévy und dem Eindeutigkeitssatz genügt es für die Verteilungskovergenz, $\varphi^{S_n^*}(u) \to \varphi^{N(0,1)}(u) = e^{-u^2/2}$ für alle $u \in \mathbb{R}$ zu zeigen. Setze $\tilde{X}_i = (X_i - \mu)/\sigma$. Nach den Rechenregeln für charakteristische Funktionen gilt (benutze $(\tilde{X}_i)$ i.i.d.)

$$\varphi^{S_n^*}(u) = \varphi^{\sum_{i=1}^n \tilde{X}_i}(u/\sqrt{n}) = \Big(\varphi^{\tilde{X}_1}(u/\sqrt{n})\Big)^n.$$

Wegen $\tilde{X}_1 \in \mathscr{L}^2$ mit $\mathbb{E}[\tilde{X}_1] = 0$, $\mathrm{Var}(\tilde{X}_1) = 1$ erhalten wir nach Lemma 4.41 $\varphi^{\tilde{X}_1}(u/\sqrt{n}) = 1 - \frac{u^2}{2n}(1 + r_2(u/\sqrt{n}))$ mit $r_2(u/\sqrt{n}) \to 0$ für $n \to \infty$. Mit der Approximation $\log(1+h) = h + O(h^2)$ für $h \to 0$ (wähle für $h \in \mathbb{C}$, $|h| < 1$ den Zweig des Logarithmus mit $\log 1 := 0$, also $\log(1+h) = -\sum_{k\geqslant 1}(-h)^k/k$) folgt

$$\log(\varphi^{S_n^*}(u)) = n \log\Big(1 - \frac{u^2}{2n}(1 + r_2(u/\sqrt{n}))\Big) = -\frac{u^2}{2}(1 + r_2(u/\sqrt{n})) + O\Big(\frac{1}{n}\Big),$$

und die rechte Seite konvergiert gegen $-u^2/2$, wie zu zeigen war. □

4.49 Bemerkung. Standardisierung bedeutet gerade, dass $\mathbb{E}[S_n^*] = 0$, $\mathrm{Var}(S_n^*) = 1$. Der zentrale Grenzwertsatz ist ein Universalitätsprinzip: wie auch immer die Ausgangsverteilung der X_i ist (unter der Minimalbedingung $X_i \in \mathscr{L}^2$), so ergibt sich durch standardisierte Mittelung stets asymptotisch eine Standardnormalverteilung. Bei komplexeren physikalischen Messapparaturen (mit vielen ähnlichen Fehlerquellen) werden daher die Messfehler standardmäßig als normalverteilt angenommen. Ähnliches gilt in fast allen Anwendungsfeldern, was oft (aber durchaus nicht immer) sehr gut mit der Empirie übereinstimmt. Das folgende Korollar wird häufig benutzt, um daraus asymptotische Konfidenzintervalle über Quantile der Normalverteilung zu gewinnen.

4.50 Korollar. *Unter den Voraussetzungen von Satz 4.48 gilt für $a < b$*

$$P(S_n^* \in (a,b)) \to \Phi(b) - \Phi(a), \quad P(S_n^* \in [a,b]) \to \Phi(b) - \Phi(a)$$

mit der Verteilungsfunktion Φ der Standardnormalverteilung $N(0,1)$; insbesondere

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^n X_i - \mu\Big| > 1{,}96\frac{\sigma}{\sqrt{n}}\Big) \to 2 - 2\Phi(1{,}96) \approx 0{,}05.$$

Beweis. Nach Satz 4.48 und der Charakterisierung der Verteilungskonvergenz gemäß Satz 4.25 folgt direkt $P(S_n^* \in (a,b]) \to \Phi(b) - \Phi(a)$, da Φ stetig ist. Für jedes $\delta \in (0, b-a)$ gilt wegen Monotonie

$$\liminf_{n\to\infty} P(S_n^* \in (a,b)) \geqslant \lim_{n\to\infty} P(S_n^* \in (a, b-\delta]) = \Phi(b-\delta) - \Phi(a).$$

Mit $\delta \downarrow 0$ und Stetigkeit von Φ folgt daher auch $\lim_{n\to\infty} P(S_n^* \in (a,b)) = \Phi(b) - \Phi(a)$. Die zweite Konvergenz wird analog gezeigt oder folgt durch Komplementbildung.

Mit $b = 1{,}96$, $a = -1{,}96$ und $\Phi(a) = 1 - \Phi(b)$ aus Symmetriegründen erhalten wir $P(|S_n^*| \leqslant 1{,}96) \to 2\Phi(1{,}96) - 1$. Wegen $\{|\frac{1}{n}\sum_{i=1}^n X_i - \mu| > 1{,}96\frac{\sigma}{\sqrt{n}}\} = \{|S_n^*| \not\leqslant 1{,}96\}$ folgt die letzte Aussage durch Komplementbildung sowie die numerische Näherung $\Phi(1{,}96) \approx 0{,}975$. □

Ende 18. Vorlesung

4.51 Satz (Cramér-Wold)**.** *Für Zufallsvektoren X_n, X im $\mathbb{R}^d$ gilt $X_n \xrightarrow{d} X$ genau dann, wenn für alle $v \in \mathbb{R}^d$ die eindimensionale Konvergenz $\langle X_n, v\rangle \xrightarrow{d} \langle X, v\rangle$ vorliegt.*

Beweis. '⇒' folgt sofort aus der Stetigkeit von $x \mapsto \langle x, v\rangle$. '⇐': Aus der Verteilungskonvergenz von $\langle X_n, v\rangle$ folgt die Konvergenz der charakteristischen Funktionen $\mathbb{E}[e^{iw\langle X_n, v\rangle}] \to \mathbb{E}[e^{iw\langle X, v\rangle}]$ für alle $w \in \mathbb{R}$, $v \in \mathbb{R}^d$. Das impliziert aber die Konvergenz der d-dimensionalen charakteristischen Funktionen $\mathbb{E}[e^{i\langle u, X_n\rangle}] \to \mathbb{E}[e^{i\langle u, X\rangle}]$ für alle $u \in \mathbb{R}^d$. Nach dem Stetigkeitssatz von Lévy folgt $X_n \xrightarrow{d} X$. □

4.52 Definition. Für einen Vektor $\mu \in \mathbb{R}^d$ und eine symmetrische, positiv semi-definite Matrix $\Sigma \in \mathbb{R}^{d\times d}$ ist die Verteilung $N(\mu, \Sigma)$ gegeben als die Verteilung von $\mu + \Sigma^{1/2}Z$ mit $Z \sim N(0, E_d)$ standard-normalverteilt.

4.53 Bemerkung. Dies verallgemeinert die bisherige Definition für invertierbare (strikt positiv-definite) Σ. Beachte, dass für symmetrische positiv semi-definite Matrizen $\Sigma \in \mathbb{R}^{d\times d}$ stets eine symmetrisch positiv semi-definite Matrix $\Sigma^{1/2} \in \mathbb{R}^{d\times d}$ existiert mit $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$.

Für $d = 1$ ist $N(\mu, 0)$ also die Verteilung von $\mu + 0Z$, also das Einpunktmaß δ_μ. Ist allgemein Σ nicht invertierbar, so besitzt $N(\mu, \Sigma)$ keine Dichte mehr und hat Träger auf dem echten (affinen) Unterraum $\{\mu + \Sigma^{1/2}z \,|\, z \in \mathbb{R}^d\} \subseteq \mathbb{R}^d$. Diese natürliche Verallgemeinerung der Normalverteilung erlaubt eine einfache Formulierung des mehrdimensionalen zentralen Grenzwertsatzes.

4.54 Lemma. *Ist W ein $N(\mu, \Sigma)$-verteilter Zufallsvektor im $\mathbb{R}^d$, so gilt $\langle W, v\rangle \sim N(\langle \mu, v\rangle, \langle \Sigma v, v\rangle)$ für jedes $v \in \mathbb{R}^d$.*

Beweis. Schreibe $W = \mu + \Sigma^{1/2}Z$ mit $Z \sim N(0, E_d)$. Für die charakteristische Funktion erhalten wir damit

$$\begin{aligned}\varphi^{\langle W,v\rangle}(u) &= \mathbb{E}[e^{iu\langle v,\mu+\Sigma^{1/2}Z\rangle}] = e^{iu\langle\mu,v\rangle}\,\mathbb{E}[e^{i\langle\Sigma^{1/2}(uv),Z\rangle}]\\ &= e^{iu\langle\mu,v\rangle}\varphi^Z(\Sigma^{1/2}(uv)) = e^{iu\langle\mu,v\rangle}e^{-u^2|\Sigma^{1/2}v|^2/2},\end{aligned}$$

was nach Lemma 4.37 gerade die charakteristische Funktion von $\langle\mu, v\rangle + |\Sigma^{1/2}v|\tilde{Z}$ mit $\tilde{Z} \sim N(0, 1)$ ist. Wegen $|\Sigma^{1/2}v|^2 = \langle\Sigma v, v\rangle$ und dem Eindeutigkeitssatz folgt also, dass $\langle W, v\rangle$ $N(\langle\mu, v\rangle, \langle\Sigma v, v\rangle)$-verteilt ist. □

4.55 Satz. *(Zentraler Grenzwertsatz, allgemein) Ist $(X_i)_{i\geqslant 1}$ eine Folge von i.i.d.-Zufallsvektoren im $\mathbb{R}^d$ mit $\mathbb{E}[|X_i|^2] < \infty$, $\mu = \mathbb{E}[X_i] \in \mathbb{R}^d$, $\Sigma = (\mathrm{Cov}((X_i)_k, (X_i)_\ell))_{1\leqslant k,\ell\leqslant d} \in \mathbb{R}^{d\times d}$, so gilt*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \xrightarrow{d} N(0, \Sigma).$$

Beweis. Im Fall $d = 1$ setze $\Sigma = \sigma^2$ für $\sigma \geqslant 0$. Im Fall $\sigma > 0$ liefert der zentrale Grenzwertsatz $S_n^* := \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \xrightarrow{d} Z$ mit $Z \sim N(0, 1)$. Daraus folgt $\sigma S_n^* \xrightarrow{d} \sigma Z \sim N(0, \sigma^2)$, da $x \mapsto f(\sigma x)$ für $f \in C_b(\mathbb{R})$ wiederum in $C_b(\mathbb{R})$ liegt ('continuous mapping theorem', vergleiche Übung). Für $\sigma = 0$ gilt $\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) = 0$ fast sicher (da die Varianz Null ist) und die Behauptung ist gerade $\delta_0 \xrightarrow{w} \delta_0$, was trivial ist.

Nach dem Satz von Cramér-Wold und dem Lemma genügt es, für $d \geqslant 2$ die Konvergenz $\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \mu, v\rangle \xrightarrow{d} N(0, \langle\Sigma v, v\rangle)$ für alle $v \in \mathbb{R}^d$ nachzuweisen.

Nun sind $\tilde{X}_i = \langle X_i, v\rangle$ unabhängige reellwertige Zufallsvariablen mit $\mathbb{E}[\tilde{X}_i] = \langle\mu, v\rangle$ sowie $\mathrm{Var}(\tilde{X}_i) = \sum_{k,\ell=1}^d \mathrm{Cov}(X_{i,k}v_k, X_{i,\ell}v_\ell) = \langle\Sigma v, v\rangle$. Aus dem Fall $d = 1$ folgt daher $\frac{1}{\sqrt{n}}\sum_{i=1}^n (\tilde{X}_i - \langle\mu, v\rangle) \xrightarrow{d} N(0, \langle\Sigma v, v\rangle)$, was zu zeigen war. □

4.4 Asymptotik der empirischen Verteilung

4.56 Definition. Es seien $X_1, \ldots, X_n$ unabhängige, identisch verteilte Zufallsvariablen (*Beobachtungen*) mit Werten in $\mathbb{R}$. Dann heißt das Wahrscheinlichkeitsmaß $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ empirische Verteilung oder empirisches Maß sowie seine Verteilungsfunktion $F_n(x) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty,x]}(X_i)$, $x \in \mathbb{R}$, empirische Verteilungsfunktion. Im folgenden setze $F^X := F^{X_i}$, $P^X := P^{X_i}$. Beachte dabei, dass μ_n und F_n zufällige Objekte sind.

4.57 Satz. *Für alle $x \in \mathbb{R}$, $x_1 < \ldots < x_m \in \mathbb{R}$ und $n \to \infty$ gilt:*

(a) $F_n(x) \to F^X(x)$ *P-fast sicher;*

(b) $\sqrt{n}(F_n(x) - F^X(x)) \xrightarrow{d} N(0, F^X(x)(1 - F^X(x)))$;

(c) mit der Kovarianzmatrix $\Sigma = (F^X(x_k \wedge x_\ell) - F^X(x_k)F^X(x_\ell))_{1\leqslant k,\ell\leqslant m} \in \mathbb{R}^{m\times m}$ *gilt mehrdimensionale Konvergenz in Verteilung:*

$$\sqrt{n}\begin{pmatrix} F_n(x_1) - F^X(x_1) \\ \vdots \\ F_n(x_m) - F^X(x_m) \end{pmatrix} \xrightarrow{d} N(0, \Sigma).$$

Beweis. Da $Z_i(x) := \mathbf{1}_{(-\infty,x]}(X_i)$, $i \geqslant 1$, i.i.d.-Zufallsvariablen sind mit $\mathbb{E}[Z_i(x)] = P(X_i \leqslant x) = F^X(x)$ und $\mathrm{Var}(Z_i(x)) = F^X(x) - F^X(x)^2$ folgt mit dem starken Gesetz der großen Zahlen Teil (a) und mit dem zentralen Grenzwertsatz 4.55 in Dimension Eins Teil (b). Außerdem ist $\mathbb{E}[Z_i(x_k)Z_i(x_\ell)] = P(X_i \leqslant x_k \wedge x_l) = F^X(x_k \wedge x_\ell)$, so dass

$$\mathrm{Cov}(Z_i(x_k), Z_i(x_\ell)) = F^X(x_k \wedge x_\ell) - F^X(x_k)F^X(x_\ell) = \Sigma_{k,\ell}$$

gilt. Nach dem mehrdimensionalen ZGWS 4.55 folgt

$$\sqrt{n}\begin{pmatrix} F_n(x_1) - F^X(x_1) \\ \vdots \\ F_n(x_m) - F^X(x_m) \end{pmatrix} \xrightarrow{d} N(0, \Sigma),$$

wobei wir für die Euklidische Norm $|(Z_i(x_k))_{1\leqslant k\leqslant m}| \leqslant \sqrt{m} \in \mathscr{L}^2$ beachten. □

4.58 Bemerkung. Die empirische Verteilungsfunktion ist ein schönes Beispiel für einen *stochastischen Prozess*, also eine mit $x \in \mathbb{R}$ indizierte Familie von Zufallsvariablen. Alternativ kann diese als eine zufällige Funktion angesehen werden (deren Eigenschaften man studieren kann). Teil (c) zeigt, dass der standardisierte stochastische Prozess (*empirischer Prozess* genannt) $(\sqrt{n}(F_n(x) - F^X(x)), x \in \mathbb{R})$ gegen einen sogenannten Gauß-Prozess $(G^X(x), x \in \mathbb{R})$ mit Erwartungswertfunktion $\mathbb{E}[G^X(x)] = 0$ und Kovarianzfunktion $\mathrm{Cov}(G^X(x), G^X(y)) = F^X(x \wedge y) - F^X(x)F^X(y)$ konvergiert in dem Sinne, dass endlich-dimensionale Randverteilungen an beliebigen Stellen $x_1, \ldots, x_m$ in Verteilung konvergieren. Im Fall $X_i \sim U([0,1])$ ist $(G^X(x), x \in [0,1])$ eine sogenannte *Standard-Brownsche Brücke*, eine Brownsche Bewegung darauf bedingt, zur 'Zeit' $x = 1$ den Wert Null anzunehmen.

Dies wird in Stochastik II im Detail studiert werden, hier werden wir nur das starke Gesetz der großen Zahlen verschärfen, indem wir P-f.s. gleichmäßige Konvergenz der empirischen Verteilungsfunktion gegen die wahre Verteilungsfunktion nachweisen. Dies wird manchmal auch Hauptsatz der Statistik genannt, weil es zeigt, dass durch unendlich viele unabhängige Wiederholungen desselben Zufallsexperiments, das zugehörige Wahrscheinlichkeitsmaß eindeutig aus den empirischen Beobachtungen rekonstruiert werden kann (mit Wahrscheinlichkeit Eins). Über diese Grenzwerte relativer Häufigkeiten wurden ursprünglich Wahrscheinlichkeiten intuitiv und dann sogar formal (von Mises, HU Berlin, 1919) definiert, bevor Kolmogorov (1933) axiomatisch Wahrscheinlichkeiten in die Maßtheorie einbettete und deduktiv die Konvergenz der relativen Häufigkeiten herleitete.

4.59 Satz (Glivenko-Cantelli)**.** *Die empirische Verteilungsfunktion konvergiert P-f.s. gleichmäßig gegen die wahre Verteilungsfunktion:*

$$P\Big(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x)-F^X(x)|=0\Big)=1.$$

Beweis. Da der Schnitt über abzählbar viele Einsmengen wieder eine Einsmenge ist, zeigt Satz 4.57(a) $P(\forall r\in\mathbb{Q}:\lim_{n\to\infty}F_n(r)=F^X(r))=1$. Daraus folgt bereits die gleichmäßige Konvergenz für Verteilungsfunktionen. Wir führen einen rein analytischen Widerspruchsbeweis.

Sonst existieren $\varepsilon>0$ und Folgen (x_k), (n_k) mit $|F_{n_k}(x_k)-F^X(x_k)|>\varepsilon$ für alle $k\geqslant 1$. Da es $a,b\in\mathbb{Q}$ gibt mit $F^X(a)<\varepsilon/2$, $F^X(b)>1-\varepsilon/2$ implizieren die Grenzwerte $\lim_{k\to\infty}F_{n_k}(a)<\varepsilon/2$, $\lim_{k\to\infty}F_{n_k}(b)>1-\varepsilon/2$, dass wegen Monotonie $\liminf_{k\to\infty}x_k\geqslant a$, $\limsup_{k\to\infty}x_k\leqslant b$ gelten muss. Insbesondere existiert wegen Kompaktheit von $[a,b]$ eine Teilfolge (k_ℓ) mit $x_{k_\ell}\to x$ für ein $x\in[a,b]$.

Setze $F^X(x-):=\lim_{y\uparrow x}F^X(y)=\sup_{q<x}F^X(q)$ für $q\in\mathbb{Q}$ wegen Monotonie von F^X. Deshalb und wegen Rechtsstetigkeit existieren $r_1,r_2\in\mathbb{Q}$ mit $r_1<x<r_2$ und $F^X(r_2)<F^X(x)+\varepsilon$, $F^X(r_1)>F^X(x-)-\varepsilon$. Wir erhalten

$$\begin{aligned}\limsup_{\ell\to\infty}|F_n(x_{k_\ell})-F^X(x_{k_\ell})|\mathbf{1}(x_{k_\ell}\geqslant x)&\leqslant F^X(r_2)-F^X(x)<\varepsilon,\\ \limsup_{\ell\to\infty}|F_n(x_{k_\ell})-F^X(x_{k_\ell})|\mathbf{1}(x_{k_\ell}<x)&\leqslant F^X(x-)-F^X(r_1)<\varepsilon,\end{aligned}$$

im Widerspruch zu $|F_{n_k}(x_k)-F^X(x_k)|>\varepsilon$ für alle $k\geqslant 1$. Also konvergiert F_n gegen F gleichmäßig auf einem Ereignis von Wahrscheinlichkeit Eins. □

4.60 Korollar. *Für $n\to\infty$ konvergiert das empirische Maß μ_n P-f.s. schwach gegen die Verteilung P^X von X:*

$$P(\{\omega\in\Omega\,|\,\mu_n(\omega)\xrightarrow{w}P^X\})=1.$$

Beweis. Nach dem Satz von Glivenko-Cantelli hat das Ereignis $\{\omega\in\Omega\,|\,F_n(\omega,x)\to F^X(x)$ für alle $x\in\mathbb{R}\}$ (wir schreiben das Argument ω, um den Zufallsvariablencharakter zu betonen) Wahrscheinlichkeit Eins. Da F_n die Verteilungsfunktion von μ_n ist, folgt nach Satz 4.25 insbesondere die schwache Konvergenz der zugehörigen Maße. □

5 Einführung in Statistik

5.1 Hypothesentests und Neyman-Pearson-Lemma

5.1 Definition. Ein statistisches Modell ist ein Tripel $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$ bestehend aus einer Menge $\mathcal{X}$ mit einer σ-Algebra $\mathscr{F}$ (dem Stichprobenraum) und einer Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ von Wahrscheinlichkeitsmaßen auf $\mathscr{F}$. Die mindestens zwei-elementige Menge Θ heißt Parametermenge und jedes $\vartheta \in \Theta$ Parameter.

5.2 Bemerkung. Ein statistisches Modell ist also eine durch ϑ parametrisierte Familie von Wahrscheinlichkeitsräumen (bloß wird kanonisch $\mathcal{X}$ statt Ω als Notation verwendet), die alle denselben Messraum $(\mathcal{X}, \mathscr{F})$ besitzen. Die Grundidee der mathematischen Statistik ist, dass wir eine Realisierung eines Zufallsexperiments $(\mathcal{X}, \mathscr{F}, P_\vartheta)$ beobachten und daraus Rückschlüsse auf den unbekannten Parameter ϑ ziehen.

Wir beginnen mit dem Problem, zwei Hypothesen gegeneinander zu testen. Typische Beispiele sind zu testen, ob eine Münze fair ist, ob Mädchen- und Jungengeburten gleichhäufig sind, ob Kinder genauso oder weniger ansteckend sind als Erwachsene oder ob die vorliegende Email Spam ist oder nicht.

5.3 Definition. Aufbau eines Testverfahrens:

(a) Wahl eines statistischen Modells $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$

(b) Formulierung von Hypothese und Alternative: $\Theta = \Theta_0 \dot\cup \Theta_1$
$\vartheta \in \Theta_0$: ϑ entspricht der Null-Hypothese H_0
$\vartheta \in \Theta_1$: ϑ entspricht der Alternative H_1

(c) Wahl eines Irrtumsniveaus $\alpha \in (0, 1)$ für den Fehler erster Art, sich bei Vorliegen der Nullhypothese für die Alternative zu entscheiden.

(d) Konstruktion eines Tests φ zum Niveau α:
$\varphi : \mathcal{X} \to \{0, 1\}$ ist eine messbare Funktion der Beobachtungen mit
$\varphi(x) = 0$: Entscheidung für H_0,
$\varphi(x) = 1$: Entscheidung für H_1,
$\sup_{\vartheta\in\Theta_0} P_\vartheta(\varphi = 1) \leqslant \alpha$.

(e) Durchführen des Experiments

Ende 18. Vorlesung

5.4 Beispiel. Wir wollen testen, ob Mädchen- und Jungengeburten in Berlin gleichwahrscheinlich sind. Dazu sei $\vartheta \in [0, 1]$ die Wahrscheinlichkeit für eine Mädchengeburt und $1 - \vartheta$ für eine Jungengeburt (nicht gemeldete oder diverse Geschlechter seien hier weggelassen). Als Modell nehmen wir an, dass bei n Kindern jedes Kind mit Wahrscheinlichkeit ϑ, unabhängig von den anderen, als Mädchen geboren wird. Dann ist bei Betrachtung der Summe der Mädchengeburten unser statistisches Modell gerade $\mathcal{X} = \{0, \dots, n\}$, $\mathscr{F} = \mathscr{P}(\mathcal{X})$ und $P_\vartheta = \mathrm{Bin}(n, \vartheta)$ für $\vartheta \in \Theta = [0, 1]$. Als Null-Hypothese wählen wir $H_0 : \vartheta = 1/2$, also $\Theta_0 = \{1/2\}$, als Alternative $H_1 : \vartheta \neq 1/2$, also $\Theta_1 = [0, 1] \setminus \{1/2\}$ und als

Irrtumsniveau $\alpha = 0,05$ (häufige Werte für das Irrtumsniveau sind 1%, 5% und 10%, je nach Konsequenzen bei einem Fehler erster Art).

Da unser Testproblem symmetrisch ist und die Binomialverteilung unter der Nullhypothese den Erwartungswert $n/2$ besitzt, wählen wir den Test $\varphi(x) = \mathbf{1}(|x - n/2| > \kappa_\alpha)$ für ein geeignetes $\kappa_\alpha > 0$ mit $P_{1/2}(\varphi = 1) \leqslant \alpha = 0,05$. Idealerweise nehmen wir einen maximalen Wert von κ_α, der das Niveau α noch einhält (κ_α heißt dann kritischer Wert zum Niveau α).

Da die Anzahl n der Geburten recht groß sein wird, verwenden wir eine Normalapproximation der Binomialverteilung gemäß ZGWS. Da $\mathrm{Bin}(n, 1/2)$ die Varianz $n/4$ besitzt, gilt

$$\lim_{n\to\infty} P_{1/2}(|x - n/2| > 1,96\sqrt{n/4}) = (1 - \Phi(1,96)) + \Phi(-1,96) \approx 0,05.$$

Wir setzen daher $\kappa_{0,05} = 1,96\sqrt{n/4}$ und haben damit den Test φ vollständig konstruiert.

Jetzt erst betrachten wir die Daten; denn sonst sind Manipulationen Tür und Tor geöffnet und jede datenabhängige Wahl von Verfahren müsste mit geeigneten Zufallsmechanismen (wiederum im vorhinein!) modelliert werden.

Das Amt für Statistik Berlin-Brandenburg meldet für Juli 2022 in Berlin $n = 3216$ Lebendgeborene, darunter $x = 1556$ Mädchen. Also gilt

$$\varphi(x) = \mathbf{1}(|1556 - 1608| > 0,98\sqrt{3216}) = \mathbf{1}(52 > 55,58) = 0.$$

Zum Niveau 5% können wir die Nullhypothese gleichwahrscheinlicher Mädchen- und Jungengeburten nicht ablehnen. Betrachten wir aber den Jahreswert 2022, so gab es $n = 35731$ Lebendgeborene und darunter $x = 17247$ Mädchen. Hier gilt also

$$\varphi(x) = \mathbf{1}(|17247 - 17865,5| > 0,98\sqrt{35731}) = \mathbf{1}(618,5 > 185,25) = 1.$$

Hier wird die Nullhypothese zum Niveau $\alpha = 5\%$ abgelehnt. In beiden Fällen sind es etwas mehr als 48% Mädchengeburten.

Durch die mathematische Modellierung haben wir es aber erreicht, die Unsicherheit in den Daten zu quantifizieren und können so im zweiten Fall von einer *signifikanten* Abweichung von der Nullhypothese sprechen, die weiter wissenschaftlich verstanden werden sollte, vgl. Spiegel-Artikel vom 30.3.2015.

5.5 Bemerkung. Bislang haben wir Tests nur unter der Null-Hypothese H_0 betrachtet, und die Wahl eines spezifischen Tests scheint zum Teil arbiträr (z.B. erfüllt $\varphi(x) = 0$ alle bisherigen Anforderungen). Um über Optimalität zu sprechen und eine ertragreiche mathematische Theorie zu entwickeln, betrachten wir nun den Fehler 2. Art, die Null-Hypothese zu akzeptieren, obwohl die Alternative gilt. Ziel ist es, unter allen Tests vom Niveau α möglichst solche auszuwählen, die geringe Fehlerwahrscheinlichkeit 2. Art besitzen.

Die Asymmetrie zwischen Null-Hypothese und Alternative bei diesem Ansatz entspringt der Erkenntnistheorie und ist Grundlage modernen wissenschaftlichen Arbeitens: Daten können eine Theorie nie beweisen, sondern nur widerlegen. Die Null-Hypothese sollte also eine etablierte Theorie oder allgemeine Einsichten kodieren, während die Alternative gewisse Abweichungen erlaubt.

Wenn die Daten nur mit sehr kleiner Wahrscheinlichkeit (nur bei geringem Niveau des Tests) durch die Theorie erklärt werden können, lehnt man sie (vorerst; beachte Datenfehler) ab.

Ein Beispiel dafür ist die 'Entdeckung' des Higgs-Bosons: die Null-Hypothese, dass die Messergebnisse durch zufällige Schwankungen ohne dieses Elementarteilchen verursacht worden sind, wurde zum Niveau $1 : 3500000$ abgelehnt (Wahrscheinlichkeit für 22-mal hintereinander 'Kopf' beim Münzwurf) und führte zum Nobelpreis 2013, auch wenn eine minimale Wahrscheinlichkeit existiert, dass nur 'Zufall' beobachtet wurde, vergleiche Wikipedia. In gewissen Anwendungen wird die Asymmetrie auch mit den schweren Folgen eines Fehlers 1. Art begründet. In der pharmazeutischen Industrie muss beispielsweise ein neues Medikament signifikant (besser als andere Therapien) wirken, da bei Neueinführung immer die Gefahr nicht beachteter Nebenwirkungen besteht. Hier ist also die Null-Hypothese H_0 stets, dass keine (bessere) Wirkung vorliegt, und die Aufsichtsbehörde schreibt ein sehr geringes Testniveau vor.

Bei anderen Fragestellungen, insbesondere in großen Datenanwendungen, liegt eher Symmetrie vor, und es werden Verfahren entwickelt, die die Summe der Fehlerwahrscheinlichkeiten 1. und 2. Art minimieren, zum Beispiel bei der Bildklassifikation zwischen den Hypothesen 'Tier ist Hund' und 'Tier ist Katze'.

5.6 Definition. Für $\vartheta_1 \in \Theta_1$ bezeichnet $\beta_\varphi(\vartheta_1) = P_{\vartheta_1}(\varphi = 0)$ die Wahrscheinlichkeit für den Fehler 2. Art der Entscheidung für H_0, obwohl $\vartheta_1 \in \Theta_1$ vorliegt.

5.7 Definition. Ein Test φ von $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$ heißt gleichmäßig bester Test zum Niveau α (englisch UMP, *uniformly most powerful*), falls φ ein Test zum Niveau α ist und für jeden anderen Test ψ zum Niveau α gilt:

$$\forall\, \vartheta_1 \in \Theta_1 :\ P_{\vartheta_1}(\varphi = 1) \geqslant P_{\vartheta_1}(\psi = 1).$$

5.8 Bemerkung. Im Fall einfacher Null-Hypothese und einfacher Alternative, das heißt für $\Theta = \{0, 1\}$ sind $\Theta_0 = \{0\}$, $\Theta_1 = \{1\}$ ein-elementig, kann man beste Tests stets konstruieren, im Allgemeinen existieren sie nicht immer und man wählt beispielsweise asymptotisch beste Tests oder schränkt die Familie der in Frage kommenden Tests durch weitere Eigenschaften ein.

Der sogenannte *zweiseitige Binomialtest* aus Beispiel 5.4 ist beispielsweise gleichmäßig bester Tests unter allen Tests, wo unter der Alternative $P_{\vartheta_1}(\varphi = 1) \geqslant \alpha$ gilt. Das scheint sehr vernünftig, muss aber gefordert werden, weil sonst einseitige Tests für $\vartheta_1 > \vartheta_0$ besser sind (Fehlerwahrscheinlichkeit 2. Art ist kleiner), während sie für $\vartheta_1 < \vartheta_0$ sehr schlecht sind oder umgekehrt. Wir hätten also einen andere Test gewählt, wenn wir die Nullhypothese 'Mädchengeburten sind mindestens so häufig wie Jungengeburten' gegen die Alternative 'Mädchengeburten sind seltener als Jungengeburten' getestet hätten. Ein solcher einseitiger Binomialtest der Form $\varphi(x) = \mathbf{1}(x - n/2 > \tilde{\kappa}_\alpha)$ hätte bereits bei den Quartalsdaten zum Niveau 5% abgelehnt. In der konkreten Anwendung spielt die Modellierung also eine wichtige Rolle.

5.9 Definition. Für ein statistisches Modell $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$, bei dem jedes P_ϑ eine Dichte p_ϑ bezüglich einem Maß μ auf $(\mathcal{X}, \mathscr{F})$ besitzt, ist die

Likelihood-Funktion die (unter jedem P_ϑ) zufällige Funktion $L : \Theta \to [0,\infty)$ mit

$$L(\vartheta) := p_\vartheta, \text{ das heißt } L(\vartheta, x) = p_\vartheta(x), \quad \vartheta \in \Theta,\, x \in \mathcal{X}.$$

Im Fall $\Theta = \{\vartheta_0, \vartheta_1\}$, $\Theta_0 = \{\vartheta_0\}$, $\Theta_1 = \{\vartheta_1\}$ heißt jeder Test φ der Form

$$\varphi = \mathbf{1}(L(\vartheta_1) > \kappa L(\vartheta_0)), \text{ d.h. } \varphi(x) = \mathbf{1}(L(\vartheta_1, x) > \kappa L(\vartheta_0, x))$$

mit beliebigem $\kappa \geqslant 0$ Neyman-Pearson-Test für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_1$.

5.10 Bemerkung. Typische Fälle von Likelihood-Funktionen sind im diskreten Fall das Zählmaß μ auf $\mathbb{Z}$, die Zähldichten p_ϑ und beim Binomialmodell $(\mathrm{Bin}(n,\vartheta))_{\vartheta\in[0,1]}$, zum Beispiel $L(\vartheta) = \binom{n}{X}\vartheta^X(1-\vartheta)^{n-X}$, $\vartheta \in [0,1]$. Hier schreiben wir X statt k, um den Charakter der über X zufälligen Funktion in ϑ zu betonen. Ein Neyman-Pearson-Test hat im Binomialmodell für $\vartheta_0, \vartheta_1 \in (0,1)$ die Form

$$\varphi = \mathbf{1}\Big((\vartheta_1/(1-\vartheta_1))^X > \kappa(\vartheta_0/(1-\vartheta_0))^X\Big),$$

wobei sich Faktoren unabhängig von ϑ in der Likelihood-Funktion herausgekürzt haben. Im Fall $\vartheta_1 > \vartheta_0$ lässt sich das zu $\varphi = \mathbf{1}(X > \tilde\kappa)$ mit $\tilde\kappa = \log(\kappa)/\log(\vartheta_1(1-\vartheta_0)/((\vartheta_0(1-\vartheta_1)))$ vereinfachen. Wir lehnen also $H_0 : \vartheta = \vartheta_0$ zugunsten von $H_1 : \vartheta = \vartheta_1$ ab, wenn der Versuchsausgang X größer als ein kritischer Wert $\tilde\kappa$ ist.

Im Fall von Wahrscheinlichkeitsdichten f_ϑ der Maße P_ϑ auf $(\mathbb{R}^d, \mathfrak{B}_{\mathbb{R}^d})$ ist mit dem Lebesguemaß μ und $L(\vartheta) = f_\vartheta(X)$ beim Normalverteilungsmodell $(N(\vartheta, E_d))_{\vartheta\in\mathbb{R}^d}$ beispielsweise $L(\vartheta) = (2\pi)^{-d/2}e^{-|\vartheta-X|^2/2}$ eine Glockenkurve in ϑ, die beim zufälligen Wert X zentriert ist.

Bei der Definition der Likelihood-Funktion ändert sich im Vergleich zu den Dichten p_ϑ allein der Fokus auf den interessierenden Parameter ϑ anstatt x. Formal hängt L vom Maß μ ab, für die statistischen Aussagen, die $L(\vartheta)$ immer unter einer Wahrscheinlichkeit P_{ϑ_0} studieren, wird dies jedoch keine Rolle spielen.

5.11 Satz. *(Neyman-Pearson-Lemma) Es sei φ^* ein Neyman-Pearson-Test für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_1$ und $\alpha := P_{\vartheta_0}(\varphi^* = 1)$. Dann ist φ^* gleichmäßig bester Test zum Niveau α: es gilt $P_{\vartheta_1}(\varphi^* = 1) = \sup_\varphi P_{\vartheta_1}(\varphi = 1)$, wobei das Supremum über alle Tests φ vom Niveau α gebildet wird.*

Beweis. Betrachte einen beliebigen Test φ vom Niveau α. Der Beweis beruht auf einer geschickten Zerlegung der Dichteintegrale mit der Beobachtung $\varphi^*(x) = 1 \Rightarrow L(\vartheta_1, x) > \kappa L(\vartheta_0, x)$ bzw. $\varphi^*(x) = 0 \Rightarrow L(\vartheta_1, x) \leqslant \kappa L(\vartheta_0, x)$, was einen Maßwechsel von P_{ϑ_1} zu P_{ϑ_0} gestattet. Wir erhalten über Subtraktion und Addition von $P_{\vartheta_1}(\varphi^* = 1, \varphi = 1)$:

$$\begin{aligned}
P_{\vartheta_1}(\varphi^* = 1) - P_{\vartheta_1}(\varphi = 1) &= P_{\vartheta_1}(\varphi^* = 1, \varphi = 0) - P_{\vartheta_1}(\varphi^* = 0, \varphi = 1) \\
&= \int_{\mathcal{X}} \big(\mathbf{1}(\varphi^* = 1,\, \varphi = 0) - \mathbf{1}(\varphi^* = 0, \varphi = 1)\big)\, L(\vartheta_1)\, d\mu \\
&\geqslant \int_{\mathcal{X}} \mathbf{1}(\varphi^* = 1,\, \varphi = 0)\, \kappa L(\vartheta_0)\, d\mu - \int_{\mathcal{X}} \mathbf{1}(\varphi^* = 0,\, \varphi = 1)\, \kappa L(\vartheta_0)\, d\mu \\
&= \kappa\big(P_{\vartheta_0}(\varphi^* = 1, \varphi = 0) - P_{\vartheta_0}(\varphi^* = 0, \varphi = 1)\big) \\
&= \kappa\big(P_{\vartheta_0}(\varphi^* = 1) - P_{\vartheta_0}(\varphi = 1)\big) \geqslant 0,
\end{aligned}$$

wobei zuletzt die Definition von α und das Niveau α von φ eingegangen sind. $\square$

5.12 Beispiel. Wir betrachten den Fall, dass wir über n unabhängige Beobachtungen einer $U([0,\vartheta])$-Verteilung verfügen, also das statistische Modell $(\mathbb{R}^n, \mathfrak{B}_{\mathbb{R}^n}, (U([0,\vartheta])^{\otimes n})_{\vartheta>0})$. Bezüglich dem Lebesguemaß μ erhalten wir die Likelihood-Funktion

$$L(\vartheta) = \prod_{i=1}^{n} \vartheta^{-1} \mathbf{1}_{[0,\vartheta)}(X_i) = \vartheta^{-n} \mathbf{1}\Big(\vartheta \geqslant \max_{i=1,\dots,n} X_i\Big).$$

Testen wir $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_1$ für $\vartheta_0 < \vartheta_1$, so hat jeder Neyman-Pearson-Test die Gestalt

$$\varphi(x) = \mathbf{1}\Big(\mathbf{1}\Big(\max_{i=1,\dots,n} x_i \leqslant \vartheta_1\Big) > \kappa(\vartheta_1/\vartheta_0)^n \mathbf{1}\Big(\max_{i=1,\dots,n} x_i \leqslant \vartheta_0\Big)\Big), \quad x \in \mathbb{R}^n\,.$$

Für alle $\kappa(\vartheta_1/\vartheta_0)^n \geqslant 1$ gilt also $\varphi = \mathbf{1}(\max_{i=1,\dots,n} X_i > \vartheta_0)$ und für alle $\kappa(\vartheta_1/\vartheta_0)^n < 1$ gilt bloß $\varphi = 1$, wobei wir $\max_i X_i \leqslant \vartheta_1$ als Bedingung weglassen können, da unter P_{ϑ_0} und P_{ϑ_1} dies fast sicher erfüllt ist. Für die großen Werte von κ besitzt φ stets Niveau $\alpha = 0\%$ wegen $X_i \leqslant \vartheta_0$ P_{ϑ_0}-f.s., für kleine Werte von κ besitzt φ nur Niveau $\alpha = 100\%$, verwirft also immer. Es existiert also kein Neyman-Pearson-Test φ in unserem Sinn mit $P_{\vartheta_0}(\varphi = 1) = \alpha \in (0,1)$.

Für eine vollständige Theorie erlaubt man sogenannte randomisierte Tests $\varphi : \mathcal{X} \to [0,1]$ (statt $\{0,1\}$!) mit der Interpretation, dass ein unabhängiges Zufallsexperiment durchgeführt wird, dass mit Wahrscheinlichkeit $\varphi(x)$ die Null-Hypothese ablehnt. Mathematisch wird die Menge der Tests konvexifiert, was eine weitaus befriedigerende Theorie gestattet, insbesondere existiert zu jedem Niveau $\alpha \in [0,1]$ ein randomisierter Neyman-Pearson-Test, der weiter optimal ist. Für kompliziertere (*zusammengesetzte*) Null-Hypothesen oder Alternativen lässt sich die Idee des Neyman-Pearson-Tests verallgemeinern.

5.13 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit Likelihood-Funktion L bezüglich einem Maß μ. Für ein beliebiges Testproblem $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$ mit $\Theta = \Theta_0 \dot\cup \Theta_1$ heißt ein Test der Form

$$\varphi(x) = \mathbf{1}\Big(\frac{\sup_{\vartheta\in\Theta_1} L(\vartheta, x)}{\sup_{\vartheta\in\Theta_0} L(\vartheta, x)} > \kappa\Big)$$

mit $\kappa > 0$ Likelihood-Quotienten-Test.

5.14 Bemerkung. Bei Vorliegen einer Beobachtung x wählen wir also diejenigen Parameter $\vartheta_1 \in \Theta_1$ und $\vartheta_0 \in \Theta_0$, die jeweils die Likelihood $L(\vartheta, x)$ maximieren (falls die Suprema angenommen werden) und testen dann entsprechend des Neyman-Pearson-Ansatzes. Dies führt meist auf vernünftige und in vielen Fällen sogar auf in geeignetem Sinne optimale Tests.

5.15 Beispiel. In n unabhängigen Experimenten soll der Wert μ_0 einer physikalischen Naturkonstante überprüft werden. Dies modellieren wir mit einem Gaußmodell $(N(\mu,\sigma^2)^{\otimes n})_{\mu\in\mathbb{R},\sigma>0}$, wo der unbekannte Parameter $\vartheta = (\mu,\sigma)$ zwei-dimensional ist. Wir testen $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$, setzen daher

$\Theta_0 = \{(\mu_0, \sigma) \,|\, \sigma > 0\}$, $\Theta_1 = \{(\mu, \sigma) \,|\, \mu \in \mathbb{R} \setminus \{\mu_0\}, \sigma > 0\}$. Die Fehlervarianz σ^2 ist dabei ein unbekannter, sogenannter Störparameter (nur μ interessiert uns). Wir wählen μ als n-dimensionales Lebesguemaß und maximieren die sogenannte *Loglikelihood-Funktion* zunächst über Θ_1 (setze $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$)

$$\begin{aligned}
\sup_{(\mu,\sigma)\in\Theta_1} \log(L(\mu,\sigma)) &= \sup_{(\mu,\sigma)\in\Theta_1} \Big(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}\Big) \\
&= \sup_{\sigma^2>0} \Big(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (X_i-\bar{X})^2}{2\sigma^2}\Big) \\
&= -\frac{n}{2}\log(2\pi\bar{\sigma}^2) - \frac{\sum_{i=1}^n (X_i-\bar{X})^2}{2\bar{\sigma}^2}\Big) \\
&= \frac{n}{2}\Big(-\log(2\pi\bar{\sigma}^2) - 1\Big).
\end{aligned}$$

Analog erhält man über Θ_0

$$\begin{aligned}
\sup_{(\mu,\sigma)\in\Theta_0} \log(L(\mu,\sigma)) &= \sup_{\sigma^2>0} \Big(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (X_i-\mu_0)^2}{2\sigma^2}\Big) \\
&= \frac{n}{2}\Big(-\log\Big(2\pi\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2\Big) - 1\Big).
\end{aligned}$$

Ein Likelihood-Quotiententest hat also die Form

$$\begin{aligned}
\varphi &= \mathbf{1}\Big(\frac{n}{2}\Big(-\log(2\pi\bar{\sigma}^2) - 1\Big) - \frac{n}{2}\Big(-\log\Big(2\pi\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2\Big) - 1\Big) \geqslant \log(\kappa)\Big) \\
&= \mathbf{1}\Big(-\log(2\pi\bar{\sigma}^2)) + \log\Big(2\pi\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2\Big) \geqslant \frac{2}{n}\log(\kappa)\Big) \\
&= \mathbf{1}\Big(\frac{\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2}{\bar{\sigma}^2} \geqslant \kappa^{2/n}\Big).
\end{aligned}$$

Mit $\tilde{\kappa} = \kappa^{2/n} \geqslant 0$ ergibt sich der Likelihoodquotiententest

$$\varphi = \mathbf{1}\Big(\frac{\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2}{\bar{\sigma}^2} \geqslant \tilde{\kappa}\Big).$$

Zu einem Niveau $\alpha \in (0,1)$ wird also $\tilde{\kappa}_\alpha$ so gewählt, dass

$$P_{\vartheta_0}\Big(\frac{\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2}{\bar{\sigma}^2} \geqslant \tilde{\kappa}_\alpha\Big) = \alpha \text{ für alle } \vartheta_0 \in \Theta_0$$

gilt. Schreibt man $X_i = \mu_0 + \sigma Z_i$ mit unabhängigen $Z_i \sim N(0,1)$ unter P_{ϑ_0}, so muss

$$P\Big(\frac{\sum_{i=1}^n Z_i^2}{\sum_{i=1}^n (Z_i - \bar{Z})^2} \geqslant \tilde{\kappa}_\alpha\Big) = \alpha$$

gelöst werden, was unabhängig von μ_0 und σ ist. Man erhält $\tilde{\kappa}_\alpha$ über die sogenannte t-Verteilung, weshalb φ t-Test heißt. Dieser ist von großer Bedeutung in allen Anwendungen.

5.2 Der χ^2-Anpassungstest

5.16 Beispiel. Als typisches Beispiel für einen Multinomial- und χ^2-Test, betrachte einen Zufallszahlengenerator der die Ziffern '0', '1',...,'9' alle gleichwahrscheinlich (und unabhängig voneinander) ziehen soll. Bei n Aufrufen liefert er jeweils X_j-mal die Ziffer j, $j = 0, \dots, 9$. Wie können wir die Null-Hypothese der Gleichverteilung testen?

Unter der Nullhypothese gilt für $X = (X_0, \dots, X_9)$ gemäß Definition 1.25 $X \sim \mathrm{Mult}(n, 10, p_0, \dots, p_9)$ mit $p_0 = \cdots = p_9 = 1/10$. Unter der Alternative gelte immer noch $X \sim \mathrm{Mult}(n, 10, p_0, \dots, p_9)$, aber mit anderen Klassenwahrscheinlichkeiten $p_0, \dots, p_9$. Wir betrachten dazu den Likelihood-Quotienten-Test, den wir im folgenden bestimmen und asymptotisch für große n approximieren werden.

5.17 Lemma. *Betrachte das durch* $\mathrm{Mult}(n, r, p_1, \dots, p_r)$ *gegebene statistische Modell mit* $p_j \geqslant 0$, $\sum_{j=1}^r p_j = 1$ *und einen festen Vektor von Klassenwahrscheinlichkeiten* p^0 *mit* $p_j^0 > 0$ *und* $\sum_{j=1}^r p_j^0 = 1$. *Dann hat der Likelihood-Quotienten-Test für die Null-Hypothese* $H_0 : p = p^0$ *gegen die Alternative* $H_1 : p \neq p^0$ *die Form*

$$\varphi(x) = \mathbf{1}\Big(\sum_{j=1}^r x_j \log\Big(\frac{x_j}{np_j^0}\Big) \geqslant \tilde\kappa \Big), \quad \tilde\kappa \in \mathbb{R},$$

wobei jeder Summand bei $x_j = 0$ *stetig durch Null fortgesetzt sei.*

Beweis. Die Likelihood-Funktion (bezüglich Zählmaß μ) ist gerade die Zähldichte aus Definition 1.25

$$L(p) = \frac{n!}{X_1! \cdots X_r!} p_1^{X_1} \cdots p_r^{X_r}.$$

Diese ist stetig in p, so dass das Supremum über alle $p \neq p^0$ gleich dem Maximum über alle p ist. Wir behaupten, dass das Maximum bei $p = (X_1/n, \dots, X_r/n)$ angenommen wird. In der Tat ist

$$\begin{aligned} L(p) &= L(X_1/n, \dots, X_r/n) \prod_j \Big(\frac{np_j}{X_j}\Big)^{X_j} \\ &\leqslant L(X_1/n, \dots, X_r/n) \exp\Big(\sum_j X_j \Big(\frac{np_j}{X_j} - 1\Big)\Big) \\ &= L(X_1/n, \dots, X_r/n) \exp(n - n) = L(X_1/n, \dots, X_r/n), \end{aligned}$$

wobei sich wegen $x^0 := 1$ für alle $x \geqslant 0$ das Produkt und die Summe nur über alle j mit $X_j \geqslant 1$ erstreckt und wir die Ungleichung

$$\prod_j A_j^{\rho_j} \leqslant \prod_j (e^{A_j - 1})^{\rho_j} = \exp\Big(\sum_j \rho_j (A_j - 1)\Big), \quad A_j \in \mathbb{R},\ \rho_j \geqslant 0,$$

benutzt haben. Alternativ kann man das Maximum auch durch Ableiten mit einem Lagrangemultiplikator für die Nebenbedingung $\sum_j p_j = 1$ bestimmen.

Wir erhalten also

$$\frac{\sup_{p\neq p^0} L(p)}{L(p^0)} = \frac{L(X_1/n,\dots,X_r/n)}{L(p_1^0,\dots,p_r^0)} = \prod_{j=1}^{r} \Big(\frac{X_j}{np_j^0}\Big)^{X_j}.$$

Logarithmieren ergibt für den Likelihood-Quotienten-Test mit kritischem Wert $\kappa \geqslant 0$ (beachte $\log(X_j^{X_j}) = 0$ im Fall $X_j = 0$ wegen $0^0 := 1$)

$$\varphi(x) = \mathbf{1}\Big(\sum_{j=1}^{r} x_j \log\Big(\frac{x_j}{np_j^0}\Big) \geqslant \log(\kappa)\Big).$$

Mit $\tilde{\kappa} = \log(\kappa) \in \mathbb{R}$ folgt die Behauptung. □

5.18 Definition. Der in Lemma 5.17 hergeleitete Test heißt Multinomialtest.

5.19 Bemerkung. Die Bestimmung des kritischen Werts $\tilde{\kappa}$, um ein gegebenes Niveau α im Multinomialtest einzuhalten, ist schwierig und kann in der Praxis durch Simulation (unter der $\mathrm{Mult}(n, r, p^0)$-Verteilung) erreicht werden.

Unter der Asymptotik $n \to \infty$ ergibt sich eine weitreichende Vereinfachung. Zunächst betrachten wir dazu die quadratische Taylorentwicklung

$$x\log(x/x_0) = (x - x_0) + \frac{(x-x_0)^2}{2x_0} - \frac{(x-x_0)^3}{6(x_0 + h(x-x_0))^2} \text{ für ein } h \in [0,1]$$

und erwarten für $X \sim \mathrm{Mult}(n, r, p^0)$ und große n

$$\sum_{j=1}^{r} X_j \log\Big(\frac{X_j}{np_j^0}\Big) \approx \sum_{j=1}^{r} \Big((X_j - np_j^0) + \frac{(X_j - np_j)^2}{2np_j^0}\Big) = \sum_{j=1}^{r} \frac{(X_j - np_j^0)^2}{2np_j^0},$$

wobei wir $\sum_j X_j = \sum_j np_j^0 = n$ ausgenutzt haben.

5.20 Definition. Ist $X \sim \mathrm{Mult}(n, r, p)$ und p^0 ein Vektor von Klassenwahrscheinlichkeiten mit $p_j^0 > 0$, $j = 1, \dots, r$, so heißt

$$V_n^2 := \sum_{j=1}^{r} \frac{(X_j - np_j^0)^2}{np_j^0}$$

Pearsons χ^2-Teststatistik. Der χ^2-Anpassungstest für die Hypothese $H_0 : p = p^0$ gegen $H_1 : p \neq p^0$ ist gegeben durch $\varphi = \mathbf{1}(V_n^2 > \kappa)$ für einen kritischen Wert $\kappa > 0$.

5.21 Bemerkung. In V_n^2 werden also die gewichteten quadratischen Abweichungen von X_j zum Erwartungswert np_j^0 unter der Null-Hypothese aufsummiert. Die Form $\frac{1}{np_j^0}$ der Gewichte stellt sich dabei als optimal heraus (es ist nicht die Varianz von X_j!). Wir zeigen jetzt rigoros, dass Pearsons Teststatistik die Teststatistik des Multinomialtests approximiert, und beweisen dann, dass V_n^2 asymptotisch χ^2-verteilt ist.

5.22 Lemma. *Für $X \sim \mathrm{Mult}(n, r, p^0)$ und $n \to \infty$ gilt*

$$\tfrac{1}{2} V_n^2 - \sum_{j=1}^{r} X_j \log\Big(\frac{X_j}{np_j^0}\Big) \xrightarrow{P} 0.$$

Beweis. Aus $X \sim \mathrm{Mult}(n, r, p^0)$ folgt ja $X_j \sim \mathrm{Bin}(n, p_j^0)$ und nach dem ZGWS $n^{-1/2}(X_j - np_j^0) \xrightarrow{d} N(0, p_j^0(1-p_j^0))$. Insbesondere sind die Verteilungen gemäß Lemma 4.31 straff, und für jedes $\varepsilon > 0$ existiert ein $K_\varepsilon > 0$ mit

$$\forall n \geqslant 1 : \; P\Big(n^{-1/2}|X_j - np_j^0| > K_\varepsilon\Big) \leqslant \varepsilon.$$

Für das Restglied der Taylorentwicklung bei hinreichend großem n gilt mit Wahrscheinlichkeit mindestens $1 - \varepsilon$ (genauer: auf dem Ereignis $\{n^{-1/2}|X_j - np_j^0| \leqslant K_\varepsilon\}$)

$$\begin{aligned}\Big|X_j \log\Big(\frac{X_j}{np_j^0}\Big) - \frac{(X_j - np_j^0)^2}{2np_j^0}\Big| &\leqslant \sup_{h\in[0,1]} \frac{|X_j - np_j^0|^3}{6(np_j^0 + h(X_j - np_j^0))^2} \\ &\leqslant \frac{|X_j - np_j^0|^3}{(np_j^0 - |X_j - np_j^0|)^2} \leqslant n^{-1/2} \frac{K_\varepsilon^3}{(p_j^0 - n^{-1/2}K_\varepsilon)^2}.\end{aligned}$$

Dies zeigt (wie genau?) stochastische Konvergenz $X_j \log(\frac{X_j}{np_j^0}) - \frac{(X_j - np_j^0)^2}{2np_j^0} \xrightarrow{P} 0$ für jedes j. Mittels Dreiecksungleichung folgt damit auch die Konvergenz der Summe über j. □

Ende 19. Vorlesung

5.23 Definition. Die χ^2-Verteilung mit $k \in \mathbb{N}$ Freiheitsgraden ist die Verteilung von $|Z|^2 = Z_1^2 + \cdots + Z_k^2$ auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$, wobei $Z \sim N(0, E_k)$ ein k-dimensionaler standardnormalverteilter Vektor ist. Notation: $\chi^2(k)$.

5.24 Bemerkung. Mit Dichtetransformation haben wir in Beispiel 1.70 die $\chi^2(1)$-Dichte bestimmt. Durch k-fache Faltung erhalten wir die $\chi^2(k)$-Dichte

$$f_{\chi^2(k)}(x) = \frac{2^{-k/2}}{\Gamma(k/2)} x^{(k-2)/2} e^{-x/2} \mathbf{1}(x > 0), \quad x \in \mathbb{R}\,.$$

Die $\chi^2(k)$-Verteilung ist ein Spezialfall der sogenannten Γ-Verteilungen.

5.25 Satz. *Für $X \sim \mathrm{Mult}(n, r, p^0)$ und $n \to \infty$ erhalten wir Konvergenz der χ^2-Teststatistik gegen die χ^2-Verteilung mit $r - 1$ Freiheitsgraden:*

$$V_n^2 \xrightarrow{d} \chi^2(r-1).$$

Beweis. Die wichtigste Beobachtung für den Beweis ist, dass

$$V_n^2 = \sum_{j=1}^{r} \frac{(X_j - np_j^0)^2}{np_j^0} = |Y_n|^2 \text{ mit } Y_n := \Big(\frac{X_1 - np_1^0}{\sqrt{np_1^0}}, \ldots, \frac{X_r - np_r^0}{\sqrt{np_r^0}}\Big)^\top$$

gilt. Unten zeigen wir $Y_n \xrightarrow{d} N(0,\Sigma)$ mit Kovarianzmatrix $\Sigma = E_r - ww^\top \in \mathbb{R}^{r\times r}$, wobei $w = (\sqrt{p_1^0},\dots,\sqrt{p_r^0})^\top \in \mathbb{R}^r$ ein Einheitsvektor ist: $|w|^2 = \sum_j p_j^0 = 1$. Insbesondere ist w ein Eigenvektor von Σ zum Eigenwert 0, und jeder Vektor $v \in \mathbb{R}^r$ orthogonal zu w ist Eigenvektor von Σ zum Eigenwert 1 wegen $ww^\top v = w\langle w, v\rangle = 0$. Also erhalten wir die Diagonalisierung $\Sigma = ODO^\top$ mit der Diagonalmatrix $D = \operatorname{diag}(1,\dots,1,0)$ und einer Orthogonalmatrix O (d.h. $O^\top O = E_r$). Da die Euklidische Norm invariant unter orthogonalen Transformationen und stetig ist, folgt mit $Z \sim N(0,\Sigma)$ und $W := O^\top Z \sim N(0,D)$ (beachte $\mathbb{E}[W] = O^\top \mathbb{E}[Z] = 0$, $\mathbb{E}[WW^\top] = O^\top \mathbb{E}[ZZ^\top]O = O^\top \Sigma O = D$), d.h. $W_1,\dots,W_{r-1} \sim N(0,1)$, $W_r = 0$ und unabhängig:

$$|Y_n|^2 \xrightarrow{d} |Z|^2 = |W|^2 = \sum_{j=1}^{r-1} W_j^2 \sim \chi^2(r-1)$$

nach Definition der χ^2-Verteilung. Dies zeigt die Behauptung.

Es bleibt, $Y_n \xrightarrow{d} N(0,\Sigma)$ mittels charakteristischer Funktionen nachzuweisen (beachte, dass die X_j nicht unabhängig sind!). Für die Berechnung verwenden wir die verallgemeinerte binomische Formel $(a_1 + \cdots + a_r)^n = \sum_k \frac{n!}{k_1!\cdots k_r!} a_1^{k_1}\cdots a_r^{k_r}$ mit $a_j \in \mathbb{C}$ und Summation über $k = (k_1,\dots,k_r) \in \mathbb{N}_0^r$ mit $\sum_j k_j = n$, vergleiche die Multinomialzähldichte. Wir erhalten für $u \in \mathbb{R}^r$ und $v = (u_1/\sqrt{np_1^0},\dots,u_r/\sqrt{np_r^0})^\top$

$$\begin{aligned}
\varphi^{Y_n}(u) &= \mathbb{E}\left[e^{i\langle u, Y_n\rangle}\right] = \mathbb{E}\left[e^{i(v_1X_1+\cdots+v_rX_r)}\right] e^{-in(v_1p_1^0+\cdots+v_rp_r^0)} \\
&= \sum_k e^{i(v_1k_1+\cdots+v_rk_r)} \frac{n!}{k_1!\cdots k_r!} (p_1^0)^{k_1}\cdots(p_r^0)^{k_r} e^{-in(v_1p_1^0+\cdots+v_rp_r^0)} \\
&= \left(e^{iv_1}p_1^0 + \cdots + e^{iv_r}p_r^0\right)^n e^{-in(v_1p_1^0+\cdots+v_rp_r^0)}.
\end{aligned}$$

Eine Taylorentwicklung der Exponentialfunktionen zeigt also für $n \to \infty$

$$\begin{aligned}
\varphi^{Y_n}(u) &= \Big(\sum_{j=1}^r \big(e^{iu_j/(np_j^0)^{1/2}} p_j^0\big) e^{-i\sum_{j=1}^r u_j (p_j^0/n)^{1/2}}\Big)^n \\
&= \Big(\Big(\sum_{j=1}^r p_j^0\big(1 + iu_j(p_j^0)^{-1/2}n^{-1/2} - u_j^2(2p_j^0)^{-1}n^{-1} + O(n^{-3/2})\big)\Big) \\
&\quad \times \Big(1 - n^{-1/2}\sum_{j=1}^r iu_j(p_j^0)^{1/2} - (2n)^{-1}\Big(\sum_{j=1}^r u_j(p_j^0)^{1/2}\Big)^2 + O(n^{-3/2})\Big)\Big)^n \\
&= \Big(1 + \frac{1}{2n}\Big(\sum_{j=1}^r u_j(p_j^0)^{1/2}\Big)^2 - \frac{1}{2n}\sum_{j=1}^r u_j^2 + O(n^{-3/2})\Big)^n \\
&\to \exp\Big(-\frac{1}{2}\Big(\sum_{j=1}^r u_j^2 - \Big(\sum_{j=1}^r u_j(p_j^0)^{1/2}\Big)^2\Big)\Big).
\end{aligned}$$

Die quadratische Form im Exponenten ist gerade gleich $-\frac{1}{2}(|u|^2 - \langle u, w\rangle^2) = -\frac{1}{2}\langle \Sigma u, u\rangle$. Mit Blick auf Lemma 4.37 haben wir damit $\varphi^{Y_n}(u) \to \varphi^{N(0,\Sigma)}(u)$ gezeigt. Nach dem Stetigkeitssatz von Lévy gilt also $Y_n \xrightarrow{d} N(0,\Sigma)$. □

5.26 Bemerkung. Die geneigten Leser mögen einen Blick auf Abschnitte 11.1 und 11.2 im Georgii werfen, wo diese Beweise elementarer (z.B. ohne charakteristische Funktionen), aber sehr viel aufwändiger geführt werden. Im χ^2-Anpassungstest können wir jetzt also zu einem gegebenen *asymptotischen* Niveau α einfach den kritischen Wert $\kappa_\alpha > 0$ mit $P(Y > \kappa_\alpha) = \alpha$ für $Y \sim \chi^2(r-1)$ wählen (κ_α is das $(1-\alpha)$-Quantil der $\chi^2(r-1)$-Verteilung). Überraschenderweise hängt die asymptotische Verteilung nicht von p^0 ab. Man sagt, der χ^2-Anpasssungstest sei asymptotisch *verteilungsfrei*.

5.27 Korollar. *Der χ^2-Anpassungstest mit dem $(1-\alpha)$-Quantil κ_α der $\chi^2(r-1)$-Verteilung hat asymptotisches Niveau $\alpha \in (0,1)$:*

$$\lim_{n\to\infty} P_{\mathrm{Mult}(n,r,p^0)}(V_n^2 > \kappa_\alpha) = \alpha.$$

Beweis. Dies folgt direkt aus dem Satz, wenn man beachtet, dass die $\chi^2(r-1)$-Verteilung eine stetige Verteilungsfunktion besitzt, so dass Konvergenz in Verteilung die Konvergenz der Verteilungsfunktionen an jedem Punkt impliziert. □

5.28 Beispiel. Klassisch ist die Verwendung des χ^2-Anpassungstests zur Analyse des grundlegenden Experiments zur Vererbungslehre. Mendel beobachtete bei den Früchten bestimmter Erbsenpflanzen die Ausprägungen rund (A) und kantig (a) sowie gelb (B) und grün (b). Gemäß der Theorie werden die Merkmale A und B dominant, die Merkmale a und b rezessiv vererbt. Die Genotypen 'AA', 'Aa', 'aA' führen also zum Phänotyp A und der Genotyp 'aa' zum Phänotyp a. Entsprechendes gilt für das Farbmerkmal. Betrachtet man Nachkommen des (heterozygoten) Genotyps 'AaBb', so sollten daher die Merkmale AB, Ab, aB, ab im Verhältnis 9:3:3:1 auftreten. Mendel (1865) publizierte folgende Daten bei $n = 556$ Erbsenpflanzen:

Merkmale	gelb	grün
rund	315	108
kantig	101	32

Wir wenden den χ^2-Anpassungstest mit $r = 4$ und Klassenwahrscheinlichkeiten $p^0 = (9/16, 3/16, 3/16, 1/16)$ an. Wir wählen $\alpha = 0,1$ (recht groß, um Abweichungen von der Theorie möglichst sicher aufzudecken) und erhalten $\kappa_\alpha = 6,3$ als $0,9$-Quantil der $\chi^2(3)$-Verteilung (Statistik-Software). Einsetzen der Daten liefert $V_n^2 = 0,47$, so dass der Test die Vererbungstheorie akzeptiert. Die Übereinstimmung der Daten ist so groß, dass sogar ein Test zum Niveau 90% die Null-Hypothese akzeptiert hätte. Daher wurde und wird spekuliert, ob Mendel die Daten manipuliert haben könnte, allerdings wurde der χ^2-Test von Pearson erst im Jahr 1900 entwickelt, diese Analyse stand Mendel also nicht zur Verfügung.

Ende 20. Vorlesung

5.3 Einführung in die Schätztheorie

5.29 Bemerkung. Häufig ist ein Modell bis auf gewisse Parameter unbekannt, und anhand von Daten sollen diese Parameter geschätzt werden. Im folgenden werden wir nur das Problem betrachten, wie ein reeller Parameter geschätzt werden kann. Dabei stellt sich die grundsätzliche Frage, welches von vielen möglichen Schätzverfahren minimalen Schätzfehler aufweist. Zu diesem Zweck muss natürlich zunächst erst einmal der Begriff eines Schätzverfahrens und seines Schätzfehlers geklärt werden. Wir beschränken uns auf die Untersuchung des mittleren quadratischen Fehler.

5.30 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit Parametermenge $\Theta \subseteq \mathbb{R}$. Jede messbare Funktion $\hat{\vartheta} : \mathcal{X} \to \mathbb{R}$ heißt Schätzer (oder Schätzverfahren, Schätzmethode) von ϑ. Für eine Realisierung (konkrete Beobachtung, Stichprobe) $x \in \mathcal{X}$ ist $\hat{\vartheta}(x)$ der zugehörige Schätzwert.

Der mittlere quadratische Fehler MSE (*mean squared error*, quadratisches Risiko) eines Schätzers $\hat{\vartheta}$ von ϑ ist gegeben durch

$$R(\hat{\vartheta}, \vartheta) := \mathbb{E}_\vartheta[(\hat{\vartheta} - \vartheta)^2] := \int_{\mathcal{X}} (\hat{\vartheta}(x) - \vartheta)^2 P_\vartheta(dx) \in [0, \infty], \quad \vartheta \in \Theta.$$

5.31 Bemerkung. Als Schätzer wird also zunächst jede beliebige (messbare) Methode zugelassen, die die Beobachtungen x auf die reellen Zahlen abbildet. Für den MSE betrachten wir die quadratische Abweichung des Schätzwertes $\hat{\vartheta}(x)$ vom wahren (aber unbekannten) Parameter ϑ. Diese Abweichung $(\hat{\vartheta}(x) - \vartheta)^2$ ist allerdings zufällig, und der MSE ist der Erwartungswert der Abweichung unter dem wahren (datenerzeugenden) Wahrscheinlichkeitsmaß P_ϑ. Insbesondere hängt der Schätzfehler nicht nur von der Schätzmethode, sondern im Allgemeinen auch vom wahren, aber unbekannten Parameter ab.

5.32 Beispiel (Normalverteilungsmodell)**.** Für ein n-fach wiederholtes physikalisches Experiment, wo die Messfehler als $N(0, \sigma^2)$-verteilt für ein (bekanntes) $\sigma > 0$ angenommen werden können, ist $(\mathbb{R}^n, \mathfrak{B}_{\mathbb{R}^n}, (P_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta = \mathbb{R}$ und $P_\vartheta = N(\vartheta, \sigma^2)^{\otimes n}$ ein oft adäquates statistisches Modell für die Messergebnisse.

Schätzer von ϑ sind $\hat{\vartheta}_1(x) = 15$ (egal, welche Daten vorliegen, wird immer der Wert 15 geschätzt), $\hat{\vartheta}_2(x) = x_1$ (schätze ϑ durch den ersten Messwert), $\hat{\vartheta}_3(x) = \frac{1}{n}\sum_{i=1}^n x_i$ (schätze ϑ durch den Mittelwert der Messwerte), $\hat{\vartheta}_4(x) =$Median$(x_1, \ldots, x_n)$ (schätze ϑ durch den Median der Messwerte). Dann gilt $R(\hat{\vartheta}_1, \vartheta) = (15 - \vartheta)^2$, $R(\hat{\vartheta}_2, \vartheta) = \sigma^2$, $R(\hat{\vartheta}_3, \vartheta) = \frac{\sigma^2}{n}$ und $R(\hat{\vartheta}_4, \vartheta)$ ist nicht explizit berechenbar. Der MSE von $\hat{\vartheta}_2$ ist stets größer als der MSE von $\hat{\vartheta}_3$, aber je nach Parameterwert ϑ kann der MSE von $\hat{\vartheta}_1$ kleiner oder auch viel größer als der MSE von $\hat{\vartheta}_3$ sein.

5.33 Definition. Gilt $\hat{\vartheta} \in \mathscr{L}^1(P_\vartheta)$ für einen Schätzer $\hat{\vartheta}$ von ϑ (also $\mathbb{E}_\vartheta[|\hat{\vartheta}|] < \infty$), so heißt

$$B(\hat{\vartheta}, \vartheta) := \mathbb{E}_\vartheta[\hat{\vartheta}] - \vartheta$$

Bias (Verzerrung) von $\hat{\vartheta}$ bei ϑ. Gilt $B(\hat{\vartheta}, \vartheta) = 0$ (also $\mathbb{E}_\vartheta[\hat{\vartheta}] = \vartheta$) für alle $\vartheta \in \Theta$, so heißt der Schätzer $\hat{\vartheta}$ erwartungstreu (unverzerrt).

5.34 Lemma (Bias-Varianz-Zerlegung)**.** *Für einen Schätzer $\hat{\vartheta} \in \mathscr{L}^2(P_\vartheta)$ gilt die Bias-Varianz-Zerlegung des MSE*

$$R(\hat{\vartheta}, \vartheta) = B(\hat{\vartheta}, \vartheta)^2 + \mathrm{Var}_\vartheta(\hat{\vartheta}).$$

Insbesondere ist $R(\hat{\vartheta}, \vartheta) = \mathrm{Var}_\vartheta(\hat{\vartheta})$ für erwartungstreue Schätzer $\hat{\vartheta} \in \mathscr{L}^2(P_\vartheta)$.

Beweis. Wende die Bias-Varianz-Zerlegung aus Satz 3.15 mit $X = \hat{\vartheta}$, $x = \vartheta$ und $P = P_\vartheta$ an. □

5.35 Beispiele.

(a) Im Normalverteilungsbeispiel sind $\hat{\vartheta}_2$ und $\hat{\vartheta}_3$ offensichtlich erwartungstreu, auch der Median $\hat{\vartheta}_4$ ist erwartungstreu (folgt mit einem Symmetrieargument). Der Schätzer $\hat{\vartheta}_1$ ist nicht erwartungstreu (es gilt $\mathbb{E}_\vartheta[\hat{\vartheta}_1] = \vartheta$ nur im Fall $\vartheta = 15$). Die beiden Summanden in der Bias-Varianz-Zerlegung sind nicht-negativ, und $\hat{\vartheta}_2$, $\hat{\vartheta}_3$, $\hat{\vartheta}_4$ minimieren den Bias-Anteil $B(\hat{\vartheta}_i, \vartheta)^2$, während $\hat{\vartheta}_1$ den Varianzanteil $\mathrm{Var}_\vartheta(\hat{\vartheta}_4)$ minimiert.

(b) Wir beobachten unabhängige $X_1, \dots, X_n$ mit $X_i \sim U([0, \vartheta])$ für $\vartheta > 0$ unbekannt. Das statistische Modell kann als $(\mathbb{R}_+^n, \mathfrak{B}_{\mathbb{R}_+^n}, (U([0,\vartheta])^{\otimes n})_{\vartheta \in \Theta})$. Dies ist eine stetige Version des berühmten *Taxiproblems*, wo die Registriernummern von n Taxis am Times Square beobachtet wird und die Gesamtzahl der Taxis in Manhattan geschätzt werden soll.

Eine Idee für einen Schätzer ist, den Durchschnitt der X_i zu betrachten. Wegen $\mathbb{E}_\vartheta[\frac{1}{n}\sum_{i=1}^n X_i] = \vartheta/2$, wählen wir den erwartungstreuen Schätzer $\hat{\vartheta}_1 = \frac{2}{n}\sum_{i=1}^n X_i$. Es gilt dann

$$R(\hat{\vartheta}_1, \vartheta) = \frac{4}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{4}{n}\frac{\vartheta^2}{12} = \frac{1}{3n}\vartheta^2.$$

Alternativ können wir die Likelihoodfunktion

$$L(\vartheta) = \prod_{i=1}^n \vartheta^{-1}\mathbf{1}_{[0,\vartheta]}(X_i) = \vartheta^{-n}\mathbf{1}\Big(\max_{i=1,\dots,n} X_i \leqslant \vartheta\Big), \quad \vartheta > 0,$$

verwenden. Das Maximum-Likelihood-Prinzip führt auf den Maximum-Likelihood-Schätzer $\hat{\vartheta}_2 = \mathrm{argmax}_{\vartheta>0} L(\vartheta) = \max_{i=1,\dots,n} X_i$. Dieser ist nicht erwartungstreu, sein Bias ist $B(\hat{\vartheta}_2, \vartheta) = -\frac{1}{n+1}\vartheta$, seine Varianz ist $\mathrm{Var}_\vartheta(\hat{\vartheta}_2) = \frac{n}{(n+1)^2(n+2)}\vartheta^2$. Damit ergibt die Bias-Varianz-Zerlegung

$$R(\hat{\vartheta}_2, \vartheta) = \frac{1}{(n+1)^2}\vartheta^2 + \frac{n}{(n+1)^2(n+2)}\vartheta^2 = \frac{2}{(n+1)(n+2)}\vartheta^2.$$

Für große n sehen wir einen dramatischen Unterschied: $R(\hat{\vartheta}_2, \vartheta) \sim n^{-2}\vartheta^2 \ll n^{-1}\vartheta^2 \sim R(\hat{\vartheta}_1, \vartheta)$ und zwar für alle $\vartheta > 0$.

Wir können $\hat{\vartheta}_3$ weiter verbessern, indem wir eine Bias-Korrektur einführen und $\hat{\vartheta}_3 = \frac{n+1}{n}\hat{\vartheta}_2$ setzen. Aus den Rechnungen für $\hat{\vartheta}_2$ ergibt sich

$$R(\hat{\vartheta}_3, \vartheta) = \frac{(n+1)^2}{n^2}\frac{n}{(n+1)^2(n+2)}\vartheta^2 = \frac{1}{n(n+2)}\vartheta^2.$$

Wir sehen, dass $\hat{\vartheta}_3$ einen kleineren mittleren quadratischen Fehler als $\hat{\vartheta}_1$ und $\hat{\vartheta}_2$ besitzt. Ziel der mathematischen Statistik ist es, dies zu verstehen und zu klären, ob es noch bessere Schätzer geben kann.

5.36 Bemerkung. Oft ist (näherungsweise) Erwartungstreue eine gewünschte Eigenschaft, die insbesondere artifizielle Schätzer wie $\hat{\vartheta}_1$ im Normalverteilungsmodell ausschließt. Es sei jedoch darauf hingewiesen, dass in hochdimensionalen Problemen ($\Theta \subseteq \mathbb{R}^d$ mit d groß) meist der Varianzanteil im Schätzfehler überwiegt und auf Erwartungstreue verzichtet werden muss.

Für die Klasse der erwartungstreuen Schätzer können wir untere Schranken für den MSE beweisen. Wenn der MSE eines Schätzers diese untere Schranke erreicht, ist dieser also optimal. Wenn man bedenkt, dass wir jede messbare Funktion der Daten als Schätzer zulassen, ist allein die Existenz solcher unteren Schranken überraschend.

5.37 Lemma (Chapman-Robbins-Ungleichung)**.** *Es sei $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$, $\Theta \subseteq \mathbb{R}$, ein statistisches Modell mit Likelihood-Funktion $L(\vartheta)$ bezüglich einem Maß μ. Ist $\hat{\vartheta}$ ein erwartungstreuer Schätzer von ϑ, so gilt für alle $\vartheta, \vartheta_0 \in \Theta$ mit $P_\vartheta \neq P_{\vartheta_0}$, $\forall x \in \mathcal{X} : L(\vartheta_0, x) = 0 \Rightarrow L(\vartheta, x) = 0$ und $L(\vartheta)/L(\vartheta_0) \in \mathscr{L}^2(P_{\vartheta_0})$*

$$R(\hat{\vartheta}, \vartheta_0) \geqslant \frac{(\vartheta - \vartheta_0)^2}{\operatorname{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0))}.$$

5.38 Bemerkung. Gemäß der Konvention in der Maßtheorie setzen wir im Fall $L(\vartheta_0, x) = L(\vartheta, x) = 0$ den Quotienten $L(\vartheta, x)/L(\vartheta_0, x) := 0$. Man kann sich überlegen, dass die Bedingung $\forall x \in \mathcal{X} : L(\vartheta_0, x) = 0 \Rightarrow L(\vartheta, x) = 0$ äquivalent ist zur Absolutstetigkeit $P_\vartheta \ll P_{\vartheta_0}$ gemäß Bemerkung 1.59.

Beweis. Der Beweis fußt auf einer geschickten Anwendung der Cauchy-Schwarz-Ungleichung in $\mathscr{L}^2(P_{\vartheta_0})$. Da $\hat{\vartheta}$ erwartungstreu ist, gilt nach Definition der Likelihood-Funktion

$$\begin{aligned}
\vartheta - \vartheta_0 &= \mathbb{E}_\vartheta[\hat{\vartheta} - \vartheta_0] - \mathbb{E}_{\vartheta_0}[\hat{\vartheta} - \vartheta_0] \\
&= \int_{\mathcal{X}} (\hat{\vartheta}(x) - \vartheta_0)(L(\vartheta, x) - L(\vartheta_0, x))\,\mu(dx) \\
&= \int_{\{x \,|\, L(\vartheta_0,x)>0\}} (\hat{\vartheta}(x) - \vartheta_0)\Big(\frac{L(\vartheta, x)}{L(\vartheta_0, x)} - 1\Big) L(\vartheta_0, x)\,\mu(dx) \\
&= \mathbb{E}_{\vartheta_0}\Big[(\hat{\vartheta} - \vartheta_0)\Big(\frac{L(\vartheta)}{L(\vartheta_0)} - 1\Big)\Big].
\end{aligned}$$

Beachte dabei, dass wir $\{x \,|\, L(\vartheta, x) > 0\} \subseteq \{x \,|\, L(\vartheta_0, x) > 0\}$ verwendet haben. Aus der Cauchy-Schwarz-Ungleichung folgt daher

$$(\vartheta - \vartheta_0)^2 \leqslant \mathbb{E}_{\vartheta_0}\Big[(\hat{\vartheta} - \vartheta_0)^2\Big]\,\mathbb{E}_{\vartheta_0}\Big[\Big(\frac{L(\vartheta)}{L(\vartheta_0)} - 1\Big)^2\Big] = R(\hat{\vartheta}, \vartheta_0)\operatorname{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0)),$$

wobei wir $\mathbb{E}_{\vartheta_0}[L(\vartheta)/L(\vartheta_0)] = \int (L(\vartheta)/L(\vartheta_0))L(\vartheta_0)\,d\mu = \int L(\vartheta)\,d\mu = 1$ benutzt haben. Wäre $\mathrm{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0)) = 0$, so würde $L(\vartheta)/L(\vartheta_0) = 1$ P_{ϑ_0}-fast sicher gelten, was aber der Annahme $P_\vartheta \neq P_{\vartheta_0}$ widerspricht. Daher können wir durch $\mathrm{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0))$ teilen und erhalten die Chapman-Robbins-Ungleichung. □

5.39 Beispiele.

(a) Im Normalverteilungsbeispiel ist die Likelihoodfunktion $L(\vartheta, x) = (2\pi\sigma^2)^{-n/2}\exp(-\sum_{i=1}^n (x_i-\vartheta)^2/(2\sigma^2))$ bezüglich dem Lebesguemaß μ. Insbesondere ist $L(\vartheta, x) > 0$ für alle $\vartheta \in \mathbb{R}$, $x \in \mathbb{R}^n$. Weiterhin gilt

$$\begin{aligned}\frac{L(\vartheta)}{L(\vartheta_0)} &= \exp\Big(\frac{1}{2\sigma^2}\sum_{i=1}^n \Big((X_i-\vartheta_0)^2 - (X_i-\vartheta)^2\Big)\Big)\\ &= \exp\Big(\frac{1}{\sigma^2}\sum_{i=1}^n (\vartheta-\vartheta_0)X_i - \frac{n}{2\sigma^2}(\vartheta^2-\vartheta_0^2)\Big).\end{aligned}$$

Da Exponentialfunktionen bezüglich der Normalverteilungsdichte integrierbar sind, gilt $L(\vartheta)/L(\vartheta_0) \in \mathscr{L}^2(P_{\vartheta_0})$. Die Voraussetzungen der Chapman-Robbins-Ungleichung sind daher erfüllt. Schreiben wir $X_i = \vartheta_0 + \sigma Z_i$ mit unabhängigen $Z_i \sim N(0,1)$ unter P_{ϑ_0}, so erhalten wir

$$\begin{aligned}\mathbb{E}_{\vartheta_0}\Big[\Big(\frac{L(\vartheta)}{L(\vartheta_0)}\Big)^2\Big] &= \mathbb{E}_{\vartheta_0}\Big[\exp\Big(\frac{2}{\sigma^2}\sum_{i=1}^n(\vartheta-\vartheta_0)X_i\Big)\Big]\exp\Big(-\frac{n(\vartheta^2-\vartheta_0^2)}{\sigma^2}\Big)\\ &= \mathbb{E}\Big[\exp\Big(2\sum_{i=1}^n\frac{\vartheta-\vartheta_0}{\sigma}Z_i\Big)\Big]\exp\Big(-\frac{n(\vartheta-\vartheta_0)^2}{\sigma^2}\Big)\\ &= \mathbb{E}\Big[\exp\Big(2\frac{\vartheta-\vartheta_0}{\sigma}Z_1\Big)\Big]^n\exp\Big(-\frac{n(\vartheta-\vartheta_0)^2}{\sigma^2}\Big)\\ &= \exp\Big(\frac{n(\vartheta-\vartheta_0)^2}{\sigma^2}\Big),\end{aligned}$$

wobei wir die Formel $\mathbb{E}[e^{\gamma Z}] = e^{\gamma^2/2}$ für $\gamma \in \mathbb{R}$, $Z \sim N(0,1)$ benutzt haben. Wegen $\mathbb{E}_{\vartheta_0}[L(\vartheta)/L(\vartheta_0)] = 1$ zeigt die Chapman-Robbins-Ungleichung also die (über $\vartheta \neq \vartheta_0$ maximierte) untere Schranke

$$R(\hat\vartheta, \vartheta_0) \geqslant \sup_{\vartheta\in\mathbb{R}\setminus\{\vartheta_0\}} \frac{(\vartheta-\vartheta_0)^2}{\exp(n(\vartheta-\vartheta_0)^2\sigma^{-2})-1} = \frac{\sigma^2}{n}.$$

Das Supremum haben wir dabei erhalten für $(\vartheta-\vartheta_0)^2 \to 0$ (Berechnung mit L'Hopital-Regel). Der MSE jedes erwartungstreuen Schätzers ist also mindestens $\frac{\sigma^2}{n}$. Das Stichprobenmittel $\hat\vartheta_3$ besitzt genau diesen MSE und ist also (im Normalverteilungsmodell!) optimal. Der Stichprobenmedian $\hat\vartheta_4$ besitzt übrigens einen leicht höheren MSE, hat aber bessere *Robustheitseigenschaften*, falls die Fehler nicht entsprechend einer Normalverteilung um den Erwartungswert streuen.

(b) Beim Taxiproblem betrachten wir $0 < \vartheta < \vartheta_0$, so dass die Bedingung $L(\vartheta_0, x) = 0 \iff \max_i x_i > \vartheta_0 \Rightarrow L(\vartheta, x) = 0$ für $x \in \mathbb{R}_+^n$ folgt. Dann

gilt

$$\mathrm{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0)) = \mathrm{Var}_{\vartheta_0}((\vartheta_0/\vartheta)^n \mathbf{1}(\max_i X_i \leqslant \vartheta)) = (\vartheta_0/\vartheta)^n - 1.$$

Wir können die Chapman-Robbins-Ungleichung anwenden und erhalten für $n \geqslant 2$

$$R(\hat{\vartheta}, \vartheta_0) \geqslant \sup_{\vartheta < \vartheta_0} \frac{(\vartheta - \vartheta_0)^2}{(\vartheta_0/\vartheta)^n - 1} \geqslant \frac{n^{-2}}{(1-1/n)^{-n} - 1}\vartheta_0^2 \approx \frac{1}{e-1} n^{-2}\vartheta_0^2,$$

wobei wir $\vartheta = (1 - \frac{1}{n})\vartheta_0$ eingesetzt haben. Diese untere Schranke ist kleiner als $R(\hat{\vartheta}_3, \vartheta_0) = \frac{1}{n(n+2)}\vartheta_0^2$ oben, aber die Asymptotik $n^{-2}\vartheta_0^2$ stimmt überein. In der Tat kann die mathematische Statistik mit anderen Methoden beweisen, dass $\hat{\vartheta}_3$ den kleinsten Fehler unter allen erwartungstreuen Schätzern besitzt.

5.40 Bemerkung. Ist die Likelihoodfunktion $L(\vartheta)$ fast sicher differenzierbar in ϑ (wie im Normalverteilungsmodell, aber nicht beim Taxiproblem), so wird die untere Schranke von Chapman-Robbins meist maximal für $\vartheta \to \vartheta_0$. Falls dann Grenzwert und Erwartungswert vertauscht werden können, erhalten wir für $\vartheta \to \vartheta_0$

$$\begin{aligned}
\frac{(\vartheta - \vartheta_0)^2}{\mathrm{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0))} &= \Big(\mathbb{E}_{\vartheta_0} \Big[\Big(\frac{L(\vartheta) - L(\vartheta_0)}{L(\vartheta_0)(\vartheta - \vartheta_0)} \Big)^2 \Big] \Big)^{-1} \\
&\to \Big(\mathbb{E}_{\vartheta_0} \Big[\Big(\frac{\frac{d}{d\vartheta} L(\vartheta_0)}{L(\vartheta_0)} \Big)^2 \Big] \Big)^{-1} \\
&= \Big(\mathbb{E}_{\vartheta_0} \Big[\Big(\frac{d}{d\vartheta} \log(L(\vartheta_0)) \Big)^2 \Big] \Big)^{-1} =: \frac{1}{I(\vartheta_0)}.
\end{aligned}$$

Die Zahl

$$I(\vartheta_0) = \mathbb{E}_{\vartheta_0} \Big[\Big(\frac{d}{d\vartheta} \log(L(\vartheta_0)) \Big)^2 \Big]$$

heißt Fisher-Information des Modells bei ϑ_0 und besitzt viele für ein Informationsmaß wünschenswerte Eigenschaften. So ist beispielsweise die Fisher-Information beim n-fachen Produktmodell gerade n-mal die Fisher-Information beim einfachen Modell. Dies und die erforderliche Regularität des statistischen Modells wird in Abschnitt 7.5 bei Georgii (elementarer auch in Abschnitt 4.5 bei Krengel) dargestellt. Wir geben ohne Beweis die daraus erhaltene berühmte untere Schranke an.

5.41 Satz (Cramér-Rao-Unlgeichung, Informationsungleichung)**.** *In einem regulären statistischen Modell $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta \in \Theta})$, $\Theta \subseteq \mathbb{R}$, mit Likelihood-Funktion $L(\vartheta)$ und Fisher-Information $I(\vartheta)$ gilt für den MSE jedes erwartungstreuen Schätzers $\hat{\vartheta}$ von ϑ*

$$\forall \vartheta \in \Theta : \; R(\hat{\vartheta}, \vartheta) \geqslant \frac{1}{I(\vartheta)}.$$

Beweis. Satz 7.19 in Georgii mit $\tau(\vartheta) = \vartheta$. □

5.42 Beispiel. Im Normalverteilungsmodell $(\mathbb{R}^n, \mathfrak{B}_{\mathbb{R}^n}, (N(\vartheta, \sigma^2)^{\otimes n})_{\vartheta\in\mathbb{R}})$ mit $\sigma > 0$ bekannt gilt

$$\frac{d}{d\vartheta}\log(L(\vartheta)) = \frac{d}{d\vartheta}\Big(- n\log(2\pi\sigma^2)/2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\vartheta)^2\Big) = \sum_{i=1}^n \frac{X_i-\vartheta}{\sigma^2}.$$

Wegen Unabhängigkeit der X_i erhalten wir die Fisher-Information

$$I(\vartheta) = \mathbb{E}_\vartheta\Big[\Big(\frac{d}{d\vartheta}\log(L(\vartheta))\Big)^2\Big] = \mathrm{Var}_\vartheta\Big(\sum_{i=1}^n \frac{X_i-\vartheta}{\sigma^2}\Big) = \sigma^{-4}\sum_{i=1}^n \mathrm{Var}_\vartheta(X_i) = \frac{n}{\sigma^2},$$

entsprechend der obigen Herleitung aus der Chapman-Robbins-Ungleichung. Das Normalverteilungsmodell erfüllt auch die notwendigen Regularitätsbedingungen, so dass die Cramér-Rao-Ungleichung direkt anwendbar ist.

Auch im Binomialmodell $(\{0,\dots,n\}, \mathscr{P}(\{0,\dots,n\}), (\mathrm{Bin}(n,\vartheta))_{\vartheta\in(0,1)})$ lässt sich nachrechnen, dass das Stichprobenmittel (die relative Häufigkeit) $\hat\vartheta = X/n$ ein erwartungstreuer Schätzer von ϑ ist, dessen MSE $R(\hat\vartheta,\vartheta) = \frac{\vartheta(1-\vartheta)}{n} = I(\vartheta)^{-1}$ minimal ist. Beachte, dass hier der MSE und die Fisher-Information von ϑ abhängen.