Stochastik I
Skript zur Vorlesung
im Sommersemester 2020

Prof. Dr. Markus Reiß
Humboldt-Universität zu Berlin

Version vom 13. Juli 2020

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 Interpretation $(b_1, \ldots, b_n) \in \Omega$ kodiert Ergebnis in den Würfen 1 bis n via $b_i = 1$ für Kopf im i-ten Wurf, $b_i = 0$ für Zahl im i-ten Wurf. Wir betrachten die Ereignisse $A_1 = \{(1, \ldots, 1)\}$ (n-mal Kopf), $A_2 = \{b \in \Omega \,|\, b_1 + \cdots + b_n = n-1\}$ ($(n-1)$-mal Kopf), $A_3 = \{b \in \Omega \,|\, b_1 + \cdots + b_n \leqslant n - 1\}$ (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 $A_3 = A_1^{\complement} := \Omega \setminus A_1$ und $P(A_1) + P(A_3) = 1$ gilt. A_1 und A_3 heißen auch <u>komplementäre Ereignisse</u> oder <u>Gegenereignisse</u>.

(c) Wir nehmen nun an, dass die Münze beliebig oft hintereinander geworfen wird. Die Versuchsausgänge modellieren wir dann durch 0-1-Folgen $B = (b_n)_{n\in\mathbb{N}} = (b_1, b_2, \ldots)$, d.h. $\Omega = \{0, 1\}^{\mathbb{N}}$. Ereignisse sind dann $A_1 = \{(1, 1, \ldots)\}$ (immer nur Kopf), $A_2 = \{b \in \Omega \,|\, \forall M \in \mathbb{N} \,\exists k \in \mathbb{N} : b_k = b_{k+1} = \cdots = b_{k+M-1} = 1\}$ (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. $A_n \subseteq A_{n+1}$, $\bigcup_n A_n = A$)*

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

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 $B_n \times \{0,1\} \times \{0,1\} \times \cdots = \{(\omega_i)_{i\geqslant 1} \,|\, (\omega_1,\dots,\omega_n) \in B_n, \omega_i \in \{0,1\}, i > n\}$ 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 (gehe zu Komplementen über!) $P(A_1) = 0$ 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.

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 Wahrscheinlichkeitsram $(\Omega, \mathscr{F}, P)$ 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 $P^X = P(A^{\complement})\delta_0 + P(A)\delta_1$.

▷ **Kontrollfragen**

(a) $\mathscr{F} = \{\varnothing, \mathbb{Q}, \mathbb{R} \setminus \mathbb{Q}, \mathbb{R}\}$ bildet eine σ-Algebra auf den reellen Zahlen $\mathbb{R}$. Beweis? Welche Funktionen $g : \mathbb{R} \to \mathbb{R}$ sind $(\mathscr{F}, \mathscr{P}(\mathbb{R}))$-messbar?

Es gilt $\mathbb{R} \in \mathscr{F}$, und $\mathscr{F}$ ist abgeschlossen unter Komplementen und Vereinigungen, so dass $\mathscr{F}$ eine σ-Algebra ist. Jede Funktion $f(x) = a\mathbf{1}_{\mathbb{Q}}(x) + b\mathbf{1}_{\mathbb{R} \setminus \mathbb{Q}}(x)$, $x \in \mathbb{R}$, für beliebige $a, b \in \mathbb{R}$ ist $(\mathscr{F}, \mathscr{P}(\mathbb{R}))$-messbar, weil $f^{-1}(A) \in \{\mathbb{Q}, \mathbb{R} \setminus \mathbb{Q}, \mathbb{R}\ \varnothing\}$ gilt, je nachdem, ob nur $a \in A$, nur $b \in A$, $\{a, b\} \subseteq A$ oder $\{a, b\} \cap A = \varnothing$ gilt. Jede $(\mathscr{F}, \mathscr{P}(\mathbb{R}))$-messbare Funktion muss diese Gestalt haben, weil für eine Funktion, die auf $\mathbb{Q}$ (oder $\mathbb{R} \setminus \mathbb{Q}$) zwei oder mehr Werte annimmt, das Urbild eines dieser Werte eine echte Teilmenge von $\mathbb{Q}$ (bzw. $\mathbb{R} \setminus \mathbb{Q}$) und damit nicht in $\mathscr{F}$ wäre.

(b) Geben Sie alle möglichen σ-Algebren auf der Grundmenge $\Omega = \{0, 1, 2\}$ an.

$\mathscr{F}_0 = \{\varnothing, \Omega\}$, $\mathscr{F}_i = \{\{i-1\}, \{i-1\}^c\} \cup \mathscr{F}_0$ für $i = 1, 2, 3$ und $\mathscr{F}_4 = \mathscr{P}(\Omega)$.

(c) Beweisen Sie formal, dass P^X wieder ein Wahrscheinlichkeitsmaß ist.

Da X messbar ist, ist P^X wohldefiniert. P^X ist Wahrscheinlichkeitsmaß wegen $P^X(S) = P(X^{-1}(S)) = P(\Omega) = 1$ sowie (Mengenlehre!) für paarweise disjunkte $A_n \in \mathscr{S}$

$$\begin{aligned} P^X\Big(\bigcup_{n \geqslant 1} A_n\Big) &= P\Big(X^{-1}\Big(\bigcup_{n \geqslant 1} A_n\Big)\Big) = P\Big(\bigcup_{n \geqslant 1} X^{-1}(A_n)\Big) \\ &= \sum_{n \geqslant 1} P(X^{-1}(A_n)) = \sum_{n \geqslant 1} P^X(A_n). \end{aligned}$$

(d) Wieso ist beim 3-fachen fairen Münzwurf die Anzahl von 'Kopf' eine Zufallsvariable? Was ist die Verteilung der 'Kopf'-Anzahl? (saubere Modellierung!)

$\Omega = \{0, 1\}^3$, $\mathscr{F} = \mathscr{P}(\Omega)$, $X : \Omega \to \mathbb{R}$, $X(\omega_1, \omega_2, \omega_3) = \omega_1 + \omega_2 + \omega_3$. X ist messbar und damit Zufallsvariable, weil $\mathscr{F}$ die Potenzmenge ist (klar?). Es gilt

$$\begin{aligned} &P^X(\{0\}) = P(\{(0,0,0)\}) = \tfrac{1}{8}, \quad P^X(\{3\}) = P^X(\{0\}), \\ &P^X(\{1\}) = P(\{(1,0,0), (0,1,0), (0,0,1)\}) = \tfrac{3}{8}, \quad P^X(\{2\}) = P^X(\{1\}) \end{aligned}$$

und für allgemeine Mengen $A \subseteq \mathbb{R}$ gilt $P^X(A) = \frac{1}{8}(\mathbf{1}_A(0) + \mathbf{1}_A(3)) + \frac{3}{8}(\mathbf{1}_A(1) + \mathbf{1}_A(2))$.

(e) Wie kann man durch ein (intuitives) Grenzwertargument die Wahrscheinlichkeit von A_2 in Beispiel 1.1(c) bestimmen?

Es gilt $A_2 = \bigcap_{M \geqslant 1} \bigcup_{k \geqslant 1} B_{k,M}$ mit $B_{k,M} = \{b \in \Omega \,|\, b_k = \cdots = b_{k+M-1} = 1\}$. Wegen σ-Stetigkeit gilt also $P(A_2) = \lim_{M \to \infty} P(\bigcup_{k \geqslant 1} B_{k,M})$. Wir wollen $P(\bigcup_{k \geqslant 1} B_{k,M}) = 1$ für alle M zeigen, wofür wegen Monotonie $P(\bigcup_{j \geqslant 1} B_{jM,M}) = 1$ ausreicht (beachte, dass die runs in $B_{jM,M}$ nicht überlappen!). Wegen σ-Stetigkeit reicht es also, $\lim_{J \to \infty} P(\bigcup_{j=1}^J B_{jM,M}) = 1$ zu zeigen. Betrachte dazu das Komplement $\bigcap_{j=1}^J (B_{jM,M})^c$, also dass jeweils zwischen jM und $(j+1)M - 1$, $1 \leqslant j \leqslant J$, kein run vorliegt.

Dies hängt nur von den ersten JM Münzwürfen ab. Die dafür günstigen Möglichkeiten sind alle Würfe, wo die ersten M Würfe nicht alle gleich '1' sind, die nächsten M Würfe nicht alle gleich '1' sind usw. Es gilt deshalb $P(\bigcap_{j=1}^J (B_{jM,M})^c) = \frac{\prod_{j=1}^J (2^M-1)}{2^{JM}} = (1-2^{-M})^J$. Wir erhalten somit

$$\begin{aligned} P(A_2) &= \lim_{M\to\infty} P(\bigcup_{k\geqslant 1} B_{k,M}) \geqslant \lim_{M\to\infty} \lim_{J\to\infty} P(\bigcup_{j=1}^J B_{jM,M}) \\ &= \lim_{M\to\infty} \lim_{J\to\infty} (1-(1-2^{-M})^J) = 1. \end{aligned}$$

Später werden wir dies mit dem Borel-Cantelli-Lemma einfacher herleiten können.

Ende der 1. Vorlesung ______________________________

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 endliche oder abzählbar unendliche 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 $1, 2, \dots, N$. Es werden n Kugeln gezogen. Dann gilt für die Anzahl verschiedener Versuchsausgänge:*

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

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

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

Mit Zurücklegen, ohne Betrachtung der Reihenfolge:
$\Omega_4 = \{(s_1, \dots, s_n) \mid 1 \leqslant s_1 \leqslant s_2 \leqslant \dots \leqslant s_n \leqslant N\}$, $|\Omega_4| = \binom{N+n-1}{n}$.

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 $\Omega = \{0, \dots, W\}$ gegeben durch die Zähldichte

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

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

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

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 $\Omega = \{0, \dots, n\}$ ist gegeben durch die Zähldichte

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

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 $\{0, \dots, n\}$-wertige Zufallsvariable $X(\omega_1, \dots, \omega_n) = \omega_1 + \dots + \omega_n$. Dann gilt für $k \in \{0, \dots, n\}$

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

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

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

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

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

1.26 Beispiel. Die Multinomialverteilung $\mathrm{Mult}(n, r, p_1, \dots, p_r)$ zählt die Anzahl der Versuchsausgänge in r Klassen bei n Versuchen und Klassenwahrscheinlichkeiten $p_1, \dots, p_r$. Werden die Ziffern '0' bis '9' rein zufällig n-mal erzeugt, so kann man die erhaltenen Häufigkeiten der Ziffern mit der $\mathrm{Mult}(n, 10, \frac{1}{10}, \dots, \frac{1}{10})$-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 $\log((1-p_n)^n) = n\log(1-p_n) = -np_n + O(np_n^2)$):

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

Wir erhalten also $p_{Pois(\lambda)}(k)$, 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 heute neu Corona-Infizierten in Berlin.

▷ **Kontrollfragen**

(a) Die Zufallsvariable X modelliere die Augensumme von zwei fairen Würfeln, wie in Beispiel 1.11(a). Was ist p^X?

$p^X(2) = p^X(12) = \frac{1}{36}$, $p^X(3) = p^X(11) = \frac{2}{36}$, $p^X(4) = p^X(10) = \frac{3}{36}$, $p^X(5) = p^X(9) = \frac{4}{36}$, $p^X(6) = p^X(8) = \frac{5}{36}$, $p^X(7) = \frac{6}{36}$.

(b) Wieso ist $p_{Geo(p)}$ eine Zähldichte?

Offensichtlich ist $p_{Geo(p)}(k) \geqslant 0$. Außerdem folgt mit der geometrischen Reihe

$$\sum_{k\geqslant 1} p_{Geo(p)}(k) = p\sum_{k\geqslant 1}(1-p)^{k-1} = p\frac{1}{1-(1-p)} = 1.$$

(c) Finde weitere Anwendungsbeispiele für jede der angegebenen diskreten Verteilungen.

Beim Roulette ergibt sich die Gleichverteilung auf $\{0,\dots,36\}$. Wählen wir n der N Bundesbürger rein zufällig aus und schauen wir, wie viele eine bestimmte Krankheit besitzen, so ergibt sich die hypergeometrische Verteilung $\mathrm{Hyp}(N,W,n)$, wobei W die Anzahl der Bundesbürger mit dieser Krankheit bezeichnet. Spielen wir n-mal Roulette und setzen jeweils auf eine einzige Zahl, so ergibt sich das Bernoulli-Schema $\mathrm{Bern}(n, \frac{1}{37})$. Die Anzahl der Gewinne bei diesem Roulettespiel ist $\mathrm{Bin}(n, \frac{1}{37})$-verteilt. Zählen wir wie oben die Anzahl der kranken Bundesbürger, wobei wir diese 'mit Zurücklegen' ziehen (d.h. mit möglicherweise mehrfacher Auswahl), ergibt sich eine Binomialverteilung, nämlich $\mathrm{Bin}(N, W/N)$. Unterscheiden wir bei den kranken Bundesbürgen noch, ob sie krank mit oder ohne Symptome sind, ergibt sich eine Multinomialverteilung mit den $r = 3$ Klassen 'gesund', 'krank mit Symptomen', 'krank ohne Symptome'. Die Anzahl der S-Bahn-Fahrten bis zur ersten Kontrolle ist geometrisch verteilt. Die Anzahl der mit einem Geigerzähler gemessenen Atomzerfälle ist Poisson-verteilt.

(d) Zeige $A_n \asymp B_n, C_n \asymp D_n \Rightarrow A_nC_n \asymp B_nD_n$ und arbeite sauber alle Zwischenschritte im Beweis des Poissonschen Grenzwertsatzes aus.

Es gilt $\lim_{n\to\infty} \frac{A_nC_n}{B_nD_n} = \lim_{n\to\infty} \frac{A_n}{B_n} \lim_{n\to\infty} \frac{C_n}{D_n} = 1$. Verwenden wir dies mit $\frac{n(n-1)\cdots(n-k+1)}{n^k} \asymp 1$ und $\frac{(1-p_n)^n}{e^{np_n}} = \exp(n\log(1-p_n) - np_n) = \exp(O(np_n^2)) \asymp 1$, folgt die erste Asymptotik. Mit $(np_n)^k \asymp \lambda^k$ und $e^{-np_n} \asymp e^{-\lambda}$ ergibt sich damit die zweite.

(e) Zeichne einige $\mathrm{Bin}(n,p)$-Zähldichten und vergleiche mit den zugehörigen Poisson-Zähldichten.

Siehe z.B. den Eintrag 'Poisson-Approximation' in Wikipedia.

Ende der 2. Vorlesung ______________________________

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 $T_n(\omega) = T_n(\omega_1, \omega_2, \dots) = (\omega_1, \dots, \omega_{n-1}, 1-\omega_n, \omega_{n+1}, \dots)$ 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 $S = \{n_1, \dots, n_k\} \in \mathscr{S}$ 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. 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 $\mathfrak{B}_S := \sigma(\{O \subseteq S \,|\, O \text{ offen}\})$ 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) $\mathscr{E}_1 := \{(a, b) \,|\, a, b \in \mathbb{R}\}$;

(ii) $\mathscr{E}_2 := \{[a, b] \,|\, a, b \in \mathbb{R}\}$;

(iii) $\mathscr{E}_3 := \{(a, b] \,|\, a, b \in \mathbb{R}\}$;

(iv) $\mathscr{E}_4 := \{(-\infty, b] \,|\, b \in \mathbb{R}\}$;

(v) $\mathscr{E}_5 := \{(-\infty, b) \,|\, b \in \mathbb{R}\}$.

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

(i) $\mathscr{E}_1^d := \{(a_1, b_1) \times \cdots \times (a_d, b_d) \,|\, a_k, b_k \in \mathbb{R}, k = 1, \dots, d\}$;

(ii) $\mathscr{E}_2^d := \{[a_1, b_1] \times \cdots \times [a_d, b_d] \,|\, a_k, b_k \in \mathbb{R}, k = 1, \dots, d\}$;

(iii) $\mathscr{E}_3^d := \{(a_1, b_1] \times \cdots \times (a_d, b_d] \,|\, a_k, b_k \in \mathbb{R}, k = 1, \dots, d\}$;

(iv) $\mathscr{E}_4^d := \{(-\infty, b_1] \times \cdots \times (-\infty, b_d] \,|\, b_k \in \mathbb{R}, k = 1, \dots, d\}$;

(v) $\mathscr{E}_5^d := \{(-\infty, b_1) \times \cdots \times (-\infty, b_d) \,|\, b_k \in \mathbb{R}, k = 1, \dots, d\}$.

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

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 $F(x) = \mathbf{1}_{[x_0,\infty)}(x)$. Für $P = \sum_{n\geqslant 1} p_n \delta_{x_n}$ mit $p_n \geqslant 0$, $\sum_{n\geqslant 1} p_n = 1$ und $x_1 < x_2 < \cdots$ gilt $F(x) = \sum_{n=1}^{N} p_n$ für $x \in [x_N, x_{N+1})$ sowie $F(x) = 0$ für $x < x_1$, $F(x) = 1$ für $x \geqslant \lim_{n\to\infty} x_n$ (falls dieser Grenzwert endlich ist).

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 in Verbindung mit Komplementbildung impliziert die Rechtsstetigkeit von F wegen $(-\infty, x] = \bigcap_{n\geqslant 1}(-\infty, x_n]$ 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 $\bigcap_{n\geqslant 1}(-\infty, x_n] = \varnothing$ und $F(x_n) \to 1$ für $x_n \uparrow \infty$ aus $\bigcup_{n\geqslant 1}(-\infty, x_n] = \mathbb{R}$. □

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.

Ende der 3. Vorlesung

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

Betrachte

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

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

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

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. klar! □

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

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

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 $\text{Bin}(1,p)$-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 $\mathbf{1}(U > 1-p)$ $\text{Bin}(1,p)$-verteilt ist für eine $U((0,1))$-verteilte Zufallsvariable U.

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.

▷ **Kontrollfragen**

(a) Wieso ist jede monoton wachsende (nicht notwendigerweise stetige) Funktion $f:\mathbb{R}\to\mathbb{R}$ Borel-messbar? (Tipp: Korollar 1.41)

Nach Korollar 1.41(b) reicht es, $\{x\in\mathbb{R}\mid f(x)\leqslant a\}\in\mathfrak{B}_{\mathbb{R}}$ für alle $a\in\mathbb{R}$ nachzuweisen. Wegen Monotonie ist $\{x\in\mathbb{R}\mid f(x)\leqslant a\}$ ein Intervall $(-\infty,x_a]$ oder $(-\infty,x_a)$ mit $x_a=\sup\{x\in\mathbb{R}\mid f(x)\leqslant a\}$ (ggf. leer für $x_a=-\infty$), je nach dem, ob das Supremum angenommen wird oder nicht. Intervalle sind stets Borelmengen. Also ist f Borel-messbar.

(b) Zeichne die Verteilungsfunktionen von Beispiel 1.47 sowie von der $U([a,b])$- und der Dreiecks-Verteilung.

In Beispiel 1.47 ergeben sich stückweise konstante Verteilungsfunktionen mit Sprüngen der Höhe p_n bei x_n. Weiterhin ist $F_{U([a,b])}(x)=\frac{x-a}{b-a}$ für $x\in[a,b]$ und $F_{U([a,b])}(x)=0$ für $x<a$, $F_{U([a,b])}(x)=1$ für $x>b$. Die Verteilungsfunktion der Dreiecksverteilung ist null für $x\leqslant 0$, eins für $x\geqslant 2$, gleich der Parabel $\frac{1}{2}x^2$ für $x\in[0,1]$ und gleich der Parabel $1-\frac{1}{2}(2-x)^2$ für $x\in[1,2]$.

(c) Welche der folgenden Funktionen $f:\mathbb{R}\to\mathbb{R}$ sind Dichten: $f(x)=e^{-x}\mathbf{1}(x>0)$, $f(x)=\frac{1}{2}e^{-|x|}$, $f(x)=|\sin(x)/x|$? Wie lauten gegebenenfalls die Verteilungsfunktionen?

$f(x)=e^{-x}\mathbf{1}(x>0)$ ist Dichte ($\mathrm{Exp}(1)$-Verteilung) mit Verteilungsfunktion $(1-e^{-x})_+$, $f(x)=\frac{1}{2}e^{-|x|}$ ist Dichte (*Laplace-Verteilung*) mit Verteilungsfunktion $\frac{1}{2}e^{x}\mathbf{1}_{(-\infty,0]}(x)+(1-\frac{1}{2}e^{-x})\mathbf{1}_{(0,\infty)}(x)$. $f(x)=|\sin(x)/x|$ ist zwar Borel-messbar und nicht-negativ, aber nicht integrierbar ($\int_{-\infty}^{\infty}f(x)dx=+\infty$), also keine Wahrscheinlichkeitsdichte.

(d) Wie ist die exakte Begründung dafür, dass Q im Beweis von Satz 1.56 ein Wahrscheinlichkeitsmaß ist?

Es gilt $Q(\mathbb{R})=\int_{\mathbb{R}}f(x)\,dx=1$ nach Voraussetzung sowie für paarweise disjunkte Borelmengen B_n

$$Q\Big(\bigcup_{n\geqslant 1}B_n\Big)=\int_{\bigcup_{n\geqslant 1}B_n}f(x)\,dx=\int_{\mathbb{R}}\lim_{N\to\infty}\sum_{n=1}^{N}\mathbf{1}_{B_n}(x)f(x)\,dx.$$

Wegen $\sum_{n\geqslant 1}Q(B_n)=\lim_{N\to\infty}\int_{\mathbb{R}}\sum_{n=1}^{N}\mathbf{1}_{B_n}(x)f(x)\,dx$ müssen wir begründen, warum Grenzwert und Integral vertauscht werden können. Da $g_N(x):=\sum_{n=1}^{N}\mathbf{1}_{B_n}(x)f(x)$ die Monotonie $g_{N+1}(x)\geqslant g_N(x)$ erfüllt, folgt dies aus dem Satz über monotone Konvergenz (*Satz von Beppo Levi*).

(e) Durch $F(a,b)=P((-\infty,a]\times(-\infty,b])$, $a,b\in\mathbb{R}$, wird die Verteilungsfunktion eines Wahrscheinlichkeitsmaßes P auf $\mathfrak{B}_{\mathbb{R}^2}$ definiert. Wie können die Eigenschaften vom eindimensionalen Fall übertragen werden?

Die Monotonie gilt gemäß $F(a,b)\leqslant F(c,d)$, wenn $a\leqslant c$ und $b\leqslant d$. Rechtsstetigkeit überträgt sich als $F(a_n,b_n)\downarrow F(a,b)$ für alle Folgen $a_n\downarrow a$, $b_n\downarrow b$. Außerdem gilt $\lim_{a\to-\infty}F(a,b)=\lim_{b\to-\infty}F(a,b)=0$ und $\lim_{a,b\to\infty}F(a,b)=1$. Eine genauere Übertragung der Monotonie zeigt $F(c,d)-F(a,d)-F(c,b)+F(a,b)=P((a,b]\times(c,d])\geqslant 0$ für $a\leqslant c$, $b\leqslant d$. Für jede solche Funktion F kann umgekehrt wieder gezeigt werden, dass ein entsprechendes Lebesgue-Stieltjes-Maß P auf $\mathfrak{B}_{\mathbb{R}^2}$ existiert mit der letzten Eigenschaft.

Ende der 4. Vorlesung

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 Lemma. *Für die Verteilungsfunktion $\Phi_{\mu,\sigma^2} = F_{N(\mu,\sigma^2)}$ von $N(\mu, \sigma^2)$ gilt:*

(a) $\Phi_{\mu,\sigma^2}(x) = \Phi_{0,1}((x-\mu)/\sigma)$, $x \in \mathbb{R}$;

(b) $\frac{1}{\sqrt{2\pi}(x+x^{-1})} e^{-x^2/2} \leqslant 1 - \Phi_{0,1}(x) \leqslant \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}$, $x > 0$.

Beweis. Die Substitutionsregel mit $y = (z-\mu)/\sigma$ zeigt $\int_{-\infty}^x \varphi_{\mu,\sigma^2}(z)dz = \int_{-\infty}^{(x-\mu)/\sigma} \varphi_{0,1}(y)dy$, also Teil (a).

Für die untere Schranke in (b) benutze $(x^{-1}e^{-x^2/2})' = -(1+x^{-2})e^{-x^2/2}$ und schließe damit

$$\int_x^\infty e^{-y^2/2}dy \geqslant \int_x^\infty \frac{1+y^{-2}}{1+x^{-2}} e^{-y^2/2}dy = \frac{x^{-1}}{1+x^{-2}} e^{-x^2/2} = \frac{1}{x+x^{-1}} e^{-x^2/2}.$$

Wegen $1 - \Phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2}dy$ liefert dies die Behauptung. Für die obere Schranke berechne analog $\int_x^\infty e^{-y^2/2}dy \leqslant \int_x^\infty \frac{y}{x} e^{-y^2/2}dy = x^{-1}e^{-x^2/2}$. □

1.68 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. □

1.69 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 $f^Y = f_{N(\mu,\sigma^2)}$. □

1.70 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 $x_1, \ldots, x_m \in I$, 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 ($\{x_1, \ldots, x_m\}$ 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

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

1.71 Beispiel. Es sei X eine reellwertige Zufallsvariable mit Dichte f^X. Die Dichte von $Y = X^2$ ist nach Bemerkung 1.70(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.

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

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

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. □

1.73 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.

1.74 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:

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

1.75 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.

▷ **Kontrollfragen**

(a) Welche Wahrscheinlichkeitsdichte ist auf dem 10DM-Schein abgebildet?
Die Normalverteilung („Glockenkurve“) zusammen mit Carl Friedrich Gauß.

(b) Nach dem Dichtetransformationssatz gilt $f^{X+a}(x) = f^X(x-a)$ für $a \in \mathbb{R}$. Wie kann man die Umkehrung des Vorzeichens '+' nach '-' heuristisch erklären? Formuliere ein ähnliches Resultat für Zähldichten auf $\mathbb{Z}$.
Für kleines $\varepsilon > 0$ gilt $P(X+a \in [x-\varepsilon, x+\varepsilon]) = P(X \in [x-a-\varepsilon, x-a+\varepsilon]) \approx (2\varepsilon)^{-1}f^X(x-a)$. Die Wahrscheinlichkeit, dass $X+a$ nahe x ist, ist also gerade die Wahrscheinlichkeit, dass X nahe $x-a$ ist. Für $\mathbb{Z}$-wertige Zufallsvariablen X und $a \in \mathbb{Z}$ gilt

$$p^{X+a}(x) = P(X+a = x) = P(X = x-a) = p^X(x-a), \quad a \in \mathbb{Z}.$$

(c) Wie folgt mittels Dichtetransformation, dass $Y = \varphi(U)$ mit $\varphi(x) = -\log(1-x)$, $x \in (0,1)$, und $U \sim U((0,1))$ eine $\mathrm{Exp}(1)$-Verteilung besitzt? Beachte auch $\varphi = F^{-1}_{\mathrm{Exp}(1)}$ und vergleiche mit Beispiel 1.54.

Die Umkehrfunktion ist $\varphi^{-1}(y) = 1 - e^{-y}$ und $(\varphi^{-1})'(y) = e^{-y}$, so dass nach dem Dichtetransformationssatz gilt $f^Y(y) = \mathbf{1}_{(0,1)}(1-e^{-y})e^{-y} = e^{-y}\mathbf{1}_{(0,\infty)}(y)$. Also ist Y $\mathrm{Exp}(1)$-verteilt, was alternativ mit der Quantilstransformation aus Beispiel 1.54 folgt.

(d) Wie kann man eine Zähldichte als μ-Dichte gemäß Bemerkung 1.73 verstehen? (Tipp: betrachte das Zählmaß μ)

Ist $\Omega = \{\omega_1, \omega_2, \ldots\}$ abzählbar, so ist $\mu = \sum_{i \geqslant 1} \delta_{\omega_i}$ das Zählmaß, alternativ definiere $\mu(A) = |A|$ für $A \subseteq \Omega$. Ist P ein Wahrscheinlichkeitsmaß auf $\mathscr{P}(\Omega)$ mit Zähldichte p, so gilt nach Definition des Maßintegrals

$$P(A) = \sum_{\omega \in A} p(\omega)\mu(\{\omega\}) = \int_A p(\omega)\mu(d\omega), \quad A \subseteq \Omega.$$

Also ist p die μ-Dichte von P.

(e) Wie kann man $P_f([a_1,b_1] \times [a_2,b_2])$ aus f_1, f_2 berechnen für eine zweidimensionale Produktdichte $f(x_1,x_2) = f_1(x_1)f_2(x_2)$?

Mit Fubini gilt

$$P_f([a_1,b_1]\times[a_2,b_2]) = \int_{a_1}^{b_1}\int_{a_2}^{b_2} f_1(x_1)f_2(x_2)dx_2dx_1 = P_{f_1}([a_1,b_1])P_{f_2}([a_2,b_2]).$$

Ende der 5. Vorlesung

1.76 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*

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

Ist allgemeiner $\varphi : U \to V$ und $U = \bigcup_{j=1}^m U_j$ eine Zerlegung in paarweise disjunkte, offene Mengen U_j derart, dass $P(X \in U) = 1$ und $\varphi_j = \varphi|_{U_j} : U_j \to \varphi(U_j)$, $j = 1, \ldots, m$, jeweils einen C^1-Diffeomorphismus definiert, so besitzt $Y = \varphi(X)$ die Dichte

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

1.77 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 behandeln zunächst den Fall ohne Zerlegung und 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.

Die Erweiterung auf stückweise Diffeomorphismen ergibt sich wie im Eindimensionalen, vergleiche Bemerkung 1.70(b). □

1.78 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 $\det(D(\varphi^{-1}(y))) = \det(A^{-1}) = \det(A)^{-1}$. □

1.79 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

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

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 $P(A\,|\,B) = \frac{3/36}{9/36} = \frac{1}{3}$. 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 $P(A\,|\,B) > P(A)$ gilt und das Eintreten des Ereignisses B die Wahrscheinlichkeit des Ereignisses A beeinflusst (später: A und B sind nicht unabhängig).

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

(a) *Durch $Q(A) := P(A\,|\,B)$ 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*

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

(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*

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

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):

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

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

2.5 Bemerkungen.

(a) Bei Anwendern werden oft $P(A\,|\,B)$ und $P(A\cap B)$ vermischt. Mathematisch sind beide Wahrscheinlichkeiten nur gleich, wenn $P(B)=1$ gilt. Während $P(A\,|\,B)$ 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 2.7) 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 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 Beispiel (Test auf eine seltene Krankheit)**.** Der Standard-IgA-Elisa-Schnelltest auf Corona-Antikörper besitzt eine *Spezifität* von 93% und eine *Sensitivität* von 97%, vgl. auch den Spiegel-Artikel vom 27.4.20. Das heißt, dass er bei 93% der Patienten mit Antikörpern, aber auch bei 3% der Patienten ohne Antikörper ein positives Testergebnis zeigt. Alternativ sagt man auch, dass der Test 3% *falsch-positive* und 7% *falsch-negative* Resultate aufweist. Es sei angenommen, dass $0,5\%$ der Bevölkerung Corona-Antikörper entwickelt haben. Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person mit positivem Testergebnis in der Tat Corona-Antikörper besitzt?

Setzen wir $A=$'Test positiv', $B_1=$'Antikörper vorhanden', $B_2=B_1^c=$'keine Antikörper', so liefert die Bayes-Formel

$$P(B_1\,|\,A) = \frac{P(B_1)P(A\,|\,B_1)}{P(B_1)P(A\,|\,B_1)+P(B_2)P(A\,|\,B_2)} = \frac{0,005\times 0,93}{0,005\times 0,93+0,995\times 0,03},$$

also $P(B_1 \mid A) \approx 13\%$. Obwohl der Test im Vergleich zur Konkurrenz gut abschneidet, wird ein positives Ergebnis bei ca. 87% der Getesteten falsch sein, also womöglich eine Immunität nur vorgaukeln. Grund dafür ist der relativ kleine Anteil von $0,5\%$, bei dem *a priori* Antikörper in der Bevölkerung vorliegen. Durch das positive Testergebnis wird die *a priori*-Wahrscheinlichkeit $0,5\%$ zur *a posteriori*-Wahrscheinlichkeit 13% erhöht (*Bayesian updating*). Um dies signifikant zu verbessern, sollte die falsch-positiv-Rate des Tests verringert werden. Eine Halbierung der Falsch-Positiven ergäbe fast eine Verdopplung von $P(B_1 \mid A)$ auf 24%, während eine Halbierung der Falsch-Negativen nur zu einer leichten Erhöhung $P(B_1 \mid A) \approx 14\%$ führen würde.

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

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

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 $P(A_1 \cap \dots \cap A_{n-1}) > 0 \Rightarrow P(A_1 \cap \dots \cap A_{n-2}) > 0$)

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

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

2.9 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*).

▷ **Kontrollfragen**

(a) An welchem Punkt ist die Dichte $f_{N(\mu,\Sigma)}$ maximal? Welche geometrische Form besitzen die Höhenlinien von $f_{N(\mu,\Sigma)}$ in $\mathbb{R}^2$ (d.h. $\{x \in \mathbb{R}^2 \mid f_{N(\mu,\Sigma)}(x) = y\}$ für geeignete $y > 0$)?

$f_{N(\mu,\Sigma)}$ ist maximal bei μ. Die Höhenlinien in $\mathbb{R}^2$ sind von der Form $\{x \in \mathbb{R}^2 \mid \langle \Sigma^{-1}(x-\mu), x-\mu \rangle = a\}$ für $a \geqslant 0$, also Ellipsen mit Schwerpunkt μ und Halbachsen bestimmt durch die positiv-definite Matrix Σ^{-1}, vgl. lineare Algebra.

(b) Wie groß ist beim fairen zweifachen Münzwurfexperiment die Wahrscheinlichkeit, dass die zweite Münze 'Kopf' zeigt, gegeben dass die erste 'Zahl' zeigt? (rigorose Herleitung!)

Betrachte das Bernoulli-Schema der Länge 2 mit $p = 1/2$, d.h. P sei die Laplaceverteilung auf $\Omega = \{0,1\}^2$. Setze $X_k : \Omega \to \{0,1\}$, $X_k(\omega_1, \omega_2) := \omega_k$ für $k \in \{1,2\}$. Interpretiere 'Kopf' als '1', 'Zahl' als '0'. Dann ist $P(X_1 = 0, X_2 = 1) = P(\{(0,1\}) = \frac{1}{4}$, $P(X_1 = 0) = P(\{(0,1),(0,0)\}) = \frac{1}{2}$ und somit ist die gesuchte Wahrscheinlichkeit $P(X_2 = 1 \,|\, X_1 = 0) = \frac{1/4}{1/2} = \frac{1}{2}$.

(c) Gilt $P(A^c \,|\, B) = 1 - P(A \,|\, B)$? Gilt $P(A \,|\, B^c) = 1 - P(A \,|\, B)$? (alles sei wohldefiniert)

Da $P(\bullet \,|\, B)$ ein Wahrscheinlichkeitsmaß ist, gilt die erste Identität. Die zweite ist im Allgemeinen falsch, z.B. ist für $A = \Omega$ stets $P(A \,|\, B^c) = P(A \,|\, B) = 1$.

(d) Führe die Rechnungen in Beispiel 2.6 explizit aus.

Für die Ereignisse $A =$'zugelasssen', $B =$'männlich', $C_i =$'Bewerbung auf Fach i ergibt sich, wenn man die Wahrscheinlichkeiten über die relativen Häufigkeiten definiert:

$$\begin{aligned} P(A \,|\, B) &= \sum_{i=1}^{3} P(A \,|\, B \cap C_i) P(B \cap C_i) \\ &= 0,62 \frac{825}{1710} + 0,63 \frac{560}{1710} + 0,34 \frac{325}{1710} \approx 0,57. \\ P(A \,|\, B^c) &= \sum_{i=1}^{3} P(A \,|\, B^c \cap C_i) P(B^c \cap C_i) \\ &= 0,82 \frac{108}{726} + 0,68 \frac{25}{726} + 0,37 \frac{593}{726} \approx 0,45. \end{aligned}$$

(e) Gib alle Wahrscheinlichkeiten für die Versuchsausgänge nach drei Ziehungen in Beispiel 2.9 an. Addieren sie sich zu eins auf?

Wir nehmen einfachheitshalber $S \geqslant 2$, $W \geqslant 2$ an. Es gilt $P('SSW') = P('SWS') = P('WSS') = \frac{S(S-1)W}{N(N-1)(N-2)}$, $P('WWS') = P('WSW') = P('SWW') = \frac{SW(W-1)}{N(N-1)(N-2)}$, $P('SSS') = \frac{S(S-1)(S-2)}{N(N-1)(N-2)}$, $P('WWW') = \frac{W(W-1)(W-2)}{N(N-1)N(-2)}$. Durch Ausmultiplizieren überprüft man

$$\begin{aligned} &N(N-1)(N-2) = (S+W)(S+W-1)(S+W-2) \\ &= 3S(S-1)W + 3SW(W-1) + S(S-1)(S-2) + W(W-1)(W-2), \end{aligned}$$

so dass sich die Wahrscheinlichkeiten zu Eins aufsummieren. Man kann auch entlang der Pfade aufsummieren, z.B. $P('SSW') + P('SSS') = \frac{S(S-1)W + S(S-1)(S-2)}{N(N-1)(N-2)} = \frac{S(S-1)}{N(N-1)} = P('SS')$, so dass man induktiv über die Anzahl der Ziehungen argumentieren kann.

Ende 6. Vorlesung

2.2 Unabhängige Ereignisse und Lemma von Borel-Cantelli

2.10 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 $P(B)$).

2.11 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

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

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

2.13 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.

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

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

2.15 Bemerkung. 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 $\{\omega \in \Omega \,|\, (X_n(\omega))_{n \geqslant 1}$ 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.

2.16 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 die Folge $(A_n)_{n \geqslant 1}$ 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 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. $\square$

2.17 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

2.18 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.

2.19 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.

2.20 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. $\square$

2.21 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 beliebige endliche Indexmengen $J \subseteq I$ 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 o.B.d.A. eine Indexmenge J mit $|J| \geqslant 2$. Betrachte nun $A_i \in \mathscr{S}_i$ 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 $A_i \in \mathscr{S}_i$. $\square$

2.22 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.

2.23 Korollar. *Es seien $X_1, \ldots, X_n$ 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) *Hat jedes X_k Werte in $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$, so sind $X_1, \dots, X_n$ 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 $\mathscr{E}_k = \{\{s_k\} \,|\, s_k \in S_k\} \cup \{\varnothing\}$. 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 2.21 sind $X_1, \dots, X_n$ unabhängig.

Für (b) folgt auf dem $\cap$-stabilen Erzeuger $\mathscr{E} = \{(-\infty, b] \,|\, b \in \mathbb{R}\}$ 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 2.21 zeigt daher die Unabhängigkeit von $X_1, \dots, X_n$. □

2.24 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.

2.25 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 Randdichte*

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

(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

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

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 2.23(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. □

▷ **Kontrollfragen**

(a) Beim Würfelwurf sind '1. Augenzahl 6', '2. Augenzahl 6', 'Pasch' drei paarweise unabhängige Ereignisse, die aber nicht gemeinsam unabhängig sind. Überprüfe!

Alle drei Ereignisse haben Wahrscheinlichkeit $1/6$. Die Ereignisse '1. und 2. Augenzahl 6', '1. Augenzahl 6 und Pasch', '2. Augenzahl 6 und Pasch' sind alle drei gleich und haben Wahrscheinlichkeit $\frac{1}{36} = \frac{1}{6} \bullet \frac{1}{6}$, so dass paarweise Unabhängigkeit folgt. Der Schnitt aller drei Ereignisse ist dann aber gleich dem Schnitt von je zwei Ereignissen und besitzt Wahrscheinlichkeit $\frac{1}{36} \neq \frac{1}{6} \bullet \frac{1}{6} \bullet \frac{1}{6}$, so dass die drei Ereignisse nicht unabhängig sind. Das ist ja auch intuitiv, weil z.B. die Wahrscheinlichkeit für 'Pasch' gegeben '1. und 2. Augenzahl ist 6' gleich eins ist.

(b) Was bedeutet das Ereignis $\liminf_{n\to\infty} A_n := \bigcup_{m\geqslant 1} \bigcap_{n \geqslant m} A_n$? Was sagt das Lemma von Borel-Cantelli darüber aus?

$\liminf_{n\to\infty} A_n = \{\omega \in \Omega \,|\, \exists m \geqslant 1 \forall n \geqslant m : \omega \in A_n\}$ bedeutet gerade, dass alle bis auf endlich viele A_n eintreten. Wegen $(\liminf_{n\to\infty} A_n)^c = \limsup_{n\to\infty} A_n^c$ folgt aus dem Lemma von Borel-Cantelli, dass $P(\liminf_{n\to\infty} A_n) = 1$ im Fall $\sum_{n\geqslant 1}(1 - P(A_n)) < \infty$ und $P(\liminf_{n\to\infty} A_n) = 0$ im Fall $\sum_{n\geqslant 1}(1 - P(A_n)) = \infty$ und $(A_n)_{n\geqslant 1}$ unabhängig gilt. Beachte dazu, dass nach Beispiel 2.22 $(A_n)_{n\geqslant 1}$ genau dann unabhängig ist, wenn $(A_n^c)_{n\geqslant 1}$ unabhängig ist.

(c) Finde Zufallsvariablen X_1, X_2 und messbare Funktionen g_1, g_2, so dass $g_1(X_1)$, $g_2(X_2)$ unabhängig sind, aber X_1, X_2 nicht.

Die einfachsten unabhängigen Zufallsvariablen sind deterministisch. Setze daher $g_1(x) = g_2(x) = 0$ und zum Beispiel $X_1 = X_2 \sim N(0,1)$. Dann gilt $P(X_1 \geqslant 0, X_2 \geqslant 0) = P(X_1 \geqslant 0) = 1/2 \neq P(X_1 \geqslant 0)P(X_2 \geqslant 0)$, so dass X_1, X_2 nicht unabhängig sind. Andererseits ist $P(g_1(X_1) \in A_1, g_2(X_2) \in A_2) = \mathbf{1}_{A_1 \cap A_2}(0) = P(g_1(X_1) \in A_1)P(g_2(X_2) \in A_2)$ für beliebige Borelmengen A_1, A_2, so dass $g_1(X_1)$, $g_2(X_2)$ unabhängig sind.

(d) Wie folgt durch explizite Rechnung, dass A, B unabhängig A^c, B^c unabhängig impliziert?

$P(A^c \cap B^c) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A \cap B)$
$= P(A^c) - P(B) + P(A)P(B) = P(A^c)(1 - P(B)) = P(A^c)P(B^c)$.

(e) Beim Bernoulli-Schema $(\Omega, \mathscr{F}, P)$ mit $\Omega = \{0,1\}^n$ betrachte die Zufallsvariablen $X_k(\omega) = \omega_k$, $k = 1, \ldots, n$. Bestimme die Zähldichten von $(X_1, \ldots, X_n)$ und von X_k, $k = 1, \ldots, n$. Sind $X_1, \ldots, X_n$ unabhängig?

Da $X = (X_1, \ldots, X_n)$ die Identität auf Ω ist, gilt $p^X(\omega) = p_{\mathrm{Bern}(n,p)}(\omega)$. Wegen $\{X_k = 1\} = \{\omega \in \Omega \,|\, \omega_k = 1\}$ gilt

$$\begin{aligned} P(X_k = 1) &= \sum_{\omega \in \Omega, \omega_k = 1} p^{\sum_i \omega_i}(1-p)^{\sum_i (1-\omega_i)} \\ &= p \prod_{1 \leqslant i \leqslant n, i \neq k} \Big(\sum_{\omega_i = 0}^{1} p^{\omega_i}(1-p)^{1-\omega_i} \Big) = p. \end{aligned}$$

Wir erhalten $p^{X_k}(1) = p$, $p^{X_k}(0) = 1 - p$, also $p^{X_k}(\omega_k) = p^{\omega_k}(1-p)^{1-\omega_k}$ und daher $p^X(\omega) = \prod_{k=1}^n p^{X_k}(\omega_k)$. Also sind die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig. Natürlich ist genau das die Idee des Bernoulli-Schemas: die Erfolge in jeder 'Runde' treten unabhängig mit Wahrscheinlichkeit p ein.

Ende 7. Vorlesung

2.26 Beispiele.

(a) Sind $f_1, \ldots, f_n$ 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 $X_k(x_1, \ldots, x_n) = x_k$, $k = 1, \ldots, n$, sind Borel-messbar, also Zufallsvariablen. Gemäß Satz 2.25(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 $P^{(X_1,\ldots,X_n)}$ 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.

(b) Ist $X = (X_1, \ldots, X_n) \sim N(0, E_n)$, 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.

(c) Es sei $X \sim U(B)$ mit $B := \{x \in \mathbb{R}^2 \mid |x| \leqslant 1\}$ ein gleichmäßig auf der Kreisscheibe B verteilter Zufallsvektor. Betrachte seine Polarkoordinaten

$$R := (X_1^2 + X_2^2)^{1/2} \geqslant 0, \quad \Phi := \arctan(X_2/X_1) \in [0, 2\pi),$$

wobei Φ (d.h. der „Zweig“ von arctan) so definiert sei, dass $X_1 = R\cos\Phi$, $X_2 = R\sin\Phi$ gilt (z.B. $\Phi = 0$ für $X_1 = X_2 = 0$). Da R und Φ (gemäß Konstruktion) Borel-messbare Funktionen von X sind, sind R, Φ Zufallsvariablen. Für $r \in [0,1]$, $\varphi \in [0, 2\pi]$ gilt nach Gleichverteilung

$$P(R \leqslant r, \Phi \leqslant \varphi) = \frac{\varphi r^2/2}{\pi} \quad \Big(\frac{\text{Fläche des Kreissegments mit Radius } r}{\text{Fläche von } B}\Big).$$

Insbesondere gilt $P(R \leqslant r) = \frac{2\pi r^2/2}{\pi} = r^2$ und $P(\Phi \leqslant \varphi) = \frac{\varphi}{2\pi}$. Damit folgt $R^2 \sim U([0,1])$ (beachte $P(R^2 \leqslant x) = x$ für $x \in [0,1]$), $\Phi \sim U([0, 2\pi])$, und wegen $P(R \leqslant r, \Phi \leqslant \varphi) = P(R \leqslant r)P(\Phi \leqslant \varphi)$ für alle r, φ (betrachte $r < 0$, $r > 1$, $\varphi < 0$, $\varphi > 2\pi$ gesondert) sind R, Φ unabhängige Zufallsvariablen. Nach Lemma 2.20 sind übrigens auch R^2, Φ unabhängige Zufallsvariablen. Man mache sich klar, dass X_1, X_2 in diesem Fall nicht(!) unabhängig sind.

(d) 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 $D := \{(x,x) \mid x \in \mathbb{R}\}$, aber D besitzt zweidimensionales Lebesguemaß Null. Damit kann $P^{(X_1,X_2)}$ keine Dichte (bezüglich dem Lebesguemaß) besitzen.

2.27 Bemerkung. Gemäß Beispiel 2.26(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. Dann ist $X_1(\omega) := \pi_1(\omega)$ eine Zufallsvariable mit Verteilung

$$(P_1 \otimes P_2)^{X_1}(A_1) = (P_1 \otimes P_2)(\pi_1^{-1}(A_1)) = P_1(A_1)P_2(\Omega_2) = P_1(A_1), \quad A_1 \in \mathscr{F}_1.$$

Analog gilt $(P_1 \otimes P_2)^{X_2} = P_2$. Außerdem sind X_1, X_2 (unter $P_1 \otimes P_2$) unabhängig, da

$$\begin{aligned}(P_1 \otimes P_2)(X_1 \in A_1, X_2 \in A_2) &= (P_1 \otimes P_2)(A_1 \times A_2) = P_1(A_1)P_2(A_2)\\ &= (P_1 \otimes P_2)^{X_1}(A_1)(P_1 \otimes P_2)^{X_2}(A_2)\end{aligned}$$

gilt. Diese Konstruktion erlaubt dann die Konstruktion von Produkträumen und entsprechend verteilten unabhängigen Zufallsvariablen aus endlich vielen vorgegebenen Wahrscheinlichkeitsräumen. Wie das unendlich lange Münzwurfexperiment lehrt, führen einfache Modelle bereits auf unendlich viele unabhängige Zufallsvariablen, deren Existenz a priori weniger klar ist.

2.28 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 <u>Produkt-σ-Algebra</u>

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

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 <u>Produktmaß</u> der P_i, Schreibweise $Q = \bigotimes_{i\in I} P_i$.

2.29 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 $\mathbb{R}^{\mathbb{N}}$.

2.30 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. X ist $(\mathscr{F}, \bigotimes_{i\in I} \mathscr{F}_i)$-messbar, weil nach Voraussetzung die Komposition $\pi_i \circ X = X_i$ für jedes i $(\mathscr{F}, \mathscr{F}_i)$-messbar ist, was nach Konstruktion der Produkt-σ-Algebra hinreichend ist (Messbarkeit auf Erzeuger!). Das System

$$\mathscr{Z} = \Big\{ \bigcap_{j\in J} \pi_j^{-1}(A_j) \,\Big|\, J \subseteq I \text{ endlich, } A_j \in \mathscr{F}_j \Big\}$$

der <u>Zylindermengen</u> ist ein $\cap$-stabiler Erzeuger von $\bigotimes_{i\in I} \mathscr{F}_i$ (klar nach Definition!). Für jedes Ereignis $\bigcap_{j\in J} \pi_j^{-1}(A_j)$ in $\mathscr{Z}$ gilt nach Definition und wegen Unabhängigkeit

$$\begin{aligned}P^X\Big(\bigcap_{j\in J} \pi_j^{-1}(A_j)\Big) &= P\Big(\bigcap_{j\in J} \{X_j \in A_j\}\Big) = \prod_{j\in J} P(X_j \in A_j) = \prod_{j\in J} P^{X_j}(A_j)\\ &= \prod_{j\in J} \Big(\bigotimes_{i\in I} P^{X_i}\Big)\Big(\pi_j^{-1}(A_j)\Big) = \Big(\bigotimes_{i\in I} P^{X_i}\Big)\Big(\bigcap_{j\in J} \pi_j^{-1}(A_j)\Big).\end{aligned}$$

Nach dem Eindeutigkeitssatz folgt $P^X = \bigotimes_{i\in I} P^{X_i}$. $\square$

2.31 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.

2.32 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 2.27). 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 $A_1, \dots, A_1, A_2, \dots, A_2, \dots$). Aus $A_{n+1} \subseteq A_n$ folgt dann $A'_{n+1} \subseteq A'_n \times \Omega_{n+1}$. Für $m > n$ setze

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

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

$$\begin{aligned}
&\int_{\Omega_{n+1}} h_{n+1}(\bar{\omega}_1, \ldots, \bar{\omega}_n, \omega_{n+1}) P_{n+1}(d\omega_{n+1}) \\
&\qquad = \lim_{m\to\infty} \int_{\Omega_{n+1}} h_{m,n+1}(\bar{\omega}_1, \ldots, \bar{\omega}_n, \omega_{n+1}) P_{n+1}(d\omega_{n+1}) \\
&\qquad = \lim_{m\to\infty} h_{m,n}(\bar{\omega}_1, \ldots, \bar{\omega}_n) = h_n(\bar{\omega}_1, \ldots, \bar{\omega}_n) \geqslant \delta > 0.
\end{aligned}$$

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. □

2.33 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.

2.34 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 $i \in I$. □

▷ **Kontrollfragen**

(a) Wie zeigt man, dass die Koordinaten eines $N(\mu, \Sigma)$-verteilten Zufallsvektors unabhängig sind, wenn Σ eine Diagonalmatrix ist? Gilt auch die umgekehrte Implikation?
Mit $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ gilt $\Sigma^{-1} = \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_d^{-2})$ und (einfache Rechnung)

$$\varphi_{\mu,\Sigma}(x) = \prod_{i=1}^{d} \Big((2\pi\sigma_i^2)^{-1/2} e^{-(x_i-\mu_i)^2/(2\sigma_i^2)}\Big).$$

Also ist die $N(\mu,\Sigma)$-Dichte in diesem Fall eine Produktdichte der $N(\mu_i,\sigma_i^2)$-Dichten. Nach dem Satz sind die einzelnen Koordinaten also unabhängig.

Sind die Koordinaten andererseits unabhängig, so muss $\varphi_{\mu,\Sigma}$ eine Produktdichte sein. Betrachte den Logarithmus und bezeichne mit C einen nicht von x abhängenden Term:

$$\log(\varphi_{\mu,\Sigma}(x)) = C - \tfrac{1}{2}\langle \Sigma^{-1}(x-\mu), x-\mu\rangle = C - \frac{1}{2}\sum_{i,j=1}^{d}(\Sigma^{-1})_{i,j}(x_i-\mu_i)(x_j-\mu_j).$$

Als Produktdichte muss andererseits $\log(\varphi_{\mu,\Sigma})(x) = \sum_{i=1}^{d} g_i(x_i)$ mit geeigneten Funktionen g_i gelten. Wegen der Kreuzterme $x_i x_j$ impliziert dies $(\Sigma^{-1})_{i,j} = 0$ für $i \neq j$ (Σ^{-1} ist symmetrisch), also Σ^{-1} diagonal und somit Σ diagonal.

(b) Zeige formal, dass X_1, X_2 nicht unabhängig sind in Beispiel 2.26(c).

Dies folgt beispielsweise aus $P(X_1 > \frac{1}{\sqrt{2}}, X_2 > \frac{1}{\sqrt{2}}) \leqslant P(|X|^2 > 1) = 0$, aber $P(X_1 > \frac{1}{\sqrt{2}}) = P(X_2 > \frac{1}{\sqrt{2}}) > 0$ (Fläche des Kreissegments $\{(x_1,x_2)\,|\,x_1^2 + x_2^2 \leqslant 1, x_1 \geqslant \frac{1}{\sqrt{2}}\}$ ist größer Null).

(c) Wie kann ein Bernoulli-Schema der Länge n als Produktraum von n Wahrscheinlichkeitsräumen konstruiert werden?

Setze $\Omega_i = \{0,1\}$, $\mathscr{F}_i = \mathscr{P}(\Omega_i)$ und $P_i = \mathrm{Bern}(1,p)$. Dann ist $\Omega = \prod_{i=1}^{n}\Omega_i = \{0,1\}^n$ und $\mathscr{F} = \bigotimes_{i=1}^{n}\mathscr{F}_i = \mathscr{P}(\Omega)$, weil die Potenzmenge von Ω durch die Zylindermengen $\{\omega \in \Omega\,|\,\omega_i = b_i\}$, $i = 1,\dots,n$, $b_i \in \{0,1\}$ erzeugt wird. Das Produktmaß $P = \bigotimes_{i=1}^{n} P_i$ erfüllt $P(\{b\}) = P(\bigcap_{i=1}^{n}\{\omega_i = b_i\}) = \prod_{i=1}^{n} P_i(\{b_i\})$ für $b \in \{0,1\}^n$. Wie in Kontrollfrage 7(e) leitet man her, dass die Zähldichte von P gegeben ist durch $p_{\mathrm{Bern}(n,p)}$.

(d) Betrachte $\{0,1\}^{\mathbb{N}}$, die Menge der 0-1-Folgen, mit Produkt-σ-Algebra $\mathscr{P}(\{0,1\})^{\otimes\mathbb{N}}$. Liegt das Ereignis $A = \{(b_i)_{i\geqslant 1}\,|\,\text{nur endliche viele } b_i = 1\}$ in der Produkt-σ-Algebra?

Es gilt $A^c = \limsup_{i\to\infty}\{b_i = 1\}$, Als Zylindermenge ist $\{b_i = 1\} \in \mathscr{P}(\{0,1\})^{\otimes\mathbb{N}}$ für jedes i. Damit liegen auch abzählbare Vereinigungen und Schnitte sowie Komplemente in der σ-Algebra $\mathscr{P}(\{0,1\})^{\otimes\mathbb{N}}$. Mit der Definition von $\limsup$ zeigt dies $A \in \mathscr{P}(\{0,1\})^{\otimes\mathbb{N}}$.

(e) Wie kann man nun sauber den Wahrscheinlichkeitsraum des unendlich langen Münzwurfs konstruieren? Lässt sich der Beweis von Satz 2.32 in diesem Fall vereinfachen?

Mit $(\Omega_i,\mathscr{F}_i,P_i)$ wie in (c) mit $p = 1/2$ betrachte den unendlichen Produktraum $(\prod_{i\geqslant 1}\Omega_i, \bigotimes_{i\geqslant 1}\mathscr{F}_i, \bigotimes_{i\geqslant 1}P_i)$. Nach Definition von Produkt-σ-algebra und -maß ist die Wahrscheinlichkeit von Ereignissen, die nur von n Würfen abhängen (also von Zylindermengen), gerade die Wahrscheinlichkeit im Bernoulli-Schema der Länge n, was wir bislang 'intuitiv' vorausgesetzt haben. Der Satz von Vitali zeigt übrigens $\mathscr{P}(\{0,1\})^{\otimes\mathbb{N}} \subsetneqq \mathscr{P}(\{0,1\}^{\mathbb{N}})$.

Der Beweisteil von Satz 2.32 für '$P(A_n) \to \delta > 0$, $A_{n+1} \subseteq A_n$ impliziert $\bigcap_{n\geqslant 1} A_n \neq \varnothing$' lässt sich konkreter fassen. Wir nehmen wieder $A_n = \Pi_n^{-1}(A_n')$ mit $A_n' \in \{0,1\}^n$ an. Setze noch $P^n = \bigotimes_{i=1}^{n} P_i$. Die bedingten Wahrscheinlichkeiten erfüllen $\frac{1}{2}P^n(A\,|\,\omega_1 = 0) + \frac{1}{2}P^n(A\,|\,\omega_1 = 1) = P^n(A)$ (Formel der totalen Wahrscheinlichkeit). Wegen $P^n(A_n') = P(A_n) \downarrow \delta$ und $P^{n+1}(A_{n+1}'|\,\omega_1 = b) \leqslant P^n(A_n'|\,\omega_1 = b)$ (beachte $A_{n+1}' \subseteq A_n' \times \{0,1\}$) gibt es ein $\bar\omega_1 \in \{0,1\}$ mit $\lim_{n\to\infty} P^n(A_n'|\,\omega_1 = \bar\omega_1) \geqslant \delta$. Dasselbe Argument, angewendet auf $A_n' \cap \{\omega_1 = \bar\omega_1\}$ liefert ein $\bar\omega_2$ mit $\lim_{n\to\infty} P^n(A_n'|\,\{\omega_1 = \bar\omega_1, \omega_2 = \bar\omega_2\}) \geqslant \delta$. Induktiv erhalten wir so eine Folge $(\bar\omega_m)_{m\geqslant 1}$ mit

$\lim_{n\to\infty} P^n(A'_n| \bigcap_{1\leqslant i\leqslant m}\{\omega_i = \bar\omega_i\}) \geqslant \delta$. Wegen Inklusion ist insbesondere $(\bar\omega_i)_{1\leqslant i\leqslant m} \in \Pi_m(\bigcap_{n\geqslant 1} A_n)$ für alle $m \geqslant 1$. Dies bedeutet gerade, dass die Folge $(\bar\omega_i)_{i\geqslant 1}$ in $\bigcap_{n\geqslant 1} A_n$ liegt.

Ein anderer Beweis für Freunde der Topologie beruht auf der Produkttopologie von $\{0,1\}^{\mathbb{N}}$, die den 0-1-Folgenraum kompakt macht, und einem Kompaktheitsargument genau wie beim Prämaßnachweis für Lebesgue-Stieltjes-Maße. Eine elementare (einfache und elegante) Konstruktion des unendlichen Produktmaßes ist mir selbst beim einfachen Münzwurfexperiment nicht bekannt, für Hinweise oder Ideen wäre ich dankbar.

Ende 8. Vorlesung

2.35 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

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

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

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

2.37 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 $\bigcap_{m\geqslant 1}\bigcup_{n\geqslant m}\{X_n \in B\} = \bigcap_{m\geqslant k}\bigcup_{n\geqslant m}\{X_n \in B\}$ (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 $\mathscr{A}_X$ nicht leer, 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 $\bigcup_{N=1}^{N_{max}} B_N = B_{N_{max}} \in \sigma(\bigcup_{k\geqslant n} \mathscr{F}^{X_k})$ für alle $N_{max} \geqslant n$. Dies impliziert $\bigcup_{N=1}^{\infty} B_N \in \sigma(\bigcup_{k\geqslant n} \mathscr{F}^{X_k})$ für alle $n \in \mathbb{N}$ und somit, dass A ein asymptotisches Ereignis ist.

2.38 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 $P(A) = 1$.*

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

2.40 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 2.30 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 2.21 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. $\square$

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 2.21 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 $P(A) \in \{0, 1\}$. □

2.41 Beispiel (Harmonische Reihe mit zufälligen Vorzeichen)**.** Sind $(\varepsilon_k)_{k\geqslant 1}$ unabhängige $\{-1, 1\}$-wertige Zufallsvariablen, so konvergiert nach Beispiel 2.37(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.

2.4 Faltung

2.42 Beispiel. Es seien X und Y unabhängige $\mathbb{Z}$-wertige Zufallsvariablen mit Zähldichten p^X und p^Y. Dann ist $X+Y$ wiederum eine $\mathbb{Z}$-wertige Zufallsvariable (Messbarkeit ist trivial wegen Potenzmenge als σ-Algebra). Die Zähldichte der Summe berechnet sich durch Fallunterscheidung

$$\begin{aligned} p^{X+Y}(m) &= P\Big(\bigcup_{k\in\mathbb{Z}} \{X = m-k, Y = k\}\Big) = \sum_{k\in\mathbb{Z}} P(X = m-k, Y = k) \\ &= \sum_{k\in\mathbb{Z}} p^X(m-k) p^Y(k), \quad m \in \mathbb{Z}, \end{aligned}$$

wobei wir im letzten Schritt die Unabhängigkeit von X und Y verwendet haben.

Für Folgen $(a_m)_{m\in\mathbb{Z}}$, $(b_m)_{m\in\mathbb{Z}}$ ist ihre <u>Faltung</u> $a * b$ definiert als die Folge $((a * b)_m)_{m\in\mathbb{Z}}$ mit

$$(a * b)_m := \sum_{k\in\mathbb{Z}} a_{m-k} b_k.$$

Diese ist wohldefiniert für ℓ^1-Folgen, das heißt falls $\sum_m |a_m| < \infty$, $\sum_m |b_m| < \infty$, und somit insbesondere für Zähldichten. Durch Indexsubstitution kann man sich leicht klarmachen, dass $a * b = b * a$ sowie $(a * b) * c = a * (b * c)$ gilt, die Faltung also kommutativ und assoziativ ist. Die Identität bezüglich Faltung ist gerade die Folge $e_m = \mathbf{1}(m = 0)$.

Für unabhängige $\mathbb{Z}$-wertige Zufallsvariablen X, Y berechnet sich die Zähldichte der Summe $p^{X+Y} = p^X * p^Y$ also gerade als Faltung der Zähldichten. Die Kommutativität und Assoziativität der Faltung ergibt sich stochastisch aus $X+Y = Y+X$ sowie $X+(Y+Z) = (X+Y)+Z$ für Zufallsvariablen X, Y, Z. Wegen $X+0 = X$ ist $p^0(m) = \mathbf{1}(m=0)$ als Zähldichte von der Einpunktverteilung δ_0 das neutrale Element.

Konkret erhalten wir für $X \sim \operatorname{Pois}(\lambda_X)$ und $Y \sim \operatorname{Pois}(\lambda_Y)$ unabhängig

$$\begin{aligned} p^{X+Y}(m) &= \sum_{k\in\mathbb{Z}} e^{-\lambda_X} \frac{\lambda_X^{m-k}}{(m-k)!} \mathbf{1}(m-k \geqslant 0) e^{-\lambda_Y} \frac{\lambda_Y^k}{k!} \mathbf{1}(k \geqslant 0) \\ &= e^{-(\lambda_X+\lambda_Y)} \sum_{k=0}^m \lambda_X^{m-k} \lambda_Y^k \binom{m}{k} \frac{1}{m!} \\ &= e^{-(\lambda_X+\lambda_Y)} \frac{(\lambda_X+\lambda_Y)^m}{m!}, \end{aligned}$$

wobei wir $m \geqslant 0$ betrachtet haben (für $m < 0$ zeigt die erste Zeile $p^{X+Y}(m) = 0$). Wir haben bewiesen, dass die Summe zweier unabhängiger Poisson-verteilter Zufallsvariablen wieder Poisson-verteilt ist. Der Parameter der Summe ist gerade die Summe der Parameter.

Der Zusammenhang zwischen der Summe unabhängiger Zufallsvariablen und der Faltung ihrer Verteilungen gilt sehr viel allgemeiner als auf der Gruppe $(\mathbb{Z}, +)$. Wir werden uns nur der wichtigsten Gruppe $(\mathbb{R}, +)$ widmen.

2.43 Definition. Sind P, Q Wahrscheinlichkeitsmaße auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$, so ist ihre Faltung $P * Q$ definiert als das Wahrscheinlichkeitsmaß

$$(P*Q)(B) = \int_\mathbb{R} P(B - \{x\})\, Q(dx), \quad B \in \mathfrak{B}_\mathbb{R}, \text{ mit } B - \{x\} = \{b - x \,|\, b \in B\}.$$

2.44 Bemerkung. Die Abbildung $x \mapsto P(B - \{x\})$ ist Borel-messbar wegen $P(B - \{x\}) = \int_\mathbb{R} \mathbf{1}_B(x+y)\, P(dy)$ und der Messbarkeitsaussage im Satz von Fubini. Insbesondere gilt

$$(P*Q)(B) = \int_\mathbb{R} \int_\mathbb{R} \mathbf{1}_B(x+y)\, P(dy)\, Q(dx). \tag{2.1}$$

In dieser Darstellung folgt die σ-Additivität einfach mit monotoner Konvergenz.

▷ **Kontrollfragen**

(a) Begründe $\bigcap_{m\geqslant 1} \bigcup_{n\geqslant m} \{X_n \in B\} = \bigcap_{m\geqslant k} \bigcup_{n\geqslant m} \{X_n \in B\}$ in Beispiel 2.37(a).

Wegen $\bigcup_{n\geqslant m+1} \{X_n \in B\} \subseteq \bigcup_{n\geqslant m} \{X_n \in B\}$ gilt $\bigcap_{m\leqslant k} \bigcup_{n\geqslant m} \{X_n \in B\} = \bigcup_{n\geqslant k} \{X_n \in B\}$ und damit die Identität.

(b) Wieso liegen für reellwertige Zufallsvariablen $(X_n)_{n\geqslant 1}$ die Ereignisse $\{\lim_{n\to\infty} X_n = a\}$ für $a \in \mathbb{R}$ in der asymptotischen σ-Algebra $\mathscr{A}_X$? [Freiwillig: Verstehen Sie die weiteren Aussagen von Korollar 2.39 im Klenke]

Es gilt $\{\lim_{n\to\infty} X_n = a\} = \bigcap_{\varepsilon>0, \varepsilon\in\mathbb{Q}} \bigcup_{N\geqslant 1} \bigcap_{n\geqslant N} \{|X_n - a| < \varepsilon\} \in \mathscr{A}_X$ analog zu Beispiel 2.37(b). Aus dem 0-1-Gesetz folgert man dann sofort, dass $P(\lim_{n\to\infty} X_n = a) \in \{0,1\}$ für unabhängige $(X_n)_{n\geqslant 1}$ gilt. Klenke zeigt aber sogar, dass in diesem Fall $\liminf_{n\to\infty} X_n$, $\limsup_{n\to\infty} X_n$ fast sicher gleich einem deterministischen Wert $a \in \mathbb{R}\cup\{-\infty,+\infty\}$ sind. Der Zufall verschwindet also gewissermaßen im Unendlichen. Das Gleiche gilt für $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i$, $\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i$, was wir bei den Gesetzen der großen Zahl noch genauer untersuchen werden.

(c) Gib für die zufällige harmonische Reihe in Beispiel 2.41 mit $P(\varepsilon_k = +1) = P(\varepsilon_k = -1) = 1/2$ unendlich viele Versuchsausgänge (d.h. Werte/Realisierungen von ε_k) an, für die die Reihe konvergiert bzw. divergiert.

Das Ereignis $\liminf_{k\to\infty}\{\varepsilon_k = 1\}$ enthält alle Folgen (ε_k), für die endliche viele $\varepsilon_k = -1$ sind. Das sind abzählbar unendlich viele Versuchsausgänge, für die alle $\sum_{k\geqslant 1} \varepsilon_k \frac{1}{k} = \infty$ gilt, die Reihe also divergiert. Ebenso konvergiert $\sum_{k\geqslant 1} \varepsilon_k \frac{1}{k}$ für alle Versuchsausgänge in $\liminf_{k\to\infty}\{\varepsilon_k = (-1)^{k+1}\}$, da jeweils die Reihenreste ab einem Index mit denen der alternierenden harmonischen Reihe übereinstimmen. Eine der beiden unendlichen Ereignismengen muss nach dem 0-1-Gesetz eine Nullmenge unter dem Prosuktmaß sein.

(d) Gilt auch $X + X \sim \mathrm{Pois}(2\lambda)$ für $X \sim \mathrm{Pois}(\lambda)$?

Nein, natürlich nicht (z.B. $P(X + X = 1) = P(X = 1/2) = 0 \neq p_{\mathrm{Pois}(\lambda)}(1)$).

(e) Wieso gilt $X_1 + X_2 \sim \mathrm{Bin}(n_1 + n_2, p)$ für unabhängige Zufallsvariablen $X_1 \sim \mathrm{Bin}(n_1, p)$, $X_2 \sim \mathrm{Bin}(n_2, p)$?

Wegen der Halbgruppeneigenschaft (Assoziativität) reicht es, den Fall $n_1 = n$ und $n_2 = 1$ zu betrachten. Wir erhalten für die Faltung

$$\begin{aligned}
p^{X_1+X_2}(m) &= \sum_{k\in\mathbb{Z}} \binom{n}{m-k} p^{m-k}(1-p)^{n-m+k}\binom{1}{k}p^k(1-p)^{1-k}\\
&= \Big(\binom{n}{m} + \binom{n}{m-1}\Big)p^m(1-p)^{n-m+1}\\
&= \binom{n+1}{m}p^m(1-p)^{n+1-m} = p_{\mathrm{Bin}(n+1,p)}(m),
\end{aligned}$$

wobei wir $\binom{n}{k} := 0$ für $k \notin \{0,\dots,n\}$ setzen.

Ende 9. Vorlesung

2.45 Satz. *Es seien X und Y unabhängige reellwertige Zufallsvariablen. Dann besitzt $X + Y$ die Verteilung $P^{X+Y} = P^X * P^Y$.*

Beweis. Unter Beachtung von $P^{(X,Y)} = P^X \otimes P^Y$ gemäß Lemma 2.30, dem Satz von Fubini und (2.1) erhalten wir für die Verteilungsfunktion

$$\begin{aligned}
F^{X+Y}(z) = P(X + Y \leqslant z) &= P^{(X,Y)}(\{(x,y)\in\mathbb{R}^2 \mid x + y \leqslant z\})\\
&= \int_{\mathbb{R}^2} \mathbf{1}_{(-\infty,z]}(x+y)\,(P^X\otimes P^Y)(dx,dy)\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbf{1}_{(-\infty,z]}(x+y)\,P^X(dx)\,P^Y(dy) = (P^X * P^Y)((-\infty,z])
\end{aligned}$$

für alle $z \in \mathbb{R}$. Nach Korollar 1.51 folgt $P^{X+Y} = P^X * P^Y$. □

2.46 Korollar. *Die Faltung ist kommutativ und assoziativ.*

Beweis. Dies folgt aus dem Satz wegen $X+Y=Y+X$ und $X+(Y+Z)=(X+Y)+Z$ für unabhängige Zufallsvariablen X,Y,Z. Alternativ sieht man die Kommutativität auch direkt mit der Darstellung (2.1) und dem Satz von Fubini. □

2.47 Korollar. *Es seien X und Y unabhängige reellwertige Zufallsvariablen und X besitze eine Dichte f^X. Dann besitzt $X+Y$ die Dichte*

$$f^{X+Y}(z)=\int_{\mathbb{R}} f^X(z-y)\,P^Y(dy),\quad z\in\mathbb{R}\,.$$

Falls auch Y eine Dichte besitzt, so gilt

$$f^{X+Y}(z)=f^X*f^Y(z):=\int_{\mathbb{R}} f^X(z-y)f^Y(y)\,dy,\quad z\in\mathbb{R}\,.$$

Beweis. Wir überprüfen die Verteilungsfunktion. Mit dem Satz von Fubini gilt für $x\in\mathbb{R}$

$$\begin{aligned}\int_{-\infty}^x\int_{\mathbb{R}} f^X(z-y)\,P^Y(dy)\,dz&=\int_{\mathbb{R}}\int_{-\infty}^x f^X(z-y)\,dz\,P^Y(dy)\\&=\int_{\mathbb{R}} P^X((-\infty,x-y])\,P^Y(dy)\\&=(P^X*P^Y)((-\infty,x])\end{aligned}$$

gemäß Definition von P^X*P^Y und $(-\infty,x]-\{y\}=(-\infty,x-y]$. Dies zeigt, dass $f^X(\bullet-y)\,P^Y(dy)$ die Dichte von P^X*P^Y ist. Im Fall einer Dichte f^Y folgt die zweite Formel mit der Integrationsregel aus Satz 3.7(b) unten (bzw. Maßtheorie). □

2.48 Beispiel. Für $\lambda,p>0$ definiere die <u>Gamma-Verteilung</u> $\Gamma(\lambda,p)$ über die Dichte

$$f_{\lambda,p}(x)=\frac{\lambda^p}{\Gamma(p)}x^{p-1}e^{-\lambda x}\mathbf{1}_{(0,\infty)}(x),\quad x\in\mathbb{R},$$

mit der Gamma-Funktion $\Gamma(p)=\int_0^\infty t^{p-1}e^{-t}dt$, so dass $\Gamma(n+1)=n!$ für $n\in\mathbb{N}_0$ und $\Gamma(1/2)=\sqrt{\pi}$. Spezialfälle sind $\Gamma(\lambda,1)=\mathrm{Exp}(\lambda)$, $\Gamma(1/2,1/2)=\chi^2(1)$. Wir erhalten für $z>0$ und mit einer Konstanten $C>0$ für die Faltung

$$\begin{aligned}(f_{\lambda,p_1}*f_{\lambda,p_2})(z)&=C\int_{\mathbb{R}}(z-y)^{p_1-1}e^{-\lambda(z-y)}y^{p_2-1}e^{-\lambda y}\mathbf{1}(y\geqslant 0,z-y\geqslant 0)\,dy\\&=Ce^{-\lambda z}\int_0^z(z-y)^{p_1-1}y^{p_2-1}\,dy\\&=Ce^{-\lambda z}\int_0^1(z-zw)^{p_1-1}(zw)^{p_2-1}z\,dw\\&=Ce^{-\lambda z}z^{p_1+p_2-1}\int_0^1(1-w)^{p_1-1}w^{p_2-1}\,dw\\&=f_{\lambda,p_1+p_2}(z).\end{aligned}$$

Beachte, dass eine Wahrscheinlichkeitsdichte sich zu Eins integriert und selbst dann festgelegt ist, wenn ein konstanter Faktor unbekannt ist. Wir haben $\Gamma(\lambda, p_1) * \Gamma(\lambda, p_2) = \Gamma(\lambda, p_1 + p_2)$ bewiesen.

Setzt man zusätzlich $\Gamma(\lambda, 0) = \delta_0$, so bildet $(\Gamma(\lambda, p))_{p \geqslant 0}$ für festes $\lambda > 0$ eine Faltungshalbgruppe, d.h. es gilt $\Gamma(\lambda, p_1) * \Gamma(\lambda, p_2) = \Gamma(\lambda, p_1 + p_2)$. Insbesondere ist die Summe von m unabhängigen $\chi^2(1)$-verteilten Zufallsvariablen (äquivalent: die quadrierte Norm $\|Z\|^2$ eines standard-normalverteilten Zufallsvektors $Z \sim N(0, E_m)$ im $\mathbb{R}^m$) gemäß $\chi^2(m) := \Gamma(1/2, m/2)$-verteilt ($\chi^2$-Verteilung mit m Freiheitsgraden). Im Spezialfall $m = 2$ ergibt sich $\chi^2(2) = \operatorname{Exp}(1/2)$, was für die Box-Müller-Methode zur Simulation der Normalverteilung (Übung!) genutzt wird.

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) $\mathbb{E}[X] = \sum_{x \in X(\Omega)} x P(X = x)$*; 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

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

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 $X \bullet Y$

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

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

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

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

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

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

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.

▷ **Kontrollfragen**

(a) Welche Dichte besitzt $U_1 + U_2$ für $U_1, U_2 \sim U([0,1])$ unabhängig?
Mit der Faltungsformel für Dichten erhalten wir

$$f^{U_1+U_2}(x) = \int_\mathbb{R} \mathbf{1}_{[0,1]}(x-y)\mathbf{1}_{[0,1]}(y)\,dy = \int_\mathbb{R} \mathbf{1}_{[0,1]\cap[x-1,x]}(y)\,dy = x_+ \wedge (2-x)_+.$$

$U_1 + U_2$ besitzt also die Dreiecksverteilung aus Beispiel 1.58. Als Dichte von $U_1 + \cdots + U_n$ für unabhängige $U([0,1])$-verteilte Zufallsvariablen $U_1, \ldots, U_n$ ergibt sich übrigens stets ein *B-Spline* der Ordnung $n-1$ (vergleiche Numerik).

(b) Betrachte eine deterministische Zahl $a \in \mathbb{R}$ als Zufallsvariable A mit Verteilung $P^A = \delta_a$. Welche Dichte besitzt $X + A$, wenn X die Dichte f^X besitzt?
Wir bestimmen die Dichte gemäß Faltungsformel

$$f^{X+A}(x) = \int_{\mathbb{R}} f^X(x-y)\delta_a(dy) = f^X(x-a), \quad x \in \mathbb{R}\,.$$

Dies hatten wir bereits aus dem Dichtetransformationssatz geschlossen.

(c) Es sei $X \sim \mathrm{Pois}(\lambda)$. Finde einfache X_n mit $X_n \uparrow X$. Wie folgt $\mathbb{E}[X] = \lambda$?
Die Zufallsvariablen $X_n = X \wedge n$ nehmen nur die Werte $0, 1, \dots, n$ an, sind also einfach. Ferner gilt $X_n(\omega) \uparrow X(\omega)$ für jedes $\omega \in \Omega$. Nun ist

$$\mathbb{E}[X_n] = \sum_{k=0}^{n} kP(X_n = k) = \sum_{k=1}^{n-1} e^{-\lambda}\frac{\lambda^k}{(k-1)!} + n\sum_{k \geqslant n} e^{-\lambda}\frac{\lambda^k}{k!}.$$

Für $n \to \infty$ konvergiert $\mathbb{E}[X_n]$ gegen $\sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-1)!} = \lambda$. Also gilt $\mathbb{E}[X] = \lambda$. Direkter kann man das natürlich über Satz 3.7(c) erhalten.

(d) Was ist $\mathbb{E}[X]$ für $X \sim \mathrm{Exp}(\lambda)$?
Nach Satz 3.7(b) berechne mit partieller Integration

$$\mathbb{E}[X] = \mathbb{E}[|X|] = \int_0^\infty x\lambda e^{-\lambda x}\,dx = -xe^{-\lambda x}|_{x=0}^{\infty} + \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}.$$

(e) X besitze die Cauchydichte $f^X(x) = \pi^{-1}\frac{1}{1+x^2}$, $x \in \mathbb{R}$. Was ist mit $\mathbb{E}[X]$?
Auch wenn man intuitiv X gerne den Erwartungswert null zuordnen möchte, existiert er nicht. Es gilt nämlich

$$\mathbb{E}[|X|] = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}dx \geqslant \int_1^\infty \frac{1}{2\pi x}dx = \infty,$$

so dass $X \notin \mathscr{L}^1$. Der Vorteil der strikten Forderung $\mathbb{E}[|X|] < \infty$ (statt z.B. $\mathbb{E}[X] = \lim_{R\to\infty}\int_{-R}^{R} xf^X(x)dx$ zu setzen) ist, dass wir ohne Bedenken alle Hilfsmittel der Lebesgueschen Integrationstheorie zur Verfügung haben.

Ende 10. Vorlesung

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

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

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). □

3.9 Beispiel. Für $X \sim N(0,1)$ gilt mit $(-e^{-x^2/2})' = xe^{-x^2/2}$

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\, P^X(dx) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} dx = \frac{-1}{\sqrt{2\pi}} e^{-x^2/2}\Big|_{x=-\infty}^{\infty} = 0$$

sowie mittels partieller Integration

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2\, P^X(dx) = \frac{-1}{\sqrt{2\pi}} \int_{\mathbb{R}} x\big(- xe^{-x^2/2}\big) dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} 1 \bullet e^{-x^2/2} dx = 1.$$

Alternativ kann man verwenden, dass $Y = X^2$ $\chi^2(1)$-verteilt ist, so dass $\mathbb{E}[X^2] = \int_0^\infty y(2\pi y)^{-1/2} e^{-y/2} dy = 1$.

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

(a) Der Erwartungswert ist linear: ist Y eine weitere Zufallsvariable in $\mathscr{L}^1$ und 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 Y eine weitere Zufallsvariable in $\mathscr{L}^1$ mit $X \leqslant Y$, so gilt $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$. Aus $X \leqslant Y$ und $\mathbb{E}[X] = \mathbb{E}[Y]$ folgt $P(X = Y) = 1$.

(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. $\square$

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; für $X \in \mathscr{L}^p$ und $p > 0$ heißt $\mathbb{E}[|X|^p]$ das p-te absolute Moment von X.

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

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

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. $\square$

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 $X \in \mathscr{L}^q$.*

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. $\square$

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

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

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

Die Funktion $\varphi(x) = \mathbb{E}[(X-x)^2]$ 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 φ die Summe aus dem quadrierten Bias, der genau für $x = \mathbb{E}[X]$ offensichtlich minimal ist, und der Varianz, die nicht von x abhängt. $\square$

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 $\mathbb{E}[X] = \operatorname{argmin}_x \varphi(x)$ 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 Wahrscheinlichkeitsmaßen, 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 <u>konvex</u>, 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 $\varphi'(x-)$, $\varphi'(x+)$ mit $\varphi'(x-) \leqslant \varphi'(x+)$. Es gilt weiterhin*

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

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

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

Für $x_n \downarrow x$ ist $D(x_n,x)$ also monoton fallend mit einem Grenzwert $\varphi'(x+) \in [-\infty,\infty)$. Dieselbe Ungleichung zeigt $D(x_n,x) \geqslant D(x_n',x)$ für $x_n' < x < x_n$. Für $x_n' \uparrow x$ ist also insbesondere $D(x_n',x)$ monoton wachsend mit Grenzwert $\varphi'(x-) \leqslant \varphi'(x+)$.

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 $D(y,x) \leqslant \varphi'(x-) \leqslant \varphi'(x+)$. □

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

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

die aus dem Lemma und dem Einsetzen von $x = y$ folgt (zeichne für die Beispiele oben!). An jeder Stelle y ist φ das Supremum über lineare Funktionen.

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* <u>*Jensensche Ungleichung*</u>

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

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)

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

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 $X \in \mathscr{L}^p \cap \mathscr{L}^q$.

▷ **Kontrollfragen**

(a) Wie folgt aus der Bias-Varianz-Zerlegung $\mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ für $X \in \mathscr{L}^2$?
Setze $x = 0$ in der Bias-Varianz-Zerlegung und stelle um.

(b) Es sei X $\mathrm{Exp}(\lambda)$-verteilt. In welchen $\mathscr{L}^p$-Räumen liegt X? Bestimme die Momente $\mathbb{E}[X^m]$, $m \in \mathbb{N}$.
Es gilt $X \geqslant 0$ und $\mathbb{E}[X^p] = \int_0^\infty x^p \lambda e^{-\lambda x} dx < \infty$, so dass $X \in \mathscr{L}^p$ für alle $p \in (0,\infty)$ gilt. Durch Substitution erhalten wir die Gamma-Funktion:

$$\mathbb{E}[X^p] = \lambda^{-p} \int_0^\infty z^p e^{-z} dz = \lambda^{-p}\Gamma(p+1).$$

Für $m \in \mathbb{N}$ folgt also $\mathbb{E}[X^m] = \lambda^{-m} m!$, insbesondere $\mathbb{E}[X] = \frac{1}{\lambda}$, $\mathbb{E}[X^2] = \frac{2}{\lambda^2}$.

(c) Gibt es für jedes $q > p > 0$ eine Zufallsvariable $X \in \mathscr{L}^p \setminus \mathscr{L}^q$? (Tipp: $\int_0^1 x^{-\alpha} dx < \infty \iff \alpha < 1$)
Es sei $U \sim U((0,1))$ und $X = U^{-\beta}$. Dann gilt $\mathbb{E}[|X|^p] = \mathbb{E}[U^{-p\beta}] < \infty \iff p\beta < 1$. Für $\beta \in [q^{-1}, p^{-1})$ gilt also $X \in \mathscr{L}^p \setminus \mathscr{L}^q$.

(d) Wie erhält man die Eigenschaften aus Beispiel 3.18 rigoros?
Die Monotonie der Ableitung impliziert für $x < y$, $t \in (0,1)$

$$\begin{aligned}(1-t)\varphi(x) + t\varphi(y) &= \varphi(x) + t\int_0^{y-x} \varphi'(x+s)\,ds\\ &\geqslant \varphi(x) + \int_0^{t(y-x)} \varphi'(x+s)\,ds = \varphi(x+t(y-x)),\end{aligned}$$

was Konvexität zeigt. Bilden der Ableitung zeigt daher die Konvexität von $e^{\alpha x}$ für $\alpha \in \mathbb{R}$ und $|x|^p$ für $p > 1$. Der Betrag ist konvex nach Dreiecksungleichung: $|\alpha x + (1-\alpha)y| \leqslant \alpha|x| + (1-\alpha)|y|$ für $\alpha \in [0,1]$.

(e) Gilt $\mathbb{E}[X \wedge R] \leqslant \mathbb{E}[X] \wedge R$ oder $\mathbb{E}[X \wedge R] \geqslant \mathbb{E}[X] \wedge R$ für alle $X \in \mathscr{L}^1$ und $R \in \mathbb{R}$? Argumentiere mit Jensenscher Ungleichung und elementar.

Die Abbildung $x \mapsto x \wedge R$ ist konkav, d.h. $x \mapsto -(x \wedge R)$ ist konvex (klar wegen $(\alpha x+(1-\alpha)y)\wedge R \geqslant \alpha(x\wedge R)+(1-\alpha)(y\wedge R)$). Die Jensensche Ungleichung zeigt also $\mathbb{E}[-(X \wedge R)] \geqslant -(\mathbb{E}[X] \wedge R)$ und daher $\mathbb{E}[X \wedge R] \leqslant \mathbb{E}[X] \wedge R$. Das ist natürlich auch elementar klar wegen Monotonie des Erwartungswerts: $\mathbb{E}[X\wedge R] \leqslant \mathbb{E}[X]$ und $\mathbb{E}[X\wedge R] \leqslant \mathbb{E}[R] = R$. Für $X \sim N(0,1)$ gilt $\mathbb{E}[X\wedge 0] < 0$, aber $\mathbb{E}[X] = 0$, so dass $\mathbb{E}[X \wedge R] \not\geqslant \mathbb{E}[X] \wedge R$ für $R = 0$.

Ende 11. Vorlesung

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*

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

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

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

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 $\operatorname{Var}(X) = 0$, so ist $X = \mathbb{E}[X]$ P-fast sicher und somit $\operatorname{Var}(Y - aX) = \operatorname{Var}(Y)$, so dass a^* beliebig gewählt werden kann. Wegen

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

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

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

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

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

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

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

Beweis. Übung! □

3.31 Beispiele.

(a) Eine $\operatorname{Bin}(n, p)$-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 $\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = np$, $\operatorname{Var}(X) = \sum_{i=1}^n \operatorname{Var}(X_i) = np(1 - p)$. Da Erwartungswert und Varianz nur von der Verteilung abhängen, sagt man, dass die $\operatorname{Bin}(n, p)$-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 $\mathrm{Var}(X) = \mathrm{Var}(\sigma Z + \mu) = \sigma^2 \mathrm{Var}(Z) = \sigma^2$.

(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: $P(D = m \mid S = k) = P(D = -m \mid S = k)$.

▷ **Kontrollfragen**

(a) Kann $\mathrm{Var}(X + Y) < \mathrm{Var}(X) + \mathrm{Var}(Y)$ für $X, Y \in \mathscr{L}^2$ gelten?
Ja, natürlich. Z.B. für $Y = -X$ gilt $\mathrm{Var}(X + Y) = 0$ und im Allgemeinen $\mathrm{Var}(X) = \mathrm{Var}(Y) > 0$. Allgemein gilt die Ungleichung immer, wenn $\mathrm{Cov}(X, Y) < 0$, also X und Y *negativ korreliert* sind.

(b) Was ist die Varianz einer Poisson-verteilten Zufallsvariablen?
Wir wissen bereits $\mathbb{E}[X] = \lambda$ für $X \sim \mathrm{Pois}(\lambda)$. Ferner ist

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k\geqslant 0} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k\geqslant 1} \Big(\frac{\lambda^k}{(k-1)!} + \frac{\lambda^k}{(k-2)!} \Big) \\ &= e^{-\lambda}(\lambda e^{\lambda} + \lambda^2 e^{\lambda}) = \lambda + \lambda^2,\end{aligned}$$

wobei wir $\frac{k^2}{k!} = \frac{k-1+1}{(k-1)!} = \frac{1}{(k-2)!} + \frac{1}{(k-1)!}$ benutzt haben und den ersten Summanden im Fall $k = 1$ als Null definieren. Also folgt

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

(c) Was sind Beispiele für Zufallsvariablen X, Y mit $\rho(X, Y) = 0$, $\rho(X, Y) = 1$ bzw. $\rho(X, Y) = -1$?
$\rho(X, Y) = 0$ gilt z.B., wenn X und Y unabhängig sind; $\rho(X, Y) = 1$ gilt, wenn $Y = aX + b$ ist mit $a > 0$, $b \in \mathbb{R}$; $\rho(X, Y) = -1$ gilt, wenn $Y = aX + b$ ist mit $a < 0$, $b \in \mathbb{R}$.

(d) Es sei $\mathscr{L}_z^2 = \{X \in \mathscr{L}^2(\Omega, \mathscr{F}, P) \mid \mathbb{E}[X] = 0\}$ die Menge der zentrierten $\mathscr{L}^2$-Zufallsvariablen auf einem Wahrscheinlichkeitsraum (*zentriert* heißt $\mathbb{E}[X] = 0$). Ist $\mathscr{L}_z^2$ ein Vektorraum und die Kovarianz ein Skalarprodukt?

$\mathscr{L}^2_z$ ist in der Tat ein Vektorraum, da skalare Vielfache und Summen von $\mathscr{L}^2$-Zufallsvariablen wieder in $\mathscr{L}^2$ liegen und weil der Erwartungswert dabei Null bleibt. Die Kovarianz ist im Allgemeinen kein Skalarprodukt, da $\mathrm{Cov}(X,X) = 0$ nicht unbedingt $X = 0$ impliziert. Identifiziert man jedoch Zufallsvariablen X, Y mit $P(X = Y) = 1$, so gilt $\mathbb{E}[X^2] = \mathrm{Cov}(X,X) = 0 \iff X = 0 P$-f.s. Die Kovarianz ist weiterhin symmetrisch und linear in jedem Argument sowie stets nicht-negativ, so dass sie unter Identifizierung in der Tat ein Skalarprodukt definiert, dessen zugehörige Norm die Standardabweichung $\sigma(X)$ ist.

(e) Beim n-fachen Münzwurf mit $X_1, \dots, X_n \in \{0,1\}$ bestimme die beste lineare Vorhersage von X_1 aus der Kenntnis der Gesamtzahl der Erfolge $S_n = X_1 + \cdots + X_n$. Ist dies auch die beste Vorhersage durch $Z = g(S_n)$ mit $g : \mathbb{R} \to \mathbb{R}$ beliebig im Fall $n = 2, 3, \dots$?

Es ist $\mathbb{E}[X_1] = 1/2$, $\mathbb{E}[S_n] = n/2$, $\mathrm{Var}(X_1) = 1/4$, $\mathrm{Var}(S_n) = \sum_{i=1}^n \mathrm{Var}(X_i) = n/4$ sowie

$$\mathrm{Cov}(X_1, S_n) = \sum_{i=1}^n \mathrm{Cov}(X_1, X_i) = \mathrm{Var}(X_1) = 1/4.$$

Also ist $a^* S_n + b^*$ beste lineare Vorhersage mit $a^* = \frac{1/4}{n/4} = \frac{1}{n}$, $b^* = 1/2 - a^* n/2 = 0$. Wir sagen X_1 also durch S_n/n voraus, was wegen der Austauschbarkeit von $X_1, \dots, X_n$ auch intuitiv ist.

Im nichtlinearen Fall für $n = 2$ müssen wir $g(0), g(1), g(2)$ so bestimmen, dass

$$\mathbb{E}[(X_1 - g(S_2))^2] = \tfrac{1}{4}\big((1 - g(1))^2 + (1 - g(2))^2 + g(0)^2 + g(1)^2\big)$$

minimal ist. Das ergibt $g(0) = 0$, $g(2) = 1$, $g(1) = 1/2$, also $g(S_2) = S_2/2$ wie im linearen Fall. Eine ähnliche Rechnung ergibt $g(S_3) = S_3/3$ im Fall $n = 3$. Für beliebige n kommt man auf die Idee, für jedes $k = 0, \dots, n$ einzeln $g(k)$ zu bestimmen als Minimierer von

$$\begin{aligned}
&\mathbb{E}[(X_1 - g(S_n))^2 \mathbf{1}(S_n = k)] \\
&= \sum_{\ell=0}^{1} (\ell - g(k))^2 P(X_2 + \cdots + X_n = k - \ell, X_1 = \ell) \\
&= \sum_{\ell=0}^{1} (\ell - g(k))^2 \binom{n-1}{k-\ell} 2^{-n},
\end{aligned}$$

wobei wir die Unabhängigkeit von $X_2 + \cdots + X_n \sim \mathrm{Bin}(n-1, 1/2)$ und $X_1 \sim \mathrm{Bin}(1, 1/2)$ ausgenutzt haben. Im Fall $k = n$ gilt natürlich $g(n) = 1$ und ebenso $g(0) = 0$, im Fall $k = 1, \dots, n-1$ betrachten wir

$$F(\gamma) := \gamma^2 \binom{n-1}{k} + (1-\gamma)^2 \binom{n-1}{k-1} \text{ mit } F'(\gamma) = 2\gamma \binom{n}{k} - 2\binom{n-1}{k-1}.$$

Dann wird F bei $\gamma^* = k/n$ minimal, so dass $\mathbb{E}[(X_1 - g(S_n))^2 \mathbf{1}(S_n = k)]$ für $g(S_n) = S_n/n$ minimal wird. Summieren über k zeigt daher insgesamt, dass $g(S_n) = S_n/n$ auch die beste Vorhersage von X_1 unter allen nichtlinearen Funktionen ist. In Stochastik II wird man $g(S_n)$ als bedingte Erwartung $\mathbb{E}[X_1 \,|\, S_n]$ studieren.

Ende 12. Vorlesung

3.3 Mehrdimensionale Normalverteilung

3.32 Definition. Es seien $\mu \in \mathbb{R}^d$ ein Vektor, $\Sigma \in \mathbb{R}^{d\times d}$ eine symmetrische, positiv semi-definite Matrix und Z ein standard-normalverteilter Vektor im $\mathbb{R}^d$. Die Verteilung des Zufallsvektors $X = \mu + \Sigma^{1/2}Z$ heißt dann $\underline{d\text{-dimensionale}}$ Normalverteilung mit Mittelwert(vektor) μ und Kovarianzmatrix Σ, Notation $X \sim N(\mu, \Sigma)$.

Im Fall einer strikt positiv-definiten Matrix Σ besitzt $N(\mu, \Sigma)$ (gemäß Beispiel 1.79) die Dichte

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

3.33 Bemerkung. Σ ist positiv-semidefinit, falls $\langle \Sigma v, v\rangle \geqslant 0$ für alle $v \in \mathbb{R}^d$ gilt. Da jede symmetrische Matrix diagonalisierbar ist, ist diese Eigenschaft äquivalent dazu, dass alle Eigenwerte λ_i nicht-negativ sind. Es gilt also $\Sigma = T^{-1}DT$ mit einer invertierbaren (sogar orthogonalen) Matrix T und einer Diagonalmatrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$. Man setzt dann $\Sigma^{1/2} = T^{-1}D^{1/2}T$ mit $D^{1/2} = \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2})$. Man überprüft leicht, dass $\Sigma^{1/2}$ ebenfalls symmetrisch, positiv-semidefinit ist mit $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. Wenn sogar $\lambda_i > 0$ für $i = 1, \ldots, d$ gilt, so heißt Σ (strikt) positiv-definit.

3.34 Beispiel. Ist Σ nicht invertierbar, so nimmt $X = \mu + \Sigma^{1/2}Z$ Werte in einem echten (affinen) Unterraum des $\mathbb{R}^d$ an, dessen Lebesguemaß Null ist. In diesem Fall besitzt $N(\mu, \Sigma)$ also keine Dichte. Für $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ beispielsweise gilt $\Sigma^{1/2} = \Sigma$ und $P(X_2 = \mu_2) = 1$, der Träger von $P^X = N(\mu, \Sigma)$ ist die Gerade $\{(x, \mu_2) \,|\, x \in \mathbb{R}\} \subseteq \mathbb{R}^2$.

3.35 Satz. *Für einen $N(\mu, \Sigma)$-verteilten Zufallsvektor $X = (X_1, \ldots, X_d)$ und $1 \leqslant k, \ell \leqslant d$ gilt*

$$\mathbb{E}[X_k] = \mu_k, \quad \operatorname{Cov}(X_k, X_\ell) = \Sigma_{k,\ell}.$$

3.36 Bemerkung. In diesem Sinne ist μ Erwartungswert und Σ Kovarianzmatrix von $N(\mu, \Sigma)$.

Beweis. Aus der Linearität des Erwartungswerts und $Z_\ell \sim N(0, 1)$, $\ell = 1, \ldots, d$, folgt

$$\mathbb{E}[X_k] = \mathbb{E}[\mu_k + (\Sigma^{1/2}Z)_k] = \mu_k + \sum_{\ell=1}^d \Sigma_{k,\ell}^{1/2}\,\mathbb{E}[Z_\ell] = \mu_k$$

sowie wegen $\mathbb{E}[Z_iZ_j] = \mathbf{1}(i = j)$ (für $i \neq j$ ist $\mathbb{E}[Z_iZ_j] = \mathbb{E}[Z_i]\,\mathbb{E}[Z_j] = 0$, da Z_i, Z_j unabhängig)

$$\begin{aligned}\operatorname{Cov}(X_k, X_\ell) &= \mathbb{E}[(X_k - \mu_k)(X_\ell - \mu_\ell)] = \mathbb{E}[(\Sigma^{1/2}Z)_k(\Sigma^{1/2}Z)_\ell] \\ &= \sum_{i,j=1}^d \Sigma_{k,i}^{1/2}\Sigma_{\ell,j}^{1/2}\,\mathbb{E}[Z_iZ_j] = \sum_{i=1}^d \Sigma_{k,i}^{1/2}\Sigma_{\ell,i}^{1/2} \\ &= (\Sigma^{1/2}\Sigma^{1/2})_{k,\ell} = \Sigma_{k,\ell}.\end{aligned}$$

Beachte bei den Rechnungen die Notation $\Sigma_{k,\ell}^{1/2} := (\Sigma^{1/2})_{k,\ell}$. $\square$

3.37 Korollar. *Sind $X_1, \dots, X_n$ gemeinsam normalverteilt (d.h. $X = (X_1, \dots, X_n)$ ist n-dimensional normalverteilt) und sind $X_1, \dots, X_n$ (paarweise) unkorreliert, so sind $X_1, \dots, X_n$ sogar unabhängig.*

Beweis. Nach Voraussetzung gilt $\operatorname{Cov}(X_k, X_\ell) = 0$ für alle $k \neq \ell$ und $X \sim N(\mu, \Sigma)$. Nach dem Satz folgt $\Sigma_{k,\ell} = 0$ für $k \neq \ell$, so dass Σ Diagonalmatrix ist. Also gilt $X_k = \mu_k + \Sigma_{k,k} Z_k$, $1 \leqslant k \leqslant d$. Nun sind $Z_1, \dots, Z_n$ unabhängige Zufallsvariablen und daher nach Lemma 2.20 auch $X_1, \dots, X_n$. □

3.38 Satz. *Ist X ein $N(0, \Sigma)$-verteilter Zufallsvektor im $\mathbb{R}^d$ und ist $A \in \mathbb{R}^{m \times d}$ eine deterministische Matrix, so ist $Y = AX$ ein $N(0, A\Sigma A^\top)$-verteilter Zufallsvektor im $\mathbb{R}^m$.*

Beweis. Wir müssen zeigen, dass sich Y darstellen lässt als $Y = (A\Sigma A^\top)^{1/2} Z$ mit einer geeigneten Zufallsvariablen $Z \sim N(0, E_m)$. Aus der Darstellung $X = \Sigma^{1/2} W$ mit $W \sim N(0, E_d)$ ergibt sich die zu erfüllende Bedingung als $A\Sigma^{1/2} W = (A\Sigma A^\top)^{1/2} Z$.

Der Satz zur orthogonalen Normalform (z.B. in M. Koecher, Lineare Algebra und Analytische Geometrie, Seite 199) zeigt, dass es orthogonale Matrizen $T_1 \in \mathbb{R}^{m \times m}, T_2 \in \mathbb{R}^{d \times d}$ und eine Diagonalmatrix $D \in \mathbb{R}^{r \times r}$, $r \leqslant \min(m, d)$, mit strikt positiven Diagonaleinträgen gibt, so dass in Blockmatrixnotation (beachte jeweils die Dimensionen!) $A\Sigma^{1/2} = T_1 \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} T_2$ gilt. Dies impliziert (beachte $T_2^\top T_2 = E_d$ da T_2 orthogonale Matrix)

$$\begin{aligned}(A\Sigma A^\top)^{1/2} = (A\Sigma^{1/2}(A\Sigma^{1/2})^\top)^{1/2} &= \left(T_1 \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}^\top T_1^\top\right)^{1/2} \\ &= T_1 \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} T_1^\top.\end{aligned}$$

Wir müssen also $Z \sim N(0, E_m)$, $W \sim N(0, E_d)$ auf einem gemeinsamen Wahrscheinlichkeitsraum finden mit

$$\begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} T_1^\top Z = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} T_2 W. \tag{3.2}$$

Im Fall $m \leqslant d$ sei $W \sim N(0, E_d)$ gegeben und setze $Z = T_1 \begin{pmatrix} E_m & 0 \end{pmatrix} T_2 W$. Dann gilt (3.2). Nach Beispiel 1.79 ist die Standard-Normalverteilung invariant unter orthogonalen Transformationen, so dass $T_2 W \sim N(0, E_d)$. $\begin{pmatrix} E_m & 0 \end{pmatrix} T_2 W$ ist dann ein Vektor von m unabhängigen $N(0,1)$-verteilten Zufallsvariablen, also $N(0, E_m)$-verteilt, so dass auch $Z \sim N(0, E_m)$ gilt wie gefordert. Im Fall $m > d$ sei $Z \sim N(0, E_m)$ gegeben, setze $W = T_2^\top \begin{pmatrix} E_d & 0 \end{pmatrix} T_1^\top Z$ und schließe analog, dass (3.2) gilt sowie $W \sim N(0, E_d)$. □

3.39 Korollar. *Ist X ein $N(\mu, \Sigma)$-verteilter Zufallsvektor im $\mathbb{R}^d$ und sind $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$ deterministisch, so ist $Y = AX + b$ ein $N(A\mu + b, A\Sigma A^\top)$-verteilter Zufallsvektor im $\mathbb{R}^m$.*

Beweis. Nach Definition ist $X - \mu \sim N(0, \Sigma)$ (schreibe $X = \mu + \Sigma^{1/2} Z$) und nach dem Satz $A(X - \mu) \sim N(0, A\Sigma A^\top)$. Also ist $Y = A(X - \mu) + A\mu + b \sim N(A\mu + b, A\Sigma A^\top)$. □

3.40 Korollar. *Sind X_1 und X_2 unabhängige $N(\mu_1,\Sigma_1)$- bzw. $N(\mu_2,\Sigma_2)$-verteilte d-dimensionale Zufallsvektoren, so gilt $X_1+X_2 \sim N(\mu_1+\mu_2,\Sigma_1+\Sigma_2)$. Insbesondere bildet $(N(\mu t,\sigma^2 t))_{t\geqslant 0}$ für $\mu\in\mathbb{R}$, $\sigma\geqslant 0$ eine Faltungshalbgruppe.*

Beweis. Es sei $Y\sim N(\mu,\Sigma)$ ein $2d$-dimensionaler Zufallsvektor mit

$$\mu=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\quad \Sigma=\begin{pmatrix}\Sigma_1 & 0\\ 0 & \Sigma_2\end{pmatrix}.$$

Nach dem vorigen Korollar gilt dann für die Projektionen $\tilde X_1 := \begin{pmatrix}E_d & 0\end{pmatrix} Y \sim N(\mu_1,\Sigma_1)$, $\tilde X_2 := \begin{pmatrix}0 & E_d\end{pmatrix} Y \sim N(\mu_2,\Sigma_2)$. Aus $Y=\mu+\Sigma^{1/2}Z$ mit $Z\sim N(0,E_{2d})$ und $\Sigma^{1/2} = \begin{pmatrix}\Sigma_1^{1/2} & 0\\ 0 & \Sigma_2^{1/2}\end{pmatrix}$ (überprüfe!) folgt, dass $(Y_i)_{1\leqslant i\leqslant d}$ eine messbare Funktion von $(Z_i)_{1\leqslant i\leqslant d}$ ist und ebenso $(Y_i)_{d+1\leqslant i\leqslant 2d}$ eine messbare Funktion von $(Z_i)_{d+1\leqslant i\leqslant 2d}$. Nun sind $(Z_i)_{1\leqslant i\leqslant d}$ und $(Z_i)_{d+1\leqslant i\leqslant 2d}$ unabhängig, vergleiche Lemma 2.40, und somit auch $\tilde X_1$ und $\tilde X_2$. Wiederum nach dem Korollar mit $A=\begin{pmatrix}E_d & E_d\end{pmatrix}\in\mathbb{R}^{d\times 2d}$ gilt $\tilde X_1+\tilde X_2 = AY\sim N(\mu_1+\mu_2,\Sigma_1+\Sigma_2)$. Da $(\tilde X_1,\tilde X_2)$ wie (X_1,X_2) verteilt sind, ist die erste Aussage bewiesen.

Die zweite Aussage, also $N(\mu t,\sigma^2 t)*N(\mu s,\sigma^2 s)=N(\mu(t+s),\sigma^2(t+s))$ für $t,s\geqslant 0$, folgt aus der ersten mit $d=1$, $\mu_1=\mu t$, $\mu_2=\mu s$, $\Sigma_1=\sigma^2 t$, $\Sigma_2=\sigma^2 s$ unter Beachtung von Satz 2.45. □

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 $K>0$ 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 $\varphi(X(\omega))\geqslant \varphi(K)\mathbf{1}(X(\omega)\geqslant K)$ für alle $\omega\in\Omega$ (Grundmenge, wo X definiert). 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 $\mathbb{E}[\varphi(X)]\in[0,\infty]$ wegen $\varphi(X)\geqslant 0$, und dass die Ungleichung im Fall $\mathbb{E}[\varphi(X)]=\infty$ trivial ist. □

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.

▷ **Kontrollfragen**

(a) Überprüfe, dass $\Sigma^{1/2}$ wiederum symmetrisch, positiv-semidefinit ist und dass $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$ gilt. [Man kann zeigen, dass diese Eigenschaften die Wurzel $\Sigma^{1/2}$ eindeutig charakterisieren.]
Mit $\Sigma^{1/2} = T^{-1}D^{1/2}T$ und $T^{-1} = T^\top$ (T ist orthogonal) folgt $(\Sigma^{1/2})^\top = T^\top D^{1/2}(T^{-1})^\top = \Sigma^{1/2}$, $\langle \Sigma^{1/2}v, v\rangle = \langle D^{1/2}Tv, Tv\rangle = \sum_{i=1}^d \lambda_i^{1/2}(Tv)_i^2 \geqslant 0$ sowie $\Sigma^{1/2}\Sigma^{1/2} = T^{-1}D^{1/2}T^{-1}TD^{1/2}T = \Sigma$.

(b) Es sei $X_\sigma \sim N(0,\sigma^2)$ und $f : \mathbb{R} \to \mathbb{R}$ stetig und beschränkt. Wieso gilt $\lim_{\sigma\downarrow 0} \mathbb{E}[f(X_\sigma)] = \mathbb{E}[f(X_0)]$, so dass wir $N(0,0)$ als stetige Fortsetzung von $N(0,\sigma^2)$ für $\sigma > 0$ verstehen können?
Schreibe $X_\sigma = \sigma Z$ mit $Z \sim N(0,1)$. Dann gilt $f(X_\sigma) = f(\sigma Z) \to f(0)$ für $\sigma \to 0$, weil f stetig ist. Mit dominierter Konvergenz ($\|f\|_\infty < \infty$) folgt $\mathbb{E}[f(X_\sigma)] \to \mathbb{E}[f(0)] = f(0)$. Später werden wir daher sagen, dass X_σ gegen X_0 in Verteilung konvergiert oder $N(0,\sigma^2)$ schwach gegen $N(0,0)$ konvergiert für $\sigma \to 0$.

(c) Es seien X, Y unabhängige Zufallsvariablen mit $X \sim N(0,1)$ und $P(Y=1) = P(Y=-1) = 1/2$. Setze $X_1 = X$, $X_2 = XY$. Zeige: $X_1, X_2 \sim N(0,1)$, $\operatorname{Cov}(X_1,X_2) = 0$, aber X_1 und X_2 sind nicht unabhängig. Wieso widerspricht das nicht Korollar 3.37?
Für die Verteilungsfunktion von X_2 gilt

$$F^{X_2}(x) = P(XY \leqslant x) = \tfrac{1}{2}P(X \leqslant x) + \tfrac{1}{2}P(-X \leqslant x) = P(X \leqslant x),$$

weil auch $-X$ $N(0,1)$-verteilt ist. Das zeigt $X_2 \sim N(0,1)$. Außerdem gilt wegen Unabhängigkeit von X und Y sowie $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[YX] = 0$:

$$\operatorname{Cov}(X_1,X_2) = \operatorname{Cov}(X,YX) = \mathbb{E}[X\bullet(YX)] = \mathbb{E}[X^2]\,\mathbb{E}[Y] = 0.$$

Natürlich sind X_1 und X_2 nicht unabhängig, z.B. ist $P(|X_1| \leqslant 1, |X_2| \leqslant 1) = P(|X| \leqslant 1) \in (0,1)$, aber

$$P(|X_1| \leqslant 1)P(|X_2| \leqslant 1) = P(|X| \leqslant 1)^2 \neq P(|X_1| \leqslant 1, |X_2| \leqslant 1).$$

Der Vektor (X_1, X_2) ist nicht gemeinsam normalverteilt. Sonst wäre $X_1 + X_2 = (1+Y)X$ normalverteilt mit Erwartungswert 0 und Varianz $\mathbb{E}[(1+Y)^2X^2] = \mathbb{E}[1+2Y+Y^2]\,\mathbb{E}[X^2] = 2$, was wegen $P(X_1+X_2 = 0) = P(Y=-1) = 1/2$ falsch ist.

(d) Wie ist X_1 verteilt, falls $(X_1,\dots,X_d) \sim N(\mu,\Sigma)$? Wie ist $(X_1,\dots,X_m)$ verteilt für $m \leqslant d$?
Schreibe $X_1 = AX$ mit $A = (1,0,\dots,0) \in \mathbb{R}^{1\times d}$. Dann gilt $X_1 \sim N(A\mu, A\Sigma A^\top) = N(\mu_1, \Sigma_{1,1})$. Entsprechend folgt mit $A = (E_m\ 0) \in \mathbb{R}^{m\times d}$, dass $(X_1,\dots,X_m) \sim N((\mu_1,\dots,\mu_m), (\Sigma_{i,j})_{1\leqslant i,j\leqslant m})$.

(e) Wie lässt sich Satz 3.38 im Fall von $m = d$ und invertierbaren Matrizen Σ, A elementar über die Dichteformel beweisen?
Nach dem Dichtetransformationssatz gilt

$$\begin{aligned}
f^Y(y) &= f^X(A^{-1}y)|\det(A^{-1})| \\
&= (2\pi)^{-d/2}\det(\Sigma)^{-1/2}\exp\big(-\tfrac{1}{2}\langle \Sigma^{-1}A^{-1}y, A^{-1}y\rangle\big)\det(A^\top A)^{-1/2} \\
&= (2\pi)^{-d/2}\det(A\Sigma A^\top)^{-1/2}\exp\big(-\tfrac{1}{2}\langle (A\Sigma A^\top)^{-1}y, y\rangle\big).
\end{aligned}$$

Also ist Y $N(0, A\Sigma A^\top)$-verteilt.

Ende 13. Vorlesung

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

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

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

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

□

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 $S_n \sim \operatorname{Bin}(n, 1/2)$. Wir erhalten $\mathbb{E}[\hat{p}] = \mathbb{E}[S_n]/n = 1/2$, $\operatorname{Var}(\hat{p}) = \operatorname{Var}(S_n)/n^2 = 1/(4n)$. Die Tschebyschev-Ungleichung gibt die Abschätzung

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

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.

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 $\sup_i \operatorname{Var}(X_i) < \infty$. 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

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

Mit der Tschebyschev-Ungleichung folgt also

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

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 $X_1, \dots, X_n$ unabhängige $\mathrm{Bin}(1,p)$-verteilte Zufallsvariablen mit $p \in [0,1]$. Dann gilt $\sum_{i=1}^n X_i \sim \mathrm{Bin}(n,p)$ und somit

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

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

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

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 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! $\square$

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.

▷ **Kontrollfragen**

(a) Wieso zeigt der Fall $X \sim \mathrm{Bin}(1, 1/2)$, dass die Tschebyschev-Ungleichung 'scharf' ist?

Es gilt $P(|X - \mathbb{E}[X]| \geqslant 1/2) = 1$ sowie $\mathrm{Var}(X) = 1/4$, also Gleichheit in der Tschebyshev-Ungleichung mit $K = 1/2$.

(b) Es seien A_n, A Ereignisse mit $P(A_n) \to P(A)$. Gilt dann $\mathbf{1}_{A_n} \xrightarrow{P} \mathbf{1}_A$? Was ist für $A = \varnothing$?

Ist A ein Ereignis mit $P(A) = 1/2$ und $A_n = A^c$, so gilt $P(A_n) = P(A)$, aber $|\mathbf{1}_{A_n} - \mathbf{1}_A| = 1$, so dass $\mathbf{1}_{A_n} \not\to \mathbf{1}_A$ stochastisch. Für $P(A) = 0$, also auch $A = \varnothing$, und $P(A_n) \to 0$ hingegen gilt für $\varepsilon \in (0,1)$, dass $P(|\mathbf{1}_{A_n} - \mathbf{1}_A| > \varepsilon) = P(A_n) \to 0$ und somit $\mathbf{1}_{A_n} \xrightarrow{P} \mathbf{1}_A = 0$.

(c) Zeigen Sie mindestens eine Richtung der Äquivalenz $d_0(X_n, X) \to 0 \iff X_n \xrightarrow{P} X$ in Bemerkung 4.12.

Aus $d_0(X_n, X) \to 0$ folgt für $\varepsilon \in (0,1)$ mittels Markov-Ungleichung:

$$P(|X_n - X| > \varepsilon) = P(|X_n - X| \wedge 1 > \varepsilon) \leqslant \varepsilon^{-1}\,\mathbb{E}[|X_n - X| \wedge 1] \to 0,$$

also stochastische Konvergenz. Die andere Richtung lässt sich wegen $|X_n - X| \wedge 1 \in \mathscr{L}^1$ aus Satz 4.19(d) unten mit $p = 1$ gewinnen oder aber nachrechnen; denn für $\varepsilon \in (0,1)$ gilt

$$\mathrm{d}_0(X_n, X) \leqslant \mathbb{E}[\mathbf{1}(|X_n - X| > \varepsilon) + \varepsilon\mathbf{1}(|X_n - X| \leqslant \varepsilon)] \leqslant P(|X_n - X| > \varepsilon) + \varepsilon$$

und die rechte Seite konvergiert gegen ε, was $d_0(X_n, X) \to 0$ zeigt, da ε beliebig war.

(d) Wieso folgt aus $X_n \to X$ P-f.s., dass auch $X_n^2 \to X^2$ P-f.s.?

Dies folgt aus der Inklusion $\{X_n \to X\} \subseteq \{X_n^2 \to X^2\}$ ($x \mapsto x^2$ ist stetig) und $P(X_n \to X) = 1$.

(e) Im Beweis des starken Gesetzes wurde die Teilfolge $(A_{n^2})_{n\geqslant 1}$ betrachtet. Für welche Teilfolgen der Form $(A_{\lfloor n^\alpha \rfloor})_{n\geqslant 1}$ mit $\alpha \in [1, \infty)$ geht der Beweis genauso durch?

Für jedes $\alpha > 1$ konvergiert $A_{\lfloor n^\alpha \rfloor} \to \mu$ P-fast sicher wegen $\sum_{n\geqslant 1} \frac{1}{\lfloor n^\alpha \rfloor} < \infty$. Mit $\lfloor n(m)^\alpha \rfloor \leqslant m < \lfloor (n(m)+1)^\alpha \rfloor$ folgt dann

$\frac{1}{\lfloor n(m)^\alpha \rfloor}\left|\sum_{i=1}^m X_i - \sum_{i=1}^{\lfloor n(m)^\alpha \rfloor} X_i\right| \xrightarrow{P\text{-f.s.}} 0$ für $m \to \infty$, weil wir für die Summierbarkeit der Wahrscheinlichkeiten im Borel-Cantelli-Argument nur $\sum_{n\geqslant 1} n^{\alpha-1}\frac{n^{\alpha-1}}{n^{2\alpha}} < \infty$ benötigen, was für alle $\alpha > 1$ gilt. Daraus erhält man wie gehabt das starke Gesetz der großen Zahlen. Also tut es jedes $\alpha > 1$.

Ende 14. Vorlesung

4.15 Definition. Eine Zahl $x \in [0,1]$ heißt normal, falls für jedes $m \in \mathbb{N}$ jede Ziffernfolge der Länge m in ihrer Dezimalentwicklung gleich häufig auftritt, genauer falls $x = \sum_{k=1}^\infty x_k 10^{-k}$ gilt mit $x_k \in \{0, \dots, 9\}$ und

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n \mathbf{1}\Big(x_{(k-1)m+1}10^{m-1} + x_{(k-1)m+2}10^{m-2} + \cdots + x_{km} = z\Big) = 10^{-m}$$

für alle $z = 0, 1, \dots, 10^m - 1$ und $m \in \mathbb{N}$.

4.16 Bemerkung. Es gibt nur wenige Konstruktionen normaler Zahlen, die einfachste ist $x = 0{,}12345678910111213\dots$ Es ist nicht bekannt, ob beispielsweise die Nachkommastellen von π normal sind, vergleiche Spiegel vom 29.4.13. Der nächste Satz besagt, dass im stochastischen Sinn fast alle Zahlen normal sind. Im Georgii wird übrigens auch der Fall von Überschneidungen für $m \geqslant 2$ behandelt.

4.17 Korollar (Borels Gesetz der normalen Zahlen)**.** *Eine auf* $[0,1]$ *gleichverteilte Zahl ist fast sicher normal.*

Beweis. Wir betrachten allgemein q-adische Entwicklungen für $q \in \mathbb{N}\setminus\{1\}$. Dazu seien $(X_k)_{k\geqslant 1}$ unabhängige, auf $\{0, 1, \dots, q-1\}$ gleichverteilte Zufallsvariablen. Wir zeigen zunächst, dass dann $X := \sum_{k\geqslant 1} X_k q^{-k}$ $U([0,1])$-verteilt ist. Dazu reicht es, für Intervalle $I(j,k) := [jq^{-k}, (j+1)q^{-k})$, $j \in \{0, 1, \dots, q^k - 1\}$, $k \in \mathbb{N}$, die Identität $P(X \in I(j,k)) = |I(j,k)| = q^{-k}$ nachzuweisen ($\cap$-stabiler Erzeuger!). Dies ist wegen $\{X \in I(j,k)\} = \{\sum_{l=1}^k X_l q^{k-l} = j\}$ und der Wahrscheinlichkeit $(1/q)^k$, dass jedes X_l die $(k-l)$-te Ziffer, $l = 1, \dots, k$, in der q-adischen Darstellung von j annimmt, klar.

In der q-adischen Entwicklung von X ist fast sicher jede Ziffernfolge gleich häufig; denn nach dem starken Gesetz der großen Zahlen gilt $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n \mathbf{1}(X_k = z) = \mathbb{E}[\mathbf{1}(X_k = z)] = 1/q$ fast sicher für $z = 0, \dots, q-1$. Beachtet man noch, dass die nicht-eindeutigen Darstellungen (Perioden) mit $X_k = q-1$ für alle $k \geqslant k_0$ und ein $k_0 \in \mathbb{N}$ mit Wahrscheinlichkeit Null auftreten, so gilt also für $X \sim U([0,1])$, dass $X = \sum_{k\geqslant 1} X_k q^{-k}$ mit $X_k \in \{0, \dots, q-1\}$ und $P(\forall\, z = 0, \dots, q-1 :\ \frac{1}{n}\sum_{k=1}^n \mathbf{1}(X_k = z) \to 1/q) = 1$ (Schnitt über q Einsmengen ist Einsmenge).

Nun können wir aber auch noch den Schnitt über alle diese Ereignisse mit $q = 10^m$, $m \in \mathbb{N}$, nehmen, der gerade das Ereignis 'X ist normale Zahl' ergibt und als abzählbarer Schnitt wiederum eine Einsmenge ist. □

4.18 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.

4.19 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 $P(|X_{n_k} - X| > \varepsilon_k$ 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!

□

4.20 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, das 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).

4.21 Beispiele.

(a) Im schwachen Gesetz der großen Zahlen gilt sogar $A_n \xrightarrow{L^2} \mu$ wegen

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

(b) (Harmonische Reihe mit zufälligen Vorzeichen, Beispiel 2.41) 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 $\mathbb{E}[S_n] = 0$, $\operatorname{Var}(S_n) = \sum_{k=1}^n \frac{1}{k^2}$. Wegen $\operatorname{Var}(S_n - S_m) \leqslant \sum_{k \geqslant m} \frac{1}{k^2}$ 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 $A_n \leqslant C B_n$ für eine Konstante $C > 0$. Nach der Tschebyshev-Ungleichung gilt

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

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

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

wobei wir $\lfloor (n(m)+1)^\alpha \rfloor - \lfloor n(m)^\alpha \rfloor \lesssim n(m)^{\alpha-1}$ für $\alpha > 1$ verwendet haben. Wir schließen für $\alpha \in (1, 3/2)$

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

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

4.2 Konvergenz in Verteilung

4.22 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.

4.23 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 $X_n \xrightarrow{d} P :\iff P^{X_n} \xrightarrow{w} P$.

4.24 Beispiel. Ist 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 $\delta_{b_n}(\{b\}) = 0 \neq 1 = \delta_b(\{b\})$.

▷ **Kontrollfragen**

(a) Wie folgt aus $\mathbb{E}[X_n] \to 0$ und $\mathrm{Var}(X_n) \to 0$ für $X_n \in \mathscr{L}^2$, dass $X_n \xrightarrow{L^2} 0$?
Das folgt direkt mittels $\mathbb{E}[X_n^2] = \mathrm{Var}(X_n) + \mathbb{E}[X_n]^2 \to 0 + 0 = 0$.

(b) Finde ein Gegenbeispiel, das zeigt, dass $\mathbb{E}[|X|^p]^{1/p}$ für $p \in (0,1)$ keine Norm auf L^p ist.
Sind $X, Y \sim \mathrm{Bin}(1, 1/2)$ unabhängig, so gilt $\mathbb{E}[|X|^p]^{1/p} = \mathbb{E}[|Y|^p]^{1/p} = 2^{-1/p}$ für alle $p > 0$, aber $\mathbb{E}[|X+Y|^p]^{1/p} = (0{,}25{\bullet}0^p + 0{,}5{\bullet}1^p + 0{,}25{\bullet}2^p)^{1/p} = (2^{-1} + 2^{p-2})^{1/p} > 2^{(p-1)/p} = 2^{-1/p} + 2^{-1/p}$ für $p \in (0,1)$, so dass die Dreiecksungleichung verletzt ist.

(c) Welche der beiden Implikationen $X_n \xrightarrow{L^2} X \Rightarrow X_n \xrightarrow{L^1} X$ bzw. $X_n \xrightarrow{L^1} X \Rightarrow X_n \xrightarrow{L^2} X$ ist wahr und welche falsch?
Die erste ist wahr, wie mit der Cauchy-Schwarz- oder Hölder-Ungleichung folgt. Die zweite ist falsch, es muss nicht einmal $X_n - X \in \mathscr{L}^2$ gelten.

(d) Wie kann $X_n \xrightarrow{P} X$ und $X_n \xrightarrow{P\text{-f.s.}} X$ für diskrete Zufallsvariablen X_n, X mit Werten in $\mathbb{Z} \subseteq \mathbb{R}$ genauer charakterisiert werden?

$X_n \xrightarrow{P} X$ bedeutet dann gerade (wähle $\varepsilon \in (0,1)$ in der Definition) $P(X_n = X) \to 1$. Da $X_n(\omega) \to X(\omega)$ für ganze Zahlen gerade $X_n(\omega) = X(\omega)$ für alle $n \geqslant N(\omega)$ mit einem $N(\omega) \in \mathbb{N}$ bedeutet, ist fast sichere Konvergenz gleichbedeutend mit $P(\exists N \forall n \geqslant N : X_n = X) = 1$.

(e) Wogegen konvergieren $X_n \sim \mathrm{Exp}(\lambda_n)$ mit $\lambda_n \to \infty$ in Verteilung?

Es gilt $X_n \xrightarrow{d} 0$ oder Äquivalent $\mathrm{Exp}(\lambda_n) \xrightarrow{w} \delta_0$, weil für stetige und beschränkte Funktionen φ wegen dominierter Konvergenz gilt

$$\int_0^\infty \varphi(x)\lambda_n e^{-\lambda_n x} dx = \int_0^\infty \varphi(\lambda_n^{-1} y) e^{-y} dy \to \int_0^\infty \varphi(0) e^{-y} dy = \varphi(0).$$

Ende 15. Vorlesung

4.25 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 4.19(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. □

4.26 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.

4.27 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 (<u>Stetigkeitspunkte</u> von F^X).

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 $x_0 < \dots < x_K$ von F^X mit $F^X(x_0) < \delta$, $F^X(x_K) > 1-\delta$ und $\forall x \in [x_{k-1}, x_k] : |\varphi(x)-\varphi(x_k)| < \delta$, $k = 1, \dots, K$. 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 $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$. □

4.28 Beispiele.

(a) 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 $np_n \to \lambda > 0$.

(b) 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 $n\min(U_1, \dots, U_n) \xrightarrow{d} \mathrm{Exp}(1)$.

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

4.30 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 von F.*

Beweis. Wir verwenden ein Diagonalfolgenargument. sei dazu $(q_n)_{n\geqslant 1}$ eine Abzählung von $\mathbb{Q}$, d.h. $\mathbb{Q} = \{q_n \mid n \in \mathbb{N}\}$. Da $(F_n(q_1))_{n\geqslant 1}$ eine beschränkte Folge ist, existiert eine Teilfolge $(n_1(k))_{k\geqslant 1}$ 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 $(n_\ell(k))_{k\geqslant 1}$ 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 $\ell \geqslant 1$ 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. □

4.31 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.

4.32 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 $\sup_{n\geqslant 1} P_n([-K_\varepsilon, K_\varepsilon]^c) \leqslant \varepsilon$.

4.33 Beispiel. Es gelte $P_n \xrightarrow{w} P$ und 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 $n = 1, \dots, N-1$ (möglich wegen σ-Stetigkeit). Dann gilt $P_n([-y, y]^c) \leqslant \varepsilon$ für alle n. (P_n) ist also straff.

4.34 Korollar (Satz von Prokhorov)**.** *Ist (P_n) eine straffe Folge von Wahrscheinlichkeitsmaßen, 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 4.27. 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. □

4.35 Bemerkung. Nach Beispiel 4.33 und dem Satz von Prokhorov ist die Straffheit der Folge (P_n) äquivalent dazu, dass jede Teilfolge (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.

▷ **Kontrollfragen**

(a) Es seien X_n, X reellwertige Zufallsvariablen und $f : \mathbb{R} \to \mathbb{R}$ stetig. Wieso gelten dann die Folgerungen (continuous mapping theorem) (i) $X_n \xrightarrow{P\text{-f.s.}} X \Rightarrow f(X_n) \xrightarrow{P\text{-f.s.}} f(X)$, (ii) $X_n \xrightarrow{P} X \Rightarrow f(X_n) \xrightarrow{P} f(X)$, (iii) $X_n \xrightarrow{d} X \Rightarrow f(X_n) \xrightarrow{d} f(X)$?

Wegen Stetigkeit gilt die Inklusion $\{X_n \to X\} \subseteq \{f(X_n) \to f(X)\}$ und somit $P(X_n \to X) = 1 \Rightarrow P(f(X_n) \to f(X)) = 1$. Aus dieser fast sicheren Konvergenz folgt stochastische Konvergenz am besten über ein Teilteilfolgenargument. Für jede Teilfolge (n_k) existiert eine Teilteilfolge (n_{k_ℓ}) mit $X_{n_{k_\ell}} \to X$ P-fast sicher, also auch $f(X_{n_{k_\ell}}) \to f(X)$ P-fast sicher und somit stochastisch. Würde nicht $f(X_n) \xrightarrow{P} f(X)$ gelten, so gäbe es eine Teilfolge (n_k) und $\varepsilon > 0$ mit $\liminf_{k\to\infty} P(|f(X_{n_k}) - f(X)| > \varepsilon) > 0$, im Widerspruch zu $f(X_{n_{k_\ell}}) \xrightarrow{P} f(X)$.

Für Verteilungskonvergenz beachte in der Definition nur, dass mit φ auch die Komposition $\varphi \circ f$ stetig und beschränkt ist.

(b) Es sei Φ die $N(0,1)$-Verteilungsfunktion. Wieso folgt aus $X_n \xrightarrow{d} N(0,1)$, dass $P(X_n \in [a,b]) \to \Phi(b) - \Phi(a)$ für alle reellen Zahlen $a \leqslant b$?

Vergleiche Korollar 4.52 unten.

(c) Finde ein Beispiel mit $X_n \xrightarrow{d} X$, aber $F^{X_n}(x) \not\to F^X(x)$ für ein $x \in \mathbb{R}$.

Betrachte die deterministischen Zufallsvariablen $X_n = 1/n$ und $X = 0$. Dann gilt $X_n \xrightarrow{d} X$ aber $F_{X_n}(0) = 0 \neq 1 = F^X(0)$.

(d) Gib Beispiele von Folgen diskreter Verteilungen an, die straff bzw. nicht straff sind. Finde gegebenenfalls die Grenzverteilung.

Nicht straff sind $(\mathrm{Pois}(n))_{n\geqslant 1}$, $(\mathrm{Lap}(\{1,\dots,n\}))_{n\geqslant 1}$; straff sind $(\mathrm{Poiss}(1/n))_{n\geqslant 1}$ und $(\mathrm{Bin}(m,p_n))_{n\geqslant 1}$ für beliebige $m\in\mathbb{N}$, $p_n\in[0,1]$. Es gilt $\mathrm{Poiss}(1/n)\xrightarrow{w}\delta_0$, da die Zähldichten konvergieren. Ist p ein Häufungspunkt von (p_n), so existiert eine Teilfolge (n_k) mit $p_{n_k}\to p$ und dann auch $\mathrm{Bin}(m,p_{n_k})\xrightarrow{w}\mathrm{Bin}(m,p)$ wegen Konvergenz der Zähldichten.

(e) Zeige anhand eines Beispiels, dass H und F im Beweis des Auswahlsatzes von Helly nicht immer auf $\mathbb{Q}$ übereinstimmen.

Wie bei der Modellierung in (c) gilt für $F_n=\mathbf{1}_{[1/n,\infty)}$, dass $H(0)=0$, aber $F(0)=\inf_{q>0}H(q)=1$. Der punktweise Grenzwert rechtsstetiger Funktionen muss also nicht selbst rechtsstetig sein.

Ende 16. Vorlesung

4.3 Charakteristische Funktionen und Zentrale Grenzwertsätze

4.36 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.

4.37 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.

4.38 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)

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

4.39 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 4.38(e) und $|\Sigma^{1/2}u|^2 = \langle \Sigma u, u\rangle$. $\square$

4.40 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

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

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. $\square$

4.41 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. $\square$

4.42 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): $\varphi^{P_1 * P_2}(u) = \varphi^{P_1}(u)\varphi^{P_2}(u)$.

4.43 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 $\sup_{u\in\mathbb{R}} |r_m(u)| \leqslant 2\,\mathbb{E}[|X|^m]$ 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! $\square$

4.44 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 $\frac{d^2}{du^2}\varphi^{\mathrm{Bin}(n,p)}(u) = -n(n-1)(pe^{iu} + 1 - p)^{n-2} p^2 e^{2iu} - n(pe^{iu} + 1 - p)^{n-1} pe^{iu}$ 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.

4.45 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!), $\int_{-U}^{U} \frac{\cos(uy)}{u} du = 0$ 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. $\square$

4.46 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. $\square$

▷ **Kontrollfragen**

(a) Welche charakteristische Funktion besitzt die $\mathrm{Exp}(\lambda)$-Verteilung?
Es gilt $\varphi^{\mathrm{Exp}(\lambda)}(u) = (1 - i\lambda^{-1}u)^{-1}$ wegen

$$\int_0^\infty e^{iux}\lambda e^{-\lambda x}dx = \frac{-\lambda}{\lambda - iu}e^{-(\lambda - iu)x}|_{x=0}^\infty = \frac{\lambda}{\lambda - iu}.$$

(b) Wie folgt $\mathrm{Bin}(n,p) = \mathrm{Bin}(1,p)^{*n}$ und $\mathrm{Pois}(\lambda n) = \mathrm{Pois}(\lambda)^{*n}$ mittels charakteristischer Funktionen?
Es gilt $\varphi^{\mathrm{Bin}(1,p)}(u)^n = (1 - p + pe^{iu})^n = \varphi^{\mathrm{Bin}(n,p)}(u)$, also nach Lemma 4.41 $\varphi^{\mathrm{Bin}(1,p)^{*n}}(u) = \varphi^{\mathrm{Bin}(n,p)}(u)$ und der Eindeutigkeitssatz 4.45 zeigt $\mathrm{Bin}(n,p) = \mathrm{Bin}(1,p)^{*n}$. Für Poissonverteilungen folgt dies analog aus $\varphi^{\mathrm{Pois}(\lambda)}(u)^n = (e^{\lambda(e^{iu}-1)})^n = \varphi^{\mathrm{Pois}(\lambda n)}(u)$.

(c) Wie geht der formale Beweis, dass $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ für unabhängige $\mathbb{C}$-wertige Zufallsvariablen X, Y gilt mit $|X|, |Y| \in \mathscr{L}^1$?
Zerlege in Real- und Imaginärteil, beachte, dass mit X, Y auch die Zufallsvektoren $(\mathrm{Re}(X), \mathrm{Im}(X))$, $(\mathrm{Re}(Y), \mathrm{Im}(Y))$ unabhängig und koordinatenweise in $\mathscr{L}^1$ sind, und überprüfe mittels Satz 3.10(c):

$$\begin{aligned}\mathbb{E}[XY] &= \mathbb{E}[\mathrm{Re}(X)\,\mathrm{Re}(Y) - \mathrm{Im}(X)\,\mathrm{Im}(Y)] + i\,\mathbb{E}[\mathrm{Re}(Y)\,\mathrm{Im}(X) + \mathrm{Im}(X)\,\mathrm{Re}(Y)]\\ &= \mathbb{E}[\mathrm{Re}(X)]\,\mathbb{E}[\mathrm{Re}(Y)] - \mathbb{E}[\mathrm{Im}(X)]\,\mathbb{E}[\mathrm{Im}(Y)]\\ &\quad + i\big(\mathbb{E}[\mathrm{Re}(Y)]\,\mathbb{E}[\mathrm{Im}(X)] + \mathbb{E}[\mathrm{Im}(X)]\,\mathbb{E}[\mathrm{Re}(Y)]\big)\\ &= \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

(d) Wie erhält man Erwartungswert und Varianz einer Poissonverteilung aus der charakteristischen Funktion?
Berechne $\frac{d}{du}\varphi^{\mathrm{Pois}(\lambda)}(u) = i\lambda e^{iu}\varphi^{\mathrm{Pois}(\lambda)}(u)$, $\frac{d^2}{du^2}\varphi^{\mathrm{Pois}(\lambda)}(u) = -(\lambda e^{iu} + \lambda^2 e^{2iu})\varphi^{\mathrm{Pois}(\lambda)}(u)$, so dass $\frac{d}{du}\varphi^{\mathrm{Pois}(\lambda)}(0) = i\lambda$, $\frac{d^2}{du^2}\varphi^{\mathrm{Pois}(\lambda)}(0) = -(\lambda + \lambda^2)$ und somit Erwartungswert λ, zweites Moment $\lambda + \lambda^2$ und Varianz λ folgen.

(e) Wie lässt sich Korollar 3.39 zur Transformation von Normalverteilungen in einer Zeile mit charakteristischen Funktionen beweisen?
Nach den Rechenregeln gilt

$$\varphi^{AX+b}(u) = \varphi^X(A^\top u)e^{i\langle u,b\rangle} = \exp\Big(-\frac{\langle \Sigma A^\top u, A^\top u\rangle}{2} + i\big(\langle A^\top u, \mu\rangle + \langle u, b\rangle\big)\Big),$$

was gerade $\varphi^{N(A\mu+b, A\Sigma A^\top)}(u)$ ist und mit dem Eindeutigkeitssatz $AX + b \sim N(A\mu + b, A\Sigma A^\top)$ impliziert.

Ende 17. Vorlesung

4.47 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.

4.48 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 $P(\mathbb{R}^d \setminus [-\frac{2}{u}, \frac{2}{u}]^d) \leqslant \frac{2}{(2u)^d} \int_{[-u,u]^d} (1 - \varphi^P(v))dv$ 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

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

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

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

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

Nach dem Satz von Prokhorov (Korollar 4.34 bzw. Bemerkung 4.35 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 eine stetige beschränkte Funktion φ und eine Teilfolge (n_m) mit $\liminf_{m\to\infty} |\int \varphi\, dP_{n_m} - \int \varphi\, dP| > 0$. Wegen Straffheit gibt es eine Teilteilfolge (n_{m_k}) und ein Wahrscheinlichkeitsmaß Q mit $P_{n_{m_k}} \xrightarrow{w} Q$, so dass $\varphi^{P_{n_{m_k}}}(u) \to \varphi^Q(u)$. Also gilt $\varphi^Q = \psi = \varphi^P$ und nach dem Eindeutigkeitssatz $P = Q$, was $\liminf_{m\to\infty} |\int \varphi\, dP_{n_m} - \int \varphi\, dP| > 0$ zum Widerspruch führt. Es folgt $P_n \xrightarrow{w} P$. □

4.49 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 $u/\sigma_n = 0$

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

$\varphi^{S_n^*}(u)$ 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.

4.50 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 4.43 $\varphi^{\tilde{X}_1}(u/\sqrt{n}) = 1 - \frac{u^2}{2n}(1 + r_m(u/\sqrt{n}))$ mit $r_m(u/\sqrt{n}) \to 0$ 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

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

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

4.51 Bemerkung. 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.

4.52 Korollar. *Unter den Voraussetzungen von Satz 4.50 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 4.50 und der Charakterisierung der Verteilungskonvergenz gemäß Satz 4.27 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 $\{|\frac{1}{n}\sum_{i=1}^n X_i - \mu| > 1{,}96\frac{\sigma}{\sqrt{n}}\} = \{|S_n^*| \nleqslant 1{,}96\}$ folgt die letzte Aussage durch Komplementbildung sowie die numerische Näherung $\Phi(1{,}96) \approx 0{,}975$. □

4.53 Korollar. *Unter den Voraussetzungen von Satz 4.50, aber auch für $\sigma = 0$ gilt*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Beweis. Für $\sigma > 0$ folgt dies aus dem zentralen Grenzwertsatz; denn $S_n^* \xrightarrow{d} Z$ mit $Z \sim N(0,1)$ impliziert $\sigma S_n^* \xrightarrow{d} \sigma Z \sim N(0,\sigma^2)$ (mit φ ist auch $x \mapsto \varphi(\sigma x)$ stetig und beschränkt). Für $\sigma = 0$ gilt $P(X_i = \mu) = 1$ und somit $\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \sim \delta_0 = N(0,0)$. □

4.54 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. '$\Rightarrow$' folgt sofort aus der Stetigkeit von $x \mapsto \langle x, v\rangle$. '$\Leftarrow$': 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 $X_n \xrightarrow{d} X$. □

4.55 Satz. *(Zentraler Grenzwertsatz, Vektorversion) 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. Nach dem Satz von Cramér-Wold genügt es, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \mu, v\rangle \xrightarrow{d} \langle Z, v\rangle$ für $Z \sim N(0,\Sigma)$ und alle $v \in \mathbb{R}^d$ nachzuweisen. Zunächst beachte $\langle Z, v\rangle \sim N(0, \langle \Sigma v, v\rangle)$ gemäß Korollar 3.39. Nun sind aber $\tilde{X}_i = \langle X_i - \mu, v\rangle$ unabhängige reellwertige Zufallsvariablen mit $\mathbb{E}[\tilde{X}_i^2] \leqslant \mathbb{E}[|X_i - \mu|^2]|v|^2 \leqslant \mathbb{E}[|X_i|^2]|v|^2 < \infty$, $\mathbb{E}[\tilde{X}_i] = 0$ 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 Korollar 4.53 folgt daher $\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{X}_i \xrightarrow{d} N(0, \langle \Sigma v, v\rangle)$ und somit die Behauptung. □

▷ **Kontrollfragen**

(a) Im Beweis des Stetigkeitssatzes von Lévy wurde im Fall $d = 1$ folgende Aussage verwendet: wenn für alle $\varepsilon > 0$ Werte $K_\varepsilon > 0$ existieren mit der Eigenschaft $\limsup_{n\to\infty} P_n([-K_\varepsilon, K_\varepsilon]^c) \leqslant \varepsilon$, so ist (P_n) straff. Beweis?

Wähle N mit $P_n([-K_{\varepsilon/2}, K_{\varepsilon/2}]^c) \leqslant \varepsilon$ für alle $n \geqslant N$ und dann für $P_1, \dots, P_{N-1}$ jeweils $C_1, \dots, C_{N-1} > 0$ mit $P_i([-C_i, C_i]^c) \leqslant \varepsilon$ (möglich da jedes einzelne W-Maß straff ist, betrachte seine Verteilungsfunktion). Mit $\tilde K_\varepsilon = \max(K_{\varepsilon/2}, C_1, \dots, C_{N-1})$ folgt dann $\sup_{n\geqslant 1} P_n([-\tilde K_\varepsilon, \tilde K_\varepsilon]^c) \leqslant \varepsilon$, also Straffheit.

(b) Für unabhängige $U([-1,1])$-verteilte Zufallsvariablen X_j, $j \in \mathbb{N}$, gilt $\mathbb{E}[X_j] = 0$, $\operatorname{Var}(X_j) = 1/3$. Überprüfe, dass die *standardisierte Summe* $S_n^* = \sqrt{3/n}\sum_{j=1}^n X_j$ die charakteristische Funktion

$$\varphi^{S_n^*}(u) = \Big(\frac{\sin(\sqrt{3}u/\sqrt{n})}{\sqrt{3}u/\sqrt{n}}\Big)^n = \big(1 - u^2/(2n) + o(1/n)\big)^n$$

besitzt und wiederum $S_n^* \xrightarrow{d} N(0,1)$ für $n \to \infty$ folgt. Suche im Internet entsprechende Graphiken zur Konvergenz.

Es gilt $\varphi^{X_j}(u) = \frac{1}{2}\int_{-1}^1 e^{iux}dx = \frac{\sin(u)}{u}$ (mit stetiger Ergänzung bei $u = 0$), und die Formel für $\varphi^{S_n^*}$ folgt direkt mit den Rechenregeln für Faltung und Multiplikation. Wegen der Potenzreihe $\frac{\sin(u)}{u} = \sum_{k\geqslant 0} \frac{(-1)^k}{(2k+1)!}u^{2k}$ folgt die Ordnung $(1 - u^2/(2n) + o(1/n))^n$. Die Konvergenz gegen $e^{-u^2/2}$ ergibt sich aus $n\log(1 - u^2/2n + o(1/n)) \to -u^2/2$, vergleiche den Beweis des ZGWS. Die Konvergenz ist (visuell) extrem schnell, siehe z.B. Link.

(c) In Berlin gab es bis zum 18.6.20 3732 männliche und 3862 weibliche Corona-Infizierte. Unter der Annahme, dass die Ansteckungswahrscheinlichkeit für beide Geschlechter gleich ist, wie groß ist die Wahrscheinlichkeit für eine solche Abweichung um 65 vom Mittelwert oder eine noch größere Abweichung? (verwende numerische Werte von Φ)

Unter der Annahme, dass jede Ansteckung unabhängig erfolgt, können wir die Anzahl X männlicher Infizierter als $\operatorname{Bin}(n, 1/2)$-verteilt modellieren mit $n = 3732 + 3862 = 7594$. Dann ist $\mathbb{E}[X] = 3797$, $\sigma(X) = \sqrt{7594/4} \approx 43,57$. Wir verwenden mit $Z \sim N(0,1)$ die Normalapproximation der Binomialverteilung:

$$P(|X - \mathbb{E}[X]| \geqslant 65) \approx P(|Z| \geqslant \tfrac{65}{\sigma(X)}) \approx (1 - \Phi(1,49)) + \Phi(-1,49) \approx 0,134.$$

In der Statistik werden Abweichungen, die wie hier in über 10% der Fälle durch reinen Zufall entstehen können, gewöhnlicherweise nicht als signifikant eingestuft.

(d) Wie oft muss eine Münze mindestens geworfen werden, damit die Wahrscheinlichkeit, dass die relative Häufigkeit $\hat p$ von 'Kopf' maximal um $0,001$ vom Erwartungswert $p \in [0,1]$ abweicht, ca. 95% beträgt? Gibt es eine Schranke unabhängig vom unbekannten p? (verwende Approximation durch die Normalverteilung)

Bezeichnet X die Anzahl von 'Kopf' bei n Würfen, so modelliere $X \sim \operatorname{Bin}(n,p)$. Dann gilt für $\hat p = X/n$ mit Normalapproximation gemäß Satz von de Moivre-Laplace bzw. ZGWS:

$$P(|\hat{p}-p|\leqslant 0{,}001)=P\Big(\frac{|X-np|}{\sqrt{np(1-p)}}\leqslant\frac{0{,}001\sqrt{n}}{\sqrt{p(1-p)}}\Big)$$
$$\approx\Phi\Big(\frac{0{,}001\sqrt{n}}{\sqrt{p(1-p)}}\Big)-\Phi\Big(-\frac{0{,}001\sqrt{n}}{\sqrt{p(1-p)}}\Big).$$

Gemäß Korollar 4.52 erhalten wir den Wert $0{,}05$, falls $\frac{0{,}001\sqrt{n}}{\sqrt{p(1-p)}}\approx 1{,}96$, also $n\approx 1960^2p(1-p)$. Wegen $p(1-p)\leqslant 1/4$ reicht in jedem Fall $n\approx 960400$. Die Konvergenzrate im ZGWS beträgt $\frac{1}{\sqrt{n}}$, so dass die erforderliche Stichprobengröße quadratisch mit der Inversen der geforderten Präzision wächst.

(e) Satz von Cramér-Wold: Gibt es Beispiele von Zufallsvektoren X_n, X im $\mathbb{R}^d$, wo jede Koordinate $X_{n,i}\xrightarrow{d} X_i$, $i=1,\dots,d$, erfüllt, aber trotzdem nicht $X_n\xrightarrow{d} X$ gilt?

Ja, natürlich. Die gemeinsame Verteilung der Koordinaten X_i von X ist ja nicht festgelegt. Betrachte beispielsweise $X\sim N(0,\Sigma)$ und $X_n\sim N(0,\operatorname{diag}(\Sigma_{11},\dots,\Sigma_{dd}))$, wobei Σ eine nicht-diagonale Kovarianzmatrix ist, dann ist $X_{n,i}\sim X_i\sim N(0,\Sigma_{ii})$, aber X_n und X sind nicht identisch verteilt.

Ende 18. Vorlesung

4.56 Definition. Eine Familie $(X_j^{(n)})_{1\leqslant j\leqslant n, n\geqslant 1}$ von Zufallsvariablen in $\mathscr{L}^2$ bildet ein standardisiertes Dreiecksschema, falls für jedes $n\in\mathbb{N}$ die Zufallsvariablen $X_1^{(n)},\dots,X_n^{(n)}$ („in einer Zeile") unabhängig sind sowie $\mathbb{E}[X_j^{(n)}]=0$, $j=1,\dots,n$, und $\sum_{j=1}^n \operatorname{Var}(X_j^{(n)})=1$ gilt. $(X_j^{(n)})_{1\leqslant j\leqslant n,n\geqslant 1}$ genügt der Lindeberg-Bedingung, falls gilt

$$\forall\delta>0:\ \lim_{n\to\infty}\sum_{j=1}^n \mathbb{E}\left[(X_j^{(n)})^2\mathbf{1}(|X_j^{(n)}|\geqslant\delta)\right]=0.$$

4.57 Bemerkung. Man stellt sich ein Dreiecksschema graphisch vor:

$$\begin{array}{llll} X_1^{(1)} & & & \\ X_1^{(2)} & X_2^{(2)} & & \\ & \vdots & & \\ X_1^{(n)} & X_2^{(n)} & \cdots & X_n^{(n)} \end{array}$$

Bei der Standardversion des zentralen Grenzwertsatzes (ZGWS) haben die Summanden für jedes n dieselbe Verteilung und wir können $X_j^{(n)}=\frac{X_j-\mu}{\sigma\sqrt{n}}$ setzen. Oft treten jedoch unterschiedlich Verteilungen innerhalb der standardisierten Summe oder auch für verschiedene n auf. In Anwendungen denke man an einen Messfehler, der durch Überlagerung vieler, meist nicht identisch verteilter Fehler innerhalb der Messapparatur entsteht. Auch in diesem Fall können wir einen ZGWS formulieren und somit ein noch allgemeineres Universalitätsprinzip erreichen, allerdings müssen wir sicherstellen, dass nicht einzelne Summanden dominieren. Genau dies stellt die Lindeberg-Bedingung sicher, vergleiche auch die folgenden Beispiele. Letzlich wird der Beweis dieses verallgemeinerten ZGWS auf einer genaueren Abschätzung der Taylorentwicklung der charakteristischen Funktion beruhen.

4.58 Beispiele.

(a) Ist $(X_j)_{j\geqslant 1}$ eine Folge von i.i.d. Zufallsvariablen in $\mathscr{L}^2$ mit $\mu = \mathbb{E}[X_j]$, $\sigma^2 = \operatorname{Var}(X_j) > 0$, so bildet $X_j^{(n)} = (X_j - \mu)/(\sigma\sqrt{n})$ ein standardisiertes Dreiecksschema, das der Lindeberg-Bedingung genügt (wegen identischer Verteilung sind die Summanden alle gleich!):

$$\sum_{j=1}^n \mathbb{E}\left[(X_j^{(n)})^2 \mathbf{1}(|X_j^{(n)}| \geqslant \delta)\right] = \mathbb{E}\left[\frac{(X_1-\mu)^2}{\sigma^2}\mathbf{1}(|X_1-\mu| \geqslant \delta\sigma\sqrt{n})\right] \to 0$$

für $n \to \infty$ folgt mit dominierter Konvergenz (Majorante $\sigma^{-2}(X_1-\mu)^2$).

(b) Ist $(X_j^{(n)})_{1\leqslant j\leqslant n, n\geqslant 1}$ ein standardisiertes Dreiecksschema und gilt für ein $p > 2$ die Lyapunov-Bedingung

$$\lim_{n\to\infty} \sum_{j=1}^n \mathbb{E}\left[|X_j^{(n)}|^p\right] = 0,$$

so genügt $(X_j^{(n)})_{1\leqslant j\leqslant n, n\geqslant 1}$ auch der Lindeberg-Bedingung. Dies folgt aus

$$\sum_{j=1}^n \mathbb{E}\left[(X_j^{(n)})^2 \mathbf{1}(|X_j^{(n)}| \geqslant \delta)\right] \leqslant \sum_{j=1}^n \mathbb{E}\left[(X_j^{(n)})^2 (|X_j^{(n)}|/\delta)^{p-2}\right] \to 0,$$

wobei die rechte Seite gerade nach der Lyapunov-Bedingung für jedes $\delta > 0$ gegen Null konvergiert.

Ist insbesondere $(Y_j)_{j\geqslant 1}$ eine Folge unabhängiger und standardisierter (d.h. $\mathbb{E}[Y_j] = 0$, $\operatorname{Var}(Y_j) = 1$) Zufallsvariablen, die in L^p beschränkt ist (d.h. $\sup_{j\geqslant 1} \mathbb{E}[|Y_j|^p] < \infty$) für ein $p > 2$, so ist $X_j^{(n)} = Y_j/\sqrt{n}$ ein standardisiertes Dreiecksschema, das der Lyapunov- und somit auch der Lindeberg-Bedingung genügt:

$$\sum_{j=1}^n \mathbb{E}\left[|X_j^{(n)}|^p\right] = n^{-p/2}\sum_{j=1}^n \mathbb{E}[|Y_j|^p] \leqslant n^{1-p/2} \sup_{j\geqslant 1} \mathbb{E}[|Y_j|^p] \xrightarrow{n\to\infty} 0.$$

Dieser Fall deckt sehr viele Anwendungen mit unabhängigen, aber nicht identisch verteilten Summanden ab.

(c) In der harmonischen Reihe mit zufälligen Vorzeichen $S_n = \sum_{j=1} \varepsilon_j j^{-1}$ mit $(\varepsilon_j)_{j\geqslant 1}$ unabhängig, $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = 1/2$, bilden $X_j^{(n)} = \varepsilon_j/(j\sigma_n)$ mit $\sigma_n^2 = \sum_{j=1}^n j^{-2} \to \sigma_\infty^2 = \pi^2/6$ ein standardisiertes Dreiecksschema, das der Lyapunov-Bedingung *nicht* genügt; denn für $\delta \in (0, \sigma_\infty^{-1})$ erhält man

$$\sum_{j=1}^n \mathbb{E}[(X_j^{(n)})^2 \mathbf{1}(|X_j^{(n)}| \geqslant \delta)] = \sum_{j=1}^n j^{-2}\sigma_n^{-2}\mathbf{1}(j\sigma_n\delta \leqslant 1) \geqslant \sigma_n^{-2} \geqslant \sigma_\infty^{-2}.$$

In der Tat konvergiert ja S_n bereits in Verteilung (wegen stochastischer Konvergenz, siehe oben), was wir nun auch über charakteristische Funktionen nachweisen können:

$$\varphi^{S_n}(u) = \prod_{j=1}^{n} \varphi^{\varepsilon_j j^{-1}}(u) = \prod_{j=1}^{n} \cos(uj^{-1}) \to \prod_{j=1}^{\infty} \cos(j^{-1}u).$$

Das unendliche Produkt konvergiert, da $1-\cos(j^{-1}u) = \frac{j^{-2}u^2}{2} + O(j^{-4}u^4)$ (für $j \to \infty$) in j summierbar ist. Es ist auch bei $u = 0$ stetig, so dass der Stetigkeitssatz von Lévy anwendbar ist. Über die charakteristische Funktion lassen sich die Momente von S_∞ bestimmen. Man kann auch zeige, dass S_∞ eine Dichte besitzt und interessante Eigenschaften schlussfolgern, siehe den Artikel von Byron Schmuland. Es liegt definitiv keine Normalverteilung vor.

4.59 Lemma. *Bei einem standardisierten Dreiecksschema $(X_j^{(n)})_{1\leqslant j\leqslant n, n\geqslant 1}$ folgt aus der Lindeberg-Bedingung die Feller-Bedingung*

$$. \lim_{n\to\infty} \max_{1\leqslant j\leqslant n} \mathbb{E}[(X_j(n))^2] = 0.$$

Beweis. Setze $\sigma_{j,n}^2 = \mathbb{E}[(X_j^{(n)})^2]$. Aus

$$\sigma_{j,n}^2 = \mathbb{E}[(X_j^{(n)})^2(\mathbf{1}(|X_j^{(n)}| \leqslant \delta) + \mathbf{1}(|X_j^{(n)}| > \delta))] \leqslant \delta^2 + \mathbb{E}[(X_j^{(n)})^2\mathbf{1}(|X_j^{(n)}| > \delta)]$$

für $\delta > 0$ folgt mit der Lindeberg-Bedingung und einer Abschätzung des Maximums durch die Summe

$$\limsup_{n\to\infty} \max_{1\leqslant j\leqslant n} \sigma_{j,n}^2 \leqslant \delta^2 + \lim_{n\to\infty} \sum_{j=1}^{n} \mathbb{E}[(X_j^{(n)})^2\mathbf{1}(|X_j^{(n)}| > \delta)] = \delta^2.$$

Da $\delta > 0$ beliebig war, folgt $\max_j \sigma_{j,n}^2 \to 0$. □

4.60 Satz (Zentraler Grenzwertsatz nach Lindeberg, 1922)**.** *Ist $(X_j^{(n)})_{1\leqslant j\leqslant n, n\geqslant 1}$ ein standardisiertes Dreiecksschema, das der Lindeberg-Bedingung genügt, so gilt für (die Zeilensummen) $S_n^* = \sum_{j=1}^n X_j^{(n)}$ und $n \to \infty$*

$$S_n^* \xrightarrow{d} N(0,1).$$

4.61 Bemerkung. Man kann umgekehrt zeigen, dass ein standardisiertes Dreiecksschema mit $S_n^* \xrightarrow{d} N(0,1)$ die Lindeberg-Bedingung erfüllt, sofern es der Feller-Bedingung genügt, siehe Satz 28.3 in Bauer. Die Feller-Bedingung schließt einfache Fälle aus wie $X_1^{(n)} \sim N(0,1)$, $X_j^{(n)} = 0$ für $j \geqslant 2$, wo trotzdem $S_n^* \xrightarrow{d} N(0,1)$ (hier sogar mit Gleichheit) gilt.

Beweis. Nach Lemma 4.43 gilt $\varphi^X(u) = 1 - \frac{u^2}{2}(\sigma^2 + r(u))$ für $X \in \mathscr{L}^2$, wobei wir den Restterm anders abschätzen: wegen $|e^{iy} - 1| \leqslant 2$ und gleichmäßiger

Stetigkeit von $y \mapsto e^{iy}$ existiert für jedes $\varepsilon > 0$ und $u \in \mathbb{R}$ ein $\delta > 0$, so dass

$$\begin{aligned}|r(u)| &\leqslant \sup_{0\leqslant h\leqslant 1} |(\varphi^X)''(hu) - (\varphi^X)''(0)| = \sup_{0\leqslant h\leqslant 1} \mathbb{E}[X^2|e^{iuhX} - 1|] \\ &\leqslant \mathbb{E}\Big[X^2\Big(\varepsilon\mathbf{1}(|X| \leqslant \delta) + 2\mathbf{1}(|X| > \delta)\Big)\Big] \leqslant \varepsilon\,\mathbb{E}[X^2] + 2\,\mathbb{E}[X^2\mathbf{1}(|X| > \delta)].\end{aligned}$$

Für beliebige $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{C}$ mit $|a_i| \leqslant 1$, $|b_i| \leqslant 1$ nutzen wir die Ungleichung (Teleskopsumme!)

$$\begin{aligned}\Big|\prod_{j=1}^n a_j - \prod_{j=1}^n b_j\Big| &= \Big|(a_1 - b_1)\prod_{j=2}^n a_j + b_1(a_2 - b_2)\prod_{j=3}^n a_j + \dots + (b_1 \cdots b_{n-1})(a_n - b_n)\Big| \\ &\leqslant \sum_{j=1}^n |a_j - b_j|.\end{aligned}$$

Setze $\sigma_{j,n}^2 := \mathbb{E}[(X_j^{(n)})^2]$ sowie $r_{j,n}$, $\tilde{r}_{j,n}$ für den obigen Restterm r bei $X = X_j^{(n)}$ bzw. $X \sim N(0, \sigma_{j,n}^2)$. Nach der Faltungsformel für charakteristische Funktionen gilt $\varphi^{N(0,1)}(u) = \prod_{j=1}^n \varphi^{N(0,\sigma_{j,n}^2)}(u)$ und wir erhalten mit den Rechenregeln sowie der gerade gezeigten Ungleichung

$$\begin{aligned}|\varphi^{S_n^*}(u) - \varphi^{N(0,1)}(u)| &= \Big|\prod_{j=1}^n \varphi^{X_j^{(n)}}(u) - \prod_{j=1}^n \varphi^{N(0,\sigma_{j,n}^2)}(u)\Big| \\ &\leqslant \sum_{j=1}^n |\varphi^{X_j^{(n)}}(u) - \varphi^{N(0,\sigma_{j,n}^2)}(u)| \\ &\leqslant \sum_{j=1}^n \frac{u^2}{2}|r_{j,n}(u) - \tilde{r}_{j,n}(u)| \\ &\leqslant \sum_{j=1}^n u^2\Big(\varepsilon\sigma_{j,n}^2 + \mathbb{E}[(X_{j,n})^2\mathbf{1}(|X_j^{(n)}| > \delta)] + \mathbb{E}[\sigma_{j,n}^2 Z^2\mathbf{1}(|\sigma_{j,n}Z| > \delta)]\Big),\end{aligned}$$

wobei $Z \sim N(0,1)$, also $\sigma_{j,n}Z \sim N(0, \sigma_{j,n}^2)$ gilt. Wegen der Standardisierung $\sum_{j=1}^n \sigma_{j,n}^2 = 1$ ist die Summe über den ersten Klammerterm εu^2. Die Lindeberg-Bedingung zeigt, dass die Summe über den zweiten Term gegen Null konvergiert. Mit $\mathbf{1}(|Z| > t) \leqslant \frac{Z^2}{t^2}$, $t > 0$ (vergleiche auch Beispiel 4.58(b)) ist die Summe über den dritten Klammerterm kleiner oder gleich $u^2 \sum_{j=1}^n \sigma_{j,n}^4 \delta^{-2}\,\mathbb{E}[Z^4]$. Wegen der Standardisierung gilt $\sum_{j=1}^n \sigma_{j,n}^4 \leqslant \max_j \sigma_{j,n}^2 \to 0$ gemäß Lemma 4.59. Insgesamt schließen wir $\varphi^{S_n^*}(u) \to \varphi^{N(0,1)}(u)$ (wähle $\varepsilon \downarrow 0$), so dass $S_n^* \xrightarrow{d} N(0,1)$. □

▷ **Kontrollfragen**

(a) Es X eine Cauchy-verteilte Zufallsvariable, das heißt $f^X(x) = \pi^{-1}\frac{1}{1+x^2}$, $x \in \mathbb{R}$. Wie kann man mit dem Residuensatz $\varphi^X(u) = e^{-|u|}$ einsehen?
Nach Definition gilt $\varphi^X(u) = \pi^{-1}\int_{-\infty}^{\infty} g_u(x)\,dx$ mit $g_u(x) = (1+x^2)^{-1}e^{iux}$. g_u lässt sich zu einer meromorphen Funktion auf den komplexen Zahlen fortsetzen mit Polen bei $\pm i$. Für $z \in \mathbb{C}$ gilt $|g_u(z)| = |1+z^2|^{-1}e^{u\,\mathrm{Im}(z)}$. Betrachte

$u > 0$ und als Integrationsweg den unteren Halbkreis vom Radius $R > 0$, also den Weg γ_R von $-R$ nach $+R$ und dann entlang $Re^{i\varphi}$ von $\varphi = 2\pi$ nach $\varphi = \pi$. Dies umschließt den Pol bei $z = -i$ und nach dem Residuensatz ist

$$\int_{\gamma_R} g_u(z)dz = 2\pi i \mathrm{Res}_{z=-i}(g_u(z)) = 2\pi i(2i)^{-1}e^{-u} = \pi e^{-u}.$$

Für $R \to \infty$ und $\varphi \in (\pi, 2\pi)$ konvergiert $g_u(Re^{i\varphi}) \to 0$ wegen $\mathrm{Im}(e^{i\varphi}) < 0$ und man kann überprüfen, dass auch das gesamte Integral über den unteren Kreisrand gegen Null konvergiert. Es folgt $\int_{-\infty}^{\infty} g_u(x)dx = \lim_{R\to\infty} \int_{\gamma_R} g_u(z)dz = \pi e^{-u}$. Das zeigt $\varphi^X(u) = e^{-u}$ für $u > 0$. Für $u = 0$ gilt stets $\varphi^X(u) = 1$ und für $u < 0$ zeigt eine analoge Integration über den oberen Halbkreis $\int_{-\gamma_R} g_u(z)dz = \pi e^u$ und dasselbe Argument $\varphi^X(u) = e^u$. Alternativ sieht man sofort ein, dass φ^X symmetrisch um Null sein muss, da die Cauchydichte symmetrisch ist (substituiere $x \mapsto -x$ im Integral).

(b) Wie kann man mit (a) folgern, dass für unabhängige Cauchy-verteilte Zufallsvariablen $X_1, \dots, X_n$ der Mittelwert $\frac{1}{n}\sum_{i=1}^n X_i$ wieder Cauchy-verteilt ist?

Setze $A_n = \frac{1}{n}\sum_{i=1}^n X_i$. Dann ist $\varphi^{A_n}(u) = \prod_{i=1}^n \varphi^{X_i}(u/n) = (e^{-|u|/n})^n = e^{-|u|}$, so dass nach dem Eindeutigkeitssatz A_n Cauchy-verteilt ist.

(c) Wieso gilt in (b) weder der zentrale Grenzwertsatz noch ein Gesetz der großen Zahlen?

Für den ZGWS wird $X_i \in \mathscr{L}^2$ vorausgesetzt, was hier nicht gilt. In der Tat ist $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$, was in etwa der standardisierten Summe beim ZGWS entspricht, für $n \geqslant 1$ nicht einmal straff, da es wie $\sqrt{n}$ mal einer Cauchy-verteilten Zufallsvariable wächst.

Für das Gesetz der großen Zahlen wird die Existenz des Erwartungswerts (und bei uns auch $X_i \in \mathscr{L}^2$) gefordert, aber $X_i \notin \mathscr{L}^1$. Würde A_n stochastisch gegen einen deterministischen Wert konvergieren, so auch in Verteilung, was hier ebenfalls nicht zutrifft (die Grenzverteilung ist ja Cauchy).

(d) Was ist ein Beispiel für ein Dreiecksschema, wo die Feller-Bedingung, aber nicht die Lindeberg-Bedingung gilt?

Betrachte unabhängige $X_j^{(n)}$ mit $P(X_j^{(n)} = 1) = P(X_j^{(n)} = -1) = 1/n$ und $P(X_j^{(n)} = 0) = 1 - 2/n$. Dann ist $\mathbb{E}[X_j^{(n)}] = 0$, $\mathrm{Var}(X_j^{(n)}) = 1/n$, und es liegt ein standardisiertes Dreiecksschema vor, das die Feller-Bedingung erfüllt. Für jedes $\delta \in (0,1)$ ist jedoch die Lindeberg-Bedingung nicht erfüllt, da $(X_j^{(n)})^2 \mathbf{1}(|X_j^{(n)}| \geqslant \delta) = (X_j^{(n)})^2$ gilt. In der Tat gilt $P(S_n^* = 0) = (1 - 2/n)^n \to e^{-2} > 0$, S_n^* kann also nicht asymptotisch normalverteilt sein (es entsteht eine *zusammengesetzte Poissonverteilung* im Grenzwert).

(e) Wo geht im Beweis des ZGWS (Standardversion) die Unabhängigkeit ein? Kann Unabhängigkeit durch Unkorreliertheit ersetzt werden?

Die Unabhängigkeit wird benötigt, um die charakteristische Funktion der Summe als Produkt der charakteristischen Funktionen zu schreiben. Unkorreliertheit reicht hier nicht. Ein Gegenbeispiel sind die unkorrellierten Zufallsvariablen $X_i = \varepsilon_i Z$ mit $Z \sim N(0,1)$ und einer i.i.d.-Folge (ε_i) mit $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$. Die standardisierte Summe ist dann $S_n^* = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i Z$. Mit dem ZGWS für die Summe über ε_i und mit direkter Rechnung folgt $S_n^* \xrightarrow{d} YZ$ mit $Y, Z \sim N(0,1)$ unabhängig. Dieses Produkt YZ ist nicht mehr normalverteilt (z.B. $\mathbb{E}[(YZ)^4] = \mathbb{E}[Y^4]\,\mathbb{E}[Z^4] = 9$ ist nicht 3).

Ende 19. Vorlesung

4.4 Asymptotik der empirischen Verteilung

4.62 Definition. Es seien $X_1, \dots, X_n$ 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 $F^X := F^{X_i}$, $P^X := P^{X_i}$.

4.63 Satz. *Für alle $x \in \mathbb{R}$, $x_1 \leqslant \ldots \leqslant x_m \in \mathbb{R}$ 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 $\operatorname{Var}(Z_i(x)) = F^X(x) - F^X(x)^2$ folgt mit dem starken Gesetz der großen Zahlen Teil (a) und mit dem zentralen Grenzwertsatz nach Korollar 4.53 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

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

gilt. Nach dem mehrdimensionalen ZGWS aus Satz 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. □

4.64 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 $\operatorname{Cov}(G^X(x), G^X(y)) = F^X(x \wedge y) - F^X(x)F^X(y)$ konvergiert in dem Sinne, dass endlich-dimensionale Randverteilungen an beliebigen Stellen $x_1, \dots, x_m$ 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.

4.65 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 4.63(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. □

4.66 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 4.27 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

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 $\mathcal{X} = \{0, \ldots, n\}$, $\mathscr{F} = \mathscr{P}(\mathcal{X})$ 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 meldet für das 4. Quartal 2019 in Berlin $n = 10099$ Lebendgeborene, darunter $x = 4956$ Mädchen. Also gilt

$$\varphi(x) = \mathbf{1}(|4956 - 5049{,}5| > 0{,}98\sqrt{10099}) = \mathbf{1}(93{,}5 > 98{,}48) = 0.$$

Zum Niveau 5% können wir die Nullhypothese gleichwahrscheinlicher Mädchen- und Jungengeburten nicht ablehnen. Betrachten wir aber den letzten bekannten Jahreswert 2018, so gab es $n = 40203$ Lebendgeborene und darunter $x = 19591$ Mädchen. Hier gilt also

$$\varphi(x) = \mathbf{1}(|19591 - 20101{,}5| > 0{,}98\sqrt{40203}) = \mathbf{1}(410{,}5 > 196{,}50) = 1.$$

Hier wird die Nullhypothese zum Niveau $\alpha = 5\%$ abgelehnt. In beiden Fällen sind es etwas weniger als 49% 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.

▷ **Kontrollfragen**

(a) Modellieren Sie die weiteren Testprobleme aus Bemerkung 5.2 rigoros.

Das Testen, ob eine Münze fair ist, führt auf dasselbe mathematische Testproblem wie bei den Mädchen- und Jungengeburten. Wir schlagen eine Modellierung vor zum Testen, ob Kinder genauso oder weniger infektiös als Erwachsene sind. Dabei ist natürlich die Datenlage von entscheidender Bedeutung. Idealerweise kennen wir von m kranken Kindern und n kranken Erwachsenen die Anzahl $K_1, \dots, K_m$ bzw. $E_1, \dots, E_n$ der jeweils angesteckten Personen. Weil von den Ansteckungsgefährdeten sich jeweils nur ein kleiner Anteil wirklich ansteckt, modellieren wir gemäß Poissonschem Grenzwertsatz $K_1, \dots, K_m \sim \mathrm{Poiss}(\lambda_K)$, $E_1, \dots, E_n \sim \mathrm{Poiss}(\lambda_E)$. Außerdem nehmen wir an, dass $K_1, \dots, K_m, E_1, \dots, E_n$ unabhängig sind. Dann können wir das statistische Modell schreiben als

$$\Big(\mathbb{N}_0^{m+n}, \mathcal{P}(\mathbb{N}_0^{m+n}), \big(\mathrm{Poiss}(\lambda_K)^{\otimes m} \otimes \mathrm{Poiss}(\lambda_E)^{\otimes n}\big)_{\lambda_K, \lambda_E > 0}\Big).$$

Aus einer Risikoabschätzung bezüglich Regelbetrieb in Schulen wird als Nullhypothese $H_0 : \lambda_K \geqslant \lambda_E$, also $\Theta_0 = \{(x_K, x_E) \in \mathbb{R}_+^2 \mid x_K \geqslant x_E\}$ gewählt. Die Alternative ist dann das Komplement $H_1 : \lambda_K < \lambda_E$, $\Theta_1 = \mathbb{R}_+^2 \setminus \Theta_0$. Die Wahrscheinlichkeit für den Fehler erster Art, Kinder als weniger ansteckend anzusehen, obwohl sie mindestens so ansteckend wie Erwachsene sind, wird dann durch das Niveau α eines Tests beschränkt. Bei diesem sogenannten Zweistichprobentest kann man zeigen, dass Tests der Form $\varphi = \mathbf{1}(\frac{1}{m}\sum_{i=1}^m K_i \leqslant \kappa \frac{1}{n}\sum_{i=1}^n E_i)$ für geeignete $\kappa > 0$ nicht nur intuitiv, sondern auch in gewissem Sinn optimal sind.

Ende 20. Vorlesung

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 <u>Fehler 2. Art</u> 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:

$$\begin{aligned}
P_{\vartheta_1}(\varphi^* = 1) - P_{\vartheta_1}(\varphi = 1) &= P_{\vartheta_1}(\varphi^* = 1, \varphi = 0) - P_{\vartheta_1}(\varphi^* = 0, \varphi = 1)\\
&= \int_{\mathfrak{X}} \big(\mathbf{1}(\varphi^* = 1,\, \varphi = 0) - \mathbf{1}(\varphi^* = 0, \varphi = 1)\big)\, L(\vartheta_1)\, d\mu\\
&\geqslant \int_{\mathfrak{X}} \mathbf{1}(\varphi^* = 1,\, \varphi = 0)\, \kappa L(\vartheta_0)\, d\mu - \int_{\mathfrak{X}} \mathbf{1}(\varphi^* = 0,\, \varphi = 1)\, \kappa L(\vartheta_0)\, d\mu\\
&= \kappa\big(P_{\vartheta_0}(\varphi^* = 1, \varphi = 0) - P_{\vartheta_0}(\varphi^* = 0, \varphi = 1)\big)\\
&= \kappa\big(P_{\vartheta_0}(\varphi^* = 1) - P_{\vartheta_0}(\varphi = 1)\big) \geqslant 0,
\end{aligned}$$

wobei zuletzt die Definition von α und das Niveau α von φ eingegangen sind. □

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.

▷ **Kontrollfragen**

(a) Geben Sie alle Neyman-Pearson-Tests an im diskreten Modell mit $\mathcal{X} = \{1, 2, 3\}$ und folgenden Zähldichten:

$p_\vartheta(x)$	x=1	x=2	x=3
$\vartheta = \vartheta_0$	0,2	0,5	0,3
$\vartheta = \vartheta_1$	0,1	0,4	0,5

Welche Fehlerwahrscheinlichkeiten erster Art ergeben sich?

Betrachte Neyman-Pearson-Tests $\varphi_\kappa(x) = \mathbf{1}(\frac{p_{\vartheta_1}(x)}{p_{\vartheta_0}(x)} > \kappa)$ (der Nenner ist nie null). Für $\kappa \in [0, 1/2)$ gilt $\varphi_\kappa(x) = 1$ (verwerfe H_0 immer), für $\kappa \in [1/2, 4/5)$ gilt $\varphi_\kappa(x) = \mathbf{1}(x \in \{2, 3\})$, für $\kappa \in [4/5, 5/3)$ gilt $\varphi_\kappa(x) = \mathbf{1}(x = 3)$ und für $\kappa \geqslant 5/3$ gilt $\varphi_\kappa(x) = 0$ (akzeptiere H_0 immer). Das Niveau dieser Tests ist (in der Reihenfolge) $\alpha = 1$, $\alpha = 0,5 + 0,3 = 0,8$, $\alpha = 0,3$, $\alpha = 0$.

(b) Bestimmen Sie die Neyman-Pearson-Tests für das Normalverteilungsmodell $(N(\vartheta, 1)^{\otimes n})_{\vartheta\in\mathbb{R}}$, wo n unabhängige $N(\vartheta, 1)$-verteilte Beobachtungen vorliegen.

Die Likelihood-Funktion (bezüglich Lebesguemaß im $\mathbb{R}^n$) ist

$$L(\vartheta) = (2\pi)^{-n/2} \exp\Big(-\frac{1}{2}\sum_{i=1}^n (X_i - \vartheta)^2\Big).$$

Für $\vartheta_0 < \vartheta_1$ hat also ein Neyman-Pearson-Test auf $H_0 : \vartheta = \vartheta$ gegen H_1 :

$\vartheta = \vartheta_1$ die Form

$$\begin{aligned}
\varphi &= \mathbf{1}\Big(\exp\Big(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\vartheta_1)^2\Big) > \kappa\exp\Big(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\vartheta_0)^2\Big)\Big)\\
&= \mathbf{1}\Big(\frac{1}{2}\sum_{i=1}^{n}\Big((X_i-\vartheta_0)^2-(X_i-\vartheta_1)^2\Big) > \log(\kappa)\Big)\\
&= \mathbf{1}\Big(\sum_{i=1}^{n}X_i > \frac{\log(\kappa)+n(\vartheta_1^2-\vartheta_0^2)/2}{\vartheta_1-\vartheta_0}\Big).
\end{aligned}$$

Da $\kappa > 0$ gewählt wird, können wir auch gleich die rechte Seite der letzten Ungleichung in $\mathbb{R}$ geeignet wählen, um das Niveau einzuhalten. Jeder Neyman-Pearson-Test hat also die Form $\varphi = \mathbf{1}(\sum_{i=1}^{n} X_i > \tilde{\kappa})$ mit $\tilde{\kappa} \in \mathbb{R}$ geeignet. Dies ist auch sehr intuitiv, weil eine große Summe $\sum_i X_i$ eher für ϑ_1 als für ϑ_0 als Erwartungswert spricht. Im Fall $\vartheta_1 < \vartheta_0$ dreht sich übrigens auch im Neyman-Pearson-Test nur die Ungleichung um.

(c) Wie geht der Beweis des Neyman-Pearson-Lemmas im Spezialfall von Zähldichten?

Genauso, wir ersetzen nur Integrale durch Summen und schreiben hier $p_\vartheta(x)$ statt $L(\vartheta, x)$:

$$\begin{aligned}
&P_{\vartheta_1}(\varphi^*=1)-P_{\vartheta_1}(\varphi=1) = P_{\vartheta_1}(\varphi^*=1,\varphi=0)-P_{\vartheta_1}(\varphi^*=0,\varphi=1)\\
&= \sum_{k\in\mathbb{Z}}\big(\mathbf{1}(\varphi^*(k)=1,\,\varphi(k)=0)-\mathbf{1}(\varphi^*(k)=0,\varphi(k)=1)\big)\,p_{\vartheta_1}(k)\\
&\geqslant \sum_{k\in\mathbb{Z}}\mathbf{1}(\varphi^*(k)=1,\,\varphi(k)=0)\,\kappa p_{\vartheta_0}(k)-\sum_{k\in\mathbb{Z}}\mathbf{1}(\varphi^*(k)=0,\,\varphi(k)=1)\,\kappa p_{\vartheta_0}(k)\\
&= \kappa\big(P_{\vartheta_0}(\varphi^*=1,\varphi=0)-P_{\vartheta_0}(\varphi^*=0,\varphi=1)\big)\\
&= \kappa\big(P_{\vartheta_0}(\varphi^*=1)-P_{\vartheta_0}(\varphi=1)\big) \geqslant 0,
\end{aligned}$$

(d) Wieso ist der Binomialtest zum Niveau α für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_1$ mit $\vartheta_1 > \vartheta_0$ aus Bemerkung 5.10 auch gleichmäßig bester Test vom Niveau α für das einseitige Testproblem $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$?

Für jedes $\vartheta_1 > \vartheta_0$ hat der Neyman-Pearson-Test zum Niveau α die gleiche Form (vergleiche Bemerkung 5.10 und beachte, dass für das Niveau nur ϑ_0 entscheidend ist). Damit besitzt er kleinere Fehlerwahrscheinlichkeit zweiter Art an jeder stelle $\vartheta_1 > \vartheta_0$ als jeder andere Test. Das bedeutet gerade, dass es sich um einen gleichmäßig besten Test handelt.

(e) Wie ergibt sich der zweiseitige Binomialtest aus Beispiel 5.4 als Likelihood-Quotienten-Test?

Es ist (bezüglich Zählmaß)

$$\frac{\sup_{\vartheta\neq 1/2} L(\vartheta,x)}{L(1/2,x)} = \frac{\sup_{\vartheta\in[0,1]}\binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}}{\binom{n}{x}(1/2)^n} = 2^n(x/n)^x(1-x/n)^{n-x},$$

wie eine einfache Kurvendiskussion ergibt. Der Likelihoodquotient ist in x symmetrisch um $n/2$ und wächst in $|x-n/2|$, so dass der Likelihoodquotiententest in der Tat auf den zweiseitigen Binomialtest in Beispiel 5.4 führt.

Ende 21. Vorlesung

5.2 Der χ^2-Anpassungstest

5.15 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, $j = 0, \ldots, 9$. Wie können wir die Null-Hypothese der Gleichverteilung testen?

Unter der Nullhypothese gilt für $X = (X_0, \ldots, X_9)$ gemäß Definition 1.25 $X \sim \mathrm{Mult}(n, 10, p_0, \ldots, p_9)$ mit $p_0 = \cdots = p_9 = 1/10$. Unter der Alternative gelte immer noch $X \sim \mathrm{Mult}(n, 10, p_0, \ldots, p_9)$, aber mit anderen Klassenwahrscheinlichkeiten $p_0, \ldots, p_9$. Wir betrachten dazu den Likelihood-Quotienten-Test, den wir im folgenden bestimmen und asymptotisch für große n approximieren werden.

5.16 Lemma. *Betrachte das durch* $\mathrm{Mult}(n, r, p_1, \ldots, p_r)$ *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 $p = (X_1/n, \ldots, X_r/n)$ angenommen wird. In der Tat ist

$$\begin{aligned}
L(p) &= L(X_1/n, \ldots, X_r/n) \prod_j \Big(\frac{np_j}{X_j}\Big)^{X_j} \\
&\leqslant L(X_1/n, \ldots, X_r/n) \exp\Big(\sum_j X_j \Big(\frac{np_j}{X_j} - 1 \Big) \Big) \\
&= L(X_1/n, \ldots, X_r/n) \exp(n - n) = L(X_1/n, \ldots, X_r/n),
\end{aligned}$$

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. $\square$

5.17 Definition. Der in Lemma 5.16 hergeleitete Test heißt Multinomialtest.

5.18 Bemerkung. Die Bestimmung des kritischen Werts $\tilde\kappa$, um ein gegebenes Niveau α im Multinomialtest einzuhalten, ist schwierig und wird in der Praxis durch Simulation (unter der $\mathrm{Mult}(n,r,p^0)$-Verteilung) erreicht.

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.

5.19 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 $\kappa > 0$.

5.20 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.

5.21 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 Beispiel 4.33 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. □

5.22 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 $w = (\sqrt{p_1^0}, \ldots, \sqrt{p_r^0})^\top \in \mathbb{R}^r$ 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 $D = \mathrm{diag}(1,\ldots,1,0)$ 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 $W := O^\top Z \sim N(0,D)$ gemäß Satz 3.38, d.h. $W_1, \ldots, W_{r-1} \sim N(0,1)$, $W_r = 0$ und unabhängig:

$$|Y_n|^2 \xrightarrow{d} |Z|^2 = |W|^2 = \sum_{j=1}^{r-1} W_j^2 \sim \chi^2(r-1);$$

denn gemäß Beispiel 2.48 ist die Norm eines m-dimensionalen standardnormalverteilten Zufallsvektors gerade $\chi^2(m)$-verteilt. 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 $u\in\mathbb{R}^d$ und $v=(u_1/\sqrt{np_1^0},\dots,u_r/\sqrt{np_r^0})^\top$

$$\begin{aligned}
\mathbb{E}\left[e^{i\langle u,Y_n\rangle}\right] &= \mathbb{E}\left[e^{i(v_1X_1+\cdots+v_rX_r)}\right]e^{-in(v_1p_1^0+\cdots+v_rp_r^0)}\\
&= \sum_k e^{i(v_1k_1+\cdots+v_rk_r)}\frac{n!}{k_1!\cdots k_r!}(p_1^0)^{k_1}\cdots(p_r^0)^{k_r}e^{-in(v_1p_1^0+\cdots+v_rp_r^0)}\\
&= \Big(e^{iv_1}p_1^0+\cdots+e^{iv_r}p_r^0\Big)^n e^{-in(v_1p_1^0+\cdots+v_rp_r^0)}.
\end{aligned}$$

Eine Taylorentwicklung der Exponentialfunktionen zeigt also für $n\to\infty$

$$\begin{aligned}
&\varphi^{Y_n}(u) = \Big(\big(e^{iu_1/\sqrt{np_1^0}}p_1^0+\cdots+e^{iu_r/\sqrt{np_r^0}}p_r^0\big)e^{-i(u_1\sqrt{p_1^0}+\cdots+u_r\sqrt{p_r^0})/\sqrt{n}}\Big)^n\\
&= \Big(\Big(\sum_{j=1}^r p_j^0\big(1+iu_j(p_j^0)^{-1/2}n^{-1/2}-u_j^2(2p_j^0)^{-1}n^{-1}+O(n^{-3/2})\big)\Big)\\
&\quad\times\Big(1-n^{-1/2}\sum_{j=1}^r iu_j(p_j^0)^{1/2}-(2n)^{-1}\Big(\sum_{j=1}^r u_j(p_j^0)^{1/2}\Big)^2+O(n^{-3/2})\Big)\Big)^n\\
&= \Big(1+\frac{1}{2n}\Big(\sum_{j=1}^r u_j(p_j^0)^{1/2}\Big)^2-\frac{1}{2n}\sum_{j=1}^r u_j^2+O(n^{-3/2})\Big)^n\\
&\to \exp\Big(-\frac{1}{2}\Big(\sum_{j=1}^r u_j^2-\Big(\sum_{j=1}^r u_j(p_j^0)^{1/2}\Big)^2\Big)\Big).
\end{aligned}$$

Die quadratische Form im Exponenten ist gerade gleich $-\frac12(|u|^2-\langle u,w\rangle^2) = -\frac12\langle\Sigma u,u\rangle$. Mit Blick auf Lemma 4.39 haben wir damit $\varphi^{Y_n}(u)\to\varphi^{N(0,\Sigma)}(u)$ gezeigt. Nach dem Stetigkeitssatz von Lévy gilt also $Y_n\xrightarrow{d}N(0,\Sigma)$. □

5.23 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*.

5.24 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. $\square$

5.25 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.

▷ **Kontrollfragen**

(a) Im Fall von zwei Klassen entspricht ja die $\mathrm{Mult}(n, 2, (p, 1-p))$-Verteilung einer $\mathrm{Bin}(n, p)$-Verteilung. Wie vereinfachen sich in diesem Fall Multinomial- und χ^2-Test? Vollziehe die Beweise in diesem einfacheren Spezialfall nach.

Ist $X \sim \mathrm{Bin}(n, p)$, so ist $(X, n-X) \sim \mathrm{Mult}(n, 2, p, 1-p)$. Der Multinomialtest ist dann

$$\varphi = \mathbf{1}\Big(X \log\Big(\frac{X}{np^0}\Big) + (n-X)\log\Big(\frac{n-X}{n(1-p^0)}\Big) \geqslant \tilde{\kappa}\Big), \quad \tilde{\kappa} \in \mathbb{R},$$

was gerade einem Binomialtest für $H_0 : p = p^0$ gegen $H_1 : p \neq p^0$ entspricht. Vergleiche Kontrollfrage 20(e).

Die χ^2-Teststatistik ist

$$V_n^2 = \frac{(X-np^0)^2}{np^0} + \frac{(n-X-n(1-p^0))^2}{n(1-p^0)} = \frac{(X-np^0)^2}{np^0(1-p^0)}.$$

nach dem Satz von de Moivre-Laplace (oder ZGWS) gilt $\frac{X-np^0}{\sqrt{np^0(1-p^0)}} \xrightarrow{d} N(0,1)$ für $n \to \infty$ und $X \sim \mathrm{Bin}(n, p^0)$. Wegen Stetigkeit der Quadratfunktion und $Z^2 \sim \chi^2(1)$ für $Z \sim N(0,1)$ folgt also

$$V_n^2 = \Big(\frac{X - np^0}{\sqrt{np^0(1-p^0)}}\Big)^2 \xrightarrow{d} \chi^2(1).$$

Wegen $r = 2$ ist das genau die Aussage von Satz 5.22. Die Beweise werden erheblich expliziter. So besitzt die Kovarianzmatrix $\Sigma \in \mathbb{R}^{2\times 2}$ gerade die normierten Eigenvektoren $w = (\sqrt{p^0}, \sqrt{1-p^0})^\top$ und $w^\perp = (\sqrt{1-p^0}, -\sqrt{p^0})^\top$, so dass man konkret diagonalisieren kann.

(b) Angenommen in Beispiel 5.15, liefert der Zufallszahlengenerator $n = 10\,000$ Ziffern und folgende Häufigkeiten

'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
988	1014	990	995	1010	1004	994	987	1019	999

Schätzen Sie zunächst rein intuitiv, ob die Daten eher für oder gegen gleichwahrscheinliche Ziffern sprechen. Bestimmen Sie nun den χ^2-Anpassungstest und testen Sie zum Niveau $\alpha = 0,1$ die Null-Hypothese gleichwahrscheinlicher Ziffern (Quantile in Mathematik-Software und Tabelle).

Die Abweichungen vom Erwartungswert sind bei jeder Ziffer im Bereich von $0-2\%$. Jede Anzahl ist $\mathrm{Bin}(10\,000, 0,1)$-verteilt, hat also Standardabweichung $\sqrt{np(1-p)} = 30$. Das scheint einzeln also noch keine signifikante Abweichung vom Erwartungswert zu sein, aber bei simultanen Abweichungen im Zehndimensionalen mag das weniger klar sein.

Berechne die χ^2-Statistik mit $np_j^0 = 1000$, $j = 0,\dots,9$, und erhalte $V_n^2 = 0,567$. Wegen $r = 10$ wird das $0,9$-Quantil der $\chi^2(9)$-Verteilung benötigt, das $q_{0,9} = 14,68$ beträgt. Wegen $V_n^2 < q_{0,9}$ lehnen wir die Nullhypothese also nicht ab, dass die Ziffern mit gleicher Wahrscheinlichkeit (und unabhängig) gezogen wurden. Wie bei Mendels Erbsen ist auch hier die Übereinstimmung eher zu groß (auch für $\alpha = 0,995$ hätten wir H_0 noch akzeptiert). Wir sollten also sehr misstrauisch sein, ob der Zufallszahlengenerator nicht vielleicht 'zu gleichverteilt' ist.

(c) Leite formal die stochastische Konvergenz im Beweis von Lemma 5.21 aus der gezeigten Ungleichung her.

Gezeigt wurde für $Y_n := X_j \log(X_j/(np_j^0)) - (X_j - np_j^0)^2/(2np_j^0)$

$$P(|Y_n| > n^{-1/2}K_\varepsilon^3(p_j^0 - n^{-1/2}K_\varepsilon)^{-2}) \leqslant \varepsilon$$

für alle $\varepsilon > 0$ mit jeweils einem $K_\varepsilon > 0$. Für jedes $\varepsilon > 0$ und $\delta > 0$ existiert also ein hinreichend großes $N(\delta,\varepsilon) \in \mathbb{N}$ mit

$$\forall n \geqslant N(\delta,\varepsilon):\ P(|Y_n| > \delta) \leqslant \varepsilon.$$

Das heißt aber gerade, dass die Konvergenz $P(|Y_n| > \delta) \to 0$ für alle $\delta > 0$ vorliegt, also stochastische Konvergenz $Y_n \xrightarrow{P} 0$ (die vertauschte Notation mit δ und ε mag zunächst verwirren).

(d) Wie kann eine orthogonale Matrix O mit $\Sigma = ODO^\top$ im Beweis von Satz 5.22 konstruiert werden?

In den Spalten von O stehen gerade die (orthonormalen) Eigenvektoren $v_1,\dots,v_r$, also $O = \begin{pmatrix} v_1 & \cdots & v_r \end{pmatrix}$. Wähle $v_1 := w$ und bestimme (z.B. mit Gram-Schmidt-Verfahren) eine Orthonormalbasis $v_2,\dots,v_r$ vom orthogonalen Komplement $U = \{u \in \mathbb{R}^r \mid \langle u, w\rangle = 0\}$.

(e) Gilt die Asymptotik in Korollar 5.24 auch für die entsprechende Multinomial-Teststatistik anstelle von V_n^2?

Sei $M_n := 2\sum_{j=1}^r X_j \log(X_j/(np_j^0))$ die Multinomial-Teststatistik, für die Lemma 5.21 $V_n^2 - M_n \xrightarrow{P} 0$ zeigt. Die Frage ist, ob aus $V_n^2 \xrightarrow{d} \chi^2(r-1)$ dann auch $M_n \xrightarrow{d} \chi^2(r-1)$ folgt. Die Antwort ist ja und ist eine Form des *Lemma von Slutsky* (Stochastik II). Hier argumentieren wir so: es genügt, für die charakteristischen Funktionen $\mathbb{E}[e^{iuM_n}] - \mathbb{E}[e^{iuV_n^2}] \to 0$ zu zeigen (da ja dann auch die charakteristische Funktion.von M_n gegen die charakteristische Funktion von $\chi^2(r-1)$ konvergiert). Da $x \mapsto e^{iux}$ Lipschitz-stetig (mit Konstante $|u|$) und durch Eins beschränkt ist, gilt für jedes $u \in \mathbb{R}$

$$\mathbb{E}[|e^{iuM_n} - e^{iuV_n^2}|] \leqslant \mathbb{E}[\min(|u(M_n - V_n^2)|, 2)] \to 0,$$

Dies folgt aus dominierter Konvergenz gemäß Satz 4.19(d), da das Argument des Erwartungswerts stochastisch gegen Null konvergiert (stetige Funktion von $M_n - V_n^2$) und durch 2 beschränkt ist.

Ende 22. Vorlesung

5.3 Einführung in die Schätztheorie

5.26 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.

5.27 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

$$R(\hat{\vartheta}, \vartheta) := \mathbb{E}_\vartheta[(\hat{\vartheta} - \vartheta)^2] := \int_{\mathcal{X}} (\hat{\vartheta}(x) - \vartheta)^2 P_\vartheta(dx) \in [0, \infty], \quad \vartheta \in \Theta.$$

5.28 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.

5.29 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$(x_1, \dots, x_n)$ (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.

5.30 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).

5.31 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. □

5.32 Beispiel (Normalverteilungsmodell)**.** 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.

5.33 Bemerkung. Oft ist (näherungsweise) Erwartungstreue eine gewünschte Eigenschaft, die insbesondere artifizielle Schätzer wie $\hat\vartheta_1$ 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.

5.34 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 $L(\vartheta)/L(\vartheta_0) \in \mathscr{L}^2(P_{\vartheta_0})$

$$R(\hat{\vartheta}, \vartheta_0) \geqslant \frac{(\vartheta - \vartheta_0)^2}{\operatorname{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0))}.$$

5.35 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

$$\begin{aligned}
\vartheta - \vartheta_0 &= \mathbb{E}_\vartheta[\hat{\vartheta} - \vartheta_0] - \mathbb{E}_{\vartheta_0}[\hat{\vartheta} - \vartheta_0] \\
&= \int_{\mathcal{X}} (\hat{\vartheta}(x) - \vartheta_0)(L(\vartheta, x) - L(\vartheta_0, x))\, \mu(dx) \\
&= \int_{\mathcal{X}} (\hat{\vartheta}(x) - \vartheta_0)\Big(\frac{L(\vartheta, x)}{L(\vartheta_0, x)} - 1\Big) L(\vartheta_0, x)\, \mu(dx) \\
&= \mathbb{E}_{\vartheta_0}\Big[(\hat{\vartheta} - \vartheta_0)\Big(\frac{L(\vartheta)}{L(\vartheta_0)} - 1\Big)\Big].
\end{aligned}$$

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 $\operatorname{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0)) = 0$, 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 $\operatorname{Var}_{\vartheta_0}(L(\vartheta)/L(\vartheta_0))$ teilen und erhalten die Chapman-Robbins-Ungleichung. $\square$

5.36 Beispiel (Normalverteilungsmodell)**.** 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

$$\begin{aligned}
\mathbb{E}_{\vartheta_0}\Big[\Big(\frac{L(\vartheta)}{L(\vartheta_0)}\Big)^2\Big] &= \mathbb{E}_{\vartheta_0}\Big[\exp\Big(\frac{2}{\sigma^2}\sum_{i=1}^n(\vartheta-\vartheta_0)X_i\Big]\exp\Big(-\frac{n(\vartheta-\vartheta_0)^2}{\sigma^2}\Big)\\
&= \mathbb{E}\Big[\exp\Big(2\sum_{i=1}^n\frac{\vartheta-\vartheta_0}{\sigma}Z_i\Big)\Big]\exp\Big(-\frac{n(\vartheta-\vartheta_0)^2}{\sigma^2}\Big)\\
&= \mathbb{E}\Big[\exp\Big(2\frac{\vartheta-\vartheta_0}{\sigma}Z_1\Big)\Big]^n\exp\Big(-\frac{n(\vartheta-\vartheta_0)^2}{\sigma^2}\Big)\\
&= \exp\Big(\frac{n(\vartheta-\vartheta_0)^2}{\sigma^2}\Big),
\end{aligned}$$

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.

5.37 Bemerkung. Auch allgemein wird die untere Schranke von Chapman-Robbins meist maximal für $\vartheta \to \vartheta_0$. Falls die Likelihood-Funktion bei ϑ_0 differenzierbar in ϑ ist und alle folgenden Operationen erlaubt sind, so 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.

5.38 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 $\tau(\vartheta) = \vartheta$. □

5.39 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 $\mathbb{E}_\vartheta[X_i - \vartheta] = 0$ und 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 $R(\hat{\vartheta}, \vartheta) = \frac{\vartheta(1-\vartheta)}{n}$ minimal ist. Beachte, dass hier der MSE in der Tat von ϑ abhängt.

Markus Reiß

Vorlesung *Stochastik I*
Sommersemester 2020
Humboldt-Universität zu Berlin

Fragenkatalog zur Prüfung

(a) Formuliere und beweise: Poissonscher Grenzwertsatz, Eigenschaften von Verteilungsfunktionen, Dichtetransformationssatz (einfachste Version, $d = 1$), Bayesformel, Lemma von Borel-Cantelli, Jensensche Ungleichung, Rechenregeln von Erwartungswert, Varianz und Kovarianz, Markovungleichung, Tschebyschevungleichung, schwaches Gesetz der großen Zahlen, zentraler Grenzwertsatz (Standardversion).

(b) Formuliere und gebe die Hauptargumente im Beweis an: Satz von Vitali, 0-1-Gesetz von Kolmogorov, starkes Gesetz der großen Zahlen, Auswahlsatz von Helly, Lévy-Stetigkeitssatz, Grenzwertsatz nach Lindeberg, Satz von Glivenko-Cantelli.

(c) Was ist die Verteilung P^X einer Zufallsvariablen X? Wie ist der Erwartungswert von einer reellwertigen Zufallsvariablen X allgemein definiert? Für welche Funktionen h gilt $\mathbb{E}[h(X)] = \int h(x)P^X(dx)$ und wie beweist man dies? Wie kann man den Erwartungswert bei Existenz von Dichten bzw. Zähldichten berechnen? Welches Fehlerkriterium minimiert der Erwartungswert, welches ein Median? In welchem Sinn ist ein Neyman-Pearson-Test optimal?

(d) Welche Konvergenzarten von Zufallsvariablen gibt es? Welche Implikationen gelten und welche nicht (Gegenbeispiele!)?

(e) Welche Kriterien für Unabhängigkeit von Zufallsvariablen gibt es? Was ist der Zusammenhang zum Produktmaß/Produktraum? Was gilt unter Unabhängigkeit und was bereits für unkorrelierte Zufallsvariablen?

(f) Gib Dichte bzw. Zähldichte an von Gleichverteilung, Binomialverteilung, hypergeometrischer Verteilung, geometrischer Verteilung, Poissonverteilung, Normalverteilung, Exponentialverteilung, χ^2-Verteilung, Γ-Verteilung. Wo treten diese Verteilungen auf? Berechne Erwartungswert, Varianz und charakteristische Funktion von Binomialverteilung, Poissonverteilung, Normalverteilung. Welche dieser Verteilungen bilden Faltungshalbgruppen?

(g) Löse die Probeklausur und Übungsaufgaben erneut. Gib insbesondere bei den Textaufgaben jeweils eine exakte Definition des Wahrscheinlichkeitsraums und der Zufallsvariablen und begründe gut die Modellierung.

(h) Überlege zu den Resultaten in (a) und (b) Beispiele sowie Gegenbeispiele, wo Voraussetzungen verletzt sind und die Aussage nicht gilt.

Markus Reiß

Vorlesung *Stochastik I*
Sommersemester 2020
Humboldt-Universität zu Berlin

Informationen zur Klausur am 21.8.
21/08 Exam Information

(a) Es gibt eine deutsche und englische Version des Aufgabenblatts. Unabhängig davon dürfen Sie Englisch oder Deutsch für die Antworten wählen.

There will be available an English and a German version of the exam. In any case, you may write solutions in English or German.

(b) Sie dürfen (und sollten!) ein handbeschriebenes DIN A4-Blatt (beidseitig) als Gedächtnisstütze mitbringen sowie Stift und (außer Namen, Matrikelnr.) leere DIN A4-Blätter zum Schreiben. Andere Hilfsmittel (Bücher, Rechner, Handy) sind nicht erlaubt.

You may (and should!) bring one hand-written a4 sheet of paper (double-sided) as a memory aid for you as well as a pen and (besides your name and matriculation number) void a4 pages for your solutions. Other tools (like books, calculator, mobile phone) are not permitted.

(c) Am Dienstag 18.8., 9:30 Uhr, biete ich eine Zoom-Sprechstunde an, wo Fragen von Ihnen sowie aus dem Fragenkatalog diskutiert werden.

On Tuesday 18/08, 9h30, I'll offer a zoom meeting where questions from you and the above question catalogue will be discussed.