Einführung in die Stochastik

Volker Betz

1. Wahrscheinlichkeitsräume und Zufallsvariable

Zunächst sollen kurz und nichtrigoros einige Beispiele für Fragestellungen gegeben werden, mit denen die Wahrscheinlichkeitstheorie (W-Theorie) sich befasst.

Typische Fragestellungen der Wahrscheinlichkeitstheorie

(1.1) Einfache Irrfahrt in d Dimensionen

Starte bei $x = 0 \in \mathbb{R}^d$, mache N Schritte mit Schrittlänge r in zufällige Richtungen und zwar

$d = 1$: Nach links und rechts je mit W'keit 50%,
$d = 2$: Nach links, rechts, oben, unten mit je 25%
$d = 3$: in jede Richtung mit $1/6 * 100\%$ etc.

Anwendungen:

$d = 1$: Glücksspiel mit Münzwurf.
$d \geqslant 2$: Diffusion eines Teilchens in einem Medium.

Fragestellungen:

a) Wie weit kommt man „im Durchschnitt“ oder „typischerweise“ nach N Schritten für große N?
Antworten: Gesetz der großen Zahl, Zentraler Grenzwertsatz, später in dieser Vorlesung.
b) Wie wahrscheinlich ist es, dass man *wesentlich* weiter kommt als in a) berechnet?
Antworten: Theorie der großen Abweichungen, nächstes Semester.
c) Wie sieht ein „typischer Pfad“ der Irrfahrt aus, wenn man ihn lange laufen lässt und von weit weg betrachtet, d.h. Grenzwert $N \to \infty$ und $r \to 0$? Gibt es ein Grenzobjekt, und was hat es für Eigenschaften?
Antwort: Theorie der Brown'schen Bewegung, Vertiefungszyklus.

(1.2) Selbstvermeidende Irrfahrt

Situation wie in (1.1), aber nun darf man keine Stelle betreten, die man zuvor betreten hat. Ist nur in $d \geqslant 2$ interssant. **Achtung:** man kann einen Pfad der Länge N nicht sukzessive aufbauen, da man sich festfahren kann. Stattdessen wählt man unter allen zulässigen Pfaden der Länge N „zufällig“ (d.h.: gleichverteilt) einen aus.

Anwendung: Räumliche Konfiguration von Polymeren

Fragestellungen:

a) Wie weit kommt man nach N Schritten (siehe Problem (1.1.a) oben!).
Antwort: für $2 \leqslant d \leqslant 4$ nicht bekannt! Wichtiges (und schwieriges) ungelöstes Problem. Für $d > 4$ Antwort wie in (1.1 a), ist aber schwer zu beweisen (sogenannte Lace Expansion).

b) Wie sieht ein typischer Pfad aus (siehe Problem (1.1.c) oben): **Antwort:** In $d = 2$ glaubt man zu wissen, was herauskommt (sog. Schramm-Löwner Kurve, im Zusammenhang mit der Theorie dieser Kurven gab es 2 Fields-Medaillen (2006 und 2010)), man kann es aber nicht beweisen. In $d = 3, 4$ weiß man sehr wenig. In $d \geqslant 5$ Antwort wie in (1.1.c), aber schwer zu beweisen.

(1.3) Perkolation

Auf einem unendlich großen karierten Papier wirft man für jedes Kästchen eine Münze, die mit Wahrscheinlichkeit $p \cdot 100\%$ „Kopf“ zeigt, und mit $(1-p) \cdot 100\%$ „Zahl“. Jedes Kästchen, für das „Kopf“ geworfen wurde, färbt man schwarz.

Fragestellung: Wie groß muss p sein, damit man eine Chance hat, einen Weg von 0 nach ∞ zu finden, dessen Schritte immer nur zu benachbarten Kästchen führen, und der nur schwarze Kästchen benutzt?
Antwort (und noch mehr): Perkolationstheorie, siehe z.B. Buch von G. Grimmett (447 Seiten).
Anwendung: Modell für poröse Materialien

(1.4) Diffusion limited aggregation (DLA)

Wie in (1.3) betrachte ein unendliches kariertes Papier. Färbe das Kästchen am Koordinatenursprung schwarz. Nun starte eine einfache Irrfahrt (siehe (1.1)) „bei ∞“ und warte, bis sie eines der Kästchen trifft, die neben dem schwarz eingefärbten liegt. Färbe dieses Kästchen ebenfalls schwarz. Dann starte eine weitere Irrfahrt bei ∞ und lasse diese laufen, bis ein an das bereits schwarz eingefärbte Gebilde angrenzendes Teilchen trifft, und färbe dieses Kästchen schwarz. Dann immer so weiter.

Fragestellungen: Wie sieht das entstehende Gebilde aus schwarzen Kästchen aus?
Welchen typischen Durchmesser hat es nach N Läufen?
Wenn man es $N \to \infty$ gehen lässt und gleichzeitig Kästchen von immer kleinerem Durchmesser wählt, erhält man dann ein geometrisches Objekt als Grenzwert? Wenn ja, was sind seine Eigenschaften, z.B. fraktale Dimension?

Antworten: Sehr wenig ist bekannt - fast nur Simulationen. Neue Entwicklungen in einem eng verwandten Modell: Sidovaricius, Stauffer: Invent. math. (2019) 218:491–571. Hieraus stammt auch die abgebildete Simulation.

(1.5) Fazit und Agenda dieser Vorlesung

Aus den vorigen Beispielen wird eine Besonderheit der Wahrscheinlichkeitstheorie innerhalb der Mathematik klar, die sonst nur noch in der Zahlentheorie zu beobachten ist: es gibt sehr viele natürliche und leicht zu beschreibende Fragestellungen, deren Lösung sehr schwierig und oft

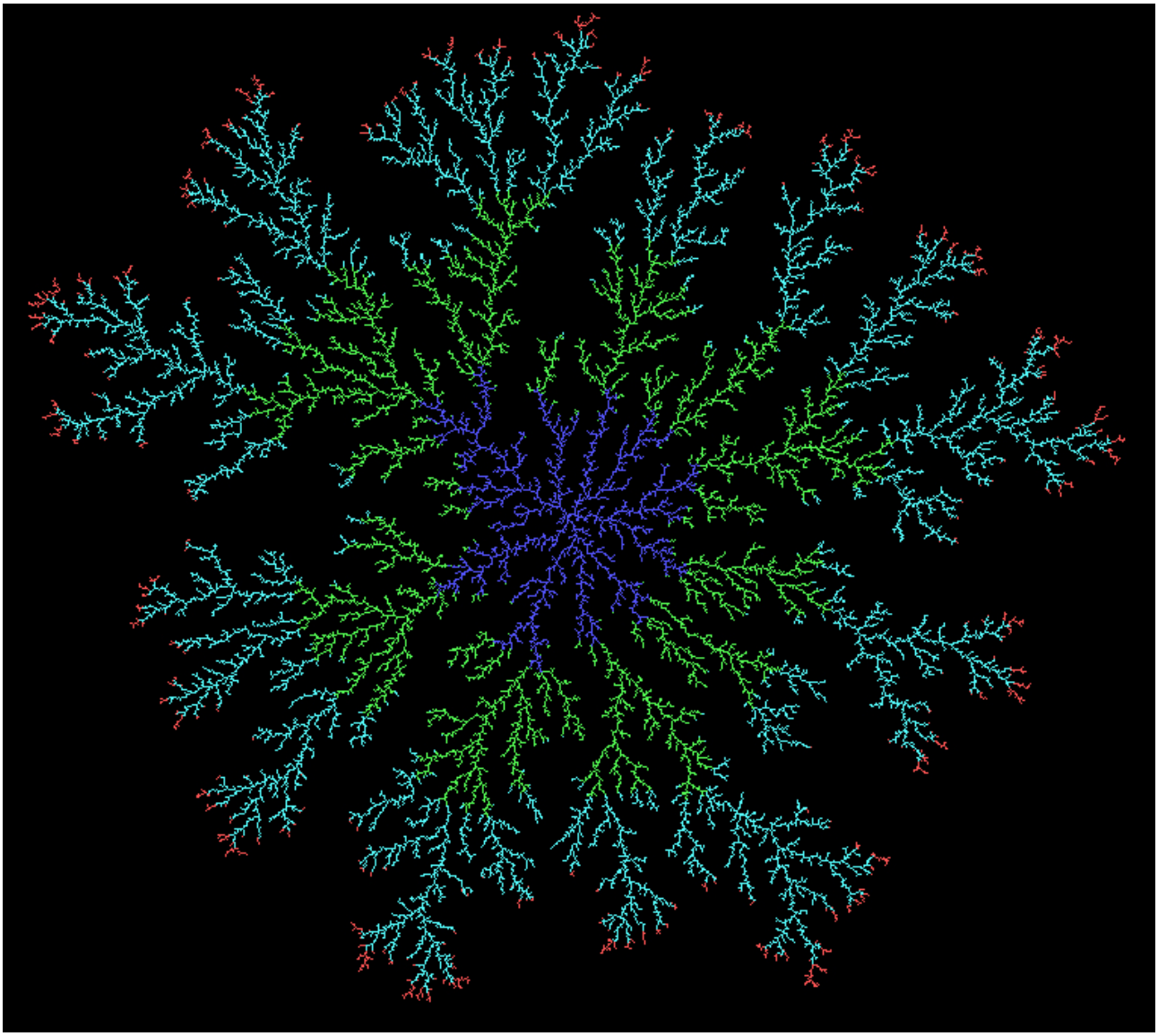


ABBILDUNG 1. Simulation eines DLA-Clusters. Quelle: Wikipedia, public domain

noch unbekannt ist. Bevor wir uns allerdings solchen Fragestellungen weiter widmen können, müssen wir Sprache und Techniken der W-Theorie erlernen. Im Einzelnen:

a) Wir brauchen rigorose mathematische Modelle für Situationen wie die oben beschriebenen, d.h. Formulierungen von zufälligen Situationen mit Hilfe von Mengen, Abbildungen, Axiomen etc. Ebenso für Begriffe wie „im Durchschnitt“, oder „typischerweise“. Dies machen wir als nächstes in dieser Vorlesung.
b) Alle obigen Beispiele betrachten Grenzwerte, zB in der Anzahl der Schritte einer Irrfahrt, oder implizit durch Annahme eines unendlich großen karierten Papiers. Wir müssen lernen, solche Grenzwerte sauber zu formulieren und etwas über sie herauszufinden. Dies machen wir im zweiten Teil der Vorlesung in den einfachsten Fällen.

c) So wie in anderen mathematischen Disziplinen gibt es auch in der W-Theorie grundlegende Beispiele, Bausteine und Techniken, auf denen man aufbaut. In der Analysis sind die Bausteine z.B. Polynome, Exponential- und Winkelfunktionen, Folgen, Reihen, und die Techniken sind ε-δ Beweise von Stetigkeit und Grenzwerten, Differentialrechnung, Integralsätze etc. Wir lernen in dieser Vorlesung sowohl die wichtigsten Wahrscheinlichkeitsmaße (Bausteine) als auch die grundlegendsten Techniken kennen.

Wahrscheinlichkeitsräume

Wir geben zunächst einige abstrakte Definitionen, deren Sinn vielleicht nicht sofort klar wird. Im Anschluss werden wir dann aber klären, was wir uns unter diesen Objekten vorzustellen haben.

(1.6) Definition

Die **Potenzmenge** $\mathcal{P}(\Omega)$ einer Menge Ω ist die Menge, deren Elemente alle Teilmengen von Ω sind.
Beispiel: $\Omega = \{1, 2, 3\}$, dann

$$\mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

(1.7) Definition

Ω sei eine Menge. Eine **σ-Algebra $\mathcal{F}$ über** Ω ist eine Teilmenge von $\mathcal{P}(\Omega)$ (man sagt auch: ein „Mengensystem“), das folgende Eigenschaften hat:

a) $\Omega \in \mathcal{F}$.
b) Falls $A \in \mathcal{F}$, dann ist auch $A^c \in \mathcal{F}$.
c) Für abzählbar viele Mengen $A_1, A_2, \ldots$ mit $A_i \in \mathcal{F}$ für alle i ist auch $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Das Paar $(\Omega, \mathcal{F})$ heißt dann auch **messbarer Raum** (oder auch **Ereignisraum**).
Beispiele: $\mathcal{F}_{\max} = \mathcal{P}(\Omega)$ ist eine σ-Algebra, die größtmögliche.
$\mathcal{F}_{\min} = \{\emptyset, \Omega\}$ ist die kleinstmögliche σ-Algebra.
Für $\Omega = \{1, 2, 3\}$ ist $\mathcal{F} = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$ eine σ-Algebra, aber $\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2, 3\}\}$ ist keine.

Bemerkung: in der Definition oben ist es equivalent, in der Situation von c) zu fordern, dass $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$ ist (Beweis: Übung!).

(1.8) Definition

$(\Omega, \mathcal{F})$ sei messbarer Raum. Eine Abbildung $\mathbb{P} : \mathcal{F} \to [0,1]$ heißt **Wahrscheinlichkeitsmaß** (kurz: **W-Maß**), falls sie folgende Eigenschaften hat:

a) $\mathbb{P}$ ist *normiert*, d.h. $\mathbb{P}(\Omega) = 1$

b) $\mathbb{P}$ ist *σ-additiv*, d.h. für abzählbar viele disjunkte Mengen $A_1, A_2, \ldots$ mit $A_i \in \mathcal{F}$ (disjunkt bedeutet: $A_i \cap A_j = \emptyset$ falls $i \neq j$) gilt:

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

Das Tripel $(\Omega, \mathcal{F}, \mathbb{P})$ heißt dann **Wahrscheinlichkeitsraum** (kurz: W-Raum).
Beispiele: Auf $\mathcal{F}_{\min}\{\emptyset, \Omega\}$ gibt es nur das triviale W-Maß mit $\mathbb{P}(\Omega) = 1$ und $\mathbb{P}(\emptyset) = 0$.
Für $\Omega = \mathbb{N}$ und $\mathcal{F} = \mathcal{P}(\mathbb{N})$ kann man jedes W-Maß mit Hilfe einer Folge schreiben: Sei $(p_n)_{n\in\mathbb{N}}$ eine Folge mit $p_n \geqslant 0$ für alle n und $\sum_{n=1}^{\infty} p_n = 1$. Dann ist $\mathbb{P} : \mathcal{F} \to [0,1]$ mit

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} p_n \mathbb{1}_A(n) \equiv \sum_{n\in A} p_n \qquad \text{(wobei } \mathbb{1}_A(n) = 1 \text{ falls } n \in A \text{ und } \mathbb{1}_A(n) = 0 \text{ sonst)}$$

ein W-Maß auf $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Jedes W-Maß auf $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ kann mit Hilfe einer geeigneten Folge dargestellt werden (Beweis: Übung, siehe auch Satz (1.12) unten).

(1.9) Bezeichnung

Sei Ω eine Menge und $A \subset \Omega$. Die Abbildung

$$\mathbb{1}_A : \Omega \to \{0,1\}, \qquad \mathbb{1}_A(\omega) = \begin{cases} 1 & \text{falls } \omega \in A \\ 0 & \text{sonst} \end{cases}$$

heißt **Indikatorfuntion** von A.

(1.10) Interpretation und Nomenklatur

a) Ω modelliert alle möglichen Zustände, die das uns interessierende System einnehmen kann. Beim Standard-Würfel kann man also $\Omega = \{1, 2, \ldots, 6\}$ wählen, bei viermal Würfeln $\Omega = \{(i,j,k,l) : 1 \leqslant i,j,k,l \leqslant 6\}$, bei zweimal würfeln ohne Beachtung der Reihenfolge $\Omega = \{\{i,j\} : 1 \leqslant i,j \leqslant 6\}$. Bei der Perkolation (siehe Beispiel (1.3)) ist eine gute Wahl $\Omega = \{\eta : \mathbb{Z}^d \to \{0,1\}\} = \{0,1\}^{\mathbb{Z}^d}$ (Menge aller Abbildungen von $\mathbb{Z}^d$ nach $\{0,1\}$. Hier bedeutet $\eta(z) = 1$, dass das Kästchen bei z schwarz gefärbt ist.

b) Jedes $A \in \mathcal{P}(\Omega)$ entspricht der Antwort auf eine Ja-Nein-Frage, die wir in unserem System stellen können.
Beispiel Würfel: $A = \{1,2,3\}$ bedeutet „es wurde eine eins, zwei oder drei geworfen“.
Beispiel Perkolation: $A = \{\eta \in \{0,1\}^{\mathbb{Z}^d} : \eta(z) = \eta(w)\}$ bedeutet: Kästchen z und Kästchen w haben die gleiche Farbe.

c) Wenn man $\mathcal{F} \neq \mathcal{P}(\Omega)$ wählt, dann bedeutet das, dass man gewisse Fragen an das System nicht erlaubt (nämlich diejenigen, deren „Antwortmengen“ A nicht in $\mathcal{F}$ sind). Warum sollte man das wollen?

Grund 1: Modellierung von mangelndem Wissen, z.B.: Beim Würfeln mit zwei Würfeln ist $\Omega = \{(i,j) : 1 \leqslant i,j \leqslant 6\}$. Setzt man nun $\Omega_0 = \{1,\dots,6\}$ und wählt

$$\mathcal{F}_1 := \big\{A \times \Omega_0 : A \subset \{1,2,\dots,6\}\big\} \equiv \big\{\{(i,j) : i \in A, j \in \Omega_0\} : A \subset \{1,2,\dots,6\}\big\},$$

dann bedeutet das, dass man über den zweiten Würfelwurf nichts wissen kann oder will (obwohl es ihn gibt, das hat man ja in Ω modelliert). Genauso kann man im Fall von

$$\mathcal{F}_2 := \big\{\Omega_0 \times A : A \subset \{1,2,\dots,6\}\big\} \equiv \big\{\{(i,j) : i \in \Omega_0, j \in A\} : A \subset \{1,2,\dots,6\}\big\}$$

nichts über den ersten Würfelwurf sagen. Da man manchmal beide σ-Algebren gleichzeitig auf Ω betrachtet, sollte man nicht einfach $\Omega = \{1,2,\dots,6\}$ wählen, um die Modellierung zu vereinfachen.
Eine weitere interessante σ-Algebra ist diejenige, die entsteht, wenn man nur die Summe der beiden Würfel kennen darf - wie sieht die aus?

Grund 2: technische Schwierigkeiten mit der Potenzmenge - dazu später mehr.

d) **Nomenklatur:** Die Elemente einer σ-Algebra heißen auch **Ereignisse**. Ereignisse, die einelementige Mengen sind (also $A = \{\omega\}$ für $\omega \in \Omega$) nennt man auch **Elementarereignisse**.

e) **Konsistenz:** Die Bedingungen an $\mathcal{F}$ in (1.7) sorgen dafür, dass die obige Interpretation logisch konsistent ist: Wenn man wissen kann, ob Ereignis A eintritt ($A \in \mathcal{F}$), muss man auch wissen können, ob es nicht eintritt ($A^c \in \mathcal{F}$). Wenn man für Ereignisse A und B jeweils wissen kann, ob sie eintreten, so muss man auch wissen können, ob beide gleichzeitig eintreten ($A \cap B$) oder ob mindestens eines davon eintritt ($A \cup B$). Ebenso für mehr als zwei und als Abstraktion sogar für abzählbar viele. **Achtung:** Für überabzählbar viele verlangt man das nicht - das hat wichtige technische Gründe!
Für ein W-Maß bedeutet $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ falls $A \cap B = \emptyset$: wenn zwei Ereignisse sich gegenseitig ausschließen, dann müssen sich ihre Eintreffwahrscheinlichkeiten addieren. Ebenso für endlich viele und (als Abstraktion) für abzählbar viele sich gegenseitig ausschließende Ereignisse.
Bevor wir weitere Beispiele betrachten, sammeln wir zunächst die wichtigsten Rechenregeln für W-Maße.

(1.11) Satz

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum, alle unten erwähnten Mengen seien in $\mathcal{F}$.

a) $\mathbb{P}(\emptyset) = 0$, und $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.

b) $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$.

c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

d) Falls $A \subset B$, dann ist $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.

e) Es gelte $A_n \nearrow A$ wenn $n \to \infty$, d.h. $A_1 \subset A_2 \subset A_3 \subset \dots$ und $\bigcup_{n=1}^{\infty} A_n = A$. Dann gilt

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A) \qquad \text{(Stetigkeit von unten)}.$$

f) Es gelte $A_n \searrow A$ wenn $n \to \infty$, d.h. $A_1 \supset A_2 \supset A_3 \supset \dots$ und $\bigcap_{n=1}^{\infty} A_n = A$. Dann gilt

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A) \qquad \text{(Stetigkeit von oben)}.$$

g) Für alle Folgen von Mengen (A_n) gilt

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n) \qquad (\sigma\text{-Subadditivität}).$$

Beweis: a) ist klar mit folgendem Trick: Setze $A_1 = A$, $A_2 = A^c$, $A_i = \emptyset$ für alle anderen i. Damit ist

$$1 \overset{(1.8a)}{=} \mathbb{P}(\Omega) \overset{(1.8b)}{=} \mathbb{P}(A) + \mathbb{P}(A^c) + \sum_{i=1}^{\infty} \mathbb{P}(\emptyset).$$

Damit dies gelten kann, muss $\mathbb{P}(\emptyset) = 0$ sein, und dann folgt die andere Aussage auch.

b) gilt wegen $A = (A \setminus B)\dot\cup(A \cap B)$;
(**Notation:** $\dot\cup$ bedeutet, dass die vereinigten Mengen (paarweise) disjunkt sind).
Daher ist mit (1.8 b) (Version für endliche Vereinigungen, siehe a) oben): $\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$, wie behauptet.

c) gilt wegen $A \cup B = (A \setminus B)\dot\cup(B \setminus A)\dot\cup(A \cap B)$ und weiter wie in b).

d) gilt wegen $B = A\dot\cup(B \setminus A)$ und $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geqslant \mathbb{P}(A)$.

e) Wir zerlegen

$$A = A_1\dot\cup(A_2 \setminus A_1)\dot\cup(A_3 \setminus A_2)\dot\cup\ldots = A_1\dot\cup\dot\bigcup_{n\in\mathbb{N}}(A_{n+1} \setminus A_n).$$

Daher ist mit (1.8 b)

$$\mathbb{P}(A) = \mathbb{P}(A_1) + \sum_{n=1}^{\infty} \mathbb{P}(A_{n+1} \setminus A_n) = \lim_{N\to\infty} \big(\mathbb{P}(A_1) + \sum_{n=1}^{N} \mathbb{P}(A_{n+1} \setminus A_n)\big) =$$
$$= \lim_{N\to\infty} \mathbb{P}\Big(A_1\dot\cup\dot\bigcup_{n\leqslant N}(A_{n+1} \setminus A_n)\Big) = \lim_{N\to\infty} \mathbb{P}(A_{N+1}).$$

f) Benutze $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$ und $\mathbb{P}(A^c) = \mathbb{P}(\bigcup_{n=1}^{\infty} A_n^c)$, dann wende e) an.

g) Setze $B_n = A_n \setminus \big(\bigcup_{i=1}^{n-1} A_i\big)$. Dann ist $B_n \subset A_n$ für alle n, und $\bigcup_{n=1}^{\infty} A_n = \dot\bigcup_{n=1}^{\infty} B_n$. Daher ist

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty} B_n\Big) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) \overset{d)}{\leqslant} \sum_{n=1}^{\infty} \mathbb{P}(A_n). \qquad \square$$

Wenn Ω eine abzählbare Menge ist, dann gibt es eine sehr einfache und nützliche Darstellung von W-Maßen über $(\Omega, \mathcal{P}(\Omega))$:

(1.12) Definition und Satz

a) Ω sei eine abzählbare Menge. Eine Funktion $p : \Omega \to \mathbb{R}$ mit den Eigenschaften

$$0 \leqslant p(\omega) \leqslant 1 \text{ für alle } \omega \in \Omega, \qquad \text{und} \quad \sum_{\omega\in\Omega} p(\omega) = 1$$

heißt **Gewichtsfunktion** auf Ω.

b) Für jede Gewichtsfunktion ist die Abbildung

$$\mathbb{P} : \mathcal{P}(\Omega) \to [0,1], \quad A \mapsto \sum_{\omega\in A} p(\omega)$$

ein W-Maß auf $(\Omega, \mathcal{P}(\Omega))$. p heißt oft auch **Zähldichte** von $\mathbb{P}$.

c) Ist $\tilde{\mathbb{P}}$ ein beliebiges W-Maß auf $(\Omega, \mathcal{P}(\Omega))$ und ist Ω abzählbar, so besitzt $\tilde{\mathbb{P}}$ eine Gewichtsfunktion, nämlich $\tilde{p}(\omega) = \tilde{\mathbb{P}}(\{\omega\})$.
Die Beweise sind recht einfach und verbleiben als Übung.

(1.13) Beispiel: Ruin des Spielers

Wir starten das Spiel mit k Goldstücken, und gewinnen oder verlieren eines pro Spielrunde mit Wahrscheinlichkeit je $1/2$. Das Spiel endet wenn wir 0 Goldstücke haben (wir sind pleite), oder wenn wir M Goldstücke haben (wir gehen mit dem Gewinn nach Hause).

Variante 1: Zusätzlich nehmen wir an, dass wir höchstens N Spiele machen. Das führt uns zu folgendem Modell. Die Menge

$$\Omega_N = \big\{(x_0, x_1, \ldots, x_N) : 0 \leqslant x_i \leqslant M \ \forall i \in \{1, \ldots, N\}\big\}$$

enthält alle denkbaren zeitlichen Entwicklungen unseres Vermögens, auch die, die nicht den Regeln des Spiels entsprechen, also Gewinn von mehr als 1 Einheit pro Runde, Weiterspielen nach Ruin etc. Als σ-Algebra nehmen wir einfach $\mathcal{F}_N = \mathcal{P}(\Omega_N)$, und das W-Maß definieren wir durch

$$\mathbb{P}_N\big(\{(x_1, \ldots, x_N)\}\big) = \begin{cases} (1/2)^N & \text{falls } x_0 = k, |x_{i+1} - x_i| = 1 \ \forall i < N, 0 < x_i < M \ \forall i < N, \\ (1/2)^{n_0} & \text{falls } x_0 = k, |x_{i+1} - x_i| = 1 \ \forall i \leqslant n_0 - 1, 0 < x_i < M \ \forall i < n_0, \\ & \text{und } x_{i+1} = x_i \in \{0, M\} \ \forall n_0 \leqslant i < N \\ 0 & \text{sonst} \end{cases}$$

und erweitern es auf ganz $\mathcal{F}$ mittels

$$\mathbb{P}(A) = \sum_{\boldsymbol{x}\in A} \mathbb{P}(\{\boldsymbol{x}\}) \qquad \text{für alle } A \in \mathcal{F}.$$

(Notation hier: $\boldsymbol{x} = (x_1, \ldots, x_N)$). Man sieht also, dass wir den Spielverläufen, die zwar in Ω enthalten sind aber gegen die Regel verstoßen, einfach die W-keit 0 zuordnen. Dadurch sind sie für uns unsichtbar.

Wir müssen aber noch prüfen, ob wir überhaupt ein W-Maß definiert haben! $p(\boldsymbol{x}) := \mathbb{P}(\{\boldsymbol{x}\})$ ist eine Gewichtsfunktion, falls wir $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1$ prüfen können. Dies zu tun scheint erst mal wenig attraktiv, aber hier hilft folgende Idee:

Betrachte die Matrix

$$P = (p_{i,j})_{0 \leqslant i,j \leqslant M} \quad \text{mit} \quad p_{i,j} = \begin{cases} 1/2 & \text{falls } 0 < i < M \text{ und } |i-j| = 1, \\ 1 & \text{falls } i \in \{0, M\}, \text{ und } i = j, \\ 0 & \text{sonst.} \end{cases}$$

Dann gilt:

$$\mathbb{P}_N(\{(k, x_1, \ldots, x_N)\}) = p_{k,x_1} p_{x_1,x_2} \cdots p_{x_{N-1},x_N}, \qquad \text{und} \qquad \sum_{y=0}^{M} p_{x,y} = 1 \text{ für alle } x \in \{0, \ldots, M\}.$$

Daher ist

$$\begin{aligned} \mathbb{P}_N(\Omega) &= \sum_{\boldsymbol{x} \in \Omega} \mathbb{P}_N(\boldsymbol{x}) = \sum_{x_1=0}^{M} \cdots \sum_{x_N=0}^{M} p_{k,x_1} p_{x_1,x_2} \cdots p_{x_{N-1},x_N} = \\ &= \sum_{x_1=0}^{M} p_{k,x_1} \sum_{x_2=0}^{M} p_{x_1,x_2} \cdots \sum_{x_{N-1}=0}^{M} p_{x_{N-2},x_{N-1}} \underbrace{\sum_{x_N=0}^{M} p_{x_{N-1},x_N}}_{=1} = \\ &= \sum_{x_1=0}^{M} p_{k,x_1} \sum_{x_2=0}^{M} p_{x_1,x_2} \cdots \underbrace{\sum_{x_{N-1}=0}^{M} p_{x_{N-2},x_{N-1}}}_{=1} = \ldots = 1. \end{aligned}$$

In Matrix-Sprache kann man das auch so ausdrücken: $v_1 = (1, \ldots, 1)^t$ (als Spaltenvektor) ist ein Rechts-Eigenvektor der Matrix P, und $\mathbb{P}(\Omega) = P^N v_1 = v_1$, wobei P^N das N-fache Matrixprodukt von P mit sich selbst bezeichnet.
Sie sollten sich davon überzeugen, dass die obige Konstruktion immer noch funktioniert, wenn man die Spielregeln durch beliebige andere Spielregeln ersetzt im folgenden Sinne: Startet man in einer Spielrunde mit dem Vermögen $\bar{x}$, so erhält man in der folgenden Runde mit Wahrscheinlichkeit $p(x, y)$ das Vermögen y, wobei $0 \leqslant p(x, y) \leqslant 1$ und $\sum_{y=0}^{M} p(x, y) = 1$ gelten muss. Wir werden unten gleich eine wichtige Verallgemeinerung dieser Ideen sehen.

Variante 2: Wenn wir keine maximale Anzahl von Spielen haben wollen, sollten wir es vermutlich mit

$$\Omega = \{(x_i)_{i \in \mathbb{N}} : 0 \leqslant x_i \leqslant M\}$$

versuchen. Das ist auch richtig. Als σ-Algebra käme $\mathcal{F} = \mathcal{P}(\Omega)$ in Frage, das aber führt zu erheblichen technischen Schwierigkeiten, die man nicht überwinden kann. Grund ist, dass die Menge Ω nun schon überabzählbar ist, wodurch die noch viel größere Potenzmenge $\mathcal{P}(\Omega)$ „zu viele“ Mengen enthält. Warum ist das potentiell problematisch? Weil die Bedingung (1.8 b) in der Definition des W-Maßes verlangt, das etwas *für alle* Folgen von disjunkten Mengen gilt - dies wird immer schwieriger zu erfüllen, wenn man mehr solche Folgen hat, d.h. wenn die σ-Algebra größer wird. Es gibt dann also tendenziell weniger W-Maße, und es stellt sich heraus, dass $\mathcal{P}(\Omega)$ im vorliegenden Fall bereits zu groß ist, so dass viele W-Maße, die man gerne hätte, nicht mehr existieren. Das gleiche gilt, wenn man versucht, ein W-Maß (oder allgemeiner ein Maß) auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ zu definieren: es gibt zwar schon W-Maße auf diesem Raum (z.B. für jedes x dasjenige mit $\mathbb{P}(A) = 1$ genau dann wenn $x \in A$, oder allgemeiner alle W-Maße, für die

$\mathbb{P}(A) = 1$ für eine abzählbare Teilmenge von $\mathbb{R}$ gilt), aber viele andere, die man gerne hätte, gibt es eben nicht.
Diese Problematik wird in der Vorlesung der Maßtheorie noch genauer thematisiert. Für uns genügt es im Moment, festzuhalten, dass wir im Moment Variante 2 noch nicht machen wollen, und uns später eine geeignetere σ-Algebra überlegen werden.
Beispiel (1.13 a) hat eine sehr wichtige Verallgemeinerung:

(1.14) Definition und Satz

E sei eine abzählbare Menge, $N \in \mathbb{N}$, $\bar{x} \in E$. $p : E \times E \to \mathbb{R}$ sei eine Funktion (auch: „Matrix“) mit folgenden Eigenschaften:

$$0 \leqslant p(x,y) \leqslant 1 \text{ für alle } x, y \in E, \qquad \text{und} \qquad \sum_{y \in E} p(x,y) = 1.$$

Setze

$$\Omega = E^{N+1} = \{(x_0, x_1, \dots, x_N) : x_i \in E \ \forall 0 \leqslant i \leqslant N\}, \qquad \mathcal{F} = \mathcal{P}(\Omega).$$

Definiere das W-Maß $\mathbb{P}^{\bar{x}}$ mittels

$$\mathbb{P}^{\bar{x}}(\{(x_0, x_1, \dots, x_N)\}) = \mathbb{1}_{\{x_0 = \bar{x}\}} p(x_0, x_1) p(x_1, x_2) \cdots p(x_{N-1}, x_N),$$

und $\mathbb{P}^{\bar{x}}(A) = \sum_{\boldsymbol{x} \in A} \mathbb{P}(\{\boldsymbol{x}\})$. Hier ist $\mathbb{1}_{\{x_0 = \bar{x}\}}$ eine Abkürzung für $\mathbb{1}_A((x_0, \dots, x_N))$ mit der Wahl $A = \{(\tilde{x}_0, \dots, \tilde{x}_N) \in \Omega : \tilde{x}_0 = \bar{x}\}$. Dass $\mathbb{P}^{\bar{x}}$ wirklich ein W-Maß ist sieht man genau wie im Beispiel (1.13 a).
$\mathbb{P}^{\bar{x}}$ heißt **Pfadmaß der Markovkette mit Übergangsmatrix** $P = (p(x,y))_{x,y \in E}$ **und Startpunkt** $\bar{x}$.
Ist μ ein W-Maß auf E, so heißt das W-Maß (!) $\mathbb{P}^\mu$ mit $\mathbb{P}^\mu = \sum_{x \in E} \mu(x) \mathbb{P}^x$ das **Pfadmaß der Markovkette mit Übergangsmatrix** P **und Startverteilung** μ**.**
Die Menge E wird oft als der **Zustandsraum** der Markovkette bezeichnet.

(1.15) Beispiel

Die einfache Irrfahrt (Bsp. 1.1), eingeschränkt auf N Schritte, ist eine Markovkette mit $E = \mathbb{Z}^d$, $\Omega = E^{N+1}$, $\bar{x} = 0$ und $p(x,y) = \frac{1}{2d} \mathbb{1}_{\{|x-y|=1\}}$.
Es fällt auf, dass $p(x,y)$ die Wahrscheinlchkeit ist, im nächsten Schritt in y zu sein wenn man weiß, dass man im Moment in x ist. Es ist nützlich, diese Konzept zu verallgemeinern und zu formalisieren.

(1.16) Definition

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum, $A, B \in \mathcal{F}$ mit $\mathbb{P}(B) > 0$. Dann heißt

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

die **bedingte Wahrscheinlichkeit von** A **gegeben** B.

(1.17) Proposition und Beispiel

$\mathbb{P}^\mu$ sei das Pfadmaß einer Markovkette mit Zustandsraum E, Übergangsmatrix P, Startverteilung μ und N Schritten. Für $k < N$ und $x \in E$ und $C_1, \dots, C_{k-1} \subset E$ setze

$$B := \{\boldsymbol{x} \in \Omega : x_1 \in C_1, \dots, x_{k-1} \in C_{k-1}, x_k = x\}.$$

Dann gilt für alle $y \in E$:

$$\mathbb{P}^\mu(\{\boldsymbol{x} \in \Omega : x_{k+1} = y\} \mid B) = p(x, y).$$

Die sehr wichtige **Interpretation** dieser Gleichung ist folgende: B ist das Ereignis, dass gewisse Informationen (nämlich die C_i) über den Aufenthaltsort zu den Zeitpunkten vor der Zeit k vorliegen, und dass man weiß, dass man zum Zeitpunkt k genau in x ist. Die behauptete Gleichung sagt dann, dass es für die Vorhersage (bzw. die Wahrscheinlichkeit) des nächsten Schrittes egal ist, welche Information man über die fernere Vergangenheit (über die C_i) hat, sondern dass nur wichtig ist, wo man sich gerade befindet (nämlich in x). Diese sogenannte **Gedächtnislosigkeit** ist die entscheidende Eigenschaft von Markovketten.
Beweis der Gedächtnislosigkeit: Es gilt

$$\mathbb{P}^\mu(\{\boldsymbol{x} \in \Omega : x_{k+1} = y\} \cap B) = \sum_{x_0 \in E} \mu(x_0) \sum_{x_1 \in C_1} \cdots \sum_{x_{k-1} \in C_{k-1}} p(x_0, x_1) \cdots p(x_{k-2}, x_{k-1}) p(x_{k-1}, x) p(x, y),$$

und

$$\mathbb{P}^\mu(B) = \sum_{x_0 \in E} \mu(x_0) \sum_{x_1 \in C_1} \cdots \sum_{x_{k-1} \in C_{k-1}} p(x_0, x_1) \cdots p(x_{k-2}, x_{k-1}) p(x_{k-1}, x).$$

Da x fest ist und darüber also nicht summiert wird, kann man in dem Bruch, der in der Definition der bedingen W-keit auftaucht, einfach fast alles kürzen und erhält die Behauptung. □

(1.18) Proposition und Definition

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum, $B \in \mathcal{F}$ mit $\mathbb{P}(B) > 0$.
a) Die Abbildung

$$\mathbb{P}_B : \mathcal{F} \to [0, 1], \qquad A \mapsto \mathbb{P}(A|B)$$

ist ein W-Maß auf $(\Omega, \mathcal{F})$ und heißt **bedingtes W-Maß unter der Bedingung** B.

b) Ist umgekehrt $\tilde{\mathbb{P}}$ irgendein W-Maß auf $(\Omega, \mathcal{F})$ mit den Eigenschaften
(i) $\tilde{\mathbb{P}}(B) = 1$,
(ii) Es existiert ein $c > 0$ so, dass für alle $A \in \mathcal{F}$ mit $A \subset B$ gilt: $\tilde{\mathbb{P}}(A) = c\mathbb{P}(A)$.

Dann ist $\tilde{\mathbb{P}} = \mathbb{P}_B$.
Beweis: a) ist klar. Zu b) sei $C \in \mathcal{F}$ beliebig gewählt, wir müssen dann zeigen dass $\tilde{\mathbb{P}}(C) = \mathbb{P}(C|B)$. Hierzu zerlegen wir $C = (C \cap B) \dot\cup (C \cap B^c)$ und finden zunächst $\tilde{\mathbb{P}}(C \cap B^c) \leqslant \tilde{\mathbb{P}}(B^c) = 0$ wegen (1.11 d) und Eigenschaft (i) von $\tilde{\mathbb{P}}$. Wegen der Additivität (1.8 b) und Eigenschaft (ii) von $\tilde{\mathbb{P}}$ bedeutet das

$$\tilde{\mathbb{P}}(C) = \tilde{\mathbb{P}}(C \cap B) = c\mathbb{P}(C \cap B),$$

und durch die Wahl von $C = B$ sieht man, dass $c = 1/\mathbb{P}(B)$ sein muss. Setzt man dies oben ein, erhält man direkt die Behauptung. □
Es folgen die wichtigsten Rechenregeln für bedingte W-keiten:

(1.19) Satz

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum.

a) Sei $N \leqslant \infty$ und $(B_i)_{i \leqslant N}$ eine **Partition** von Ω, das heißt $B_i \cap B_j = \emptyset$ falls $i \neq j$, und $\bigcup_{i=1}^N B_i = \Omega$. Außerdem seien alle $B_i \in \mathcal{F}$ und es gelte $\mathbb{P}(B_i) > 0$ für alle i. Dann gilt für alle $A \in \mathcal{F}$:

$$\mathbb{P}(A) = \sum_{i=1}^N \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i) \qquad \text{(Fallunterscheidungsformel)}.$$

b) In der gleichen Situation wie in a) gilt für alle $A \in \mathcal{F}$ mit $\mathbb{P}(A) > 0$ und alle $k < N + 1$:

$$\mathbb{P}(B_k \,|\, A) = \mathbb{P}(A \,|\, B_k)\frac{\mathbb{P}(B_k)}{\sum_{i=1}^N \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)} \qquad \text{(Bayes'sche Formel)}.$$

c) Für $n < \infty$ und $A_1, \ldots A_n \in \mathcal{F}$ mit $\mathbb{P}(A_1 \cap \ldots \cap A_n) > 0$ gilt die Produktformel

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \,|\, A_1)\mathbb{P}(A_3 \,|\, A_1 \cap A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cap \ldots \cap A_{n-1}).$$

Beweis: a) gilt wegen

$$\sum_{i=1}^N \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i) = \sum_{i=1}^N \mathbb{P}(A \cap B_i) = \mathbb{P}\Big(\dot{\bigcup}_{i \leqslant N}(A \cap B_i)\Big) = \mathbb{P}\Big(A \cap \bigcup_{i \leqslant N} B_i\Big) = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A).$$

b) gilt wegen

$$\mathbb{P}(A \,|\, B_k)\frac{\mathbb{P}(B_k)}{\sum_{i=1}^N \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)} \overset{a)}{=} \frac{\mathbb{P}(A \,|\, B_k)\mathbb{P}(B_k)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(A)} = \mathbb{P}(B_k \,|\, A).$$

Bei c) ist die rechte Seite gleich

$$\mathbb{P}(A_1)\frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)}\frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)} \cdots \frac{\mathbb{P}(A_1 \cap \ldots \cap A_n)}{\mathbb{P}(A_1 \cap \ldots \cap A_{n-1})}.$$

Kürzt man nun schräg von links oben nach rechts unten jeweils die gleichen Terme heraus, so bleibt genau die linke Seite übrig. □

Zufallsvariable und Bildmaße

(1.20) Motivation und Vorüberlegungen

Im Prinzip könnte man die gesamte W-Theorie nur mit Hilfe von W-Maßen und σ-Algebren beschreiben. Das ist aber nicht sehr elegant und praktisch, wie wir bereits gesehen haben:

a) Zum Beispiel mussten wir in Beispiel (1.17) mühsam die Menge B einführen, um zu beschreiben, was unsere Markovkette in der Vergangenheit gemacht hat. Generell ist es auf Dauer sehr unpraktisch, alles was man modellieren will explizit mit Teilmengen zu beschreiben.

b) Will man mehrere verschiedene Aspekte eines Geschehens modellieren, wird es noch mühsamer. Z.B. könnte man im Beispiel (1.13) für jedes $k \leqslant N$ das (zufällige) Vermögen zum Zeitpunkt k betrachten wollen, und gleichzeitig den höchsten Stand, den das Vermögen bis zum Zeitpunkt k jemals hatte. Zwar kann man das alles mit Teilmengen ausdrücken, aber es macht keinen Spaß.

c) Man möchte auch verschiedene Aspekte eines Geschehens kombinieren können. In Beispiel (1.13) könnte man zum Beispiel fragen, wie wahrscheinlich es ist, dass die Summe aller Vermögensstände, die man im gesamten Spiel hatte, einen gewissen Wert übersteigt. Man kann dies wieder von Hand mit Teilmengen modellieren, aber das muss man von Fall zu Fall einzeln machen und es ist mühsam.

Aus diesem Grund gibt es das Konzept der Zufallsvariable, das wir nun einführen wollen.

(1.21) Definition

Ω, Ω' seien Mengen, $X : \Omega \to \Omega'$ eine Abbildung. Die Abbildung

$$X^{-1} : \mathcal{P}(\Omega') \to \mathcal{P}(\Omega) : \quad A \mapsto X^{-1}(A) := \{\omega \in \Omega : X(\omega) \in A\}$$

heißt die von X induzierte **Urbildabbildung**.

Beispiele:

a) $\Omega = \Omega' = \mathbb{R}$, $X(\omega) = \omega^2$. Dann ist beispielsweise

$$X^{-1}(\{\omega'\}) = \{\sqrt{\omega'}, -\sqrt{\omega'}\} \qquad \text{falls } \omega' \geqslant 0 \quad \text{ und } X^{-1}(\{\omega'\}) = \emptyset \text{ sonst.}$$

Was ist $X^{-1}([2, \infty))$?

b) $\Omega = \Omega' = \mathbb{R}$, $X(\omega) = \mathbb{1}_{[0,\infty)}(\omega)$. Dann ist

$$X^{-1}(A) \in \{\emptyset, \Omega, (-\infty, 0), [0, \infty)\} \qquad \forall\ A \subset \mathbb{R}.$$

c) $\Omega = [0, 1)$, $\Omega' = \mathbb{R}$, $N \in \mathbb{N}$, $X(\omega) = \frac{k}{N}$ für $\frac{k-1}{N} \leqslant \omega < \frac{k}{N}$ und $0 \leqslant k < N$, k ganzzahlig. Dann ist

$$\{X^{-1}(A) : A \subset \mathbb{R}\} = \Big\{ \bigcup_{j=1}^{k} B_j : B_j \in \{[\tfrac{k-1}{N}, \tfrac{k}{N}) : k \leqslant N\}\ \forall j,\ \ 0 \leqslant k < N \Big\}$$

Wie sieht $X^{-1}([c, d])$ aus für $0 < c < d < 1$? Was ändert sich, wenn man $X(\omega) = \cos(\pi k/2)$ für $k - 1 \leqslant \omega \leqslant k$ wählt?

d) $\Omega = \{(i, j) : i, j \in \{1, \ldots, 6\}\}$, $\Omega' = \mathbb{N}$ und $X(i, j) = i + j$ (Würfelsumme). Dann ist z.B.

$$X^{-1}(\{8\}) = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}, \text{ und } X^{-1}(\{1\}) = \emptyset.$$

e) Markovkette: E sei Menge, $N \in \mathbb{N}$, $\Omega = E^{N+1}$, $\Omega' = E$. Schreibe $\omega \in \Omega$ als $\omega = (\omega_0, \ldots, \omega_N)$ mit $\omega_i \in E$ für alle i, und definiere $X_i(\omega) := \omega_i$, also die Position der Markovkette im i-ten Schritt. Für $C \subset E$ ist dann

$$X_i^{-1}(C) = E \times \ldots E \times \underbrace{C}_{i\text{-te Stelle}} \times E \times \ldots \times E \equiv \{\omega \in \Omega : \omega_i \in C\}.$$

(1.22) Definition

$(\Omega, \mathcal{F})$ und $(\Omega', \mathcal{F}')$ seien messbare Räume. Eine Abbildung $X : \Omega \mapsto \Omega'$ heißt **messbar** (oder $\mathcal{F}$-$\mathcal{F}'$-messbar), wenn gilt:

$$\forall A \in \mathcal{F}' \text{ ist } X^{-1}(A) \in \mathcal{F}.$$

Die obige Bedingung kann man auch kurz mit $X^{-1}(\mathcal{F}') \subset \mathcal{F}$ schreiben. In vielen Fällen (insbes. oft wenn $\Omega' = \mathbb{R}$, siehe später) ist klar, was $\mathcal{F}'$ ist. Dann schreibt man auch $X \in \mathrm{m}\mathcal{F}$, wenn X $\mathcal{F}$-messbar ist.

Bemerkungen:

a) Messbarkeit ist also eigentlich eine Eigenschaft von X^{-1} anstatt von X, aber andererseits wird X^{-1} durch X ja auch eindeutig bestimmt.

b) Messbarkeit ist um so schwieriger zu erfüllen, je größer (d.h. mehr Mengen!) $\mathcal{F}'$ ist und je kleiner $\mathcal{F}$ ist. Umgekehrt ist jede Abbildung immer $\mathcal{P}(\Omega)$-$\mathcal{F}'$ und $\mathcal{F}$-$\{\emptyset, \Omega\}$-messbar.

c) **Messbarkeit und Codierung von Information**: Eine für die W-Theorie sehr wichtige Interpretation der Messbarkeit ist folgende: Für ein $\mathcal{F}$-$\mathcal{F}'$-messbares X kann man alle durch $\mathcal{F}'$ erlaubten Antworten auf Ja-Nein-Fragen auch bezogen auf X beantworten. (siehe insbes. Beispiele unten!)

Beispiele:

a) Die Abbildung $X : \mathbb{R} \to \mathbb{R}, \omega \mapsto \omega^2$ ist $\mathcal{P}(\mathbb{R})$-$\mathcal{P}(\mathbb{R})$-messbar (klar). Sie ist aber auch $\mathcal{F}$-$\mathcal{P}(\Omega)$-messbar wenn

$$\mathcal{F} = \{A \subset \mathbb{R} : \forall x \in A \text{ ist auch } -x \in A\}.$$

Dagegen ist die Abbildung $Y(\omega) = \omega^3$ nicht $\mathcal{F}$-$\mathcal{P}(\Omega)$-messbar (warum?). Wie sieht es mit $Z(\omega) = \omega^4$ aus?

b) Die Abbildung $X : \mathbb{R} \to \mathbb{R}$, $\omega \mapsto \mathbb{1}_{[0,\infty)}(\omega)$ ist $\{\emptyset, \Omega, (-\infty, 0), [0, \infty)\}$-$\mathcal{P}(\mathbb{R})$-messbar.

c) Im Beispiel c) zu (1.21) ist die Abbildung $Y : [0, 1) \to \mathbb{R}, \omega \mapsto \omega^2$ nicht $\mathcal{F}$-$\mathcal{P}(\mathbb{R})$-messbar für $\mathcal{F} := \left\{ \bigcup_{j=1}^{k} B_j : B_j \in \{[\frac{k-1}{N}, \frac{k}{N}) : k \leqslant N\} \; \forall j, \; 0 \leqslant k < N \right\}$. Denn beispielsweise die Frage „ist ω' kleiner als $\frac{1}{4N^2}$?“ ist in $\mathcal{P}(\mathbb{R})$ erlaubt, die zugehörige Menge ist $(-\infty, \frac{1}{4N^2})$. Dagegen ist die Frage „wurde ω so gewählt, dass $X(\omega)$ kleiner als $\frac{1}{4N^2}$ ist?“ nicht in $\mathcal{F}$ erlaubt, denn sie bedeutet „ist $\omega \in [0, \frac{1}{2N})$:?“, und diese Menge ist nicht in $\mathcal{F}$.

d) Aufgabe: finde eine (möglichst kleine) σ-Algebra $\mathcal{F}$, so dass die Abbildung „Würfelsumme“ aus Beispiel (1.21 d) gerade noch messbar ist.

e) Im Beispiel (1.21 e) ist

$$\mathcal{F} = \{X_i^{-1}(C) : C \in E\} = \{E \times \ldots E \times \underbrace{C}_{i\text{-te Stelle}} \times E \times \ldots \times E : C \in E\}$$

die kleinste σ-Algebra, so dass $X_i \in \mathrm{m}\mathcal{F}$.

Bemerkung: Aus der Analysis kennt man die Umkehrabbildung f^{-1} einer bijektiven Abbildung $f : \mathbb{R}^d \to \mathbb{R}^d$, also die eindeutige Abbildung mit $f^{-1}(f(x)) = f(f^{-1}(x)) = x$. Leider benutzt die Urbildabbildung die gleiche Notation, bitte nicht verwechseln! Die Urbildabbildung ist aber viel angenehmer als die Umkerhabbildung: sie exisitert immer, auch wenn f nicht bijektiv ist, und auch wenn Start- und Zielraum der Abbildung f verschieden sind. Falls f doch bijektiv ist, dann gilt zusätzlich

$$f^{-1}_{\text{Urbild}}(\{y\}) = \{f^{-1}_{\text{Umkehr}}(y)\}.$$

Mit anderen Worten: ist f bijektiv, so ist das Ergebnis der Umkehrabbildung bei Anwendung auf y das einzige Element der nichtleeren, einelementigen Menge $f^{-1}_{\text{Urbild}}(\{y\})$. Ein weiterer Vorzug der Urbildabbildung ist, dass sie sich perfekt mit allen Mengen-Operationen verträgt:

(1.23) Proposition

Die Urbildabbildung $X^{-1} : \mathcal{F}' \to \mathcal{F}$ vertauscht mit allen Mengenoperationen: es gilt für $A \in \mathcal{F}'$ und beliebige (auch überabzählbare) Familien $(A_i)_{i\in I}$:

$$\big(X^{-1}(A)\big)^c = X^{-1}(A^c), \quad X^{-1}\Big(\bigcup_{i\in I} A_i\Big) = \bigcup_{i\in I} X^{-1}(A_i), \quad X^{-1}\Big(\bigcap_{i\in I} A_i\Big) = \bigcap_{i\in I} X^{-1}(A_i).$$

Beweis: Beispielhaft nur eine Aussage:

$$\omega \in X^{-1}\Big(\bigcup_{i\in I} A_i\Big) \Leftrightarrow X(\omega) \in \bigcup_{i\in I} A_i \Leftrightarrow \exists i \in I \text{ mit } \omega \in X^{-1}(A_i) \Leftrightarrow \omega \in \bigcup_{i\in I} X^{-1}(A_i). \quad \square$$

(1.24) Korollar und Definition

Ω sei eine Menge, $(\Omega', \mathcal{F}')$ ein messbarer Raum, $X : \Omega \to \Omega'$ eine Abbildung. Das Mengensystem

$$\sigma(X) := \{X^{-1}(A) : A \in \mathcal{F}'\}$$

ist eine σ-Algebra über Ω und heißt **die von X erzeugte σ-Algebra**.

(1.25) Definition und Satz

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum und $(\Omega', \mathcal{F}')$ ein messbarer Raum.

a) Eine $\mathcal{F}$-$\mathcal{F}'$-messbare Abbildung X heißt Zufallsvariable (kurz: ZV) mit Werten in Ω'.

b) Die Abbildung

$$\mathbb{P}_X : \mathcal{F}' \to [0,1], \qquad A' \mapsto \mathbb{P}_X(A') := \mathbb{P}(X^{-1}(A'))$$

ist ein W-Maß auf $(\Omega', \mathcal{F}')$. $\mathbb{P}_X$ heißt **Bildmaß** von $\mathbb{P}$ unter X, oder **Verteilung** von X. Oft schreibt man $\mathbb{P}(X \in A')$ statt $\mathbb{P}_X(A')$.

Beweis, dass $\mathbb{P}_X$ ein W-Maß ist:

(i): Weil X messbar ist, ist $X^{-1}(A') \in \mathcal{F}$ für alle $A' \in \mathcal{F}'$, somit darf man $\mathbb{P}$ auf $X^{-1}(A')$ anwenden und die gegebene Definition ist zulässig.

(ii): Wegen $X^{-1}(\Omega') = \Omega$ ist $\mathbb{P}_X(\Omega') = 1$.

(iii): Für eine disjunkte Familie $(A'_i)_{i\in\mathbb{N}}$ ist auch die Familie $(X^{-1}(A'_i))_{i\in\mathbb{N}}$ disjunkt: für $i \neq j$ ist

$$X^{-1}(A'_i) \cap X^{-1}(A_j)' \overset{(1.23)}{=} X^{-1}(A'_i \cap A'_j) = X^{-1}(\emptyset) = \emptyset.$$

Damit erhalten wir

$$\mathbb{P}_X\big(\bigcup_{i=1}^{\infty} A'_i\big) = \mathbb{P}\Big(X^{-1}\big(\bigcup_{i=1}^{\infty} A'_i\big)\Big) \overset{(1.23)}{=} \mathbb{P}\big(\bigcup_{i=1}^{\infty} X^{-1}(A'_i)\big) = \sum_{i=1}^{\infty} \mathbb{P}(X^{-1}(A'_i)) = \sum_{i=1}^{\infty} \mathbb{P}_X(A'_i). \quad \square$$

(1.26) Bemerkung

Ist also eine ZV „einfach nur“ eine messbare Funktion, auf deren Startraum zufällig noch ein W-Maß herumliegt? Einerseits natürlich schon, das sagt ja die Definition. Andererseits ist aber die neue Bezeichnung dadurch gerechtfertigt, dass man fast immer viel mehr am Bildmaß von $\mathbb{P}$ unter X als an X selbst interessiert ist. Dies geht oft so weit, dass man den Startraum $(\Omega, \mathcal{F}, \mathbb{P})$ der ZV gar nicht mehr erwähnt, sondern beispielsweise nur sagt „sei X eine Zufallsvariable mit Wertebereich $\{1, \dots, 6\}$ und Verteilung $\mathbb{P}(X = i) = 1/6$ für alle i“, wenn man einen Standard-Würfel modellieren will. Man kann in diesem Fall natürlich einfach $\Omega = \{1, \dots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\mathbb{P}(\{\omega\}) = 1/6$ für alle $\omega \in \Omega$ und $X(i) = i$ wählen; es gibt aber auch andere Möglichkeiten, z.B. den W-Raum für 2 mal würfeln, und dann $X = X_1$. Zwar ist es durchaus oft nützlich, einen ganz konkreten W-Raum bei der Hand zu haben, auf dem X definiert ist, aber oft braucht man den auch nicht und lässt ihn daher weg. Die folgende Definition entspricht dieser Sichtweise.

(1.27) Definition

Zwei ZVen X_1, X_2 mit Werten im gleichen messbaren Raum (ich sage auch gerne: zwei ZVen mit dem gleichen Zielraum) heißen **identisch verteilt**, falls $\mathbb{P}_{X_1} = \mathbb{P}_{X_2}$ ist.
Beachte, dass wir nicht gefordert haben, dass X_1 und X_2 den gleichen „Startraum“ haben!

(1.28) Beispiele für Bildmaße und Zufallsvariable

a) Zweimal würfeln: sei $\Omega = \{(i, j) : i, j \in \{1, \dots, 6\}\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ und $\mathbb{P}(\{\omega\}) = 1/36$ für alle ω. Betrachte die ZVen $X_1, X_2, Y, Z : (\Omega, \mathcal{P}(\Omega), \mathbb{P}) \to (\mathbb{R}, \mathcal{P}(\mathbb{R}))$ mit $X_1((i, j)) = i$ (erster Würfel), $X_2((i, j)) = j$ (zweiter Würfel), $Y = X_1 + X_2$ (Würfelsumme) und $Z = \frac{X_1}{X_2}$ (Würfelquotient). All dies sind ZVen, da ja auf den Startraum $\mathcal{P}(\Omega)$ angenommen wurde. Folgende Aussagen sollte man sich klarmachen:
(i): X_1 und X_2 sind identisch verteilt (aber nicht gleich!),
(ii): $\mathbb{P}_{X_1}(\{1\}) = \mathbb{P}_{X_1}([1/2, 3/2]) = 1/6$.
(iii): $\mathbb{P}_Y = \sum_{i=2}^{12} a_i \delta_i$, wobei $\delta_i(A) = \mathbb{1}_A(i)$ das Punktmaß bei i ist, also $\delta_i(A) = 1$ falls $i \in A$ und $= 0$ sonst; und $a_i = \mathbb{P}(X_1 + X_2 = i)$. [1)]
(iv): $\mathbb{P}_Z(\{1\}) = 1/6$, und $\mathbb{P}_Z((1, \infty)) = \mathbb{P}_Z((-\infty, 1)) = 5/12$.

b) Die einfache Irrfahrt in einer Dimension, N Schritte. Also:

$$\Omega = \{(x_0, \dots, x_N) : x_i \in \mathbb{Z}\}, \quad \mathcal{F} = \mathcal{P}(\Omega), \quad \mathbb{P}(\boldsymbol{x}) = p(0, x_1)p(x_1, x_2) \cdots p(x_{N-1}, x_N),$$

mit $p(x, y) = 1/2$ falls $|x-y| = 1$ und $p(x, y) = 0$ sonst. Betrachte für alle i die ZVen $X_i(\boldsymbol{x}) = x_i$, also die Position im i-ten Schritt. Dann ist $\mathbb{P}_{X_N}$ ein W-Maß auf $\mathbb{Z}$. Wir können nun die Frage „wie weit kommt man in N Schritten typischerweise, wenn N sehr groß wird“ aus Beispiel (1.1) präzisieren: Finde für jedes N ein (möglichst kleines) Intervall $A_N \subset \mathbb{R}$, so dass gilt

$$\lim_{N\to\infty} \mathbb{P}_{|X_N|}(A_N) \equiv \lim_{N\to\infty} \mathbb{P}(|X_N| \in A_N) = 1.$$

[1)] Da Maße ja Funktionen sind, ist klar, was die Addition bedeutet: für zwei Maße μ und ν und (positive) Zahlen c und d ist $(c\mu + d\nu)(A) := c\mu(A) + d\nu(A)$ für alle Mengen A.

Es stellt sich heraus, dass $A_N = [N^{1/2-\delta}, N^{1/2+\delta}]$ für jedes $\delta > 0$ eine gültige Wahl ist. Dies (und viel mehr) folgt aus dem zentralen Grenzwertsatz, den wir gegen Ende der Vorlesung beweisen werden.

c) Selbstvermeidende Irrfahrt, zwei Dimensionen:

$$\Omega_N = \{(x_0, x_1, \ldots, x_N) : x_0 = 0, x_i \in \mathbb{Z}^2 \text{ und } |x_i - x_{i-1}| = 1 \ \forall 1 \leqslant i \leqslant N, x_i \neq x_j \ \forall i \neq j\}.$$

$\mathcal{F} = \mathcal{P}(\Omega)$, $\mathbb{P}(\{\boldsymbol{x}\}) = \frac{1}{|\Omega|}$ für alle $\boldsymbol{x} \in \Omega$. Wieder ist X_i die Position im i-ten Schritt, und wir können wie in b) die Frage formalisieren, wie weit die selbstvermeidende Irrfahrt in N Schritten kommt. Anders als in b) ist hier eine befriedigende Antwort jedoch nicht bekannt.

d) Perkolation, siehe Beispiel (1.3). Im Moment wählen wir aus technischen Gründen noch $\Lambda \subset \mathbb{Z}^d$ endlich, setzen $\Omega = \{0,1\}^\Lambda$, $\mathcal{F} = \mathcal{P}(\Omega)$ und für $\eta = (\eta_i)_{i\in\Lambda} \in \Omega$ setzen wir

$$\mathbb{P}(\{\eta\}) \equiv \mathbb{P}_{\Lambda,p}(\{\eta\}) = p^{|\{i\in\Lambda:\eta_i=1\}|}(1-p)^{|\{i\in\Lambda:\eta_i=0\}|} = p^{\sum_{i\in\Lambda}\eta_i}(1-p)^{|\Lambda|-\sum_{i\in\Lambda}\eta_i}.$$

Die ZV $X_i : \Omega\{0,1\}$, $\eta \mapsto \eta_i$ modelliert das Kästchen an der Stelle i, und es gilt $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$ für alle i. Eine kompliziertere ZV ist

$$C_0 : \Omega \to \mathcal{P}(\Omega), \quad \eta \mapsto \max\{A \subset \Lambda : 0 \in A, \eta_i = 1 \ \forall i \in A, A \text{ zusammenhängend}\}$$

wobei letzteres bedeutet, dass es von jedem Punkt von A zu jedem anderen Punkt von A einen Weg gibt, der nur Schritte der Länge 1 macht und ganz in A verläuft. C_0 ist der sogenannte offene Cluster, der die 0 enthält. Sowohl der Zielraum als auch das Bildmaß von C_0 sind kompliziert - eine wichtige Frage der Theorie ist beispielsweise folgende: Wenn $\Lambda_N = [-N, N]^d \cap \mathbb{Z}^d$ ist, wie verhält sich dann, in Abhängigkeit von p, die Größe $\liminf_{N\to\infty} \mathbb{P}_{\Lambda_N,p}(C_0 \cap \partial\Lambda_N \neq \emptyset)$? Hier ist $\partial\Lambda_N$ der Rand von Λ_N, d.h. die Elemente, bei denen nicht mehr alle Nachbarn (in $\mathbb{Z}^d$) in Λ_N sind.

e) Allgemeine Markovketten mit N Schritten, siehe (1.14). E sei abzählbare Menge, $\Omega = E^{N+1}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $p : E \times E \mapsto [0,1]$ mit $\sum_{y\in E} p(x,y) = 1$, also $P = (p(x,y))_{x,y\in E}$ Übergangsmatrix. μ sei Maß auf E (Startmaß), und $\mathbb{P}^\mu$ sei wie in (1.14). Für $\boldsymbol{x} = (x_0, \ldots, x_N) \in \Omega$ beschreibt die Zufallsvarlable $X_i(\boldsymbol{x}) = x_i \in E$ die Position der Kette im i-ten Schritt. Es gilt für $B_1, \ldots, B_n \subset E$ und $n \leqslant N$

$$\begin{aligned}\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) &= \mathbb{P}^\mu(X_1^{-1}(B_1) \cap \ldots \cap X_n^{-1}(B_n)) = \mathbb{P}^\mu(B_1 \times \ldots \times B_n \times \Omega^{N-n}) = \\ &= \sum_{x_0\in E} \mu(x_0) \sum_{x_1\in B_1} p(x_0,x_1) \sum_{x_2\in B_2} p(x_2,x_1) \cdots \sum_{x_n\in B_n} p(x_{n-1},x_n).\end{aligned}$$

Insbesondere ist für $A \subset E$ und $i \leqslant N$

$$\begin{aligned}\mathbb{P}(X_i \in A) &= \sum_{x_0\in E} \mu(x_0) \sum_{x_1\in E} p(x_0,x_1) \sum_{x_2\in E} p(x_1,x_2) \cdots \sum_{x_{i-1}\in E} p(x_{i-2},x_{i-1}) \sum_{x_i\in A} p(x_{i-1},x_i) = \\ &= \sum_{x\in E}\sum_{y\in A} \mu(x) P^i(x,y),\end{aligned}$$

wobei P^i die i-te Potenz der Matrix P ist. Insbesondere finden wir

$$\mathbb{P}^\mu(X_i = y) = (\mu P^i)_y,$$

wobei hier die rechte Seite die y-te Komponente (genauer: die Komponente zum Index y) des Zeilenvektors μP^i bedeutet; μ wurde hierbei auch als Zeilenvektor aufgefasst, nämlich $\mu = (\mu(x))_{x\in E}$, und $\mu(x)$ ist die Gewichtsfunktion. Wir sehen also, dass wir die Verteilung von X_i bei einer Markovkette „einfach“ dadurch berechnen können, dass wir Matrixrechnung machen.

Wir sind nun in der Lage, Markovketten in der allgemein üblichen Weise zu definieren. Zunächst noch eine Vorbereitung:

(1.29) Definition

$(\Omega', \mathcal{F}')$ sei messbarer Raum, X, Y seien ZVen mit Werten in Ω' und dem gemeinsamen Start-Raum $(\Omega, \mathcal{F}, \mathbb{P})$.

(i): Für $A, B \in \mathcal{F}'$ schreiben wir $\mathbb{P}(X \in A, Y \in B) := \mathbb{P}(X^{-1}(A) \cap Y^{-1}(B))$.

(ii): Für $B \in \mathcal{F}'$ mit $\mathbb{P}(Y \in B) > 0$ heißt die Abbildung

$$\mathcal{F}' \to [0,1], \qquad A \mapsto \mathbb{P}(X \in A \,|\, Y \in B) := \frac{\mathbb{P}(X \in A, Y \in B)}{\mathbb{P}(Y \in B)}$$

bedingte Verteilung von X unter $\{Y \in B\}$.

Bemerkung: falls $X = Y$, dann ist $\mathbb{P}(X \in A \,|\, X \in B) = \mathbb{P}_X(A \,|\, B)$. Man kann auch auf die Werte mehrerer (Y_i) bedingen, dann ist

$$\mathbb{P}(X \in A \,|\, Y_i \in B_i \,\forall i) = \frac{\mathbb{P}(X^{-1}(A) \cap \bigcap_i Y^{-1}(B_i))}{\mathbb{P}\big(\bigcap_i Y^{-1}(B_i)\big)},$$

falls der Nenner strikt positiv ist.

(1.30) Definition und Satz

E sei eine abzählbare Menge.

a) Eine Funktion $p : E \times E \to [0,1]$ mit $\sum_{y\in E} p(x,y) = 1$ heißt **stochastische Matrix** indiziert mit E.

b) Sei $N \in \mathbb{N} \cup \{\infty\}$, μ ein W-Maß auf E. Eine Familie $(X_n)_{n<N+1}$ von E-wertigen ZVen (σ-Algebra auf E ist $\mathcal{P}(E)$) heißt **Markovkette** (MK) mit N Schritten, Startverteilung μ und Übergangsmatrix (ÜM) $P = (p(x,y))_{x,y\in E}$, falls für alle $n < N+1$, alle $x, y \in E$, alle $A_i \subset E$, $1 \leqslant i < n-1$ gilt:

$$\mathbb{P}(X_n = y \,|\, X_{n-1} = x, X_i \in A_i \,\forall i < n-1) = p(x,y).$$

und $\mathbb{P}(X_0 = y) = \mu(\{y\})$.

c) Ist (X_i) eine MK mit Startverteilung μ und ÜM P, so gilt für alle $n < N+1$ und alle $y \in E$: $\mathbb{P}(X_n = y) = (\mu P^n)_y$, siehe (1.28 e). Ebenso ist $\mathbb{P}(X_n = y \,|\, X_k = x) = (P^{n-k})(x,y)$ für $k \leqslant n$.

Bemerkung: a) Der Notationsmissbrauch, den gleichen Buchstaben μ sowohl für das W-Maß als auch für seine Gewichtsfunktion und für den entsprechenden Zeilenvektor zu verwenden, ist üblich und hoffentlich nicht allzu verwirrend.

b) Oben in Punkt b) hängt die rechte Seite der dargestellten Formel nicht von den A_i ab. Wählt man daher einmal alle $A_i = E$ und einmal allgemeine A_i und setzt dies gleich, so erhält man

$$\mathbb{P}(X_n = y \,|\, X_{n-1} = x, X_i \in A_i \,\forall i < n-1) = \mathbb{P}(X_n = y \,|\, X_{n-1} = x),$$

also die Gedächtnislosigkeit der Markovkette. Außerdem werden hierdurch die Begriffe Übergangsmatrix und Übergangswahrscheinlichkeit klar.

c) Der W-Raum Ω, auf dem die X_i definiert sein sollen, wurde nicht erwähnt. Da man bedingte Verteilungen betrachtet, muss es zumindest einmal für alle X_i der gleiche W-Raum sein, und für $N < \infty$ kann man ihn wie in (1.28 e) wählen: Dann ist nämlich in der Tat

$$\begin{aligned}&\mathbb{P}(X_n = y \,|\, X_{n-1} = x, X_i \in A_i \ \forall i < n-1) = \\ &= \frac{\sum_{x_0 \in E} \mu(x_0) \sum_{x_1 \in A_1} p(x_0, x_1) \dots \sum_{x_{n-1} \in E} p(x_{n-2}, x_{n-1}) p(x_{n-1}, x) p(x, y)}{\sum_{x_0 \in E} \mu(x_0) \sum_{x_1 \in A_1} p(x_0, x_1) \dots \sum_{x_{n-1} \in E} p(x_{n-2}, x_{n-1}) p(x_{n-1}, x)} = p(x, y).\end{aligned}$$

Für $N = \infty$ ist das schwieriger: $\Omega = E^{\mathbb{N}}$ ist noch relativ klar, die σ-Algebra aber wie schon früher besprochen weniger. Insbesondere erwarten wir in vielen Fällen, dass $\mathbb{P}(\{\omega\}) = 0$ für alle ω, da jedes ω ja nun ein unendlich lange Folge zufälliger Ereignisse repräsentiert. Wir werden jetzt lernen, wie man diesen Fall richtig behandelt.

(1.31) Lemma und Definition

a) Sei Ω eine Menge, I eine beliebige (also auch überabzählbare) Indexmenge und $(\mathcal{F}_i)_{i \in I}$ eine Familie von σ-Algebren über Ω. Dann ist

$$\bigcap_{i \in I} \mathcal{F}_i := \{A \subset \Omega : A \in \mathcal{F}_i \,\forall i \in I\}$$

eine σ-Algebra.

b) Sei Ω eine Menge, $\mathcal{A} \subset \mathcal{P}(\Omega)$ ein beliebiges Mengensystem. Dann ist die Schnittmenge aller σ-Algebren, die $\mathcal{A}$ vollständig enthalten, eine σ-Algebra. In Symbolen:

$$\sigma(\mathcal{A}) := \bigcap_{\mathcal{F} \subset \mathcal{P}(\Omega) : \mathcal{A} \subset \mathcal{F}, \mathcal{F} \text{ ist } \sigma\text{-Algebra}} \mathcal{F},$$

ist eine σ-Algebra. Sie heißt **die von $\mathcal{A}$ erzeugte σ-Algebra** und ist die kleinste σ-Algebra, die $\mathcal{A}$ enthält.

Beweis: a) Mit $\mathcal{F} := \bigcap_{i \in I} \mathcal{F}_i$ haben wir

$$A \in \mathcal{F} \Leftrightarrow A \in \mathcal{F}_i \,\forall i \in I \Leftrightarrow A_i^c \in \mathcal{F}_i \,\forall i \in I \Leftrightarrow A^c \in \mathcal{F}.$$

Genauso haben wir für Teilmengen $(A_j)_{j \in \mathbb{N}}$ von Ω:

$$A_j \in \mathcal{F} \,\forall j \Leftrightarrow A_j \in \mathcal{F}_i \,\forall i, j \Rightarrow \bigcup_{j=1}^{\infty} A_j \in \mathcal{F}_i \,\forall i \Leftrightarrow \bigcup_{j=1}^{\infty} A_j \in \mathcal{F}.$$

b) $\sigma(\mathcal{A}) \neq \emptyset$, denn $\mathcal{A} \subset \mathcal{P}(\Omega)$ und $\mathcal{P}(\Omega)$ ist eine σ-Algebra. $\sigma(\mathcal{A})$ ist ein σ-Algebra wegen a). $\sigma(\mathcal{A})$ ist (als Mengensystem) eine Teilmenge jeder σ-Algebra, die $\mathcal{A}$ enthält, denn es ist ja der Durchschnitt aller solchen σ-Algebren. Also ist $\sigma(\mathcal{A})$ die kleinste σ-Algebra, die $\mathcal{A}$ enthält. $\square$

(1.32) Proposition und Definition

$(E, \mathcal{E})$ sei messbarer Raum, $N \in \mathbb{N} \cup \{\infty\}$ und $\Omega = E^N = \{(x_i)_{i<N+1} : x_i \in E \ \forall i\}$.

a) Für jedes $n < N+1$ ist das Mengensystem

$$\mathcal{F}_{\{n\}} := \{E \times \ldots \times E \times \underbrace{A}_{n\text{-te Stelle}} \times E \times \ldots : A \in \mathcal{E}\} = \{\{(x_i)_{i<N+1} : x_n \in A\} : A \in \mathcal{E}\}$$

eine σ-Algebra über Ω.

b) Die σ-Algebra

$$\mathcal{F} = \sigma\Big(\bigcup_{n<N+1} \mathcal{F}_{\{n\}}\Big)$$

heißt **Produkt-σ-Algebra** zu N Faktoren von $(E, \mathcal{E})$. Wir schreiben

$$\mathcal{F} = \underbrace{\mathcal{E} \otimes \ldots \otimes \mathcal{E}}_{N \text{ Faktoren}} = \mathcal{E}^{\otimes N}.$$

$(E^N, \mathcal{E}^{\otimes N})$ heißt das **N-fache Podukt** von $(E, \mathcal{E})$ mit sich selbst. Im Fall $N = \infty$ schreibt man besser: $(E^{\mathbb{N}}, \mathcal{E}^{\otimes \mathbb{N}})$

c) Sei $A \subset \{1, \ldots, N\}$ falls $N < \infty$ bzw. $A \subset \mathbb{N}$ endlich falls $N = \infty$. Die σ-Algebra

$$\mathcal{F}_A := \sigma\Big(\bigcup_{n\in A} \mathcal{F}_{\{n\}}\Big)$$

heißt **σ-Algebra der A-Zylindermengen**. Wir schreiben $\mathcal{F}_n = \mathcal{F}_{\{1,\ldots,n\}}$ für $n < N+1$.

Bemerkungen:

a) Man kann das ganze auch machen, wenn die Räume $(E, \mathcal{E})$ nicht alle gleich sind, also für verschiedene $(E_i, \mathcal{E}_i)$, $i < N+1$. Alles geht genauso.

b) Beachte, dass für endliches $n > 1$ das Mengensystem $\mathcal{E}^n := \{A_1 \times A_2 \times \ldots \times A_n : A_i \in \mathcal{E} \ \forall i\}$ meist keine σ-Algebra ist. Es ist strikt kleiner als $\mathcal{E}^{\otimes n}$. Daher die neue Notation.

(1.33) Bemerkung

Sei E abzählbar, $\mathcal{E} = \mathcal{P}(E)$ und $n < N+1$. Dann existiert zu jedem $A \in \mathcal{F}_n$ genau ein $\tilde{A} \in \mathcal{P}(E^n)$ so dass $A = \tilde{A} \times E \times E \times \ldots$. (Beweis: Übung). Dies bedeutet, dass die σ-Algebra $\mathcal{F}_n$ diejenige Information repräsentiert, die man aus den ersten n Gliedern der Folge $(x_i)_{i<N+1} \in \Omega$ extrahieren kann.

Insbesondere kann man eine Markovkette *mit n Schritten*, mit Übergangsmatrix P und Startverteilung μ auf dem W-Raum $(E^{\mathbb{N}}, \mathcal{F}_n)$ wie folgt realisieren: definiere $\mathbb{P}_n^\mu$ mittels

$$\mathbb{P}_n^\mu(\tilde{A} \times E^{\mathbb{N}}) = \sum_{x_0 \in E} \mu(x_0) \sum_{(x_1,\ldots,x_n)\in \tilde{A}} p(x_0, x_1)p(x_1, x_2) \cdots p(x_{n-1}, x_n).$$

Dann ist nämlich mit $X_i(\boldsymbol{x}) = x_i$ die Familie von ZVen $(X_i)_{i \leqslant n}$ eine entsprechende Markovkette: es gilt beispielsweise für $A_1, \ldots, A_n \in E$:

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}_n^\mu(A_1 \times \cdots \times A_n \times E^{\mathbb{N}}) = \sum_{x_0 \in E} \mu(x_0) \sum_{x_1 \in A_1} p(x_0, x_1) \ldots \sum_{x_n \in A_n} p(x_{n-1}, x_n).$$

Wie wir in (1.28 e) gesehen haben, folgt daraus tatsächlich $\mathbb{P}(X_{k+1} = y \,|\, X_k = x) = p(x, y)$.

Für $m < n$ haben wir nun aber zwei Möglichkeiten, ein Pfadmaß für die Markovkette $(X_i)_{i \leqslant m}$ zu definieren. Einerseits können wir oben einfach n durch m ersetzen, d.h. das Maß $\mathbb{P}^\mu_m$ auf der σ-Algebra $\mathcal{F}_m$ betrachten. Andererseits können wir aber auch das vorherige Maß $\mathbb{P}^\mu_n$ weiterverwenden und in der vorigen Gleichung einfach $A_{m+1} = \ldots = A_m = E$ wählen. Es wäre besser, wenn wir mit beiden Methoden das gleiche Ergebnis erhalten würden!
Dies ist tatsächlich der Fall: für $m < n$ und $B \in \mathcal{F}_m$ ist auch $B \in \mathcal{F}_n$, nämlich $B = \tilde{B} \times E^{\mathbb{N}} = (\tilde{B} \times E^{n-m}) \times E^{\mathbb{N}}$, mit $\tilde{B} \in \mathcal{P}(E^M)$); daher ist

$$\begin{aligned}
\mathbb{P}^\mu_m(B) &= \sum_{x_0 \in E} \mu(x_0) \sum_{(x_1,\ldots,x_m)\in\tilde{B}} p(x_0,x_1)p(x_1,x_2)\cdots p(x_{m-1},x_m) = \\
&= \sum_{x_0 \in E} \mu(x_0) \sum_{(x_1,\ldots,x_m)\in\tilde{B}} p(x_0,x_1)p(x_1,x_2)\cdots p(x_{m-1},x_m) \underbrace{\sum_{x_{m+1}\in E} p(x_m,x_{m+1})}_{=1} \cdots \underbrace{\sum_{x_n\in E} p(x_{n-1},x_n)}_{=1} \\
&= \mathbb{P}^\mu_n(\tilde{B} \times E^{n-m} \times E^N) = \mathbb{P}^\mu_n(B).
\end{aligned}$$

Mit anderen Worten: $\mathbb{P}^\mu_n|_{\mathcal{F}_m} = \mathbb{P}^\mu_m$, wobei wie bei anderen Funktionen auch $\mathbb{P}_n|_{\mathcal{F}_m}$ die Einschränkung des Definitionsbereiches von $\mathbb{P}_n$ auf Elemente von $\mathcal{F}_m$ bedeutet.

Wir fassen zusammen: Wir haben gerade eben auf einem einzigen messbaren Raum $(\Omega, \mathbb{P})$ für jede Länge n ein W-Maß $\mathbb{P}^\mu_n$ konstruiert, so dass das Bildmaß von $\mathbb{P}^\mu_n$ unter der E^{n+1}-wertigen Zufallsvariablen $\boldsymbol{x} \mapsto (x_1, \ldots, x_n)$ das Pfadmaß einer Markovkette mit Startmaß μ und Übergangsmatrix P ist. Anders ausgedrückt bilden die ZVen $(X_k)_{k \leqslant n}$ mit $X_k(\boldsymbol{x}) = x_k$ eine Markovkette mit n Schritten im Sinne von Definition (1.30)
Weiter sind die $(\mathbb{P}^\mu_n)_{n\in\mathbb{N}}$ konsistent in folgendem Sinne: wenn man für $m \in \mathbb{N}$ ein „zu großes“ $n > m$ wählt, um die Markovkette mit m Schritten durch $\mathbb{P}^\mu_n$ zu modellieren, erhält man das gleiche Bildmaß unter den Abbildungen $(X_0, \ldots, X_m)$ als wenn man $\mathbb{P}^\mu_m$ genommen hätte. Das Maß $\mathbb{P}^\mu_n$ ist also das richtige Pfadmaß für alle entsprechenden Markovketten mit *höchstens* n Schritten.
Der Verdacht liegt nahe, dass man irgendwie einen Grenzwert $n \to \infty$ bilden sollte, um die Markovkette mit unentlich vielen Schritten zu modellieren. Dass dies geht, sagt uns der folgende Satz.

(1.34) Definition und Satz

Sei E abzählbar, $\mathcal{E} = \mathcal{P}(E)$ und $(\Omega, \mathcal{F}) = (E^{\mathbb{N}}, \mathcal{E}^{\otimes\mathbb{N}})$. Für jedes $n \in \mathbb{N}$ sei $\mathbb{P}_n$ ein W-Maß auf $(\Omega, \mathcal{F}_n)$, und es gelte die *Projektivitätsbedingung* (oder: *Konsistenzbedingung*)

$$\mathbb{P}_n|_{\mathcal{F}_m} = \mathbb{P}_m \qquad \text{falls } m \leqslant n.$$

Dann existiert ein eindeutiges W-Maß $\mathbb{P}$ auf $(\Omega, \mathcal{F})$ mit der Eigenschaft dass $\mathbb{P}|_{\mathcal{F}_n} = \mathbb{P}_n$ für alle $n \in \mathbb{N}$. $\mathbb{P}$ heißt der **projektive Limes** der Maße $\mathbb{P}_n$.
Den Beweis dieses Satzes (in einer viel allgemeineren Fassung) wird es erst im nächsten Semester geben. Wir geben statt dessen folgendes Korrollar:

(1.35) Korollar

E sei abzählbar. Dann existiert zu jedem Maß μ auf E und jeder mit E indizierten stochastischen Matrix eine Markovkette mit Übergangsmatrix P und Startmaß μ.

(1.36) Beispiel: Perkolation

Wir setzen $E = \{0,1\}$, $\Omega = E^{\mathbb{Z}^d}$, $\mathcal{F} = \mathcal{P}(\{0,1\})^{\otimes \mathbb{Z}^d}$. Wir schreiben $\eta = (\eta_i)_{i\in\mathbb{Z}^d}$ für die Elemente von Ω und $X_i(\eta) := \eta_i$. Wir wollen Satz 1.34 benutzen für den Beweis der Existenz eines eindeutiges W-Maßes $\mathbb{P}_p$ auf $(\Omega, \mathcal{F})$ mit der Eigenschaft, dass für alle endlichen $\Delta \subset \mathbb{Z}^d$ und alle Folgen $(\bar\eta_i)_{i\in\Delta}$ mit $\eta_i \in \{0,1\}$ für alle i gilt:

$$\mathbb{P}_p(\{\eta \in \Omega : \eta_i = \bar\eta_i \ \forall i \in \Delta\}) \equiv \mathbb{P}_p(X_i = \bar\eta_i \ \forall i \in \Delta) = p^{\sum_{i\in\Delta}\bar\eta_i}(1-p)^{|\Delta| - \sum_{i\in\Delta}\bar\eta_i} = \mathbb{P}_{\Delta,p}(\{\bar\eta\}).$$

Hier ist $\mathbb{P}_{\Delta,p}$ das in (1.28 d) definierte Maß auf der endlichen Menge $\{0,1\}^\Delta$.

Wir müssen nun die Konsistenz prüfen. Dazu prüfen wir, [2)] ob für alle endlichen Teilmengen Δ, Λ von $\mathbb{Z}^d$ mit $\Lambda \subset \Delta$ und für alle $\bar\eta \in \{0,1\}^\Lambda$ gilt:

$$\mathbb{P}_{\Delta,p}(\{\eta \in \{0,1\}^\Delta : \eta_i = \bar\eta_i \ \forall i \in \Lambda\}) = \mathbb{P}_{\Lambda,p}(\{\bar\eta\}).$$

Die Linke Seite oben ist gleich

$$\begin{aligned}
&\sum_{\eta\in\{0,1\}^\Delta : \eta_i = \bar\eta_i \forall i\in\Lambda} p^{\sum_{i\in\Delta}\eta_i}(1-p)^{|\Delta| - \sum_{i\in\Delta}\eta_i} = \\
&= \sum_{\eta\in\{0,1\}^\Delta : \eta_i = \bar\eta_i \forall i\in\Lambda} p^{\sum_{i\in\Lambda}\eta_i}(1-p)^{|\Lambda| - \sum_{i\in\Lambda}\eta_i}\ p^{\sum_{i\in\Delta\setminus\Lambda}\eta_i}(1-p)^{|\Delta\setminus\Lambda| - \sum_{i\in\Delta\setminus\Lambda}\eta_i} = \\
&= p^{\sum_{i\in\Lambda}\bar\eta_i}(1-p)^{|\Lambda| - \sum_{i\in\Lambda}\bar\eta_i} \sum_{\eta\in\{0,1\}^{\Delta\setminus\Lambda}} p^{\sum_{i\in\Delta\setminus\Lambda}\eta_i}(1-p)^{|\Delta\setminus\Lambda| - \sum_{i\in\Delta\setminus\Lambda}\eta_i} \\
&= p^{\sum_{i\in\Lambda}\bar\eta_i}(1-p)^{|\Lambda| - \sum_{i\in\Lambda}\bar\eta_i} \underbrace{\sum_{\eta\in\{0,1\}^{\Delta\setminus\Lambda}} \mathbb{P}_{\Delta\setminus\Lambda,p}(\{\eta\})}_{=1},
\end{aligned}$$

also auch gleich der rechten Seite.

Wir wollen nun mathematisch exakt definieren, was wir damit meinen, dass man von einem Punkt z nur auf schwarzen Kästchen nach ∞ gelangen kann. Hierzu schreiben wir $i \overset{\eta}{\longleftrightarrow} j$ für $i, j \in \mathbb{Z}^d$ und $\eta \in \Omega$ wenn es $k_1, \dots, k_n \in \mathbb{Z}^d$ mit $|k_m - k_{m-1}| = 1$, $k_1 = i$, $k_n = j$ und $\eta_{k_m} = 1$ für alle m gibt. Weiter definieren wir für alle $z \in \mathbb{Z}^d$ und $N \in \mathbb{N}$ die Menge („Box“) $\Lambda_N(z) = [z - N, z + N]^d \cap \mathbb{Z}^d$, und die Abbildung

$$M_z : \Omega \to \mathbb{N} \cup \{\infty\}, \quad \eta \mapsto \sup\{N \in \mathbb{N} : \exists i \in \Lambda_N(z) \setminus \Lambda_{N-1}(z) : z \overset{\eta}{\longleftrightarrow} i\}.$$

M_z ist also die Größe der größten Box um z, zu deren Rand man auf schwarzen Kästchen gerade noch kommen kann, und die Menge $\{M_z = \infty\}$ bedeutet, dass man von z aus nach ∞ entkommt.

Wir wollen zeigen, dass M_z eine ZV ist, also die $\mathcal{F}$-$\mathcal{P}(\mathbb{N}\cup\{\infty\})$- Messbarkeit. Da der Zielraum

[2)] Um uns streng an Definition (1.34) zu halten, müssten wir eigentlich zuerst eine Abzählung $(x_n)_{n\in\mathbb{N}}$ von $\mathbb{Z}^d$ wählen, $\Delta_n := \{x_1, \dots, x_n\}$ setzen, und dann die Aussage mit Λ_n statt Δ und Λ_m statt Λ für alle $n \geqslant m$ prüfen. Bitte überzeugen Sie sich, dass die im Text angegebene Vorgehensweise äquivalent dazu ist.

abzählbar ist, reicht es, $M_z^{-1}(\{n\}) \in \mathcal{F}$ für alle $n \in \mathbb{N} \cup \{\infty\}$ zu zeigen. Für $n \in \mathbb{N}$ ist $M_z^{-1}(\{n\}) \in \mathcal{F}_{\Lambda_{n+1}(z)}$, denn es gilt

$$M_z(\eta) = n \Leftrightarrow \begin{cases} z \overset{\eta}{\longleftrightarrow} i & \text{für ein } i \in \Lambda_n(z) \setminus \Lambda_{n-1}(z), \text{ und} \\ z \overset{\eta}{\not\longleftrightarrow} i & \text{für alle } i \in \Lambda_{n+1}(z) \setminus \Lambda_n(z). \end{cases}$$

Ob die rechte Seite gilt, hängt nicht von der Werten von η_i für $i \notin \Lambda_{n+1}$ ab, daher die Messbarkeit. Für $n = \infty$ benutzen wir, dass

$$M_z^{-1}(\{\infty\}) = \Omega \setminus \bigcup_{n \in \mathbb{N}} \underbrace{M_z^{-1}(\{n\})}_{\in \mathcal{F}_{\Lambda_{n+1}(z)} \subset \mathcal{F}} \in \mathcal{F}.$$

Wir wollen nun zeigen, dass die Funktion $p \mapsto \mathbb{P}_p(M_z = \infty)$ monoton wachsend in p ist. Dies scheint „offensichtlich“, ist aber nicht so einfach zu zeigen wenn man den Trick nicht kennt. In vielen ähnlich „offensichtlichen“ Situationen, wo man so einen Trick nicht hat, kann man analoge Aussagen bis heute nicht rigoros beweisen! Wir machen also die

Behauptung: Sei $0 \leqslant p' \leqslant p \leqslant 1$. Dann ist $\mathbb{P}_{p'}(M_z = \infty) \leqslant \mathbb{P}_p(M_z = \infty)$.

Beweis: Die Idee ist eine „Kopplung“ der beiden W-Maße; diese wichtige Technik der W- Theorie kann extrem mächtig und auch schwierig zu benutzen sein, hier ist sie in einem besonders einfachen Fall zu besichtigen. „Kopplung“ bedeutet, dass wir zwei ZVen $M_{z,p}$ und $M_{z,p'}$ auf einem einzigen W-Raum $(\Omega, \mathcal{F}, \mathbb{P})$ (insbes. mit einem einzigen W-Maß anstatt der zwei W-Maße $\mathbb{P}_p$ und $\mathbb{P}_{p'}$) finden, so dass gilt:
(i): $M_{z,p'}(\omega) \leqslant M_{z,p}(\omega)$ für alle $\omega \in \Omega$
(ii): $\mathbb{P} \circ M_{z,p}^{-1} = \mathbb{P}_p \circ M_z^{-1}$, und $\mathbb{P} \circ M_{z,p'}^{-1} = \mathbb{P}_{p'} \circ M_z^{-1}$. [3)]
Wenn dies gelingt, dann ist man fertig: denn wegen (i) ist $\{\omega \in \Omega : M_{z,p'}(\omega) = \infty\} \subset \{\omega \in \Omega : M_{z,p}(\omega) = \infty\}$, und damit folgt dann

$$\mathbb{P}_{p'}(M_z = \infty) = \mathbb{P}(M_{z,p'} = \infty) \leqslant \mathbb{P}(M_{z,p} = \infty) = \mathbb{P}_p(M_z = \infty).$$

Wie also findet man die richtige Kopplung? Zunächst intuitiv: statt schwarz und weiß färben wir die Kästchen nun schwarz (mit W-keit p'), grau (mit W-keit $p - p'$) und weiß (mit W-keit $1-p$). Für jedes feste „Kästchenmuster“ entspricht nun $M_{z,p'}$ der maximalen Größe der Box, aus der man auf den schwarzen Feldern entkommen kann, und $M_{z,p}$ derjenigen, die man erreicht, wenn man zusätzlich auch die grauen Felder benutzen darf. Formal geht das so:

$$E = \{0, 1/2, 1\}, \Omega = E^{\mathbb{Z}^d}, \mathcal{F} = \mathcal{P}(E)^{\otimes \mathbb{Z}^d}, \mathbb{P}(\eta_i = w) = \begin{cases} p' & \text{für } w = 1, \\ p - p' & \text{für } w = 1/2, \\ (1-p) & \text{für } w = 0. \end{cases}$$

Außerdem wieder für endliche $A \subset \mathbb{Z}^d$: $\mathbb{P}(\eta_i = w_i \; \forall i \in A) = \prod_{i \in A} \mathbb{P}(\eta_i = w_i)$. $i \overset{\eta}{\longleftrightarrow} j$ definieren wir genau wie oben, und $i \overset{\eta}{\Longleftrightarrow} j$ fast genau wie oben, aber wir ersetzen die Bedingung

[3)] Hierbei schreiben wir $\mathbb{P} \circ X^{-1}$ statt $\mathbb{P}_X$ für das Bildmaß. Dies ist eine übliche und sehr gute Schreibweise - überlegen Sie sich, dass dies sogar genau die richtige Schreibweise ist, wenn man sowohl die Urbildabildung als auch das W-Maß als Abbildungen auf Mengen auffasst.

$\eta_{k_m} = 1\ \forall m$ durch die Bedingung $\eta_{k_m} \geqslant 1/2\ \forall m$. Schließlich definieren wir

$$M_{z,p'}(\eta) = \max\{N \in \mathbb{N} : \exists i \in \Lambda_N(z) \setminus \Lambda_{N-1}(z) : z \overset{\eta}{\longleftrightarrow} i\}, \text{ und}$$
$$M_{z,p}(\eta) = \max\{N \in \mathbb{N} : \exists i \in \Lambda_N(z) \setminus \Lambda_{N-1}(z) : z \overset{\eta}{\Longleftrightarrow} i\},$$

und prüfen (i) und (ii) oben leicht nach. □
Für das nächste Beispiel brauchen wir eine stärkere Version der Markov-Eigenschaft (1.30 b), und ein wenig

(1.37) Notation und Philosophie

Es ist in der Stochastik oft üblich, das W-Maß in der Zufallsvariable „zu verstecken"; zum Beispiel schreibt man $\mathbb{P}(X \in A)$ und $\mathbb{P}(Y \in A)$ mit dem selben $\mathbb{P}$, und der Unterschied besteht in den ZVen X und Y; die (historische) Idee dahinter ist, dass sowohl X als auch auf Y (als auch alle anderen denkbaren ZVen) auf einem riesigen „Master-W-Raum" definiert sind, den man nicht zu kennen braucht da man sowieso nur an den Verteilungen der ZVen interessiert ist. Dies haben wir zum Beispiel in (1.30 c) gemacht, wo wir für eine Markovkette mit Startmaß μ und Übergangsmatrix P die Gleichung $\mathbb{P}(X_n = y) = (\mu P^n)(y)$ aufgeschrieben haben - damit diese Gleichung Sinn ergibt, muss man vorher schon gesagt haben, welches W-Maß auf dem Startraum der Kette (X_n) liegt. In der Formel steht letzteres nicht, nicht einmal als Symbol.
Diese Sichtweise ist historisch bedingt, aber schon aus prinzipiellen Gründen fragwürdig (warum will man das zentrale Objekt der Theorie nicht benennen?). Beim Studium der Markovketten stößt es aber endgültig an seine Grenzen, wie jetzt erläutert werden soll.
E sei abzählbare Menge, $p : E \times E \to [0, 1]$ sei Übergangsmatrix. Definiere zu $x \in E$ das W-Maß $\mathbb{P}^x$ als das Pfadmaß (auf $(E^{\mathbb{N}}, \mathcal{P}(E)^{\otimes\mathbb{N}})$) der in x gestarteten Markovkette mit Übergangsmatrix P, also derjenigen mit Startmaß μ_x, das $\mu_x(\{x\}) = 1$ erfüllt. In unserer obigen Sprechweise würde (1.30 c) nun also lauten: $\mathbb{P}(X_n = y) = (P^n)(x, y)$. Das ist solange ok, bis man mehrere $\mathbb{P}^x$ gleichzeitig auf $(\Omega, \mathcal{F})$ betrachten will, denn in der linken Seite der vorigen Gleichung taucht x ja nicht auf! Man könnte sich damit behelfen, dass man X^x für die Markovkette mit Start in x schreibt, und weiter bei einem einzigen „Master-W-Raum" bleibt. Technisch bedeutet das, dass man für jedes $x \in E$ noch eine Kopie von $(\Omega, \mathcal{F})$ anlegen muss, und sich dann X^x richtig zurechtdefinieren muss.
Viel einfacher ist es hier jedoch, einfach die verschiedenen W-Maße auf $(\Omega, \mathcal{F})$ auch verschieden zu benennen. Wir schreiben also nun $\mathbb{P}^x(X_n = y) = (P^n)(x, y)$ für alle x und y. Wir haben nun nur noch eine messbare Abbildung (nämlich $X_n(\omega) = \omega_n$ aber viele W-Maße. Der einzige Nachteil ist, dass wir mit dieser Modellierung Ausdrücke der Form $\mathbb{P}(X^x = y, X^w = z)$ nicht mehr darstellen können. Wenn wir aber von Anfang an wissen, dass wir solche Ausdrücke nie brauchen werden, ist das explizite Benennen der W-Maße (wie wir es ja auch schon bei de Perkolation gemacht haben) die wesentlich sauberere Lösung. Wir benutzen diese Notation auch beim folgenden Satz.

(1.38) Satz

E sei abzählbare Menge, (X_n) eine Markovkette mit Zustandsraum E und Übergangsmatrix $P = (p(x, y))_{x,y\in E}$. Für ein W-Maß μ auf E sei $\mathbb{P}^\mu$ das Pfadmaß der mit μ gestarteten Kette, und für $x \in E$ sei $\mathbb{P}^x$ das Pfadmaß der in x gestarteten Kette. Dann gilt für alle $n, k \in \mathbb{N}$, und

alle $A_1, \dots, A_k \in E$ und alle $x \in E$ mit $\mathbb{P}^\mu(X_n = x) > 0$:

$$\mathbb{P}^\mu(X_{n+1} \in A_1, \dots, X_{n+k} \in A_k \,|\, X_n = x) = \mathbb{P}^x(X_1 \in A_1, \dots, X_k \in A_k).$$

Insbesondere ist in diesem Fall für alle $x_1, \dots, x_k \in E$

$$\mathbb{P}^\mu(X_{n+1} = x_1, \dots, X_{n+k} = x_k \,|\, X_n = x) = p(x, x_1)p(x_1, x_2) \cdots p(x_{k-1}, x_k).$$

Beweis: Übung!.

(1.39) Beispiel: Ruin des Spielers

Wir betrachten die Markovkette (X_n) auf dem Zustandsraum $E = \{0, 1, \dots, M\}$, und mit Übergangsmatrix

$$p(x,y) = \begin{cases} 1 & \text{falls } x \in \{0, M\}, x = y \\ p & \text{falls } 0 < x < M, y = x+1, \\ (1-p) & \text{falls } 0 < x < M, y = x-1, \\ 0 & \text{sonst.} \end{cases}$$

Wir wollen nun (exakt) die W-keit berechnen, dass man mit dem Gewinn M nach Hause geht, wenn man mit Kapital k startet und die Gewinn-W-keit p pro Spiel hat. Dazu brauchen wir aber noch einige Vorbereitungen.

(1.40) Definition

Ein Zustand $x \in E$ einer Markovkette mit Übergangsmatrix p heißt absorbierend, wenn $p(x,x) = 1$ gilt.
Beispiel: 0 und M sind in (1.39) absorbierend.

(1.41) Definition und Proposition

$X = (X_n)$ sei Markovkette mit Zustandsraum E, $z \in E$. Dann ist die Abbildung

$$\tau_z : \Omega \to \mathbb{N} \cup \{\infty\}, \quad \omega \mapsto \inf\{n \geqslant 1 : X_n(\omega) = z\}$$

eine Zufallsvariable. Sie heißt **Trefferzeit** von z durch die Kette X.
Beweis: Wir zeigen, dass τ_x eine ZV ist. Es gilt

$$\{\tau_z = k\} = \{X_1 \neq z, \dots, X_{k-1} \neq z, X_k = z\} \in \mathcal{F}_n,$$

und $\{\tau_z = \infty\} = \Omega \setminus \bigcup_{k=1}^{\infty} \{\tau_z = k\} \in \mathcal{F}$. Dies zeigt die Behauptung. □

(1.42) Proposition

(X_n) sei Markovkette auf E, $z \in E$ absorbierend. Dann gilt

$$\mathbb{P}(\lim_{n\to\infty} X_n = z) = \mathbb{P}(\exists N \in \mathbb{N} : X_n = z \ \forall n \geqslant N) = \mathbb{P}(\tau_z < \infty).$$

Beweis: Die erste Gleichheit ist einfach: da E abzählbar ist, ist die (angenommene) Topologie auf E die diskrete. Eine Folge konvergiert daher dann und nur dann, wenn sie ab einem bestimmten Index konstant ist. Daher sind die beiden Mengen gleich, also auch ihre W'keiten.

Die zweite Gleichheit ist intuitiv klar (wenn man z trifft, bleibt man ja für immer dort sitzen), aber überraschend schwierig zu zeigen. Dies liegt daran, dass die Mengen hier eben nicht gleich sind, sondern die Gleichheit über Eigenschaften des W-Maßes (und zwar auf unendlich langen Zeithorizonten) gezeigt werden muss.
Schreibe $A = \{\exists N \in \mathbb{N} : X_n = z \ \forall n \geqslant N\}$. Für alle $N \in \mathbb{N}$ gilt

$$X_n = z \ \forall n \geqslant N \ \Rightarrow \ X_N = z \ \Rightarrow \ \tau_z \leqslant N \ \Rightarrow \ \tau_z < \infty.$$

Daher ist $\{X_n = z \ \forall n \geqslant N\} \subset \{\tau_z < \infty\}$ für alle N, und somit auch $A = \bigcup_{N\in\mathbb{N}}\{X_n = z \ \forall n \geqslant N\} \subset \{\tau_z < \infty\}$, und es folgt

$$\mathbb{P}(A) \leqslant \mathbb{P}(\{\tau_z < \infty\}).$$

Für die umgekehrte Ungleichung bemerke zunächst dass $\{\tau_z < \infty\} = \dot{\bigcup}_{n\in\mathbb{N}}\{\tau_z = n\}$, und damit (i):

$$\mathbb{P}(A) \geqslant \mathbb{P}(A \cap \{\tau_z < \infty\}) = \sum_{n\in\mathbb{N}} \mathbb{P}(A \cap \{\tau_z = n\}),$$

(ii):

$$\mathbb{P}(A \cap \{\tau_z = n\}) \geqslant \mathbb{P}(X_j = z \ \forall j \geqslant n, \tau_z = n) = \lim_{m\to\infty} \mathbb{P}(X_{n+i} = z \ \forall i \leqslant m, \tau_z = n),$$

denn die Mengenfolge in der letzten Klammer ist absteigend, und $\mathbb{P}$ ist stetig von oben.
(iii):

$$\begin{aligned}\mathbb{P}(X_{n+i} = z \ \forall i \leqslant m \,|\, \tau_z = n) &= \mathbb{P}(X_{n+1} = z, \ldots, X_{n+m} = z \,|\, X_n = z, X_j \neq z \ \forall j < n) = \\ &= \mathbb{P}(X_{n+1} = z, \ldots, X_{n+m} = z \,|\, X_n = z),\end{aligned}$$

letzte Gleichheit wegen der Markov-Eigenschaft. Wegen Satz (1.38) ist nun

$$\mathbb{P}(X_{n+1} = z, \ldots, X_{n+m} = z \,|\, X_n = z) = p(z,z)^{n-m} = 1.$$

Wir schreiben die vorletzte Gleichung um in

$$\mathbb{P}(X_{n+i} = z \ \forall i \leqslant m, \tau_z = n) = \mathbb{P}(\tau_z = n),$$

setzen dies in (ii) ein und bekommen für alle n

$$\mathbb{P}(A \cap \{\tau_z = n\}) \geqslant \mathbb{P}(\tau_z = n)$$

Schließlich setzen wir dies in (i) ein, und erhalten die Behauptung wegen $\sum_{n\in\mathbb{N}} \mathbb{P}(\tau_z = n) = \mathbb{P}(\tau_z < \infty)$. □

(1.43) Definition

$p : E \times E \to [0,1]$ sei eine stochastische Matrix. Eine Funktion $h : E \to \mathbb{R}$ heißt **harmonisch** (bezüglich p), wenn gilt

$$\sum_{y\in E} p(x,y)h(y) = h(x) \qquad \forall x \in E.$$

Wenn man von der Markovkette (X_n) mit Übergangsmatrix p spricht, so sagt man auch, h ist harmonisch für (X_n).
Bemerkung: Fasst man $P = (p(x,y))_{x,y\in E}$ als Matrix auf, so bedeutet h harmonisch, dass (die als Spaltenvektor aufgefasste Funktion) h ein Rechts-Eigenvektor von P zum Eigenwert 1 ist.

Da P stochastisch ist, gibt es also immer mindestens eine harmonische Funktion, nämlich die mit $h(x) = 1$ für alle x.

(1.44) Satz

(X_n) sei Markovkette auf E, $z \in E$ absorbierend.

a) Die Abbildung

$$h_z : E \to \mathbb{R}, \qquad x \mapsto h_z(x) := \mathbb{P}^x(\tau_z < \infty)$$

ist harmonisch, nichtnegativ und erfüllt $h_z(z) = 1$.

b) Für jede harmonische Funktion $\tilde{h}$ mit $\tilde{h}(x) \geqslant 0$ für alle $x \in E$ und $\tilde{h}(z) = 1$ gilt:

$$\tilde{h}(x) \geqslant h_z(x) \qquad \forall x \in E.$$

h_z ist also die minimale nichtnegative harmonische Funktion, die bei z den Wert 1 annimmt.

Beweis: a) Klar ist, dass h_z nichtnegativ ist, und da z absorbierend ist, gilt $\mathbb{P}^z(X_1 = z) = 1$ und daher $h_z(z) = 1 = \sum_{y\in E} p(z,y)h_z(y)$. Es bleibt zu zeigen, dass $\sum_{y\in E} p(x,y)h(y) = h(x)$ auch für $x \neq z$ gilt. Für solche x definiere

$$h_z^{(n)}(x) := \mathbb{P}^x(\tau_z = n), \qquad \text{dann ist } h_z(x) = \sum_{n\in\mathbb{N}} h_z^{(n)}(x).$$

Für $n \geqslant 2$ rechnen wir

$$\begin{aligned} h_z^{(n)}(x) &= \sum_{y\in E\setminus\{z\}} \mathbb{P}^x(X_1 = y, X_2 \neq z, \ldots, X_{n-1} \neq z, X_n = z) = \\ &= \sum_{y\in E\setminus\{z\}} \mathbb{P}^x(X_1 = y)\mathbb{P}^x(X_2 \neq z, \ldots, X_{n-1} \neq z, X_n = z \,|\, X_1 = y) \\ &\overset{(1.38)}{=} \sum_{y\in E\setminus\{z\}} p(x,y)\mathbb{P}^y(X_1 \neq z, \ldots, X_{n-2} \neq z, X_{n-1} = z) = \\ &= \sum_{y\in E\setminus\{z\}} p(x,y)h_z^{(n-1)}(y) = \sum_{y\in E} p(x,y)h_z^{(n-1)}(y) - p(x,z)h_z^{(n-1)}(z). \end{aligned}$$

Nun benutzen wir, dass $h_z^{(n-1)}(z) = 1$ falls $n = 2$ und sonst $h_z^{(n-1)}(z) = 0$, und dass $p(x,z) = h_z^{(1)}(x)$. Damit ergibt sich

$$h_z^{(1)}(x)\mathbb{1}_{\{n=2\}} + h_z^{(n)}(x) = \sum_{y\in E} p(x,y)h_z^{(n-1)}(y).$$

Dies summieren wir nun über $n \geqslant 2$, vertauschen die Summationsreihenfolge (ist erlaubt da alle Terme nichtnegativ sind) und erhalten die Behauptung.

b) Sei $\tilde{h}$ harmonisch, nichtnegativ, mit $\tilde{h}(z) = 1$. Für jedes $x \in E$ und alle $n \in \mathbb{N}$ ist dann (in Matrix-Schreibweise)

$$\tilde{h}(x) = (P\tilde{h})(x) = (P^2\tilde{h})(x) = \ldots = (P^n\tilde{h})(x).$$

Bezeichnet $P^n(x,y)$ die Einträge der Matrix P^n, dann ergibt das für alle

$$\tilde{h}(x) = \sum_{y\in E} \underbrace{P^n(x,y)}_{\geqslant 0} \underbrace{\tilde{h}(y)}_{\geqslant 0 \text{ n.V.}} \geqslant P^n(x,z) \underbrace{\tilde{h}(z)}_{=1} = \mathbb{P}^x(X_n = z),$$

und weil z absorbierend ist gilt für alle $m \in \mathbb{N}$

$$\mathbb{P}^x(X_n = z) = \mathbb{P}^x(X_n = X_{n+1} = \ldots = X_{n+m} = z) \overset{m\to\infty}{\longrightarrow} \mathbb{P}^x(X_k = z \ \forall k \geqslant n).$$

Im letzten Schritt haben wir wieder die Stetigkeit von oben von $\mathbb{P}^x$ benutzt. Insgesamt ist also

$$\tilde{h}(x) \geqslant \mathbb{P}^x(X_k = z \ \forall k \geqslant n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}^x(\exists n \in \mathbb{N} : X_k = z \ \forall k \geqslant n) \overset{(1.42)}{=} \mathbb{P}^x(\tau_z < \infty) = h_z(x).$$

Diesmal haben wir die Stetigkeit von unten von $\mathbb{P}^x$ benutzt. Der Beweis ist damit fertig. □

(1.45) Anwendung: Ruin des Spielers

Wie in (1.39) betrachte die Markovkette (X_n) auf dem Zustandsraum $E = \{0, 1, \ldots, M\}$, und mit Übergangsmatrix

$$p(x,y) = \begin{cases} 1 & \text{falls } x \in \{0, M\}, x = y \\ p & \text{falls } 0 < x < M, y = x+1, \\ (1-p) & \text{falls } 0 < x < M, y = x-1, \\ 0 & \text{sonst.} \end{cases}$$

Die Zustände 0 und M sind beide absorbierend, wir sind an der Gewinn-W'keit interessiert, also setzen wir $h(k) := \mathbb{P}^k(\lim_{n\to\infty} X_n = M) = \mathbb{P}^k(\tau_M < \infty)$. Dann ist $h(M) = 1$, und h ist der minimale nichtnegative Rechts-Eigenvektor von p mit dieser Eigenschaft. Da auch $h(0) = 0$ gilt, kann h nicht der konstante Rechts-Eigenvektor $(1, 1, 1 \ldots)^t$ sein. Das Bestimmen eines geeigneten h geht im Prinzip mit Linearer Algebra, kann aber beliebig unangenehm werden, wenn man nicht den folgenden kleinen Trick sieht: Für alle $x \in \{1, \ldots, M-1\}$ gilt

$$\begin{aligned}(1-p)h(x) + ph(x) = h(x) = \sum_{y\in E} p(x,y)h(y) = \\ = p(x,x-1)h(x-1) + p(x,x+1)h(x+1) = (1-p)h(x-1) + ph(x+1).\end{aligned}$$

Dies stellt man um und erhält

$$h(x+1) - h(x) = \frac{1-p}{p}(h(x) - h(x-1)) = \left(\frac{1-p}{p}\right)^2 (h(x-1) - h(x-2)) = \ldots = \left(\frac{1-p}{p}\right)^x (h(1) - h(0)).$$

Wir wissen schon dass $h(0) = 0$, und schreiben $h(n)$ als Teleskopsumme:

$$h(n) = \sum_{x=1}^{n-1} (h(x+1) - h(x)) + h(1) = \sum_{x=1}^{n-1} \left(\frac{1-p}{p}\right)^x h(1) + h(1) = h(1) \frac{1 - \left(\frac{1-p}{p}\right)^n}{1 - \frac{1-p}{p}}.$$

In der letzten Gleichheit haben wir die endliche geometrische Reihenformel benutzt, die aber nur für $p \neq 1/2$ gilt. Für $p = 1/2$ steht auf der rechten Seite einfach $nh(1)$. Die Bedingung

$h(M) = 1$ erzwingt nun den Wert von $h(1)$, nämlich

$$h(1) = \begin{cases} \frac{1-\frac{1-p}{p}}{1-\left(\frac{1-p}{p}\right)^M} & \text{falls } p \neq 1/2, \\ \frac{1}{M} & \text{falls } p = 1/2. \end{cases}$$

Schließlich erhalten wir

$$h(k) = \begin{cases} \frac{1-\left(\frac{1-p}{p}\right)^k}{1-\left(\frac{1-p}{p}\right)^M} & \text{falls } p \neq 1/2, \\ \frac{k}{M} & \text{falls } p = 1/2. \end{cases}$$

Da wir in unserer Rechnung gesehen haben, dass h durch die beiden Randbedingungen $h(0) = 0$ und $h(M) = 1$ eindeutig festgelegt ist, handelt es sich nach Satz (1.44) hierbei in der Tat um die Gewinnwahrscheinlichkeiten bei Start in p und Einzelspiel-Gewinn-W-Keit p.

Stochastische Unanbhängigkeit

(1.46) Definition

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum.

a) $A, B \in \mathcal{F}$ heißen (stochastisch) **unabhängig** (man schreibt $A \perp\!\!\!\perp B$), falls gilt:

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

b) Zwei Zufallsvariable X und Y in messbare Räume $(\Omega', \mathcal{F}')$ und $(\Omega'', \mathcal{F}'')$ heißen (stochastisch) **unabhängig**, wenn für alle $A' \in \mathcal{F}'$ und alle $A'' \in \mathcal{F}''$ gilt:

$$\mathbb{P}(X \in A', Y \in A'') = \mathbb{P}(X \in A')\mathbb{P}(Y \in A'').$$

Man schreibt dann $X \perp\!\!\!\perp Y$.

Bemerkungen:

(i): In der obigen Situation gilt $X \perp\!\!\!\perp Y$ genau dann, wenn $X^{-1}(A') \perp\!\!\!\perp Y^{-1}(A'')$ für alle Mengen $A' \in \mathcal{F}'$, $A'' \in \mathcal{F}''$.

(ii): Falls $A \perp\!\!\!\perp B$, dann ist

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

Die Interpretation dieser Gleichung ist, dass die Kenntnis von B keinen Einfluss darauf hat, als wie wahrscheinlich wir das Auftreten von A modellieren. Für unabhängige Zufallsvariable und einelementige Mengen gilt konkret

$$\mathbb{P}(X = x \,|\, Y = y) = \mathbb{P}(X = x).$$

(iii): Ist $\mathbb{P}(A) = 1$ oder $\mathbb{P}(A) = 0$, so ist $A \perp\!\!\!\perp B$ für alle $B \in \mathcal{F}$. (Beweis und Interpretation: Übung). Umgekehrt impliziert $A \perp\!\!\!\perp A$, dass $\mathbb{P}(A) \in \{0, 1\}$.

(iv): Unabhängigkeit ist wie die σ-Additivität eine „Rechenregel“ für W-Maße, die nicht immer gilt. Die σ-Additivität gilt nur für disjunkte Mengen, die Unabhängigkeit quasi per Definition nur für unabhängige. Gerade im Fall von Zufallsvariablen werden wir aber später sehen, wie man aus der (angenommenen) Unabhängigkeit von Zufallsvariablen auf die Unabhängigkeit anderer Zufallsvariablen schließen kann.

(v): Die Additivität $\mathbb{P}(A \dot\cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ von je zwei Mengen überträgt sich sofort (per Induktion) auf zumindest endlich viele Mengen (für abzählbar viele muss man sie aber wieder fordern...). Leider ist dies bei Unabhängigkeit nicht so. Dies ist auch der Grund für folgende etwas umständliche Definition.

(1.47) Definition

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum, I ein beliebige Indexmenge.

a) Eine Familie $(A_i)_{i\in I}$ mit $A_i \in \mathcal{F}$ für alle $i \in I$ heißt **unabhängige Familie von Mengen**, falls für alle *endlichen* Teilmengen $J \subset I$ gilt:

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

b) Eine Familie $(\mathcal{A}_i)_{i\in I}$ mit $\mathcal{A}_i \subset \mathcal{F}$ heißt **unabhängige Familie von Mengensystemen**, wenn für jede *endliche* Teilmenge $J \subset I$ und jede Wahl von genau einem $A_j \in \mathcal{A}_j$ für jedes $j \in J$ gilt:

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

c) Eine Familie $(X_i)_{i\in I}$ von ZVen mit Werten in messbaren Räumen $(\Omega_i, \mathcal{F}_i)$ heißt **unabhängige Familie von Zufallsvariablen**, wenn für jedes *endliche* $J \subset I$ und jede Wahl von genau einer Menge $A_j \in \mathcal{F}_j$ für jedes $j \in J$ gilt:

$$\mathbb{P}(X_j \in A_j \ \forall j \in J) = \prod_{j\in J} \mathbb{P}(X_j \in A_j).$$

Bemerkung: Zwischen Punkt b) und c) oben besteht der Zusammenhang

$$(X_i)_{i\in I} \text{ unabhängig} \qquad \Longleftrightarrow (\sigma(X_i))_{i\in I} = (\{X_i^{-1}(A) : A \in \mathcal{F}_i\})_{i\in I} \text{ unabhängig.}$$

b) Ein Beispiel dafür, dass die paarweise Unabhängigkeit nicht reicht, um Unabhängigkeit zu garantieren: Betrachte den W-Raum $(\Omega, \mathcal{F}, \mathbb{P})$ mit $\Omega = \{(x_1, x_2, x_3) : x_i \in \{0, 1\}\}$, $\mathcal{F} = \mathbb{P}(\Omega)$ und

$$\mathbb{P}(\{(0,0,0)\}) = \mathbb{P}(\{(0,1,1)\}) = \mathbb{P}(\{(1,0,1)\}) = \mathbb{P}(\{(1,1,0)\}) = \frac{1}{4},$$

und $\mathbb{P}(\{\omega\}) = 0$ für alle anderen $\omega \in \Omega$. Definiere $X_i(x_1, x_2, x_3) = x_i$, und prüfe nach, dass $X_i \perp\!\!\!\perp X_j$ für alle $i \neq j$ ist, aber dass (X_1, X_2, X_3) keine unabhängige Familie bilden (Übung!).

c) Leider reicht es auch im Fall von endlicher Indexmenge I nicht aus, in Definition (1.47 a) auf die Auswahl einer Teilmenge $J \subset I$ zu verzichten - es gibt Beispiele, wo die geforderte Gleichheit gilt, wenn man aus jedem $i \in I$ eine Menge nehmen muss, wo sie aber nicht gilt, wenn man welche weglassen darf (evtl. Übung). In c) dagegen darf man, falls I endlich ist, auch immer $J = I$ wählen, denn man kann ja für die j, die man nicht sehen will, einfach $A_j = \Omega_j$ wählen. Falls alle Mengensysteme in b) den Ganzraum Ω enthalten (z.B. weil sie σ-Algebren sind), dann darf man dies auch in b) so tun.

d) Die paarweise Unabhängigkeit der X_i in Punkt b) ist ein „algebraischer Zufall“. Wegen der Existenz solcher Zufälle ist die paarweise Unabhängigkeit kein sehr nützliches Konzpet (Ausnahme: Gaußmaße, kommt noch). „Echte“ Unabhängigkeit dagegen ist ein sehr robustes

und (möglicherweise überraschend) auch natürliches Konzept. Dies sieht man, wenn man sie in Beziehung zum Begriff der Produkträume setzt, den wir nun einführen wollen.

(1.48) Definition

$(E, \mathcal{E})$ sei ein messbarer Raum, $N \in \mathbb{N} \cup \{\infty\}$, $(E^N, \mathcal{E}^{\otimes N})$ der Produktraum. Für jedes $n < N+1$ sei $\mathbb{P}_n$ ein W-Maß auf $(E, \mathcal{E})$. Ein W-Maß $\mathbb{P}$ auf $(E^N, \mathcal{E}^{\otimes N})$ heißt **Produktmaß** mit den Faktoren $(\mathbb{P}_n)_{n<N+1}$, wenn für alle $n < N+1$ und alle $A_1, \ldots, A_n \in \mathcal{E}$ gilt:

$$\mathbb{P}(A_1 \times A_2 \times \ldots \times A_n \times E^{N-n}) = \prod_{i=1}^{n} \mathbb{P}_i(A_i).$$

Wir schreiben dann $\mathbb{P} = \bigotimes_{n<N+1} \mathbb{P}_n$.

Beispiele für Produktmaße:

a) Perkolation: Dort ist $\mathbb{P} = \bigotimes_{x\in\mathbb{Z}^d} \mathbb{P}_x$ mit $\mathbb{P}_x(\{1\}) = p = 1 - \mathbb{P}(\{0\})$ für alle x. Um dies exakt in die Form von Definition (1.48) zu bringen, sollte man eine Abzählung von $\mathbb{Z}^d$ wählen, man lässt das aber dann meist einfach weg.

b) Zweimal würfeln: $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$ mit $\mathbb{P}_i(\{k\}) = 1/6$ für alle $k = 1, \ldots, 6$.

c) Volumina im $\mathbb{R}^d$. Sei $\Omega = [0,1]^d \subset \mathbb{R}^d$, und sei für „geeignete“ Teilmengen (müssen wir später klären - Quader sind jedenfalls geeignet!) das W-Maß $\mathbb{P}_d$ so, dass dieser Teilmenge ihr Volumen (Rauminhalt) zugeordnet wird. Dann ist z.B.

$$\mathbb{P}_2([a,b] \times [c,d]) = (b-a)(d-c) = \mathbb{P}_1([a,b])\mathbb{P}_1([c,d]).$$

Wir vermuten also, dass $\mathbb{P}_d = \otimes_{i=1}^{d} \mathbb{P}_1 = \mathbb{P}_1^{\otimes d}$ ist, müssen aber noch die Existenz eines solchen W-Maßes klären. Das machen wir später. Schon jetzt sehen wir aber, dass das Konzept des Produktraumes (und damit der Unabhängigkeit) tatsächlich etwas sehr nahe liegendes ist, wenn man es von der richtigen Seite aus betrachtet.

Bemerkungen:

a) Definition (1.48) kann man auch mit unterschiedlichen W-Räumen $(E_i, \mathcal{E}_i, \mathbb{P}_i)$ machen. Ein Produktmaß $\mathbb{P} = \bigotimes_{i<N+1} \mathbb{P}_i$ ist dann ein W-Maß auf $(\bigtimes_{i<N+1} E_i, \bigotimes_{i<N+1} \mathcal{E}_i)$, das

$$\mathbb{P}(A_1 \times \ldots \times A_n \times \bigtimes_{i>n} E_i) = \prod_{i<N+1} \mathbb{P}_i(A_i)$$

erfüllt. Diese Allgemeinheit werden wir tatsächlich an manchen Stellen brauchen können.

b) Wir haben nicht geklärt, ob ein Produktmaß überhaupt existiert!

(i): Im Fall dass E abzählbar und $N < \infty$ ist kann man es direkt definieren: falls $\mathcal{E} = \mathcal{P}(E)$, dann setzt man einfach

$$\mathbb{P}(\{(x_1, \ldots, x_N)\}) := \prod_{i=1}^{N} \mathbb{P}_i(\{x_i\}),$$

definiert $\mathbb{P}$ auf ganz $\mathcal{F}$ mit Hilfe disjunkter Vereinigungen und prüft nach, dass die Formel aus Def. (1.48) gilt (Übung). Daraus folgt dann auch, dass $\mathbb{P}$ tatsächlich ein W-Maß ist (also Gesamtmasse 1 hat).

(ii): Falls $\mathcal{E} \neq \mathcal{P}(E)$, dann geht es fast genauso, nur nimmt man statt der Punkte die „Atome“ von $\mathcal{E}$, also der Elemente von $\mathcal{E}$ die keine nichttrivialen Teilmengen enthalten, die auch in $\mathcal{E}$ sind.

(iii): Für $N = \infty$ und abzählbares E kann man Satz (1.34) benutzen. Man mache sich insbesondere klar, welche Eigenschaften die Übergangsmatrix einer Markovkette haben muss, damit das Pfadmaß ein Produktmaß ist.

(iv): Tatsächlich existieren Produktmaße *immer*! Diese Aussage ist etwas schwieriger zu beweisen und wird vielleicht zum Ende der Vorlesung gemacht, wenn noch Zeit ist. Im Moment nehmen wir sie mal einfach so hin.

(1.49) Definition und Satz

Seien $N \in \mathbb{N} \cup \{\infty\}$, $(E_n, \mathcal{E}_n, \mathbb{P}_n)_{n<N+1}$ W-Räume, und $(\Omega, \mathcal{F}, \mathbb{P})$ der Produktraum. Dann sind die durch $X_n(\omega) = \omega_n \in E_n$ definierten Zufallsvariablen (die sogenannten **Koordinatenabbildungen**) eine unabhängige Familie. Der W-Raum $(\Omega, \mathcal{F}, \mathbb{P})$ heißt die **kanonische Darstellung** einer unabhängigen Familie von ZVen (Y_i) mit $\mathbb{P}_{Y_i} = \mathbb{P}_i$.

Beweis und Bemerkungen: Der Beweis ist fast trivial: ist J eine endliche Teilmenge der Indexmenge, und sind $A_j \in \mathcal{E}_j$ für alle $j \in J$, dann setze $M = \max J < \infty$, $\tilde{A}_j = A_j$ falls $j \in J$ und $\tilde{A}_j = E_j$ falls $j \notin J$, und erhalte

$$\begin{aligned}\mathbb{P}(X_j \in A_j\ \forall j \in J) =& \mathbb{P}(X_j \in \tilde{A}_j \forall j \leqslant M) = \mathbb{P}(\bigtimes_{n=1}^{M} \tilde{A}_j \times \bigtimes_{n>M} E_n) \overset{\text{Def Produktmaß}}{=} \\ &= \prod_{j \leqslant M} \mathbb{P}_j(\tilde{A}_j) = \prod_{j \leqslant M} \mathbb{P}_j(X_j \in \tilde{A}_j) = \prod_{j \in J} \mathbb{P}_j(X_j \in A_j).\end{aligned}$$

Das war zu zeigen. □

Die Bedeutung der kanonischen Darstellung ist folgende: die (Y_i) können ohne weiteres auf einem W-Raum Ω definiert sein, der kein Produktraum ist, und möglicherweise ist diese Darstellung sogar sehr nützlich. Wenn man es jedoch praktisch findet, dann kann man sich jederzeit eine Familie (X_i) verschaffen, die auf einem Produktraum definiert ist, und die im W-Theoretischen Sinne von der Familie der (Y_i) nicht zu unterscheiden ist: zunächst sind alle Wahrscheinlichkeiten, dass gewisse X_i gleichzeitig in gewissen Mengen A_i landen,genau die gleichen wie sie es für die entsprechenden Y_i wären. Dass dies schon reicht, um zu zeigen, dass die (X_i) und die (Y_i) wirklich die gleiche *gemeinsame Verteilung* haben, zeigen wir jetzt. Dazu brauchen wir ein paar Vorbereitungen.

Bemerkung: Sparsames Nachprüfen von Gleichheit von W-Maßen oder Unabhängigkeit von Zufallsvariablen

In vielen Bereichen der Mathematik ist es nützlich, einfacher nachzuprüfende hinreichende Kriterien dafür zu finden, dass zwei Objekte gleich sind, oder dass ein Objekt gewisse Eigenschaften hat. In der Analysis z.B. genügt es für zwei stetige Funktionen f, g nachzuprüfen, dass $f(x) = g(x)$ für alle $x \in \mathbb{Q}$ ist um zu wissen, dass $f(x) = g(x)$ für alle $x \in \mathbb{R}$ ist. In der linearen Algebra genügt es, für eine lineare Abbildung $L : V \to W$ zwischen zwei Vektorräumen die Werte Le_i für eine Basis von (e_i) von V zu kennen, um Lx für alle $x \in V$ zu kennen.

Einen ähnlichen Satz suchen wir nun für Maße und σ-Algebren. Denn um beispielsweise $X \perp\!\!\!\perp Y$ zu prüfen müssen wir laut Definition $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ für alle Mengen A, B aus den Ziel-σ-Algebren prüfen, und letztere können unangenehm viele Mengen enthalten - es wäre schön, wenn die Prüfung dieser Gleichheit auf einer kleineren Auswahl von Mengen schon ausreichen würde. Außerdem haben wir oben öfters Gleichungen der Form

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(\tilde{X} \in A, \tilde{Y} \in B)$$

für alle Mengen A, B der Ziel-σ-Algebra $\mathcal{E}$ und ZVen $X, Y, \tilde{X}, \tilde{Y}$ gesehen, und implizit (aber nicht explizit) daran gedacht, dass dann die Paare (X, Y) und $(\tilde{X}, \tilde{Y})$ stochastisch ununterscheidbar sind, d.h. die gleiche gemeinsame Verteilung haben. Damit dies stimmt, müsste allerdings $\mathbb{P}((X, Y) \in C) = \mathbb{P}((\tilde{X}, \tilde{Y}) \in C)$ für alle $C \in \mathcal{E} \otimes \mathcal{E}$ gelten, die oben dargestellte Gleichung garantiert dies aber nur für „rechteckige“ Mengen, also für solche aus $\mathcal{E} \times \mathcal{E}$. Wir brauchen ein Werkzeug, um diese Gleichheit auf die gesamte Produkt-σ-Algebra auszuweiten. Grundlage ist die folgende Definition und der dann folgende Satz.

(1.50) Definition

Ω sei eine Menge, $\mathcal{A} \subset \mathcal{P}(\Omega)$.

a) $\mathcal{A}$ heißt **π-System** (oder **durchschnittsstabiles System**, oder **$\cap$-stabiles System**), wenn gilt

$$\forall A, B \in \mathcal{A} \quad \text{ist auch} \quad A \cap B \in \mathcal{A}.$$

b) $\mathcal{A}$ heißt **λ-System** (oder **Dynkin-System**) wenn gilt:

(i): $\Omega \in \mathcal{A}$

(ii): Falls $A \in \mathcal{A}$, dann auch $A^c \in \mathcal{A}$,

(iii): Für paarweise disjunkte Mengen $A_i \in \mathcal{A}$, $i \in \mathbb{N}$ (also: $A_i \cap A_j \ \forall i \neq j$) gilt $\dot{\bigcup}_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

Bemerkung: Jede σ-Algebra ist ein π-System und ein λ-System. Ein λ-System ist „beinahe“ eine σ-Algebra, aber die Abgeschlossenheit gegenüber Bildung von Vereinigungen gilt nur für *disjunkte* Mengen. Dagegen ist ein λ-System im Allgemeinen aber kein π-System und daher auch keine σ-Algebra.

(1.51) Satz: Dynkins π-λ-Theorem

Ω sei eine Menge, $\mathcal{A}, \mathcal{B} \subset \mathcal{P}(\Omega)$. Es gelte:

(i): $\mathcal{A}$ ist ein π-System,

(ii): $\mathcal{B}$ ist ein λ-System,

(iii): $\mathcal{A} \subset \mathcal{B}$.

Dann ist sogar $\sigma(\mathcal{A}) \subset \mathcal{B}$.

Der **Beweis** ist Inhalt der Vorlesung Maß- und Integrationstheorie. Der Nutzen dieses Satzes ist erst mal nicht offensichtlich, wir werden aber gleich sehen, dass er sehr mächtig ist.

(1.52) Satz

$(\Omega, \mathcal{F})$ sei ein messbarer Raum, $\mathbb{P}$ und $\mathbb{P}'$ seien W-Maße auf $(\Omega, \mathcal{F})$. Für $\mathcal{A} \subset \mathcal{F}$ gelte:

(i): $\mathcal{A}$ ist ein π-System,

(ii): $\sigma(\mathcal{A}) = \mathcal{F}$,
(iii): $\mathbb{P}(A) = \mathbb{P}'(A)$ für alle $A \in \mathcal{A}$.
Dann ist sogar $\mathbb{P} = \mathbb{P}'$, also $\mathbb{P}(A) = \mathbb{P}'(A)$ für alle $A \in \mathcal{F}$.
Beweis: Wir definieren das Mengensystem

$$\mathcal{B} := \{B \in \mathcal{F} : \mathbb{P}(B) = \mathbb{P}'(B)\}.$$

Dann ist $\mathcal{B}$ ein λ-System, denn $\Omega \in \mathcal{B}$ wegen $\mathbb{P}(\Omega) = 1 = \mathbb{P}'(\Omega)$, $A^c \in \mathcal{B}$ für $A \in \mathcal{B}$ wegen $\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 1 - \mathbb{P}'(A) = \mathbb{P}'(A^c)$, und falls $A_i \in \mathcal{B}$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$, dann ist

$$\mathbb{P}(\dot{\bigcup}_{n\in\mathbb{N}} A_i) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i) = \sum_{i\in\mathbb{N}} \mathbb{P}'(A_i) = \mathbb{P}'(\dot{\bigcup}_{n\in\mathbb{N}} A_i),$$

und daher $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{B}$. Nach Voraussetzung ist außerdem $\mathcal{A} \subset \mathcal{B}$, und ein π-System. Deswegen ist

$$\mathcal{F} \overset{\text{n.V.}}{=} \sigma(\mathcal{A}) \overset{(1.51)}{\subset} \mathcal{B} \subset \mathcal{F},$$

und somit $\mathcal{B} = \mathcal{F}$. □

Bemerkung: Ein Mengensystem mit den Eigenschaften (1.52) (i) und (ii) heißt auch **durchschnittsstabiler Erzeuger** (oder: **$\cap$-stabiler Erzeuger**) der σ-Algebra $\mathcal{F}$. Der obige Satz sagt also aus, dass zwei W-Maße übereinstimmen, wenn sie auf einem durchschnittsstabilen Erzeuger der σ-Algebra übereinstimmen.

(1.53) Korollar

$(E, \mathcal{E})$ sei ein messbarer Raum, $N \in \mathbb{N} \cup \{\infty\}$, $(\Omega, \mathcal{F}) = (E^N, \mathcal{E}^{\otimes N})$. $\mathbb{P}$ und $\mathbb{P}'$ seien W-Maße auf $(\Omega, \mathcal{F})$, und es gelte für alle $n < N + 1$ und alle $A_1, \ldots, A_n \in \mathcal{E}$ die Gleichheit

$$\mathbb{P}(A_1 \times A_2 \times \ldots \times A_n \times E^{N-n}) = \mathbb{P}'(A_1 \times A_2 \times \ldots \times A_n \times E^{N-n}).$$

Dann ist $\mathbb{P}(B) = \mathbb{P}'(B)$ für alle $B \in \mathcal{F}$.
In Worten: W-Maße auf Produkträumen sind durch ihr Verhalten auf Zylindermengen eindeutig festgelegt.
Beweis: Das Mengensystem

$$\mathcal{Z} := \{A_1 \times A_2 \times \ldots \times A_n \times E^{N-n} : n < N + 1, A_1, \ldots, A_n \in \mathcal{E}\}$$

ist $\cap$-stabil, denn für $n, m \in \mathbb{N}$ mit oBdA $m \leqslant n$ und $A_i, B_i \in \mathcal{E}$ gilt

$$\begin{aligned}(A_1 \times A_2 \times \ldots \times A_n \times E^{N-n}) \cap (B_1 \times B_2 \times \ldots \times B_m \times E^{N-m}) = \\ = (A_1 \cap B_1) \times \ldots \times (A_m \cap B_m) \times A_{m+1} \times \ldots \times A_n \times E^{N-n} \in \mathcal{Z}.\end{aligned}$$

Nach Definition ist $\mathcal{E}^{\otimes N} = \sigma(\mathcal{Z})$, damit greift Satz (1.52). □

Wichtiger Spezialfall: Seien (X_n) und (Y_n) Markovketten mit gleichem Zustandsraum $(E, \mathcal{E})$, gleicher Übergangsmatrix p und gleichem Startmaß μ. Dann gilt wegen Satz (1.38) die Gleichheit

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \sum_{x\in E} \mu(x)p(x, x_1)p(x_1, x_2)\cdots p(x_{n-1}, x_n) = \mathbb{P}(Y_1 = x_1, \ldots, Y_n = x_n),$$

und somit nach Korollar (1.53) auch $\mathbb{P} \circ (X_i)_{i\in\mathbb{N}}^{-1} = \mathbb{P} \circ (Y_i)_{i\in\mathbb{N}}^{-1}$.
In Worten: die Bildmaße unter den Abbildungen $\omega \mapsto (X_i(\omega))_{i\in\mathbb{N}}$ und $\omega \mapsto (Y_i(\omega))_{i\in\mathbb{N}}$, sind gleich. Diese Abbildungen nehmen Werte in den E-wertigen Folgen an, und der Bildpunkt $(X_i(\omega))_{i\in\mathbb{N}}$ eines einzelnen $\omega \in \Omega$ wird auch als ein **Pfad** der Markovkette bezeichnet. Die Bildmaße sind dann genau die (eindeutig bestimmten) Pfadmaße der Markovkette mit Übergangsmatrix p und Startverteilung μ, und auf dem W-Raum $(E^{\mathbb{N}}, \mathcal{E}^{\otimes\mathbb{N}}, \mathbb{P} \circ (X_i)_{i\in\mathbb{N}}^{-1})$ sind die Koordinatenabbildungen $Z_i(\boldsymbol{x}) = x_i$ genau die kanonische Darstellung (oder Pfadraumdarstellung) dieser Markovkette.

Die Konsequenz dieser Gleichheit ist, dass alle ZV'en, die man aus den (X_i) konstruieren kann (also auch nichttriviale wie z.B. Trefferzeiten) genau die gleichen Verteilungen haben, als hätte man sie aus den (Y_i) konstruiert.

In ähnlicher Weise kann man den kanonischen W-Raum für Familien unabhängiger ZVen konstruieren. Zunächst:

(1.54) Notation und Nomenklatur

X, Y seien ZVen mit gleichem Zielraum $(E, \mathcal{E})$. Wir schreiben $X \stackrel{d}{=} Y$ und sagen „$X = Y$ in Verteilung“, wenn $\mathbb{P}_X = \mathbb{P}_Y$ gilt. „d“ kommt vom englischen Wort „distribution“ für Verteilung. Oft schreibt man auch $X \sim Y$ statt $X \stackrel{d}{=} Y$.

(1.55) Satz

Sei $N \in \mathbb{N} \cup \{\infty\}$. X_i und Y_i, $i < N+1$, seien ZV'en mit Werten in $(E_i, \mathcal{E}_i)$. Wir schreiben $X = (X_i)_{i<N+1}$ und $Y = (Y_i)_{i<N+1}$ für die Folgen-wertigen Zufallsvariablen, die durch gleichzeitiges Betrachten aller X_i bzw. aller Y_i entstehen. Es gelte

(i): $X_i \stackrel{d}{=} Y_i$ für alle $i < N + 1$,

(ii): $(X_i)_{i<N+1}$ und $(Y_i)_{i<N+1}$ seien unabhängige Familien.

Dann gilt $X \stackrel{d}{=} Y$.
Beweis: Für $J \subset \{1, 2, \ldots, N\}$ endlich und $A_i \in \mathcal{E}_i$ für alle $i \in \mathcal{J}$ ist

$$\mathbb{P}_X(\bigtimes_{j\in J} A_j \times \bigtimes_{j\notin J} E_j) = \mathbb{P}(X_j \in A_j \ \forall j \in J) \stackrel{(ii)}{=} \prod_{j\in J} \mathbb{P}(X_j \in A_j) =$$

$$\stackrel{(i)}{=} \prod_{j\in J} \mathbb{P}(Y_j \in A_j) \stackrel{(ii)}{=} \mathbb{P}(Y_j \in A_j \ \forall j \in J) = \mathbb{P}_Y(\bigtimes_{j\in J} A_j \times \bigtimes_{j\notin J} E_j).$$

Wegen (1.53) ist also $\mathbb{P}_X = \mathbb{P}_Y$. □

Insbesondere gilt im Kontext von (1.49), dass die kanonische Darstellung $(X_i)_{i\in\mathbb{N}}$ der unabhängigen ZVen $(Y_i)_{i\in\mathbb{N}}$ wirklich die gleiche gemeinsame Verteilung hat wie die (Y_i) selbst. Für alle Aussagen, die nur Wahrscheinlichkeiten betreffen (also alle W-theoretischen Aussagen), kann man die (Y_i) also durch die (X_i) ersetzen.

(1.56) Satz

a) $(Y_i)_{i\in\mathbb{N}}$ seien ZVen mit Zielräumen $(E_i, \mathcal{E}_i)$. Für jedes i sei $\mathcal{G}_i$ ein $\cap$-stabiler Erzeuger von $\mathcal{E}_i$. Es gelte für alle endlichen $J \subset \mathbb{N}$ und für jede Wahl von $G_i \in \mathcal{G}_i$, dass

$$\mathbb{P}(Y_j \in G_j \ \forall j \in J) = \prod_{j\in J} \mathbb{P}(Y_j \in G_j). \qquad (*)$$

Dann gilt $(*)$ sogar für jede Wahl von $G_i \in \mathcal{E}_i$, d.h. die (Y_i) sind unabhängig. Es genügt also, die Unabhängigkeit auf $\cap$-stabilen Erzeugern der Ziel-σ-Algebren zu prüfen.

b) Formulierung mit σ-Algebren: Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein W-Raum, und seien $(\mathcal{F}_i)_{i\in\mathbb{N}}$ σ-Algebren mit $\mathcal{F}_i \subset \mathcal{F}$ für alle i. Für jedes i sei $\mathcal{G}_i$ ein $\cap$-stabiler Erzeuger von $\mathcal{F}_i$. Falls die Mengensysteme $(\mathcal{G}_i)_{i\in\mathbb{N}}$ unabhängig sind, so sind auch die Mengensysteme $(\mathcal{F}_i)_{i\in\mathbb{N}}$ unabhängig.

Beweis: Der Beweis wird mit Induktion geführt. Wir beweisen nur a), der Beweis von b) ist völlig analog und verbleibt als Übung.

Für J endlich und $F_i \in \mathcal{E}_i$ sei n die Anzahl der Indizes, für die F_i nicht auch in $\mathcal{G}_i$ liegt, also

$$n := |\{i \in J : F_i \notin \mathcal{G}_i\}|.$$

Für $n = 0$ gilt $(*)$ nach Annahme. Nun nehmen wir an, $(*)$ gelte schon für alle $m \leqslant n$ für ein $n \geqslant 0$. Falls nun $n+1$ der Mengen F_i nicht in $\mathcal{G}_i$ liegen, dann wähle ein k mit $G_k \notin \mathcal{G}_i$ und definiere

$$A = \bigcap_{j\in J\setminus\{k\}} Y_j^{-1}(F_j).$$

Nach Induktionsvoraussetzung ist

$$\mathbb{P}(A) = \prod_{j\in J\setminus\{k\}} \mathbb{P}(Y_j \in F_j).$$

Falls nun $\mathbb{P}(A) = 0$, dann gilt

$$\mathbb{P}(A \cap Y_k^{-1}(F_k)) = 0 = \prod_{j\in J\setminus\{k\}} \mathbb{P}(Y_j \in F_j)\mathbb{P}(Y_k \in F_k),$$

also gilt $(*)$. Falls $\mathbb{P}(A) > 0$, dann kann man das bedingte W-Maß

$$\mathbb{P}_k : \mathcal{E}_k \to [0,1], \qquad B \mapsto \mathbb{P}(Y_k^{-1}(B)|A) = \frac{\mathbb{P}(Y_k^{-1}(B) \cap A)}{\mathbb{P}(A)}$$

definieren. Wir definieren außerdem

$$\tilde{\mathbb{P}}_k(B) := \mathbb{P}(Y_k \in B).$$

Für $B \in \mathcal{G}_k$ gilt nun (nach Induktionsvoraussetzung)

$$\mathbb{P}_k(B) = \frac{\mathbb{P}(Y_k \in B)\prod_{j\in J\setminus\{k\}} \mathbb{P}(Y_j \in F_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(Y_k \in B)\mathbb{P}(A)}{\mathbb{P}(A)} = \tilde{\mathbb{P}}_k(B).$$

Satz (1.52) liefert nun $\mathbb{P}_k(F_k) = \tilde{\mathbb{P}}_k(F_k)$ für alle $F_k \in \mathcal{F}_k$, mit anderen Worten

$$\mathbb{P}(Y_j \in F_j \ \forall j \in J) = \mathbb{P}_k(F_k)\mathbb{P}(A) = \tilde{\mathbb{P}}_k(F_k)\mathbb{P}(A) = \prod_{j\in J} \mathbb{P}(Y_j \in F_j).$$

Also gilt $(*)$ auch für $n+1$. □

Beispiel: $X_1, \ldots, X_n$ seien ZVen mit Zielraum $(E, \mathcal{P}(E))$, E sei abzählbar. Dann gilt:

$$X_1, \ldots, X_n \text{ unabhängig} \quad \Longleftrightarrow \quad \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \ \forall x_1, \ldots, x_n \in E.$$

Denn (1.56) gilt mit $\mathcal{G}_i = \{\{x\} : x \in E\} \cup \emptyset$.

(1.57) Definition und Proposition

$X_1, \ldots, X_n$ seien unabhängige ZVen mit Zielraum $(E, \mathcal{P}(E))$, E sei abzählbar. p_i sei die Gewichtsfunktion (oder: Zähldichte) von $\mathbb{P}_{X_i}$, also $\mathbb{P}(X_i = x_i) = p_i(x_i)$. Dann ist die Funktion

$$p : E^n \to [0,1], \quad (x_1, \ldots, x_n) \mapsto \prod_{i=1}^{n} p_i(x_i)$$

die Gewichtsfunktion der gemeinsamen Vertilung der (X_i), also des Bildmaßes von $\mathbb{P}$ unter dem n-Tupel $(X_1, \ldots, X_n) \in E^n$. p heißt **Produkt-Zähldichte** der Zähldichten p_i.
Den Beweis, dass p tatsächlich die Gewichtsfunktion ist, haben wir in obigem Beispiel geführt.

Beispiele:
a) Perkolation auf $\Lambda \subset \mathbb{Z}^d$, Λ endlich. Die ZVen sind $(\eta_i)_{i \in \Lambda}$ mit

$$p_i(\{1\}) = \mathbb{P}(\eta_i = 1) = p = 1 - p_i(\{0\}) = 1 - \mathbb{P}(\eta_i = 0).$$

Dann ist für $x = (x_i)_{i \in \Lambda}$ die Gewichtsfunktion gegeben durch

$$p(x) = \prod_{i \in \Lambda} p_i(x_i) = p^{\sum_{i \in \Lambda} x_i} (1-p)^{|\Lambda| - \sum_{i \in \Lambda} x_i}.$$

b) n-mal Würfeln: $p_i(\{k\}) = \mathbb{P}(X_i = k) = 1/6$ für $k = \{1, \ldots, 6\}$. Dann ist

$$p(k_1, \ldots, k_n) = \prod_{i-1}^{n} p_i(k) = (1/6)^n.$$

(1.58) Satz

Sei $N \in \mathbb{N} \cup \{\infty\}$. $(Y_i)_{i<N+1}$ seien unabhängige ZVen mit Zielräumen $(E_i, \mathcal{E}_i)$.
a) Stabilität der Unabhängigkeit unter Einsetzen in messbare Abbildungen:
$(\Omega_i, \mathcal{F}_i)$ seien messbare Räume, und $\phi_i : E_i \to \Omega_i$ messbare Abbildungen. Dann sind auch die Zufallsvariablen

$$Z_i = \phi_i \circ Y_i : \Omega \to \Omega_i, \quad \omega \mapsto \phi_i(Y_i(\omega))$$

unabhängig.
b) Stabilität unter Zusammenfassung in Blöcke:
Seien $J_1, J_2, \ldots$ disjunkte Teilmengen der Indexmenge. Für $k \in \mathbb{N}$ sei

$$W_k(\omega) = (Y_i(\omega))_{i \in J_k} \in \bigtimes_{i \in J_k} E_i.$$

Dann sind die $(W_k)_{k\in\mathbb{N}}$ eine Familie unabhängiger ZVen.
Beweis:
a) Sei J endliche Teilmenge der Indexmenge, und $A_j \in \mathcal{F}_j$ für alle $j \in J$. Dann ist

$$\begin{aligned}\mathbb{P}(Z_j \in A_j\ \forall j \in J) &= \mathbb{P}\Big(\bigcap_{j\in J} Z_j^{-1}(A_j)\Big) = \mathbb{P}\Big(\bigcap_{j\in J}(\phi_j \circ Y_j)^{-1}(A_j)\Big) = \\ &= \mathbb{P}\Big(\bigcap_{j\in J} Y_j^{-1} \circ \phi_j^{-1}(A_j)\Big) = \mathbb{P}\Big(\bigcap_{j\in J} Y_j^{-1}\big(\underbrace{\phi_j^{-1}(A_j)}_{\in\mathcal{E}_j}\big)\Big) = \\ &= \prod_{j\in J} \mathbb{P}(Y_j^{-1}(\phi_j^{-1}(A_j))) = \prod_{j\in J} \mathbb{P}(Z_j \in A_j).\end{aligned}$$

b) Wir stellen die (Y_i) auf dem Produktraum dar, also $\Omega = \bigtimes_{i<N+1} E_i$, $\mathcal{F} = \bigotimes_{i<N+1} \mathcal{E}_i$, und $\mathbb{P} = \bigotimes_{i<N+1} \mathbb{P}_{Y_i}$. Dann ist $Y_i(\omega) = \omega_i$, die Koordinatenprojektion. Weiter definieren wir

$$\Omega' = \bigtimes_{k\in\mathbb{N}} \Big(\bigtimes_{j\in J_k} \Omega_j\Big), \quad \mathcal{F}' = \bigotimes_{k\in\mathbb{N}} \Big(\bigotimes_{j\in J_k} \mathcal{E}_j\Big), \quad \text{und } \mathbb{P}' = \bigotimes_{k\in\mathbb{N}} \mathbb{P}_{W_k},$$

wobei $W_k(\omega) = \omega_{J_k} := (\omega_j)_{j\in J_k} \in \bigtimes_{j\in J_k} E_k$. Für eine endliche Teilmenge K der Indexmenge und jedes $k \in K$ betrachte nun Mengen der Form

$$B_k = \bigtimes_{j\in J_k} A_j \qquad \text{mit } A_j \in \mathcal{E}_j.$$

Für solche (speziellen) B_j gilt

$$\begin{aligned}\mathbb{P}'(W_k \in B_k\ \forall k \in K) &= \mathbb{P}'\big(\bigcap_{k\in K} W_k^{-1}(B_k)\big) = \mathbb{P}\big(\bigcap_{k\in K}\bigcap_{j\in J_k} Y_j^{-1}(A_j)\big) = \\ &= \prod_{k\in K}\prod_{j\in J_k} \mathbb{P}(Y_j^{-1}(A_j)) = \prod_{k\in K} \mathbb{P}\big(\bigcap_{j\in J_k} Y_j^{-1}(A_j)\big) = \prod_{k\in K} \mathbb{P}'(W_k \in B_k).\end{aligned}$$

Da für jedes $k \in K$ die B_k einen $\cap$-stabilen Erzeuger der σ-Algebra $\bigotimes_{j\in J_k} \mathcal{E}_j$ bilden, folgt die Behauptung nun aus Satz 1.56. □

Beispiel:
a) Perkolation: Seien $A, B \subset \mathbb{Z}^d$ mit $A \cap B = \emptyset$. Dann ist $(\eta_i)_{i\in A} \perp\!\!\!\perp (\eta_i)_{i\in B}$.
b) 3-mal Würfeln: Unabhängige ZVen X_1, X_2, X_3; dann ist z.B. $X_1X_2 \perp\!\!\!\perp X_3$. Denn $X_1X_2 = \phi(X_1, X_2)$ mit $\phi(x, y) = xy$.
c) Einfache Irrfahrt - Stopp, das ist ja eine Markovkette und keine Familie unabhängiger ZVen. Um zu sehen, dass sie trotzdem etwas mit Unabhängigkeit zu tun hat, brauchen wir:

(1.59) Definition und Satz

a) p und q seien Zähldichten auf $\mathbb{Z}^d$. Dann ist

$$p * q(x) := \sum_{y\in\mathbb{Z}^d} p(x-y)q(y) = \sum_{y\in\mathbb{Z}^d} q(x-y)p(y) = \ q * p(x)$$

eine Zähldichte auf $\mathbb{Z}^d$. $p * q$ heißt **Faltungsprodukt** der Zähldichten p und q.

b) Seien X, Y unabhängige Zufallsvariable mit Werten in $\mathbb{Z}^d$, p die Zähldichte von X und q die Zähldichte von Y. Dann hat die $\mathbb{Z}^d$-wertige Zufallsvariable $X+Y$ die Zähldichte $p*q$; in Formeln:

$$\mathbb{P}(X+Y=x)=\sum_{y\in\mathbb{Z}^d} p(x-y)q(y)=p*q(x).$$

Merksatz: Die Zähldichte der Summe unabhängiger ZVen ist die Faltung der einzelnen Zähldichten.

Beweis:
a) Es gilt

$$0\leqslant \sum_{y\in\mathbb{Z}^d} p(x-y)q(y)\leqslant \sum_{y\in\mathbb{Z}^d} q(y)=1,$$

und

$$\sum_{x\in\mathbb{Z}^d}\sum_{y\in\mathbb{Z}^d} p(x-y)q(y)=\sum_{y\in\mathbb{Z}^d}\sum_{x\in\mathbb{Z}^d} p(x-y)q(y)=\sum_{y\in\mathbb{Z}^d}\underbrace{\sum_{z+y:z\in\mathbb{Z}^d} p(z)\,q(y)}_{=1}=1.$$

Also ist $p*q$ eine Zähldichte. Mit einer Um-Indizierung der Summation ähnlich wie in der obigen Rechnung sieht man auch die behauptete Gleichheit $p*q=q*p$.
b) $X+Y$ ist tatsächlich eine ZVe (d.h. ist messbar), denn

$$\{\omega\in\Omega: X(\omega)+Y(\omega)=z\}=\bigcup_{y\in\mathbb{Z}}\{\omega\in\Omega: X(\omega)=y, Y(\omega)=z-y\}=$$
$$=\bigcup_{y\in\mathbb{Z}}\Big(X^{-1}(\{y\})\cap Y^{-1}(\{z-y\})\Big)\in\mathcal{F}.$$

Da $X^{-1}(\{y\})\cap X^{-1}(\{y'\})=\emptyset$ für $y\neq y'$ ist die obige Vereinigung außerdem disjunkt, und es folgt

$$\mathbb{P}(X+Y=x)=\sum_{y\in\mathbb{Z}^d}\mathbb{P}\big(X^{-1}(\{y\})\cap Y^{-1}(\{x-y\})\big)\overset{X\perp\!\!\!\perp Y}{=}\sum_{y\in\mathbb{Z}^d}\mathbb{P}(X=y)\mathbb{P}(Y=x-y)=$$
$$=\sum_{y\in\mathbb{Z}^d} p(y)q(x-y)=q*p(x).$$ □

Beispiele:
a) Die Intuition der einfachen Irrfahrt in einer Dimension ist, dass sie ein faires Münzwurfspiel repräsentiert, also $(X_i)_{i\in\mathbb{N}}$ unabhängige ZVen mit $\mathbb{P}(X_i=1)=\mathbb{P}(X_i=-1)=1/2$, und $S_n=\sum_{i=1}^n X_i$ als Position der einfachen Irrfahrt im n-ten Schritt. Definiert haben wir sie aber über die Übergangsmatrix als Markovkette. Wir wollen nun sehen, dass diese Definition und die mit dem fairen Münzwürfen übereinstimmen.
Die Zähldichte der X_i ist gegeben durch p mit $p(z)=1/2$ falls $z=\pm1$ und $p(z)=0$ sonst. Wir setzen $S_0=0$ und benutzen die Gleichheit $S_j-S_{j-1}=X_j$, um zu rechnen:

$$\mathbb{P}(S_1=z_1,\dots,S_n=z_n)=\mathbb{P}(S_j-S_{j-1}=z_j-z_{j-1}\ \forall 1\leqslant j\leqslant n)=$$
$$=\mathbb{P}(X_j=z_j-z_{j-1}\ \forall j\leqslant n)=\prod_{j=1}^n p(z_j-z_{j-1}).$$

Sei nun P die ÜM der Markovkette, also $P(x,y) = 1/2$ falls $|x-y| = 1$ und $= 0$ sonst. Dann ist $P(x,y) = p(y-x)$, und daher

$$\mathbb{P}(S_1 = z_1, \dots, S_n = z_n) = P(0,z_1)P(z_1,z_2), \cdots P(z_{n-1}, z_n).$$

Somit ist (S_j) die einfache Irrfahrt wegen Satz 1.38 und Korrolar 1.53.

b) Dies gilt viel allgemeiner: Sei (Y_n) eine Markovkette auf $\mathbb{Z}^d$, und die Übergangsmatrix sei *translationsinvariant*, d.h. es gebe eine Funktion $p : \mathbb{Z}^d \to [0,1]$ mit $P(x,y) = p(y-x)$ für alle $x, y \in \mathbb{Z}^d$. Dann gilt $Y_n = Y_0 + \sum_{i=1}^n X_i$, wobei die X_i unanbhängige ZVen sind, die auch unabhängig von Y_0 sind, und jedes X_i die Zähldichte $\mathbb{P}(X_n = z) = p(z)$ hat. Mit anderen Worten: Hängt in einer mit $\mathbb{Z}^d$ indizierten Übergangsmatrix der Eintrag bei (x,y) nur von der Differenz $x-y$ ab, so ist die Markovkette in Wirklichkeit eine Summe unabhängiger ZVen. Der Beweis steht wörtlich in a).

c) Zweimal würfeln: X_1, X_2 unabhängig mit Zähldichte $p(x) = 1/6$ für $x \in \{1, \dots, 6\}$ und $p(x) = 0$ sonst. Dann ist

$$\mathbb{P}(X_1 + X_2 = y) = \sum_{z \in \mathbb{Z}} p(z)p(y-z) = \frac{1}{36} \sum_{z \in \mathbb{Z}} \mathbb{1}_{\{1,\dots,6\}}(z) \mathbb{1}_{\{1,\dots,6\}}(y-z).$$

Wertet man dies aus, bekommt man die bekannte Verteilung; im Prinzip haben wir die gleiche Rechnung ja die ganze Zeit schon gemacht, wenn wir 2-mal würfeln besprochen haben. Hier ist sie eben noch mal im größeren formalen Rahmen.

Verhalten im Unendlichen

Wir haben Markovketten und unabhängige ZVen bevorzugt auf W-Räumen modelliert, deren Elemente (unendliche) Folgen sind. Wir wollen uns nun ansehen, was wir über W'keiten sagen können, die mit dem Verhalten dieser Folgen im Unendlichen zusammenhängen.

(1.60) Definition

I sei eine abzählbare (Index-) Menge $(E_i, \mathcal{E}_i)$ für jedes $i \in I$ ein messbarer Raum, $(\Omega, \mathcal{F}) = (\bigtimes_{i\in I} E_i, \bigotimes_{i \in I} \mathcal{E}_i)$ der Produktraum. In leichter Verallgemeinerung von (1.32) sei für $j \in I$ die σ-Algebra

$$\mathcal{F}_{\{j\}} := \Big\{ \{(x_i)_{i\in I} \in \bigtimes_{i \in I} E_i : x_j \in A\} : A \in \mathcal{E}_j \Big\}$$

und für (nicht notwendigerweise endliches) $J \subset I$ die σ-Algebra

$$\mathcal{F}_J = \bigotimes_{j \in J} \mathcal{F}_{\{j\}} := \sigma\Big(\bigcup_{j \in J} \mathcal{F}_{\{j\}} \Big)$$

definiert.

a) Für $J \subset I$ endlich heißt

$$\mathcal{T}_J := \mathcal{F}_{I \setminus J}$$

die **σ-Algebra der Ereignisse außerhalb von J.**

b) Die σ-Algebra

$$\mathcal{T} := \bigcap_{J\subset I: J \text{ endlich}} \mathcal{T}_J$$

heißt **terminale σ-Algebra**, oder **σ-Algebra der Ereignisse im Unendlichen**

(1.61) Bemerkung und Beispiele

a) Oft ist $I = \mathbb{N}$; dann existiert für jedes endliche $A \subset \mathbb{N}$ ein $n \in \mathbb{N}$ mit $A \subset \{1, \dots, n\}$. Da $\mathcal{T}_A \subset \mathcal{T}_B$ falls $B \subset A$ gilt, ist somit

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{T}_{\{1,\dots,n\}}$$

Wir schreiben auch $\mathcal{T}_n := \mathcal{T}_{\{1,\dots,n\}}$, gleiche Konvention wie für $\mathcal{F}_n$.

b) Um zu sehen, welche Art von Mengen in $\mathcal{T}$ sind, betrachten wir $I = \mathbb{N}$, $E_i = E$ und $\mathcal{E}_i = \mathcal{P}(E)$ für alle i, mit E abzählbar; wie immer setzen wir $\Omega = (E)^{\mathbb{N}}$, $\mathcal{F} = (\mathcal{P}(E))^{\otimes N}$.
Wie schon oben erwähnt ist es sehr hilfreich, sich Ω als Raum aller E-wertigen Folgen vorzustellen. Für $\omega = (\omega_n)_{n\in\mathbb{N}} \in \Omega$ definieren wir wie schon früher $X_n(\omega) = \omega_n$, die Koordinatenprojektion.
(i): Für $z \in E$ betrachte die Menge

$$A := \{\exists N \in \mathbb{N} : X_n = z \ \forall n > N\}$$

(vergleiche Proposition (1.42)!). Dann ist $A \in \mathcal{T}$; um dies zu sehen, setze

$$A_M := \{\exists N \geqslant M : X_n = z \ \forall n > N\} = \bigcup_{N \geqslant M} \underbrace{\bigcap_{n>N} \{X_n = z\}}_{=:B_N}.$$

Es gilt $B_N \in \mathcal{T}_N \subset \mathcal{T}_M$ für alle $N \geqslant M$, und daher ist $A_M \in \mathcal{T}_M$ für alle M. Außerdem ist (B_N) monoton wachsend in N, denn für kleinere N wird der Durchschnitt über mehr Mengen gebildet. Daher gilt

$$A = \bigcup_{N\in\mathbb{N}} B_N = \bigcup_{N \geqslant M} B_N = A_M \in \mathcal{T}_M$$

für alle M. Daher ist $A \in \mathcal{T}$.
(ii): Sei τ_z die Trefferzeit des Punktes z, siehe (1.41). Die Menge $\{\tau_z < \infty\}$ ist *nicht* in $\mathcal{T}$. Zusammen mit a) ist dies auch eine Erklärung dafür, dass wir uns im Beweis von Satz (1.42) so anstrengen mussten.
Wie sieht man $\{\tau_z < \infty\} \notin \mathcal{T}$? Dazu erinnert man sich zunächst daran, dass jede Menge $A \in \mathcal{F}_{\{2,3,\dots,n\}}$ in der Form

$$A = E \times \tilde{A} \times E^{\mathbb{N}} \qquad \text{mit } \tilde{A} \in \mathcal{E}^{\otimes(n-1)}$$

geschrieben werden kann. Eine andere Art, dies auszudrücken, ist, dass für alle $(x_i)_{i\in\mathbb{N}}$ mit $x_i \in E$ gelten muss

$$(x_1, x_2, \dots, x_n, x_{n+1}, \dots) \in A \quad \Longleftrightarrow \quad (y_1, x_2, \dots, x_n, y_{n+1}, \dots) \in A \ \forall (y_i)_{i\in\mathbb{N}}, y_i \in E.$$

Indem man hier $n \to \infty$ schickt, sieht man, dass für alle $A \in \mathcal{T}_1$ gelten muss:

$$(x_1, x_2, \dots,) \in A \quad \Longleftrightarrow \quad (y, x_2, x_3, \dots) \in A \; \forall y \in E.$$

Dies ist aber für $\{\tau_z < \infty\}$ nicht erfüllt, denn beispielsweise ist $(z, y, y, \dots) \in \{\tau_z < \infty\}$, aber $(y, y, y, \dots) \notin \{\tau_z < \infty\}$. Daher ist $\{\tau_z < \infty\} \notin \mathcal{T}_1$, und damit auch $\notin \mathcal{T}$.

c) Sei nun $E = \mathbb{Z}$, $\mathcal{E} = \mathcal{P}(\mathbb{Z})$ und $(\Omega, \mathcal{F}) = (\mathbb{Z}^{\mathbb{N}}, \mathcal{E}^{\otimes \mathbb{N}})$. Für jede Funktion $Z : \mathbb{N} \to \mathbb{R}^+$ mit $\lim_{n\to\infty} Z(n) = \infty$ und jedes Intervall $I \subset \mathbb{R}$ gilt

$$\{\omega \in \Omega : \limsup_{n\to\infty} \frac{1}{Z(n)} \sum_{i=1}^{n} X_i(\omega) \in I\} \in \mathcal{T},$$

und ebenso

$$\{\omega \in \Omega : \liminf_{n\to\infty} \frac{1}{Z(n)} \sum_{i=1}^{n} X_i(\omega) \in I\} \in \mathcal{T}.$$

Beweis: Übung!

d) Perkolation: In Beispiel (1.36) wurde die Zufallsvariable $\eta \mapsto M_z(\eta)$ definiert, die den Radius der größten Box um z darstellt, an deren Rand man gerade noch auf jeweils benachbarten schwarzen Kästchen gehen kann. Wir behaupten, dass

$$\{\exists z \in \mathbb{Z}^d : M_z = \infty\} \in \mathcal{T}.$$

Um dies zu sehen, definieren wir zunächst für endliches $A \subset \mathbb{Z}^d$ und $u \notin A$ die Zufallsvariable $\eta \mapsto M_{u,A^c}(\eta)$, die die (halbe) Seitenlänge der größten Box beschreiben soll, deren Rand man von u aus auf jeweils benachbarten schwarzen Kästchen gerade noch erreichen kann, ohne aber Kästchen in A zu benutzen (zur Übung versuche man, M_{u,A^c} exakt zu definieren analog zur Definition von M_z in (1.36)). Nun sei $A \subset \mathbb{Z}^d$ endlich, $z \in \mathbb{Z}^d$ und $M_z = \infty$. Dann existiert ein $u \in A^c$ mit $M_{u,A^c} = \infty$. Dies ist anschaulich „klar“, man sollte aber zur Übung versuchen einen rigorosen Beweis zu konstruieren. Als Konsequenz finden wir

$$\{M_z = \infty\} \subset \{\exists u \in \mathbb{Z}^d \setminus A : M_{u,A^c} = \infty\} \in \mathcal{T}_A.$$

Die letzte Aussage stimmt deshalb, weil man durch Abändern von Kästchenfarben in A die Aussage $M_{u,A^c}(\eta) = \infty$ nicht wahr machen kann, wenn sie vorher falsch war, oder umgekehrt. Nun haben wir

$$\{\exists u \in \mathbb{Z}^d \backslash A : M_{u,A^c} = \infty\} \subset \{\exists z \in \mathbb{Z}^d : M_z = \infty\} = \bigcup_{z\in\mathbb{Z}^d} \{M_z = \infty\} \subset \{\exists u \in \mathbb{Z}^d \backslash A : M_{u,A^c} = \infty\},$$

und daher

$$\{\exists u \in \mathbb{Z}^d \setminus A : M_{u,A^c} = \infty\} = \{\exists z \in \mathbb{Z}^d : M_z = \infty\}.$$

Dies zeigt, dass $\{\exists z \in \mathbb{Z}^d : M_z = \infty\} \in \mathcal{T}_A$ für alle endlichen A ist, also auch in $\mathcal{T}$.

Beachte übrigens, dass für jedes $z \in \mathbb{Z}^d$ das Ereignis $\{M_z = \infty\}$ *nicht* in $\mathcal{T}$ ist (warum?).

e) Die folgende wichtige Intuition sollte aus obigen Beispielen klar geworden sein, wird aber noch mal explizit wiederholt: Für $A \in \mathcal{F}$ ist $A \in \mathcal{T}$ genau dann, wenn es für alle $\omega \in \Omega$ unmöglich ist, durch Betrachten endlich vieler Komponenten $\omega_i \in E$ zu entscheiden, ob $\omega \in A$ ist. Eine andere Art, dies auszudrücken ist, dass für $A \in \mathcal{T}$ gilt: ist $\omega \in A$ und unterscheidet sich $\bar{\omega}$ von ω nur in endlich vielen Komponenten ($\omega_i \neq \bar{\omega}_i$ nur endlich oft), dann ist auch $\bar{\omega} \in A$.

Der rigorose Nachweis dieser Eigenschaft kann im Einzelfall technisch sein, aber ob sie gilt kann man oft nach kurzem Nachdenken entscheiden.
Wir manchen die gleiche Definition für Zufallsvariable:

(1.62) Definition

I sei eine abzählbare Menge, $(X_i)_{i\in I}$ seien ZVe. Die **terminale σ-Algebra** der X_i ist gegeben durch

$$\mathcal{T} = \bigcap_{J\subset I: J \text{ endlich}} \sigma((X_i)_{i\in J^c}).$$

Wenn man die (X_i) auf dem Produktraum realisiert, kommt man exakt zu Definition (1.60) zurück. Der folgende Satz liefert eine interessante Aussage in dem Fall, dass die (X_i) unabhängig sind (oder: dass auf dem Produktraum ein Produktmaß liegt).

(1.63) Satz: (0-1-Gesetz von Kolmogorov)

Sei I eine abzählbare Menge, und seien $(X_i)_{i\in I}$ unabhängige ZVe, definiert auf $(\Omega, \mathcal{F})$. Sei $\mathcal{T} \subset \mathcal{F}$ die terminale σ-Algebra der X_i. Dann gilt

$$\forall A \in \mathcal{T}: \quad \mathbb{P}(A) \in \{0,1\}.$$

Beweis: Sei $A \in \mathcal{T}$. Zunächst zeigen wir
a) Für alle $n \in \mathbb{N}$ und alle $B \in \mathcal{F}_n$ ist $A \perp\!\!\!\perp B$.
Dazu sei $(E_i, \mathcal{E}_i)$ der Zielraum von X_i. Für $1 \leqslant i \leqslant n+m$ und $C_i \in \mathcal{E}_i$ gilt dann wegen der Unabhängigkeit der (X_i)

$$\mathbb{P}(X_i \in C_i\ \forall i \leqslant n, X_j \in C_j \forall n < j \leqslant n+m) = \mathbb{P}(X_i \in C_i\ \forall i \leqslant n)\mathbb{P}(X_j \in C_j \forall n < j \leqslant n+m).$$

Nun ist $\{\{X_i \in C_i\ \forall i \leqslant n\} : C_i \in \mathcal{E}_i\ \forall i \leqslant n\}$ ein $\cap$-stabiler Erzeuger von $\mathcal{F}_n$, und $\{\{X_j \in C_j\ \forall n < j \leqslant n+m\} : m \in \mathbb{N}, C_j \in \mathcal{E}_j\ \forall n < j \leqslant n+m\}$ ein $\cap$-stabiler Erzeuger von $\mathcal{T}_n$. Daher gilt wegen (1.56 b) $\mathcal{F}_n \perp\!\!\!\perp \mathcal{T}_n$, und wegen $\mathcal{T} \subset \mathcal{T}_n$ daher auch $\mathcal{F}_n \perp\!\!\!\perp \mathcal{T}$ für alle n. Dies zeigt die Behauptung a).
b) Andererseits ist $\bigcup_{n\in\mathbb{N}} \mathcal{F}_n$ ein $\cap$-stabiler Erzeuger von $\mathcal{F}$. Daher ist sogar $\mathcal{F} \perp\!\!\!\perp \mathcal{T}$, und da $\mathcal{T} \subset \mathcal{F}$ somit $\mathcal{T} \perp\!\!\!\perp \mathcal{T}$. Mit anderen Worten gilt für alle $A, B \in \mathcal{T}$: $A \perp\!\!\!\perp B$. Man darf hier natürlich auch $A = B$ wählen, das führt zu $A \perp\!\!\!\perp A$ für alle $A \in \mathcal{T}$. Dies bedeutet $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2$, also $\mathbb{P}(A) \in \{0,1\}$. □.

Anwendungen:

a) Perkolation:
(i): Wir haben in (1.36) gesehen, dass $p \mapsto \mathbb{P}_p(M_z = \infty)$ monoton wächst.
(ii): Wegen Satz (1.63) und Beispiel (1.61 d) ist $\mathbb{P}_p(\exists z \in \mathbb{Z} : M_z = \infty) \in \{0,1\}$.
(iii): Wegen der Translationsinvarianz der Verteilung von η gilt $\mathbb{P}_p(M_z = \infty) = \mathbb{P}_p(M_y = \infty)$ für alle $y, z \in \mathbb{Z}$.
(iv): Damit ergeben sich zwei Möglichkeiten: Falls $\mathbb{P}_p(M_0 = \infty) > 0$, so ist $\mathbb{P}_p(\exists z : M_z = \infty) \geqslant \mathbb{P}_p(M_0 = \infty) > 0$, daher $\mathbb{P}_p(\exists z : M_z = \infty) = 1$. Falls andererseits $\mathbb{P}_p(M_0 = \infty) = 0$,

dann ist wegen der σ-Subadditivität

$$\mathbb{P}_p(\exists z : M_z = \infty) = \mathbb{P}_p(\bigcup_{z\in\mathbb{Z}} M_z = \infty) \leqslant \sum_{z\in\mathbb{Z}} \mathbb{P}_p(M_z = \infty) = \sum_{z\in\mathbb{Z}} \mathbb{P}_p(M_0 = \infty) = 0.$$

Es existiert also ein sogenanntes *kritisches* $p_c \in [0,1]$ so dass

$$\mathbb{P}_p(\exists z : M_z = \infty) = 0 \text{ für } p < p_c \qquad \text{und} \qquad \mathbb{P}_p(\exists z : M_z = \infty) = 1 \text{ für } p > p_c.$$

Für $d = 1$ werden wir sehr bald zeigen können, dass $p_c = 1$ ist, d.h. $M_z = \infty$ ist nur in der trivialen Situation möglich, wo man alle Kästchen schwarz macht. Für $d > 2$ stellt sich heraus, dass $0 < p_c < 1$ gilt, dies zu zeigen ist aber schon relativ schwierig.
Was dagegen leicht ist, ist folgende Aussage: Sei $p_c(d)$ das kritische p für die Perkolation in $\mathbb{Z}^d$. Dann ist $p_c(d+1) \leqslant p_c(d)$ für alle d. Der Beweis bleibt als Übung.
Den *Wert* von p_c zu bestimmen oder abzuschätzen ist noch schwieriger. Für einige wenige Abwandlungen der Perkolation (andere Gitter, leicht andere Regeln) gibt es analytische Ergebnisse, im allgemeinen aber nur numerische. Im Fall von $\mathbb{Z}^d$ gibt es nur numerische Resultate, hier ist $p_c(2) \approx 0.59$ und $p_c(3) \approx 0.22455$.

b) Einfache Irrfahrt, oder: wie fair ist ein faires Münzwurfspiel?
Seien $(X_i)_{i\in\mathbb{N}}$ unabhängige ZVen mit $\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = 1) = 1/2$. $S_n = \sum_{i=1}^n X_i$ ist dann also die Position der einfachen Irrfahrt in einer Dimension, oder das Vermögen von Spieler 1 nach n Spielen eines fairen Münzwurfspiels mit Gewinn oder Verlust von 1 pro Runde. Wir wollen untersuchen, wie wahrscheinlich es ist, dass ein Spieler während eines unendlich langen Spieles „immer wieder“ signifikant im Vorteil ist.
Um diese Aussage zu präzisieren, betrachten wir eine (nicht-zufällige) Folge $(Z_n)_{n\in\mathbb{N}}$ mit $Z_n > 0$ für alle n; interessant sind insbesondere Folgen mit $\lim_{n\to\infty} Z_n = \infty$. Das Ereignis $\{S_n/Z_n \geqslant 1$ für unendlich viele $n\}$ bedeutet dann, dass der Vermögensstand des ersten Spielers unendlich oft über der Grenze Z_n liegt. Wenn $\lim_{n\to\infty} Z_n = \infty$, so könnte man das so interpretieren, dass das Spiel „immer unfairer wird“ (aber: sollte man das auch?).
Wir sind nun für vorgegebenes (Z_n) interessiert an $\mathbb{P}(S_n/Z_n \geqslant 1$ für unendlich viele $n)$. Wegen Satz (1.63) gilt jedenfalls

$$p := \mathbb{P}(S_n/Z_n \geqslant 1 \text{ für unendlich viele } n) \in \{0,1\}$$

für jede vorher festgelegte Folge (Z_n) (die genaue Begründung sollte man als Übung durchführen!).
(i): Sogar ohne Benutzung dieser Tatsache sieht man sofort dass $p = 0$ falls $\liminf_{n\to\infty} Z_n/n > 1$ (warum?).
(ii): Nicht viel schwieriger ist es, ebenfalls ohne Benutzung von Satz (1.63) zu zeigen, dass $p = 0$ falls $Z_n = n$ (Übung).
(iii): Mit Hilfe von Satz (1.63) kann man relativ leicht zeigen (Übung!), dass

$$\mathbb{P}(S_n \geqslant 0 \text{ unendlich oft}) = \mathbb{P}(S_n \leqslant 0 \text{ unendlich oft}) = 1.$$

Dies bedeutet, dass mit W'keit 1 kein Spieler einen so großen Vorsprung erhält, dass er „für immer“ in Führung bleibt. In diesem Sinne ist das Spiel also fair.
(iv): Um genauer zu untersuchen, welchen Wert p für welche (Z_n) hat, brauchen wir noch einige weitere Vorbereitungen.

(1.64) Definition und Diskussion: liminf und limsup

Aus der Analysis kennen wir die Begriffe lim sup und lim inf für reelle Folgen (a_n):

$$\limsup_{n\to\infty} a_n := \lim_{n\to\infty} \sup_{m \geqslant n} a_m, \qquad \liminf_{n\to\infty} a_n := \lim_{n\to\infty} \inf_{m \geqslant n} a_m.$$

Ausgehend von diesen Definitionen mache man sich klar, dass für alle $c \in \mathbb{R}$ gilt:

$$\limsup_{n\to\infty} a_n > c \quad \Longrightarrow \quad a_n > c \text{ für unendlich viele } n \quad \Longrightarrow \quad \limsup_{n\to\infty} a_n \geqslant c,$$

während

$$\liminf_{n\to\infty} a_n > c \quad \Longrightarrow \quad a_n \leqslant c \text{ nur für endlich viele } n \quad \Longrightarrow \quad \liminf_{n\to\infty} a_n \geqslant c.$$

Die etwas komplizierten Aussagen sind der Tatsache geschuldet, dass $\lim_{n\to\infty} a_n = c$ gelten kann obwohl $a_n < c$ für alle n ist. Falls die Folge (a_n) nur Werte in $\mathbb{Z}$ annehmen kann dann haben wir die einfacheren Beziehungen

$$\limsup_{n\to\infty} a_n \geqslant c \quad \Longleftrightarrow \quad a_n \geqslant c \text{ für unendlich viele } n,$$

und

$$\liminf_{n\to\infty} a_n \geqslant c \quad \Longleftrightarrow \quad a_n < c \text{ nur für endlich viele } n.$$

Als nächstes können wir die Definition des lim inf und lim sup problemlos auf Funktionen übertragen, indem wir sie *punktweise* interpretieren: für eine Menge E und eine Folge von Funktionen $(f_n)_{n\in\mathbb{N}}$ mit $f_n : E \to \mathbb{R}$ sind die Funktionen $\liminf_{n\to\infty} f_n$ und $\limsup_{n\to\infty} f_n$ definiert durch

$$\big(\liminf_{n\to\infty} f_n\big)(x) := \liminf_{n\to\infty} f_n(x), \qquad \big(\limsup_{n\to\infty} f_n\big)(x) := \limsup_{n\to\infty} f_n(x)$$

für alle $x \in E$. Für uns ist der Fall der Indikatorfunktionen interessant, also wenn $f_n(x) = \mathbb{1}_{A_n}(x)$ mit $A_n \subset E$ ist. Dann gilt

$$\limsup_{n\to\infty} \mathbb{1}_{A_n}(x) = 1 \iff \mathbb{1}_{A_n}(x) = 1 \text{ unendlich oft} \iff x \in A_n \text{ für unendlich viele } n,$$

und

$$\liminf_{n\to\infty} \mathbb{1}_{A_n}(x) = 1 \iff \mathbb{1}_{A_n}(x) = 0 \text{ nur endlich oft} \iff x \notin A_n \text{ nur für endlich viele } n.$$

Dies führt direkt auf die Definition von lim sup und lim inf für Mengen: für Mengen $A_n \subset E$, $n \in \mathbb{N}$ ist

$$\begin{aligned}\limsup_{n\to\infty} A_n &:= \{x \in E : x \in A_n \text{ für unendlich viele } n\} = \\ &= \{x \in E : \forall m \in \mathbb{N}\ \exists n \geqslant m \text{ mit } x \in A_n\} = \bigcap_{m\in\mathbb{N}} \bigcup_{n \geqslant m} A_n,\end{aligned}$$

und

$$\begin{aligned}\liminf_{n\to\infty} A_n &:= \{x \in E : x \notin A_n \text{ für nur endlich viele } n\} = \\ &= \{x \in E : \exists m \in \mathbb{N}\ \forall n \geqslant m \text{ ist } x \in A_n\} = \bigcup_{m\in\mathbb{N}} \bigcap_{n \geqslant m} A_n.\end{aligned}$$

Aus der obigen Diskussion sehen wir, dass

$$\mathbb{1}_{\liminf_{n\to\infty} A_n}(x) = \liminf_{n\to\infty} \mathbb{1}_{A_n}(x), \quad \mathbb{1}_{\limsup_{n\to\infty} A_n}(x) = \limsup_{n\to\infty} \mathbb{1}_{A_n}(x)$$

für alle $x \in E$. Weiter sollte man sich klar machen, dass gilt:

$$\bigcap_{n\in\mathbb{N}} A_n \subset \liminf_{n\to\infty} A_n \subset \limsup_{n\to\infty} A_n \subset \bigcup_{n\in\mathbb{N}} A_n.$$

Ebenso sollte man sich im Kontext des letzten Beispiels klarmachen, dass

$$\{\limsup_{n\to\infty} S_n/Z_n \geqslant 1\} = \limsup_{n\to\infty}\{S_n/Z_n \geqslant 1\},$$

ebenso für $\liminf$.

(1.65) Satz: Lemmata von Borel-Cantelli

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum, $A_j \in \mathcal{F}$ für alle $j \in \mathbb{N}$.

a) Erstes Lemma von Borel-Cantelli:

Falls $\sum_{j=1}^{\infty} \mathbb{P}(A_j) < \infty$, dann ist $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$.

b) Zweites Lemma von Borel-Cantelli:

Die Mengen (A_n) seien eine unabhängige Familie, und es gelte $\sum_{j=1}^{\infty} \mathbb{P}(A_j) = \infty$. Dann ist $\mathbb{P}(\limsup_{n\to\infty} A_n) = 1$.

Beweis: Setze $A := \limsup_{n\to\infty} A_n$.

a) Es gilt

$$\bigcup_{m\geqslant n} A_m \underset{n\to\infty}{\searrow} \bigcap_{n\in\mathbb{N}} \bigcup_{m\geqslant n} A_m = A,$$

und daher

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{m\geqslant n} A_m\Big) \leqslant \lim_{n\to\infty} \sum_{m\geqslant n} \mathbb{P}(A_m) = 0,$$

denn nach Voraussetzung konvergiert ja die Reihe $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

b) Mit A wie oben gilt $A^c = \bigcup_{m\in\mathbb{N}} \bigcap_{n\geqslant m} A_n^c$, und es gilt $\bigcap_{n=m}^{N} A_n^c \underset{N\to\infty}{\searrow} \bigcap_{n=m}^{\infty} A_n^c$ für alle $m \in \mathbb{N}$. Daher ist

$$\mathbb{P}(A^c) \leqslant \sum_{m=1}^{\infty} \mathbb{P}\Big(\bigcap_{n=m}^{\infty} A_n^c\Big) = \sum_{m=1}^{\infty} \lim_{N\to\infty} \mathbb{P}\Big(\bigcap_{n=m}^{N} A_n^c\Big) = \sum_{m=1}^{\infty} \lim_{N\to\infty} \prod_{n=m}^{N} \mathbb{P}(A_n^c).$$

Mit Hilfe der sehr nützlichen Ungleichung $1 - x \leqslant \mathrm{e}^{-x}$ für alle $x \in \mathbb{R}$ erhält man

$$\prod_{n=m}^{N} \mathbb{P}(A_n^c) = \prod_{n=m}^{N} (1 - \mathbb{P}(A_n)) \leqslant \prod_{n=m}^{N} \mathrm{e}^{-\mathbb{P}(A_n)} = \mathrm{e}^{-\sum_{n=m}^{N} \mathbb{P}(A_n)} \overset{N\to\infty}{\longrightarrow} 0,$$

da ja $\sum_{n=m}^{\infty} \mathbb{P}(A_n) = \infty$ direkt aus der Annahme folgt. Daher ist $\mathbb{P}(A^c) = \sum_{m=1}^{\infty} 0 = 0$. $\square$

(1.66) Beispiele

a) Perkolation:
Wir zeigen hier, dass $p_c = 1$ im Fall $d = 1$ gilt. Hierzu muss man nicht zwingend das Borel-Cantelli-Lemma benutzen, man kann es aber wie folgt tun: sei $p < 1$ und $A_n = \{\eta_n = \eta_{-n} = 0\}$, dann gilt $M_0(\eta) < n$ für alle $\eta \in A_n$, also

$$\{M_0 < \infty\} \supset \bigcup_{n\in\mathbb{N}} A_n \supset \limsup_{n\to\infty} A_n.$$

Es ist $A_n \perp\!\!\!\perp A_m$ für alle $n \neq m$ und $\mathbb{P}(A_n) = (1-p)^2 > 0$ für alle n, somit $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty$. Mit dem zweiten Broel-Cantelli Lemma folgt daher $\mathbb{P}(\limsup_{n\to\infty} A_n) = 1$, also $\mathbb{P}(M_0 < \infty) = 1$.

Es ist interessant zu sehen, warum diese Strategie in zwei und mehr Dimensionen nicht mehr funktoniert. Der offensichtliche Versuch wäre, nun

$$A_n = \{\eta_x = 0 \ \forall x \in \mathbb{Z}^d \text{ mit } x_i = n \text{ für mindestens ein } i \text{ und } x_k \leqslant n \text{ für alle } i \leqslant d\}$$

A_n ist also das Ereignis, dass der gesamte Rand einer Box aus weißen Kästchen besteht. Zwar ist wie vorher $\{M_0 < \infty\} \supset \limsup_{n\to\infty} A_n$, aber diesmal ist im Unterschied zu $d = 1$ die Anzahl der Kästchen, für die in A_n eine Bedingung verlangt wird, wachsend; er besteht aus (viel) mehr als n^{d-1} Kästchen. Die W'keit, dass alle diese Kästchen weiß sind, ist

$$\mathbb{P}(A_n) \leqslant (1-p)^{n^{d-1}} \leqslant (1-p)^n,$$

und daher $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$. Daher kann man nun nicht mehr wie vorher das zweite Borel-Cantelli-Lemma anwenden um $\mathbb{P}(M_0 = \infty) = 0$ zu schließen. Das erste Borel-Cantelli liefert nun zwar $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$, aber das hilft uns nicht weiter. Denn es besagt ja nur, dass man mit W'keit 1 nur endlich oft einen „geschlossenen Ring" von weißen (blockierenden) Kästchen um die 0 findet; daraus kann man zwar schließen, dass man mit positiver W'keit gar keinen solchen Ring findet (wie?), allerdings gibt es natürlich viele andere Arten, um eine Verbindung von 0 nach ∞ auf schwarzen Kästchen zu verhindern, so dass uns diese Aussage nicht dazu berechtigt, $\mathbb{P}_p(M_0 = \infty) > 0$ zu schließen; das wäre für beliebige p auch falsch.

b) Einfache Irrfahrt und faires Münzwurfspiel
Wie im Beispiel b) zu (1.64) seien $(X_i)_{i\in\mathbb{N}}$ mit $\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = 1) = 1/2$, und $S_n = \sum_{i=1}^n X_i$. Wir suchen eine (möglichst langsam) monoton wachsende Folge (Z_n), für die wir gerade noch

$$\mathbb{P}(\limsup_{n\to\infty} S_n/Z_n \geqslant 1) = 0$$

beweisen können. Wir wissen aus dem vorigen Beispiel, dass $Z_n = n$ so eine Folge ist, aber das können wir wesentlich verbessern. Da $\{\limsup_{n\to\infty} S_n/Z_n \geqslant 1\} = \limsup_{n\to\infty}\{S_n/Z_n \geqslant 1\}$ gilt und wir das erste BC-Lemma anwenden wollen, müssen wir die Summe

$$\sum_{n=1}^{\infty} \mathbb{P}(S_n/Z_n \geqslant 1) = \sum_{n=1}^{\infty} \mathbb{P}(S_n \geqslant Z_n)$$

untersuchen. Dies kann man entweder recht mühsam „von Hand" tun, indem man sich mit Hilfe von Binomialkoeffizienten überlegt, wie wahrscheinlich es ist, dass die einfache Irrfahrt nach n Schritten größer als eine gewisse Schranke ist. Im übernächsten Kapitel werden wir jedoch eine relativ einfache und elegante Methode kennen lernen (die sogenannte Chernoff-Schranke),

die es uns erlaubt, für viele Zufallsvariable sehr gute Abschätzungen für W'keiten der Form $\mathbb{P}(X > c)$ zu erhalten. Für S_n ergibt sich daraus

$$\mathbb{P}(S_n \geqslant c) \leqslant \mathrm{e}^{-\frac{c^2}{2n}}$$

für alle $c > 0$ und alle $n \in \mathbb{N}$. Wenn wir diese Schranke für den Moment als gegeben hinnehmen, erhalten wir beispielsweise mit $Z_n = n^{1/2+\varepsilon}$, $\varepsilon > 0$ die Abschätzung

$$\mathbb{P}(S_n \geqslant Z_n) \leqslant \mathrm{e}^{-\frac{n^{2\varepsilon}}{2}} =: b_n$$

Um zu sehen, dass die Folge (b_n) summierbar ist, reicht es zu sehen, dass $n^2 b_n$ eine beschränkte Folge ist. Dies ist wegen

$$\log(n^2 b_n) = 2\log n - n^{2\varepsilon}/2 \overset{n\to\infty}{\longrightarrow} -\infty$$

der Fall, da dann $n^2 b_n \to 0$ gehen muss und damit beschränkt ist. Mit dem ersten BC Lemma folgt nun

$$\mathbb{P}(\limsup_{n\to\infty} \frac{S_n}{n^{1/2+\varepsilon}} \geqslant 1) = 0.$$

Dies sagt uns, dass es unmöglich ist, dass ein Spieler unendlich oft um mehr in Führung liegt, als die $1/2+\varepsilon$-te Potenz aus der Anzahl der gespielten Runden. Die Frage ist natürlich, wie gut diese Abschätzung ist. Im Kapitel über den zentralen Grenzwertsatz werden wir beweisen, dass $\mathbb{P}(\limsup_{n\to\infty} S_n/\sqrt{n} = \infty) = 1$ gilt. Dies lässt nur noch einen kleinen Korridor der Unsicherheit, welcher durch den berühmten Satz vom iterierten Logarithmus geschlossen wird: es gilt nämlich

$$\mathbb{P}(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log(\log n)}} = 1\} = 1.$$

Während man die obere Schranke (also $\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log(\log n)}} \leqslant 1$) mit einer Verfeinerung der Chernoff-Schranke noch mit Mitteln dieser Vorlesung hinbekommen kann, ist die untere Schranke wesentlich aufwendiger zu beweisen.

2. Wichtige Wahrscheinlichkeitsmaße

Diskrete Verteilungen

(2.1) Die Gleichverteilung

Ω sei eine endliche Menge. Das W-Maß

$$\mathbb{P}_U : \mathcal{P}(\Omega) \to [0,1], \qquad A \mapsto \frac{|A|}{|\Omega|}$$

heißt **Gleichverteilung** auf Ω.

Beispiele:

a) Würfel: $\Omega = \{1,\dots,6\}$, $\mathbb{P}_U(\{\omega\}) = 1/6$ für alle $\omega \in \Omega$.

b) Einfache Irrfahrt mit N Schritten und Start 0, dargestellt als Gleichverteilung:

$$\Omega_N = \{(x_1,\dots,x_N) : x_i \in \mathbb{Z}^d, |x_i - x_{i-1}| = 1 \ \forall i \leqslant N\}$$

Dann ist $|\Omega_N| = (2d)^N$, und unter der Gleichverteilung $\mathbb{P}_U$ auf Ω sind die „Koordinatenabbildungen" (Achtung: Ω ist kein Produktraum!!) $(X_n)_{n\leqslant N}$ mit $X_i(x) = x_i$ eine einfache Irrfahrt

mit N Schritten. Übung: man rechne das konkret nach!

c) Selbstvermeidende Irrfahrt mit N Schritten und Start in 0:

$$\Omega_N = \{(x_1, \dots, x_N) : x_i \in \mathbb{Z}^d, |x_i - x_{i-1}| = 1 \ \forall i \leqslant N, x_i \neq x_j \text{ falls } i \neq j\},$$

wieder $\mathbb{P}_U$ die Gleichverteilung auf Ω_N. Außer für den (trivialen) Fall $d = 1$ gibt es für die Anzahl $|\Omega_N|$ der möglichen selbstvermeidenden Irrfahrten (und damit für die W'keit $1/|\Omega_N|$ einer einzelnen Irrfahrt) keine geschlossene Formel. Man kann zwar beweisen, dass der Grenzwert $\mu := \lim_{N\to\infty} \Omega_N^{1/N}$ existiert (evtl: Übung); die Anzahl der selbstvermeidenden Irrfahrten wächst also asymptotisch wie eine gewisse Potenz μ ihrer Länge; allerdings weiß man über die Größe von μ in fast allen Fällen nicht viel. Man sieht also, dass sogar die banale Gleichverteilung zu schwierigen Problemen führen kann, wenn der W-Raum Ω, auf dem sie definiert ist, kombinatorisch schwierig ist.

Im Folgenden werden wir oft die alternative Schreibweise $X \sim Y$ statt $X \stackrel{d}{=} Y$ verwenden, wenn X und Y die gleiche Verteilung haben.

(2.2) Die Bernoulli-Verteilung

a) Eine Zufallsvariable X mit Werten in $\{0, 1\}$ und $\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$ heißt **Bernoulli-verteilte ZV**. Man schreibt $X \sim \mathrm{Ber}_p$. Eine solche ZV modelliert ein Experiment (z.B. Münzwurf) mit Erfolgswahrscheinlichkeit p.

b) Die gemeinsame Verteilung von n unabhängigen ZVem $(X_i)_{i \leqslant n}$ mit $X_i \sim \mathrm{Ber}_p$ für alle i heißt **Bernoulli-Verteilung für n Versuche**. Wegen der Formel für die Produkt-Zähldichte ist für alle $x_1, \dots, x_n \in \{0, 1\}$

$$\mathbb{P}(X_1 = x_1, \dots X_n = x_n) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

(2.3) Die Binomialverteilung

a) Sei $n \in \mathbb{N}$, $p \in [0, 1]$. Die Funktion

$$b_{n,p} : \{0, \dots n\} \to [0, 1], \qquad k \mapsto \binom{n}{k} p^k (1-p)^{n-k}$$

ist eine Zähldichte auf $\{0, \dots, n\}$.

b) Eine ZV X mit Zähldichte $b_{n,p}$ heißt **binomialverteilt mit Erfolgswahrscheinlichkeit p und n Versuchen**. Man schreibt $X \sim B_{n,p}$.

c) Für unabhängige ZVen $(X_i)_{i \leqslant n}$ mit $X_i \sim \mathrm{Ber_p}$ für alle i gilt: $\sum_{i=1}^n X_i \sim B_{n,p}$. Die Binomialverteilung modelliert also die **Anzahl der Erfolge** bei n unabhängigen Versuchen und Wahrscheinlichkeit p für einen Erfolg.

d) Es gilt: falls $X \sim B_{n,p}$ und $Y \sim B_{m,p}$, und falls $X \perp\!\!\!\perp Y$, dann ist $X + Y \sim B_{n+m,p}$.
Die Beweise der obigen Aussagen verbleiben als Übung.

Beispiele:

a) Urne mit N Kugeln, davon w weiße und s schwarze, ziehen mit Zurücklegen. Dann ist die Anzahl der gezogenen weißen Kugeln nach n Ziehungen gemäß $B_{n,w/(s+w)}$ verteilt.

b) Einfache Irrfahrt: Seien $Y_n \sim \mathrm{Ber}_{1/2}$, unabhängig, und sei $X_n = 2Y_n - 1$. Dann ist $S_n = \sum_{i=1}^n X_i$ die einfache Irrfahrt, daher gilt für diese: $S_n \sim 2B_{n,1/2} - n$. Hier bezeichnet n die konstante ZVe, die für jedes $\omega \in \Omega$ den Wert n annimmt.

c) Das **Galton-Brett** ist ein schönes Hilfsmittel zur Veranschaulichung der $B_{n,1/2}$-Verteilung. Schauen Sie sich eine Simulation davon im Internet an und versuchen Sie, das Gesehene mit den Inhalten der Vorlesung in Verbindung zu bringen.

(2.4) Bezeichnung

Λ sei eine abzählbare Menge. Eine Abbildung $\eta : \Lambda \to \mathbb{N}_0$ (oft auch geschrieben als $\eta \in \mathbb{N}_0^\Lambda$) heißt **Besetzungszahl-Abbildung** oder auch nur **Besetzungszahl** der Elemente von Λ.
Beispiele:
Für $x \in \Lambda$ und $\eta \in \mathbb{N}^\Lambda$ wird η_x als die Anzahl von „ Objekten“ vom Typ x (oder: von Objekten an der Stelle x) interpretiert.

a) Ziehen aus einer Urne mit $|\Lambda|$ verschiedenfarbigen Kugeln. Dann ist η_x die Anzahl der gezogenen Kugeln der Farbe x.

b) Perkolation: $\Lambda = \mathbb{Z}^d$, $\eta_x \in \{0, 1\}$ für alle x. Interpretation: statt ein Kästchen schwarz zu machen, besetzt man es mit höchstens einer Kugel. Der W-Raum der Kopplung von Beispiel (1.36) benutzte dagegen Besetzungszahlen mit den Werten $0, 1, 2$.

c) Gasmodell: Ein großer Würfel $W \subset \mathbb{R}^3$ wird in N^3 kleinere Würfel aufgeteilt. Λ ist die Menge der Mittelpunkte dieser Würfel. η_x für $x \in \Lambda$ ist die Anzahl der Gasatome, die sich im Würfel mit Mittelpunkt x befinden.

(2.5) Die Multinomialverteilung

a) Λ sei eine endliche Menge, p eine Zähldichte auf Λ, $N \in \mathbb{N}$. Auf der Menge der $\Omega_N := \{\eta \in \mathbb{N}_0^\Lambda : \sum_{x\in\Lambda} \eta_x = N\}$ der Besetzungszahlen von Λ mit Gesamtbesetzung N ist die Funktion $m_{N,p}$ mit

$$m_{N,p}(\eta) = \frac{N!}{\prod_{x\in\Lambda} \eta_x!} \prod_{x\in\Lambda} p(x)^{\eta_x}$$

eine Zähldichte. Die zugehörige Verteilung heißt **Multinomialverteilung** für N Stichproben und Einzewahrscheinlichkeiten $p(x)_{x\in\Lambda}$. Für eine ZV X mit dieser Verteilung schreiben wir $X \sim \mathcal{M}_{N,p}$.

b) Für $\Lambda = \{1, \dots, n\}$ kann man $m_{N,p}$ auch mit Hilfe des Multinomialkoeffizienten

$$\binom{N}{\eta_1, \dots, \eta_n} = \frac{N!}{\eta_1!\eta_2! \cdots \eta_n!}$$

schreiben. Daher der Name der Verteilung.

c) $\mathcal{M}_{N,p}$ modelliert die W'keit, bei N Versuchen jeweils genau η_x Objekte vom Typ x zu

erhalten, wenn die Einzelw'keit für ein Objekt vom Typ x bei einer Durchführung einer Ziehung $p(x)$ ist.
Beweis, dass $m_{N,p}$ eine Zähldichte ist: die W'keit, jeweils η_x-mal das Ergebnis x in einer vorher festgelegten Reihenfolge zu erhalten, ist bei unabhängigen Versuchen wegen der Produktformel der Zähldichten genau $\prod_{x\in\Lambda} p(x)^{\eta_x}$. Da es auf die Reihenfolge der Ziehung aber nicht ankommt, muss man nun nachzählen, auf wie viele Weisen ein Ergebnis $(\eta_x)_{x\in\Lambda}$ erzielt werden kann. Für den ersten Typ x_1 muss man η_{x_1} Zeitpunkte der Ziehung wählen, das gibt $\binom{N}{x_1} = \frac{N!}{x_1!(N-x_1)!}$ Möglichkeiten. Für den Typ x_2 muss man aus den verbleibenden $N - x_1$ Zeitpunkten nun x_2 wählen, das ergibt $\binom{N-x_1}{x_2} = \frac{(N-x_1)!}{x_2!(N-x_2)!}$ Möglichkeiten. Da man diese Größen multiplizieren muss, sieht man schon, dass sich hier etwas kürzt. Setzt man dies fort, erhält man den Multinomialkoeffizienten. Da der ursprüngliche Raum der zeitlich geordneten Ziehungen ein W-Raum war, ist die angegebene Funktion eine Zähldichte, nämlich die Zähldichte des Bildmaßes unter der Abbildung

$$\Lambda^N \to \mathbb{N}^\Lambda, \qquad (x_i)_{i\leqslant N} \mapsto \Big(\big|\{i\in\{1,\dots,N\} : x_i = x\}\big|\Big)_{x\in\Lambda},$$

die zählt, wie oft jedes Element von Λ in der Folge $(x_1,\dots,x_N)$ vertreten ist. □

Bemerkung: für $\Lambda = \{0,1\}$ erhält man übrigens wieder die Binomialverteilung.

(2.6) Hypergemoetrische Verteilung

Λ sei eine endliche Menge. Für $N\in\mathbb{N}$ und $k\in\mathbb{N}_0^\Lambda$ mit $K := \sum_{x\in\Lambda} k_x \geqslant N$ ist die Abbildung $h_{N,(k_x)_{x\in\Lambda}} : \mathbb{N}_0^\Lambda \to [0,1]$ mit

$$h_{N,(k_x)_{x\in\Lambda}}(\eta) := \begin{cases} \frac{1}{\binom{K}{N}} \prod_{x\in\Lambda} \binom{k_x}{\eta_x} & \text{falls } \eta_x \leqslant k_x\ \forall x\in\Lambda, \sum_{x\in\Lambda}\eta_x = N \\ 0 & \text{sonst} \end{cases}$$

eine Zähldichte. Das zugehörige W-Maß $H_{N,(k_x)_{x\in\Lambda}}$ heißt **hypergeometrische Verteilung** für N Versuche und k_x vorhandene Exemplare vom Typ x. Sie modelliert das Ziehen ohne Zurücklegen aus einer Urne mit ursprünglich k_x Kugeln vom Typ x. $h_{N,(k_x)_{x\in\Lambda}}(\eta)$ ist hierbei die W'keit, genau η_x Kugeln vom Typ x zu ziehen.
Der Beweis, dass eine Zähldichte vorliegt, benutzt wieder Kombinatorik und wird hier weggelassen. Die hypergeometrische Verteilung ist für uns im Folgenden von untergeordneter Bedeutung. Allerdings hat sie beim Lotto eine konkrete Anwendung:

Beispiel: Setze $\Lambda = \{0,1\}$. Dann ist, für $\eta_1 \leqslant k_1$, $\eta_2 \leqslant k_2$ und $\eta_1+\eta_2 = N$,

$$h_{N,(k_1,k_2)}(\eta) = \frac{\binom{k_1}{\eta_1}\binom{k_1}{\eta_2}}{\binom{k_1+k_2}{N}}$$

die W'keit, bei N-maligem Ziehen aus k_1 „guten“ und k_2 „schlechten“ Optionen genau η_1 gute und $\eta_2 = N-\eta_1$ schlechte zu bekommen. Beim Lotto 6 aus 49 sind die 6 angekreuzten Zahlen die guten Optionen, und wir haben beispielsweise

$$\mathbb{P}(\text{ vier Richtige }) = h_{6,(6,43)}(4) = \frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}} \approx 9.686\cdot 10^{-4}.$$

(2.7) Die geometrische Verteilung

Für $0 \leqslant p \leqslant 1$ ist

$$g_p : \mathbb{N}_0 \to [0,1], \qquad k \mapsto p(1-p)^k$$

eine Zähldichte auf $\mathbb{N}_0$. Das zugehörige W-Maß heißt **geometrische Verteilung zur Erfolgswahrscheinlichkeit** p. Man schreibt $X \sim \mathrm{Geo}_p$ wenn X geometrisch verteilt ist.
Bemerkungen:

a) Der Beweis der Zähldichteneigenschaft ist kurz: $\sum_{k=0}^{\infty} p(1-p)^k = p\frac{1}{1-(1-p)} = 1$ wegen der geometrischen Summenformel; daher auch der Name der Verteilung.

b) $g_p(k)$ modelliert die W'keit, dass man in einem unabhängig wiederholten Experiment (z.B. Münzwurf), welches die Erfolgsw'keit p hat, genau k Misserfolge sieht, bevor man den ersten Erfolg erzielt. (Aus diesem Grund nennt man die geometrische Verteilung auch eine *Wartezeitverteilung*.) Wenn man will, kann man die obige Behauptung mit Hilfe der Zufallsvariablen „Trefferzeit“ zeigen: wir setzen

$$\Omega = \{0,1\}^{\mathbb{N}}, \quad \mathcal{F} = \mathcal{P}(\{0,1\})^{\otimes\mathbb{N}}, \quad \mathbb{P} = \mathrm{Ber_p}^{\otimes\mathbb{N}}, \quad X_i(x) = x_i \ \forall x \in \Omega, \quad \tau_1 = \inf\{n \in \mathbb{N}_0 : X_n = 1\}$$

Dann ist $\mathbb{P}(k \text{ Misserfolge vor erstem Erfolg}) = \mathbb{P}(\tau_1 = k+1)$, und

$$\begin{aligned}\mathbb{P}(\tau_1 = k+1) &= \mathbb{P}(X_1 = 0, \dots, X_k = 0, X_{k+1} = 1) = \\ &= \mathbb{P}(X_1 = 0) \cdots \mathbb{P}(X_k = 0)\mathbb{P}(X_{k+1} = 1) = (1-p)^k p = g_p(k).\end{aligned}$$

Vornehm ausgedrückt: Geo_p ist das Bildmaß des unendlichen Produktmaßes von $(\mathrm{Ber}_p)^{\otimes\mathbb{N}}$ unter der Abbildung $\tau_1 - 1$.

c) Man kann auch fragen, wie wahrscheinlich es ist, vor dem r-ten Erfolg genau k Misserfolge erleiden zu müssen. Das Ergebnis ist

$$\bar{\mathcal{B}}_{r,p}(k) = \frac{r(r+1)(r+2)\cdots(r+k-1)}{k!} p^r (1-p)^k$$

Dies ist eine Zähldichte (Beweis und Rechtfertigung der Interpretation: Kombinatorik) und die zugehörige Verteilung heißt **negative Binomialverteilung** zu den Parametern r, k und p. Für $r = 1$ ist dies wieder die geometrische Verteilung.

(2.8) Die Poisson-Verteilung

Sei $\lambda \geqslant 0$. Dann ist die Abbildung $p_\lambda : \mathbb{N}_0 \to [0,1]$ mit

$$p_\lambda(k) := \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}$$

eine Zähldichte (Konvention: $0^0 = 0! = 1$). Das zugehörige W-Maß heißt **Poisson-Verteilung zum Parameter** λ. Ist eine ZV X Possion-verteilt, schreiben wir $X \sim \mathrm{Poi}_\lambda$.
Da $\sum_{k=0}^{\infty} \lambda^k/k! = \mathrm{e}^\lambda$ ist, ist die p_λ tatsächlich eine Zähldichte.

(2.9) Lemma

Für $X \sim \mathrm{Poi}_\lambda$ und $Y \sim \mathrm{Poi}_\mu$ ist $X + Y \sim \mathrm{Poi}_{\lambda+\mu}$.
Beweis: Übung.

(2.10) Satz: Poisson-Approximation

Sei $\lambda \geqslant 0$ und $(p_n)_{n\in\mathbb{N}}$ eine Folge mit Werten in $[0,1]$ und mit der Eigenschaft $\lim_{n\to\infty} np_n = \lambda$. Dann gilt für alle $k \in \mathbb{N}_0$:

$$\lim_{n\to\infty} \mathrm{B}_{n,p_n}(\{k\}) = \mathrm{Poi}_\lambda(\{k\}).$$

Hier ist $\mathrm{B}_{n,p}$ die Binomialverteilung.

Beweis: Zu zeigen ist

$$\lim_{n\to\infty} b_{n,p_n}(k) = \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \mathrm{e}^{-\lambda}\,\frac{\lambda^k}{k!} = p_\lambda(k).$$

Den Fall $\lambda = 0$ macht man getrennt, indem man in diesem Fall $\lim_{n\to\infty} b_{n,p_n}(0) = 1$ zeigt. Dies geht ähnlich wie das Argument am Schluss des jetzt folgenden Beweises und verbleibt als Übung.

Sei also nun $\lambda > 0$. Wir zeigen, dass der Quotient aus den beiden Ausdrücken oben gegen 1 konvergiert. Da der rechte Ausdruck strikt positiv ist, reicht das. Wir rechnen also

$$\begin{aligned} q_n &:= \binom{n}{k} p_n^k (1-p_n)^{n-k}\,\mathrm{e}^\lambda\,\lambda^{-k}k! = \frac{n(n-1)\cdots(n-k+1)}{k!}\,\frac{1}{n^k}(np_n)^k(1-p_n)^{n-k}\,\mathrm{e}^\lambda\,\lambda^{-k}k! = \\ &= \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{(*)_1}\underbrace{\Big(\frac{np_n}{\lambda}\Big)^k}_{(*)_2}(1-p_n)^n\,\mathrm{e}^\lambda\,\underbrace{(1-p_n)^{-k}}_{(*_3)}. \end{aligned}$$

Nach Voraussetzung gilt $\lim_{n\to\infty} np_n = \lambda$, und dies impliziert insbesondere $\lim_{n\to\infty} p_n = 0$. Da außerdem k endlich und fest ist, konvergieren die Ausdrücke $(*)_1$, $(*)_2$ und $(*)_3$ alle gegen 1 wenn $n \to \infty$. Der verbleibende Ausdruck $(1-p_n)^n$ ist etwas weniger einfach; es ist vorteilhaft, hier zunächst zu logarithmieren:

$$\log(1-p_n)^n = n\log(1-p_n),$$

und dann die Taylor-Reihe mit Restglied des Logartihmus zu benutzen: für $|x| < 1$ ist

$$\log(1-x) = 0 - x - \frac{x^2}{2}\frac{1}{(1-\xi)^2} \qquad \text{für ein } \xi \text{ mit } |\xi| \leqslant |x|,$$

und daher ist für $|x| < 1/2$

$$-x - 2x^2 \leqslant \log(1-x) \leqslant -x$$

Wegen $\lim_{n\to\infty} p_n = 0$ gilt dies für hinreichend große n und $x = p_n$, und es folgt

$$-np_n - 2np_n^2 \leqslant n\log(1-p_n) \leqslant -np_n,$$

und folglich $\lim_{n\to\infty} n\log(1-p_n) = -\lambda$. Da die Exponentialfunktion stetig ist, dürften wir sie auf dieses Ergebnis anwenden und den Limes hindurchziehen, das ergibt

$$\lim_{n\to\infty} (1-p_n)^n = \mathrm{e}^{-\lambda}.$$

Setzt man das noch in den Ausdruck für q_n ein, dann erhält man $\lim_{n\to\infty} q_n = 1$, wie benötigt. □

Bemerkung und Beispiele

a) Satz (2.10) ist zunächst einmal einfach eine Aussage über das Verhalten eine gewissen reellen

Folge. Wichtig ist jedoch die **W-Theoretische Interpretation:** hat man bei einem Experiment zwar nur eine winzige Wahrscheinlichkeit p auf Erfolg, dafür aber eine riesige Anzahl n an Versuchen, dann kann man zur Modellierung der (zufälligen) Anzahl der erzielten Erfolge statt der $\mathrm{B}_{n,p}$ Binomalverteilung die Poi_{np}-Poissonverteilung verwenden. Dies ist zunächst einmal ein großer rechnerischer Vorteil, da die in der Binomialverteilung vorkommenden Binomialkoeffizienten gerade für großes n sehr unhandlich werden.

b) Aus dem in a) erläuterten Grund heißt die Poisson-Verteilung auch manchmal **„Verteilung der seltenen Ereignisse“**. Ihr Auftreten als Grenzwert von Verteilungen, ihre sehr guten analytischen Eigenschaften (z.B. Lemma 2.9) und ihre einfache Zähldichte machen sie zur vermutlich zweitwichtigsten speziellen Verteilung der W-Theorie - die wichtigste ist mit großem Abstand die Normalverteilung, die wir bald kennen lernen werden.

c) **Beispiel: Versicheurung.** n ist die Anzahl de Versicherten, p die Schadenswahrscheinlichkeit für jeden Versicherten in einem Jahr. Weitere Annahmen sind Unabhängigkeit der Schadensereignisse, sowie höchstens ein Schadensereignis pro Versichertem (z.B.: Lebensversicherung). In der Praxis ist typischerweise n sehr groß und p sehr klein, also darf man modellieren:

$$\mathbb{P}(\ k \text{ Schäden in diesem Jahr }) = \binom{n}{k} p^k (1-p)^{n-k} \approx \mathrm{e}^{-pn}\, \frac{(pn)^k}{k!}.$$

d) **Beispiel: radioaktiver Zerfall.** In einem gegebenen Zeitintervall registriert ein Geigerzähler eine zufällige Anzahl von N radioaktiven Zerfällen (Klicks). Ein quantenmechanisches Modell der Atome besagt, dass jedes einzelne Atom in jedem winzigen Zeitintervall aufgrund von *Quantenfluktuationen* eine sehr kleine Chance hat, eine gewisse Energiebarriere zu überwinden und zu zerfallen. Wenn es dies nicht schafft, so befindet es sich zu Beginn des nächsten kleinen Zeitintervalls wieder im *Grundzustand*, was bedeutet, dass es seine gesamte vergangene Bewegung vergessen hat und im kommenden Zeitintervall einen unabhängigen Versuch macht, die Energiebarriere zu überwinden.
Wir beginnen mit einem Zeitintervall $[0, t_0]$. In der Probe sollen sich ein sehr große Anzahl $M \gg 1$ Atome finden, jedes davon kann genau einmal zerfallen und tut dies mit Wahrscheinlichkeit $p_0 = p_0(t_0) \ll 1$. Die zufällige Gesamtzahl N_{t_0} der Zerfälle ist daher gegeben als Summe von M unabhängigen, B_{p_0}-verteilten Zufallsvariablen, und daher

$$\mathbb{P}(N_{t_0} = k) \approx \mathrm{Poi}_{Mp_0}(\{k\}).$$

Nun wählen wir t_0 sehr klein und betrachten ein größeres Zeitintervall $t = [0, nt_0]$ mit $n \in \mathbb{N}$. Da die Gesamtzahl der Atome M sich im Laufe unserer Beobachtung nur unwesentlich verringert (wenige Zerfälle im Vergleich zu M), und da jedes Atom nach jedem Zeitintervall der Länge t_0 wieder im Grundzustand ist, ist dann die Zahl der zum Zeitpunkt t zerfallenen Atome eine Summe aus n unabhängigen Poi_{Mp_0}–verteilten ZVen. Wegen Lemma (2.9) ist daher

$$\mathbb{P}(N_t = k) \approx \mathrm{Poi}_{nMp_0}(\{k\}) \overset{n=t/t_0}{=} \mathrm{Poi}_{t\alpha}(\{k\}) \qquad \text{mit } \alpha = Mp_0/t_0$$

Zunächst gilt dies nur wenn t ein ganzzahliges Vielfaches von t_0 ist, aber indem man t_0 „beliebig klein“ macht, kann man es (nicht-rigoros) für alle $t \in \mathbb{R}$ postulieren. Eine rigorose Begründung werden wir (eventuell) erst später kennen lernen.

Nach unserem Modell sollte also die Anzahl der in $[0,t]$ gemessenen Klicks einer Poisson-Verteilung folgen, deren Parameter proportional zur Beobachtungszeit t ist (die Proportionalitätskonstante α wird auch die *Rate* des Zerfalls genannt). Man kann nun im Experiment (mit Hilfe der Statistik) nachmessen, dass dies tatsächlich der Fall ist, und damit verifizieren, dass die gemachten Modellannahmen zumindest gut genug sind, um radioaktiven Zerfall auf atomarem Niveau zu erklären. Außerdem kann man, ebenfalls mittels Statistik, die Rate α aus Experimenten messen und so z.B. Erkenntnisse über die Größe der relevanten Energiebarriere gewinnen.
Eine Frage zum Nachdenken: angenommen, man hat zwei verschiedene radioaktive Substanzen (mit Raten α_1 und α_2) in einer Probe. Ist es durch Beobachtung der Anzahlen von Klicks in beliebig vielen und beliebig gewählten Zeitintervallen möglich, dies zu erkennen und evtl. sogar α_1 und α_2 zu schätzen?

(2.11) Asymptotische Gestalt der Poisson-Verteilung für große Parameter

Wir wollen wissen, welche Werte eine Poi_λ-verteilte ZV „bevorzugt“ annimmt, wenn λ sehr groß ist. Dies bedeutet, dass wir die Funktion

$$k \mapsto p_\lambda(k) = \mathrm{e}^{-\lambda}\,\frac{\lambda^k}{k!}$$

für sehr große λ untersuchen müssen. Kurzes Nachdenken ergibt, dass in diesem Fall $p_\lambda(k) \approx 0$ für kleine k, denn dann ist $\mathrm{e}^\lambda \gg \lambda^k$, und somit $\mathrm{e}^{-\lambda}\,\lambda^k \ll 1$. Ebenso ist $p_\lambda(k) \approx 0$ für sehr große k, denn dann ist $k! \gg \lambda^k$ und somit $\lambda^k/k! \ll 1$. Als Funktion von k sollte also p_λ sein Maximum irgendwo weit weg von $k=0$ und (natürlich) auch nicht bei $k=\infty$ annehmen - letzteres ist ohnehin klar, denn die Folge $(p_\lambda(k))_{k\in\mathbb{N}_0}$ ist ja summierbar. Um dieses Maximum zu finden, müssen wir den Ausdruck $k!$ besser verstehen, und das Allheilmittel hierfür ist **Stirlings Formel**, die wir in der folgenden für alle k gültigen Form benutzen:

$$\sqrt{2\pi}k^{k+\frac12}\,\mathrm{e}^{-k}\,\mathrm{e}^{\frac{1}{12k+1}} \leqslant k! \leqslant \sqrt{2\pi}k^{k+\frac12}\,\mathrm{e}^{-k}\,\mathrm{e}^{\frac{1}{12k}}\,.$$

Der Beweis (einer Version) von Stirlings Formel kommt eventuell in den Übungen, hier benutzen wir sie einfach.
Da $\mathrm{e}^{\frac{1}{12k+1}} \approx \mathrm{e}^{\frac{1}{12k}} \approx 1$ ist, lassen wir diesen Term in Zukunft einfach weg und betrachten statt ρ_k nur noch die Größe

$$\tilde{p}_\lambda(k) = \frac{\mathrm{e}^{-\lambda}\,\lambda^k}{\sqrt{2\pi}k^{k+\frac12}\,\mathrm{e}^{-k}}$$

Die Rückschlüsse auf das wahre p_λ sind nicht schwer zu ziehen und bleiben als Übung. Wie so oft ist es einfacher, den logarithmierten Ausdruck zu untersuchen, der wegen der Monotonie der Exponentialfunktion genau dort maximal ist, wo auch $\tilde{p}_\lambda$ maximal ist:

$$\log\tilde{p}_\lambda(k) = -\lambda + k\log\lambda - \log\sqrt{2\pi} - (k+\frac12)\log k + k.$$

Zum Auffinden der Extremalstelle differenzieren wir nach k und setzen gleich Null:

$$0 = \log\lambda - \log k - \frac{k+\frac12}{k} + 1 = \log\frac{\lambda}{k} - \frac{1}{2k}.$$

Diese Gleichung ist zwar nicht in geschlossener Form lösbar, da wir aber an sehr großen λ interessiert sind, muss $k \approx \lambda$ sein, um den Betrag es ersten Terms nicht zu groß werden zu

lassen (der zweite ist ja immer betragsmäßig $\leqslant 1$), und dann kann man $\frac{1}{2k}$ vernachlässigen. Das Maximum von $\tilde{p}_\lambda(k)$ liegt also ungefähr bei $k = \lambda$, und dort ist

$$p_\lambda(\lambda) \approx \tilde{p}_\lambda(\lambda) = \frac{\mathrm{e}^{-\lambda}\,\lambda^\lambda}{\sqrt{2\pi}\lambda^{\lambda+\frac{1}{2}}\,\mathrm{e}^{-\lambda}} = \frac{1}{\sqrt{2\pi\lambda}}.$$

Wir stellen also fest, dass (für große λ) der Wert von p_λ am Maximum nicht wirklich groß ist, sondern die Größenordnung von $1/\sqrt{\lambda}$ hat. Um auf eine Gesamtwahrscheinlichkeit von der Größenordnung 1 zu kommen, müssen in einer Menge $A \subset \mathbb{N}$ also mindestens etwa $\sqrt{\lambda}$ Punkte enthalten sein. Vermutlich sollten diese Punkte auch nicht zu weit weg vom Ort $k = \lambda$ des Maximums liegen.
Um diese Vermutung noch etwas zu untermauern, betrachten wir noch $r \in \mathbb{Z}$ mit $|r| \ll \lambda$, und setzen dann $\lambda + r$ in $\log \tilde{p}_\lambda$ ein:

$$\log \tilde{p}_\lambda(\lambda + r) = -\lambda + (\lambda + r)\log\lambda - \log\sqrt{2\pi} - (\lambda + r + \frac{1}{2})\log(\lambda + r) + (\lambda + r) =$$
$$= r - \log\sqrt{2\pi} + (\lambda + r)\log\frac{\lambda}{\lambda + r} - \frac{1}{2}\log(\lambda + r) = (*).$$

Die Taylor-Entwicklung und die Annahme $|r| \ll \lambda$ ergeben

$$\log\frac{\lambda}{\lambda + r} = \log\Big(1 - \frac{r}{\lambda + r}\Big) \approx -\frac{r}{\lambda + r} - \frac{r^2}{2(\lambda + r)^2}.$$

Somit ist

$$(*) = -\log\sqrt{2\pi} - \frac{r^2}{2(\lambda + r)} - \frac{1}{2}\log(\lambda + r).$$

Wählt man nun ein festes $c \in \mathbb{R}$, dann ist, für hinreichend große λ, $r = c\sqrt{\lambda} \ll \lambda$, und wir erhalten

$$p_\lambda(\lambda + c\sqrt{\lambda}) \approx \tilde{\rho}_\lambda(\lambda + c\sqrt{\lambda}) \leqslant \mathrm{e}^{-\frac{c^2\lambda}{2(\lambda + c\sqrt{\lambda})}} \leqslant \mathrm{e}^{-\frac{c^2}{4}}.$$

Daraus sieht man, dass $p_\lambda(k)$ exponentiell klein wird, wenn k mehr als etwa $\sqrt{\lambda}$ von $k_{\max} = \lambda$ entfernt ist. p_λ ist also in einem „Streifen“ der Breite $\sqrt{\lambda}$ um $k_{\max}$ ungefähr (wir haben eigentlich nur gezeigt: höchstens) von der Größe $1/\sqrt{\lambda}$, außerhalb dieses Streifens jedoch vernachlässigbar klein. Man kann das so zusammenfassen: für $X \sim \mathrm{Poi}_\lambda$ und alle $q < 1/2$ gilt

$$\lim_{\lambda\to\infty} \mathbb{P}\Big(\Big|\frac{X}{\lambda} - 1\Big| \geqslant \frac{1}{\lambda^q}\Big) = 0.$$

(2.12) Poissonisierung

Unter Poissonisierung versteht man eine Technik, die es erlaubt, durch eine „Aufweichung“ von harten Randbedingungen aus einer Familie von abhängigen Zufallsvariablen eine Familie unabhängiger ZVen zu machen. Da letztere meist viel einfacher zu untersuchen sind, ist dies ein sehr großer Vorteil. Das Prinzip der „Aufweichung“ kann man in Analogie zu den Lagrange-Multiplikatoren in der Analysis sehen, die zum Auffinden von Minima unter Nebenbedingungen benutzt werden.

Die Poissonisierung soll hier am Beispiel der Multinomialverteilung erklärt werden. Wir erinnern uns, dass für eine endliche Menge Λ, und für eine Funktion $p : \Lambda \to [0,1]$ mit $\sum_{x\in\Lambda} p_x = 1$ auf dem Raum

$$\Omega_N := \{\eta : \Lambda \to \mathbb{N}_0 : \sum_{x\in\Lambda} \eta_x = N\}$$

der Besetzungszahlen, die Multinomialverteilung $\mathbb{P}_N$ durch ihre Zähldichte

$$m_N(\eta) = \frac{N!}{\prod_{x\in\Lambda} \eta_x!} \prod_{x\in\Lambda} p_x^{\eta_x} = N! \prod_{x\in\Lambda} \frac{p_x^{\eta_x}}{\eta_x!}$$

definiert ist. In der letzten Formel sieht man, dass m_N eigentlich eine Produkt-Zähldichte ist, aber leider ist der W-raum Ω_N kein Produktraum. Daher sind die Zufallsvariablen η_x auch nicht unabhängig für verschiedene x; beispielsweise impliziert ja $\eta_x = N$ schon $\eta_y = 0$ für alle $y \neq x$, also ist $\mathbb{P}(\eta_x = N, \eta_y = N) = 0 < \mathbb{P}(\eta_x = N)\mathbb{P}(\eta_y = N)$.

Die angesprochene „Aufweichung“ besteht nun darin, dass man nicht auf der Bedingung $\sum_{x\in\Lambda} \eta_x = N$ besteht, sondern nur fordert, dass sie „ungefähr“ gilt. In (2.11) haben wir gesehen, dass für große N (und meist sind dies diejenigen, die uns interessieren) gilt: $\mathrm{Poi}_N(X/N \approx 1) \approx 1$. Wir definieren daher

$$\Omega := \{\eta : \Lambda \to \mathbb{N}_0 : \sum_{x\in\Lambda} \eta_x < \infty\},$$

den (abzählbaren!) Raum aller endlichen Besetzungszahlen. Wir setzen $\mathcal{F} = \mathcal{P}(\Omega)$ und definieren, für $\lambda > 0$, das poissonisierte Maß $\tilde{\mathbb{P}}_\lambda$ durch seine Zähldichte

$$\tilde{q}_\lambda(\eta) := \sum_{n=0}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^n}{n!} m_n(\eta) \mathbb{1}_{\{\sum_{x\in\Lambda} \eta_x = n\}};$$

hierbei ist wieder die Konvention, dass die 0, die durch $\mathbb{1}_{\{\sum_{x\in\Lambda} \eta_x = n\}}$ im Fall $\sum_{x\in\Lambda} \eta_x \neq n$ erzeugt wird, gegen die dann nicht definierte Größe $m_n(\eta)$ gewinnt und der Term gleich 0 ist. Die obige Summe hat also immer nur einen Term.

Weiter ist

$$\tilde{\mathbb{P}}_\lambda(\sum_{x\in\Lambda} \eta_x = k) = \sum_{\eta\in\Omega:\sum_x \eta_x = k} \tilde{\mathbb{P}}_\lambda(\{\eta\}) = e^{-\lambda} \frac{\lambda^k}{k!},$$

und daher hat man für $\lambda = N \gg 1$ mit hoher W'keit etwa N „Teilchen“ im System. Andererseits ist

$$\mathbb{P}_N(\{\eta\}) = \tilde{\mathbb{P}}_\lambda(\{\eta\} \mid \sum_{x\in\Lambda} \eta_x = N)$$

für alle λ und alle N, also ist $\mathbb{P}_N$ tatsächlich die „harte“ Bedingung auf Gesamtbesetzungszahl N, und zwar für alle λ, nicht nur für $\lambda \approx N$. Das besondere an $\lambda \approx N$ ist nur, dass in diesem Fall mit hoher Wahrscheinlichkeit das Maß $\tilde{\mathbb{P}}_\lambda$ schon „von selbst“ Konfigurationen η auswürfelt, deren Gesamtbesetzungszahl (in einem gewissen Sinn) nicht so unterschiedlich zu N ist.

Die wichtigste Eigenschaft von $\tilde{\mathbb{P}}_N$ ist jedoch die, dass für alle η mit $\sum_{x\in\Lambda} \eta_x = N$ gilt:

$$\tilde{q}_\lambda(\eta) = \mathrm{e}^{-\lambda} \frac{\lambda^N}{N!} \lambda^N m_N(\eta) = \mathrm{e}^{-\lambda \sum_{x\in\Lambda} p_x} \frac{\lambda^{\sum_{x\in\Lambda} \eta_x}}{N!} N! \prod_{x\in\Lambda} \frac{p_x^{\eta_x}}{\eta_x!} = \prod_{x\in\Lambda} \mathrm{e}^{-\lambda p_x} \frac{(\lambda p_x)^{\eta_x}}{\eta_x!}.$$

Da die rechte Seite nicht mehr explizit von N abhängt, gilt diese Formel für alle $\eta \in \Omega$, und somit ist $\tilde{\mathbb{P}}_\lambda = \bigotimes_{x\in\Lambda} \mathrm{Poi}_{\lambda p_x}$, mit anderen Worten, die η_x sind nun unabhängig und selbst Poisson-verteilt mit Parametern λp_x. Der Parameter λ regelt „weich“ die erwünschte Gesamtbesetzungszahl.

Stetige Verteilungen

(2.13) Motivation und Vorüberlegungen

a) In Beispiel d) zu (2.10) haben wir für die Wahrscheinlichkeit, dass in einem *festen* Zeitintervall $[0, t]$ die *zufällige* Anzahl N_t der radioaktiven Zerfälle den Wert k annimmt, die Formel

$$\mathbb{P}(N_t = k) = \mathrm{Poi}_{\alpha t}(\{k\})$$

hergeleitet. Wir können aber natürlich auch die umgekehrte Frage stellen: wie groß ist die *zufällige* Zeit, zu der (für *festes* k) der k-te Zerfall gemessen wird?

b) Ähnliche Probleme sind: der Wert einer zufällig gemessenen Temperatur oder einer anderen physikalischen Größe, oder die Länge der Kreissehne, die eine zufällig auf ein Papier geworfene (lange) Nadel mit einem dort gezeichneten Kreis bildet, falls sie den Kreis trifft (Bertrand-Problem, siehe unten).

c) Oft ist es auch analytisch einfacher, diskrete Größen durch reelle zu approximieren. Dies ist beispielsweise bei Aktienkursen der Fall.

d) Um diese Fragen mathematisch untersuchen zu können, müssen wir uns zunächst Wahrscheinlichkeitsmaße auf $\mathbb{R}$ (bzw. Zufallsvariablen mit einer nicht diskreten Wertemenge) verschaffen; natürlich schließt unsere ursprüngliche Definition solche W-Maße nicht aus, Ω war ja einfach eine Menge. Allerdings müssen wir bei der Definition schon etwas mehr aufpassen, denn

(i): Vermutlich ist in den meisten interessanten Fällen $\mathbb{P}(\{x\}) = 0$ für alle x; das kennen wir schon von Folgenräumen (z.B. einfache Irrfahrt mit unendlich vielen Schritten).
(ii): ist leider $\mathcal{P}(\mathbb{R})$ viel zu groß. Dank des Auswahlaxioms enthält es Mengen, die man sich nicht vorstellen kann oder will, und die verhindern, dass man vernünftige W-Maße auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ definieren kann. Wer sich dafür interessiert, kann im Buch von Georgii den S Satz (1.5) nachschlagen oder bei Wikipedia nach „Satz von Vitali“ suchen. Ein noch drastischeres Beispiel für Auswahlaxiom-Mengen, die sehr unintuitive Eigenschaften haben, findet man unter dem Stichwort „Banach Tarski Paradox“; hierzu braucht man jedoch mindestens den $\mathbb{R}^3$.

Das Problem (ii) wird pragmatisch gelöst, indem man die σ-Algebra kleiner macht.

(2.14) Definition und Satz

a) Die von den halbunendlichen, offenen Intervallen $\{(-\infty, a) : a \in \mathbb{R}\}$ erzeugte σ-Algebra

$$\mathcal{B} \equiv \mathcal{B}(\mathbb{R}) := \sigma(\{(-\infty, a) : a \in \mathbb{R}\})$$

heißt **Borel'sche σ-Algebra**, oder **Borel-σ-Algebra**.

b) Es gilt:

$$\mathcal{B} = \sigma(\{A \subset \mathbb{R} : A \text{ offen}\}) = \sigma(\{C \subset \mathbb{R} : C \text{ abgeschlossen}\}) = \sigma(\{(-\infty, c] : c \in \mathbb{R}\}).$$

c) Für (beliebiges!) $\Omega_0 \subset \mathbb{R}$, $\Omega_0 \neq \emptyset$ ist

$$\mathcal{B}_{\Omega_0} := \{A \cap \Omega_0 : A \in \mathcal{B}\}$$

eine σ-Algebra. Sie heißt **Spur** von $\mathcal{B}$ auf Ω_0.
Beweis: Übung; für b) verwende etwa $[a,b) = (-\infty, b) \setminus (-\infty, a)$, und $[a,b] = \bigcap_{n=1}^{\infty} [a, b+1/n)$, und ähnliches. □

(2.15) Nomenklatur

Eine ZV mit Zielraum $(\mathbb{R}, \mathcal{B})$ heißt **reelle Zufallsvariable**. Eine ZV mit Zielraum $(\mathbb{R} \cup \{\infty\} \cup \{-\infty\}, \sigma(\mathcal{B} \cup \{\infty\} \cup \{-\infty\}))$ heißt **numerische Zufallsvariable**.

(2.16) Definition

a) μ sei ein W-Maß auf $(\mathbb{R}, \mathcal{B})$. Die Abbildung

$$F_\mu : \mathbb{R} \to [0,1], \qquad x \mapsto \mu((-\infty, x])$$

heißt **Verteilungsfunktion** (VF) von μ. Oft schreibt man auch nur F statt F_μ.
b) Sei X eine reelle ZV. Dann heißt

$$F_X : \mathbb{R} \to [0,1], \qquad x \mapsto F_X(x) := \mathbb{P}(X \leqslant x)$$

Verteilungsfunktion (VF) von X. Man schreibt oft auch F statt F_X.
Bemerkung: Klarerweise ist $F_X = F_{\mathbb{P}_X}$, die beiden Definitionen sind also ein wenig redundant.

(2.17) Satz

Sei X eine reelle ZV, F ihre Verteilungsfunktion. Dann gilt:

a) F ist monoton wachsend.

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

c) F ist stetig von rechts, d.h. für $(x_n) \subset \mathbb{R}$ und $x \in \mathbb{R}$ mit $x_n \geqslant x_{n+1}$ für alle n und $\lim_{n\to\infty} x_n = x$ gilt:

$$\lim_{n \to \infty} F(x_n) = F(x).$$

d) F hat linksseitige Grenzwerte, d.h. für $(x_n) \subset \mathbb{R}$ und $x \in \mathbb{R}$ mit $x_n \leqslant x_{n+1}$ für alle n und $\lim_{n\to\infty} x_n = x$ gilt:

$$F(x^-) := \lim_{n \to \infty} F(x_n) \text{ existiert.}$$

e) Es gelten für alle $x < y$ die Gleichungen

$$F(y) - F(x) = \mathbb{P}(X \in (x,y]), \quad F(y^-) = \mathbb{P}(X < y), \quad F(y) - F(y^-) = \mathbb{P}(X = y).$$

Die analogen Aussagen gelten enstprechend für ein W-Maß μ und seine VF.
Beweis: Übung. □

(2.18) Definition

Eine Abbildung, die Bedingungen (2.17 a) - d)) erfüllt, heißt **Verteilungsfunktion**.

(2.19) Lesen von Verteilungsfunktionen

$F : \mathbb{R} \to [0,1]$ sei die Verteilungsfunktion einer Zufallsvariable X. Dann gilt:

a) Sprungstellen von F entsprechen Werten, die X mit positiver W'keit annimmt. Die entsprechende W'keit ist genau die Höhe des Sprunges.

b) Ist F stetig an einer Stelle x, so ist $\mathbb{P}(X = x) = 0$.

c) Intervalle, auf denen F konstant ist, werden von X gemieden: die W'keit, dass X Werte in einem solchen Intervall annimmt, ist gleich 0.

d) Insbesondere ist

$$\inf\{x \in \mathbb{R} : F(x) > 0\} = \sup\{x \in \mathbb{R} : \mathbb{P}(X < x) = 0\},$$

d.h. falls $F(a) = 0$ für ein $a > -\infty$, dann nimmt X nur mit W'keit 0 Werte kleiner als a an. Ebenso:

$$\sup\{x \in \mathbb{R} : F(x) < 1\} = \inf\{x \in \mathbb{R} : \mathbb{P}(X > x) = 0\},$$

d.h. falls $F(a) = 1$ für ein $a < \infty$, dann nimmt X nur mit W'keit 0 Werte größer als a an.

Alle Aussagen folgen ziemlich direkt aus der Definition und verbleiben als Übung.

(2.20) Satz

Die Verteilungsfunktion charakterisiert die Verteilung einer reellen ZV, das heißt: sind X, Y reelle ZVen mit $F_X = F_Y$, dann ist auch $\mathbb{P}_X = \mathbb{P}_Y$.

Beweis: Nach Definition der VF und nach Annahme ist

$$\mathbb{P}_X((-\infty, c]) = F_X(c) = F_Y(c) = \mathbb{P}_Y((-\infty, c]).$$

für alle $c \in \mathbb{R}$. Das Mengensystem $\{(-\infty, c] : c \in \mathbb{R}\}$ ist $\cap$-stabil, und erzeugt die Borel-σ-Algebra. Satz 1.52 liefert die Behauptung. □

Was wir bisher nicht geklärt haben ist, ob es zu gegebener Verteilungsfunktion F überhaupt eine Zufallsvariable X *gibt*, so dass $F = F_X$. Das müssen wir nachholen; hierzu zunächst:

(2.21) Definition und Satz

a) Sei $(\Omega, \mathcal{F})$ ein messbarer Raum. Eine σ-additive Abbildung $\mu : \mathcal{F} \to [0, \infty]$ heißt **Maß** auf $\mathcal{F}$. Ein Maß hat also genau die gleichen Eigenschaften wie ein W-Maß, außer dass $\mu(\Omega) \in \mathbb{R}^+ \cup \{\infty\}$ sein darf und nicht gleich 1 sein muss.

b) Es existiert ein eindeutiges Maß λ auf $\mathcal{B}(\mathbb{R})$ mit der Eigenschaft, dass $\lambda([a, b]) = b - a$ für alle $a < b$. Für dieses Maß und alle $a < b$ gilt dann

$$b - a = \lambda([a, b]) = \lambda((a, b]) = \lambda([a, b)) = \lambda((a, b)).$$

c) Das Maß λ aus b) heißt **Lebesgue-Maß** auf $\mathbb{R}$.

d) Für $A \in \mathcal{B}(\mathbb{R})$ heißt das Maß

$$\lambda|_A : \mathcal{B}_A \to [0, \infty], \quad B \to \lambda(B)$$

Lebesgue-Maß auf A. Für $\lambda(A) = 1$ ist dies ein W-Maß.
Der Beweis von b) ist Inhalt der Vorlesung Maßtheorie und wird hier nicht gemacht. Die Aussage in d) sollte klar sein.

(2.22) Satz

F sei eine Verteilungsfunktion in Sinn von Definition (2.18). Dann ist die Abbildung

$$X : (0,1) \to \mathbb{R}, \qquad x \mapsto \inf\{c \in \mathbb{R} : F(c) \geqslant x\}$$

$\mathcal{B}((0,1))$-$\mathcal{B}(\mathbb{R})$-messbar, und das Bildmaß von $\lambda|_{(0,1)}$ unter X hat die Verteilungsfunktion F.
Beweis: Wir prüfen zunächst, dass X tatsächlich nur Werte in $\mathbb{R}$ (also nicht $\pm\infty$) annimmt: Sei $u \in (0,1)$. Wegen $\lim_{y\to-\infty} F(y) = 0$ finden wir ein $\bar{c} \in \mathbb{R}$ mit $F(\bar{c}) \leqslant u$. Daher (und weil F monoton ist) ist dann

$$X(u) = \inf\{c \in \mathbb{R} : F(c) > u\} \geqslant \bar{c} > -\infty.$$

Ebenso schließt man $X(u) < \infty$ aus $\lim_{y\to\infty} F(y) = 1$.
Um zu zeigen, dass F die VF von X ist, rechnen wir $\mathbb{P}(X \leqslant a)$ aus. Dazu wieder $u \in (0,1)$, und $a \in \mathbb{R}$. Dann gilt:

$$X(u) \leqslant a \overset{\text{Def. of } X}{\iff} \inf\{c \in \mathbb{R} : F(c) \geqslant u\} \leqslant a \overset{F \text{ rechtsstetig}}{\iff} \min\{c \in \mathbb{R} : F(c) \geqslant u\} \leqslant a$$
$$\iff F(a) \geqslant u \iff u \leqslant F(a).$$

Das heißt:

$$X^{-1}((-\infty, a]) = \{u \in (0,1) : X(u) \leqslant a\} = \{u \in (0,1) : u \leqslant F(a)\} = (0, F(a)] \in \mathcal{B}((0,1)),$$

womit wir schon mal gezeigt haben, dass X tatsächlich die behauptete Messbarkeit hat (denn $\{(-\infty, a] : a \in \mathbb{R}\}$ erzeugt ja $\mathcal{B}(\mathbb{R})$). Indem wir auf diese Gleichung nun λ anwenden, sehen wir

$$\mathbb{P}(X \leqslant a) = \lambda(X^{-1}((-\infty, a]))) = \lambda((0, F(a)]) = F(a) - 0 = F(a).$$

Dies zeigt die Behauptung. □

Bemerkung: Wir wissen jetzt also: jede Verteilungsfunktion erzeugt genau ein W-Maß: höchstens eines wegen (2.21) und mindestens eines wegen (2.22). Es besteht also eine Eins-zu-eins-Beziehung zwischen VFen und W-Maßen auf $\mathbb{R}$, d.h. es gibt „genau so viele“ W-Maße auf $\mathbb{R}$ wie es Funktionen mit den Eigenschaften (2.17 a-d) gibt.

Beispiele:
a) $\lambda|_{[0,1]}$ ist ein W-Maß, seine VF ist

$$F(x) = \begin{cases} 0 & \text{falls } x \leqslant 0, \\ x & \text{falls } 0 < x < 1, \\ 1 & \text{falls } x \geqslant 1. \end{cases}$$

b) Ein alternativer W-Raum für die geometrische Verteilung: Erinnerung:

$$X \sim \mathrm{Geo}_p \quad \overset{\text{Def}}{\iff} \mathbb{P}(X = k) = (1-p)^k p \ \forall k \in \mathbb{N}_0.$$

Wir können das auch in $\mathbb{R}$ „einbetten“ und X als reelle ZV auffassen. Denn die Verteilungsfunktion

$$F : \mathbb{R} \to [0,1], \qquad x \mapsto \begin{cases} 0 & \text{falls } x < 0 \\ p\sum_{j=0}^{k-1}(1-p)^j & \text{falls } k-1 \leqslant x \leqslant k, k \in \mathbb{N} \end{cases}$$

hat bei $k \in \mathbb{N}_0$ jeweils einen Sprung der Höhe $p(1-p)^k$, beschreibt also das Bildmaß von X. Natürlicherweise definiert man X auf dem W-Raum $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0), \mathrm{Geo}_p)$, dann ist $X : \mathbb{N}_0 \to \mathbb{R}$ mit $X(k) = k$. Man kann jedoch auch die Abbildung aus Satz (2.22) benutzen: dann ist der W-Raum $((0,1), \mathcal{B}((0,1)), \lambda|_{(0,1)})$, und die Abbildung ist

$$Y : (0,1) \to \mathbb{R}, \quad s \mapsto Y(s) = \begin{cases} 0 & \text{falls } s < p, \\ k & \text{falls } \sum_{j=0}^{k-1} p(1-p)^j \leqslant s < \sum_{j=0}^{k} p(1-p)^j \end{cases}$$

Y ist also eine Stufenfunktion mit Sprunghöhe jeweils 1 und Sprungpunkten bei p, $(1-p)p$, $(1-p)^2 p$ etc.

c) Übung: man überlege sich, wie die Gleichverteilung auf $\{1, \dots, 6\}$ (Würfel) und die Poissonverteilung als Bildmaß einer ZV auf dem W-Raum $((0,1), \mathcal{B}((0,1)), \lambda|_{(0,1)})$, analog zu b), erzeugt werden können.

(2.23) Verteilungsfunktionen mittels Riemann-Integration

a) $\rho : \mathbb{R} \to \mathbb{R}$ sei Riemmann-integrierbar, und es gelte

$$\rho(x) \geqslant 0 \ \forall x \in \mathbb{R}, \qquad \text{und} \qquad \int_{-\infty}^{\infty} \rho(x)\,\mathrm{d}x = 1.$$

Dann ist durch

$$F : \mathbb{R} \to [0,1], \qquad x \mapsto \int_{-\infty}^{x} \rho(y)\,\mathrm{d}y$$

eine Verteilungsfunktion definiert. Für eine ZV mit dieser Verteilungsfunktion gilt: $\mathbb{P}(X \in [a,b]) = \int_a^b \rho(y)\,\mathrm{d}y$ für alle $a < b$.

b) In der Situation von a) heißt ρ **Dichte** des W-Maßes $\mathbb{P}_X$, oder **Verteilungsdichte** von X.

c) Sei F eine Verteilungsfunktion, stetig und stückweise stetig differenzierbar mit endlich vielen Stellen der Nicht-Differenzierbarkeit. Dann ist durch

$$\rho(x) = \begin{cases} F'(x) & \text{falls } F \text{ diff'bar bei x,} \\ 0 & \text{sonst} \end{cases}$$

die Dichte des durch F definierten W-Maßes gegeben.

Beweis:

a) F ist monoton wegen $\rho \geqslant 0$, $\lim_{x\to-\infty} F(x) = 0$ weil ρ integrierbar ist, und $\lim_{x\to\infty} F(x) = 1$ wegen $\int_{\mathbb{R}} \rho(y)\,\mathrm{d}y = 1$. Da F stetig ist, gelten auch ((2.17) c) und d).

c) Nach dem Hauptsatz der Differential- und Integralrechnung ist

$$F(b) - F(a) = \int_a^b F'(y)\,\mathrm{d}y = \int_a^b \rho(y)\,\mathrm{d}y,$$

falls zwischen a und b keine Stelle liegt, an der F nicht differenzierbar ist. Sind solche Stellen vorhanden, zerlegt man $[a, b]$ in endliche viele Stücke anhand dieser Stellen, macht die Rechnung auf jedem Stück, und summiert. Mit $a \to -\infty$ erhält man die Behauptung. □

(2.24) Die (stetige) Gleichverteilung

a) Seien $a, b \in \mathbb{R}$, $a < b$. Das W-Maß $\mathbb{P}_U : \mathcal{B}([a, b]) \to [0, 1]$, welches durch die Bedingungen

$$\mathbb{P}_U([c, d]) = \frac{d - c}{b - a} \quad \forall c, d \in [a, b] \text{ mit } c < d$$

eindeutig bestimmt ist, heißt **Gleichverteilung** auf $[a, b]$.

b) Die Gleichverteilung hat die VF F_U mit

$$F_U(x) = \begin{cases} 0 & \text{für } x \leqslant a \\ \frac{x-a}{b-a} & \text{für } a \leqslant x \leqslant b \\ 1 & \text{für } x > b. \end{cases}$$

und Dichte $\rho_U(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$.

c) Die Gleichverteilung ist translationsinvariant auf $[a, b]$, d.h. falls $c < d$ und $a \leqslant \min\{c, c + x\} \leqslant \max\{d, d + x\} \leqslant b$, dann ist $\mathbb{P}_U([c, d]) = \mathbb{P}_U([c + x, d + x])$. Das gleich gilt allgemeiner, wenn man ein Intervall „mit periodischen Randbedingungen“ verschiebt, d.h. der Teil der rechts aus $[a, b]$ herausrutscht kommt links wieder hinein; das ist allerdings etwas unhandlicher aufzuschreiben.

d) Auf unendlich langen Intervallen gibt es kein translationsinvariantes W-Maß, d.h. keines, das die in c) beschriebene Eigenschaft hat. Überlegen Sie sich, warum das so ist.

(2.25) Beispiel: Das Bertrand'sche Nadelproblem

Problemstellung: In einen Kreis vom Radius 1 sei ein gleichseitiges Dreieck einbeschrieben (das hat dann Seitenlänge $\sqrt{3}$). Nun werde eine Nadel zufällig auf das Papier geworfen, auf dem sich die Zeichnung befindet. Unter der Bedingung, dass die Nadel den Kreis trifft, berechne man die Wahrscheinlichkeit p, dass die Kreissehne, die durch die Nadel entsteht, länger ist als eine Seite des Dreiecks.

Lösung 1: Jede Sehne schneidet genau eine der Radius-Strecken vom Mittelpunkt zum Kreisrand senkrecht; nur wenn die Sehne genau durch den Mittelpunkt geht, stimmt das nicht, dies hat aber Wahrscheinlichkeit 0. Wegen der Symmetrie spielt der Winkel dieser Radius-Linie keine Rolle, und der Abstand des Schnittpunktes x vom Mittelpunkt ist zufällig. Um mit der Länge der Dreiecksseite zu vergleichen, dreht man nun das Dreieck auch so, dass seine Seite parallel zur Sehne ist. Siehe Abbildung A, oben. Aus geometrischen Gründen (ein gleichseitiges Dreieck hat einen Inkreis mit halbem Radius des Umkreises) ist daher $\mathbf{p = 1/2}$.

Lösung 2: Jede Sehne schneidet genau zwei Punkte auf der Kreislinie. Der erste Punkt kann wegen der Symmetrie beliebig und fest gewählt werden, der zweite ist dann zufällig auf der Kreislinie zu wählen. Wieder darf man das Dreieck so drehen, dass man es leicht hat, diesmal mit der Spitze auf den ersten Punkt (Abbildung A, unten). Die Sehne ist länger als die Dreiecksseite genau dann, wenn sie die dem Punkt A gegenüberliegende Seite schneidet. Es ist sofort klar, dass daraus folgt: $\mathbf{p = 1/3}$.

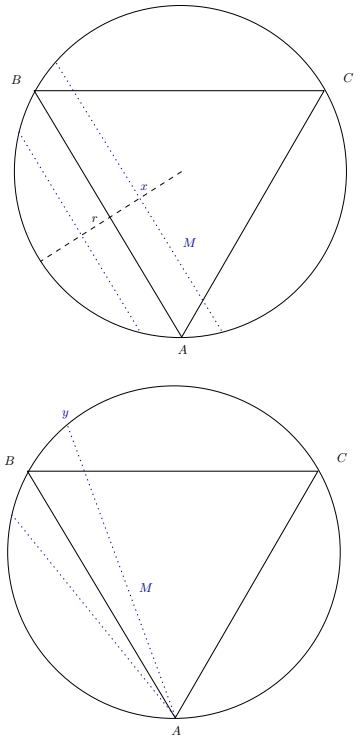


Abbildung A: Drei Möglichkeiten, eine Nadel „zufällig“ auf einen Kreis zu werfen. [4)]

Lösung 3: Jede Sehne hat genau einen Punkt x, der dem Mittelpunkt am nächsten ist. Da sie zufällig ist, ist es auch dieser Punkt. Die Wahrscheinlichkeit, dass er im Inkreis des gleichseitigen Dreiecks liegt, ist also proportional zum Quotient aus der Fläche des Inkreises und des Umkreises (Gleichverteilung!). Da der Inkreis den halben Radius, also ein Viertel der Fläche des Umkreises hat, bekommt man $\mathbf{p = 1/4}$.

Diskussion: Dieses „Paradoxon“ wurde im Jahr 1889 von Joseph Bertrand veröffentlicht und trug nicht zum Ansehen der Wahrscheinlichkeitstheorie als mathematische Wissenschaft bei. Bertrand argumentierte sogar, dass es wegen der oben gezeigten „Widersprüche“ sinnlos sei, Wahrscheinlichkeiten für nicht-diskrete Probleme zu formulieren.
Aus heutiger Sicht ist die Wurzel des Problems aber klar; sie liegt in der ungenauen Modellierung; zwar nimmt man in allen drei Lösungsansätzen die Gleichverteilung an, aber eben nicht die Gleichverteilung auf der gleichen Menge!
In Lösung 1 nimmt man die Gleichverteilung auf der Menge $[0, 1]$ der möglichen Schnittpunkte der Geraden mit der Radiuslinie des Kreises. Die relevante Zufallsvariable ist dann

$$X : [0, 1] \to \{0, 1\}, \qquad x \mapsto \mathbb{1}_{[1/2,1]}(x),$$

[4)] Die Abbildungen stammen aus den Lecture notes von Ariel Yadin, die ich sehr empfehlen kann: `https://www.math.bgu.ac.il/~yadina/teaching/probability/prob_notes.pdf`

d.h. X bildet den Punkt auf der Radiuslinie auf 1 ab, wenn die Sehne länger ist als die Seite des Dreiecks, sonst auf 0.
In Lösung 2 nimmt man die Gleichverteilung auf $[0, 2\pi]$ auf der Kreislinie. Die Zufallsvariable

$$X : [0, 2\pi) \to \{0, 1\}, \qquad \phi \mapsto \mathbb{1}_{[2\pi/3, 4\pi/3]}(\phi)$$

erfüllt den selben Zweck wie oben, hat aber natürlich eine andere Verteilung.
In Lösung 3 nimmt man die Gleichverteilung auf der Kreisscheibe $B_1(0) \subset \mathbb{R}^2$ (die wir erst im Kapitel 3 rigoros einführen werden). Die relevante Zufallsvariable ist dann

$$X : B_1(0) \to \{0, 1\}, \qquad (x, y) \mapsto \mathbb{1}_{[0,1/2]}\big(\sqrt{x^2 + y^2}\big).$$

Mathematisch sind wir jetzt fertig - es stellt sich (wenig überraschend) heraus, dass das Ergebnis davon abhängt, wie wir die warmen Worte „zufällig geworfene Nadel“ mathematisch modellieren.

Prinzipiell stellt sich aber natürlich weiterhin die Frage, welche Modellierung nun die angemessene ist, wenn wir ein „ideales Experiment“ voraussetzen; d.h. beispielsweise: der Akt des „Werfens“ der Nadel sollte möglichst nicht von der Geschicklichkeit des Experimentators beeinflusst werden! Hier gibt es verschiedene Meinungen (siehe Wikipedia-Artikel zu dem Thema!), meine eigene ist folgende: das Problem ist, dass wir (implizit) die *bedingte Wahrscheinlichkeit* ausrechnen wollen, dass die Sehne länger als die Dreiecksseite ist, unter der Bedingung, dass die Nadel den Kreis überhaupt trifft. Stellt man sich nun eine unendlich große (oder besser, sehr große und im Grenzwert unendliche) Fläche vor, auf der man eine unendlich lange Nadel (vulgo: eine Gerade) beispielsweise mit zufälligem Aufpunkt und Winkel plaziert, dann sieht man leicht, dass die W'keit, den Kreis zu schneiden, gleich Null ist (bzw. besser: mit der Größe des Gebietes gegen Null geht). Man bedingt hier also auf ein Ereignis der W'keit 0, was man zwar tun kann (wir können es noch nicht, werden es aber in den künftigen Semestern lernen), was aber wie man sieht nicht immer einfach ist.

Ein (m.E. vernünftiger) Ausweg besteht darin, *zuerst* die zufällige Nadel und dann (ebenfalls zufällig) einen Kreis auf eine große Fläche zu werfen: aus Symmetriegründen dürfen wir annehmen, dass die Nadel auf der x-Achse liegt. Der Vorteil ist nun, dass der später hingeworfene Kreis durch eine einzige Größe, nämlich seinen Mittelpunkt $(x, y) \in \mathbb{R}^2$, vollständig bestimmt ist. Zwar wird man auch hier (bei immer größeren Gebieten) die Linie immer öfter verfehlen, aber hier ist klar, was man damit meint, wenn man darauf bedingt, dass der Kreis die Nadel trifft: der Mittelpunkt muss sich im Abstand von 1 oder weniger von der x-Achse befinden. Da es egal ist, welchen x-Wert der Mittelpunkt hat, kann man $x = 0$ setzen. Man es also mit einer auf $[-1, 1]$ gleichverteilten y-Koodinate des Mittelpunktes zu tun. Die Sehne ist länger als die Dreiecksseite, wenn $|y| \geqslant 1/2$ ist, wie in Lösung 1. Die Lösung $p = 1/2$ ist also die richtige, jedenfalls ist es die einzige, die das „Umkehren der Reihenfolge der Erzeugung von Kreis und Linie“ überlebt.

(2.26) Die Exponentialverteilung

Sei $\alpha > 0$. Die Funktion

$$\rho_\alpha : \mathbb{R} \to \mathbb{R}_0^+, \qquad x \mapsto \rho_\alpha(x) := \alpha\, \mathrm{e}^{-\alpha x}\, \mathbb{1}_{[0,\infty)}(x)$$

ist eine Wahrscheinlichkeitsdichte. Das zugehörige W-Maß heißt **Exponentialverteilung** zum Parameter α. Für eine entsprechend verteilte ZV X schreiben wir $X \sim \mathrm{Exp}_\alpha$.

Bemerkungen:

a) Partielle Integration zeigt, dass ρ_α eine Dichte ist.

b) Die Verteilungsfunktion zur Exponentialverteilung ist identisch 0 für $x < 0$, und für $x \geqslant 0$ ist

$$F_\alpha(x) = \int_0^x \rho_\alpha(y)\,\mathrm{d}y = 1 - \mathrm{e}^{-\alpha x}\,.$$

Wir beobachten, dass für $X \sim \mathrm{Exp}_\alpha$ gilt:

$$\mathbb{P}(X \geqslant t) = \mathrm{e}^{-\alpha t} = \mathbb{P}(W = 0) \qquad \text{mit } W \sim \mathrm{Poi}_{\alpha t}.$$

Dies ist kein Zufall, wie wir jetzt sehen werden.

c) Dazu zeigen wir zunächst, dass die Exponentialverteilung der Grenzwert der (diskreten) Wartezeitverteilung Geo_p für kleine p und lange Wartezeiten ist, genauer: Sei $Y \sim \mathrm{Geo}_{p_n}$ mit $p_n = \alpha/n$, und sei $t_n = tn$. Dann gilt

$$\mathbb{P}(Y > t_n) = 1 - \mathbb{P}(Y \leqslant t_n) = 1 - \sum_{k=0}^{\lfloor t_n \rfloor} p_n(1-p_n)^k = 1 - p_n \frac{1-(1-p_n)^{\lfloor t_n\rfloor+1}}{1-(1-p_n)} =$$

$$= \Big(1 - \frac{\alpha}{n}\Big)^{\lfloor tn\rfloor+1} \overset{n\to\infty}{\longrightarrow} \mathrm{e}^{-\alpha t} = \mathbb{P}(X \geqslant t) \qquad \text{mit } X \sim \mathrm{Exp}_\alpha.$$

d) Nun sei $(Z_k)_{k\in\mathbb{N}} \sim (\mathrm{Ber}_{p_n})^{\otimes\mathbb{N}}$ mit $p_n = \alpha/n$ (also: $(Z_k)_{k\in\mathbb{N}}$ unabhängige Familie von ZVen mit $Z_k \sim \mathrm{Ber}_{p_n}$ für alle k). Dann gilt:

(i): Die W'keit, in den ersten $\lfloor nt \rfloor$ Versuchen keinen Erfolg zu haben, ist

$$\mathbb{P}(\sum_{j=1}^{\lfloor nt\rfloor} Z_j = 0) \overset{n\to\infty}{\longrightarrow} \mathrm{Poi}_{\alpha t}(\{0\}),$$

Konvergenz wegen (2.10).

(ii): Die W'keit, länger als $\lfloor nt \rfloor$ auf den ersten Erfolg zu warten, ist ebenfalls

$$\mathbb{P}(\sum_{j=1}^{\lfloor nt\rfloor} Z_j = 0) = \mathrm{Geo}_{p_n}(\{\lfloor nt\rfloor, \lfloor nt\rfloor + 1, \ldots\}) \overset{n\to\infty}{\longrightarrow} \mathrm{Exp}_\alpha([t,\infty))$$

wegen c). Also muss b) aus prinzipiellen Gründen richtig sein und ist kein algebraischer Zufall.

e) Die Exponentialverteilung ist *gedächtnislos* in folgendem Sinn: Ist $X \sim \mathrm{Exp}_\alpha$, $s, t > 0$, dann gilt

$$\mathbb{P}(X \geqslant t+s \,|\, X \geqslant t) = \frac{\mathrm{e}^{-\alpha(t+s)}}{\mathrm{e}^{-\alpha t}} = \mathrm{e}^{-\alpha s} = \mathbb{P}(X \geqslant s).$$

Interpretation: Wenn wir bis zur Zeit t auf das Ereignis gewartet haben, dessen Wartezeit X modelliert, und es ist nicht eingetreten, dann bekommen wir für die bereits gewartete Zeit keinerlei „Bonus“: die W'keit, dass es noch weitere s Zeiteinheiten dauert, bis das Ereignis eintritt, ist genauso hoch, als hätte man gerade erst angefangen zu warten.

Aufgabe: Definiere (Z_j) wie in d) und zeige, dass $S_n = \sum_{j=1}^n Z_j$ eine Markovkette ist (was ist

die Übergangsmatrix?). Schließe dann die Gedächtnislosigeit aus der Markov-Eigenschaft und einem Grenzübergang analog zu d).

(2.27) Die Gamma-Verteilung

Seien $\alpha, r > 0$. Die Funktion

$$\gamma_{\alpha,r} : \mathbb{R} \to \mathbb{R}_0^+, \qquad x \mapsto \gamma_{\alpha,r}(x) := \frac{\alpha^r}{\Gamma(r)} x^{r-1} \,\mathrm{e}^{-\alpha x}\, \mathbb{1}_{[0,\infty)}(x)$$

(mit $\Gamma(r) := \int_0^\infty y^{r-1} \,\mathrm{e}^{-y}\, \mathrm{d}y$, Gammafunktion) ist eine Wahrscheinlichkeitsdichte. Das zugehörige W-Maß heißt **Gamma-Verteilung** mit Parametern α und r. Für entsprechend verteilte ZVen schreibt man $X \sim \Gamma_{\alpha,r}$.

Bemerkungen:

a) $\gamma_{\alpha,r}$ ist eine W-Dichte, denn mit dem Koordinatenwechsel $y = \alpha x$ ist

$$\int_0^\infty x^{r-1} \,\mathrm{e}^{-\alpha x}\, \mathrm{d}x = \int_0^\infty \Big(\frac{y}{\alpha}\Big)^{r-1} \,\mathrm{e}^{-y}\, \frac{\mathrm{d}y}{\alpha} = \alpha^{-r}\Gamma(r),$$

also $\int_{-\infty}^\infty \gamma_{\alpha,r}(x)\,\mathrm{d}x = 1$.

b) Für $X \sim \Gamma_{\alpha,r}$, $r \in \mathbb{N}$ ist $\Gamma(r) = (r-1)!$, und

$$\mathbb{P}(X \leqslant t) = \frac{\alpha^r}{(r-1)!} \int_0^t x^{r-1} \,\mathrm{e}^{-\alpha x}\, \mathrm{d}x \overset{(*)}{=} \mathrm{e}^{-\alpha t} \sum_{k=r}^\infty \frac{(\alpha t)^k}{k!} = \mathrm{Poi}_{\alpha t}(\{r, r+1, \ldots\}).$$

Zum Beweis von $(*)$: den Fall $r = 1$ haben wir soeben in den Bemerkungen zu (2.26) behandelt, und für $r > 1$ definiere

$$v(s) := \mathrm{e}^{-\alpha s} \sum_{k=r}^\infty \frac{(\alpha s)^k}{k!},$$

dann ist $v(0) = 0$ und

$$\partial_s v(s) = -\alpha \,\mathrm{e}^{-\alpha s} \sum_{k=r}^\infty \frac{(\alpha s)^k}{k!} + \mathrm{e}^{-\alpha s}\, \alpha \sum_{k=r}^\infty \frac{(\alpha s)^{k-1}}{(k-1)!} = \alpha \,\mathrm{e}^{-\alpha s}\, \frac{(\alpha s)^{r-1}}{(r-1)!} = \frac{\alpha^r}{(r-1)!} \underbrace{s^{r-1} \,\mathrm{e}^{-\alpha s}}_{=:u(s)}.$$

Da auch $u(0) = 0$ ist, liefert der Hauptsatz der Differential- und Integralrechnung die Gleichheit $(*)$, indem man beide Seiten oben von 0 bis t integriert.

c) Punkt b) legt einen einfachen Zusammenhang zwischen Exp_α und $\Gamma_{\alpha,r}$ nahe. Im nächsten Kapitel werden wir mit Hilfe des Lebesgue-Integrals und der Faltungsformel für Verteilungen mit Dichte beweisen können, dass für unabhängige (X_j) mit $X_j \sim \mathrm{Exp}_\alpha$ für alle j gilt:

$$\sum_{j=1}^r X_j \sim \Gamma_{\alpha,r} \qquad \forall r \in \mathbb{N}.$$

Eine $\Gamma_{\alpha,r}$-verteilte ZV modelliert also die Wartezeit bis zum r-ten Ereignis bei einer Folge von Ereignissen, so dass Wartezeiten zwischen je zwei Ereignissen exponentialverteilt mit Parameter α sind.

Aufgabe: denken Sie darüber nach, wie man die Gamma-Verteilung sowie ihre Beziehung zur Exponentialverteilung und zur Poissonverteilung aus der nach (2.26) erwähnten Markovkette und einem Grenzübergang herleiten könnte.

(2.28) Die Normalverteilung

Sei $m \in \mathbb{R}$ und $v > 0$. Die Abbildung

$$\varphi_{m,v} : \mathbb{R} \to \mathbb{R}^+, \qquad x \mapsto \frac{1}{\sqrt{2\pi v}} \exp\Big(- \frac{(x-m)^2}{2v} \Big)$$

ist eine W-Dichte. Das zugehörige W-Maß heißt **Normalverteilung** (oder: **Gaußverteilung**) **mit Mittelwert** m **und Varianz** v. [5)] Für eine entsprechend verteilte ZV X schreibt man $X \sim \mathcal{N}_{m,v}$. Für $m = 0$ und $v = 1$ spricht man von der **Standard-Normalverteilung**.

Bemerkungen:

a) $\varphi_{m,v}$ ist eine W-Dichte, denn

$$\frac{1}{\sqrt{2\pi v}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{(x-m)^2}{2v}} \, \mathrm{d}x \overset{y=x-m}{=} \frac{1}{\sqrt{2\pi v}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{y^2}{2v}} \, \mathrm{d}y \overset{z=y/\sqrt{v}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z =: c,$$

und

$$2\pi c^2 = \int_{-\infty}^{\infty} \mathrm{d}x \, \mathrm{e}^{-\frac{x^2}{2}} \int_{-\infty}^{\infty} \mathrm{d}y \, \mathrm{e}^{-\frac{y^2}{2}} = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \, \mathrm{e}^{-\frac{x^2+y^2}{2}} = \qquad \text{(Polarkoordinaten)}$$
$$= \int_0^{2\pi} \mathrm{d}\beta \int_0^{\infty} \mathrm{e}^{-r^2} \, 2r \mathrm{d}r = -2\pi \Big[\mathrm{e}^{-\frac{r^2}{2}} \Big]_0^{\infty} = 2\pi.$$

Also folgt $c = 1$, die Behauptung.

b) Die Normalverteilung ist die mit Abstand wichtigste Verteilung der Stochastik. Warum das so ist erschließt sich im Moment noch nicht - hierzu müssen wir zunächst die Faltungsformel und später den zentralen Grenzwertsatz kennenlernen.

3. Integration, Erwartungswert, Momente

Lebesgue-Integral, W-Maße im $\mathbb{R}^n$, Faltungsformel

(3.1) Rekapitulation: Dichten und Integrale

ρ sei die W-Dichte eines W-Maßes $\mathbb{P}$ auf $\mathcal{B}(\mathbb{R})$. Stimmt es eigentlich, dass

$$\mathbb{P}(A) = \int_A \rho(x) \mathrm{d}x := \int_{-\infty}^{\infty} \rho(x) \mathbb{1}_A(x) \, \mathrm{d}x$$

gilt? Man möchte sofort und laut „ja“ sagen, aber lassen Sie uns kurz überlegen, was unsere Theorie bisher wirklich leistet. Wir hatten ρ als Riemann-integrierbar vorausgesetzt, und dann mittels $F(x) = \int_{-\infty}^{x} \rho(y) \mathrm{d}y$ eine Verteilungsfunktion konstruiert. Wegen (2.21) finden wir daher (mit Hilfe des Lebesgue-Maßes auf $\mathcal{B}((0,1))$, dessen Existenz wir nicht bewiesen haben, aber immerhin) eine ZV $X : (0,1) \to \mathbb{R}$ mit $\mathbb{P}_X((-\infty, x]) = F(x)$ für alle $x \in \mathbb{R}$. Da $\mathbb{P}_X$ als Bildmaß

[5)] Die Bedeutung der Begriffe Erwartungswert und Varianz wird im nächsten Kapitel klar werden.

einer ZV natürlich ein W-Maß ist, und da die halbunendlichen Intervalle jedes Maß eindeutig festlegen, wissen wir nun:

$$\mathbb{P}_X(A) \text{ existiert für alle } A \in \mathcal{B}(\mathbb{R}), \quad \text{und} \quad \mathbb{P}_X(A) = \int_A \rho(x)\,\mathrm{d}x \text{ für } A = (-\infty, x], x \in \mathbb{R}.$$

Mit den üblichen Methoden (Additivität, Stetigkeit von oben und unten) können wir die letzte Gleichheit auf viele weitere $A \in \mathcal{B}(\mathbb{R})$ (endliche Intervalle etc.) ausdehnen, aber leider gibt es in $\mathcal{B}(\mathbb{R})$ noch viel mehr Mengen; und die Riemann-Integration ist nicht gerade für ihre Flexibilität bekannt. Beispielsweise ist $\mathbb{Q} \in \mathcal{B}(\mathbb{R})$ (warum?), und aus unserer bisher erarbeiteten Theorie können wir schließen dass $\mathbb{P}_X(\mathbb{Q}) = 0$ (wieder: warum?) in obiger Situation. Aber die Funktion $\rho(x)\mathbb{1}_{\mathbb{Q}}(x)$ ist *nicht* Riemann-integrierbar, denn die Untersummen konvergieren nach 0 und die Obersummen nach 1 (bei Integration über $\mathbb{R}$). Das ist kein Zustand, den wir tolerieren können (bzw. wollen), daher brauchen wir ein besseres Integral. Wir machen das zunächst ganz allgemein.

(3.2) Das Integral im Sinne von Lebesgue

$(\Omega, \mathcal{F})$ sei ein messbarer Raum, μ ein Maß auf $\mathcal{F}$ (das Standardbeispiel ist $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}$ und $\mu = \lambda$, das Lebesgue-Maß).

a) Zielsetzung:

Unser Ziel ist es, ein „Integral“ $\int_\Omega f(x)\mu(\mathrm{d}x)$ für möglichst viele Funktionen f zu definieren. Zunächst sollten wir uns aber klarmachen, was wir von diesem Integral gerne hätten. Orientiert am Riemann-Integral verlangen wir:
(i): das Integral hängt mit dem Maß μ über die Flächeninhaltsformel

$$\int \mathbb{1}_A(x)\mu(\mathrm{d}x) = \mu(A) \qquad \forall A \in \mathcal{F}$$

zusammen.
(ii): Das Integral ist eine lineare Abbildung, die reelle Funkionen auf Zahlen abbildet:

$$\int_\Omega (f(x) + g(x))\mu(\mathrm{d}x) = \int_\Omega f(x)\mu(\mathrm{d}x) + \int_\Omega g(x)\mu(\mathrm{d}x).$$

Die Linearität ist notwendig für die Konsistenz von (i) falls $f = \mathbb{1}_A$ und $g = \mathbb{1}_B$ mit $A \cap B = \emptyset$, es ist aber natürlich sinnvoll, sie allgemein zu fordern.

b) Erinnerung an das Riemann-Integral:

Eine Intervallpartition eines Intervalls $[a, b)$ ist eine Menge

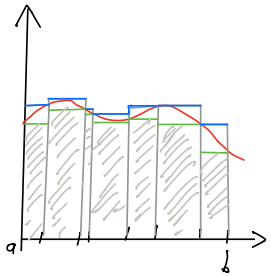
$$\mathcal{P} = \{[a_i, b_i) : i = 1, \dots, N, a_1 = a, b_i = a_{i-1}\ \forall\, 2 \leqslant i \leqslant N, b_N = b\}$$

Die Ober- und Untersummen einer Funktion $f : [a, b) \to \mathbb{R}$ zur Partition $\mathcal{P}$ sind

$$S^+_{\mathcal{P}} = \sum_{i=1}^{N} \sup\{f(x) : a_i \leqslant x < b_i\}|b_i - a_i|,$$

$$S^-_{\mathcal{P}} = \sum_{i=1}^{N} \inf\{f(x) : a_i \leqslant x < b_i\}|b_i - a_i|,$$

Das Riemann-integral existiert, falls für jede Folge $(\mathcal{P}_n)_{n\in\mathbb{N}}$ von Partitionen, deren *Gitterweite* $|\mathcal{P}_n| = \max\{|c-d| : [c,d) \in \mathcal{P}_n\}$ gegen Null konvergiert, die Ober- und Untersummen ebenfalls und gegen die gleiche Zahl konvergieren.



Ober- und Untersummen beim Riemann-Integal: vertikale Zerlegung der Fläche

Eines der Probleme, die diese Definition hat, ist, dass sie nicht gut verallgemeinert, wenn der Raum Ω, auf dem f definiert ist, keine Ordnungsrelation hat. Für $\mathbb{R}^d$ kann man das mit einiger Mühe noch reparieren, aber was sollte beispielsweise im Raum $\Omega = \mathbb{Z}^{\mathbb{N}}$, auf dem die einfache Irrfahrt definiert ist, die Intervallpartition ersetzen? Die (a posteriori naheliegende) Idee ist, die Ordnungsrelation des *Zielraumes* statt dessen zu verwenden. Das geht für alle reellwertigen Funktionen, egal, wo sie definiert sind, und zwar wie folgt:

c) Grundidee des Integrals:

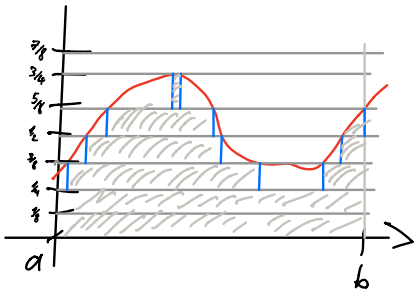
Da wir schon mit einem Maßraum starten, bekommen wir Forderung (i) aus a) geschenkt: Für *messbare* Teilmengen A ist die Funktion $\mathbb{1}_A$ messbar, und (i) gilt.
Die entscheidende Idee des allgemeinen Integrals ist es nun, die „Fläche“ unter einer Funktion f nicht vertikal, sondern horizontal zu zerlegen. Sei f messbar und (im Moment) $f(x) \geqslant 0$ für alle $x \in \Omega$.

Für alle $n, N \in \mathbb{N}_0$ definieren wir

$$A_{n,N} := \{x \in \Omega : f(x) \geqslant \frac{n}{2^N}\}.$$

Weil f messbar ist, sind es auch die Mengen $A_{n,N}$. Betrachte die Summe

$$J_N(f) := \sum_{n=1}^{\infty} \frac{1}{2^N}\mu(A_{n,N}) \overset{!}{=} \sum_{n=1}^{\infty} \frac{n}{2^N}\mu(A_{n,N} \setminus A_{n+1,N}).$$



Horizontale Diskretisierung: Der unterste Balken entspricht der zu $A_{1,3}$ gehörigen „Schicht“ der Pyramide, der nächste zu $A_{2,3}$ und so weiter.

Beide Ausdrücke veranschaulicht man sich, indem man sich vorstellt, dass Ω eine „Ebene mit ortsabhängiger Gravitation (gegeben durch μ)“ ist, auf der man eine Stufenpyramide baut. Im mittleren Ausdruck lässt man alle Stellen, an denen f nicht einmal den Wert 2^{-N} erreicht, unbebaut - der 0-te Term der Summe fehlt!
Dann baut man eine Schicht der Höhe 2^{-N} auf die Fläche $A_{1,N}$, wo f mindestens den Wert 2^{-N} erreicht. Das „Gewicht“ dieser Schicht ist $\mu(A_{1,N})2^{-N}$ („Volumen unter Berücksichtigung der Gravitation“). Auf diese Schicht baut man eine weitere oben drauf, und zwar auf die (kleinere) Fläche $A_{2,N}$, wo f mindestens die Höhe $2 \cdot 2^{-N}$ erreicht. Fährt man auf diese Weise fort, erhält man eine Stufenpyramide, die unterhalb von f liegt. Der rechte Ausdruck oben baut

diese Pyramide etwas anders auf - hier baut man zunächst nur auf die Stellen von A_1, die in der fertigen Pyramide von oben sichtbar sein (d.h. später nicht mehr bebaut) werden, eine Schicht der Höhe 2^{-N}. Dann baut man ein doppelte Schicht auf diejenigen Stellen von $A_{2,n}$, die später nicht mehr bebaut werden, dann eine dreifache auf die entsprechenden Stellen von $A_{3.n}$ und so weiter.

$J_N(f)$ beschreibt also die „horizontal diskretisierte Fläche unter f“. Man kann sich auch leicht überzeugen, das $J_{N+1}(f) \geqslant J_N(f)$ (man mach sich klar, welche Stellen man an der Stufenpyramide dazubauen muss, um von $J_{N+1}(f)$ nach $J_N(f)$ zu kommen).

Intuitiv ist klar, dass $J_N(f) \overset{N\to\infty}{\longrightarrow} \int f(x)\mu(\mathrm{d}x)$ gelten sollte. Formal definiert man etwas allgemeiner.

(3.3) Definition

$(\Omega, \mathcal{F}, \mu)$ sei ein Maßraum.

a) $g : \Omega \to \mathbb{R}$ heißt **elementar**, falls es $n \in \mathbb{N}$, eine Partition $A_1, \ldots, A_n$ von Ω mit $A_i \in \mathcal{F}$ für alle i, und Zahlen $\alpha_1, \ldots \alpha_n \geqslant 0$ gibt mit

$$g(x) = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}(x) \qquad \forall x \in \Omega.$$

b) g elementar habe die Darstellung wie oben. Die Abbildung

$$J_{\mathrm{el}} : \{g : \Omega \to \mathbb{R} : g \text{ elementar}\} \to \mathbb{R}, \qquad g \mapsto J_{\mathrm{el}}(g) \equiv \int_\Omega g(x)\mu(\mathrm{d}x) := \sum_{i=1}^{n} \alpha_i \mu(A_i)$$

heißt **Integral** von g bezüglich μ.

c) Die Abbildung

$$J : \{f : \Omega \to \mathbb{R}_0^+ : f \text{ messbar}\} \to \mathbb{R},$$
$$f \mapsto J(f) \equiv \int_\Omega f(x)\,\mu(\mathrm{d}x) := \sup\{J_{\mathrm{el}}(g) : g \text{ elementar und } g(x) \leqslant f(x)\ \forall x \in \Omega\}$$

heißt **Integral** von f bezüglich μ.

Bemerkungen:

a) Die *Darstellung* einer elementaren Funktion ist nicht eindeutig: wenn zwei oder mehr der α_i gleich groß sind, dann kann man die zugehörigen Mengen auch zusammenfassen; ebenso kann man vorhandene Mengen beliebig (messebar) zerteilen und gleiche α_i zuordnen. Es ist eine einfache Übung, sich klarzumachen, dass dies nichts am Wert von $J(g)$ ändert. Die Abbildung J ist also wohldefiniert, da sie zwei gleichen (aber anders dargestellten) elementaren Funktionen den gleichen Wert zuordnet.

b) Oft schreibt man auch $\int f\,\mathrm{d}\mu$ oder $\int f(x)\mu(\mathrm{d}x)$ statt $\int_\Omega f(x)\mu(\mathrm{d}x)$.

c) Es gilt mit J_N aus (3.2): $\lim_{N\to\infty} J_N(f) = J(F)$. Die Definition mit dem Supremum ist etwas flexibler, aber weniger anschaulich.

(3.4) Definition

$(\Omega, \mathcal{F}, \mu)$ sei ein Maßraum.

a) Für $f : \Omega \to \mathbb{R}$ setze $f^+(x) := \max\{f(x), 0\}$ (**Positivteil von** f) und $f^-(x) := -\min\{f(x), 0\}$ (**Negativteil von** f). Achtung: nach Konvention ist $f^-(x) \geqslant 0$.

b) Für f messbar mit $\int f^+ \mathrm{d}\mu < \infty$ oder $\int f^- \mathrm{d}\mu < \infty$ definiere das Integral von f als

$$\int f \,\mathrm{d}\mu = \int f^+ \,\mathrm{d}\mu - \int f^- \,\mathrm{d}\mu \in \mathbb{R} \cup \{-\infty, \infty\}.$$

Wir sagen dann, dass „das Integral existiert".

c) Falls f messbar und sowohl $\int f^+ \mathrm{d}\mu < \infty$ als auch $\int f^- \mathrm{d}\mu < \infty$, dann schreiben wir $f \in \mathcal{L}^1(\mu)$ und sagen „f ist integrierbar". In diesem Fall ist $|\int f \,\mathrm{d}\mu| < \infty$.

(3.5) Eigenschaften des Integrals

$(\Omega, \mathcal{F}, \mu)$ sei ein Maßraum. Dann gilt

a) Linearität: Für $f, g \in \mathcal{L}^1(\mu)$, $\alpha \in \mathbb{R}$ ist auch $f + \alpha g \in \mathcal{L}^1(\mu)$, und es gilt $\int (f + \alpha g) \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu + \alpha \int g \,\mathrm{d}\mu$.

b) Monotonie: Für $f, g \in \mathcal{L}^1(\mu)$ mit $f \leqslant g$ ist auch $\int f \,\mathrm{d}\mu \leqslant \int g \,\mathrm{d}\mu$.

c) Ungleichung für Betrag und Integral: Für $f \in \mathcal{L}^1$ gilt $|\int f \,\mathrm{d}\mu| \leqslant \int |f| \,\mathrm{d}\mu$.

d) Verschwinden außerhalb einer Menge von Maß Null: f sei messbar.
Falls $\mu(f \neq 0) = 0$, dann ist f integrierbar und $\int f \,\mathrm{d}\mu = 0$.
Falls $f \geqslant 0$ und $\int f \,\mathrm{d}\mu = 0$, dann ist $\mu(f > 0) = 0$. (Eine „Ausnahmemenge" Ω_0 mit $f(x) > 0$ für $x \in \Omega_0$ ist also erlaubt, solange $\mu(\Omega_0) = 0$ gilt.)

e) Menge der Stellen wo $f(x) = \infty$: Falls $\int f \,\mathrm{d}\mu < \infty$, dann ist $\mu(f = \infty) = 0$ (auch hier: Ausnahmemenge erlaubt!).

f) Verhältnis zum Riemann-Integral: Für $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{B}, \lambda)$ und Riemann-integrierbares f gilt: Falls $f \geqslant 0$ oder $f \in \mathcal{L}^1$, dann ist $\int f(x)\mu(\mathrm{d}x)$ gleich dem Wert des Riemann-Integrals. Auf der Schnittmenge der jeweils erlaubten Funktionen stimmen Riemann- und Lebesgueintegral also überein.

Beweis: Inhalt der Maßtheorie.

(3.6) Konvergenzsätze

$(\Omega, \mathcal{F}, \mu)$ sei ein Maßraum, (f_n) eine Folge messbarer reeller Funktionen.

a) Satz von der monotonen Konvergenz:
Es existiere ein $\Omega_0 \in \mathcal{F}$ mit $\mu(\Omega_0) = 0$, so dass für alle $x \in \Omega_0^c$ gilt: $f_{n+1}(x) \geqslant f_n(x)$ für alle n (d.h. die Folge (f_n) ist auf Ω_0^c monoton wachsend). Außerdem existiere das Integral von f_1, und es gelte $\int f_1 \mathrm{d}\mu > -\infty$. Dann gilt

$$\lim_{n\to\infty} \int f_n(x)\, \mu(\mathrm{d}x) = \int \limsup_{n\to\infty} f_n(x) \mu(\mathrm{d}x) \leqslant \infty$$

Auf der Menge Ω_0^c kann wegen der Monotonie der lim sup durch einen lim ersetzt werden.

b) Satz von der dominierten Konvergenz:

Es existiere ein $\Omega_0 \in \mathcal{F}$ mit $\mu(\Omega_0) = 0$, so dass für alle $x \in \Omega_0^c$ der Grenzwert $f(x) := \lim_{n\to\infty} f_n(x)$ existiert. Außerdem gebe es ein $g \in \mathcal{L}^1(\mu)$, so dass für alle $x \in \Omega_0$ gilt:

$$\sup_{n\in\mathbb{N}} |f_n(x)| \leqslant g(x).$$

Dann ist f (beliebig (messbar) fortgesetzt auf Ω_0) in $\mathcal{L}^1$, und es gilt

$$\lim_{n\to\infty} \int f_n(x)\mu(\mathrm{d}x) = \int f(x)\mu(\mathrm{d}x).$$

Beweis: Inhalt der Maßtheorie.

Bemerkung:
Beide Sätze geben Bedingungen an, unter denen Integral und Grenzwert vertauscht werden dürfen. Diese Bedingungen braucht man auch - es ist sehr nützlich, sich zu überlegen, was bei dieser Vertauschung eigentlich schief gehen kann. Hier gibt es zwei grundlegende Effekte: „Masse“ kann „ins Unendliche“ entkommen, oder sie kann sich „auf Mengen von Maß 0 konzentrieren“. Konkretere Beispiele machen wir in den Übungen. Das Lemma von Fatou (unten) ist die beste Aussage, die ohne weitere Voraussetzungen gilt.

Nomenklatur:
a) f sei eine messbare Funktion. Es existiere eine „Ausnahmemenge“ $\Omega_0 \in \mathcal{F}$ mit $\mu(\Omega_0) = 0$, so dass für alle alle $x \notin \Omega_0$ die Zahlen $f(x)$ eine gewisse Eigenschaft haben (siehe oben!). Wir sagen dann, f habe diese Eigenschaft **μ-fast überall**.
b) X sei eine Zufallsvariable. Es existiere eine „Ausnahmemenge“ $\Omega_0 \in \mathcal{F}$ mit $\mathbb{P}(\Omega_0) = 0$, so dass für alle alle $\omega \notin \Omega_0$ die Zahlen $X(\omega)$ eine gewisse Eigenschaft haben. Wir sagen dann, X habe diese Eigenschaft **$\mathbb{P}$-fast sicher**.

(3.7) Lemma von Fatou

(f_n) sei eine Folge messbarer Funktionen, und es gelte $f_n \geqslant 0$ μ-fast überall für alle n. Dann gilt

$$\int \liminf_{n\to\infty} f_n(x)\,\mu(\mathrm{d}x) \leqslant \liminf_{n\to\infty} \int f_n(x)\,\mu(\mathrm{d}x).$$

Beweis: Die Folge $(g_n)_{n\in\mathbb{N}}$ mit $g_n(x) := \inf_{m\geqslant n} f_m(x)$ ist monoton wachsend, und es gilt $g_n(x) \leqslant f_n(x)$ für alle n. Daher ist

$$\begin{aligned}\int \liminf_{n\to\infty} f_n(x)\,\mu(\mathrm{d}x) &= \int \lim_{n\to\infty} g_n(x)\,\mu(\mathrm{d}x) = \qquad \text{(Monotone Konvergenz)}\\ &= \lim_{n\to\infty} \int g_n(x)\,\mu(\mathrm{d}x) = \liminf_{n\to\infty} \int g_n(x)\,\mu(\mathrm{d}x) \leqslant\\ &\leqslant \liminf_{n\to\infty} \int f_n(x)\,\mu(\mathrm{d}x).\end{aligned}$$

□

(3.8) Satz von Fubini

$(\Omega_i, \mathcal{F}_i, \mu_i)$ seien Maßräume für $i = 1, 2$, und $(\Omega, \mathcal{F}, \mu) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mu_1 \otimes \mu_2)$ der Produktraum. $f : \Omega \to \mathbb{R}$ sei messbar, und es gelte

$$\text{entweder } f \geqslant 0 \ \mu\text{-fast überall}, \qquad \text{oder } f \in \mathcal{L}^1(\mu).$$

Dann gilt:

a) Die Abbildungen $J_1 f : \Omega_2 \to \mathbb{R} \cup \{\infty\}$ und $J_2 f : \Omega_1 \to \mathbb{R} \cup \{\infty\}$ mit

$$[J_1 f](x_2) = \int_{\Omega_1} f(x_1, x_2)\mu(\mathrm{d}x_1), \qquad [J_2 f](x_1) = \int_{\Omega_2} f(x_1, x_2)\mu(\mathrm{d}x_2)$$

sind messbar (bezüglich der jeweils relevanten σ-Algebra).

b) Es gilt (mit $x = (x_1, x_2) \in \Omega$):

$$\begin{aligned}\int_\Omega f(x)\mu(\mathrm{d}x) &= \int_{\Omega_2} J_1 f(x_2)\mu_2(\mathrm{d}x_2) \equiv \int_{\Omega_2} \Big(\int_{\Omega_1} f(x_1, x_2)\mu_1(\mathrm{d}x_1) \Big)\mu_2(\mathrm{d}x_2) = \\ &= \int_{\Omega_1} J_2 f(x_1)\mu_1(\mathrm{d}x_1) \equiv \int_{\Omega_1} \Big(\int_{\Omega_2} f(x_1, x_2)\mu_2(\mathrm{d}x_2) \Big)\mu_1(\mathrm{d}x_1).\end{aligned}$$

Beweis: Für das Lebesgue-Maß macht man den Beweis in der Maßtheorie; wir machen hier einen Beweis, der so glatt nur für W-Maße (oder endliche Maße) funktioniert, der aber eine wichtige Grundtechnik (π-λ-Theorem) wiederholt und eine andere („Leiter“ von elementaren über nichtnegative zu integrierbaren Funktionen) einführt.

Schritt 1: Zunächst beweisen wir die Aussage für $f(x) = \mathbb{1}_A(x)$ mit $A = A_1 \times A_2$, und $A_i \in \mathcal{F}_i$ (messbare Rechtecke). Hier gilt sie, denn $\mathbb{1}_{A_1 \times A_2}(x_1, x_2) = \mathbb{1}_{A_1}(x_1)\mathbb{1}_{A_2}(x_2)$, und daher

$$(J_1 \mathbb{1}_A)(x_2) = \int_{\Omega_1} \mathbb{1}_{A_1}(x_1)\mathbb{1}_{A_2}(x_2)\mu(\mathrm{d}x_1) = \mathbb{1}_{A_2}(x_2)\mu(A_1).$$

$J_1 \mathbb{1}_A$ ist also $\mathcal{F}_2$-messbar. Weiter folgt (wegen des Produktmaßes)

$$\int_\Omega \mathbb{1}_A(x)\mu(\mathrm{d}x) = \mu_1(A_1)\mu_2(A_2) = \mu(A_1) \int_{\Omega_2} \mathbb{1}_{A_2}(x_2)\mu(\mathrm{d}x_2) = \int_{\Omega_2} (J_1 \mathbb{1}_A)(x_2)\, \mu(\mathrm{d}x_2).$$

Mit Vertauschung der Indizes 1 und 2 folgt die andere Aussage.

Schritt 2: Wir zeigen nun die Aussage für alle Indikatorfunktionen. Wir erinnern uns zunächst, dass die Menge $\mathcal{R} := \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\} \subset \mathcal{F}$ der messbaren Rechtecke ein π-System ist. Wie schon früher definieren wir

$$\mathcal{G} := \{B \in \mathcal{F} : \text{ die Aussagen des Satzes gelten für } \mathbb{1}_B\}.$$

Nach Schritt 1 ist $\mathcal{R} \subset \mathcal{G}$. Wir wollen zeigen, dass $\mathcal{G}$ ein λ-System ist, das bedeutet für $A \in \mathcal{G}$ ist $A^c \in \mathcal{G}$, und disjunkte Vereinigungen von Mengen aus $\mathcal{G}$ sind in $\mathcal{G}$. Sei also $A \in \mathcal{G}$, dann ist

$$(J_1 \mathbb{1}_{A^c})(x_2) = \int (1 - \mathbb{1}_A(x_1, x_2))\mu_1(\mathrm{d}x_1) = \mu_1(\Omega_1) - (J_1 \mathbb{1}_A)(x_2) = 1 - (J_1 \mathbb{1}_A)(x_2).$$

An dieser Stelle kämen wir direkt nicht weiter, wenn $\mu_1(\Omega_1) = \infty$ wäre, da wir nicht ausschließen könnten, dass rechts $\infty - \infty$ steht. So aber sehen wir dass $J_1 \mathbb{1}_{A^c}$ messbar ist, denn $J_1 \mathbb{1}_A \in \mathcal{G}$

war ja vorausgesetzt. Ebenso sehen wir

$$\int_\Omega \mathbb{1}_{A^c}(x)\mu(\mathrm{d}x) = \int_\Omega (1-\mathbb{1}_A(x))\mu(\mathrm{d}x) = 1 - \int_{\Omega_2} (J_1\mathbb{1}_A)(x_2)\mu(\mathrm{d}x_2) = \int_{\Omega_2} (J_1\mathbb{1}_{A^c})(x_2)\mu(\mathrm{d}x_2).$$

Die letzte Gleichheit folgt aus der vorigen Rechnung.
Nun sei (A_i) eine disjunkte Folge von Mengen aus $\mathcal{G}$. Dann ist mit $A = \dot{\bigcup}_{j\in\mathbb{N}} A_j$

$$\Big(J_1\mathbb{1}_A\Big)(x_2) \overset{(A_j)\text{ dijunkt}}{=} \Big(J_1(\sum_{j\in\mathbb{N}} \mathbb{1}_{A_j})\Big)(x_2) \overset{\mu_1\text{ ist }\sigma\text{-additiv}}{=} \sum_{j\in\mathbb{N}} J_1\mathbb{1}_{A_j}(x_2).$$

Daher ist $J_1\mathbb{1}_A$ messbar als monotoner Grenzwert der messbaren Abbildungen $\sum_{j=1}^N J_1\mathbb{1}_{A_j}$. Außerdem gilt:

$$\int \mathbb{1}_A(x)\mu(\mathrm{d}x) = \int_\Omega \sum_{j\in\mathbb{N}} \mathbb{1}_{A_j}(x)\mu(\mathrm{d}x) \overset{\text{Mon. Konv.}}{=} \sum_{j\in\mathbb{N}} \int_\Omega \mathbb{1}_{A_j}(x)\mu(\mathrm{d}x) =$$

$$= \sum_{j\in\mathbb{N}} \int_{\Omega_2} J_1\mathbb{1}_{A_j}(x_2)\mu_2(\mathrm{d}x_2) \overset{\text{Mon. Konv.}}{=} \int_{\Omega_2} \sum_{j\in\mathbb{N}} J_1\mathbb{1}_{A_j}(x_2)\mu_2(\mathrm{d}x_2) = \int_{\Omega_2} J_1\mathbb{1}_A(x_2)\mu(\mathrm{d}x_2).$$

Damit ist gezeigt, dass $\mathcal{G}$ ein λ-System ist, und Satz (1.51) beendet Schritt 2.

Schritt 3: Wegen der Additivität der Integrale in allen Behauptungen gelten diese somit für alle elementaren Funktionen im Sinne von Definition (3.3 a).

Schritt 4: Sei nun $f(x) \geqslant 0$ für alle $x \in \Omega$. Es existiert eine monoton wachsende Folge (f_n)von elementaren Funktionen mit $\lim_{n\to\infty} f_n(x) = f(x)$ für alle x; in den Übungen konstruieren wir eine, in (3.2 c) ist ebenfalls eine angedeutet. Mit dem Satz von der monotonen Konvergenz folgt dann

$$J_1 f(x_2) = \lim_{n\to\infty} \int_{\Omega_1} f_n(x_1, x_2)\mu_1(\mathrm{d}x_1) = \lim_{n\to\infty} J_1 f_n(x_2),$$

$J_1 f$ ist also messbar als Grenzwert der monoton wachsenden Folge $(J_1 f_n)$ messbarer Funktionen. Ebenfalls mit monotoner Konvergenz folgt dann

$$\int_\Omega f(x)\mu(\mathrm{d}x) = \lim_{n\to\infty} \int_\Omega f_n(x)\mu(\mathrm{d}x) \overset{\text{Schritt 3}}{=} \lim_{n\to\infty} \int_{\Omega_2} J_1 f_n(x_2)\mu(\mathrm{d}x_2) = \int_{\Omega_2} J_1 f(x_2)\mu(\mathrm{d}x_2).$$

Schritt 4: Für $f \in \mathcal{L}^1$ zerlegt man $f = f^+ - f^-$ und benutzt die Linearität der behaupteten Aussagen. □

Bemerkung: Die Methode des obigen Beweises ist wichtig und wird an vielen Stellen in der Theorie benutzt: will man eine Aussage über ZVen zeigen, die „linear“ und „monoton“ ist, d.h. die für die Summe und den monotonen Grenzwert von ZVen wahr bleibt, wenn sie für die einzelnen ZVen wahr ist, so reicht es, diese Aussage für Indikatorfunktionen zu zeigen. In vielen Fällen erlaubt einem das π-λ-Theorem sogar, die Aussage auf einer noch kleineren Menge (etwa Indikatoren von Zylindermengen oder messbaren Rechtecken) zu zeigen.

(3.9) Definition

a) Das d**-dimensionale Lebesgue-Maß** λ^d ist das eindeutige Produktmaß von d eindimensionalen Lebesgue-Maßen, also das eindeutig bestimmte Maß auf auf $(\mathbb{R}^d, \mathcal{B}^{\otimes d})$ mit

$$\lambda^d(A_1 \times \cdots \times A_d) = \prod_{j=1}^{d} \lambda(A_j) \qquad \forall A_1, \dots A_d \in \mathcal{B}.$$

Statt $\int f(x)\lambda^d(\mathrm{d}x)$ schreibt man auch $\int f(x)\,\mathrm{d}x$.

b) Sei $\rho \in \mathcal{L}^1(\lambda^d)$, $\rho \geqslant 0$ und $\int \rho(x)\lambda^d(\mathrm{d}x) = 1$. Man sagt dann, das (eindeutige) W-Maß μ auf $(\mathbb{R}^d, \mathcal{B}^{\otimes d})$ mit $\mu(A) = \int_A \rho(x)\mathrm{d}x$ habe die **Dichte** ρ.

c) Für eine $\mathbb{R}^d$-wertige X habe $\mathbb{P}_X$ die Dichte ρ. Dann heißt ρ **Verteilungsdichte** von X.

Bemerkungen:

a) wie auch bei W-Maßen zeigen wie die Existenz des d-dimensionalen Lebesgue-Maßes in dieser Vorlesung nicht. Der Beweis findet sich in Büchern der Maßtheorie.

b) Wir haben nun (etwas mehr als) das in (3.1) formulierte Ziel erreicht, und insbesondere W-Maße mit Dichten auf $\mathbb{R}^d$ definiert. Diese W-Maße sind mit Abstand der wichtigste Typ von W-Maßen auf $\mathbb{R}^d$, und alle Beispiele aus Kapitel 2 waren von diesem Typ. Die Dichte verhält sich in vielen Aspekten genau wie die Zähldichte für diskrete W-Maße (daher der Name von letzterer). Das wollen wie als nächstes thematisieren.

(3.10) Definition

$X_1, \dots, X_n$ seien reelle ZVen auf $(\Omega, \mathcal{F}, \mathbb{P})$. Das Bildmaß von $\mathbb{P}$ unter der $\mathbb{R}^n$-wertigen ZV $\omega \mapsto (X_1(\omega), \dots, X_n(\omega))$ heißt **gemeinsame Verteilung** der ZVen $X_1, \dots, X_n$.

(3.11) Produktdichte und Faltungsformel

Die reellen ZVen $X_1, \dots, X_n$ seien unabhängig, und ρ_i sei die Verteilungsdichte von X_i.

a) Die gemeinsame Verteilung von $X_1, \dots, X_n$ hat auf $\mathbb{R}^n$ die **Produktdichte**

$$\boldsymbol{\rho} : \mathbb{R}^n \to \mathbb{R}_0^+, \qquad (x_1, \dots, x_n) \mapsto \rho_1(x_1)\rho_2(x_2)\cdots\rho_n(x_n).$$

b) Die Verteilungsdichte von $X_1 + X_2$ ist gegeben durch das **Faltungsprodukt** von ρ_1 und ρ_2, d.h. durch $\rho : \mathbb{R} \to \mathbb{R}_0^+$ mit

$$\rho(x) = \rho_1 * \rho_2(x) := \int_{\mathbb{R}} \rho_1(y)\rho_2(x-y)\,\mathrm{d}y = \int_{\mathbb{R}} \rho_2(y)\rho_1(x-y)\,\mathrm{d}y = \rho_2 * \rho_1(x).$$

Beweis:

a) Wir führen den Beweis nur für $d = 2$. Für größere d muss man den Satz von Fubini mehrmals anwenden, es erhöht sich vor allem der Notationsaufwand sehr stark. Seien $c_1, c_2 \in \mathbb{R}$. Dann

gilt

$$\mathbb{P}(X_1 \leqslant c_1, X_2 \leqslant c_2) \overset{\text{unabh.}}{=} \mathbb{P}(X_1 \leqslant c_1)\mathbb{P}(X_2 \leqslant c_2) = \int_{-\infty}^{c_1} \rho_1(x)\,\mathrm{d}x \int_{-\infty}^{c_2} \rho_2(y)\,\mathrm{d}y =$$
$$= \int_{-\infty}^{c_1} \rho_1(x)\Big(\int_{-\infty}^{c_2} \rho_2(y)\,\mathrm{d}y\Big)\,\mathrm{d}x = \int_{\mathbb{R}} \mathbb{1}_{(-\infty,c_1]}(x)\rho_1(x)\Big(\int_{\mathbb{R}} \mathbb{1}_{(-\infty,c_2]}(y)\rho_2(y)\,\mathrm{d}y\Big)\mathrm{d}x =$$
$$= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} \mathbb{1}_{(-\infty,c_1]\times(-\infty,c_2]}(x,y)\rho_1(x)\rho_2(y)\,\mathrm{d}y\Big)\,\mathrm{d}x \overset{\text{Fubini}}{=} \int_{\mathbb{R}^2} \mathbb{1}_{(-\infty,c_1]\times(-\infty,c_2]}(x,y)\rho_1(x)\rho_2(y)\,\mathrm{d}x\,\mathrm{d}y.$$

Da $\mathcal{B}^{\otimes 2}$ von den Mengen $(-\infty, c_1] \times (-\infty, c_2]$ erzeugt wird, zeigt dies

$$\mathbb{P}((X_1, X_2) \in A) = \int_A \rho_1(\mathrm{d}x)\rho_2(\mathrm{d}y)\,\mathrm{d}x\mathrm{d}y \qquad \forall A \in \mathcal{B}^{\otimes 2}.$$

b) Wir rechnen

$$\mathbb{P}(X_1 + X_2 \leqslant c) \overset{a)}{=} \int_{\{(x,y)\in\mathbb{R}^2 : x+y \leqslant c\}} \rho_1(x)\rho_2(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbb{R}^2} \mathbb{1}_{(-\infty,c]}(x+y)\rho_1(x)\rho_2(y)\,\mathrm{d}x\,\mathrm{d}y =$$
$$\overset{\text{Fubini}}{=} \int_{\mathbb{R}} \rho_1(x)\Big(\int_{\mathbb{R}} \mathbb{1}_{(-\infty,c]}(x+y)\rho_2(y)\,\mathrm{d}y\Big)\,\mathrm{d}x \overset{z=x+y}{=} \int_{\mathbb{R}} \rho_1(x)\Big(\int_{\mathbb{R}} \mathbb{1}_{(-\infty,c]}(z)\rho_2(z-x)\,\mathrm{d}z\Big)\,\mathrm{d}z =$$
$$\overset{\text{Fubini}}{=} \int_{\mathbb{R}} \mathbb{1}_{(-\infty,c]}(z)\Big(\rho_1(x)\rho_2(z-x)\,\mathrm{d}x\Big)\,\mathrm{d}z = \int_{-\infty}^{c} \rho_1 * \rho_2(z)\,\mathrm{d}z.$$

Wie oben folgt die Behauptung dann für allgemeine Mengen. Die Symmetrie sieht man durch Vertauschen der Rollen von ρ_1 und ρ_2 in ab dem dritten $=$. □

(3.12) Beispiele

a) X und Y seien normalverteilt und unabhängig, genauer:

$$X \sim \mathcal{N}(m_1, v_1), \qquad Y \sim \mathcal{N}(m_2, v_2), \qquad X \perp\!\!\!\perp Y.$$

Dann ist auch $X + Y$ normalverteilt, genauer:

$$X + Y \sim \mathcal{N}(m_1 + m_1, v_1 + v_2).$$

Diese sehr starke algebraische Eigenschaft ist ein erster Hinweis darauf, dass die Normalverteilung etwas ganz besonderes ist.

b) $(X_1, \ldots, X_n)$ seien Exp_α-verteilt und unabhängig. Dann ist $\sum_{i=1}^n X_i \sim \Gamma_{\alpha,n}$.
Beweise: Übung mit Hilfe der Faltungsformel.

(3.13) Integration bezüglich des Bildmaßes

$(\Omega, \mathcal{F}, \mu)$ sei ein Maßraum, $(\Omega', \mathcal{F}')$ ein messbarer Raum, $X : \Omega \to \Omega'$ sowie $f : \Omega' \to \mathbb{R}$ seien messbar. μ_X bezeichne das Bildmaß von μ unter X. Falls f nichtnegativ oder $f \circ X \in \mathcal{L}^1(\mu)$ ist, dann gilt

$$\int_\Omega f(X(\omega))\,\mu(\mathrm{d}\omega) = \int_{\Omega'} f(y)\mu_X(\mathrm{d}y).$$

Beweis:

Für Indikatorfunktionen f stimmt die Aussage nach Definition des Bildmaßes:

$$\Big[\mathbb{P}(X \in A) = \Big] \int \mathbb{1}_A(X(\omega))\,\mu(\mathrm{d}\omega) = \mu_X(A) = \int \mathbb{1}_A(y)\mu_X(\mathrm{d}y) \Big[= \mathbb{P}_X(A)\Big].$$

(Die Ausdrücke in Klammern nur, wenn μ ein W-Maß ist.) Da die behauptete Gleichung für $f + \alpha g$, $\alpha \in \mathbb{R}$, wahr ist, wenn sie für (integrierbare oder positive) f, g wahr ist und $\alpha \in \mathbb{R}$, und weil sie unter monotonen Limiten von (f_n) wahr ist, wenn sie für die einzelnen f_n gilt, greift die Methode aus dem Beweis von (3.8). □

Beispiel: Der Transformationssatz für Integrale in einer Dimension:
Sei $[a, b] \subset \mathbb{R}$ und $h : [a, b] \to \mathbb{R}$ stetig differenzierbar und streng monoton wachsend (also invertierbar). Betrachte das Maß μ auf $[a, b]$ mit Lebsgue-Dichte $x \mapsto h'(x) = \partial_x h(x)$. Nach (3.13) ist für integrierbare f

$$\int_a^b f(h(x))h'(x)\,\mathrm{d}x = \int \mathbb{1}_{[a,b]}\big(h^{-1}(h(x))\big) f(h(x))\,\mu(\mathrm{d}x) = \int f(y)\mathbb{1}_{[a,b]}(h^{-1}(y))[\mu \circ h^{-1}](\mathrm{d}y).$$

Was ist $\mu \circ h^{-1}$? Zunächst ist für Intervalle (c, d) mit $(c, d) \cap [h(a), h(b)] = \emptyset$

$$h^{-1}((c, d)) = \{x \in \mathbb{R} : h(x) \in (c, d)\} = \emptyset), \qquad \text{daher } \mu \circ h^{-1}((c, d)) = 0.$$

Also lebt $\mu \circ h^{-1}$ auf $[h(a), h(b)]$. Nach dem Hauptsatz der Differential- und Integralrechnung ist für $x \leqslant h(b)$

$$\mu \circ h^{-1}([h(a), x]) = \mu(\{y \in [a, b] : h(a) \leqslant h(y) \leqslant x\}) = \mu(\{y \in [a, b] : a \leqslant y \leqslant h^{-1}(x)\}) =$$
$$= \mu\big([a, h^{-1}(x)]\big) = \int_a^{h^{-1}(x)} h'(y)\,\mathrm{d}y = h(h^{-1}(x)) - h(a) = x - h(a).$$

$\mu \circ h^{-1}$ ist also das Lebesgue-Maß auf $[h(a), h(b)]$, und wegen $\mathbb{1}_{[a,b]}(h^{-1}(y)) = \mathbb{1}_{[h(a),h(b)]}(y)$ folgt der Transformationssatz

$$\int_a^b f(h(x))h'(x)\,\mathrm{d}x = \int_{h(a)}^{h(b)} f(y)\,\mathrm{d}y.$$

(3.14) Der Poisson-Prozess

a) Vorüberlegungen und Wiederholung

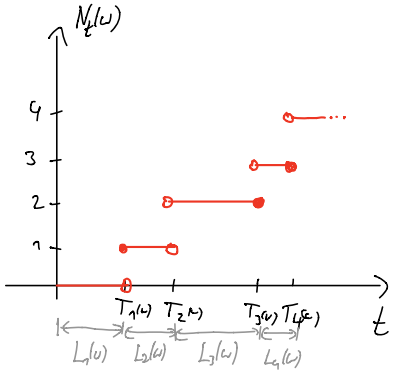
(i): In (2.26) haben wir gesehen, dass für $L \sim \mathrm{Exp}_\alpha$ und $X \sim \mathrm{Poi}_{\alpha t}$ gilt: $\mathbb{P}(L \geqslant t) = \mathbb{P}(X = 0)$.
(ii): In (2.27) haben wir gesehen, dass für $S \sim \Gamma_{\alpha,r}$ und $X \sim \mathrm{Poi}_{\alpha t}$ gilt: $\mathbb{P}(S \geqslant t) = \mathbb{P}(X \leqslant r-1)$.
(iii): Mit Beispiel (3.12 b) gilt also für unabhängige (L_n) mit $L_n \sim \mathrm{Exp}_\alpha$ und $X \sim \mathrm{Poi}_{\alpha t}$:

$$\mathbb{P}(\sum_{k=1}^{r} L_k \geqslant t) = \mathbb{P}(X \leqslant r - 1) \qquad \forall r \in \mathbb{N}.$$

(iv): Die obige Aussage zeigt die zwei Aspekte des gleichen zufälligen Geschehens: $S \sim \sum_{k=1}^r L_k$ beschreibt die *zufällige Wartezeit* bis zum r-ten Erfolg bei festem r, während X die zufällige Anzahl der Erfolge bis zu einer festen Zeit beschreibt. Der Poisson-Prozess vereinigt diese beiden Aspekte in ein Objekt.

b) Die Definition: $(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum, auf dem eine unabhängige Familie (L_n) von Exp_α-verteilten Zufallsvariablen definiert ist. Für $t \in \mathbb{R}$ definiere

$$N_t(\omega) := \sum_{r=1}^{\infty} \mathbb{1}_{(0,t]}(T_r(\omega)), \qquad \text{mit} \quad T_r(\omega) := \sum_{k=1}^{r} L_k(\omega).$$



Der Ausdruck $N_t(\omega)$ zählt also, wie viele Glieder der monoton wachsenden Folge $(T_r(\omega))_{r\in\mathbb{N}}$ einen Wert von höchstens t haben. Die Funktion $t \mapsto N_t(\omega)$ beschreibt für gegebenes ω den zu diesem ω gehörigen Ablauf z.B. der Zählung der Klicks eines Geigerzählers. Die Familie von ZVen $(N_t)_{t\in\mathbb{R}^+}$ heißt **Poisson-Prozess mit Intensität** α.

c) Unabhängige und Poisson-verteilte Zuwächse: Wegen (2.27) und (3.12 b) gilt $\mathbb{P}(N_t = k) = \mathrm{Poi}_{\alpha t}(k)$. Es gilt sogar viel mehr: für $0 = t_0 < t_1 < \ldots < t_n$ bilden die **Zuwächse** (oder **Inkremente**) $(N_{t_{i+1}} - N_{t_i})_{i \leqslant n}$ eine unabhängige Familie von ZVen, und $N_{t_{i+1}} - N_{t_i} \sim \mathrm{Poi}_{\alpha(t_{i+1}-t_i)}$. Die Interpretation dieser Tatsache ist, dass die im Zeitintervall $[t_i, t_{i+1}]$ beobachtete Anzahl von „Ereignissen“ (z.B. Geigerzähler-Klicks) von den Anzahlen in allen anderen Zeitintervallen nicht beeinflusst wird. Plausibel wird dies, indem man auf der Zeitachse diskrete Zeitpunkte $(s_k)_{k\in\mathbb{N}}$ mit Abstand $s_{k+1} - s_k = N$ einführt, und zu diesen Zeitpunkten jeweils unabhängig eine $\mathrm{Ber}_{\alpha/N}$-verteilte Münze wirft, die dann mit Wahrscheinlichkeit α/N zu einem „Ereignis“ bei diesem Zeitpunkt führt. Lässt man nun $N \to \infty$, so ist aus der Poisson-Konvergenz die Verteilung von $N_{t_{i+1}} - N_{t_i}$ plausibel, und die Unabhängigkeit der Inkremente gilt wegen der Eigenschaft (1.58 b) (Zusammenfassung in Blöcke) zumindest für endliche N. Da die rigorosen Grenzwertbetrachtungen aber lästig sind, geht man zum Beweis der Aussagen einen direkteren Weg.

Beweis der Aussage über die Zuwächse:
Wir zeigen nur den Fall von zwei Zeitpunkten, der allgemeine Fall ist ähnlich aber mit (noch) mehr Notationsaufwand verbunden. Seien $0 < s < t$ und $k, l \in \mathbb{Z}$. Wir werden zeigen, dass

$$(*) \qquad \mathbb{P}(N_s = k, N_t - N_s = l) = \mathrm{e}^{-\alpha s}\frac{(\alpha s)^k}{k!}\,\mathrm{e}^{-\alpha(t-s)}\frac{(\alpha(t-s))^l}{l!} = \mathrm{Poi}_{\alpha s}(\{k\})\mathrm{Poi}_{\alpha(t-s)}(\{l\}).$$

Mit $(*)$ folgt dann die Behauptung aus der Tatsache, dass die rechte Seite die korrekte Produkt-Zähldichte ist. $(*)$ selbst zeigt man mit expliziter Rechnung: Die Verteilung der $(L_i)_{1 \leqslant i \leqslant k+l+1}$ hat die Produkt-Zähldichte

$$\rho(x_1, \ldots, x_{k+l+1}) = \prod_{j=1}^{k+l+1} (\alpha\,\mathrm{e}^{-\alpha x_j}),$$

und somit ist

$$\mathbb{P}(N_s = k, N_t - N_s = l) = \mathbb{P}\Big(\sum_{j=1}^{k} L_j \leqslant s < \sum_{j=1}^{k+1} L_j, \sum_{j=1}^{k+l} L_j \leqslant t < \sum_{j=1}^{k+l+1} L_j \Big) =$$

$$= \int_0^\infty \mathrm{d}x_1 \cdots \int_0^\infty \mathrm{d}x_{k+l+1}\ \alpha^{k+l+1} \prod_{k=1}^{k+l+1} \mathrm{e}^{-\alpha x_j}\ \mathbb{1}_{(0,s]}(\sum_{j=1}^{k} x_j) \mathbb{1}_{(s,\infty)}(\sum_{j=1}^{k+1} x_j) \mathbb{1}_{(0,t]}(\sum_{j=1}^{k+l} x_j) \mathbb{1}_{(t,\infty)}(\sum_{j=1}^{k+l+1} x_j) =$$

$$= \alpha^{k+l} \int_{[0,\infty)^k} \mathrm{d}x_1 \cdots \mathrm{d}x_k\ \mathbb{1}_{(0,s]}(\sum_{j=1}^{k} x_j) \int_{[0,\infty)^l} \mathrm{d}x_{k+1} \cdots \mathrm{d}x_{k+l} \mathbb{1}_{(s,\infty)}(\sum_{j=1}^{k+1} x_j) \times$$

$$\times\ \mathbb{1}_{(0,t]}(\sum_{j=1}^{k+l} x_j) \int_0^\infty \mathrm{d}x_{k+l+1} \alpha\, \mathrm{e}^{-\alpha \sum_{j=1}^{k+l+1} x_j}\ \mathbb{1}_{(t,\infty)}(\sum_{j=1}^{k+l+1} x_j).$$

Im Integral der der letzten Zeile macht man die Substitution $y = x_{k+l+1} + \sum_{j=1}^{k+l} x_j$, wobei die $(x_j)_{j \leqslant k+l}$ als Konstanten behandelt werden. Dann ist $\mathrm{d}y = \mathrm{d}x_{k+l+1}$, und aus der Integralgrenze $x_{k+l+1} = 0$ wird die Integralgrenze $y = \sum_{j=1}^{k+1} x_j$. Die letzte Zeile ist dann gleich

$$\mathbb{1}_{(0,t]}(\sum_{j=1}^{k+l} x_j) \int_{\sum_{j=1}^{k+1} x_j}^\infty \alpha\, \mathrm{e}^{-\alpha y}\ \mathbb{1}_{(t,\infty)}(y) \mathrm{d}y = \mathbb{1}_{(0,t]}(\sum_{j=1}^{k+l} x_j) \int_t^\infty \alpha\, \mathrm{e}^{-\alpha y}\, \mathrm{d}y = \mathbb{1}_{(0,t]}(\sum_{j=1}^{k+l} x_j)\, \mathrm{e}^{-\alpha t}\,.$$

Die untere Integralgrenze durfte durch t ersetzt werden, da der Faktor $\mathbb{1}_{(0,t]}(\sum_{j=1}^{k+l} x_j)$ dafür sorgt, dass für $\sum_{j=1}^{k+l} x_j > t$ beide Seiten gleich Null sind. Es bleibt das Integral

$$\int_{[0,\infty)^{k+l}} \mathrm{d}x_1 \cdots \mathrm{d}x_{k+l}\ \mathbb{1}_{(0,s]}(\sum_{j=1}^{k} x_j) \mathbb{1}_{(s,\infty)}(\sum_{j=1}^{k+1} x_j) \mathbb{1}_{(0,t]}(\sum_{j=1}^{k+l} x_j)$$

zu berechnen. Man rechnet (ebenfalls mit Substitution etc.) nach, dass der Wert dieses Integrals genau $\frac{s^k}{k!} \frac{(t-s)^l}{l!}$ ist. Wegen $\alpha^{k+l}\, \mathrm{e}^{-\alpha t} = \alpha^k\, \mathrm{e}^{-\alpha s}\, \alpha^l\, \mathrm{e}^{-\alpha(t-s)}$ setzen sich die einzelnen Teile dann genau zur rechten Seite von $(*)$ zusammen.

d) Das Pfadmaß des Poisson-Prozesses: Wir haben den Poisson-Prozess oben einfach als Sammlung von (überabzählbar vielen) ZVen $(N_t)_{t \geqslant 0}$ definiert. Die Abbildung suggeriert aber eine viel natürlichere Beschreibung: für festes $\omega \in \Omega$ ist ja $t \mapsto N_t(\omega)$ eine (stückweise konstante) reelle Funktion. Sie heißt der zu $\omega \in \Omega$ gehörige **Pfad** des Poisson-Prozeses. (vergleiche dies mit der Situation bei den Markovketten!). Das Bildmaß von $\mathbb{P} = \mathrm{Exp}_\alpha^{\otimes \mathbb{N}}$ unter der Abbildung $\omega \mapsto (N_t(\omega))_{t \geqslant 0}$ sollte also ein W-Maß auf einem Wahrscheinlichkeitsraum sein, dessen Elemente Funktionen sind. Hier tauchen jedoch technische Schwierigkeiten auf: welchen Raum von Funktionen sollte man nehmen (der Raum *aller* reellen Funktionen scheint viel zu groß)? Welche σ-Algebra ist angemessen, damit die Abbildung von ω auf den dazugehörigen Pfad überhaupt messbar ist? Diese Probleme kann man lösen, und das resultierende W-Maß ist das Pfadmaß eines sogenannten **stochastischen Prozesses** (in stetiger Zeit). Die Theorie solcher Prozesse wird in der gleichnamigen Vertiefungsveranstaltung behandelt.

e) Der Poisson-Punktprozess: Eine weitere Sichtweise auf die in b) beschriebene Situation ist diejenige, dass man in der Abbildung die senkrechte Achse ignoriert und sich nur auf die Position der Punkte (T_k) konzentriert. Da man hieraus N_t rekonstruieren kann, ist das eine äquivalente Sichtweise. Man betrachtet also eine zufällige Sammlung von Punkten $(T_k)_{k \in \mathbb{N}}$ mit

den Eigenschaten:

(i): Die Anzahl der Punkte in einem Intervall $[a, b]$ ist $\mathrm{Poi}_{\alpha(b-a)}$-verteilt.

(ii): Die Anzahlen der Punkte in disjunkten Intervallen sind unabhängige ZVen.

Die eleganteste mathematische Formulierung dieser Ansammlung von Punkten ist als zufälliges (diskretes) Maß; man betrachtet also das Bildmaß von $\mathbb{P} = \mathrm{Exp}_\alpha^{\otimes\mathbb{N}}$ unter der Abbildung $\omega \mapsto \sum_{j=1}^{\infty} \delta_{T_j(\omega)}$, wobei $\delta_x(A) = \mathbb{1}_A(x)$ für $A \in \mathcal{B}$, also δ_x das Dirac-Maß bei x ist. Ähnliche technische Probleme wie oben treten auf und können gelöst werden. Das enstehende W-Maß heißt **Possion'scher Punktprozess** und ist der wichtigste (und einfachste) Vertreter der Klasse der Punktprozesse.

Erwartungswert, Varianz, Momente

(3.15) Definition

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum und X eine reelle ZV. Es gelte $X \geqslant 0$ oder $X \in \mathcal{L}^1(\mathbb{P})$. Die Größe

$$\mathbb{E}(X) = \int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega)$$

heißt **Erwartungswert** von X.

(3.16) Bemerkungen zu Notation und Nomenklatur

Der Erwartungswert ist offenbar mathematisch nichts anderes als das Integral der Funktion $X : \Omega \to \mathbb{R}$ bezüglich des Maßes $\mathbb{P}$. Warum (abgesehen von historischen Gründen) benutzen wir dann eine andere Bezeichnung und eine andere Notation?

a) Motivation der Nomenklatur: Wie schon beim Begriff der ZV, die ja auch einfach eine messbare Funktion ist, liegt der Unterschied bei der Interpretation der Objekte. Klassischerweise wird das Integral mit der Fläche unter einer Kurve assoziiert, oder mit dem Rauminhalt eines Körpers. Diese Interpretation bleibt zwar in der W-Theorie manchmal nützlich, aber wichtiger ist eine andere: der Erwartungswert beschreibt den Wert, den eine reelle ZV „im Durchschnitt" annehmen sollte, also den Wert, den man von ihr „erwarten" kann. Hierzu ein

Beispiel: ein *unfairer* Würfel mit den Zahlen 1 bis 6 wird durch $\Omega = \{1, \ldots, 6\}$ und die Zähldichte $(p_i)_{i=1,\ldots,6}$ mit $\sum_i p_i = 1$ modelliert. Bei einem Spiel bekommt man das Quadrat der gewürfelten Zahl in Goldstücken ausbezahlt. Mit welchem Gewinn sollte man rechnen, d.h. wie viel sollte man höchstens bezahlen, um an dem Spiel teilnehmen zu dürfen?

Bei einmaligem Würfeln ist man natürlich auf das Glück angewiesen, aber wenn man das Spiel sehr oft spielen darf, sollte sich doch ausrechnen lassen, wie viel man „im Durchschnitt" bekommt. Eine erste Idee: für n unabhängig ZVen $(X_i)_{i \leqslant n}$ (die nun natürlich auf dem Produktraum Ω^n mit Produktmaß definiert sein sollten!) setzt man schlicht

$$\bar{S}_n(\omega) := \frac{1}{n} \sum_{i=1}^{n} (X_i(\omega))^2. \qquad (*)$$

Für „die meisten" Spielverläufe (d.h. „die meisten" $\omega \in \Omega^n$) sollte man hieraus eine Zahl bekommen, die den durchschnittlichen Gewinn beschreibt. Natürlich sind auch „Glückssträhnen"

mit $X_i(\omega) = 6$ für alle ω möglich, aber meist eben nicht sehr wahrscheinlich. In der Praxis würde man natürlich genau so vorgehen, und n „Probespiele“ machen, die man dann auswertet.
Da wir aber mit den $(p_i)_{i \leqslant 6}$ ein Modell für unser Spiel gemacht haben, können wir eventuell weniger aufwändig und exakter vorgehen: wenn wir daran glauben, dass p_i die W-keit ist, mit der man i erhält, dann solle man in obiger Situation (bei großem n) in ziemlich genau $100 \cdot p_i$ Prozent aller Würfe die Zahl i bekommen haben. Analog zur Situation mit Riemann- und Lebesgue-Integral können wir daher versuchen, die Summe $(*)$ statt „vertikal“ lieber „horizontal“ zu betrachten:

$$\bar{S}_n(\omega) = \frac{1}{n}\sum_{k=1}^{6} k^2 |\{i \leqslant n : X_i(\omega) = k\}| = \sum_{k=1}^{6} k^2 \frac{|\{i \leqslant n : X_i(\omega) = k\}|}{n} \overset{(?!)}{\approx} \sum_{k=1}^{6} k^2\, p_k = \mathbb{E}(X^2),$$

wobei die ZV X die Verteilungs-Zähldichte (p_i) hat. Für die letzte Gleichheit siehe nochmal konkret (3.17) unten, sie folgt aber direkt aus (3.13). Das $\approx$ ersetzt die relative Häufigkeit des Auftretens von k durch ihren „theoretischen Wert“ p_k. Dies stimmt wie gesagt nicht immer (Glückssträhne!). Ein zentraler Satz der Theorie (das Gesetz der großen Zahl), den wir (in abgeschwächter Form) noch kennenlernen werden, sagt jedoch, dass es im Grenzwert $n \to \infty$ in allen Fällen stimmt, wo $\mathbb{E}(X)$ existiert.
Die Interpretation des Erwartungswertes ist also als gewichtete Summe (oder im Kontinuumslimes: Integral) der möglichen Ergebnisse eines zufälligen Geschehens, und zwar gewichtet mit der Wahrscheinlichkeit ihres Auftretens.

b) Zur Notation: Die Notation $\mathbb{E}(X)$ (mit $\mathbb{E}$ für „erwarteter Wert“) soll einerseits die obige Interpretation unterstreichen, andererseits ist sie von der Prämisse geleitet, dass es nicht nötig sein sollte, die genaue Form des W-Raums zu kennen, um Aussagen über die ZV X machen zu können - in der Integraldarstellung $\int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega)$ kommt er ja explizit vor. Dies hat jedoch auch einige Nachteile:

(i): zunächst ist die Verteilung von X entscheidend, aber in der Notation selbst nicht vorhanden: beispielsweise ist für $X : \mathbb{R} \to \mathbb{R}, x \mapsto x^2$ die Zahl $\mathbb{E}(X)$ davon abhängig, ob man auf $(\mathbb{R}, \mathcal{B})$ etwa die Normalverteilung $\mathcal{N}(0,1)$ oder die Exponentialverteilung Exp_α hat. Das ist natürlich beim Integral genauso, aber dort steht es in der Formel: $\int x^2 \mathrm{Exp}_\alpha(\mathrm{d}x) \neq \int x^2 \mathcal{N}_{0,1}(\mathrm{d}x)$ ist direkt aus dem Kalkül augenfällig. Dagegen muss man bei der $\mathbb{E}$-Notation immer vorher sagen „sei X eine $\mathcal{N}_{0,1}$-verteilte ZV“ etc. und sich später daran erinnern. Besonders schwierig wird es, wenn man auf dem selben Maßraum $(\Omega, \mathcal{F})$ mehrere W-Maße hat - wir haben diese Problematik bereits in (1.37) im Kontext der Markovketten angesprochen. Man behilft sich dann damit, Indizes an den Erwartungswert zu schreiben, etwas $\mathbb{E}_\mu(X) = \int X(\omega)\mu(\mathrm{d}\omega)$, oder $\mathbb{E}^x(X_n) = \int X_n(\omega)\mathbb{P}^x(\mathrm{d}\omega)$ im Falle der Markovketten.

(ii): die $\mathbb{E}$-Notation erlaubt es nicht, die (stumme) Integrationsvariable aufzuschreiben. Während man bei $\int_\Omega f(\omega)\mu(\mathrm{d}\omega) = \int f\,\mathrm{d}\mu$ die Wahl zwischen der Kurz- und Langform hat, gibt es bei der $\mathbb{E}$-Notation nur die Kurzform. In fast allen Fällen kann man durch sorgfältiges Nachdenken trotzdem einen formal korrekten Ausdruck aufschreiben, aber es ist oft sehr einfach, sich mit der Notation aufs Glatteis zu begeben. Hat man beispielsweise eine Familie $X_s : \mathbb{R} \times \Omega \to \mathbb{R}$ von ZVen, die noch von einem Parameter abhängen (wie etwa der Poisson-Prozess), dann ist es manchmal interessant, den Wert des Parameters selbst zufällig zu wählen, während er vielleicht kurz davor noch als fest angenommen war. Man hätte also im einem Fall die Abbildung

$\omega \mapsto X_s(\omega)$, und im anderen Fall die Abbildung $\omega \mapsto X_{s(\omega)}(\omega)$. In der Integralnotation sieht man direkt an der Formel, mit welcher Situation man es zu tun hat:

$$\int X_s(\omega)\mu(\mathrm{d}\omega) \neq \int X_{s(\omega)}(\omega)\mu(\mathrm{d}\omega).$$

In der $\mathbb{E}$-Notation steht beide Male einfach $\mathbb{E}(X_s)$ da, weil man die Integrationsvariable ja nicht aufschreibt. Man kann sich hier behelfen, indem man s anders benennt, wenn es fest ist (vielleicht ist das ingesamt ohnehin die sauberere Lösung), aber das Hauptproblem ist, dass die Gelegenheiten für Missverständnisse und Verwirrung wesentlich größer sind.

Trotz dieser Probleme hat sich die $\mathbb{E}$-Notation so fest etabliert, dass wir nicht darauf warten können, bis sie abgeschafft wird, und sie daher erlernen müssen. Tatsächlich hilft sie (insbesondere der Aspekt a) bei der Entwicklung einer „probabilistischen Intuition“.

(3.17) Berechnung von Erwartungswerten

a) Integration bezüglich des Bildmaßes: Sei X eine reelle ZV, $\mathbb{P}_X$ ihr Bildmaß, und $f : \mathbb{R} \to \mathbb{R}$ eine messbare Funktion so, dass $\omega \mapsto f(X(\omega))$ nichtnegativ oder integrierbar ist. Dann gilt:

$$\mathbb{E}(f(X)) = \int f(x)\mathbb{P}_X(\mathrm{d}x), \qquad \text{insbesondere} \quad \mathbb{E}(X) = \int x\mathbb{P}_X(\mathrm{d}x).$$

Hier benutzen wir die Notation $f(X) = f \circ X$, also $[f(X)](\omega) := f(X(\omega))$.

b) Wichtige Spezialfälle:

(i) Es gebe eine abzählbare Menge $M \subset \mathbb{R}$ mit $\mathbb{P}(X \in M) = 1$. Dann ist

$$\mathbb{E}(f(X)) = \sum_{z\in M} f(z)\mathbb{P}(X = z), \qquad \text{insbesondere} \quad \mathbb{E}(X) = \sum_{z\in M} z\mathbb{P}(X = z).$$

(ii) Für $M = \mathbb{Z}$ und Zähldichte $p_z = \mathbb{P}(X = z)$ für alle $z \in \mathbb{Z}$ wird das zu

$$\mathbb{E}(f(X)) = \sum_{z=-\infty}^{\infty} f(z)p_z, \qquad \text{insbesondere} \quad \mathbb{E}(X) = \sum_{z=-\infty}^{\infty} zp_z.$$

(iii): Hat die Verteilung von X eine Lebesgue-Dichte ρ, dann gilt

$$\mathbb{E}(f(X)) = \int f(x)\rho(x)\,\mathrm{d}x, \qquad \text{insbesondere} \quad \mathbb{E}(X) = \int x\rho(x)\,\mathrm{d}x.$$

c) Alternative Formeln für den Erwartungswert:

(i): Sei X eine integrierbare, nichtnegative reelle ZV. Dann ist

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t$$

(ii): Insbesondere ist, falls X eine ZV mit Werten in $\mathbb{N}_0$ ist,

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geqslant k).$$

Beweis: a) Dies ist (3.13) in anderer Notation. Die zweite Gleichung gilt jeweils mit der Wahl $f(x) = x$.

b) Folgt direkt aus a) durch konkretes Einsetzen des jeweiligen Bildmaßes.
c), (i): Es gilt mit dem Satz von Fubini für nichtnegative Funktionen:

$$\int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t = \int_0^\infty \mathrm{d}t \int_\Omega \mathbb{P}(\mathrm{d}\omega)\,\mathbb{1}_{(t,\infty)}(X(\omega)) = \int_\Omega \mathbb{P}(\mathrm{d}\omega) \int_0^\infty \mathrm{d}t\,\mathbb{1}_{(t,\infty)}(X(\omega)) =$$
$$= \int_\Omega \mathbb{P}(\mathrm{d}\omega) \int_0^\infty \mathrm{d}t\,\mathbb{1}_{(0,X(\omega))}(t) = \int_\Omega \mathbb{P}(\mathrm{d}\omega) \int_0^{X(\omega)} \mathrm{d}t = \int_\Omega \mathbb{P}(\mathrm{d}\omega)\,X(\omega) = \mathbb{E}(X).$$

c), (ii): Folgt aus (i), da in diesem Fall $\mathbb{P}(X > t) = \mathbb{P}(X \geqslant k)$ für $k - 1 \leqslant t < k$; das Integral $\int \mathbb{P}(X > t)\,\mathrm{d}t$ geht dann also über eine Stufenfunktion mit Stufen der Breite 1 und der Höhe $\mathbb{P}(X \geqslant k)_{k\in\mathbb{N}}$. □

(3.18) Momente und Varianz

X sei eine reelle ZV.
a) Falls $|X|^p \in \mathcal{L}^1$ für ein $p > 0$, so schreibt man $X \in \mathcal{L}^p$. Für $m \in \mathbb{N}$ und $X \in \mathcal{L}^m$ heißt $\mathbb{E}(X^m)$ das m**-te Moment von** X.
b) Sei $X \in \mathcal{L}^2$. Die Zahl

$$\mathbb{V}(X) := \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

heißt **Varianz** von X. Die Größe $\sqrt{\mathbb{V}(X)}$ heißt **Standardabweichung** von X.
c) Für alle $c \in \mathbb{R}$ gilt

$$\mathbb{V}(X + c) = \mathbb{V}(X), \qquad \text{und} \qquad \mathbb{V}(cX) = c^2\mathbb{V}(X).$$

Beweis: zu der b) behaupteten Gleichheit: da $\mathbb{E}(X)$ eine Zahl ist, darf man sie vor das Integral (den Erwartungswert) ziehen, damit rechnet man

$$\mathbb{E}((X-\mathbb{E}(X))^2) = \mathbb{E}\big(X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2\big) = \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

c) rechnet man einfach nach: $(X+c)-\mathbb{E}(X+c) = X-\mathbb{E}(X)$ und $(cX-\mathbb{E}(cX))^2 = c^2(X-\mathbb{E}(X))^2$. Nun wende den Erwartungswert darauf an. □

Bemerkungen:

(i): $X \in \mathcal{L}^p$ bedeutet, dass große Werte von X umso unwahrscheinlicher sein müssen, je größer p ist. Dies sieht man direkt, wenn X eine Verteilungsdichte ρ hat, denn dann muss ja die Funktion $\int \rho(y)|y|^p\,\mathrm{d}y$ noch integrierbar sein, dafür muss aber ρ für große y schnell genug klein werden. Diese Intuition liegt der einfachen, aber sehr nützlichen Chebyshev-Markov-Ungleichung zu Grunde, die wir gleich kennenlernen werden.

(ii): Wegen $|x|^p \leqslant 1 + |x|^q$ für $p \leqslant q$ gilt $X \in \mathcal{L}^p$ für alle $X \in \mathcal{L}^q$. Man schreibt dies kurz als $\mathcal{L}^q \subset \mathcal{L}^p$. **Achtung:** diese Beziehung gilt nicht, wenn man $\mathbb{P}$ durch ein Nicht-W-Maß μ mit $\mu(\Omega) = \infty$ (z.B. das Lebesgue-Maß) ersetzt (warum?).

(iii): Die Varianz misst, wie gut eine Verteilung um ihren Mittelwert konzentriert ist: dies wird im Fall einer Verteilungsdichte aus der Formel $\mathbb{V}(X) = \int \rho(x)(x - \mathbb{E}(X))^2\,\mathrm{d}x$ sichtbar: Die Funktion $x \mapsto (x-\mathbb{E}(X))^2$ ist nahe bei $\mathbb{E}(X)$ klein; wenn $\rho(x)$ sich stark in der Nähe von $\mathbb{E}(X)$ konzentriert, ist $\mathbb{V}(X)$ ebenfalls klein, falls $\rho(x)$ allerdings für x weit weg von $\mathbb{E}(X)$ langsam abfällt, dann kann $\mathbb{V}(X)$ auch sehr groß werden.

Beispiele:
a) Um eine Vorstellung davon zu bekommen, was Varianz intuitiv aussagt, stelle man sich die folgenden beiden Spiele: Beim einen Spiel wirft man eine Münze, und der Gewinner zahlt dem Verlierer 1 Euro; also $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Beim anderen wirft man die gleiche Münze noch mal, aber diesmal zahlt der Gewinner dem Verlierer 1.000.000 Euro; also $\mathbb{P}(Y = 10^6) = \mathbb{P}(Y = 10^{-6}) = 1/2$. Bei beiden Spielen ist der Erwartungswert des Gewinns gleich Null, aber es fühlt sich vermutlich ganz anders an, das zweite zu spielen als das erste. Der „Grund" ist, dass $\mathbb{V}(X) = \frac{1}{2}(1^2) + \frac{1}{2}(-1)^2 = 1$, aber $\mathbb{V}(Y) = \frac{1}{2}(10^6)^2 + \frac{1}{2}(-10^6)^2 = 10^{12}$.

b) Eine subtilere Variante des obigen Phänomens: Ein Spieler bietet Ihnen folgendes Spiel an: Sie zahlen 1 Euro, dann wird eine Zufallszahl zwischen 0 und 1 ermittelt. Ist diese Zahl kleiner als $\varepsilon > 0$, dann bekommen Sie $100/\varepsilon$ Euro ausbezahlt, andernfalls nichts. Also ist Ihr Gewinn durch X modelliert mit

$$\mathbb{P}(X = -1) = 1 - \varepsilon, \quad \mathbb{P}(X = \frac{100}{\varepsilon}) = \varepsilon, \qquad \text{daher } \mathbb{E}(X) = (1-\varepsilon)\cdot(-1) + \varepsilon\frac{100}{\varepsilon} = 99 + \varepsilon > 0$$

„Im Durchschnitt" gewinnen Sie also mehr als 99 Euro pro Spiel. Bei $\varepsilon = 1/2$ würden Sie sicher mitspielen - aber wie sieht es mit $\varepsilon = \frac{1}{10000}$ aus? Und wie bei $\varepsilon = 10^{-100}$? Beachten Sie: der mögliche Gewinn wird immer größer, je kleiner ε ist!

c) Es geht noch extremer: Der Spieler bietet Ihnen nun an, eine faire Münze so lange zu werfen, bis zum ersten mal „Kopf" erscheint; dann endet das Spiel. Für jede Runde, die das Spiel bis dahin gelaufen ist (also für jedes Auftreten von „Zahl"), erhalten Sie Geld, und zwar für das erste Auftreten 1 Euro, für das zweite 2 Euro, dann 4, dann 8, allgemein 2^{k-1} Euro für das k-te Mal „Zahl". Wenn „Kopf" bei der n-ten Runde erstmals auftritt, bekommen Sie also insgesamt $\sum_{k=1}^{n-1} 2^{k-1} = 2^{n-1} - 1$ Euro. Die Verteilung der Länge des Spiels kennen wir, es ist die Wartezeit auf das erste Mal „Kopf", also durch eine ZV Z mit $Z \sim \mathrm{Geo}_{1/2}$ modelliert. Der Gewinn des Spiels ist somit durch $Y = 2^{Z-1} - 1$ modelliert, und der Erwartungswert unseres Gewinns („erwarteter Gewinn") ist

$$\mathbb{E}(Y) = \mathbb{E}(2^{Z-1} - 1) = \sum_{k=0}^{\infty} 2^{k-1}\mathrm{Geo}_{1/2}(\{k\}) - 1 = \sum_{k=0}^{\infty} 2^{k-1}(\frac{1}{2})^{k+1} - 1 = \infty.$$

Wären Sie daher bereit, einen unendlich hohen Betrag zu zahlen, um hier spielen zu dürfen? Beachten Sie: die Wahrscheinlichkeit, dass Sie einen Einsatz von $2^{N-1} - 1$ Euro zumindest wieder herausbekommen, liegt bei $\mathbb{P}(Z \geqslant N) = 2^{-N}$. Wenn Sie aber einmal bis Runde N gekommen sind, bringt Ihnen schon der nächste glückliche Münzwurf die unfassbare Summe von 2^N an Reingewinn! Dieses Spiel ist schon 1738 von Daniel Bernoulli untersucht worden und trägt den Namen **Petersburg-Spiel**, nach der Zeitschrift Commentarii Academiae Scientiarum Imperialis Petropolitanae, in der der entsprechende Artikel veröffentlicht wurde.

Beispiele b) und c) zeigen, dass bei hoher Varianz (oder gar, wenn der Erwartungswert unendlich ist), diese Größen nicht sehr nützlich sind, um Entscheidungen zu treffen. Probleme dieser Art sind nicht rein akademisch, sondern treten beispielsweise beim Bau von Atomkraftwerken oder bei der Gesetzgebung zum Brandschutz eines Gebäudes auf - hier hat man es mit sehr hohen Schäden zu tun, die mit sehr geringer Wahrscheinlichkeit auftreten.
Dass wir hier mit dem Erwartungswert nicht wirklich weiter kommen, ist nicht schlimm, denn wir haben ja das ganze W-Maß, das wir analysieren können. Auch wenn wir unten sehen werden,

welche überraschenden Mengen and Information man allein mit Erwartungswert, Momenten und ähnlichen Integralgrößen oft bekommt, so können wir nicht erwarten, dass sie uns alles über ein zufälliges Geschehen sagen.

(3.19) Rechenregeln für Erwartungswert und Varianz

$(\Omega, \mathcal{F}, \mathbb{P})$ sei ein W-Raum, X, Y seien reelle ZVen auf Ω, $c \in \mathbb{R}$.

a) Falls $X, Y \in \mathcal{L}^1$ oder $X, Y \geqslant 0$, dann gilt

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y), \qquad \mathbb{E}(cX) = c\mathbb{E}(X) \qquad \text{(Linearität)}.$$

und

$$\mathbb{E}(X) \geqslant \mathbb{E}(Y) \qquad \text{falls } X \geqslant Y \ \mathbb{P}\text{-fast sicher.} \qquad \text{(Monotonie)}.$$

b) Falls $X \in \mathcal{L}^2$ und $c \in \mathbb{R}$, dann gilt

$$\mathbb{V}(X+c) = \mathbb{V}(X), \qquad \mathbb{V}(cX) = c^2\mathbb{V}(X) \qquad \text{(Varianz ist zentriert und quadratisch)}.$$

c) Falls $X, Y \in \mathcal{L}^1$ und X, Y unabhängig sind, dann ist auch $XY \in \mathcal{L}^1$ und es gilt

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \qquad \text{(Multiplikativität, falls } X, Y \text{ unabhängig)}.$$

d) Falls $X, Y \in \mathcal{L}^2$ und X, Y unabhängig sind, dann ist auch $X + Y \in \mathcal{L}^2$ und

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) \qquad \text{(Varianz ist additiv, falls } X, Y \text{ unabhängig)}.$$

Beweis: a) und b) haben wir bereits in (3.5) und (3.18) gesehen.
Zu c) sei $\mathbb{P}_{(X,Y)}$ die gemeinsame Verteilung von X und Y, also das Bildmaß von $\mathbb{P}$ unter $\omega \mapsto (X(\omega), Y(\omega))$. Satz (1.49) impliziert, dass X, Y genau dann unabhängig sind, wenn $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \otimes \mathbb{P}_Y$. Daher ist

$$\begin{aligned}\mathbb{E}(|X|\,|Y|) &= \int_{\mathbb{R}^2} |x||y| \mathbb{P}_{(X,Y)}(\mathrm{d}x\,\mathrm{d}y) = \int_{\mathbb{R}^2} |x||y| \mathbb{P}_X(\mathrm{d}x)\mathbb{P}_Y(\mathrm{d}y) \overset{\text{Fubini}}{=} \\ &= \int_{\mathbb{R}} |x| \mathbb{P}_X(\mathrm{d}x) \int_{\mathbb{R}} |y| \mathbb{P}_Y(\mathrm{d}y) = \mathbb{E}(|X|)\mathbb{E}(|Y|).\end{aligned}$$

Daher ist $XY \in \mathcal{L}^1$, und die gleiche Rechnung ohne Betragsstiche liefert die behauptete Gleichheit.
d) Übung! □

Bemerkung: Es ist sehr wichtig festzuhalten, dass die Multiplikativität aus Punkt c) oben eine *viel weniger starke* Eigenschaft ist als die Unabhängigkeit: es ist sehr leicht, Beispiele von ZVen zu finden wo $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ gilt, wo X und Y aber nicht unabhängig sind. Das folgende Beispiel ist nicht das „kleinste“ (man kann ein Beispiel mit $\Omega = \{1, 2, 3\}$ finden), aber es zeigt anschaulich wo das Problem liegt.

Zwei Spieler spielen gegen die Bank. Das Spiel ist ein faires Münzwurfspiel, und man darf vor Beginn des Spiels entscheiden, ob man bei Kopf 1 Euro gewinnt und bei Zahl 1 Euro verliert, oder umgekehrt. Die Münze, die hochgeworfen wird, bestimmt den Ausgang immer für beide Spieler zugleich. Spieler 1 entscheidet sich für „Kopf“ als Gewinnseite, und sein Vermögen nach k Runden wird durch den Wert einer einfachen Irrfahrt $(X_n)_{n\in\mathbb{N}}$ für $n = k$ modelliert. Spieler

2 wirft erst eine eigene Münze (natürlich unabhängig von den späteren Ereignissen im Spiel), die durch die ZV Z mit $\mathbb{P}(Z=1)=\mathbb{P}(Z=-1)=1/2$ modelliert wird, und wählt „Kopf“ als Gewinnseite falls $Z=1$, und sonst „Zahl“. Das Vermögen von Spieler 2 nach k Runden ist dann durch $Y_k := ZX_k$ modelliert, ist also (abhängig von Z) immer gleich groß oder immer genau das negative von demjenigen von Spieler 1. Diese Symmetrie ist es auch, die zur Multiplikativität des Erwartungswertes führt: wegen $\mathbb{1}_{\{-1\}}(Z)+\mathbb{1}_{\{1\}}(Z)=1$ ist

$$\begin{aligned}\mathbb{E}(X_kY_k) &= \mathbb{E}\Big(X_kY_k\big(\mathbb{1}_{\{-1\}}(Z)+\mathbb{1}_{\{1\}}(Z)\big)\Big) = \mathbb{E}(X_kY_k\mathbb{1}_{\{-1\}}(Z)) + \mathbb{E}(X_kY_k\mathbb{1}_{\{1\}}(Z)) = \\ &= \tfrac{1}{2}\mathbb{E}(X_k(-X_k)) + \tfrac{1}{2}\mathbb{E}(X_k^2) = 0 = \mathbb{E}(X_k)\mathbb{E}(Y_k).\end{aligned}$$

Intuitiv ist andererseits ganz klar, dass X_k und Y_k nicht unabhängig sein können - sie sind ja dem Betrag nach gleich! Formal sieht man das z.B. über

$$\mathbb{P}(X_2=2,Y_2=0)=0\neq\mathbb{P}(X_2=2)\mathbb{P}(Y_2=0)=\frac{1}{4}\cdot\frac{1}{2}=\frac{1}{8}.$$

(X_1 und Y_1 sind tatsächlich unabhängig, weil sie „zu einfach“ sind - ein algebraischer „Zufall“).

Bevor wir mit Betrachtungen um die Varianz weiter machen, brauchen wir die Cauchy-Schwarz-Ungleichung. Wir nehmen dies zum Anlass, gleich die wichtigsten Ungleichungen rund um den Erwartungswert alle zu behandeln.

(3.20) Die Chebyshev-Markov-Ungleichung

X sei eine reelle ZV, und $f:\mathbb{R}_0^+\to\mathbb{R}_0^+$ monoton wachsend und messbar. Dann gilt für alle $c>0$ mit $f(c)>0$ die Ungleichung

$$\mathbb{P}(|X|\geqslant c)\leqslant\frac{\mathbb{E}\big(f\big(|X|\big)\big)}{f(c)}.$$

Insbesondere folgt

$$\mathbb{P}\Big(\big|X-\mathbb{E}(X)\big|\geqslant c\Big)\leqslant\frac{\mathbb{V}(X)}{c^2}$$

Beweis: Auf der Menge $\{\omega\in\Omega: f(|X(\omega)|)\geqslant f(c)\}$ gilt (trivialerweise) $\frac{f(|X(\omega)|)}{f(c)}\geqslant 1$. Weil f monoton ist folgt aus $|X(\omega)|\geqslant c$ die Ungleichung $f(|X(\omega)|)\geqslant f(c)$. Diese beiden fast schon tautologischen Beobachtungen erlauben die Rechnung

$$\mathbb{P}(|X|\geqslant c)\leqslant\mathbb{P}(f(|X|)\geqslant f(c))=\mathbb{E}(\mathbb{1}_{f(|X|)\geqslant f(c)})\leqslant\mathbb{E}\Big(\mathbb{1}_{f(|X|)\geqslant f(c)}\frac{f(|X|)}{f(c)}\Big)\leqslant\frac{1}{f(c)}\mathbb{E}(f(|X|)).$$

Die zweite Gleichung folgt durch Anwendung der ersten auf $Y=X-\mathbb{E}(X)$ mit der Funktion $f(x)=x^2$. □

(3.21) Anwendungen der Chebyshev-Markov-Ungleichung

a) Die Chernoff-Cramer-Schranke: Sie folgt direkt aus der Chebyshev-Markov-Ungleichung: Sei X eine reelle ZV, es gelte $\mathbb{E}(\,\mathrm{e}^{\beta X}\,) < \infty$ für alle $\beta \in (-M, M)$ für ein $M > 0$. Man wendet die Ungleichung auf die nichtnegative ZV $Y = \mathrm{e}^{\beta X}$ mit der Funktion $f(x) = x$ an, und bekommt

$$\mathbb{P}(X \geqslant c) = \mathbb{P}(Y \geqslant \mathrm{e}^{\beta c}) \leqslant \frac{\mathbb{E}(Y)}{\mathrm{e}^{\beta c}} = \mathbb{E}(\,\mathrm{e}^{\beta X}\,)\,\mathrm{e}^{-\beta c} = \exp\big(\log(\mathbb{E}(\,\mathrm{e}^{\beta X}\,)) - \beta c\big).$$

Falls man $\mathbb{E}(\,\mathrm{e}^{\beta X}\,)$ explizit ausrechnen kann, so kann man in dieser Ungleichung das optimale β suchen, welches die rechte Seite möglichst klein macht - die linke hängt ja nicht von β ab. Dies ermöglicht oft überraschend starke Resultate. Insbesondere wenn $S = \sum_{i=1}^n Z_i$ und die Z_i unabhängige ZVen sind, die man gut kennt, dann ist

$$\log \mathbb{E}(\,\mathrm{e}^{\beta S}\,) = \log\Big(\mathbb{E}\big(\prod_{i=1}^n \mathrm{e}^{\beta Z_i}\,\big)\Big) = \log\Big(\prod_{i=1}^n \mathbb{E}\big(\,\mathrm{e}^{\beta Z_i}\,\big)\Big) = \sum_{i=1}^n \log \mathbb{E}(\,\mathrm{e}^{\beta Z_i}\,),$$

damit kann man oft etwas anfangen. Beispielsweise:

b) Untypische Werte der einfachen Irrfahrt:

In (1.66 b) hatten wir für die einfache Irrfahrt $(S_n)_{n\in\mathbb{N}}$ die Schranke

$$\mathbb{P}(S_n \geqslant c) \leqslant \mathrm{e}^{-\frac{c^2}{2n}}$$

behauptet. Diese können wir nun beweisen. Wir wählen die Darstellung $S_n = \sum_{i=1}^n X_i$ mit unabhängigen X_i und mit $\mathbb{P}(X_i = \pm 1) = 1/2$, und wenden die Chernoff-Cramer-Schranke an. Aus b) wissen wir, dass

$$\mathbb{P}(S_n \geqslant c) \leqslant \exp\Big(n \log \mathbb{E}\big(\,\mathrm{e}^{\beta X_1}\,\big) - \beta c\Big)$$

Es gilt

$$\mathbb{E}(\,\mathrm{e}^{\beta X_1}\,) = \frac{1}{2}\,\mathrm{e}^{\beta} + \frac{1}{2}\,\mathrm{e}^{-\beta} = \cosh(\beta).$$

Wenn wir die Chernoff-Schranke optimieren wollen, müssen wir also den Ausdruck

$$\beta \mapsto n \log(\cosh(\beta)) - \beta c$$

minimieren. Mit anderen Worten, wir müssen die transzendente Gleichung

$$0 = \partial_\beta\big(n \log\cosh(\beta) - \beta c\big) = n \sinh(\beta)/\cosh(\beta) - c = n \tanh(\beta) - c$$

lösen. Das kann man (in guter Näherung) tun, uns genügt aber die gröbere Abschätzung

$$\mathbb{E}(\,\mathrm{e}^{\beta X_j}\,) = \frac{1}{2}\,\mathrm{e}^{\beta} + \frac{1}{2}\,\mathrm{e}^{-\beta} = \sum_{k=0}^{\infty} \frac{\beta^{2k}}{(2k)!} \leqslant \sum_{k=0}^{\infty} \frac{(\beta^2/2)^k}{k!} = \mathrm{e}^{\beta^2/2}$$

ergibt. Hiermit ist dann

$$\log \mathbb{E}(\,\mathrm{e}^{\beta X_1}\,) \leqslant \log\,\mathrm{e}^{\beta^2/2} = \beta^2/2,$$

und

$$\mathbb{P}(S_n \geqslant c) \leqslant \mathrm{e}^{n\beta^2/2 - \beta c}$$

Hier ist das optimale β leicht zu finden und liegt bei $\beta = c/n$. Einsetzen dieses Wertes ergibt die behauptete Ungleichung.

c) Poissonverteilung für große Parameter: In (2.11) hatten wir mit einigem Aufwand nachgerechnet, dass für eine Poi_λ-verteilte ZV X gilt:

$$\lim_{\lambda\to\infty} \mathbb{P}\Big(\big|\frac{X}{\lambda} - 1\big| \geqslant \frac{1}{\lambda^q}\Big) = 0$$

falls $q < \frac{1}{2}$. Die Chebyshev-Ungleichung liefert dies (und mehr) direkt: zunächst berechnen wir Erwartungswert und Varianz der Poisson-Verteilung:

$$\mathbb{E}(X) = \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\,\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \mathrm{e}^{-\lambda}\,\lambda\,\mathrm{e}^{\lambda} = \lambda,$$

und

$$\mathbb{E}(X(X-1)) = \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\,\lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.$$

Es folgt

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X(X-1)) + \mathbb{E}(X) - \mathbb{E}(X)^2 = \lambda.$$

Nun liefert die spezielle Form der Chebyshev-Ungleichung

$$\mathbb{P}\Big(\big|\frac{X}{\lambda} - 1\big| \geqslant \frac{1}{\lambda^q}\Big) = \mathbb{P}(|X - \mathbb{E}(X)| \geqslant \lambda^{1-q}) \leqslant \frac{\mathbb{V}(X)}{\lambda^{2-2q}} = \lambda^{-1+2q}.$$

Dies verschwindet im Grenzwert $\lambda \to \infty$, falls $q < 1/2$.

(3.22) Die Jensen-Ungleichung

$\varphi : \mathbb{R} \to \mathbb{R}$ sei eine konvexe Funktion, X eine reelle ZV mit $X \in \mathcal{L}^1$ und $\varphi(X) \in \mathcal{L}^1$. Dann gilt

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

Beweis: Aus der Analysis kennen Sie vermutlich die Äquivalenz (manchmal auch als Definition von Konvexität von Funktionen benutzt):

$$\varphi \text{ ist konvex} \quad\Longleftrightarrow\quad A := \{(x,y) \in \mathbb{R}^2 : y > \varphi(x)\} \text{ ist eine konvexe Menge.}$$

Dies bedeutet, dass man für jedes $x \in \mathbb{R}$ eine Gerade durch den Punkt $(x, \varphi(x))$ finden kann, die A nicht schneidet. Formal:

$$\forall x \in \mathbb{R} : \exists a \in \mathbb{R} : \forall y \in \mathbb{R} : \ \ell(y) := \varphi(x) + a(y-x) \leqslant \varphi(y).$$

Für die Wahl $x = \mathbb{E}(X)$ erhält man also ein passendes a, wählt dann $y = X(\omega)$ und bekommt die Ungleichung

$$\varphi(X(\omega)) \geqslant \ell((X(\omega)) = \varphi(\mathbb{E}(X)) + a(X(\omega) - \mathbb{E}(X)),$$

gültig für alle $\omega \in \Omega$. Integriert man diese Ungleichung, so ergibt sich

$$\mathbb{E}(\varphi(X)) \geqslant \mathbb{E}(\varphi(\mathbb{E}(X))) + a\mathbb{E}(X - \mathbb{E}(X)) = \varphi(\mathbb{E}(X)) + 0,$$

wie behauptet. □

(3.23) Hölder-Ungleichung und Cauchy-Schwarz-Ungleichung

X, Y seien reelle ZVen.

a) Für alle $p, q \geqslant 1$ mit $\frac{1}{p} + \frac{1}{q} = 1$ gilt die **Hölder-Ungleichung**

$$\mathbb{E}(|XY|) \leqslant \Big(\mathbb{E}(|X|^p)\Big)^{1/p} \Big(\mathbb{E}(|Y|^q)\Big)^{1/q}.$$

b) Für $p = q = 2$ heißt die Hölder-Ungleichung auch **Chauchy-Schwarz-Ungleichung**, sie wird oft in der folgenden Form formuliert:

$$\mathbb{E}(|XY|)^2 \leqslant \mathbb{E}(|X|^2)\mathbb{E}(|Y|^2).$$

c) Man schreibt oft $\|X\|_p := (\mathbb{E}(|X|^p))^{1/p}$, und $\|X\|_\infty := \inf\{M > 0 : |X| \leqslant M \ \mathbb{P}\text{-fast sicher}\}$. Mit diesen Bezeichnungen (und der Konvention $\frac{1}{\infty} = 0$) gilt für alle $1 \leqslant p, q \leqslant \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$

$$\mathbb{E}(|XY|) \leqslant \|X\|_p \|Y\|_q.$$

Beweis: a) Falls $\mathbb{E}(|X|^p) = 0$, dann ist, wegen (3.5 d), $|X| = 0$ $\mathbb{P}$-fast sicher, und somit, ebenfalls nach (3.5 d), $\mathbb{E}(|XY|) = 0$, und die Ungleichung stimmt. Sei daher nun $\mathbb{E}(|X|^p) > 0$. Wir definieren das W-Maß $\tilde{\mathbb{P}}$ durch

$$\tilde{\mathbb{P}} : \mathcal{F} \to \mathbb{R}, \qquad A \mapsto \tilde{\mathbb{P}}(A) := \frac{1}{\mathbb{E}(|X|^p)} \mathbb{E}(\mathbb{1}_A |X|^p).$$

Wir schreiben $\tilde{\mathbb{E}}$ für den Erwartungswert bezüglich $\tilde{\mathbb{P}}$, und setzen $Z := |Y||X|^{1-p}\mathbb{1}_{|X|>0}$. Dann ist $|X(\omega)Y(\omega)| = |X(\omega)^p Z(\omega)|$, und

$$\begin{aligned}
\mathbb{E}(|XY|) = \mathbb{E}(|X|^p Z) = \mathbb{E}(|X|^p)\tilde{\mathbb{E}}(Z) = \mathbb{E}(|X|^p)\big(\tilde{\mathbb{E}}(Z)^q\big)^{1/q} &\overset{\text{Jensen}}{\leqslant} \\
\leqslant \mathbb{E}(|X|^p)\tilde{\mathbb{E}}(Z^q)^{1/q} = \mathbb{E}(|X|^p)\mathbb{E}\Big(\frac{|Y|^q}{|X|^{q(p-1)}} \frac{|X|^p}{\mathbb{E}(|X|^p)}\Big)^{1/q} &= \\
= \mathbb{E}\big(|Y|^q |X|^{p-q(p-1)}\big)^{1/q} \ \mathbb{E}\big(|X|^p\big)^{1-1/q}.&
\end{aligned}$$

Wegen $\frac{1}{p} + \frac{1}{q} = 1$ ist $q + p = pq$, und somit $p - q(p-1) = p - pq + q = 0$, und außerdem $1 - 1/q = 1/p$. Die Behauptung folgt daraus.

c) Es fehlt noch der Fall $p = 1, q = \infty$. Dieser folgt aus der Monotonie der Integrals:

$$\mathbb{E}(|XY|) \leqslant \mathbb{E}(|X|\|Y\|_\infty) = \|X\|_1 \|Y\|_\infty. \qquad \square$$

(3.24) Kovarianz und Korrelationskoeffizient

Seien X, Y Zufallsvariablen, $X, Y \in \mathcal{L}^2$.

a) Die Größe

$$\mathrm{Cov}(X, Y) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

heißt **Kovarianz** von X und Y.

b) Die Größe

$$\varrho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$$

heißt **Korrelationskoeffizient** von X und Y. Es gilt $-1 \leqslant \varrho(X,Y) \leqslant 1$.

c) Falls $\varrho(X,Y) > 0$, so heißen die ZVen X und Y **positiv korreliert**.
Falls $\varrho(X,Y) < 0$, so heißen die ZVen X und Y **negativ korreliert**.
Falls $\varrho(X,Y) = 0$, so heißen die ZVen X und Y **unkorreliert**.

Bemerkungen:

a) Die behauptete Gleichheit für die Kovarianz zeigt man genau wie die in (3.18). Die Tatsache, dass $XY \in \mathcal{L}^1$ ist (d.h. dass die Formel für $\mathrm{Cov}(X,Y)$ überhaupt erlaubt ist) folgt aus der Cauchy-Schwarz-Ungleichung.

b) Die behauptete Ungleichung folgt direkt aus der Cauchy-Schwarz-Ungleichung. Die Intuition zum Begriff der Korrelation: sind X und Y positiv korreliert, so bedeutet das, dass Werte von X, die (im Vergleich zu $\mathbb{E}(X)$) groß sind, mit höherer W'keit angenommen werden, wenn man weiß, dass Y große Werte annimmt, ebenso bei kleinen Werten. Bei negativer Korellation ist es umgekehrt: große Werte von X machen kleine Werte von Y wahrscheinlicher, und kleine Werte von X machen große Werte von Y wahrscheinlicher. Diese Aussagen sind etwas zu schwammig, um sie ohne große Schwierigkeiten zu formalisieren. Den Mechanismus, der dahinter steckt, sieht man jedoch gut, wenn man annimmt, dass die gemeinsame Verteilung von $X - \mathbb{E}(X)$ und $Y - \mathbb{E}(Y)$ eine Lebesgue-Dichte $(x,y) \mapsto \rho(x,y)$ hat: dann ist

$$\mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big) = \int_{\mathbb{R}^2} \mathbb{P}_{(X,Y)}(\mathrm{d}x\,\mathrm{d}y) xy\rho(x,y),$$

und da die Funktion $(x,y) \mapsto xy$ im ersten und dritten Quadranten positiv und in den anderen beiden negativ ist, sieht man, dass der obige Ausdruck eher dann positiv ist, wenn ρ entweder mehr Masse auf den ersten und dritten Quadranten legt als auf die beiden anderen, oder diese Masse weiter weg von der 0 platziert. In diesen beiden Quadranten haben aber x und y, und damit die Werte von X und Y, die zu $\rho(x,y)$ gehören, das gleiche Vorzeichen.

c) **Beispiele** für Korrelationen im Sinne „große Werte von X begünstigen große Werte von Y" findet man im Alltag in großen Mengen: in der Medizin (z.B. Tragen von Gesichtsmasken und Infektionszahlen), in der Soziologie (sozialer Status und Lebenserwartung), etc... in diesen Fällen kommt die Korrelation aus den Daten, und nicht wie in Definition (3.23) aus dem Modell, und wird zunächst empirisch gemessen. Dann sollte man versuchen, dann ein möglichst gutes Modell zu machen, das diese Korrelationen erklärt. Beide dieser Aufgaben sind Inhalt der *Statistik*, die wir am Ende dieses Kurses noch kennenlernen werden. Wohlbekannt aber wichtig ist die Tatsache, dass aus Korrelation allein keine Kausalität geschlossen werden kann, also weder dass große Werte von X ein Grund für große Werte von Y sind, noch umgekehrt. Man kann nur schließen, dass solche Werte oft gemeinsam auftreten, dies könnte einen gemeinsamen Grund haben, den man nicht kennt.

d) Die Kovarianz kann man als **Skalarprodukt** auf einem Raum von Funktionen auffassen: zunächst ist $\mathcal{L}^2$ ein Vektorraum wegen der Rechenregeln für das Integral. Wir definieren die Abbildungen $\|\cdot\| : \mathcal{L}^2 \to \mathbb{R}_0^+$ und $\langle\cdot,\cdot\rangle : \mathcal{L}^2 \times \mathcal{L}^2 \to \mathbb{R}$ durch

$$\|X\| := \sqrt{\mathbb{V}(X)}, \qquad \langle X, Y\rangle := \mathrm{Cov}(X,Y).$$

Die Abbildung $\langle\cdot,\cdot\rangle$ ist bilinear (linear im ersten und zweiten Argument), und $\langle X, X\rangle = \|X\|^2$. Die Abbildung $\|.\|$ erfüllt die Dreiecksungleichung: $\|X+Y\| \leqslant \|X\|+\|Y\|$ (Beweis: Übung!), und ist homogen, d.h. $\|cX\| = c\|X\|$ für $c \geqslant 0$. Sie ist keine Norm auf $\mathcal{L}^2$, denn $\|X\| = 0$ ist nicht nur

dann der Fall, wenn $X = 0$ (also $X(\omega) = 0$ für alle $\omega \in \Omega$) gilt, sondern auch, wenn dies nur fast sicher gilt, und wegen der Subtraktion des Erwartungswertes sogar dann, wenn X fast sicher konstant ist. Will man eine Norm haben, muss man (formal mittels Äquivalenzrelation) alle ZVen X, Y identifiziert, die sich fast überall nur um eine Konstante unterscheiden, für die es also ein $c \in \mathbb{R}$ gibt mit $X = Y + c$ $\mathbb{P}$-fast überall. Auf dem Raum dieser Äquivalenzklassen ist dann $\|\cdot\|$ eine Norm, und $\langle \cdot, \cdot \rangle$ ist das Skalarprodukt, das diese Norm erzeugt. Eine **geometrische Interpretation** der Unkorreliertheit von X und Y ist daher, dass X auf Y senkrecht steht bezüglich dieses Skalarproduktes.

e) Aus der Maßtheorie kennen Sie vielleicht die L^p-Räume, mit der Norm $\|f\|_p = (\int |f|^p \, \mathrm{d}\mu)^{1/p}$, und auf Äquivalenzklassen von Funktionen bezüglich der Äquivalenzrelation wo $f \sim g$ falls $f = g$ μ-fast überall. Diese Räume spielen auch in der W-Theorie eine große Rolle, der feinen Unterschied zu den in d) eingeführten ist jedoch, dass dort die Äquivalenzklassen noch größer sind: es wurde ja $\mathbb{E}(X)$ abgezogen (d.h. auf Mittelwert 0 normiert), das tut man bei den L^p-Räumen nicht. In diesem Zusammenhang ist die folgende Aussage interessant: Es gilt

$$\mathbb{V}(X) = \min\big\{\mathbb{E}\big((X-a)^2\big) : a \in \mathbb{R}\big\}.$$

(Beweis: Übung).

f) Aus der Diskussion in b) ist man versucht zu sagen: je größer der Betrag von $\mathrm{Cov}(X, Y)$, desto eher glaubt man, dass große Werte von X und Y gemeinsam auftreten. Das stimmt aber nicht ganz: nimmt man etwa zwei ZVen X, Y mit $\mathrm{Cov}(X, Y) = 0.01$ (also scheinbar sehr schwach positiv korreliert), dann gilt $\mathrm{Cov}(1000X, 1000Y) = 10^6 \cdot 0.01 = 10^4$ (also scheinbar sehr stark positiv korelliert). Dabei sollte das Multiplizieren mit einer positiven Zahl nichts daran ändern, welche Werte man gleichzeitig erwartet. Dies wird im Korrelationskoeffizienten $\varrho(X, Y)$ korrigiert, dieser ist so normiert, dass man aus seiner Größe tatsächlich sehen kann, ob eine starke oder schwache Tendenz zum gleichzeitigen Annehmen großer bzw. kleiner Werte besteht.

(3.25) Rechenregeln für Varianz und Kovarianz

X, Y und X_n, $n \in \mathbb{N}$, seien in $\mathcal{L}^2$, $a, b \in \mathbb{R}$. Dann gilt:

a) $\mathrm{Cov}(X + a, Y + b) = \mathrm{Cov}(X, Y)$.

b) $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$.

c) $\sum_{i=1}^n X_i \in \mathcal{L}^2$, und

$$\mathbb{V}\Big(\sum_{i=1}^n X_i\Big) = \sum_{i=1}^n \mathbb{V}(X_i) + \sum_{i,j\in\{1,\dots,n\}, i\neq j} \mathrm{Cov}(X_i, X_j)$$

d) Falls die (X_i) paarweise unkorreliert sind, ergibt sich aus c) die **Gleichheit von Bienaimé**:

$$\mathbb{V}\Big(\sum_{i=1}^n X_i\Big) = \sum_{i=1}^n \mathbb{V}(X_i)$$

Beweis: Übung

Wir haben schon betont, dass Unabhängigkeit eine viel stärkere Eigenschaft ist als Unkorreliertheit. Der folgende Satz zeigt, um wieviel stärker sie ist - beachte dass $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ in Punkt b) unten dem Fall $f(x) = g(x) = x$ entspricht.

(3.26) Satz

X, Y seien reelle ZVen. Dann sind die folgenden Aussagen äquivalent:
(i): $X \perp\!\!\!\perp Y$.
(ii): Für alle messbaren Abbildungen $f, g : \mathbb{R} \to \mathbb{R}$, für die $f(X)$ und $g(Y)$ in $\mathcal{L}^1$ sind, gilt

$$\mathbb{E}\big(f(X)g(Y)\big) = \mathbb{E}\big(f(X)\big)\mathbb{E}\big(g(Y)\big)$$

Beweis: Übung.

Es folgen zwei etwas komplexere Beispiele, die den Erwartungswert benutzen:

(3.27) Beispiel: Das Ising-Modell

Im Experiment beobachtet man das Phänomen der **Hysterese**: Bringt man ein magnetisierbares Material wie Eisen in ein äußeres Magnetfeld der Stärke μ, und lässt man μ langsam gegen 0 gehen, dann behält das Material auch im Punkt $\mu = 0$ eine Magnetisierung bei, es „erinnert“ sich also an das vorher angelegte äußere Magnetfeld. Dies ist aber nur dann der Fall, wenn die Temperatur $1/\beta$ (gemessen in Kelvin) nicht zu groß ist. Das Ising-Modell wurde geschaffen, um dieses Verhalten an einem einfachen mathematischen Modell zu verstehen und ist heute eines der grundlegenden Modelle der statistischen Mechanik mit vielen anderen Anwendungsbereichen.
Der mathematische Rahmen ist zunächst ähnlich wie bei der Perkolation: wir wählen

$$\Lambda := \mathbb{Z}^d \cap [-N, N]^d, \qquad \Omega = \{-1, 1\}^\Lambda = \{(x_i)_{i\in\Lambda} : x_i \in \{-1, 1\}\, \forall i\},$$

und $\mathcal{F} = \mathcal{P}(\Omega)$. Wir schreiben $\boldsymbol{x} = (x_i)_{i\in\Lambda}$ für die Elemente von Ω, und $X_i(\boldsymbol{x}) = x_i$ für die Koordinatenabbildung, die $\boldsymbol{x}$ an der Stelle $i \in \Lambda$ auswertet. X_i ist eine der zwei Richtungen, in die der „Elementarmagnet“ an der Stelle i zeigen kann; die Beschränkung auf zwei Richtungen ist eine mathematische Vereinfachung. Aus historischen Gründen (Zusammenhang mit quantenmechanischer Theorie des Magnetismus) wird $X_i(\boldsymbol{x})$ als **Spin** an der Stelle i in der Konfiguration $\boldsymbol{x}$ bezeichnet.

Das W-Maß $\mathbb{P}_{\beta,\mu,N}$ ist für $\beta \geqslant 0$ und $\mu \in \mathbb{R}$ gegeben durch die Zähldichte

$$p_{\beta,\mu,N}(\boldsymbol{x}) = \mathbb{P}_{\beta,\mu,N}(X_i = x_i\, \forall i \in \Lambda) := \frac{1}{Z_N(\beta,\mu)} \exp\Big(\beta \sum_{i,j\in\Lambda:|i-j|=1} x_i x_j + \mu \sum_{i\in\Lambda} x_i\Big)$$

Hierbei ist

$$Z_N(\beta,\mu) = \sum_{\boldsymbol{x}\in\Omega} \exp\Big(\beta \sum_{i,j\in\Lambda:|i-j|=1} x_i x_j + \mu \sum_{i\in\Lambda} x_i\Big),$$

so dass $p_{\beta,N}$ tatsächlich eine Zähldichte ist. Die Interpretation dieses Ausdruckes ist die folgende: benachbarte Spins, die den gleichen Wert haben („in die gleiche Richtung zeigen“), tragen

eine Term $+1$ zu der Summe bei, solche, die in unterschiedliche Richtung zeigen, eine -1. Die Summe (und damit die Zähldichte bei $\boldsymbol{x}$) wird also um so größer, je mehr die Paare mit „gleicher Richtung“ in der Mehrheit sind, und Elemente von Ω, in denen die viele Paare mit gleicher Richtung vorkommen, sind somit wahrscheinlicher als solche, in denen es nur wenige dieser Paare gibt - benachbarte Elementarmanete richten sich bevorzugt in die gleiche Richtung aus! Der Parameter β (aus physikalischer Motivation heraus „inverse Temperatur“ genannt) regelt, wie stark dieser Effekt ist. μ modelliert die Stärke des äußeren Magnetfeldes und gibt den Spins eine bevorzugte Richtung: ist $\mu > 0$, so sind Konfigurationen wahrscheinlicher, in denen viele der x_i positiv sind, für $\mu < 0$ solche mit vielen negativen x_i. Physikalisch ist μ die Stärke eines von außen angelegten Magnetfeldes, das den Elementarmagneten eine bevorzugte Richtung gibt. Für $\mu = \beta = 0$ sind alle Konfigurationen gleich wahrscheinlich, und $\mathbb{P}_{\beta,\mu,N}$ ist dann einfach das Produktmaß aus Gleichverteilungen auf $\{-1, 1\}$.

Das Interessante an dem W-Maß $\mathbb{P}_{\beta,\mu,N}$ ist die folgende Konkurrenz zweier Effekte: Sei zunächst $\mu = 0$. Die Zähldichte $p_{\beta,N}(\boldsymbol{x})$ ist dann (für $\beta > 0$) für solche $\boldsymbol{x}$ am größten, die sehr „geordnet“ aussehen. Man sagt auch, die **Energie** $H(\boldsymbol{x}) := -\sum_{i,j\in\Lambda:\sim=1} x_i x_j$ ist für geordnete Zustände klein. Andererseits gibt es viel mehr „ungeordnete“ als geordnete Zustände: diese etwas schwammige Aussage kann man sich klar machen, indem man den komplett geordneten Zustand $\boldsymbol{x}$ mit $x_i = 1$ für alle i ansieht - bei ihm (und dem Zustand $-\boldsymbol{x}$) ist $p_{\beta,\mu,N}$ maximal. Aber es gibt nur zwei solche Zustände, und schon von denjenigen Zuständen, wo auch nur ein einziger Spin „falsch“ ist, gibt es $(2N + 1)^d$ Stück. Die W'keit, wenigstens ein Paar benachbarter Spins zu finden, die nicht übereinstimmen, wird also für festes β und sehr große N (und der Grenzwert $N \to \infty$ wird uns interessieren) viel größer sein als die, kein solches Paar zu finden, aber wiederum viel kleiner als die, mindestens zwei solche Paare zu finden, etc. Für den Fall $\mu \neq 0$ ist das Bild ähnlich, außer dass hier eine der beiden Richtungen bevorzugt wird.

Wir wollen uns die *erwartete mittlere Magnetisierung*

$$m_N(\beta, \mu) := \mathbb{E}_{\beta,\mu,N}\Big(\frac{1}{|\Lambda|}\sum_{i\in\Lambda} x_i\Big)$$

ansehen; ist sie ungleich 0, so hat das Material einen Überschuss an Elementarmagneten einer Richtung und ist selbst magnetisch. Sie kann, möglicherweise etwas überraschend, allein aus der Normierungskonstante $Z_N(\beta, \mu)$ berechnet werden:

$$\begin{aligned}\frac{1}{|\Lambda|}\partial_\mu \log Z_N(\beta,\mu) &= \frac{1}{|\Lambda|}\frac{1}{Z_N(\beta,\lambda)}\sum_{\boldsymbol{x}\in\Omega}\partial_\mu \exp\Big(\beta\sum_{i,j\in\Lambda:|i-j|=1} x_i x_j + \mu\sum_{i\in\Lambda} x_i\Big) = \\ &= \frac{1}{|\Lambda|}\sum_{\boldsymbol{x}\in\Omega}\Big(\sum_{i\in\Lambda} x_i\Big)p_{\beta,\mu,N}(\boldsymbol{x}) = \mathbb{E}_{\beta,\mu,N}\Big(\frac{1}{|\Lambda|}\sum_{i\in\Lambda} x_i\Big) = m_N(\beta,\mu).\end{aligned}$$

Wir betrachten nun $m_N(\beta, \mu)$ und variieren μ, also das äußere Magnetfeld. Die Annahme dabei ist, dass wir das Magnetfeld so langsam variieren, dass das System immer Zeit hat, sich auf den jeweils zu einem festen μ gehörigen Zustand, der durch $\mathbb{P}_{\lambda,\mu,N}$ beschrieben wird, einzuschwingen. Hysterese (im endlichen Volumen $|\Lambda|$) würde man also beobachten, wenn $\lim_{\mu\searrow 0} m_N(\beta, \mu) > 0$ ist. Dies ist allerdings erst einmal sicher nicht der Fall, denn $\mu \mapsto m_N(\beta, \mu)$ ist eine stetige Funktion ist (denn es ist eine endliche Summe stetiger Funktionen, wie man aus den obigen Formeln sieht), und daher gilt $\lim_{\mu\to 0} m_N(\beta, \mu) = m_N(\beta, 0) = 0$. Andererseits ist

$m_N(\beta,\mu) = -m_N(\beta,-\mu)$ aus Symmetriegründen (ersetze jedes x_i durch ein $-x_i$), und daher ist $m_N(\beta,0) = 0$.

Anders sieht es aus, wenn man den Grenzübergang zum Kontinuum (zum unendlich großen System macht): zunächst muss man zeigen, dass $m(\beta,\mu) = \lim_{N\to\infty} m_N(\beta,\mu)$ existiert, aber wenn man das hat, dann kann man fragen, ob $\lim_{\mu\searrow 0} m(\beta,\mu) > 0$ ist. Dies ist für $d \geqslant 2$ und hinreichend große β der Fall, ist aber nicht ganz einfach zu zeigen, und wird in der Vertiefungsveranstaltung „mathematische statistische Mechanik“ behandelt.

(3.28) Beispiel: Stationäre Verteilung einer Markovkette

(X_n) sei eine Markovkette mit Zustandsraum E (endlich oder abzählbar), und mit Übergangsmatrix $p : E \times E \to [0,1]$. Wir interessieren uns dafür, ob wir die Markovkette mit einem W-Maß μ starten können, so dass gilt: $\mathbb{P}^\mu(X_n = x) = \mu(\{x\})$ für alle $x \in E$; mit anderen Worten ob wir ein Startmaß μ finden, so dass sich die Aufenthaltsw'keiten in allen Zuständen im Lauf der Zeit nicht ändern, die Kette also im „Gleichgewicht“ ist. Ein solches μ nennen wir **stationäre Verteilung** der Markovkette.

Mathematisch suchen wir (wenn wir wieder μ mit seiner Zähldichte identifizieren) eine Funktion

$$\mu : E \to \mathbb{R}^+ \qquad \text{mit } \sum_{x\in E} \mu(x) = 1 \quad \text{und} \quad \sum_{x\in E} \mu(x)p(x,y) = \mu(y)\,\forall x,y \in E.$$

Denn die zweite Gleichung bedeutet ja $\mathbb{P}^\mu(X_1 = y) = \mu(y)$ für alle y, und impliziert (durch Iteration) $\mathbb{P}^\mu(X_n = y) = \mu(y)$ für alle n und alle y. In Matrixschreibweise liest sich das (mit $P = p(x,y)_{x,y\in E}$): $\mu P = \mu$, wir suchen also einen **Links-Eigenvektor** von P (mit speziellen Zusatzeigenschaften) - ein klassisches Problem der der linearen Algebra (zumindest für endliches E)! Wir erinnern uns an die Definition der Trefferzeit

$$\tau_x(\omega) = \inf\{n \geqslant 1 : X_n(\omega) = x\},$$

und definieren die erwartete Anzahl λ_y an Treffern an einer Stelle $y \in E$ zwischen zwei Besuchen von $x \in E$, nämlich

$$\lambda_x(y) := \mathbb{E}^x\Big(\sum_{n=1}^{\infty} \mathbb{1}_{\{X_n=y\}}\mathbb{1}_{\{\tau_x \geqslant n\}}\Big) = \sum_{n=1}^{\infty} \mathbb{P}^x(X_n = y, \tau_x \geqslant n).$$

Behauptung: Es existiere ein $x \in E$ mit $\mathbb{E}^x(\tau_x) < \infty$. Dann existiert eine stationäre Verteilung (mit Zähldichte μ) der Markovkette, und es gilt

$$\mu(y) = \frac{\lambda_x(y)}{\mathbb{E}^x(\tau_x)} \qquad \forall y \in E.$$

Zum **Beweis** rechnen wir zunächst nach, dass $(\lambda_x(y))_{y\in E}$ tatsächlich ein Links-Eigenvektor von P ist, also dass $\sum_{z\in E} \lambda_x(z)p(z,y) = \lambda_x(y)$ gilt. Zunächst ist $\{\tau_x \geqslant n\} = \{X_1 \neq x, \ldots, X_{n-1} \neq x\}$, und die Markov-Eigenschaft liefert uns für alle z und n mit $\mathbb{P}^x(X_n = z, \tau_x \geqslant n) > 0$ die Gleichung

$$\begin{aligned}\mathbb{P}^x(X_{n+1} = y \,|\, X_n = z, \tau_x \geqslant n) &= \mathbb{P}^x(X_{n+1} = y \,|\, X_n = z, X_1 \neq x, \ldots X_{n-1} \neq x) = \\ &= \mathbb{P}^x(X_{n+1} = y \,|\, X_n = z) = p(z,y)\end{aligned}$$

Somit erhalten wir für alle $z \in E$

$$\lambda_x(z)p(z,y) = \sum_{n=1}^{\infty} \mathbb{P}^x(X_n = z, \tau_x \geqslant n)\mathbb{P}^x(X_{n+1} = y \,|\, X_n = z, \tau_x \geqslant n) =$$
$$= \sum_{n=1}^{\infty} \mathbb{P}^x(X_n = z, X_{n+1} = y, \tau_x \geqslant n).$$

Da die Ereignisse $\{X_n = z\}$ disjunkt sind und $\bigcup_{z\in E}\{X_n = z\} = \Omega$ gilt, liefert uns die σ-Additivität von $\mathbb{P}^x$ nun

$$\sum_{z\in E} \lambda_x(z)p(z,y) = \sum_{n=1}^{\infty}\sum_{z\in E} \mathbb{P}^x(X_n = z, X_{n+1} = y, \tau_x \geqslant n) = \sum_{n=1}^{\infty} \mathbb{P}^x(X_{n+1} = y, \tau_x \geqslant n).$$

Es gilt

$$\mathbb{P}^x(X_{n+1} = y, \tau_x \geqslant n) = \mathbb{P}^x(X_{n+1} = y, \tau_x = n) + \mathbb{P}(X_{n+1} = y, \tau_x \geqslant n+1),$$

und einsetzen der beiden Terme in die Summe über n gibt für den ersten

$$\sum_{n=1}^{\infty} \mathbb{P}^x(X_{n+1} = y, \tau_x = n) = \sum_{n=1}^{\infty} \mathbb{P}^x(X_{n+1} = y \,|\, \tau_x = n)\mathbb{P}^x(\tau_x = n) = p(x,y)\sum_{n=1}^{\infty} \mathbb{P}^x(\tau_x = n) = p(x,y);$$

die letzte Gleichheit folgt aus $\mathbb{P}^x(\tau_x = \infty) = 0$, was wiederum eine Folge der Annahme $\mathbb{E}^x(\tau_x) < \infty$ ist. Für den zweiten Term erhalten wir

$$\sum_{n=1}^{\infty} \mathbb{P}^x(X_{n+1} = y, \tau_x \geqslant n+1) = \sum_{n=2}^{\infty} \mathbb{P}^x(X_n = y, \tau_x \geqslant n) = \lambda_x(y) - \mathbb{P}^x(X_1 = y, \tau_x \geqslant 1).$$

Da der negative Term in der letzten Zeile gleich $p(x,y)$ ist, folgt wie behauptet $\sum_{z\in E} \lambda_x(z)p(z,y) = \lambda_x(y)$. Außerdem ist nach der Definition klar, dass $\lambda_x(y) \geqslant 0$ ist. Schließlich ist

$$\sum_{y\in E} \lambda_x(y) = \sum_{y\in E}\sum_{n=1}^{\infty} \mathbb{P}^x(X_{n+1} = y, \tau_x \geqslant n) = \sum_{n=1}^{\infty} \mathbb{P}^x(\tau_x \geqslant n) = \mathbb{E}(\tau_x),$$

das letzte = wegen (3.17 c). Der Beweis ist damit beendet. □

Kommentare: a) Wenn E endlich ist, so lässt sich die Voraussetzung $\mathbb{E}^x(\tau_x) < \infty$ immer für mindestens ein $x \in E$ erfüllen (Beweis: Übung!)

b) Wir haben gezeigt, dass (unter den gegebenen Bedingungen) eine stationäre Verteilung existiert, aber nicht, dass sie eindeutig ist. Dies ist auch nicht immer der Fall. Zur Übung verdeutliche man sich das am Beispiel des Ruins des Spielers (was sind hier die Zustände mit $\mathbb{E}^x(\tau_x) < \infty$?).

c) Man kann relativ einfach zeigen, dass es keine stationäre Verteilung für die einfache Irrfahrt geben kann (Übung!). Daraus folgt insbesondere, dass $\mathbb{E}^x(\tau_x) = \infty$ für alle $x \in \mathbb{Z}^d$, eine Gleichung, die man auch (allerdings mit Aufwand) kombinatorisch beweisen kann.

Erzeugende Funktionen

(3.29) Definition

a) X sein eine ZV mit Werten in $\mathbb{N}_0$. Die Funktion

$$\varphi_X : [0,1] \to [0,1], \qquad u \mapsto \varphi_X(u) := \mathbb{E}(u^X) = \sum_{k=0}^{\infty} u^k \mathbb{P}(X=k)$$

heißt **erzeugende Funktion** der Verteilung von X.

b) Ist $\mathbb{P}$ ein W-Maß auf $\mathbb{N}_0$ mit Zähldichte $(p_n)_{n\in\mathbb{N}_0}$, dann heißt

$$\varphi_{\mathbb{P}} : [0,1] \to [0,1]. \qquad u \mapsto \varphi_{\mathbb{P}}(u) := \int_{\mathbb{N}_0} u^k \mathbb{P}(\mathrm{d}k) = \sum_{k=0}^{\infty} u^k p_k$$

erzeugende Funktion von $\mathbb{P}$.

Beispiele:

a) Für $X \sim \mathrm{Poi}_\alpha$ ist $\varphi_X(u) = \mathrm{e}^{-\alpha} \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} u^k = \mathrm{e}^{\alpha(u-1)}$.

b) Für $Y \sim \mathrm{Geo}_p$ ist $\varphi_Y(u) = \sum_{k=0}^{\infty} p(1-p)^k u^k = \frac{p}{1-(1-p)u}$.

c) Für $Z \sim B_{n,p}$ ist $\varphi_Z(u) = \sum_{k=0}^{n} u^k \binom{n}{k} p^k (1-p)^{n-k} = (1-p+pu)^n$.

(3.30) Eigenschaften der erzeugenden Funktion

X, Y seien $\mathbb{N}_0$-wertige ZVen, φ_X und φ_Y ihre erzeugenden Funktionen.

a) φ_X ist monoton wachsend auf $[0,1]$, und $\varphi_X \in C^\infty([0,1))$.

b) φ_X charakterisiert die Verteilung von X, d.h. wenn $\varphi_X(u) = \varphi_Y(u)$ für ZVen X, Y, dann ist $X \sim Y$. Konkret gilt

$$\mathbb{P}(X=k) = \frac{1}{k!} \frac{\mathrm{d}^k}{\mathrm{d}u^k} \varphi_X(u)\Big|_{u=0} \qquad \forall k \in \mathbb{N}_0,$$

und insbesondere $\mathbb{P}(X=0) = \varphi_X(0)$.

c) Es gilt

$$\mathbb{E}(X) = \varphi'_X(1) := \lim_{u \nearrow 1} \varphi'_X(u)$$

(beide Seiten können $= \infty$ sein).

d) Allgemeiner gilt für $k \in \mathbb{N}$: $X \in \mathcal{L}^k$ genau dann wenn $\lim_{u\nearrow 1} \partial^k \varphi_X(u) < \infty$, und in jedem Fall gilt

$$\mathbb{E}(X(X-1)\cdots(X-k+1)) = \partial_u^k \varphi_X(1) := \lim_{u\nearrow 1} \partial_u^k \varphi_X(u).$$

e) Seien X und Y unabhängig, und sei φ_{X+Y} die erzeugende Funktion von $X+Y$. Dann ist

$$\varphi_{X+Y}(u) = \varphi_X(u)\varphi_Y(u) \qquad \forall u \in [0,1].$$

Beweis:

a) $\varphi(u) = \sum_{j=0}^{\infty} \mathbb{P}(X=j) u^j$ ist eine Potenzreihe mit nichtnegativen Koeffizienten, und wegen $\varphi(1) = 1$ ist ihr Konvergenzradius $\geqslant 1$. Aus der ersten Aussage folgt die Monotonie, aus der zweiten die C^∞-Eigenschaft.

b) Das folgt aus der Taylor-Entwicklung: der j-te Koeffizient einer Potenzreihe mit strikt positivem Konvergenzradius ist $1/j!$ mal die j-te Ableitung der dargestellten Funktion an der Stelle 0.

c) Für $u < 1$ ist

$$\mathbb{E}(Xu^X) = \sum_{j \geqslant 0} ju^j \mathbb{P}(X = j) = u \sum_{j \geqslant 0} (\partial_u u^j)\mathbb{P}(X = j) \overset{(*)}{=} u\partial_u \sum_{j \geqslant 0} u^j \mathbb{P}(X = j) = u\varphi'_X(u)$$

In $(*)$ wurde benutzt, dass die Reihe $\sum_{j \geqslant 0} ju^j \mathbb{P}(X = j)$ auf $[0, u+(1-u)/2]$ gleichmäßig konvergiert und man daher Summe und Ableitung vertauschen darf. Wegen monotoner Konvergenz gilt $\lim_{u \nearrow 1} \mathbb{E}(Xu^X) = \mathbb{E}(X)$, also fogt die Behauptung durch durchführen des Grenzwertes $u \nearrow 1$ auf beiden Seiten.

d) Das gleiche Argument wie in c) macht man mit der Reihe

$$\mathbb{E}(X(X-1)(X-2)\cdots(X-k+1)u^X) = \sum_{j=0}^{\infty} j(j-1)\cdots(j-k+1)u^j \mathbb{P}(X = j) =$$

$$= \sum_{j=k}^{\infty} j(j-1)\cdots(j-k+1)u^j \mathbb{P}(X = j) = u^k \sum_{j=k}^{\infty} \partial_u^k u^j \mathbb{P}(X = j).$$

Die vorletzte Darstellung zeigt, dass $u \mapsto X(X-1)(X-2)\cdots(X-k+1)u^X)$ weitherhin monoton steigend ist, also kann man tatsächlich wie oben argumentieren.

e) Da X und Y unabhängig sind, sind wegen (1.58 a) auch u^X und u^Y unabhängig. Somit ist

$$\varphi_{X+Y}(u) = \mathbb{E}(u^{X+Y}) = \mathbb{E}(u^X u^Y) = \mathbb{E}(u^X)\mathbb{E}(u^Y) = \varphi_X(u)\varphi_Y(u). \qquad \square$$

(3.31) Beispiel: Galton-Watson-Prozess

a) Intuition und Geschichte:
Der Galton-Watson-Prozess beschreibt die Entwicklung einer Population, in der ein Individuum eine zufällige Anzahl Z an Nachkommen erzeugt, und dann stirbt. Alle diese Nachkommen bekommen selbst wieder eine zufällige Anzahl von Nachkommen. Diese Anzahl hat die gleiche Verteilung wie Z, ist aber unabhängig von Z. Mutationen oder Veränderungen finden also nicht statt. Dieser grundlegende Mechanismus hat z.B. Anwendungen in der Genealogie (das Überleben von Familiennamen im England des 19. Jahrhunderts, Galton und Watson 1874), der Physik (Anzahl freier Neutronen bei einer nuklearen Kettenreaktion, Manhattan Projekt, Leo Slizard ca. 1940), sowie in der Biologie und Epidemiologie (Überleben von seltenen Mutationen oder von Krankheitserregern in einer Population, heutzutage das Hauptanwendungsgebiet - siehe auch weiter unten!).

b) Definition:
Sei Z eine $\mathbb{N}_0$-wertige ZV mit $\mathbb{P}(Z = 0) > 0$. Z modelliert die Anzahl der Nachkommen eines Individuums. $(Z_{j,k})_{j,k\in\mathbb{N}}$ sei eine Familie von unabhängigen ZVen mit $Z_{j,k} \sim Z$ für alle j, k. Die doppelte Indizierung brauchen wir später für eine saubere Notation: $Z_{j,k}$ gibt dann die Anzahl der Nachkommen des j-ten Individuums in der k-ten Generation an. Der **Galton-Watson-Prozess** mit Nachkommenverteilung Z ist die Markovkette (X_n) mit Zustandsraum $E = \mathbb{N}_0$

und Übergangsmatrix

$$p(n,m) = \mathbb{P}(X_{k+1} = m \,|\, X_k = n) := \mathbb{P}\Big(\sum_{j=1}^{n} Z_{j,k} = m\Big).$$

Eine äquivalente Beschreibung von (X_n) ist die rekursive Definition

$$X_{k+1}(\omega) := \sum_{j=1}^{X_k(\omega)} Z_{j,k}(\omega).$$

X_n ist also die Größe der Gesamtpopulation nach n Generationen. Wir nehmen immer $X_0 = 1$ an; ist man an dem Fall interessiert, wo man mit m Individuen startet, so kann man m-mal unabhängig den Fall $X_0 = 1$ betrachten und die Ergebnisse addieren.

c) Die Aussterbe-Wahrscheinlichkeit:
Wegen $p(0,0) = 1$ ist 0 ein absorbierender Zustand von (X_n). Wie im Beispiel (1.45) interessieren wir uns für die Trefferzeit

$$\tau_0 = \inf\{k \geqslant 1 : X_k = 0\}.$$

Die **Aussterbe-Wahrscheinlichkeit** ist durch $\mathbb{P}(\tau_0 < \infty)$ gegeben, und falls $\mathbb{P}(\tau_0 < \infty) = 1$, dann kann man nach der **erwarteten Lebensdauer** $\mathbb{E}(\tau_0)$ der Population fragen.

d) Einfache Vorüberlegungen: Für den Fall $\mathbb{P}(Z \geqslant 2) = 0$ ist (X_k) sehr einfach: dann existiert $p > 0$ mit $\mathbb{P}(Z = 1) = p = 1 - \mathbb{P}(Z = 0)$, und wir erhalten

$$\mathbb{P}(X_k = 1) = \mathbb{P}(X_j = 1 \,\forall j \leqslant k) = p^k, \qquad \text{und} \qquad \mathbb{P}(\tau_0 = k+1) = p^k(1-p) \,\forall k \geqslant 1.$$

Daher ist $\tau_0 \sim \mathrm{Geo}_{1-\mathrm{p}} + 1$, $\mathbb{P}(\tau_0 < \infty) = 1$, und

$$\mathbb{E}(\tau_0) = 1 + (1-p)\sum_{j=0}^{\infty} jp^j = 1 + p(1-p)\partial_p \sum_{j=0}^{\infty} p^j = 1 + p(1-p)\partial_p \frac{1}{1-p} = \frac{1}{1-p}.$$

Für den allgemeinen Fall mit $\mathbb{P}(Z \geqslant 2) > 0$ machen wir folgende **Annahme:** es gelte

$$\mu := \mathbb{E}(Z) < \infty.$$

Dann können wir eine Rekursion für den Erwartungswert von X_k berechnen: es gilt

$$\mathbb{E}(X_{k+1}) = \sum_{m=0}^{\infty} \mathbb{E}(X_{k+1} \mathbb{1}_{\{m\}}(X_k)) = \sum_{m=0}^{\infty} \mathbb{E}\Big(\sum_{j=1}^{m} Z_{j,k} \mathbb{1}_{\{m\}}(X_k)\Big) = (*).$$

Da X_k nur von den $(Z_{j,l})_{j \in \mathbb{N}, l \leqslant k-1}$ abhängt, ist X_k unabhängig von den $Z_{j,k}$, und es ist

$$(*) = \sum_{m=0}^{\infty} \sum_{j=1}^{m} \mathbb{E}(Z_{j,k})\mathbb{P}(X_k = m) = \sum_{m=0}^{\infty} m\mu \mathbb{P}(X_k = m) = \mu \mathbb{E}(X_k).$$

Es folgt $\mathbb{E}(X_k) = \mu^k \mathbb{E}(X_0) = \mu^k$, und mit der Chebyshev-Ungleichung erhalten wir

$$\mathbb{P}(\tau_0 \geqslant n) = \mathbb{P}(Z_n \geqslant 1) \leqslant \mathbb{E}(Z_n) = \mu^n.$$

Für $\mu < 1$ haben wir also bereits $\mathbb{P}(\mu = \infty) = \lim_{n\to\infty} \mathbb{P}(\mu \geqslant n) = 0$ gezeigt, sowie die Abschätzung

$$\mathbb{E}(\tau_0) = \sum_{n=1}^{\infty} \mathbb{P}(\tau_0 \geqslant n) \leqslant \sum_{n=1}^{\infty} \mu^n = \frac{\mu}{1-\mu}.$$

Die Zahl μ entspricht übrigens dem R-Wert in der Epidemiologie.

e) Genauere Untersuchung mit erzeugenden Funktionen:
Wir definieren für $0 \leqslant u \leqslant 1$

$$\varphi_0(u) := \mathbb{E}(u^{X_0}) = u, \quad \varphi_1(u) := \mathbb{E}(u^{X_1}) = \mathbb{E}(u^Z), \qquad \varphi_k(u) := \mathbb{E}(u^{X_k})\, \forall k \geqslant 2.$$

Dann gilt (ähnlich wie in d)):

$$\varphi_{k+1}(u) = \mathbb{E}(u^{X_{k+1}}) = \sum_{m=0}^{\infty} \mathbb{E}(u^{X_{k+1}} \mathbb{1}_{\{m\}}(X_k)) = \sum_{m=0}^{\infty} \mathbb{E}\Big(u^{\sum_{j=1}^m Z_{j,k}}\Big)\mathbb{P}(X_k = m) =$$

$$= \sum_{m=0}^{\infty} \prod_{j=1}^{m} \mathbb{E}(u^{Z_{j,k}})\mathbb{P}(X_k = m) = \sum_{m=1}^{\infty} \varphi_1(u)^m \mathbb{P}(X_k = m) = \mathbb{E}\Big(\big(\varphi_1(u)\big)^{X_k}\Big) = \varphi_k\big(\varphi_1(u)\big).$$

Wir haben also $\varphi_k(u) = \varphi_1^{\circ k}(u)$ gezeigt, in Worten: die erzeugende Funktion von X_k erhält man, indem man die erzeugende Funktion von Z k-mal iteriert. Außerdem gilt

$$\xi := \mathbb{P}(\tau_0 < \infty) = \lim_{n\to\infty} \mathbb{P}(\tau_0 \leqslant n) = \lim_{n\to\infty} \mathbb{P}(X_n = 0) = \lim_{n\to\infty} \varphi_n(0),$$

und wegen der bewiesenen Rekursion für die erzeugenden Funktionen und deren Stetigkeit finden wir

$$\xi = \lim_{n\to\infty} \varphi_1(\varphi_{n-1}(0)) = \varphi_1(\lim_{n\to\infty} \varphi_{n-1}(0)) = \varphi_1(\xi).$$

In Worten: $\xi = \mathbb{P}(\tau_0 < \infty)$ ist ein **Fixpunkt** der Funktion φ_1, und tatsächlich ist ξ der kleinste nichtnegative Fixpunkt, denn für jeden Fixpunkt ζ von φ_1 ist (wegen der Monotonie von φ_n) $\varphi_n(0) \leqslant \varphi_n(\zeta) = \zeta$, und somit $\xi = \lim_{n\to\infty} \varphi_n(0) \leqslant \zeta$.

Wegen $\varphi(1) = 1$ für alle erzeugenden Funktionen ist 1 immer ein Fixpunkt von φ_1. Wann gibt es einen kleineren? Hierzu beachte man, dass unter der Annahme $\mathbb{P}(Z \geqslant 2) > 0$ die Funktion φ_1 strikt konvex ist, und dass wegen der Annahme $\mathbb{P}(Z = 0) > 0$ die Ungleichung $\varphi_1(0) > 0$ gilt. Der Graph der Funktion φ_1 beginnt also (bei $u = 0$) oberhalb der Diagonalen $f(x) = x$, endet (bei $u = 1$) genau auf dieser Diagonalen, und jeder Schnittpunkt mit der Diagonalen ist ein Fixpunkt. Wegen der strikten Konvexität gibt es höchstens einen solchen Schnittpunkt neben dem trivialen bei $u = 1$, und wegen der Stetigkeit gibt es genau einen solchen genau dann, wenn die Steigung von φ_1 an der Stelle $u = 1$ größer ist als die der Diagonalen, also größer als 1. Wegen (3.30 c) gilt $\mathbb{E}(Z) = \varphi_1'(1)$, und wir haben bewiesen:

Für einen Galton-Watson-Prozess gilt $\mathbb{P}(\tau_0 < \infty) < 1$ genau dann, wenn $\mathbb{E}(Z) > 1$ ist.

4. Grenzwertsätze

(4.1) Motivation und Überblick

Bei der Motivation des Erwartungswertes in (3.16) hatten wir argumentiert, dass etwa für eine $\mathbb{N}_0$-wertige ZV Y mit $\mathbb{P}(Y = k) = p_k$ die Formel $\mathbb{E}(Y) = \sum_{k=0}^{\infty} kp_k$ wie folgt gerechtfertigt werden kann: der Erwartungswert Y sollte das sein, was man bekommt, wenn man das mit Y modellierte Experiment oft unabhängig ausführt Ergebnisse mittelt. Formal: sind $(Y_n)_{n\in\mathbb{N}}$ unabhängige Kopien von Y, so sollte die Rechnung

$$\frac{1}{N}\sum_{n=1}^{N} Y_n(\omega) = \sum_{k=0}^{\infty} k\frac{|\{n \leqslant N : Y_n(\omega) = k\}|}{N} \overset{N\to\infty\,(!?)}{\longrightarrow} \sum_{k=0}^{\infty} kp_k = \mathbb{E}(Y)$$

zu einer beweisbaren Aussage verbessert werden können. Erstes Ziel dieses Kapitels ist es, dieses Ziel in recht guter Allgemeinheit zu erreichen - für *alle* $\omega \in \Omega$ kann das allerdings nicht wahr sein, wie ebenfalls in (3.16) diskutiert wurde - eine extreme Glückssträhne kann eben nie ganz ausgeschlossen werden. Daher werden wir uns zuerst überlegen, welche Konvergenzbegriffe hier angemessen sind. Dann folgt unter verschiedenen Voraussetzungen und in verschiedenen Konvergenzarten ein Beweis der Gleichung

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} Y_n = \mathbb{E}(Y).$$

Aussagen dieser Art (Konvergenz gemittelter Ergebnisse von Zufallsvariablen gegen eine Zahl) heißen **Gesetze der großen Zahl**.

Das zweite Ziel ist dann eine wesentliche Verfeinerung solcher Gesetze: schreibt man diese leicht um, erhält man Aussagen vom Typ

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \big(Y_n - \mathbb{E}(Y)\big) = 0.$$

Man kann sich nun fragen, ob die Normierung mit $\frac{1}{N}$ optimal ist - schließlich ist ja, falls die Aussage mit $1/N$ gilt, auch $\lim_{N\to\infty} \frac{1}{N^2}\sum_{n=1}^{N} \big(Y_n - \mathbb{E}(Y)\big) = 0$ und so weiter. Es wird sich herausstellen, dass (zumindest wenn $\sigma^2 := \mathbb{V}(Y)$ existiert und die ZVen unabhängig sind) die optimale Normierung diejenige mit Vorfaktor $1/\sqrt{N}$ ist - allerdings ist der Grenzwert dann nicht mehr eine Zahl, sondern selbst eine Zufallsvariable, und zwar eine **Normalverteilung** mit Mittelwert 0 und Varianz σ^2. Die Tatsache, dass diese Grenzverteilung *universell* ist, d.h. nur von der Varianz von Y aber nicht von der Verteilung von Y abhängt, begründet die zentrale Bedeutung der Normalverteilung. Sätze, in denen die Konvergenz einer optimal normierten Folge von Verteilungen gegen eine Normalverteilung bewiesen wird, nennt man **zentrale Grenzwertsätze**.

(4.2) Definition

(X_n) sei eine Folge von reellen ZVen, X eine reelle ZV.

a) Wir sagen, das (X_n) **stochastisch gegen X konvergiert**, wenn gilt:

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

b) Wir sagen, dass (X_n) **$\mathbb{P}$-fast sicher gegen X konvergiert**, wenn es eine Ausnahmemenge $\Omega_0 \in \mathcal{F}$ mit $\mathbb{P}(\Omega_0) = 0$ gibt, so dass gilt:

$$\forall \omega \notin \Omega_0 : \quad \lim_{n\to\infty} X_n(\omega) = X(\omega)$$

Eine äquivalente Formulierung ist: (X_n) konvergiert gegen X $\mathbb{P}$-fast sicher, falls

$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| = 0) = 1.$$

Bemerkungen: a) Die stochastische Konvergenz sagt also, dass man für jeden beliebig klein gewählten „Sicherheitsabstand" ε die W'keit, diesen Sicherheitsabstand mit $|X_n - X|$ zu überschreiten, so klein machen kann wie man will, wenn man nur n groß genug wählt. Dies heißt aber nicht, dass $|X_n(\omega) - X(\omega)|$ „im Durchschnitt" besonders klein sein muss - für diejenigen $\omega \in \Omega$, für die $|X_n(\omega) - X(\omega)|$ den Sicherheitsabstand überschreitet, hat man überhaupt keine Kontrolle darüber, wie groß $|X_n(\omega) - X(\omega)|$ ist. Somit ist es beispielsweise möglich, dass zwar $X_n \to X$ stochastisch, aber $\lim_{n\to\infty} \mathbb{E}(|X_n - X|) = \infty$ gilt (Übung). In der Integrationstheorie lernen Sie allgemein die Konvergenz in L^p kennen, die dann gilt, wenn $\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0$. Die stochastische Konvergenz impliziert also nicht die L^p-Konvergenz, umgekehrt aber folgt aus der L^p-Konvergenz die stochastische (wegen der Chebyshev-Ungleichung).

b) Die stochastische und die fast sichere Konvergenz unterscheiden sich „nur" in der Platzierung eines Grenzwertes: zunächst ist

$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| = 0) = 1 \qquad \text{genau dann, wenn} \qquad \mathbb{P}(\limsup_{n\to\infty} |X_n = X| > \varepsilon) = 0 \quad \forall \varepsilon > 0$$

für alle $\varepsilon > 0$. Zum Beweis setze $A_m := \{\omega \in \Omega : \limsup_{n\to\infty} |X_n(\omega) - X(\omega)| \geqslant \frac{1}{m}\}$, und benutze die Steitigkeit des Maßes von unten, Details als Übung. Somit gilt

$$X_n \to X \text{ stochastisch, genau dann wenn} \quad \forall \varepsilon > 0 : \limsup_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$
$$X_n \to X \ \mathbb{P}\text{-fast sicher, genau dann wenn} \quad \forall \varepsilon > 0 : \mathbb{P}(\limsup_{n\to\infty} |X_n - X| > \varepsilon) = 0.$$

c) Aus b) kann man schnell sehen, dass die fast sichere Konvergenz stärker als die stochastische Konvergenz ist: zunächst rechnen wir

$$\begin{aligned}\mathbb{P}(\limsup_{n\to\infty} |X_n - X| \leqslant \varepsilon) &\leqslant \mathbb{P}(\liminf_{n\to\infty} |X_n - X| \leqslant \varepsilon) = \mathbb{E}(\liminf_{n\to\infty} \mathbb{1}_{\{|X_n - X| \leqslant \varepsilon\}}) \\ &\leqslant \liminf_{n\to\infty} \mathbb{P}(|X_n - X| \leqslant \varepsilon),\end{aligned}$$

wobei die letze Ungleichung das Lemma von Fatou war. Daher gilt

$$\begin{aligned}\mathbb{P}(\limsup_{n\to\infty} |X_n - X| > \varepsilon) &= 1 - \mathbb{P}(\limsup_{n\to\infty} |X_n - X| \leqslant \varepsilon) \geqslant 1 - \liminf_{n\to\infty} \mathbb{P}(|X_n - X| \leqslant \varepsilon) \\ &= \limsup_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon).\end{aligned}$$

Falls also $\mathbb{P}(\limsup_{n\to\infty} |X_n - X| > \varepsilon) = 0$ gilt, so muss auch $\limsup_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$ gelten, somit folgt aus der fast sicheren die stochastische Konvergenz.

d) Andersherum folgt aus der stochastischen Konvergenz nicht die fast sichere. Der Grund ist, dass die „Ausnahmemenge" bei der stochastischen Konvergenz für jedes n eine andere sein kann, und diese Mengen zusammengenommen sogar wieder ganz Ω ergeben können. Beispiele

hierfür gibt es in den Übungen.

(4.3) Nomenklatur

Im Folgenden werden wir es sehr oft mit einer Folge (X_n) von unabhängigen ZVen zu tun haben, die alle die gleiche Verteilung haben, also $X_1 \sim X_2 \sim X_3 \cdots$. Hier sagen wir kurz, die (X_n) seien **unabhängig und identisch verteilt**, und kürzen das mit **uiv** ab.

(4.4) Schwaches Gesetz der großen Zahl, einfachste Variante

(X_i) seien reelle, uiv ZVen, und es gelte $\mathbb{V}(X_1) < \infty$. Dann gilt für alle $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\Big(\big|\frac{1}{n}\sum_{j=1}^{n} X_i - \mathbb{E}(X_1)\big| \geqslant \varepsilon\Big) = 0,$$

mit anderen Worten: die Folge $(\bar{S}_n)$ mit $\bar{S}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ konvergiert stochastisch gegen die Zahl $\mathbb{E}(X_1)$.
Beweis: Mit $\bar{S}_n$ wie oben gilt $\mathbb{E}(\bar{S}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mathbb{E}(X_1)$, und

$$\mathbb{V}(\bar{S}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n}\mathbb{V}(X_1).$$

Die Chebyshev-Ungleichung liefert nun

$$\mathbb{P}\Big(\big|\frac{1}{n}\sum_{j=1}^{n} X_i - \mathbb{E}(X_1)\big| \geqslant \varepsilon\Big) = \mathbb{P}(|\bar{S}_n - \mathbb{E}(\bar{S}_n)| \geqslant \varepsilon) \leqslant \frac{1}{\varepsilon^2}\mathbb{V}(\bar{S}_n) = \frac{1}{n\varepsilon^2},$$

und im Grenzwert $n \to \infty$ folgt die Behauptung. □

Beispiel: Für die einfache Irrfahrt (S_n) bedeutet das, dass $\frac{1}{n}S_n$ stochastisch gegen 0 konvergiert. Dies (und mehr) wissen wir allerdings auch schon aus (3.21 b), wenn wir dort $c = \varepsilon n$ setzen. Mit dem schwachen Gesetz der großen Zahl können wir dies allerdings auch für andere (nicht-einfache) Irrfahrten leicht zeigen, also für solche, wo die Verteilung des nächsten Schrittes Erwartungswert 0 und endliche Varianz hat, aber ansonsten beliebig ist, und insbesondere keine diskrete ZV sein muss.

Die folgenden zwei Versionen des schwachen GGZ schwächen dessen Voraussetzungen in verschiedene Richtungen ab.

(4.5) Schwaches Gesetz der großen Zahl, verallgemeinerte Fassung mit Varianzbedingung

(X_n) sei eine Folge reeller ZVen, für die gilt:
(i): Die X_i sind paarweise unkorreliert, also $\mathrm{Cov}(X_i, X_j) = 0$ falls $i \neq j$.
(ii): Die X_i haben uniform beschränkte Varianzen, also $v := \sup\{\mathbb{V}(X_i) : i \in \mathbb{N}\} < \infty$.

Dann gilt für alle $n \in \mathbb{N}$ und alle Folgen $(\varepsilon_n)_{n\in\mathbb{N}}$ mit $\varepsilon_n > 0$ für alle n:

$$\mathbb{P}\Big(\big|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))\big| \geqslant \varepsilon_n\Big) \leqslant \frac{v}{n\varepsilon_n^2}$$

Insbesondere konvergiert $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))$ stochastisch gegen 0 falls $n\varepsilon_n^2 \to \infty$.
Beweis: Übung, sehr ähnlich wie oben. □

(4.6) Schwaches Gesetz der großen Zahl, Version ohne Varianzbedingung

(X_n) sei eine Folge identisch verteilter reeller ZVen, es gelte $X_n \perp\!\!\!\perp X_m$ für $m \neq n$ (das ist eine schwächere Voraussetzung als diejenige, dass die X_i uiv sind), und es gelte $\mathbb{E}(|X_1|) < \infty$. Dann gilt für alle $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\Big(\big|\frac{1}{n}\sum_{j=1}^{n} X_i - \mathbb{E}(X_1)\big| \geqslant \varepsilon\Big) = 0,$$

mit anderen Worten: die Folge $(\bar{S}_n)$ mit $\bar{S}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ konvergiert stochastisch gegen die Zahl $\mathbb{E}(X_1)$.
Beweis: Wir zerlegen X_i in zwei Teile:

$$X_i(\omega) = \underbrace{X_i(\omega)\mathbb{1}_{\{|X_i(\omega)|<i^{1/4}\}}}_{=:Y_i(\omega)} + \underbrace{X_i(\omega)\mathbb{1}_{\{|X_i(\omega)|\geqslant i^{1/4}\}}}_{=:Z_i(\omega)}.$$

Wegen $\mathbb{E}(X_1) = \mathbb{E}(X_i) = \mathbb{E}(Y_i) + \mathbb{E}(Z_i)$ für alle i gilt mit der Dreiecksungleichung

$$\Big|\frac{1}{n}\sum_{i=1}^{n}(X_i(\omega) - \mathbb{E}(X_1))\Big| \leqslant \Big|\frac{1}{n}\sum_{i=1}^{n}(Y_i(\omega) - \mathbb{E}(Y_i))\Big| + \Big|\frac{1}{n}\sum_{i=1}^{n}(Z_i(\omega) - \mathbb{E}(Z_i))\Big|,$$

und somit

$$\underbrace{\Big\{\big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\big| > \varepsilon\Big\}}_{A_n} \subset \underbrace{\Big\{\big|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}(Y_i)\big| > \varepsilon/2\Big\}}_{B_n} \cup \underbrace{\Big\{\big|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}(Z_i)\big| > \varepsilon/2\Big\}}_{C_n}.$$

Deshalb ist $\mathbb{P}(A_n) \leqslant \mathbb{P}(B_n) + \mathbb{P}(C_n)$, und es genügt, separat $\mathbb{P}(B_n) \to 0$ und $\mathbb{P}(C_n) \to 0$ zu zeigen. Die Y_i haben nun endliche Varianzen, da sie beschränkt sind. Wegen (1.58 a) und der vorausgesetzten paarweisen Unabhängigkeit der X_i gilt $Y_i \perp\!\!\!\perp Y_j$ für $i \neq j$, und somit auch $\mathbb{V}(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n}\mathbb{V}(Y_i)$. Es folgt mit der Chebyshev-Ungleichung und der Definition $\bar{W}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$:

$$\mathbb{P}(B_n) = \mathbb{P}(|\bar{W}_n - \mathbb{E}(\bar{W}_n)| > \varepsilon/2) \leqslant \frac{4}{\varepsilon^2}\mathbb{V}(\bar{W}_n) \leqslant \frac{4}{n^2\varepsilon^2}\sum_{i=1}^{n}\mathbb{V}(Y_i) \leqslant$$

$$\leqslant \frac{4}{n^2\varepsilon^2}\sum_{i=1}^{n}\mathbb{E}(Y_i^2) \leqslant \frac{4}{n^2\varepsilon^2}\sum_{i=1}^{n} i^{1/2} \leqslant \frac{4}{n^2\varepsilon^2}n^{3/2} \overset{n\to\infty}{\longrightarrow} 0.$$

Mit $\bar{U}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ finden wir mit der Chebyshev-Ungleichung (diesmal mit $f(x) = x$):

$$\mathbb{P}(C_n) = \mathbb{P}(|\bar{U}_n - \mathbb{E}(\bar{U}_n)| > \varepsilon/2) \leqslant \frac{2}{\varepsilon}\mathbb{E}(|\bar{U}_n - \mathbb{E}(\bar{U}_n)|) =$$
$$= \frac{2}{\varepsilon n}\mathbb{E}\big(|\sum_{i=1}^n (Z_i - \mathbb{E}(Z_i))|\big) \leqslant \frac{2}{n\varepsilon}\sum_{i=1}^n \mathbb{E}(|Z_i - \mathbb{E}(Z_i)|).$$

Für alle i gilt

$$\mathbb{E}(|Z_i - \mathbb{E}(Z_i)|) \leqslant 2\mathbb{E}(|Z_i|) = 2\mathbb{E}(|X_1|\mathbb{1}_{\{|X_1| \geqslant i^{1/4}\}}),$$

und der letzte Ausdruck konvergiert gegen 0, denn $\mathbb{1}_{\{|X_1| \geqslant i^{1/4}\}} \to 0$ mit $i \to \infty$ fast überall wegen $\mathbb{P}(X_1 = \infty) = 0$, und dann kann man wegen der Annahme $\mathbb{E}(|X_1|) < \infty$ den Satz von der majorisierten Konvergenz benutzen. Wir haben also $\mathbb{P}(B_n) \leqslant \frac{1}{n}\sum_{i=1}^n a_i$ für eine Folge (a_i) mit $\lim_{i\to\infty} a_i = 0$. Man kann nun, zu gegebenem $\delta > 0$, diese Folge in einen Anfang $(a_1, \ldots, a_M)$ den Rest zerlegen, und dabei M so wählen, dass $|a_i| < \delta$ für alle $i \geqslant M$. Dann ist

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n a_i \leqslant \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^M a_i + \lim_{n\to\infty}\frac{1}{n}\sum_{i=M+1}^n \delta = 0 + \delta.$$

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

Bemerkung:
Das schwache GGZ ist in vielen Fällen alles, was man braucht. Allerdings hat es einen (zumindest philosophischen) Nachteil: es zeigt nicht die in (4.1) behauptete Konvergenz der gemittelten Summen zum Erwartungswert. Der Grund ist der gleiche, der auch schon in Bemerkung d) nach (4.2) genannt wurde: sei (Y_n) eine Folge von uiv ZVen mit $\mathbb{E}(Y_1) = 0$, und setze $\bar{S}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Es ist denkbar, dass zwar für alle n die Ausnahmemenge $\{\omega \in \Omega : |\bar{S}_n(\omega)| > \varepsilon\}$ eine sehr kleine W'keit hat (und die Folge dieser W'keiten sogar mit ε gegen Null geht), dass aber für alle $\omega \in \Omega$ der zugehörige *Pfad* $(\bar{S}_n(\omega))_{n\in\mathbb{N}}$ unendlich viele „Ausflüge“ aus dem Intervall $[-1, 1]$ heraus macht, die zwar immer seltener werden (beispielsweise könnte zwischen dem k-ten und dem $k+1$-ten Ausflug eine Zeitspanne von e^k oder sogar von $k!$ oder k^{k^k} liegen), aber eben nie ganz enden. Dann würde $\bar{S}_n(\omega)$ für kein ω gegen 0 konvergieren. Das starke GGZ sagt, dass dies zumindest für uiv ZVen, deren Erwartungswert existiert, nicht passiert. Wir beweisen hier eine nur eine etwas schwächere Variante:

(4.7) Starkes Gesetz der großen Zahl mit $\mathcal{L}^4$-Bedingung

(X_n) sei eine uiv Folge von ZVen, und es gelte $\mathbb{E}(X_1^4) < \infty$. Dann gilt für $\bar{S}_n := \frac{1}{n}\sum_{i=1}^n X_i$: $\lim_{n\to\infty} \bar{S}_n = \mathbb{E}(X_1)$ $\mathbb{P}$-fast sicher, anders ausgedrückt:

$$\mathbb{P}\Big(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n X_i = \mathbb{E}(X_1)\Big) = 1.$$

Beweis: Zunächst reduzieren wir unsere Bemühungen auf den Fall $\mathbb{E}(X_1) = 0$. Denn wenn $\mathbb{E}(X_1) \neq 0$, dann beweisen wir die Aussage zuerst für $Y_i = X_i - \mathbb{E}(X_i)$ und nutzen dann die Gleichheit $\bar{S}_n(\omega) = \frac{1}{n}\sum_{i=1}^n Y_i(\omega) + \mathbb{E}(X_1)$, gültig für alle $\omega \in \Omega$.

Sei also $\mathbb{E}(X_1) = 0$. Zunächst stellen wir fest, dass für jede Folge $(a_n)_{n\in\mathbb{N}}$ gilt:

$$\limsup_{n\to\infty} |a_n| n^{1/8} \leqslant 1 \qquad \Longrightarrow \qquad \lim_{n\to\infty} a_n = 0$$

Wenn wir also zeigen können, dass

$$\mathbb{P}(\limsup_{n\to\infty} |\bar{S}_n| n^{1/8} > 1) = 0 \qquad (*)$$

gilt, dann folgt

$$1 = \mathbb{P}(\limsup_{n\to\infty} |\bar{S}_n| n^{1/8} \leqslant 1) \leqslant \mathbb{P}(\lim_{n\to\infty} \bar{S}_n = 0) \leqslant 1,$$

und daher $\mathbb{P}(\lim_{n\to\infty} \bar{S}_n = 0) = 1$.

Um $(*)$ zu zeigen, erinnern wir uns zunächst daran (siehe (1.64)), dass gilt:

$$\{\omega \in \Omega : \limsup_{n\to\infty} |\bar{S}_n(\omega)| n^{1/8} > 1\} = \limsup_{n\to\infty} \{\omega \in \Omega : |\bar{S}_n(\omega)| n^{1/8} > 1\} = \limsup_{n\to\infty} A_n =: A$$

mit

$$A_n = \Big\{\omega \in \Omega : \big|\bar{S}_n(\omega)\big| > n^{-1/8}\Big\}.$$

Um $\mathbb{P}(A) = 0$ zu zeigen, benutzen wir das Lemma von Borel-Cantelli: nach diesem reicht es, $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ zu zeigen. Wir schätzen mit Hilfe der Chebyshev-Ungleichung für die Funktion $f(x) = x^4$ ab:

$$\mathbb{P}(A_n) = \mathbb{P}\Big(\big|\frac{1}{n}\sum_{i=1}^{n} X_i\big| > n^{-\frac{1}{8}}\Big) \leqslant \frac{1}{(n^{-\frac{1}{8}})^4}\mathbb{E}\Big((\frac{1}{n}\sum_{i=1}^{n} X_i)^4\Big) = n^{\frac{1}{2}-4}\mathbb{E}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} X_iX_jX_kX_l\Big).$$

Die Summen kann man aus dem Erwartungswert ziehen, und wegen der Unabhängigkeit und $\mathbb{E}(X_i) = 0$ sind alle Terme gleich 0, außer denjenigen, bei denen entweder alle Indizes gleich, oder je zwei Indizes gleich sind - bei allen anderen kann man ein $\mathbb{E}(X_i)$ abspalten, das den Term dann zerstört. Es gibt 3 Möglichkeiten, je 2 Indizes i_1, i_2, i_3, i_4 gleich zu wählen: $i_1 = i_2$ und $i_3 = i_4$ oder $i_1 = i_3$ und $i_2 = i_4$ oder $i_1 = i_4$ und $i_2 = i_3$. Für das Pärchen mit i_1 hat man dann n Indizes zur Auswahl, für das andere $n-1$. Daher ist

$$\mathbb{E}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} X_iX_jX_kX_l\Big) = \sum_{i=1}^{n}\mathbb{E}(X_i^4) + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}(X_i^2)\mathbb{E}(X_j^2) = n\mathbb{E}(X_1^4) + 3n(n-1)\mathbb{E}(X_1^2)^2.$$

Insgesamt ergibt sich

$$\mathbb{P}(A_n) \leqslant n^{-5/2}\mathbb{E}(X_1^4) + n^{-3/2}\mathbb{E}(X^2)^2$$

und die rechte Seite ist summierbar in n. Damit ist die Behauptung bewiesen. □

Bemerkung: Die Aussage gilt unverändert, wenn man die Voraussetzung $\mathbb{E}(X_1^4) < \infty$ durch die schwächere Voraussetzung $\mathbb{E}(|X_1|) < \infty$ ersetzt. Der Beweis ist allerdings deutlich schwieriger und wird erst in der Veranstaltung „Probability Theory“ im kommenden Semester gemacht.

(4.8) Vorüberlegungen zum zentralen Grenzwertsatz

a) In Beispiel (1.66 b) haben wir gesehen, dass für die einfache Irrfahrt (S_n) gilt:

$$\mathbb{P}(S_n \geqslant c) \leqslant \mathrm{e}^{-\frac{c^2}{2n}}.$$

Wegen den Symmetrie der einfachen Irrfahrt folgt sofort

$$\mathbb{P}(|S_n| \geqslant c) \leqslant \mathbb{P}(S_n \geqslant c) + \mathbb{P}(S_n \leqslant -c) \leqslant 2\,\mathrm{e}^{-\frac{c^2}{2n}}.$$

Setzt man $c = n^{1-\alpha}$ für $0 \leqslant \alpha \leqslant 1$, und schreibt $S_n = \sum_{i=1}^n X_i$ mit (X_i) uiv und $\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = 1) = 1/2$, dann ist $\mathbb{E}(X_1) = 0$, und die obige Ungleichung lässt sich schreiben als

$$\mathbb{P}\Big(\big|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}(X_1)\big| \geqslant n^{-\alpha}\Big) \leqslant \mathrm{e}^{\frac{-n^{1-2\alpha}}{2}}.$$

Vergleicht man dies mit der Aussage aus dem schwachen GGZ (4.6) so sehen wir, dass wir hier (im Spezialfall) eine stark verbesserte Variante des GGZ vorliegen haben, falls $\alpha < 1/2$ ist: in diesem Fall wissen wir nicht nur, dass für festes ε die W'keit dafür, dass $|\frac{1}{n}\sum_{i=1}^n X_i| < \varepsilon$ ist, gegen 0 konvergiert, sondern wir können sogar eine (sehr schnell schrumpfende) obere Schranke angeben, und gleichzeitig das feste ε durch eine Folge von Schranken ersetzen, die selbst gegen 0 geht. Eine äquivalente, aber für das Folgende etwas bessere Formulierung ist

$$\mathbb{P}\Big(\big|\frac{1}{n^{1-\alpha}}\sum_{i=1}^n (X_i - \mathbb{E}(X_i))\big| \geqslant 1\Big) \leqslant \mathrm{e}^{\frac{-n^{1-2\alpha}}{2}}.$$

b) Die Überlegung aus a) geht nur gut, solange $\alpha < 1/2$ ist. Dies ist leicht zu erklären: denn für $\alpha \geqslant 1/2$ ist

$$\mathbb{V}\Big(\frac{1}{n^{1-\alpha}}\sum_{i=1}^n (X_i - \mathbb{E}(X_i))\Big) = n^{-2+2\alpha}\sum_{i=1}^n \mathbb{V}(X_i) = n^{2\alpha-1}\mathbb{V}(X_1) \overset{n\to\infty}{\longrightarrow} \infty.$$

Diese Rechnung gilt unabhängig von der Verteilung der (X_i) für alle uiv Folgen von $\mathcal{L}^2$-ZVen und zeigt, dass die mit $1/n^{1-\alpha}$ gewichtete Summe für $\alpha < 1/2$ stochastisch gegen 0 geht (das sagt die Chebyshev-Ungleichung), für $\alpha > 1/2$ aber eine divergierende Varianz hat. Zwar gibt es ZVen, die eine divergierende Varianz haben und trotzdem stochastisch gegen 0 gehen (siehe Bemerkung a) zu (4.2), doch sind diese eher die Ausnahme, und für Summen von uiv ZVen ist dies nicht der Fall wie wir noch sehen werden.

c) Der Fall $\alpha = 1/2$ ist interessant, da hier die Varianz der gewichteten Summe endlich bleibt. Wir werden sehen, dass diese hier tatsächlich konvergiert, allerdings brauchen wir dazu noch einen dritten Konvergenzbegriff.

(4.9) Konvergenz in Verteilung

a) Eine Folge (X_n) von $\mathbb{R}^d$-wertigen von ZVen **konvergiert in Verteilung** gegen eine $\mathbb{R}^d$-wertige ZV X, wenn für jede **stetige, beschränkte** Funktion $f : \mathbb{R}^d \to \mathbb{R}$ gilt:

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

b) Eine Folge von W-Maßen (μ_n) **konvergiert schwach** gegen ein W-Maß μ, wenn für jede **stetige, beschränkte** Funktion $f : \mathbb{R}^d \to \mathbb{R}$ gilt:

$$\lim_{n\to\infty} \int f \mathrm{d}\mu_n = \int f \,\mathrm{d}\mu$$

c) (X_n) konvergiert also in Verteilung gegen X genau dann, wenn $\mathbb{P}_{X_n}$ schwach gegen $\mathbb{P}_X$ konvergiert.

Bemerkungen:

a) Die obige Definition macht deutlich, dass die Konvergenz in Verteilung eine Konvergenz von Folgen von Maßen ist. Ähnlich wie bei der stochastischen, aber im Gegensatz zur fast sicheren Konvergenz, müssen die (X_n) nicht einmal auf dem selben W-Raum definiert sein. Man kann die Definition ohne Änderung auch für einen allgemeineren Zielraum als $\mathbb{R}^d$ machen, man braucht lediglich einen Begriff dafür, was dann eine stetige Funktion ist, also eine Topologie auf dem Zielraum der ZVen (X_n).

b) Die Konvergenz in Verteilung ist ein typischer Fall des „Prinzip des Testens“: um herauszufinden, ob eine Folge von Objekten konvergiert oder zwei Objekte gleich sind, schreibt man eine gewissen Anzahl von Tests vor, die diese Objekte bestehen müssen. Beispiele sind:

(i): zwei Vektoren v und w im $\mathbb{R}^d$ sind gleich, wenn ihre Skalarprodukte mit allen Vektoren des $\mathbb{R}^d$ gleich sind. Wenn man für festes $y \in \mathbb{R}^d$ die Abbildung $f_y : v \mapsto \langle v, y\rangle \equiv v \cdot y$ als stetige lineare Abbildung von $\mathbb{R}^d$ nach $\mathbb{R}$ auffasst, dann kann man das auch so formulieren, dass die Ergebnisse $f_y(v)$ und $f_y(w)$ für alle „Test-Abbildungen“ f_y, $y \in \mathbb{R}^d$, übereinstimmen müssen. Man kann sogar zeigen, dass alle linearen Abbildungen $\mathbb{R}^d \mapsto \mathbb{R}$ die Form f_y für ein geeignetes y haben müssen.

(ii): Statt aller f_y reicht es auch, nur d Stück zu wählen, falls diese eine Basis bilden. Im Falle einer Orthogonalbasis erhält man dann als Testergebnisse die Koeffizienten der Vektoren.

(iii): Ebenso konvergiert eine Folge (v_n) von Vektoren im $\mathbb{R}^d$ gegen v genau dann, wenn alle für alle y (oder: alle y aus einer ONB) die Testergebnisse $(f_y(v_n))_{n\in\mathbb{N}}$ gegen $f_y(v)$ konvergieren.

(iv): zwei Maße μ und ν auf einer σ-Algebra $\mathcal{F}$ sind gleich, wenn $\mu(A) = \nu(A)$ für entweder alle $A \in \mathcal{F}$ gilt (dann ist das im Wesentlichen die Definition dafür, dass zwei Funktionen $\mu, \nu : \mathcal{F} \to [0, 1]$ gleich sind); man kann aber auch nur A aus einem durchschnitts-stabilen Erzeuger von $\mathcal{F}$ wählen, also auch hier die Anzahl der „Testfälle“ erheblich reduzieren.

c) Anders als bei Vektoren im $\mathbb{R}^d$ ist aber bei Folgen von W-Maßen die Auswahl der Menge der Tester sehr wichtig. Man kann leicht Beispiele von W-Maßen μ_n konstruieren, die schwach gegen μ konvergieren, wo es aber $A \in \mathcal{F}$ gibt, so dass für die unstetige Funktion $\mathbb{1}_A$ die getesteten Größen $\mathbb{P}_{\mu_n}(A) = \int \mathbb{1}_A \,\mathrm{d}\mu_n$ nicht konvergieren (Übung!). Dies bedeutet insbesondere, dass schwach konvergente Maße nicht „punktweise“ konvergent sein müssen. Eine etwas schwächere Aussage in einem für uns interessanten Spezialfall werden wir aber als nächstes zeigen.

d) Ebenso kann man in der Definition auf die Beschränktheit nicht verzichten: erstens gibt es (sehr viele, auch praxisrelevante) Beispiele von Folgen (X_n), die zwar in Verteilung gegen eine beschränkte ZV X konvergieren, wo aber etwa $\mathbb{E}(X_n^2) \overset{n\to\infty}{\longrightarrow} \infty$ oder sogar $\mathbb{E}(|X_n|) \overset{n\to\infty}{\longrightarrow} \infty$. Übung: man finde Beispiele für solche Folgen und erläutere, mit welchen Funktionen f hier im Sinne von (4.9 a) getestet wurde.

e) Allerdings ist es möglich, die Klasse der Tester noch weiter einzuschränken - in vielen Fällen

muss man nicht mit allen stetigen beschränkten Funkionen testen, um die Konvergenz in Verteilung zu beweisen. Dies wird im nächsten Semester genauer untersucht - Stichwort Lévy'scher Steteigketissatz. Schon jetzt brauchen wir aber eine solche Aussage zur Reduktion der Tester unter stärkeren Voraussetzungen:

(4.10) Proposition

(X_n) sei eine Folge von $\mathbb{R}$-wertigen ZVen, und X eine ZV, deren Bildmaß eine Lebesgue-Dichte besitzt. In diesem Fall sind äquivalent:

(i): X_n konvergiert gegen X in Verteilung.

(ii): $\lim_{n\to\infty} \mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$ für alle beschränkten C_b^∞-Funktionen, d.h. für alle Funktionen, für die beliebig viele Ableitungen existieren und beschränkt sind.

(iii): Für alle $c \in \mathbb{R}$ ist $\lim_{n\to\infty} \mathbb{P}(X_n \leqslant c) = \mathbb{P}(X \leqslant c)$

Beweis: (i) $\Rightarrow$ (ii) ist klar, da alle C_b^∞-Funktionen insbesondere stetig sind.

Für **(ii) $\Rightarrow$ (iii)** brauchen wir eine C_b^∞-Funktion f mit den Eigenschaften $f(x) = 1$ für $x \leqslant 0$, $f(x) = 0$ für $x \geqslant 1$, und f ist monoton fallend auf $[0,1]$. Wir werden am Schluss des Beweises klären, dass es so eine Funktion gibt. Angenommen, wir haben sie bereits, so definieren wir zu $c \in \mathbb{R}$ und $\delta > 0$ die Funktionen

$$f_\delta^+(x) := f(\tfrac{x-c}{\delta}), \qquad f_\delta^-(x) = f(\tfrac{x-c+\delta}{\delta}).$$

Dann ist $f_\delta^+(x) = 1$ für $x \leqslant c$, $f_\delta^+(x) = 0$ für $x \geqslant c+\delta$, $f_\delta^-(x) = 1$ für $x \leqslant c-\delta$ und $f_\delta^+(x) = 0$ für $x \geqslant c$, und beide sind C_b^∞.

Für $\delta > 0$ gilt daher nach Annahme (ii) dass

$$\lim_{n\to\infty} \mathbb{E}(f_\delta^\pm(X_n)) = \mathbb{E}(f_\delta^\pm(X)).$$

Wegen $f_\delta^-(x) \leqslant \mathbb{1}_{(-\infty,c]}(x) \leqslant f_\delta^+(x)$ ist somit

$$\begin{aligned}\mathbb{E}(f_\delta^-(X)) = \lim_{n\to\infty} \mathbb{E}(f_\delta^-(X_n)) &\leqslant \liminf_{n\to\infty} \mathbb{P}(X_n \leqslant c) \leqslant \\ &\leqslant \limsup_{n\to\infty} \mathbb{P}(X_n \leqslant c) \leqslant \lim_{n\to\infty} \mathbb{E}(f_\delta^+(X_n)) = \mathbb{E}(f_\delta^+(X)). \qquad (*)\end{aligned}$$

Andererseits gilt wegen $\mathbb{P}_X(\mathrm{d}x) = \rho(x)\,\mathrm{d}x$:

$$|\mathbb{E}(f_\delta^\pm(X)) - \mathbb{P}(X \leqslant c)| \leqslant \int |f_\delta^\pm(x) - \mathbb{1}_{(-\infty,c]}(x)|\rho(x)\,\mathrm{d}x \leqslant \int \mathbb{1}_{[c-\delta,c+\delta]}(x)\rho(x)\,\mathrm{d}x \xrightarrow{\delta\to 0} 0$$

nach dem Satz von der dominierten Konvergenz. Nehmen wir also den Grenzwert $\delta \to 0$ in $(*)$, so erhalten wir (iii).

Für **(iii) $\Rightarrow$ (i)** bemerken wir zuerst, dass aus (iii) direkt $\lim_{n\to\infty} \mathbb{P}(X_n \in (c,d]) = \mathbb{P}(X \in (c,d])$ für alle $c < d$ folgt.

Sei nun $\varepsilon > 0$ beliebig. Wegen

$$0 = \mathbb{P}(|X| = \infty) = \lim_{N\to\infty} \mathbb{P}(X \notin (-N,N])$$

existiert ein Intervall $(-N,N] \subset \mathbb{R}$ so dass $\mathbb{P}(X \notin (-N,N]) < \varepsilon$. Daher ist auch $\mathbb{P}(X_n \notin (-N,N]) < 2\varepsilon$ für hinreichend große n. Für eine beschränkte Funktion f gilt somit schon

einmal

$$\limsup_{n\to\infty} \big|\mathbb{E}(f(X_n) - f(X))\big| \leqslant \limsup_{n\to\infty} \Big|\mathbb{E}\Big([f\mathbb{1}_{(-N,N]}](X_n) - [f\mathbb{1}_{(-N,N]}](X)\Big)\Big| + 3\varepsilon\|f\|_\infty.$$

Ist nun f stetig, so ist f auf $[-N,N]$ sogar gleichmäßig stetig, es existiert also ein $\delta > 0$ so dass $\sup\{|f(x) - f(y)| : x, y \in [-N,N], |x-y| \leqslant \delta\} < \varepsilon$ ist. Wir treffen nun zu dem durch ε bestimmten δ eine Auswahl von $m = m_\delta$ Punkten $x_0, \dots, x_m \in [-N,N]$, für die $0 < x_{i+1} - x_i < \delta$, $x_0 = -N$, und $x_m = N$ gilt. Mit

$$f_\delta(y) := \sum_{i=1}^{m} f(x_i)\mathbb{1}_{(x_{i-1},x_i]}(y)$$

gilt dann $\|f_\delta - f\mathbb{1}_{(-N,N]}\|_\infty < \varepsilon$, und somit

$$\Big|\mathbb{E}\Big([f\mathbb{1}_{(-N,N]}](X_n) - [f\mathbb{1}_{(-N,N]}](X)\Big)\Big| \leqslant \Big|\mathbb{E}\big(f_\delta(X_n) - f_\delta(X)\big)\Big| + 2\varepsilon.$$

Da außerdem

$$\lim_{n\to\infty} \mathbb{E}(f_\delta(X_n) - f_\delta(X)) = \lim_{n\to\infty} \sum_{i=1}^{m} f(x_i)\big(\mathbb{P}(X_n \in (x_{i-1}, x_i]) - \mathbb{P}(X \in (x_{i-1}, x_i])\big) = 0$$

gilt, haben wir

$$\limsup_{n\to\infty} \big|\mathbb{E}(f(X_n) - f(X))\big| \leqslant 3\varepsilon\|f\|_\infty + 2\varepsilon$$

gezeigt. Da $\varepsilon > 0$ beliebig war, folgt (i).

Es bleibt die Existenz der in (ii) $\Rightarrow$ (iii) behaupteten Funktion zu zeigen. Eine solche Funktion ist zum Beispiel gegeben durch

$$f(x) = \mathbb{1}_{(-\infty,1/2]} * \psi(x) := \int \mathbb{1}_{(-\infty,1/2]}(y)\psi(x-y)\,\mathrm{d}y$$

mit

$$\psi(x) := \begin{cases} \frac{1}{Z}\exp\Big(-\frac{1}{1-4x^2}\Big) & \text{falls } |x| < 1/2, \\ 0 & \text{sonst,} \end{cases} \quad \text{und } Z = \int_{-1/2}^{1/2} \mathrm{e}^{-\frac{1}{1-4x^2}}\,\mathrm{d}x.$$

Denn man kann leicht überprüfen, dass $\psi \in C^\infty$ ist. Für jedes endliche Intervall $[a,b] \subset \mathbb{R}$ gilt außerdem die Ungleichung

$$\sup\{\mathbb{1}_{(-\infty,1/2]}(y)\partial_y^k\psi(x-y) : x \in [a,b]\} \leqslant \|\partial_y^k\psi\|_\infty \mathbb{1}_{[a-1/2,b+1/2]}(y).$$

Da die letztere Funktion integrierbar ist, darf man zur Berechnung von $\partial_x^k f$ die Ableitung unter das Integral ziehen und erhält $f \in C^\infty$. Die Forderungen $f(x) = 1$ für $x \leqslant 0$ und $f(x) = 0$ für $x \geqslant 1$ sind auch erfüllt. □

(4.11) Der zentrale Grenzwertsatz

(X_n) sei eine Folge von uiv ZVen, und es gelte $\mathbb{E}(X_1) = m \in \mathbb{R}$, $\mathbb{V}(X_1) = \sigma^2 < \infty$. Setze

$$\bar{S}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m).$$

Dann konvergiert $(\bar{S}_n)$ in Verteilung gegen eine $\mathcal{N}_{0,\sigma^2}$-verteilte ZV. Außerdem gilt für alle $c \in \mathbb{R}$:

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m) \leqslant c\Big) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{c} \mathrm{e}^{-\frac{x^2}{2\sigma^2}}\, \mathrm{d}x.$$

Beweis:
Da die Normalverteilung eine Dichte hat, genügt es, die Aussage (4.10 (ii)) zu beweisen. Außerdem genügt es, den Fall $m = 0$ und $\sigma^2 = 1$ zu behandeln, denn für allgemeine X_i hat $\tilde{X}_i = \frac{1}{\sigma}(X_i - m)$ diese Eigenschaften, und wenn die Aussage für $m = 0$ und $\sigma^2 = 1$ gilt, dann bekommen wir

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m) \leqslant c\Big) = \mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{X}_i \leqslant \frac{c}{\sigma}\Big) \overset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c/\sigma} \mathrm{e}^{-\frac{y^2}{2}}\, \mathrm{d}y = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{c} \mathrm{e}^{-\frac{x^2}{2\sigma^2}}\, \mathrm{d}x.$$

Sei nun $(Y_i)_{i\in\mathbb{N}}$ eine Folge von uiv $\mathcal{N}_{0,1}$-verteilten ZVen, und setze

$$\bar{T}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

Wegen (3.12 a) ist $\bar{T}_n \sim \mathcal{N}_{0,1}$, und daher reicht es, die Aussage

$$\lim_{n\to\infty} \mathbb{E}\Big(f(\bar{S}_n) - f(\bar{T}_n)\Big) = 0 \qquad (*)$$

für alle beschränkten $f \in C^3$ zu zeigen. Hierzu benutzen wir eine trickreiche Teleskopsumme, es ist nämlich

$$f(\bar{S}_n(\omega)) - f(\bar{T}_n(\omega)) = \sum_{i=1}^{n} f\Big(W_{i,n}(\omega) + \frac{1}{\sqrt{n}} X_i(\omega)\Big) - f\Big(W_{i,n}(\omega) + \frac{1}{\sqrt{n}} Y_i(\omega)\Big), \qquad (**)$$

wobei

$$W_{i,n}(\omega) := \frac{1}{\sqrt{n}} \sum_{j=1}^{i-1} X_i(\omega) + \frac{1}{\sqrt{n}} \sum_{j=i+1}^{n} Y_j(\omega)$$

die Summe ist, in der man das i-te Glied auslässt und danach statt der $X_i(\omega)$ die $Y_i(\omega)$ summiert. Der Vorteil dieser Darstellung ist, dass man nun versuchen kann, den Erwartungswert jeder der in $(**)$ auftretenden Differenzen viel kleiner als $\frac{1}{n}$ zu machen, woraus dann $(*)$ folgt. Das erste Hilfsmittel hierzu ist der Satz von Taylor: danach existiert ein $\vartheta_0 \in [0,1]$ mit

$$f(x+y) = f(x) + yf'(x) + \frac{y^2}{2} f''(x) - \frac{y^2}{2}(f''(x) - f''(x + \vartheta_0 y)).$$

Hierbei kann $\vartheta_0 = \vartheta_0(x,y)$ so gewählt werden, dass es stetig von x und y abhängt. Definiert man nun die messbaren, $[0,1]$-wertigen Abbildungen $\vartheta(\omega) := \vartheta_0(W_{i,n}(\omega), \frac{1}{\sqrt{n}} Y_i(\omega))$ und $\tilde{\vartheta}(\omega) :=$

$\vartheta_0(W_{i,n}(\omega), \frac{1}{\sqrt{n}}X_i(\omega))$

$$\mathbb{E}\Big(f\big(W_{i,n}+\frac{1}{\sqrt{n}}X_i\big)-f\big(W_{i,n}+\frac{1}{\sqrt{n}}Y_i\big)\Big)=\frac{1}{\sqrt{n}}\mathbb{E}\big((X_i-Y_i)f'(W_{i,n})\big)+\frac{1}{2n}\mathbb{E}\big((X_i^2-Y_i^2)f''(W_{i,n})\big)+$$
$$+\frac{1}{2n}\mathbb{E}\Big(Y_i^2\big(f''(W_{i,n})-f''(W_{i,n}+\tfrac{\vartheta}{\sqrt{n}}Y_i)\big)-X_i^2\big(f''(W_{i,n})-f''(W_{i,n}+\tfrac{\tilde{\vartheta}}{\sqrt{n}}X_i)\big)\Big).$$

Da $X_i \perp\!\!\!\perp W_{i,n}$ und $Y_i \perp\!\!\!\perp W_{i,n}$ für alle i gilt, faktorisieren die Erwartungswerte auf der rechten Seite der ersten Zeile, und da $\mathbb{E}(X_i)=\mathbb{E}(Y_i)=0$ und $\mathbb{E}(X_i^2)=\mathbb{E}(Y_i^2)=1$ gilt, verschwinden sie sogar.
Für den Term in der zweiten Zeile bemerken wir zunächst, dass f'' Lipschitz-stetig ist (mit Lipschitzkonstante $\|f'''\|_\infty$). Für $C>0$ und diejenigen ω, für die $|Y_i|(\omega)\leqslant C$ ist, gilt also

$$R_{i,n,Y_i}(\omega):=\Big|Y_i^2(\omega)\big(f''(W_{i,n}(\omega))-f''(W_{i,n}(\omega)+\tfrac{\vartheta(\omega)}{\sqrt{n}}Y_i(\omega))\big)\Big|\leqslant\tfrac{C^3}{\sqrt{n}}\|f'''\|_\infty.$$

Für die übrigen ω haben wir immerhin noch die Abschätzung $R_{i,n}(\omega)\leqslant 2Y_i^2(\omega)\|f''\|_\infty$. Zusammen erhalten wir

$$\mathbb{E}(R_{i,n,Y_i})=\mathbb{E}\Big(R_{i,n,Y_i}\big(\mathbb{1}_{\{|Y_i|\leqslant C\}}+\mathbb{1}_{\{|Y_i|>C\}}\big)\Big)\leqslant\tfrac{C^3}{\sqrt{n}}\|f'''\|_\infty+2\mathbb{E}(Y_i^2\mathbb{1}_{\{|Y_i|>C\}})\|f''\|_\infty.$$

Die gleiche Rechnung gilt, wenn man statt Y_i nun X_i nimmt, und insgesamt erhalten wir

$$\begin{aligned}|\mathbb{E}(f(\bar{S}_n)-f(\bar{T}_n))|&\leqslant\frac{1}{2n}\sum_{i=1}^n(\mathbb{E}(R_{i,n,X_i})+\mathbb{E}(R_{i,n,Y_i}))\leqslant\\&\leqslant\frac{1}{n}\sum_{i=1}^n\big(\tfrac{C^3}{\sqrt{n}}\|f'''\|_\infty+\mathbb{E}(X_i^2\mathbb{1}_{\{|X_i|>C\}})\|f''\|_\infty+\mathbb{E}(Y_i^2\mathbb{1}_{\{|Y_i|>C\}})\|f''\|_\infty\big)=\\&=\tfrac{C^3}{\sqrt{n}}\|f'''\|_\infty+\mathbb{E}(X_1^2\mathbb{1}_{\{|X_1|>C\}})\|f''\|_\infty+\mathbb{E}(Y_1^2\mathbb{1}_{\{|Y_1|>C\}})\|f''\|_\infty.\end{aligned}$$

Der erste Term konvergiert für jedes $C>0$ mit $n\to\infty$ gegen 0. Die beiden hinteren Terme konvergieren mit $C\to\infty$ gegen 0, da $\mathbb{E}(X_1^2)=\mathbb{E}(Y_1^2)=1<\infty$, und können somit durch entsprechende Wahl von C unter jedes $\varepsilon>0$ gedrückt werden. Daher muss $\lim_{n\to\infty}\mathbb{E}(f(\bar{S}_n)-f(\bar{T}_n))=0$ gelten. □

Bemerkungen:
a) Sowohl die Zerlegungstechnik (also $\mathbb{E}(X_i)=\mathbb{E}(X_i\mathbb{1}_{\{X_i\leqslant C\}})+\mathbb{E}(X_i\mathbb{1}_{\{X_i>C\}})$), als auch die Ausnutzung der Gleichverteilung der (X_i) beim der Abschätzung der Terms $\mathbb{E}(X_i\mathbb{1}_{\{X_i>C\}})$ hatten wir bereits im Beweis von Satz (4.6) gesehen. Diese Techniken sind oft nützlich.

b) Der zentrale Grenzwertsatz erlaubt zusammen mit dem 0-1-Gesetz von Kolmogorov (1.63) eine interessante Aussage: (X_i) sei eine uiv Folge von ZVen mit $\sigma^2=\mathbb{V}(X_1)<\infty$. Wegen (4.11) konvergiert die Folge $\bar{S}_n:=\frac{1}{\sqrt{n}}\sum_{i=1}^n(X_i-\mathbb{E}(X_i))$ in Verteilung gegen eine $\mathcal{N}_{0,1}$-verteilte ZV. Wegen (1.63) und der Tatsache, dass $\omega\mapsto\limsup_{n\to\infty}\bar{S}_n(\omega)$ messbar bezüglich der terminalen σ-Algebra ist, erhalten wir dass

$$\mathbb{P}\Big(\limsup_{n\to\infty}\bar{S}_n>C\Big)\in\{0,1\}\qquad\forall C>0.$$

Mit der Gleichheit

$$\{\omega \in \Omega : \limsup_{n\to\infty} \bar{S}_n(\omega) > C\} = \limsup_{n\to\infty} \{\omega \in \Omega : \bar{S}_n(\omega) > C\}$$

erhalten wir

$$\mathbb{P}(\limsup_{n\to\infty} \bar{S}_n > C) = \mathbb{E}(\mathbb{1}_{\limsup_{n\to\infty}\{\bar{S}_n > C\}}) = \mathbb{E}(1 - \liminf_{n\to\infty} \mathbb{1}_{\{\bar{S}_n \leqslant C\}}) = (*).$$

Das Lemma von Fatou liefert nun

$$(*) \geqslant 1 - \liminf_{n\to\infty} \mathbb{P}(\bar{S}_n \leqslant C) = 1 - \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{C} \mathrm{e}^{-\frac{x^2}{2\sigma^2}}\, \mathrm{d}x > 0.$$

Somit muss $\mathbb{P}(\limsup_{n\to\infty} \bar{S}_n > C) = 1$ für alle C gelten, und mit $C \to \infty$ und der Stetigkeit von oben des Maßes $\mathbb{P}$ schließen wir $\mathbb{P}(\limsup_{n\to\infty} \bar{S}_n = \infty) = 1$. Das gleiche Argument zeigt dass $\mathbb{P}(\liminf_{n\to\infty} \bar{S}_n = -\infty) = 1$.

Es stellt sich also heraus, dass die $\bar{S}_n$ zwar in Verteilung gegen die Normalverteilung konvergieren, dass aber die *Pfade* $(\bar{S}_n(\omega))_{n\in\mathbb{N}}$ sich höchst irregulär verhalten und mit W'keit 1 unendlich oft zwischen immer größeren und immer kleineren Werten hin und her pendeln. Im Kontext der einfachen Irrfahrt wissen wir also jetzt folgendes: ist (S_n) die einfache Irrfahrt, $\alpha > 0$, dann gilt jeweils $\mathbb{P}$-fast sicher:

$$\limsup_{n\to\infty} \frac{1}{n^\alpha} S_n = -\liminf_{n\to\infty} \frac{1}{n^\alpha} S_n = \begin{cases} 0 & \text{falls } \alpha > 1/2, \\ \infty & \text{falls } \alpha \leqslant 1/2. \end{cases}$$

Der erste Fall folgt aus (3.21 b) in Verbindung mit dem Lemma von Borel-Cantelli (Übung!), den zweiten haben wir gerade gezeigt. Dies ist zwar noch nicht ganz so scharf wie das in (1.66 b) zitierte Gesetz vom iterierten Logarithmus, aber immerhin schon nahe dran - auf der „algebraischen Skala“, wo man nur auf Potenzen von n schaut, haben wir das richtige α bereits gefunden!

5. Statistik

(5.1) Vergleich von Statistik und Wahrscheinlichkeitstheorie

Gemeinsamkeiten: Sowohl die Statistik als auch die W-Theorie benutzen (weitgehend) die gleichen Grundlagen, also grundlegende Definitionen wie W-Räume, Zufallsvariable, stochastische Unabhängigkeit und bedingte W'keiten, Erwartungswert und Momente, sowie grundlegende Sätze über deren Eigenschaften.

Unterschiede: Bisher haben in den Beispielen die **W-Theorie** kennengelernt. Der Ansatz ist hier der für die Mathematik typische: man definiert sich zunächst (oft motiviert von Anwendungen) ein präzises mathematisches Objekt, wie z.B. das Perkolationsmodell, die einfache Irrfahrt, den Galton-Watson Prozess, oder auch allgemeiner Summen von uiv ZVen oder Markovketten. Dann versucht man, so viel wie möglich über dieses Objekt herauszufinden, interessante Aussagen zu treffen und zu beweisen, das Objekt zu verallgemeinern oder auch weiter zu spezialisieren, um es in den Spezialfällen noch besser zu verstehen etc. Ausgangspunkt ist jedoch immer ein konkretes Modell, das untersucht wird.

Der Ansatz der **Statistik** kommt in gewisser Weise genau von der entgegengesetzten Seite: hier

geht man davon aus, dass man **Daten** gegeben hat, und dies sind zunächst einfach eine (meist sehr große) Menge an reellen Zahlen. Von diesen Daten nimmt man nun an, dass sie durch einen Mechanismus entstanden sind, der dem Zufall unterliegt, was immer man darunter verstehen möchte. Ab hier gibt es nun drei (teilweise überlappende) Bereiche:

In der **beschreibenden Statistik** versucht man, die Daten so aufzubereiten, dass man eine Übersicht über wichtige Eigenschaften bekommt, die in ihnen versteckt sein könnten. Dies kann durch Visualisierung (z.B. mittels Histogramm und Boxplot) geschehen, oder durch Berechnung von ad-hoc Kenngrößen wie empirischem Mittelwert oder empirischer Standardabweichung. Diese Erkenntnisse können dabei helfen, Modelle aufzustellen, die erklären sollen, wie die Daten entstanden sind.

In der **Regression und Faktoranalyse** arbeitet man quantitativer: in der klassischen Regressionstheorie hat man es beispielsweise mit Datenpaaren $(x_i, y_i)_{i=1,\dots,n}$ zu tun, von denen man annimmt, dass es eine Funktion f gibt, so dass $y_i = f(x_i) + \xi_i$ gilt, wobei die zufällige Größe ξ_i beispielsweise für einen Messfehler stehen kann. Man versucht dann, mittels verschiedener Ansätze die Funktion f möglichst gut zu bestimmen und für gegebenes f auch eine Maßzahl für die Qualität der Approximation der Daten (x_i, y_i) durch die Paare $(x_i, f(x_i))$ zu finden. In der Faktoranalyse geht es darum, in vielen, möglicherweise hochdimensionalen, durch zufällige Einflüsse verrauschten Daten Gesetzmäßigkeiten zu erkennen, beispielsweise dass sie sich alle nahe einer (explizit gegebenen) Hyperfläche befinden. Auch hier möchte man die Güte der Approximation der Daten durch die Hyperfläche quantifizieren.

In der Schätz- und Testtheorie (auch **schließende Statistik**) genannt, geht es darum, aus einer Menge von möglichen Modellen, aus denen die Daten entstanden sein könnten, das plausibelste auszuwählen. Die Menge der Modelle, die in Frage kommen, schränkt man oft so weit ein, dass man sie durch einen oder mehrere Parameter beschreiben kann - dies ist dann die parametrische Statistik. Beispielsweise nimmt man oft an, dass die Daten durch n-maliges unabhängiges Würfeln einer $\mathcal{N}_{m,\sigma^2}$-verteilten Zufallsvariable entstanden sind; in der **Schätztheorie** sucht man nach Methoden, um die Parameter μ und σ^2 möglichst gut zu bestimmen, und interessiert sich dafür, wie schnell die Schätzungen besser werden, wenn man mehr Daten (größeres n) zur Verfügung hat. In der **Testtheorie** dagegen will man meist wissen, welche Parameter man aufgrund der Datenlage mit hoher Sicherheit ausschließen kann. Die mathematische Begriffsbildung zu „mit hoher Sicherheit“ und „aufgrund der Datenlage“ ist ein wenig subtil und wird gemacht, wenn wir das Thema Testtheorie behandeln.

(5.2) Nomenklatur

Unter dem Begriff **Daten** verstehen wir ab jetzt immer eine endliche Menge $\{x_1, \dots, x_n\}$ oder ein n-Tupel $(x_1, \dots, x_n)$ von reellen Zahlen oder von Elementen im $\mathbb{R}^d$.

(5.3) Histogramme: Definitionen

a) Sei $(c,d] \subset \mathbb{R}$ ein halboffenes Intervall und $N \in \mathbb{N}$. Eine **Intervallpartition** $\mathcal{P}$ von $(c,d]$ ist eine Menge $(A_i)_{1 \leqslant i \leqslant N}$ von Intervallen mit $A_i = (a_i, a_{i+1}]$, wobei $a_1 = c$ und $a_{N+1} = d$. Für $A = (a,b] \in \mathcal{P}$ schreiben wir $|A| = b-a$ für die Länge von A. Die Größe $|\mathcal{P}| := \max\{|A| : A \in \mathcal{P}\}$ heißt die **Feinheit** der Partition.

b) Gegeben seien reelle Daten $\boldsymbol{x} := (x_1, \ldots, x_n)$ und eine Intervallpartition $\mathcal{P}$ von eines Intervalls $(c,d]$ so dass $x_i \in (c,d]$ für alle i. Für $A \in \mathcal{P}$ schreiben wir

$$N_A(\boldsymbol{x}) := \sum_{k=1}^{n} \mathbb{1}_A(x_k)$$

für die Anzahl der Datenpunkte, die in A liegen. Das zu $\mathcal{P}$ gehörige **Histogramm** der Datenpunkte (x_i) ist die Funktion

$$H_{\mathcal{P},\boldsymbol{x}} : (c,d] \to \mathbb{R}, \qquad y \mapsto \frac{1}{n} \sum_{A \in \mathcal{P}} \frac{N_A(\boldsymbol{x})}{|A|} \mathbb{1}_A(y).$$

Bemerkungen:
a) $H_{\mathcal{P},\boldsymbol{x}}$ ist also eine Stufenfunktion, deren Stufen den Mengen in $\mathcal{P}$ entsprechen, während die Höhe der zu A gehörigen Stufe dadurch bestimmt wird, welcher Bruchteil aller Datenpunkte in A landet, **und** wie breit A ist. Diese zweite Normierung führt dazu, dass die *Fläche* (und nicht die Höhe) der zu A gehörigen Stufe anzeigt, wie viele Datenpunkte sie enthält. Dies ist zunächst dann nützlich, wenn die Stufen nicht alle gleich breit sind: sonst würde bei gleichmäßiger Verteilung der Datenpunkte eine Stufe, die doppelt so breit ist wie eine andere, die doppelte Höhe haben, was die optische Illusion erzeugt, dass in diesem Bereich „mehr Daten zu finden“ sind. Tatsächlich gibt es auch die Variante, wo man nicht mit $|A|$ normiert, diese nennt man **Säulendiagramm**. Für uns spielt diese Variante keine Rolle.

b) Das Histogramm ist ein gutes Beispiel für *absichtlichen Informationsverlust*: Man vergisst sowohl die ursprüngliche Anordnung der Daten (falls diese jemals eine natürlich Ordnung hatten) als auch die genaue Lage - letztere insofern, dass man nur noch zählt, wie viele Datenpunkte in jeder Menge liegen. Eine sinnvolle Reduktion der zur Verfügung stehenden Information ist oft entscheidend dafür, den Daten überhaupt etwas ansehen zu können. Hier sollte man das Wort „sinnvoll“ jedoch durchaus ernst nehmen!

(5.4) Histogramme und Grenzwerte

a) Vorbemerkung: Sei $\mathcal{P}$ eine feste Partition eines Intervalls. Wir wollen jetzt annehmen, dass unsere Daten $x_1, \ldots, x_n$ von uiv ZVen erzeugt wurden. Oft sagt man auch, $x_1, \ldots, x_n$ sollen **Realisierungen** einer Folge von uiv ZVen sein. Was aber bedeutet das eigentlich?

Zunächst könnte man versuchen, folgendes zu fordern: es gebe einen W-Raum $(\Omega, \mathcal{F}, \mathbb{P})$, unabhängige ZVen $X_1, \ldots, X_n$ und ein $\omega \in \Omega$ gibt so dass $x_i = X_i(\omega)$ für dieses ω und alle i. Das Problem dabei ist, dass man dann fast nichts aussagen kann - denn wenn beispielsweise die Daten $x_i = 0$ für alle i sind, so ist diese Forderung mit dem W-Raum $\Omega = (\mathbb{R}^n, \mathcal{B}^{\otimes n}, \mathcal{N}_{0,1}^{\otimes n})$, den ZVen $X_i(\omega) = \omega_i$ und dem Element $\omega = (0, \ldots, 0) \in \Omega = \mathbb{R}^n$ erfüllt, obwohl die Daten sicher nichts mit der Normalverteilung zu tun haben.

Was wir stattdessen tun sollten ist, die Abbildungen $\boldsymbol{x} \mapsto H_{\mathcal{P},\boldsymbol{x}}$ und $\omega \mapsto (X_1(\omega), \dots, X_n(\omega))$ hintereinander zu schalten. Dann gibt es für jedes $\mathbb{P}$, das wir auf $(\Omega, \mathcal{F})$ wählen, ein Bildmaß unter der Abbildung $\omega \mapsto H_{\mathcal{P},(X_1(\omega),\dots,X_n(\omega))}$ - im speziellen Fall des Histogramms ist dieses Bildmaß ein Maß auf Stufenfunktionen. Dieses Bildmaß und seine Eigenschaften meinen wir, wenn wir sagen, die Daten seien durch unabhängige Realisierungen von X gegeben.

b) Seien also die x_i Realisierungen einer uiv Folge $(X_i)_{i\in\mathbb{N}}$, $\boldsymbol{X}_{(n)} = (X_1, \dots, X_n)$. Dann gilt für alle $A \in \mathcal{P}$ und alle $y \in A$:

$$\lim_{n\to\infty} H_{\mathcal{P},\boldsymbol{X}_{(n)}}(y) = \frac{1}{|A|}\mathbb{P}(X_1 \in A) \quad \mathbb{P}\text{-fast sicher}$$

Denn nach Definition ist

$$H_{\mathcal{P},\boldsymbol{X}_{(n)}(\omega)}(y) = \frac{1}{n}\frac{1}{|A|}\sum_{k=1}^{n} \mathbb{1}_A(X_k(\omega)),$$

die ZVen $\omega \mapsto \mathbb{1}_A(X_k(\omega))$ sind uiv und haben beschränkte Momente. Daher folgt die Behauptung aus dem starken GGZ. Schreibt man dies etwas anders, so erhält man

$$\lim_{n\to\infty} H_{\mathcal{P},\boldsymbol{X}_{(n)}}(y) = \frac{1}{|A|}\int \mathbb{1}_A(x)\mathbb{P}_X(\mathrm{d}x).$$

c) Nehmen wir nun an, dass $\mathbb{P}_X$ eine stetige Lebesgue-Dichte ρ besitzt. Man kann dann relativ leicht zeigen (Übung!), dass für Folgen $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ mit $a_n < b_n$ für alle n, $a_n \to c$ und $b_n \to c$ gilt:

$$\lim_{n\to\infty} \frac{1}{b_n - a_n}\int_{a_n}^{b_n} \rho(x)\,\mathrm{d}x = \rho(c).$$

Zusammen mit b) bekommen wir also

$$\lim_{|\mathcal{P}|\to 0}\lim_{n\to\infty} H_{\mathcal{P},\boldsymbol{X}_{(n)}}(y) = \rho(y) \qquad \mathbb{P}\text{-fast sicher.}$$

für alle $y \in [c, d]$, mit anderen Worten: das Histogramm approximiert die Dichte von X_1, wenn wir viele Daten und eine feine Auflösung haben. Hier sehen wir auch nochmal die Bedeutung der Normierung $1/|A|$.
Allerdings müssen wir mit der Reihenfolge der Grenzwerte aufpassen: vertauschen wir sie beispielsweise, so kommt nichts sinnvolles heraus (warum?).

Für die Praxis sind die Fälle relevant, wo die Anzahl der Daten und die Feinheit der Partition gleichzeitig divergiert bzw. gegen 0 geht. Auch hier muss man aufpassen: In jedem Intervall sollten noch genügend Datenpunkte sein, damit die in b) bewiesene Mischungseigenschaft nicht zerstört wird - das Gesetz der großen Zahl muss (wenigstens in seiner schwachen Form) noch gelten. Wenn Sie über die Chebyshev-Ungleichung nachdenken, sollte es Ihnen nicht allzu schwer fallen, das richtige Verhältnis der Grenzwerte zu erraten: Wenn für eine Folge $(\mathcal{P}_n)$ von Partitionen zwar $\lim_{n\to\infty} |\mathcal{P}_n| = 0$ gilt, aber auch das *Minimum* von $\{|A| : A \in \mathcal{P}\}$ nicht schneller verschwindet als $n^{-\alpha}$ mit $\alpha < 1/2$, dann kann man beweisen, dass

$$\lim_{n\to\infty} H_{\mathcal{P}_n,\boldsymbol{X}_n}(y) = \rho(y) \qquad \text{(im Sinne der stochastischen Konvergenz).}$$

Diese Aussage lässt sich in verschiedene Richtungen verschärfen (bessere Konvergenz, bessere Uniformität in y), was wir hier aber nicht machen wollen.

(5.5) Median und Quantile

a) Sei $0 < \alpha < 1$, X eine relle ZV. Jede Zahl q_α, für die gilt:
(i): $\mathbb{P}(X \leqslant q_\alpha) \geqslant \alpha$
(ii): $\mathbb{P}(X \geqslant q_\alpha) \geqslant 1 - \alpha$
heißt α**-Quantil** der Verteilung von X.

b) Sei $0 < \alpha < 1$, $\mathbb{P}$ ein W-Maß auf $\mathbb{R}$. Jede Zahl q_α, für die gilt:
(i): $\mathbb{P}((-\infty, q_\alpha]) \geqslant \alpha$
(ii): $\mathbb{P}([q_\alpha, \infty)) \geqslant 1 - \alpha$
heißt α**-Quantil** von $\mathbb{P}$.
Offensichtlich ist ein α-Quantil von X einfach ein α-Quantil von $\mathbb{P}_X$.

c) Seien $x_1, \ldots, x_n$ reelle Daten. Jede Zahl q_α, für die gilt

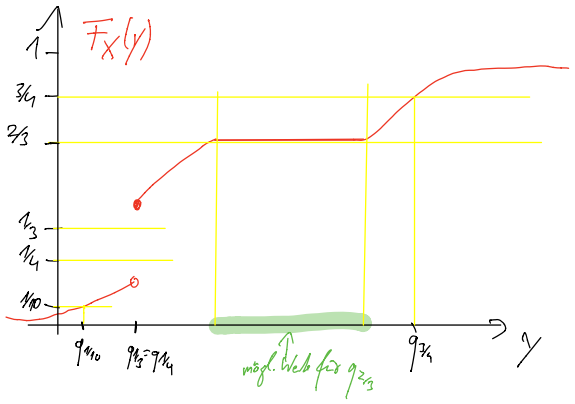
(i):Mindestens $n\alpha$ Datenwerte sind $\leqslant q_\alpha$, also $\dfrac{|\{i \leqslant n : x_i \leqslant q_\alpha\}|}{n} \geqslant \alpha$,

(ii):Mindestens $n(1-\alpha)$ Datenwerte sind $\geqslant q_\alpha$, also $\dfrac{|\{i \leqslant n : x_i \geqslant q_\alpha\}|}{n} \geqslant 1 - \alpha$;

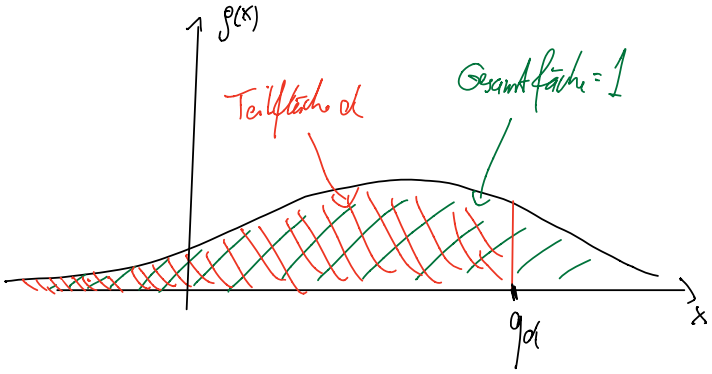
heißt α**-Quantil** der Daten $(x_1, \ldots, x_n)$.

d) in allen Fällen haben gewisse Quantile spezielle Namen:
das 1/2-Quantil heißt **Median** oder **zweites Quartil**,
das 1/4-Quantil heißt **erstes Quartil**, und
das 3/4-Quantil heißt **drittes Quartil**.
Das $(1 - \alpha)$-Quartil wird manchmal auch auch als α-Fraktil bezeichnet, wir machen diese Inflation der Nomenklatur aber nicht mit.
Bemerkungen:
a) Die Definition in (3.6 c) ist in Wirklichkeit ein Spezialfall der Definitionen in a) und b). Legt man nämlich auf die Datenwerte $x_1, \ldots, x_n$ die Gleichverteilung, betrachtet also das W-Maß $\mathbb{P}$ auf $\mathbb{R}$ mit $\mathbb{P}(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_A(x_i)$ für alle $A \in \mathcal{F}$. Dann ist $\mathbb{P}((-\infty, q]) = \frac{1}{n}|\{i \leqslant n : x_i \leqslant q\}$ etc.

b) Erinnern wir uns an die Definition der Verteilungsfunktion und an (2.17), so finden wir, dass jede Zahl q_a mit $F_X(q_a) \geqslant a$ und $F_X(q_a^-) \leqslant \alpha$ ein α-Quantil ist. Die nebenstehende Graphik verdeutlicht beispielhaft den Zusammenhang zwischen Quantilen und Verteilungsfunktionen. Im Beweis zu (2.22) wurde übrigens ebenfalls bereits ein Quantil verwendet, dort wurde es aber (per Konvention) eindeutig gemacht.



c) Falls X eine Dichte hat, dann ist jede Lösung q_α der Gleichung $\int_{-\infty}^{q_\alpha} \rho(y)\,\mathrm{d}y = \alpha$ ein α-Quantil. Falls $\{x \in \mathbb{R} : \rho(x) > 0\}$ ein (endliches oder unendliches) Intervall ist, dann ist diese Lösung eindeutig.



(5.6) Algorithmus zur Berechnung von Quantilen für gegebene Daten

Gegeben seien reelle Daten $x_1, \dots, x_n$, gesucht ist das α-Quantil q_α für $0 < \alpha < 1$.

(i): Ordne die Daten $x_1, \dots, x_n$ der Größe nach. Wir nehmen ab jetzt an, dass $x_i \leqslant x_j$ falls $i \leqslant j$.

(ii): Suche das kleinste $k \in \mathbb{N}$ mit $k/n \geqslant \alpha$, also $k = \lceil \alpha n \rceil$

(iii): Für den Wert x_k sind dann mindestens k Datenpunkte $\leqslant x_k$, also

$$\frac{|\{i \leqslant n : x_i \leqslant x_k\}|}{n} \geqslant \frac{k}{n} \geqslant \alpha.$$

Wegen der Minimalität von k ist außerdem $\frac{k-1}{n} < \alpha$, somit gilt

$$\frac{|\{i \leqslant n : x_i \geqslant x_k\}|}{n} \geqslant \frac{n-k+1}{n} = 1 - \frac{k-1}{n} > 1 - \alpha.$$

Daher ist $q_\alpha = x_k$ schon einmal eine gültige Wahl.

(iv): Falls $\frac{k}{n} > \alpha$, dann ist $q_\alpha = x_k$ das einzige α-Quantil, denn dann ist

$$\frac{|\{i \leqslant n : x_i \geqslant x_k + \delta\}|}{n} \leqslant \frac{n-k}{n} < 1 - \alpha \qquad \forall \delta > 0.$$

(v): Falls $\frac{k}{n} = \alpha$ und $x_{k+1} = x_k$, dann ist $q_\alpha = x_k = \frac{x_{k+1}+x_k}{2}$ ebenfalls das einzige α-Quantil, denn dann ist

$$\frac{|\{i \leqslant n : x_i \geqslant x_k + \delta\}|}{n} \leqslant \frac{n-k-1}{n} < 1-\alpha \qquad \forall \delta > 0.$$

(vi): Falls $\frac{k}{n} = \alpha$ und $x_{k+1} > x_k$, dann kann q_α beliebig aus $[x_k, x_{k+1}]$ gewählt werden, denn dann ist

$$\frac{|\{i \leqslant n : x_i \geqslant x_{k+1}\}|}{n} = \frac{n-k}{n} = 1-\alpha.$$

In der Praxis wählt man in diesem Fall oft $q_\alpha = \frac{x_k+x_{k+1}}{2}$.

(5.7) Boxplots

Mit einem Boxplot kann man mehrere Eigenschaften eines Datensatzes gleichzeitig visualisieren. Seien $x_1, \ldots, x_n$ reelle Daten. Zur Erstellung eines Boxplots führt man folgende Schritte durch:

1) Man berechnet den Median $q_{1/2}$, das erste und das dritte Quartil $q_{1/4}$ und $q_{3/4}$, sowie den **Interquartilsabstand** $\mathrm{IQR} := q_{3/4} - q_{1/4}$.

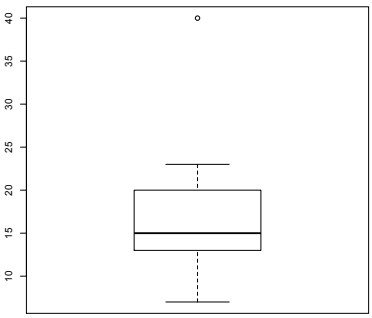
2) Man zeichnet eine Box (beliebiger Breite), deren obere Kante bei $q_{3/4}$ und deren untere Kante bei $q_{1/4}$ liegt. Den Median $q_{1/2}$ zeichnet man als waagrechte Linie in die Box.

3) Nun zeichnet man auf beiden Seiten Antennen („whiskers“) an die Box: sie beginnen an den Enden $q_{1/4}$ und $q_{3/4}$ der Box, und sie sind nie länger als bis zum letzten Datenpunkt in die jeweilige Richtung, aber auch nie länger als $1.5 \cdot \mathrm{IQR}$.

4) Für die tatsächliche Länge der Antennen gibt es mehrere Konventionen - wir benutzen die, die im Statistkproramm R eingebaut ist. Mit dieser Konvention geht die Antenne immer genau bis zum letzten Datenpunkt in die jeweilige Richtung, der noch einen Abstand $\leqslant 1.5 \cdot \mathrm{IQR}$ vom Rand der Box hat.

5) Alle Datenpunkte, die nun noch außerhalb der Spannweite der Antennen liegen, werden als Ausreißer separat durch kleine Kreise oder Punkte eingezeichnet.

Das nebenstehende Beispiel zeigt einen Boxplot aus den Daten $(7, 12, 13, 14, 14, 15, 16, 17, 20, 23, 40)$. Es ist $q_{1/4} = 13, q_{1/2} = 15, q_{3/4} = 20$, daher $\mathrm{IQR} = 7$ und $1.5 \cdot \mathrm{IQR} = 10.5$. Die obere Antenne endet bei $23 < 18 + 10.5$, die untere endet bei $7 > 13 - 10.5$. Der Ausreißer 40 ist separat eingezeichnet.



Der Plot wurde mit dem Statistik-Programm R erstellt. Dieses kennt insgesamt 9 (!) verschiedene Konventionen für die Berechnung von Quantilen, und nicht alle davon sind Spezialfälle der unseren in dem Sinne, dass man in (5.6 (vi)) eine Wahl trifft. [6)]Für sehr große Datenmengen spielt das dann meist keine Rolle, aber die Tatsache, dass es in der Statistik offenbar niemanden stört, wenn ein Objekt 9 verschiedene Definitionen hat, zeigt schon, dass man es hier nicht mehr mit Mathematik im engeren Sinne zu tun hat.

(5.8) Empirische Mittelwerte, empirische Varianz

a) Gegeben seien reelle Daten $(x_1, \ldots, x_n)$. Das arithmetische Mittel

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$$

dieser Daten heißt auch **empirischer Mittelwert**, die Größe

$$\bar{v} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

heißt **empirische Varianz**. Die Größe $\bar{\sigma} = \sqrt{\bar{v}}$ heißt **empirische Standardabweichung**.

b) Die Bedeutung der Normierung mit $\frac{1}{n-1}$ statt $\frac{1}{n}$ in der Formel für $\bar{v}$ sieht man, wenn man wieder annimmt, dass die Daten durch uiv ZVen erzeugt werden. Wir definieren dann

$$\bar{X}(\omega) := \frac{1}{n} \sum_{i=1}^{n} X_i(\omega), \qquad \bar{V}(\omega) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i(\omega) - \bar{X}(\omega))^2.$$

Unter der Annahme, dass diese ZVen einen endlichen Erwartungswert bzw. eine endliche Varianz besitzen, erhält man $\mathbb{E}(\bar{X}) = \mathbb{E}(X_1)$ und

$$\mathbb{E}(\bar{V}) = \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}\big((X_i - \bar{X})^2\big) = (*).$$

Zunächst erhalten wir nun für alle i

$$\mathbb{E}\big((X_i - \bar{X})^2\big) = \mathbb{E}(X_i^2) - \frac{2}{n} \sum_{j=1}^{n} \mathbb{E}(X_i X_j) + \frac{1}{n^2} \sum_{j,k=1}^{n} \mathbb{E}(X_j X_k) = (**),$$

[6)]Leider wählt die Funktion boxplot() bei R eine sehr obskure Konvention, die man auch nicht ändern kann - es ist also ohne Installation von Zusatzpackages nicht möglich, den nebenstehenden Plot aus den Daten zu erzeugen.

und da $\mathbb{E}(X_jX_k)$ für gleiche Indizes den Wert $\mathbb{E}(X_1^2)$ und für verschiedene den Wert $\mathbb{E}(X_1)^2$ annimmt, erhalten wir durch Zählen der entsprechenden Terme

$$(**) = \mathbb{E}(X_1^2) - 2\frac{n-1}{n}\mathbb{E}(X_1)^2 - \frac{2}{n}\mathbb{E}(X_1^2) + \frac{n}{n^2}\mathbb{E}(X_1^2) + \frac{n(n-1)}{n^2}\mathbb{E}(X_1)^2 =$$
$$= \mathbb{E}(X_1^2)\frac{n^2-2n+n}{n^2} - \frac{(n-1)}{n}\mathbb{E}(X_1)^2 = \frac{n-1}{n}\mathbb{V}(X_1).$$

Setzen wir das in $(*)$ ein erhalten wir genau $\mathbb{E}(\bar{V}) = \mathbb{V}(X_1)$. Im Mittel liefert $\bar{V}$ also (wegen der Normierung $\frac{1}{n-1}$) genau die richtige Varianz. Diese Eigenschaft, die wir später als „Erwartungstreue“ eines Schätzers kennenlernen werden, ist natürlich erwünscht, und rechtfertigt die Normierung.

c) Wie für die Histogramme kann man wieder zeigen: wenn genügend viele Momente endlich sind (nach unserem Beweis: 4 bzw. 8, mit dem optimalen Beweis: 1 bzw. 2), dann gilt mit dem starken GGZ:

$$\lim_{n\to\infty} \bar{X}(\omega) = \mathbb{E}(X_1), \qquad \lim_{n\to\infty} \bar{V}(X) = \mathbb{V}(X_1), \qquad \text{jeweils } \mathbb{P}\text{-fast sicher.}$$

d) Die Konvergenz in c) sagt uns nichts darüber aus, wie gut unsere Schätzung für endliches n ist. Hier hilft (zumindest wenn man sich mit stochastischer Konvergenz zufrieden gibt) wieder die Chebyshev-Ungleichung: beispielsweise ist für $X_1 \in \mathcal{L}^2$ und $\alpha < 1/2$

$$\mathbb{P}(|\bar{X} - \mathbb{E}(X_1)| > \varepsilon n^{-\alpha}) \leqslant \frac{n^{2\alpha}}{\varepsilon^2}\mathbb{V}(\bar{X}) = \frac{1}{\varepsilon^2 n^{2-2\alpha}}\mathbb{V}(\sum_{i=1}^n X_i) = \frac{1}{n^{1-2\alpha}\varepsilon^2}\mathbb{V}(X_1).$$

Andererseits wissen wir aus dem Zentralen Grenzwertsatz, dass wir nicht erwarten dürfen, dass $\bar{X} - \mathbb{E}(X_1)$ wesentlich kleiner als $\frac{1}{\sqrt{n}}$ ist, denn

$$\mathbb{P}(|\bar{X}-\mathbb{E}(X_1)| > \frac{D}{\sqrt{n}}) = \mathbb{P}(|\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i-\mathbb{E}(X_1))| > D) \overset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi\mathbb{V}(X_1)}}\int_{[-D,D]^c} \mathrm{e}^{-\frac{x^2}{2\mathbb{V}(X_1)}}\,\mathrm{d}x > 0.$$

e) Für festes $\varepsilon > 0$ lässt sich die Schranke aus d) jedoch noch stark verbessern: mit Hilfe der Theorie der großen Abweichungen kann man zeigen, dass in vielen Fällen ein $C_\varepsilon > 0$ und ein $r(\varepsilon) > 0$ existieren, so dass für alle n gilt:

$$\mathbb{P}(|\bar{X} - \mathbb{E}(X_1)| > \varepsilon) \leqslant C_\varepsilon\,\mathrm{e}^{-r(\varepsilon)n}\,.$$

Einen Spezialfall hiervon haben wir in (3.21 b) gesehen, wenn man dort $c = \varepsilon n$ setzt; man erhält dann dort $r(\varepsilon) = \varepsilon^2/2$. Mehr zu dieser Theorie und auch untere exponentielle Schranken an $\mathbb{P}(|\bar{X} - \mathbb{E}(X_1)| > \varepsilon)$ gibt es im nächsten Semester.

(5.9) Vergleich von Median und empirischem Mittelwert

Der Median hat gegenüber dem empirischen Mittelwert und dem Erwartungswert den Vorteil, dass er *robust* ist, was bedeutet, dass einzelne sehr große oder sehr kleine Datenpunkte (bzw. extreme Ereignisse, die mit sehr kleiner W'keit auftreten) ihn nicht stark verändern. Daher wird er sehr oft für statistische Betrachtungen bevorzugt. Beispielsweise lag das durchschnittliche Monatseinkommen der Stadt Heilbronn im Jahr 2017 bei ca. 42.000 Euro, und damit (weit) über dem von z.B. München. Der Grund ist allerdings, dass Heilbronn nicht sehr groß ist und

dort der Besitzer von Lidl und ein paar andere Leute dieser Art wohnen. Der Median des Monatseinkommens dagegen lag bei 3372 Euro, nur leicht über dem Bundesdurchschnitt.
Für theoretische Betrachtungen ist allerdings der Mittelwert und der Erwartungswert wegen der besseren algebraischen Eigenschaften zu bevorzugen - und für Verteilungen, die keine allzu großen Ausreißer haben, liegen Median und Mittelwert auch nicht allzu weit voneinander entfernt.

(5.10) Empirische Korrelation und Korrelationskoeffizient

Gegeben seien Datenpaare $(x, y) = (x_i, y_i)_{1 \leqslant i \leqslant n}$ mit $x_i, y_i \in \mathbb{R}$. $\bar{x}$ und $\bar{y}$ seien die empirischen Mittelwerte der x_i bzw. y_i, $\bar{v}_x$ und $\bar{v}_y$ die entprechenden empirischen Varianzen.
a) Die **empirische Kovarianz** der Datenpaare (x_i, y_i) ist die Zahl

$$\bar{c}_{x,y} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Der **empirische Korellationskoeffizient** der Datenpaare (x_i, y_i) ist die Zahl

$$\rho_{x,y} := \frac{\bar{c}_{x,y}}{\sqrt{\bar{v}_x \bar{v}_y}}.$$

b) Es gilt immer $-1 \leqslant \rho_{x,y} \leqslant 1$ (Übung mit Hilfe der Cauchy-Schwarz-Ungleichung). Die Daten heißen

(stark) positiv korreliert falls $\rho_{x,y}$ (wesentlich) größer als 0 ist,

(stark) negativ korreliert falls $\rho_{x,y}$ (wesentlich) kleiner als 0 ist,

schwach korreliert bzw. unkorreliert falls $\rho_{x,y}$ nahe bei 0 bzw. (im wesentlichen) gleich 0 ist.

c) $\rho_{x,y}$ ist dann wesentlich größer als 0, wenn für „viele“ Punktepaare gilt: ist $x_i - \bar{x}$ positiv (bzw. negativ), also x im Vergleich zum Durchschnitt $\bar{x}$ eher groß, so ist auch $y_i - \bar{y}$ positiv (bzw. negativ). $\rho_{x,y}$ ist dann wesentlich kleiner als 0, wenn für „viele“ Punktepaare gilt: $x_i - \bar{x}$ und $-(y_i - \bar{y})$ sind gleichzeitig positiv oder negativ. Das Anführungszeichen um das Wort „viele“ kommt daher, dass für positive Korrelation im Extremfall nur ein ein einziges Punktepaar $(x_i - \bar{x})(y_i - \bar{y}) > 0$ erfüllen muss, wenn diese Zahl riesig ist und alle anderen $(x_j - \bar{x})(y_j - \bar{y}) < 0$ aber betragsmäßig relativ klein sind. Das liegt daran, dass $\rho_{x,y}$ ebenso wie der empirische Mittelwert und die empirische Varianz nicht *robust* ist, d.h. anfällig für Beeinflussung durch extreme Ausreißer.

(5.11) Lineare Regression

a) Problemstellung:
Gegeben seien Paare $(x, y) = (x_i, y_i)_{i=1,\dots,n}$ von reellen Datenpunkten. Bei der Regression geht man davon aus, dass diese Datenpunkte „verrauschte“ Messungen eines funktionalen Zusammenhangs sind, d.h. man vermutet dass es eine Funktion $f : \mathbb{R} \to \mathbb{R}$ und uiv Zufallsvariablen ξ_i gibt mit

$$y_i = f(x_i) + \xi_i \qquad \forall i \leqslant n.$$

Zunächst einmal muss dafür $\bar{v}_x > 0$ sein, denn sonst sind alle x_i gleich, und man kann nicht sinnvoll entscheiden, welche Funktion man bestimmen soll. Außerdem muss man zusätzlich annehmen, dass f gewisse Zusatzeigenschaften hat, sonst könnte man ja etwa einfach eine stückweise lineare Funktion durch die Datenpunkte legen, falls die x_i alle verschieden sind.
Bei der linearen Regression nimmt man an, dass $f(u) = au + b$, also dass f linear ist, und versucht, a und b zu bestimmen.

b) Methode der kleinsten Quadrate:
In den meisten Fällen gibt es keine Funktion $f(u) = au + b$, deren Graph mehr als zwei Datenpaare (x_i, y_i) enthält - schließlich sind die Daten ja verrauscht! Um trotzdem die beste Wahl für a und b zu treffen, betrachtet man die Größen $|f(x_i) - y_i|$, also die Unterschiede zwischen den „idealen“ Werten $f(x_i)$ und den tatsächlich gemessenen Werten y_i. Die **quadratische Fehlerfunktion** ϕ ist dann definiert durch

$$\phi \equiv \phi(f; (x_i, y_i)_{i=1,\dots,n}) := \sum_{i=1}^{n} (f(x_i) - y_i)^2.$$

Für fest gegebene Daten ist dies eine Funktion von f, d.h. der Definitionsbereich von ϕ ist die Klasse der erlaubten Funktionen. Wir suchen nun innerhalb dieser Klasse diejenige (bzw. eine) Funktion, welche ϕ minimiert. Für die Klasse der linearen Funktionen $f(u) = au + b$ hat die quadratische Fehlerfunktion die Gestalt

$$\phi(a, b; (x_i, y_i)) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

Dies ist nun nur noch eine (quadratische) Funktion der zwei reellen Variablen a, b, so dass wir das absolute Minimum einfach durch Ableiten bestimmen können; die minimierenden a, b sind also Lösungen der linearen Gleichungen

$$0 = \frac{1}{2n} \partial_b \phi(a, b; (x_i, y_i)) = \frac{1}{n} \sum_{i=1}^{n} (ax_i + b - y_i)$$

und

$$0 = \frac{1}{2n} \partial_a \phi(a, b; (x_i, y_i)) = \frac{1}{n} \sum_{i=1}^{n} x_i (ax_i + b - y_i)$$

Die Lösung dieser Gleichung kann man mit folgendem Trick mit Hilfe bekannter Größen aufschreiben: Man legt die Gleichverteilung auf die Datenpaare (x_i, y_i), betrachtet also ZVen X, Y mit $\mathbb{P}(X = x_i, Y = y_i) = \frac{1}{n}$ für alle i. Dann lauten die beiden Gleichungen

$$0 = a\mathbb{E}(X) + b - \mathbb{E}(Y)$$
$$0 = a\mathbb{E}(X^2) + b\mathbb{E}(X) - \mathbb{E}(XY).$$

Also ist $b = \mathbb{E}(Y) - a\mathbb{E}(X)$, und setzt man dies in die zweite Gleichung ein, dann erhält man

$$0 = a\mathbb{E}(X^2) + \big(\mathbb{E}(Y) - a\mathbb{E}(X)\big)\mathbb{E}(X) - \mathbb{E}(XY) = a\mathbb{V}(X) - \mathrm{Cov}(X, Y).$$

Nun ist

$$\mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n-1}{n} \bar{v}_x,$$

und ebenso $\mathrm{Cov}(X,Y) = \frac{n-1}{n}\bar{c}_{x,y}$. Somit ist $a = \frac{\bar{c}_{x,y}}{\bar{v}_x}$, und wir erhalten die geschlossene Formel

$$f(u) = \frac{\bar{c}_{x,y}}{\bar{v}_x}u + \bar{y} - \frac{\bar{c}_{x,y}}{\bar{v}_x}\bar{x} = \frac{\bar{c}_{x,y}}{\bar{v}_x}(u-\bar{x}) + \bar{y}.$$

Für diese optimale Wahl a^*, b^* der Parameter hat die quadratische Fehlerfunktion den Wert

$$\phi(a^*,b^*;(x_i,y_i)) = \sum_{i=1}^{n}\Big(\frac{\bar{c}_{x,y}}{\bar{v}_x}(x_i-\bar{x})+(\bar{y}-y_i)\Big)^2 = n-1\Big(\frac{\bar{c}_{x,y}}{v_x}-2\frac{\bar{c}_{x,y}}{v_x}\bar{c}_{x,y}+v_y\Big) = (n-1)\bar{v}_y(1-\rho^2_{x,y}).$$

Man sieht daraus, dass die Approximationsgüte dann hoch ist, wenn die Daten stark korreliert sind; falls die y_i im Wesentlichen auf einer waagrechten Geraden liegen (also $\bar{v}_y$ klein ist), dann macht das die Approximationsgüte noch mal besser. Die Asymmetrie des Ausdruckes für die Approximationsgüte in den (x_i) und (y_i) kommt daher, dass nur der senkrechte Abstand $|y_i - f(x_i)|$ in die Schätzung eingeht. Dies ist jedoch im Anwendungsfall auch genau die richtige Annahme!

c) Andere Fehlerfunktionen:
Die Wahl der Fehlerfunktion ϕ wirkt etwas willkürlich - man könnte ja beispielsweise auch genauso gut

$$\phi_p \equiv \phi_p(f;(x_i,y_i)_{i=1,\dots,n}) := \sum_{i=1}^{n}|f(x_i)-y_i|^p$$

für jedes $p > 0$ nehmen. Die resultierende Regressionsgerade ist auch nicht unabhängig von dieser Wahl - große p sorgen dafür, dass man mehr kleinere Abweichungen $|f(x_i)-y_i|$ akzeptiert, um dadurch wenige sehr große solche Abweichungen zu vermeiden! Ein Vorteil der Wahl $p = 2$ ist, dass man damit viel besser rechnen kann als mit allen anderen Optionen.

Schließende Statistik: Schätz- und Testtheorie

Im Rest des Kapitels behandeln wir die parametrische Statistik: hier hat man Daten $(x_1,\dots,x_n)$ vorliegen, die aus einem von mehreren, durch einen Parameter θ unterschiedenen Zufallsmodellen stammen können. Das Ziel der Schätztheorie ist es, möglichst vernünftige Aussagen darüber zu machen, was der Parameter θ gewesen sein könnte, der die Daten $x_1,\dots,x_n$ erzeugt hat. Das Ziel der Testtheorie ist es, mit gutem Recht behaupten zu können, dass die Daten nicht von Modellen mit bestimmten Parametern stammen können. Beide Ziele sind sehr vage formuliert, was nicht daran liegt, dass wir uns zu wenig angestrengt haben, sondern daran, dass die Begriffsbildung hier nicht so einfach ist. Diese werden wir unten eingehende besprechen. Dazu kommt, dass traditionell die Nomenklatur in der Statistik besonders verwirrend und oft auch mathematisch mangelhaft ist. Das letztere habe ich versucht, in dieser Vorlesung anders und besser zu machen, mit dem kleinen Nachteil, dass man etwas mehr schreiben muss, und dem etwas größeren Nachteil, dass meine Notation von der Notation fast aller anderen Quellen abweicht.

Um den zweiten Nachteil etwas auszugleichen, wird weiter unten (wenn wir einige Beispiele gemacht haben) auf die Unterschiede zwischen der hier verwendeten und der klassischen Notation eingegangen werden.

(5.12) Das statistische Modell

a) Θ sei eine Menge, der sogenannte **Parameterraum**, $(E, \mathcal{E})$ ein messbarer Raum. Für jedes $\theta \in \Theta$ sei eine Familie $(X_{\theta,n})_{n\in\mathbb{N}}$ von E-wertigen ZVen gegeben. Die Familie $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ heißt **durch $\theta \in \Theta$ parametrisiertes statistisches Modell mit Datenmenge** E.

b) Sei $(E, \mathcal{E})$ ein messbarer Raum. Eine Familie $(S_n)_{n\in\mathbb{N}}$ von (bezüglich der jeweiligen Produkt-σ-Algebra) messbaren Abbildungen $S_n : E^n \to \mathbb{R}^d$ heißt **$\mathbb{R}^d$-wertige Statistik** für Datenmengen in E.

Beispiele:

a) Für $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ ist der **Mittelwertschätzer** die Statistik $(M_n)_{n\in\mathbb{N}}$ mit $M_n : \mathbb{R}^n \to \mathbb{R}$ und

$$M_n(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

b) Der **Varianzschätzer** ist die Statistik $(V_n)_{n\in\mathbb{N}}$ mit $V_n : \mathbb{R}^n \to \mathbb{R}$ und

$$V_n(x_1, \ldots, x_n) = \begin{cases} 0 & \text{falls } n = 1, \\ \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j)^2 & \text{sonst.} \end{cases}$$

c) Für jedes $p \in \Theta = [0, 1]$ sei $(X_{p,n})_{n\in\mathbb{N}}$ eine uiv Folge von Ber_p-verteilten ZVen. Das statistische Modell $(X_{p,n})_{p\in[0,1],n\in\mathbb{N}}$ heißt **Bernoulli-Modell**.

d) Für jedes $m \in \mathbb{R}$ und $v \geqslant 0$ seien uiv ZVen $(X_{(m,v),n})_{n\in\mathbb{N}}$ mit $X_{(m,v),1} \sim \mathcal{N}_{m,v}$ gegeben. Für $\Theta = \{(m, v) : m \in \mathbb{R}, v \geqslant 0\}$ ist $(X_{(m,v),n})_{(m,v)\in\Theta,n\in\mathbb{N}}$ das **Gauß-Produktmodell**.

(5.13) Schätzer

Sei $(E, \mathcal{E})$ ein messbarer Raum und $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ ein statistisches Modell mit Datenmenge E. Sei $h : \Theta \to \mathbb{R}^d$ eine Abbildung. Eine $\mathbb{R}^d$-wertige Statistik $T = (T_n)_{n\in\mathbb{N}}$ auf E heißt **Schätzer** (oder **Punktschätzer**) für h, wenn wir glauben, dass das Bildmaß der ZVen $T_n(X_{\theta,1}, \ldots, X_{\theta,n})$ für immer größere n den Wert $h(\theta)$ in sinnvoller Weise beschreibt.

Nachdem das aber keine vernünftige Definition ist, sagt man stattdessen: jede Statistik $T = (T_n)_{n\in\mathbb{N}}$ auf E heißt **Schätzer** für h. Das ist zwar auch absurd, aber wenigstens mathematisch rigoros. Die obere (sinnvolle aber mathematisch nicht greifbare) Definition werden wir gleich mit Hilfe von Eigenschaften des Schätzers einzufangen versuchen.

Beispiele:

a) Sei $\Theta = \mathbb{R} \times \Theta_0$ mit Θ_0 beliebig, und sei $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ eine beliebige Familie von $\mathbb{R}^d$-wertigen ZVen mit $\mathbb{E}(X_{(m,\theta_0),n}) = m$ für alle n und alle θ_0. Dann hat der Mittelwertschätzer M_n aus Beispiel (5.12 a) tatsächlich etwas mit dem Mittelwert zu tun, denn zumindest ist für $\theta = (m, \theta_0)$

$$\mathbb{E}(M_n(X_{\theta,1}, \ldots, X_{\theta,n}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_{\theta,i}) = m.$$

Wenn die (X_i) auch noch unabhängig sind (und für die von uns bewiesene Theorie: vierte Momente haben), dann gilt nach dem Gesetz der großen Zahlen sogar:

$$\lim_{n\to\infty} M_n(X_{\theta,1}, \ldots, X_{\theta,n}) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_{\theta,i} = m \quad \mathbb{P}\text{-fast sicher.}$$

Wenn wir also davon ausgehen, dass unsere Daten aus dem statistischen Modell $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ stammen, aber nicht wissen, die zu welchem θ gehörige Folge von ZVen die Daten nun konkret erzeugt hat, dann können wir bei Beobachtung genügend vieler Datenpunkte zumindest die erste Komponente von θ (also den Erwartungswert der $(X_{n,\theta})_{n\in\mathbb{N}}$) herausfinden. Somit ist M_n nicht nur ein Schätzer für die Abbildung $h : \Theta \to \mathbb{R}, (m, \theta_0) \mapsto m$ (denn jede $\mathbb{R}$-wertige Statistik ist ein Schätzer für m, nach Definition!), sonder sogar ein sinnvoller. Wir werden hierfür gleich noch andere Namen lernen.

b) Sei $\Theta = \mathbb{R}_0^+ \times \Theta_0$ mit Θ_0 beliebig, und sei $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ eine beliebige Familie von $\mathbb{R}^d$-wertigen ZVen mit $\mathbb{V}(X_{(v,\theta_0),n}) = v$ für alle n und alle θ_0. Zusätzlich seien nun die $(X_{\theta,n})_{n\in\mathbb{N}}$ für alle θ uiv ZVen. Der Varianzschätzer $V = (V_n)_{n\in\mathbb{N}}$ erfüllt dann für alle $\theta = (v, \theta_0)$ ebenfalls

$$\mathbb{E}(V_n(X_{\theta,1}, \ldots, X_{\theta,n})) = \mathbb{V}(X_{\theta,1}) = v,$$

das haben wir in (5.8 b) nachgerechnet. Dort haben wir auch erwähnt, dass bei der Existenz von genügend vielen Momenten wieder die Gleichung

$$\lim_{n\to\infty} V_n(X_{\theta,1}, \ldots, X_{\theta,n}) = v \qquad \mathbb{P}\text{-fast sicher}$$

gilt. Auch der Varianzschätzer ist also ein „sinnvoller“ Schätzer für die Abbildung $(v, \theta_0) \mapsto v$.

c) Für ein etwas konkreteres Beispiel kann man oben für die das stochastische Modell jeweils das Gaußmodell aus Beispiel (5.12 c) nehmen, einmal ist dann $\theta_0 = v$ und einmal $\theta_0 = m$.

(5.14) Bias und Erwartungstreue von Schätzern

Sei $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ ein statistisches Modell mit Daten in E, $h : \Theta \to \mathbb{R}^d$ eine Abbildung und $T = (T_n)_{n\in\mathbb{N}}$ ein Schätzer für h.

a) Die Größe

$$\mathrm{Bias}_{\theta,n}(T) := \mathbb{E}(T_n(X_{\theta,1}, \ldots, X_{\theta,n})) - h(\theta)$$

heißt **Bias (oder Verzerrtheit)** von T bezüglich der Schätzung von h im statistischen Modell $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$, bei n Datenpunkten.

b) T heißt **erwartungstreuer Schätzer von** h (bezüglich des statistischen Modells $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$) falls gilt:

$$\mathrm{Bias}_{\theta,n}(T) = 0 \qquad \forall n \in \mathbb{N}, \forall\theta \in \Theta.$$

c) T heißt **asymptotisch erwartungstreuer Schätzer von** h (bezüglich des statistischen Modells $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$) falls gilt:

$$\lim_{n\to\infty} \mathrm{Bias}_{\theta,n}(T) = 0 \qquad \forall\theta \in \Theta.$$

Bemerkung: wir haben bereits gesehen, dass sowohl der Mittelwertschätzer als auch der Vairanzschätzer erwartungstreu sind. Der „falsch normierte“ Varianzschätzer $\tilde{S}$ mit $\tilde{S}_n = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ ist nur asymptotisch erwartungstreu. Asymptotische Erwartungstreue sollte das mindeste sein, was man von einem Schätzer verlangt - eigentlich will man aber mehr, nämlich die Konsistenz:

(5.15) Konsistenz von Schätzern

Sei $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ ein statistisches Modell mit Daten in E, $h:\Theta\to\mathbb{R}$ eine Abbildung und $T=(T_n)_{n\in\mathbb{N}}$ ein Schätzer für h.

a) T heißt **schwach konsistent** (bezüglich der Schätzung von h im statistischen Modell $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$), falls gilt:

$$\forall\theta\in\Theta,\forall\varepsilon>0: \qquad \lim_{n\to\infty}\mathbb{P}(|T_n(X_{\theta,1},\dots,X_{\theta,n})-h(\theta)|>\varepsilon)=0.$$

b) T heißt **konsistent im quadratischen Mittel** (bezüglich der Schätzung von h im statistischen Modell $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$), falls gilt:

$$\forall\theta\in\Theta: \qquad \lim_{n\to\infty}\mathbb{E}\Big(\big|T_n(X_{\theta,1},\dots,X_{\theta,n})-h(\theta)\big|^2\Big)=0.$$

Fall T zusätzlich asymptotisch erwartungstreu ist, dann ist dies äquivalent zu der Bedingung

$$\lim_{n\to\infty}\mathbb{E}\Big(\big|T_n(X_{\theta,1},\dots,X_{\theta,n})-\mathbb{E}(T_n(X_{\theta,1},\dots,X_{\theta,n}))\big|^2\Big)\equiv\lim_{n\to\infty}\mathbb{V}(T_n(X_{\theta,1},\dots,X_{\theta,n}))=0 \qquad \forall\theta\in\Theta.$$

c) T heißt **stark konsistent** (bezüglich der Schätzung von h im statistischen Modell $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$), falls für alle $\theta\in\Theta$ gilt:

$$\lim_{n\to\infty}T_n(X_{\theta,1},\dots X_{\theta,n})=h(\theta) \qquad \mathbb{P}\text{-fast sicher.}$$

Bemerkungen: a) Seien Daten $x_1,\dots,x_n\in E$ und ein Schätzer $T=(T_n)_{n\in\mathbb{N}}$ gegeben, von dem man weiß, dass er für ein statistisches Modell $(X_{\theta,n})$ und die Funktion $h(\theta)=\theta$ schwach konsistent ist. Sei n sehr groß, $\Theta=\mathbb{R}$, und man nehme an, dass man sicher weiß, dass die Daten von einer der Folgen $(X_{\theta,n})_{n\in\mathbb{N}}$ erzeugt wurden. Wir nehmen sogar an, dass wir quantitative Schranken haben, dass wir also zu unserem n und gegebenem $\varepsilon>0$ ein $\delta>0$ finden können so dass gilt:

$$\mathbb{P}(|T_n(X_{\theta,1},\dots,X_{\theta,n})-\theta|>\varepsilon)\leqslant\delta \qquad \forall\theta\in\Theta.$$

Man kann nun versucht sein, diese Ungleichung folgendermaßen zu interpretieren:
„Das wahre θ liegt mit W'keit $1-\delta$ im Intervall $\{\theta\in\mathbb{R}:|T(x_1,\dots,x_n)-\theta|\leqslant\varepsilon\}$".

Das ist aber falsch: denn auf der Parametermenge Θ liegt kein W-Maß, daher ergibt die Aussage schlicht und einfach keinen Sinn[7)]: sowohl die x_i als auch das wahre θ liegen fest, und somit ist das wahre Theta entweder („mit Wahrscheinlichkeit 1") im betreffenden Intervall oder eben nicht.

Die richtige Interpretation der Gleichung ist folgende: die Wahrscheinlichkeit dass man von mit $(X_{n,\theta})_{n\in\mathbb{N}}$ erzeugten Daten um mehr als ε in die Irre geführt wird, dass man also Aufgrund des Ergebnisses von $T_n(X_{\theta,1},\dots,X_{\theta,n})$ auf ein θ' schließt, das weiter als ε von θ entfernt ist, diese W'keit ist kleiner als δ, und zwar egal welchen Wert das wahre θ hat. δ gibt also eine Schranke für die W'keit, mit den Daten „Pech zu haben" und mittel T um mehr als ε in die Irre zu gehen.

b) Die Namen der Konsistenzbegriffe orientieren sich an den Namen der Konvergenzbegriffe für ZVen, die für sie relevant sind; der Begriff starke Konsistenz orientiert sich am starken GGZ.

[7)] Man kann durchaus auf den Parameterraum auch ein W-Maß legen, dann ändert sich die Situation - dies ist die Bayes'sche Statistik.

c) Wir haben oben bereits gesehen, dass der Mittelwertschätzer und der Varianzschätzer unter geeigneten Bedingungen stark konsistent sind. Andere Konsistenzarten werden in den Übungen (ebenfalls unter geeigneten Bedingungen) behandelt.

(5.16) Maximum Likelihood (ML) Schätzer

a) Sei $(E, \mathcal{E})$ ein messbarer Raum, Θ ein messbarer Raum, und für jedes $n \in \mathbb{N}$ und jedes $x = (x_1, \dots, x_n) \in E^n$ sei eine messbare Funktion $\rho_{x,n} : \Theta \to \mathbb{R}_0^+$ gegeben. Für alle $n \in \mathbb{N}$ und alle $x \in E^n$ gelte

$$\operatorname{argmax}(\rho_{x,n}) := \{\theta \in \Theta : \rho_{x,n}(\theta) \geqslant \rho_{x.n}(\theta') \ \forall \theta' \in \Theta\} \neq \emptyset.$$

$h : \Theta \to \mathbb{R}^d$ sei eine Funktion, und $(T_n)_{n\in\mathbb{N}}$ eine $\mathbb{R}^d$-wertige Statistik, für die gilt:

$$T_n(x) \in \{h(\theta) : \theta \in \operatorname{argmax}(\rho_{x,n})\} \qquad \forall n \in \mathbb{N}, \forall x \in E^n.$$

Dann bezeichnen wir T als $(\rho_{x,n})$**-Maximums-Schätzer** für die Funktion h.[8)]

In Worten: T_n wählt also zu jedem $x \in E^n$ aus den möglicherweise mehreren $\theta \in \Theta$, bei denen die Abbildung $\theta \mapsto \rho_{x,n}(\theta)$ maximal ist, eines aus, und setzt es in die Funktion h ein.

b) Sei E endlich oder abzählbar, $\mathcal{E} = \mathcal{P}(E)$, $\Theta \subset \mathbb{R}^d$, und $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ ein statistisches Modell mit Daten in E. Zu jedem $n \in \mathbb{N}$ und jedem $\theta \in \Theta$ sei $\rho_{x,n}(\theta)$ die Zähldichte der Verteilung von $(X_{\theta,1}, \dots, X_{\theta,n})$. Dann erfüllt die Funnktion $\theta \mapsto \rho_{x,n}(\theta)$ die Bedingungen in a).
Ein $(\rho_{x,n})$-Maximumsschätzer mit dieser speziellen definierenden Funktion $\rho_{x,n}$ heißt **Maximum-Likelihood-Schätzer** für das statistische Modell $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ und zu schätzende Funktion h.

c) Sei $(E, \mathcal{E}) = (\mathbb{R}^m, \mathcal{B}^{\otimes m})$, $\Theta \subset \mathbb{R}^d$, und $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ ein E-wertiges statistisches Modell. Zu jedem $n \in \mathbb{N}$ und jedem $\theta \in \Theta$ habe die Verteilung von $(X_1, \dots, X_n)$ eine stetige Zähldichte $x = (x_1, \dots, x_n) \mapsto \rho_{x,n}(\theta)$. Dann erfüllt die Funktion $\theta \mapsto \rho_{x,n}(\theta)$ die Bedingungen in a); wir nehmen zusätzlich an, dass man eine Wahl für $T(x) \in \operatorname{argmax}(\rho_{x,n}) \subset \Theta$ treffen kann, so dass die Abbildung $x \mapsto T(x)$ messbar ist.
Ein $(\rho_{x,n})$-Maximumsschätzer mit dieser speziellen definierenden Funktion $\rho_{x,n}$ heißt **Maximum-Likelihood-Schätzer** für das statistische Modell $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ und zu schätzende Funktion h.

d) **Interpretation** (im Fall $h(\theta) = \theta$): Man hat die Daten $x = (x_1, \dots, x_n)$ gegeben, weiß, dass sie durch ZVen $(X_{\theta,1}, \dots, X_{\theta,n})$ für ein $\theta \in \Theta$ erzeugt wurden, und sucht das plausibelste $\theta \in \Theta$. Im Fall wo E endlich oder abzählbar ist, gilt für alle x und alle θ:

$$\mathbb{P}((X_{\theta,1}, \dots, X_{\theta,n}) = x) = \rho_{x,n}(\theta).$$

Der ML-Schätzer wählt dann unter allen Modellen dasjenige aus, das den Punkt x mit der höchsten W'keit hervorbringt; typischerweise ist diese W'keit für alle Modelle eher winzig, aber beim ausgewählten eben etwas größer als bei den anderen.
Im Fall von $\mathbb{R}^m$-wertigen ZVen ist typischerweise

$$\mathbb{P}((X_{\theta,1}, \dots, X_{\theta,n}) = x) = 0 \qquad \text{für alle } x \in E^n, \theta \in \Theta.$$

Daher übernimmt nun die Verteilungsdichte der $(X_{\theta,i})_{i=1,\dots,n}$ die Rolle dieser W-keiten. In den praxisrelevanten Fällen funktioniert das gut, aber man kann sich leicht künstliche Gegenbeispiele bauen, wo der ML-Schätzer sich dadurch ein θ aussucht, wo die Dichte der X_θ beim

[8)] Man beachte, dass wir insbesondere die Messbarkeit der Abbildung $x \mapsto T_n(x)$ gefordert haben.

gemessenen Datenpunkt x zwar groß ist, aber ringsherum viel schneller abfällt als bei anderen, nicht gewählten θ', so dass in diesen Fällen vielleicht systematisch das falsche θ gewählt wird.
Beispiel: $(X_{\theta,n})_{n\in\mathbb{N}}$ seien uiv Poi_θ-verteilte ZVen. Wir suchen für das statistische Modell $(X_{\theta,n})_{\theta\in[0,\infty),n\in\mathbb{N}}$ den ML-Schätzer. Hierzu berechnen wir zunächst die Zähldichte der Verteilung von $(X_{\theta,1},\dots,X_{\theta,n})$:

$$\rho_{(x_1,\dots,x_n),n}(\theta)=\prod_{i=1}^{n}\mathrm{e}^{-\theta}\,\frac{\theta^{x_i}}{x_i!}$$

Um das θ zu finden, das dies maximiert, könnte man diesen Ausdruck ableiten und nullsetzen - weniger Rechenarbeit ist es aber hier und in vielen anderen Fällen, wenn man ihn zunächst logarithmiert. Wir haben

$$\ln\rho_{(x_1,\dots,x_n),n}(\theta)=\sum_{i=1}^{n}(-\theta+x_i\ln\theta-\ln x_i!)=-n\theta+\ln\theta\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\ln x_i!,$$

und daraus errechnet man ganz leicht

$$\mathrm{argmax}(\rho_{(x_1,\dots,x_n),n})=\Big\{\frac{1}{n}\sum_{i=1}^{n}x_i\Big\}.$$

Somit ist der ML-Schätzer $T(x)=\frac{1}{n}\sum_{i=1}^{n}x_i$ gerade der Mittelwertschätzer. Dies ist wenig überraschend, da ja der $\mathbb{E}(X_{\theta,1})=\theta$ gilt, zeigt aber doch, dass die ML-Methode vernünftige Schätzer erzeugt.

Bemerkung: ML-Schätzer sind in vielen Fällen schwach konsistent, z.B. wenn für alle n und alle x die Funktion $\theta\mapsto\rho_{x,n}(\theta)$ unimodal ist, also ein eindeutiges globales Maximum besitzt, welches gleichzeitig auch das einzige lokale Maximum ist. Wir beweisen das aber in dieser Vorlesung nicht.

(5.17) Vergleich unserer Nomenklatur zur sonst üblichen

a) Unsere Definition des statistischen Modells weicht etwas von der allgemein üblichen ab: oft (z.B. im Buch von Georgii) ist ein statistisches Modell als Tripel $(E,\mathcal{E},(\mathbb{P}_\theta)_{\theta\in\Theta})$ definiert, man verzichtet also auf die Abhängigkeit von n und auf die Zufallsvaraiblen und arbeitet direkt auf den Bildmaßen: in unserer Sprache wäre das $(E,\mathcal{E},(\mathbb{P}_{X_{\theta,1}})_{\theta\in\Theta})$. Man spezialisiert dann auf den Fall wo $E=E_0^n,\mathcal{E}=\mathcal{E}_0^{\otimes n}$ und $\mathbb{P}_\theta$ ein Produktmaß ist, und nennt dies ein statistisches Produktmodell.

Wir haben also in unserer Definition sofort das Produktmodell eingeführt (allerdings ohne die Bedingung der Unabhängigkeit), und zwar gleich für alle $n\in\mathbb{N}$ auf einmal. Außerdem unterscheiden wir zwischen dem Raum E (auf dem die Daten leben) und dem W-Raum $(\Omega,\mathcal{F},\mathbb{P})$, auf dem zufällige Daten erzeugt werden. Diese zusätzliche Unterscheidung sorgt (hoffentlich) für mehr konzeptionelle Klarheit.

b) Ebenso haben wir die Definition der Statistik etwas anders aufgebaut: in der genannten Quelle ist eine Statistik einfach eine messbare Abbildung $S:E\to\Sigma$ für einen anderen messbaren Raum $(\Sigma,\mathscr{S})$. Die für uns unnötige Allgemeinheit dieses allgemeinen Raumes haben wir weggelassen, und wir haben darauf bestanden, dass man immer gleich für jedes n eine solche

messbare Abbildung angeben muss. Außerdem haben wir den Begriff der Statistik von dem des statistischen Modells entkoppelt, d.h. das beispielsweise die Mittelwertschätzer oder der Varianzschätzer immer die gleichen sind, egal welches statistische Modell man ansieht.

c) Die Philosophie hinter dieser leicht geänderten Definition ist die **strikte Trennung von Daten und Modellen**: In dem Raum $(E, \mathcal{E})$ liegen die Datenpunkte, und ein Element aus E^n ist ein Tupel mit n verschiedenen Datenpunkten. Jede der Zufallsvariablen $X_{\theta,n}$ entspricht der Erzeugung oder Messung eines dem Zufall unterworfenen Datenpunktes. Oft sind die $(X_{\theta,n})_{n\in\mathbb{N}}$ unabhängig. Wenn wir unter der Annahme, dass gegebene Datenpunkten $x_1, \dots, x_n \in E$ durch einen mit $X_{\theta,1}, \dots, X_{\theta,n}$ modellierten zufälligen Prozess erzeugt wurden, etwas über den Parameter θ herausfinden wollen, dann hilft (mir) die konzeptionelle Trennung von Datenpunkten und Zufallsvariablen sehr.

d) Hinzu kommt, dass in der klassischen Definition eine Statistik einfach nur eine messbare Abbildung ist - das ist also völlig redundant, und man redet sich dann damit heraus, dass man angeblich eine andere Interpretation dafür hat usw. In unserer Definition ist eine Statistik eine Folge von messbaren Abbildungen, d.h. man sollte von vorne herein angeben können, was man zu tun gedenkt, wenn man n Datenpunkte vorgelegt bekommt. Da man in so gut wie allen Situationen am Grenzwert vieler Datenpunkte interessiert ist, scheint mir das die richtige Definition - wenn man unbedingt will, kann man ja auch S_n für $n > 1$ trivial wählen, dann ist man zurück bei der klassischen Definition.

(5.18) Konfidenzintervalle: Vorüberlegungen

Ein Schätzer für den Parameter $\theta \in \Theta$ bekommt einen Datensatz $x_1, \dots, x_n$ und rät („schätzt“) daraus den Wert von $h(\theta)$, beispielsweise den Parameter selbst. Wenn allerdings etwa $\Theta = \mathbb{R}$, dann ist die Schätzung $T_n(x_1, \dots, x_n) \in \mathbb{R}$ eigentlich immer falsch in dem Sinne, dass in allen praxisrelevanten statistischen Modellen $(X_{\theta,n})$ gilt: $\mathbb{P}(T_n(X_{\theta,1}, \dots, X_{\theta,n}) \neq \theta) = 1$.
Als Gütekriterium für den Schätzer haben wir lediglich die Konsistenz benannt, die aussagt, dass der geschätzte Wert in einem gewissen (von der Art der Konsistenz abhängigen) Sinne nicht „allzu falsch“ ist, wenn man die Anzahl n der Datenpunkte groß werden lässt - für endliches n hilft das aber immer noch nicht.
Andererseits haben wir in Bemerkung a) zu (5.15) bereits den Fall angesprochen, dass wir die schwache Konsistenz durchaus auch auf quantitative Weise fordern können, d.h. wir können für jedes n und jedes $\varepsilon > 0$ die W'keit $\mathbb{P}(|T(X_{\theta,1}, \dots, X_{\theta,n}) - \theta| > \varepsilon)$, dass der Schätzer bei einem neuen, vom statistischen Modell erzeugten Datensatz um mehr als ε falsch liegt, durch ein (vermutlich von n und ε abhängiges, aber explizites) $\delta > 0$ begrenzen. Dies ist die Grundidee der Konfidenzintervalle.

(5.19) Bereichsschätzer

Sei $(E, \mathcal{E})$ ein messbarer Raum und $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ ein statistisches Modell mit Datenmenge E. Sei $h : \Theta \to \mathbb{R}^d$ eine Abbildung. Eine Folge von Abbildungen $T = (T_n)_{n\in\mathbb{N}}$ mit $T_n : E^n \mapsto \mathcal{B}(\mathbb{R}^d)$ heißt **Bereichschätzer** für h, wenn für jedes $a \in \mathbb{R}^d$ und alls $n \in \mathbb{N}$ gilt:

$$\{x \in E^n : T_n(x) \ni a\} \in \mathcal{E}^{\otimes n}.$$

Wie in der Definition des Punktschätzers kommt die zu schätzende Größe h nicht in der Definition vor. Wie dort implizieren wir: wenn die Daten $x \in E^n$ vorliegen und wir annehmen, dass diese Daten durch eine der Folgen von ZVen $(X_{\theta,n})_{n\in\mathbb{N}}$ aus dem statistischen Modell erzeugt wurden (wir aber θ nicht kennen), dann glauben wir daran, dass die Menge $T_n(x) \subset \mathcal{B}(\mathbb{R}^d)$ den wahren Wert von $h(\theta)$ mit einiger Sicherheit enthält.

(5.20) Konfidenzbereiche und Konfidenzintervalle

$(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ sei ein statistisches Modell mit Daten in $(E, \mathcal{E})$, $h : \Theta \to \mathbb{R}^d$, und $T = (T_n)_{n\in\mathbb{N}}$ ein Bereichsschätzer für h. $\alpha = (\alpha_n)_{n\in\mathbb{N}}$ sei eine reelle Folge mit $0 < \alpha_n < 1$ für alle n.

a) T heißt **Konfidenzbereich** für die Schätzung von h im statistischen Modell $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ zum Irrtumsniveau α (bzw. zum Sicherheitsniveau $1 - \alpha$), wenn gilt

$$\forall\theta \in \Theta, \forall n \in \mathbb{N} : \qquad \mathbb{P}(h(\theta) \in T_n((X_{\theta,1}, \ldots, X_{\theta,n}))) \geqslant 1 - \alpha_n.$$

b) Falls in der obigen Situation $d = 1$ und $T_n(x)$ für alle x ein (offenes, geschlossenes oder halboffenes) Intervall ist, dann heißt ein Konfidenzbereich auch **Konfidenzintervall**. Dies ist mit Abstand der wichtigste Spezialfall.

c) Irgendwelche Konfidenzintervalle zu finden ist nicht so schwierig - denn beispielsweise ist $T_n(x) = \mathbb{R}$ immer ein Konfidenzintervall. Das Ziel ist allerdings, möglichst kleine Konfidenzintervalle zu finden, da nur sie viel Aussagekraft über den wahren Wert von θ haben.

d) Sei $(\Omega, \mathcal{F}, \mathbb{P})$ der W-Raum, auf dem die $(X_{\theta,n})_{n\in\mathbb{N}}$ definiert sind. Wegen der geforderten Messbarkeit der T_n ist dann

$$\{\omega \in \Omega : h(\theta) \in T_n((X_{\theta,1}(\omega), \ldots, X_{\theta,n}(\omega)))\} = (X_{\theta,1}, \ldots, X_{\theta,n})^{-1}(\{x \in E^n : h(\theta) \in T(x)\}) \in \mathcal{F},$$

somit ist die in a) geforderte W'keit auch wirklich definiert.

e) In der klassischen Theorie wird sowohl die Messbarkeitsforderung oft weggelassen oder verschämt in einer Bemerkung hinterhergeschoben, also auch die Abhängigkeit von n nicht gemacht; für letzteres mag es gute Gründe geben (man hat eben in der Praxis eine feste Anzahl von Datenpunkten und nicht zu jedem n einen Satz Daten), aber für die Theorie scheint es mir klarer, wenn man wieder fordert, dass es im Prinzip möglich sein sollte, mit jedem beliebigen n zu arbeiten. Wenn man auf die klassische Theorie zurück will, kann man noch immer alle Begriffe außerhalb desjenigen n, das einen interessiert, trivial machen. Außerdem ist in der klassische Theorie, da man kein n hat, das α auch immer fest und nicht eine Folge (α_n) wie bei uns. Wenn man $\alpha_n = \alpha$ für alle n setzt, dann kann man das bei uns natürlich ebenso machen, in diesem Fall wird dann oft das Konfidenzintervall immer kürzer, je größer n ist. Eine andere Möglichkeit ist es, α_n mit wachsendem n schrumpfen zu lassen, dann bleibt (wenn man es richtig macht) das Konfidenzintervall beispielsweise für alle n gleich lang, aber das Irrtumsniveau wird immer besser (d.h. kleiner). Wir werden in den Beispielen unten sehen, wie das funktioniert.

f) Im Falle des Konfidenzintervalls kann man statt der Abbildung $T_n : E^n \to \mathcal{B}(\mathbb{R})$ auch die

bei den Abbildungen $T_n^{\pm} : E^n \to \mathbb{R} \cup \{-\infty, \infty\}$ betrachten, die die obere bzw. untere Intervallgrenze beschreiben. Dies sind dann wieder Statistiken im Sinne von Definition (5.12) (bis auf die Tatsache dass man $\pm\infty$ als Wert zulässt), und die Messbarkeit kann man im konkreten Fall dann meist ganz leicht überprüfen.

g) Die Bemerkung a) nach (5.15) gilt auch hier: Für gegebene Daten $(x_1, \dots, x_n)$ kann man nicht schließen, dass das wahre θ mit Wahrscheinlichkeit mindestens $1 - \alpha_n$ in $T_n(x)$ liegt, denn für den Parameter θ haben wir kein W-theoretisches Modell aufgestellt, und die Daten sind einfach, was sie sind, also n fest gegebene Zahlen. Die richtige Interpretation ist wieder: falls wir annehmen, dass für ein (uns unbekanntes) θ eine der Folgen $(X_{\theta,n})_{n\in\mathbb{N}}$ das relevante Experiment beschreibt, und falls man nun mittels dieser ZVen n neue Daten $X_{\theta,1}, \dots, X_{\theta,n}$ erzeugt, dann liegt mit W'keit $1 - \alpha_n$ der wahre Wert von θ in dem von diesen neuen Daten definierten Konfidenzbereich $T_n(X_{\theta,1}, \dots, X_{\theta,n})$, und zwar unabhängig davon, was das wahre θ ist. Natürlich kann es immer noch sein, dass gerade die Daten x, die uns vorliegen, dies nicht leisten, aber das kann eben nur mit viel Pech passieren (falls α_n klein ist).

Noch anschaulicher: stellen Sie sich ein Spiel vor, wo der Gegner sich ein beliebiges θ aussuchen darf, dann auf dem zu θ gehörigen Würfel n Zufallszahlen $X_{\theta,1}, \dots, X_{\theta,n}$ erzeugen und Ihnen zeigen muss - Ihre Aufgabe ist es, einen Bereich festzulegen, in dem Sie das wahre θ vermuten. Wenn (T_n) ein Konfidenzintervall zum Irrtumsniveau α ist und Sie raten, dass $\theta \in T_n(X_{\theta,1}, \dots, X_{\theta,n})$, dann werden Sie (auf lange Sicht) mindestens $100(1 - \alpha)$ Prozent aller gespielten Spiele gewinnen.

(5.21) Konstruktion von Konfidenzintervallen

a) Das Standardverfahren zur Konstruktion von Konfidenzintervallen ist folgendes: Sei $(X_{\theta,n})$ ein beiliebiges statistisches Modell mit Daten in E, $h : \Theta \to \mathbb{R}$, und (S_n) eine beliebige reellwertige Statistik auf E. Das zentrale Objekt unserer Betrachtungen sind die Bildmaße der Abbildungen

$$\omega \mapsto S_n(X_{\theta,1}(\omega), \dots, X_{\theta,n}(\omega)) - h(\theta)$$

für alle $\theta \in \Theta$ und alle $n \in \mathbb{N}$.

Seien $0 < \alpha < 1$, $\alpha_1, \alpha_2 > 0$ mit $\alpha_1 + \alpha_2 = \alpha$. Für gegebenes θ sei $q_{\alpha_1,\theta,n}$ sei das α_1-Quantil des zu θ und n gehörigen Bildmaßes, also der Verteilung von $S_n(X_{\theta,1}, \dots, X_{\theta,n}) - h(\theta)$, und $q_{1-\alpha_2}$ das $1 - \alpha_2$-Quantil ebendieser Verteilung. Dann gilt für alle $\theta \in \mathbb{R}$:

$$\begin{aligned}
&\mathbb{P}\big(h(\theta) \notin [S_n(X_{\theta,1}, \dots, X_{\theta,n}) - q_{1-\alpha_2,\theta,n}\,,\, S_n(X_{\theta,1}, \dots, X_{\theta,n}) - q_{\alpha_1,\theta,n})\big) = \\
&= \mathbb{P}\big(S_n(X_{\theta,1}, \dots, X_{\theta,n}) - q_{1-\alpha_2,\theta,n} > h(\theta)\big) + \mathbb{P}\big(S_n(X_{\theta,1}, \dots, X_{\theta,n}) - q_{\alpha_1,\theta,n} \leqslant h(\theta)\big) = \\
&= \mathbb{P}\big(S_n(X_{\theta,1}, \dots, X_{\theta,n}) - h(\theta) > q_{1-\alpha_2,\theta,n}\big) + \mathbb{P}\big(S_n(X_{\theta,1}, \dots, X_{\theta,n}) - h(\theta) \leqslant q_{\alpha_1,\theta,n}\big) = \alpha_2 + \alpha_1 = \alpha.
\end{aligned}$$

Für dieses spezielle θ erfüllt also die Abbildung $x \mapsto [S_n(x) - q_{1-\alpha_2,\theta,n}, S_n(x) - q_{\alpha_1,\theta,n}]$ die Bedingungen eines Konfidenzinetrvalls - allerdings reicht das nicht, denn das Konfidenzintervall zu den Daten x darf ja nicht von dem θ abhängen, das wir nicht kennen! Daher muss man das Intervall noch über alle θ maximieren: man definiert

$$q_{\alpha_1,n} := \inf_{\theta\in\Theta} q_{\alpha_1,n,\theta} < \sup_{\theta\in\Theta} q_{1-\alpha_2,n,\theta} =: q_{1-\alpha_2,n},$$

und dann ist für alle $\theta \in \Theta$

$$\mathbb{P}\big(h(\theta) \notin [S_n(X_{\theta,1},\dots,X_{\theta,n}) - q_{1-\alpha_2,n}\,,\, S_n(X_{\theta,1},\dots,X_{\theta,n}) - q_{\alpha_1,n})\big) \leqslant$$
$$\leqslant \mathbb{P}\big(h(\theta) \notin [S_n(X_{\theta,1},\dots,X_{\theta,n}) - q_{1-\alpha_2,\theta,n}\,,\, S_n(X_{\theta,1},\dots,X_{\theta,n}) - q_{\alpha_1,\theta,n})\big) \leqslant \alpha,$$

Daher ist in diesem Fall die Folge von Abbildungen $T_n : x \mapsto [S_n(x) - q_{1-\alpha_2,n}, S_n(x) + q_{\alpha_1,n})$ ein Konfidenzintervall zum Irrtumsniveau α für das statistische Modell $(X_{\theta,n})$ und die Funktion h. Natürlich kann bei ungeschickter Wahl der Statistik (S_n) wegen des sup und inf sehr leicht $T_n(x) = \mathbb{R}$ für alle x herauskommen - hier kommt es auf die richtige Wahl von S_n an! In sehr vielen Beispielen ist $q_{\alpha,\theta,n}$ sogar unabhängig von θ, so auch im folgenden.

(5.22) Konfidenzintervalle für den Mittelwert

a) $(X_{\theta,n})_{n\in\mathbb{N}}$ für jedes θ eine uiv Folge von $\mathcal{N}_{\theta,1}$-verteilten ZVen, und (S_n) mit $S_n(x) = \frac{1}{n}\sum_{i=1}^n x_i$ sei der Mittelwertschätzer. Dann ist

$$Y_{n,\theta} := S_n(X_{\theta,1},\dots X_{\theta,n}) - \theta = \frac{1}{n}\sum_{i=1}^n (X_{\theta,i} - \theta) = \frac{1}{n}\sum_{i=1}^n (X_{\theta,i} - \mathbb{E}(X_{\theta,i})) \qquad (*)$$

normalverteilt, mit Mittelwert 0 und Varianz $\frac{1}{n}$; also hängt die Verteilung von $Y_{n,\theta}$ gar nicht von θ ab. Bezeichnet $\bar{q}_\alpha$ das α-Quantil der Standardnormalverteilung, dann ist für alle θ und alle $\alpha, \alpha_1, \alpha_2$ wie in (5.20) und $Z \sim \mathcal{N}_{0,1}$:

$$\mathbb{P}(S_n(X_{\theta,1},\dots X_{\theta,n}) - \theta \leqslant \frac{\bar{q}_{\alpha_1}}{\sqrt{n}}) = \mathbb{P}(Y_{n,\theta} \leqslant \frac{\bar{q}_{\alpha_1}}{\sqrt{n}}) = \mathbb{P}(Z \leqslant \bar{q}_{\alpha_1}) = \alpha_1,$$

und ebenso für $\alpha_2 > 0$

$$\mathbb{P}(S_n(X_{\theta,1},\dots X_{\theta,n}) - \theta > \frac{\bar{q}_{1-\alpha_2}}{\sqrt{n}}) = \alpha_2.$$

Somit ist durch

$$T_n(x) = [S_n(x) - \frac{\bar{q}_{1-\alpha_2}}{\sqrt{n}}, S_n - \frac{\bar{q}_{\alpha_1}}{\sqrt{n}})$$

ein Konfidenzintervall zum Niveau $\alpha = \alpha_1 + \alpha_2$ definiert. Oft wählt man $\alpha_1 = \alpha_2 = \alpha/2$, und weil die Normalverteilung eine Dichte hat, braucht man bei den Intervallgrenzen nicht aufzupassen. Damit bekommt dann das Konfidenzintervall die symmetrische Form

$$T_n(x) = [S_n - \frac{\bar{q}_{1-\alpha/2}}{\sqrt{n}}, S_n + \frac{\bar{q}_{1-\alpha/2}}{\sqrt{n}}].$$

Hier sieht man nun auch, dass für $\alpha_n = \alpha$ für alle n das Konfidenzintervall immer kürzer wird, je größer n wird, d.h. je mehr Daten man hat. Alternativ kann man auch das Niveau verschärfen, also α_n mit wachsendem n immer kleiner machen. Richtet man z.B. α_n so ein, dass $\bar{q}_{1-\alpha_n/2} = c\sqrt{n}$ für ein $c > 0$, dann bleiben die Konfidenzintervalle gleich lang, aber das Irrtumsniveau geht gegen 0.

b) Für ein allgemeineres statistisches Modell, wo die $(X_{\theta,n})_{n\in\mathbb{N}}$ uiv ZVen mit endlicher Varianz und Mittelwert θ sind, bietet sich das gleiche Verfahren an: man wählt S_n wie in a), bestimmt die Verteilung von $Y_{n,\theta}$ und hat die berechtigte Hoffnung, dass diese nicht allzu sehr von θ abhängen wird. Für bestimmte konkrete statistische Modelle (z.B. Bernoulli oder Poisson) kann man das machen, mit unterschiedlich viel Rechenaufwand. Approximativ aber kann man genau so vorgehen wie in a), denn die ZV $Y_{n,\theta}$ ist nun zwar nicht mehr normalverteilt und könnte auch

von θ abhängen, aber wegen des zentralen Grenzwertsatzes kann man sie zumindest für große n recht gut durch eine Normalverteilung mit Mittelwert 0 und Varianz $\frac{1}{n}\mathbb{V}(X_{\theta,1})$ approximieren. Man kann ab dann wieder genau so vorgehen wie in a).

(5.23) Konfidenzintervalle für die Varianz, erster Versuch

$(X_{\theta,n})$ sei ein statistisches Modell mit Daten in $\mathbb{R}$, $\Theta = \mathbb{R}_0^+ \times \Theta_0$. Die $(X_{\theta,n})_{n\in\mathbb{N}}$ seien uiv und es gelte $\mathbb{V}(X_{\theta,1}) = v$ für $\theta = (v, \theta_0) \in \Theta$. In (5.8 a) haben wir den Varainzschätzer

$$V_n(x_1, \dots, x_n) := \frac{1}{n-1}\sum_{i=1}^n \Big(x_i - \frac{1}{n}\sum_{i=1}^n x_i\Big)^2$$

kennengelernt; wir wollen versuchen, mit diesem gemäß der Methode aus (5.21) ein Konfidenzintervall zum Niveau α konstruieren. Seien $\alpha_1, \alpha_2 > 0$ mit $\alpha_1 + \alpha_2 = \alpha < 1$, und sei $q_{\alpha_1,n,v}$ das α_1-Quantil der ZV $V_n(X_{\theta,1}, \dots, X_{\theta,n}) - v$, sowie $q_{1-\alpha_2,n,v}$ das $(1-\alpha_2)$-Quantil dieser ZV. Dann ist mit $q_{\alpha_1,n} := \inf_{v\in\mathbb{R}_0^+} q_{\alpha_1,n,\theta}$ und $q_{1-\alpha_2,n} = \sup_{v\in\mathbb{R}_0^+} q_{1-\alpha_2,n,v}$ das Intervall

$$J(x_1, \dots, x_n) := [V_n(x_1, \dots, x_n) - q_{1-\alpha_2,n}, V_n(x_1, \dots, x_n) - q_{\alpha_1,n})$$

ein Konfidenzintervall zum Niveau α, gemäß (5.21). Um die Quantile zu bestimmen, müssten wir jedoch etwas über die Verteilung von $V_n(X_{\theta,1}, \dots X_{\theta_n}) - v$ wissen. Im Fall von normalverteilten ZVen kommt man hier weiter:

(5.24) Satz von Student

$(X_1, \dots, X_n)$ seien uiv mit $X_1 \sim \mathcal{N}_{\mu,\sigma^2}$. Mit

$$M_n(\omega) := \frac{1}{n}\sum_{i=1}^n X_i(\omega), \qquad V_n(\omega) := \frac{1}{n-1}\sum_{i=1}^n (X_i(\omega) - M(\omega))^2$$

gilt:

a) $M_n \perp\!\!\!\perp V_n$, und $M_n \sim \mathcal{N}_{\mu,n^{-1}\sigma^2}$.

b) Die Verteilung von $\frac{n-1}{\sigma^2}V_n$ ist also die Gamma-Verteilung mit Parametern $(1/2, n/2)$, in Formeln:

$$\frac{n-1}{\sigma^2}V_n \sim \Gamma_{1/2,n/2}, \qquad \text{also} \qquad \mathbb{P}\Big(\tfrac{n-1}{\sigma^2}V_n \in [a,b]\Big) = \int_{[a,b]} \frac{2^{-n/2}}{\Gamma(n/2)} x^{n/2}\, \mathrm{e}^{-x/2}\, \mathbb{1}_{[0,\infty)}(x)\, \mathrm{d}x$$

In der Statistik heißt die $\Gamma_{1/2,n/2}$-Verteilung auch **χ^2-Verteilung zum Parameter** $n-1$, oder mit $n-1$ Freiheitsgraden. Man schreibt $\Gamma_{1/2,n/2} \sim \chi^2_{n-1}$.

c) Die Verteilung mit der Lebesguedichte

$$\tau_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(n/2)\Gamma(1/2)}\Big(1 + \frac{x^2}{n}\Big)^{-\frac{n+1}{2}}$$

heißt **(Student'sche) t-Verteilung mit n Freiheitsgraden** und wird mit t_n bezeichnet. Es gilt:

$$\text{Für } T_{n,\mu}(\omega) := \frac{\sqrt{n}(M_n(\omega) - \mu)}{\sqrt{V_n(\omega)}} \quad \text{ ist } \quad T_{n,\mu} \sim t_{n-1}.$$

Bemerkungen:

a) Der Beweis dieses Satzes besteht vor allem im geduldigen und fehlerfreien Anwenden der Transformationsformel für Integrale. Teile a) und b) machen wir als Übung, für Teil c) sei auf die Literatur verwiesen.

b) Schaut man die Dichte τ_n genau an, so erkennt man im x-abhängigen Teil im Wesentlichen die Approximation der Exponentialfunktion $\mathrm{e}^{-x^2/2}$ durch das Polynom $(1+\frac{x^2/2}{n/2})^{-n/2}$. Die t-Verteilung ist also für große n der Normalverteilung sehr ähnlich, für kleinere n ist sie ein wenig breiter.

(5.25) Konfidenzintervall für die Varianz im Gauß-Modell, zweiter Versuch

Wir wenden Satz (5.24) auf die ZV $V_n \equiv V_n(X_{\theta,1},\dots,X_{\theta,n})$ mit $\theta=(v,\theta_0)$ aus (5.23) an, unter der Annahme, dass die $X_{\theta,n}$ normalverteilt mit Varianz v sind. Dann ist für $Y\sim\chi^2_{n-1}$ und $0<\beta<1$

$$\beta=\mathbb{P}(V_n-v\leqslant q_{\beta,n,v})=\mathbb{P}\Big(\frac{V_n}{v}(n-1)\leqslant\big(\frac{q_{\beta,n,v}}{v}+1\big)(n-1)\Big)=\mathbb{P}\Big(Y\leqslant\big(\frac{q_{\beta,n,v}}{v}+1\big)(n-1)\Big)$$

Wenn also $q_{\beta,Y}$ das β-Quantil der χ^2_{n-1}-Verteilung ist, dann gilt somit

$$q_{\beta,n,v}=\Big(\frac{q_{\beta,Y}}{n-1}-1\Big)v.$$

Für gar nicht einmal so kleine $\beta>0$ ist also $\inf_{v\in\mathbb{R}_0^+} q_{\beta,n,v}=-\infty$, und genauso für β einigermaßen nahe bei 1 dann $\inf_{v\in\mathbb{R}_0^+} q_{\beta,n,v}=\infty$. Das macht unser Konfidenzintervall aus ((5.24)) leider ziemlich nutzlos.

Was ist passiert? Das Problem ist, dass wir das Bildmaß der $(X_{\theta,1},\dots,X_{\theta,n})$ unter unserer Statistik $x\mapsto V_n(x)-v$ zwar dank (5.24) nun ausrechnen können, dass es aber leider sehr stark von v abhängt. Das hätten wir vorher schon erraten können, und es liegt auch nicht an der Normalverteilung: wenn man für ZVen mit einer riesigen Varianz ($v\to\infty$) die empirische Varianz ausrechnet, dann darf man sich nicht wundern, wenn die sehr stark schwankt, also die Quantile sehr negativ oder sehr groß sind.

Die Lösung des Problems ist die Wahl einer besseren Statistik. Wir weichen daher von der in (5.23) gegebenen einfachsten Lösung ab und setzen

$$S_n(x)=\frac{(n-1)V_n(x)}{v}$$

Sei $\tilde q_{\beta,n,v}$ das β-Quantil der Verteilung von $S_n(X_{\theta,1},\dots,X_{\theta,n})$. Dann ist, wenn $Y\sim\chi^2_{n-1}$,

$$\beta=\mathbb{P}\Big(\frac{(n-1)V_n(X_{\theta,1},\dots,X_{\theta,n})}{v}\leqslant\tilde q_{\beta,n,v}\Big)=\mathbb{P}(Y\leqslant q_{\beta,n,v})$$

Daher gilt $\tilde q_{\beta,n,v}=q_{\beta,Y}$, also unabhängig von v. Somit ist

$$C(x):=\Big[\frac{(n-1)V_n(x)}{q_{1-\alpha_2,Y}},\frac{(n-1)V_n(x)}{q_{\alpha_1,Y}}\Big]$$

ein Konfidenzintervall, denn es gilt beispielsweise

$$\mathbb{P}\Big(v < \frac{(n-1)V_n(X_{\theta,1},\dots,X_{\theta,n})}{q_{1-\alpha_2,Y}}\Big) = \mathbb{P}\Big(\frac{(n-1)V_n(X_{\theta,1},\dots,X_{\theta,n})}{v} > q_{1-\alpha_2,Y}\Big) < \alpha_2,$$

ebenso mit α_1.
Übung: überzeugen Sie sich, dass anders als es aufgrund dieser Formel auf den ersten Blick der Fall zu sein scheint, dieses Konfidenzintervall für große n kürzer wird, nämlich von der Größenordnung $\frac{1}{\sqrt{n}}$ ist.

(5.26) Konfidenzintervall für den Mittelwert bei unbekannter Varianz und Normalverteilungen

Sei $\Theta = \mathbb{R} \times \mathbb{R}_0^+$, $\theta = (m, v)$ und $(X_{\theta,n})_{n\in\mathbb{N}}$ uiv mit $X_{\theta,1} \sim \mathcal{N}_{m,v}$. Ziel ist es, ein Konfindenzintervall für $h(\theta) = m$ zu bestimmen. Das Konfidenzintervall aus (5.22) funktioniert nicht mehr; denn wenn M_n der Mittelwertschätzer und $\bar{q}_\alpha$ das α-Quantil von $\mathcal{N}_{0,1}$ ist, dann ist für festes v zwar

$$K_v(x) := [M_n(x) - \frac{\sqrt{v}}{\sqrt{n}}q_{1-\alpha/2}, M_n(x) + \frac{\sqrt{v}}{\sqrt{n}}q_{1-\alpha/2}],$$

da wir das aber über alle v maximieren müssen, ist es wieder nutzlos. Mit (5.24 c) können wir aber eine Statistik finden, wo wir das relevante Bildmaß kennen und es nicht von m abhängt: Ist V_n der Varianzschätzer, dann setzen wir

$$S_n(x) = \frac{\sqrt{n}M_n(x) - m}{\sqrt{V_n(x)}},$$

somit ist $S_n(X_{\theta,1},\dots,X_{\theta,n}) \sim t_{n-1}$. Wenn $\bar{q}_\beta$ das β-Quantil der t-Verteilung ist, dann gilt

$$\beta = \mathbb{P}(S_n(X_{\theta,1},\dots,X_{\theta,n}) \leqslant \bar{q}_\beta) = \mathbb{P}\Big(M_n(X_{\theta,1},\dots,X_{\theta,n}) - \frac{\sqrt{V_n(X_{\theta,1},\dots,X_{\theta,n})}}{\sqrt{n}}\bar{q}_\beta \leqslant m\Big),$$

und analog für mit $\bar{q}_{1-\beta}$. Es folgt, dass

$$C(x) := \Big[M_n(x) + \frac{\sqrt{V_n(x)}}{\sqrt{n}}\bar{q}_{1-\alpha_2}, M_n(x) + \frac{\sqrt{V_n(x)}}{\sqrt{n}}\bar{q}_{\alpha_1}\Big]$$

ein Konfidenzintervall zum Niveau $\alpha_1 + \alpha_2$ ist.
Bemerkungen:
a) Man beachte die formale Ähnlichkeit des Konfidenzintervalls mit dem für bekannte und feste Varianz: formal ersetzen wir einfach diese bekannte Varianz durch die geschätzte, zu dem Preis, dass wir statt der Quantile der Normalverteilung die Quantile der (breiteren) t-Verteilung benutzen müssen. Dadurch wird das Konfidenzintervall etwas größer.

b) Die Standard-Methode (5.21) zum Berechnen von Konfidenzintervallen funktioniert in den letzten beiden Beispielen nicht mehr und musste erweitert werden, in folgendem Sinne: Wir verschaffen uns zunächst eine Statistik S_n, die sowohl von den Daten als auch von $h(\theta)$ abhängen darf, wo aber das Bildmaß der $X_{\theta,i}$ unter dieser Statistik nicht von θ abhängt. Dies erlaubt es uns, ohne Maximierung über θ die Quantile dieses Bildmaßes zu bestimmen. Nun müssen wir in der Lage sein, diese Statistik für alle x zu invertieren, d.h. den Ausdruck nach $h(\theta)$

aufzulösen. Dann bekommen wir durch diese Invertierung und durch die berechneten Quantile ein Konfidenzinetrvall, ähnlich wie es oben gemacht wurde.

(5.27) Tests: Grundideen und Beispiele

In vielen Situationen des Lebens sind wir gezwungen, binäre (d.h. Ja-Nein) Entscheidungen auf Grundlage von kontinuierlichen und möglicherweise mit Unsicherheiten behafteten Informationen zu treffen. Eines der ersten Beispiele, das einem vielleicht einfällt, ist hier das Eingehen von Beziehungen oder Ehen, was wir aber aus verschiedenen Gründen lieber nicht mathematisch formalisieren wollen.

Sehr wichtige Anwendungen dagegen sind gesetzliche Regelungen, zum Beispiel bei der Medikamentenzulassung. Zwar ist es klar, dass die Wirksamkeit und Verträglichkeit eines neuen Medikaments nicht entweder vorhanden oder abwesend ist, sondern sich auf einem weiten und nicht leicht zu messenden Spektrum bewegt. Trotzdem muss der Gesetzgeber letztlich eine binäre Entscheidung treffen, nämlich ob er das Medikament freigibt und bereit ist, mit den Geldern der Krankenkassen zu kaufen, oder nicht. Für diese Entscheidung muss es objektive Spielregeln geben, die vorher bekannt gemacht werden. Dies und ähnliche regulatorische Entscheidungen (z.B. Grenzwerte für Schadstoffe) sind das wichtigste Einsatzgebiet statistischer Tests.

Ein weiteres Einsatzgebiet sind die Untermauerung wissenschaftlicher oder anderer Zusammenhänge, deren Messung relative großen Unsicherheiten unterliegt. Beispiele hierfür wären:

a) die Reaktion eines kleinen Ökosystems auf gewisse Einflüsse von außen (z.B. Abnahme von Individuen einer Spezies bei Eingriffen in Fließgewässer);

b) die Qualitätsprüfung von Produkten aufgrund kleiner Stichproben; denken Sie beispielsweise an eine Lieferung von 10000 Feuerwerkskörpern, bei der der Lieferant garantiert, dass höchstens 3% nicht funktionieren; wenn Sie das kurz nach dem Kauf überprüfen wollen, sollten Sie nicht alle testen, sondern nur einen kleinen Teil, und daraus Schlüsse ziehen.

c) der „Beweis“ der Wirksamkeit eines neuen „Superfood“ für das persönliche Wohlbefinden.

Der Einsatz von Tests ist hier wieder vor allem dann sinnvoll, wenn die Ergebnisse in binäre Entscheidungen münden - beim Beispiel a) etwa die Renaturierung von Bächen, die mit hohen Kosten verbunden ist, bei Beispiel b) die Reklamation der Lieferung mit eventuell folgendem Rechtsstreit und teuren Gutachten. In anderen Fällen sind Tests weit weniger sinnvoll, denn man postuliert damit eine binäre Wirklichkeit, die es so vielleicht gar nicht gibt. Trotzdem werden sie auch dort oft eingesetzt, entweder zu Werbezwecken (Beispiel c), oder aufgrund des menschlichen Unbehagens gegen Aussagen, die nicht mit einem einfachen „gilt / gilt nicht“ zu erledigen sind.

Aus den obigen Beispielen sieht man übrigens, dass oft eine der zwei Handlungsoptionen risikoreicher oder teurer ist als die andere: Bei Medikamenten ist eine fehlerhafte Zulassung eines schädlichen Medikaments eine Katastrophe (suchen Sie im Netz nach Contaganskandal), die fehlerhafte Nicht-Zulassung eines nützlichen Medikaments lediglich eine verpasste Chance. Vergleichbares gilt für die Beispiele Renaturierung (hohe Kosten ohne Wirkung, Ende der

politischen Laufbahn vs. Beibehaltung des Status quo und niemand nimmt Notiz), Reklamation (Ärger, Gutachter, Gerichtsprozesse vs. kleiner wirtschaftlicher Schaden durch Nicht-Bemerken), und im Prinzip auch Superfood (Blamage und PR-Gau bei belegter Wirkungslosigkeit vs. etwas weniger Umsatz), wobei die Praxis im letzten Beispiel oft eine andere ist. Diese Asymmetrie spiegelt sich auch in der Definition von Tests wieder.

(5.28) Tests: Definitionen

Sei $(E, \mathcal{E})$ ein messbarer Raum, Θ sei eine Menge und $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ ein statistisches Modell mit Daten in E. Sei $H_0 \subset \Theta$, $H_1 = \Theta \setminus H_0$, und $\alpha = (\alpha_n)_{n\in\mathbb{N}}$ mit $0 < \alpha_n < 1$ für alle $n \in \mathbb{N}$.

a) Die Aussage $\theta \in H_0$ heißt **Nullhypothese** und die Aussage $\theta \in H_1$ **Alternative** oder **Alternativhypothese**.

b) Ein **Test mit Nullhypothese** H_0 **zu Niveau** α ist eine $\{0,1\}$-wertige Statistik für die Datenmenge E, für die gilt:

$$\mathbb{P}(T_n(X_{\theta,1},\dots,X_{\theta,n}) = 1) \leqslant \alpha_n \qquad \forall\theta \in H_0.$$

c) Die **Gütefunktion** eines Tests T mit Nullhypothese H_0 ist die Funktionenfolge $G = (G_n)_{n\in\mathbb{N}}$ mit

$$G_n : \Theta \to [0,1], \qquad \theta \mapsto G(\theta) = \mathbb{P}(T_n(X_{\theta,1},\dots,X_{\theta,n}) = 1) = \mathbb{E}(T_n(X_{\theta,1},\dots,X_{\theta,n}))$$

Für $\theta_1 \in H_1$ heißt $G(\theta_1)$ auch die **Macht** des Tests T beim Parameter θ_1.

d) Ein Test T mit Nullhypothese H_0 heißt **unverfälscht** zum Niveau α, wenn für alle $n \in \mathbb{N}$ gilt:

$$G_n(\theta_0) \leqslant \alpha_n \leqslant G_n(\theta_1) \qquad \forall\theta_0 \in H_0, \forall\theta_1 \in H_1.$$

Bemerkungen:

a) $T_n(x) = 1$ bedeutet, dass man aufgrund der Daten x vermutet, dass (für kleine α) sehr wahrscheinlich die Alternativhypothese gilt, der wahre Parameter, der die Daten erzeugt hat, also in H_1 liegt. Formalisiert wird der Ausdruck „sehr wahrscheinlich“ genau wie bei den Konfidenzintervallen: Wenn man für irgend einen Parameter $\theta_0 \in H_0$ die ZVem $X_{\theta_0,1},\dots,X_{\theta_0,n}$ in den Test einsetzt, erhält man eine ZV; von dieser verlangt man, dass sie höchstens mit W'keit α den „falschen“ Wert 1 annimmt. Man denkt also wieder so: falls jemand mit einem zu $\theta_0 \in H_0$ gehörigen Zufallsmechanismus Daten erzeugt, mir vorlegt, und mich fragt, ob der Parameter, der die Daten erzeugt hat, in H_0 liegt, dann werde ich mich aufgrund dieser Daten in $(1-\alpha)\cdot 100$ Prozent aller Fälle für die richtige Antowrt („Ja“) entscheiden.

b) Für fest vorgegebene Daten kann man wieder, aus dem gleichen Grund wie früher, nichts darüber sagen, mit welcher W'keit man Recht hat. Ebenso sagt das Niveau α nichts darüber aus, mit welcher W'keit man die Tatsache $\theta \in H_1$ aus zufällig erzeugten Daten erkennt, wenn sie wahr ist. Letzteres ist die Aufgabe der Gütefunktion. Bei einem Test vom Niveau α ist notwendigerweise $G_n(\theta_0) \leqslant \alpha$ für $\theta_0 \in H_0$, aber man möchte Tests so konstruieren, dass $G(\theta_1)$ für möglichst viele $\theta_1 \in H_1$ möglichst nahe bei 1 ist. Falls $\theta = \mathbb{R}$ und $H_0 = (-\infty, c]$, dann möchte man also, dass $G_n(\theta)$ der Indikatorfunktion $\mathbb{1}_{(c,\infty)}(\theta)$ möglichst ähnlich wird. Da $\theta \mapsto G_n(\theta)$ oft stetig ist, muss man sich mit einer Approximation zufrieden geben, und das Niveau α schreibt zusätzlich vor, dass (falls G_n monoton ist, was im gegebenen Beispiel meist so ist) $G_n(c) \leqslant \alpha$ sein muss.

c) Die in (5.27) angesprochene Asymmetrie schlägt sich in der unterschiedlichen Behandlung von H_0 und H_1 nieder: man *erzwingt* $\mathbb{P}(T_n(X_{\theta,1},\dots,X_{\theta,n})=1)\leqslant\alpha_n$ für $\theta\in H_0$, während man für $\theta\in H_1$ zunächst keine Forderungen stellt. Da α_n oft sehr klein ist (0.05 oder 0.01 sind beliebte Werte), bedeutet das, dass eine (fälschliche) Entscheidung für H_1 selten ist, wenn $\theta\in H_0$ gilt, während man bezüglich einer fälschlichen Enscheidung für H_0 bei $\theta\in H_1$ weit toleranter ist. Hier gibt es einige Nomenklatur:
(i): Man sagt, **die Nullhypothese H_0 wird aufgrund der Daten x abgelehnt** wenn $T(x)=1$.
(ii): Man sagt, man begeht einen **Fehler erster Art**, wenn man die Nullhypothese ablehnt, obwohl sie wahr ist.
(iii): Man sagt, man begeht einen **Fehler zweiter Art**, wenn man die Nullhypothese nicht ablehnt, obwohl sie falsch ist.

In den obigen Beispielen ist die Nullhypothese also die Unwirksamkeit oder Gefährlichkeit eines neuen Medikaments, die Einhaltung vertragsgemäßer Qualitätsstandards bei Feuerwerkskörpern, oder die Wirkungslosigkeit von Superfood. Ein Fehler erster Art ist somit die (katastrphale) Zulassung eines schädlichen oder nutzlosen Medikaments, die ungerechtfertigte Reklamation etc.

d) Die Bedeutung der Unverfälschtheit sieht man am besten an Beispielen - zunächst halten wir nur fest, dass Unverfälschtheit bedeutet, dass man ein $\theta_1\in H_1$ mindestens mit der gleichen W'keit also solches erkennt, wie man ein $\theta_0\in H_0$ fälschlicherweise nach H_1 einordnet.

e) In der Testtheorie setzt man eigentlich immer $\alpha_n=\alpha_1=\alpha$ für alle n. Der Grund ist, dass man sich von vorne herein für ein Niveau entscheidet, mit dem man dann zufrieden ist. Durch Erhöhen der Datenmenge n verbessert sich die Macht des Tests.

f) Mathematisch sind Tests eng verwandt mit Konfidenzbereichen: Sei $h:\Theta\to\{0,1\}$ mit $h(\theta)=\mathbb{1}_{H_1}(\theta)$, und T_n ein Test zur Nullhypothese H_0 zum Niveau α. Betrachte den Bereichsschätzer $\bar{T}_n(x):=\{T_n(x)\}$. Dann gilt für alle $\theta_0\in H_0$:

$$\mathbb{P}(h(\theta_0)\notin\bar{T}_n(X_{\theta_0,1},\dots,X_{\theta_0,n}))=\mathbb{P}(0\neq T_n(X_{\theta_0,1},\dots,X_{\theta_0,n}))=\mathbb{P}(T_n(X_{\theta_0,1},\dots,X_{\theta_0,n})=1)\leqslant\alpha_n.$$

Die Abbildung $\bar{T}_n$ erfüllt also die Bedingung für ein Konfindenzintervall für die Schätzung von h, allerdings nicht für alle θ sondern nur für $\theta\in H_0$.

(5.29) Konstruktion von Tests aus Konfidenzbereichen

Sei $(E,\mathcal{E})$ ein messbarer Raum, $\Theta\subset\mathbb{R}^d$, und $(X_{\theta,n})_{\theta\in\Theta,n\in\mathbb{N}}$ ein statistisches Modell mit Daten in E. Sei $H_0\subset\Theta$, $H_1=\Theta\setminus H_0$, und $\alpha=(\alpha_n)_{n\in\mathbb{N}}$ mit $0<\alpha_n<1$ für alle $n\in\mathbb{N}$.

Sei $C=(C_n)_{n\in\mathbb{N}}$ ein Konfidenzbereich für die Abbildung $h(\theta)=\theta$ zum Irrtumsniveau α. Dann ist die Abbildung $T=(T_n)_{n\in\mathbb{N}}$ mit

$$T_n(x)=\begin{cases}1 & \text{falls } C_n(x)\cap H_0\neq\emptyset,\\ 0 & \text{sonst}\end{cases}$$

ein Test zum Niveau α mit Nullhypothese H_0.
Beweis: Sei $\theta\in H_0$. Dann ist

$$\begin{aligned}&\mathbb{P}(T_n(X_{\theta,1},\dots,X_{\theta,n})=1)=\mathbb{P}(C_n(X_{\theta,1},\dots,X_{\theta,n})\cap H_0=\emptyset)\leqslant\\ &\leqslant\mathbb{P}(C_n(X_{\theta,1},\dots,X_{\theta,n})\cap\{\theta\}=\emptyset)=\mathbb{P}(\theta\notin C_n(X_{\theta,1},\dots,X_{\theta,n}))\leqslant\alpha_n,\end{aligned}$$

also ist T_n ein Test zum Niveau α. □

Bemerkung: Für jeden Konfidenzbereich für die Schätzung von θ können wir also den zugehörigen Test konstruieren. Natürlich gibt es viele Konfidenzbereiche, die man geschickt wählen sollte, damit zunächst einmal überhaupt nützliche, und dann möglichst mächtige Tests herauskommen - insbesondere haben wir in unseren Beispielen immer die rechte und linke Grenze durch α_1 und α_2 festgelegt, und bei der Konstruktion von Tests spielt es eine Rolle, wie man α_1 und α_2 wählt, je nachdem welches H_0 man vorliegen hat.

(5.30) Beispiele

a) Gaußtest, zweiseitig:

Das stochastische Modell sei $(X_{\theta,n})_{n\in\mathbb{N}}$ uiv mit $X_{\theta,1} \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 fest und bekannt, $\theta \in \Theta = \mathbb{R}$. Die Nullhypothese sei $H_0 = \{\theta_0\}$; relevant ist dies beispielsweise, wenn θ_0 die Lehrmeinung für eine naturwissenschaftliche Größe darstellt und man mittels verrauschter Messungen belegen will, dass diese Lehrmeinung falsch ist.

Analog zu (5.22) kann man das symmetrische Konfindenzintervall

$$C_n(x) = [M_n - \frac{\sigma \bar{q}_{1-\alpha/2}}{\sqrt{n}}, M_n + \frac{\sigma \bar{q}_{1-\alpha/2}}{\sqrt{n}}]$$

konstruieren, wobei $\bar{q}_\beta$ das β-Quantil von $\mathcal{N}_{0,1}$ ist. Wegen

$$C_n(x) \cap H_0 = \emptyset \quad \Longleftrightarrow \quad \theta_0 \notin C_n(x) \quad \Longleftrightarrow \quad |M_n(x) - \theta_0| > \frac{\sigma}{\sqrt{n}} \bar{q}_{1-\alpha/2}$$

ist durch

$$T_n(x) = \begin{cases} 1 & \text{falls } |M_n(x) - \theta_0| > \frac{\sigma}{\sqrt{n}} \bar{q}_{1-\alpha/2}, \\ 0 & \text{sonst.} \end{cases}$$

ein Test zu Niveau α gegeben.

b) Gaußtest, einseitig:

Das stochastische Modell und Θ sei wie in a), aber nun $H_0 = (-\infty, \theta_0]$. Eine Anwendung ist beispielsweise die Behauptung, das mittels $\theta \in \mathbb{R}$ gemessene und in der Bevölkerung (mit Varianz σ^2) normalverteilte individuelle Wohlbefinden werde durch Einnahme eines Superfoods über den in der Gesamtbevölkerung üblichen Wert θ_0 hinaus gesteigert, was man durch Messen dieses Wertes (wie auch immer) bei n Testpersonen, die das Produkt benutzen, zu belegen versucht.

Man kann hier zwar das gleiche Konfidenzintervall wie in a) wählen; man bekommt über die Bedingung $C_n(x) \cap H_0 = \emptyset$ den Test $T_n(x) = 1$ genau dann wenn $M_n(x) - \theta > \frac{\sigma}{\sqrt{n}} \bar{q}_{1-\alpha/2}$. Das ist aber nicht optimal, denn es gibt hier keinen Grund, das Konfindenzintervall nach oben zu beschränken; und je länger man es oben werden lässt, desto kürzer kann man es am unteren Ende machen. Man sollte also im Kontext von (5.22) $\alpha_1 = 0$ und $\alpha_2 = \alpha$ wählen, dann ist $\bar{q}_{\alpha_1} = -\infty$, und das Konfidenzintervall bekommt die Form

$$C_n(x) = [M_n - \frac{\sigma \bar{q}_{1-\alpha}}{\sqrt{n}}, \infty).$$

Damit ist durch

$$T_n(x) = \begin{cases} 1 & \text{falls } M_n(x) > \theta_0 + \frac{\sigma}{\sqrt{n}}\bar{q}_{1-\alpha}, \\ 0 & \text{sonst.} \end{cases}$$

ein Test zum Niveau α gegeben.

c) χ^2-Test, einseitig:

Das stochastische Modell ist $(X_{\theta,n})$ mit $(X_{\theta,n})_n$ uiv, $\theta = (m, v) \in \mathbb{R} \times \mathbb{R}_0^+ = \Theta$, und $X_{\theta,1} \sim \mathcal{N}_{m,v}$. Die Nullhypothese ist $H_0 = \{(m, v) : v \leqslant v_0\}$ für ein $v_0 > 0$. Eine Anwendung ist etwa die Behauptung, die Variabilität (Varianz) eines Merkmals einer Population von Individuen sei klein, woraus man schließen möchte, dass die Population seit langer Zeit isoliert lebt. In (5.25) haben wir das Konfidenzintervall

$$C_n(x) := \Big[\frac{(n-1)V_n(x)}{\bar{q}_{1-\alpha_2}}, \frac{(n-1)V_n(x)}{\bar{q}_{\alpha_1}}\Big]$$

für die Schätzung der Abbildung $h(\theta) = v$ ausgerechnet, wobei V_n der Varianzschätzer und $\bar{q}_\beta$ das β-Quantil der χ^2_{n-1}-Verteilung ist. Da uns die rechte Grenze dieses Intervalls wie oben nicht interessiert, setzen wir wieder $\alpha_2 = \alpha$ und $\alpha_1 = 0$, und erhalten den Test

$$T_n(x) = \begin{cases} 1 & \text{falls } V_n(x) > \frac{v_0\bar{q}_{1-\alpha}}{n-1}, \\ 0 & \text{sonst.} \end{cases}$$

Ein zweiseitiger Test geht natürlich auch, man überlege sich was man hier ändern muss.

d) t-Test, einseitig

Das statistische Modell sei wie in c), aber nun sei $H_0 = \{(m, v) : m \geqslant m_0\}$. Die Anwendung ist ähnlich wie in b), mit dem Unterschied, dass man nun nicht davon ausgeht, die Varianz a priori zu kennen. Außerdem ist nun die Nullhypothese ein in m von unten (statt von oben) beschränktes Gebiet, wir suchen also nach Konfidenzintervallen, die nach oben hin möglichst kurz sind. In (5.26) haben wir das Konfidenzintervall

$$C(x) = \Big[M_n(x) + \frac{\sqrt{V_n(x)}}{\sqrt{n}}\bar{q}_{1-\alpha_2}, M_n(x) + \frac{\sqrt{V_n(x)}}{\sqrt{n}}\bar{q}_{\alpha_1}\Big]$$

gefunden, wobei M_n und V_n Mittelwert- und Varainzschätzer, une $\bar{q}_\beta$ nun das Quantil der t_{n-1}-Verteilung ist. Wir setzen nun $\alpha_2 = 0$ und $\alpha_1 = \alpha$, und erhalten

$$T_n(x) = \begin{cases} 1 & \text{falls } M(x) + \frac{1}{\sqrt{n}}\sqrt{V_n(x)}\bar{q}_\alpha < m_0, \\ 0 & \text{sonst} \end{cases}$$

als Test zum Nivea α.

e) Binomialtest: Das stochastische Modell sei $(X_{\theta,n})$ mit $(X_{\theta,n})_{n\in\mathbb{N}}$ uiv, $\theta = p \in [0, 1] = \Theta$, und $X_{p,1} \sim \mathrm{Ber}_p$. Die Nullhypothese ist $H_0 = [0, p_0]$. Anwendungen sind „Experimente“, die mit einer W'keit von p gelingen, und man möchte diese W'keit testen. So ein Experiment kann zum Beispiel sein, einen Feuerwerkskörper einer Lieferung zu testen, und ein „Erfolg“ wäre dann, dass er nicht funktioniert. Für $p_0 = 0.03$ bedeutet die Nullhypothese, dass höchstens 3 Prozent aller gelieferten Feuerwerkskörper defekt sind.

Wir haben hier wieder einen einseitigen Mittelwert-Test, allerdings sind die Zufallsvariablen nun nicht mehr Gaußverteilt. Statt zu versuchen, den Test nach unserem Schema zu konstruieren,

gehen wir hier ad hoc vor. Die einzig sinnvolle Vorgehensweise scheint es, ab einer gewissen Anzahl n_0 von Erfolgen die Nullhypothese abzulehnen. Setzt man

$$T_n(x) := \begin{cases} 1 & \text{falls } \sum_{i-1}^n x_i \geqslant n_0, \\ 0 & \text{sonst} \end{cases}$$

so muss man noch n_0 bestimmen. Es ist klar dass der Fehler ester Art für $p = p_0$ unter allen $p \in [0, p_0]$ am größten ist, daher brauchen wir und diesen Fall ansehen.Es gilt

$$\mathbb{P}(T_n(X_{p_0,1}, \dots X_{p_0,n})) = 1) = \mathbb{P}(\sum_{i-1}^n X_{p_0,i} \geqslant n_0) = \sum_{k=n_0}^n \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

Das Auftreten der Binomialverteilung gibt dem Test seinen Namen.

Für $\mathbb{P}(T_n(X_{p_0,1}, \dots X_{p_0,n})) = 1) \leqslant \alpha$ muss man n_0 groß genug wählen. Im Fall $n = 100$, $\alpha = 0.05$ und $\theta_0 = 0.03$ erhält man $n_0 = 7$. Hat man also 10000 (oder eine andere beliebige sehr große Anzahl) Feuerwerkskörper gekauft, 100 getestet und dabei 7 Nieten gefunden, dann kann man zum Niveau 0.05 die Nullhypothese ablehnen, dass von den 10000 weniger als 3 Prozent fehlerhaft sind; bei nur 6 gefundenen Nieten kann man das nicht tun.

Eigentlich gilt das alles nur im Grenzwert unendlich großer Lieferung, denn ansonsten müsste man statt der Bionimalverteilung die wesentlich weniger erfreuliche hypergeometrische Verteilung nehmen. Zum Glück besteht für große n kaum ein Unterschied.

Bemerkungen:

a) Zur Unverfälschtheit:

Betrachten Sie den einseitigen Gaußtest im Beispiel b) oben. Nehmen wir an, dass wir vermuten, das das wahre θ nicht viel größer als θ_0 ist, und wir wollen unsere Chancen erhöhen, zum Niveau $\alpha = 0.05$ die Nullhypothese abzulehnen. Wir können den „Sicherheitsabstand“ $\sigma\bar{q}_{1-\alpha}/\sqrt{n}$ zwischen dem aus den Daten errechneten Wert $S_n(x)$ und θ_0 tatsächlich ein wenig reduzieren, indem wir die Nullhypothese statt dessen auch dann ablehnen, wenn die Werte von $S_n(x)$ *zu groß* sind, also oberhalb einer Zahl $c_{\max}$ liegen. Nehmen wir beispielsweise an, wir wollen sie auch noch bei Ergebnissen ablehnen, die nur um $\frac{1}{2}\sigma\bar{q}_{1-\alpha}/\sqrt{n}$ oberhalb von θ_0 liegen. Im kritischen Fall $\theta = \theta_0$, wo der Fehler erster Art am größten ist, hat dann die W'keit, die Nullhypothese abzulehnen, die Form

$$\mathbb{P}\Big(S_n(X_{\theta_0,1}, \dots, X_{\theta_0,n}) \in \big[\theta_0 + \frac{1}{2}\sigma\bar{q}_{1-\alpha}/\sqrt{n}, c_{\max}\big]\Big),$$

und durch geschickte Wahl von $c_{\max} > \theta_0 + \frac{1}{2}\sigma\bar{q}_{1-\alpha}/\sqrt{n}$ bekommt man diesen Ausdruck $\leqslant \alpha$.

Der Nachteil, den man sich erkauft, ist zwar, dass dieser Test nun im Fall, wo das wahre θ viel größer als θ_0 ist, die Nullhypothese mit ziemlicher Sicherheit nicht ablehnt obwohl sie falsch ist; wenn man aber sowieso vermutet dass das wahre θ nicht weit von θ_0 entfernt ist kann einem das egal sein. Trotzdem ist das ganze natürlich hochgradig willkürlich und ganz und gar nicht im Sinne des Erfinders. Daher fordert man die Unverfälschtheit bei Tests; diese erfüllt unser alternativer Test nicht, da ja seine Macht bei sehr großen Werten von θ beliebig klein wird.

b) Zum Betrug bei Tests:

Gerade weil es in der Praxis bei statistischen Tests oftmals um viel Geld oder Ansehen geht (beispielsweise bei Medikamenten), ist die Versuchung sehr groß, hier zu tricksen. Möglichkeiten

dafür gibt es viele: man kann versuchen, Daten verschwinden zu lassen, die nicht gut aussehen, oder man kann so viele Fragen gleichzeitig testen, dass man eine davon mit hoher Wahrscheinlichkeit positiv beantworten können wird, und dann in der Publikation so tun, als hätte man nur diese Frage untersucht. Eine andere Methode ist es, das Niveau α nachträglich noch mal zu ändern: zum Beispiel nehme man an, man wolle zum Niveau $\alpha = 0.05$ zeigen, dass das neueste Superfood das Wohlbefinden steigert; leider darf man aber aufgrund der Daten die Nullhypothese (keine Steigerung) zu diesem Niveau nicht ablehnen. Nun rechnet man nach und findet heraus, dass man sie zum Niveau $\alpha = 0.07$ aufgrund der gleichen Daten hätte ablehnen dürfen - man findet nun eine 93-prozentige Zuverlässigkeit der Aussage auch nicht viel schlechter als eine 95-prozentige, und ändert kurzerhand das Niveau. Schon kann man mit Fug und Recht behaupten, was man immer schon geglaubt hat.

Dieses letzte Beispiel zeigt ein Grundproblem bei Studien: zunächst ist ja tatsächlich ein Niveau von 0.07 auch nicht zu verachten, aus physikalischer Sicht ist die Vermutung immer noch angemessen, dass es hier etwas zu entdecken gibt. Andererseits ist dann natürlich nicht klar, wo die Grenze ist - bei 10 % oder doch erst bei 15 %? So kann man also nicht seriös vorgehen. Wenn andererseits ein naturwissenschaftlicher Effekt so subtil ist, dass man ihn nur mit größtem Aufwand und langen Studien entdecken kann, wie nützlich ist es dann, ihn zu kennen? Solche Fragen sind nicht immer leicht zu beantworten. Volle Berechtigung haben die Studien jedoch im Gebiet der gesetzlichen Regulierung, da man hier einfach klare und objektive Spielregeln braucht, die für alle gelten.

c) der p-Wert:

Oft wird bei Tests auch der p-Wert angegeben: dies ist die Abbildung, die zu gegebenen Daten $x_1, \dots, x_n$ das kleinste Niveau errechnet, bei dem man aufgrund dieser Daten die Nullhypothese gerade noch ablehnen würde. Im Fall des einseitigen Gaußtests aus (5.30 b) ist beispielsweise

$$p(x) = \min\{\alpha \in [0,1] : T_n(x) = 1\} = \min\{\alpha \in [0,1] : M(x) > \theta_0 + \frac{\sigma}{\sqrt{n}}\bar{q}_{1-\alpha}\} =$$

$$= \min\{\alpha \in [0,1] : (M(x) - \theta_0)\frac{\sqrt{n}}{\sigma} > \bar{q}_{1-\alpha}\}.$$

Da die Abbildung $\alpha \mapsto \bar{q}_{1-\alpha}$ monoton fallend ist, ist $p(x)$ somit die Lösung α der Gleichung

$$\bar{q}_{1-\alpha} = (M(x) - \theta_0)\frac{\sqrt{n}}{\sigma},$$

und daher

$$p(x) = \frac{1}{\sqrt{2\pi}} \int_{(M(x)-\theta_0)\sqrt{n}/\sigma}^{\infty} \mathrm{e}^{-\frac{z^2}{2}}\, \mathrm{d}z.$$

Man sieht, dass $p(x) = 1/2$ falls $M(x) = \theta_0$, und dass $p(x)$ relativ schnell schrumpft, wenn $M(x) - \theta_0$ groß wird. Im p-Wert steckt natürlich die gleiche Information wie im Test selbst: man erzeugt die Daten und vergleicht dann den errechneten p-Wert mit dem vorher festgelegten Niveau α; ist er kleiner als dieses α, so kann man die Nullhypothese ablehnen. Gleichzeitig sagt einem der p-Wert gleich auch noch, zu welchem noch kleineren Niveau der Test die Nullhypothese auch noch abgelehnt hätte, oder wie weit man das Niveau hätte heben müssen, um anhand der Daten die Nullhypothese ablehnen zu können. Letzteres ist natürlich im Sinne von b) sehr verlockend, aber nicht erlaubt.

(5.31) Statistik: Zusammenfassung uns Ausblick

Wir sind am Ende unseres kleinen Überblicks über die Statistik. Als letztes möchte ich einen kleinen Ausblick auf einige Aspekte geben, die es in diesem Gebiet sonst noch gibt. Da ich selbst kein Statistiker bin ist dieser Ausblick sehr wahrscheinlich stark unvollständig.

a) In unseren Beispielen für Schätzer, Tests und Konfidenzintervalle war ein gewisses Muster zu erkennen: immer wenn man für ein statistisches Modell $(X_{\theta,n})$ eine (meist von θ abhängige) Statistik $T_{n,\theta}$ finden kann, so dass man die Verteilung von $T_{n,\theta}(X_{\theta,1},\dots,X_{\theta,n})$ kennt und diese nicht von θ abhängt, dann kann man Tests und Konfidenzintervalle mit Hilfe dieser Statistik konstruieren. Dieses Thema kann man weiterverfolgen, und auf diese Weise Schätzer für den Median, Tests auf Unabhängigkeit und vieles mehr konstruieren. Wichtig sind auch die Fälle, wo die $(X_{\theta,n})$ nicht unabhängig sind und beispielsweise ein zeitliches Geschehen (wie etwa eine Markovkette) beschreiben; relevant ist für solche Situationen auch die sogenannte Zeitreihenanalyse.

b) Abgesehen davon beschäftigt sich die mathematische Statistik mit der Optimalität von Tests; man versucht hierbei zu beweisen, dass gewisse Tests für einige oder alle $\theta \in H_1$ eine größere Macht haben als alle anderen Tests des gleichen Niveaus. Ein anderes Thema ist die Konvergenzgeschwindigkeit (im Sinne der Konsistenz) bei Zunahme der Datenmenge - hier ist in den meisten Fällen der zentrale Grenzwertsatz relevant, was bedeutet, dass die bei der Konsistenz berechnete Größe oft wie $1/\sqrt{n}$ verschwindet.

c) Schließlich ist die Statistik allgemein eine eigene, angewandte Wissenschaft, die beispielsweise in England traditionell auch eigene Institute hat und nicht Teil der Mathematik ist. Zielsetzung ist es immer, aus großen oder riesigen Mengen von Daten möglichst sinnvolle Schlüsse zu ziehen, und wie in anderen Naturwissenschaften steht die mathematische Beweisbarkeit dieser Schlüsse oft nur an zweiter Stelle. Dies erklärt auch zum Teil die etwas eigenwillige Notation und Nomenklatur, die in der Statistik üblich ist. In den letzten Jahren hat die angewandte Statistik durch die Buzzwords „big data“ und „machine learning“ erheblichen Auftrieb erhalten.