Wahrscheinlich?

Manfred Jäger-Ambrożewicz
www.mathfred.de
www.mathstat.de
11. Juli 2026

Das Skript ist teilweise fragmentarisch und hat bestimmt etliche Mängel/Fehler. Auch die Reihenfolge/Gliederung ist noch nicht final. Über Hinweise würde ich mich freuen; z.B. per Email an jaegera ät htw-berlin Punkt de.

Das Skript genügt (noch?) nicht den Anforderungen, die an eine wissenschaftliche Arbeit gestellt werden, denn die Quellen zu vielen Aussagen werden (noch) nicht vor Ort angegeben.

Inhaltsverzeichnis

1 Basis der Wahrscheinlichkeitstheorie

1.1 Grundlegendes: Axiome für Ereignisse deren Wahrscheinlichkeiten

In diesen Kapitel werden einige Grundbegriffe und dazugehörige Resultat eingeführt. Die Erläuterungen, Definitionen und Ergebnisse sind so grundlegend, dass sie sich in nahezu jeder Quelle finden, z.B. in Billingsley [2], Blitzstein und Hwang [4], Georgii [13], Henze [19] und Krengel [24]. Aus meiner Sicht eignen sich vor allem Henze [19] und Blitzstein und Hwang [4].

1.1.1 Bemerkung: i.) In der Wahrscheinlichkeitstheorie wird ein Rahmen für die systematische mathematische Analyse von *unsicheren Sachverhalten* entwickelt. Für solche Sach-

verhalte verwenden wir den Begriff **Experiment**.

Die **Menge der möglichen Ergebnisse** des Experiments bezeichnen wir mit

$$\Omega \neq \emptyset \quad \text{ist die Menge der Ergebnisse.}$$

Ich stelle mir unter Ω regelmäßig die Menge der möglichen **präzisen Protokolle** vor, die ein gewissenhafter Protokollant anfertigen könnte.

Für ein bestimmtes **Ergebnis** $\omega \in \Omega$ verwenden wir typischerweise das Symbol

$$\omega \quad \text{ist ein typisches Ergebnis.}$$

Für **bestimmte Teilmengen**

$$A \subset \Omega$$

verwenden wir den Begriff **Ereignis**. Ereignisse sind also „zusammengesetzte“ Ergebnisse; Ereignisse sind in einer Menge zusammengefasste Ergebnisse.

ii.) Im Mittelpunkt der Wahrscheinlichkeitstheorie steht (natürlich) der Begriff der Wahrscheinlichkeit. Es sei $A \subset \Omega$ ein **Ereignis**. Wir geben zwei denkbare **Interpretationen** der Wahrscheinlichkeit $\mathbb{P}(A)$ von A an (vgl. z.B. Georgii (2015, Seite 14)).

- **Frequentistische Interpretation:** Wir stellen uns vor, dass ein Experiment *zu gleichen Bedingungen* unendlich oft wiederholt wird. Dann ist $\mathbb{P}(A)$ die relative Häufigkeit (Frequenz) mit der das Ereignis A *typischerweise* eintritt.

- **Grad der Gewissheit:** Die Wahrscheinlichkeit von A ist der **Grad der Gewissheit** mit der A eintritt. Wenn ich auf das Eintreten des Ereignisses A wetten würde, dann würde ich mich dabei an dem Wert $\mathbb{P}(A)$ orientieren.

c.) Für beide Interpretationen (sowie für andere Interpretationen) lassen sich gute Gründe bzw. passende Beispiele angeben, leider aber auch jeweils unpassende. Die Mathematik (in engeren Sinn) *funktioniert* unabhängig von der Interpretation, so dass wir uns mit der Interpretation *vorläufig* nicht beschäftigen müssen. Wir verwenden nämlich den sogenannten **axiomatischen Ansatz**, der die *vernünftigen* **Regeln für den Umgang mit Wahrscheinlichkeiten festlegt**, ohne eine Definition für Wahrscheinlichkeit anzugeben.[1] Bezogen auf die Wahrscheinlichkeitstheorie geht dieses Vorgehen auf den Mathematiker Andrei Kolmogorow zurück.[2]

Sie werden aber später im Studium und insbesondere in der

[1] Wahrscheinlichkeit ist dann ein sogenannter undefinierter Grundbegriff vergleichbar zu Punkt in der Geometrie.

[2] Quelle ....

Statistik feststellen, dass die Interpretationen des Begriffs der Wahrscheinlichkeit für die Entwicklung bzw. die Auswahl mathematischer Modelle ausschlaggebend sein kann; insbesondere für die Unterscheidung zwischen frequentistischer und bayesiansicher Statistik.

1.1.2 Beispiel: Wir betrachten das Experiment, das darin besteht, einen Würfel (mit sechs Seiten) zu werfen. Die **möglichen Ergebnisse** erfassen wir durch

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Dabei erfasst $\omega = 3$ den Sachverhalt, dass der Würfel so liegt, dass die Seite mit drei Augen oben liegt. Wir wählen die Potenzmenge $\mathcal{P}(\Omega)$ als **Menge der Ereignisse**. Das Ereignis "eine gerade Zahl wurde gewürfelt" wird durch $A = \{2, 4, 6\}$ repräsentiert. Das Ereignis „eine 1 oder eine 6 wurde gewürfelt, wird durch $B = \{1\} \cup \{6\} = \{1, 6\}$ repräsentiert. Wir beobachten hier, dass es eine Entsprechung zwischen der (normal-)sprachlichen Verwendung des Wortes „oder" und der Verwendung der Verknüpfung $\cup$ aus der Mengenlehre gibt. Das Ereignis „eine gerade Zahl und eine Zahl größer als 2" wird also durch $\{4, 6\} = \{2, 4, 6\} \cap \{3, 4, 5, 6\}$ repräsentiert. Die Verknüpfung $\cap$ hat also ihre Entsprechung im „und". Das Ereignis „eine gerade Zahl wurde nicht gewürfelt" wird durch $\{1, 3, 5\} = \{2, 4, 6\}^c$ repräsentiert. Das „nicht"

aus der Umgangssprache gehört also zum ^{c} der Mengenlehre.

1.1.3 Bemerkung: Die im vorhergehenden Beispiel gefundenen Entsprechungen gelten allgemein. Zu den bekannten Verknüpfungen von Aussagen gehören ebenfalls bekannte Verknüpfungen von Mengen (also Ereignissen).

- Wir haben die Paare:
  - $\cup$ gehört zum *oder*
  - $\cap$ gehört zum *und*
  - ^{c} gehört zum *nicht*

1.1.4 Beispiel: Angenommen Sie würfeln so lange, bis Sie schließlich eine 6 würfeln. Wir wollen davon ausgehen, dass Sie unter Umständen unendlich Lange würfeln, also trotz andauernder Versuche immer eine von 6 verschiedene Zahl würfeln. Etwaige praktische Einwände gegen die Unmöglichkeit $\infty-$vieler Würfe dürfen wir als Mathematiker*innen ignorieren. Die Ergebnismenge ist dann $\Omega = \{\infty, 1, 2, 3, ...\}$. An diesen Beispiel erkennen wir, dass wir auch Mengen Ω mit unendlich vielen Elementen betrachten sollen.

1.1.5 Definition: 1.) Es sei $\Omega \neq \emptyset$ eine Menge. Ein Teil-

mengensystem $\mathcal{F} \subset \mathcal{P}(\Omega)$ heißt **σ-Algebra** auf Ω, falls:

i.) $\Omega \in \mathcal{F}$.

ii.) $A \in \mathcal{F}$ impliziert $A^c := \Omega \setminus A \in \mathcal{F}$.

iii.) $A_i \in \mathcal{F}, i \in \mathbb{N}$ impliziert $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

2.) Es sei $\Omega \neq \emptyset$ eine Menge und $\mathcal{F}$ eine σ-Algebra auf Ω. Dann heißt das Paar $(\Omega, \mathcal{F})$ **messbarer Raum**.

3.) Es sei $(\Omega, \mathcal{F})$ ein messbarer Raum. Eine Abbildung

$$\mathbb{P} : \mathcal{F} \to [0, 1]$$

heißt **Wahrscheinlichkeitsmaß** auf $(\Omega, \mathcal{F})$, falls:

i.) $\mathbb{P}(\Omega) = 1$.

ii.) Für jede Folge $A_i, i \in \mathbb{N}$ von paarweise disjunkten Ereignisse $A_i \in \mathcal{F}$ gilt

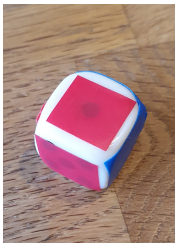
$$\mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

4.) Es sei $(\Omega, \mathcal{F})$ ein messbarer Raum und $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{F})$. Dann heißt das Triple $(\Omega, \mathcal{F}, \mathbb{P})$ **Wahrscheinlichkeitsraum**.

► Wäre $\Omega = \emptyset$, dann einerseits $\mathbb{P}(\emptyset) = \mathbb{P}(\Omega) = 1$. Andererseits $\mathbb{P}(\emptyset) = \mathbb{P}(\emptyset \uplus \emptyset \uplus \emptyset ...) = \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + ...$ Dann muss also andererseits $\mathbb{P}(\emptyset) = 0$ gelten. Die Voraussetzung $\Omega \neq \emptyset$

ist also notwendig, um den Widerspruch $0 = 1$ zu vermeiden.

1.1.6 Bemerkung: In der einführenden Bemerkung (1.1.1) haben wir (nahezu) beiläufig angemerkt, dass wir **nur bestimmte** Teilmengen von Ω **Ereignisse nennen**. Wir werden in der Tat nur die Elemente $A \in \mathcal{F}$ einer σ-Algebra $\mathcal{F}$ als Ereignis bezeichnen. Dafür gibt es zwei Gründe, von denen wir den ersten erst später – wenn wir etwas Maßtheorie besprochen haben – erläutern werden. Es gibt jedoch einen elementaren Grund, warum wir nicht alle Teilmengen als Ereignisse auffassen.



Wir verwenden die σ-Algebra $\mathcal{F}$ zur Erfassung des **Informationsstands**[3] (eventuell Zwischenstands) des Beobachters des Experiments nach dem Ende des Experiments. Für den Fall des Würfels: Die σ-Algebra $\mathcal{F} = \{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}$ erfasst, dass der Beobachter lediglich beobachtet, ob eine gerade Zahl oder eine ungerade Zahl gewürfelt wurde.

Um das zu veranschaulichen, betrachten wir den abgebildeten **abgeklebten Würfel** erklären.[4] Die ungeraden Zahlen sind rot abgeklebt und die geraden Zahlen blau. Man kann

[3] Wir beachten dabei den Hinweis von Billingsley (1995, S. 57 ff), dass wir streng zwischen mathematischen Aussagen und deren Interpretation unterscheiden müssen.

[4] Diese Illustration ist originär- :-).

die Zahlen unter dem Klebeband erkennen, wir sollen/wollen uns aber vorstellen, dass wir die Zahlen unter dem Klebeband nicht erkennen können. Wir wissen lediglich, dass sich unter rot eine ungerade und unter blau eine gerade Augenzahl verbirgt.

Angenommen eine 3 wurde gewürfelt. Der Beobachter sieht *rot*. Der Beobachter mit der **Teilinformation** $\mathcal{F}$ weiß nicht, ob eine 3, eine 1 bzw. eine 5 gewürfelt wurde. Er weiß, dass $A = \{1, 3, 5\}$ eingetreten ist. Die $\omega \in A$ sind alle rot abgeklebt und nicht unterscheidbar.

1.1.7 Bemerkung: Die folgende Übersicht fasst noch mal die Begriffe zusammen

- Ω ist die **Menge der möglichen Ergebnisse (Protokolle)**. Die Elemente von Ω werden mit ω bezeichnet. Ein ω ist also **ein mögliches Ergebnis (Protokoll)**. Wenn man WT mit Blick auf die Anwendung in der Statistik betreibt, dann nennt man Ω manchmal auch den **Stichprobenraum**.

- Das Teilmengensystem $\mathcal{F}$ der messbaren Mengen bildet eine **σ-Algebra**. Die Mengen $A \in \mathcal{F}$ heißen **Ereignisse**; und nur diese Teilmengen $A \in \mathcal{F}$ von Ω heißen Ereignisse. A ist also die Zusammenfassung von Ergebnissen zu einer eine Menge.

- Auf der Menge $\mathcal{F}$ definieren wir das **Wahrscheinlichkeitsmaß** $\mathbb{P}$. Der Definitionsbereich von $\mathbb{P}$ ist also $\mathcal{F}$ und nicht Ω. Die Zielmenge von $\mathbb{P}$ ist $[0,1]$. Also

$$\mathbb{P} : \mathcal{F} \to [0,1].$$

Die charakteristischen Eigenschaften eines Wahrscheinlichkeitsmaßes sind:

i.) $\mathbb{P}(\Omega) = 1$

ii.) $\mathbb{P}$ ist σ-additiv.

1.2 Erste Regeln

1.2.1 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann gilt:

1.) $\mathbb{P}(\emptyset) = 0$.

2.) Für disjunkte[5] $A, B \in \mathcal{F}$ gilt:

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

[5]Disjunkt bedeutet, dass sich die Ereignisse ausschließen. Also $A \cap B = \emptyset$

3.) Für alle $A \in \mathcal{F}$ gilt

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

4.) Für alle $A, B \in \mathcal{F}$ mit $A \subset B$ gilt:

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

5.) Für alle $A, B \in \mathcal{F}$ gilt:

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

6.) Für alle $A, B \in \mathcal{F}$ gilt:

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

7.) Für $A, B \in \mathcal{F}$ mit $A \subset B$ gilt

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

Beweis: Zu 1.) Es gilt $\mathbb{P}(\emptyset) = \mathbb{P}(\emptyset \uplus \emptyset \uplus ...) = \sum_{i=1}^{\infty} \mathbb{P}(\emptyset)$. Wäre $\mathbb{P}(\emptyset) = \varepsilon > 0$, dann würde $\sum_{i=1}^{\infty} \mathbb{P}(\emptyset)$ divergieren und nicht gleich ε. Nur für $\mathbb{P}(\emptyset) = 0$ kann $\mathbb{P}(\emptyset) = \sum_{i=1}^{\infty} \mathbb{P}(\emptyset)$ gelten.

Zu 2.) Es gilt $\mathbb{P}(A \cup B \cup \emptyset \cup \emptyset ...) = \mathbb{P}(A) + \mathbb{P}(B) + 0 + 0 + ... = \mathbb{P}(A) + \mathbb{P}(B)$.

Zu 3.) Es gilt $A \uplus A^c = \Omega$. Also $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \uplus A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$.

Zu 4.) Es gilt $A \uplus (B \setminus A) = B$ (beachte: $A \subset B$ wird genutzt!). Also $\mathbb{P}(B) = \mathbb{P}(A \uplus (B \setminus A)) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$.

Zu 5.) Es gilt $A \cup B = (A \setminus (A \cap B)) \uplus (B \setminus (A \cap B)) \uplus (A \cap B))$. DIY

Zu 6.) Es gilt $A = A \cap \Omega = A \cap (B \uplus B^c) = (A \cap B) \uplus (A \cap B^c)$. Also $\mathbb{P}(A) = \mathbb{P}((A \cap B) \uplus (A \cap B^c)) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$.

1.2.2 Satz: Für jede Folge $A_i \in \mathcal{F}, i \in \mathbb{N}$ von Ereignisse in $(\Omega, \mathcal{F})$ gilt die Sub-σ-Additivität

$$\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} A_i\right) \leq \sum_{i \in \mathbb{N}} \mathbb{P}(A_i).$$

1.3 Laplace

1.3.1 Definition: Ein $(\Omega, \mathcal{F}, \mathbb{P})$ heißt **Laplace Experiment** oder **Laplace Wahrscheinlichkeitsraum**, falls:

i.) $\Omega \neq \emptyset$ ist eine endliche Menge mit n Elementen.

ii.) $\mathcal{F} = \mathcal{P}(\Omega)$.

iii.) Für alle $A \in \mathcal{F}$ ist

$$\mathbb{P}(A) = \frac{\#(A)}{\#(\Omega)} = \frac{\#(A)}{n} = \frac{\text{Anzahl der für } A \text{ günstigen Ergebnisse}}{\text{Anzahl aller Ergebnisse}}.$$

1.4 Bedingte Wahrscheinlichkeiten und Unabhängigkeit

1.4.1 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{F}$ 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**.

1.4.2 Satz: Es sei $B \in \mathcal{F}$ ein Ereignis mit $\mathbb{P}(B) > 0$. Dann ist die Funktion

$$\mathbb{P}^B : \begin{cases} \mathcal{F} \to [0,1] \\ A \mapsto \mathbb{P}(A|B) \end{cases}$$

ein (weiteres) Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{F})$.

1.4.3 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{F}$. A und B heißen **unabhängig**, falls

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

1.4.4 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{F}$ mit $\mathbb{P}(B) > 0$. Dann gilt: A und B sind genau dann unabhängig, wenn $\mathbb{P}(A|B) = \mathbb{P}(A)$ gilt.

1.4.5 Bemerkung: Die Konzepte disjunkt bzw. unabhängig werden von Anfängern leider oft verwechselt. Dabei bedeuten die Konzepte etwas grundsätzlich verschiedenes. Unabhängig bedeutet, dass wir die Wahrscheinlichkeitseinschätzung für A bei Eintreten von B nicht revidieren. Wenn A und B disjunkt sind, dann revidiert man die Wahrscheinlichkeitseinschätzung aber sehr wohl (jedenfalls typischerweise). Wenn das Eintreten von B, das Eintreten von A ausschließt, so ist $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B) = \mathbb{P}(\emptyset)/\mathbb{P}(B) = 0$. Wenn also $\mathbb{P}(A) > 0$ ist, dann wird die Wahrscheinlichkeitseinschätzung *radikal* revidiert. Die Eigenschaft der Disjunktheit bzw. des sich Ausschließens, stellt also **eine extreme Form der Abhängigkeit dar**.

1.4.6 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{F}$. Wenn A und B unabhängig sind, dann sind auch A und B^c unabhängig.

1.4.7 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und I eine abzählbare Indexmenge. Eine Folge $B_i \in \mathcal{F}, i \in I$ heißt **Nullmengenlose abzählbare Partition**

von Ω, falls:

i.) Die Menge B_i sind paarweise disjunkt.

ii.) Für alle $i \in I$ gilt $\mathbb{P}(B_i) > 0$.

iii.) Es gilt

$$\Omega = \biguplus_{i \in I} B_i.$$

1.4.8 Satz von der totalen Wahrscheinlichkeit oder Cocktailformel oder Fallunterscheidungsformel: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $B_i, i \in I$ eine **Nullmengenlose abzählbare Partition** von Ω. Dann gilt für alle $A \in \mathcal{F}$

$$\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

[Wie viel Vitamin C ist in B_i und wie viel B_i ist in A.]

1.4.9 Satz von Bayes: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $B_i, i \in I$ eine **Nullmengenlose abzählbare Partition** von Ω. Dann gilt für alle $A \in \mathcal{F}$ mit $\mathbb{P}(A) > 0$

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i \in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i)} = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)}.$$

Für $\Omega = B \uplus B^c$ erhalten wir insbesondere

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}.$$

1.4.10 Definiton: Es sei I eine nichtleere Indexmenge und $A_i \in \mathcal{F}, i \in I$ eine Familie von Ereignissen.

i.) Die Familie $A_i \in \mathcal{F}, i \in I$ heißt **unabhängig**, wenn für alle endlichen $J \subset I$

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

ii.) Die Ereignisse heißen **paarweise unabhängig**, wenn für alle $i, j \in I, i \neq j$ die Ereignisse A_i und A_j unabhängig sind.

1.4.11 Bemerkung: Für den Nachweis, dass A, B, C unabhängig sind, müssen die folgenden Gleichungen verifiziert werden:

$$\begin{aligned}
\mathbb{P}(A \cap B \cap C) &= \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C),\\
\mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B),\\
\mathbb{P}(B \cap C) &= \mathbb{P}(B)\mathbb{P}(C),\\
\mathbb{P}(A \cap C) &= \mathbb{P}(A)\mathbb{P}(C)
\end{aligned}$$

Für die Unabhängigkeit ist weder $\mathbb{P}(A\cap B\cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ noch die paarweise Unabhängigkeit $\mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B), \mathbb{P}(B\cap C) = \mathbb{P}(B)\mathbb{P}(C), \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C)$ hinreichend.

▶ Vgl. die entsprechenden Aufgaben ▶ Für das Konzept der bedingten Unabhängigkeit Vgl. Blitzstein und Hwang [4]

1.4.12 Satz (Multiplikationsformel): Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A_1, A_2, ..., A_n \in \mathcal{F}$ mit $\mathbb{P}(A_1 \cap A_2 \cap ... \cap A_{n-1}) > 0$. Dann gilt

$$\begin{aligned}
&\mathbb{P}(A_1 \cap A_2 \cap ... \cap A_n) \\
=&\mathbb{P}(A_n|A_1 \cap ... \cap A_{n-1}) \cdot \\
&\mathbb{P}(A_{n-1}|A_1 \cap ... \cap A_{n-2}) \cdot \\
&\cdot ... \\
&\cdot \mathbb{P}(A_2|A_1) ... \\
&\cdot \mathbb{P}(A_1)
\end{aligned}$$

1.4.13 Satz (Unabhängige Kopplung / Produktexperimente, diskreter Fall, Krengel [24, S. 27f]): Gegeben sind für $n \geq 2$

- n Experimente $(\Omega_1, \mathcal{P}(\Omega_1), \mathbb{P}_1), ..., (\Omega_n, \mathcal{P}(\Omega_n), \mathbb{P}_n)$ mit abzählbaren nicht-leeren Mengen Ω_i.

Auf diesen *separaten* Räumen können wir keine gemeinsame Wahrscheinlichkeits*modellierung* betreiben. Wir wollen **einen gemeinsamen Raum** für diese Experimente definieren, **der deren unabhängige Durchführung erfasst**. Man spricht von der **unabhängigen Kopplung**.

Wir definieren

$$\begin{aligned}\Omega &= \Omega_1 \times ... \times \Omega_n,\\ \mathbb{P}(\{\omega\}) &= \prod_{i=1}^{n} \mathbb{P}_i(\{\omega_i\}), \quad \omega = (\omega_1, ..., \omega_n),\\ \mathbb{P}(A) &= \sum_{\omega \in A} \mathbb{P}(\omega).\end{aligned}$$

Dann ist $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein Wahrscheinlichkeitsraum, d.h. durch die Vorgabe der Ereigniswahrscheinlichkeiten ergibt sich ein Wahrscheinlichkeitsmaß.

Für diesen Wahrscheinlichkeitsraum gilt dann

$$\begin{aligned}&\mathbb{P}(\bigcap_{i=1}^{n}\{\Omega_1 \times ... \times A_i \times ... \times \Omega_n\})\\ &= \prod_{i=1}^{n} \mathbb{P}(\Omega_1 \times ... \times A_i \times ... \times \Omega_n)\end{aligned}$$

oder mit der Notation (S^{A_i} für Streifen A_i)

$$S^{A_i} = \Omega_1 \times ... \times A_i \times ... \times \Omega_n$$

die Gleichung

$$\mathbb{P}(\bigcap_{i=1}^{n} S^{A_i}) = \prod_{i=1}^{n} \mathbb{P}(S^{A_i})$$

Also: Die Wahrscheinlichkeit $\mathbb{P}(\bigcap_{i=1}^{n} S^{A_i})$, dass im 1-ten Teilexperiment A_1 **und** im 2-ten Teilexperiment in A_2 **und** ...,

die n-ten Teilexperiment A_n eintritt, ist gleich dem **Produkt** $\prod_{i=1}^{n} \mathbb{P}(A_i \text{ im } i\text{-ten Teilexperiment})$ der Wahrscheinlichkeiten $\mathbb{P}(A_i \text{ im } i\text{-ten Teilexperiment})$. Da man dabei auch $A_i = \Omega_i, i \notin J, J \subset \{1, ..., n\}$ wählen kann, sind die **Teilereignisse** $\Omega_1 \times \Omega_2 \times ... \times A_i \times ... \times \Omega_n$ **unabhängig** (im Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$) im Sinn der Definition (1.4.10). Diese Beobachtung rechtfertigt (endgültig) die Bezeichnung **unabhängige Kopplung** für dieses Wahrscheinlichkeitsmodell. In der Tat erhält man *nur so* einen Wahrscheinlichkeitsraum, so dass

- $\mathbb{P}(\Omega_1 \times \Omega_2 \times ... \times A_i \times ... \times \Omega_n) = \mathbb{P}_i(A_i)$ und
- $\Omega_1 \times \Omega_2 \times ... \times A_i \times ... \times \Omega_n$ bilden eine unabhängige Familie.

Der so definierte Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ heißt **Produktwahrscheinlichkeitsraum** der Experimente $(\Omega_i, \mathcal{P}(\Omega_i), \mathbb{P}_i)$ und beschreibt die **unabhängige Kopplung** der Experimente $(\Omega_i, \mathcal{P}(\Omega_i), \mathbb{P}_i)$.

Die obigen Aussagen/Behauptungen werden im folgenden er-

läutert/belegt. Für $A_i \in \Omega_i$ ist

$$\begin{aligned}
\mathbb{P}(A_1 \times A_2 \times ... \times A_n) &= \sum_{\omega \in A_1 \times A_2 \times ... \times A_n} \mathbb{P}(\{\omega\}) \\
&= \sum_{\omega \in A_1 \times A_2 \times ... \times A_n} \mathbb{P}_1(\{\omega_1\}) \cdot ... \cdot \mathbb{P}_n(\{\omega_n\}) \\
&= \sum_{\omega_1 \in A_1} ... \sum_{\omega_n \in A_n} \mathbb{P}_1(\{\omega_1\}) \cdot ... \cdot \mathbb{P}_n(\{\omega_n\}) \\
&= \left(\sum_{\omega_1 \in A_1} \mathbb{P}_1(\{\omega_1\}) \right) \cdot ... \cdot \left(\sum_{\omega_n \in A_n} \mathbb{P}_n(\{\omega_n\}) \right) \\
&= \prod_{i=1}^{n} \mathbb{P}_i(A_i).
\end{aligned}$$

Also

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

Diese Gleichung *sieht so aus wie* eine entsprechende Gleichung aus der Definition der Unabhängigkeit. **Aber Vorsicht:** Auf der rechten Seite stehen die $\mathbb{P}_i$ und auf der linken Seite $\mathbb{P}$; anders als in der Definition für Unabhängigkeit.

Wir beobachten jetzt (so dass links und rechts $\mathbb{P}$ stehen wird):

$$\begin{aligned}\mathbb{P}(\bigcap_{i=1}^{n}\{\Omega_1 \times ... \times A_i \times ... \times \Omega_n\}) &= \mathbb{P}(A_1 \times ... \times A_n) \\ &= \prod_{i=1}^{n} \mathbb{P}_i(A_i) \\ &= \prod_{i=1}^{n} \mathbb{P}_1(\Omega_1) \cdot ... \cdot \mathbb{P}_i(A_i) \cdot ... \cdot \mathbb{P}_n(\Omega_n) \\ &= \prod_{i=1}^{n} \mathbb{P}(\Omega_1 \times ... \times A_i \times ... \times \Omega_n)\end{aligned}$$

Also

$$\begin{aligned}&\mathbb{P}(\bigcap_{i=1}^{n}\{\Omega_1 \times ... \times A_i \times ... \times \Omega_n\}) \\ &= \prod_{i=1}^{n} \mathbb{P}(\Omega_1 \times ... \times A_i \times ... \times \Omega_n)\end{aligned}$$

Beachte, dass es $\mathbb{P}(A_i)$ nicht definiert ist; deshalb die unübersichtliche rechte Seite.

1.4.14 Satz (Konstruktion mehrstufiger Modelle / Bedingte Kopplung): Gegeben sind für $n \geq 2$

- n abzählbare nicht-leere Ergebnismengen $\Omega_1, ..., \Omega_n$.
- Eine Zähldichte $\mathbb{P}_1$ auf Ω_1.
- Für alle $k = 2, ..., n$ und beliebige $\omega_i \in \Omega_i, i = 1, ..., k-$

1 seien (bedingte) **Zähldichten** $\omega_k \mapsto \mathbb{P}_k(\omega_k|\omega_1, ..., \omega_{k-1})$ auf Ω_k gegeben.

Wir definieren:

$$\begin{aligned}\Omega &= \Omega_1 \times ... \times \Omega_n,\\ \mathbb{P}(\omega) &= \mathbb{P}_1(\omega_1) \cdot \mathbb{P}_2(\omega_2|\omega_1) \cdot ... \cdot \mathbb{P}_n(\omega_n|\omega_1, ..., \omega_{n-1}),\\ \mathbb{P}(A) &= \sum_{\omega \in A} \mathbb{P}(\omega).\end{aligned}$$

Dann ist $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein Wahrscheinlichkeitsraum.

Dieser Wahrscheinlichkeitsraum heißt **mehrstufige bedingte Kopplung** der Räume $(\Omega_k, \mathcal{P}(\Omega_k), \mathbb{P}_k), k = 1, ..., n$.

$(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ist der einzige Wahrscheinlichkeitsraum mit: in Worten ...

$\mathbb{P}(\{\eta_1\} \times \Omega_2 \times ... \times \Omega_n) = \mathbb{P}_1(\eta_1)$ und

$\mathbb{P}(\Omega_1 \times \Omega_2 \times ... \times \{\eta_j\} \times ... \times \Omega_n \quad | \quad \{\eta_1\} \times \Omega_2 \times ... \times \Omega_n \cap$

$... \cap \ \Omega_1 \times \Omega_2 \times ... \times \{\eta_{j-1}\} \times ... \times \Omega_n)$

$= \mathbb{P}_j(\eta_j|\eta_1, ..., \eta_{j-1})$ falls

$\mathbb{P}(\{\eta_1\} \times \Omega_2 \times ... \times \Omega_n \cap \quad ... \cap \Omega_1 \times \Omega_2 \times ... \times \{\eta_{j-1}\} \times ... \times \Omega_n) > 0$

▶ Wir beachten, dass die Ergebnismengen für z.B. die zweite Stufe für alle Ergebnisse auf der ersten Stufe gleich sind (obwohl sich die Wahrscheinlichkeiten natürlich ändern können). Das sieht nach einer merklichen Einschränkung der Allgemeinheit aus. Wir werden in Beispielen sehen, dass dem nicht

so ist (durch das Hinzufügen von Ergebnissen mit Wahrscheinlichkeit Null).

1.5 Unendlich oft

1.5.1 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A_i \in \mathcal{F}, i \in \mathbb{N}$ und $A \in \mathcal{F}$. Wir definieren den **von-unten-Limes** $A_i \nearrow A$, wenn für alle $i \in \mathbb{N}$ $A_i \subset A_{i+1}$ und $A = \bigcup_{i\in\mathbb{N}} A_i$. Wir definieren den **von-oben-Limes** $A_i \searrow A$, wenn für alle $i \in \mathbb{N}$ $A_i \supset A_{i+1}$ und $A = \bigcap_{i\in\mathbb{N}} A_i$.

In beiden Fällen schreiben wir dann $A = \lim A_i$, wenn die *Art* des Limes (also $\searrow$ bzw. $\nearrow$) aus dem Zusammenhang eindeutig hervor geht.

1.5.2 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A_i \in \mathcal{F}, i \in \mathbb{N}$ und $A \in \mathcal{F}$. Dann gilt:

i.) $A_i \nearrow A$ impliziert $\mathbb{P}(A_i) \to \mathbb{P}(A)$ (also $\mathbb{P}(\lim A_i) = \lim \mathbb{P}(A_i)$) (Stetigkeit von unten)
ii.) $A_i \searrow A$ impliziert $\mathbb{P}(A_i) \to \mathbb{P}(A)$ (also $\mathbb{P}(\lim A_i) = \lim \mathbb{P}(A_i)$) (Stetigkeit von oben).

Beweis: Henze [19, S. 24].

1.5.3 Bemerkung: Die **Stetigkeitseigenschaft ist keine technische Randnotiz**. Angenommen wir betrachten das

Experiment des fortgesetzten fairen Würfelns. Wir interessieren uns für das Ereignis, dass wir keine 6 würfeln. A_i sei das Ereignis, dass beim i-ten Wurf keine 6 gewürfelt wird. Es gilt natürlich $\mathbb{P}(A_i) = 5/6$. Wir wollen auch annehmen, dass die Ereignisse A_i unabhängig sind. Dann gilt $n \in \mathbb{N}$

$$\mathbb{P}(A_1 \cap A_2 \cap ... \cap A_n) = \left(\frac{5}{6}\right)^n.$$

Für $n \to \infty$ konvergiert diese Wahrscheinlichkeit gegen Null. Bedeutet das, dass die Wahrscheinlichkeit

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \mathbb{P}(\text{"nie eine 6"})$$

nie eine 6 zu würfeln, Null ist? Wegen der bewiesenen Stetigkeit können wir uns davon überzeugen: Wir definieren $B_n = \bigcap_{i=1}^{n} A_i, B = \bigcap_{i=1}^{\infty} A_i$. Es gilt $B_n \supset B_{n+1}$ und $B = \bigcap_{i\in\mathbb{N}} B_i$.

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\lim B_i\right) = \lim \mathbb{P}(B_i) = 0.$$

Es ist *beruhigend*, dass die unterstellten Eigenschaften eines Wahrscheinlichkeitsmaßes tatsächlich die Stetigkeit von Maßen implizieren.

1.5.4 Satz (Stetigkeit und σ-additiv): Es sei $\mathcal{F}$ eine $\sigma-$Algebra und $\mathbb{P} : \mathcal{F} \to [0, 1]$ eine Abbildung mit $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1$ und $\mathbb{P}(A \uplus B) = \mathbb{P}(A)+\mathbb{P}(B), A, B \in \mathcal{F}$ disjunkt.

Dann gilt: $\mathbb{P}$ ist genau dann $\sigma-$additiv, wenn $\mathbb{P}$ stetig ist.

Beweis: .....

1.5.5 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A_i \in \mathcal{F}, i \in \mathbb{N}$. Dann definieren wir:

$$\begin{aligned}
\limsup A_i &:= \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k = \{\omega | \text{ für unendlich viele } i \text{ gilt } \omega \in A_i\} \\
&= \text{„unendlich viele der Ereignisse } A_i \text{ treten ein”} \\
&=: \infty\text{-oft } A_i \\
\liminf A_i &:= \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k = \{\omega | \text{ für alle bis auf endlich viele } i \text{ ist } \omega \in A_i\} \\
&= \text{„alle bis auf endlich viele der Ereignisse } A_i \text{ treten ein”} \\
&=: \text{schließlich alle } A_i
\end{aligned}$$

1.5.6 Satz (Erstes Lemma von Borel & Cantelli): Wenn $\sum_{i=1}^{\infty} \mathbb{P}(A_i)$ konvergiert, dann $\mathbb{P}(\limsup A_i) = 0$. M.a.W: Wenn $\sum_{i=1}^{\infty} \mathbb{P}(A_i)$ konvergiert, dann ist die Wahrscheinlichkeit, dass unendlich viele A_i eintreten, Null.

Beweis: Vgl. Henze [19, S. 64].

1.5.7 Satz (Zweites Lemma von Borel & Cantelli): Wenn A_i eine **unabhängige** Folge von Ereignisse ist und $\sum_{i=1}^{\infty} \mathbb{P}(A_i)$ divergiert, dann $\mathbb{P}(\limsup A_i) = 1$. M.a.W: Wenn

für unabhängige A_i die Reihe $\sum_{i=1}^{\infty} \mathbb{P}(A_i)$ divergiert, dann ist die Wahrscheinlichkeit, dass unendlich viele A_i eintreten, Eins.

Beweis: Vgl. Henze [19, S. 64].

1.5.8 Satz (Ein einfaches 0-1 Gesetz): Wenn A_i eine unabhängige Folge von Ereignisse ist, dann gilt $\mathbb{P}(\limsup A_i) = 1$ oder $\mathbb{P}(\limsup A_i) = 0$.

1.6 Wetten und Odds

1.6.1 Definition: Bei einer Wette auf das Ergebnis ω^* mit $odds = \alpha$ und einem Wetteinsatz von 1 ist der Gewinn

$$X = \begin{cases} \alpha - 1, & \text{falls } \omega = \omega^* \\ -1, & \text{sonst} \end{cases}$$

Die Wahrscheinlichkeit $\tilde{p}_{\omega^*}$. so dass

$$\begin{aligned} & \tilde{p}_{\omega^*}(\alpha - 1) + (1 - \tilde{p}_{\omega^*})(-1) = 0 \\ \Leftrightarrow \quad & \tilde{p}_{\omega^*}\alpha - 1 = 0 \\ \Leftrightarrow \quad & \tilde{p}_{\omega^*} = \frac{1}{\alpha} \end{aligned}$$

heißt **implizite Wahrscheinlichkeit** von ω^*. Das ist die *theoretische* Wahrscheinlichkeit, so dass odds $= \alpha$ *fair* ist.

1.6.2 Proposition: Wir betrachten eine Wette mit odds $= \alpha$. Es sei $p = \mathbb{P}(\{\omega^*\})$. Der erwartete Gewinn ist

$$\mathbb{E}(X) = p(\alpha - 1) + (1 - p)(-1) = ... = \frac{p - \tilde{p}_{\omega^*}}{p}.$$

Der erwartete Gewinn ist genau dann Null, wenn $p = \tilde{p}_{\omega^*} = \frac{1}{\alpha}$ ist.

1.6.3 Proposition: Wenn man $\tilde{p}_\omega$ für alle ω bestimmt, dann ist typischerweise $\sum_\omega p_\omega > 1$.

Dann definieren wir

$$\tilde{q}_{\omega^*} = \frac{\tilde{q}_{\omega^*}}{\sum_\omega p_\omega}$$

Diese Wahrscheinlichkeiten fassen wir als Wahrscheinlichkeitseinschätzung der Wettanbieter auf. .....

1.7 Exkurs Kombinatrik

Meine Favoriten zum Thema Kombinatorik sind Mazur [28] und Biggs [3].

1.7.1 Bemerkung: Insbesondere für die Analyse von Laplace Experimenten benötigt man Ergebnisse aus der Kombinatorik. Wir betrachten die folgenden Regeln:

1. Die Anzahl der Möglichkeiten n unterschiedliche Objekte anzuordnen ist $n!$. Hier bezeichnet

$$n! = n \cdot (n-1) \cdot ... \cdot 2 \cdot 1$$

n **Fakultät**.

Für den ersten Platz gibt es n Möglichkeiten. Wenn der erste Platz bestimmt ist, dann gibt es für den zweiten Platz $(n-1)$ Möglichkeiten .....

2. Wir betrachten wieder die Anzahl der Möglichkeiten n Objekten anzuorden. Allerdings können wir bei den Anordnungen n_1 Objekte der ersten Art, n_2 Objekte der zweiten Art, ..., n_K Objekte der k-ten Art nicht unterscheiden. Die Anzahl derartiger unterscheidbarer Anordnungen ist $\frac{n!}{n_1!n_2!\cdot...\cdot n_k!}$; der Koeffizient

$$\binom{n}{n_1, ..., n_K} := \frac{n!}{n_1!n_2!...n_K!}$$

heißt **Multinomialkoeffizient**. Beachte, dass $\sum_{k=1}^{K} n_k = n$ gilt.

Angenommen wir haben 4 rote Kugel, 3 grüne Kugel

und 2 blaue Kugeln. Zunächst stellen wir uns vor, dass wir die Kugel markieren, so dass wir die roten Kugel, die grünen Kugel und die blauen Kugel doch unterscheiden können. Wir haben also 9 unterscheidbare Kugeln. Wir wissen schon, dass es 9! mögliche Anordnungen gibt. Angenommen wir haben die Kugeln in einer konkreten Reihenfolge angeordnet. Wenn wir jetzt die roten Kugeln untereinander neu anordnen, dann erhalten wir bei nicht unterscheidbaren Kugeln **keine** neue Anordnung. Wir müssen für diese Zuviel-Zählung korrigieren. Das geschieht indem man durch 4!3!2! dividiert.

Für den Multinomialkoeffizienten gibt es eine weitere **Interpretation**. Wir wollen n Bälle **färben** und haben K Farben. Dann ist $\binom{n}{n_1,\ldots,n_K}$ die Anzahl der Möglichkeiten n Kugeln so zu färben, dass n_1 die Farbe 1, ..., n_K die Farbe K bekommen. Noch eine sehr ähnliche **Interpretation**: Es gibt K (unterscheidbare) Behälter auf die n Bälle **verteilt** werden sollen; vgl. Mazur [28, Seite 142] oder Biggs [3, Chapter 12].[6]

Warum ist das so? Wir stellen uns vor, dass die Behälter die Farbe flüssig enthalten und nebeneinander aufgereiht sind. Wir stellen uns auch vor, dass die Ku-

[6] Wenn die Behälter nicht unterschieden werden können, dann erhält man Partitionen (in Mengen/Teams). Hier besteht eine Verwechselungsgefahr; vgl. Mazur [28, 145] oder Biggs [3, Chapter 12].

gel nummeriert sind. Wir legen jetzt die Kugeln so vor den Behältern abm dass n_1 vor dem ersten Behälter liegen, n_2 vor dem zweiten Behälter liegen, ... Es gibt für die Anordnung $n!$ Möglichkeiten. Bezogen auf die Färbung bzw. Verteilung der Kugel im ersten bzw. in den Behälter, kommt es aber nicht auf die Anordnung der n_1 Kugeln vor dem ersten Behälter an. Analog für die Kugeln vor den anderen Behältern. Wir müssen demnahc wieder korrigieren und erreichen das, in dem wir durch $n_1!n_2! \cdot ... \cdot n_K!$ dividieren.

3. **Partitionen** mit Vorgaben. Wenn die Behälter nicht unterschieden werden können, dann erhält man Partitionen (in Mengen/Teams). Hier besteht eine Verwechselungsgefahr; vgl. Mazur [28, 145] oder Biggs [3, Chapter 12].

$$\frac{\binom{n}{n_1,...,n_K}}{p_1! \cdot ... \cdot p_n!}$$

Dabei bezeichnet p_1 die Anzahl der Teams mit einer Kugel, p_2 die Anzahl der Teams mit zwei Kugeln, ...

Warum ist das so? Warum muss man durch $p_1! \cdot ... \cdot p_n!$ dividieren? Dazu stellen wir uns vor, dass die Behälter vorab so sortiert sind, dass $n_1 \leq n_2 \leq ... \leq n_K$ gilt. Zuerst kommen die p_1 Behälter, die nur einen Ball erhalten sollen. Dann die p_2 Behälter, die zwei Bälle enthalten

sollen, ... Dann ordnen wir die Bälle vor den Behältern wie eben an. Wenn man die Behälter unterscheiden kann, dann erhalten wir $\binom{n}{n_1,\ldots,n_K}$ mögliche Anordnungen, so dass der erste Behälter n_1, der zweite n_2 Bälle ... enthält. Jetzt können wir für alle j die Anordnung Blöcke mit p_j Bälle ändern und erhalten eine gleichartige Partition! Es gibt $p_1! \cdot \ldots \cdot p_n!$ Anordnungen solche Umordnungen.

Es gilt

$$\frac{\binom{n}{n_1,\ldots,n_K}}{p_1! \cdot \ldots \cdot p_n!} = \frac{n!}{\prod_{j=1}^{n}(j!)^{p_j} p_j!}$$

Diese Zahl gibt an, wie viele Partitionen/Team-Einteilungen von n Spielern es gibt, so dass es p_j Team mit j Spielern gibt.

1.7.2 Bemerkung (Objekte/Subjekte ziehen): Wir ziehen k Objekte/Subjekte aus einer Gesamtheit von n Objekten/Subjekten. Wir müssen dann beachten, ob die gezogen Objekte zurück gelegt werden oder nicht. Und wir müssen beachten, ob die Reihenfolge für das Zählen relevant ist oder nicht.

1. Es gibt n^k **geordnete Stichproben mit Zurücklegen** der Größe k aus einer Grundgesamtheit von n Objekten.

2. Es gibt $(n)_k := n(n-1)(n-2)\cdot\ldots\cdot(n-k+1) = \frac{n!}{(n-k)!}$ **geordnete Stichproben ohne Zurücklegen** der Größe k aus einer Grundgesamtheit von n Objekten. Man nennt

$$(n)_k := n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

fallende Faktorielle.

3. Es gibt $\binom{n}{k}$ **ungeordnete** Stichproben **ohne Zurücklegen** der Größe k aus einer Grundgesamtheit von n Objekten. Man nennt

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomialkoeffizienten.

4. Es gibt $\binom{n+k-1}{k}$ **ungeordnete** Stichproben **mit Zurücklegen** der Größe k aus einer Grundgesamtheit von n Objekten. Man nennt

$$\left(\!\binom{n}{k}\!\right) := \binom{n+k-1}{k}$$

Multimengenkoeffizient.

Die Stichprobenmodelle sind zum Vergleich noch mal in der folgenden Tabelle angegeben.

	mit Wiederholung	ohne Wiederholung
geordnet	n^k	$(n)_k$
ungeordnet	$\left(\!\binom{n}{k}\!\right) := \binom{n+k-1}{k}$	$\binom{n}{k}$

1.7.3 Bemerkung (Objekte/Subjekte verteilen): Es gibt noch eine andere[7] *Interpretation* für die Formeln in der vorgehenden Tabelle: **Verteilung von k Bällen auf n Behälter.** Manchmal ist diese Auslegung praktischer. Wegen der Verwechslungsgefahr wähle ich Bälle, die auf Behälter verteilt werden und Kugeln, die gezogen werden. Es ist in der Tat auch anschaulich, sich die Behälter als Farbtöpfe vorzustellen. Die Verwechslungsgefahr besteht insbesondere bezüglich n und k. k ist die Anzahl der Objekte/Subjekte mit denen etwas *gemacht* wird (verteilt bzw. gezogen).

Bei dieser Interpretation werden k Bälle auf n Behälter verteilt. Die Behälter sind in allen vier Varianten unterscheidbar – z.B. farblich oder durch Nummern. Die Bälle u.U. nicht.

	mit Mehrfachbelegung	ohne Mehrfachbeleg.
unterschiedliche Kugeln	n^k	$(n)_k$
nicht-unterscheidbare Kugeln	$\left(\!\binom{n}{k}\!\right)$	$\binom{n}{k}$

[7] Die wir teilweise in der vorhergehenden Bemerkung betrachtet haben.

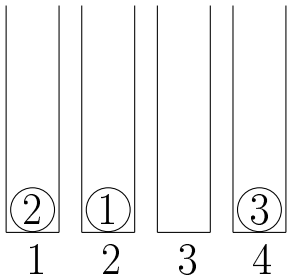
1.7.4 Bemerkung (Ziehen und Verteilen): Wir besprechen kurz den **Zusammenhang der beiden Perspektiven**. Die Bälle seien nummeriert (also unterscheidbar) und entsprechend seien auch die Kugeln nummeriert. Die Bijektion zwischen Ziehung und Verteilung ist dann die folgenden: Wenn wir zuerst die Kugel mit der Nummer l ziehen, dann legen wir den Ball mit der Nummer 1 in den l-ten Behälter. Also:

zufällig Kugel l bei Zug **1** $\leftrightarrow$ Ball **1** in **zufällig Behälter** l

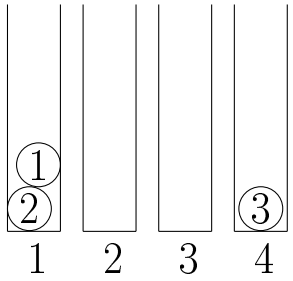
zufällig Kugel m bei Zug **2** $\leftrightarrow$ Ball **2** in **zufällig Behälter** m

Eselsbrücke für den Übergang von Ziehungen auf Belegungen: Man *zieht* die Behälter für die nummerierten bzw. nichtnummerierten Bälle.

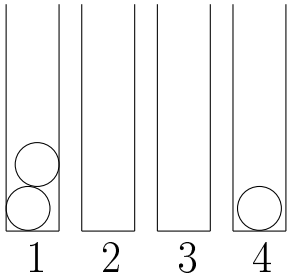
Beispiele: i.) Die folgende Abbildung zeigt eine Verteilung von 3 verschiedenen Bällen auf 4 Behälter, wobei in jeden Behälter maximal ein Ball liegt. Die Bälle sind nummeriert also unterschiedlich. Die Behälter sind ebenfalls unterschiedlich. Gemäß der Tabelle gibt es $(4)_3 = 4 \cdot 3 \cdot 2 = 24$ *solche* Verteilungen. Warum? DIY!



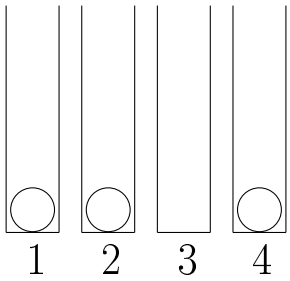
ii.) Die folgende Abbildung zeigt eine Verteilung von 3 verschiedenen Bällen auf 4 Behälter, wobei im ersten Behälter 2 Kugeln liegen. Gemäß der Tabelle gibt es $4^3 = 64$ *solche* Verteilungen. Warum? DIY!



iii.) Die folgende Abbildung zeigt eine Verteilung von 3 nicht-unterscheidbaren Bällen (nicht-unterscheidbar, da nicht nummeriert) auf 4 Behälter, wobei im ersten Behälter 2 Bälle liegen. Gemäß der Tabelle gibt es $\left(\binom{4}{3}\right) = \binom{4+3-1}{3} = \binom{6}{3} = 20$ *solche* Verteilungen. Warum? DIY!



iv.) Die folgende Abbildung zeigt eine Verteilung von 3 nicht-unterscheidbaren Bällen (nicht-unterscheidbar, da nicht nummeriert) auf 4 Behälter, wobei in jedem Behälter nur maximal ein Ball liegt. Gemäß der Tabelle gibt es $\binom{4}{3} = \binom{4}{3} = 4$ *solche* Verteilungen. Warum? DIY!


1
2
3
4

2 Diskrete Modelle

2.1 Allgemeines

Die im Folgenden erläuterten Modelle werden in vielen Quellen diskutiert. Wir verwenden Goergii [13], Henze [19], Tappe [36] und Dobrow und Wagamann [40]

2.1.1 Definition: Ein Wahrscheinlichkeitsmaß $\mathbb{P}$ heißt **diskrete Verteilung oder diskretes Wahrscheinlichkeitsmaß** auf $(\Omega, \mathcal{F})$, falls

i.) Ω eine nicht-leere abzählbare Menge und
ii.) $\mathcal{F} = \mathcal{P}(\Omega)$ ist.

P.S. Abzählbar bedeutet endlich oder abzählbar unendlich.

2.1.2 Definition: Es sei Ω eine nicht-leere abzählbare Menge und $p_j \in [0,1], j \in \Omega$ seien reelle Zahlen im Intervall $[0,1]$

mit

$$\sum_{j\in\Omega} p_j = 1.$$

Dann heißt $(p_j)_{j\in\Omega}$ **Zähldichte** oder **Wahrscheinlichkeitsfunktion**.

Eine Zähldichte/Wahrscheinlichkeitsfunktion ist (noch) kein Wahrscheinlichkeitsmaß! Insbesondere stimmen die Definitionsbereiche nicht überein. Die Wahrscheinlichkeitsfunktion liefert nur Wahrscheinlichkeiten für die Er**geb**nisse; das Wahrscheinlichkeitsmaß für alle Er**eig**nisse. Aber aus einer Wahrscheinlichkeitsfunktion ergibt sich – wie der folgende Satz zeigt – ein Wahrscheinlichkeitsmaß.

2.1.3 Satz: Es sei Ω eine nicht-leere abzählbare Menge und $(p_k)_{k\in\Omega}$ eine Zähldichte auf $(\Omega, \mathcal{P}(\Omega))$. Dann gibt es genau ein diskretes Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $(\Omega, \mathcal{P}(\Omega))$ mit $\mathbb{P}(\{\omega_k\}) = p_k, k \in \Omega$.

Für dieses Wahrscheinlichkeitsmaß gilt

$$A \mapsto \mathbb{P}(A) = \sum_{\omega_k\in A} p_k, \quad A \in \mathcal{P}(\Omega)$$

Wir nennen $\mathbb{P}$ das durch die Zähldichte $(p_k)_{k\in\Omega}$ definierte Wahrscheinlichkeitsmaß.

2.1.4 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum (nicht notwendigerweise diskret) und $X : \Omega \to \mathbb{R}$ eine Abbildung. X heißt **diskrete Zufallsvariable**, falls:

i.) $\text{Bild}(X) = X(\Omega) = \{X(\omega)|\omega \in \Omega\}$ ist abzählbar. $X(\Omega)$ heißt Bild von X.
ii.) Für alle $A \subset X(\Omega)$ ist $X^{-1}(A) \in \mathcal{F}$, d.h. X ist **messbar**.

2.1.5 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable. Dann heißt

$$\mathbb{P}^X : \begin{cases} \mathcal{P}(X(\Omega)) \to [0,1] \\ A \mapsto \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) \end{cases}$$

die Verteilung von X. Dabei ist $\mathbb{P}^X(A) = \mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \in A\}) = \mathbb{P}(X^{-1}(A))$ die Wahrscheinlichkeit das X einen Wert in der Menge A annimmt.

2.1.6 Bemerkung: Wir wollen $\mathbb{P}^X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$ berechnen können. Das geht aber nur, wenn $X^{-1}(A) \in \mathcal{F}$ gilt, denn nur für Mengen in $\mathcal{F}$ ist $\mathbb{P}$ definiert. Aus diesen Grund fordern wir bei Zufallsvariablen die Messbarkeit.

2.1.7 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable mit Vertei-

lung $\mathbb{P}^X$. Dann ist $\mathbb{P}^X$ ein diskretes Wahrscheinlichkeitsmaß auf $(X(\Omega), \mathcal{P}(X(\Omega))$.

2.1.8 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable mit Verteilung $\mathbb{P}^X$. Wir schreiben dann

$$X \sim \mathbb{P}^X.$$

2.1.9 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable mit Verteilung $\mathbb{P}^X$. Es sei ferner $p(x) = \mathbb{P}^X(\{x\}), x \in \text{Bild}(X)$ die Zähldichte für $\mathbb{P}^X$. Dann definieren wir

$$\mathbb{E}(X) = \sum_{x \in \text{Bild}(X)} p(x) \cdot x,$$

falls die Reihe – so es eine ist – absolut konvergiert. Wir sagen dann, dass der Erwartungswert von X existiert.

2.1.10 Satz (Gesetz des unbewussten Statistikers): Es gilt:

$$\mathbb{E}(g(X)) = \sum_{x \in \text{Bild}(X)} p(x)g(x).$$

Beweis: Vgl. Grimmett und Stirzaker [15, Seite 228]. Wir

setzen $Y = g(X)$ und beobachten

$$\mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x).$$

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{y \in Y(\Omega)} \mathbb{P}(Y = y)y \\
&= \sum_{y \in Y(\Omega)} \mathbb{P}(g(X) = y)y \\
&= \sum_{y \in Y(\Omega)} \sum_{x:g(x)=y} \mathbb{P}(X = x)y \\
&= \sum_{y \in Y(\Omega)} \sum_{x:g(x)=y} \mathbb{P}(X = x)g(x) \\
&= \sum_{x \in X(\Omega)} \mathbb{P}(X = x)g(x)
\end{aligned}$$

2.1.11 Satz: Es gilt:

i.) $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$. Der Erwartungswert ist ein linearer Operator.

ii.) $X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$. Der Erwartungswert ist ein monotoner Operator.

Beweis: Vgl. Grimmett und Stirzaker [14, Seite 55].

2.1.12 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable, so dass

der Erwartungswerte von X und von $(X - \mathbb{E}(X))^2$ existiert. Dann definieren wir

$$\mathbb{V}(X) := \sigma_X^2 := \mathbb{E}((X - \mathbb{E}(X))^2)$$

$\mathbb{V}(X)$ heißt die **Varianz** der Zufallsvariable X. Wir sagen, dass die Varianz von X existiert. $\sigma_X = \sqrt{\mathbb{V}(X)}$ heißt die **Standardabweichung** von X.

2.1.13 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine diskrete Zufallsvariable, so dass der Erwartungswerte von X und von $(X-\mathbb{E}(X))^2$ existiert. Dann

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

▶ Der Erwartungswert und die Varianz sind sehr wichtige Konzepte mit vielen schönen Eigenschaften, die wir genauer untersuchen, wenn wir den geeigneten allgemeinen Integrationsbegriff eingeführt haben. Die Eigenschaften kann man auch direkt analysieren, das ist jedoch mühsamer.

2.1.14 Bemerkung: Auf der Formelsammlung sind die Erwartungswerte und die Varianzen für die diskreten Standardmodelle angegeben.

2.2 Dirac und Laplace

2.2.1 Definition: Es sei Ω eine nicht-leere abzählbare Menge und $k^* \in \Omega$. Dann ist

$$p_k = \begin{cases} 1 \text{ für } k = k^* \\ 0 \text{ sonst} \end{cases}$$

eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt das im Punkt k^* konzentrierte **Dirac-Wahrscheinlichkeitsmaß**.

2.2.2 Definition: Es sei Ω eine endliche Menge mit $n \in \mathbb{N}$ Elementen. Dann ist $p_k = \frac{1}{n}, k = 1, ..., n$ eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt **Laplace-Wahrscheinlichkeitsmaß** und $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein **Laplace-Experiment**.

Ein Experiment $(\Omega, \mathcal{F}, \mathbb{P})$ heißt also **Laplace Experiment** oder **Laplace Wahrscheinlichkeitsraum**, falls:

i.) $\Omega \neq \emptyset$ ist eine endliche Menge mit $n \in \mathbb{N}$ Elementen.

ii.) $\mathcal{F} = \mathcal{P}(\Omega)$.

iii.) Für alle $A \in \mathcal{F}$ ist

$$\mathbb{P}(A) = \frac{\#(A)}{\#(\Omega)} = \frac{\#(A)}{n} = \frac{\text{Anzahl der für } A \text{ günstigen Ergebnisse}}{\text{Anzahl aller Ergebnisse}}.$$

2.3 Bernoulli, Binomial und Multinominal

2.3.1 Definition: Es sei $\Omega = \{\circ, \square\}$ und $p \in (0,1)$. Dann ist $p(\circ) = p, p(\square) = 1-p$ eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt **Bernoulli-Wahrscheinlichkeitsmaß** mit Erfolgswahrscheinlichkeit p. Man nennt $(\Omega, \mathcal{F}, \mathbb{P})$ ein **Bernoulli-Experiment**. Das Ergebnis $\circ$ nennt man "Erfolg" und das Ergebnis $\square$ "Misserfolg". p wird **Erfolgswahrscheinlichkeit** genannt.

Anstatt $\Omega = \{\circ, \square\}$ wird typischerweise $\Omega = \{1, 0\}$ verwendet, wobei 1 für Erfolg steht. In der Finanzmathematik wird regelmäßig $\Omega = \{1, -1\}$, wobei 1 wieder für Erfolg steht.

2.3.2 Definition und Satz: Es sei $\Omega = \{0, 1, 2, ..., n\}$ und $p \in (0,1)$. Dann ist

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt **Binomialverteilung** mit **Erfolgswahrscheinlichkeit** p und n Wiederholungen.

Für den Beweis von $\sum_{i=1}^{n} p_i = 1$ benötigt man den Binomi-

schen Lehrsatz:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

und eine magische Null $1 = 1 - p + p$.

2.3.3 Bemerkung: Die Binominalverteilung hat extrem viele Anwendungen. Man stellt sich dabei die unabhängige n-fache Wiederholung eines Bernoulli-Experiments vor. p_k ist dann die Wahrscheinlichkeit bei n unabhängigen Wiederholungen **genau k Erfolge und $n-k$ Misserfolge** zu beobachten.

2.3.4 Satz (Additionsgesetz für die Binomial-Verteilung): Wenn X eine Binomial-Verteilung mit m Wiederholungen und Erfolgswahrscheinlichkeit p und Y eine Binomial-Verteilung mit n Wiederholungen und Erfolgswahrscheinlichkeit p hat, dann hat $X + Y$ eine Binomial Verteilung mit $m + n$ Wiederholungen und Erfolgswahrscheinlichkeit p.

Beweis: Vgl. Henze [19, S. 94].

2.3.5 Definition und Satz: Es sei $\Omega = \{(n_1, n_2, ..., n_K) \in \mathbb{N}_0^K \mid \sum_{j=1}^{K} n_j = n\}$ und $p_i \in (0,1), i = 1, .., K$, wobei $\sum_{i=1}^{K} p_i =$

1 gilt. Dann ist

$$p((n_1, n_2, ..., n_K)) = \binom{n}{n_1, n_2, ..., n_K} p_1^{n_1} \cdot p_2^{n_2} \cdot ... \cdot p_K^{n_K}$$

eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt **Multinomialverteilung**.

Für den Nachweis, dass sich die Wahrscheinlichkeiten zu Eins addieren, benötigt man die Verallgemeinerung des Binomischen Lehrsatzes für Multinomialkoeffizienten (vgl. Mazur [28, S. 143]).

2.3.6 Bemerkung: Die Multinomialverteilung ist eine Verallgemeinerung der Binomialverteilung. Hier wird ein Experiment mit K **möglichen Ergebnissen** n**-mal unabhängig wiederholt**. Bei jeder Wiederholung ist die Wahrscheinlichkeit, das Ergebnis $i \in \{1, ..., K\}$ zu erhalten, gleich p_i. Der Multinomialkoeffizient tritt an die Stelle des Binomialkoeffizienten.

$p((n_1, n_2, ..., n_K))$ ist dann die Wahrscheinlichkeit n_1 mal das Ergebnis 1, n_2 mal das Ergebnis 2, ... und n_K das Ergebnis K zu erhalten.

Die Multinomialverteilung kann man beispielsweise auf die zufällige **Verteilung** von n Teilchen auf K Zustände anwenden; oder auf die zufällige **Färbung** von n Bälle in K Farben.

Für jeden Ball ist dabei die Wahrscheinlichkeit, in der i-ten Farbe gefärbt zu werden, gleich p_i. $p((n_1, n_2, ..., n_K))$ erfasst dabei nicht, welche der Bälle wie gefärbt werden, sondern nur wie viele Bälle wie gefärbt sind: n_1 Bälle in der Farbe 1, n_2 in der Farbe 2 ....

Wir können uns hier auch folgendes **Ziehungsmodell** vorstellen: In einer Urne liegen verschieden farbige Kugeln; K verschiedene Farben. p_i bezeichnet den Anteil der i-farbigen Kugeln. Wir ziehen n Kugeln mit zurücklegen.

2.4 Hypergeometrisch

2.4.1 Satz und Definition: Es seien $N, M, k \in \mathbb{N}$ mit $M \leq N$ und $k \leq N$ und $\Omega = \{0, 1, ..., k\}$. Dann ist $p_l = \frac{\binom{M}{l}\binom{N-M}{k-l}}{\binom{N}{k}}$ eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt **hypergeometrische Verteilung** mit Parametern N, M, k.

2.4.2 Bemerkung: Die hypergeometrische Verteilung wird für die Modellierung des gleichzeitigen/**ungeordneten Ziehens ohne Zurücklegen von k Kugeln** aus einer Urne mit N Kugel, von denen M Kugel **grün** bzw. $N - M$ **blau** sind, verwendet. p_l ist dann **die Wahrscheinlichkeit l grüne Kugeln zu ziehen**.

Wir wollen nachrechnen, dass sich die **genannte Wahrscheinlichkeiten** ergibt, wenn man die Kugeln **nacheinander zunächst geordnet ohne Zurücklegen** zieht und dann die Reihenfolge ignoriert. Es gibt $(N)_k$ mögliche geordnete Ziehungen von k Kugel aus einer Urne mit N Kugeln. Das erklärt den Nenner in der folgenden abgesetzten Gleichung. Wir betrachten dann die Anzahl der Möglichkeiten l der M grünen Kugel bzw. $k-l$ der $N-M$ blauen Kugel zu ziehen; wobei wir die Reihenfolge/Nummer zunächst ignorieren. Für solche Auswahlen gibt es $\binom{M}{l}\binom{N-M}{k-l}$ Möglichkeiten. Für jede dieser Auswahlen gibt es $k!$ Anordnungsmöglichkeiten.

Jetzt setzen wir alles zusammen und beobachten schließlich

$$\frac{\binom{M}{l}\binom{N-M}{k-l}\cdot k!}{(N)_k} = \frac{\binom{M}{l}\binom{N-M}{k-l}}{\frac{N!}{(N-k)!k!}} = \frac{\binom{M}{l}\binom{N-M}{k-l}}{\binom{N}{k}}.$$

2.5 Belegungsmodelle, Maxwell-Boltzmann, Bose-Einstein, Fermi-Dirac

2.5.1 Bemerkung: Nützlich ist die hypergeometrische Verteilung z.B. in der **Qualitätskontrolle**. Ein Produzent erhält eine Lieferung von 10000 Bauteilen und will die Qualität prüfen. Es ist zu teuer alle Bauteile zu überprüfen. Man entnimmt eine Stichprobe von z.B. 30 Bauteilen und prüft diese.

Die Situation gleicht dann offenbar der Entnahme von Kugel (die jeweils eine von zwei Farben haben) aus einer Schale.

2.5.2 Bemerkung (Belegungsmodelle): Belegungsmodelle modellieren die **Verteilung von Kugeln auf Behälter**[1]. Zwei Aspekte muss man beachten: **Beobachtungstiefe** und **Restriktionen**.

Die mögliche **Verringerung der Beobachtungstiefe** bei Belegungen besteht darin, dass/ob man die Nummern auf den Kugeln ignoriert. Die Nummern auf den Kugeln werden *unsichtbar*. Relevant ist nicht, welche Kugeln in den Behältern liegen, sondern wie viele. Die mögliche **Restriktion** ist gegebenenfalls, dass maximal eine Kugel in einem Behälter liegen darf.

Bei **maximaler Beobachtungstiefe** und **ohne Restriktion** ist die Ergebnismenge die Menge der n-Tupel mit Einträgen aus $\{1, ..., K\}$. n Kugeln werden auf die Behälter verteilt. K ist die Anzahl Behälter/Farben und n die Anzahl der Kugeln. Das Ergebnis wird als n-Tupel $(x_1, ..., x_n)$ protokolliert. Dabei bezeigt $x_i = k$ den Behälter k in den die Kugel i gelegt.

$$\Omega_1 = \{(x_1, ..., x_n) | x_i \in \{1, ..., K\}, i = 1, ..., n\}$$

[1] Die Behälter sind in dieser Bemerkung stets unterscheidbar, also mit Nummern oder Farben versehen.

und alle Ergebnisse haben die Wahrscheinlichkeit (jedenfalls ist das eine *natürliche/denkbare* Voraussetzung)

$$\mathbb{P}_1(\{(x_1, ..., x_n)\}) = \left(\frac{1}{K}\right)^n = \frac{1}{K^n}.$$

Wir stellen uns jetzt vor, dass die Nummern auf den Kugeln verschwinden. Das **Experiment soll dabei unverändert** bleiben. Es soll sich **nur die Beobachtbarkeit ändern**. Bestimmte Ergebnisse, die wir bei genauer Beobachtungstiefe unterscheiden (beobachten) können, sind dann nicht mehr unterscheidbar und werden zu einen neuen Ergebnis **zusammengelegt**. Durch den Übergang von nummeriert zu nicht-nummeriert werden bestimmte Elemente sozusagen *identifiziert*. Aus nummerierten Belegungen werden nicht-nummerierte Belegungen. Angenommen wir betrachten 5 Kugeln und 7 Behälter. Die Ergebnisse $(2, 2, 2, 2, 3)$, $(2, 2, 2, 3, 2)$, $(2, 2, 3, 2, 2)$, $(2, 3, 2, 2, 2)$ und $(3, 2, 2, 2, 2)$ beispielsweise werden zu $\langle 2, 2, 2, 2, 3\rangle$.[2] Bei verringerter Beobachtungstiefe beobachten wir, dass eine Kugel im dritten Behälter liegt und vier Kugeln im zweiten Behälter. Wir beobachten jedoch nicht, welche Kugel in den zweiten Behälter gelegt wird. Die anderen Behälter sind leer. Bei genauer Beobachtungstiefe können wir z.B. $(2, 2, 2, 2, 3)$ und $(2, 2, 2, 3, 2)$ unterscheiden können. Bei $(2, 2, 2, 2, 3)$ liegt die Kugel 5 im Behälter 3; bei $(2, 2, 2, 3, 2)$ die Kugel 4 im Behälter 3. Die Anzahl der Ergebnisse, die jeweils zusam-

[2] Wir verwenden $\langle, \rangle$ für Multimengen.

mengelegt werden entspricht dem Multinomialkoeffizienten: $\binom{5}{0,4,1,0,0,0,0} = \frac{5\cdot4\cdot3\cdot2\cdot1}{4!\cdot1!} = 5$. Die Zahl der Ergebnisse, die zusammengelegt werden, variiert: Das Ergebnis $\langle 2,2,2,2,2\rangle$ ergibt sich genau aus einem Ergebnis $(2,2,2,2,2)$. Das Ergebnis $\langle 2,2,2,2,3\rangle$ ergibt sich aus fünf: $(2,2,2,2,3)$, $(2,2,2,3,2)$, $(2,2,3,2,2)$, $(2,3,2,2,2)$ und $(3,2,2,2,2)$. Wie viele Ergebnisse zusammengelegt werden, wird durch den Multinomialkoeffizienten angegeben.

Durch dieses Zusammenlegen erhalten wir dementsprechend auf der Menge der Multimengen $\Omega_4 = \{\, \langle x_1,...,x_n\rangle \,|x_i \in \{1,...,K\}\}$ *eine*[3] **Multinomialverteilung**.

$$\mathbb{P}_4(\langle 2,2,2,2,3\rangle) = \binom{5}{0,4,1,0,0,0,0} \cdot \left(\frac{1}{K}\right)^5,$$

wobei

$$\Omega_4 = \{\, \langle x_1,...,x_n\rangle \,|x_i \in \{1,...,K\}\}.$$

Es gilt $|\Omega_4| = \left(\binom{K}{n}\right)$; Ω_4 hat $\left(\binom{K}{n}\right) = \binom{K+n-1}{n}$ Elemente.

In der Physik kennt man diese Verteilung auch als **Maxwell-Boltzmann-Statistik**[4] (siehe z.B. Feller [10, S. 39, 41f]).

Wir erhalten die *Konsistenz* zur geordneten Ziehung mit Zu-

[3] Wir haben vorne schon Multinomialverteilungen allgemein betrachtet. Hier ist $p_i = 1/K$ für $i = 1,...,K$.

[4] Mathematiker wählen nicht den Begriff Statistik.

rücklegen aus der Formel (vgl. Mazur [28, S. 143]) aus

$$K^n = \sum_{\substack{(t_1,....,t_K) \\ t_1+...+t_K=n}} \binom{n}{t_1, ..., t_K},$$

wobei die Summe über alle $(t_1, ..., t_K)$ mit $t_1 + ... + t_K = n$ läuft.

Auf Ω_4 wird durch die Multinomialverteilung (durch das Zusammenlegung) **kein Laplace Experiment definiert, sondern ein sogenanntes Maxwell-Boltzmann-Experiment**. Wenn man allen n-Multimengen die Wahrscheinlichkeit

$$\frac{1}{\left(\binom{K}{n}\right)}$$

zuweist, dann ergibt sich ein Laplace-Experiment auf Ω_4. Die Verteilung wird in diesem Zusammenhang auch **Bose-Einstein-Statistik** genannt (siehe z.B. Feller [10, S. 41f]).

Wir erhalten ein weiteres Belegungsmodell, wenn wir **Mehrfachbelegungen** ausschließen. Dann erhalten wir ein **Laplace-Experiment** auf den **n-Tupeln** $\Omega_2 \subset \Omega_1$ mit paarweise unterschiedlichen Einträgen

$$\Omega_2 = \{(x_1, ..., x_n) \,|\, x_i \in \{1, ..., K\}, x_i \neq x_j, i \neq j\}$$

mit Wahrscheinlichkeiten (für alle Ergebnisse gleich)

$$\mathbb{P}_2(\{(x_1, ..., x_n)\}) = \frac{1}{(K)_n}.$$

Wenn wir dabei auch noch die Nummern nicht beobachten/beachten, dann betrachten wir auf der Menge der n-**Mengen**

$$\Omega_3 = \{\{x_1, ..., x_n\} \,|\, x_i \in \{1, ..., K\}, x_i \neq x_j, i \neq j\}$$

die Wahrscheinlichkeiten (für alle Ergebnisse gleich)

$$\mathbb{P}_3(\{x_1, ..., x_n\}) = \frac{1}{\binom{K}{n}}.$$

Diese Verteilung wird in der Physik **Fermi-Dirac-Statistik** genannt; vgl. Feller [10, S. 41] und Blitzstein und Hwang [4, Seite 18f]. Bei Übergang von Ω_2 auf Ω_3 entsteht diesmal ein **Laplace-Experiment**, denn immer eine gleiche Anzahl von Ereignisse kann man – wenn die Nummern verschwinden – nicht mehr unterschieden (nämlich $n!$). Also rechnen wir nach

$$\frac{(K)_n}{n!} = \frac{\frac{K!}{(K-n)!}}{n!} = \frac{K!}{(K-n)!K!} = \binom{K}{n}.$$

2.6 Poisson

2.6.1 Satz und Definition: Es sei $\Omega = \{0, 1, 2, 3, ...\} = \mathbb{N}_0$ und $\lambda > 0$. Dann ist

$$p_k = e^{-\lambda}\frac{\lambda^k}{k!}$$

eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt **Poisson-Verteilung** mit Parameter λ. λ wird manchmal **Eintrittsrate** oder **Ankunftsrate** genannt.

2.6.2 Satz und Definition: Wir betrachten ein Experiment, bei dem über einen bestimmten Zeitraum (z.B. 90 Minuten) ein Ergebnis (Erfolg/Tor) mit einer *homogenen/konstanten* Rate λ eintritt. Weiter nehmen an: Wenn wir die 90 Minuten in n gleiche Teilzeiträume zerlegen, dann tritt das Ereignis mit der Rate $\frac{\lambda}{n}$ ein. Wenn das n hinreichend groß ist, dann kann man sich vorstellen, dass das Ereignis keinmal oder nur einmal im kurzen Zeitintervall nach Teilung durch n eintritt. In den kleinen Zeitintervallen haben wir dann Bernoulli-Experimente von denen wir auch annehmen wollen, dass Sie unabhängig sind. Wenn wir die *Erfolge* über 90 Minuten hinweg zählen, können wir somit das Experiment als **Binominal-Experiment mit n Wiederholung und $p = \frac{\lambda}{n}$ auffassen**.

Es gilt (vgl. z.B. Wagaman und Dobrow [40, S. 108f]):

$$\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \to e^{-\lambda}\,\frac{\lambda^k}{k!}.$$

Da $\left(\frac{\lambda}{n}\right)$ klein ist, spricht man vom **Gesetz der kleinen Zahlen** (law of rare events).

▶ Die Anzahl der Soldaten, die durch einem Pferdetritt getötet werden.

▶ Die Anzahl der Tore in einem Fußballspiel.

2.6.3 Satz (Additionsgesetz für die Poisson-Verteilung): Wenn X eine Poisson-Verteilung mit Parameter λ_X und Y eine Poisson-Verteilung mit Parameter λ_Y hat, dann hat $X+Y$ eine Poisson-Verteilung mit Parameter $\lambda_X + \lambda_Y$.

Beweis: Henze [19, S. 97]

▶ Tore der Heimteams und der Auswärtsteams. Tore insgesamt.

2.7 Geometrisch und Pascal

2.7.1 Satz und Definition: Es sei $\Omega = \{0, 1, 2, 3, ...\} = \mathbb{N}_0$ und $p \in (0, 1)$. Dann ist

$$p_k = (1-p)^k p$$

eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt **geometrische Verteilung** mit Parameter p.

2.7.2 Bemerkung: Die geometrische Verteilung verwendet man für das idealisierte Experiment des **Wartens** auf den ersten Erfolg. Ein Bernoulli-Experiment wird so lange unabhängig wiederholt bis der erste Erfolg eintritt. Dann erfasst p_k die Wahrscheinlichkeit von zunächst k Misserfolgen bis zum ersten Erfolg.

2.7.3 Definition und Satz: Es sei $\Omega = \{0, 1, 2, 3, ...\} = \mathbb{N}_0$, $p \in (0, 1)$ und $n \in \mathbb{N}$. Dann ist

$$p_k = \binom{n+k-1}{k} (1-p)^k p^n$$

eine Zähldichte. Das durch diese Zähldichte definierte diskrete Wahrscheinlichkeitsmaß heißt **Pascal-Verteilung** oder **negative Binomialverteilung** mit Parametern n und p.

2.7.4 Bemerkung: Wir betrachten die unabhängige Wiederholung eines Bernoulli-Experiments bis erstmals n Erfolge eingetreten sind p_k ist dann die Wahrscheinlichkeit $n+k$ Versuchen erstmals n Erfolge eingetreten.

3 Standardmodelle mit Dichten

3.1 Überabzählbar ist indiskret

3.1.1 Bemerkung: Es gibt **keinen** Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ mit den folgenden Eigenschaften:

i.) Ω ist **überabzählbar unendlich**,
ii.) $\mathcal{F}$ enthält alle ein-elementigen Mengen $\{\omega\}, \omega \in \Omega$,
iii.) $\mathbb{P}(\{\omega\}) > 0$ für alle $\omega \in \Omega$.

Bei diskreten Wahrscheinlichkeitsräumen konnten wir auf Basis der Zähldichte/ Wahrscheinlichkeitsfunktion $p_j = \mathbb{P}(\omega_j)$ die Wahrscheinlichkeit für ein Ereignis $A \subset \Omega$ durch Summation ermitteln: $\mathbb{P}(A) = \sum_{\omega_j \in A} \mathbb{P}(\omega_j) = \sum_{\omega_j \in A} p_j$.

Wenn Ω überabzählbar unendlich ist, dann geht das so nicht mehr. Wir können weder mit Wahrscheinlichkeitsfunktionen

noch mit der Summation arbeiten. Wir benötigen **Dichten** und **Integrale**.

Für den Widerspruchsbeweis betrachtet man

$$\Omega = \biguplus_{k=1}^{\infty} \left\{ \omega \,|\, \frac{1}{k+1} < \mathbb{P}(\omega) \leq \frac{1}{k} \right\} = \biguplus_{k=1}^{\infty} A_k$$

Mindestens eine der Mengen A_k muss unendlich viele Elemente enthalten, sonst wäre Ω abzählbar; sagen wir A_k. Dann wäre $\mathbb{P}(A_k) > 1$.

3.2 Verteilung mit Dichte

3.2.1 Definition: Es sei $\mathcal{B}(\mathbb{R}) = \sigma(\{(a,b] : a \leq b, a, b \in \mathbb{R}\})$ die kleinste σ-Algebra, die alle Loravalle $(a,b]$ enthält, d.h. alle links offenen und rechts abgeschlossenen Intervalle. Die Mengen $B \in \mathcal{B}(\mathbb{R})$ heißen **Borelmengen** (in $\mathbb{R}$). Jede andere σ-Algebra, die die **Loravalle** enthält, umfasst $\mathcal{B}(\mathbb{R})$. Man sagt, dass $\mathcal{B}$ von den Loravallen erzeugt wird. Das Konzept der „kleinsten“ σ-Algebra wird später genauer betrachtet.

3.2.2 Definition: Es sei $f : \mathbb{R} \to \mathbb{R}$ integrierbare[1] Funktion. f heißt **Dichte**, falls:

[1] Was das genau bedeutet wird erst später präzise definiert. Wir gehen hier *naiv* vor.

i.) $f \geq 0$,
ii.) $\int_{-\infty}^{\infty} f(x)dx = 1$.

3.2.3 Definition und Satz: Es sei $f : \mathbb{R} \to \mathbb{R}$ eine Dichte. Dann existiert genau ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf dem $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit

$$\mathbb{P}((-\infty, x]) = \int_{-\infty}^{x} f(z)dz, x \in \mathbb{R}.$$

In diesem Fall sagen wir: f **ist die Dichte von** $\mathbb{P}$.

3.2.4 Satz: Es sei $(\mathbb{R}, \mathcal{B}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und f die Dichte von $\mathbb{P}$. Dann gilt für alle $a, b \in \mathbb{R}, a < b$:

$$\begin{aligned}
\mathbb{P}(\{a\}) &= 0 \\
\mathbb{P}((-\infty, a]) = \mathbb{P}((-\infty, a)) &= \int_{-\infty}^{a} f(x)dx \\
\mathbb{P}((a, \infty)) = \mathbb{P}([a, \infty)) &= \int_{a}^{\infty} f(x)dx \\
\mathbb{P}((a, b)) = \mathbb{P}((a, b]) = \mathbb{P}([a, b)) = \mathbb{P}([a, b]) &= \int_{a}^{b} f(x)dx
\end{aligned}$$

3.2.5 Definition und Satz: i.) Es sei $(\Omega, \mathcal{F})$ ein messbarer Raum und $X : \Omega \to \mathbb{R}$ mit $X^{-1}(B) \in \mathcal{F}$ für alle $B \in \mathcal{B}(\mathbb{R})$. Dann heißt X eine reellwertige Zufallsvariable. Die Eigenschaft $X^{-1}(B) \in \mathcal{F}$ für alle $B \in \mathcal{B}(\mathbb{R})$ heißt **Messbarkeit** von X. [Erinnerung $X^{-1}(B) = \{\omega | X(\omega) \in B\}$]

ii.) Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Es sei $X : \Omega \to \mathbb{R}$ eine **Zufallsvariable**. Dann heißt

$$\mathbb{P}^X = \begin{cases} \mathcal{B}(\mathbb{R}) \to [0,1] \\ B \mapsto \mathbb{P}(X^{-1}(B)) =: \mathbb{P}(X \in B) \end{cases}$$

die **Verteilung von** X und wir schreiben $X \sim \mathbb{P}^X$.

Ist $F^X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leq x)$ die Verteilungsfunktion von X, dann schreiben wir auch $X \sim F^X$.

$\mathbb{P}^X$ ist ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}^X)$ ist ein Wahrscheinlichkeitsraum.

iii.) Es sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Wenn f die Dichte von $\mathbb{P}^X$ ist, dann heißt f auch die Dichte von X und wir schreiben $X \sim f$.

3.2.6 Definition: Es sei X eine Zufallsvariable mit Dichte f. Wenn das folgende Integral

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

existiert, dann heißt $\mathbb{E}(X)$ der **Erwartungswert** von X.

3.2.7 Satz: Es gilt:

i.) $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$. Der Erwartungswert ist ein **linearer** Operator.

ii.) $X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$. Der Erwartungswert ist ein **monotoner** Operator.

3.2.8 Definition: Es sei X eine Zufallsvariable mit Dichte f, so dass die Erwartungswerte von X und $(X - \mathbb{E}(X))^2$ existieren. Dann heißt

$$\sigma_X^2 = \mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

die **Varianz** von X. σ_X heißt **Standardabweichung** von X.

▶ Der Erwartungswert und die Varianz sind sehr wichtige Konzepte mit vielen schönen Eigenschaften, die wir genauer untersuchen, wenn wir den geeigneten allgemeinen Integrationsbegriff eingeführt haben. Die Eigenschaften kann man auch direkt analysieren, das ist jedoch mühsamer.

3.2.9 Bemerkung: Die folgenden schönen Eigenschaften werden hier aus dem oben genannten Grund ohne Beweis und ohne detaillierte Angaben der Voraussetzungen angeben.

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2,$$
$$\mathbb{V}(aX + b) = a^2\mathbb{V}(X) = a\mathbb{V}a.$$

Wenn X und Y unabhängig (eigentlich reicht un korreliert),

dann

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y).$$

3.3 Gleich-, Normal und Exponentialverteilung

3.3.1 Satz und Definition: Es sei $a < b$ und

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ falls } x \in (a,b) \\ 0 \text{ sonst} \end{cases}$$

f ist eine Dichte. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ mit dieser Dichte heißt **Gleichverteilung** auf dem Intervall (a,b).

3.3.2 Satz und Definition: Es sei $\mu \in \mathbb{R}, \sigma > 0$ und

$$\phi_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\phi_{\mu,\sigma}(x)$ ist eine Dichte. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ mit dieser Dichte heißt **Normalverteilung** oder **Gauß-Verteilung** mit den Parametern μ und σ. Für $\mu = 0, \sigma = 1$ erhalten wir

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

In diesem Fall heißt $\mathbb{P}$ die **Standardnormalverteilung**.

Siehe Henze [19, S. 127, 137].

3.3.3 Bemerkung: Es wird sich zeigen, dass die Normalverteilung eine Sonderrolle in der Wahrscheinlichkeitstheorie einnimmt. Die soll hier schon angedeutet werden. Wir betrachten dazu eine **Monte-Carlo Studie** im R Code `Demo_N_ZgS.R`.

Wir betrachten jeweils für jeden MC-Durchlauf n unabhängige Wiederholungen eines Bernoulli-Experiments mit Erfolgswahrscheinlichkeit p. Wir erhalten also eine Binomialverteilung für die Anzahl der Erfolge. Dieses Experiment – d.h. die n Wiederholungen – wiederholen wir m mal. Wir betrachten dann die relativen Häufigkeiten der Anzahl der Erfolge jeweils bei den m Durchläufen und vergleichen diese mit der Binomialverteilung und mit einer *angepassten* Normalverteilung. Für *großes* m erkennt man kaum einen Unterschied zwischen den relativen Häufigkeiten, der Binomialverteilung bzw. der **Normalverteilung**. Unser Experiment hat also zunächst nichts mit der Normalverteilung zu tun. Im *Grenzübergang* erhalten wir aber die Normalverteilung. **SEHR BEMERKENSWERT!** Mit diesem Zusammenhängen beschäftigen wir uns genauer, wenn wir den Zentralen Grenzwertsatz behandeln.

3.3.4 Satz und Definition: Es sei $\mu \in \mathbb{R}, \sigma > 0$. Dann ist

$$f(x) = \begin{cases} \frac{1}{x}\phi_{\mu,\sigma}(\ln(x)) & \text{falls } x > 0 \\ 0 & \text{sonst} \end{cases}$$

eine Dichte. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ mit dieser Dichte heißt **logarithmische Normalverteilung** (bzw. **Log-Normalverteilung**) mit den Parametern μ und σ.

3.3.5 Bemerkung: Es sei $X \sim N(\mu, \sigma)$. Wir werden später nachweisen, dass dann $Y = e^X$ log-normalverteilt ist. Anders formuliert: Ist Y log-normalverteilt, dann ist $\ln(Y)$ normalverteilt.

Siehe Henze [19, S. 159].

3.3.6 Satz und Definition: Es sei $\lambda > 0$. Dann ist

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{falls } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

eine Dichte. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ mit dieser Dichte heißt **Exponentialverteilung** mit Parameter λ.

3.3.7 Bemerkung: Die **Exponentialverteilung** ist das stetige Pendant zur diskreten **geometrischen Verteilung**. Dementsprechend wird sie auch für die Modellierung von **Wartezeiten** verwendet. Bei der geometrischen Verteilung wird die

Zahl der Perioden bis zum ersten Erfolg betrachtet. Bei der Exponentialverteilung die Dauer in Zeiteinheiten bis zum ersten Erfolg betrachtet. Man kann ein Experiment mit Exponentialverteilung als Grenzwert von Experimenten mit geometrischen Verteilungen auffassen. Wir betrachten dazu die Wahrscheinlichkeit $\mathbb{P}^{\mathrm{EXP}(\lambda)}((t,\infty))$, d.h. der erste Erfolg tritt nach t ein. Wir beobachten

$$\mathbb{P}([0,t]) = \int_0^t \lambda e^{-\lambda x} dx = [-e^{-\lambda x}]_0^t = 1 - e^{-\lambda t}$$

Also ist die Wahrscheinlichkeit des ersten Erfolgs nach t gleich

$$\mathbb{P}((t,\infty)) = e^{-\lambda t}.$$

Wir werden jetzt nachweisen, dass sich diese Wahrscheinlichkeit näherungsweise für ein Experiment mit geometrischer Verteilung ergibt. Zudem wird diese Approximation genauer, wenn n größer wird.

Wir betrachten Epochen E_i der Länge $\frac{t}{n}$, wobei o.B.d.A. $\mathbb{N} \ni n > t$ gelten soll. Wir zerlegen die nicht-negative Achse in Intervalle/Epochen der Länge $\frac{t}{n}$.

$$\begin{aligned}\mathbb{R}_{>0} &= \left(0, 1\cdot\frac{t}{n}\right] \cup \left(1\cdot\frac{t}{n}, 2\cdot\frac{t}{n}\right] \cup \ldots \cup \left((n-1)\cdot\frac{t}{n}, n\cdot\frac{t}{n}\right] \cup \ldots \\ &= E_1 \cup E_2 \cup \ldots \cup E_n \cup \ldots\end{aligned}$$

Wir modellieren die zufällige **Anzahl** X der Epochen, die

ohne Erfolg vergehen. Wir unterstellen, dass X geometrisch verteilt mit Erfolgswahrscheinlichkeit p ist und setzen $W := X \cdot \frac{t}{n}$. Dann erfasst W die Zeit bis zum ersten Erfolg; wobei die Zeit nur in diskreten Schritten erfasst wird.

Wir betrachten

$$W > t \Leftrightarrow X \cdot \frac{t}{n} > t \Leftrightarrow X > n.$$

Wir betrachten die Wahrscheinlichkeit, dass der Erfolg irgendwann nach mehr als n Perioden eintritt. Also

$$\mathbb{P}^{\mathrm{geo}(n)}(W > t) = \mathbb{P}(X > n) = (1-p)^n.$$

Wir wählen jetzt $p = \lambda\frac{t}{n}$; wenn $\lambda\frac{t}{n}$ nicht kleiner 1 ist, dann müssen wir ein hinreichend großes n wählen. Diese Wahrscheinlich ist proportional zur Länge der Epochen $\frac{t}{n}$, wobei der Proportionalitätsfaktor λ ist. Der Proportionalitätsfaktor erfasst die **Intensität** mit der der Erfolg eintritt. Dann ist

$$\begin{aligned}\mathbb{P}^{\mathrm{geo}(n)}(W > t) &= \mathbb{P}(X > n) = (1-p)^n \\ &= \left(1 - \frac{\lambda t}{n}\right)^n \to e^{-\lambda t} = \mathbb{P}^{\mathrm{EXP}(\lambda)}((t,\infty)).\end{aligned}$$

Also erhalten wir in der Tat die Exponentialverteilung, wenn wir eine Sequenz von mit n indizierten Experimenten mit geometrischer Verteilung betrachten.

3.3.8 Bemerkung: Eine bemerkenswerte Eigenschaft der Exponentialverteilung ist die sogenannte **Gedächtnislosigkeit**

$$\mathbb{P}(X \geq t + s | X \geq s) = \mathbb{P}(X \geq t).$$

3.3.9 Definitionen: Es sei $\tau_1, \tau_2, ...$ eine Folge von unabhängigen Zufallsvariablen mit $\tau_i \sim \text{Exp}(\lambda)$. Die **Ankunftszeit** für n Sprünge ($n \in \mathbb{N}$) ist

$$S_n = \sum_{k=1}^{n} \tau_k.$$

Die τ_i heißen dann **Zwischenankunftszeiten**.

Für $t \geq 0$ zählt die **Zufallsvariable**

$$N_t = \begin{cases} 0 & \text{falls } 0 \leq t < S_1 \\ k & \text{falls } S_k \leq t < S_{k+1} \end{cases}$$

die Sprünge (Ankünfte) bis t. Wir bemerken, dass die Pfade von $N_t, t \geq 0$ von **rechts-stetig** sind.

Man sagt, dass N_t ein Poisson Prozess mit **Intensität** λ ist.

3.3.10 Satz: N_t hat eine Poisson Verteilung:

$$\mathbb{P}(N_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad k = 0, 1, ...$$

Beweis: vgl. Shreve [33]

3.4 Transformationen

▶ Die folgenden beiden Sätze beschäftigen sich mit **Transformationen Zufallsvariablen mit Dichten**. Transformationen von Zufallsvariablen ergeben sich sehr oft; insbesondere werden viele in der Statistik zentralen Verteilungen durch Transformation anderer - z.B. der Normalverteilung – Verteilungen definiert. Eine ausführliche Darstellung einschließlich vieler hier nicht erläuterter technischer Einzelheiten findet man in Tappe [36, Kapitel 8]. Die Beispiele aus Tappe [36, Kapitel 8] bearbeiten wir als Übungsaufgaben. Eine weitere besonders interessante Quelle ist Blitzstein und Hwang [4]

3.4.1 Satz: Es sei X ein zufälliges Ereignis mit Werten im Intervall $I = (a, b), a < b$ und Dichte f_X; also $X \sim f_X$. Es sei ferner $g : (a, b) \to \mathbb{R}$ eine streng monoton wachsende C^1-Funktion; mit h bezeichnen wir die Inverse von g auf $g(I)$. Dann ist

$$f_T(t) = f_X(h(t))|h'(t)|$$

die Dichte von $T = g(X)$, d.h. $T \sim f_T$. Beweis: Tappe [36, Seite 180f]

3.4.2 Korrelar: Es sei $X \sim f_X$ und $a, b \in \mathbb{R}, a \neq 0$. Dann

hat die Transformation $Y = aX + b$ die Dichte

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

Beweis: Tappe [36, Seite 181]

3.4.3 Satz: Es sei X eine Zufallsvariable mit Dichte $f_X \sim X$ und mit Werten in $I = I_1 \biguplus I_2$, wobei die I_i disjunkte offene reelle Intervalle sind. Ferner sei $g : I \to \mathbb{R}$ eine Funktion mit $g \in C^1(I_i)$ mit $g'(x) \neq 0$ für $x \in I_i$. Dann ist

$$f_T(t) = \mathbb{1}_{g(I_1)}(t) f_X(h_1(t)) \, |h_1'(t)| + \mathbb{1}_{g(I_2)}(t) f_X(h_2(t)) \, |h_2'(t)|$$

die Dichte von $T = g(X)$, wobei $h_k := g_k^{-1}$, $g_k = g_{|I_k}$

3.5 Gamma, Chi und Beta-Verteilung

▶ Das Skript zu den folgenden vier Verteilungen ist noch sehr fragmentarisch. Vergleiche für die **Gamma**-Verteilung, die **Chi-Quadrat**-Verteilung, die t-Verteilung und die **Beta-Verteilung** die entsprechenden Abschnitte in Blitzstein und Hwang [4] und Henze [19].

3.5.1 Definition: Wir definieren die **Gamma-Funktion**

$$\Gamma(a) = \int_0^\infty x^a \frac{e^{-x}}{x} dx$$

Für alle $a > 0$ gilt $\Gamma(a+1) = a\Gamma(a)$ und insbesondere für $n \in \mathbb{N}$ gilt $\Gamma(n) = n!$.

Siehe Blitzstein und Hwang [4, Seite 357]

3.5.2 Satz und Definition: Es seien $\alpha, \lambda > 0$. Dann ist

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{falls } x > 0 \\ 0 & \text{sonst} \end{cases}$$

eine Dichte. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ mit dieser Dichte heißt **Gammaverteilung** mit Parametern λ, α (siehe z.B. Tappe [36, S. 59], Henze [19, S. 157]) und Blitzstein [4, Seite 357].

Bei $\lambda = 1, \alpha > 0$ spricht man von einer **Standard-Gammaverteilung**.

Bei $\alpha = \frac{n}{2}, \lambda = \frac{1}{2}$ mit $n \in \mathbb{N}$ spricht man von einer **Chi-Quadrat-Verteilung** mit n Freiheitsgraden (FG) (siehe Georgii [13, S.277]).

3.5.3 Bemerkung: Die Gammaverteilung mit Parameter $0 < \lambda, n, \alpha = n \in \mathbb{N}$ beschreibt die *Wartezeit* bis n Erfolge bzw. Schäden eingetreten sind (siehe Georgii [13, Seite 47] oder Blitzstein [4, Seite 361]). Für $n = 1$ erhält man die Exponentialverteilung. In der Tat gilt: Wenn $X_1, X_2, ...X_n \sim$

i.u.Exp(λ) sind, dann gilt

$$X_1 + X_2 + ... + X_n \sim \text{Gamma}(n, \lambda).$$

3.5.4 Bemerkung: Wenn $X_i \sim N(0,1), i = 1, .., k$, dann ist $X_1^2 + X_2^2 + ... + X_k^2 \sim \chi^2$ mit k Freiheitsgraden. Also eine Gammaverteilung mit $\alpha = \frac{k}{2}, \lambda = \frac{1}{2}$. Die Verteilung ist in der **Statistik** sehr oft relevant.

Siehe Blitzstein [4, Seite 477 ff] und Henze [19, Seite 157].

3.5.5 Definition: t-Verteilung ..... Henze [19, Seite 247] und Blitzstein [4, Seite 477 ff]

3.5.6 Definition: Beta-Verteilung ..... Blitzstein [4, Seite 351 ff]

4 Verteilungs- und Quantilfunktionen

Verteilungsfunktionen werden in allen Bücher zur Wahrscheinlichkeitstheorie behandelt; also auch in Henze [19]. Quantilsfunktionen werden dort auch behandelt und vergleichsweise detailliert in Föllmer und Schied [12] untersucht. Quantile und Quantilsfunktionen sind insbesondere für die **Finanzmathematik (Risikomanagement)** von SEHR großer Bedeutung. In der Tat ist das bekannteste Risikmaß – der Value at Risk – ein Quantil.

4.1 Verteilungsfunktion

4.1.1 Definition: Eine Funktion $F : \mathbb{R} \to [0, 1]$ heißt **Verteilungsfunktion**, falls:

- F ist nicht fallend, d.h. $x \leq y$ impliziert $F(x) \leq F(y)$.
- F ist von rechts stetig, d.h. $\lim_{y \to x, y > x} F(y) = F(x)$.
- $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$.

4.1.2 Satz: Es sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$. Dann ist

$$F = \begin{cases} \mathbb{R} \to [0,1] \\ x \mapsto \mathbb{P}((-\infty, x]) \end{cases}$$

eine **Verteilungsfunktion**. $F(x) = \mathbb{P}((-\infty, x])$ ist also die Wahrscheinlichkeit, beim Experiment $(\Omega, \mathcal{B}, \mathbb{P})$ ein Ergebnis kleiner oder gleich x zu beobachten.

4.1.3 Satz: Es sei $F : \mathbb{R} \to [0,1]$ eine Verteilungsfunktion. Dann gibt es genau ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $(\mathbb{R}, \mathcal{B})$, so dass $\mathbb{P}((-\infty, x]) = F(x)$ für alle $x \in \mathbb{R}$ gilt.

4.1.4 Satz: i.) Es sei $F : \mathbb{R} \to [0,1]$ eine Verteilungsfunktion. Dann existiert für alle $x \in \mathbb{R}$ der linksseitige Grenzwert

$$\lim_{y \to x, y < x} F(y) =: F(x-).$$

Für diesen Grenzwert schrieben wir $F(x-)$.

Beweis: Vgl. Rudin [30, Seite 95]

ii.) Es sei F eine Verteilungsfunktion, dann ist die Menge der Unstetigkeitsstellen höchstens abzählbar unendlich.

4.1.5 Bemerkung: Die Unstetigkeitsstellen einer Verteilungsfunktion sind nicht notwendigerweise *isoliert* (vgl. Rudin [30, Seite 97]).

4.1.6 Satz: Es sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$ und $F : \mathbb{R} \to [0,1]$, die durch $\mathbb{P}$ wie in (4.1.2) definierte Verteilungsfunktion mit $F(x) = \mathbb{P}((-\infty, x])$. Dann gilt für alle $x, y \in \mathbb{R}, x < y$:

i.) $F(x) = \mathbb{P}((-\infty, x])$

ii.) $F(x-) = \mathbb{P}((-\infty, x))$

iii.) $F(x) = 1 - \mathbb{P}((x, \infty))$

iv.) $F(x-) = 1 - \mathbb{P}([x, \infty))$

v.) $\mathbb{P}(\{x\}) = F(x) - F(x-)$ [Sprunghöhe bei x]

vi.) $\mathbb{P}((x, y]) = F(y) - F(x)$

vii.) $\mathbb{P}([x, y]) = F(y) - F(x-)$

viii.) $\mathbb{P}((x, y)) = F(y-) - F(x)$

ix.) $\mathbb{P}([x, y)) = F(y-) - F(x-)$

4.1.7 Satz: Es sei $f : \mathbb{R} \to \mathbb{R}$ eine Dichte. Dann ist F :

$\mathbb{R} \to [0,1]$ mit

$$F(x) = \int_{-\infty}^{x} f(z)dz$$

eine Verteilungsfunktion.

Man spricht von der Verteilungsfunktion F mit der **Dichte** f.

4.1.8 Satz: Es sei $\Omega \subset \mathbb{R}$ höchstens abzählbar, $\mathcal{F} = \mathcal{P}(\Omega)$ und $\mathbb{P}$ ein **diskretes** Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{F})$ mit Zähldichte p. Dann ist $F : \mathbb{R} \to [0,1]$ mit

$$F(x) = \sum_{k \in (-\infty, x] \cap \Omega} p_k$$

eine Verteilungsfunktion. Das ist eine Treppenfunktion mit Stufen an den Stellen k mit Stufenhöhe p_k.

4.1.9 Bemerkung: Es gibt auch Verteilungsfunktionen, die weder eine Dichte haben noch diskret sind: Verteilungsfunktionen mit Sprungstellen, die aber keine (reinen) Treppenfunktionen sind.

4.2 Quantile, Quantilsfunktionen und Verallgemeinerte Inverse

4.2.1 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ eine Wahrscheinlichkeitsraum, X eine reellwertige Zufallsvariable mit Verteilungsfunktion F und $\lambda \in (0,1)$. Eine reelle Zahl $q^X(\lambda) = q(\lambda) = x_\lambda$ heißt $\lambda-$**Quantil**[1] der Zufallsvariable X, falls

$$\mathbb{P}(X \leq x_\lambda) \geq \lambda \quad \textbf{und}$$
$$\mathbb{P}(X \geq x_\lambda) \geq 1 - \lambda \ .$$

In Worten:

i.) Die Wahrscheinlichkeit, dass X kleiner-gleich der Grenze x_λ ist, ist mindestens λ **und**

ii.) die Wahrscheinlichkeit, dass X größer-gleich der Grenze x_λ ist, ist mindestens $1 - \lambda$.

▶ Bei einem Jahrhundertereignis ist $\lambda = \frac{1}{100}$. Wenn das 0.01-Quantil $x_\lambda = -10000$ beträgt, dann bedeutet das: Einen Wert von -10000 oder noch geringer beobachten wir statistisch mindestens 1-mal in Hundert Jahren und einen Wert von -10000 oder mehr, beobachten wir statistisch mindestens 99-mal in Hundert Jahren.

[1] Je nach Zusammenhang verwenden $q, q^X(\lambda), q(\lambda), x_\lambda$ für ein $\lambda-$**Quantil**

▶ Das ist vielleicht eine/die Gelegenheit in z.B. Fahrmeir et al. [11, Seite 59 ff] die Abschnitte zu empirischen Quantilen nachzulesen.

4.2.2 Definition: Es sei $0 < \lambda < 1$ und $x_1, ..., x_n$ eine Urliste. Jeder Wert x_λ mit:

i.) Mindestens der Anteil λ der Daten ist kleiner/gleich x_λ und

ii.) mindestens der Anteil $1-\lambda$ der Daten ist größer/gleich x_λ.

heißt **empirisches λ-Quantil**. Also

$$\frac{\#(x\text{-Werte} \leq x_\lambda)}{n} \geq \lambda \quad \textbf{und}$$
$$\frac{\#(x\text{-Werte} \geq x_\lambda)}{n} \geq 1-\lambda$$

Wir beachten die Analogie zur Definition 4.2.1 eines Quantils einer Zufallsvariable

4.2.3 Beispiel: Urliste = 2, 4, 4, 5, 6, 6, 7, 8, 8, 9. Es ist also $n = 10$.

Es gilt $x_{0.05} = 2$.

Es gilt $x_{0.25} = 4$. Es ist

$$\frac{\#(x\text{-Werte} \leq 4)}{10} = \frac{3}{10} = \frac{6}{20} \geq \frac{5}{20} = 0.25 \text{ und}$$
$$\frac{\#(x\text{-Werte} \geq 4)}{10} = \frac{9}{10} = \frac{18}{20} \geq \frac{15}{20} = 0.75$$

Es gilt $x_{0.3} \in [4, 5]$. Es ist z.B.

$$\frac{\#(x\text{-Werte} \leq 4.7)}{10} = \frac{3}{10} \geq \frac{3}{10} = 0.30 \text{ und}$$
$$\frac{\#(x\text{-Werte} \geq 4.7)}{10} = \frac{7}{10} \geq \frac{7}{10} = 0.70$$

Also ist 4.7 ein 0.3-Quantil.

Es gilt $x_{0.5} = 6$.

Überzeugen Sie sich, dass man empirische Quantile besonders leicht an der empirischen Verteilungsfunktion *ablesen* kann.

Zwei Beobachtungen sind merkwürdig. i.) Wenn λ sehr klein ist (hier kleiner als 0.1), dann entspricht das λ Quantil dem kleinsten Beobachtungswert. ii.) Für $\lambda = 0.3$ sind alle Werte in einen angeschlossenen Intervalle 0.3-Quantile. Mindestens drei Werte des Intervalls werden regelmäßig verwendet, nämlich das Minimum, die Mitte und das Maximum. In der Tat gibt es gemäß Hyndman and Fan [20] in Statistikpaketen 9 Varianten, um das Quantil einer Zufallsvariable auf Basis eines Datensatzes zu schätzen

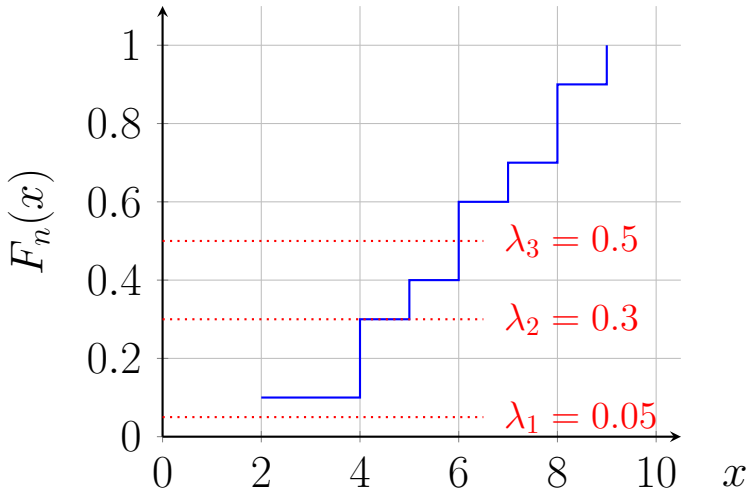


Abbildung 4.1: An der empirischen Verteilungsfunktion kann man Quantile leicht ablesen. Man trägt die λ-Linien ein. Wo diese die VF treffen, lotet man nach unten.

4.2.4 Bemerkung: Es gilt (wir beginnen mit der zweiten Bedingung aus der Definition 4.2.1)

$$\begin{aligned} & 1-\lambda \leq \mathbb{P}(X \geq x_\lambda) \\ \Leftrightarrow\ & 1-\mathbb{P}(X \geq x_\lambda) \leq \lambda \\ \Leftrightarrow\ & F(x_\lambda^-) = \mathbb{P}(X < x_\lambda) \leq \lambda \end{aligned}$$

Dementsprechend ist x_λ ein $\lambda-$**Quantil** der Zufallsvariable X, falls

$$\begin{aligned} & F(x_\lambda) = \mathbb{P}(X \leq x_\lambda) \geq \lambda \textbf{ und} \\ & F(x_\lambda^-) = \mathbb{P}(X < x_\lambda) \leq \lambda. \end{aligned}$$

▶ Bei einem Jahrhundertereignis ist $\lambda = \frac{1}{100}$. Wenn das 0.01-

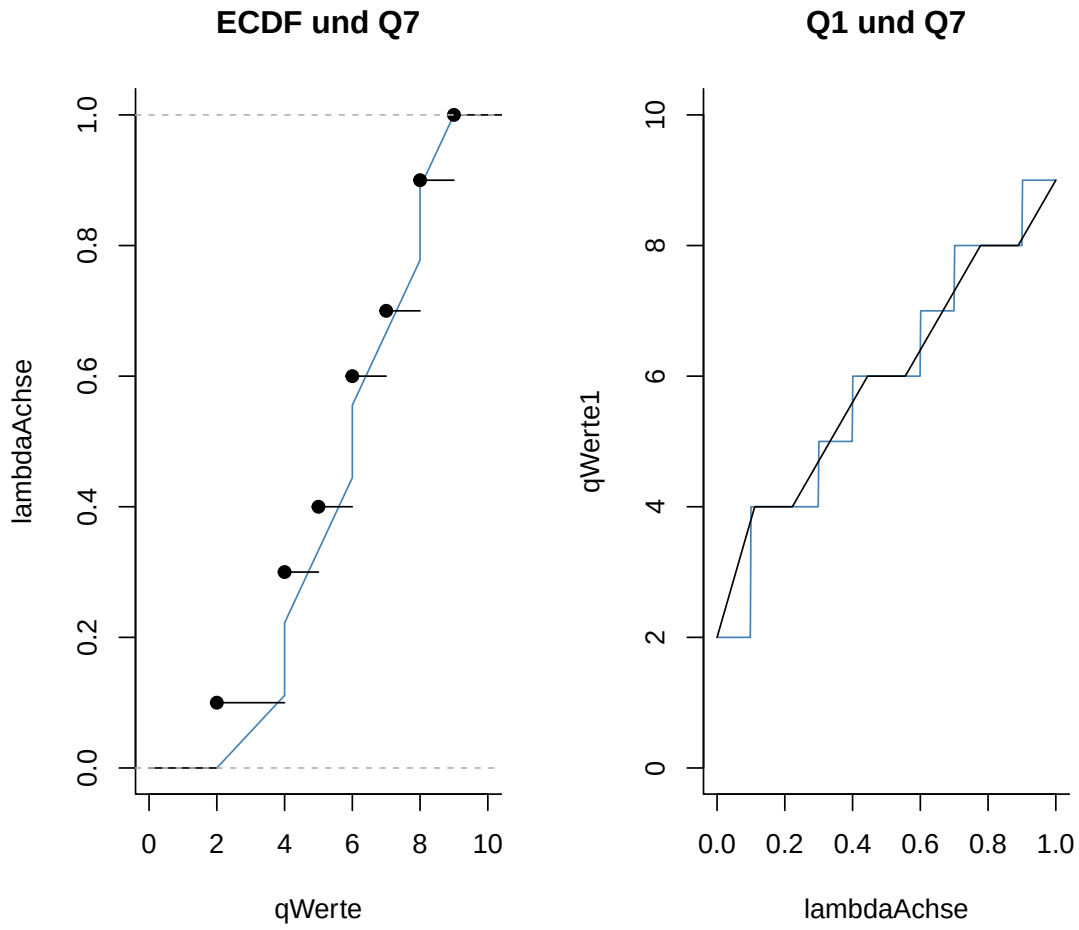


Abbildung 4.2: Die empirische Verteilungsfunktion und die Q7-Quantile im linken Panel. Im rechten Panel die Q1 und die Q7 Quantile.

Quantil $x_\lambda = -10000$ beträgt, dann bedeutet das: Einen Wert von -10000 oder noch geringer beobachten wir statistisch mindestens 1-mal in Hundert Jahren und einen Wert strickt kleiner als -10000, beobachten wir statistisch höchstens 1-mal in Hundert Jahren.

▶ Wenn $F(x_\lambda) = F(x_\lambda^-)$, dann $F(x_\lambda) = \lambda$.

4.2.5 Satz: Es sei X eine Zufallsvariable und $X \sim F$. Es sei $\lambda \in (0,1)$. $q \in \mathbb{R}$ ist genau dann ein λ-Quantil von X, wenn

$$F(q-) \leq \lambda \leq F(q+) = F(q)$$

ist.

4.2.6 Definition: Es sei F eine Verteilungsfunktion. Dann heißt $(0,1) \ni \lambda \mapsto q^+(\lambda)$ mit

$$\begin{aligned} q^+(\lambda) &= \inf\{q \in \mathbb{R} : F(q) > \lambda\} \\ &= \sup\{q \in \mathbb{R} : F(q) \leq \lambda\} \end{aligned}$$

die **obere Quantilsfunktion** von F und $(0,1) \ni \lambda \mapsto q^-(\lambda)$

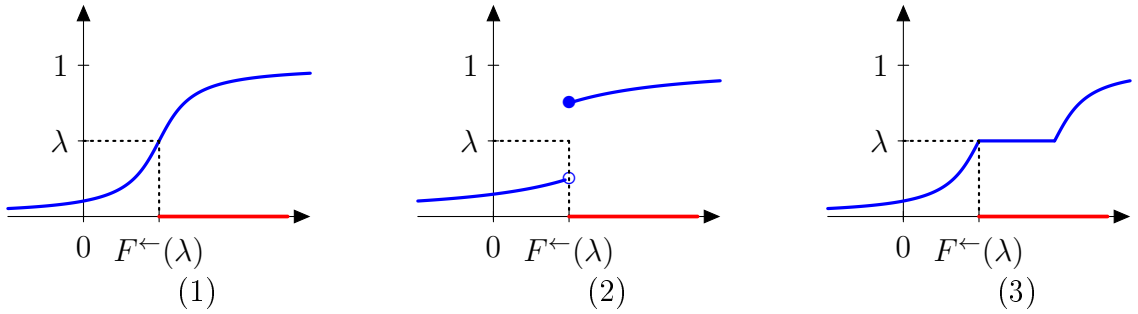
$$\begin{aligned} q^-(\lambda) &= \sup\{q \in \mathbb{R} : F(q) < \lambda\} \\ &= \inf\{q \in \mathbb{R} : F(q) \geq \lambda\} \end{aligned}$$

die **untere Quantilsfunktion** von F.

Insbesondere im **Risikomanagement** nennt man die untere Quantilsfunktion auch **Verallgemeinerte Inverse** und schreibt

$$F^{\leftarrow}(\lambda) = q^{-}(\lambda)$$

Drei Situationen können dabei auftreten, die in den Panels (1) bis (3) illustriert sind.



4.2.7 Bemerkung: Wie sieht die Menge $A := \{q \in \mathbb{R} : F(q) \geq \lambda\}$ aus? Wenn $q_1 \in A$ ist und $q_2 > q_1$ gilt, dann ist $F(q_2) \geq F(q_1) \geq \lambda$. Also ist auch $q_2 \in A$. Also ist A ein Intervall der Form (z, ∞) oder der Form $[z, \infty)$. In der Tat hat das Interval wegen der von-rechts-Stetigkeit von F die Form $[z, \infty)$. Also gilt

$$\begin{aligned} q^{-}(\lambda) &= \inf\{q \in \mathbb{R} : F(q) \geq \lambda\} \\ &= \min\{q \in \mathbb{R} : F(q) \geq \lambda\}. \end{aligned}$$

Wir können also anstatt des Infimums das Minimum bilden. Das Minimum wird angenommen: Es gibt also ein q^* mit der

Eigenschaft $F(q^*) \geq \lambda$ und für alle $q < q^*$ gilt $F(q^*) < \lambda$.

Beweis: Angenommen $z = \inf\{q \in \mathbb{R} : F(q) \geq \lambda\}$, aber $F(z) \geq \lambda$ gilt nicht. Also $F(z) < \lambda$. Wir betrachten eine Folge (q_i) mit $q_i > z$ und $q_i \to z$ (eine Folge die von links gegen z konvergiert). Wegen der Stetigkeit von rechts für F gilt $F(q_i) \to F(z)$. Dann muss es – da $F(q) < \lambda$ – ein j mit $F(q_j) < \lambda$ und $q_j > z$ geben. Das ist jedoch ein Widerspruch, denn für alle q im Intervall (z, ∞) gilt $F(q) \geq \lambda$.

4.2.8 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und X eine reellwertige Zufallsvariable mit Verteilungsfunktion F.

(1) Dann ist q^- nicht-fallend und von links stetig und

(2) q^+ nicht-fallend von rechts stetig.

(3) Es gilt $q^-(\lambda) \leq q^+(\lambda)$.

Beweis: Siehe Föllmer und Schied [12, S. 538f]

4.2.9 Satz: Es sei F eine Verteilungsfunktion. Für jedes $\lambda \in (0, 1)$ ist **die Menge der λ-Quantile das abgeschlossene Intervall** $[q^-(\lambda), q^+(\lambda)]$.

4.2.10 Bemerkung: Wenn F eine strikt monotone stetige

Verteilungsfunktion ist, dann ist für alle $\lambda \in (0,1)$:

$$q^+(\lambda) = q^-(\lambda) = F^{\leftarrow}(\lambda).$$

4.2.11 Lemma: Es sei F eine Verteilungsfunktion und

$$F^{\leftarrow}(\lambda) = q^-(\lambda) = \inf\{x|F(x) \geq \lambda\} = \min\{x|F(x) \geq \lambda\}$$

die untere Quantilsfunktion bzw. **eine Verallgemeinerte Inverse** von F. Dann gilt für $x \in \mathbb{R}$, $\lambda \in (0,1)$

$$F(x) \geq \lambda \Leftrightarrow x \geq F^{\leftarrow}(\lambda).$$

Beweis (Siehe Henze [19, 151]): Für $\Rightarrow$ müssen wir nur bemerken, dass $\tilde{x} = F^{\leftarrow}(\lambda) = \min\{x|F(x) \geq \lambda\}$ gemäß Definition das kleinste x mit $F(x) \geq \lambda$ ist. Also $x \geq \tilde{x} = F^{\leftarrow}(\lambda)$.

Für $\Leftarrow$ nehmen wir an, dass $x \geq F^{\leftarrow}(\lambda) = \tilde{x}$ gibt, aber $F(x) < \lambda$ ist. Wir betrachten eine Folge (x_i) mit $x_i > x$ und $x_i \to x$. Wegen der Stetigkeit von rechts von F gilt $F(x_i) \to F(x)$. Aus der Konvergenz $F(x_i) \to F(x)$ und $F(x) < \lambda$ folgt, dass es ein $x_j > x \geq \tilde{x}$ mit $F(x_j) < \lambda$ gibt. Es gäbe also $x_j > x \geq F^{\leftarrow}(\lambda)$ mit $F(x_j) < \lambda$. Das ist aber ein Widerspruch, denn aus $x_j \geq \tilde{x}$ folgt $F(x_j) \geq F(\tilde{x}) = \lambda$.

4.2.12 Lemma: Es sei F eine Verteilungsfunktion und $U \sim$

Unif((0,1)). Dann hat die Zufallsvariable

$$X = F^{\leftarrow}(U)$$

die Verteilungsfunktion F; $X \sim F$.

Beweis: Siehe Henze [19, 153]. Es gilt gemäß des obigen Lemmas

$$\begin{aligned} \mathbb{P}(X \leq x) &\leq \mathbb{P}(F^{\leftarrow}(U) \leq x) && X = F^{\leftarrow}(U) \\ &= \mathbb{P}(U \leq F(x)) && \text{Lemma} \\ &= F(x). && \text{VF von Unif} \end{aligned}$$

▶ Auf der Grundlage des vorhergehenden Lemmas, kann man Zufallszahlen gemäß einer Verteilung F erzeugen, wenn man Zufallszahlen gemäß einer Gleichverteilung erzeuge kann.

4.2.13 Lemma: Es sei X eine Zufallsvariable mit Verteilungsfunktion F. Ferner sei F stetig. Dann gilt $F(X) \sim$ Unif($[0,1]$). Also für $u \in [0,1]$

$$\mathbb{P}(F(X) \leq u) = u$$

Beweis: Siehe Henze [19, 153].

4.3 Risikomessung

▶ Der Rest des Paragrafen ist als Vertiefung bzw. Anwendung für Risikomanager gedacht und noch arg fragmentarisch.

4.3.1 Terminologie und Einführung: Im folgenden stellen wir uns vor, dass ein Investor oder Manager das Risiko einer Vermögensposition analysiert. Den Wert dieser Vermögensposition bezeichnen wir mit V. Wir stellen uns ferner vor, dass der Wert in mindestens zwei Zeitpunkten betrachtet wird. Wir verwenden $t = 0$ für den Zeitpunkt zu dem der Vermögenswert fix und bekannt ist. Wir bezeichnen diesen Wert mit V_0. Den Wert V der Vermögensposition zu einem zukünftigen Zeitpunkt sehen wir als eine Zufallsvariable an, dessen Wert uns für den Zeitpunkt $t = 1$ interessiert. Wir verzichten manchmal bei der Variable V für den Zeitpunkt $t = 1$ auf den Index, d.h. wir schreiben einfach V anstatt V_1. Wir untersuchen meistens nicht V, sondern die **Veränderung** $G = V - V_0$ oder den **Verlust** $L = V_0 - V$. Trotzdem sprechen wir vom Value at Risk der Vermögensposition. G bezeichnet einen Gewinn (der natürlich auch negativ sein kann). Da wir Risikonanalyse betreiben, werden wir – der Literatur folgenden – oft die Variable $L = -G$ betrachten; also den Verlust.

4.3.2 Defintion: Es sei V ein Vermögenswert und $G = V - V_0$ und $L = -G$. Dann ist der **Value-at-Risk (V@R)** von V zur **Risikotoleranz** λ (eine kleine Zahl, z.B. $\lambda = 0.01$) die reelle Zahl

$$\begin{aligned} \text{V@R}_\lambda(G) &= -q^+_{F_G}(\lambda) \\ &= q^-_{F_{-G}}(1-\lambda) \\ &= q^-_{F_L}(1-\lambda). \end{aligned}$$

Wenn wir Verluste betrachten (α nahe 1, z.B. $\alpha = 0.99$)

$$\begin{aligned} \text{V@R}_\alpha(L) &= q^-_{F_L}(\alpha) \\ &= F_L^{\leftarrow}(\alpha) \\ &= \inf\{x | F_L(x) \geq \alpha\} \\ &= \min\{x | F_L(x) \geq \alpha\} \end{aligned}$$

dabei heißt α das **Sicherheitsniveau (Konfidenzniveau)**.

Die Abbildung

$$F^{\leftarrow}(\lambda) = \inf\{x | F(x) \geq \lambda\} = \min\{x | F(x) \geq \lambda\}$$

heißt die **Verallgemeinerte Inverse** von F

► In diesem Text ist λ typischerweise eine kleine Zahl nahe 0. λ bezeichne ich als **Risikotoleranz**. Wenn $\lambda = 0.01$ ist, dann toleriert man, dass es einmal in Hundert Jahren

noch schlimmer als x_λ kommt. Dieses Risiko wird toleriert. $1-\lambda = \alpha$ ist eine Zahl nahe 1. $1-\lambda$ bezeichne ich als **Sicherheitsniveau**, denn mit einer Wahrscheinlichkeit von 0.99 bin ich mir sicher, dass der Wert größer/gleich x_λ ist.

4.3.3 Satz: Es gilt

$$\mathrm{V@R}_\lambda(G) = \inf\{m \,|\, \mathbb{P}(G+m<0) \leq \lambda\}.$$

bzw.

$$\mathrm{V@R}_\alpha(L) = \inf\{m | \mathbb{P}(L-m>0) \leq 1-\alpha\}$$

Beweis: Es gilt gemäß Definition

$$\mathrm{V@R}_\lambda(G) = q^-_{F_L}(1-\lambda) = \inf\{m \in \mathbb{R} : F_L(m) \geq 1-\lambda\}.$$

Wir zeigen, dass $F_L(m) \geq 1-\lambda$ gdw $\mathbb{P}(G+m<0) \leq \lambda$. Dann muss natürlich auch $\inf\{m \in \mathbb{R} : F_L(m) \geq 1-\lambda\} = \inf\{m \,|\, \mathbb{P}(G+m<0) \leq \lambda\}$ gelten.

Für m gilt $F_L(m) = \mathbb{P}(L \leq m)$. Also gilt für m

$$\begin{aligned} & F_L(m) \geq 1-\lambda \\ \Leftrightarrow \quad & \mathbb{P}(L \leq m) \geq 1-\lambda \\ \Leftrightarrow \quad & \lambda \geq 1-\mathbb{P}(L \leq m) \end{aligned}$$

und

$$1 - \mathbb{P}(L \leq m) = \mathbb{P}(L > m) = \mathbb{P}(G < -m) = \mathbb{P}(G + m < 0).$$

Also gilt für m

$$F_L(m) \geq 1 - \lambda \Leftrightarrow \mathbb{P}(G + m < 0) \leq \lambda.$$

Damit haben wir den ersten Teil des Satzes bewiesen.

Für den zweiten Teil beobachten wir

$$\begin{aligned} & G + m < 0 \\ \Leftrightarrow & -L + m < 0 \\ \Leftrightarrow & m < L \\ \Leftrightarrow & 0 < L - m. \end{aligned}$$

Also auch

$$\mathbb{P}(G + m < 0) = \mathbb{P}(0 < L - m).$$

Also auch

$$\inf\{m \,|\, \mathbb{P}(G + m < 0) \leq \lambda\} = \inf\{m \,|\, \mathbb{P}(0 < L - m) \leq \lambda\}.$$

4.3.4 Bemerkung: (1) Wir reden vom Value at Risk von V, obwohl die Defintion auf G bzw. L Bezug nimmt. Wir werden auch vom Value at Risk von G oder von L sprechen. Gemeint

ist das negative des oberen λ-Quantil der Zufallsvariable G bzw. das untere $\alpha = (1 - \lambda)$-Quantil des Verlustes L.

(2) Der Value at Risk wird in den Einheiten gemessen, in denen die Zufallsvariable G gemessen wird, d.h. in der Regel in **Geldeinheiten**.

(3) Typischerweise wird das für die Risikonanalyse relevante λ-Quantil $q^+_{F_G}(\lambda)$ von G negativ sein.

(4) Der Value at Risk ist wegen der Multiplikation mit -1 so definiert, dass er die Grundlage für eine Kapitalanforderung ist; das wird weiter unter noch verdeutlicht.

(5) Wenn man anstatt des Zuwachses G die Verlustvariable L betrachtet, dann sind für den Übergang **drei Anpassung** nötig: (i) Anstatt der oberen Quantilsfunktion betrachtet man die untere Quantilsfunktion. (ii) Anstatt der Risikotoleranz λ berachtet man das **Sicherheitsniveau** $\alpha = 1-\lambda$. (iii) Die Multiplikation mit -1 entfällt.

(6) Es ist $G + m = -L + m$. Also ist $G + m < 0$ genau dann, wenn $m < L$. Wenn wir m als **Eigenkapital** auffassen, dann erkennen wir, dass $\mathrm{V@R}_\lambda(G)$ der kleinste Eigenkapitalwert ist, der ausreicht eine Insolvenz – hier definiert als Verluste größer als das Eigenkapital – mit Wahrscheinlichkeit λ zu vermeiden. Man kann also den V@R im Kontext der **Eigenkapitalregulierung** unmittelbar anwenden. Der

V@R ist auch aus diesem Grund populär.

4.3.5 Beispiel: Es sei $G \sim N(\mu, \sigma^2)$ Normalverteilt und die Risikotoleranz $\lambda \in (0, 1)$.

$$\begin{aligned} \mathrm{V@R}_\lambda(G) &= -\mathbb{E}(G) - \Phi^{-1}(\lambda) \cdot \sigma && (4.1) \\ &= -\mu - \Phi^{-1}(\lambda) \cdot \sigma && (4.2) \end{aligned}$$

Für $\lambda = 0.01$ erhalten wir $\Phi^{-1}(0.01) = -2.33$, so dass $\mathrm{V@R}_\lambda(X) = -\mu + 2.33 \cdot \sigma$. Für $\lambda = 0.0001$ ist $\mathrm{V@R}_\lambda(X) = -\mu + 3.09 \cdot \sigma$. Praktisch: Wenn wir den Erwartungswert und die Varianz von X geschätzt haben und die Annahme der Normalverteilung angemessen ist, dann können wir für den Fall einer Normalverteilung den V@R leicht ermitteln.

4.3.6 Satz: Der Vermögenswert V sei eine Zufallsvariable und $V_0 \neq 0$ eine reelle Zahl. Ferner sei $R = \frac{V - V_0}{V_0}$ eine Zufallsvariable mit endlicher Varianz σ^2 und Erwartungswert μ und die Verteilungsfunktion $F_{R^{sz}}$ der Zufallsvariable $R^{sz} = \frac{R - \mu}{\sigma}$ sei strikt monoton steigend und stetig. Dann gilt

$$\mathrm{V@R}_\lambda(V) = -(\mu + \sigma \cdot F^{-1}_{R^{sz}}(\lambda))V_0.$$

Beweis: Da die Verteilungsfunktion annahmegemäß strikt monoton steigend und stetig ist gilt für den V@R die Glei-

chung

$$\lambda \;=\; \mathbb{P}\left(V - V_0 \leq -\mathrm{V@R}_\lambda(V)\right)$$

Dann folgt

$$\begin{aligned}\lambda \;&=\; \mathbb{P}\left(\frac{R-\mu}{\sigma} \leq -\frac{\mathrm{V@R}_{t,\lambda}(V)}{V_0\sigma} - \frac{\mu}{\sigma}\right)\\ &=\; \mathbb{P}\left(R^{sz} \leq -\frac{\mathrm{V@R}_{t,\lambda}(V)}{V_0\sigma} - \frac{\mu}{\sigma}\right)\end{aligned}$$

Also

$$\begin{aligned}F^{-1}_{R^{sz}}(\lambda) &= -\frac{\mathrm{V@R}_{t,\lambda}(V)}{V_0\sigma} - \frac{\mu}{\sigma}\\ \Rightarrow\; \mathrm{V@R}_\lambda(V) &= -(\mu + \sigma \cdot F^{-1}_{R^{sz}}(\lambda))V_0\end{aligned}$$

4.3.7 Bemerkung: Der Betrachtungszeitpunkt sei $\tau - 1$. Wir wollen eine Risikoeinschätzung für τ ermitteln. In der Praxis geht man regelmäßig so vor.

i.) Zunächst erstellt man für die Rendite ein Zeitreihenmodell der Form

$$R_t = \mu_t + \sigma_t z_t, t \in \mathbb{Z},$$

wobei $\mu_t = \mathbb{E}_{t-1}(R_t), \mathbb{V}_{t-1}(R_t) = \sigma_t^2$. Ferner wird angenommen, dass $(z_t) \sim \mathrm{iid}(0,1)$ mit Verteilungsfunktion $F_{R^{sz}}$ ist ((z_t) ist also striktes **weißes Rauschen**).

ii.) Wir verwenden den Satz (4.3.6) und schätzen den aktu-

ellen Value at Risk mit

$$\text{V@R}_{\lambda,\tau}(V) \;=\; -(\mu_\tau + \sigma_\tau \cdot F^{-1}_{R^{sz}}(\lambda))V_{\tau-1}.$$

Bei dieser Methoden muss man also zwei statistische Probleme lösen:

- Ein Zeitreihenmodell $R_t = \mu_t + \sigma_t z_t$ schätzen; mindestens muss man μ_t und σ_t schätzen.
- Die Verteilungsfunktion $F_{R^{sz}}$ schätzen.

4.3.8 Bemerkung: Value at Risk ist nicht unumstritten. Zwei Eigenschaften sind problematisch:

- V@R ist im Allgemeinen nicht sub-additiv.
- V@R beachtet nicht, wie die Verteilung von X links von V@R gestaltet ist.

4.3.9 Behauptung: Es sei X eine Zufallsvariable mit Verteilungsfunktion $F = F^X$. Dann gilt

$$\mathbb{E}(X) = \int X dF = \int_0^1 F^{\leftarrow}(u)du.$$

Beweis: (den Beweis können wir eigentlich noch nicht würdigen, da wir Integration bezüglich dF und den Umgang mit solchen Lebesgue-Stieltjes-Integralen noch nicht kennen).

Es sei U eine Zufallsvariable mit $U \sim \text{Unif}((0,1))$. Dann hat (auch) die Zufallsvariable $Y = F^{\leftarrow}(U)$ die Verteilungsfunktion F. Also gilt

$$\begin{aligned}\mathbb{E}(X) &= \int Y dF \\ &= \int F^{\leftarrow}(U) dF^U \\ &= \int F^{\leftarrow}(u) du.\end{aligned}$$

4.3.10 Beispiel: Wir betrachten $X \sim$ Bernoulli mit $\text{Bild}(X) = \{-b, a\}$ für $a, b > 0$ und $\mathbb{P}(X = a) = p$. Dann ist für $\lambda \in (0,1)$

$$F^{\leftarrow}(\lambda) = \begin{cases} -b \text{ falls } \lambda \leq 1-p \\ a \text{ falls } \lambda > 1-p \end{cases}$$

Es gilt einerseits

$$\mathbb{E}(X) = p\,a - (1-p)b.$$

Andererseits

$$\begin{aligned}\int F^{\leftarrow}(u) du &= \int_0^{1-p} (-b) du + \int_{1-p}^1 a du \\ &= (-b)(1-p) + a(1-(1-p)) \\ &= (-b)(1-p) + a\,p \\ &= p\,a - (1-p)b.\end{aligned}$$

4.3.11 Bemerkung: Es sei X eine Zufallsvariable mit strikt monoton wachsender C^1-Verteilungsfunktion F und $f = F'$. Dann ist F invertierbar. Wir betrachten die Substitution $u = F(x), x = F^{-1}(u)$ und verwenden die Substitutionsregel der Integration $du = F'(x)dx = f(x)dx$

$$\mathbb{E}(X) = \int xf(x)dx = \int F^{-1}(u)du.$$

4.3.12 Definition: Es sei L eine Zufallsvariable mit $L \sim F$ und $\mathbb{E}(|L|) < \infty$. Wir definieren

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}\int_\alpha^1 F^{\leftarrow}(u)du$$
$$\mathrm{TVaR}_\alpha(L) = \mathbb{E}(L|L > F^{\leftarrow}(\alpha))$$

ES steht für **Expected Shortfall** und TVaR für **Tail-Value-at-Risk**.

▶ Die folgende Bemerkung stellt Resultate über den Zusammenhang zwischen ES und TVaR zusammen. TVaR ist anschaulicher, jedoch nicht sub-additiv. ES ist sub-additiv.

4.3.13 Bemerkung (McNeil et al. [7, Seite 283]): Es sei L eine Verteilungsfunktion und $L \sim F$. Dann gilt

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}\mathbb{E}\left[[L - F^{\leftarrow}(\alpha)]^+\right] + F^{\leftarrow}(\alpha)$$

bzw.

$$\begin{aligned}\mathrm{ES}_\alpha(L) &= \frac{1}{1-\alpha}\left[\mathbb{E}\left[\mathbb{1}_{L>F^{\leftarrow}(\alpha)}\cdot L\right] + F^{\leftarrow}(\alpha)(1-\alpha-\mathbb{P}(L>F^{\leftarrow}(\alpha)))\right]\\ &= \mathbb{E}(L|L>F^{\leftarrow}(\alpha))\cdot\gamma + F^{\leftarrow}(\alpha)\cdot(1-\gamma)\\ &= \mathrm{TVaR}_\alpha(L)\cdot\gamma + \mathrm{VaR}_\alpha(L)\cdot(1-\gamma).\end{aligned}$$

Dabei ist

$$\begin{aligned}\gamma &= \frac{\mathbb{P}(L>F^{\leftarrow}(\alpha))}{1-\alpha}\\ &= \frac{1-\mathbb{P}(L\leq F^{\leftarrow}(\alpha))}{1-\alpha} = \frac{1-F(F^{\leftarrow}(\alpha))}{1-\alpha}.\end{aligned}$$

Ferner gilt

$$\mathbb{E}\left[[L-F^{\leftarrow}(\alpha)]^+\right] = \mathbb{E}\left[\mathbb{1}_{L>F^{\leftarrow}(\alpha)}\cdot L\right] - F^{\leftarrow}(\alpha)\cdot\mathbb{P}(L>F^{\leftarrow}(\alpha)).$$

Wenn F^L an der Stelle $F^{\leftarrow}(\alpha)$ stetig ist, dann gilt

$$\mathrm{ES} = \mathbb{E}(L|L>F^{\leftarrow}(\alpha)) = \mathrm{TVaR}.$$

Beweis: Wir beachten

$$\begin{aligned}\frac{1}{1-\alpha}\mathbb{E}\left[[L-F^{\leftarrow}(\alpha)]^+\right] &= \frac{1}{1-\alpha}\int_0^1 [F^{\leftarrow}(u)-F^{\leftarrow}(\alpha)]^+\,du\\ &= \frac{1}{1-\alpha}\int_\alpha^1 [F^{\leftarrow}(u)-F^{\leftarrow}(\alpha)]^+\,du\\ &= \frac{1}{1-\alpha}\int_\alpha^1 F^{\leftarrow}(u)du - \frac{F^{\leftarrow}(\alpha)}{1-\alpha}\int_\alpha^1 1\,du\\ &= \frac{1}{1-\alpha}\int_\alpha^1 F^{\leftarrow}(u)du - F^{\leftarrow}(\alpha).\end{aligned}$$

Also

$$\begin{aligned}\mathrm{ES}_\alpha(X) &= \frac{1}{1-\alpha}\int_\alpha^1 F^{\leftarrow}(u)du\\ &= \frac{1}{1-\alpha}\mathbb{E}\left[[L-F^{\leftarrow}(\alpha)]^+\right] + F^{\leftarrow}(\alpha)\end{aligned}$$

Weiterhin beachten wir

$$\begin{aligned}\mathbb{E}\left[[L-F^{\leftarrow}(\alpha)]^+\right] &= \mathbb{E}\left[\mathbb{1}_{L>F^{\leftarrow}(\alpha)}\cdot(L-F^{\leftarrow}(\alpha))\right]\\ &= \mathbb{E}\left[\mathbb{1}_{L>F^{\leftarrow}(\alpha)}\cdot L\right] - \mathbb{E}\left[\mathbb{1}_{L>F^{\leftarrow}(\alpha)}\cdot F^{\leftarrow}(\alpha)\right]\\ &= \mathbb{E}\left[\mathbb{1}_{L>F^{\leftarrow}(\alpha)}\cdot L\right] - F^{\leftarrow}(\alpha)\cdot\mathbb{P}(L>F^{\leftarrow}(\alpha))\end{aligned}$$

Also weiterhin

$$\begin{aligned}
\mathrm{ES}_\alpha(X) &= \frac{1}{1-\alpha}\mathbb{E}\left[[L - F^{\leftarrow}(\alpha)]^+\right] + F^{\leftarrow}(\alpha) \\
&= \frac{\mathbb{E}\left[\mathbb{1}_{L>F^{\leftarrow}(\alpha)} \cdot L\right] - F^{\leftarrow}(\alpha) \cdot \mathbb{P}(L > F^{\leftarrow}(\alpha))}{1-\alpha} + F^{\leftarrow}(\alpha) \\
&= \frac{\mathbb{E}\left[\mathbb{1}_{L>F^{\leftarrow}(\alpha)} \cdot L\right] - F^{\leftarrow}(\alpha)\mathbb{P}(L > F^{\leftarrow}(\alpha)) + (1-\alpha)F^{\leftarrow}(\alpha)}{1-\alpha}
\end{aligned}$$

4.3.14 Bemerkung: Man kann den ES als Erwartungswert einer Zufallsvariable eines zwei-stufigen Experiments auffassen: Zunächst mit Wahrscheinlichkeit $1-\gamma$ der fixe Wert $F^{\leftarrow}(\alpha)$ oder mit Wahrscheinlichkeit γ die Zufallsvariable L mit Verteilung unter der Bedingung $L > F^{\leftarrow}(\alpha)$. Dann erhalten wir den Erwartungswert

$$\mathrm{ES} = \gamma \cdot \mathbb{E}(L|L > F^{\leftarrow}(\alpha)) + (1-\gamma) \cdot F^{\leftarrow}(\alpha)$$

mit

$$\gamma = \frac{\mathbb{P}(L > F^{\leftarrow}(\alpha))}{1-\alpha}.$$

4.3.15 Definition und Behauptung (Rockafellar und Uryasev [29, Seite 1449]): Es sei F eine Verteilungsfunk-

tion. Wir definieren

$$F^{\alpha\text{-Tail}}(x) = \begin{cases} 0 & x < F^{\leftarrow}(\alpha) \\ \frac{F(x)-\alpha}{1-\alpha} & x \geq F^{\leftarrow}(\alpha) \end{cases}$$

$F^{\alpha\text{-Tail}}(x)$ ist eine Verteilungsfunktion.

4.3.16 Satz (Rockafellar und Uryasev [29, Seite 1448]): Es sei L eine Verteilungsfunktion und $L \sim F$. Es gilt

$$\begin{aligned} \text{TVaR} &= \int L \, dF^{\alpha\text{-Tail}} \\ &= \mathbb{E}(L|L \text{ im } \alpha\text{-Tail}). \end{aligned}$$

4.3.17 Beispiel: Es sei $\text{Bild}(L) = \{..., 2, 3, 4\}$ mit $\mathbb{P}(L = 2) = \mathbb{P}(L = 3) = \mathbb{P}(L = 4) = 0.02$. Es sei $\alpha = 0.95$. Dann ist $F^{\leftarrow}(\alpha) = 2$. Es ist $\mathbb{P}(L > F^{\leftarrow}(\alpha)) = 0.96$ und $\mathbb{P}(L = F^{\leftarrow}(\alpha)) = 0.02$.

Dann gilt einerseits

$$\begin{aligned} \int L \, dF^{\alpha\text{-Tail}} &= 2 \cdot \frac{0.96 - 0.95}{1 - 0.95} + 3 \cdot \frac{0.98 - 0.96}{1 - 0.95} + 4 \cdot \frac{1 - 0.98}{1 - 0.95} \\ &= 2 \cdot \frac{0.01}{0.05} + 3 \cdot \frac{0.02}{0.05} + 4 \cdot \frac{0.02}{0.05} \\ &= 2 \cdot \frac{1}{5} + 3 \cdot \frac{2}{5} + 4 \cdot \frac{2}{5} = \frac{2 + 6 + 8}{5} \\ &= \frac{16}{5} = 3.2 \end{aligned}$$

Andererseits gilt

$$F^{\leftarrow}(u) = \begin{cases} 4 & \text{für } 0.98 \leq u \\ 3 & \text{für } 0.98 \leq u < 0.96 \\ 2 & \text{für } 0.94 \leq u < 0.96 \\ \ldots & \ldots \quad \ldots \end{cases}$$

Also

$$\begin{aligned} &\frac{1}{1-\alpha}\int_{0.95}^{1} F^{\leftarrow}(u)du \\ &= \frac{1}{0.05}\left(2[x]_{0.95}^{0.96} + 3[x]_{0.96}^{0.98} + 4[x]_{0.98}^{1}\right) \\ &= \frac{1}{0.05}\left(2\cdot(0.96-0.95) + 3\cdot(0.98-0.96) + 4\cdot(1-0.98)\right) \\ &= \frac{1}{0.05}\left(2\cdot 0.01 + 3\cdot 0.02 + 4\cdot 0.02\right) = \frac{0.16}{0.05} \\ &= 3.2 \end{aligned}$$

Es gibt noch eine dritte Möglichkeit TVaR berechnen. Es ist

$$\gamma = \frac{\mathbb{P}(L > F^{\leftarrow}(\alpha))}{1-\alpha} = \frac{0.04}{0.05} = \frac{4}{5}$$

und weiterhin

$$
\begin{aligned}
&(1-\gamma)\cdot F^{\leftarrow}(\alpha)+\gamma\cdot\mathbb{E}(L|L>F^{\leftarrow}(\alpha))\\
&=\frac{1}{5}\cdot 2+\frac{4}{5}\cdot\left(3\cdot\frac{\mathbb{P}(L=3)}{\mathbb{P}(L>2)}+4\cdot\frac{\mathbb{P}(L=4)}{\mathbb{P}(L>2)}\right)\\
&=\frac{1}{5}\cdot 2+\frac{4}{5}\cdot\left(3\cdot\frac{0.02}{0.04}+4\cdot\frac{0.02}{0.04}\right)\\
&=\frac{2}{5}+\frac{4}{5}\cdot\left(3\cdot\frac{2}{4}+4\cdot\frac{2}{4}\right)\\
&=\frac{2}{5}+\frac{4}{5}\cdot\left(\frac{6}{4}+\frac{8}{4}\right)=\frac{2}{5}+\frac{4}{5}\cdot\left(\frac{14}{4}\right)=\frac{16}{5}\\
&=3.2
\end{aligned}
$$

5 Lebesgue-Stieltjes Integral - Erwartungswert bezüglich $\mathbb{P}$ bzw. F

5.1 Grundlagen und Maßtheoretische Induktion

5.1.1 Definition und Satz: i.) Es seien $(\Omega_1, \mathcal{F}_1)$ und $(\Omega_2, \mathcal{F}_2)$ messbare Räume. Eine Abbildung $X : \Omega_1 \to \Omega_2$ heißt **messbar**, falls $X^{-1}(B) \in \mathcal{F}_1$ für alle $B \in \mathcal{F}_2$. Das Urbild jeder $\mathcal{F}_2$-Menge ist eine $\mathcal{F}_1$-Menge.

ii.) Es sei $(\Omega_1, \mathcal{F}_1, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $(\Omega_2, \mathcal{F}_2)$ ein messbarer Raum und $X : \Omega_1 \to \Omega_2$ messbar. Dann heißt

X eine **Zufallsvariable**.

iii.) Es sei $(\Omega_1, \mathcal{F}_1, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $(\Omega_2, \mathcal{F}_2)$ ein messbarer Raum und $X : \Omega_1 \to \Omega_2$ eine Zufallsvariable. Dann heißt die Abbildung

$$\mathbb{P}^X = \begin{cases} \mathcal{F}_2 \to [0,1] \\ B \mapsto \mathbb{P}(X^{-1}(B)) =: \mathbb{P}(X \in B) \end{cases}$$

die **Bildmaß von** X. $\mathbb{P}^X$ ist ein Wahrscheinlichkeitsmaß auf $(\Omega_2, \mathcal{F}_2)$.

Insbesondere falls $(\Omega_2, \mathcal{F}_2) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ist, nennen wir $\mathbb{P}^X$ auch die **Verteilung** von X.

iv.) Es sei $(\Omega_1, \mathcal{F}_1, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega_1 \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ eine reellwertige Zufallsvariable. Dann heißt $F^X : \mathbb{R} \to [0,1]$

$$F^X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leq x)$$

die **Verteilungsfunktion von** X. $F^X(x)$ entspricht also der Wahrscheinlichkeit, dass X die Grenze x nicht überschreitet.

▶ **Beispiel:** Es sei $\Omega_1 = \{1,2,3,4,5,6\}, \mathcal{F}_1 = \{\emptyset, \Omega, \{1,3,5\}, \{2,4,6\}\}$ und $\Omega_2 = \{0,1\}, \mathcal{F}_2 = \{\emptyset, \Omega_2, \{0\}, \{1\}\}$. Es sei $X = \mathbb{1}_{\{1,2,3\}}$. X ist nicht messbar. Für die Auswertung von X benötigt man Informationen, die in $\mathcal{F}_1$ nicht verfügbar sind.

5.1.2 Definition: i.) Es sei $\Omega \neq \emptyset$ eine Menge und $A \subset \Omega$. Die Abbildung $\mathbb{1}_A : \Omega \to \mathbb{R}$ mit

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{für } \omega \in A \\ 0 & \text{für } \omega \in \Omega \setminus A \end{cases}$$

heißt **Indikatorfunktion** der Menge A.

ii.) Es sei $\Omega \neq \emptyset$ eine Menge. Eine Familie $(A_i)_{i \in I}$ von Teilmengen von Ω heißt **endliche Zerlegung** von Ω, falls:

- Für ein $n \in \mathbb{N}$ ist $I = \{1, ..., n\}$ ist eine endliche Menge mit n Elementen.
- Es gilt $\Omega = \biguplus_{i=1,...,n} A_i$.

iii.) Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ heißt **elementar**, wenn es eine Zerlegung $(A_i)_{i \in I}$ von Ω mit $A_i \in \mathcal{F}, i \in I$ sowie reelle Zahlen $\alpha_i \in \mathbb{R}, i \in I$ gibt, so dass

$$X(\omega) = \sum_{i=1,...,n} \alpha_i \mathbb{1}_{A_i}(\omega).$$

5.1.3 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und X eine elementare Zufallsvariable mit

$$X(\omega) = \sum_{i=1,...,n} \alpha_i \mathbb{1}_{A_i}(\omega).$$

Dann heißt

$$\mathbb{E}(X) = \sum_{i=1,\ldots,n} \alpha_i \mathbb{P}(A_i)$$

der **Erwartungswert** von X (oder das **Lebesgue-Integral** von X bezüglich $\mathbb{P}$).

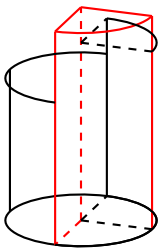


Abbildung 5.1: Die Abbildung zeigt eine elementare Zufallsvariable die drei Werte annehmen kann (die drei Höhen 2, 2.5, 3). Der Zufall wählt ein ω im Einheitskreis; mit $\mathbb{P}(A_1)$ im linken Halbkreis, mit $\mathbb{P}(A_2)$ im hinteren Viertel und mit $\mathbb{P}(A_3)$ im vorderen Viertel. Dann $\mathbb{E}(A) = \mathbb{P}(A_1) \cdot 2 + \mathbb{P}(A_2) \cdot 2.5 + \mathbb{P}(A_2) \cdot 3$

5.1.4 Definition: Es sei $X : \Omega \to \bar{\mathbb{R}}_{\geq 0}$ eine **numerische** nicht-negative Zufallsvariable. Dann heißt

$$\mathbb{E}(X) = \sup\{\mathbb{E}(Y) | X \geq Y, Y \text{ elementare ZV } \}$$

der **Erwartungswert** von X.

Zu beachten ist, dass $X(\omega) = \infty$ bzw. $\mathbb{E}(X) = \infty$ zugelassen sind. Die folgenden Regeln werden dabei unterstellt: $\infty \cdot 0 =$

$0 \cdot \infty = 0, \infty \cdot \infty = \infty$ und für alle $x > 0$ gilt $\infty \cdot x = x \cdot \infty = \infty$.

Wir verzichten hier auf die Erläuterung *technischer* Aspekte, die sich einerseits dadurch ergeben, dass wir $\bar{\mathbb{R}}$ anstatt $\mathbb{R}$ verwenden. Wir können andererseits andere technische Aspekte vermeiden, die sich ergeben würde, wenn wir den Erwartungswert nur für reellwertige Zufallsvariablen definieren würden. In $\bar{\mathbb{R}}$ beispielsweise hat jede nicht-leere Teilmenge ein Supremum. Für Details vergleiche Elstrodt [8] oder Henze [19, Seite 320].

5.1.5 Satz: Es sei $X : \Omega \to \bar{\mathbb{R}}_{\geq 0}$ eine numerische nicht-negative Zufallsvariable. Dann gilt

i.) Es gibt eine Folge $(X_i)_{i \in \mathbb{N}}$ von elementaren nicht-negativen Zufallsvariablen mit $X_i \nearrow X$.
ii.) Für jede Folge $(X_i)_{i \in \mathbb{N}}$ von elementaren nicht-negativen Zufallsvariablen mit $X_i \nearrow X$ gilt

$$\mathbb{E}(X_i) \nearrow \mathbb{E}(X).$$

Beweishinweis: Für $n \in \mathbb{N}$ betrachten wir die Zerlegung

$$A_{j,n} = \begin{cases} \{\frac{j}{2^n} \leq X < \frac{j+1}{2^n}\} & \text{für } j = 0, ..., n2^n - 1 \\ \{X \geq n\} & \text{für } j = n2^n \end{cases}$$

sowie die Funktion

$$X_n := \sum_{j=0}^{n2^n} \frac{j}{2^n} \mathbb{I}_{A_{j,n}}.$$

Beachte, dass die Ordinate in Intervalle zerlegt wird und von diesen Intervalle die Urbilder gebildet werden.

Vgl. Henze [19, Seite 328] für Details.

5.1.6 Bemerkung: Das bemerkenswerte an ii.) des vorhergehenden Satzes ist, dass sich für jede Folge der gleiche Limes ergibt. Das Integral ist also unabhängig von der Wahl der (monotonen) Folge.

5.1.7 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

i.) [**Linearität** von $\mathbb{E}$] Für zwei nicht-negative numerische Zufallsvariablen X, Y und $\alpha, \beta \in \mathbb{R}$ gilt

$$\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y).$$

ii.) [**Monotonie** von $\mathbb{E}$] Für zwei nicht-negative numerische Zufallsvariablen X, Y mit $X \leq Y$ gilt

$$\mathbb{E}(X) \leq \mathbb{E}(Y).$$

5.1.8 Defintion: Es sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Es sei $X^+ = \max\{X, 0\}, X^- = \max\{0, -X\}$; also $X = X^+ -$

X^-. X heißt **integrierbar** bezüglich $\mathbb{P}$, falls $\mathbb{E}(X^+) < \infty$ und $\mathbb{E}(X^-) < \infty$ gilt. In diesem Fall schreiben wir $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ (oder kurz $X \in \mathcal{L}^1$, wenn Missverständnisse ausgeschlossen sind). Wenn X integrierbar ist, dann definieren wir

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

Für $X \in \mathcal{L}^1$ sind auch die folgenden Schreibweisen üblich

$$\mathbb{E}(X) = \mathbb{E}^{\mathbb{P}}(X) = \int_\Omega X d\mathbb{P} = \int_\Omega X(\omega) d\mathbb{P}(\omega) = \int_\Omega X(\omega) \mathbb{P}(d\omega)$$

5.1.9 Satz: Es sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Dann ist äquivalent:

i.) $X \in \mathcal{L}^1$.

ii.) $|X| \in \mathcal{L}^1$.

iii.) $X^+, X^- \in \mathcal{L}^1$.

iv.) Es gibt $Y \in \mathcal{L}^1, Y \geq 0$ mit $|X| \leq Y$. Y heißt integrierbare **Majorante**.

5.1.10 Bemerkung: Wenn **X** beschränkt ist, dann ist **X** integrierbar. Insbesondere ist die Dirichlet Funktion χ auf dem Intervall $(0, 1]$ integrierbar und es gilt

$$\mathbb{E}(\chi) = 0$$

5.2 Konvergenzsätze

5.2.1 Satz von der majorisierten Konvergenz: Es sei $X_n : \Omega \to \bar{\mathbb{R}}$ eine Folge von Zufallsvariablen, so dass $X_n(\omega)$ mit Wahrscheinlichkeit 1 konvergiert. Ferner sei $g \in \mathcal{L}^1$ eine nicht-negative Zufallsvariable mit $|X_n| \leq g$ mit Wahrscheinlichkeit 1. Dann gilt:

i.) Für alle $n \in \mathbb{N}$ ist $X_n \in \mathcal{L}^1$ und **es gibt** $X \in \mathcal{L}^1$ mit $X_n \to X$ mit Wahrscheinlichkeit 1.
ii.) **Für jedes** $X \in \mathcal{L}^1$ mit $X_n \to X$ mit Wahrscheinlichkeit 1 gilt

$$\lim_{n\to\infty} \int X_n \, d\mathbb{P} = \int X \, d\mathbb{P} = \mathbb{E}(X) = \int (\lim_{n\to\infty} X_n) \, d\mathbb{P}$$

und

$$\lim_{n\to\infty} \int |X_n - X| \, d\mathbb{P} = 0.$$

5.3 Linearität und Monotonie

5.3.1 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

i.) (**Linearität** von $\mathbb{E}$) Für zwei Zufallsvariablen $X, Y \in \mathcal{L}^1$ und $\alpha, \beta \in \mathbb{R}$ gilt

$$\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y).$$

ii.) (**Monotonie** von $\mathbb{E}$) Für zwei Zufallsvariablen $X, Y \in \mathcal{L}^1$ mit $X \leq Y$ gilt

$$\mathbb{E}(X) \leq \mathbb{E}(Y).$$

5.4 Dreiecksungleichung und Cauchy-Bunjakowski-Schwarz-Ungleichung

5.4.1 Satz (Dreiecksungleichung): Für eine Zufallsvariable $X \in \mathcal{L}^1$ gilt

$$|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$$

▶ Blitzstein und Hwang [4, Theorem 4.4.2, Seite 164] nenn das folgende Resultat Fundamentalbrücke zwischen Wahrscheinlichkeit und Erwartung.

5.4.2 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A \in \mathcal{F}$. Dann gilt $\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A)$.

5.4.3 Definition: Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ heißt **quadrat-integrierbar**, falls $\mathbb{E}(X^2) < \infty$ ist. In diesem Fall schreiben wir $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ [oder kurz $X \in \mathcal{L}^2$, falls Missverständnisse ausgeschlossen werden können].

5.4.4 Satz (Cauchy-Bunjakowski-Schwarz-Ungleichung): Es sei $X, Y \in \mathcal{L}^2$. Dann $XY \in \mathcal{L}^1$ und

$$(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$
$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}$$

Beweis: Vgl. Tappe [36, S. 122]

5.4.5 Satz: $\mathcal{L}^2(\mathbb{P}) \subset \mathcal{L}^1(\mathbb{P})$. Also, wenn eine Zufallsvariable quadrat-integrierbar ist, dann ist sie auch integrierbar (bezüglich eines Wahrscheinlichkeitsmaßes).

5.4.6 Bemerkung: Der Satz gilt für Wahrscheinlichkeitsmaße, jedoch i.A. nicht für beliebige Maße.

5.5 Varianz

5.5.1 Definition: Es sei $X \in \mathcal{L}^2$. Dann heißt

$$\mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

die **Varianz** von X.

5.5.2 Satz: Es sei $X \in \mathcal{L}^2$. Dann gilt:

i.) $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

ii.) $\mathbb{V}(\alpha X + \beta) = \alpha^2 \mathbb{V}(X)$.

5.5.3 Satz: Es sei $X : \Omega \to [0, \infty]$ eine nicht-negative numerische Zufallsvariable. Dann gilt

$$\mathbb{E}(X) = 0 \quad \Leftrightarrow \quad X \equiv 0 \quad \mathbb{P}\text{-f.ü..}$$

$X \equiv 0 \quad \mathbb{P}$-f.ü. bedeutet $\mathbb{P}(X \equiv 0) = 1$.

Beweis: Vgl. Tappe [36, S. 111]. Wir betrachten für die Hinrichtung

$$\frac{1}{n}\mathbb{P}\left(X > \frac{1}{n}\right) = \frac{1}{n}\mathbb{E}\left[\mathbb{1}_{X>\frac{1}{n}}\right] = \mathbb{E}\left[\frac{1}{n}\mathbb{1}_{X>\frac{1}{n}}\right] \leq \mathbb{E}(X) = 0$$

Wegen der Stetigkeit von $\mathbb{P}$ gilt

$$\mathbb{P}\left(X > 0\right) = \lim \mathbb{P}\left(X > \frac{1}{n}\right) = 0.$$

Für die Rückrichtung: Es sei $\mathbb{P}(X > 0) = 0$. Dann

$$0 \leq \mathbb{E}\left(X \mathbb{1}_{X \leq n}\right) = \mathbb{E}\left(X \mathbb{1}_{0 < X \leq n}\right) \leq \mathbb{E}\left(n \mathbb{1}_{0 < X \leq n}\right)$$
$$n\mathbb{P}(0 < X \leq n) = 0$$

Also $\mathbb{E}\left(X \mathbb{1}_{X \leq n}\right) = 0$. Es gilt $\lim X \mathbb{1}_{X \leq n} = X$. Dann aus dem Satz über die monotone Konvergenz

$$\mathbb{E}(X) = \mathbb{E}(\lim_{n \to \infty} X \mathbb{1}_{X \leq n}) = \lim_{n \to \infty} \mathbb{E}(\mathbb{1}_{X \leq n}) = 0$$

5.5.4 Bemerkung: $X \equiv 0$ **fast sicher** bedeutet $\mathbb{P}(\{\omega \in \Omega | X(\omega) = 0\}) = \mathbb{P}(X = 0) = 1$.

5.5.5 Beispiel: Für die Dirichlet Funktion χ gilt $\mathbb{E}(\chi) = 0$, obwohl χ unendlich oft den Wert 1 annimmt.

5.5.6 Satz: Es sei $X \in \mathcal{L}^2$. Dann gilt

$$\mathbb{V}(X) = 0 \quad \Leftrightarrow \quad X \equiv \mathbb{E}(X) \quad \mathbb{P}\text{-f.ü..}$$

5.6 Ungleichungen: Markow, Chebycehv, Jensen

5.6.1 Satz (Allgemeine Markow-Ungleichung): Es sei $a > 0$ und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable sowie $h : \mathbb{R} \to \mathbb{R}_{\geq 0}$, sodass $h \circ X \in \mathcal{L}^1$. Dann gilt

$$\mathbb{P}(h(X) \geq a) \leq \frac{\mathbb{E}(h(X))}{a}.$$

Beweis: Es gilt

$$\begin{aligned} \mathbb{E}(h(X)) &= \mathbb{E}(\mathbb{1}_{h(X)\geq a} h(X) + \mathbb{1}_{h(X)<a} h(X)) \\ &\geq a\mathbb{P}(h(X) \geq a) \end{aligned}$$

5.6.2 Satz (Ungleichung von Markow): Es sei $X \in \mathcal{L}^1$.

Dann gilt

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}.$$

5.6.3 Satz (Ungleichung von Chebycehv): Es sei $X \in \mathcal{L}^2$. Dann gilt

i.) Für alle $a > 0$

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(X^2)}{a^2}.$$

ii.) Für alle $a > 0$

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2},$$

mit $\mu = \mathbb{E}(X)$ und $\sigma^2 = \mathbb{V}(X)$.

5.6.4 Satz (Ungleichung von Jensen): Es sei $I \subset \mathbb{R}$ ein Intervall, $g : I \to \mathbb{R}$ eine konvexe Funktion, $X \in \mathcal{L}^1$ eine integrierbare Zufallsvariable mit Bild$(X) \subset I$ und $g(X) \in \mathcal{L}^1$. Dann gilt

$$g(\mathbb{E}(X)) \leq \mathbb{E}(g(X)).$$

6 Zufallsvektoren

6.1 Unabhängigkeit

6.1.1 Definition: i.) Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine $\sigma-$Algebra $\mathcal{G} \subset \mathcal{F}$ heißt **Sub-σ-Algebra**.

ii.) Es sei I eine Indexmenge und $\mathcal{G}_i, i \in I$ eine Familie von Sub-σ-Algebren. Die Familie $\mathcal{G}_i, i \in I$ heißt **unabhängig**, falls für alle endlichen Teilmengen $J \subset I$ und $A_j \in \mathcal{G}_j, j \in J$

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

gilt.

iii.) Es sei I eine Indexmenge und $(\Omega_i, \mathcal{F}_i), i \in I$ messbare Räume und $X_i : \Omega \to \Omega_i$ Zufallsvariablen. Die Zufallsvariablen $(X_i), i \in I$ heißen **unabhängig**, falls die $\sigma(X_i) = X_i^{-1}(\mathcal{F}_i)$ unabhängig sind.

6.1.2 Bemerkung: Beachte, dass alle Zufallsvariablen im vorhergehenden Satz den gleichen Definitionsbereich jedoch möglicherweise unterschiedliche Zielmengen haben.

▶ Die Definition für unabhängige Zufallsvariablen ist *arg abstrakt*. Der folgende Satz enthält eine anschauliche äquivalente Charakterisierung.

6.1.3 Satz: Es seien $X : \Omega \to \Omega_1$ und $Y : \Omega \to \Omega_2$ Zufallsvariablen. Dann ist äquivalent:

i.) X und Y sind unabhängig.

ii.) Für alle $A \in \mathcal{F}_1$ und $B \in \mathcal{F}_2$ gilt

$$\mathbb{P}(X \in A \text{ und } Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

iii.) Für beliebige messbare Abbildungen $f : \Omega_1 \to \Omega_1'$ und $g : \Omega_2 \to \Omega_2'$ mit Werten in messbaren Räumen $(\Omega_1', \mathcal{F}_1')$ bzw. $(\Omega_2', \mathcal{F}_2')$ sind die Zufallsvariablen $f \circ X$ und $g \circ Y$ unabhängig.
Also: Die Unabhängigkeit bleibt bei jeder messbaren Transformation erhalten!

Beweis: Siehe Tappe [36, Seite 152].

6.1.4 Satz: Es seien $X : \Omega \to \Omega_1$ und $Y : \Omega \to \Omega_2$ Zufallsvariablen. Dann ist äquivalent:

i.) X und Y sind unabhängig.

ii.) Für zwei beliebige messbare Funktionen $f : \Omega_1 \to \mathbb{R}$ und $g : \Omega_2 \to \mathbb{R}$ mit $f \circ X \in \mathcal{L}^1$ und $g \circ X \in \mathcal{L}^1$ gilt

ii.) ist in Tappe etwas anders. Beweis angeben!

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y)).$$

[dann ist insbesondere $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$]

Beweis: Vgl. Schilling [31, Seite 152] für den Beweis der schwierigen Richtung i.) $\Rightarrow$. ii.)

ii.) $\Rightarrow$. i.). Angenommen $A \in \mathcal{F}_1, B \in \mathcal{F}_2$. Dann gilt

$$\begin{aligned}\mathbb{P}(X \in A, Y \in B) &= \mathbb{E}(\mathbb{1}_A(X)\mathbb{1}_B(Y)) = \mathbb{E}(\mathbb{1}_A(X))\mathbb{E}(\mathbb{1}_B(Y)) \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B).\end{aligned}$$

6.1.5 Bemerkung: Es seien $X : \Omega \to \mathbb{R}$ und $Y : \Omega \to \mathbb{R}$ **diskrete** Zufallsvariablen mit Zähldichten $p(x) = \mathbb{P}^X(\{x\}), x \in X(\Omega)$ bzw. $q(y) = \mathbb{P}^Y(\{y\}), y \in Y(\Omega)$. Dann sind X und Y genau dann unabhängig, wenn für alle $x \in X(\Omega)$ und $y \in Y(\Omega)$

$$\mathbb{P}(X = x \text{ und } Y = y) = p(x)q(y)$$

gilt.

6.2 Gemeinsame Verteilung

6.2.1 Definiton: i.) Es sei $\mathcal{B}(\mathbb{R}^2) = \sigma(\{(a_1, b_1] \times (a_2, b_2] : a_1 \leq b_1, a_2 \leq b_2, a_1, a_2, b_1, b_2 \in \bar{\mathbb{R}}\})$. Die Mengen $B \in \mathcal{B}(\mathbb{R}^2)$ heißen **Borelmengen** (in $\mathbb{R}^2$).

ii.) Es seien $X : \Omega \to \mathbb{R}$ und $Y : \Omega \to \mathbb{R}$ reellwertige Zufallsvariablen. Die Abbildung

$$\mathbb{P}^{(X,Y)}(B) = \begin{cases} \mathcal{B}(\mathbb{R}^2) \to [0,1] \\ B \mapsto \mathbb{P}((X,Y) \in B) \end{cases}$$

heißt die **gemeinsame Verteilung** von X und Y. $\mathbb{P}^{(X,Y)}(B)$ ist also die Wahrscheinlichkeit ein Ergebnis im 2-dim Bereich $B \subset \mathbb{R}^2$ zu beobachten, wobei $B \in \mathcal{B}(\mathbb{R}^2)$ eine Borelmenge ist.

ii.) Die Abbildung

$$F^{(X,Y)} = \begin{cases} \mathbb{R}^2 \to [0,1] \\ (x,y) \mapsto \mathbb{P}(X \leq x \wedge Y \leq y) = \mathbb{P}^{(X,Y)}((-\infty, x] \times (-\infty, y]) \end{cases}$$

heißt **gemeinsame Verteilungsfunktion** der Zufallsvariablen X, Y. $F(x,y)$ ist also die Wahrscheinlichkeit, dass X einen Wert kleiner gleich x **und** Y einen Wert kleiner gleich y annimmt.

iii.) Die Verteilungen $\mathbb{P}^X$ und $\mathbb{P}^Y$ von X bzw. Y nennt man in diesem Zusammenhang auch **Randverteilungen**.

6.2.2 Satz: Es seien $X : \Omega \to \mathbb{R}$ und $Y : \Omega \to \mathbb{R}$ reellwertige Zufallsvariablen. Dann ist äquivalent:

i.) X und Y sind unabhängig.

ii.) Für alle $x, y \in \mathbb{R}$ gilt $F^{(X,Y)}(x, y) = F^X(x)F^Y(y)$.

6.2.3 Definition: Es sei F die gemeinsame Verteilungsfunktion der Zufallsvariablen $X, Y : \Omega \to \mathbb{R}$. Wenn es eine Funktion $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ mit

$$F(a, b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y)\, dx\, dy \quad \text{für } a, b \in \mathbb{R}$$
$$F(\infty, \infty) := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

gibt, dann heißt f die **gemeinsame Dichte** der Zufallsvariablen X und Y.

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

heißt die **Randdichte** von X und

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

heißt die **Randdichte** von Y.

6.2.4 Satz: Es seien $X, Y : \Omega \to \mathbb{R}$ Zufallsvariablen mit gemeinsamer Dichte f. X und Y sind genau dann unabhängig,

wenn für fast alle x, y gilt:

$$f(x,y) = f_X(x) f_Y(y).$$

6.3 Abhängigkeit (linear)

$\triangleright$ Bisher wurde Unabhängigkeit untersucht. Wie ist es mit der **Nicht-Unabhängigkeit (Abhängigkeit)**? Da fängt man am besten mit der **linearen** Nicht-Unabhängigkeit (Abhängigkeit) an. ... Kovarianz! $\triangleleft$

6.3.1 Definition: Es seien $X, Y \in \mathcal{L}^2$. Dann heißt

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}((X - \mu_X)(Y - \mu_Y))\end{aligned}$$

die **Kovarianz** von X, Y. Wenn $\operatorname{cov}(X,Y) = 0$ gilt, dann heißen X und Y **unkorreliert**.

Natürlich $\operatorname{cov}(X,X) = \mathbb{V}(X)$.

6.3.2 Satz: Für $X, Y \in \mathcal{L}^2$ gilt

$$\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(XY) - \mu_X \mu_Y.$$

6.3.3 Satz: Es seien $X, Y \in \mathcal{L}^2$. Wenn X und Y unabhängig sind, dann sind X und Y unkorreliert.

6.3.4 Bemerkung: Während die Unabhängigkeit bei **jeder** messbaren **Transformation** erhalten bleibt (siehe vorne), bleibt – wie der vorhergehende Satz zeiht – die Unkorreliertheit i.A. **nur bei linearen/affinen Transformationen** erhalten.

6.3.5 Satz: Es seien $X, Y \in \mathcal{L}^2$. Wenn X und Y unkorreliert sind, dann sind auch $aX + b$ und $cY + d$ unkorreliert.

6.3.6 Satz: Für $X, Y \in \mathcal{L}^2$ und $a, b, c, d \in \mathbb{R}$ gilt

$$\mathrm{cov}(aX + b, cY + d) = ac \cdot \mathrm{cov}(X, Y)$$

6.3.7 Bemerkung: Es sei $X \sim N(0, 1)$ und $Y = X^2$. Natürlich sind X und Y nicht unabhängig. Aber:

$$\begin{aligned} \mathrm{cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(XX^2) \\ &= \mathbb{E}(X^3) \\ &= 0 \end{aligned}$$

Dieses Beispiel funktioniert natürlich auch für andere Verteilungen als die Normalverteilung.

6.3.8 Satz: Die Abbildung

$$\mathrm{cov} : \mathcal{L}^2 \times \mathcal{L}^2 \to \mathbb{R}$$

ist eine symmetrische, bilineare und positiv semi-definite Abbildung/Operator, d.h.

i.) Für alle $Y \in \mathcal{L}^2$ ist

$$\text{cov}(\cdot, Y) : \mathcal{L}^2 \to \mathbb{R}$$

linear. Also gilt für alle $a_1, a_2 \in \mathbb{R}$ und $X_1, X_2 \in \mathcal{L}^2$

$$\text{cov}(a_1 X_1 + a_2 X_2, Y) = a_1 \text{cov}(X_1, Y) + a_2 \text{cov}(X_2, Y).$$

Analog ist für alle $X \in \mathcal{L}^2$

$$\text{cov}(X, \cdot) : \mathcal{L}^2 \to \mathbb{R}$$

linear.

ii.) Für alle $X, Y \in \mathcal{L}^2$ gilt $\text{cov}(X, Y) = \text{cov}(Y, X)$.

iii.) Für alle $X \in \mathcal{L}^2$ gilt $\text{cov}(X, X) \geq 0$.

6.3.9 Satz: Wenn die Zufallsvariablen $X_1, X_2, ..., X_n \in \mathcal{L}^2$ **paarweise unkorreliert** sind, dann gilt:

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i).$$

6.3.10 Satz: Es seien $X_1, X_2 \in \mathcal{L}^2$. Dann gilt:

$$\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + 2\,\mathrm{cov}(X_1, X_2)$$
$$\mathbb{V}(X_1 - X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2) - 2\,\mathrm{cov}(X_1, X_2)$$

6.3.11 Definition: Es seien $X_1, X_2, ..., X_n \in \mathcal{L}^2$. Dann heißt die $n \times n$ Matrix

$$\mathbf{\Sigma}_X = (\mathrm{cov}(X_i, X_j))_{i=1,...,n, j=1,...,n}$$

die **Kovarianzmatrix** von des Zufallsvektor X.

$\triangleright$ Vorsicht, obwohl auf der Hauptdiagonalen der Matrix $\mathbf{\Sigma}_X$ die Varianzen $\mathbb{V}(X_i) = \sigma_i^2$ von X_i stehen, verwenden wir für die Matrix der Ko-Varianzen die Notation $\mathbf{\Sigma}_X$ und nicht $\mathbf{\Sigma}_X^2$.

6.3.12 Satz: Die Kovarianzmatrix ist symmetrisch und positiv-semidefinit, d.h. $\mathbf{\Sigma}_X = (\mathbf{\Sigma}_X)^T$ und für alle $\mathbf{b} \in \mathbb{R}^n$ gilt

$$\mathbf{b}^T \mathbf{\Sigma}_X \mathbf{b} = \langle \mathbf{\Sigma}_X \mathbf{b}, \mathbf{b} \rangle \geq 0.$$

6.3.13 Satz (Sandwich-Formel): Es seien $X_1, X_2, ..., X_n \in \mathcal{L}^2, \mathbf{X}^T = (X_1, X_2, ..., X_n), \mathbf{A} \in \mathrm{M}(m, n, \mathbb{R})$ und $\mathbf{b} \in \mathbb{R}^m$. Dann gilt für die affine Transformation $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$

$$\mathbf{\Sigma_Y} = \mathbf{A}\mathbf{\Sigma_X}\mathbf{A}^T.$$

6.3.14 Satz (Cauchy-Bunjakowski-Schwarz-Ungleichung): Es seien $X, Y \in \mathcal{L}^2$.

i.) Dann $XY \in \mathcal{L}^1$ und

$$(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$
$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}$$

ii.) Ferner

$$|\mathrm{cov}(X,Y))| \leq \sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)} = \mathrm{std}(X)\mathrm{std}(Y)$$
$$(\mathrm{cov}(X,Y))^2 \leq \mathbb{V}(X)\mathbb{V}(Y)$$

6.3.15 Defintion: Es seien $X, Y \in \mathcal{L}^2$ mit $\mathbb{V}(X), \mathbb{V}(Y) > 0$. Dann heißt

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\mathrm{std}(X) \cdot \mathrm{std}(Y)}$$

der **Korrelationskoeffizient nach von Bravais und Pearson** von X und Y.

6.3.16 Satz: Es seien $X, Y \in \mathcal{L}^2$ mit $\mathbb{V}(X), \mathbb{V}(Y) > 0$. Dann gilt

i.) $-1 \leq \rho(X,Y) \leq 1$.

ii.) Es gilt $\rho(X,Y) = 1$ genau dann, wenn es $a, b \in \mathbb{R}, a > 0$ mit $\mathbb{P}(Y = aX + b) = 1$ gibt.

iii.) Es gilt $\rho(X,Y) = -1$ genau dann, wenn es $a, b \in \mathbb{R}, a <$

0 mit $\mathbb{P}(Y = aX + b) = 1$ gibt.

iv.) X, Y unabhängig impliziert $\rho(X, Y) = 0$.

v.) Für alle a, b, c, d gilt

$$\rho(aX + b, cY + d) = \frac{ac}{|ac|}\rho(X, Y).$$

6.3.17 Bemerkung: Wenn ρ nahe 1 ist, dann ist **der lineare Zusammenhang eng**. Die etwaigen Beobachtungen liegen nahe an einer Geraden. Man kann an ρ die **Steigung dieser Geraden nicht ablesen**.

6.4 Abhängigkeit: Copulas

6.4.1 Definition: Eine ***d*-dimensionale Copula** ist die gemeinsame Verteilungsfunktion C eines d-dimensionalen Zufallsvektors $\mathbf{U} = (U_1, \ldots, U_d)^\top$ mit $U_i \sim \text{Unif}([0, 1])$ für $i = 1, ..., d$.

6.4.2 Satz: Eine Abbildung $C : [0, 1]^d \to [0, 1]$ ist genau dann eine Copula, falls

- $C(u_1, ..., u_d) = 0$ falls $u_i = 0$ für (mindestens) ein i.
- $C(1, ..., 1, u_i, 1, ..., 1) = u_i$ für $i = 1, ..., d$ und $u_i \in [0, 1]$.

- Für alle $(a_1, ..., a_d), (b_1, ..., b_d) \in [0,1]^d$ mit $a_i \leq b_i$

$$0 \leq \sum_{i_1=1}^{2} ... \sum_{i_d=1}^{2} (-1)^{i_1+...+i_d} C(u_{1,i_1}, ..., u_{d,i_d}) \leq 1,$$

wobei $u_{j,1} = a_j$ und $u_{j,2} = b_j$ für $j = 1, ..., d$.

Wenn für einen Zufallsvektor $\mathbf{U} \sim C$ gilt, dann muss natürlich $\mathbb{P}(a_1 \leq U_1 \leq b_1, ..., \mathbb{P}(a_1 \leq U_1 \leq b_1) \geq 0$ gelten. Die Rechteckgleichung stellt sicher, dass das auch der Fall ist.

6.4.3 Bemerkung:

- Man nennt: Für alle $(a_1, ..., a_d), (b_1, ..., b_d) \in [0,1]^d$ mit $a_i \leq b_i$

$$0 \leq \sum_{i_1=1}^{2} ... \sum_{i_d=1}^{2} (-1)^{i_1+...+i_d} C(u_{1,i_1}, ..., u_{d,i_d}) \leq 0$$

wobei $u_{j,1} = a_j$ und $u_{j,2} = b_j$ für $j = 1, ..., d$
die d-dimensionale **Rechteckgleichung**.

- Die Rechteckgleichung für $d = 2$: Für $0 \leq a_k \leq b_k$, $k = 1, 2$ gilt

$$0 \leq C(b_1, b_2) + C(a_1, a_2) - C(a_1, b_2) - C(b_1, a_2) \leq 1$$

Vgl. Henze Seite 133, 317

6.4.4 Satz: Es sei $\mathbf{X} = (X_1, \ldots, X_d)^\top \in \mathcal{F}(F_1, \ldots, F_d)$ ein Zufallsvektor mit gemeinsamer Verteilungsfunktion $F^{\mathbf{X}}$. Dann **existiert** ein Zufallsvektor $\mathbf{U} = (U_1, \ldots, U_d)^\top$ mit $U_i \sim \text{Unif}([0, 1])$ mit gemeinsamer Verteilungsfunktion $C = C_{\mathbf{X}}$, so dass

$$F^{\mathbf{X}}(t_1, \ldots, t_d) = C(F_1(t_1), \ldots, F_d(t_d)) \qquad \forall\, t_i \in \mathbb{R}.$$

Sind die Funktionen F_i stetig, dann ist C **eindeutig**. In der Tat

$$C(u_1, ..., u_d) = F(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d)).$$

Die rechte Seite $C(F_1(t_1), \ldots, F_d(t_d))$ beschreibt die Zerlegung der gemeinsamen Verteilung $F^{\mathbf{X}}$ in Ränder $F_1, ..., F_d$ und Copula C.

6.4.5 Satz: Sind Verteilungsfunktionen $F_1, \ldots, F_d$ und eine Copula C gegeben, so existiert ein Zufallsvektor $\mathbf{X} = (X_1, \ldots, X_d)^\top \in \mathcal{F}(F_1, \ldots, F_d)$ mit

$$F^{\mathbf{X}}(t_1, \ldots, t_d) = C(F_1(t_1), \ldots, F_d(t_d)).$$

6.4.6 Bemerkung: Was bedeutet die Gleichung

$$F^{\mathbf{X}}(t_1, t_2) = C(F_1(t_1), F_2(t_2))\,?$$

Wir definieren $u_i = F_i(t_i) = \mathbb{P}(X_i \leq t_i)$. u_i ist also die

Wahrscheinlichkeit, dass X_i nicht über der Schwelle t_i ist. Wir wollen

$$\begin{aligned} F^{\mathbf{X}}(x_1, x_2) &= \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) \\ &= C(F_1(x_1), F_2(x_2)) \\ &= C(\mathbb{P}(X_1 \leq x_1), \mathbb{P}(X_2 \leq x_2)) \\ &= C(u_1, u_2) \end{aligned}$$

interpretieren!

Nehmen wir zur Illustration an, dass $t_1 = 1.89\,\text{cm}, u_1 = 0.90$ und $t_2 = 1.64\,\text{cm}, u_2 = 0.95$ und $F^{\mathbf{X}}(t_1, t_2) = C(F_1(t_1), F_2(t_2)) = 0.04$. Auf Basis der Zahlen $t_1 = 1.89$ und $t_2 = 1.64$ können wir nicht erkennen, ob es sich um gemeinsam relativ große Werte handelt. Die gemeinsame Wahrscheinlichkeit, dass $X_1 \leq 1.89$ und $X_2 \leq 1.64$, ist (nur) für 0.04. Aber was bedeutet das? Ist $X_1 \leq 1.89$ und $X_2 \leq 1.64$ bemerkenswert? Schon, denn die Wahrscheinlichkeit größere Werte für X_1 bzw. X_2 zuhalten, sind klein: $\mathbb{P}(X_1 > t_1) = 1 - 0.9 = 0.1$ bzw. $\mathbb{P}(X_1 > t_1) = 1 - 0.9 = 0.05$. Der Wert $C(0.9, 0.95) = 0.045$ besagt, dass die Wahrscheinlichkeit für ein gemeinsames Übertreffen der 90%-Schwelle (für X_1) bzw. der 95%-Schwelle (für X_2), immerhin 0.045 ist. Anstatt die Schwellen in cm zu messen, werden die Schwelle in Wahrscheinlichkeiten gemessen. In diesem Sinn ermöglichen Copulas einen Wechsel der der Einheiten. Aus "größer als 1.64" wird "unwahrscheinlich" (nämlich ein Ereig-

nis mit einer Wahrscheinlichkeit 0.05).

Für den Fall stetiger F_i erhalten wir eine besonders einfache Interpretation, für diesen **Wechsel der Einheiten**: $C(u_1, u_2)$ entspricht der **Wahrscheinlichkeit**, dass X_1, X_2 unter den u_1- bzw. u_2-Quantilen bleibt.

Die Wahrscheinlich, dass X_1 ihr 0.95-Quantil und X_2 sein 0.9-Quantil übertrifft, ist immerhin 0.045.

6.4.7 Satz: Es sei $\mathbf{X} = (X_1, ..., X_d)^T$ ein Zufallsvektor mit stetigen Verteilungsfunktionen F_i und Copula C. Ferner seien $T_1, ..., T_d$ strikt monoton wachsende Transformationen. Dann ist C auch eine $(T_1(X_1), ..., T_d(X_d))^T$.

Beweis: Vgl. McNeil et al. [7, Seite 224].

6.4.8 Bemerkung: Copulas ermöglichen eine **Zerlegung** in die Randverteilungen und der Zusammenhangstruktur

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

Die Zerlegung erkennt man einerseits direkt an der Gleichung. Die Argumente von $C(\cdot, \ldots, \cdot)$ sind separat die univariaten Verteilungen F_i; sonst nichts.

Andererseits ist auch der letzte Satz erhellend. Die Copula bleibt erhalten, wenn wir die Zufallsvariable monoton trans-

formieren. ...

Das erleichtert die Arbeitsteilung. ...

6.4.9 Bemerkung: $\mathfrak{Z}$ F eine mVF. Im allgemeinen gibt es ∞-viele Copula C mit

$$F^{\mathbf{X}}(t_1, \ldots, t_d) = C(F_1(t_1), \ldots, F_d(t_d)) \qquad \forall\, t_i \in \mathbb{R}.$$

Wenn die Ränder steig sind, dann ist die Copula eindeutig und es gilt:

$$C(u_1, ..., u_d) = F(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d)).$$

Die Ursache für die Nicht-Eindeutigkeit kann man sich an einen einfachen bivariaten Beispiel klarmachen (vgl. McNeil [7, Seite 224])). Die gemeinsame Dichte für $\mathbf{X}$ sei wie in der folgenden Tabelle

Tab1	$X_1 = 2$	$X_2 = 4$	X_1
$X_1 = 2$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$
$X_2 = 4$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{5}{8}$
X_2	$\frac{3}{8}$	$\frac{5}{8}$	

Wir wissen aus dem Existenzsatz, dass es ein C mit

$$F^{\mathbf{X}}(t_1, t_2) = C(F_1(t_1), F_2(t_2)) \qquad \forall\, t_i \in \mathbb{R}.$$

gibt. Es gilt Bild$(F_i) = \{0, \frac{3}{8}, 1\}$. Die obige Funktionalgleichung für C liefert hier nur eine einzige Restriktion für C; nämlich $C(\frac{3}{8}, \frac{3}{8}) = \frac{1}{8}$. Diese Restriktion entsteht an der Stelle $t_1 = 2 = t_2$. Dort ist einerseits $F^{\mathbf{X}}(2,2) = \frac{1}{8}$ und $F_1(2) = \frac{3}{8} = F_2(2)$. Wenn wir jetzt z.B. $t_1 = 3 = t_2$ betrachten, dann erhalten wir wieder $F^{\mathbf{X}}(3,3) = \frac{1}{8}$ und $F_1(3) = \frac{3}{8} = F_2(3)$. Also wieder: $C(\frac{3}{8}, \frac{3}{8}) = \frac{1}{8}$. Wir können aus der Gleichung $F^{\mathbf{X}}(t_1, t_2) = C(F_1(t_1), F_2(t_2))$ nur an den Stellen Bild $F_1 \times$ Bild F_2 ermitteln. Den Wert der Copula an der Stelle $u_1 = \frac{1}{2} = u_2$ können (zwar nicht beliebig, wir müssen insb. die Rechteckungleichung beachten) *wählen*.

Angenommen wir wählen zwei Copula C^1 und C^2, die $F^{\mathbf{X}}(t_1, t_2) = C(F_1(t_1), F_2(t_2))$ erfüllen, und simulieren dann gemäß

- $(U_1^i, U_2^i) \sim C^i$
- $\mathbf{X}^i = (F^{\leftarrow}(U_1), F^{\leftarrow}(U_2))$

Gilt dann $F^{\mathbf{X}^1} = F^{\mathbf{X}^2}$.

6.4.10 Bemerkung:

- Betrachte eine Stichprobe $\mathbf{X}_i = (x_{i1}, ..., x_{id})^T, i = 1, ..., n$ des Zufallsvektors $\mathbf{X}$.
- Es sei $u_{ij} = \frac{r_{ij}}{n+1}$. Dabei ist r_{ij} der Rang von x_{ij} in $x_{kj}, k = 1, ..., n$.

- Die u_{ij} erfassen die Abhängigkeitsstruktur.
- Die Ränder werden auf auf die Gleichverteilung standardisiert.

6.5 Transformationsformel für multivariate Dichten

6.5.1 Satz: Es sei $S \subset \mathbb{R}^n$ eine offene Menge und $\mathbf{X}$ ein Zufallsvektor mit gemeinsamer Dichte $f_{\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}$. Ferner sei $g : S \to \mathbb{R}^n$ eine injektive Abbildung mit $g \in C^1$ und

$$\det J_g(\mathbf{x}) \neq 0$$

für alle $\mathbf{x} \in S$. Dabei bezeichnet

$$J_g(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}) & \ldots & \frac{\partial g_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial g_2}{\partial x_1}(\mathbf{x}) & \ldots & \frac{\partial g_2}{\partial x_n}(\mathbf{x}) \\ \ldots & \ldots & \ldots \\ \frac{\partial g_n}{\partial x_1}(\mathbf{x}) & \ldots & \frac{\partial g_n}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

die Jacobi-Matrix von g an der Stelle $\mathbf{x}$. Dann ist die Umkehrabbildung $h : g(S) \to \mathbb{R}^n$ stetig differenzierbar und $\mathbf{Y} = g(\mathbf{X})$ hat die Dichte

$$f_{\mathbf{Y}} = f_{\mathbf{X}}(h(\mathbf{y}))|\det J_g(\mathbf{x})|$$

Beweis: Tappe [36, Seite 188f]

7 Konvergenz von Folgen von Zufallsvariablen

7.1 Konvergenzformen

7.1.1 Bemerkung: Es sei (x_i) eine Folge reeller Zahlen und x eine reelle Zahl. Wir sagen: (x_i) konvergiert gegen x, falls es für alle $\varepsilon > 0$ ein $N_\varepsilon \in \mathbb{N}$ gibt, so dass $|x_i - x| < \varepsilon$ für alle $i \geq N$ gilt. Egal wie klein wir die ε-Umgebung wählen, wir können ein N finden, so dass die Folge **ab N_ε komplett in der ε-Umgebung von** x liegt.

Wir wollen nun das Konzept der *Konvergenz* auf Folgen von Zufallsvariablen (anstatt von reellen Zahlen) übertragen. Dafür gibt es mindestens vier wichtige Formen. Es liegt auf der Hand, dass der Konvergenzbegriff komplexer wird, denn Zufallsvariablen sind Abbildungen und nicht bloß Zahlen. Da

ist der Begriff der Umgebung/Nähe deutlich *schwieriger* zu erfassen.

7.1.2 Defintion: Es sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable und $X_i, i \in \mathbb{N}$ eine Folge von Zufallsvariablen $X_i : \Omega \to \mathbb{R}$.

i.) $(X_i) \to X$ **fast sicher** $:\Leftrightarrow \mathbb{P}(X_i \to X) = \mathbb{P}(\{\omega \in \Omega | X_i(\omega) \to X(\omega)\}) = 1$. In diesem Fall schreiben wir

$$\text{f.s.-}\lim X_i = X \text{ oder } X_i \xrightarrow{\text{f.s.}} X$$

ii.) Es sei $p = 1$ oder $p = 2$. $(X_i) \to X$ **im p-ten Mittel** $:\Leftrightarrow X_i \in \mathcal{L}^p, i \in \mathbb{N}$ und $\mathbb{E}(|X_i - X|^p) \to 0$ für $i \to \infty$. In diesem Fall schreibt man

$$\mathcal{L}^p\text{-}\lim X_i = X \text{ oder } X_i \xrightarrow{\mathcal{L}^p} X$$

iii.) $(X_i) \to X$ **nach Wahrscheinlichkeit** $:\Leftrightarrow$ Für alle $\varepsilon > 0$ gilt: für $i \to \infty$ gilt $\mathbb{P}(|X_i - X| > \varepsilon) \to 0$. In diesem Fall schreibt man

$$\text{n.W.-}\lim X_i = X \text{ oder } X_i \xrightarrow{\text{n.W.}} X$$

iv.) $(X_i) \to X$ **in Verteilung** $:\Leftrightarrow$ Für alle $x \in \mathbb{R}$, in denen F^X stetig ist, gilt $F^{X_i}(x) \to F^X(x)$. In diesem Fall

schreiben wir

$$\text{i.V.-}\lim X_i = X \text{ oder } X_i \xrightarrow{\text{i.V.}} X$$

7.1.3 Satz (Eindeutigkeit): Wenn n.W.-$\lim X_i = X$ und n.W.-$\lim X_i = Y$, dann $\mathbb{P}(X = Y) = 1$.

Beweis: Es sei $\varepsilon > 0$. Es gilt

$$\begin{aligned}
\mathbb{P}(|X - Y| > \varepsilon) &= \mathbb{P}(|X - X_n + X_n - Y| > \varepsilon) \\
&\leq \mathbb{P}(|X - X_n| + |X_n - Y| > \varepsilon) \\
&\leq \mathbb{P}\left(|X - X_n| > \frac{\varepsilon}{2} \text{ oder } |X_n - Y| > \frac{\varepsilon}{2}\right) \\
&\leq \mathbb{P}\left(|X - X_n| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|X_n - Y| > \frac{\varepsilon}{2}\right) \to 0
\end{aligned}$$

Die dritte Zeile ist erklärungsbedürftig. Wir müssen zeigen, dass für ω mit $|X - X_n| + |X_n - Y| > \varepsilon$ auch $|X - X_n| > \frac{\varepsilon}{2}$ oder $|X_n - Y| > \frac{\varepsilon}{2}$ gilt. Wenn $|X - X_n| > \frac{\varepsilon}{2}$ oder $|X_n - Y| > \frac{\varepsilon}{2}$ nicht gilt, dann gilt $|X - X_n| \leq \frac{\varepsilon}{2}$ und $|X_n - Y| \leq \frac{\varepsilon}{2}$. Dann gilt $|X - X_n| + |X_n - Y| \leq \varepsilon$. Also gilt ω mit $|X - X_n| + |X_n - Y| > \varepsilon$ nicht.

Es folgt $\mathbb{P}(|X - Y| > \varepsilon) = 0$. Also $\mathbb{P}(X = Y) = 1$.

7.1.4 Bemerkung: Wir werden gleich sehen, dass die f.s.-Konvergenz die n.W.-Konvergenz impliziert und ebenso die $\mathcal{L}^p$-Konvergenz die n.W.-Konvergenz impliziert. **Also gilt**

die obige Eindeutigkeit auch für die f.s.- und die $\mathcal{L}^p$ Grenzwerte.

7.1.5 Bemerkung: i.) Es gelte $X = \text{f.s.-}\lim X_i$. Wenn wir zufällig ein ω (einen *Pfad*) ziehen, dann ist das mit Wahrscheinlichkeit 1 ein Pfad mit Eigenschaft: Der gesamte Pfad $X_i(\omega)$ bleibt ab einer Stelle N in einer beliebig gewählten ε-Umgebung von $X(\omega)$ [also für alle $i \geq N$ gilt $|X_i(\omega) - X(\omega)| < \varepsilon$]. Mit Wahrscheinlichkeit 1 erhalten wir also die Konvergenz in $\mathbb{R}$.

Gilt *nur* $X = \text{n.W.-}\lim X_i$, dann kann es sogar sein, dass die pfadweise Konvergenz für **keinen** Pfad gilt. Es kann also sein, dass für alle Pfade gilt: Es gibt ein $\varepsilon > 0$, so dass es für alle $N \in \mathbb{N}$ noch Stellen $k \geq N$ gibt, so dass $|X_k(\omega) - X(\omega)| > \varepsilon$ gilt.

ii.) Bei der i.V.-Konvergenz werden nur x betrachtet, in denen F^X stetig ist. Warum das zweckmäßig ist, wird in einer Übungsaufgabe behandelt.

7.1.6 Satz: Für $\varepsilon > 0$, $n \in \mathbb{N}$ definieren wir

$$\begin{aligned} A_n(\varepsilon) &= \{|X_n - X| > \varepsilon\}, \\ B_n(\varepsilon) &= \bigcup_{m \geq n} A_m(\varepsilon). \end{aligned}$$

Dann ist äquivalent

i.) f.s.- $\lim X_i = X$.

ii.) Für alle $\varepsilon > 0$ gilt $\mathbb{P}(B_n(\varepsilon)) \to 0, n \to \infty$.

$A_n(\varepsilon)$ umfasst ε-Abweichungen in der Periode n
$B_n(\varepsilon)$ umfasst ε-Abweichungen in der Periode n oder danach.

Beweis: Der folgende Beweis ist aus Grimmett und Stirzaker [14, Seite 356f].[1].

Wir stellen zunächst ein paar Vorüberlegungen, die uns dann die eigentliche Argumentation erleichtern.

(1) Die Menge $B_n(\varepsilon)$ werden monoton nicht-größer: $B_{n_2}(\varepsilon) \subset B_{n_1}(\varepsilon)$ für $n_2 \geq n_1$. Also kann man den Grenzwert als Schnitt definieren:

$$A(\varepsilon) := \lim_{n\to\infty} B_n(\varepsilon) := \bigcap_{n\in\mathbb{N}} B_n(\varepsilon) \quad .$$

Es gilt: $\omega \in A(\varepsilon)$ gilt genau dann, wenn für alle n gibt es $m \geq n$ mit $|X_m(\omega) - X(\omega)| > \varepsilon$.

(2) Wir definieren die Ergebnisse in denen punktweise Konvergenz vorliegt:

$$C = \{\omega \in \Omega | X_n(\omega) \to X(\omega)\}.$$

Dann gilt: $\omega \in C$ gdw für alle $\varepsilon > 0$ gilt $\omega \notin A(\varepsilon)$. C^c ist die

[1] Vgl. auch Chung [5, S. 69]

Menge der ω bei denen keine Konvergenz vorliegt. Dann ist

$$C^c = \bigcup_{\varepsilon>0} A(\varepsilon).$$

Jetzt könne wir doe folgenden Äquivalenz zeigen:

$$\mathbb{P}(C^c) = 0 \Leftrightarrow \forall \varepsilon > 0 : \mathbb{P}(A(\varepsilon)) = 0.$$

i.) Dann gilt: Wenn $\mathbb{P}(C^c) = 0$ gilt, dann muss für alle $\varepsilon > 0$ auch $\mathbb{P}(A(\varepsilon)) = 0$ gelten.

ii.) Umgekehrt gilt auch: Gilt für alle $\varepsilon > 0$ auch $\mathbb{P}(A(\varepsilon)) = 0$, dann ist auch $\mathbb{P}(C^c) = 0$. In er Tat

$$\begin{aligned}
\mathbb{P}(C^c) &= \mathbb{P}\left(\bigcup_{\varepsilon>0} A(\varepsilon)\right) \\
&= \mathbb{P}\left(\bigcup_{m=1}^{\infty} A(1/m)\right) \qquad \text{denn } A(\varepsilon) \text{ ist monoton in } \varepsilon \\
&\leq \sum_{m=1}^{\infty} \mathbb{P}(A(1/m)) = 0
\end{aligned}$$

Zusammen:

f.s.-Konvergenz gilt $\Leftrightarrow \mathbb{P}(C^c) = 0 \Leftrightarrow \forall \varepsilon > 0 : \mathbb{P}(A(\varepsilon)) = 0.$

(3) Wir setzen jetzt die obigen Argumente und **die Stetigkeit von $\mathbb{P}$ ein** und erhalten das gewünschte Resultat

$$
\begin{aligned}
& X_n \xrightarrow{\text{f.s.}} X \\
\Leftrightarrow \quad & \mathbb{P}(C^c) = 0 \\
\Leftrightarrow \quad & \forall \varepsilon > 0 : \mathbb{P}(A(\varepsilon)) = 0 \\
\Leftrightarrow \quad & \forall \varepsilon > 0 : \mathbb{P}(\lim_{n\to\infty} B_n(\varepsilon)) = 0 \\
\Leftrightarrow \quad & \forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}(B_n(\varepsilon)) = 0.
\end{aligned}
$$

7.1.7 Satz: $(X_i) \to X$ **fast sicher** gilt genau dann, wenn für alle $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(\sup_{m\geq n} |X_m - X| > \varepsilon) = 0$$

gilt.

Beweis: Siehe Henze [19, Seite 196].

7.1.8 Bemerkung: Die f.s.-Konvergenz ist äquivalent zu $\mathbb{P}(B_n(\varepsilon)) \to 0, n \to \infty$. Die n.W.-Konvergenz ist äquivalent zu $\mathbb{P}(A_n(\varepsilon)) \to 0, n \to \infty$

7.1.9 Satz: Für $\varepsilon > 0$, $n \in \mathbb{N}$ definieren wir wie oben

$$A_n(\varepsilon) \; = \; \{|X_n - X| > \varepsilon\}\,.$$

Dann gilt: Wenn für alle $\varepsilon > 0$ die Reihe $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n(\varepsilon))$ konvergiert, dann gilt f.s.- $\lim X_i = X$.

Das $\mathbb{P}(A_n(\varepsilon))$ eine Nullfolge ist, ist zunächst nicht hinreichend für die f.s. Konvergenz. Wenn die Konvergenz so schnell ist, dass $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n(\varepsilon))$ konvergiert, dann gilt die f.s.-Konvergenz.

Beweis: Siehe Grimmett und Stirzaker [14, Seite 356f].

7.2 Eigenschaften von und Zusammenhänge zwischen Konvergenzformen

7.2.1 Satz: f.s.-$\lim X_i = X$ impliziert n.W.-$\lim X_i = X$.

Beweis: Folgt mit Satz 7.1.6.

7.2.2 Satz: $\mathcal{L}^p$-$\lim X_i = X$ impliziert n.W.-$\lim X_i = X$.

Beweis: Siehe Grimmett und Stirzaker [14, Seite 356].

7.2.3 Satz: n.W.-$\lim X_i = X$ impliziert i.V.-$\lim X_i = X$.

Beweis: Siehe Grimmett und Stirzaker [14, Seite 355].

7.2.4 Bemerkung: i.) Die Rückrichtungen zu den Implikationen der vorhergehenden Sätze gelten im Allgemeinen nicht. Wir werden/haben für jede Rückrichtung ein Beispiel für die Nichtgültigkeit angeben.

ii.) Im Allgemeinen kann man weder von der Gültigkeit der

f.s.-Konvergenz auf die Gültigkeit der $\mathcal{L}^p$-Konvergenz schließen, noch von der Gültigkeit der $\mathcal{L}^p$-Konvergenz auf die Gültigkeit der f.s.-Konvergenz.

iii.) Es gelten "bedingte" Rückrichtungen.

7.2.5 Beispiel (wandernde Türme, z.B. Vgl. Henze [19, S. 197f]): Es sei $(\Omega, \mathcal{A}, \mathbb{P}) = ((0,1], \mathcal{B}((0,1]), \lambda_{(0,1]})$. Für $m = 1, 2, 3, ..$ und $k = 1, 2, 3, ..., 2^m$ definieren wir

$$B_{m,k} = ((k-1)2^{-m}, k2^{-m}] \text{ und } X_{2^m+k-2} = \mathbb{I}_{B_{m,k}}.$$

Wenn man die obigen Konstruktionsanweisungen ausschreibt, dann erhält man

$$X_0 = \mathbb{1}_{(0,\frac{1}{2}]}, X_1 = \mathbb{1}_{(\frac{1}{2},1]}, X_2 = \mathbb{1}_{(0,\frac{1}{4}]}, X_3 = \mathbb{1}_{(\frac{1}{4},\frac{2}{4}]}, X_4 = \mathbb{1}_{(\frac{2}{4},\frac{3}{4}]},$$
$$X_5 = \mathbb{1}_{(\frac{3}{4},1]}, X_6 = \mathbb{1}_{(0,\frac{1}{8}]}, X_7 = \mathbb{1}_{(\frac{1}{8},\frac{2}{8}]}, ...$$

Dann:

i.) Es gilt n.W.- $\lim X_i = 0$.

ii.) Es gilt $\mathcal{L}^p$- $\lim X_i = 0$.

iii.) Es gilt nicht f.s.- $\lim X_i = 0$. In der Tat ist die Wahrscheinlichkeit einen konvergenten Pfad mit ziehen Null!

Wir beobachten : (1) Also folgt aus der n.W. Konvergenz i.A. die f.s.-Konvergenz nicht.

(2) Zudem folgt aus der $\mathcal{L}^p$-Konvergenz i.A. die f.s.-Konvergenz nicht.

7.2.6 Beispiel: Es sei X_i eine Folge von unabhängigen Zufallsvariablen mit

$$X_i = \begin{cases} i^3 \text{ mit der Wahrscheinlichkeit } \frac{1}{i^2} \\ 0 \text{ mit der Wahrscheinlichkeit } 1 - \frac{1}{i^2}. \end{cases}$$

Dann gilt n.W.-$\lim X_i = 0$, jedoch gilt $\mathbb{E}(|X_i|) = \mathbb{E}(X_i) = i \to \infty$. Also gilt die Rückrichtung von Satz (7.2.2) nicht: Aus der n.W.-Konvergenz folgt i.A. die $\mathcal{L}^P$-Konvergenz nicht.

7.2.7 Beispiel: Es sei X eine Bernoulli Zufallsvariable mit $p = 1/2$. Es sei ferner $X_i = X$ für alle $i \in \mathbb{N}$ und $Y = 1 - X$. Dann gilt i.V.-$\lim X_i = Y$, jedoch gilt n.W.-$\lim X_i = Y$ nicht. Also gilt die Rückrichtung von Satz (7.2.3) nicht.

7.2.8 Beispiel: Es sei X_i eine Folge von unabhängigen Zufallsvariablen mit

$$X_i = \begin{cases} i^3 \text{ mit der Wahrscheinlichkeit } \frac{1}{i^2} \\ 0 \text{ mit der Wahrscheinlichkeit } 1 - \frac{1}{i^2}. \end{cases}$$

i.) Betrachte $A_n(\varepsilon) = \{|X_n - 0| > \varepsilon\}$. Dann ist $\mathbb{P}(A_n(\varepsilon)) < 1/n^2$. Da die Reihe $\sum_{n=1}^{\infty} 1/n^2$ konvergiert, konvergiert auch $\sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon))$. Mit Satz (7.1.9) folgt f.s.-$\lim X_i = 0$.

ii.) Wir haben schon erläutert, dass $\mathcal{L}^1\text{- }\lim X_i = 0$ **nicht** gilt. Also: Im Allgemeinen impliziert die f.s.-Konvergenz die $\mathcal{L}^1$-Konvergenz nicht.

Das ist bemerkenswert: Obwohl für fast alle Pfade $X_i(\omega) \to 0$ gilt, divergiert der *Durchschnitt*[2] $\mathbb{E}(X_i) = \mathbb{E}(|X_i|)$.

7.2.9 Beispiel: Es sei X_i eine Folge von unabhängigen Zufallsvariablen mit

$$X_i = \begin{cases} 1 \text{ mit der Wahrscheinlichkeit } \frac{1}{i} \\ 0 \text{ mit der Wahrscheinlichkeit } 1 - \frac{1}{i}. \end{cases}$$

Dann gilt $\mathcal{L}^1\text{- }\lim X_i = 0$, jedoch gilt f.s.- $\lim X_i = 0$ nicht. Im Allgemeinen impliziert die $\mathcal{L}^1$-Konvergenz die f.s.-Konvergenz also **nicht**.

Beweis: Die Wahrscheinlichkeit für $X_i = 0$ konvergiert gegen 1. Allerdings ist die Konvergenz zu *langsam*.

Für den Nachweis, dass f.s.- $\lim X_i = 0$ nicht gilt, folgen wir Grimmett und Stirzacker [14, Seite 357].

Es sei

$$\begin{aligned} B_n(\varepsilon)) &= \{\exists m \geq n \text{ mit } |X_m(\omega) - 0| > \varepsilon\} \\ &= \{\exists m \geq n \text{ mit } X_m(\omega) = 1\}. \end{aligned}$$

[2] Eigentlich ist das kein Durschnitt ...

Wir zeigen, dass für alle n und $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(B_n(\varepsilon)) = 1.$$

In der Tat

$$\begin{aligned}
\mathbb{P}(B_n(\varepsilon)) &= \mathbb{P}\left(\bigcup_{m\geq n} A_m(\varepsilon)\right) \\
&= 1 - \mathbb{P}\left(\left[\bigcup_{m\geq n} A_m(\varepsilon)\right]^c\right) \\
&= 1 - \mathbb{P}\left(\left[\bigcap_{m\geq n} A_m^c(\varepsilon)\right]\right) \\
&= 1 - \mathbb{P}\left(\lim_{r\to\infty}\left[\bigcap_{r\geq m\geq n} A_m^c\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&= 1 - \lim_{r\to\infty} \mathbb{P}\left(\left[\bigcap_{r\geq m\geq n} A_m^c(\varepsilon)\right]\right) \\
&= 1 - \lim_{r\to\infty} \prod_{r\geq m\geq n} \mathbb{P}\left(A_m^c(\varepsilon)\right) \\
&= 1 - \lim_{r\to\infty} \prod_{r\geq m\geq n} \left(1 - \frac{1}{m}\right) \\
&= 1 - \lim_{r\to\infty} \prod_{r\geq m\geq n} \frac{m-1}{m} \\
&= 1 - \lim_{r\to\infty} \frac{n-1}{n} \cdot \frac{n+1-1}{n+1} \cdot \ldots \cdot \frac{r-1}{r} \\
&= 1 - \lim_{r\to\infty} \frac{n-1}{r} \\
&= 1.
\end{aligned}
$$

7.2.10 Bemerkung: Das Beispiel (7.2.8) zeigt, wie **verzwickt** es mit der Konvergenz von Zufallsvariablen sein kann. Gemäß der starken Form der fast sicheren Konvergenz liegt Konvergenz vor, trotzdem konvergiert der Erwartungswert $\mathbb{E}(X_i)$ nicht gegen $\mathbb{E}(X)$. Obwohl mit Wahrscheinlichkeit 1 also $X_i(\omega) \to X(\omega)$ gilt, gilt $\mathbb{E}(X_i) \to \mathbb{E}(X)$ nicht. Die naheliegende Vertauschungsregel $\mathbb{E}(\text{f.s.-}\lim X_i) = \lim \mathbb{E}(X_i)$ gilt also i.A. nicht.

Wenn es eine integrierbare Majorante für die Zufallsvariablen X_i gibt, dann kann man die Grenzwerte gemäß Satz (5.2.1) vertauschen. Eine genaue Zusatzannahme ist die Annahme der gleichgradigen Integrierbarkeit.

7.2.11 Definition: Es sei X_i eine Folge von $\mathcal{L}^1$-Zufallsvariablen. Die Familie (X_i) heißt **gleichgradig integrierbar**, falls

$$\lim_{A\to\infty} \sup_{i\geq 1} \mathbb{E}(\mathbb{1}_{|X_i|\geq A} \cdot |X_i|) = 0.$$

7.2.12 Satz: Es sei $X_i, i \in \mathbb{N}$ eine Folge von $\mathcal{L}^1$-Zufallsvariablen, X eine $\mathcal{L}^1$-Zufallsvariable und $X = \text{n.W.-lim}_{i\to\infty} X_i$. Dann ist äquivalent:

i.) (X_i) ist gleichgradig integrierbar.

ii.) $\mathcal{L}^1\text{-lim}\, X_i = X$.

Beweis: Vgl. Durrett [6, Seite 245] bzw. Grimmet und Stirzacker [14, Seite 396]

7.2.13 Satz: Es gilt:

i.) Aus $\mathcal{L}^1\text{-lim}_{n\to\infty}\, X_i = X$ folgt $\lim_{n\to\infty} \mathbb{E}(X_i) = \mathbb{E}(X)$ und $\lim_{n\to\infty} \mathbb{E}(|X_i|) = \mathbb{E}(|X|)$.

ii.) Aus $\mathcal{L}^2\text{-lim}_{n\to\infty}\, X_i = X$ folgt $\lim_{n\to\infty} \mathbb{E}(X_i^2) = \mathbb{E}(X^2)$

Beweis: Der erste Teil von i.) sieht fast selbstverständlich aus. Wir haben aber schon gesehen, dass wir bei ZV vorsichtig sein müssen. Für den Beweis verwendet man die **Drei-**

ecksungleichung

$$|\mathbb{E}(X_i - X)| \leq \mathbb{E}(|X_i - X|).$$

Gemäß Definition der $\mathcal{L}^1$-Konvergenz gilt $\mathbb{E}(|X_i - X|) \to 0$, d.h. die rechte Seite konvergiert gegen 0. Also konvergiert auch die linke Seite gegen 0.

Für den zweiten Teil von i.) nutzen wir die umgekehrte Dreiecksungleichung: Für $x, y \in \mathbb{R}$ gilt

$$\big||x| - |y|\big| \leq |x - y|.$$

Dann folgt aus der **Monotonie** des Erwartungswertes:

$$\mathbb{E}(\big||X_i| - |X|\big|) \leq \mathbb{E}(|X_i - X|).$$

Gemäß Definition der $\mathcal{L}^1$-Konvergenz gilt $\mathbb{E}(|X_i - X|) \to 0$. Also konvergiert auch die linke Seite gegen 0.

Für den Beweis von ii.) beachten wir den Hinweis in Jacod und Protter [21, Seite 149]

7.2.14 Satz: Es gilt:

i.) f.s.-lim $X_i = X$ und f.s.-lim $Y_i = Y$ impliziert f.s.-lim $X_i + Y_i = X + Y$.

ii.) $\mathcal{L}^p$-lim $X_i = X$ und $\mathcal{L}^p$-lim $Y_i = Y$ impliziert $\mathcal{L}^p$-lim

$X_i + Y_i = X + Y$.

iii.) n.W.-lim $X_i = X$ und n.W.-lim $Y_i = Y$ impliziert n.W.-lim $X_i + Y_i = X + Y$.

Für die i.V.-Konvergenz gilt die Implikation im Allgemeinen nicht.

Beweis: Vgl. Grimmett und Stirzaker [15, Ex. 7.11.2] und Henze [19, Seite 210]

7.2.15 Satz (Sluzki): Es gilt:

i.) Aus i.V.-lim $X_i = X$ und n.W.-lim $|X_i - Y_i| = 0$ folgt i.V.-lim $Y_i = X$.

i.a) Aus i.V.-lim $X_i = X$ und n.W.-lim $Y_i = a$ folgt i.V.-lim $X_i + Y_i = X + a$.

ii.) Aus i.V.-lim $X_i = X$ und n.W.-lim $Y_i = c$ folgt i.V.-lim $X_i Y_i = cX$.

iii.) Aus i.V.-lim $X_i = X$ und n.W.-lim $Y_i = c, c \neq 0$ folgt i.V.-lim $X_i/Y_i = X/c$.

Beweis: Vgl. Henze [19, Seite 209]

7.2.16 Satz: Es sei $f : \mathbb{R} \to \mathbb{R}$ eine stetige Funktion.

i.) Aus i.V.-lim $X_i = X$ folgt i.V.-lim $f(X_i) = f(X)$.

ii.) Aus n.W.-lim $X_i = X$ folgt n.W.-lim $f(X_i) = f(X)$.

iii.) Aus f.s.-lim $X_i = X$ folgt f.s-lim $f(X_i) = f(X)$.

Beweis: Siehe Jacod und Protter [21, S. 146] für ii.) und iii.) und Grimmett und Stirzaker [14, S. 360] für i.)

8 Gesetze der großen Zahlen

8.1 Definitionen GgZ

8.1.1 Definition: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von Zufallsvariablen mit $\mathbb{E}(X_i) = \mu$ für alle $i \in \mathbb{N}$. Man sagt, dass (X_i) dem **schwachen Gesetz der großen Zahlen** genügt, falls

$$\text{n.W.-}\lim \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

Der Durchschnitt konvergiert n.W. gegen den Erwartungswert.

8.1.2 Definition: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von Zufallsvariablen mit $\mathbb{E}(X_i) = \mu$ für alle $i \in \mathbb{N}$. Man sagt, dass (X_i) dem **starken Gesetz der großen Zahlen** genügt, falls

$$\text{f.s.-}\lim \frac{1}{n}\sum_{i=1}^{n} X_i = \mu.$$

8.1.3 Beispiel: Es sei $\mathbb{E}(X) = \mu$ und $X_i = X$. Dann

$$\text{f.s.-}\lim \frac{1}{n}\sum_{i=1}^{n} X_i = X.$$

Also gilt das GgZ hier i.A. nicht.

8.1.4 Bemerkung: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von Zufallsvariablen mit $\mathbb{E}(X_i) = \mu$ für alle $i \in \mathbb{N}$. Es ist naheliegend die Stichprobenfunktion $Y_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ als Schätzfunktion für den Parameter μ zu verwenden. Es gilt immerhin für alle $N \in \mathbb{N}$, dass

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{N\mu}{N} = \mu.$$

Hier haben wir folgende Situation: Wir entnehmen eine zufällige Stichprobe $(X_1, ..., X_N)$ der Größe N und betrachten den Erwartungswert von $Y_N = \frac{1}{N}\sum_{i=1}^{N} X_i$. Die Tatsache, dass $\mathbb{E}(Y_N) = \mu$ gilt, nennt man **Erwartungstreue** der Stichprobenfunktion Y_N. In den vorhergehenden Definitionen betrachten wir das Verhalten von $Y_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ für beliebig groß werdendes N. Wenn (8.1.1) gilt, dann sagt (Y_N) ist

(schwach) konsistent für μ. Wenn (8.1.2) gilt, dann sagt (Y_N) ist **stark konsistent für** μ.

8.2 Gesetze der großen Zahlen

8.2.1 Satz: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von unkorrelierten $\mathcal{L}^2$-Zufallsvariablen mit $\mathbb{E}(X_i) = \mu$ und $\mathbb{V}(X_i) = \sigma^2$ für alle $i \in \mathbb{N}$.

$$\text{n.W.-}\lim \frac{1}{n} \sum_{i=1}^{n} X_i = \mu$$

und

$$\mathcal{L}^2\text{-}\lim \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

Beweis: Es sei $Y_n = \frac{1}{n} \sum_{i=1}^{n} X_n$. Aus der Ungleichung von Chebychev (5.6.3) folgt

$$\mathbb{P}(|Y_n - \mu| \geq \varepsilon) \leq \frac{\mathbb{V}(Y_n)}{\varepsilon^2}.$$

Ferner folgt:

$$\mathbb{V}(Y_n) = \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^{n} X_n\right) = \frac{1}{n^2} \sigma^2 n = \frac{\sigma^2}{n}.$$

Somit gilt:

$$\mathbb{P}(|Y_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ für } n \to \infty,$$

also die n.W.-Konvergenz.

Es gilt

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \mu.$$

Ferner gilt:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \\ &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \to 0\end{aligned}$$

für $n \to \infty$.

Wir wissen schon, dass aus der $\mathcal{L}^2$-Konvergenz die n.W.-Konvergenz folgt. Auf den Beweis der n.W.-Konvergenz hätten wir eigentlich verzichten können. □

Wir geben noch ohne Beweis die allgemeine Aussage an, gemäß derer unter den obigen Bedingungen sogar die f.s.-Konvergenz gilt. Für den Beweis vergleiche z.B. Chung [5, Theorem 5.1.2] oder Krengel [24, Satz 12.4]

8.2.2 Satz (Rajchman[1]): Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von **unkorrelierten quadrat-integrierbaren $\mathcal{L}^2$-**Zufallsvariablen mit $\mathbb{E}(X_i) = \mu$ und $\mathbb{V}(X_i) = \sigma^2$ für alle $i \in \mathbb{N}$. Dann gilt

$$\text{f.s.-}\lim \frac{1}{n} \sum_{i=1}^{n} X_i = \mu$$

und

$$\mathcal{L}^2\text{-}\lim \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

Beweis: Krengel [24] oder Chung [5]

8.2.3 Satz (Etemadi): Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von paarweise unabhängigen und identisch verteilten $\mathcal{L}^1$-Zufallsvariablen. Dann gilt

$$\text{f.s.-}\lim \frac{1}{n} \sum_{i=1}^{n} X_i = \mu$$

und

$$\mathcal{L}^1\text{-}\lim \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

Beweis: Vgl. Billingsley [2, Section 22] oder Klenke [23, Satz 5.17] oder Henze [19, Seite 201] für den Beweis der f.s.-Konvegenz.

[1] https://de.wikipedia.org/wiki/Aleksander_Rajchman

Die $\mathcal{L}^1$-Konvergenz folgt dann wegen i.i.d. relativ einfach; bekanntlich folgt die $\mathcal{L}^1$-Konvergenz i.a. nicht aus der f.s.-Konvergenz.

8.2.4 Bemerkung: Wir betrachten ein Beispiel einer Folge von Zufallsvariablen für die das schwache Gesetz der großen Zahlen gilt, das starke Gesetz jedoch nicht; vgl. Grinstead und Snell [16, Seite 314f, Nr. 16].

Wir betrachten eine Folge unabhängiger Zufallsvariablen (X_i) mit

$$\mathbb{P}(X_i = i) = \frac{1}{2i \log i},\ \mathbb{P}(X_i = -i) = \frac{1}{2i \log i} \text{ und}$$
$$\mathbb{P}(X_i = 0) = 1 - \frac{1}{i \log i}$$

8.2.5 Bemerkung: i.) Gesetze der großen Zahlen sind von **grundsätzlicher Bedeutung** für die Wahrscheinlichkeitstheorie. Angenommen wir betrachten die fortgesetzte unabhängige Wiederholungen eines Experiments, dessen Ergebnis Erfolg oder Misserfolg sein kann – also $\Omega = \{\text{Erfolg, Misserfolg}\}$ – und bei dem die Erfolgswahrscheinlichkeit $p = \mathbb{P}(\{\text{Erfolg}\})$ beträgt. Es sei $X_i(\text{“Erfolg bei der } i\text{-ten Wiederholung”}) = 1$. Dann ist $S_n = \sum_{i=1}^n X_i$ die Anzahl der Erfolge nach n Wiederholungen. Der *gesunde Menschenverstand* besagt, dass relative Häufigkeit S_n/n gegen p „konvergiert“. In der Tat, wird eine entsprechende Aussage oft als „Definition“ von Wahr-

scheinlichkeit angesehen.[2] Gemäß des Satzes von Kolmogorow bzw. Etemati gilt die Konvergenzaussage in der Tat fast sicher bzw. in $\mathcal{L}^1$ (also in den denkbar stärksten Formen). Das Gesetz ist jedoch hier nicht Ausgangspunkt (Axion/Definition), sondern das Resultat der Überlegungen! Der Ausgangspunkt – das Axiomensystem nach Kolmogoroff – ist also auch diesbezüglich gut gewählt.

ii.) Für unabhängig wiederholte Bernoulli Experimente geht das schwache Gesetz der großen Zahlen auf Jakob Bernoulli zurück (veröffentlicht 1713). Das entsprechende starke Gesetz hat Borel (veröffentlicht 1909) bewiesen.

iii.) Für eine Folge von unabhängigen und identisch verteilten $\mathcal{L}^1$-Zufallsvariablen $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ hat Kolmogorow das Gesetz der großen Zahlen 1930 bewiesen. Nasrollah Etemadi [9] hat die *Verallgemeinerung* (paarweise unabhängig) 1981 *elementar* bewiesen.

iv.) Für unkorrelierte Zufallsvariablen deren Varianz es eine gemeinsame Schranke gibt, gilt gemäß des Satzes von Rajchman auch das starke Gesetz der großen Zahl (vgl. auch die nächste Ordnungsnummer). In diesem Sinn ist also nicht der Unterschied zwischen Unkorreliertheit und Unabhängigkeit für die Gültigkeit des schwachen anstatt des starken Gesetzes ausschlagend. ◁

[2] Vgl. den Wikipediaeintrag `https://de.wikipedia.org/wiki/Wahrscheinlichkeit` unter Häufigkeitsprinzip – statistische Wahrscheinlichkeitsauffassung.

8.2.6 Bemerkung (vgl. Chung [5, S. 107]): Wir betrachten zunächst den Fall $X_i \in \mathcal{L}^2$. Wir erlauben, dass die Zufallsvariablen X_i abhängig sind und unterschiedliche Varianzen $\sigma_i^2 = \mathbb{V}(X_i) = \mathbb{E}(X_i^2)$ haben. Ohne wesentliche Einschränkung der Allgemeinheit betrachten wir zentrierte Zufallsvariablen $\mathbb{E}(X_i) = 0$.

Wir definieren die Partialsummen

$$S_n = \sum_{i=1}^{n} X_i$$

und untersuchen die Konvergenz von $\frac{S_n}{n}$. Für die $\mathcal{L}^2$-Konvergenz gilt

$$\mathcal{L}^2\text{-lim } \frac{S_n}{n} = 0 \;\Leftrightarrow\; \lim \mathbb{E}\left(\left(\frac{S_n}{n}\right)^2\right) = 0 \;\Leftrightarrow\; \mathbb{E}\left(S_n^2\right) = o(n^2).$$

Mit wachsenden n wird $\mathbb{E}\left(S_n^2\right)$ anwachsen. Wenn dieses Wachstums hinreichend langsam ist, dann gilt die $\mathcal{L}^2$-Konvergenz. Aus der $\mathcal{L}^2$-Konvergenz folgt bekanntlich die Konvergenz nach Wahrscheinlichkeit.

Also: Wenn $\mathbb{E}\left(S_n^2\right)$ hinreichend langsam wächst, dann gilt das schwache Gesetz der großen Zahlen. Wir erhalten die Implikation

$$\mathbb{E}\left(S_n^2\right) = o(n^2) \Rightarrow \text{n.W.-lim } \frac{S_n}{n} = 0.$$

Uns interessiert also die Größenordnung von $\mathbb{E}\left(S_n^2\right)$. Wir be-

achten dazu

$$\mathbb{E}(S_n^2) = \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \sum_{j \neq k} \mathbb{E}(X_j X_k)$$

und betrachten zwei Fälle:

- Gilt $\mathbb{E}(X_j X_k) = 0, i \neq j$ (man sagt, dass die Zufallsvariablen X_i orthogonal[3] sind) und $\sigma_i^2 \leq M, i \in \mathbb{N}$, so gilt

$$\begin{aligned} \mathbb{E}(S_n^2) &= \sum_{i=1}^{n} \mathbb{E}(X_i^2) \leq nM \\ \Rightarrow \frac{\mathbb{E}(S_n^2)}{n^2} &\leq \frac{M}{n} \end{aligned}$$

  Also $\mathbb{E}(S_n^2) = o(n^2)$ und es folgt das schwache Gesetz der großen Zahlen.

- Wenn wir lediglich voraussetzen, dass die Varianzen und Kovarianzen eine gemeinsame Schranke $\operatorname{cov}(X_i, X_j) = \mathbb{E}(X_i X_j) \leq M$ haben, dann folgt im allgemeinen das Gesetz der großen Zahlen jedoch nicht. Dazu betrachten wir

$$\mathbb{E}(S_n^2) = \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \sum_{j \neq k} \mathbb{E}(X_j X_k).$$

  Auf der rechten Seite stehen n^2 Terme. Die rechte Seite

[3] Gilt $\mathbb{E}(X_i) = 0$, dann $\mathbb{E}(X_i X_j) = \operatorname{cov}(X_i, X_j)$.

ist dann $O(n^2)$, d.h.

$$\begin{aligned} \mathbb{E}(S_n^2) &\leq Mn^2 \\ \Rightarrow \frac{\mathbb{E}(S_n^2)}{n^2} &\leq M. \end{aligned}$$

Im Allgemeinen folgt so $\mathbb{E}\left(S_n^2\right) = o(n^2)$ nicht; dazu müsste $\frac{\mathbb{E}(S_n^2)}{n^2}$ nicht nur beschränkt sein, sondern sogar gegen Null gehen.

Wenn wir uns die Gleichung

$$\mathbb{E}(S_n^2) = \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \sum_{j \neq k} \mathbb{E}(X_j X_k).$$

nochmal ansehen, dann erkennen wir, dass sich für denn Fall $\mathbb{E}(X_j X_k) \neq 0$ i.A. zu *viele Terme* ergeben. Gilt $\mathbb{E}(X_j X_k) = 0$, dann sind es n Terme; für $\mathbb{E}(X_j X_k) \neq 0$ jedoch n^2.

Der folgende Satz fasst die obige Bemerkung zusammen. Anschließend entwickeln wir Argumente für die Gültigkeit des Gesetzes der großen Zahlen für abhängige Zufallsvariablen. Solche Sätze nennt man **Ergodensätze**.

8.2.7 Satz: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von **unkorrelierten quadrat-integrierbaren** $\mathcal{L}^2$-Zufallsvariablen mit $\mathbb{E}(X_i) = \mu$ und $\mathbb{V}(X_i) = \sigma_i^2 \leq M$ für alle $i \in \mathbb{N}$. Dann

gilt

$$\mathcal{L}^2\text{- }\lim \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

$$\text{n.W.- }\lim \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

$$\text{f.s.- }\lim \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

Die Verallgemeinerung gegenüber Satz (8.2.1) besteht darin, dass die Zufallsvariablen unterschiedliche Varianzen haben dürfen. Unter den genannten Voraussetzungen gilt – wie in der dritten Gleichung angegeben – sogar das starke Gesetz der großen Zahlen. Siehe Chung [5, Seite 108]

Für korrelierte Zufallsvariablen deren zweite Momente eine gemeinsame Schranke haben – also $\mathbb{E}(X_i X_j) \leq M$ –, gilt somit das Gesetz der großen Zahlen **im Allgemeinen nicht**. Solche Gesetze gelten jedoch unter als nächstes erläuterten passenden Bedingungen. Diese stellen sicher, dass Abhängigkeiten hinreichend *schwach* sind.

8.3 GgZ bei Abhängigkeiten und Ergodizität

8.3.1 Definition: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von $\mathcal{L}^2$-Zufallsvariablen. (X_i) heißt **schwach stationärer stochastischer Prozess**, falls

- $\mathbb{E}(X_i) = \mu$ für alle $i \in \mathbb{N}$. Alle Zufallsvariablen haben den gleichen Erwartungswert.
- $\text{cov}(X_i, X_j) = \gamma(i-j)$ für alle $i, j \in \mathbb{N}$. Die Kovarianz $\text{cov}(X_i, X_j)$ von X_i und X_j hängt nur von der Differenz $i - j$ ab.

8.3.2 Bemerkung: Es sei $X_i = X$ für eine Zufallsvariable X mit $\mathbb{E}(X) = \mu, \mathbb{V}(X) = \sigma^2$. Dann gilt $\mathbb{E}(X_i) = \mu$ und $\gamma(i-j) = \text{cov}(X_i, X_j) = \text{cov}(X, X) = \sigma^2$. Also ist X_i stationär. Das Gesetz der großen Zahlen gilt natürlich nicht:

$$\text{n.W.-}\lim \frac{1}{n} \sum_{i=1}^{n} X_i = X \neq \mu.$$

8.3.3 Satz: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ ein schwach statio-

närer Prozess. Wenn $\gamma_k \to 0, k \to \infty$, dann

$$\mathcal{L}^2\text{- }\lim \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

$$\text{n.W.- }\lim \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

(Man sagt, dass X_i mittelwert-ergodisch ist.)

Beweis: Karlin und Taylor [22, Seite 476] beweisen eine stärkere Aussage. Der Beweis von dort ist auf die hier betrachtet Situation angepasst. Wir unterstellen o.B.d.A., dass $\mathbb{E}(X_i) = 0$. Wir definieren $\text{cov}(X_i, X_j) = \gamma(i-j) = \gamma_h, h = i-j$. Dann folgt

$$\begin{aligned}
\mathbb{E}(S_n^2) &= \sum_{i=1}^{n} \mathbb{E}(X_i^2) + 2 \sum_{0 \le i < k \le n} \mathbb{E}(X_i X_k) \\
&= n\sigma^2 + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n} \gamma(k-j) \\
&= n\sigma^2 + 2\sum_{j=1}^{n-1}(n-j)\gamma_j = n\sigma^2 + 2n\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\gamma_j
\end{aligned}$$

Wir erhalten als Bedingung für die $\mathcal{L}^2$-Konvergenz:

$$\frac{1}{n}\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\gamma_j \to 0.$$

Wir weise jetzt nach, dass $\gamma_j \to 0$ dafür hinreichend ist.

Zunächst wählen wir m so groß, so dass $|\gamma_i| \leq \frac{\varepsilon}{2}$ für alle $i \geq m$. Das geht, denn $\gamma_i \to 0$. Dann gilt für $n-1 \geq m$ bzw. für $n \geq m+1$

$$\frac{1}{n}\sum_{j=m}^{n-1} |\gamma_j| \leq \frac{1}{n}\sum_{j=m}^{n-1} \frac{\varepsilon}{2} \leq \frac{1}{n} n \frac{\varepsilon}{2} = \frac{\varepsilon}{2}.$$

Jetzt wählen wir $n \geq m+1$ (m bleibt fest), so dass

$$\frac{1}{n}\sum_{j=1}^{m-1} |\gamma_j| \leq \frac{\varepsilon}{2}.$$

Das geht, denn m ist fest.

Jetzt setzen wir die beiden Abschätzungen zusammen:

$$\frac{1}{n}\sum_{j=1}^{n-1} |\gamma_j| = \frac{1}{n}\sum_{j=1}^{m-1} |\gamma_j| + \frac{1}{n}\sum_{j=m}^{n-1} |\gamma_j| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Also

$$\frac{1}{n}\sum_{j=1}^{n-1} |\gamma_j| \to 0.$$

Wir erhalten die gewünschte Konvergenz

$$\left|\frac{1}{n}\sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) \gamma_j\right| \leq \frac{1}{n}\sum_{j=1}^{n-1} \left|\left(1 - \frac{j}{n}\right) \gamma_j\right| \leq \frac{1}{n}\sum_{j=1}^{n-1} |\gamma_j| \to 0$$

8.4 Hauptsatz der Statistik – Glivenko-Cantelli

8.4.1 Definition: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Familie von identisch und unabhängig verteilten Zufallsvariablen. Dann heißt

$$F_n(x,\omega) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(X_i(\omega)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i(\omega)\leq x}$$

die **empirische Verteilungsfunktion** von $X_1, ..., X_n$.

Beachte: Für festes x ist $\omega \mapsto F_n(x,\omega)$ eine Zufallsvariable!

8.4.2 Satz: Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von unabhängig und identisch verteilten Zufallsvariablen mit Verteilungsfunktion F. Dann gilt für alle $x \in \mathbb{R}$

$$\text{f.s.-}\lim_{n\to\infty} F_n(x) = F(x)$$

Es gilt sogar (**Satz von Glivenko-Cantelli**)

$$\text{f.s.-}\lim_{n\to\infty} \left(\sup_{x\in\mathbb{R}} |F_n(x,\omega) - F(x)| \right) = 0.$$

Beweis: Vgl. Henze [19, Seite 277]

9 Asymptotische Verteilung

9.1 Zentraler Grenzwertsatz

9.1.1 Satz (Lindeberg-Levy): Es sei $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$ eine Folge von unabhängig und identisch verteilten $\mathcal{L}^2$-Zufallsvariablen mit $\mathbb{E}(X_i) = \mu$ und $\mathbb{V}(X_i) = \sigma^2 > 0$ für alle $i \in \mathbb{N}$. Es sei

$$\begin{aligned} S_n &= \sum_{k=1}^{n} X_k \\ S_n^* &= \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - \mathbb{E}(S_n)}{\mathrm{std}(S_n)} \end{aligned}$$

Dann konvergiert S_n^* für $n \to \infty$ in Verteilung gegen eine Zufallsvariable mit Standardnormalverteilung; es gilt also

$$\text{i.V.-}\lim S_n^* = Z,\ Z \sim \mathcal{N}(0,1),$$

dabei hat Z die Verteilungsfunktion

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dz.$$

Beweis: Siehe Henze [19, S. 215f]

9.1.2 Satz (ZGS-MW-Approximation[1]): Es sei $\mathbb{E}(X_i) = \mu$ und $\mathbb{V}(X_i) = \sigma^2$. Für großes N ist

$$\bar{X}_N := \frac{1}{N}(X_1 + ... + X_N)$$

approximativ $N\left(\mu, \frac{\sigma^2}{N}\right)$:

$$\bar{X}_N \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right).$$

Beweis: Zunächst beobachten wir

$$S_N^* = \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{N(\bar{X} - \mu)}{\sigma\sqrt{N}} = \sqrt{N}\left(\frac{\bar{X}_N - \mu}{\sigma}\right).$$

Gemäß ZGS ist $S_N^* = \sqrt{N}\left(\frac{\bar{X}_N - \mu}{\sigma}\right)$ näherungsweise $\mathcal{N}(0,1)$. Also insbesondere $\mathbb{E}(S_N^*) = 0$ und $\mathbb{V}(S_N^*) = 1$. Ferner gilt

$$\bar{X}_N = \sigma S_N^* \frac{1}{\sqrt{N}} + \mu$$

[1] MW steht für Mittelwert und ZGS für Zentraler Grenzwertsatz.

Dann folgt $\mathbb{E}(\bar{X}_N) = \mu$ und

$$\mathbb{V}(\bar{X}_N) = \mathbb{V}\left(\sigma S_N^* \frac{1}{\sqrt{N}} + \mu\right) = \sigma^2 \frac{1}{N}.$$

▶ Vielleicht ist die Aussage $\bar{X}_N \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$ insbesondere für Anwender noch interessanter als die, die man ZGS nennt. Da die Varianz aber von N abhängt, ist sie *unbequemer*.

9.1.3 Bemerkung: Gemäß GgZ gilt

$$\left(\frac{\bar{X}_N - \mu}{\sigma}\right) \to 0.$$

Gemäß ZGS: Wenn man mit $\sqrt{N}$ dann ergibt sich eine nicht-triviale Verleitung

$$\sqrt{N}\left(\frac{\bar{X}_N - \mu}{\sigma}\right) \to Z, Z \sim \mathcal{N}(0,1)$$

9.1.4 Bemerkung: Wir beachten

$$\mathbb{V}(\bar{X}_N - \mu) = \mathbb{V}(\bar{X}_N) = \frac{\sigma^2}{N}$$

also auch

$$\mathrm{sd}(\bar{X}_N - \mu) = \frac{\sigma}{\sqrt{N}}$$

Die Standardabweichung wir also für größeres N immer kleiner. **Der MW liegt immer näher am Erwartungswert.**

Allerdings ist die Konvergenz langsam; nämlich nur von Ordnung $N^{1/2}$.

9.1.5 Bemerkung: i.) Egal welches reellwertige Experiment (mit endlicher Varianz) unabhängig wiederholt wird, für hinreichend viele Wiederholungen entspricht die **Verteilung der standardisierten Summe der Zufallsergebnisse** näherungsweise der Verteilung der **Standardnormalverteilung**.

ii.) Wenn sich das Ergebnis eines Zufallsexperiments durch unabhängige additive Überlagerung einer großen Zahl von Zufallsgrößen ergibt, dann erhalten wir näherungsweise normalverteilte Beobachtungen.

iii.) Salopp (vgl. Henze [19, 216]): Eine Summe S_n aus vielen unabhängig und identisch verteilten Summanden *vergisst* – bis auf Erwartungswert und Varianz vergisst – im Limes für $n \to \infty$ die Verteilung der einzelnen Summanden.

9.1.6 Bemerkung: Regelmäßig verwendet man den Zentralen Grenzwertsatz als Grundlage für Approximationen.

i.) Die erste Variante betrachtet $S_n = \sum_{k=1}^{n} X_k$.

S_n^* ist für großes n näherungsweise $\mathcal{N}(0,1)$. Wir können also die Wahrscheinlichkeiten $\mathbb{P}(S_n^* \leq k) \approx \Phi(k)$ gemäß der

Standardnormalverteilung bestimmen. Ferner

$$\begin{aligned} & S_n^* \leq k \\ \Leftrightarrow \quad & \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq k \\ \Leftrightarrow \quad & S_n - n\mu \leq k\sigma\sqrt{n} \\ \Leftrightarrow \quad & S_n \leq n\mu + k\sigma\sqrt{n} \end{aligned}$$

Wir erhalten dann die nützliche Formel

$$\mathbb{P}(S_n \leq n\mu + k\sigma\sqrt{n}) \approx \Phi(k).$$

Analog für um Null symmetrische Intervalle

$$\begin{aligned} & -k \leq S_n^* \leq k \\ \Leftrightarrow & -k \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq k \\ \Leftrightarrow & -k\sigma\sqrt{n} \leq S_n - n\mu \leq k\sigma\sqrt{n} \\ \Leftrightarrow & n\mu - k\sigma\sqrt{n} \leq S_n \leq n\mu + k\sigma\sqrt{n} \end{aligned}$$

bzw.

$$\begin{aligned} & -k \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq k \\ \Leftrightarrow & k\sigma\sqrt{n} \geq n\mu - S_n \geq -k\sigma\sqrt{n} \\ \Leftrightarrow & S_n + k\sigma\sqrt{n} \geq n\mu \geq S_n - k\sigma\sqrt{n} \end{aligned}$$

Gemäß ZGS gilt dann

$$\mathbb{P}\left(S_n + k\sigma\sqrt{n} \geq n\mu \geq S_n - k\sigma\sqrt{n}\right) \approx 2\Phi(k) - 1 \text{ bzw}$$
$$\mathbb{P}\left(n\mu - k\sigma\sqrt{n} \leq S_n \leq n\mu + k\sigma\sqrt{n}\right) \approx 2\Phi(k) - 1$$

wobei wir $\Phi(k) - \Phi(-k) = 2\Phi(k) - 1$ verwenden.

ii.) Die zweite (eigentlich äquivalente) Variante betrachtet stattdessen den **Durchschnitt**. Es gilt

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{n(\frac{1}{n}S_n - \mu)}{\sigma\sqrt{n}} = \frac{\frac{1}{n}S_n - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

Also ist $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \dot{\sim} \mathcal{N}(0,1)$

Wir beobachten (analog zu oben) um symmetrische Intervalle

$$\begin{aligned} -k &\leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq k \\ \Leftrightarrow -k\sigma/\sqrt{n} &\leq \bar{X}_n - \mu \leq k\sigma/\sqrt{n} \\ \Leftrightarrow \mu - k\sigma/\sqrt{n} &\leq \bar{X}_n \leq \mu + k\sigma/\sqrt{n} \end{aligned}$$

bzw.

$$\begin{aligned} -k &\leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq k \\ \Leftrightarrow k\sigma/\sqrt{n} &\geq \mu - \bar{X}_n \geq -k\sigma/\sqrt{n} \\ \Leftrightarrow \bar{X}_n - k\sigma/\sqrt{n} &\geq \mu \geq \bar{X}_n + k\sigma/\sqrt{n} \end{aligned}$$

Gemäß ZGS gilt

$$\mathbb{P}\left(\bar{X}_n - k\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + k\sigma/\sqrt{n}\right) \approx 2\Phi(k) - 1 \text{ bzw.}$$
$$\mathbb{P}\left(\mu - k\sigma/\sqrt{n} \leq \bar{X}_n \leq \mu + k\sigma/\sqrt{n}\right) \approx 2\Phi(k) - 1$$

9.1.7 Satz von Moivre-Laplace: Es sei (X_i) eine Folge unabhängiger Bernoulli-verteilter Zufallsvariablen und

$$S_n = \sum_{i=1,\ldots,n} X_i$$

und

$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}. \tag{9.1}$$

Dann gilt

$$\text{i.V.-}\lim S_n^* = Z, Z \sim \mathcal{N}(0,1). \tag{9.2}$$

9.1.8 Bemerkung: Der Satz von Moivre-Laplace ist zwar einerseits lediglich ein Spezialfall des Satzes von Lindeberg-Levy. Andererseits wahrscheinlich aber die am häufigsten verwendete Variante; vgl. Blitzstein und Hwang [4, S. 475]. Praktisch verwendet man dabei die sogenannte *Stetigkeitskorrektur*.

Also, wenn $Y \sim \text{Binomial}(n,p)$, dann ist $Y = X_1 + \ldots + X_n$

mit $X_i \sim$ iiBernuolli(p). Dann gilt die Approximation

$$Y \dot{\sim} \mathcal{N}(np, np(1-p))$$

Dann

$$\begin{aligned}\mathbb{P}(Y = k) &= \mathbb{P}\left(k - \frac{1}{2} < Y < k + \frac{1}{2}\right) \\ &\approx \Phi\left(\frac{k + 1/2 - np}{\sqrt{np(1-p)}} -\right) - \Phi\left(\frac{k - 1/2 - np}{\sqrt{np(1-p)}} -\right)\end{aligned}$$

9.1.9 Bemerkung: Wir betrachten eine Umfrage, die ermitteln soll, wie viele Personen die Alternative A der Alternative B vorziehen; vgl. Grinstead und Snell [16, S. 333]. Wir fassen jede Antwort als das Ergebnis eines **Bernoulli-Experiments** mit Erfolgswahrscheinlichkeit p auf. *Erfolg* bedeutet, dass wir zufällig jemanden *gezogen/angerufen* haben, der/die die Alternative A bevorzugt. Der naheliegende Schätzer für p ist $\hat{p} = \bar{p} = \frac{S_n}{n}$.

Wir beobachten zunächst

$$\begin{aligned}\bar{p} = \frac{S_n}{n} &= \frac{S_n}{n} + p - p = \frac{S_n - pn}{n} + p \\ &= \frac{S_n - pn}{\sqrt{pq}n}\sqrt{pq} + p = \frac{S_n - pn}{\sqrt{pq}\sqrt{n}\sqrt{n}}\sqrt{pq} + p \\ &= \frac{S_n - pn}{\sqrt{pqn}}\sqrt{\frac{pq}{n}} + p \\ &= S_n^*\sqrt{\frac{pq}{n}} + p\end{aligned}$$

Jetzt können wir den zentralen Grenzwertsatz verwenden, denn $S_n^* := \frac{S_n - pn}{\sqrt{pqn}}$ ist für großes n näherungsweise standardnormal, also

$$\mathbb{P}\left(\bar{p} - 2\sqrt{\frac{pq}{n}} < p < \bar{p} + 2\sqrt{\frac{pq}{n}}\right) \approx 0.954$$

Mit dieser können wir noch nicht *konkret* rechnen, p und q sind nicht bekannt. Wir setzen noch $\bar{p}$ für p und $\bar{q}$ für q ein (das ist eine weitere Approximation) und erhalten:

$$\mathbb{P}\left(\bar{p} - 2\sqrt{\frac{\bar{p}\bar{q}}{n}} < p < \bar{p} + 2\sqrt{\frac{\bar{p}\bar{q}}{n}}\right) \approx 0.954$$

Das **zufällige** Intervall

$$\left(\bar{p} - 2\sqrt{\frac{\bar{p}\bar{q}}{n}}\,,\, \bar{p} + 2\sqrt{\frac{\bar{p}\bar{q}}{n}}\right)$$

heißt das 95%-**Konfidenzintervall** für den unbekannten (aber festen) Parameter p. Mit einer Wahrscheinlichkeit von 0.95 enthält dieses zufällige Intervall den Parameter p. Man darf dabei nicht den Fehler machen und p als zufällig aufzufassen. p liegt zufällig im Intervall, weil das Zufalls-Intervall den fixen Wert p enthät. Die Abbildung in Grinsead und Snell [16, S. 333] ist diesbezüglich erhellend!

Die Länge des oben genannten Konfidenzintervall ist von ab-
hängig. Wenn das Konfidenzintervall zu lang ist, dann ist die

Umfrage nicht sehr *informativ*. Das Umfrageinstitut möchte, dass das Konfidenzintervall nicht länger als 0.06 ist. Es soll ein Abschätzung für n ermittelt werden, so dass

$$2\sqrt{\frac{\bar{p}\bar{q}}{n}} \leq 0.03$$

Verwendet man noch die Ungleichung $\bar{p}\bar{q} \leq 1/4$. Dann ist $n \geq 1111$ hinreichend.

9.1.10 Beispiel (vgl. Blitzstein und Hwang [4]): Die ZV X_i seien unabhängig Bernoulli mit $\text{Bild}(X_i) = \{0, 1\}$ und $\mathbb{P}(X_i = 1) = \frac{1}{2}$. Gemäß des GgZ gilt f.s.- $\lim \bar{X}_N = \mu$.

Es ist $\mathbb{E}(X_i) = \frac{1}{2}$ und $\mathbb{V}(X_i) = \frac{1}{4}$.

Also gemäß ZGS-Approximation

$$\bar{X}_N \dot{\sim} \mathcal{N}\left(\frac{1}{2}, \frac{1}{4N}\right)$$

Für $N = 100$ liegt mit Wahrscheinlichkeit 0.95 $\bar{X}_N$ im Intervall $[0.4, 0.6]$.

Auf Basis dieser Beobachtung kann man einen *Schnelltest*[2] für die Fairness einer Münze konstruieren.

[2] Das kann man auch anders besser machen. Deshalb liefert die obige Methode nur einen approximativen *Schnell*test.

10 Maßtheorie

10.1 Einführung zur Maßtheorie

10.1.1 Bemerkung: i.) DAS großartige Buch zur Maß- und Integrationstheorie ist von **Elstrodt** [8]. **Für dieses Kapitel besonders geeignet ist aber unbedingt Henze** [19, Kapitel 8]. Ich habe auch andere Bücher geprüft, aber meine Favoriten sind Elstrodt und Henze. Eine andere sehr gute Quelle ist **Küchler** [25]. Eine ebenfalls sehr gute Quelle ist das berühmte Buch von Billingsley [2].

ii.) Wir haben schon (in WT1) etwas Maßtheorie betrieben. Ein Wahrscheinlichkeitsmaß ist nämlich ein Maß mit der zusätzlichen Eigenschaft $\mathbb{P}(\Omega) = 1$.

Wir betreiben jetzt Maßtheorie allgemeiner. **Wir wollen systematisch (axiomatisch) erklären, wie man Mengen messen kann.** Es geht dabei nicht mehr nur um Ereignisse und deren Wahrscheinlichkeiten, sondern z.B. auch um

Anzahl, Länge, Fläche und Volumen von passenden Mengen.

Dabei gehen wir so vor. Wir definieren zunächst **möglichst einfach** sogenannte **Inhalte auf möglichst einfachen Strukturen (einfachen Teilmengensystemen)** (sogenannten **Halbringen**). Dann **setzen** wir so einen **Inhalt** zu einem **Maß** auf der erzeugten **σ-Algebra fort**. Wir geben schließlich Bedingungen an, so dass diese Fortsetzung eindeutig ist.

Für uns – mit Blick auf die Anwendung in der WT – am wichtigsten sind die folgenden beiden Maße:

- Das **Lebesgue-Maß** auf $(\mathbb{R}, \mathcal{B})$ bzw. auf $(\mathbb{R}^n, \mathcal{B}^n)$. Wikipedia über Henri Lebesque.
- **Lebesgue-Stieltjes-Maße** auch auf $(\mathbb{R}, \mathcal{B})$ bzw. auf $(\mathbb{R}^n, \mathcal{B}^n)$.
  Lebesgue-Stieltjes-Maße bei Wikipedia. Wikipedia über Thomas Stietjes

10.2 Mengensysteme

10.2.1 Definition: Es sei $\Omega \neq \emptyset$ eine nicht-leere Menge. Ein Teilmengensystem $\mathcal{H} \subset \mathcal{P}(\Omega)$ heißt **Halbring** (oder **Semi-Ring**) auf Ω, wenn gilt:

i.) $\emptyset \in \mathcal{H}$,

ii.) $A, B \in \mathcal{H} \Rightarrow A \cap B \in \mathcal{H}$,

iii.) $A, B \in \mathcal{H} \Rightarrow$ es gibt disjunkte $C_1, ..., C_n \in \mathcal{H}$ mit $A \setminus B = C_1 \uplus ... \uplus C_n$.

10.2.2 Bemerkung: Bei der Bedingung iii.) wird **nicht** gefordert, dass $A \setminus B \in \mathcal{H}$ gilt! Es wird (nur) gefordert, dass sich $A \setminus B$ als disjunkte Vereinigung von Mengen aus $\mathcal{H}$ schreiben lässt.

10.2.3 Beispiel: i.) Die **Loravalle**[1] $\{(a, b] \,|\, a, b \in \mathbb{R}, a \leq b\}$ bilden einen Halbring.

ii.) Die Menge der Kreise $\{B_\varepsilon(x) \,|\, x \in \mathbb{R}^2, \varepsilon \geq 0\}$ bilden **keinen Halbring**, denn der Schnitt von zwei Kreisen ist i.A. kein Kreis.

10.2.4 Definition: a.) Es sei Ω eine nicht-leere Menge. Ein Teilmengensystem $\mathcal{R} \subset \mathcal{P}(\Omega)$ heißt **Ring**, wenn gilt:

i.) $\emptyset \in \mathcal{R}$,
ii.) $A, B \in \mathcal{R} \Rightarrow A \cup B \in \mathcal{R}$,
iii.) $A, B \in \mathcal{R} \Rightarrow A \setminus B \in \mathcal{R}$.

b.) Ist $\mathcal{A}$ ein Ring mit $\Omega \in \mathcal{A}$, dann ist $\mathcal{A}$ eine **Algebra**.

[1] Ich glaube außer mir, verwendet niemand den Begriff Loravalle. Mal sehen ...

c.) Eine Algebra $\mathcal{A}$ bzw. ein Ring $\mathcal{R}$ heißt σ-**Algebra** respektive σ-**Ring**, falls die Abgeschlossenheit ii.) bezülich $\cup$ auch für abzählbare Vereinigungen gilt.

10.3 Inhalt, Prämaß, Maß

10.3.1 Definiton: Es sei $\Omega \neq \emptyset$ und $\mathcal{H} \subset \mathcal{P}(\Omega)$ ein Halbring.

i.) Eine Abbildung $\mu : \mathcal{H} \to [0, \infty]$ heißt **Inhalt**, falls $\mu(\emptyset) = 0$ und für alle $A_i \in \mathcal{H}, i = 1, ..., n$ mit

$$\bigcup_{i=1}^{n} A_i \in \mathcal{H}, A_i \cap A_j = \emptyset, i \neq j$$

gilt:

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i)$$

(Beachte: Bei der Eigenschaft der σ-Additivität wird $A_i \in \mathcal{H}$ und $\biguplus A_i \in \mathcal{H}$ vorausgesetzt).

ii.) Ein σ-additiver Inhalt auf $\mathcal{H}$ heißt **Prämaß** auf $\mathcal{H}$.

iii.) Ein **Maß** ist ein auf einer σ-Algebra definiertes Prämaß.

iv.) Wenn μ ein Maß auf einer σ-Algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$ ist, dann heißt das Triple $(\Omega, \mathcal{A}, \mu)$ ein **Maßraum**.

10.3.2 Bemerkung: Wenn $(\Omega, \mathcal{A}, \mu)$ ein Maßraum mit $\mu(\Omega) = 1$ ist, dann ist $(\Omega, \mathcal{A}, \mu)$ ein Wahrscheinlichkeitsraum

10.3.3 Bemerkung: Es sei Ω eine Menge. Dann ist $\mathcal{P}(\Omega)$ eine σ-Algebra. Warum suchen wir nicht nach einem Maß auf ganz $\mathcal{P}(\Omega)$? Für den Fall, dass Ω endlich oder abzählbar unendlich ist, haben wir (in WT1) regelmäßig $\mathcal{P}(\Omega)$ verwendet. Warum jetzt so kompliziert? Weil es im Allgemeinen nicht geht; jedenfalls nicht so wie wir es gerne hätten. Die Aussage wollen wir jetzt substantivieren. Drei berühmte Resultate sollen diese Unmöglichkeit vorführen. Wir betrachten $\Omega = \mathbb{R}^n$, da dieser Raum für viele praktische Probleme (Länge, Fläche, Volumen) angemessen ist. Ein berühmtes Buch über die Probleme beim mathematischen Messen ist Tomkowicz und Wagon [37]. Dort werden die unten genannten Aussagen auch mathematisch genau untersucht (und bewiesen).

> **Bemerkung:** Eine Abbildung $\beta : \mathbb{R}^n \to \mathbb{R}^n$ heißt **Bewegung** (**Isometrie**), falls $||x-y||_2 = ||\beta(x)-\beta(y)||_2$ für alle x, y gilt.
>
> Eine Bewegung erhält also alle Abstände. Wenn wir eine Menge $M \subset \mathbb{R}^n$ messen wollen, dann ist es (wohl) vernünftig, dass sich das Maß der Menge nicht ändert, wenn wir die Menge (nur) bewegen.
>
> **Satz (Hausdorff)**: Das Inhaltsproblem ist auf

$\mathbb{R}^n, n \geq 3$ unlösbar. D.h. Es gibt **keine** Abbildung $m : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ mit

- m ist endlich additiv
- m ist bewegungsinvariant, d.h. für alle Bewegungen β ist $m(\beta(A)) = m(A))$
- m ist normiert, d.h. $m([0,1]^n) = 1$

Über Felix Hausdorff bei Wikipeada

Satz (Vitali): Das Maßproblem ist unlösbar.[2]

Über Giuseppe Vitali bei Wikipeada

Satz (Banach, Tarski): Es sei $n \geq 1$ und $A, B \subset \mathbb{R}^n$ seien **beliebige** Mengen mit nicht-leerem Inneren. Dann gibt es abzählbar viele Mengen $C_k \subset \mathbb{R}^n$ und Bewegungen $\beta_k : \mathbb{R}^n \to \mathbb{R}^n$, so dass

$$A = \biguplus C_k$$
$$B = \biguplus \beta_k(C_k)$$

Über Stefan Banach bei Wikipeada
Über Alfred Tarski bei Wikipeada

[2] Für alle n.

10.4 Fortsetzungssätze

10.4.1 Definiton: Es sei $\mathcal{K} \subset \mathcal{P}(\Omega)$ ein Teilmengensystem. Dann heißt

$$\bigcap_{\substack{\mathcal{A} \text{ ist eine } \sigma-\text{Algebra} \\ \mathcal{A} \supset \mathcal{K}}} \mathcal{A} =: \sigma(\mathcal{K})$$

die von $\mathcal{K}$ **erzeugte** $\sigma-$**Algebra**. $\sigma(\mathcal{K})$ ist die kleinste σ-Algebra, die $\mathcal{K}$ enthält.

Wenn also $\tilde{\mathcal{A}}$ eine σ-Algebra mit $\mathcal{K} \subset \tilde{\mathcal{A}}$ ist, dann gilt $\sigma(\mathcal{K}) \subset \tilde{\mathcal{A}}$.

10.4.2 Bemerkung: i.) Das obige Konstruktionsprinzip funktioniert, denn der Schnitt beliebig vieler σ-Algebren ist eine σ-Algebra (siehe Elstrodt [8, Seite 17]). Für Halbringe gilt das nicht. Da wir mit Halbringen anfangen, ist das aber unproblematisch. Denn: Wir suchen nicht nach erzeugten Halbringen, sondern wir erzeugen mit Halbringen.

ii.) Es gibt eine Reihe anderer Mengensysteme, die je nach Fragestellung (und insbesondere in Beweisen) wichtig sind: Monotone Klassen, π-System, λ-System (Dynkin-Systeme). Auf diese Mengensystem sei hier nur aber nur kurz hingewiesen. Für Details vergleiche Elstrodt [8] und Küchler [25].

10.4.3 Bemerkung: Wir haben den folgenden Plan. Wir starten mit einem Inhalt ν auf einem Halbring $\mathcal{H}$. Das ist nämlich eine relativ übersichtliche Situation. Wir wollen ν zu einem Maß μ auf der von $\mathcal{H}$ erzeugten σ-Algebra **fortsetzen**.

10.4.4 Satz (Fortsetzungssatz von Caratheodory): Es sei $\mathcal{H}$ ein Halbring auf Ω. ν ein Inhalt auf $\mathcal{H}$.

Für $A \in \mathcal{P}(\Omega)$ definieren wir das äußere Maß von A

$$\mu(A) = \inf \left\{ \sum_{j=1}^{\infty} \nu(A_j) \mid A_j \in \mathcal{H},\, A \subset \bigcup_{j=1}^{\infty} A_j \right\}$$

Ist μ ein Prämaß, so ist μ ein Maß auf der von $\mathcal{H}$ erzeugten σ-Algebra $\mathcal{A} := \sigma(\mathcal{H})$ und es gilt $\mu_{|\mathcal{H}} = \nu$.

Beweis: Henze [19, Seite 313]

Über Caratheodory bei Wikipeada

10.4.5 Bemerkung: Die obige Definition von μ durch die summierten Inhalte der Überdeckungen nennt man die **Approximation von außen**.

10.4.6 Definition: Ein Inhalt ν heißt **σ-endlich** (σ-finit), falls es eine Folge von Mengen $E_j \in \mathcal{H}$ mit $\nu(E_j) < \infty$ und $\Omega = \bigcup E_j$.

10.4.7 Satz (Eindeutigkeit der Fortsetzung): Es sei $\mathcal{H}$ ein Halbring auf Ω und ν ein σ-endlicher Inhalt auf $\mathcal{H}$, dann ist die Fortsetzung gemäß des Fortsetzungssatz von Caratheodory eindeutig.

10.4.8 Bemerkung: Der Beweis der Eindeutigkeit bedient sich des Konzept der Dynkin-Systeme.

10.4.9 Bemerkung: Das **Programm nach Caratheodory**[3] geht also so:

1. Auf einem **Halbring** $\mathcal{H}$ definieren wir einen **Inhalt** ν.
2. Wir weisen nach, dass der Inhalt ν sogar σ-additiv ist; ν ist also ein **Prämaß**.
3. Die Sätze von Caratheodory beschreiben wie wir ein **Maß** auf $\sigma(\mathcal{H})$ erhalten.

Also:

Inhalt auf einem Halbring $\curvearrowright$ Prämaß $\curvearrowright$ Maß auf einer σ-Algebra

10.4.10 Satz: Es gilt

i.) Die Menge der **Loravalle** $\{(a, b] \,|\, a, b \in \mathbb{R}, a \leq b\}$ bilden einen Halbring.

[3] Auch diesen Begriff verwendet außer mir wohl keiner.

ii.) Die Abbildung

$$\nu : \mathcal{H} \to \mathbb{R}, \nu((a,b]) \mapsto b - a$$

ist ein Inhalt.

iii.) ν ist sogar ein Prämaß.

Beweis: Vgl. Billingsley [2, Seite 13ff]

10.4.11 Bemerkung: Der Beweis des vorhergehneden Satzes verwendet den Überdeckungssatz von Heine-Borel: Eine Teilmenge des $\mathbb{R}^n$ ist genau dann kompakt (also beschränkt und abgeschlossen), wenn jede offene Überdeckung eine endliche Teilüberdeckung enthält.

10.4.12 Definition: Die von den Loravallen erzeugte $\sigma-$Algebra heißt **Borel'sche $\sigma-$Algebra**; sie wird mit $\mathcal{B}(\mathbb{R})$ bezeichnet. Die Mengen in $\mathcal{B}$ heißen **Borelmengen**. Auf den Loravallen definieren wir den Inhalt ν wie im vorhergehenden Satz. Das dann wie im Satz von Caratheodory definierte Maß $\lambda : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ heißt **Borel-Lebesgue-Maß**.

10.5 Maße mit F definieren

▶ Für die flexible Definition des Erwartungswertes einer Zufallsvariable X benötigen wir sogenannte **Stieltjes-Intregrale** (bzw. eigentlich Lebesgue-Stieltjes-Intregrale; dazu später mehr). Die Idee kann man an der folgenden Approximation erfassen:

$$\begin{aligned}\mathbb{E}(X) &= \int X dF \\ &\approx x_1 \cdot (F(b_1) - F(a_1)) + x_2 \cdot (F(b_2) - F(a_2)) + \dots .\end{aligned}$$

Wir stellen uns vor, dass der Integrationsbereich in Intervalle unterteilt wird: $(a_1, b_1] \cup (a_2, b_2] \cup \dots$ und Zwischenwerte $x_1 \in (a_1, b_1], \dots$ entnommen werden. Um diese Idee nachzuvollziehen betrachten wir die entsprechende Approximation beim Riemann-Integral:

$$\int x dx = x_1 \cdot (b_1 - a_1) + x_2 \cdot (b_2 - a_1) + \dots .$$

Während die Zwischenwerte $x_i \in (a_i, b_i]$ beim Riemann-Integral mit den **Lebesgue-Inhalten** $b_i - a_i$ multipliziert werden, werden die Zwischenwerte beim Stieltjes-Integral mit den **Stieltjes-Inhalten** $F(b_i) - F(a_i)$ multipliziert. Wenn wir uns weiter vorstellen, dass F eine Verteilungsfunktion einer Zufallsvariable X ist, dann erhalten wir

$$F(b_i) - F(a_i) = \mathbb{P}^X((a_i, b_i]).$$

Der Stieltjes Inhalt erfasst also **die Wahrscheinlichkeit** für einen Wert im Intervall $(a_i, b_i]$ anstatt *nur* die Länge des Intervalls $(a_i, b_i]$. Wann wäre die Länge angemessen?

10.5.1 Satz: i.) Es sei $F : \mathbb{R} \to \mathbb{R}$ nicht-fallend. Dann ist $\nu(\,(a,b]\,) = F(b) - F(a)$ **ein endlicher Inhalt** auf den Lovallen $\mathcal{H}$.[4] Für zwei nicht-fallende Funktionen $F, G : \mathbb{R} \to \mathbb{R}$ gilt $\nu_F = \nu_G$ genau dann, wenn $F - G$ konstant ist.
ii.) Ist $\nu : \mathcal{H} \to \mathbb{R}$ ein endlicher Inhalt und $F : \mathbb{R} \to \mathbb{R}$ definiert durch

$$F(x) = \begin{cases} \nu(\,(0,x]\,) & \text{für } x \geq 0 \\ -\nu(\,(x,0]\,) & \text{für } x < 0 \end{cases} .$$

Dann ist F nicht-fallend und $\nu = \nu_F$.

Beweis: Siehe Elstrodt [8, Seite 39]

▶ Der Lebesgue-Inhalt λ ergibt sich für $F(x) = x$.

10.5.2 Definition: Für nicht-fallendes $F : \mathbb{R} \to \mathbb{R}$ heißt ν_F der zu F gehörende **Stieltjes Inhalt**.

10.5.3 Satz: Es sei $F : \mathbb{R} \to \mathbb{R}$ nicht-fallend und ν_F der zu F gehörende **Stieltjes-Inhalt**. ν_F ist genau dann ein **Prämaß**, wenn F von rechts stetig ist.

[4] Ein Inhalt heißt endlich, falls $\nu(H) < \infty$ für alle $H \in \mathcal{H}$.

Beweis: Siehe Elstrodt [8, Seite 39]

10.5.4 Bemerkung: Die beiden vorhergehenden Sätze charakterisieren alle endlichen Inhalte bzw. alle endlichen Prämaße auf den Loravallen. Ein endlicher Inhalt wird durch eine bis auf Konstante eindeutig bestimmte nicht fallende Funktion $F : \mathbb{R} \to \mathbb{R}$ charakterisiert. Ein endliches Prämaß wird durch eine bis auf Konstante eindeutig bestimmte nicht fallende von rechts stetig Funktion $F : \mathbb{R} \to \mathbb{R}$ charakterisiert.

10.5.5 Definition: Es sei F nicht-fallend und von rechts stetig und ν_F der zu F gehörende **Stieltjes Inhalt**. Das dann wie im Satz von Caratheodory definierte Maß $\lambda_F : \mathcal{B}(\mathbb{R}) \to [0, 1]$ heißt **Borel-Lebesgue-Stieltjes-Maß**.

10.5.6 Bemerkung: Wir haben ein bemerkenswertes Programm abgearbeitet. Es ist ganz offensichtlich, dass die Loravalle für die Anwendung sehr wichtig sind. **Wir wollen regelmäßig mindestens Loravalle messen.**

Wir haben auch gesehen, wie alle Inhalte beschrieben werden können, nämlich durch eine nicht-fallende Funktion $F : \mathbb{R} \to \mathbb{R}$. Wenn wir noch fordern, dass F von rechts stetig ist, dann erhalten so nach dem Programm nach Caratheodory passende Maße (**Fortsetzungen**) auf $\mathcal{B}$.

▶ Agenda: Wir wollen F ist einen stetigen und einen diskreten (Sprunganteil) zerlegen.

10.5.7 Definition: Eine Funktion $J : \mathbb{R} \to \mathbb{R}$ heißt **Sprungfunktion**, wenn es

- eine abzählbare Menge der Sprungstellen $\Delta \subset \mathbb{R}$ gibt und
- eine positive Funktion $p : \Delta \to (0, \infty)$ der Sprunghöhen mit $\sum_{y \in \Delta \cap [-n,n]} p(y) < \infty$ für alle $n \in \mathbb{N}$ und
- $\alpha \in \mathbb{R}$,

so dass

$$J(x) = \begin{cases} \alpha + \sum_{y \in \Delta \cap (0,x]} p(y) & \text{für } x \geq 0 \\ \alpha - \sum_{y \in \Delta \cap (x,0]} p(y) & \text{für } x < 0 \end{cases} .$$

10.5.8 Satz: Es sei $F : \mathbb{R} \to \mathbb{R}$ nicht-fallend und von rechts stetig. Dann ist die Menge Δ der **Unstetigkeitsstellen** von F abzählbar.

Beweis: Vgl. Elstrodt [8, S. 43]

10.5.9 Satz: i.) Es sei $F : \mathbb{R} \to \mathbb{R}$ nicht-fallend und von rechts stetig. Dann gibt es eine **nicht-fallende und von rechts stetige Sprungfunktion** J und eine **nicht-fallende**

und stetige Funktion $H : \mathbb{R} \to \mathbb{R}$, so dass $F = J + H$.
ii.) Es gilt $\nu_F = \nu_J + \nu_H$.
iii.) Bis auf additive Konstanten sind die Funktionen J, H eindeutig bestimmt.

Beweis: Elstrodt [8, Seite 43f]

10.5.10 Bemerkung: i.) Wir können gemäß des vorhergehenden Satzes also insbesondere Verteilungsfunktion in einen stetigen und einen Sprungteil zerlegen.

ii.) Man kann den stetigen Teil in einen absolut-stetigen (mit Dichte) und eine singulären Teil zerlegen.

10.6 Vollständigkeit

10.6.1 Definition: Es sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum. Dieser Maßraum heißt **vollständig**, wenn jede Teilmenge $A \subset N$ einer Nullmenge N (d.h. $\mu(N) = 0$) auch zu $\mathcal{A}$ gehört: $A \in \mathcal{A}$.

10.6.2 Satz: Es sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum. Es sei $\mathcal{N}$ das System aller Nullmengen. Wir betrachten das Mengensystem

$$\widetilde{\mathcal{A}} = \mathcal{A} \cup \mathcal{N}$$

und auf $\widetilde{\mathcal{A}}$, die durch

$$\widetilde{\mu}(\widetilde{A}) := \mu(A), \widetilde{A} = A \cup N, A \in \mathcal{A}, N \in \mathcal{N}$$

definierte Abbildung $\widetilde{\mu}$.

Dann ist $(\Omega, \widetilde{\mathcal{A}}, \widetilde{\mu})$ ein vollständiger Maßraum.

Beweis: Vgl. Elstrodt [8, Seite 71]

10.6.3 Definition: Das Maß $\widetilde{\mu}$ aus dem vorhergehenden Satz heißt **Vervollständigung** von μ.

10.6.4 ...

11 Messbare Abbildungen

Quellen: Henze [19], Elstrodt [8], Küchler [25].

11.1 Grundlegende Begriffe

Für die Integration (im nächsten Kapitel) benötigen wir unbedingt den Begriff der **Messbarkeit von Abbildungen**. Die Messbarkeit steht zudem in einem unmittelbaren Zusammenhang zur Information, die in $\mathcal{A}$ enthalten ist. Salopp formuliert: Eine $\mathcal{A}$-messbare Abbildung kann nicht mehr Information als $\mathcal{A}$ anzeigen.

11.1.1 Definition: Es sei $f : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ eine Abbildung zwischen messbaren Räumen, dann heißt f **messbar** falls $f^{-1}(A') \in \mathcal{A}$ für alle $A' \in \mathcal{A}'$ gilt [also $f^{-1}(\mathcal{A}') \subset \mathcal{A}$]. Also: Das Urbild jeder im Zielraum messbaren Menge ist im Startraum Raum messbar.

▶ Wir benötigen den Begriff *messbar* insbesondere, wenn wir auf der Zielmenge Ω' messen wollen, aber *zunächst* nur ein Maß auf dem Definitionsbereich Ω haben. Das ist bei Zufallsvariablen der Fall. Die sind auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ definiert und nehmen Werte in $\mathbb{R}$ an. Um das *Verhalten* der **Werte** $f(\omega)$ der Zufallsvariablen zu erfassen, benötigen wir das Bildmaß (die Verteilung). Um das Bildmaß zu definieren, benötigen wir die Messbarkeit.

▶ Wir erinnern uns $f^{-1}(A')$ ist die Menge aller $\omega \in \Omega$, die von f in die Menge A' abgebildet werden: $f(\omega) \in A'$.

11.1.2 Definition und Satz: Es sei $f : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ eine messbare Abbildung und μ ein Maß auf $(\Omega, \mathcal{A})$. Dann ist

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

ein Maß auf $(\Omega', \mathcal{A}')$.

ν heißt das **Bildmaß** von μ bezüglich f.

Wesentlich vereinfacht wird der Nachweis der Messbarkeit durch den folgenden Satz.

11.1.3 Satz: Es sei $f : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ eine Abbildung und $\mathcal{E}'$ ein Erzeuger von $\mathcal{A}'$; also $\sigma(\mathcal{E}') = \mathcal{A}'$. Dann gilt: f ist

genau dann messbar, wenn $f^{-1}(\mathcal{E}') \subset \mathcal{A}$.

▶ Wir müssen also nur $f^{-1}(E') \in$ für die $E' \in \mathcal{E}'$ untersuchen!

11.1.4 Definition: Es sei $f : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ eine Abbildung. Die kleinste σ-Algebra auf Ω, so dass f messbar ist, heißt die von f **erzeugte σ-Algebra**. Sie wird mit $\sigma(f)$ bezeichnet.

Kleinste σ-Algebra bedeutet: Wenn $\tilde{\mathcal{A}}$ eine σ-Algebra auf Ω ist, so dass f messbar ist, dann gilt $\sigma(f) \subset \tilde{\mathcal{A}}$.

11.1.5 Satz: Es gilt

$$\sigma(f) = \{f^{-1}(A') | A' \in \mathcal{A}'\}.$$

Beweis: Natürlich muss $\{f^{-1}(A') | A' \in \mathcal{A}'\} \subset \sigma(f)$ gelten, denn sonst wäre f nicht $\sigma(f)$-messbar. Außerdem bleiben alle Eigenschaften einer σ-Algebra bei der Bildung der Urbilder erhalten. Also ist $\{f^{-1}(A') | A' \in \mathcal{A}'\}$ eine σ-Algebra. Dann kann $\{f^{-1}(A') | A' \in \mathcal{A}'\}$ keine echte Teilmenge von $\sigma(f)$ sein. Sonst wäre $\{f^{-1}(A') | A' \in \mathcal{A}'\}$ eine kleinere σ-Algebra bezüglich der f messbar ist.

11.1.6 Beispiel: Wir betrachten $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A} = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$ und $\Omega' = \mathbb{R}$, $\mathcal{A}' = \mathcal{B}$. Wir betrach-

ten die Abbildung $f : \Omega \to \mathbb{R}$ mit

$$f(\omega) = \begin{cases} 8, \text{ falls } \omega \in \{1,2,3\} \\ 11, \text{ falls } \omega \in \{4,5,6\}. \end{cases}$$

a.) f ist nicht messbar.
b.) Es ist $\sigma(f) = \{\emptyset, \{1,2,3\}, \{4,5,6\}, \Omega\}$.

Wir beobachten auch, dass $f(1) = 8 \neq 11 = f(5)$ ist. Die Abbildung f ist auf der Menge $\{1,3,5\}$ nicht konstant.

▶ $\Omega = \{1,2,3,4,5,6\}, \mathcal{A} = \{\emptyset, \Omega, \{1,3,5\}, \{2,4,6\}\}$ erfasst das Experiment des abgeklebten Würfels. f zeigt insbesondere an, ob ein Ergebnis aus $\{1,2,3\}$ gewürfelt wurde. Das *funktioniert* jedoch nicht, wenn die Information eigentlich $\mathcal{A} = \{\emptyset, \Omega, \{1,3,5\}, \{2,4,6\}\}$ ist. Wir können 1 und 5 eigentlich – also bei der Information $\mathcal{A}$ – nicht unterscheiden. Die Abbildung zeigt aber unterschiedliche Werte für $\omega = 1$ und $\omega = 5$ an. Mit der Beobachtung von f kann man $\omega = 1$ versus $\omega = 5$ doch entschieden. Messbarkeit von f würde bedeutet, dass so was **nicht passiert**: An den Werten von f kann man keine über $\mathcal{A}$ hinausgehende Information ablesen.

▶ Lösung im Unterricht!

11.1.7 Satz: i.) Es sei c eine reelle Zahl. Die Abbildung $f : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}), \omega \mapsto c$ ist messbar.

ii.) Die Indikatorabbildung $\mathbb{1}_A : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ für eine Menge $A \subset \Omega$ ist genau dann messbar, wenn $A \in \mathcal{A}$. Die Messbarkeit der Indikatorfunktion entspricht also der Messbarkeit der Menge, die angezeigt werden soll.
iii.) Es sei $A \subset \Omega$. Die kleinste σ-Algebra, so dass $\mathbb{1}_A : \Omega \to (\mathbb{R}, \mathcal{B})$ messbar ist, ist $\{\emptyset, \Omega, A, A^c\}$.

Beweis: i.) Es sei $B \in \mathcal{B}$. Wenn $c \in B$ gilt, dann $f(\omega) = c \in B$ für alle $\omega \in \Omega$. Also $f^{-1}(B) = \Omega$. Wenn $c \notin B$ gilt, dann $f(\omega) = c \notin B$ für alle $\omega \in \Omega$. Also $f^{-1}(B) = \emptyset$. Natürlich gilt $\emptyset, \Omega \in \mathcal{A}$ (das gilt für jede σ-Algebra). Also ist f messbar.

ii.) "$\Rightarrow$": Die Abbildung $\mathbb{1}_A$ sei messbar. Das Loravall $(0.5, 1.5]$ ist natürlich eine Borelmenge (denn Borelmengen werden mit den Loravallen erzeugt). Wegen der Messbarkeit von $\mathbb{1}_A$, muss $A = \mathbb{1}_A^{-1}((0.5, 1.5]) \in \mathcal{A}$ gelten.

"$\Leftarrow$": Es sei $A \in \mathcal{A}$; dann ist auch $A^c \in \mathcal{A}$. Natürlich sind auch $\emptyset, \Omega \in \mathcal{A}$. Wir halten schon mal fest, dass unter der Voraussetzung $A \in \mathcal{A}$ die vier Mengen $\emptyset, \Omega, A, A^c \in \mathcal{A}$ in $\mathcal{A}$ liegen.

Es sei $B \in \mathcal{B}$. Wenn $0, 1 \notin B$, dann $\mathbb{1}_A^{-1}(B) = \emptyset$. Wenn $0 \notin B$ und $1 \in B$, dann $\mathbb{1}_A^{-1}(B) = A$. Wenn $0 \in B$ und $1 \notin B$, dann $\mathbb{1}_A^{-1}(B) = A^c$. Wenn $0, 1 \in B$, dann $\mathbb{1}_A^{-1}(B) = \Omega$. Also ist $\mathbb{1}_A^{-1}(B) \in \mathcal{A}$ für alle $B \in \mathcal{B}$.

iii.) Es sei $A \subset \Omega$. $\{\emptyset, \Omega, A, A^c\}$ ist eine σ-Algebra; in der Tat ist $\{\emptyset, \Omega, A, A^c\}$ die kleinste σ-Algebra, die A enthält. Wir haben im Beweis von ii.) gesehen, dass es genau vier Möglichkeiten für $\mathbb{1}_A^{-1}(B), B \in \mathcal{B}$ gibt und zwar genau die Mengen $\{\emptyset, \Omega, A, A^c\}$. Also ist $\mathbb{1}_A^{-1}$ messbar bezüglich $\{\emptyset, \Omega, A, A^c\}$ und eine kleinere σ-Algebra kann es nicht geben.

11.1.8 Satz: Es sei K endlich und $\mathcal{Z} = \{Z_j | j \in K\}$ eine Zerlegung von $\Omega \neq \emptyset$. Dann gilt

$$\sigma(\mathcal{Z}) = \left\{ \bigcup_{k \in J} Z_k \,|\, J \subset K \right\}.$$

Es folgt insbesondere, dass Mengen $\emptyset \neq A \subsetneq Z_i$ nicht in $\sigma(\mathcal{Z})$ liegen können; $A \notin \sigma(\mathcal{Z})$.

▶ Feinere Information als Z_i gibt es in $\sigma(\mathcal{Z})$ nicht.

Beweis: Wir beobachten zunächst, dass

$$\mathcal{A} = \left\{ \bigcup_{k \in J, J \subset K} Z_k \,|\, J \subset K \right\}$$

eine σ-Algebra ist. Dann beobachten wir: Wenn $\mathcal{A}'$ eine σ-Algebra ist, die $\mathcal{Z}$ enthält. Dann sind auch alle Menge aus $\mathcal{A}$ in $\mathcal{A}'$; die Mengen aus $\mathcal{A}$ ergeben sich als Vereinigung aus Mengen von $\mathcal{Z}$ und $\mathcal{A}'$ ist vereinigungsstabil. Also ist – wie behauptet – $\mathcal{A}$ die kleinste σ-Algebra, die $\mathcal{Z}$ enthält.

11.1.9 Satz: Wenn f eine Abbildung mit endlichem Bild$(f) = \{c_1, ..., c_n\}$ ist (o.B.d.A. c_i verschieden), dann definieren die Mengen $Z_i = f^{-1}(\{c_i\})$ eine Zerlegung $\mathcal{Z} = \{Z_1, ..., Z_n\}$ und es gilt

$$\sigma(f) = \sigma(\mathcal{Z}) = \left\{ \bigcup_{k \in J} Z_k \,|\, J \subset K \right\}.$$

Beweis: Die Mengen $Z_i = f^{-1}(\{c_i\})$ bilden (natürlich) eine Zerlegung von Ω, denn Bild$(f) = \{c_1, ..., c_n\}$.

Wir müssen nachweisen, dass alle Mengen $f^{-1}(B), B \in \mathcal{B}$ in $\left\{\bigcup_{k \in J} Z_k \,|\, J \subset K\right\}$ vorkommen und alle Mengen in $\left\{\bigcup_{k \in J} Z_k \,|\, J \subset K\right\}$ auch in $\{f^{-1}(B) | B \in \mathcal{B}\}$ sind.

Es sei B eine Borelmeng. Zunächst: Wenn $f^{-1}(B) = \emptyset$, dann ist $c_k \notin B$ alle k. Also ist auch $\bigcup_{j \text{ mit } c_j \in B} Z_j = \emptyset$. Umgekehrt, ist $\bigcup_{j \text{ mit } c_j \in B} Z_j = \emptyset$, dann ist $c_k \notin B$ alle k. Dann gilt auch $f^{-1}(B) = \emptyset$.

Es sei B eine Borelmenge mit $f^{-1}(B) \neq \emptyset$. Wir wollen

$$f^{-1}(B) = \bigcup_{j \text{ mit } c_j \in B} Z_j$$

zeigen.

Es sei $\omega \in f^{-1}(B)$. Es sei $B \ni f(\omega) = c_l$ für ein geeignetes l. Dann ist $\omega \in Z_l$ und folglich auch $\omega \in \bigcup_{j \text{ mit } c_j \in B} Z_j$.

Ist $\omega \in \bigcup_{j \text{ mit } c_j \in B} Z_j$, dann gibt es ein l mit $\omega \in Z_l$ und $c_l \in B$. Also $\omega \in f^{-1}(B)$.

Wir haben nachgewiesen, dass

$$f^{-1}(B) = \bigcup_{j \text{ mit } c_j \in B} Z_j$$

gilt.

Auf der linken Seite stehen alle Mengen aus $\sigma(f)$ und auf der rechten alle Mengen aus $\sigma(\mathcal{Z})$.

11.1.10 Satz: Es sei $\mathcal{Z} = \{Z_1, ..., Z_n\}$ eine Zerlegung von Ω und $f : \Omega \to \mathbb{R}$ eine Abbildung. f ist genau dann $\sigma(\mathcal{Z})$-messbar, wenn f auf den Mengen Z_k der Zerlegung konstant ist.

An den Werten von $f(\omega) = c_j$ kann man nicht ablesen, welches $\omega \in Z_j$ eingetreten ist; nur das irgendein $\omega \in Z_j$ eingetreten ist.

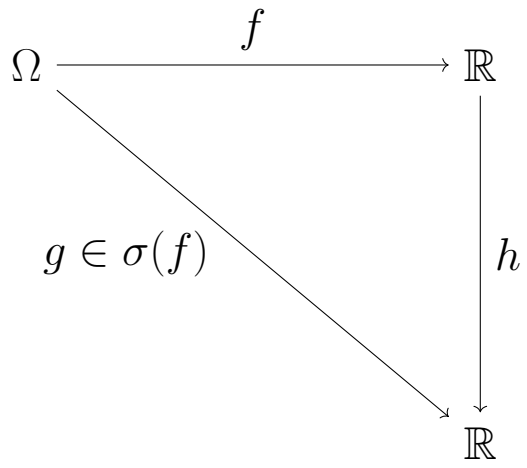
Beweis: Angenommen f ist auf Z_1 nicht konstant, d.h. es gibt $c_1, c_2, c_1 \neq c_2$ und $\omega_1, \omega_2 \in Z_1$ mit $f(\omega_1) = c_1 \neq c_2 = f(\omega_2)$. Wäre f $\sigma(\mathcal{Z})$-messbar, dann müsste $f^{-1}(\{c_1\}) \in \sigma(\mathcal{Z})$ gelten (denn $\{c_1\}$ ist eine Borelmenge). Also auch $A = f^{-1}(\{c_1\}) \cap Z_1 \in \sigma(\mathcal{Z})$; denn $\sigma(\mathcal{Z})$ ist eine σ-Algebra. Dann ist $\emptyset \neq A \subsetneq Z_1$; die strikte $\subsetneq$ Relation gilt wegen $\omega_2 \in Z_1$.

Ein solches A kann es in $\sigma(\mathcal{Z})$ nicht geben (das haben wir vorne beobachtet). Also kann f nicht $\sigma(\mathcal{Z})$-messbar sein.

Angenommen f ist auf den Mengen Z_j konstant. Für jede Borelmenge B gilt dann (das zeigt man wie im vorhergehenden Beweis)

$$f^{-1}(B) = \bigcup_{j \text{ mit } c_j \in B} Z_j.$$

Also $f^{-1}(B) \in \sigma(\mathcal{Z})$.



$\lightning\ g \in \sigma(f) \Rightarrow \exists h : \mathbb{R} \to \mathbb{R}$ mit $g(\omega) = h(f(\omega))$

Abbildung 11.1: Schema zum Faktorisierungslemma

11.1.11 Satz (Faktorisierungslemma): Es sei $f : \Omega \to \mathbb{R}$ eine Abbildung und $\sigma(f) = f^{-1}(\mathcal{B})$, die von f auf Ω erzeugte σ-Algebra. Ferner sei $g : \Omega \to \mathbb{R}$ eine $\sigma(f)$-messbare Abbildung. Dann gibt es eine messbare Abbildung $h : \mathbb{R} \to \mathbb{R}$ mit $g(\omega) = h(f(\omega))$.

▶ Anschaulich: $\sigma(f)$ erfasst die Information (denn es ist die kleinste σ-Algebra, so dass f messbar ist), die zu f gehört. Wenn auch g zu dieser Information passt, dann ist g *nur* eine messbare Transformation von f. g kann keine weitere Information *anzeigen*.

▶ Wenn wir den Wert von g an der Stelle ω bestimmen wollen (was wir je nach Informationslage nicht können), dann bestimmen wir den Wert von f an der Stelle ω und dann von $f(\omega)$ die Transformation $h(f(\omega))$

Beweis: Vgl. Küchler [25, Seite 115] und Übungsaufgabe .... oder Shao [32, Seite 37] oder Henze [19, Seite 175].

Beweis: Wir betrachten zunächst eine elementare Funktion g, d.h.

$$g = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}.$$

Gemäß Voraussetzung ist g messbar bezüglich der von f erzeugten σ-Algebra $\sigma(f) = f^{-1}(\mathcal{B})$. Dann muss $A_i \in \sigma(f)$ gelten. Da $\sigma(f) = f^{-1}(\mathcal{B})$ muss es Borelmengen $B_i \in \mathcal{B}$ geben, so dass $A_i = f^{-1}(B_i)$ gilt.

Dann definieren wir

$$h = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{B_i}$$

und rechnen nach, dass

$$g = h \circ f$$

also

$$g(\omega) = h(f(\omega)) \text{ für alle } \omega \in \Omega$$

gilt.

.... weiter in der Mitschrift

11.1.12 Beispiel: Es sei $\Omega = \{1, 2, 3, 4, 5, 6\}$ und $f(1) = f(3) = f(5) = 8$ und $f(2) = f(4) = f(6) = 11$. Dann ist $\sigma(f) = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$.

Ferner sei $g(1) = 1$ und $g(k) = 0, k = 2, ..., 6$. Die Abbildung g ist auf $\{1, 3, 5\}$ nicht konstant und deshalb auch nicht $\sigma(f)$-messbar. Gemäß des vorherigen Satzes kann es ein h nicht geben. Angenommen doch. Angenommen es gibt ein h mit

$$g(\omega) = h(f(\omega)) \text{ für alle } \omega \in \Omega.$$

Dann

$$1 = g(1) = h(f(1)) = h(8)$$
$$0 = g(3) = h(f(3)) = h(8)$$

Wir erhalten den Widerspruch $0 = 1$. Es kann so ein h nicht geben.

11.1.13 Beispiel: Es sei $\Omega = \{1, 2, 3, 4, 5, 6\}$, $f(1) = 17$ und $f(3) = f(5) = 8$ und $f(2) = f(4) = f(6) = 11$. Dann ist $\sigma(f) = \{\emptyset, \{1\}, \{3, 5\}, \{2, 4, 6\}, ..., \Omega\}$.

Ferner sei $g(1) = 1$ und $g(k) = 0, k = 2, ..., 6$.

11.2 Treppenfunktionen

11.2.1 Defintion: Eine messbare Funktion $f : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ mit endlicher Bildmenge heißt **Treppenfunktion**. Beachte, dass die Messbarkeit Teil der Definition ist!

11.2.2 Satz: Wenn $f : \Omega \to \mathbb{R}$ eine Treppenfunktion ist, dann gibt es reelle Zahlen $\alpha_1, ..., \alpha_m$ und $A_1, ..., A_m \in \mathcal{A}$ mit

$$f = \sum_{j=1}^{m} \alpha_j \mathbb{1}_{A_j}.$$

Hinweis: Die Messbarkeit von f und $A_1, ..., A_m \in \mathcal{A}$ sind Teil der Definition der Treppenfunktion. Eine Abbildung, die exakt so *aussieht* wie eine Treppe, deren $A_1, ..., A_m$ nicht alle in $\mathcal{A}$ liegen, nennen wir nicht Treppenfunktion.

11.2.3 Definition: f ist eine Treppenfunktion in **kanonischer Form**, wenn die α_j paarweise verschieden sind.

11.2.4 Satz: Man kann eine Treppenfunktion immer in kanonischer Form darstellen.

11.3 $\bar{\mathbb{R}}$ und numerische Funktionen

▶ Für die Entwicklung des Integrationsbegriff ist es vorteilhaft, auch ∞ und $-\infty$ als Funktionswerte zu zulassen. Dafür benötigen wir eine entsprechende **Erweiterung** der reellen Zahlen/Achse sowie der Borelmengen.

11.3.1 Definition: Wir definieren den Maßraum $(\bar{\mathbb{R}}, \bar{\mathcal{B}})$

$$\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$$
$$\bar{\mathcal{B}} = \{B \cup E | B \in \mathcal{B}, E \subset \{-\infty, \infty\}\}.$$

Eine Funktion $f : (\Omega, \mathcal{A}) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ heißt **numerische Funktion**.

11.3.2 Bemerkung: Es sei (a_i) eine nicht fallende Folge in $\mathbb{R}$. Wenn die Folge beschränkt ist, dann konvergiert so eine Folge. Wenn sie unbeschränkt ist dann *konvergiert* sie **nur auf der erweiterten Achse** und zwar gegen ∞.

▶ Bezogen auf die Messarbeit wird es immerhin nicht komplizierter:

11.3.3 Satz: Eine numerische Funktion $f : (\Omega, \mathcal{A}) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ ist genau dann messbar, wenn $f^{-1}(B) \in \mathcal{A}$ für alle $B \in \mathcal{B}$.

Beweis: Vgl. Küchler [25, Seite 92]

Der folgende wichtige Satz deutet an, wie wir später (insbesondere bei der Entwicklung des Integrationsbegriffs) vorgehen wollen: Wir betrachten zunächst Treppenfunktionen und bilden dann anschließend Grenzübergang. Ob dabei *Eigenschaften* erhalten bleiben, **müssen** wir natürlich überprüfen bzw. durch passende Voraussetzungen sicherstellen.

11.3.4 Satz (Approximationssatz): Eine nicht-negative numerische Funktion $f : (\Omega, \mathcal{A}) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ ist genau dann messbar, wenn es eine Folge (u_n) nicht negativer Treppenfunktionen mit $(u_n) \nearrow f$ gibt.

Beweis: vgl. Küchler [25, Seite 100] oder Tappe [36, Seite 104]

12 Integration

12.1 Integral nicht-negativer Treppenfunktionen

12.1.1 Definition: Für eine nicht-negative Treppenfunktion

$$f = \sum_{j=1}^{m} \alpha_j \mathbb{1}_{A_j}$$

auf einem Maßraum $(\Omega, \mathcal{A}, \mu)$ definieren wir das **Integral von** f bezüglich μ (über Ω):[1]

$$\int f \, d\mu = \int f(x) \, d\mu(x) = \sum_{j=1}^{m} \alpha_j \mu(A_j)$$

[1] Das Integralzeichen wurde gemäß Elstrodt [8, S. 135] von Leibniz eingeführt und stellt ein stilisiertes S (das für Summe steht) dar. In Elstrodt [8] findet sich dort auch der Hinweis, dass der Begriff Integral von Johann Bernoulli geprägt wurde und erstmals in einer Arbeit von Jacob Bernoulli erschien.

12.1.2 Behauptung: Für alle $A \in \mathcal{A}$ gilt

$$\int \mathbb{1}_A d\mu = \mu(A).$$

Wenn $\mu = \mathbb{P}$ ein Wahrscheinlichkeitsmaß ist, dann hat die Behauptung die Form

$$\mathbb{E}(\mathbb{1}_A) = \int \mathbb{1}_A d\mathbb{P} = \mathbb{P}(A).$$

12.2 Maß-Integral

12.2.1 Definition: i.) Für eine messbare nicht-negative numerische Funktion $f : (\Omega, \mathcal{A}, \mu) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ definieren wir das **Integral bezüglich** μ durch

$$\int f d\mu = \sup \left\{ \int u \, d\mu \,|\, \text{nicht-negative Treppenfunktion } u \leq f \right\}.$$

ii.) Es sei f eine messbare numerische Funktion und $f^+ = \max\{f, 0\}$, $f^- = \max\{-f, 0\}$. Gilt

$$\int f^+ d\mu < \infty, \int f^- d\mu < \infty,$$

dann heißt f **integrierbar** und wir schreiben $f \in \mathcal{L}^1(\mu)$. Wir definieren dann das **Integral bezüglich** μ

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu.$$

12.2.2 Hilfssatz: Es sei $f : (\Omega, \mathcal{A}, \mu) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ eine messbare nicht-negative numerische Funktion. Dann gilt für jede Folge nicht-negativer **Treppenfunktionen** $(g_i)_{i\in\mathbb{N}}$ mit $g_i \nearrow f$

$$\int g_i \, d\mu \nearrow \int f \, d\mu.$$

Mit anderen Worten:

$$\lim \int g_i \, d\mu = \int f \, d\mu = \int \lim g_i \, d\mu.$$

12.2.3 Bemerkung: i.) Wenn wir ein Integral durch Grenzübergang berechnen/definieren wollen, dann können wir **irgendeine Folge** von Treppenfunktionen mit $g_i \nearrow f$ wählen. ii.) Unter bestimmten Voraussetzungen können wir also $\int$ und lim vertauschen. Vgl. dazu auch den folgenden Satz sowie später den Satz über die majorisierte Konvergenz.

12.2.4 Satz von der monotonen Konvergenz: Es sei $(f_n)_{n\in\mathbb{N}}$ eine monoton wachsende Folge von messbaren **nichtnegativen** numerischen Funkionen. Dann gilt:

$$\int (\lim f_n) \, d\mu = \lim \int f_n d\mu.$$

Beweis: Vgl. Elstrodt [8, Seite 139]

12.2.5 Satz von der majorisierten Konvergenz: Es sei

$f_n : \Omega \to \bar{\mathbb{R}}$ eine Folge von messbaren Abbildungen, die fast überall konvergiert. Ferner sei $g \in \mathcal{L}^1$ eine nicht-negative Abbildung mit $|f_n| \leq g$ fast überall. Dann gilt:

i.) Für alle $n \in \mathbb{N}$ ist $f_n \in \mathcal{L}^1$ und es gibt $f \in \mathcal{L}^1$ mit $f_n \to f$ fast überall.

ii.) Für jedes $f \in \mathcal{L}^1$ mit $f_n \to f$ fast überall gilt

$$\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu = \int (\lim_{n\to\infty} f_n) \, d\mu$$

und

$$\lim_{n\to\infty} \int |f_n - f| \, d\mu = 0.$$

Beweis: Siehe für i.) Bauer [1, 95f].

12.2.6 Satz: Es seien $f, g \in \mathcal{L}^1$ und $\alpha, \beta \in \mathbb{R}$. Dann ist auch $\alpha f + \beta g \in \mathcal{L}^1$ und es gilt die **Linearität der Integration:**

$$\int (\alpha f + \beta g) \, d\mu = \alpha \int f \, d\mu + \beta \int g \, d\mu.$$

12.2.7 Satz (Monotonie der Integration): Für $f, g \in \mathcal{L}^1$ mit $f \leq g$ gilt

$$\int f \, d\mu \leq \int g \, d\mu.$$

12.2.8 Satz (Dreiecksungleichung für die Integrati-

on): Für $f \in \mathcal{L}^1$ gilt

$$\left| \int f d\mu \right| \leq \int |f| d\mu.$$

12.2.9 Satz: Es sei $f : \Omega \to \bar{\mathbb{R}}$ eine messbare Funktion. Dann sind die folgenden Aussagen äquivalent:

i.) $f \in \mathcal{L}^1$ (d.h. f ist integrierbar).

ii.) Es gibt eine nicht-negative integrierbare Funktion $g \in \mathcal{L}^1$ mit $|f| \leq g$. (so ein g heißt **Majorante**)

iii.) $f^+, f^- \in \mathcal{L}^1$.

iv.) Es gibt nicht-negative integrierbare Funktionen $g, h \in \mathcal{L}^1$ mit $f = g - h$.

v.) $|f| \in \mathcal{L}^1$.

12.2.10 Folgerung: Wenn $\mu(\Omega) < \infty$, dann ist jede beschränkte messbare Abbildung integrierbar.

Beweis: Eine integrierbare Majorante ist $g = M$.

12.2.11 Bemerkung: Die Funktion

$$f(x) = \begin{cases} 1 \text{ falls } x \in \mathbb{Q} \\ 0 \text{ falls } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

heißt **Dirichlet Funktion** und ist berühmt (berüchtigt). Für die Einschränkung $g = f_{|[0,1]}$ von f auf das Intervall $[0,1]$ gilt $f \in \mathcal{L}^1([0,1], \mathcal{B}_{[0,1]}, \lambda_{[0,1]})$. Diese Funktion ist jedoch nicht Riemann-integrierbar.

12.2.12 Bemerkung: $f \in \mathcal{L}^1$ impliziert, dass $\mu(\{x : |f| = \infty\}) = 0$.

12.3 Parameterintegrale

12.3.1 Satz: Es sei $f : (a,b) \times \Omega \to \mathbb{R}$ eine Abbildung mit:

i.) Für alle *Parameter* $t \in (a,b)$ ist $f(t,\cdot) \in \mathcal{L}^1$

ii.) Für fast alle $x \in \Omega$ ist $f(\cdot, x) : (a,b) \to \mathbb{R}$ stetig in $t_0 \in (a,b)$.

iii.) Es gibt ein Intervall $(c,d) \ni t_0$ und eine nicht-negative Majorante $g \in \mathcal{L}^1$, so dass für alle $t \in (c,d)$ fast überall $|f(t,\cdot)| \leq g$ gilt.

Dann ist $F : (a,b) \to \mathbb{R}$ mit

$$F(t) = \int f(t,x)\, d\mu(x), t \in (a,b)$$

in t_0 stetig. Also

$$\lim_{t\to t_0} F(t) = \lim_{t\to t_0} \int f(t,x)\,d\mu(x) \;= \int \lim_{t\to t_0} f(t,x)\,d\mu(x)$$
$$= \int f(t_0,x)\,d\mu(x)$$

Beweis: Vgl. Bauer [1, Seite 101f] oder Henze [19, Seite 337]

12.3.2 Satz: Es sei $f : (a,b)\times\Omega \to \mathbb{R}$ eine Abbildung mit:

i.) Für alle $t \in (a,b)$ ist $f(t,\cdot) \in \mathcal{L}^1$

ii.) Die partielle Ableitung

$$\left(\frac{\partial}{\partial t} f(t,x)\right)_{|t_0}$$

existiert für alle $x \in \Omega$.

iii.) Es gibt ein Intervall $(c,d) \ni t_0$ und eine nicht-negative Majorante $g \in \mathcal{L}^1$, so dass für alle $t \in (c,d)\cap(a,b), t \neq t_0$

$$\left|\frac{f(t,x) - f(t_0,x)}{t - t_0}\right| \leq g(x) \text{ f.ü..}$$

Dann ist $F : (a,b) \to \mathbb{R}$

$$F(t) = \int f(t,x)\,d\mu(x)$$

in t_0 differenzierbar und $\frac{\partial}{\partial t} f(t,x)_{|t_0} \in \mathcal{L}^1$ integrierbar und es

gilt

$$F'(t_0) = \frac{d}{dt}F(t)_{|t_0} = \frac{d}{dt}\int f(t_0, x)\, d\mu(x) = \int \frac{\partial}{\partial t} f(t_0, x)\, d\mu(x).$$

Beweis: Vgl. Bauer [1, Seite 102] oder Henze [19, Seite 337]

12.4 Riemann-Integral

12.4.1 Satz: i.) Eine beschränkte Funktion $f : [a, b] \to \mathbb{R}$ ist genau dann **Riemann-integrierbar**, wenn die Menge der Unstetigkeitstellen eine Lebesgue-Nullmenge bilden.

Eine Menge $N \subset \mathbb{R}$ heißt **Lebesgue-Nullmenge**, falls $\lambda(N) = 0$.

ii.) Wenn eine beschränkte Funktion $f : [a, b] \to \mathbb{R}$ Riemann-integrierbar ist, dann stimmen Riemann-Integral und Lebesgue-Integral überein.

Beweis: Vgl. Elstrodt [8, S. 166]

12.4.2 Satz: Es sei $I \subset \mathbb{R}$ ein Intervall und $f : I \to \mathbb{R}$ auf jedem kompakten Intervall in I Riemann-integrierbar. Dann gilt:
i.) f ist genau dann Lebesgue-integrierbar, wenn $|f|$ uneigentlich Riemann-integrierbar über I ist.

ii.) Wenn $|f|$ uneigentlich Riemann-integrierbar über I ist, dann stimmen das Lebesgue-Integral und das uneigentliche Riemann-Integral überein.

Beweis: Vgl. Elstrodt [8, S. 168]

Wir können trotz dieser scheinbaren Überlegenheit nicht auf das Riemann-Integral verzichten.

12.4.3 Bemerkung: Das uneigentliche Riemann-Integral

$$R\text{-}\int_0^\infty \frac{\sin(x)}{x} dx$$

existiert (Elstrodt [8, S. 168f]). Aber $|\frac{\sin(x)}{x}|$ ist nicht über $(0, \infty)$ uneigentlich Riemann-integrierbar. Also ist $\frac{\sin(x)}{x}$ auch nicht über $(0, \infty)$ Lebesgue-integrierbar. Ausgerechnet das Integral über $\frac{\sin(x)}{x}$ benötigt man für die Inversionsformel von Levy für charakteristische Funktionen

12.4.4 Bemerkung: Das vorliegende Kapitel liefert nur einen CRASH-Kurs der Integrationstheorie. Details finden Sie in der **Maß- und Integrationsbibel** Elstrodt [8]. Für unsere Zwecke sind insbesondere noch die Bücher von Küchler [25] und Henze [19] sehr empfehlenswert.

13 Bedingte Erwartung – bedingte Wahrscheinlichkeit

Das Thema dieses Kapitels ist relativ schwer. Meine favorisierten Quellen sind die jeweiligen Kapitel aus Billingsley [2], Küchler [25], Williams [38], Williams [39] und die immer gute Quelle **Henze** [19].

13.1 Bedingte Erwartungen gegeben B

Dieser Abschnitt ist eine *elementare* Vorbereitung und weder ausreichend kommentiert noch genau (d.h. teilweise fehlen die exakten Voraussetzungen für Formel und Aussagen).

13.1.1 Definition: Es sei X eine integrierbare Zufallsvaria-

ble und eine B eine Ereignis mit $\mathbb{P}(B) > 0$. Wir definieren

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(X \cdot \mathbb{1}_B)}{\mathbb{P}(B)}$$

Vgl. die Hinweise in Williams [39, Seite 385]

13.1.2 Bemerkungen: Zwei Fälle sind für die Anwendung besonders wichtig. Wir geben die Ergebnisse ohne Beweis an.

Y diskret

$$\mathbb{E}(Y|B) = \sum_y y\mathbb{P}(y|B)$$

Y mit bedingter Dichte

$$\mathbb{E}(Y|B) = \int yf(y|B)dy$$

Vgl. die Hinweise in Blitzstein [4, Seite 416]

13.1.3 Bemerkungen: Auch für die bedingten Erwartungen gibt es eine Fallunterscheidungsformel

$$\mathbb{E}(Y) = \sum_{i=1}^{n} \mathbb{E}(Y|B_i)\mathbb{P}(B_i)$$

Vgl. die Hinweise in Blitzstein [4, Seite 416]

13.1.4 Bemerkungen: Varianzzerlegung ... Williams [39,

Seite 392]

$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))$$

13.2 Radon-Nikodym

Zunächst führen wir den für den späteren Gebrauch (insbesondere in Beweisen) wichtigen Satz von Radon und Nikodym ein.

13.2.1 Definition: Es seien μ, ν zwei σ-finite Maße auf dem Maßraum $(\Omega, \mathcal{F})$. μ heißt **absolut stetig** bezüglich ν, falls für alle $B \in \mathcal{F}$ mit $\nu(B) = 0$ auch $\mu(B) = 0$ gilt. In diesem Fall schreib en wir $\mu \ll \nu$.

Gilt $\mu \ll \nu$ und $\nu \ll \mu$, dann heißen μ und ν **äquivalent**.

Bevor Sie die folgende Proposition lesen, sollte Sie sich die $\epsilon - \delta$ der Stetigkeit aus der Analysis anschauen: f ist gemäß Definition stetig in x_0, falls es für alle $\varepsilon > 0$ ein $\delta > 0$, gibt, so dass aus $x - x_0 < \delta$ die Ungleichung $f(x) - f(x_0) < \varepsilon$ folgt.

Mit dieser Definition im Sinn ist die folgenden Begriffsbildung angemessen.

13.2.2 Proposition: Es seien μ, ν zwei **endliche** Maße[1] auf dem Maßraum $(\Omega, \mathcal{F})$. Dann ist μ genau dann absolut stetig bezüglich ν, falls es für alle $\epsilon > 0$ ein $\delta > 0$ gibt, so dass für alle $A \in \mathcal{F}$ mit $\nu(A) < \delta$ ist $\mu(A) < \epsilon$.

13.2.3 Satz (Radon-Nikodym): Es seien μ, ν zwei σ-finite Maße auf dem Maßraum $(\Omega, \mathcal{F})$.

i.) Wenn $\mu \ll \nu$, dann gibt es eine nicht-negative numerische Funktion $f : \Omega \to \bar{\mathbb{R}}_{\geq 0}$ mit

a.) f ist messbar.

b.) Für alle $B \in \mathcal{F}$ gilt

$$\mu(B) = \int_B f d\nu = \int \mathbb{1}_B f d\nu.$$

ii.) f ist ν-fast überall eindeutig.

13.2.4 Definition: Die Funktion f aus dem Satz von Radon-Nikodym heißt **Radon-Nikodym-Ableitung** von μ bezüglich ν. Man schreibt:

$$f = \frac{d\mu}{d\nu}.$$

[1] Wahrscheinlichkeitsmaße sind endlich!

13.3 Bedingte Erwartung bezüglich $\mathcal{F}$

Im folgenden benötigen wir regelmäßig mehrere σ-Algebren. Wir beachten dabei folgende Konventionen: $\mathcal{A}$ bezeichnet die *maximale* Information des Experiments $(\Omega, \mathcal{A}, \mathbb{P})$; $\mathcal{A}$ wie *A*lles.

Für Teilinformationen verwenden wir Sub-σ-Algebren, für die wir die typischerweise die Notation $\mathcal{F}$ verwenden. Was so eine Sub-σ-Algebra erfasst, wollen wir an dem abgebildeten **abgeklebten Würfel** erklären.

Die ungeraden Zahlen sind rot abgeklebt und die geraden Zahlen blau. Man kann die Zahlen unter dem Klebeband erkennen, wir sollen/wollen uns aber vorstellen, dass wir die Zahlen unter dem Klebeband nicht erkennen können. Wir wissen lediglich welche Zahlen mit welcher Farbe abgeklebt sind.

i.) Bestimmen Sie den bedingten Erwartungswert $\mathbb{E}(X|\,\text{rot}\,)$ des Augenwerts, wenn Sie rot sehen! ii.) Wie hoch ist die bedingte Wahrscheinlichkeit $\mathbb{P}(\{1,3\}|\,\text{rot}\,)$für $\{1,3\}$, wenn Sie rot sehen?

13.3.1 Bemerkung: Wir betrachten eine integrierbare Zufallsvariable $X \in \mathcal{L}^1$. Als **Lagemaß** enthält der Erwartungs-

wert $\mathbb{E}(X)$ Information über die **Lage** von X: Gemäß des GgZ verteilen sich bei unabhängiger identischer Wiederholung die Realisierungen **durchschnittlich** um $\mathbb{E}(X)$. Diese Information über die Verteilung ist aber grob; nur eine einzelne Zahl. Oft haben wir **Zusatzinformation/Teilinformation/Fallunt** über die Werte von X.

Wir betrachten beispielsweise die Augenzahl X eines fairen Würfels. Dann ist bekanntlich $\mathbb{E}(X) = 3.5$. Angenommen die geraden Zahlen auf dem Würfel werden blau und die ungeraden Zahlen rot angeklebt. Dann können wir zwei *Werte* ermitteln; je nachdem, welche Farbe wir sehen (je nach Ergebnis ω) :

$$\mathbb{E}(X|\text{Farbinfo})(\omega) = \begin{cases} 3 & \text{falls } \omega \in \{1,3,5\} \quad \text{[wir sehen rot]} \\ 4 & \text{falls } \omega \in \{2,4,6\} \quad \text{[wir sehen blau]} \end{cases}$$

Wir bemerken, dass **die bedingte Erwartung eine Zufallsvariable** ist, den ihr Wert hängt von der zufälligen Ziehung ω ab!

Diese Perspektive wollen wir formalisieren. Es ist $\Omega = \{1,2,3,...,6\}$ und $\mathcal{A} = \mathcal{P}(\Omega)$. Wir wissen schon, dass man die Teilinformation (die Farbinformation bei abgeklebten Würfel) durch die folgende **Sub-σ-Algebra** erfassen kann:

$$\mathcal{F} = \{\emptyset, \Omega, \{1,3,5\}, \{2,4,6\}\}.$$

Wenn die Teilinformation $\mathcal{F}$ vorliegt, dann können wir für die Mengen in $\mathcal{F}$ *entscheiden*, ob sie eingetreten sind oder nicht.[2] Wenn wir beispielsweise *rot* sehen, dann wissen wir, dass $\{1, 3, 5\}$ eingetreten ist. Wir wissen auch, dass $\{2, 4, 6\}$ nicht eingetreten ist. Wir können jedoch nicht entscheiden, ob 3 eingetreten bzw. nicht eingetreten ist. So genau/fein ist die Teilinformation $\mathcal{F}$ nicht.

Angenommen Y ist eine Zufallsvariable, die **nur die Teilinformation** in $\mathcal{F}$ **verwendet**. Mathematisch fordern wir dann: Y ist $\mathcal{F}$-messbar.

13.3.2 Satz (Kolmogorov): Es sei $X : \Omega \to \mathbb{R}$ eine integrierbare Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Es sei $\mathcal{F}$ eine Sub-σ-Algebra von $\mathcal{A}$. Dann gibt es eine $\mathbb{P}$-f.s. eindeutig bestimmte integrierbare Zufallsvariable Y mit:

a.) Y ist $\mathcal{F}$-messbar.

b.) Für alle $F \in \mathcal{F}$ gilt

$$\mathbb{E}(\mathbb{1}_F Y) = \mathbb{E}(\mathbb{1}_F X).$$

Beweis: Henze [19, Seite 171]

[2] Wir beachten aber den Hinweis von Billingsley (1995, S. 57 ff), dass wir streng zwischen mathematischen Aussagen und deren Interpretation unterscheiden müssen.

13.3.3 Definition: Eine Zufallsvariable Y wie aus dem vorhergehenden Satz heißt **bedingte Erwartung**[3] von X gegeben/bezüglich $\mathcal{F}$. Für die bedingte Erwartung schreiben wir $\mathbb{E}(X|\mathcal{F})$.

13.3.4 Bemerkung: Was bedeuten die beiden Bedingungen in der Definition der bedingten Erwartung?

Die Bedingung a.) bedeutet, dass Y sich nur auf die Information in $\mathcal{F}$ bezieht. Mit diesem Aspekt haben wir uns relativ ausführlich im Kapitel 11 beschäftigt.

Die Bedingung b.) bedeutet, dass sich $Y = \mathbb{E}(X|\mathcal{F})$ auf **allen Ereignissen** $F \in \mathcal{F}$ durchschnittlich wie X verhält:

$$\mathbb{E}(Y; F) = \mathbb{E}(\mathbb{1}_F Y) = \mathbb{E}(\mathbb{1}_F X) = \mathbb{E}(X; F).$$

▶ Diese Aussage wird deutlicher durch die nächste Proposition.

13.3.5 Proposition: Angenommen $\mathcal{F}$ wird von einer Zerlegung $F_i, i = 1, ..., n, \Omega = \biguplus F_i$ erzeugt; wir unterstellen zudem $\mathbb{P}(F_i) > 0$. $\mathbb{E}(X|\mathcal{F})$ ist $\mathcal{F}$-messbar und deshalb auf den Mengen F_i konstant. Es gilt in der Tat

$$\mathbb{E}(X|\mathcal{F})(\omega) = \frac{\mathbb{E}(\mathbb{1}_{F_i} X)}{\mathbb{P}(F_i)} = \frac{\mathbb{E}(X; F_i)}{\mathbb{P}(F_i)} \text{ für } \omega \in F_i.$$

[3] Vorsicht: bedingte Erwartung und nicht bedingter Erwartungswert. Die gleichen sich, sind aber nicht dasselbe. Details folgen.

Wenn X zudem diskret ist, dann gilt

$$\mathbb{E}(X|\mathcal{F})(\omega) = \sum_{x \in \text{Bild}(X)} x \cdot \mathbb{P}(X = x \,|\, F_i) \text{ für } \omega \in F_i.$$

Beweis: Da Y $\mathcal{F}$-messbar ist, ist Y konstant auf F_i; sagen wir $= \alpha_i$. Gemäß Satz von Kolmogorov gilt

$$\mathbb{E}(\mathbb{1}_{F_i} Y) = \mathbb{E}(\mathbb{1}_{F_i} X).$$

Es ist

$$\mathbb{E}(\mathbb{1}_{F_i} Y) = \alpha_i \mathbb{P}(F_i)$$

Also gilt

$$\mathbb{E}(X|\mathcal{F})(\omega) = \alpha_i = \frac{\mathbb{E}(\mathbb{1}_{F_i} X)}{\mathbb{P}(F_i)}, \omega \in F_i$$

Wenn X diskret ist, dann $\text{Bild}(X) = \{x_1, x_2, ...\}$ (x_i unterschiedlich).

Wir beobachten

$$X = \sum_{x \in \text{Bild}(X)} \mathbb{1}_{\{X = x_k\}} x_k$$

und weiterhin

$$\begin{aligned}\mathbb{E}(\mathbb{1}_{F_i}X) &= \mathbb{E}\left(\mathbb{1}_{F_i}\sum_{x_k\in\text{Bild}(X)}\mathbb{1}_{\{X=x_k\}}x_k\right)\\ &= \mathbb{E}\left(\sum_{x_k\in\text{Bild}(X)}x_k\mathbb{1}_{F_i}\mathbb{1}_{\{X=x_k\}}\right)\\ &= \mathbb{E}\left(\sum_{x_k\in\text{Bild}(X)}x_k\mathbb{1}_{F_i\cap\{X=x_k\}}\right)\end{aligned}$$

$$\begin{aligned}&= \sum_{x_k\in\text{Bild}(X)}x_k\mathbb{E}(\mathbb{1}_{F_i\cap\{X=x_k\}})\\ &= \sum_{x_k\in\text{Bild}(X)}x_k\mathbb{P}(F_i\cap\{X=x_k\})\\ &= \sum_{x_k\in\text{Bild}(X)}x_k\mathbb{P}(F_i\cap\{X=x_k\})\\ &= \sum_{x_k\in\text{Bild}(X)}x_k\mathbb{P}(\{X=x_k\}|F_i)\mathbb{P}(F_i)\end{aligned}$$

Also

$$\mathbb{E}(X|\mathcal{F})(\omega)=\alpha_i=\sum_{x_k\in\text{Bild}(X)}x_k\mathbb{P}(\{X=x_k\}|F_i),\omega\in F_i$$

13.3.6 Bemerkung: Wir müssen den Unterschied zwischen

$$\mathbb{E}(X|\mathcal{F})(\omega),\omega\in F_i$$

und

$$\mathbb{E}(\mathbb{1}_{F_i}X)$$

beachten. Der Wert $\mathbb{E}(X|\mathcal{F})(\omega), \omega \in F_i$ berücksichtigt im Nenner die Wahrscheinlichkeit, dass $\omega \in F_i$ (also $\mathbb{P}(F_i)$) ist. Angenommen eine 1 wurde gewürfelt, aber die ungeraden Zahlen sind rot abgeklebt. Wenn wir mit diesem Wissen die bedingte Erwartung bestimmen, dann rechnen wir $\frac{\frac{1}{6}}{\frac{1}{2}} \cdot 1 + \frac{\frac{1}{6}}{\frac{1}{2}} \cdot 3 + \frac{\frac{1}{6}}{\frac{1}{2}} \cdot 5 = \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 3 + \frac{1}{3} \cdot 5$ entspricht nicht der Wahrscheinlichkeit für eine 1 [die wäre $\frac{1}{6}$]. Die Wahrscheinlichkeit $\frac{1}{3} = \frac{1}{6}/\frac{1}{2}$ berücksichtigt im Nenner die Wahrscheinlichkeit für eine ungerade Zahl $\mathbb{P}(\{1,3,5\}) = \frac{1}{2}$, denn wir wissen, dass $\{1,3,5\}$ eingetreten ist, denn wir sehen rot.

Der Wert[4] $\mathbb{E}(\mathbb{1}_{F_i}X)$ hingegen verwendet die ursprünglichen (a-priori) Wahrscheinlichkeiten $\mathbb{P}$ und *nullt* alle Werte mit Urbildern außerhalb F_i. Wir erhalten also für $\mathbb{E}(\mathbb{1}_{F_i}X)$ ***fast*** so etwas wie einen *bedingten* Erwartungswert; die ungenullten Wahrscheinlichkeiten geben untereinander die richtigen Wahrscheinlichkeits-Proportionen an, addieren sich aber nicht zu Eins. Wenn wir diese *a-priori* Wahrscheinlichkeiten durch $\mathbb{P}(F_i) = \frac{1}{2}$ teilen, dann erhalten wir die bedingte Erwartung $\mathbb{E}(X|\mathcal{F})(\omega), \omega \in F_i$ berechnet mit den bedingten Wahrscheinlichkeiten.

Der *Vorteil* von $\mathbb{E}(\mathbb{1}_F X)$ im Vergleich zu $\mathbb{E}(X|F_i)$ besteht

[4] Ohne |.. . Manchmal schreibt man $\mathbb{E}(X; F_i)$.

darin, dass wir die Voraussetzung $\mathbb{P}(F) > 0$ nicht benötigen. Für die allgemeine Definition ist das komfortabler. In der allgemeinen Definition erstrecken sich die Gleichung über ganz $\mathcal{F}$ und die Bedingung $\mathbb{P}(F) > 0$ ist dann für mit ausschlaggebende Mengen $\mathbb{P}(F) > 0$ nicht erfüllt

▶ Der folgende Satz fasst die wichtigsten Eigenschaften der bedingten Erwartung zusammen. **Die Formeln sind SEHR wichtig für das Verständnis und für das Rechnen mit bedingten Erwartungen.** Die Eigenschaft f.) nennt man manchmal das **Gesetz von der iterierten Erwartung** oder die **Turmeigenschaft** der bedingten Erwartung. Die Turmeigenschaft und die Regel bei Unabhängigkeit g.) habe ich beim praktischen Rechnen mehr als 1000 mal angewendet!

Die Eigenschaft i.) zeigt die große Bedeutung der bedingten Erwartung für die Anwendung. Die bedinge Erwartung ist die **beste Prognose**, wenn der Prognostiker über die Information in $\mathcal{F}$ verfügt [und der **mittlere quadratische Fehler** das Qualitätskriterium ist].

13.3.7 Satz: Es seien X, X_1, X_2 integrierbare Zufallsvariablen und $\mathcal{F} \subset \mathcal{A}$ eine Sub-σ-Algebra von $\mathcal{A}$. Dann gilt

a.) Für alle $\alpha_1, \alpha_2 \in \mathbb{R}$ gilt $\mathbb{P}$-f.ü.

$$\mathbb{E}(\alpha_1 X_1 + \alpha_2 X_2|\mathcal{F}) = \alpha_1 \mathbb{E}(X_1|\mathcal{F}) + \alpha_2 \mathbb{E}(X_2|\mathcal{F})$$

b.) Ist $X_1 \leq X_2$, so gilt $\mathbb{P}$-f.ü.

$$\mathbb{E}(X_1|\mathcal{F}) \leq \mathbb{E}(X_2|\mathcal{F})$$

c.) Ist $X(\omega) = a$, $\mathbb{P}$-f.ü., dann ist $\mathbb{P}$-f.ü.

$$\mathbb{E}(X|\mathcal{F}) = a$$

d.) Ist $\mathcal{F} = \{\emptyset, \Omega\}$, dann ist $\mathbb{P}$-f.ü.

$$\mathbb{E}(X|\mathcal{F}) = \mathbb{E}(X)$$

e.) Ist $\mathcal{F} = \mathcal{A}$, dann ist $\mathbb{P}$-f.ü.

$$\mathbb{E}(X|\mathcal{F}) = X$$

f.) Ist $\mathcal{G} \subset \mathcal{F}$ eine Sub-σ-Algebra von $\mathcal{F}$, dann gilt $\mathbb{P}$-f.ü.

$$\mathbb{E}(\mathbb{E}(X|\mathcal{F})|\mathcal{G}) = \mathbb{E}(X|\mathcal{G}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{F})$$

g.) Ist V eine $\mathcal{F}$-messbare integrierbare Zufallsvariable, dann

ist $\mathbb{P}$-f.ü.

$$\mathbb{E}(VX|\mathcal{F}) = V\mathbb{E}(X|\mathcal{F})$$

h.) Sind $\sigma(X)$ und $\mathcal{F}$ unabhängig, dann gilt $\mathbb{P}$-f.ü.

$$\mathbb{E}(X|\mathcal{F}) = \mathbb{E}(X)$$

i.) Ist $X \in \mathcal{L}^2$, dann ist $\mathbb{E}(X|\mathcal{F}) \in \mathcal{L}^2$ und es gilt

$$\mathbb{E}[\,(X - \mathbb{E}(X|\mathcal{F}))^2\,] = \min_Y\{\mathbb{E}[(X-Y)^2] \,|\, Y \in \mathcal{L}^2, Y \in \mathcal{F}\}$$

13.3.8 Definition: Es sei X eine integrierbare Zufallsvariable und $Z : \Omega \to \mathbb{R}^n$ ein Zufallsvektor. Es sei $\mathcal{F} = Z^{-1}(\mathcal{B}^n)$, d.h. $\mathcal{F}$ ist die von Z erzeugte σ-Algebra. . Dann heißt $\mathbb{E}(X|Z) = \mathbb{E}(X|\mathcal{F})$ die **bedingte Erwartung** von X gegeben Z.

▶ In der vorhergehenden Definition stammt die Teilinformation also aus einem Zufallsvektor, aber indirekt aus $\sigma(Z)$ und nicht direkt aus dem Wert den Z angenommen hat. Die naheliegende Idee: Die Teilinformation kann direkt **dem angenommenen Wert des Zufallsvektor** Z entnommen werden (und nicht *indirekt* der σ-Algebra $\mathcal{F} = Z^{-1}(\mathcal{B}^n)$). Das stimmt, wie das folgende sogenannte Faktorisierungslemma zeigt. Dieser Zugang ist sogar *natürlicher*.

$$\text{Z } \mathcal{F} = \sigma(Z) \text{ dann } \mathbb{E}(X|\mathcal{F})(\omega) = \mathbb{E}(X|Z = Z(\omega))$$

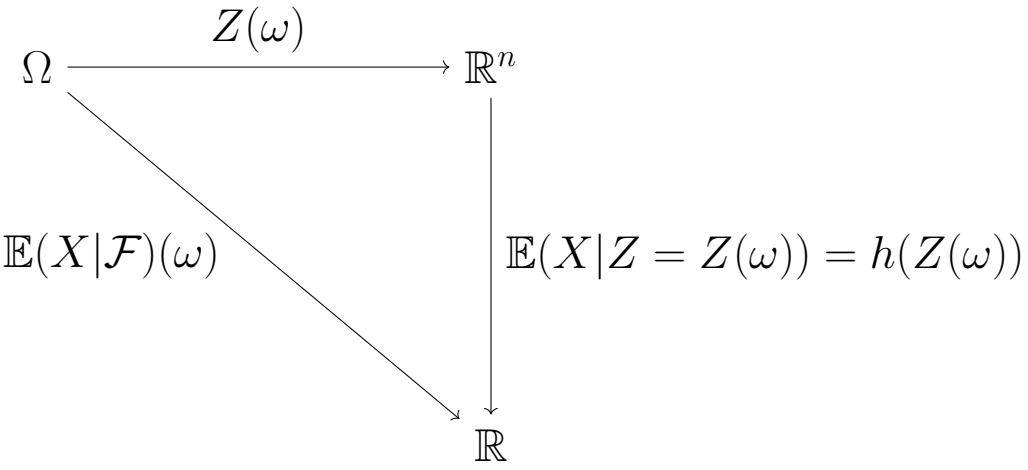


Abbildung 13.1: Faktorisierungslemma für bedingte Erwartungen bzw. bedingter Erwartungswert

13.3.9 Satz (Faktorisierungslemma für die bedingte Erwartung): Es sei X eine integrierbare Zufallsvariable und $Z : \Omega \to \mathbb{R}^n$ ein Zufallsvektor. Es sei $\mathcal{F} = \sigma(Z) = Z^{-1}(\mathcal{B}^n)$ die durch Z erzeugte σ-Algebra. Es gibt eine $\mathcal{B}^n$-$\mathcal{B}$ messbare Abbildung $h : \mathbb{R}^n \to \mathbb{R}$ mit

$$\mathbb{E}(X|Z)(\omega) = h(Z(\omega)) \ \mathbb{P}\text{-f.ü. .}$$

▶ Wir wollen Erwartungen bezüglich X ermitteln. Dazu können wir die Information beachten, die von Z erzeugt wird. Wenn ω eingetreten ist, dann ist die Erwartung $\mathbb{E}(X|Z)(\omega)$ (und damit besser informiert als $\mathbb{E}(X)$). Alternativ können

wir den Wert $Z(\omega)$ verwenden und $h(Z(\omega))$ ermitteln. h liefert uns also die bedingte Erwartung als Funktion des Wertes $Z(\omega)$, den die Bedingung annimmt. Es fügt sich also alles schön zusammen. Wir definieren dementsprechend den bedingten Erwartungswert wie folgt.

13.3.10 Definition: Die im vorhergehenden Satz genannte Abbildung $h : \mathbb{R}^n \to \mathbb{R}$ heißt der **bedingte Erwartungswert** von X gegeben $Z = z$. Wir schreiben dafür

$$h(z) = \mathbb{E}(X|Z = z), z \in \mathbb{R}^n.$$

$z \mapsto \mathbb{E}(X|Z = z)$ ist also eine Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}$. Das Argument ist z (hier blau markiert).

13.3.11 Bemerkung: i.) Es gilt gemäß Definition $\mathbb{P}$-f.ü.

$$\mathbb{E}(X|\sigma(Z))(\omega) = \mathbb{E}(X|Z)(\omega) = \mathbb{E}(X|Z = Z(\omega)).$$

ii.) Wir müssen trotzdem[5] die **bedingten Erwartungen** $\mathbb{E}(X|\mathcal{F})$ von den **bedingten Erwartungswerten** $\mathbb{E}(X|Z = z)$ unterscheiden. In beiden Fällen werden Erwartungen von X gebildet und es kommt das *gleiche* heraus. Aber:

- Der Definitionsbereich von $\mathbb{E}(X|\mathcal{F})$ ist Ω. Was erwarten wir für X, wenn wir den Information $\sigma(Z)$ haben und

[5] Trotz des Gleichheitszeichens.

ω eingetreten ist.

- Der Definitionsbereich von $\mathbb{E}(X|Z = z)$ ist $\mathbb{R}^n$. Was erwarten wir für X, wenn wir den Information haben, dass z eingetreten ist.

- Die gleiche Antwort erhält man, wenn $z = Z(\omega)$ ist.

▶ Zwei für die angewandte besonders wichtige Fälle – die man zudem auch elementarer hätte behandeln können – sollen in den anschließenden Propositionen behandelt werden. (1) Das Bild von Z ist endlich und (2) X und Z haben eine gemeinsame Dichte.

13.3.12 Proposition: Es seien X und Z reellwertige Zufallsvariablen auf $(\Omega, \mathcal{F}, \mathbb{P})$ und $\text{Bild}(Z) = \{z_1, ..., z_n\}$ mit $\mathbb{P}(Z = z_i) > 0$. Die von Z erzeugte σ-Algebra wird durch die Zerlegung $\{Z = z_i\}$ erzeugt. Ferner, eine Zufallsvariable Y ist genau dann bezüglich $\sigma(Z)$ messbar, wenn Y auf den Mengen $\{Z = z_i\}$ konstant ist.

Es sei für $z_i \in \text{Bild}(Z)$

$$\mathbb{P}(X \in B|Z = z_i) = \frac{\mathbb{P}(X \in B \wedge Z = z_i)}{\mathbb{P}(Z = z_i)}$$

Dann gilt für $z_i \in \text{Bild}(Z)$

$$\mathbb{E}(X|Z = z_i) = \int X d\,\mathbb{P}(X \in dx|Z = z_i).$$

Ist auch X diskret, d.h. gilt $\text{Bild}(X) = \{x_1, ..., x_k\}$, dann ist

$$\begin{aligned}\mathbb{E}(X|Z = z_i) &= \sum x_i\, \mathbb{P}(X = x_i \,|\, Z = z_i) \\ &= \sum x_i\, \frac{\mathbb{P}(X = x_i \wedge Z = z_i)}{\mathbb{P}(Z = z_i)}\end{aligned}$$

13.3.13 Beispiel: $\Omega = \{\clubsuit, \spadesuit, \diamondsuit, \heartsuit\}$ mit den Wahrscheinlichkeiten (in der entsprechenden Reihenfolge) $\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}$. Die Zufallsvariable $X : \Omega \to \mathbb{R}$ habe die Werte (wieder in der Reihenfolge) $3, 4, 5, 6$ und die die Zufallsvariable $Z : \Omega \to \mathbb{R}$ habe die Werte (wieder in der Reihenfolge) $10, 10, 11, 11$.

Bestimmen Sie:

a.) $\mathbb{E}(X|\sigma(Z))$ und
b.) $\mathbb{E}(X|Z = z)$.

13.3.14 Proposition: Es seien X und Z zwei (reellwertige) Zufallsvariablen mit gemeinsamer Dichte f. Es gilt also für $B_1, B_2 \in \mathcal{B}$:

$$\mathbb{P}((X, Z) \in B_1 \times B_2) = \int_{B_2} \int_{B_1} f(x, z)\, dx\, dz.$$

Wir bezeichnen die Randdichten mit f_X bzw. f_Z und wir

definieren

$$f(x|z) = \begin{cases} \frac{f(x,z)}{f_Z(z)} & \text{falls } f_Z(z) \neq 0 \\ 0 & \text{sonst.} \end{cases}$$

Dann gilt für den bedingten Erwartungswert von X gegebenen $Z = z$

$$\mathbb{E}(X|Z = z) = \int_{\mathbb{R}} x f(x|z) dx.$$

Beweis: Es sei

$$g(z) = \int_{\mathbb{R}} x f(x|z) dx.$$

Wir müssen beweisen, dass für alle $B \in \mathcal{B}$

$$\mathbb{E}(\mathbb{1}_B(Z) g(Z)) = \mathbb{E}(\mathbb{1}_B(Z) X)$$

gilt. Es gilt in der Tat

$$\begin{aligned} \mathbb{E}(\mathbb{1}_B(Z) g(Z)) &= \int \mathbb{1}_B(z) \left(\int_{\mathbb{R}} x f(x|z) dx \right) f_Z(z) dz \\ &= \int \mathbb{1}_B(z) \left(\int x \frac{f(x,z)}{f_Z(x)} dx \right) f_Z(z) dz \\ &= \int \int \mathbb{1}_B(z) x f(x,z) dx dz \\ &= \mathbb{E}(\mathbb{1}_B(Z) X) \end{aligned}$$

13.3.15 Beispiel: ...

$$f(x,y) = \frac{1}{2}\,\mathbb{1}_D(x,y)$$

13.4 Bedingte Wahrscheinlichkeit und bedingte Verteilung

13.4.1 Bemerkung: Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Es seien $A, B \in \mathcal{A}$ Ereignisse, wobei $0 < \mathbb{P}(B) < 1$. Wir interessieren uns für die Wahrscheinlichkeit von A. Ohne *Information* ist die gleich $\mathbb{P}(A)$. Wenn A und B nicht unabhängig sind, dann können wir die Wahrscheinlichkeitseinschätzung für das Eintreten von A anpassen, wenn uns Information über B bzw. B^c vorliegt:

$$f(\omega) = \begin{cases} \mathbb{P}(A|B) \text{ falls } \omega \in B \\ \mathbb{P}(A|B^c) \text{ falls } \omega \in B^c \end{cases}$$

Wir betrachten wieder den abgeklebten Würfel (rot ungerade, blau gerade) und das Ereignis *mindestens eine vier* gewürfelt zu haben; $A = \{4,5,6\}$. Ohne Information ist $\mathbb{P}(A) = \frac{1}{2}$. Wenn wir rot sehen, dann ist $\mathbb{P}(A|\text{ rot }) = \frac{1}{3}$ und $\mathbb{P}(A|\text{ blau }) = \frac{2}{3}$

Beachte: $f(\omega)$ ist eine Zufallsvariable, denn sie hängt vom Ergebnis ab (über das wir aber nicht volle Kenntnis haben).

$f(\omega)$ beantwortet die Frage: Welche Erwartung haben wir, wenn ω eingetreten ist, wir aber nur $\mathcal{F}$ (die Farbe) wissen.

Wir wollen die Situation formalisieren. Die Vorinformation wird wieder durch eine σ-Algebra repräsentiert, also $\mathcal{F} = \{\emptyset, \Omega, B, B^c\}$. Wir schreiben $f = \mathbb{P}(A|\mathcal{F})$. f ist eine Zufallsvariable, denn je nach Ergebnis ergibt sich eine andere bedingte Wahrscheinlichkeit

f hat zwei Eigenschaften

a.) f ist $\mathcal{F}$-messbar

b.) Für alle $F \in \mathcal{F}$ gilt

$$\int_F \mathbb{P}(A|\mathcal{F}) d\,\mathbb{P} = \mathbb{P}(A \cap F)$$

Es ist für $F = \{1, 3, 5\}$

$$\mathbb{P}(A \cap F) = \frac{1}{6}$$

und mit $x_F = \mathbb{P}(A|\mathcal{F})(\{1\}) = \mathbb{P}(A|\mathcal{F})(\{3\}) = \mathbb{P}(A|\mathcal{F})(\{5\})$ [diese Werte müssen wegen der geforderten $\mathcal{F}$-Messbarkeit

gleich sein]

$$\begin{aligned}\int_F \mathbb{P}(A|\mathcal{F})d\,\mathbb{P} &= \mathbb{E}(\mathbb{1}_F\mathbb{P}(A|\mathcal{F})) \\ &= \mathbb{1}_F(1)\mathbb{P}(\{1\})x_F + \mathbb{1}_F(2)\mathbb{P}(\{2\})x_F + ... + \mathbb{1}_F(6)\mathbb{P}(\{6\})x_F \\ &= \mathbb{P}(\{1\})x_F + \mathbb{P}(\{3\})x_F + \mathbb{P}(\{5\})x_F \\ &= (\mathbb{P}(\{1\}) + \mathbb{P}(\{3\}) + \mathbb{P}(\{5\}))x_F = \frac{1}{2}x_F\end{aligned}$$

Also gilt

$$x_F = \frac{1/6}{1/2} = \frac{1}{3} = \mathbb{P}(A|F)$$

13.4.2 Satz: Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $\mathcal{F} \subset \mathcal{A}$ eine Sub-σ-Algebra von $\mathcal{A}$ und $A \in \mathcal{A}$ ein Ereignis. Dann gilt

i.) $\nu(F) = \mathbb{P}(A \cap F), F \in \mathcal{F}$ ist ein endliches Maß auf $\mathcal{F}$.

ii.) $\nu \ll \mathbb{P}$.

iii.) Es gibt eine $\mathcal{F}$-messbare $\mathbb{P}$-integrierbare Abbildung f mit

$$\mathbb{P}(A \cap F) = \nu(F) = \int_F f d\,\mathbb{P},\ F \in \mathcal{F}$$

13.4.3 Definition: Die Funktion f aus dem vorhergehenden Satz heißt die **bedinge Wahrscheinlichkeit** von A bezüg-

lich $\mathcal{F}$. Sie wird mit

$$f = \mathbb{P}(A|\mathcal{F})$$

bezeichnet.

13.4.4 Satz: Die folgenden beiden Bedingungen charakterisieren die bedinge Wahrscheinlichkeit:

a.) $\mathbb{P}(A|\mathcal{F})$ ist $\mathcal{F}$-messbar

b.) Für alle $F \in \mathcal{F}$ gilt

$$\int_F \mathbb{P}(A|\mathcal{F}) d\,\mathbb{P} = \mathbb{P}(A \cap F)$$

13.4.5 Satz:

$$\mathbb{P}(A|\mathcal{F}) = \mathbb{E}(\mathbb{1}_A|\mathcal{F})$$

13.4.6 Bemerkung: Manchmal ist die folgende Formel *leichter*

$$\int_F \mathbb{P}(A|\mathcal{F}) d\,\mathbb{P} = \mathbb{E}(\mathbb{1}_F \mathbb{P}(A|\mathcal{F}))$$

13.4.7 Bemerkung: Für die Interpretation der Bedingung b.) beachten wir für den diskreten Fall $\Omega = \{\omega_1, ..., \omega_k\}$ ei-

nerseits

$$\begin{aligned}\int_F \mathbb{P}(A|\mathcal{F})d\,\mathbb{P} &= \int_\Omega \mathbb{1}_F\,\mathbb{P}(A|\mathcal{F})d\,\mathbb{P}\\ &= \sum_{i=1}^{k} \mathbb{1}_F(\omega_k)\mathbb{P}(A|\mathcal{F})(\omega_k)\mathbb{P}(\omega_k)\end{aligned}$$

und andererseits die diskrete Cocktail-Formel

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A|\{\omega_1\})\mathbb{P}(\{\omega_1\}) + ... + \mathbb{P}(A|\{\omega_k\})\mathbb{P}(\{\omega_k\})\\ &= \mathbb{E}(\mathbb{P}(A|\{\omega_1\}))\end{aligned}$$

Also lautet die Bedingungen b.)

$$\begin{aligned}&\mathbb{P}(A\cap F)\\ &= \mathbb{P}(A\cap F|F)\mathbb{P}(F)\\ &= \mathbb{P}(A\cap F|F)(\mathbb{P}(F|\{\omega_1\})\mathbb{P}(\omega_1) + ... + \mathbb{P}(F|\{\omega_k\})\mathbb{P}(\omega_k))\\ &= \mathbb{1}_F(\omega_1)\mathbb{P}(A\cap F|\{\omega_1\})\mathbb{P}(\{\omega_1\}) + ... + \mathbb{1}_F(\omega_k)\mathbb{P}(A\cap F|\{\omega_k\})\mathbb{P}(\{\omega_k\})\\ &= \mathbb{E}(\mathbb{1}_F(\omega)\mathbb{P}(A|\{\omega\}))\\ &= \int_\Omega \mathbb{1}_F\,\mathbb{P}(A|\mathcal{F})d\,\mathbb{P}.\end{aligned}$$

Bei dieser Bedingung wird also das **Prinzips der Cocktail-Formel** auf alle Ereignisse $A\cap F, F\in\mathcal{F}$ angewendet.

Wir betrachten nochmal $\mathcal{F} = \{\emptyset, \Omega, \{1,3,5\}, \{2,4,6\}\}$ und

$A = \{4, 5, 6\}$. Dann ist (diesmal) für $\tilde{F} = \{2, 4, 6\}$

$$\mathbb{P}(A \cap \tilde{F}) = \frac{2}{6}$$

und mit $x_{\tilde{F}} = \mathbb{P}(A|\mathcal{F})(\{2\}) = \mathbb{P}(A|\mathcal{F})(\{4\}) = \mathbb{P}(A|\mathcal{F})(\{6\})$ [diese Werte müssen wegen der geforderten $\mathcal{F}$-Messbarkeit gleich sein]

$$\begin{aligned}\int_\Omega \mathbb{1}_{\tilde{F}}\, \mathbb{P}(A|\mathcal{F}) d\,\mathbb{P} &= \mathbb{P}(\{2\})\, x_{\tilde{F}} + \mathbb{P}(\{4\}) x_{\tilde{F}} + \mathbb{P}(\{6\}) x_{\tilde{F}} \\ &= (\mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\})) x_{\tilde{F}} = \frac{1}{2} x_{\tilde{F}}.\end{aligned}$$

Also

$$\frac{2}{6} = \frac{1}{2} x_{\tilde{F}}$$

bzw.

$$x_{\tilde{F}} = \frac{2/6}{1/2} = \frac{2}{3} = \mathbb{P}(A|\tilde{F}).$$

13.4.8 Satz: Es gilt

a.) Es ist $\mathbb{P}$-f.ü.

$$\mathbb{P}(\emptyset|\mathcal{F}) = 0$$

b.) Es ist $\mathbb{P}$-f.ü.

$$\mathbb{P}(\Omega|\mathcal{F}) = 1$$

c.) Für alle $A \in \mathcal{A}$ ist $\mathbb{P}$-f.ü.

$$0 \leq \mathbb{P}(A|\mathcal{F}) \leq 1$$

d.) Für $(A_i)_{i\in\mathbb{N}}$ mit $A_i \cap A_j = \emptyset, i \neq j$ gilt $\mathbb{P}$-f.ü.

$$\mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i|\mathcal{F}\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|\mathcal{F})$$

13.4.9 Bemerkung: i.) Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $\mathcal{F} \subset \mathcal{A}$ eine Sub-σ-Algebra von $\mathcal{A}$, $\omega \in \Omega$. Dann ist

$$A \mapsto \mathbb{P}(A|\mathcal{F})(\omega)$$

i.A. kein Wahrscheinlichkeitsmaß, obwohl die Eigenschaften von $\mathbb{P}(A|\mathcal{F})$ die Vermutung begründen. Die Eigenschaften gelten nur $\mathbb{P}$-f.ü. und die Ausnahmemengen hängen i.A. von A ab. Man kann in der Tat Beispiele konstruieren, so dass $\mathbb{P}(A|\mathcal{F})$ kein WM ist. Unter bestimmten Voraussetzungen – die zudem in der Anwendung besonders relevant sind – kann man aber doch aus den $\mathbb{P}(A|\mathcal{F}), A \in \mathcal{A}$ ein Wahrscheinlichkeitsmaß konstruieren.

ii.) In diesem Paragraphen haben wir die bedingte Erwartung $\mathbb{E}(X|\mathcal{F})$ definiert. (Unbedingte) Erwartung sind $\mathbb{P}$-Integrale. **Sind bedingte Erwartungen Integrale bezüglich be-**

dingter Wahrscheinlichkeiten? In der Tat haben wir gerade bedingte Verteilugnen $\mathbb{P}(A|\mathcal{F})$ definiert. Die Vermutung ist naheliegend, dass

$$\mathbb{E}(X|\mathcal{F}) = \int X d\,\mathbb{P}(\cdot|\mathcal{F}).$$

gilt. Die Aussage aus i.) zeigt, dass man vorsichtig argumentieren muss, denn $\mathbb{P}(\cdot|\mathcal{F})$ ist im allgemeinen kein WM ist.

13.4.10 Satz: Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, X eine Zufallsvariable, $\mathcal{F} \subset \mathcal{A}$ eine Sub-σ-Algebra von $\mathcal{A}$. Dann gibt es eine Abbildung (einen sogenannten **Markovkern**)

$$\mu : \mathcal{B} \times \Omega \to \mathbb{R}$$

mit

a.) Für alle $\omega \in \Omega$ ist

$$\mu(\cdot, \omega) = \begin{cases} \mathcal{B} \to [0,1] \\ B \mapsto \mu(B, \omega) \end{cases}$$

ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$.

b.) Für alle $B \in \mathcal{B}$ ist

$$\mu(B, \cdot) = \mathbb{P}(X \in B|\mathcal{F})(\cdot).$$

μ heißt die **bedingte Verteilung** von X bezüglich $\mathcal{F}$.

Beweis: Siehe Billingsley [2, S. 339]

13.4.11 Satz: Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, X eine reellwertige Zufallsvariable, $\mathcal{F} \subset \mathcal{A}$ eine Sub-σ-Algebra von $\mathcal{A}$ und $\mu(\cdot, \cdot) = \mathbb{P}(X \in \cdot\, | \mathcal{F})(\cdot)$ eine bedingte Verteilung von X bezüglich $\mathcal{F}$. Es sei ferner $f : \mathbb{R} \to \mathbb{R}$ eine $\mathcal{B}$-messbare Abbildung und $f \circ X$ eine $\mathbb{P}$-integrierbare Zufallsvariable. Dann gilt

$$\begin{aligned}\mathbb{E}(f(X)|\mathcal{F})(\omega) &= \int_{\mathbb{R}} f(X)\, \mu(d\lambda, \omega) \\ &= \int_{\mathbb{R}} f(X)\, d\,\mathbb{P}(\cdot|\mathcal{F})(\omega) \\ &= \int_{\mathbb{R}} f(X)\, \mathbb{P}(d\lambda|\mathcal{F})(\omega).\end{aligned}$$

Beweis: Vgl. Mürmann [27, Abschnitt 13.3]

13.4.12 Beispiel: Prototypisches Beispiel ... und *Formel*

13.5 Totale Wahrscheinlichen

13.5.1 Satz (Gesetz(e) von der totalen Wahrscheinlichkeit):

i.) X, Y haben Dichten

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy \quad .$$

ii.) Y hat eine Dichte, X diskret

$$\mathbb{P}(X = x) = \int_{-\infty}^{\infty} \mathbb{P}(X = x|Y = y) f_Y(y) dy \quad .$$

iii.) X hat eine Dichte, Y diskret

$$f_X(x) = \sum_y f_X(x|Y = y) \mathbb{P}(Y = y) \quad .$$

Beweis: Blitzstein und Hwang [4].

13.5.2 Satz und Definition: Faltung ...

14 Erzeugende Funktionen

15 Charakteristische Funktionen

15.1 Komplexe Zahlen

15.1.1 Bemerkung: i.) **Komplexe Zahlen** geben wir regelmäßig in der Form

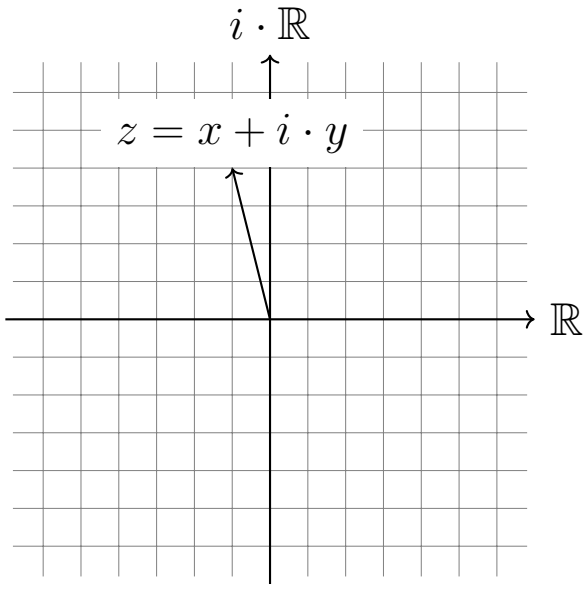
$$z = x + iy$$

an. Dabei ist $i = \sqrt{-1}$. x heißt **Realteil** von z und y der **Imaginärteil** von z. Die kompelxe Zahl $\bar{z} = x - iy$ heißt die zu z **konjugiert kompelxe** Zahl.

ii.) Unter einer komplexen Zahl kann man sich auch einen Punkt $(x, y)^T$ der komplexen Ebene vorstellen:

$$z = x + iy = \begin{pmatrix} x \\ y \end{pmatrix}$$

Die Abbildung zeigt die komplexe Ebene mit der komplexen Zahl $z = -\frac{1}{2} + 2\,i$.

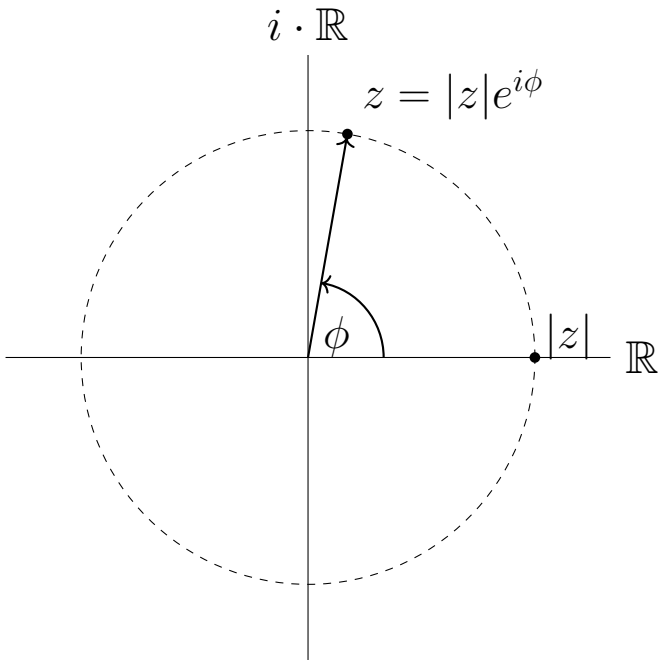


iii.) Es gilt: $||z||^2 = z\bar{z}, ||z|| = \sqrt{x^2 + y^2}$ und $z = \bar{z} \Leftrightarrow z \in \mathbb{R}$.

iv.) Es gilt

$$\begin{aligned} e^{i\phi} &= \cos(\phi) + i\sin(\phi) \\ e^{-i\phi} &= \overline{e^{i\phi}} \\ ||e^{i\phi}|| &= 1 \end{aligned}$$

Bekanntlich kann man dann komplexe Zahlen auch mittels Betrag und Winkeln darstellen.



15.2 Definition und grundlegende Eigenschaften

15.2.1 Definition: Es sei X eine Zufallsvariable. Die Abbildung

$$\psi_X : \mathbb{R} \to \mathbb{C}, t \mapsto \mathbb{E}(e^{itX})$$

heißt **charakteristische Funktion** von X.

15.2.2 Bemerkung: i.) Eigentlich wäre es besser von der charakteristische Funktion einer Verteilung zu sprechen, denn die Verteilungsfunktion legt die charakteristische Funktion

fest.

ii.) Die Abbildung $t \mapsto \mathbb{E}(e^{itX})$ kann man sich als Graph in $\mathbb{R}^3$ veranschaulichen. Für jedes t erhält man eine komplexe Zahl $\mathbb{E}(e^{itX}) = \mathbb{E}\cos(tX) + i \cdot \mathbb{E}\sin(tX)$ (die man sich als 2-dimensionalen Vektor vorstellen kann). Insgesamt erhält man Kurven wie sie in beiden Panels der Abbildung 15.1 dargestellt sind. Dabei ist t auf der vertikalen Achse abgetragen. Die blaue Kurve zeigen die CF der **Standardnormalverteilung** mit $\mu = 0$ und $\sigma = 1$; die Kurve verläuft komplett in der reellen Ebene. Die rote Kurve zeigt die CF der **Normalverteilung** mit $\mu = 5$ und $\sigma = 1$. Die beiden Panels zeigen die gleichen CFen; der Blickwinkel ist variiert.

Die Abbildung 15.2 zeigt die CFen der Gleichverteilungen auf $[-1, 1]$ (blau) bzw. auf $[0, 1]$ (rot). Die CF der Gleichverteilung auf $[-1, 1]$ ist reellwertig. Es gilt $\Psi_{\text{Unif}[-1,1]}(t) = \frac{\sin(t)}{t}$.

iii.) Die für die charakteristische Funktion relevante Information über die Zufallsvariable X ist bekanntlich komplett durch die Verteilungsfunktion, also in der Abbildung

$$F^X : \mathbb{R} \to [0, 1], x \mapsto \mathbb{P}(X \leq x)$$

enthalten.

Nach der **Transformation zur charakteristische Funk-**

tion erhalten wir die Abbildung

$$\psi_X : \mathbb{R} \to \mathbb{C}, t \mapsto \mathbb{E}(e^{itX}).$$

Später werden wir sehen, dass wir die Transformation eindeutig umkehren können. Schematisch:

$$F_X \xrightarrow{\text{cf-Transformnation}} \psi_X$$
$$F_X \xleftarrow{\text{Inverse-cf-transformnation}} \psi_X$$

Fazit: Man kann die Verteilung einer Zufallsvariablen wahlweise durch die **Verteilungsfunktion oder die charakteristische Funktion** repräsentieren.

15.2.3 Beispiel: Es sei $X \sim$ Bernoulli(p). Dann gilt

$$\psi_X(t) = \mathbb{E}(e^{itX}) = (1-p)e^{it\cdot 0} + pe^{it\cdot 1} = (1-p) + pe^{it}.$$

$e^{it\cdot 0}$ und $e^{it\cdot 1}$ sind Punkte auf dem Einheitskreis. Von diesen Punkten bilden wir die Konvexkombination Erwartungswertes.

15.2.4 Satz. Es sei X eine Zufallsvariable mit charakteristischer Funktion ψ_X. Dann gilt

i.) $\psi_X(0) = 1$ und $|\psi_X(t)| \leq 1$ für alle $t \in \mathbb{R}$.

ii.) ψ_X ist gleichmäßig stetig.

iii.) $\psi_X(-t) = \overline{\psi_X()}$ für alle $t \in \mathbb{R}$.

iv.) $\psi_{a+bX}(t) = e^{itb} \cdot \psi_X(at)$ für alle $a, b, t \in \mathbb{R}$.

15.3 Inversion

15.3.1 Satz (Inversionssatz von Levy): Es sei $\psi_X(t)$ die charakteristische Funktion der Zufallsvariable X mit der Vertilung $\mathbb{P}^X$. Dann

$$\lim_{T \nearrow \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \psi_X(t) dt = \frac{1}{2}\mathbb{P}^X(\{a\}) + \mathbb{P}^X((a,b)) + \frac{1}{2}\mathbb{P}^X(\{b\})$$

Wenn zudem $\int_{\mathbb{R}} |\Psi_X(t)| dt < \infty$, dann hat X eine stetige Dichte und es gilt

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \psi_X(t) dt,$$

so dass wir eine schöne *Dualität*

$$\psi_X(t) = \int_{\mathbb{R}} e^{-itx} f(x) dt$$

gilt.

15.4 Stetigkeit

15.4.1 Satz (Stetigkeitssatz von Levy-Cramer): Es sei X_j eine Folge von Zufallsvariablen mit Verteilungsfunktionen $X_j \sim F_j$ und charakteristischen Funktionen ψ_j. Dann ist äquivalent

i.) Es gibt eine Verteilungsfunktion F mit $F_j \xrightarrow{i.V.} F$.

ii.) Für jedes $t \in \mathbb{R}$ existiert $\psi(t) = \lim_{j\to\infty} \psi_j(t)$ und die Funktion $\psi(t)$ ist stetig im Nullpunkt.

Falls i.) oder ii.) gilt, dann ist ψ die charakteristische Funktion von F:

$$\psi(t) = \int e^{itx} dF(x).$$

15.5 Unabhängigkeit

15.5.1 Satz: Wenn die Zufallsvariablen $X_1, ..., X_n$ **unabhängig** sind, dann gilt

$$\psi_{X_1+...+X_n} = \prod_{j=1}^{n} \psi_{X_j}$$

15.5.2 Bemerkung:

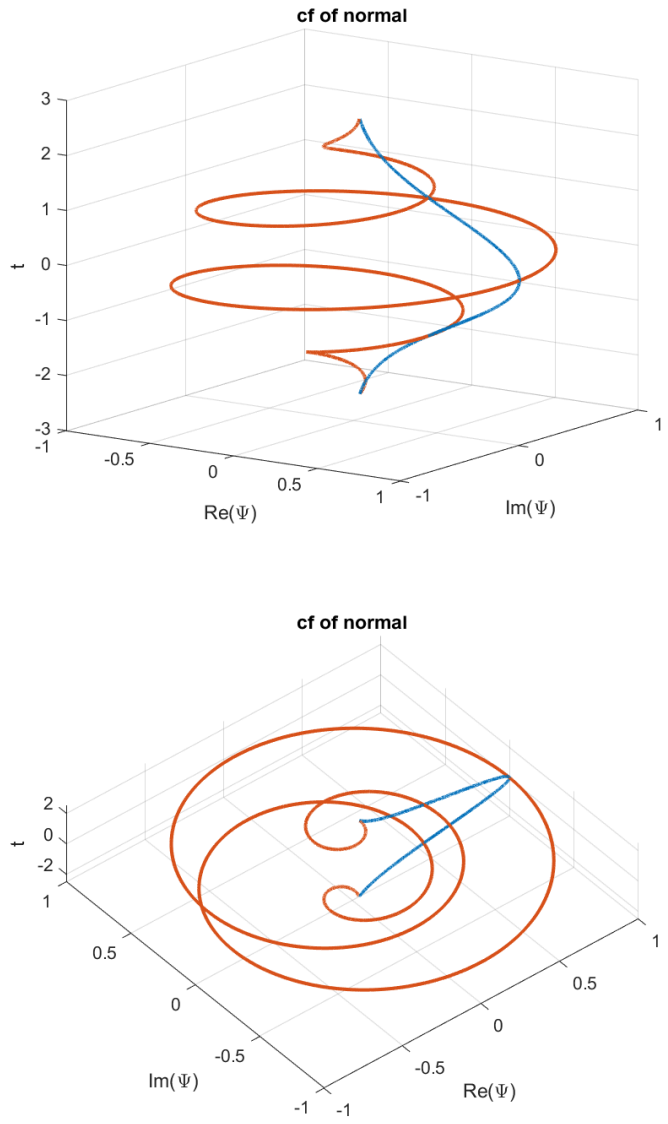


Abbildung 15.1: CF der Standardnormalverteilung (blau) und der Normalverteilung mit $\mu = 5$ und $\sigma = 1$. Quelle: Eigene Darstellung mit dem Matlab Code `plot3dSomeCFs.m`.

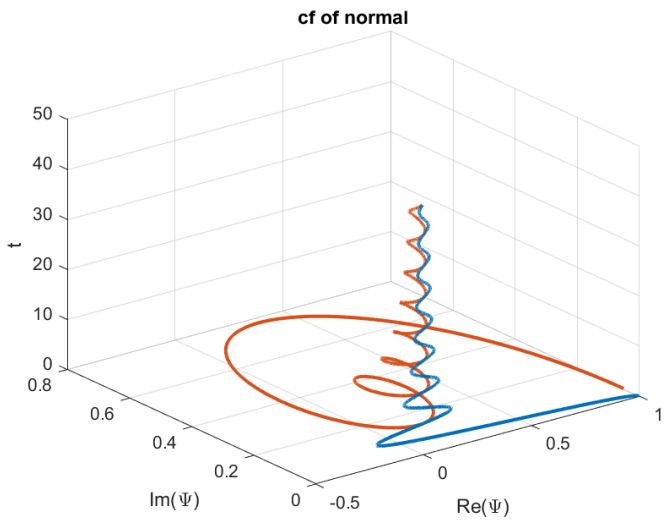


Abbildung 15.2: CFen der Gleichverteilungen auf $[-1, 1]$ (blau) bzw. auf $[0, 1]$ (rot). Quelle: Eigene Darstellung mit dem Matlab Code `plot3dSomeCFs.m`.

16 Statistik

16.1 Vorbereitung

16.1.1 Bemerkung: Von einem *Experiment* haben wir **Daten** erhoben. Wir wollen auf Basis dieser Daten und mit **statischen Methoden** Erkenntnisse über den **Daten generierenden Mechanismus** gewinnen.

Die folgende **Einteilung** der statistischen Methoden ist üblich und nützlich:

1. **Beschreiben**
2. **Schätzen**
3. **Testen**

Bevor man mit der statistischen Analyse anfängt, muss man sich mit grundlegende Aspekten von Daten, der Datenerhe-

bung und der Beschreibung von Daten beschäftigen.

Daten: Von **statistischen Einheiten** (Merkmalsträger) werden **Merkmale** erhoben/gemessen/beobachtet.[1] Die Merkmale werden dabei anhand einer **Skala** erfasst/gemessen.

*Die Messung einer konkreten Ausprägung eines Merkmals beruht auf einer **Skala**, die die möglichen Merkmalsausprägungen (z.B. Messergebnisse) vorgibt. Eine **Skala** repräsentiert eine Vorschrift, die jeder statistischen Einheit der Stichprobe einen **Beobachtungswert** zuordnet. Dieser Wert gibt die Ausprägung des jeweils interessierenden Merkmals an (vgl. Burkschart [?, S. 10f.]).*

Bei Skalen unterscheidet verschiedene **Skalenniveau**.

nominal Die möglichen Werte sind *nur* Namen; beispielsweise rot, grün, blau, ...

ordinal Die möglichen Werte haben eine Ordnung; Schulnoten (Die Prädikate sehr gut, gut, befriedigend, ...); Umfrage über Grad der Zustimmung zu einer Aussage (Stimme gar nicht zu, Stimme etwas zu, Stimme zu, Stimme ausdrücklich zu)

intervalskaliert Die möglichen Werte des Merk-

[1] Übersicht in Fahrmeier [11, S. 14].

mals können durch Zahlen erfasst werden und die Abstände zwischen diesen Zahlen haben eine feste Bedeutung. Diese Bedeutung ist bis auf eine affine Umrechnung gleich. [Temperatur in Grad Celsius oder in Fahrenheit]

verhältnisskaliert Es gibt für alle Skalen in denen das Merkmal gemessen wird einen gemeinsamen Nullpunkt. Die Quotienten der Zahlen haben die gleiche *Die Existenz eines absoluten Nullpunktes erlaubt Aussagen über Größenverhältnisse. Deshalb spricht man von der "Verhältnisskala". Ist etwa ein Objekt doppelt so lang wie ein zweites Objekt, so bleibt diese Aussage korrekt, egal ob die Längen in Meter, Zentimeter oder in Zoll ausgedrückt werden, da auf diesen Skalen der Wert Null die gleiche Bedeutung besitzt Schuster [**?**, S. 15].* [Länge in cm oder inch]

Daten werden zudem in **diskret** (endlich oder abzählbar unendlich viele Ausprägungen) versus **stetig** (alle Werte eines Intervalls sind mögliche Ausprägungen) eingeteilt.

16.2 Beschreibende Statistik

Je nach Skalierung können **Verfahren** bzw. **Kennziffern** zur **Beschreibung** der Daten verwendet werden.

Häufigkeiten **Absolute Häufigkeiten** und **relative Häufigkeiten** (eventuell nach Gruppierung)

$$H_i = \# i$$
$$h_i = \frac{\# i}{n}$$

Können auch für nominal skalierte Merkmal angegeben werden.

Quantile x_p heißt p-**Quantil**, wenn mindestens der Anteil p der Daten ist kleiner oder gleich x_p **und** mindestens der Anteil $1 - p$ ist größer oder gleich x_p.

Dafür müssen Daten ordinal skaliert sein.

Boxplot In dieser Grafik werden Median, unteres Quartil, oberes Quartil, zwei Whisker[2] und gegebenenfalls Ausreißer eingetragen. Die Daten müssen mindestens ordinal skaliert sein. Der obere Whisker bei der maximale Beobachtung eingetragen, außer das Maximum ist um 1.5 größer als das 75 Prozent Quantil. ...

[2] https://de.wikipedia.org/wiki/Box-Plot

Histogramm Die Daten werden gruppiert (in **Klassen** eingeteilt):

$$[c_0, c_1), [c_1, c_2), ..., [c_{k-1}, c_k).$$

Da das Auge primär die Fläche der Rechtecke bzw. Säulen wahrnimmt, wird das Histogramm so konstruiert, dass die Fläche über den Intervallen gleich oder proportional zu den absoluten bzw. relativen Häufigkeiten ist (vgl. Fahrmeier [11, S. 38]). Die **Klassenbreite** ist $d_j = c_j - c_{j-1}$. Die Höhe der Rechtecke wird dann gleich (oder proportional) zu H_j/d_j bzw. h_j/d_j. Die Fläche ist gleich (oder proportional) zu H_j bzw. h_j.

Mittelwert Die Daten müssen mindestens intervalskaliert sein.

$$\bar{x} = \frac{1}{n}(x_1 + ... + x_n)$$

Empirische Varianz, Standardabweichung, Stichprobenvarianz

Um die Streuung zu messen verwendet man ... (Die Daten müssen dabei mindestens intervalskaliert sein.)

- Empirische Varianz

$$\begin{aligned}\tilde{s}^2 &= \frac{1}{n}\left((x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2\end{aligned}$$

- (empirische) Standardabweichung

$$\tilde{s} = \sqrt{\frac{1}{n}\left((x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2\right)}$$

- Stichprobenvarianz

$$s^2 = \frac{1}{n-1}\left((x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2\right)$$

- Stichprobenstandardabweichung

$$s = \sqrt{\frac{1}{n-1}\left((x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2\right)}$$

- Umrechnung

$$\tilde{s}^2 \frac{n}{n-1} = s^2$$
$$\tilde{s} \cdot \sqrt{\frac{n}{n-1}} = s$$

16.3 Statistische Modelle

Schätzen und Testen sind die Grundpfeiler der sogenannten **schließenden Statistik** (Induktiven Statistik). Auf Basis von Daten soll etwas über den Daten generierenden Prozess gelernt werden.

Der folgende Abschnitt orientiert sich an Henze [19, Abschnitte 7.1, 7.2, 7.4]

16.3.1 Definition: Ein **statistisches Modell** besteht aus

- Dem **Stichprobenraum** $\mathcal{X} \neq \emptyset$. In dieser Menge liegen die denkbaren Ergebnisse der Datenerhebung/Stichprobe.
- Einer σ-Algebra $\mathcal{B}$ auf $\mathcal{X}$; damit wir auf Ereignisses analysieren können.
- Einem **Parameterraum** $\Theta \neq \emptyset$; typischerweise $\Theta \subset \mathbb{R}^d$ (es gibt d Parameter).
- Wahrscheinlichkeitsmaßen $\mathbb{P}_\theta, \theta \in \Theta$, wobei die **Parametrisierung** $\theta \mapsto \mathbb{P}_\theta$ injektiv ist.

16.3.2 Bemerkung: Wir haben also jetzt **viele Wahrscheinlichkeitsmaße** (zur *Auswahl*, die es sein könnten). Wir stellen uns vor, dass wir das Wahrscheinlichkeitsmaß, gemäß dessen die Stichprobe erzeugt wird (die Daten generiert wurden), nicht genau kennen. Wir haben uns mit dem Sachverhalt (dem Daten generierenden Prozess) vertraut gemacht und eine grundsätzliche Vorstellung, welche Art der Verteilung in Frage kommt. Wir kennen aber den/die *wahren Parameter* θ nicht. Wir wollen aus einer beobachteten Stichprobe *etwas* über die $\mathbb{P}_\theta$ lernen. Konkret: Wir wollen

(1) den Parameter θ schätzen und (2) Hypothesen über den Parameter θ testen (prüfen).

16.4 Schätzen

16.4.1 Definition: Gegeben sei ein statistisches Modell $\mathcal{X}, \mathcal{B}, (\mathbb{P}_\theta), \theta \in \Theta, \Theta \subset \mathbb{R}^d$. Ein Schätzer (Punktschätzer) für θ ist eine messbare Abbildung $T : \mathcal{X} \to \mathbb{R}^l$.

Ist $x \in \mathcal{X}$ eine konkrete Stichprobe, dann heißt $\hat{\theta} = T(x)$ Schätzwert für θ zur Beobachtung $\mathbf{x}$.

▷ Wann ist ein Schätzer gut? ◁

16.4.2 Definition: Es sei T ein Schätzer für θ.

i.) $\mathrm{MQA}_T(\theta) = \mathbb{E}_\theta(T-\theta)^2$ heißt **mittlere quadratische Abweichung** von T (an der Stelle θ).

ii.) T heißt **erwartungstreu**, falls für alle $\theta \in \Theta$

$$\mathbb{E}_\theta(T) = \theta$$

iii.) $b_T(\theta) = \mathbb{E}_\theta(T) - \theta$ heißt **Verzerrung** von T.

16.4.3 Satz: Es gilt

$$\mathrm{MQA}_T(\theta) = \mathbb{V}_\theta(T) + b_T(\theta)^2$$

16.4.4 Definition: Es sei $\alpha \in (0,1)$. Ein Abbildung

$$C : \mathcal{X} \to \mathcal{P}(\mathbb{R}^l) \tag{16.1}$$

heißt **Konfidenzbereich** für $\gamma(\theta)$ zur Vertrauenswahrscheinlichkeit (Konfidenzwahrscheinlichkeit) α, falls

$$\mathbb{P}_\theta(\{x \in \mathcal{X} | C(x) \ni \theta\}) \geq 1 - \alpha$$

16.5 Testen

Regelmäßig ist es so, dass es einen Wert $\{\theta_0\}$ im Parameterraum bzw. einen Bereich $\Theta_0 \subset \Theta$ des Parameterraum gibt, der von spezifischen Interesse ist. Beispielsweise: Die Münze ist fair, wenn $p = \frac{1}{2}$ ist. Das Medikament ist unwirksam, wenn $\mu = 0$ gilt. Statistische Tests sind wie folgt ausgelegt: **Ist die konkrete Stichprobe gut oder schlecht mit der Hypothese vereinbar, dass der wahre Parameter in Θ_0 liegt.**

16.5.1 Definition: Gegeben sei ein statistisches Modell $\mathcal{X}, \mathcal{B}, (\mathbb{P}_\theta), \theta \in \Theta, \Theta \subset \mathbb{R}^d$. Für den Parameterraum betrachten wir eine Zer-

legung

$$\Theta = \Theta_0 \uplus \Theta_1$$

wobei $\Theta_0, \Theta_1 \neq \emptyset$.

Das **Testproblem** besteht darin, auf Grundlage einer konkreten Stichprobe x zwischen den Alternativen $\theta \in \Theta_0$ und $\theta \in \Theta_1$ zu entscheiden.

Ein statistischer Test etabliert also eine Regel, die für jede Stichprobe $x \in \mathcal{X}$ angibt, ob man sich für

$$\text{Hypothese } H_0 : \quad \theta \in \Theta_0$$

oder für

$$\text{Hypothese } H_1 : \quad \theta \in \Theta_1$$

entscheiden soll. Oft nennt man H_0 auch die **Nullhypothese** und H_1 die **Alternativhypothese**.

Es gibt bei einem Test eine messbare Menge $\mathcal{K} \subset \mathcal{X}$, so dass für alle Stichproben $x \in \mathcal{K}$ die **Nullhypothese abgelehnt** wird. Die Menge $\mathcal{K}$ heißt **kritischer Bereich (Ablehnungsbereich)**. **Fällt die Stichprobe also in den Ablehnungsbereich, dann wird die Nullhypothese verworfen/abgelehnt (nullify H_0)**.

Fällt die Stichprobe hingegen x nicht in $\mathcal{K}$ – ist also $x \notin \mathcal{K}$ –, dann entscheidet man sich für die Nullhypothese.

Fehler sind möglich!

Gilt $\theta \in \Theta_0$, ist aber $x \in \mathcal{K}$, dann liegt eine **Fehler erster Art** vor. Man begeht also einen Fehler erster Art, wenn H_0 **fälschlicherweise ablehnt**. Tatsächlich gilt $\theta \in \Theta_0$, aber die Beobachtung liegt im Ablehnungsbereich.

Gilt $\theta \in \Theta_1$, aber $x \notin \mathcal{K}$, dann liegt eine **Fehler zweiter Art** vor. Man begeht also einen Fehler zweiter Art, wenn man fälschlicherweise H_0 **nicht verwirft**.

Wirkungstabelle/Konfusionstabelle

	in Wirklichkeit ist	
Entscheidung ↓	$\theta \in H_0$	$\theta \in H_1$
für H_0	richtige Entscheidung	Fehler zweiter Art
für H_1	Fehler erster Art	richtige Entscheidung

Regelmäßig kann man den kritischen Bereich durch eine **Testfunktion** $T : \mathcal{X} \to \mathbb{R}$ angeben:

$$\{T \geq c\} = \{x \in \mathcal{X} \,|\, T(x) \geq c\} = \mathcal{K}.$$

Die Konstante c heißt **kritischer Wert**. Wenn also die Testfunktion ausgewertet mit der Beobachtung einen Wert größer

als c annimmt, dann lehnt man die Nullhypothese ab. **Die Logik: Der beobachtete Wert von T ist zu groß, als dass die beobachtete Stichprobe mit der Nullhypothese vereinbar sein könnte.**

Am besten wäre es, man könnte beide Fehler vermeiden; mindestens kontrollieren. Das geht im Allgemeinen nicht. **Typischerweise kontrolliert man den Fehler erster Art.** Man legt im Vorhinein fest, wie hoch der Fehler erster Art höchsten sein soll. Man spricht dann **von einem Test zum Niveau α**. In der konkreten Anwendung muss die Asymmetrie sachlogisch begründet werden und der Fehler zweiter Art sollte untersucht werden.

Wir definieren dazu zunächst die **Gütefunktion**:

$$g_{\mathcal{K}}(\theta) = \mathbb{P}_\theta(X \in \mathcal{K}).$$

$g_{\mathcal{K}}(\theta)$ ist also die Wahrscheinlichkeit, dass die Stichprobe in den kritischen Bereich fällt, wenn der wahre Parameter θ ist. Die **ideale Gütefunktion** ist die, die 0 auf Θ_0 und 1 auf Θ_1. Das ist aber nie der Fall.

Wir **kontrollieren den Fehler erster Art**, in dem wir dafür sorgen, dass die Gütefunktion auf Θ_0 jedenfalls nicht grö-

ßer als α ist: Gilt

$$g_{\mathcal{K}}(\theta) \leq \alpha \text{ für alle } \theta \in \Theta_0$$

dann spricht man von **einen Test zum Niveau** α. Wenn man sich auf solche Test beschränkt, dann wird man über die Zeit, die Nullhypothese allenfalls in $\alpha \cdot 100$ Prozent verwerfen, obwohl sie gültig ist.

In **Statistikprogrammen** – insbesondere in R – wird regelmäßig der sogenannten p**-Wert** verwendet, um dem Nutzer die Entscheidung über Ablehnung und Nicht-Ablehnung zu erleichtern. Der p-Wert wird für gegebene Beobachtung x so ermittelt: Man berechnet $T(x)$ und löst die Gleichung

$$p = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T > T(x))$$

Man lehnt dann die Nullhypothese ab, wenn der p-Wert kleiner-gleich α ist. Wenn man so vorgeht, dann testet man zum Niveau α (typischerweise $= 0.05$).

16.5.2 Beispiel: Wenn X und Y normal verteilt sind und die gleiche Varianz haben, dann kann man für den Test mit der Nullhypothese $\mathbb{E}(X) - \mathbb{E}(Y)$ die Testfunktion

$$T_{m,n} = \frac{\sqrt{\frac{mn}{m+n}}\left(\bar{X}_m - \bar{Y}_n - (\mathbb{E}(X) - \mathbb{E}(Y))\right)}{S_{m,n}}.$$

verwenden. Die $T_{m,n}$ ist dann t_{m+n-2}-verteilt. Man benötigt dann nur noch das 2.5 Prozent Quantil; die t_{m+n-2}-Verteilung ist symmetrisch.

17 Regression

17.1 Querschnittsregression

17.1.1 Definitionen und Bemerkungen: Wir betrachten einen $\mathbb{R}^{K+1}$ Zufallsvektor

$$\curlyvee = (Y, X_1, ..., X_K) \sim F.$$

Dieser Zufallsvektor gehört zu **einer typischen Beobachtungseinheit** (eine Person, ein Haushalt, ein Unternehmen, eine Immobilieneinheit, vieles ist denkbar .....). Später werden wir Zufallsstichproben aus einer Grundgesamtheit ziehen: Für $i = 1, ..., N$ Beobachtungseinheiten ermitteln wir $\curlyvee_i$. Jeder Zug wird dabei durch einen Zufallsvektor $\curlyvee_i$ repräsentiert, die identisch und unabhängig verteilt sein sollen, d.h. $\curlyvee_i \sim uiF$.

Y heißt **Label, Ergebnisvariable, Zielvariable, Regressand oder erklärte Variable**.

$X_1, ..., X_K$ heißen **Featurevariablen, Features, Regressoren oder erklärende Variablen**.

Wir interessieren uns besonders für den **bedingten Erwartungswert** von Y gegeben $\mathbf{X}$

$$\begin{aligned} m(x_1, ..., x_K) &= \mathbb{E}(Y|X_1 = x_1, ..., X_K = x_K) \\ &= \mathbb{E}(Y|X_1, ..., X_K)(x_1, ..., x_K) \end{aligned}$$

Im betrachteten Zusammenhang heißt $m(x_1, ..., x_K)$ **Regressionsfunktion**.

In der Wahrscheinlichkeitstheorie lernt man, dass der bedingte Erwartungswert die beste **Prognose** von Y ist, wenn man die Werte von $\mathbf{X}$ beobachtet.[1]

Wir betrachten in diesem Skript **lineare Regressionsfunktionen bzw. lineare Regressionsmodelle**[2]:

$$\mathbb{E}(Y|X_1, ..., X_K) = \gamma_0 + \gamma_1 X_1 + ... + \gamma_K X_K.$$

Auch für den Zufallsvektor $\mathcal{Y}$ selbst betrachten wir **lineare Modelle** (für den Daten generierenden Prozess (DGP))

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K + u. \tag{17.1}$$

[1] Das ist salopp formuliert. Details bitte in einer geeigneten Quellen nachsehen; z.B. in Henze [19, Seite 175, Seite 107] oder im WT2 Skript.

[2] Linear bezüglich der Parameter.

mit den **Koeffizienten** $\beta_0, \beta_1, ..., \beta_K$ und **Restterm** u.

Wir betrachten insbesondere (später) lineare Modelle bei den die Annahme der **schwachen Exogenität (mean independence, unconfoundedness)** erfüllt ist:

$$\mathbb{E}(u|X_1, ..., X_K) = 0. \tag{17.2}$$

17.1.2 Bemerkungen: Wenn (17.1) und (17.2) gilt, dann gilt

$$\begin{aligned}\mathbb{E}(Y|X_1, ..., X_K) &= \mathbb{E}(\beta_0 + \beta_1 X_1 + ... + \beta_K X_K + u|X_1, ..., X_K) \\ &= \beta_0 + \beta_1 X_1 + ... + \beta_K X_K + \mathbb{E}(u|X_1, ..., X_K) \\ &= \beta_0 + \beta_1 X_1 + ... + \beta_K X_K.\end{aligned}$$

Die bedingte Erwartung ist also linear, wenn (17.1) und (17.2) gelten. Zudem sind die Koeffizienten der Regressionsfunktion die des linearen Modells (17.1).

Andererseits: Angenommen die bedingte Erwartung ist linear:

$$\mathbb{E}(Y|X_1, ..., X_K) = \gamma_0 + \gamma_1 X_1 + ... + \gamma_K X_K.$$

Wir definieren $e = Y - \gamma_0 - \gamma_1 X_1 - ... - \gamma_K X_K$. Dann gilt

$$\begin{aligned}\mathbb{E}(e|X_1, ..., X_K) &= \mathbb{E}(Y - \gamma_0 - \gamma_1 X_1 - ... - \gamma_K X_K | X_1, ..., X_K) \\ &= \mathbb{E}(Y|X_1, ..., X_K) - \gamma_0 - \gamma_1 X_1 - ... - \gamma_K X_K \\ &= 0\end{aligned}$$

Das bedeutet jedoch nicht, dass ein konsistenter Schätzer der Parameter γ_k auch die Parameter β_k des linearen Modells schätzt. Das ist so, wenn $\mathbb{E}(u|X_1, ..., X_N) = 0$. Die Fehler e *verhalten* scheinbar wie die Reste u. Der Schein kann trügen.

17.1.3 Bemerkungen: Es sei Y eine Zufallsvariable und $\mathbf{X}$ ein $(K+1)$-dimensionaler Zufallsvektor mit

$$\begin{gathered}\mathbb{E}(Y^2) < \infty \\ \mathbb{E}(\mathbf{X}^2) < \infty \\ \mathbb{E}(\mathbf{X}\mathbf{X}^T) \text{ positiv definit.}\end{gathered}$$

Wir wollen die Features in $\mathbf{X}$ nutzen, um den Wert von Y zu prognostizieren. Wir wollen dafür eine **lineare Abbildung** $g(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$ verwenden. Der erste Eintrag von $\mathbf{X}$ ist dabei $X_0 = 1$ und $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_K)^T$.

Wir betrachten den **erwarteten quadratischen Fehler**

$$S(\boldsymbol{\beta}) = \mathbb{E}([Y - \mathbf{X}^T\boldsymbol{\beta}]^2)$$

und bestimmen das *beste* $\boldsymbol{\beta}$, also:

$$\boldsymbol{\beta}^p = \text{argmin}_{\boldsymbol{\beta}} S(\boldsymbol{\beta}).$$

Damit erhalten wir eine Prognoseabbildung $\mathcal{P}(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}^p$, die sogenannte **Projektion der Zufallsvariable Y auf den von den Features aufgespannten linearen Raum**:

$$\mathcal{P}(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}^p.$$

Wir können $\boldsymbol{\beta}^p$ sogar explizit angeben[3]

$$\boldsymbol{\beta}^p = \left(\mathbb{E}(\mathbf{X}\mathbf{X}^T)\right)^{-1}\mathbb{E}(\mathbf{X}Y).$$

Der *kleinste* Fehler ist

$$e^p = Y - \mathbf{X}^T\boldsymbol{\beta}^p.$$

Es gilt

$$\mathbb{E}(X_i e^p) = 0\,,\ i = 0, ..., K.$$

e^p und X_i sind also zueinander **orthogonal**.[4] Da $X_0 = 1$ gilt, gilt zudem

$$\mathbb{E}(e^p) = 0.$$

[3] Und der Nachweis ist nicht zu schwer.

[4] Der Begriff orthogonal ist angemessen, denn durch $\langle X, Y\rangle = \mathbb{E}(X, Y)$ definiert ein (quasi-Skalarprodukt.)

Dann folgt

$$\text{cov}(e^p, X_i) = \mathbb{E}(e^p X_i) - \mathbb{E}(e^p)\mathbb{E}(X_i) = 0$$

e^p und X_i sind also unkorreliert.

Fazit: Die Abbildung

$$\mathcal{P}(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}^p.$$

ist die beste lineare Approximation von Y, wenn wir (nur) die Information aus $\mathbf{X}$ haben bzw. nutzen. Wenn $\mathbb{E}(Y|\mathbf{X})$ linear ist, dann gilt $\mathbb{E}(Y|\mathbf{X}) = \mathcal{P}(Y|\mathbf{X})$.

Die Parameter $\boldsymbol{\beta}^p$ haben im Allgemeinen mindestens erst mal *keine strukturelle/kausale* Bedeutung. $\boldsymbol{\beta}^p$ ist der Parametervektor, der die beste lineare Approximation **des statistischen Zusammenhangs (der Assoziation)** darstellt. Wenn man nur **prognostizieren** will, dann ist das angemessen; jedenfalls wenn die Linearität akzeptabel ist. Wenn ein sachbezogener struktureller Zusammenhang (z.B. ein kausaler Zusammenhang) untersucht werden soll, dann benötigt man weitere Annahmen, die wir im nächsten Abschnitt erläutern.

▶ Bisher haben wir nur über die Grundgesamtheit bzw. den Daten generierenden Prozess gesprochen, aber noch nicht über Daten. Damit beginnen wir jetzt. Wir stellen uns Daten

als **Stichprobe** aus der Grundgesamtheit vor.

17.1.4 Definitionen und Bemerkungen: Bisher haben wir **abstrakt eine prototypische Beobachtungseinheit** betrachtet (einen Zufallsvektor). Wir stellen uns für die **statistische** Analyse vor, dass N **Einheiten** gezogen oder **beobachtet** wurden:

$$\begin{aligned}\Upsilon_1 &= (Y_1, X_{11}, ..., X_{1K})\\ \Upsilon_2 &= (Y_2, X_{21}, ..., X_{2K})\\ &...\\ \Upsilon_N &= (Y_N, X_{N1}, ..., X_{NK})\end{aligned}$$

Wir haben somit N Beobachtungen und sprechen von einer **Stichprobe** der Größe N. Jede Zeile **der Stichprobe** entspricht einer Instanz des Zufallsexperiments, das für eine **Beobachtungseinheit** analysiert wird.

Vorsicht: Vorne hatten wir die Feature in einer Spalte **X** zusammen gefasst. Jetzt stecken die Feature zur einer Bobachtungseinheit in einer Zeile.

Wir benötigen für die statistische Analyse Annahmen über die gemeinsame Verteilung *Zeilen* Υ_i. Wir betrachten zunächst den Standardfall für die **Querschnittsanalyse**.

Die Stichprobe heißt **Zufallsstichprobe**, falls $\Upsilon_1, ..., \Upsilon_N$ **iden-**

tisch und unabhängig verteilt sind.

17.1.5 Definitionen und Bemerkungen: Wir betrachten ein lineares Regressionsmodell

$$\mathbb{E}(Y|X_1, ..., X_K) = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K$$

und eine Stichprobe der Größe N

$$\begin{aligned} \mathcal{Y}_1 &= (Y_1, X_{11}, ..., X_{1K}) \\ \mathcal{Y}_2 &= (Y_2, X_{21}, ..., X_{2K}) \\ &... \\ \mathcal{Y}_N &= (Y_N, X_{N1}, ..., X_{NK}). \end{aligned}$$

Wir **kennen die Koeffizienten** $\beta_0, \beta_1, ..., \beta_K$ **nicht** und wollen diese auf Basis der Stichprobe **schätzen**. An dieser Stelle treffen wir keine Annahme über u und die β's haben keine **strukturelle** Bedeutung. Wir suchen (nur) nach einen quantitativen Erfassung der Assoziation zwischen den Features und dem Label.

Wir betrachten für irgendwelche Koeffizienten $\tilde{\beta}_0, ..., \tilde{\beta}_K$ die **Fehler**

$$\tilde{e}_i = Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{i1} - ... - \tilde{\beta}_K X_{iK}, \quad i = 1, ..., N$$

Wir schätzen die Koeffizienten so, dass der *Gesamtfehler* mi-

nimal ist:

$$(\hat{\beta}_0, ..., \hat{\beta}_K) = \arg \min_{\tilde{\beta}_0,...,\tilde{\beta}_K} \sum_{i=1}^{N} (Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{i1} - ... - \tilde{\beta}_K X_{iK})^2$$

$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K)^T$ heißt **kleinste Quadrate Schätzer** der Koeffizienten $\boldsymbol{\beta}$.

17.1.6 Definition und Satz: Gegeben sei eine Stichprobe der Größe N. Wir definieren die **Regressormatrix**

$$\mathbb{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & .... & X_{1K} \\ 1 & X_{21} & X_{22} & .... & X_{2K} \\ ... & & & & \\ 1 & X_{N1} & X_{N2} & .... & X_{NK} \end{pmatrix}.$$

Die erste Spalte erfasst in Gleichungen mit Matrizen nahezu immer den Achsenabschnitt. Die erste Zufallsvariable ist dann $\mathbb{X}_{\bullet,0} \equiv 1$ und **die erste Spalte hat nur 1'en.** Wenn man mit Matrizen arbeitet, dann erfasst der Koeffizient β_0 den Achsenabschnitt.

Erinnerung **X** ist ein **Zufallsvektor** und Vektoren sind üblicherweise Spalten. $\mathbb{X}$ ist eine **Matrix** in der **jede Zeile** zu den Feature einer Beobachtungseinheit gehört, die wir vorher als Spalte aufgeschrieben wurden.

verwirrende Notation!

Wenn die Matrix $\mathbb{X}^T\mathbb{X}$ invertierbar ist, dann können wir den

dann eindeutig bestimmten **kleinste Quadrate Schätzer** kompakt in der Form

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\frac{1}{N}\mathbb{X}^T\mathbb{X}\right)^{-1}\frac{1}{N}\mathbb{X}^T\mathbb{Y}\\ &= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}\end{aligned}$$

schreiben, wobei

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}.$$

17.1.7 Bemerkung Wir beschäftigen uns mit der kleinste Quadrate Schätzung

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\frac{1}{N}\mathbb{X}^T\mathbb{X}\right)^{-1}\frac{1}{N}\mathbb{X}^T\mathbb{Y}\\ &= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}.\end{aligned}$$

In der Matrix $\mathbb{X}^T\mathbb{X}$ *stecken* die **empirischen zweiten Mo-**

mente der Features (aber ohne $\frac{1}{N}$):

$$\mathbb{X}^T\mathbb{X} = \begin{pmatrix} X_{10} & X_{20} & .... & X_{N0} \\ X_{11} & X_{21} & .... & X_{N1} \\ ... & & & \\ X_{1K} & X_{2K} & .... & X_{NK} \end{pmatrix} \begin{pmatrix} X_{10} & X_{11} & X_{12} & .... & X_{1K} \\ X_{20} & X_{21} & X_{22} & .... & X_{2K} \\ ... & & & & \\ X_{N0} & X_{N1} & X_{N2} & .... & X_{NK} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^N X_{i0}X_{i0} & ... & \sum_{i=1}^N X_{i0}X_{iK} \\ & ... & \\ \sum_{i=1}^N X_{iK}X_{i0} & ... & \sum_{i=1}^N X_{iK}X_{iK} \end{pmatrix}$$

$$= \sum_{i=1}^N \mathbf{X}_i\mathbf{X}_i^T$$

Etwas anschaulicher als empirische Momente

$$\frac{1}{N}\mathbb{X}^T\mathbb{X} = \frac{1}{N}\sum_{i=1}^N \mathbf{X}_i\mathbf{X}_i^T.$$

17.1.8 Satz: Wir betrachten ein **lineares Regressionsmodell**, d.h. die Regressionsfunktion (die bedingte Erwartungsfunktion) sein linear. Also es gelte

$$\mathbb{E}(Y|X_1, ..., X_K) = \beta_0 + \beta_1 X_1 + ...\beta_K X_K.$$

Gegeben sei zudem eine **Zufallsstichprobe** und die Matrix $\mathbb{X}^T\mathbb{X}$ sei invertierbar.

Dann ist der kleinste Quadrate Schätzer **unverzerrt**, d.h.

$$\mathbb{E}(\hat{\boldsymbol{\beta}}|\mathbb{X}) = \boldsymbol{\beta}.$$

Beweis: Vgl. Hansen [17, Abschnitt 4.5]. Der KQS ist

$$\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}$$

Aus der Annahme der Linearität der Erwartung $\mathbb{E}(Y|X_1,...,X_K) = \beta_0+\beta_1X_1+...\beta_KX_K$ folgt $\mathbb{E}[\mathbb{Y}|\mathbb{X}] = \mathbb{X}\boldsymbol{\beta}$. Damit folgt weiter

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}|\mathbb{X}) &= \mathbb{E}\left[\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}|\mathbb{X}\right]\\ &= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{E}[\mathbb{Y}|\mathbb{X}]\\ &= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{X}\boldsymbol{\beta}\\ &= \boldsymbol{\beta}.\end{aligned}$$

17.1.9 Satz: Wir betrachten ein **lineares Regressionsmodell**, d.h. die Regressionsfunktion (die bedingte Erwartungsfunktion) sei linear. Es gelte also

$$\mathbb{E}(Y|X_1,...,X_K) = \beta_0 + \beta_1X_1 + ...\beta_KX_K,$$

wobei für die zweiten Momente

$$\begin{aligned}\mathbb{E}(Y^2) &< \infty\\ \mathbb{E}(X_k^2) &< \infty, k = 1,...,K\end{aligned}$$

gelten soll. Ferner sei die Matrix der theoretischen zweiten Momente der X-Variablen (der Merkmale)

$$\mathbf{Q}_{XX} = \mathbb{E}\left(\begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_K \end{pmatrix}\Big(X_1, X_2, ..., X_K\Big)\right) = \mathbb{E}(XX^T)$$

$$= \mathbb{E}\begin{pmatrix} X_1 \cdot X_1 & X_1 \cdot X_2 & ... & X_1 \cdot X_K \\ X_2 \cdot X_1 & X_2 \cdot X_2 & ... & X_2 \cdot X_K \\ ... & & & \\ X_K \cdot X_1 & X_K \cdot X_2 & ... & X_K \cdot X_K \end{pmatrix}$$

positiv definit (also insbesondere invertierbar).

Durch die Gleichung

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_K X_K + e$$

sei der *Fehler* e definiert und es sei $\sigma^2 = \mathbb{V}(e|X_1, ..., X_K)$.

Gegeben sei zudem eine **Zufallsstichprobe** und die Matrix $\mathbb{X}^T\mathbb{X}$ sei invertierbar.

Für die Varianz der kleinste Quadrate Schätzung gilt

$$\mathbb{V}(\hat{\boldsymbol{\beta}}|\mathbb{X}) = (\mathbb{X}^T\mathbb{X})^{-1}(\mathbb{X}^T\mathbf{D}\mathbb{X})(\mathbb{X}^T\mathbb{X})^{-1}$$

mit

$$\mathbf{D} = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0... \\ 0 & 0 & ... & \sigma_N^2 \end{pmatrix} = \mathbb{V}(\mathbb{e}|\mathbb{X}).$$

Beweis: Es gilt

$$\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}.$$

Dann erhalten wir

$$\begin{aligned} \mathbb{V}(\hat{\boldsymbol{\beta}}|\mathbb{X}) &= \mathbb{V}(\hat{\boldsymbol{\beta}}|\mathbb{X}) \\ &= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{V}(\mathbb{Y}|\mathbb{X})\left(\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\right)^T \\ &= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{V}(\mathbb{Y}|\mathbb{X})\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}. \end{aligned}$$

Wir beobachten

$$\mathbb{V}(\mathbb{Y}|\mathbb{X}) = \mathbb{V}(\mathbb{X}\boldsymbol{\beta} + \mathbb{e}|\mathbb{X}) = \mathbb{V}(\mathbb{e}|\mathbb{X})$$

Zusammen

$$\mathbb{V}(\hat{\boldsymbol{\beta}}|\mathbb{X}) = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\,\mathbb{V}(\mathbb{e}|\mathbb{X})\,\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$$

17.1.10 Bemerkung: Die Formel

$$\mathbb{V}(\hat{\boldsymbol{\beta}}|\mathbb{X}) = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\,\mathbb{V}(\mathbb{e}|\mathbb{X})\,\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$$

sieht kompliziert aus; oft findet man einfachere Formeln für

den Fall der Homoskedastizität (alle σ_i^2 sind gleich). Die Formel erfasst aber auch den Fall der (bedingten) Heteroskedastizität.

17.1.11 Bemerkung: Das vorherigen Ergebnisse sind sehr bemerkenswert. Typischerweise benötigen wir aber nicht nur einen Schätzer, sondern wir wollen diesen auch *beurteilen*; beispielsweise mittels eines Hypothesentests. Die Varianz $\mathbb{V}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ wird dabei erst dann richtig nützlich, wenn wir mehr über die Verteilung von $\hat{\boldsymbol{\beta}}$ ermitteln. In der Literatur wird dann oft der Fall untersucht, dass die Fehler Normal verteilt sind. Leider ist das in Anwendungen in der Regel unangemessen. Alternativ kann man **asymptotische Eigenschaften** (also für N groß) von $\hat{\boldsymbol{\beta}}$ ermitteln. Diese Resultate kann man auf große Datensätze approximativ anwenden. Mit dem Zentralen Grenzwertsatzes ergibt sich in der Tat Normalität. Dieser Weg wird im folgenden Satz eingeschlagen.

Es ist in konkreten Anwendung oft nicht klar, ob der Datensatz groß genug ist oder ob ein sogenannter finite sample bias droht. Zu diesen Thema gibt es zahlreiche wissenschaftliche Quellen, die insbesondere Monte Carlo Methoden verwenden.

17.1.12 Satz: i.) Wir betrachten für $N \in \mathbb{N}$ eine Zufallsstichprobe $(\Upsilon_1, ..., \Upsilon_N)^T$ für ein lineares Regressionsmodell mit dem kleinste quadrate Schätzer $\hat{\boldsymbol{\beta}}^N = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$. Für die

zweiten Momente gelte

$$\mathbb{E}(Y^2) < \infty$$
$$\mathbb{E}(X_k^2) < \infty, k = 1, ..., K$$

Dann gilt

$$\text{n.W.-}\lim \hat{\boldsymbol{\beta}}^N = \boldsymbol{\beta}.$$

Der KQS ist also konsistent.

Gilt sogar

$$\mathbb{E}(Y^4) < \infty, \mathbb{E}(X_k^4) < \infty,$$

dann gilt

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}^N - \boldsymbol{\beta}) \xrightarrow{\text{i.V.}} N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}}),$$

wobei

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{XX}^{-1}\, \boldsymbol{\Omega}\, \boldsymbol{Q}_{XX}^{-1},$$

und

$$\mathbf{Q}_{XX} = \mathbb{E}(XX^T)$$
$$\boldsymbol{\Omega} = \mathbb{E}(XX^T e^2).$$

Der kleinste quadrate Schätzer $\hat{\boldsymbol{\beta}}^N$ ist konsistent und asym-

ptotisch normal. Vorsicht: $\mathbf{V}_{\boldsymbol{\beta}}$ ist nicht der Grenzwert von $\mathbb{V}(\hat{\boldsymbol{\beta}}^N)$! Unten wird dazu noch mal Stellung genommen.

ii.) Für die Inferenz (Signifikanz-Test) müssen wir die Varianz von $\hat{\boldsymbol{\beta}}^N$ schätzen! Es seien

$$\widehat{\mathbf{Q}}_{XX} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^T,\ \widehat{\mathbf{Q}}_{XY} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i Y_i$$

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^T \hat{e}_i^2 \ .$$

Der Schätzer

$$\widehat{\mathbf{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}} = \widehat{\mathbf{Q}}_{XX}^{-1}\, \widehat{\boldsymbol{\Omega}}\, \widehat{\mathbf{Q}}_{XX}^{-1}$$

für $\mathbf{V}_{\boldsymbol{\beta}}$ ist konsistent:

$$\widehat{\mathbf{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}} \xrightarrow{\text{n.W.}} \mathbf{V}_{\boldsymbol{\beta}} \quad .$$

und die Varianz des KQS können wir schließlich so konsistent schätzen:

$$\widehat{\mathbb{V}}(\hat{\boldsymbol{\beta}}^N) = \frac{1}{N}\widehat{\mathbf{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}}$$

Beweis: Vgl. Hansen [17, Kapitel 7].

17.1.13 Bemerkung: Unter den oben genannten Voraus-

setzungen gilt

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}}^N - \boldsymbol{\beta}) \xrightarrow{\text{i.V.}} \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}}).$$

$\sqrt{N}\,(\hat{\boldsymbol{\beta}}^N - \boldsymbol{\beta})$ ist näherungsweise $\mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$. Dann ist $\hat{\boldsymbol{\beta}}^N$ näherungsweise

$$\hat{\boldsymbol{\beta}}^N \dot{\sim} \mathcal{N}\left(\boldsymbol{\beta}, \frac{1}{N}\mathbf{V}_{\boldsymbol{\beta}}\right).$$

In der Tat

$$\mathbf{V}_{\boldsymbol{\beta}} \approx \mathbb{V}\left(\sqrt{N}\,(\hat{\boldsymbol{\beta}}^N - \boldsymbol{\beta})\right) = \mathbb{V}\left(\sqrt{N}\,\hat{\boldsymbol{\beta}}^N\right) = N\,\mathbb{V}(\hat{\boldsymbol{\beta}}^N).$$

Also

$$\mathbb{V}(\hat{\boldsymbol{\beta}}^N) \approx \frac{1}{N}\mathbf{V}_{\boldsymbol{\beta}}.$$

Vorsicht, hier droht eine Verwechselung: $\mathbf{V}_{\boldsymbol{\beta}}$ ist nicht die Varianz $\mathbb{V}(\hat{\boldsymbol{\beta}}^N)$ des Schätzers $\hat{\boldsymbol{\beta}}^N$, sondern $\frac{1}{N}\mathbf{V}_{\boldsymbol{\beta}}$. Vgl. Hansen [17, Abschnitt 7.8] für eine Übersicht zur Notation im Zusammenhang mit den Varianz-Kovarianz-Matrizen

17.1.14 Bemerkungen: Die Markierung HC zeigt an, dass **Heteroskedastizität** *erlaubt* ist. Es gibt eine Reihe von Schätzern für $\mathbf{V}_{\boldsymbol{\beta}}$, die trotz Heteroskedastizität konsistent sind; vgl. Hansen [17, Abschnitt 7.9].

17.1.15 Bemerkungen (Fazit für die Anwendung): Wenn die **Zufallsstichprobe** (einigermaßen) groß ist und Linearität für die bedingte Erwartung gilt, dann können wir die Regressionsfunktion mit KQ schätzen. Der Schätzer ist unter den genannten Voraussetzungen erwartungstreu und konsistent. Man erhält auch eine Grundlagen für die Inferenzanalyse (t-Statistik, p-Werte, Standardfehler), denn jedenfalls für große N ist der KQS normal.

Bezogen auf Linearität: Es gibt Tests für eine etwaige funktionale Fehlspezifikationen (Nicht-Linearität); vgl. z.B. Verbeek [?]. Durch Hinzunahme von Regressoren z.B. der Form $X_k^2, X_1X_2, ...$ lassen sich manche Nicht-Linearitäten in den X erfassen. Üblich sind in Zusammenhang mit Nicht-Linearität auch Transformationen (insbesondere log).

17.2 Kausalität

▶ Quellen: Stock und Watson [?]. Wasserman [?].

17.2.1 Bemerkung: Regelmäßig sind wir an **kausalen Zusammenhängen** interessiert, die wir empirisch untersuchen wollen. Wir müssen solche Fragestellungen von Prognosen unterscheiden, bei denen der Parametervektor der Projektion *reicht*.

Wir wollen beispielsweise wissen, ob (und um wieviel) ein besseres Betreuungsverhältnis (Schulklassengröße) in Schulen die Leistungen der Schüler kausal verbessert. Das ist kompliziert, denn die Leistungen sind von vielen Faktoren (individuelle Faktoren, familiärer Hintergrund, sonstige Schulausstattung, ...)abhängig, von denen etliche nicht mal (gut) gemessen werden können.

Oder, wir wollen wissen, welchen Effekt eine Preiserhöhung auf die Nachfrage nach dem Gut hat (z.B weil eine Steuer erhoben werden soll, um den Zuckerkonsum zu reduzieren). Auch das ist kompliziert, denn die beobachteten Preis-Menge-Kombination entstehen aus Nachfrage und Angebot. Wir wollen aber die Elastizität der Nachfrage ermitteln.

17.2.2 Definition: Wir betrachten ein lineares Modell[5]

$$Y = \beta_0 + \alpha X + u,$$

mit $\mathbb{E}(u) = 0$. Wir wollen den kausalen Effekt von X auf Y schätzen, d.h. wir wollen den Parameter α schätzen.

[5]Konvention: Wenn der Parameter einen kausalen bzw. strukturellen Effekt erfassen soll, dann verwenden wir den Buchstaben α. Wenn der Koeffizient (nur) zur bedingten Erwartung gehört, dann verwenden wir den Buchstaben ein β.

Wir definieren die **Potential Outcome Variablen**

$$Y(1) = \beta_0 + \alpha \cdot 1 + u$$
$$Y(0) = \beta_0 + \alpha \cdot 0 + u.$$

Wir stellen uns also **hypothetisch** vor, dass entweder alle Einheiten der Grundgesamtheit $X = 1$ (treatment) oder alle $X = 0$ (non-treatment) erhalten.

Dann heißt

$$\theta = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0))$$

der **mittlere kausale Effekt** von X auf Y. Wenn wir den wahren Parameter α kennen würden, dann wäre $\theta = \alpha$.

Realiter werden wir **eine der beiden** Variablen $Y(1)$ oder $Y(0)$ beobachten **und die andere nicht**! Wenn für eine konkrete Beobachtungseinheit z.B. $X = 1$ ist, dann beobachten wir $Y(1)$ und $Y(0)$ bleibt für diese Beobachtungseinheit unbeobachtet.

17.2.3 Bemerkung: Bei den echten Daten wird X stochastisch *zugeteilt* bzw. die Probanden entscheiden selbst über treatment und non-treatment. Wir beobachten für die gezogene Beobachtungseinheiten nur eine der beide Zufallsvariablen $Y(1)$ bzw. $Y(0)$. Den **beobachteten** Label bezeichnen wir wie gehabt mit Y. Ebenfalls wie gehabt betrachten wir

ein lineares Regressionsmodell

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$$

Wir betrachten

$$\gamma = \mathbb{E}(Y|X_1 = 1) - \mathbb{E}(Y|X_1 = 0)$$

γ heißt **mittlerer empirischer Effekt**. Diesen Effekt können wir *wie gehabt* schätzen und es gilt $\gamma = \beta_1$. Wir hoffen $\alpha = \gamma$. Im allgemeinen gilt jedoch (vgl. Wasserman [**?**, Seite 253])

$$\alpha \neq \gamma$$

Wie können wir sicherstellen, dass $\alpha = \gamma$ gilt?

Es gilt (vgl. Wasserman [**?**, Seite 254]): Wenn $(Y(0), Y(1)) \perp\!\!\!\perp X$, dann $\alpha = \gamma$. Die Annahme $(Y(0), Y(1)) \perp\!\!\!\perp X$ heißt **Random Assignment**. Dieses Resultat wollen wir auf Regressonsanalyse übertragen. Dazu betrachten wir lineare Modelle und die Voraussetzung der schwachen Exogenität.

17.2.4 Bemerkungen: Der beschriebene Ansatz (Potential Outcome) geht auf Rubin[6] zurück, der Ideen von Neyman aufgegriffen hat. Wasserman [**?**, Kapitel 16] ist eine sehr gu-

[6] https://en.wikipedia.org/wiki/Rubin_causal_model

te Einstiegsquelle. In Stock und Watson [**?**, Kapitel 13] wird dieser Ansatz im Zusammenhang mit Regressionsanalyse besprochen.

17.2.5 Definitionen und Bemerkungen: Angenommen für den DGP (für die Grundgesamtheit) gilt ein lineares Modell und schwache Exogenität, d.h.

$$Y = \beta_0 + \alpha X + u.$$
$$\mathbb{E}(u|X) = 0.$$

Dann folgt: der mittlere empirische Effekt gleich dem mittleren kausalen Effekt ist:

$$\alpha = \gamma.$$

Beweis: Es gilt $Y = Y(1)X + Y(0)(1-X)$. Wir beobachten

$$\begin{aligned}
\gamma &= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \\
&= \mathbb{E}(Y(1)X + Y(0)(1-X)|X=1) - \mathbb{E}(Y(1)X + Y(0)(1-X)|X=0 \\
&= \mathbb{E}(Y(1)|X=1) - \mathbb{E}(Y(0)|X=0) \\
&= \mathbb{E}(\beta_0 + \alpha \cdot 1 + u|X=1) - \mathbb{E}(\beta_0 + \alpha \cdot 0 + u|X=0) \\
&= \alpha + \mathbb{E}(u|X=1) - \mathbb{E}(u|X_1=0) \\
&= \alpha + 0
\end{aligned}$$

17.2.6 Bemerkung: Wenn

$$Y = \beta_0 + \alpha X + u$$
$$\mathbb{E}(u|X) = 0$$

dann

$$\mathbb{E}(Y|X) = \beta_0 + \alpha X.$$

17.2.7 Bemerkung: Im vorherigen Abschnitt haben wir gesehen, dass man unter passenden Momentenbedingungen die Parameter β_0, α konsistent schätzen kann und die asymptotische Verteilung normal ist.

17.2.8 Definition: Wir betrachten jetzt ein lineares Modell mit sogenannten **Kontrollvariablen** $W_1, ..., W_M$

$$Y = \beta_0 + \alpha_1 X_1 + \gamma_1 W_1 + ... + \gamma_M W_M + u.$$

Wir wollen den kausalen Effekt von X_1 auf Y schätzen, d.h. wir wollen den Parameter α_1 schätzen. Bezogen auf die anderen Effekte/Parameter sind wir *agnostisch*. Deshalb haben wir denen auch einen anderen Buchstaben gegeben.

Wir definieren die **Potential Outcome Variablen**

$$Y(1) = \beta_0 + \alpha_1 \cdot 1 + \gamma_1 W_1 + ... + \gamma_M W_M + u$$
$$Y(0) = \beta_0 + \alpha_1 \cdot 0 + +\gamma_1 W_1 + ... + \gamma_M W_M + u.$$

Wir stellen uns also **hypothetisch** vor, dass entweder alle Einheiten der Grundgesamtheit $X_1 = 1$ (treatment) oder alle $X_1 = 0$ (non-treatment) erhalten. Dann heißt

$$\theta = \mathbb{E}(Y(1)|\mathbf{W}) - \mathbb{E}(Y(0)|\mathbf{W})$$

der **mittlere kausale Effekt** von X_1 auf Y gegeben $\mathbf{W}$. Wenn wir die wahren Parameter kennen würden, dann wäre $\theta = \alpha_1$.

Realiter werden wir **eine der beiden** Variablen $Y(1)$ oder $Y(0)$ beobachten **und die andere nicht**! Wenn für eine konkrete Beobachtungseinheit z.B. $X = 1$ ist, dann beobachten wir $Y(1)$ und $Y(0)$ bleibt für diese Beobachtungseinheit unbeobachtet.

17.2.9 Bemerkung: Bei den echten Daten wird X_1 stochastisch *zugeteilt* bzw. die Probanden entscheiden selbst über treatment und non-treatment. Wir beobachten für den gezogene Beobachtungseinheiten nur eine der beide Zufallsvariablen $Y(1)$ bzw. $Y(0)$. Den beobachteten Label bezeichnen wir wie gehabt mit Y. Ebenfalls wie gehabt betrachten wir

ein lineares Regressionsmodell

$$\mathbb{E}(Y|X_1, ..., X_K) = \beta_0 + \beta_1 X_1 + \gamma_1 W_1 + ... + \gamma_M W_M$$

Wir betrachten

$$\begin{aligned}\gamma = &\mathbb{E}(Y|X_1 = 1, W_1 = w_1, ..., W_M = w_M) \\ &- \mathbb{E}(Y|X_1 = 0, W_1 = w_1, ..., W_M = w_M)\end{aligned}$$

γ heißt **mittlerer empirischer Effekt**, wobei wir für die Effekte von $W_1, ..., W_M$ **kontrollieren**. Diesen Effekt können wir *wie gehabt* schätzen und es gilt $\gamma = \beta_1$. Wir hoffen $\alpha_1 = \gamma$. Im allgemeinen gilt jedoch (vgl. Wasserman [**?**, Seite 253]).

$$\alpha_1 \neq \gamma$$

Wie können wir sicherstellen, dass $\alpha_1 = \gamma$ gilt?

Es gilt: Wenn $(Y(0), Y(1)) \perp\!\!\!\perp X|W_1, ..., W_M$, dann $\alpha_1 = \gamma$. Die Annahme $(Y(0), Y(1)) \perp\!\!\!\perp X|W_1, ..., W_M$ heißt **Conditinal Independence Assumption (CIA)**. Dieses Resultat wollen wir auf Regressonsanalyse übertragen. Dazu betrachten wir lineare Modelle.

17.2.10 Definitionen und Bemerkungen: Angenommen für den DGP (für die Grundgesamtheit) gilt ein lineares Mo-

dell, d.h.

$$Y = \beta_0 + \alpha_1 X_1 + \gamma_1 W_1 + ... + \gamma_M W_M + u.$$

Dann folgt aus der **mean conditional independence**

$$\mathbb{E}(u|X_1, W_1, ..., W_M) = \mathbb{E}(u|W_1, ..., W_M),$$

dass der mittlere empirische Effekt gleich dem mittleren kausalen Effekt ist:

$$\alpha_1 = \gamma.$$

Beweis: Es gilt $Y = Y(1)X_1 + Y(0)(1 - X_1)$.

$$\begin{aligned}
\gamma &= \mathbb{E}(Y|X_1 = 1, W_1 = w_1, ..., W_M = w_M) \\
&\quad - \mathbb{E}(Y|X_1 = 0, W_1 = w_1, ..., W_M = w_M) \\
&= \mathbb{E}(Y(1)X_1 + Y(0)(1 - X_1)|X_1 = 1, W_1 = w_1, ..., W_M = w_M) \\
&\quad - \mathbb{E}(Y(1)X_1 + Y(0)(1 - X_1)|X_1 = 0, W_1 = w_1, ..., W_M = w_M) \\
&= \mathbb{E}(Y(1)|X_1 = 1, W_1 = w_1, ..., W_M = w_M) \\
&\quad - \mathbb{E}(Y(0)|X_1 = 0, W_1 = w_1, ..., W_M = w_M) \\
&= \mathbb{E}(\beta_0 + \alpha_1 \cdot 1 + \gamma_1 W_1 + ... + \gamma_M W_M + u|X_1 = 1, W_1 = w_1, ..., W_M = \\
&\quad - \mathbb{E}(\beta_0 + \alpha_1 \cdot 0 + \gamma_1 W_1 + ... + \gamma_M W_M + u|X_1 = 0, W_1 = w_1, ..., W_M \\
&= \alpha_1 + \mathbb{E}(u|X_1 = 1, W_1 = w_1, ..., W_M = w_M) \\
&\quad - \mathbb{E}(u|X_1 = 0, W_1 = w_1, ..., W_M = w_M) \\
&= \alpha_1 + \mathbb{E}(u|W_1 = w_1, ..., W_M = w_M) \\
&\quad - \mathbb{E}(u|W_1 = w_1, ..., W_M = w_M) \\
&= \alpha_1 + 0
\end{aligned}$$

17.2.11 Bemerkung: Wenn die Zuweisung von X_1 die Annahme der mean conditional indepedence erfüllt, dann ist $\alpha_1 = \gamma$.

17.2.12 Satz: Angenommen für den DGP (für die Grundgesamtheit) gilt ein lineares Modell, d.h.

$$Y = \beta_0 + \alpha_1 X_1 + \gamma_1 W_1 + ... + \gamma_M W_M + u.$$

Es gelte zudem die **mean conditional independence** Bedingung

$$\mathbb{E}(u|X_1, W_1, ..., W_M) = \mathbb{E}(u|W_1, ..., W_M)$$

und für die rechts stehende bedingte Erwartung gelte Linearität

$$\mathbb{E}(u|W_1, ..., W_M) = \gamma_0 + \gamma_1 W_1 + ... + \gamma_M X_M.$$

Dann gilt

$$\mathbb{E}(Y|X_1, X_2, ..., X_K) = \delta_0 + \alpha_1 X_1 + \delta_1 W_1 + ... + \delta_M W_M$$

und

$$\begin{aligned} Y &= \delta_0 + \alpha_1 X_1 + \delta_1 W_1 + ... + \delta_M W_M + \nu \\ 0 &= \mathbb{E}(\nu|X_1, W_1, ..., W_M) \end{aligned}$$

Beweis: Vgl. Stock und Watson [**?**, S. .....]

17.2.13 Bemerkung: Im vorherigen Abschnitt haben wir gesehen, dass man unter passenden Momentenbedingungen die Parameter $\delta_0, \delta_2, ..., \delta_K$ und α_1 konsistent schätzen kann und die asymptotische Verteilung normal ist.

▶ Die Ausführungen dieses Abschnitt erklären, dass

$$\mathbb{E}(u|X) = 0$$

bzw.

$$\mathbb{E}(u|X, \mathbf{W}) = \mathbb{E}(u|\mathbf{W})$$

hinreichende Voraussetzungen für die konsistente Schätzung ist.

Die folgenden Bemerkungen erklären etwas detailliert, was geschieht, wenn diese Bedingungen nicht erfüllt sind. ... und wie man die Bedingungen (noch) interpretieren kann. Letzteres ist nützlich, wenn man in konkreten Zusammenhängen beurteilen muss, ob die Bedingungen erfüllt sind.

17.2.14 Definitionen und Bemerkungen: Angenommen **es gibt einen kausalen/strukturellen Zusammenhang**

$$Y = \gamma_0 + \alpha X + u,$$

wobei wir o.B.d.A $\mathbb{E}(u) = 0$ gilt. Wir betrachten nur ein Feature, um die Notation einfach zu halten. Wir stellen uns vor, dass wir die sachbezogene Bedeutung des Parameters α detailliert abgeleitet und erläutert haben!

Wir wollen den strukturellen Parameter α **schätzen**.

Um die Parameter α zu **schätzen**, benötigen wir Daten. Am besten wäre es, wir könnten ein **kontrolliertes Experiment** durchführen. Wir würden dabei $X = 1$ und $X = 0$ zufällig auf die Beobachtungseinheiten $i = 1, ..., N$ *verteilen.*

Bei **Beobachtungsdaten** wird den Beobachtungseinheiten der Wert von X nicht zufällig zugewiesen. Dementsprechend müssen wir überprüfen/inspizieren, ob $\mathbb{E}(u|X) = 0$ eine Eigenschaft des Daten generierten Prozesses (DGP) ist. Gilt für den DGP tatsächlich $\mathbb{E}(u|X) = 0$, dann ist das *so gut wie die Annahmen zufälligen Zuweisung* in einen kontrollierten Experiment (Stock und Watson [**?**, ...]).

$\mathbb{E}(u|X) = 0$ bedeutet salopp formuliert, dass wir aus den Realisierung X *nix* über den Wert von u lernen.Genau an dieser Stelle wird es relevant das betrachtete Experiment auch **sachlich inhaltlich** zu verstehen. Der Fehlerterm u enthält alle Determinanten von Y, die nicht vom Modellteil $\beta_0 + \alpha X$ erfasst werden. Wenn es z.B. eine Zufallsgröße Z mit $\text{cov}(X, Z) > 0, \text{cov}(u, Z) > 0$ gibt, dann ergibt sich das Problem, dass eine beobachtete gemeinsame Bewegung von Y und X in eine Richtung (die durch den Term αX erfasst werden soll) auf Z zurückgeführt sein könnte. Wenn man diesen Effekt beim **Schätzen** von α nicht herausrechnet (erfasst), dann schätzt man α u.U. zu groß.

Angenommen die Daten werden z.B gemäß

$$\begin{aligned} Y &= \beta_0 + \alpha X + Z + e_2 \\ &= \beta_0 + \alpha X + e_1 \end{aligned}$$

erzeugt, wobei $\text{cov}(X, Z) > 0, \text{cov}(X, e_2) = 0$ ist. Die Zufallsvariable Z sei latent. Sie kann nicht beobachtet werden. Wenn wir Beobachtungen von Y und X plotten und die beste Gerade (mit kleinste Quadrate geschätzt) durch die Daten legen würden, dann **würden wir nicht die Steigung** α erhalten, **sondern** $\beta_1 = \alpha + (\mathbb{E}(XX^T))^{-1}\mathbb{E}(Xe_1)$. Wenn X um eine Einheit steigt und Z konstant bliebe, dann würde Y um α steigen (das ist c.p.-Effekt, den wir strukturell erfassen wollen). In **Beobachtungsdaten** Daten steigt aber wegen $\text{cov}(X, Z) > 0$ mit X auch Z an. Deshalb steigt Y um mehr als α_1; nämlich um β_1.

Wenn man die Parameter mit der Methode der kleinsten Quadrate schätzt, dann ist der Schätzer $\hat{\alpha}$ inkonsistent, denn es gilt

$$\hat{\beta}_1 \to \alpha + (\mathbb{E}(XX^T))^{-1}\mathbb{E}(Xe_1).$$

17.2.15 Beispiel: Der Parameter, der uns interessiert sei β_1

des folgenden Modells.

$$\begin{aligned} Y &= \beta_0 + \alpha X + Z + e_1 \\ Y &= \beta_0 + \alpha X + u \\ Z &= \gamma X + e_2 \end{aligned}$$

Wenn $\gamma \neq 0$, dann gilt $\mathbb{E}(u|X) = \gamma X \neq 0$. Welche Geschichte könnte dahinterstecken? Wir wollen den Einfluss von X auf Y erfassen. Der DGP hat die Eigenschaft: wenn mehr X gemacht wird, dann wird auch mehr Z gemacht. Wenn X steigt, dann steigt Y. Das ist aber nicht nur wegen des höheren X so, sondern mit X erhöht sich auch Z und es gilt einen *Begleiteffekt* auf Y.

Wenn wir Z *einsetzen*:

$$\begin{aligned} Y &= \beta_0 + \alpha X + \gamma X + e_2 + e_1 \\ &= \beta_0 + \beta_1 X + e_2 + e_1 \\ &= \beta_0 + \beta_1 X + \tilde{u} \end{aligned}$$

und es gilt $\mathbb{E}(\tilde{u}|X) = 0$. Jetzt haben wir ein lineares Modell mit $\mathbb{E}(\tilde{u}|X) = 0$, aber nicht für den Parameter α bzw. den strukturellen Effekt, der uns eigentlich interessiert. Wir haben ein Als-ob-Modell für den Effekt von X einschließlich des Begleiteffekts durch Z.

17.3 Endogenität

Bei einen kontrollierten Experiment sorgt die zufällige Zuweisung/Verteilung für

$$\mathbb{E}(u|X_1, ..., X_K) = 0.$$

Dann gilt auch $\text{cov}(u, X_k) = 0$. Bei Beobachtungsdaten werden wir regelmäßig vermuten, dass es Effekte mit $\text{cov}(u, X_k) \neq 0$ gibt.

17.3.1 Bemerkungen: Wenn

$$\text{cov}(u, X_k) \neq 0$$

gilt, dann können wir die Parameter β_k mit KQS nicht konsistent schätzen. Wir über- oder unterschätzen den Zusammenhang zwischen Y und X_k. Wenn wir *nur* an einer Prognose interessiert sind, dann ist das eher unerheblich. Wenn wir uns für strukturelle Zusammenhänge interessieren, dann haben wir im Allgemeinen ein veritables Problem!

Gilt

$$\text{cov}(u, X_k) \neq 0,$$

dann sprechen wir von **Endogenität**.

17.3.2 Bemerkungen: Siehe Hansen [17, S. 341f]. Angenommen wir betrachten ein lineares Modell

$$Y = X^T\boldsymbol{\beta} + u$$

aber[7]

$$\mathbb{E}(Xu) \neq 0$$

Wenn wir die beste lineare Approximation, d.h. die lineare Projektion bestimmten, dann verschwindet das Problem *scheinbar*. Dabei betrachtet man[8]

$$\begin{aligned} Y &= X^T\boldsymbol{\beta}^p + e, \\ \mathbb{E}(Xe) &= 0. \end{aligned}$$

Weiterhin gilt

$$\begin{aligned} \boldsymbol{\beta}^p &= \left(\mathbb{E}(XX^T)\right)^{-1}\mathbb{E}(XY) \\ &= \left(\mathbb{E}(XX^T)\right)^{-1}\mathbb{E}(XX^T\boldsymbol{\beta} + Xu) \\ &= \boldsymbol{\beta} + \left(\mathbb{E}(XX^T)\right)^{-1}\mathbb{E}(Xu) \neq \boldsymbol{\beta} \end{aligned}$$

Der kleinste Quadrate Schätzer konvergiert gegen den Projektionskoeffizienten

$$\hat{\boldsymbol{\beta}} \xrightarrow{\text{n.W.}} \boldsymbol{\beta}^p \neq \boldsymbol{\beta}$$

[7] Dann gilt auch $\mathbb{E}(u|X_1, ..., X_K) = 0$ nicht.

[8] $\boldsymbol{\beta}^p$ ist der Koeffizient der Projektion.

Wenn wir lediglich an der linearen Projektion interessiert sind (z.B. für lineare Prognosen bzw. für eine quantitative Erfassung der Assoziation), dann ist Nicht-Exogenität kein Problem. Wenn wir aber an einem strukturellen Parameter interessiert sind – weil wir z.B. eine finanzwirtschaftliche Hypothese testen wollen – dann müssen wir das Problem der Endogenität adressieren!

17.3.3 Bemerkungen: Es gibt verschiedene Ursachen für Endogenität.

- **Vergessene Regressoren.** Vgl. Stock und Watson [**?**, S. 211]. Das Problem haben wir vorn mit Z erläutert. Eine Lösung haben wir ebenfalls schon erläutert. Wir kontrollieren in dem wir weitere Feature X_k bzw. Kontrollvariablen W_i berücksichtigen.

- Die *zwei* Zufallsvariablen Y und X beeinflussen sich über zwei Kanäle (z.B. **Angebot und Nachfrage**) gegenseitig **endogen** und wir interessieren uns nicht nur für den resultierenden Zusammenhang, sondern für einen der beiden Wege (z.B. für die Nachfragekurve).

- **Messfehler.** Ein Regressor wird mit einen Fehler gemessen. Vgl. Hansen [17, S. 342]. Gerade dieses Problem ist in der Finanzmarktökonometrie typischerweise akut. Volatilitäten sind latent, Marktindizes sind nur Proxys,

β's sind geschätzt.

17.3.4 Bemerkungen: Lösungsansätze sind vor allem die folgenden drei.

- **Kontrollvariablen.** Für weitere Effekte kontrollieren, die mutmaßlich für die Korrelation verantwortlich sind. Also mehr Regressoren berücksichtigen und insbesondere bivariate Regressionen mit sehr großer Skepsis betrachten (eigentlich für die strukturelle Analysen irrelevant).

- **Instrumente.** Angenommen wir haben Daten zu einer Zufallsvariable Z mit $\text{cov}(Z, u) = 0$ und $\text{cov}(Z, X) \neq 0$ (Stock und Watson [**?**, S. 428ff]). Solche Variablen heißen **Instrumente**. Jetzt gehen wir in **zwei Stufen** vor. Zunächst schätzen wir auf der ersten Stufe die Regressionsgleichung

$$X = \gamma_0 + \gamma_1 Z + e.$$

Auf der zweiten Stufe betrachten wir die Regression

$$Y = \beta_0 + \alpha \hat{X} + u.$$

wobei $\hat{X} = \gamma_0 + \gamma_1 Z$.

Wir beobachten

$$\begin{aligned}\mathrm{cov}(Z,Y) &= \mathrm{cov}(Z, \beta_0 + \beta_1 X + u) \\ &= \beta_1 \mathrm{cov}(Z,X) + \mathrm{cov}(Z,u)\end{aligned}$$

Also

$$\beta_1 = \frac{\mathrm{cov}(Z,Y)}{\mathrm{cov}(Z,X)}.$$

- Warum funktioniert diese Methode? Die Regression auf der ersten Stufe **filtert** aus X den Teil $\hat{X}$ heraus, **der nicht mit u korreliert ist**.

- Es gilt (vgl. Stock und Watson [**?**, Seite 434, 467])

$$\hat{\beta}_1^{\mathrm{2TSLS}} = \frac{s_{ZY}}{s_{ZX}}$$

- **Panel.** Eine sehr gute (auch sehr gut nachvollziehbare) Darstellung der Panel-Regression ist Stock und Watson [**?**, S. 361ff].

  Fragmentarisch: ʒ

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + \tilde{u}_{it}$$

  mit $\mathrm{cov}(a,x) \neq 0$ und $\mathrm{cov}(x,\tilde{u}) = 0$. Die Endogenität entsteht hier also wegen der Kovarianz zwischen a

und x. Dabei ist a_i eine **zeitlich invariante Querschnittseigenschaft der Beobachtungseinheiten**. Dann betrachten man die zeitliche Differenz

$$\begin{aligned} y_{i,t} - y_{i,t-1} &= \beta_1(x_{i,t} - x_{i,t-1}) + \tilde{u}_{it} - \tilde{u}_{i,t-1} \\ &= \beta_1(x_{i,t} - x_{i,t-1}) + u_{it} \end{aligned}$$

Jetzt gilt $\text{cov}(u, x) = 0$. Allerdings gilt $\text{cov}(u_{i,t}, u_{i,t-1}) \neq 0$. Man ist das Problem wegen der Endogenität los, hat sich aber ein Problem bezüglich der intertemporalen Korrelation der Fehler eingehandelt. Wie man damit umgeht ist aber gut bekannt.

17.4 Verallgemeinerte Momentenmethode (GMM)

Quelle: Verbeek [**?**, Seite 175 ff] und Hansen [17].

17.4.1 Bemerkungen: In der **Finanzmarktökonomik** ist das folgende Optimierungsproblem[9] **generisch**:

$$\max \mathbb{E}_t \left(\sum_{s=1,\ldots,S} \delta^s u(C_{t+s}) \right)$$

$$C_{t+s} + V_{t+s} = w_{t+s} + (1 + r_{t+s})V_{t+s-1}$$

[9] Es handelt sich um ein intertemporales Optimierungsproblem. .....

$\mathbb{E}_t(\cdot)$ bezeichnet den bedingten Erwartungswert. C bezeichnet den Konsum, V das Anlagevermögen, r die Rendite und w das Arbeitseinkommen. Der Parameter δ erfasst den Nutzen-Diskontfaktor und $u(\cdot)$ den (Perioden-)Nutzen aus Konsum.

Eine mögliche Wahl für u ist

$$u(C) = \frac{C^{1-\gamma}}{1-\gamma}$$

Die Bedingung erster Ordnung kann man wie folgt angeben

$$\mathbb{E}_t \left(\delta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} (1 + r_{t+1}) - 1 \right) = 0.$$

Man spricht von einer **Momenten-Bedingung**.

Wenn $z_t \in \mathcal{F}_t$ ist, dann gilt

$$\mathbb{E}_t \left[\left(\delta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} (1 + r_{t+1}) - 1 \right) z_t \right] = 0.$$

Man bezeichnet solche Variable in diesem Kontext als **Instrumente**. Wir erhalten also für jedes Instrument eine weitere Momenten-Bedingung.

Die verallgemeinerte Momenten Methode (GMM) passt genau zu dieser Modellkonstellation. Mit der GMM können wir insbesondere die Parameter δ und γ schätzen und Hypothesen testen.

17.4.2 Bemerkungen: GMM betrachtet zu den theoretischen Momenten-Bedingungen

$$\mathbb{E}[f(x_t, z_t, \theta)] = 0$$

die empirischen Momenten-Funktionen

$$g_T(\theta) := \frac{1}{T} \sum_{t=1}^{T} f(x_t, z_t, \theta).$$

Es sei M die Anzahl der Parameter und K die Anzahl der Instrumente (die Anzahl der Momenten-Bedingungen).

Wenn $K = M$ gilt, dann spricht man von just-identified und löst die Gleichungen

$$g_T(\theta) := \frac{1}{T} \sum_{t=1}^{T} f(x_t, z_t, \theta) = 0.$$

In diesem Fall ist es eine Momenten-Methode.

Gilt hingegen $M > K$, dann betrachten man das Minimierungsproblem

$$\min_{\theta} Q_T(\theta) = \min_{\theta} (g_T(\theta))^T \, \mathbf{W}_T \, g_T(\theta)$$

Dabei ist $\mathbf{W}_T$ eine (geeignet zu wählende) positiv definite Gewichtungsmatrix. Die Lösung $\hat{\theta}$ dieses Minimierungsproblem heißt GMM-Schätzer der Parameter θ.

Der GMM-Schätzer ist unter geeigneten Regularitätsbedingugnen konsistent und asymptotisch normal. Je nach Wahl der Gewichtungsmatrix $\mathbf{W}_T$ ergibt sich ein anderer GMM-Schätzer. Die *optimale* Wahl der Gewichungsmatrix $\mathbf{W}_T^{\text{opt}}$ *minimiert* die Kovarianzmatrix des GMM-Schätzers.

Ohne Autokorrelation z.B. gilt

$$\mathbf{W}_T^{\text{opt}} = \left(\mathbb{E}[f(x_t, z_t, \theta)(f(x_t, z_t, \theta))^T]\right)^{-1}$$

Es ist naheliegend die optimale Gewichtungsmatrix so zu schätzen

$$\left(\frac{1}{T}\sum_{t=1}^{T} f(x_t, z_t, \hat{\theta}_{[1]})(f(x_t, z_t, \hat{\theta}_{[1]}))^T\right)^{-1}$$

Dabei ist $\hat{\theta}_{[1]}$ irgendein konsistenter GMM-Schätzer (z.B. mit $\mathbf{W}_T = \mathbf{E}$).

17.5 Ridge Regression

Quelle: [18, Seite 237ff] und [18, Seite 33ff]

17.5.1 Bemerkungen: Die folgende Zerlegung ist fundamental für das Maschine Learning. Wir betrachten einen Test-Datum (y_0, x_0) mit Label y_0 und Features $\mathbf{x}_0$. Es sei $\hat{f}_{\mathcal{T}}$ die Prediction-Funktioin auf Basis der zufälligen Trainingsmenge

$\mathcal{T}$. Für den Test-MSE ermitteln wir die Zerlegung

$$\begin{aligned}\mathbb{E}\left(y_0 - \hat{f}_{\mathcal{T}}(\mathbf{x}_0)\right)^2 &= \mathbb{V}(y_0 - \hat{f}_{\mathcal{T}}(\mathbf{x}_0)) + \left(\mathbb{E}(y_0 - \hat{f}_{\mathcal{T}}(\mathbf{x}_0))\right)^2 + \mathbb{V}(\varepsilon) \\ &= \mathbb{V}(\hat{f}_{\mathcal{T}}(\mathbf{x}_0)) + \left(\mathbb{E}(y_0 - \hat{f}_{\mathcal{T}}(\mathbf{x}_0))\right)^2 + \mathbb{V}(\varepsilon) \\ &= \mathbb{V}(\hat{f}_{\mathcal{T}}(\mathbf{x}_0)) + \left(\mathrm{Bias}(\hat{f}_{\mathcal{T}}(\mathbf{x}_0))\right)^2 + \mathbb{V}(\varepsilon).\end{aligned}$$

$\mathbb{V}(\varepsilon)$ ist der unvermeidliche Fehler. Diesen Fehler können wir auch durch eine kluge Wahl von $\hat{f}_{\mathcal{T}}$ nicht vermeiden. Bei den beiden Termen $\mathbb{V}(\hat{f}_{\mathcal{T}}(\mathbf{x}_0))$, $\left(\mathrm{Bias}(\hat{f}_{\mathcal{T}}(\mathbf{x}_0))\right)^2$ ergibt sich der sogenannte **Bias-Variance Trade-off**.

Wenn man die **Flexibilität** von $\hat{f}_{\mathcal{T}}$ erhöhen, dann verringert das typischerweise die Verzerrung und erhöht die Varianz. Warum? Bei einer höheren Flexibilität erhalten wir eine bessere Anpassung an die Daten. Allerdings riskieren wir, dass das Modell auch an Noise angepasst wird. Wenn man dann ein neues Datum betrachtet, dann ergibt sich ein *großer* out-of-sample Fehler (denn der Noise ist anders). Die höhere Flexibilität führt zu **over-fitting**.

17.5.2 Bemerkungen: Bei der KQ-Methode minimiert man

$$\mathrm{RSS} = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{k=1}^{K}\beta_k x_{ik}\right)^2.$$

Bei der Ridge-Regression (vgl. James et al. [18, Seite 237])

minimiert man

$$\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{k=1}^{K}\beta_k x_{ik}\right)^2+\lambda\sum_{k=1}^{K}\beta_k^2.$$

$\lambda > 0$ nennt man Tuning Parameter. Der Term $\lambda\sum_{k=1}^{K}\beta_k^2$ heißt Shrinkage Strafe.

Warum funktioniert die Ridge-Regression (gegebenenfalls)? Wenn λ größer ist, dann verringert das die Flexibilität des Prediction-Modells. Dadurch sinkt die Varianz. Wenn die Varianz stärker sinkt als der Bias steigt, dann erhält einen kleineren Test-MSE.

17.6 Quantile Regression

17.6.1 Satz: Für die reellwertige ZV Y gelte $\mathbb{E}(Y^2) < \infty$. Die folgenden Aussagen sind äquivalent:

- Für die reelle Zahl d^* gilt

$$\mathbb{E}((Y-d^*)^2)=\inf_{-\infty<d<\infty}\mathbb{E}((Y-d)^2)$$

- $d^*=\mathbb{E}(Y)$.

- Anschaulich: Der Erwartungswert minimiert den durchschnittlichen quadrierten Abstand zu den Daten.

17.6.2 Satz: Für die reellwertige ZV Y gelte $\mathbb{E}(|Y|) < \infty$. Die folgenden Aussagen sind äquivalent:

- Für die reelle Zahl d^* gilt

$$\mathbb{E}(|Y - d^*|) = \inf_{-\infty<d<\infty} \mathbb{E}(|Y - d|)$$

- d^* ist ein Median von Y.

17.6.3 Definition: Für $\tau \in (0, 1)$ und $Y \sim F$ heißt jede reelle Zahl q_τ mit

$$F(u) \geq \tau \text{ und } F(u-) \leq \tau$$

ein τ-Quantil von Y. Für $\tau \in (0, 1)$ definieren wir die Funktion $\rho_\tau : \mathbb{R} \to \mathbb{R}$ wie folgt

$$\rho_\tau(x) = \begin{cases} -x(1-\tau) & \text{falls } x < 0 \\ x\tau & \text{sonst} \end{cases}$$

17.6.4 Satz: Für die reellwertige ZV Y gelte $\mathbb{E}(|Y|) < \infty$. Die folgenden Aussagen sind äquivalent:

- Für die reelle Zahl d^* gilt

$$\mathbb{E}(\rho_\tau(Y - d^*)) = \inf_{-\infty<d<\infty} \mathbb{E}(\rho_\tau(Y - d))$$

- d^* ist ein τ-Quantil von Y.

17.6.5 Least absolute Deviation – LAD

- Wir betrachten das folgende Modell

$$\begin{aligned}(y_i, \mathbf{x}_i) &\sim \text{iid} \\ y_i &= \mathbf{x}_i^T \boldsymbol{\beta}_{1/2} + \varepsilon_i \\ \text{med}(\varepsilon_i | X = \mathbf{x}_i) &= 0\end{aligned}$$

- Dann ist $\text{med}(y_i | X = \mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_{1/2}$ der bedingte Median.

- Der LAD Schätzer von $\boldsymbol{\beta}_{1/2}$ ist

$$\sum_{k=1}^{N} |y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{1/2}| = \min_{\boldsymbol{\beta}} \sum_{k=1}^{N} |y_i - \mathbf{x}_i^T \boldsymbol{\beta}|.$$

17.6.6 Quantil Regression

- Wir betrachten das folgende Modell

$$(y_i, \mathbf{x}_i) \sim \text{iid} \tag{17.3}$$

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_\tau + \varepsilon_i \tag{17.4}$$

$$Q_\tau(\varepsilon_i | X = \mathbf{x}_i) = 0 \tag{17.5}$$

- Dann ist $Q_\tau(y_i | X = \mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_\tau$ das bedingte Quantil.

- Der QR Schätzer von $\boldsymbol{\beta}_\tau$ ist

$$\sum_{k=1}^{N} \rho_\tau(y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_\tau) = \min_{\boldsymbol{\beta}} \sum_{k=1}^{N} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}).$$

17.6.7 Quantil Regression

- **Satz:** Für das Modell (17.3) – (17.5) gelte

  - $\mathbb{E}(||\mathbf{x}_i||^2) < \infty$.
  - Die bedingten Fehler $\varepsilon_i|\mathbf{x}_i$ haben eine Lebesgue-Dichte $f(\varepsilon_i|\mathbf{x}_i)$.
  - $\mathbb{E}(f(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T)$ ist positive definit.

  Dann ist der QR Schätze $\widehat{\boldsymbol{\beta}}_\tau$ asymptotisch normal-verteilt mit

$$\sqrt{n} \cdot (\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \to N(\mathbf{0}, \mathbf{V}_\tau) \text{ i.V.} \tag{17.6}$$

$$\mathbf{V}_\tau = \tau(1-\tau)(\mathbb{E}(f(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T))^{-1}\, \mathbb{E}\left(\mathbf{x}_i\mathbf{x}_i^T\right)\, \mathbb{E}(f(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T)^{-1} \tag{17.7}$$

17.6.8 V@R mit QR

- Wir betrachten QR-Modelle für die Rendite r_i mit ge-

eigneten Prediktoren

$$\begin{aligned}(r_t, \mathbf{x}_t) &\sim \text{iid} \\ r_t &= \mathbf{x}_t^T \boldsymbol{\beta}_\tau + \varepsilon_t \\ Q_\tau(\varepsilon_t | X = \mathbf{x}_t) &= 0 \\ Q_\tau(y_t | X = \mathbf{x}_t) &= \mathbf{x}_t^T \boldsymbol{\beta}_\tau\end{aligned}$$

- Schätze QR Modell. Dann

$$\widehat{\text{V@R}_{\tau,t^*}(V)} = -(\mathbf{x}_{t^*}^T \widehat{\boldsymbol{\beta}}_\tau) V_{t^*-1}$$

18 Zustandsraummodelle und Kalman-Filter

18.1 Zustandsraummodelle

18.1.1 Definition: Ein Zustandsraummodell modelliert zwei stochastische Prozesse $(\mathbf{x}_t), (\mathbf{y}_t), t \geq 0$ und besteht aus zwei Gleichungen. Letztlich ergibt sich ein Modell für den stochastischen Prozess $(\mathbf{y}_t), t \geq 0$. Vorgelagert wird ein Modell für den Prozess $(\mathbf{x}_t), t \geq 0$.

Die erste Gleichung (**Zustandsgleichung**) beschreibt die **Dynamik** der **Zustandsvariabeln** $\mathbf{x}_t$. Der stochastische Prozess $\mathbf{x}_t$ der Dimension $K \in \mathbb{N}$ genügt der Zustandsgleichung

$$\mathbf{x}_t = \mathbf{a} + \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\varepsilon}_t, \ \boldsymbol{\varepsilon}_t \sim \text{i.i.N}(0, \mathbb{I}).$$

$\mathbf{a}$ ist ein Vektor der Dimension K. $\mathbf{A}, \mathbf{\Sigma_x}$ sind $K \times K$ Matrizen. K ist die Dimension des Zustandsraums. Die Zustände $\mathbf{x}_t$ – und natürlich die Störungen $\boldsymbol{\varepsilon}_t$ – sind i.A. latent, d.h. sie werden i.A. nicht *direkt* beobachtet. Typischerweise wird ein Startwert $\mathbf{x}_0 \in \mathbb{R}^K$ oder eine Startverteilung $\mathbf{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \mathbf{Q}_0)$ unterstellt.

Die zweite Gleichung (**Beobachtungsgleichung**) verbindet die Zustände $\mathbf{x}_t$ und die beobachteten Variablen $\mathbf{y}_t$:

$$\mathbf{y}_t = \mathbf{b} + \mathbf{B}\mathbf{x}_t + \mathbf{D}\mathbf{d}_t + \mathbf{\Sigma_y}\boldsymbol{\nu}_t, \ \boldsymbol{\nu}_t \sim \mathrm{i.i.N}(0, \mathbb{I}).$$

$\mathbf{a}$ ist ein Vektor der Dimension N, $\mathbf{B}$ ist eine $N \times K$ Matrix und $\mathbf{\Sigma_y}$ eine $N \times N$ Matrix. $\mathbf{y}_t$ ist ein stochastischer Prozess der Dimension N; N nennen wir die Dimension des Beobachtungsraum.

$\mathbf{d}_t \in \mathbb{R}^m$ enthält nicht-stochastische von Zeit anhängige Variablen, die z.B. deterministische Trends erfassen. $\mathbf{D}$ ist eine $N \times m$ Matrix.

Ferner unterstellen wir, dass die Zustandsinnovationen $\boldsymbol{\varepsilon}_t$ und die Beobachtungsfehler $\boldsymbol{\nu}_t$ unkorreliert sind:

$$\mathbb{E}(\boldsymbol{\nu}_t \boldsymbol{\epsilon}_s^T) = 0.$$

Außerdem sind die Zustandsinnovationen und Beobachtungs-

fehler unkorreliert zum Startwert

$$\mathbb{E}(\boldsymbol{\nu}_t \mathbf{x}_0^T) = 0,$$
$$\mathbb{E}(\boldsymbol{\epsilon}_t \mathbf{x}_0^T) = 0.$$

Zusammengefasst lautet das Zustandsraummodell also:

$$\mathbf{x}_t = \mathbf{a} + \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\Sigma}_\mathbf{x}\boldsymbol{\varepsilon}_t \qquad \boldsymbol{\varepsilon}_t \sim \text{i.i.N}(0, \mathbb{I})$$
$$\mathbf{y}_t = \mathbf{b} + \mathbf{B}\mathbf{x}_t + \mathbf{D}\mathbf{d}_t + \boldsymbol{\Sigma}_\mathbf{y}\boldsymbol{\nu}_t \qquad \boldsymbol{\nu}_t \sim \text{i.i.N}(0, \mathbb{I})$$

wobei

$$\mathbb{E}(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_s^T) = \delta_{t,s}\mathbb{I}$$
$$\mathbb{E}(\boldsymbol{\nu}_t \boldsymbol{\nu}_s^T) = \delta_{t,s}\mathbb{I}$$
$$\mathbb{E}(\boldsymbol{\nu}_t \boldsymbol{\epsilon}_s^T) = 0$$
$$\mathbb{E}(\boldsymbol{\nu}_t \mathbf{x}_0^T) = 0$$
$$\mathbb{E}(\boldsymbol{\epsilon}_t \mathbf{x}_0^T) = 0.$$

Die Parameter des Zustandsraummodell sind $\mathbf{a}, \mathbf{A}, \mathbf{b}, \mathbf{B}, \boldsymbol{\Sigma}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{y}$. Der Startwert $\mathbf{x}_0$ und die Störterme $\boldsymbol{\epsilon}_t, \boldsymbol{\nu}_t$ sind **exogene Werte** (bzw. die **Inputs** oder **Eingaben**). Die (unbeobachteten) Zustände ergeben sich aus diesen Inputs und den Systemgleichungen. Die Beobachtungen ergeben sich schließlich affine aus den Zuständen und gestört durch die Beobachtungsfehler.

18.2 Kalman-Filter

18.2.1 Definition: Wir definieren:

$$\begin{aligned}
\mathbf{x}_{t|s} &:= \mathbb{E}(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_s) \\
\mathbb{V}_{\mathbf{x}}(t|s) &:= \mathbb{V}(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_s) \\
\mathbf{y}_{t|s} &:= \mathbb{E}(\mathbf{y}_t|\mathbf{y}_1, ..., \mathbf{y}_s) \\
\mathbb{V}_{\mathbf{y}}(t|s) &:= \mathbb{V}(\mathbf{y}_t|\mathbf{y}_1, ..., \mathbf{y}_s)
\end{aligned}$$

18.2.2 Satz (Kalman-Filter & Kalman-Glätter): Sind die Startwerte $\mathbf{x}_{0|0}$ und $\mathbb{V}_{\mathbf{x}}(0|0)$ gegeben, dann gelten für die bedingten Erwartungswerte bzw. Varianzen die Rekursionsgleichungen:

Prognoseschritt

$$\begin{aligned}
\mathbf{x}_{t|t-1} &= \mathbf{A}\mathbf{x}_{t-1|t-1} \\
\mathbb{V}_{\mathbf{x}}(t|t-1) &= \mathbf{A}\mathbb{V}_{\mathbf{x}}(t-1|t-1)\mathbf{A}^T + \mathbf{Q}_{\mathbf{x}} \\
\mathbf{y}_{t|t-1} &= \mathbf{b} + \mathbf{B}\mathbf{x}_{t|t-1} + \mathbf{D}\mathbf{d}_t \\
\mathbb{V}_{\mathbf{y}}(t|t-1) &= \mathbf{B}\mathbb{V}_{\mathbf{x}}(t|t-1)\mathbf{B}^T + \mathbf{R}
\end{aligned}$$

Korrekturschritt

$$\begin{aligned}
\mathbf{x}_{t|t} &= \mathbf{x}_{t|t-1} + \mathbf{G}_t(\mathbf{y}_t - \mathbf{y}_{t|t-1}) \\
\mathbb{V}_{\mathbf{x}}(t|t) &= \mathbb{V}_{\mathbf{x}}(t|t-1) - \mathbf{G}_t\mathbb{V}_{\mathbf{y}}(t|t-1)\mathbf{G}_t^T \\
\mathbf{G}_t &= \mathbb{V}_{\mathbf{x}}(t|t-1)\mathbf{B}^T\mathbb{V}_{\mathbf{y}}(t|t-1)^{-1}
\end{aligned}$$

Glättung

$$\begin{aligned} x_{t|T} &= x_{t|t} + \mathbf{S}_t(x_{t+1|T} - x_{t+1|t}) \\ \mathbb{V}_{\mathbf{x}}(t|T) &= \mathbb{V}_{\mathbf{x}}(t|t) - \mathbf{S}_t(\mathbb{V}_{\mathbf{x}}(t+1|t) - \mathbb{V}_{\mathbf{x}}(t+1|T))\mathbf{S}_t^T \\ \mathbf{S}_t &= \mathbb{V}_{\mathbf{x}}(t|t)\mathbf{A}^T\mathbb{V}_{\mathbf{x}}(t+1|t)^{-1} \end{aligned}$$

18.2.3 Satz (Likelihood):

$$\begin{aligned} \log \mathcal{L} = &-\frac{NT}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{T}\log|\mathbb{V}_{\mathbf{y}}(t|t-1)| \\ &- \frac{1}{2}\sum_{j=1}^{T}(\mathbf{y}_t - \mathbf{y}_{t|t-1})^T\mathbb{V}_{\mathbf{y}}(t|t-1)^{-1}(\mathbf{y}_t - \mathbf{y}_{t|t-1}) \end{aligned}$$

wobei $\mathbf{y}_{1|0} = \mathbb{E}(\mathbf{y}_1), \mathbb{V}_{\mathbf{y}}(1|0) = \mathbb{V}(\mathbf{y}_1)$. Alle Werte – bis auf die Startwerte – werden im Kalman-Filter berechnet.

Die Beweise der Rekursionen beruhen auf dem folgenden Satz. Die Beweise werden insbesondere in Lütkepohl [26] angegeben.

19 Einfache Irrfahrten

Quellen: Feller [10], Grimmett & Stirzaker [14], [15], Henze [?], Steele [34], Schilling [?, Kap.12], Privault [?]

19.1 Einführung

19.1.1 Defintion: i.) Es sei $(X_i)_{i\in\mathbb{N}}$ eine Folge von unabhängigen Bernoulli Zufallsvariablen[1], mit

$$\mathbb{P}(X_i = 1) = p > 0$$
$$\mathbb{P}(X_i = -1) = q = 1 - p > 0$$

[1] Typischerweise in der WT haben Bernoulli ZV das Bild $\{0, 1\}$. In Stoch Proz wählen wir Bild$(X) = \{1, -1\}$. Warum?

Ferner sei $S_0 = a$ (der Anfangswert, oft 0). Der durch

$$S_n = S_0 + \sum_{i=1}^{n} X_i \quad (n = 1, 2, 3, ...)$$
$$= a + \sum_{i=1}^{n} X_i \quad (n = 1, 2, 3, ...)$$

definierte stochastische Prozess heißt **einfache Irrfahrt** mit Anfangswert a.

ii.) Gilt $p = 1/2$, dann heißt $(S_n)_{n\in\mathbb{N}_0}$ **einfache symmetrische Irrfahrt**.

19.1.2 Bemerkung: i.) Man kann sich die Irrfahrt als zufällige Bewegung auf $\mathbb{Z}$ vorstellen. Zufällig bewegt sich der *betrunkene Punkt* nach links oder rechts. Machen Sie eine Skizze!

ii.) Zur Illustration betrachten wir auch die Punkte $(n, S_n) \in \mathbb{R}^2, n \in \mathbb{N}_0$. Für jedes $\omega \in \Omega$ erhalten wir einen Pfad $(n, S_n(\omega)) \in \mathbb{N}_0 \times \mathbb{Z}, n \in \mathbb{N}_0$.

n zählt die Anzahl der Schritte (die (horizontale) **Länge des Pfads**). S_n erfasst die vertikale Abweichung.

Schreiben Sie ein Python Skript oder ein R Skript, das symmetrische Irrfahrten simuliert und graphisch darstellt.

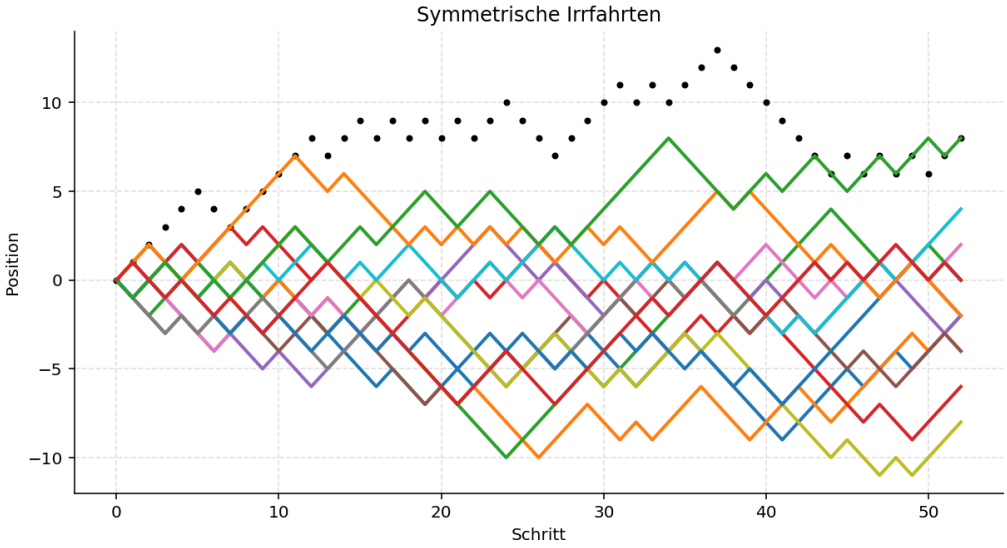


Abbildung 19.1: Pfade der symmetrischen Irrfahrt mit 52 Schritten.

19.1.3 Proposition: Es sei $(S_n)_{n\in N_0}$ eine einfache Irrfahrt. Dann gilt

$$\begin{aligned}\mathbb{E}(S_n) &= S_0 + n\,\mathbb{E}(X_1) = S_0 + n(p-q),\\ \mathbb{V}(S_n) &= 4npq,\\ \mathrm{sd}(S_n) &= \sqrt{4pq}\cdot\sqrt{n}\end{aligned}$$

Beweis: DIY (:= Do it yourself)

▶ Die Streuung gemessen in der Standardabweichung wächst proportional zu $\sqrt{n}$.

19.1.4 Proposition: Es sei $(S_n)_{n\in N_0}$ eine einfache Irrfahrt

mit $S_0 = a$. Dann gilt

$$\mathbb{P}(S_n = b) = \binom{n}{\frac{1}{2}(n+b-a)} p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)},$$

wobei wir die Binomialkoeffizienten Null sind, falls $\frac{1}{2}(n+b-a) \notin \{0,1,2,...,n\}$.

Beweis: Wir beachten

$$b = S_n = a + \sum_{i=1}^{n} X_i$$

$$b - a = \sum_{i=1}^{n} X_i = u - d$$

Aus den beiden Gleichungen

$$u + d = n$$

$$u - d = b - a$$

folgt

$$u = \frac{n + (b-a)}{2}$$

$$d = \frac{n - (b-a)}{2}$$

Um von a auf b zu kommen, muss $b - a = u - d$ gelten. Wir benötigen also u Einsen und d Minus-Einsen. Bei $n = u + d$ Schritten gibt es dafür $\binom{u+d}{d}$ Möglichkeiten.

Also gilt

$$\mathbb{P}(S_n = b) = \binom{u+d}{d} p^u q^d.$$

Jetzt muss man nur noch die Gleichungen für u und d einsetzen.

19.2 Einfache symmetrische Irrfahrt

Die einfache symmetrische Irrfahrt (mit Anfangswert 0) ist besonders wichtig. Deshalb werden wir diese ausführlicher betrachten. Später werden wir uns ausführlich mit der **Brown'schen Bewegung** beschäftigen, der wohl der **wichtigste stochastische Prozess** ist. Die Brown'sche Bewegung kann gut durch eine geraffte skalierte symmetrische Irrfahrt approximiert werden.

Feller [10] hat eine besonders einflussreiche Analyse der symmetrischen Irrfahrt vorgelegt. Das Lehrbuch von **Feller** [10] ist berühmt! Hier `https://de.wikipedia.org/wiki/William_Feller` können Sie etwas über William Feller lesen.

19.2.1 Proposition (Feller [10, S. 68]): Es sei $(X_i)_{i\in\mathbb{N}}$ eine Folge von unabhängigen Bernoulli Zufallsvariablen, wobei

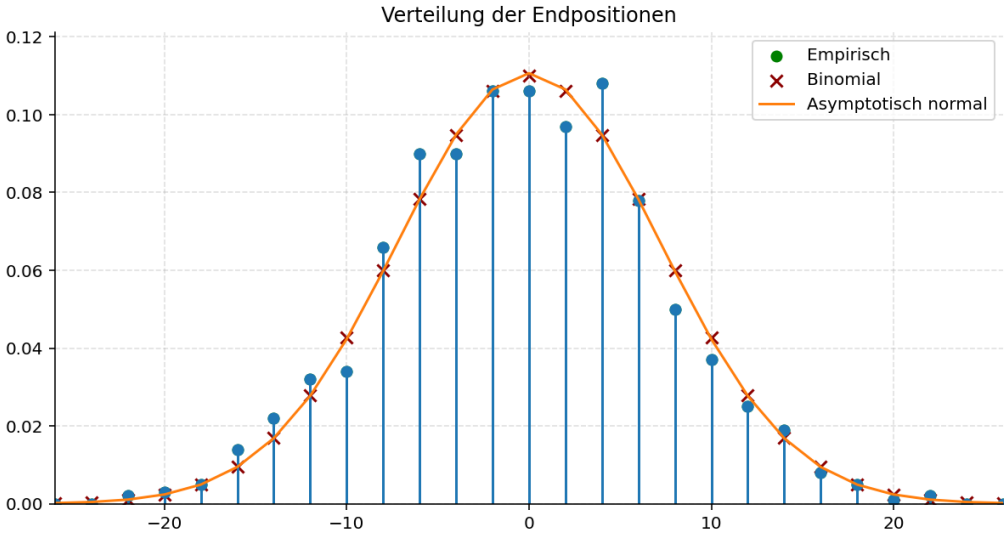


Abbildung 19.2: Verteilung von S_{52}. Empirisch: Monte Carlo mit 1000 Pfaden.

$\mathbb{P}(X_i = 1) = \frac{1}{2} = \mathbb{P}(X_i = -1)$ und

$$S_n = 0 + \sum_{i=1}^{n} X_i \quad (n = 1, 2, 3, ...)$$

die durch X_i definierte einfache symmetrische Irrfahrt mit Anfangswert $S_0 = 0$. Dann gilt

$$p_{n,x} := \mathbb{P}(S_n = x) = \left(\frac{1}{2^n}\right) \cdot \binom{u+d}{u}$$

dabei ist $n = u + d$ und $x = u - d$.

▶ u bezeichnet hier die Anzahl der 1'en und d die Anzahl der -1'en in der Summe S_n.

Bekanntlich zählt der Binomialkoeffizient

$$\binom{n}{u} = \frac{n!}{u!d!} = \binom{n}{d}$$

die Anzahl der Möglichkeiten u 1'en und d (-1)'en anzuordnen.

Wir beachten: u und d ergeben sich aus n und x. Wir können x als **Vorsprung** auffassen.[2]

Wir notieren noch für den späteren Gebrauch

$$n = u + d$$
$$x = u - d$$

und durch Umformung noch

$$d = \frac{n - x}{2}$$
$$u = \frac{n + x}{2}.$$

Wir können die obige Wahrscheinlichkeit deshalb so angeben

$$\mathbb{P}(S_n = x) = \left(\frac{1}{2^n}\right) \cdot \binom{n}{d} = \left(\frac{1}{2^n}\right) \cdot \binom{n}{\frac{n+x}{2}}$$
$$= \left(\frac{1}{2^n}\right) \cdot \binom{n}{u}$$

19.2.2 Proposition: Für die einfache symmetrische Irrfahrt

[2] Feller [10] schreibt p anstatt u und q anstatt d. Wir brauchen aber p und q für die Erfolgswahrscheinlichkeit bzw. die Misserfolgswahrscheinlichkeit.

mit Anfangswert $S_0 = 0$ gilt u_{2n}

$$\mathbb{P}(S_{2n} = 0) = \binom{2n}{n} \cdot \left(\frac{1}{2^{2n}}\right) =: u_{2n}$$

► Das folgende sogenannte Spiegelungsprinzip kann erstens noch verallgemeinert werden und hat sich zweitens als außergewöhnlich nützlich erwiesen. In Henze [**?**] und in Feller [10] kann man dazu sehr viele Argumente finden. ◄

19.2.3 Spiegelungsprinizip (Feller [10, S. 72]): Es sei $A = (a, \alpha) \in \mathbb{N}_0 \times \mathbb{Z}, B = (b, \beta) \in \mathbb{N}_0 \times \mathbb{Z}$ mit $b > a \geq 0, \alpha > 0, \beta > 0$. Dann gilt: Die Anzahl der Pfade von A nach B, die die Nullachse berühren oder schneiden, ist gleich der Anzahl aller Pfade von $A' = (a, -\alpha)$ nach B.

In Worten: Die Anzahl der Pfade von $A = (a, \alpha)$ nach $B = (b, \beta)$ mit einer Nullstelle, ist gleich der Anzahl aller Pfade von $A' = (a, -\alpha)$ nach $B = (b, \beta)$.

Beweis: Mitschrift SL 20260409

19.2.4 Ballot Theorem (Henze [?, S. 14], Feller [10, S. 73]): Für $n, x \in \mathbb{N}$ gibt es genau $\left(\frac{x}{n}\right) \cdot N_{n,x}$ Pfade der einfachen Irrfahrt mit $S_0 = 0$, $S_1, ..., S_{n-1} > 0$ und $S_n = x$.

Es gibt also $\left(\frac{x}{n}\right) \cdot N_{n,x}$ nullstellenfreie Pfade von $(0, 0)$ nach (n, x).

Dabei bezeichnet $N_{n,x}$ die Anzahl der Pfade von $(0,0)$ nach (n,x). Es ist

Hier gibt es unausgesprochene Konventionen. Welche?

$$N_{n,x} = \binom{n}{u} = \binom{n}{d} = \binom{u+d}{u},$$

wobei wie gehabt $u+d=n$ und $x=u-d$ bzw. $d=\frac{n-x}{2}, u=\frac{n+x}{2}$. u ist die Anzahl der 1'en, die wir benötigen um von $(0,0)$ nach (n,x) zu kommen, und die Anzahl der -1'en ist d.

Beweis: Zunächst beobachten wir, dass die Anzahl der im Theorem genannten Pfade gleich der Anzahl der Pfade **ohne** Nullstellen von $(1,1)$ nach (n,x) ist. Die Anzahl der Pfade ohne Nullstellen von $(1,1)$ nach (n,x) ist dann gleich der Anzahl **aller** Pfade von $(1,1)$ nach (n,x) **abzüglich** der Pfade **mit** Nullstellen von $(1,1)$ nach (n,x).

Jetzt bestimmen wir die **Anzahl aller Pfade** von $(1,1)$ nach (n,x). Wir **verschieben dazu das Koordinatensystem** so, dass $(1,1)$ in den Punkt $(0,0)$ übergeht. Dann sehen wir, dass die Anzahl aller Pfade von $(1,1)$ nach (n,x) gleich der Anzahl aller Pfade von $(0,0)$ nach $(n-1,x-1)$ ist; also gleich $N_{n-1,x-1}$.

Es gilt

$$N_{n-1,x-1} = \binom{u+d-1}{u-1},$$

denn, wir benötigen für einen Pfad von $(0,0)$ nach $n-1, x-1$

im Vergleich zu einem Pfad von $(0,0)$ zu (n,x) eine $+1$ weniger und einem Schritt weniger. Das können wir uns anhand einer Skizze klar machen oder nachrechnen: Wir bestimmten die Anzahl der nötigen 1'en und (-1)'en. Die beide Werte kann man an den Indizes von $N_{n-1,x-1}$ ablesen.

$$u_{\text{verschoben}} = \frac{n-1+(x-1)}{2} = \frac{n+x-2}{2} = \frac{n+x}{2} - 1 = u - 1$$
$$d_{\text{verschoben}} = \frac{n-1-(x-1)}{2} = \frac{n-x}{2} - 1 = d - 1$$

Jetzt bestimmen wir die **Anzahl aller Pfade** von $(1,1)$ nach (n,x) **mit Nullstellen**. Gemäß des **Spiegelungsprinzips** ist das gleich der Anzahl aller Pfade von $(1,-1)$ nach (n,x). Jetzt verschieben wir das Koordinatensystem so, dass der Punkt $(1,-1)$ in den Punkt $(0,0)$ übergeht. Wir erkennen jetzt, dass die Anzahl aller Pfade von $(1,-1)$ nach (n,x) gleich der Anzahl aller Pfade von $(0,0)$ nach $(n-1,x+1)$ ist; also gleich $N_{n-1,x+1}$.

Es gilt

$$N_{n-1,x+1} = \binom{u+d-1}{u}$$

wobei wir uns vergewissern, dass es tatsächlich u unten im Binomialkoeffizienten sein muss! Davon können wir uns in einer Skizze vergewissern. Wir können es aber auch wieder

nachrechnen:

$$u_{\text{verschoben}} = \frac{n-1+(x+1)}{2} = \frac{n+x}{2} = u$$
$$d_{\text{verschoben}} = \frac{n-1-(x+1)}{2} = \frac{n-x-2}{2} = d-1$$

Dann rechnen wir nach (siehe Mitschrift Mathfred StochProz SL 20260416), dass

$$\binom{u+d-1}{u-1} - \binom{u+d-1}{u} = \frac{x}{n} \cdot N_{n,x}$$

ist.

19.2.5 Bemerkung: Wir betrachten eine symmetrische Irrfahrt (S_n) mit Anfangswert $S_0 = 0$. Wir haben vorne ermittelt:

$$p_{n,x} := \mathbb{P}(S_n = x) = \frac{N_{n,x}}{2^n} = \frac{\binom{n}{\frac{n+x}{2}}}{2^n} = \frac{\binom{n}{u}}{2^n}$$

und

$$u_{2n} := \mathbb{P}(S_{2n} = 0) = \binom{2n}{n} \cdot \left(\frac{1}{2^{2n}}\right)$$

19.2.6 Hauptlemma (Feller [10, S. 76]): Es gilt

$$\mathbb{P}(S_1 \neq 0, ..., S_{2n} \neq 0) = \mathbb{P}(S_{2n} = 0) = u_{2n}.$$

[das zweite = ist eigentlich nicht Teil der Behauptung, son-

dern schon bekannt/definiert] Mit Worten: Die Wahrscheinlichkeit einen Pfad der Länge n ohne Nullstellen zu erhalten ist genauso hoch, wie die Wahrscheinlichkeit, dass der Pfad nach n Schritten 0 ist.

Beweis: Aus Symmetriegründen

$$\mathbb{P}(S_1 \neq 0, ..., S_{2n} \neq 0) = 2 \cdot \mathbb{P}(S_1 > 0, ..., S_{2n} > 0)$$

Wir beachten/betrachten deshalb

$$\begin{aligned}
\mathbb{P}(S_1 > 0, ..., S_{2n} > 0) &= \sum_{r=1}^{\infty} \mathbb{P}(S_1 > 0, ..., S_{2n-1} > 0, S_{2n} = 2r) \\
&= \sum_{r=1}^{\infty} (N_{2n-1,2r-1} - N_{2n-1,2r+1}) \cdot 2^{-2n} \\
&= \sum_{r=1}^{\infty} (N_{2n-1,2r-1} - N_{2n-1,2r+1}) \cdot 2^{-2n+1} 2^{-1} \\
&= \sum_{r=1}^{\infty} (N_{2n-1,2r-1} - N_{2n-1,2r+1}) \cdot 2^{-(2n-1)} 2^{-1}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{r=1}^{\infty} \frac{1}{2} \left(N_{2n-1,2r-1} \cdot 2^{-(2n-1)} - N_{2n-1,2r+1} \cdot 2^{-(2n-1)} \right) \\
&= \frac{1}{2} \left(N_{2n-1,1} \cdot 2^{-(2n-1)} \right) \\
&= \frac{1}{2} \mathbb{P}\left(S_{2n-1} = 1\right) = \frac{1}{2} \mathbb{P}\left(S_{2n} = 0\right) \qquad \text{[siehe unten]} \\
&= \frac{1}{2} u_{2n}
\end{aligned}
$$

Wir rechnen $\mathbb{P}\left(S_{2n-1} = 1\right) = \mathbb{P}\left(S_{2n} = 0\right)$ nach:

$$
\begin{aligned}
\mathbb{P}\left(S_{2n-1} = 1\right) &= \binom{2n-1}{\frac{2n-1+1}{2}} \frac{1}{2^{2n-1}} = \binom{2n-1}{n} \frac{1}{2^{2n-1}} \\
&= \binom{2n-1}{n} \frac{2}{2^{2n}} \\
&= \frac{(2n-1)!}{n!(2n-1-n)!} \frac{2}{2^{2n}} \\
&= \frac{2n(2n-1)!}{n!(2n-1-n)!\,2n} \frac{2}{2^{2n}} \\
&= \frac{2n(2n-1)!}{n!(n-1)!\,n} \frac{1}{2^{2n}} \\
&= \binom{2n}{n} \frac{1}{2^{2n}} \\
&= \mathbb{P}\left(S_{2n} = 0\right)
\end{aligned}
$$

19.2.7 Definition: Wir definieren die **Erstwiederkehrzeit**

$$W = \inf\{2k \mid k \in \mathbb{N}, S_{2k} = 0\}.$$

Wenn die ZV W (für einen Pfad) den Wert $2k$ annimmt, dann

ist $2k$ der erste Zeitpunkt zudem der Pfad eine Nullstelle hat; also zur 0 zurückkehrt.

19.2.8 Proposition: Es gilt für $n \geq 1$

$$f_{2n} := \mathbb{P}(W = 2n) = \frac{\binom{2(n-1)}{n-1}}{2^{2(n-1)}} \frac{1}{2n} = \frac{u_{2(n-1)}}{2n}$$
$$= \frac{1}{2n-1} u_{2n}.$$

Beweis: Feller [10, S. 78] oder Henze [**?**, S. 43] Wir beobachten und nutzen das Hauptlemma

$$\begin{aligned} f_{2n} &= \mathbb{P}(\{S_1 \neq 0, ..., S_{2n-2} \neq 0\} \setminus \{S_1 \neq 0, ..., S_{2n} \neq 0\}) \\ &= \mathbb{P}(\{S_1 \neq 0, ..., S_{2n-2} \neq 0\}) - \mathbb{P}(\{S_1 \neq 0, ..., S_{2n} \neq 0) \\ &= u_{2n-2} - u_{2n}. \end{aligned}$$

Die Behauptung folgt dann durch *direktes/mühsames* Rechnen.

$$\binom{2n-2}{n-1} \cdot \left(\frac{1}{2^{2n-2}}\right) - \binom{2n}{n} \cdot \left(\frac{1}{2^{2n}}\right) = ...$$

19.2.9 Satz: Der Zustand $S = 0$ der einfachen symmetri-

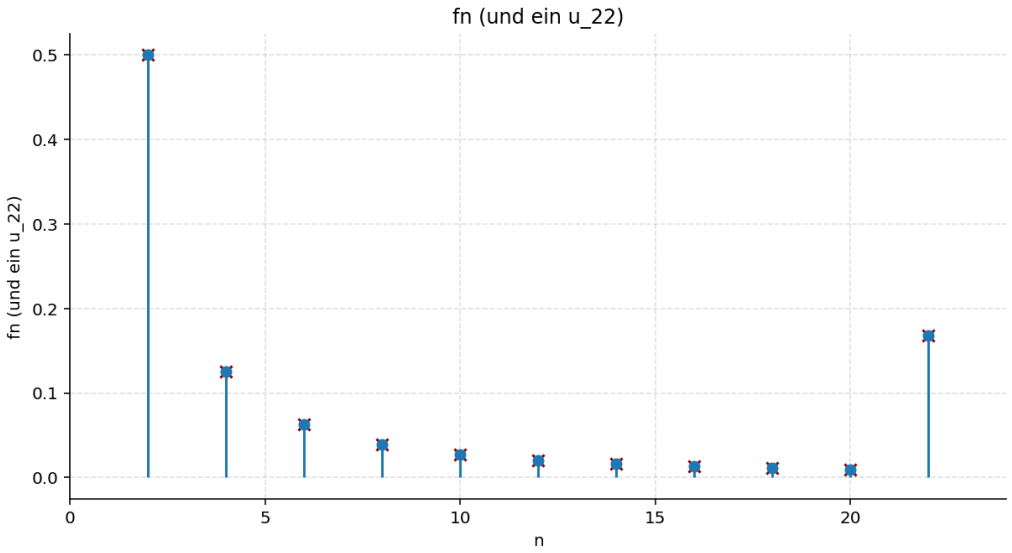


Abbildung 19.3: Die Wahrscheinlichkeiten $f_n, n = 2, 4, ..., 20$ und u_{22}.

schen Irrfahrt ist **null-rekurrent**, d.h.

$$\mathbb{P}(W < \infty) = \sum_{n=1}^{\infty} \mathbb{P}(W = 2n) = 1$$
$$\mathbb{E}(W) = \infty$$

Beweis: Vgl. Henze [**?**, S. 43] (aber auch Feller [10, S. 78]). Die erste Formel haben wir bereits vorne festgestellt.

Wir zeigen $\mathbb{P}(W = \infty) = 0$. Es gilt $\{W = \infty\} = \cap_{k=1}^{\infty}\{W \geq 2k\}$. Ferner ist $\{W \geq 2(k+1)\} \subset \{W \geq 2k\}$. Also gilt

$$\{W = \infty\} = \lim_{k\to\infty} \{W \geq 2k\}$$

Da Wahrscheinlichkeitsmaße stetig sind, folgt

$$\begin{aligned}\mathbb{P}(\{W=\infty\}) &= \mathbb{P}(\lim_{k\to\infty}\{W\geq 2k\})\\ &= \lim_{k\to\infty}\mathbb{P}(\{W\geq 2k\})\\ &= \lim_{k\to\infty} u_{2(k-1)} = 0,\end{aligned}$$

wobei wir

$$\lim_{k\to\infty} u_{2k}\sqrt{\pi k} = 1$$

verwenden (vgl. Henze [**?**, S. 27]).

Für die dritte Gleichung der Behauptung

$$\begin{aligned}\mathbb{E}(W) &= \sum_{k=1}^{\infty} 2k\cdot\mathbb{P}(W=2k) = \sum_{k=1}^{\infty} 2k\cdot\frac{u_{2(k-1)}}{2k}\\ &= \sum_{k=1}^{\infty} u_{2(k-1)} = \infty.\end{aligned}$$

Vgl. Henze [**?**, S. 43f] für einige fehlende Schritte.

▶ Wir können uns also bei der einfachen symmetrischen Irrfahrt sicher sein, dass sie zur Null zurückfindet, aber wir erwarten, dass es unendlich lange dauert. Das strapaziert und erweitert unsere Anschauung! ◀

▶ Am Rande: Gemaß Durrett [6, S. 250] findet auch ein betrunkener Mensch in $\mathbb{Z}^2$ mit Wahrscheinlichkeit 1 nach Hause

zurück, ein betrunkener Vogel in $\mathbb{Z}^3$ nicht.[3] ◀

19.3 Homogenität und Markov-Eigenschaft

Für diesen Abschnitt verweise ich besonders auf Grimmett und Stirzacker [14, S. 18f, 79ff].

19.3.1 Satz: Eine einfache Irrfahrt ist **räumlich und zeitlich homogen**:

$$\begin{aligned}\mathbb{P}(S_n = j \,|\, S_0 = a) &= \mathbb{P}(S_n = j + b \,|\, S_0 = a + b),\\ \mathbb{P}(S_n = j \,|\, S_0 = a) &= \mathbb{P}(S_{m+n} = j \,|\, S_m = a).\end{aligned}$$

Beweis: Zu räumlichen Homogenität: Man kommt in n Schritten von a nach j gdw $\sum_{i=1,\ldots,n} X_i = j - a$ gilt. Genau dann ist aber auch $\sum_{i=1,\ldots,n} X_i = j + b - (a + b)$. Also kommt man in n Schritten von $a + b$ nach $j + b$.

Zur zeitlichen Homogenität: Es gilt $\mathbb{P}(S_n = j \,|\, S_0 = a) = \mathbb{P}(\sum_{i=1}^{n} X_i = j - a) = \mathbb{P}(\sum_{i=m+1}^{n+m} X_i = j - a)$, wobei das letzte Gleichheitszeichen gilt, denn die X_i haben alle die gleiche Verteilung.

19.3.2 Satz: Eine einfache Irrfahrt hat die **Markov Eigen-**

[3] Durrett [6, S. 250]: To steal a joke from Kakutani (U.C.L.A. colloquium talk): *A drunk man will eventually find his way home, but a drunk bird may get lost forever.*

schaft:

$$\mathbb{P}(S_{m+n} = j \mid S_0 = s_0, S_1 = s_1, ..., S_m = s_m) = \mathbb{P}(S_{m+n} = j \mid S_m = s_m), n$$

Also: Für die Zukunft ist von der Geschichte nur die Gegenwart relevant. Oder: The future is independent of the past given the present.

Beweis: Es gilt $\mathbb{P}(S_{m+n} = j \mid S_0 = s_0, S_1 = s_1, ..., S_m = s_m) = \mathbb{P}(\sum_{i=m+1}^{n+m} X_i = j - s_m) = \mathbb{P}(S_{m+n} = j \mid S_m = s_m)$.

19.4 Des Spielers Ruin

Für diesen Abschnitt verweise ich besonders auf Feller [10, 342ff], Grimmett und Stirzacker [14, S. 18f, 79ff] sowie auf Steele [**?**, S. 1ff].

19.4.1 Defintion: Es sei $(X_i)_{i\in\mathbb{N}}$ eine Folge von unabhängigen Bernoulli Zufallsvariablen, wobei $\mathbb{P}(X_i = 1) = p > 0, \mathbb{P}(X_i = -1) = 1 - p = q > 0$. Ferner sei $S_0 = k$ und $S_n = S_0 + \sum_{i=1}^{n} X_i, n = 1, 2, 3, ...$ die durch $(X_i)_{i\in\mathbb{N}}$ erzeugte einfache Irrfahrt und $A, B \in \mathbb{N}_0, -B \leq k \leq A$. Es sei

$$\tau = \min\{n \geq 0 \,:\, S_n = A \text{ oder } S_n = -B\}$$

der zufällige Zeitpunkt zudem die Irrfahrt A oder B erreicht.

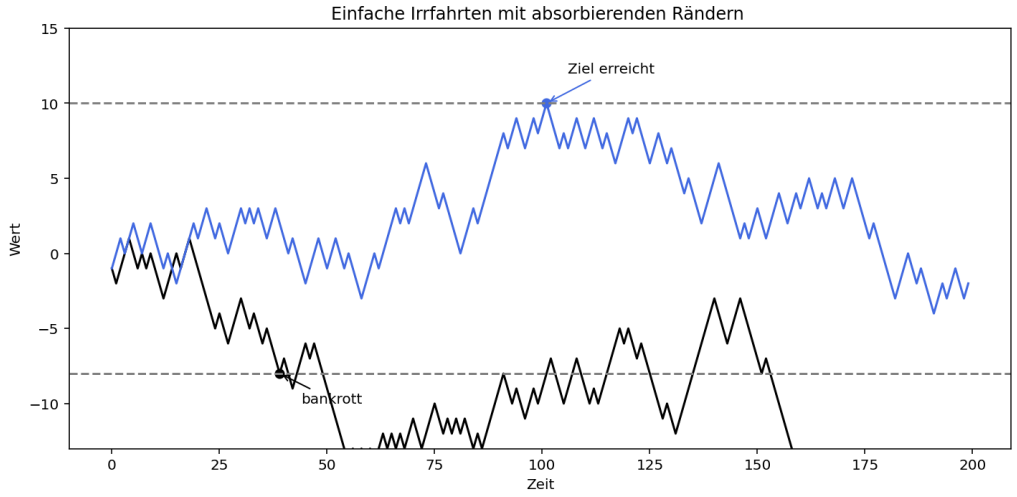


Abbildung 19.4: Die blaue Irrfahrt erreicht das Ziel (ohne vorher bankrott zu gehen). Der schwarze nicht :-(

τ heißt **Stoppzeit der einfachen Irrfahrt mit den absorbierenden Rändern** A und $-B$.

19.4.2 Satz: Die Stoppzeit einer einfachen Irrfahrt mit absorbierenden Rändern A und $-B$ ist mit Wahrscheinlichkeit 0 unendlich, also

$$\mathbb{P}(\tau = \infty) = 0$$

Beweis (vgl. Steele [34, Seite 3]): Wenn $S_0 = -B$, dann stoppt die Irrfahrt sofort und $\tau = 0 < \infty$. Wir können also o.B.d.A. davon ausgehen, dass $S_0 > -B$ gilt. Wir betrachten die Ereignis $E_1 = \{X_1 = 1, X_2 = 1, ..., X_{A+B} = 1\}$, d.h. eine Serie von $A + B$ aufeinanderfolgenden Gewinnen. Wenn E_1 eintritt, dann stoppt die Kette in A. Wir betrachten weiterhin

$E_2 = \{X_{A+B+1} = 1, X_{A+B+2} = 1, ..., X_{2(A+B)} = 1\}$ und allgemein $E_k = \{X_{(k-1)(A+B)+1} = 1, ..., X_{k(A+B)} = 1\}$.

Es gilt $\mathbb{P}(E_k) = p^{A+B}$, $\mathbb{P}(E_i^c) = 1 - p^{A+B}$, die Ereignisse E_k sind unabhängig und es gilt

$$\mathbb{P}(\tau > n(A+B) | \, S_0 = k) \leq \mathbb{P}\left(\bigcap_{i=1}^{n} E_i^c\right) = (1-p)^n.$$

Dann gilt also für alle n und $k > -B$

$$\mathbb{P}(\tau = \infty \, | \, S_0 = k) \leq \mathbb{P}(\tau > n(A+B) | \, S_0 = k) \leq (1-p)^n.$$

Dann muss $\mathbb{P}(\tau = \infty \, | \, S_0 = k) = 0$ gelten, denn wir können n beliebig groß und damit $(1-p)^n$ klein wählen.

19.4.3 Satz: Für die einfache symmetrische Irrfahrt – also mit $p = \frac{1}{2}$ – und absorbierenden Rändern A und $-B$ und Anfangswert $k \in \{-B, ..., 0, ..., A\}$ gilt:

$$\begin{aligned} \mathbb{P}(S_\tau = A | S_0 = k) &= \mathbb{P}((S_n)_{n\geq 0} \text{ endet in } A \, | \, S_0 = k) \\ &= \frac{k+B}{A+B}. \end{aligned}$$

Beweis (vgl. Steele [34, S. 2]): Wir betrachten

$$f(k) = \mathbb{P}(S_\tau = A | S_0 = k).$$

Dann gilt die Differenzengleichung

$$f(k) = \frac{1}{2}f(k-1) + \frac{1}{2}f(k+1)$$

und die Randbedingungen

$$f(-B) = 0$$
$$f(A) = 1.$$

Die Aufgabe besteht jetzt darin, die Differenzengleichung mit Randbedingungen zu lösen. Dafür gibt es systematische Methoden. Wir machen einen Ansatz und setzen $f(-B+1) = \alpha$. Jetzt verwenden wir die Differenzengleichungen und erhalten

$$f(-B+1) = \frac{1}{2}f(-B) + \frac{1}{2}f(-B+2)$$
$$\alpha = \frac{1}{2} \cdot 0 + \frac{1}{2}f(-B+2)$$

also

$$2\alpha = f(-B+2).$$

Jetzt setzen wir $f(-B+2), f(-B+1)$ in die Differenzen-

gleichung ein und erhalten

$$\begin{aligned} f(-B+2) &= \frac{1}{2}f(-B+1) + \frac{1}{2}f(-B+3) \\ 2\alpha &= \frac{1}{2}\alpha + \frac{1}{2}f(-B+3) \\ 4\alpha &= \alpha + f(-B+3) \\ 3\alpha &= f(-B+3) \end{aligned}$$

Allgemein erhalten wir

$$f(-B+k) = k\alpha.$$

Wir haben die zweite Randbedingung noch nicht verwendet! Damit folgt

$$1 = f(A) = f(-B+B+A) = \alpha(B+A)$$

Also

$$\alpha = \frac{1}{A+B}$$

Also

$$f(-B+k) = \frac{k}{A+B}$$

bzw.

$$f(k) = f(-B+B+k) = \frac{B+k}{A+B}$$

19.4.4 Satz: Für die Stoppzeit einer einfachen symmetrischen Irrfahrt mit absorbierenden Rändern A und $-B$ gilt

$$\mathbb{E}(\tau \mid S_0 = k) = (A - k)(k + B).$$

Beweis (vgl. Steele [34, S. 2]): Wir betrachten

$$g(k) = \mathbb{E}(\tau \mid S_0 = k).$$

Dann gilt[4]

$$\begin{aligned} g(k) &= \frac{1}{2}(1 + g(k-1)) + \frac{1}{2}(1 + g(k+1)) \\ &= \frac{1}{2}g(k-1) + \frac{1}{2}g(k+1) + 1 \end{aligned}$$

und

$$g(-B) = 0 \text{ und } g(A) = 0.$$

Dann beobachten wir

$$\begin{aligned} & g(k) = \frac{1}{2}g(k-1) + \frac{1}{2}g(k+1) + 1 \\ \Rightarrow \quad & g(k) - g(k-1) = g(k+1) - g(k) + 2 \\ \Rightarrow \quad & g(k) - g(k-1) = g(k+1) - g(k) + 2 \\ \Rightarrow \quad & \Delta g(k-1) = \Delta g(k) + 2 \\ \Rightarrow \quad & \frac{1}{2}\Delta^2 g(k-1) = -1 \end{aligned}$$

[4]Dabei beachten wir das Gesetz der totalen Wahrscheinlichkeit für Erwartungswerte (Grimmett und Stirzaker [14, 72]): $\mathbb{E}(Y) = \sum_{i \in I} \mathbb{E}(Y|A_i)\mathbb{P}(A_i)$

Auf Basis dieser Gleichung ist es naheliegend, dass g quadratisch in k ist (die zweite *Ableitung* ist konstant). Zudem kennen wir zwei Nullstellen A und $-B$. Dann kann man $(A-k)(k+B)$ ansetzen und nachweisen, dass dieser Ansatz alle drei Bedingungen erfüllt.

19.4.5 Satz: Für die einfache asymmetrische Irrfahrt $p \neq 1/2$ und absorbierenden Rändern A und $-B$ und Anfangswert $k \in \{-B, ..., 0, ..., A\}$ gilt:

$$\mathbb{P}(S_\tau = A \,|\, S_0 = k) = \mathbb{P}(S_n \text{ endet in } A \,|\, S_0 = k) = \frac{(q/p)^{k+B} - 1}{(q/p)^{A+B} - 1}.$$

Beweis: Übung bzw. Steele [34, Seite 6]

19.4.6 Satz: Für die Stoppzeit einer einfachen asymmetrischen Irrfahrt mit absorbierenden Rändern A und $-B$ gilt

$$\mathbb{E}(\tau \,|\, S_0 = k) = \frac{B+k}{q-p} - \frac{A+B}{q-p} \cdot \frac{1-(q/p)^{B+k}}{1-(q/p)^{A+B}}.$$

Beweis: Übung

19.5 Nachtrag zur Wiederkehr

Dieser Abschnitt ist kurz und behandelt nur einen Satz. Für den Beweis verweise ich auf Grimmett und Stirzacker [14, S. 183ff]. Der Beweis dort verwendet erzeugende Funktionen.

19.5.1 Satz: Es sei $(S_n)_{n\in\mathbb{N}_0}$ eine einfache nicht notwendigerweise symmetrische Irrfahrt und $S_0 = 0$. Dann gilt:

i.) Die Wahrscheinlichkeit, dass die Irrfahrt jemals zu Null zurückkehrt ist $1 - |p - q|$.

Also: Die einfache symmetrische und nur die einfache symmetrische Irrfahrt kehrt sicher (d.h. mit Wahrscheinlichkeit $= 1$) zur Null zurück. Bei einer asymmetrischen Irrfahrt tritt die Rückkehr zur Null nur mit einer Wahrscheinlichkeit < 1.

ii.) Für die einfache symmetrische Irrfahrt tritt die Rückkehr zur Null mit Wahrscheinlichkeit 1 ein, aber der Erwartungswert für die Dauer bis zur Rückkehr ist ∞.

19.5.2 Bemerkung: Man sagt, dass 0 für $p = 1/2$ **rekurrent** ist. Für $p \neq \frac{1}{2}$ ist 0 flüchtig, vorübergehend oder **transient**.

20 Markov-Ketten

Dieses Kapitel beruht auf **Grimmett und Stirzaker** [14] und **Henze** [19] sowie Billingsley [2]. Norris [**?**], Tolver [35] und Privault [**?**] sind weitere gute Quellen. Die beiden Bücher Korosteleva [**?**] und Dobrow [40] sind anwendungsorientiert und es gibt zudem nützlichen R Code, den Sie gut als Ausgangspunkt für eine Programmierungen nehmen können.

20.1 Definition

20.1.1 Definiton: Ein **zeit-diskreter stochastischer Prozess** $(X_n)_{n\in\mathbb{N}_0}$ mit **Werten in einer höchstens abzählbaren Menge** S heißt **Markov-Kette**, falls er die Markov-Eigenschaft

$$\mathbb{P}(X_n = s|X_0 = x_0, ..., X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = s|X_{n-1} = x_{n-1})$$

für alle $n \geq 1, x_0, ..., x_{n-1}, s \in S$ hat.

Für die Zukunft ist von der Vergangenheit nur die Gegenwart relevant.

Da die **Zustandsmenge** S abzählbar ist, werden wir

$$s_i \in S, i \in I, I \subset \mathbb{N} \text{ mit } i \in I, I \subset \mathbb{N}$$

identifizieren. Wir schreiben also $X_n = i$ anstatt $X_n = s_i$. Wir sagen dann, dass die Kette im Zustand i ist (anstatt im Zustand s_i).

20.1.2 Definiton: Eine Markov-Kette $(X_n)_{n\in\mathbb{N}_0}$ heißt (zeitlich) **homogen**, falls

$$\mathbb{P}(X_n = j|X_{n-1} = i) = \mathbb{P}(X_1 = j|X_0 = i)$$

für alle n, i, j gilt. Homogenität bedeutet also, dass der Zeitpunkt n für die Übergangswahrscheinlichkeit $\mathbb{P}(X_n = j|X_{n-1} = i)$ irrelevant ist

In diesem homogenen Fall definieren wir die $\#(S) \times \#(S)$ **Übergangsmatrix P** durch

$$p_{ij} = \mathbb{P}(X_1 = j|X_0 = i).$$

Diese Übergangsmatrix hat u.U. unendlich viele Zeilen und Spalten.

Wir werden nur homogene Markov-Ketten betrach-

ten.

20.2 Übergangsdynamiken

20.2.1 Satz: Für die Übergangsmatrix einer homogenen Markov-Kette gilt:

i.) Für alle i, j gilt $p_{ij} \geq 0$.

ii.) Für alle i gilt $\sum_{j \in S} p_{ij} = 1$. Die Zeilensumme ist 1.

Man sagt manchmal, dass $\mathbf{P}$ eine stochastische Matrix ist. Das ist jedoch missverständlich, da die Einträge in der Matrix keineswegs stochastisch (Zufallsgrößen) sind.

20.2.2 Definition: Die **n-Schritt Übergangsmatrix** $\mathbf{P}(m, m+n) = (p_{ij}(m, m+n))$ einer homogenen Markov-Kette ist durch

$$p_{ij}(m, m+n) = \mathbb{P}(X_{m+n} = j | X_m = i)$$

definiert.

Beachte, dass $p_{ij}(m, m+n)$ eine **bedingte** Wahrscheinlichkeit bezeichnet.

Wir werden sehen, dass wegen Homogenität der Zeitpunkt m für die Übergangswahrscheinlichkeit $p_{ij}(m, m+n)$ irrelevant

ist, so dass die Notation $p_{ij}(m, m+n) = p_{ij}^{(n)}$ gerechtfertigt ist (siehe unten).

Wir beachten ferner $\mathbf{P}(m, m+1) = \mathbf{P}$.

20.2.3 Satz: Für die Übergangsmatrizen einer homogenen Markov-Kette gelten die **Chapman-Kolmogorov Gleichungen**:

$$p_{ij}(m, m+n+r) = \sum_k p_{ik}(m, m+n) p_{kj}(m+n, m+n+r).$$

Also gilt

$$\mathbf{P}(m, m+n+r) = \mathbf{P}(m, m+n)\mathbf{P}(m+n, m+n+r)$$

und

$$\mathbf{P}(m, m+n) = \mathbf{P}^n.$$

Wir erhalten die Übergangsdynamiken also durch **einfaches Potenzieren der Übergangsmatrix**.

Die i-te Zeile von $\mathbf{P}^n$ enthält an der j-ten Stelle die (bedingte) Wahrscheinlichkeiten, dass die Kette ausgehend von i mach n Schritten im Zustand j ist.

Beweis: Es gilt

$$p_{ij}(m, m+n+r) = \mathbb{P}(X_{m+n+r} = j | X_m = i)$$
$$= \frac{\mathbb{P}(X_{m+n+r} = j, X_m = i)}{\mathbb{P}(X_m = i)}$$

$$= \sum_k \frac{\mathbb{P}(X_{m+n+r} = j, X_{m+n} = k, X_m = i)}{\mathbb{P}(X_m = i)}$$
$$= \sum_k \frac{\mathbb{P}(X_{m+n+r} = j | X_{m+n} = k, X_m = i)\mathbb{P}(X_{m+n} = k, X_m = i)}{\mathbb{P}(X_m = i)}$$
$$= \sum_k \frac{\mathbb{P}(X_{m+n+r} = j | X_{m+n} = k)\mathbb{P}(X_{m+n} = k, X_m = i)}{\mathbb{P}(X_m = i)}$$
$$= \sum_k \mathbb{P}(X_{m+n+r} = j | X_{m+n} = k)\mathbb{P}(X_{m+n} = k | X_m = i)$$
$$= \sum_k p_{ik}(m, m+n)p_{kj}(m+n, m+n+r)$$

20.2.4 Bemerkung: Wegen des vorhergehenden Satzes gilt also $\mathbf{P}(m, m+n) = \mathbf{P}^n = \mathbf{P}(0, n)$, d.h. die Matrix der n-Schritt Übergangswahrscheinlichkeiten $\mathbf{P}(m, m+n)$ ist unabhängig von m. Deshalb ist die **Notation**

$$p_{ij}^{(n)} = p_{ij}(m, m+n)$$

gerechtfertigt.

Wir definieren $p_{ii}^{(0)} := 1$, dann kann man einige Formeln einfacher angeben.

In Worten:

$$p_{ij}^{(n)} = \text{ „ in } n \text{ Schritten von } i \text{ nach } j \text{ “}$$

Man kann diese Wahrscheinlichkeit zu Fuß ausrechnen. Das ist aber unübersichtlich. Man bestimmt $\mathbf{P}^n$ und dann wählt man den Eintrag an der Stelle ij: $p_{ij}^{(n)} = \mathbf{P}_{ij}^n$

▶ Bisher haben wir *nur* Übergangswahrscheinlichkeiten (also bedingte Wahrscheinlichkeiten) betrachtet. *Natürlich* kann man auch Wahrscheinlichkeiten betrachten. Der folgende Satz betrachtet die Verteilung (die Wahrscheinlichkeiten) zum Zeitpunkt n angeben. Wir stellen uns vor, dass die Anfangsverteilung $\mathbb{P}(X_0 = i)$ gegeben ist und erhalten eine Formel für die Verteilung $\mathbb{P}(X_n = i)$ zum Zeitpunkt n

20.2.5 Satz: Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine homogenen Markov-Kette und $\boldsymbol{\pi}_i^{(n)} = \mathbb{P}(X_n = i)$ **die Verteilung von** X_n. . Dann gilt – wir fassen die $\boldsymbol{\pi}^{(k)}$ als **Zeilenvektoren** auf –

$$\boldsymbol{\pi}^{(m+n)} = \boldsymbol{\pi}^{(m)}\mathbf{P}^n$$

und insbesondere

$$(\mathbb{P}(X_n = i))_{s\in S}^T = \boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)}\mathbf{P}^n.$$

Beachte, dass die Einträge in $\boldsymbol{\pi}^{(n)}$ **unbedingte Wahrschein-**

lichkeiten für die Periode n sind.

In Worten

$\boldsymbol{\pi}_i^{(n)} =$ „ Wahrscheinlichkeit im Zeitpunkt n im Zustand i zu sein “

20.3 Zustandseigenschaften

▶ In diesem Abschnitt führen wir Eigenschaften von Zuständen ein.

20.3.1 Definition: Ein Zustand i einer homogenen Markov-Kette heißt **rekurrent (wiederkehrend)**, falls

$$\mathbb{P}(\text{ es gibt ein } n \in \mathbb{N} \text{ mit } X_n = i \,|X_0 = i) = 1$$

bzw. äquivalent

$$\mathbb{P}(\text{ für alle } n \in \mathbb{N} \text{ gilt } X_n \neq i \,|X_0 = i) = 0$$

Gilt hingegen *nur* $\mathbb{P}(X_n = i \text{ für ein } n \in \mathbb{N} \,|X_0 = i) < 1$, so heißt i **transient (vorübergehend)**.

Rekurrent bedeutet: Für fast alle Pfade – d.h. mit Wahrscheinlichkeit 1 – kann man einen Zeitpunkt n finden, zudem die Kette, die in i gestartet ist, zu i zurückkehrt. **Wenn wir einen Pfad zufällig ziehen, dann können wir prak-**

tisch sicher sein, dass eine Rückkehr in endlicher Zeit stattfindet.

Transient bedeutet: Nur mit einer Wahrscheinlichkeit kleiner als 1 zieht man einen Pfad, der in endlicher Zeit zu i zurückkehrt. $i \in S$ ist also **transient**, wenn

$$\mathbb{P}(X_n \neq i \text{ für alle } n \in \mathbb{N} \,|X_0 = i) > 0.$$

20.3.2 Definition: i.) Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine homogene Markov-Kette und für $i \in S$

$$T_i = \inf\{n \geq 1 | X_n = i\}$$

der zufällige Zeitpunkt zudem der Zustand erstmals gleich i ist. Wir definieren die **mittlere Rückkehrzeit**:

$$\mu_i = \mathbb{E}(T_i|X_0 = i) = \begin{cases} \sum_{k=1}^{\infty} k f_{ii}^{(k)}, \text{ falls } i \text{ rekurrent ist.} \\ \infty, \text{ falls } i \text{ transient ist.} \end{cases}$$

Dabei ist für $k \in \mathbb{N}$

$$f_{ij}^{(k)} = \mathbb{P}(X_k = j, X_{k-1} \neq j, ..., X_1 \neq j | X_0 = i).$$

die Wahrscheinlichkeit, dass der Zustand ausgehend von i erstmal nach $k \in \mathbb{N}$ Schritten gleich j ist.

Warum ist $\mu_i = \infty$, wenn i transient ist. .....

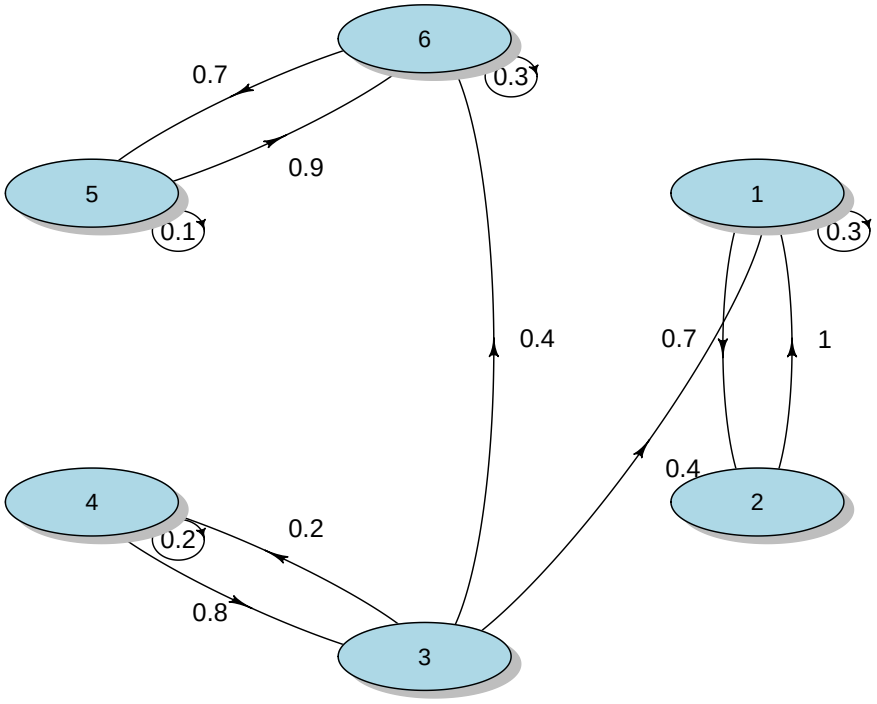


Abbildung 20.1: Es gilt $\mathbb{P}(X_1 = 6|X_0 = 3) = 0.4$. Einen Pfad der in 3 beginnt und dann in den Zustand 6 übergeht, erhalten wir mit Wahrscheinlichkeit 0.4. **Für diese Pfade** gibt es **kein** $n \in \mathbb{N}$, so dass die Kette in n zur 3 **zurückkehrt**. Der Zustand 3 ist also nicht rekurrent. Also ist der Zustand 3 transient. Es gilt aber auch nicht $\mathbb{P}(X_n = 3 \text{ für ein } n \in \mathbb{N}\,|X_0 = i) = 0$. Mit Wahrscheinlichkeit 0.2 wechselt die Kette in den Zustand 4. Mit Wahrscheinlichkeit 1 wechselt die Kette dann irgendwann in den Zustand 3 zurück (bitte vergewissern Sie sich davon). Es gilt sogar $\mathbb{P}(X_n = i \text{ für ein } n \in \mathbb{N}\,|X_0 = 3) = 0.2$, denn auch bei einem Wechsel nach 1 gibt es keine Rückkehr zur 3.

ii.) Ist der Zustand i rekurrent aber (trotzdem) $\mu_i = \infty$, dann nennen wir i **null-rekurrent**. Gilt $\mu_i < \infty$, dann nennen wir i **positiv-rekurrent**.

iii.) Für einen Zustand i heißt $d(i) = \text{ggT}\{n : p_{ii}^{(n)} > 0\}$ die **Periode** des von i.

d ist also genau dann die Periode von i, wenn: $p_{ii}(n) = 0$ außer für n, die Vielfaches von d sind und d ist die größte natürliche Zahl mit dieser Eigenschaft. Also, wenn die Periode von i 2 ist, dann kehrt die Kette mit positiver Wahrscheinlichkeit nach 2,4,6, ... Schritten nah i zurück. Nach z.B. 3 Schritten mit Wahrscheinlichkeit 0.

i heißt **periodisch**, falls $d(i) > 1$ ist. i heißt **aperiodisch**, falls $d(i) = 1$.

Aperiodisch bedeutet: Es gibt ein $N \in \mathbb{N}$, so dass $p_{ii}^{(n)} > 0$ für alle $n \geq N$ gilt.[1] Wenn also N z.B. 100 ist, dann gilt $p_{ii}^{(100)} > 0$, $p_{ii}^{(101)} > 0$, $p_{ii}^{(102)} > 0$, ... Aperiodizität ist wichtig für die Langfrist-Analyse von Zuständen beziehungsweise Ketten, wie wir später sehen werden!

iv.) Ein Zustand heißt **ergodisch**, falls er positiv-rekurrent und aperiodisch ist.

Also: Wenn i ergodisch ist, dann ist i aperiodisch, die Ket-

[1] Der Beweis benötigt etwas Zahlentheorie .....

te kehrt sicher[2] zu i zurück und die Kette benötigt für die Rückkehr im Durchschnitt nicht ewig.

v.) Ein Zustand i heißt **absorbierend**, falls $\mathbb{P}(X_{n+1} = i|X_n = i) = 1$.

20.3.3 Satz: Wir verwenden die folgenden Konzepte:

$$N_i = \#\{n \in \mathbb{N}_0 | X_n = i\}$$

bezeichnet die **Anzahl der Besuche** bei Zustand i, die in i beginnt und (wie schon vorher definiert)

$$T_i = \min\{n \geq 1 | X_n = i\}$$

bezeichnet den **Zeitpunkt des ersten Besuch des Zustands** i.

i.) Die folgenden Aussagen sind äquivalent:

a.) i ist **rekurrent**.

b.) $\mathbb{P}(T_i = \infty | X_0 = i) = 0$.

c.) $\mathbb{P}(N_i = \infty | X_0 = i) = 1$.

d.) $\sum_n p_{ii}^{(n)} = \infty$.

In Worten: Wenn ein Zustand **rekurrent** ist, dann können

[2] bedeutet mit Wahrscheinlichkeit 1

wir uns praktisch sicher sein, dass **er unendlich oft besucht wird** (deswegen wird er auch wiederkehrend genannt). Wir können uns auch praktisch sicher sein, dass es nicht unendlich lange dauert bis die Rückkehr eintritt.

ii.) Die folgenden Aussagen sind äquivalent:

a.) i ist **transient**.

b.) $\mathbb{P}(T_i = \infty | X_0 = i) > 0$.

c.) $\mathbb{P}(N_i = \infty | X_0 = i) = 0$.

d.) $\sum_n p_{ii}^{(n)} < \infty$.

In Worten: Wenn ein Zustand **transient** ist, dann können wir uns praktisch sicher sein, dass er **nur endlich oft besucht wird**. Das bedeutet, dass er **ab einem bestimmten Zeitpunkt nicht mehr besucht wird** (deshalb wird er auch manchmal flüchtig genannt). Zudem müssen wir mit positiver Wahrscheinlichkeit damit rechnen, dass die Rückkehr unendlich lange dauert.

Beweis: Billingsley [2, S. 118] für die Aussagen in c.) und d.), Grimmett und Stirzaker [14, S. 248]. Privault [**?**, Kapitel 6]. Vgl. auch Norris [**?**, Seite 26].

▶ Es gilt (vgl. Norris [?, Seite 25])

$$\mathbb{E}(N_i|X_0 = i) = \sum_{n=0}^{\infty} p_{ii}^{(n)}.$$

20.4 Klassen und Klasseneigenschaften

▶ Wir definieren auf der Menge der Zustände eine Äquivalenzrelation. Dann können wir die Zustände in Äquivalenzklassen zerlegen. Wenn eine Kette aus nur einer Klasse besteht (dann heißt sie irreduzibel) und kann für sich *übersichtlicher* untersucht werden.

Wir werden bemerken, dass die Eigenschaften positiv-rekurrent, null-rekurrent, transient und d-periodisch Klasseneigenschaften sind.

20.4.1 Definition: Wir schreiben $i \to j$ und sagen j **ist aus** i **erreichbar**, falls es ein $m \in \mathbb{N}_0$ mit $p_{ij}^{(m)} > 0$ gibt.

Es gilt: $p_{ij}^{(m)} > 0$ genau dann, wenn es eine Sequenz von Zuständen $i_0, i_1, ..., i_n$ mit $i_0 = i, i_m = j$ und stets $p_{i_k i_{k+1}} > 0$.

Wenn $i \to j$ und $j \to i$ gilt, dann schreiben wir $i \leftrightarrow j$. In diesem Fall sagen wir, dass i und j **kommunizieren**.

Bemerkung: Wegen $p_{ii}^{(0)} = 1$ (gemäß Festsetzung) gilt $i \leftrightarrow i$

für alle i.

20.4.2 Satz: $i \leftrightarrow j$ definiert auf S eine Äquivalenzrelation. Es gilt also:

- Für alle $i \in S$ ist $i \leftrightarrow i$.
- Gilt für $i, j \in S, i \leftrightarrow j$, dann gilt auch $j \leftrightarrow i$.
- Gilt für $i, j, k \in S, i \leftrightarrow j, j \leftrightarrow k$, dann gilt auch $i \leftrightarrow k$.

Bemerkung: Auf Basis der Äquivalenzrelation $\leftrightarrow$ können wir **Äquivalenzklassen** definieren. Es sei $s \in S$ in Zustand, dann heißt

$$[s] = \{x \in S \,|\, x \leftrightarrow s\}$$

Äquivalenzklasse mit **Repräsentanten** s.

20.4.3 Satz: Für $i, j \in$ mit $i \leftrightarrow j$ gilt:

i.) i und j haben die gleiche Periode.

ii.) i ist genau dann transient, wenn j transient ist.

iii.) i ist genau dann null-rekurrent, wenn j null-rekurrent ist.

iv.) i ist genau dann positiv-rekurrent, wenn j positiv-rekurrent

ist.

Bemerkung: Die vier Eigenschaften sind also **Klasseneigenschaften**: Gilt eine Eigenschaft für einen Zustand einer Klasse. dann gilt die Eigenschaft für alle Elemente dieser Klasse. Dementsprechend sprechen wir insbesondere von einer rekurrenten Klasse bzw. von einer transienten Klasse, usw. .

Beweis: Grimmet und Stirzaker [14] und Tolver [35, S.24].

Ein kompletter Beweis wäre zu lang für ein Skript. Wir zeigen (wie in Tolver [35, S.24]) folgendes: Wenn $i \leftrightarrow j$ und j rekurrent, dann ist auch i rekurrent.

Wenn $i \leftrightarrow j$ gilt, dann gibt es gemäß Definition $l, m \in \mathbb{N}$ mit $p_{ij}^{(l)} > 0$ und $p_{ji}^{(m)} > 0$.

Jetzt betrachten wir für $k \in \mathbb{N}$ die Wahrscheinlichkeit nach $l+k+m$ Schritte von i zurück nach i zulangen und dabei in den Zustand j nach l und dann nochmal nach $l+k$ Schritten zu besuchen, also:

$$p_{ij}^{(l)} p_{jj}^{(k)} p_{ji}^{(m)}.$$

Dann gilt natürlich

$$p_{ii}^{(l+k+m)} \geq p_{ij}^{(l)} p_{jj}^{(k)} p_{ji}^{(m)}$$

Jetzt beobachten wir

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} \geq \sum_{k=1}^{\infty} p_{ii}^{(l+k+m)} \geq p_{ij}^{(l)} \left(\sum_{k=1}^{\infty} p_{jj}^{(k)} \right) p_{ji}^{(m)}.$$

Wenn j rekurrent ist, dann divergiert $\sum_{k=1}^{\infty} p_{jj}^{(k)} = \infty$. Dann muss aber auch die Reihe $\sum_{i=1}^{\infty} p_{ii}^{(n)} = \infty$ divergieren. Also ist i rekurrent. Es gilt also: Wenn $i \leftrightarrow j$ und j rekurrent. Dann ist auch i rekurrent.

20.4.4 Definition: i.) Eine Menge von Zuständen $M \subset S$ heißt **abgeschlossen**, falls $p_{ij} = 0$ für alle $i \in M, j \notin M$. M heißt **kommunizierend**, falls $i \leftrightarrow j$ für alle $i, j \in M$.

ii.) Eine Markov-Kette heißt **irreduzibel**, wenn sie nur aus einer einzigen $\leftrightarrow$-**Klasse** besteht.

20.4.5 Bemerkung: i.) M ist genau dann abgeschlossen, wenn $\sum_{j \in M} p_{ij} = 1$ für alle $i \in S$ gilt.

ii.) Wenn eine Kette eine abgeschlossene Menge erreicht, dann verlässt ist die Kette diese Menge nicht mehr.

iii.) Wenn eine Menge von Zuständen kommunizierend ist, dann kommunizieren gemäß Definition alle Elemente in dieser Menge miteinander. Wenn C nicht nur eine Menge, sondern sogar eine $\leftrightarrow$-**Klasse** ist, dann liegen in C **alle** Zustände, die

miteinander kommunizieren. In diesem Sinn ist jede Klasse eine maximale kommunizierende Menge.

20.4.6 Satz: Wenn C eine rekurrente $\leftrightarrow$-Klasse ist, dann ist C abgeschlossen (vgl. Norris [**?**, Seite 27]).

Beweis: Angenommen C wäre nicht abgeschlossen. Dann gibt es $i \in C$ und $j \notin C$ mit $p_{ij} > 0$. Also $i \to j$. Würde auch $j \to i$ gelten, dann wäre $j \in C$, denn C ist eine Klasse. Also $j \not\to i$.

Jetzt betrachten wir das Ereignis, dass die Kette in i startet und nie zurückkehrt: $\{X_n \neq i \ \forall n, X_0 = i\}$. Dann gilt

$$\{X_n \neq i \ \forall n, X_0 = i\} \supset \bigcup_{k \neq i} \{X_1 = k, X_0 = i\} \supset \{X_1 = j, X_0 = i\}.$$

Also gilt auch

$$\mathbb{P}\left(\{X_n \neq i \ \forall n\} \,|X_0 = i\right) \geq \mathbb{P}(\{X_1 = j\} \,|X_0 = i) = p_{ji} > 0.$$

Dann ist i aber nicht rekurrent.

▶ Die Umkehrung gilt, wenn C endlich ist.

20.4.7 Satz: Wenn C eine endliche abgeschlossene $\leftrightarrow$-Klasse ist, dann ist sie rekurrent (vgl. Norris [**?**, Seite 27]).

20.4.8 Satz: Der Zustandsraum einer homogenen Markow Kette lässt sich eindeutig in der Form

$$S = T \uplus C_1 \uplus C_2 \uplus ...$$

zerlegen. Dabei ist T die Menge der transienten Zustände der Kette und $C_1, C_2, ...$ sind die abgeschlossenen[3] rekurrenten Klassen.

Beweis: Grimmett und Stirzaker [14, 251]

20.4.9 Bemerkung (Grimmett und Stirzaker [14, 252]): Wenn die Markov-Kette in C_j beginnt, dann bleibt die Kette für immer in dieser abgeschlossenen rekurrenten Klasse. Wenn die Kette in T beginnt, dann bleibt die Kette u.U. für immer in T oder sie wechselt in eine rekurrente Klasse und bleibt dann dort für immer. Für den Fall $\#(S) < \infty$ gilt: Wenn die Kette in T startet, dann wechselt die Kette mit Wahrscheinlichkeit 1 in eine rekurrente Klasse.

[3] *abgeschlossenen* hätte man nicht angeben müssen, denn rekurrente Klassen sind abgeschlossen.

20.5 Stationäre Verteilung

20.5.1 Definition: Ein **Zeilen**vektor $\boldsymbol{\pi}^*$ heißt **stationäre Verteilung** der homogenen Markov-Kette, falls

i.) $\pi_j^* \geq 0$ für alle j und $\sum_{j\in S} \pi_j^* = 1$.
ii.) $\boldsymbol{\pi}^* = \boldsymbol{\pi}^* \mathbf{P}$.

20.5.2 Bemerkung: $\boldsymbol{\pi}^*$ ist/heißt wegen ii.) der Definition ein **Links**eigenvektor der Matrix $\mathbf{P}$ zum Eigenwert 1.

20.5.3 Bemerkung: Gilt $\mathbb{P}(X_0 = j) = \pi_j^*$, dann gilt auch $\mathbb{P}(X_n = j) = \pi_j^*$ (für alle n). Deshalb heißt eine stationäre Verteilung so wie sie heißt.

20.5.4 Satz: i.) Eine irreduzible homogene Markov-Kette hat genau dann eine stationäre Verteilung $\boldsymbol{\pi}^*$, wenn **alle Zustände positiv rekurrent** sind.

ii.) In diesem Fall ist die stationäre Verteilung $\boldsymbol{\pi}^*$ eindeutig bestimmt und es gilt

$$\pi_i^* = \frac{1}{\mu_i},$$

wobei

$$\mu_i = \mathbb{E}(T_i | X_0 = i)$$

die erwartete Dauer bis zur nächsten Rückkehr zu in i ist.

Beweis: Grimmet und Stirzaker [14, S. 254] und Norris [?, Seite 37]

20.6 Langfristverhalten

20.6.1 Bemerkung (Grimmett & Stirzaker [14, 259]): Wir wollen das Grenzverhalten $p_{ij}(n)$ für $n \to \infty$ untersuchen. Wie das folgenden Beispiel belegt, führt Periodizität zu einer Komplikation.[4] Es sei $S = \{1, 2\}$ und

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Dann alterniert $p_{ii}(n)$. Das verkompliziert die Asymptotik.

20.6.2 Satz (Grimmet und Stirzaker [14, S. 262]): Für jeden aperiodischen Zustand j einer homogenen Markov-Kette gilt

$$\lim_{n\to\infty} p_{jj}^{(n)} = \frac{1}{\mu_j}$$

und für jeden Zustand i gilt

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{f_{ij}}{\mu_j},$$

[4] Diese Kette hat aber einer eine stationäre Verteilung.

wobei

$$\mu_i = \mathbb{E}(T_i | X_0 = i)$$

und

$$f_{ij} = \mathbb{P}(\text{ es gibt ein } n \in \mathbb{N} \text{ mit } X_n = j | X_0 = i).$$

20.6.3 Bemerkung: Die Gleichung $\lim_{n\to\infty} p_{ij}^{(n)} = f_{ij}/\mu_j$ gilt auch für $i = j$. Wenn j rekurrent ist, dann ist $f_{jj} = 1$. Wenn $f_{jj} < 1$, dann ist j flüchtig. Dann folgt $\mu_j = \infty$.

▶ Die folgenden Sätze und Bemerkungen sind weniger allgemein als die Vorhergehenden, dafür etwas übersichtlicher (und leichter zu beweisen). Außerdem beachten wir die Bemerkung 20.4.9. Gemäß derer eine Kette .....

20.6.4 Satz: Für die Zustände einer irreduziblen[5] aperiodischen homogenen Markov-Kette gilt

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j}$$
$$\lim_{n\to\infty} \mathbb{P}(X_n = j) = \frac{1}{\mu_j}.$$

Hier muss man bemerken, dass die Grenzwerte unabhängig von i sind.

[5] Vgl. aber den Satz unten für eine weitreichendere Aussage ohne die Voraussetzung der Irreduzibilität.

Für null-rekurrente und transiente Zustände gilt die Aussage auch; mit $\mu_j = \infty$ und Grenzwert $0 = \frac{1}{\infty}$.

Beweis: Grimmet und Stirzaker [14, S. 259] Wir zeigen nur die Gültigkeit des zweiten Grenzwertes, wobei den ersten verwenden.

Es gilt

$$\mathbb{P}(X_n = j) = \sum_i \mathbb{P}(X_0 = i) p_{ij}^{(n)}.$$

Dann beobachten wir

$$\begin{aligned} \lim_{n\to\infty} \sum_i \mathbb{P}(X_0 = i) p_{ij}^{(n)} &= \sum_i \mathbb{P}(X_0 = i) \lim_{n\to\infty} p_{ij}^{(n)} \\ &= \sum_i \mathbb{P}(X_0 = i) \frac{1}{\mu_j} = \frac{1}{\mu_j} \end{aligned}$$

Also

$$\lim_{n\to\infty} \mathbb{P}(X_n = j) = \frac{1}{\mu_j}$$

20.6.5 Bemerkung: i.) Wenn die irreduzible aperiodische Kette transient oder null-rekurrent ist, dann ist also

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

ii.) Wenn die irreduzible aperiodische Kette positiv-rekurrent

ist also, dann

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j} = \pi_j^*,$$

wobei $\boldsymbol{\pi}^*$ ist die eindeutig bestimmte stationäre Verteilung ist (die es bei es bei positiv-rekurrent gibt).

iii.) Wenn die irreduzible Kette die Periode d hat, dann

$$p_{jj}^{(nd)} \to \frac{d}{\mu_j}.$$

20.6.6 Zusammenfassung für irreduzible aperiodische homogene Markov-Kette: Für eine irreduzible aperiodische homogene Markov-Kette gilt einer von drei Fälle:

1. Die Kette ist **transient** und $\lim_{n\to\infty} p_{ij}^{(n)} = 0$.

2. Die Kette ist nur **null-rekurrent**; es gibt keine stationäre Verteilung und $\lim_{n\to\infty} p_{ij}^{(n)} = 0$.

3. Die Kette ist **positiv-rekurrent**, es gibt eine stationäre Verteilung und

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{\mu_j} = \pi_j^*.$$

20.7 Ergodentheorem

20.7.1 Satz: Gegeben sei eine irreduzible Kette. Dann gilt

$$\mathbb{P}\left(\frac{V_i(n)}{n} \to \frac{1}{\mu_i} \text{ für } n \to \infty\right) = 1,$$

wobei

$$V_i(n) = \sum_{k=0}^{n-1} \mathbb{I}_{X_k=i}$$

die Anzahl der Besuche vor n ist, $\frac{V_i(n)}{n}$ ist der **Anteil der Zeit**, die die Kette vor n im Zustand i ist.

Beweis: vgl. Norris [**?**, Seite 53]

20.7.2 Bemerkung: Angenommen wir wollen einen Zeitpunkt N der *fernen* die Wahrscheinlichkeit $\mathbb{P}(X_N = i) = \frac{1}{\mu_i}$ **schätzen**. Wenn die Ketten schon lange läuft (d.h. n ist groß), dann können wir den Langschnitt-Durchschnitt $\frac{V_i(n)}{n}$ als Schätzer verwenden. Die Vorgehensweise ist von fundamentaler Bedeutung für die Zeitreihenanalyse (und wird oft vernachlässigt).

20.8 Hitting Wahrscheinlichkeiten, durchschnittliche Hitting Zeiten und durchschnittliche Rückkehrzeiten

▶ Die folgenden drei Sätze sind Anwendungen der sehr nützlichen First Step Analysis. Die Darstellung beruht auf Norris [?] und Privault [?] und ist noch nicht korrigiert (über Hinweise auf Fehler würde ich mich freuen).

20.8.1 Definition: Wir definieren Hitzeit

$$H^{\{i\}} = \inf\{n \geq 0 | X_n \in \{i\}\},$$

$H^{\{i\}}$ ist der früheste Zeitpunkt zudem die Kette $\{i\}$ besucht. Zudem definieren wir

$$h_l^{\{i\}} = \mathbb{P}(H^{\{i\}} < \infty | X_0 = l)$$
$$k_l^{\{i\}} = \mathbb{E}(H^{\{i\}} | X_0 = l)$$

20.8.2 Bemerkung: Wir beachten die folgenden Randbedingungen

$$h_i^{\{i\}} = 1$$
$$k_i^{\{i\}} = 0$$

Denn: wenn $X_0 = i$ ist, dann ist $H^{\{i\}} = \inf\{n \geq 0 | X_n \in \{i\}\} = 0$.

20.8.3 Satz: Der Vektor der Hitting Wahrscheinlichkeiten ist die minimale nicht-negative Lösung des folgenden linearen Gleichungssystems

$$\begin{aligned} h_i^{\{i\}} &= 1 \\ h_k^{\{i\}} &= \sum_{j \in S} p_{kj} h_j^{\{i\}} \quad k \neq i \end{aligned}$$

Beweis: vgl. Norris [?, Seite 13]

20.8.4 Satz: Der Vektor der mittleren Hitting Zeiten Hitting Wahrscheinlichkeiten ist die minimale nicht-negative Lösung des folgenden linearen Gleichungssystems

$$\begin{aligned} k_i^{\{i\}} &= 0 \\ k_k^{\{i\}} &= 1 + \sum_{j \neq i} p_{kj} k_j^{\{i\}} \quad k \neq i \end{aligned}$$

Beweis: vgl. Norris [?, Seite 17]

20.8.5 Satz: Für die Rückkehrzeiten gilt:

$$\mu_i = 1 + \sum_{k \neq i} P_{i,k} k_k^{\{i\}}.$$

Beweis: vgl. Privault [?, Seite 126f]

20.9 Endliche Ketten

20.9.1 Definition: Eine Markov-Kette heißt **endlich**, falls $N = \#(S) < \infty$.

20.9.2 Satz: Eine endliche homogene Markov-Kette hat mindestens eine rekurrenten Klasse und alle rekurrente Zustände sind positiv-rekurrent.

20.9.3 Satz: Für die Übergangsmatrix $\mathbf{P}$ einer endlichen homogenen Markov-Kette gilt:

i.) $\lambda = 1$ ist Eigenwert von $\mathbf{P}$ zum Eigenvektor $(1, 1, ..., 1)^T$.
ii.) Für alle Eigenwerte von $\mathbf{P}$ gilt $|\lambda| \leq 1$.

20.9.4 Satz (Perron-Frobenius): Für die Übergangsmatrix $\mathbf{P}$ einer irreduziblen endlichen Markov-Kette mit Periode d gilt:

i.) Die Eigenwerte vom Betrag 1 sind: $\lambda_1 = \omega^0, \lambda_2 = \omega^1, ..., \lambda_d = \omega^{d-1}$, wobei $\omega = e^{2\pi i/d}$.
ii.) Für alle anderen Eigenwerte $\lambda_{d+1}, ..., \lambda_N$ gilt $|\lambda_i| < 1, i = d+1, ..., N$.

Insbesondere: Im Falle der Aperiodizität $d = 1$ ist $\lambda = 1$ der einzige Eigenwert mit Betrag 1. Alle anderen Eigenwerte haben einen Betrag strikt kleiner 1.

Beweis: Grimmett und Stirzaker [14, Seite 269]

20.9.5 Satz: i.) Die Anzahl der persistenten Klassen einer homogenen Markov-Kette entspricht der Vielfachheit ν_1 des Eigenwertes $\lambda_1 = 1$, wobei die algebraische und die geometrische Vielfachheit übereinstimmen. (Vgl. Karlin und Taylor [**?**, Seite 4])

ii.) Die Übergangsmatrix einer endlichen homogenen Markov-Kette kann durch die geeignete Wahl der Zustandsnummerierung in der Form

$$\mathbf{P} = \begin{pmatrix} P_1 & 0 & ... & 0 & 0 \\ 0 & P_2 & ... & 0 & 0 \\ : & : & : & : & : \\ 0 & 0 & ... & P_{\nu_1} & 0 \\ Q_1 & Q_2 & ... & Q_{\nu_1} & Q \end{pmatrix}.$$

angegeben werden. Dabei korrespondieren die Zeilen mit den Matrizen $P_1, ..., P_{\nu_1}$ zu den persistenten Klassen und die Zeilen mit den Q-Matrizen zu den flüchtigen Zuständen.

21 Brown'sche Bewegung

21.1 Brown'schen Bewegung als skalierte geraffte Irrfahrt

Dieses Kapitel orientiert sich an Shreve [33, Kapitel 3].

21.1.1 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Ein zeit-stetiger stochastischer Prozess $B_t, t \in [0, \infty)$ heißt **Brown'sche Bewegung**, falls:

i.) $B_0 = 0$.
ii.) $B_t(\omega)$ ist für alle ω eine **stetige** Funktion der Zeit t.
iii.) Für alle $0 = t_0 < t_1 < ... < t_m$ sind die Zuwächse (bzw. Inkremente) $B_{t_1} - B_{t_0}, ..., B_{t_m} - B_{t_{m-1}}$ **unabhängig** und **normal-verteilt** mit

$$\mathbb{E}(B_{t_i} - B_{t_{i-1}}) = 0,$$
$$\mathbb{V}(B_{t_i} - B_{t_{i-1}}) = t_i - t_{i-1}.$$

also

$$B_{t_i} - B_{t_{i-1}} \sim \text{uiN}(\mu = 0, \sigma = \sqrt{t_i - t_{i-1}}).$$

▶ Wir betrachten als diskrete Vorstufe zur zeit-stetigen BB die einfach symmetrische Irrfahrt.

21.1.2 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und für $n \in \mathbb{N}$

$$S_n = S_0 + \sum_{i=1}^{n} X_i$$

eine einfache symmetrische Irrfahrt mit $S_0 = 0$ und $\mathbb{P}(X_i = 1) = \frac{1}{2} = \mathbb{P}(X_i = -1)$. Dann gilt:

i.) Für alle $0 = k_0 < k_1 < ... < k_m$ sind die Zuwächse $S_{k_1} - S_{k_0}, ..., S_{k_m} - S_{k_{m-1}}$ unabhängig.
ii.) $\mathbb{E}(S_{k_i} - S_{k_{i-1}}) = 0$.
iii.) $\mathbb{V}(S_{k_i} - S_{k_{i-1}}) = k_i - k_{i-1}$.
iv.) $S_{k_i} - S_{k_{i-1}}$ ist Binomialverteilt mit $k_i - k_{i-1}$ Wiederholungen und Erfolgswahrscheinlichkeit $\frac{1}{2}$

Beweis: Übung

21.1.3 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum

und

$$S_n = S_0 + \sum_{i=1}^{n} X_i, n \in \mathbb{N}$$

eine einfache symmetrische Irrfahrt mit $S_0 = 0$. Dann gilt für alle $l > k$:

$$\mathbb{E}(S_l|\mathcal{F}_k) = S_k.$$

Dabei bezeichnet $\mathcal{F}_k$ die von $X_1, ..., X_k$ erzeugte σ-Algebra und $\mathbb{E}(S_l|\mathcal{F}_k)$ die bedingte Erwartung bezüglich $\mathcal{F}_k$.

S_n hat also die sogenannte **Martingal-Eigenschaft**.

21.1.4 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und

$$S_n = S_0 + \sum_{i=1}^{n} X_i, n \in \mathbb{N}$$

eine einfache symmetrische Irrfahrt mit $S_0 = 0$ und $\mathbb{P}(X_i = 1) = \frac{1}{2} = \mathbb{P}(X_i = -1)$.. Dann gilt für die **quadratische Variation** von $(S_n)_{n\in\mathbb{N}_0}$

$$[S, S]_k := \sum_{j=1}^{k} (S_j - S_{j-1})^2 = k.$$

Beweis: Beachte $X_j^2 = 1$.

21.1.5 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und

$$S_n = S_0 + \sum_{i=1}^{n} X_i, n \in \mathbb{N}$$

eine einfache symmetrische Irrfahrt mit $S_0 = 0$. Für alle $n \in \mathbb{N}$ definieren wir, falls $nt \in \mathbb{N}_0$ gilt,

$$W_t^{(n)} = \frac{1}{\sqrt{n}} S_{nt}, t \geq 0.$$

Ist $nt \notin \mathbb{N}_0$, dann bestimmen wir das größte $s < t$ mit $ns \in \mathbb{N}_0$ und das kleinste $u > t$ mit $nu \in \mathbb{N}_0$ und definieren $W_t^{(n)}$ durch lineare Interpolation:

$$W_t^{(n)} = W_s^{(n)} + (W_u^{(n)} - W_s^{(n)}) \cdot n \cdot (t - s)$$

bzw.

$$\begin{aligned}
W_t^{(n)} &= W_{\frac{\lfloor nt \rfloor}{n}}^{(n)} + \left(W_{\frac{\lfloor nt \rfloor + 1}{n}}^{(n)} - W_{\frac{\lfloor nt \rfloor}{n}}^{(n)} \right) \cdot n \cdot \left(t - \frac{\lfloor nt \rfloor}{n} \right) \\
&= \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor} + \left(\frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor + 1} - \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor} \right) \cdot n \cdot \left(t - \frac{\lfloor nt \rfloor}{n} \right) \\
&= \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor} + \frac{1}{\sqrt{n}} \left(X_{\lfloor nt \rfloor + 1} \right) \cdot n \cdot \left(t - \frac{\lfloor nt \rfloor}{n} \right) \\
&= W_{\frac{\lfloor nt \rfloor}{n}}^{(n)} + \frac{1}{\sqrt{n}} \left(X_{\lfloor nt \rfloor + 1} \right) \cdot (nt - \lfloor nt \rfloor) \\
&= W_{\frac{\lfloor nt \rfloor}{n}}^{(n)} + \sqrt{n} \cdot \left(t - \frac{\lfloor nt \rfloor}{n} \right) \cdot X_{\lfloor nt \rfloor + 1}
\end{aligned}$$

$W_t^{(n)}$ heißt **skalierte symmetrische einfache Irrfahrt**.

21.1.6 Bemerkung: i.) Die Skalierung und der Zeitraffer sind unübersichtlich. Um diese Transformation intuitiv zu erfassen, beobachten wir

$$W_1^{(n)} = \frac{1}{\sqrt{n}} S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

Dann ist

$$\mathbb{V}(W_1^{(n)}) = \left(\frac{1}{\sqrt{n}}\right)^2 n = 1$$

Die Varianz pro Zeiteinheit ist demnach 1.

ii.) Die Skalierung und der Zeitraffer sind unübersichtlich. Wir betrachten ein Zahlenbeispiel. Wir wählen $n = 52$ und $t = 1.5$. Dann ist $nt = 78 \in \mathbb{N}$. Also ist

$$W_{1.5}^{(52)} = \frac{1}{\sqrt{52}} S_{78}.$$

Jetzt betrachten wir $t = 1.7$. Dann ist $nt = 88.4 \notin \mathbb{N}$. Also $\lfloor nt \rfloor = 88$, $s = \frac{88}{52} = 1.692...$, $u = \frac{89}{52} = 1.711....$ Weiterhin

$$W_{1.7}^{(52)} = W_{\frac{88}{52}}^{(52)} + \left(W_{\frac{89}{52}}^{(52)} - W_{\frac{88}{52}}^{(52)}\right) \cdot 52 \cdot \left(t - \frac{88}{52}\right).$$

Wir machen noch eine Probe

$$\begin{aligned} W_{\frac{89}{52}}^{(52)} &= W_{\frac{88}{52}}^{(52)} + \left(W_{\frac{89}{52}}^{(52)} - W_{\frac{88}{52}}^{(52)}\right) \cdot 52 \cdot \left(\frac{89}{52} - \frac{88}{52}\right) \\ &= W_{\frac{88}{52}}^{(52)} + \left(W_{\frac{89}{52}}^{(52)} - W_{\frac{88}{52}}^{(52)}\right). \end{aligned}$$

21.1.7 Bemerkung: Drei *Maßnahmen* charakterisieren die skalierte symmetrische Irrfahrt im Vergleich zur einfachen symmetrischen Irrfahrt:

i.) Die Zeit wird **gerafft**. Der Wert von $W^{(n)}$ zum Zeitpunkt t ergibt sich aus dem Wert der einfachen symmetrischen Irrfahrt $S_\cdot$ zum Zeitpunkt nt.

Die Zeitraffung kann man an einem Beispiel leicht erfassen. Angenommen $n = 200$. Wenn dann die Zeit t das Einheitszeitintervall von 0 bis 1 durchläuft, dann erhalten wir 200 Beobachtungen. Also $n = 200$ Werte im Einheitszeitintervall.
ii.) Die Pfade werden durch die Multiplikation mit $\frac{1}{\sqrt{n}}$ **zusammengedrückt/skaliert**.
iii.) Die Pfade werden zu einem **zeit-stetigen Prozess fortgesetzt**.

▶ Bis auf Grenzübergang *ist* eine skalierte symmetrische einfache Irrfahrt eine Brown'sche Bewegung. Wir fassen die Eigenschaften der skalierte symmetrische einfache Irrfahrt zusammen, die für die Brown'sche Bewegung ebenfalls gelten.

21.1.8 Satz: Für die skalierte symmetrische Irrfahrt gilt:

i.) Für $0 = t_0 < t_1 < ... < t_m$ mit $nt_i \in \mathbb{N}_0, i = 0, ..., m$ sind

die **Zuwächse**

$$W_{t_1}^{(n)} - W_{t_0}^{(n)}, ..., W_{t_m}^{(n)} - W_{t_{m-1}}^{(n)}$$

unabhängig.

ii.) [**Erwartungswet Null**] Für $0 \leq s \leq t$ mit $nt, ns \in \mathbb{N}_0$:

$$\mathbb{E}(W_t^{(n)} - W_s^{(n)}) = 0.$$

iii.) [**Varianz gleich Länge Zeitintervall**] Für $0 \leq s \leq t$ mit $nt, ns \in \mathbb{N}_0$:

$$\mathbb{V}(W_t^{(n)} - W_s^{(n)}) = t - s.$$

iv.) [**Martingal-Eigenschaft**] Für $0 \leq s \leq t$ mit $nt, ns \in \mathbb{N}_0$:

$$\mathbb{E}(W_t^{(n)} | \mathcal{F}_{ns}) = W_s^{(n)}.$$

v.) [**Quadratische Variation gleich** t] Für alle $t \geq 0$ mit $nt \in \mathbb{N}_0$ gilt

$$[W^{(n)}, W^{(n)}]_t = t.$$

▶ Der folgende Satz ist ungenau formuliert. .....

21.1.9 Satz: Die skalierte symmetrische Irrfahrt *konvergiert*[1] für $n \to \infty$ gegen eine Brown'sche Bewegung.

Beweis: Shreve [33, S. 89]

▶ Für spätere Betrachtungen und für die Simulation ist eine alternative Repräsentation der skalierte symmetrische Irrfahrt nützlich, die wir jetzt einführen.

21.1.10 Definition: Es sei z_k eine **symmetrische Bernoulli** Zufallsvariablen mit $\text{Bild}(z_k) = \{-1, 1\}$ und $\mathbb{P}(z_k = 1) = \mathbb{P}(z_k = -1) = \frac{1}{2}$. Wir schreiben dann $z_k \sim \text{SymBern}$. Den kleinen Buchstaben z reservieren wir für **SymBern** verteilte Zufallsvariablen.

21.1.11 Definition und Satz: $W_t^{\Delta t}$ für $t = t_k = k\Delta t, k = 0, 1, 2, ...$ heißt **skalierte symmetrische Irrfahrt**, wenn:

- Für $k \in \mathbb{N}$ ist $z_k \sim \text{iiBern}$.
- $W_0^{\Delta t} = 0$ und iterativ für $k = 0, 1, 2, ...$

$$W_{(k+1)\Delta t}^{\Delta t} = W_{k\Delta t}^{\Delta t} + \sqrt{\Delta t} \cdot z_{k+1} = W_{k\Delta t}^{\Delta t} + X_{k+1},$$

wobei $X_j = \sqrt{\Delta t} \cdot z_j$

Wenn man ein hinreichend kleines Δt wählt, dann liefert die

[1] Das ist ungenau, denn die *Form* der Konvergenz ist unklar.

skalierte symmetrische Irrfahrt Y_t für $t_k = k\Delta t, k = 0, 1, 2, ...$ eine gute Approximation für eine Brown'sche Bewegung; also

$$W_{t_k} \approx W_{t_k}^{\Delta t}, t_k = k\Delta t, k = 0, 1, 2, ...$$

Die X_j erfassen die Zuwächse für die Zeitspanne Δt. Es gilt $\mathbb{E}(X_j) = 0$ und $\mathbb{V}(X_j) = \mathbb{V}(\sqrt{\Delta t} \cdot z_j) = \Delta t$. Bei der skalierten symmetrischen Irrfahrt sind die Zuwächse Bernoulli verteilt. Bei der Brown'schen Bewegung sind die Zuwächse normal verteilt. Erwartungswert und Varianz sind gleich. In der Wahrscheinlichkeitstheorie lernt man, dass die Summe von Bernoulli Variablen im Grenzübergang normal verteilt ist.

21.1.12 Bemerkung: Wir haben in der vorhergehenden Definition eine scheinbar alternative Definition für eine skalierte symmetrische Irrfahrt an. Wenn man $\Delta t = \frac{1}{n}$ und $k = n \cdot t$ (für solche t, so dass $n \cdot t \in \mathbb{N}$) beachtet, dann erkennt man, dass die Definitionen übereinstimmten.

Notiz:

$$\Delta t = \frac{1}{n}, \sqrt{\Delta t} = \frac{1}{\sqrt{n}}$$

Also für $t = t_k$

$$W_t^{(n)} = \frac{1}{\sqrt{n}} S_{nt} = \sqrt{\Delta t} \cdot S_k = W_{k\Delta t}^{\Delta t} = W_{t_k}^{\Delta t}$$

21.1.13 Fallstudie: Wir simulieren jeweils 25 skalierte symmetrische Irrfahrten bzw. Brown'sche Bewegungen. Zunächst generieren wir jeweils 25 Pfade mit 52 Schritten für eine Zeiteinheit [1J]. Dann generieren wir jeweils 25 Pfade mit 250 Schritten und schließlich mit 750 Schritten. Bei 25 Schritten ist der Unterschied zwischen der symmetrischen skalierten Irrfahrt und der Brown'sche Bewgung sehr auffällig.

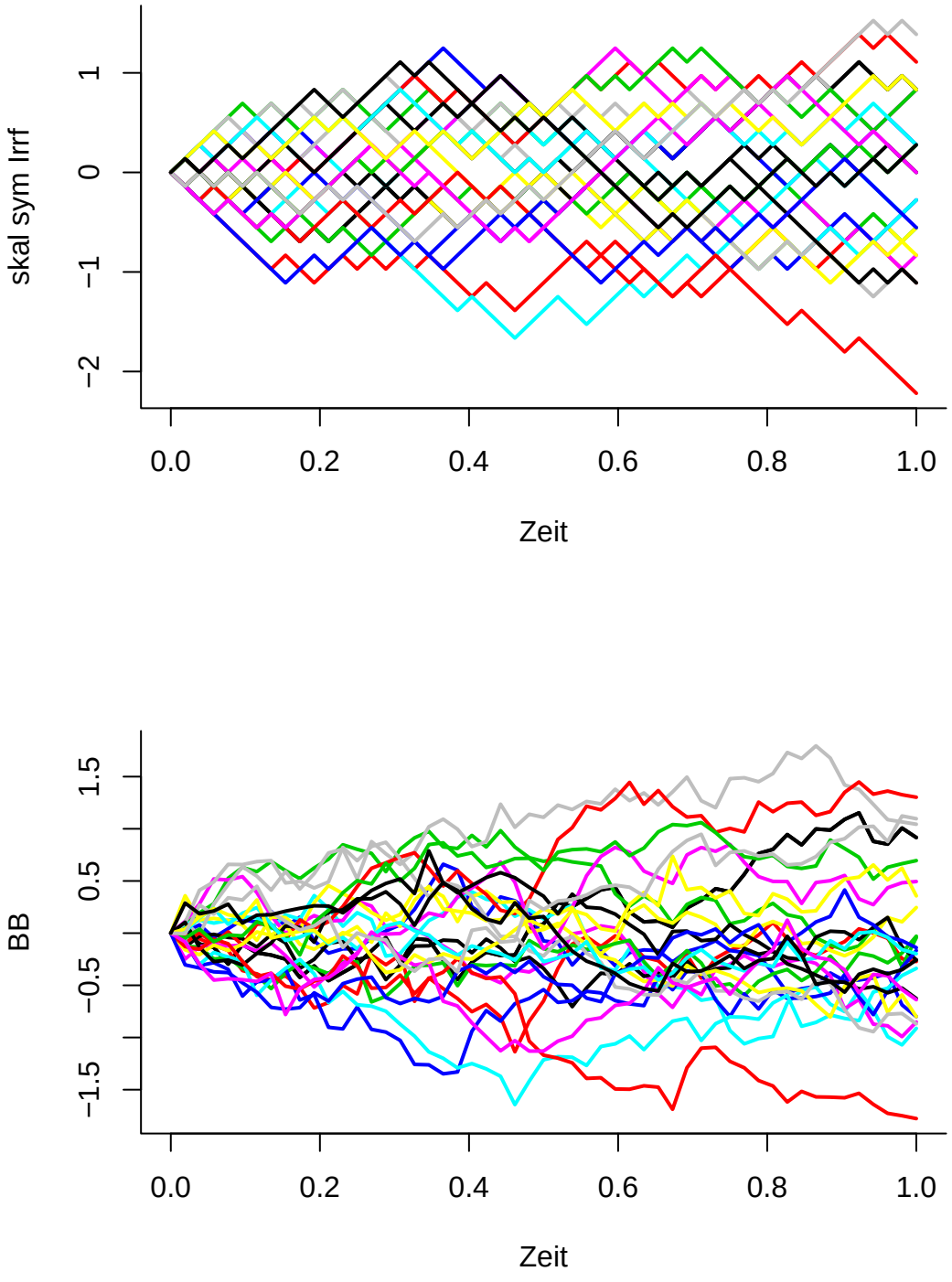


Abbildung 21.1: 25 Pfade von symmetrischen Irrfahrten bzw. von Brown'schen Bewegungen mit 52 Schritten

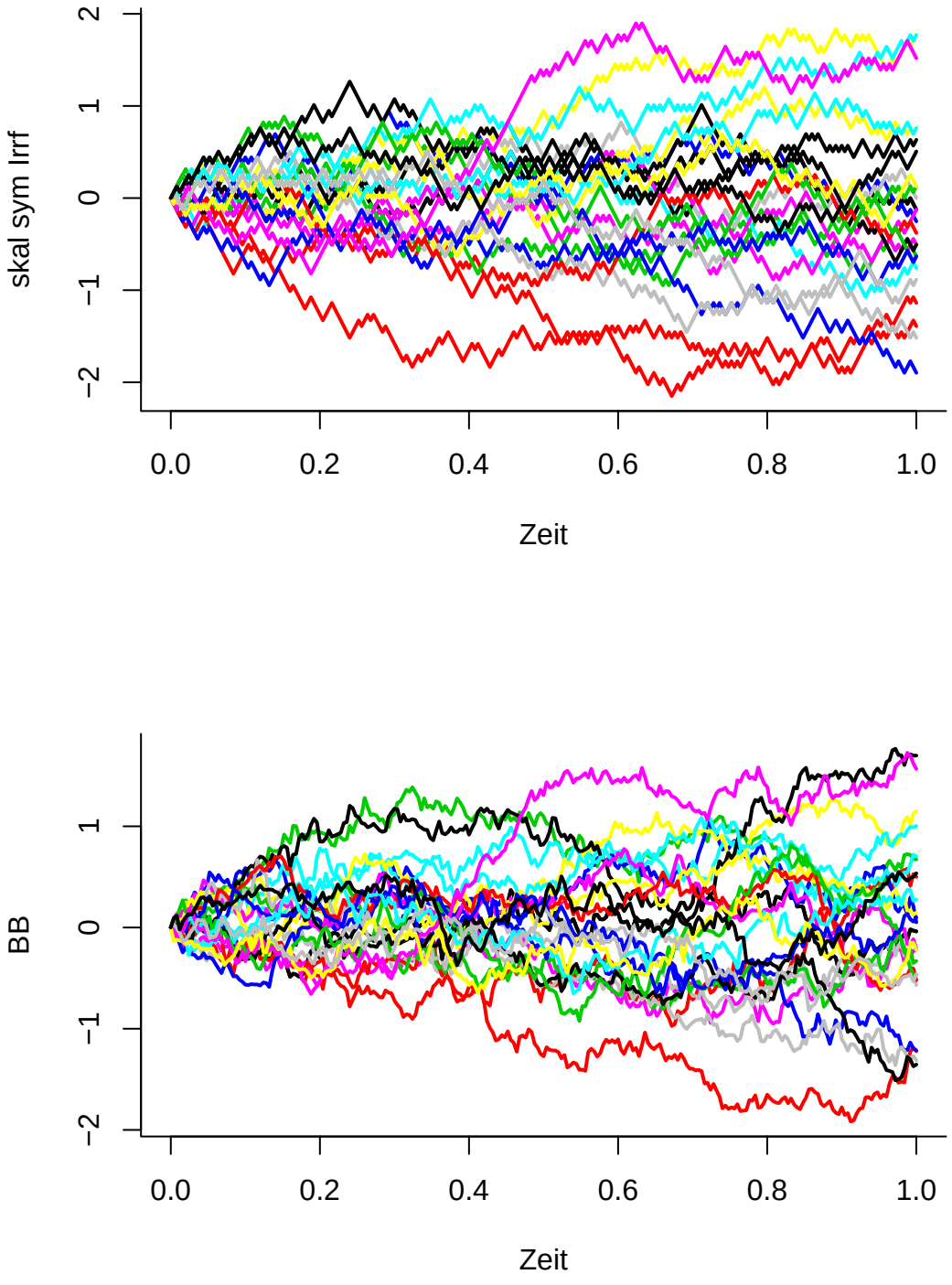


Abbildung 21.2: 25 Pfade von symmetrischen Irrfahrten bzw. von Brown'schen Bewegungen mit 250 Schritten

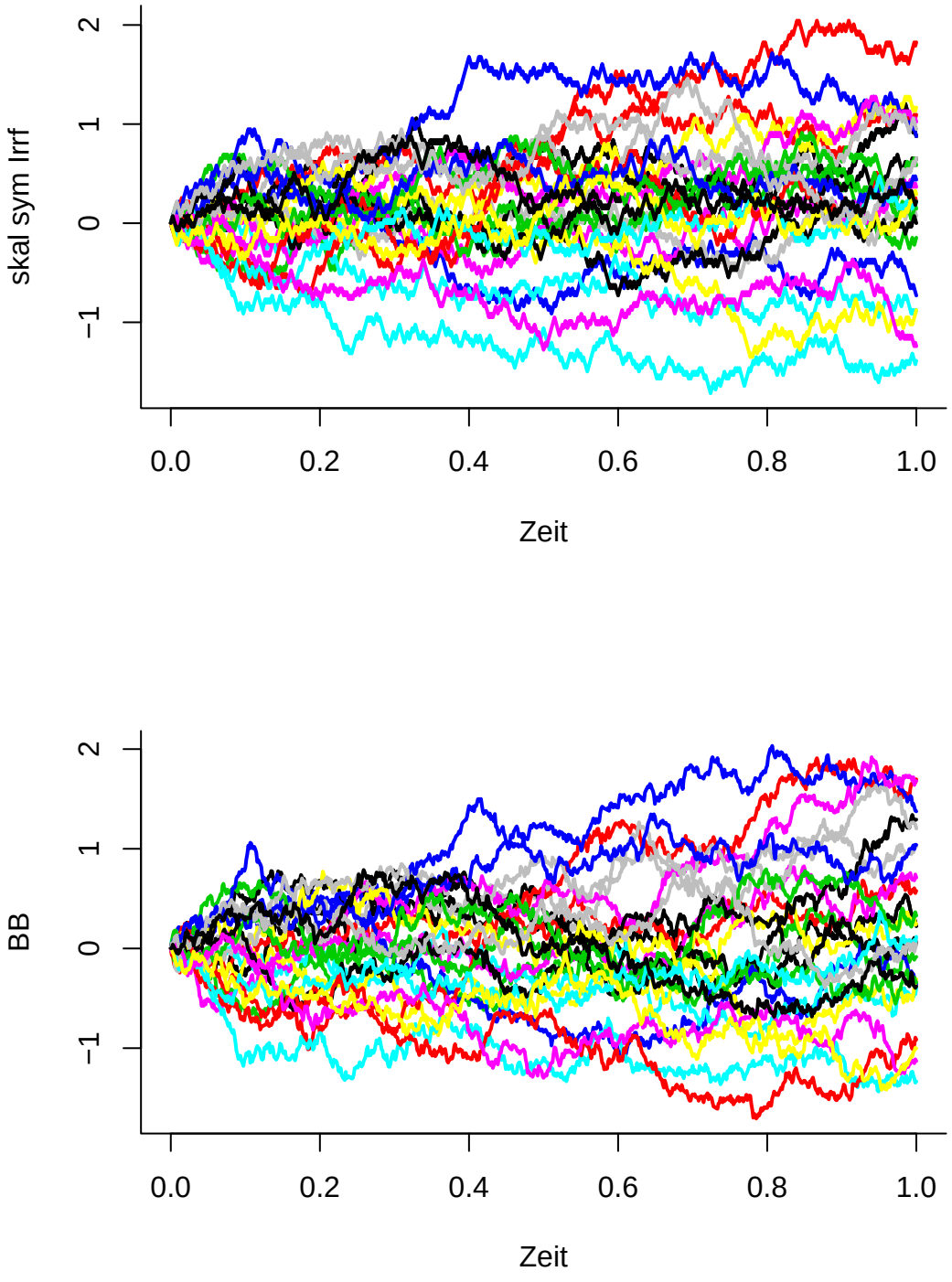


Abbildung 21.3: 25 Pfade von symmetrischen Irrfahrten bzw. von Brown'schen Bewegungen mit 750 Schritten

Gemäß Definition hat die Brown'sche Bewegung W_1 zum Zeitpunkt $t = 1$ eine Standard-Normalverteilung. Das Histogramm bzw. die geschätzte empirische Dichte der Endwerte der Pfade bei $t = 1$ müsste also näherungsweise wie die Dichte der Standard-Normalverteilung aussehen. Bisher haben wir nur 25 Pfade betrachtet. Um die zuletzt genannte Approximation zu sehen, müssen wir deutlich mehr Pfade betrachten. Der folgende Plot zeigt die Histogramme bzw. die empirische Dichte für skalierte Irrfahrt und für die simulierte Brown'sche Bewegung für 100000 Pfade und außerdem die Standard-Normalverteilung.

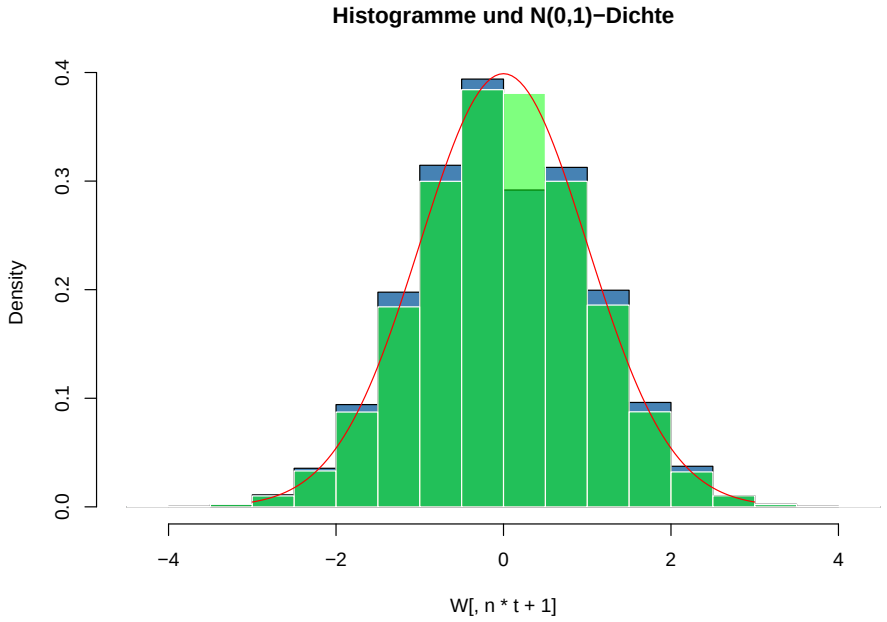


Abbildung 21.4: Histogramme der Endwerte der Pfade von W und W^{Δ} sowie die Dichte der Standard-Normalverteilung

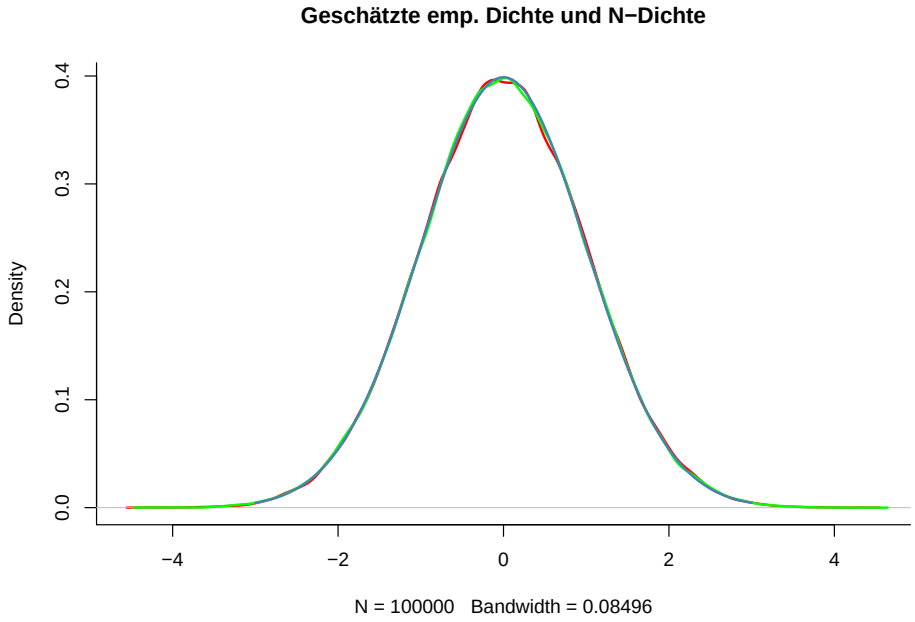


Abbildung 21.5: Geschätzte empirische Dichten der Endwerte der Pfade von W und W^{Δ} sowie die Dichte der Standard-Normalverteilung

21.1.14 Bemerkung: Aus der Fallstudie ergibt sich die folgenden naheliegende Interpretation für die Brown'sche Bewegung: W erfasst die hochfrequenten kumulierten unabhängigen **auf-und-ab Impulse**. Bei der unskalierten Irrfahrt sind die Impulse nur grob durch 1 bzw. -1 erfasst. Zudem werden die Impulse nur zu diskreten Zeitpunkte mit der *Frequenz* (Abtastrate) $1/\Delta t$ pro Jahr erfasst. Die Brown'sche Bewegung ergibt sich durch Grenzübergang $1/\Delta t \to \infty$, d.h. bei kontinuierlicher Abtastung. Durch die Skalierung der Irrfahrt – die sich durch die Multiplikation der z mit $\sqrt{\Delta t}$ ergibt –

wird die 1-Jahres Varianz auf 1 normiert.

Die Brown'sche Bewegung selbst taugt (noch) nicht als Modell für die Aktienkurs. Dafür verwendet wir die geometrische Brown'sche Bewegung.

21.2 Die Martingal-Eigenschaft der BB

21.2.1 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und (B_t) eine Brown'sche Bewegung. Eine **Filtration für die Brown'sche Bewegung** ist eine Familie $\mathcal{F}_t, t \geq 0$ von σ-Algebren mit

i.) Für $0 \leq s < t$ gilt $\mathcal{F}_s \subset \mathcal{F}_t$.
ii.) Für alle $t \geq 0$ ist B_t bezüglich $\mathcal{F}_t$ meßbar.
iii.) Für alle $0 \leq t < u$ ist der Zuwachs $B_u - B_t$ von $\mathcal{F}_t$ unabhängig.

Wenn $\mathcal{F}_t, t \geq 0$ eine Filtration für (B_t) ist, dann sagen wir auch, dass (B_t) eine Brown'sche Bewegung mit Filtration $\mathcal{F}_t, t \geq 0$ ist.

21.2.2 Satz: Es sei (B_t) eine Brown'sche Bewegung und $(\mathcal{F}_t)$ eine Filtration für (B_t). Dann erfüllt die Brown'sche Bewegung die Martingal Eigenschaft:

$$\mathbb{E}(B_t|\mathcal{F}_s) = B_s, 0 \leq s \leq t$$

Beweis: Übung

21.3 Quadratische Variation

21.3.1 Definition: Es sei $f : [0,T] \to \mathbb{R}$ eine reellwertige Funktion und $\mathcal{Z} = \{t_0, ..., t_n\}, 0 = t_0 < ... < t_n = T$ eine Zerlegung des Intervalls $[0,T]$ mit der Feinheit $||\mathcal{Z}|| := \max\{|t_i - t_{i-1}| : i = 1, ..., n\}$. Wenn der Grenzwert

$$[f,f]_T = \lim_{||\mathcal{Z}||\to\infty} \sum_{i=1,...,n} (f(t_i) - f(t_{i-1}))^2$$

existiert, dann heißt $[f,f]_T$ die **quadratische Variation** von f.

21.3.2 Satz: Ist $f \in C^1[0,T]$ stetig differenzierbar, so gilt $[f,f]_T = 0$.

Beweis: Shreve [33, S. 101].

21.3.3 Satz: Es sei (B_t) eine Brown'sche Bewegung. Dann gilt für die quadratische Variation

$$[B,B]_T = T \quad \mathbb{P}\text{-f.s.}$$

Beweis: Shreve [33, S. 102].

21.3.4 Bemerkung (Shreve [33, S. 104]): Es sei (B_t) eine Brown'sche Bewegung. Wir haben nachgewiesen (Übung oder Shreve [33, S. 103, Gl. (3.4.6) bzw. (3.6.7)]), dass

$$\mathbb{E}[(B_{t_i} - B_{t_{i-1}})^2] = t_i - t_{i-1}, \tag{21.1}$$

$$\mathbb{V}[(B_{t_i} - B_{t_{i-1}})^2] = 2(t_i - t_{i-1})^2. \tag{21.2}$$

Wenn $t_i - t_{i-1}$ klein ist, dann ist $2(t_i - t_{i-1})^2$ sehr klein. Manche argumentieren, dass deshalb – oder aus anderen (schlechten) Gründen – für kleine Intervalllängen $t_i - t_{i-1}$ näherungsweise

$$(B_{t_i} - B_{t_{i-1}})^2 = t_i - t_{i-1}.$$

gilt. Das würde bedeuten, dass für kleine Intervalllängen die zufälligen quadrierten stochastischen Zuwachse $(B_{t_i} - B_{t_{i-1}})^2$ durch fixe (nicht-stochastische) Werte $t_i - t_{i-1}$ ersetzen können. Das ist jedoch nicht der Fall. Wenn $t_i - t_{i-1}$ eine Approximation für $(B_{t_i} - B_{t_{i-1}})^2$ wäre, dann sollte für kleine $t_i - t_{i-1}$ näherungsweise

$$\frac{(B_{t_i} - B_{t_{i-1}})^2}{t_i - t_{i-1}} = 1$$

gelten. Dazu machen wir jedoch die folgende Beobachtung. Die Zufallsvariablen $B_{t_i} - B_{t_{i-1}}$ sind normalverteilt mit dem Erwartungswert 0 und die Varianz $t_i - t_{i-1}$.

Die Zufallsvariablen $\frac{B_{t_i}-B_{t_{i-1}}}{\sqrt{t_i-t_{i-1}}}$ haben deshalb die Varianz 1 und den Erwartungswert 0. Die Varianz ist 1 egal wie klein $t_i - t_{i-1}$ ist.

Der Quotient

$$\frac{(B_{t_i} - B_{t_{i-1}})^2}{t_i - t_{i-1}}$$

ist eine **Zufallsvariable** deren Verteilung nicht von der Intervalllänge $t_i - t_{i-1}$ abhängig ist. In der Tat ist diese Zufallsvariable χ^2 verteilt mit FG 1. Die Varianz ist 1, vollkommen unabhängig von der Schrittweite $t_i - t_{i-1}$.

Man kann also die einzelnen Zuwächse $(B_{t_i} - B_{t_{i-1}})^2$ keineswegs durch die nicht-stochastische Intervalllänge $t_i - t_{i-1}$ ersetzen; auch dann nicht wenn $t_i - t_{i-1}$ winzig ist.

Für die kumulierten quadrierten Zuwächse hingegen gilt (gemäß GgZ) die Approximation

$$\sum_{i=1,\ldots,k} (B_{t_i} - B_{t_{i-1}})^2 \approx t_n - t_0 \text{ f.s.}$$

für hinreichend kleine Feinheit. Das ist auch genau das, was wir brauchen. Später wird die quadratische Variation bei der Ito Integration auftauchen und bei der Integration werden kumulierte Zuwächse analysiert.

21.3.5 Bemerkung: Anstatt – insbesondere im sogenann-

ten stochastischen Kalkül, den wir später besprechen werden

–

$$\lim_{||\mathcal{Z}||\to\infty} \sum_{i=1,\ldots,n} (B(t_i) - B(t_{i-1}))^2 = [B,B]_T = T \quad \mathbb{P}\text{-f.s.}$$

schreibt man oft

$$dBdB = dt$$

bzw.

$$(dB)^2 = dt.$$

$$\int_0^T (dB)^2 = \int_0^T 1 \cdot dt = T$$

obwohl insbesondere $dBdB = (dB)^2 = dt$ irreführend ist.

21.4 Markov-Eigenschaft der BB

21.4.1 Satz (Shreve [33, Seite 107]): Wenn (B_t) eine Brown'sche Bewegung ist und $\mathcal{F}_t$ eine Foltration für (B_t), dann ist (B_t) ein **Markov Prozess**. Es gibt also (gemäß Definition) für jede messbare Funktion f und $0 \leq s \leq t$ eine messbare Funktion g mit

$$\mathbb{E}(\, f(W_t) \,|\, \mathcal{F}_s \,) = g(W_s).$$

Anschaulich: $\mathcal{F}_s$ umfasst die Inforation **bis** s (die Geschichte bis s). Für jede Transformation $f(W_t)$ benötigt man für die Prognose $\mathbb{E}(\, f(W_t) \,|\, \mathcal{F}_s \,)$ nur den aktuellen Wert W_s der Brown'schen Bewegung.

21.5 Übersicht zur Brown'sche Bewegung

21.5.1 Satz: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Ein zeit-stetiger stochastischer Prozess $B_t, t \in [0, \infty)$ heißt **Brown'sche Bewegung**, falls:

i.) $B_0 = 0$.
ii.) $B_t(\omega)$ ist für alle ω eine **stetige** Funktion der Zeit t.
iii.) Für alle $0 = t_0 < t_1 < ... < t_m$ sind die Zuwächse/Inkremente $B_{t_1} - B_{t_0}, ..., B_{t_m} - B_{t_{m-1}}$ **unabhängig** und **normal-verteilt** mit

$$\mathbb{E}(B_{t_i} - B_{t_{i-1}}) = 0, \tag{21.3}$$

$$\mathbb{V}(B_{t_i} - B_{t_{i-1}}) = t_i - t_{i-1}. \tag{21.4}$$

21.5.2 Satz: Die skalierte symmetrische Irrfahrt konvergiert für $n \to \infty$ gegen eine Brown'sche Bewegung.

21.5.3 Satz: Es sei (B_t) eine Brown'sche Bewegung und $(\mathcal{F}_t)$ eine Filtration für (B_t). Dann erfüllt die Brown'sche

Bewegung die Martingal Eigenschaft:

$$\mathbb{E}(B_t|\mathcal{F}_s) = B_s, 0 \leq s \leq t.$$

21.5.4 Satz: Es sei (B_t) eine Brown'sche Bewegung. Dann gilt für die quadratische Variation

$$[B, B]_T = T \quad \mathbb{P}\text{-f.s.}$$

Wir werden später kompakt

$$dBdB = (dB)^2 = dt$$

schreiben.

21.5.5 Satz (Shreve [33, Seite 107]): Wenn (B_t) eine Brown'sche Bewegung ist und $\mathcal{F}_t$ eine Foltration für (B_t), dann ist (B_t) ein **Markov Prozess**. Es gibt also (gemäß Definition) für jede messbare Funktion f und $0 \leq s \leq t$ eine messbare Funktion g mit

$$\mathbb{E}(\, f(W_t) \,|\, \mathcal{F}_s \,) = g(W_s).$$

Anschaulich: $\mathcal{F}_s$ umfasst die Inforation **bis** s (die Geschichte bis s). Für jede Transformation $f(W_t)$ benötigt man für die Prognose $\mathbb{E}(\, f(W_t) \,|\, \mathcal{F}_s \,)$ nur den aktuellen Wert W_s der Brown'schen Bewegung.

22 Martingale

22.1 Martingale in diskreter Zeit

Wir orientieren uns an Henze [19, S. 178f].

22.1.1 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine Familie $\mathbb{F} = (\mathcal{F}_k)_{k\in\mathbb{N}_0}, \mathcal{F}_k \subset \mathcal{F}$ von Sub-σ-Algebren heißt **Filtration**, falls $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ für alle k gilt.

Eine Folge von Zufallsvariablen $(X_n)_{n\in\mathbb{N}_0}$ heißt $\mathbb{F}$-**adaptiert**, falls $X_n \in \mathcal{F}_n$ für jedes $n \in \mathbb{N}$ gilt, d.h. X_n ist $\mathcal{F}_n$-meßbar.

Für eine Folge von Zufallsvariablen $(X_n)_{n\in\mathbb{N}_0}$ heißt $\mathcal{F}_n^X = \sigma(X_0, X_1, ..., X_n)$ die **natürliche Filtration** von $(X_n)_{n\in\mathbb{N}_0}$.

22.1.2 Definition: Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\mathbb{F} = (\mathcal{F}_k)_{k\in\mathbb{N}_0}$ eine Filtration in $\mathcal{F}$. Ein zeit-diskreter adaptierter integrierbarer stochastischer Prozess $(X_k)_{k\in\mathbb{N}_0}, X_k \in$

$\mathcal{L}^1(\mathbb{P})$ heißt $\mathbb{F}$-**Martingal**, falls für alle $m, n \in \mathbb{N}, m < n$

$$\mathbb{E}(X_n|\mathcal{F}_m) = X_m \quad \mathbb{P}\text{-f.s.}$$

22.1.3 Satz (Doob'sches Martingal): Es sei $X \in \mathcal{L}^1$ und $\mathbb{F} = (\mathcal{F}_k)_{k\in\mathbb{N}_0}$ eine Filtration. Dann ist

$$X_n = \mathbb{E}(X|\mathcal{F}_n), n \geq 0$$

ein Martingal.

Beweis: Siehe Henze [19, S. 179]

22.1.4 Definition: Ein Prozess $H = (H_k)_{k\in\mathbb{N}_0}$ heißt $\mathbb{F}$-**vorhersehbar oder prävisibel**, falls H_0 eine Konstante ist und H_n für alle $n \in \mathbb{N}$ bezüglich $\mathcal{F}_{n-1}$ messbar ist.

Was ist der Unterschied zu *adaptiert*?

22.1.5 Satz (Doobsche Zerlegung): Es sei $\mathbb{F} = (\mathcal{F}_k)_{k\in\mathbb{N}_0}$ eine Filtration und $(X_k)_{k\in\mathbb{N}_0}$ ein adaptierter integrierbarer Prozess. Dann gibt es eine eindeutig bestimmte Darstellung

$$X_n = M_n + V_n,$$

wobei M_n ein Martingal, V_n prävisibel und $V_0 = 0$ ist.

Beweis: Henze [19, S. 179]

22.1.6 Definition: Es sei $\mathbb{F} = (\mathcal{F}_k)_{k\in\mathbb{N}_0}$ eine Filtration, $(X_k)_{k\in\mathbb{N}_0}$ ein adaptierter Prozess und $(H_k)_{k\in\mathbb{N}}$ ein prävisibeler Prozess. Dann heißt

$$(H \bullet X)_{n\in\mathbb{N}} = \sum_{j=1}^{n} H_j \Delta X_j = \sum_{j=1}^{n} H_j (X_j - X_{j-1})$$

heißt H-**Transformierte**.

▶ Man kann H_k als Wetteinsatz für die Wettrunde mit Auszahlung in k. Die Voraussetzung $H_k \in \mathcal{F}_{k-1}$ ist dann ganz natürlich. Wäre $H_k \in \mathcal{F}_{k-1}$, dann könnte man den Wetteinsatz vom Ergebnis abhängig machen. Schön wär's.

22.1.7 Satz: Ist $(X_k)_{k\in\mathbb{N}_0}, X_k \in \mathcal{L}^1$ ein $\mathbb{F}$-adaptierter integrierbarer stochatischer Prozess, $H = (H_k)_{k\in\mathbb{N}_0}$ $\mathcal{F}$-prävisibel und $H_j(X_j - X_{j-1}) \in \mathcal{L}^1, n \in \mathbb{N}$.

Dann gilt: Ist X ein $\mathbb{F}$-Martingal, dann ist auch $(H \bullet X)_{n\in\mathbb{N}}$ ein $\mathbb{F}$-Martingal.

Beweis (vgl. Luschgy [?, Seite 13]): Es gilt

$$\begin{aligned}
\mathbb{E}(\,(H \bullet X)_n|\mathcal{F}_{n-1}\,) &= \mathbb{E}\left(\sum_{j=1}^{n} H_j(X_j - X_{j-1})\,|\,\mathcal{F}_{n-1}\right) \\
&= \sum_{j=1}^{n} \mathbb{E}(H_j(X_j - X_{j-1})\,|\,\mathcal{F}_{n-1}) \\
&= \sum_{j=1}^{n-1} H_j(X_j - X_{j-1})\mathbb{E}(1\,|\,\mathcal{F}_{n-1}) \\
&\quad + H_n\mathbb{E}((X_n - X_{n-1}|\mathcal{F}_{n-1}) \\
&= \sum_{j=1}^{n-1} H_j(X_j - X_{j-1}) \\
&= (H \bullet X)_{n-1}
\end{aligned}$$

22.2 Martingale in stetiger Zeit

Wir orientieren uns jetzt an Steele [34, S. 50f].

22.2.1 Definition: Eine Familie von Sub-σ-Algebren $(\mathcal{F}_t)_{t\in[0,\infty)}, \mathcal{F}_t \subset \mathcal{F}$ heißt **Filtration**, falls $\mathcal{F}_s \subset \mathcal{F}_t$ für alle $s \leq t$ gilt. Ein stochastischen Prozess $(X_t)_{t\in[0,\infty)}$ heißt **adaptiert**, falls $X_t \in \mathcal{F}_t$.

Ein integrierbarer stochastischen Prozess $(X_t)_{t\in[0,\infty)}, X_t \in \mathcal{L}$

heißt **Martingal**, falls für alle $t \geq s$

$$\mathbb{E}(X_t|\mathcal{F}_s) = X_s \quad \mathbb{P}\text{-f.s.}$$

gilt.

22.2.2 Definition (BB-Standard-Filtration): Es sei $(B_t)_{t\in[0,T]}$ eine Brown'sche Bewegung. Die natürliche Filtration ist gemäß Definition $\tilde{\mathcal{F}}_t = \sigma((B_t)_{t\in[0,T]})$.

Wir benötigen jedoch eine etwas erweiterte Filtration (siehe Steele [34, S. 50]). In der Menge $\sigma((B_s)_{s\in[0,T]})$ betrachten wir alle **Nullmengen**, d.h. die Menge $\mathcal{C} = \{B \in \sigma((B_s)_{s\in[0,T]})|\mathbb{P}(B) = 0\}$. Dann betrachten wir die Menge aller Teilmengen von $\mathcal{C}$, d.h. $\mathcal{N} = \{A|A \subset B, \mathbb{P}(B) = 0\}$. Wir setzen das Wahrscheinlichkeitsmaß durch $\mathbb{P}(A) = 0, A \in \mathcal{N}$ fort.

Die **BB-Standard-Filtration** $(\mathcal{F}_t^B)_{t\in[0,T]}$ erhalten wir dann wie folgt: $\mathcal{F}_t$ ist die kleinste σ-Algebra, die $\sigma((B_s)_{s\in[0,t]})$ und $\mathcal{N}$ umfasst.

Die **BB-Standard-Filtration** $(\mathcal{F}_t^B)_{t\in[0,T]}$ hat zwei wichtige (technische) Eigenschaften

- $\mathcal{F}_0$ umfasst alle Nullmengen $\mathcal{N}$.

- Für alle $t \geq 0$ gilt

$$\mathcal{F}_t = \bigcap_{s \text{ mit } s>t} \mathcal{F}_s$$

Diese beiden Bedingungen heißen **übliche Bedingungen (usual conditions)**.

22.2.3 Satz: Es sei $(B_t)_{t\in[0,T]}$ eine Brown'sche Bewegung. Dann sind die folgenden stochastischen Prozesse Martingale:

1. B_t
2. $B_t^2 - t$
3. $e^{\alpha B_t - \frac{\alpha^2}{2}t}$

Beweis: Vgl. die Übung

23 Itô-Döblin-Integral und Itô-Döblin-Lemma

Dieses Kapitel orientiert sich an Shreve [33] und Steele [34].

23.1 Itô-Döblin-Integral

$(\Omega, \mathcal{F}, \mathbb{P})$ Wahrscheinlichkeitsraum, Ω Zufallsergebnisse/Pfade

$\mathcal{F}_t$ eine Filtration

B_t eine BB

Itô-Döblin-Integral und Martingaltransformation.

23.1.1 Definition (Steele [34, Seite 80]): Ein zeit-stetiger stochastischer Prozess $\chi_{t,t\in[0,T]}$ heißt **einfacher** $\mathcal{F}_t$-adaptierter Prozess, falls es eine Zerlegung $0 = t_0 < ... < t_n = T$ des

Intervalls $[0,T]$ und Zufallsvariablen $h_i^s \in \mathcal{F}_{t_i}, i = 0,...,n$, $\mathbb{E}((h_i^s)^2) < \infty$ gibt, so dass $\chi_0(\omega) = h_0^s(\omega)$ und für $t \in (0,T]$

Das s bei h_i^s steht für simple

Was ist mit 0?

$$\chi_t(\omega) = \sum_{i=0}^{n-1} h_i^s(\omega) \mathbb{I}_{(t_i,t_{i+1}]}(t)$$

gilt.

$h_i^s(\omega)$ ist der Wert, den $\chi_t(\omega)$ auf dem Intervall $(t_i, t_{i+1}]$ annimmt (d.h. wenn $t \in (t_i, t_{i+1}]$ ist). Dieser Wert hängt von ω ab. Da $h_i^s \in \mathcal{F}_{t_i}$ vorausgesetzt wird, bezieht sich $h_i^s(\omega)$ nur auf Information, die bis t_i (einschließlich) vorliegt. $h_i^s(\omega)$ ist der Wert von χ **unmittelbar nach** t_i (und bis t_{i+1}) annimmt; jedoch nicht in t_i. Dort in t_i ist noch $h_{i-1}^s(\omega)$ der Wert von χ. Wir haben die etwas merkwürdige Gleichung $\chi_{t_i}(\omega) = h_{i-1}^s(\omega)$.

Die Wahl der Intervalle $(t_i, t_{i+1}]$ als rechts abgeschlossen führt dazu, dass die Pfade $\chi_t(\omega), t \in [0,T]$ (ω jeweils fest) **von links stetig** sind.

Interpretation: Wir interpretieren $h_i^s(\omega)$ als Handelsposition im Zeitpunkt für das Intervall $(t_i, t_{t+1}]$. $h_i^s(\omega)$ wird in t_i auf Basis der Information bis $\mathcal{F}_{t_i}$ festgelegt.

23.1.2 Bemerkung: Die Wahl der Intervalle $(t_i, t_{i+1}]$ als rechts abgeschlossen führt dazu, dass die Pfade $\chi_t(\omega), t \in$

$[0,T]$ (ω jeweils fest) **von links stetig** sind. Für manche Betrachtungen sind von rechts stetige Prozesse bequemer. Dafür wählt man die Intervalle $[t_i, t_{i+1})$ von links abgeschlossen. Die beiden Varianten unterscheiden sich nur an den Stellen t_i. Für die rechts-stetige Variante ist insbesondere $\chi_{t_i}(\omega) = a_i(\omega)$, während für von links stetige einfache Prozesse die verschobene Gleichung $\chi_{t_i}(\omega) = a_{i-1}(\omega)$ gilt.

Für die stochastische Integration im Allgemeinen sind von links stetige Prozesse *besser*; vgl. Durrett [6, Seite 35]. Wir betrachten die BB als Integrator. In diesem Fall ist die Unterscheidung nicht relevant, so dass wir auch von rechts stetige Prozesse betrachten könnten.

23.1.3 Definition: Es sei (B_t) eine Brown'sche Bewegung mit Filtration $(\mathcal{F}_t)$ und (χ_t) ein einfacher $\mathcal{F}_t$-adaptierter Prozess. Dann definieren wir das **Itô-Döblin-Intergral**

Muss die Filtration Eigenschaften haben.

$$\left(\int_0^T \chi_s dB_s\right)(\omega) = \sum_{i=0}^{n-1} h_i^s(\omega) \cdot (B_{t_{i+1}}(\omega) - B_{t_i}(\omega)).$$

Wir haben das Integral bis T definiert. Wir wollen aber Prozesse erhalten.

23.1.4 Definition: Wir definieren das Itô-Döblin-Integral

bis $0 < t < T, t \in (t_k, t_{k+1}]$ durch

$$\begin{aligned} I_t(\omega) &:= \left(\int_0^t \chi_s dB_s\right)(\omega) \\ &:= \left(\int_0^T \mathbb{I}_{[0,t]}(s)\chi_s dB_s\right)(\omega) \\ &= \sum_{i=0}^{k-1} h_i^s(\omega)(\omega) \cdot (B_{t_{i+1}}(\omega) - B_{t_i}(\omega)) + h_k^s(\omega) \cdot (B_t(\omega) - B_{t_k}(\omega)). \end{aligned}$$

23.1.5 Bemerkung: Es sei

$$\chi_t(\omega) = \sum_{i=0}^{n-1} h_i^s(\omega)\mathbb{I}_{(t_i,t_{i+1}]}(t)$$

und

$$\begin{aligned} \left(\int_0^T \chi_s dB_s\right)(\omega) &= \sum_{i=0}^{n-1} \chi_{t_i}(\omega) \cdot (B_{t_{i+1}}(\omega) - B_{t_i}(\omega)) \\ &= \sum_{i=0}^{n-1} h_i^s(\omega) \cdot (B_{t_{i+1}}(\omega) - B_{t_i}(\omega)) \end{aligned}$$

schreiben. In dieser Form lässt sich das Integral gut interpretieren (siehe Shreve [33, S. 127]): Wir wollen χ_t als Handelsposition zum Zeitpunkt auffassen t. Dann ist $\chi_{t_i}(\omega) \cdot (B_{t_{i+1}}(\omega) - B_{t_i}(\omega))$ der Zugewinn über das Zeitintervall von t_i bis t_{i+1}. $\left(\int_0^T \chi_s dB_s\right)(\omega)$ ist der kumulierte Gewinn bis T.

23.1.6 Satz: Es sei (B_t) eine Brown'sche Bewegung mit Filtration $(\mathcal{F}_t)$ und (χ_t) ein einfacher $\mathcal{F}_t$-adaptierter Prozess (je-

weils $t \in [0, T]$). Dann ist $I_t = \int_0^t \chi_s dB_s$ ein **Martingal**.

Beweis: Shreve [33, S. 128]. Dort werden von rechts stetige einfache Prozesses verwendet. Der Beweis sollte aber trotzdem funktionieren.

Wir bilden den Erwartungswert mit Information in s für den Zeitpunkt t, wobei $s \in (t_l, t_{l+1}]$ und $t \in (t_k, t_{k+1}]$. Wir betrachten den gleichen Fall wie Shreve: $k > l$.

Zunächst

$$\begin{aligned} I_t(\omega) = &\sum_{j=0}^{l-1} a_j(\omega) \cdot (B_{t_{j+1}}(\omega) - B_{t_j}(\omega)) \\ &+ a_l(\omega) \cdot (B_{t_{l+1}}(\omega) - B_{t_l}(\omega)) \\ &+ \sum_{j=l+1}^{k-1} a_i(\omega) \cdot (B_{t_{j+1}}(\omega) - B_{t_j}(\omega)) \\ &+ a_k(\omega) \cdot (B_t(\omega) - B_{t_k}(\omega)) \end{aligned}$$

Wir müssen nachweisen, dass

$$\mathbb{E}(I_t|\mathcal{F}_s) = I_s.$$

Es gilt für die *zweite* Zeile

$$\begin{aligned} \mathbb{E}(a_l(\omega) \cdot (B_{t_{l+1}}(\omega) - B_{t_l}(\omega))|\mathcal{F}_s) &= a_l(\omega) \cdot (\mathbb{E}(B_{t_{l+1}}(\omega)|\mathcal{F}_s) - B_{t_l}(\omega)) \\ &= a_l(\omega) \cdot (B_s(\omega) - B_{t_l}(\omega)) \end{aligned}$$

Es gilt für die *erste* Zeile ... Unterricht bzw. Shreve II

Es gilt für die *dritte* Zeile ... Unterricht bzw. Shreve II

Es gilt für die *vierte* Zeile ... Unterricht bzw. Shreve II

23.1.7 Satz (Itô-Isometrie): Es sei (B_t) eine Brown'sche Bewegung mit Filtration $(\mathcal{F}_t)$ und (χ_t) ein einfacher $\mathcal{F}_t$-adaptierter Prozess. Dann gilt

$$\mathbb{E}(I_T^2) = \mathbb{E}\left[\int_0^T \chi_s^2 ds\right].$$

bzw.

$$||I_T||_{L^2(d\mathbb{P})} = ||(\chi_s)||_{L^2(dt\times d\mathbb{P})}$$

Beweis: Shreve [33, S. 129].

23.1.8 Definition: Es sei $f : [0,T] \to \mathbb{R}$ eine reellwertige Funktion und $\mathcal{Z} = \{t_0, ..., t_n\}, 0 = t_0 < ... < t_n = T$ eine Zerlegung des Intervalls $[0,T]$ mit der Feinheit $||\mathcal{Z}|| := \max\{|t_i - t_{i-1}| : i = 1, ..., n\}$. Wenn der Grenzwert

$$[f,f]_T = \lim_{||\mathcal{Z}||\to\infty} \sum_{i=1,...,n} (f(t_i) - f(t_{i-1}))^2$$

existiert, dann heißt $[f,f]_T$ die **quadratische Variation** von f.

23.1.9 Satz (Quadratische Variation): Es sei (B_t) eine Brown'sche Bewegung mit Filtration $(\mathcal{F}_t)$ und (χ_t) ein einfacher $\mathcal{F}_t$-adaptierter Prozess. Dann gilt

$$[I, I]_t = \int_0^t \chi_s^2 ds.$$

Beweis: Shreve [33, S. 131]. Das sind ZV! Die Gleichung f.s.

23.1.10 Bemerkung: i.) Da $\mathbb{E}(I_t) = 0$ gilt $\mathbb{V}(I_t) = \mathbb{E}(I_t^2)$. Also

$$\mathbb{V}(I_t) = \mathbb{E}(I_t^2) = \mathbb{E}\left[\int_0^t \chi_s^2 ds\right] = \mathbb{E}([I, I]_t)$$

Die quadratische Variation einer ZV ist eine ZV! Erfasst wird die Variabilität für jeden Pfad (also pfadweise). Wenn wir den Erwartungswert bilden (um die Varianz zu berechnen), dann betrachten wir das erwartete Variabilität/Risiko *über* alle Pfade.

23.1.11 Definition: Es sei (B_t) eine Brown'sche Bewegung mit Filtration $(\mathcal{F}_t)$ und $h = (h_t)$ ein $(\mathcal{F}_t)$-adaptierter Prozess. Ferner sei h_T messbar bezüglich der Produkt-σ-Algebra $\mathcal{F}_T \otimes \mathcal{B}_{[0,T]}$ und es gelte

$$||h||_{L^2(dt \times d\mathbb{P})} = \int \int_0^T h_s^2 \, ds \, d\mathbb{P} = \mathbb{E} \int_0^T h_s^2 ds < \infty.$$

Es sei $\chi_n = (\chi_{n,t})$ eine Folge einfacher adaptierter Prozesse[1] mit

$$\chi_n \overset{L^2(dt\times d\mathbb{P})}{\longrightarrow} h,$$

d.h.

$$||(\chi_n - h)||_{L^2(dt\times d\mathbb{P})} = \mathbb{E}\left(\int_0^T (\chi_{n,s} - h_s)^2 ds\right) \overset{n\to\infty}{\longrightarrow} 0.$$

Für jedes χ_n ist das Integral $\int_0^T \chi_{n,s}\, dB_s$ definiert.

Das **Itô-Döblin-Integral** von h bezüglich (B_t) definieren wir als Limes[2]:

$$I_T = \int_0^T h_s\, dB_s = \lim_{n\to\infty} \int_0^T \chi_{n,s}\, dB_s,$$

dabei betrachten wir den Limes bezüglich $||\cdot||_{L^2(d\mathbb{P})}$.

Mit anderen Worten/Zeichen

$$||I_T - \int_0^T \chi_{n,t}\, dB_s||_{L^2(d\mathbb{P})} \overset{n\to\infty}{\longrightarrow} 0$$

Mit noch anderen Worten/Zeichen

$$\int_0^T \chi_{n,s}\, dB_s \overset{L^2(d\mathbb{P})}{\longrightarrow} I_T$$

23.1.12 Bemerkung: Die Konstruktion funtioniert wie folgt:

[1] So einen Prozess gibt es; vgl. ...

[2] Diesen Limes gibt es; vgl. ...

Wir wollen das Integral $\int_0^T h_s dB_s$ definieren. Wir wählen eine Folge einfacher Prozesse χ_s^n mit

$$\chi_{n,s} \overset{L^2(dt\times d\mathbb{P})}{\longrightarrow} h_s$$

wobei die Konvergenz bezüglich $||\cdot||_{L^2(dt\times d\mathbb{P})}$ vorliegt. Da die Folge $\chi_{n,s}$ konvergiert, ist diese Folge insbesondere eine Cauchy-Folge. Wegen der Itô-Isometrie ist auch $\int_0^T \chi_{n,s} dB_s$ eine Cauchy-Folge. Dann beweist man, dass die Folge $\int_0^T \chi_{n,s} dB_s$ konvergiert[3] und der Grenzwert ist dann $\int_0^T \Delta_s dB_s$:

$$\int_0^T \chi_{n,s} dB_s \overset{L^2(d\mathbb{P})}{\longrightarrow} \int_0^T \Delta_s dB_s$$

Für die technische Details siehe Steele [34, S. 81 - 84].

23.1.13 Bemerkung: Wir haben das Integral $I_T = \int_0^T h_s dB_s$ definiert. Wir müssen noch erklären, was wir unter den **Prozess** $I_t = \int_0^t \chi_s dB_s$ verstehen; vgl. Steele [34, S. 81 – 83].

Es ist naheliegend, diesen Prozess durch

$$I_t := \int_0^T \mathbb{I}_{[0,t]} h_s dB_s$$

zu definieren. Aber man muss eine *Spitzfindigkeit* beachten. Für jedes t kann man den Wert der Zufallsvariable I_t auf Nullmengen abändern. Da es überabzählbar viele t gibt, er-

[3] In welchem Raum?.

gibt sich ein substantielles Kompatibilitätsproblem: Die Vereinigung von Nullmengen ist in diesem Fall nicht unbedingt eine Nullmenge. Man setzt diese Stücke nicht irgendwie zusammensetzen, sondern so, dass sich ein stetiges Martingal ergibt.

Wir wählen eine Modifikation X_t der Zufallsvariablen I_t, die ein stetiges Martingale ist. Der folgende Satz bestätigt, dass das möglich ist.

23.1.14 Satz: Es gibt ein **stetiges Martingal** X_t, so dass für alle t

$$\mathbb{P}\left(X_t = \int_0^T \mathbb{I}_{[0,t]} \cdot h_s \, dB_s\right) = 1$$

gilt. Wir definieren

$$\int_0^t h_s dB_s := X_t.$$

Beweis: Vgl. Steele [34, S. 83f]

23.1.15 Definition: Es sei (h_t) ein $(\mathcal{F}_t)$-adaptierter Prozess. Ferner sei h_T messbar bezüglich der Produkt-σ-Algebra $\mathcal{F}_T \otimes \mathcal{B}_{[0,T]}$ und es gelte

$$||(h_t)||_{L^2(dt\times d\mathbb{P})} = \int \int_0^T h_s^2 \, ds \, d\mathbb{P} = \mathbb{E} \int_0^T h_s^2 ds < \infty.$$

Dann definieren wir

$$\int_0^t h_s dB_s := X_t,$$

wobei X_t das stetige Martingal, das gemäß des vorherigen Satz existiert.

Wir schreiben regelmäßig *nur*

$$dI_t = h_t dB_t,$$

wenn wir das Integral $I_t = \int_0^t h_s dB_s$ meinen.

23.1.16 Satz: Es sei (B_t) eine Brown'sche Bewegung mit Filtration $\mathcal{F}_t$ und $h_t, t \in [0,T]$ ein $\mathcal{F}_t$-adaptierter Prozess mit

$$||(h_s)||_{L^2(dt\times d\mathbb{P})} = \mathbb{E}\int_0^T h_s^2 ds < \infty.$$

Dann gilt folgendes für das Itô-Döblin-Integral $I_t = \int_0^t h_s dB_s$.

i.) I_t ist für alle ω eine **stetige** Funktion von t.
ii.) **Linearität:** Sind $h_t^1, t \in [0,T], h_t^2, t \in [0,T]$ zwei $\mathcal{F}_t$-adaptierte Prozesse mit $||h_s^1||_{L^2(dt\times d\mathbb{P})}$, $||h_s^2||_{L^2(dt\times d\mathbb{P})} < \infty$ und $\alpha, \beta \in \mathbb{R}$, dann ist

$$\int_0^t (\alpha h_s^1 + \beta h_s^2) dB_s = \alpha \int_0^t \Delta_s^1 dB_s + \beta \int_0^t \Delta_s^2 dB_s.$$

iii.) I_t ist $\mathcal{F}_t$-**adaptiert**.

iv.) I_t ist ein **Martingal**.

v.) **Ito-Isometrie:** Es gilt

$$\mathbb{E}(I_t^2) = \mathbb{E}\left[\int_0^t h_s^2 ds\right] = \int_0^t \mathbb{E}\left[h_s^2\right] ds.$$

vi.) **Quadratische Variation:** Es gilt

$$[I, I]_t = \int_0^t h_s^2 ds.$$

Oft schreibt man dafür

$$dI_t dI_t = (dI_t)^2 = h_t^2 dt$$

23.1.17 Bemerkung: Man kann das Itô-Döblin-Integral auch für Integranden definieren, die schwächere Integrabilitätsbedingungen erfüllen, als die oben genannte Voraussetzung $||h_s||_{L^2(dt\times d\mathbb{P})} < \infty$, dann geht jedoch im Allgemeinen die Martingaleigenschaft des Prozesses I_t verloren. Gebräuchlich ist beispielsweise (vgl. [34, Seite 95])

$$\mathbb{P}\left(\int_0^T h_t(\omega)\, dt < \infty\right) = 1.$$

I_t ist dann i.A. nur noch ein sogenanntes **lokales Martingal**. Vgl. für Einzelheiten Steele [34, Kapitel 7]

23.1.18 Satz: Es gilt

$$\int_0^t B_s dB_s = \frac{1}{2}B_t^2 - \frac{1}{2}t.$$

Beweis: Wie auch bei der Riemann-Integration bestimmt man Integrale sehr selten unter Verwendung der Definition. Wir werden später Das Itô-Döblin Lemma nutzen. Es ist aber so instruktiv mindestens ein Integral zu Fuß zu bestimmen (siehe Shreve [33, S. 134 f]). Zunächst bestimmt man elementar, dass

$$\frac{1}{2}\sum_{j=0}^{n-1}(W_{j+1}^{(n)} - W_j^{(n)})^2 = ... = \frac{1}{2}(W_n^{(n)})^2 + \sum_{j=0}^{n-1} W_j^{(n)}(W_j^{(n)} - W_{j+1}^{(n)})$$

gilt. Umgestellt

$$\sum_{j=0}^{n-1} W_j^{(n)}(W_{j+1}^{(n)} - W_j^{(n)}) = \frac{1}{2}(W_n^{(n)})^2 - \frac{1}{2}\sum_{j=0}^{n-1}(W_{j+1}^{(n)} - W_j^{(n)})^2.$$

Wenn wir jetzt den Grenzwert bilden, dann ergib sich die Behauptung. Dabei ergibt für den Subtrahend die **quadratische Variation** der Brown'schen Bewegung.

23.1.19 Bemerkung: Für das Itô-Integral $\int_0^t B_s dB_s$ erhalten wir also

$$\int_0^t B_s dB_s = \frac{1}{2}B_t^2 - \frac{1}{2}t,$$

während für das „entsprechende“ **Riemann Integral**

$$\int_0^t x dx = \frac{1}{2}t^2$$

gilt; und für das entsprechende **Stieltjes Integral** gilt und

$$\int_0^t F(x)dF(x) = \int_0^t F(x)F'(x)dx = \frac{1}{2}(F(t))^2$$
$$\int_0^t F(x)dF(x) = \frac{1}{2}F^2$$

Für das **Itô-Döblin Integral** erhalten wir also den zusätzlichen Term

$$-\frac{1}{2}\,t = -\frac{1}{2}\,([I,I]_t)^2.$$

Ein Integral ergibt sich bei der Kumulation sehr vieler sehr kleiner Schritte. Anders als bei der „normalen Integration“ fällt bei der stochastischen Integration die quadratische Variation nicht weg. Wie das passiert erkennt man, wenn man die Berechnungen in Shreve [33, S. 134 f] detailliert nachvollzieht.

23.2 Stochastischer Kalkül: Itô-Döblin Formel

23.2.1 Satz (Itô-Döblin Formel für $f = f(t, B_t)$): Es sei $f(t, x)$ eine reellwertige zweimal stetig differenzierbare Funktion und $(B_t)_{t\geq 0}$ eine Brown'sche Bewegung. Dann gilt für alle $T \geq 0$

$$f(T, B_T) = f(0, B_0) + \int_0^T f'_t(t, B_t)dt + \int_0^T f'_x(t, B_t)dB_t + \frac{1}{2}\int_0^T f''_{xx}(t, B_t)dt.$$

Das Ergebnis wird regelmäßig in Differentialform angegeben (und in dieser Form für den Kalkül benutzt)

$$\begin{aligned} df(t, B_t) &= f'_t(t, B_t)dt + f'_x(t, B_t)dB_t + \frac{1}{2}f''_{xx}(t, B_t)dt \\ &= (f'_t(t, B_t) + \frac{1}{2}f''_{xx}(t, B_t))dt + f'_x(t, B_t)dB_t. \end{aligned}$$

23.2.2 Bemerkung: Die zentrale Beobachtung des Beweises lässt sich an der folgenden Gleichung erklären; vgl. Shreve [33, S. 137, 141]. Wir betrachten eine Zerlegung $\mathcal{Z} =$

$\{t_0, ..., t_n\}, 0 = t_0 < t_1 < ... < t_n = T$. Es gilt

$$\begin{aligned}
f(T, B_T) - f(0, B_0) &= \sum_{j=0}^{n-1} f(t_{j+1}, B_{t_{j+1}}) - f(t_j, B_{t_j}) \\
&= \sum_{j=0}^{n-1} f'_t(t_j, B_{t_j})(t_{j+1} - t_j) \\
&+ \sum_{j=0}^{n-1} f'_x(t_j, B_{t_j})(B_{t_{j+1}} - B_{t_j}) \\
&+ \frac{1}{2} \sum_{j=0}^{n-1} f''_{xx}(t_j, B_{t_j})(B_{t_{j+1}} - B_{t_j})(B_{t_{j+1}} - B_{t_j}) \\
&+ \sum_{j=0}^{n-1} f''_{xt}(t_j, B_{t_j})(B_{t_{j+1}} - B_{t_j})(t_{j+1} - t_j) \\
&+ \frac{1}{2} \sum_{j=0}^{n-1} f''_{tt}(t_j, B_{t_j})(t_{j+1} - t_j)^2 \\
&+ \text{ Terme noch höherer Ordnung}
\end{aligned}$$

Wenn die Feinheit $||\mathcal{Z}||$ der Zerlegung gegen Null geht, dann bleiben drei Terme übrig. Der erste Term liefert das Integral bezüglich dt, der zweite Term das Itô-Integral bezüglich dB_t und der dritte Term das zweite Integral bezüglich dt. Das bemerkenswerte ist dabei das Verhalten des dritten Terms. Bei Grenzübergang ergibt sich für diesen Term die quadratische

Variation

$$\frac{1}{2}\sum_{j=0}^{n-1} f''_{xx}(t_j, B_{t_j})(B_{t_{j+1}} - B_{t_j})^2 \to \frac{1}{2}\int_0^T f''_{xx}(t_j, B_{t_j})dt$$

Symbolisch – und sehr einprägsam – schreibt und rechnet man

$$(dB_t)^2 = dt$$

sowie

$$\begin{aligned}(dB)\,(dt) &= 0\\ (dt)^2 &= 0\\ (dB)^3 &= 0\end{aligned}$$

23.2.3 Bemerkung: i.) Sowohl der Fehler der Approximation

$$f(W_{t_{j+1}}) - f(W_{t_j}) = f'(W_{t_j})(W_{t_{j+1}} - W_{t_j}) + \text{fehler}_1$$

als auch der Fehler der Approximation

$$\begin{aligned}f(W_{t_{j+1}}) - f(W_{t_j}) &= f'(W_{t_j})(W_{t_{j+1}} - W_{t_j})\\ &\quad + \frac{1}{2}f''(W_{t_j})(W_{t_{j+1}} - W_{t_j})^2 + \text{fehler}_2\end{aligned}$$

konvergierent gegen Null. **Wenn man jedoch Summen der ersten Approximation betrachtet** (für die Integra-

tion), dann kumulieren sich die Fehler derart, dass sich keine Konvergenz gegen Null ergibt. Bei der zweiten Approximation hingegen verschwinden beim Grenzübergang $||\mathcal{Z}|| \to 0$ auch die kumulierten Fehler (Vgl. Shreve [33, S 142]).

ii.) In der Analysis lernen wir, dass unter geeigneten Voraussetzungen

$$f(x) = f(0) + \int_0^x f'(z)dz.$$

In der stochastischen Analysis – wenn das *Argument* von $f(\cdot)$ eine Brown'sche Bewegung ist – gilt

$$f(B_T) = f(B_0) + \int_0^T f'_x(B_t)dB_t + \frac{1}{2}\int_0^T f''_{xx}(B_t)dt.$$

Das Besondere ist der Beitrag der quadratischen Variation $(dB)^2 = dt$, die den Beitrag $\frac{1}{2}\int_0^T f''_{xx}(B_t)dt$ erzeugt.

▶ Wenn man die Formel aus dem vorhergehenden Satz umstellt

$$f(T, B_T) - f(0, B_0) - \int_0^T f'_t(t, B_t)dt + \frac{1}{2}\int_0^T f''_{xx}(t, B_t)dt = \int_0^T f'_x(t, B_t)dB_t$$

dann erhält man einen Ansatz, um das Integral $\int_0^T f'_x(t, B_t)dB_t$

zu berechnen. Wir haben dazu ein paar Aufgaben gesehen. Wenn das geht, dann müssen wir das Integral nicht zu Fuß ausrechnen. Das ist analog zu Integrieren in der Analysis. Dort *rät* man Stammfunktionen und verifiziert dann das sie Stammfunktionen sind und kann dann den Hauptsatz der Differential und Integralrechnung verwenden.

▶ In Stoch Fin wird eigentlich nicht die BB als Integrator für die Handelsposition benötigt, sondern allgemeinere **Itô-Prozesse** (insbesondere die sogenannte geometrische BB, siehe unten).

23.2.4 Definition: Es sei $(B_t)_{t\geq 0}$ mit Filtration $(\mathcal{F}_t)_{t\geq 0}$. Ein stochastischer Prozess der Form[4]

$$X_t = X_0 + \int_0^t \Theta_s ds + \int_0^t \Delta_s dB_s,\ \ X_0 \in \mathbb{R},$$

heißt **Itô-Prozess**. Dabei sind $(\Theta_t)_{t\geq 0}$, $(\Delta_t)_{t\geq 0}$ adaptierte stochastische Prozesse mit

$$\mathbb{E}\int_0^T \Delta_s^2 ds < \infty,$$
$$\mathbb{E}\int_0^T |\Theta_s| ds < \infty.$$

Anstatt in Integralform geben wir **Itô-Prozess** auch in Dif-

[4] Θ wie **T**rend und Δ wie **D**iffusion.

ferentialform an:

$$dX_t = \Theta_t dt + \Delta_t dB_t.$$

Dabei heißt $\Theta_t dt$ **Driftterm/Trendterm** und $\Delta_t dB_t$ **Diffusionsterm**.

23.2.5 Satz: Es sei X_t ein **Itô-Prozess**. Dann gilt für die quadratische Variation von X_t

$$[X, X]_t = \int_0^t \Delta_s^2 ds.$$

bzw. kurz

$$dX_t dX_t = \Delta_t^2 ds$$

23.2.6 Definition: Es sei $(X_t)_{t\geq 0}$ ein **Itô-Prozess** und $(\alpha_t)_{t\geq 0}$ ein adaptierter stochastischer Prozess. Dann definieren wir das **Itô-Döblin Integral bezüglich** dX_t so

$$\int_0^t \alpha_s dX_s = \int_0^t \alpha_s \Theta_s ds + \int_0^t \alpha_s \Delta_s dB_s$$

▶ Passend für das Integral bezüglich eines **Itô-Prozesse** gibt es auch eine Itô-Döblin Formel. Hier betrachten wir $f_t = f(t, X_t)$ und wollen $df_t = df(t, X_t)$ bestimmen

23.2.7 Satz (Itô-Döblin Lemma für $f_t = f(t, X_t)$): Es

sei $(X_t)_{t\geq 0}$ ein Itô-Prozess

$$dX_t = \Theta_t dt + \Delta_t dB_t$$

und $f(t,x)$ eine reelle zweimal stetig differenzierbare Funktion. Dann gilt für alle $t \geq 0$

$$\begin{aligned} f(T,X_T) &= f(0,X_0) + \int_0^T f'_t(t,X_t)dt + \int_0^T f'_x(t,X_t)dX_t \\ &\quad + \frac{1}{2}\int_0^T f''_{xx}(t,X_t)\Theta_t^2 dt \\ &= f(0,X_0) + \int_0^T \left(f'_t(t,X_t) + f'_x(t,X_t)\Theta_t + \frac{1}{2}f''_{xx}(t,X_t)\Delta_t^2 \right) dt \\ &\quad + \int_0^T f'_x(t,X_t)\Theta_t dB_t \end{aligned}$$

In Differentialform lautet das Ergebnis

$$\begin{aligned} df(t,X_t) &= f'_t(t,X_t)dt + f'_x(t,X_t)\Theta_t dt + \frac{1}{2}f''_{xx}(t,X_t)\Delta_t^2 dt \\ &\quad + f'_x(t,X_t)\Delta_t dB_t \\ &= \left(f'_t(t,X_t) + f'_x(t,X_t)\Theta_t + \frac{1}{2}f''_{xx}(t,X_t)\Delta_t^2 \right) dt \\ &\quad + f'_x(t,X_t)\Delta_t dB_t \end{aligned}$$

bzw. noch kompakter

$$df(t,X_t) = f'_t(t,X_t)dt + f'_x(t,X_t)\, dX_t + \frac{1}{2}f''_{xx}(t,X_t)\, dX_t dX_t$$

bzw. noch kompakter

$$df = f'_t dt + f'_x \, dX_t + \frac{1}{2} f''_{xx} \, dX dX.$$

23.2.8 Bemerkung: In der Differentialform

$$df(t, X_t) = f'_t(t, X_t)dt + f'_x(t, X_t) \, dX_t + \frac{1}{2} f''_{xx}(t, X_t) \, dX_t dX_t$$

wird der **Anfangswert** $f(0, X_0)$ von $f(\cdot, \cdot)$ *unterdrückt.* An der Gleichung

$$\begin{aligned} f(T, X_T) &= f(0, X_0) + \int_0^T f'_t(t, X_t)dt + \int_0^T f'_x(t, X_t)dX_t \\ &\quad + \frac{1}{2} \int_0^T f''_{xx}(t, X_t)\Delta_t^2 dt \\ &= f(0, X_0) + \int_0^T \left(f'_t(t, X_t) + f'_x(t, X_t)\Theta_t + \frac{1}{2} f''_{xx}(t, X_t)\Delta_t^2 \right) dt \\ &\quad + \int_0^T f'_x(t, X_t)\Delta_t dB_t \end{aligned}$$

kann man den Anfangswert $f(0, X_0)$ direkt ablesen.

23.2.9 Satz (deterministischer Integrand): Es sei $(B_t)_{t \geq 0}$ eine Brown'sche Bewegung, $\Delta_t, t \geq 0$ eine deterministische Funktion und

$$I_t = \int_0^t \Delta_s dB_s$$

Dann ist I_t normal verteilt mit Erwartungswert 0 und Varianz $\int_0^t \Delta_s ds$.

Beweis: Shreve [33, S. 149].

23.2.10 Satz (geometrische Brown'sche Bewegung): Es sei

$$S_t = S_0 e^{\left(\mu t - \frac{1}{2}\sigma^2 t\right) + \sigma B_t}.$$

Dann

$$dS_t = \mu S_t dt + \sigma S_t dB_t$$
$$\frac{dS_t}{S_t} = \mu dt + \sigma dB_t$$

Beweis: Wir schreiben

$$S_t = S_0 e^{\left(\mu t - \frac{1}{2}\sigma^2 t\right) + \sigma B_t} = S_0 e^{X_t}$$

mit dem Itô-Prozess $X_t = \mu t - \frac{1}{2}\sigma^2 t + \sigma B_t$.

In der Tat: Mit $\Theta_t = \mu - \frac{1}{2}\sigma^2$ und $\Delta_t = \sigma$ ist $dX_t = \left(\mu - \frac{1}{2}\sigma^2\right) dt + \sigma dB_t$.

Dann

$$f(t,x) = S_0 e^x$$

und

$$\begin{aligned} f'_t &= 0 \\ f'_x &= S_0 e^x \\ f''_{xx} &= S_0 e^x \end{aligned}$$

Schließlich

$$\begin{aligned} df_t &= \left(f'_t(t, X_t) + f'_x(t, X_t)\Theta_t + \frac{1}{2} f''_{xx}(t, X_t)\Delta_t^2 \right) dt + f'_x(t, X_t)\Delta_t dB_t \\ &= \left(0 + S_0 e^{X_t}\Theta_t + \frac{1}{2} S_0 e^{X_t}\Delta_t^2 \right) dt + S_0 e^{X_t}\Delta_t dB_t \\ &= \left(0 + S_t \left(\mu - \frac{1}{2}\sigma^2 \right) + \frac{1}{2} S_t \sigma^2 \right) dt + S_t \sigma dB_t \\ &= \mu S_t dt + \sigma S_t dB_t \end{aligned}$$

Alternativ kann man auch die einfache Formel anwenden, denn in

$$S_t = S_0 e^{\left(\mu t - \frac{1}{2}\sigma^2 t\right) + \sigma B_t}$$

taucht B_t *separat* auf. Also betrachtet man

$$f(t, x) = S_0 e^{\left(\mu t - \frac{1}{2}\sigma^2 t\right) + \sigma x}$$

Dann

$$\begin{aligned} f'_t &= \left(\mu - \frac{1}{2}\sigma^2\right) \cdot S_t \\ f'_x &= \sigma S_t \\ f''_{xx} &= \sigma^2 S_t \end{aligned}$$

Also

$$\begin{aligned} df_t &= \left(\mu - \frac{1}{2}\sigma^2\right) \cdot S_t \cdot dt + \sigma S_t dB_t + \frac{1}{2}\sigma^2 \cdot S_t \cdot dt \\ &= \mu S_t dt + \sigma S_t dB_t \end{aligned}$$

▶ Im folgenden Beispiel geht B_t nicht separat ein. Jetzt muss man einen richtigen Ansatz für den Itô-Prozess X_t finden, so dass sich $f_t = f(t, X_t)$ ergibt.

23.2.11 Satz (Vasicek Process): Für

$$R_t = e^{-\beta t} R_0 + \frac{\alpha}{\beta}(1 - e^{-\beta t}) + \sigma e^{-\beta t} \int_0^t e^{\beta s} dB_s$$

gilt

$$dR_t = (\alpha - \beta R_t)dt + \sigma dB_t.$$

Beweis: Wir setzen $X_t = \int_0^t e^{\beta s} dB_s$. X_t ist ein Itô-Prozess mit

$$dX_t = 0\, dt + e^{\beta t} dB_t.$$

Also $\Theta_t = 0, \Delta_t = e^{\beta t}$.

Dann definieren wir

$$f(t,x) = e^{-\beta t} R_0 + \frac{\alpha}{\beta}(1 - e^{-\beta t}) + \sigma e^{-\beta t} x.$$

23.2.12 Bemerkung (Vasicek Zinsmodell): ooo. Da der Integrand von $\int_0^t e^{\beta s} dB_s$ ist, ist R_t normal verteilt. ooo.

23.3 Multivariabler stochastischer Kalkül

$B_{1,t}, B_{2,t}$ zwei unabhängige BB.

23.3.1 Satz: Für die quadratischen Variationen gilt

$$\begin{aligned} dB_{1,t} dB_{1,t} &= dt \\ dB_{2,t} dB_{2,t} &= dt \\ dB_{1,t} dB_{2,t} &= 0 \\ dB_{1,t} dt &= 0 \\ dB_{2,t} dt &= 0 \\ dt dt &= 0 \end{aligned}$$

23.3.2 Satz (Itô-Döblin-Formel): Es sei

$$dX_t = \Theta_{1,t}dt + \sigma_{11,t}dB_{1,t} + \sigma_{12,t}dB_{2,t}$$
$$dY_t = \Theta_{2,t}dt + \sigma_{21,t}dB_{1,t} + \sigma_{22,t}dB_{2,t}.$$

Ferner sei $f(t, X_t, Y_t)$ eine Funktion mit stetigen partiellen Ableitungen $f'_t, f'_x, f'_y, f''_{xx}, f''_{xy}, f''_{yy}$. Dann gilt

$$df = f'_t dt + f'_x dX + f'_y dY + \frac{1}{2}f''_{xx}dXdX + \ f''_{xy}dXdY + \frac{1}{2}f''_{yy}dYdY$$

wobei

$$dX_t dX_t = (\sigma^2_{11,t} + \sigma^2_{12,t})dt$$
$$dY_t dY_t = (\sigma^2_{21,t} + \sigma^2_{22,t})dt$$
$$dX_t dY_t = (\sigma_{11,t}\sigma_{21,t} + \sigma_{12,t}\sigma_{22,t})dt$$

23.3.3 Satz (Itô-Produkt-Regel):

$$d(X_t Y_t) = X_t dY_t + dX_t Y_t + dX_t dY_t$$

Dieses Ergebnis sollte man mit der Produktregel aus der Analysis vergleichen, damit man den zusätzlichen Term **nicht vergisst**! Das ist für das Konzept der Selbstfinanzierung relevant!

23.3.4 Satz (Produkt-Regel):

$$(uv)' = u'v + uv'$$
$$d(uv) = du\, v + u\, dv$$

23.3.5 Satz (Korrelation): Vgl. Shreve [33, S. 171]

$$dS_1 = \alpha_1 S_1 dt + \sigma_1 S_1 dB_1$$
$$dS_2 = \alpha_2 S_2 dt + \sigma_2 S_2 [\rho dB_1 + \sqrt{1-\rho^2} dB_2]$$

und

$$B_3 = \rho B_1 + \sqrt{1-\rho^2} B_2$$

Dann ist B_3 eine Brown'sche Bewegung und wir können die beiden Finanzpreise in der Form

$$dS_1 = \alpha_1 S_1 dt + \sigma_1 S_1 dB_1$$
$$dS_2 = \alpha_2 S_2 dt + \sigma_2 S_2 dB_3$$
$$d(B_1 B_3) = B_1 dB_3 + B_3 dB_1 + \rho dt$$

schreiben. Ferner

$$\mathbb{E}[d(B_1 B_3)] = \rho\, dt$$

ρ erfasst die instantane Korrelation zwischen den Finanzpreisen.

23.4 Stochastische Differentialgleichungen

Der folgende Abschnitt orientiert sich sehr an Klebaner [**?**, Abschnitt 5.1 bis 5.5]

Eine Gleichung der Form

$$y' = f(y, x)$$

nennt man **gewöhnliche Differentialgleichung**. Typischerweise betrachtet man sogenannte **Anfangswertprobleme**

$$\begin{aligned} y' &= f(y, x) \\ y(x_0) &= y_0 \end{aligned}$$

Eine differenzierbare Funktion $y = y(x)$ heißt Lösung des Anfangswertproblems, falls

$$\begin{aligned} y'(x) &= f(y(x), x), x \in [x_0, x_1] \\ y(x_0) &= y_0 \end{aligned}$$

23.4.1 Definition: Es sei $(B_t)_{t\geq 0}$ eine BB. Eine Gleichung der Form

$$dX_t = \Theta(X_t, t)\, dt + \Delta(X_t, t)\, dB_t$$

heißt **stochastische Differentialgleichung**. Dabei sind die Funktionen $\Theta(\cdot, \cdot)$ und $\Delta(\cdot, \cdot)$ und X_0 gegeben.

Gesucht wird ein Stochastischer Prozess, der diese Gleichung löst.

Ein stochastischer Prozess X_t heißt (starke) **Lösung**, falls die Integrale $\int_0^t \Theta(X_s, s)ds$ und $\int_0^t \Delta(X_s, s)dB_s$ für alle $t \in [0, T]$ existieren und

$$X_t = X_0 + \int_0^t \Theta(X_s, s)ds + \int_0^t \Delta(X_s, s)dB_s.$$

23.4.2 Beispiel: Die stochastische Differentialgleichung

$$dS_t = \mu S_t dt + \sigma S_t dB_t$$

hat die Lösung

$$S_t = S_0\, e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma B_t}$$

Das rechnet man mit Itô's Lemma leicht nach.

Es ist dabei

$$\Theta(x, t) = \mu \cdot x$$
$$\Delta(x, t) = \sigma \cdot x$$

23.4.3 Satz: Wenn die Bedingungen

1. Die Koeffizienten $\Theta(\cdot, \cdot)$ und $\Delta(\cdot, \cdot)$ sind *lokal* Lipschitz in x gleichmäßig in t. Für alle T und M gibt es eine

Konstante K (die nur von T und M abhängt) mit: Gilt $|x|, |y| < M$ und $0 \leq t \leq T$, dann gilt

$$|\Theta(x,t) - \Theta(y,t)| + |\Delta(x,t) - \Delta(y,t)| < K|x-y|$$

2. Die Koeffizienten $\Theta(\cdot,\cdot)$ und $\Delta(\cdot,\cdot)$ erfüllen die folgenden lineare Wachstumsbedingung:

$$|\Theta(x,t)| + |\Delta(x,t)| \leq K(1+|x|)$$

3. X_0 ist unabhängig von $B_t, t \in [0,T]$ und $\mathbb{E}(X_0) < \infty$

erfüllt sind, dann hat die SDGL

$$dX_t = \Theta(X_t,t)\,dt + \Delta(X_t,t)\,dB_t$$

für gegebene $\Theta(\cdot,\cdot)$, $\Delta(\cdot,\cdot)$ und X_0 eine eindeutig bestimmte strake Lösung mit stetigen Pfaden

23.4.4 Satz: Die starke Lösung einer stochastischen Differentialgleichung hat die Markov-Eigenschaft

$$\mathbb{P}(X_t|\mathcal{F}_s) = \mathbb{P}(X_t|X_s)$$

wobei $\mathcal{F}_s$ die Information **bis** s erfasst.

23.5 Satz vo Girsanov

24 Sprungprozesse

Dieses Kapitel orientiert sich eng an Shreve [33, Abschnitt 11.1 bis 11.3]

24.1 Poisson Prozess

24.1.1 Definition: Es sei $\lambda > 0$. Es sei τ eine Zufallsvariable mit der Dichte

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{falls } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Dann sagt man, dass τ eine **Exponentialverteilung** mit Intensität λ hat. In diesem Fall schreiben wir $\tau \sim \text{Exp}(\lambda)$.

24.1.2 Satz: Es sei $\tau \sim \text{Exp}(\lambda)$. Dann gilt

$$\begin{aligned} \mathbb{E}(\tau) &= \frac{1}{\lambda} \\ \mathbb{P}(\tau \leq t) &= 1 - e^{-\lambda t} \quad (\tau \geq 0) \\ \mathbb{P}(\tau > t) &= e^{-\lambda t} \quad (\tau \geq 0) \end{aligned}$$

Ferner ist τ **gedächtnislos**:

$$\mathbb{P}(\tau > t + s | \tau > s) = \mathbb{P}(\tau > t)$$

24.1.3 Definition: Es sei $\tau_1, \tau_2, ...$ eine Folge von unabhängigen Zufallsvariablen mit $\tau_i \sim \text{Exp}(\lambda)$. Die **Ankunftszeit** für n Sprünge ($n \in \mathbb{N}$) ist

$$S_n = \sum_{k=1}^{n} \tau_k.$$

Die τ_i heißen dann **Zwischenankunftszeiten**.

Der **Poisson-Prozess** N_t zählt die Sprünge (Ankünfte) bis t

$$N_t = \begin{cases} 0 & \text{falls } 0 \leq t < S_1 \\ 1 & \text{falls } S_1 \leq t < S_2 \\ \vdots & \\ n & \text{falls } S_n \leq t < S_{n+1} \end{cases}$$

Wir bemerken, dass die Pfade von N_t von **rechts-stetig** sind.

Man sagt, dass N_t ein Poisson Prozess mit **Intensität** λ ist.

24.1.4 Satz: S_n ist Gamma verteilt, d.h. S_n hat die Dichte

$$g_n(s) = \frac{(\lambda s)^{n-1}}{(n-1)!}\lambda e^{-\lambda s}, s \geq 0$$

24.1.5 Satz: N_t ist hat eine Poisson Verteilung:

$$\mathbb{P}(N_t = k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t} \quad k = 0, 1, ...$$

24.1.6 Satz: Es sei N_t ein Poisson Prozess mit Intensität $\lambda > 0$. Wir betrachten Zeitpunkte $0 = t_0 < t_1 < ... < t_n$

Dann sind die Zuwächse

$$N_{t_1} - N_{t_0}, ..., N_{t_n} - N_{t_{n-1}}$$

unabhängig und stationär.

Es gilt für $k = 0, 1, ...$

$$\mathbb{P}(N_{t_{j+1}} - N_{t_j} = k) = \frac{\lambda^k (t_{j+1} - t_j)^k}{k!}e^{-\lambda(t_{j+1} - t_j)}$$

24.1.7 Satz: Es gilt

$$\mathbb{E}(N_t - N_s) = \lambda(t - s)$$
$$\mathbb{V}(N_t - N_s) = \lambda(t - s)$$

24.1.8 Satz: Es sei N_t ein Poisson Prozess mit Intensität λ. Dann ist

$$M_t = N_t - \lambda t$$

ein **Martingal**.

24.2 Zusammengesetzte Poisson-Prozesse

24.2.1 Definition: Es sei N_t eine Poisson Prozess mit Intensität λ und $Y_1, Y_2, ...$ eine Folge unabhäniger identisch verteilter Zufallsvariablen mit Erwartungswert $\mathbb{E}(Y_i) = \beta$. Dann heißt

$$Q_t = \sum_{i=1}^{N_t} Y_i \quad t \geq 0$$

zusammengesetzter Poisson Prozess.

24.2.2 Satz: Es gilt

$$\mathbb{E}Q_t = \beta\lambda t$$

und

$$M_t = Q_t - \beta\lambda t$$

ist ein Martingal.

24.2.3 Satz: Wir betrachten Zeitpunkte $0 = t_0 < t_1 < ... < t_n$. Die Zuwächse

$$Q_{t_1} - Q_{t_0}, ..., Q_{t_n} - Q_{t_{n-1}}$$

sind unabhängig und stationär. Insbesondere haben

$$Q_{t_j} - Q_{t_{j-1}} \text{ und } Q_{t_j - t_{j-1}}$$

die gleiche Verteilung.

Literaturverzeichnis

[1] Heinz **Bauer**, 1992, Maß- und Integrationstheorie, 2. Auflage, De Gruyter

[2] Patrick **Billingsley**, 1995, Probability and Measure, 3th. Auflage; Wiley

[3] Norman **Biggs**, 2002, Discrete Mathematics, 2. ed.: Oxford University Press

[4] Joseph **Blitzstein**, Jessica Hwang, Introduction to Probability, CRC Press

[5] Kai Chung **Chung**, 2000, A Course in Probability Theory, Wiley

[6] Rick **Durrett**, 2019, Probability: Theory and Examples, Cambridge Unuiversity Press

[7] Alexander McNeil, Rüdiger Frey, Paul Embrechts, 2015, Quantitative Risk Managment, 2. Auflage, Princeton

University Press

[8] Jürgen **Elstrodt**, 2018, Maß- und Integrationstheorie, 8. Auflage, Springer

[9] Nasrollah **Etemadi**; 1981; An elementary proof of the strong law of large numbers; Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete; Volume 55, Seiten 119-122.

[10] William **Feller**; 1968; An Introduction to Probability – Theory and its Application; 3ed; Wiley.

[11] Ludwig **Fahrmeir**, Christian **Heumann**, Rita **Künstler**, Iris **Pigeot**, Gerhard **Tutz**, 2016, Statistik – Der Weg zur Datenanalyse, 8. Auflage, Springer

[12] Hans **Föllmer** und Alexander **Schied**, 2016, Stochastic Finance – An introduction in discrete time, 4. Auflage, De Gruyter

[13] Hans-Otto **Georgii**; 2015, Stochastik, Fünfte Auflage, De Gruyter

[14] Geoffrey **Grimmett** und David **Strikzaker**, 2020, Probability and Random Processes, Fourth Edition, Oxford University Press

[15] Geoffrey **Grimmett** und David **Strikzaker**, 2020, One

tousend exercises in probability, Third Edition, Oxford University Press

[16] Charles **Grinstead** und Laurie **Snell**, 2006, Introduction to Probability, https://chance.dartmouth.edu/teaching_aids/books_articles/probability_book/book.html [abgerufen am 20210612]

[17] Bruce **Hansen**; 2022; Econometrics; Princeton University Press

[18] Gareth **James**, Daniela **Witten**, Trevor **Hastie**, Robert Tibshirami; 2021; An Introduction to Statistical Learning; Springer

[19] Norbert **Henze**, 2019, Stochastik – Einführung mit Grundzügen der Maßtheorie, Springer

[20] R. J. **Hyndman** und Y. Fan, 1996, Sample quantiles in statistical packages, American Statistician 50, 361–365.

[21] Jean **Jacod** und Philip **Protter**, 2004, Probability Essentials, Springer

[22] Samuel **Karlin**, Howard Taylor; 1975; A first course in Stochastic Processes; 2ed.; Academic Press

[23] Achim **Klenke**, 2020, Wahrscheinlichkeitstheorie, 4. Auflage, Springer Spektrum, Berlin

[24] Ulrich **Krengel**, 2005, Einführung in die Wahrscheinlichkeitstheorie und Statistik, 8. Auflage, Springer

[25] Uwe **Küchler**, 2016, Maßtheorie für Statistiker, 8. Auflage, Springer

[26] **Lütkepohl, Helmut**, New Introduction to Multiple Time Series Analysis, Springer

[27] Michael **Mürmann**, 2014, Wahrscheinlichkeitstheorie und Stochastische Prozesse, Springer

[28] David **Mazur**, 2010, Combinatorics, MAA Press

[29] Ralph Tyrrell **Rockafellar**, Stanislav **Uryasev**; 2002; Conditional value-at-risk for general loss distributions; Journal of Banking & Finance; Volume 26, Seiten 1443-1471.

[30] Walter **Rudin**, 1976, Principle of Real Analysis, Springer

[31] Rene **Schilling**; 2021; Measure, Integral, Probability & Processes; Rene-Schilling-Dresden

[32] Jun **Shao**, 2003, Mathematical Statitics, Springer

[33] **Shreve, Steven**, 2004, Stochastic Calculus for Finance 2, Springer, Heidelberg

[34] Michael **Steele**; 2010; Stochastic Calculus and Financial Applications; Springer.

[35] Anders **Tolver**; An Introduction to Markov Chains; http://web.math.ku.dk/noter/filer/stoknoter.pdf.

[36] Stefan **Tappe**, 2013, Einführung in die Wahrscheinlichkeitstheorie, Springer

[37] Grzegorz **Tomkowicz** und Stan **Wagon**, 2016, The Banach-Tarski Paradox, 2. Auflage, Cambridge University Press

[38] David **Williams**, 1991, Probility with Martingales, Cambridge University Press

[39] David **Williams**, 2001, Weighting the odds, Cambridge University Press

[40] Amy **Wagaman** und Robert **Dobrow**, 2021, Probability – With Applications and §, 2. Auflage, Wiley