Einführung in die Stochastik

Notizen zu einer Vorlesung an der Johannes-Gutenberg-Universität Mainz, Winter 2020/2021

Matthias Birkner

Version vom 17. Februar 2021

Kommentare, Korrekturvorschläge, Hinweise auf (Tipp-)fehler gerne per Email an birkner@mathematik.uni-mainz.de senden

Inhaltsverzeichnis

0 Auftakt **3**

1 Grundlegendes **6**
1.1 Ereignisse und Wahrscheinlichkeiten 6
1.2 Bericht zum Fall $\Omega = \mathbb{R}^d$ 12
1.3 Klassische diskrete Verteilungen 16
1.4 Zufallsvariablen 20

2 Bedingte Wahrscheinlichkeiten und Unabhängigkeit **27**
2.1 Bedingte Wahrscheinlichkeiten 27
2.2 Mehrstufige Zufallsexperimente 30
2.3 Unabhängigkeit 35
2.4 Faltung 40
2.5 Asymptotische Ereignisse 42

3 Erwartungswert, Varianz und Kovarianz **44**
3.1 Diskreter Fall 44
3.2 Der Fall mit Dichte 51
3.3 Varianz und Kovarianz 52
3.4 Median(e) 60
3.5 Erzeugende Funktionen* 63

4 Gesetz der großen Zahlen **67**
4.1 Beweis von Satz 4.6* 70

5 Zentraler Grenzwertsatz **72**
5.1 Beweis von Satz 5.5* 78

6 Ideen und Begriffe aus der Statistik **80**
6.1 Zur deskriptiven Statistik 80
6.2 Grundlegende Begriffe, Schätzen von Parametern 80
6.3 Konfidenzintervalle (und Konfidenzbereiche) 90
6.4 Statistische Tests 99

7 Markovketten **115**
7.1 Treffwahrscheinlichkeiten und erwartete Eintrittszeiten 119
7.2 Gleichgewichte 122

7.3 Rekurrenz und Transienz* 127

Kapitel 0

Auftakt

Das Problem der Punkte

Aus dem Briefwechsel zwischen Blaise Pascal[1] und Pierre de Fermat[2] 1654, angeregt durch Fragen von Antoine Gombard, genannt Chevalier de Méré[3], siehe auch [KW, S. 95ff]:

> Spieler A und Spieler B spielen über mehrere Runden, jede einzelne Runde ist ein faires Glücksspiel (z.B. ein fairer Münzwurf).
>
> Am Anfang setzt jeder gleich viel ein, derjenige, der als erster insgesamt vier Runden gewonnen hat, bekommt alles.
>
> Nach drei Runden muss das Spiel abgebrochen werden, es steht 2 : 1 für A.

Frage: Wie soll der Einsatz nun gerecht aufgeteilt werden?

Ansatz: Aufteilung gemäß der Wahrscheinlichkeit, von diesem Spielstand aus zu gewinnen.

Wie wahrscheinlich ist es, dass A vom Spielstand 2 : 1 aus gewinnt?

Fermats Berechnungsvorschlag („Aufzählung aller Vorwärtspfade“)

Spiele im Geiste 4 Runden weiter (dann wäre das Spiel sicher entschieden), die 16 möglichen Spielweiterführungen sind

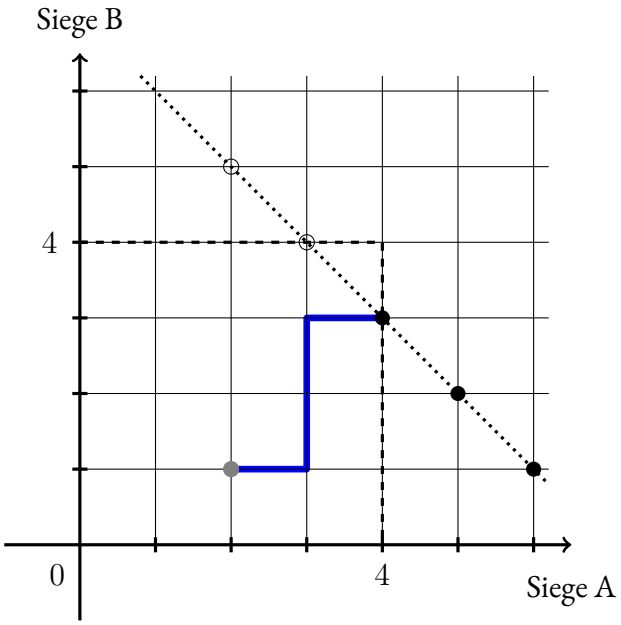
AAAA	ABAA	BAAA	BBAA
AAAB	ABAB	BAAB	BBAB
AABA	ABBA	BABA	BBBA
AABB	ABBB	BABB	BBBB

(unterstrichen $\widehat{=}$ führt zu Sieg von A)

[1] Blaise Pascal, 1623–1662, Mathematiker, Physiker, Philosoph, ...

[2] Pierre de Fermat, 1607(?)–1665, Jurist, Gelehrter, Mathematiker, ...

[3] Antoine Gombard, 1607–1684, Schriftsteller und (Amateur-)Mathematiker



Ein Spielverlauf ≙ Nord-Ost-Pfad in $\mathbb{N}_0 \times \mathbb{N}_0$, blau eingezeichnet Verlauf ABBA. Die schwarz ausgefüllten Punkte sind Endstände solcher Spielweiterführungen, bei denen A gewinnt.

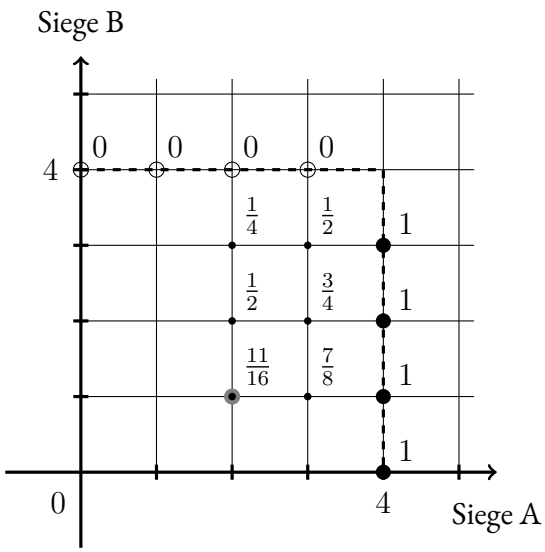
Demnach ist die

Wahrscheinlichkeit, dass A vom Stand $2:1$ aus gewinnt, gleich $\frac{11}{16}$.

Pascals Berechnungsvorschlag („Rückwärtsrechnung“)

Berechne „rückwärts“ die Wahrscheinlichkeit $f(x,y)$, dass A vom Spielstand $x:y$ aus gewinnt: Wenn man $f(x+1,y)$ und $f(x,y+1)$ schon kennt, kennt man auch

$$f(x,y) = \frac{1}{2}f(x+1,y) + \frac{1}{2}f(x,y+1).$$



Bemerkung. Pascals Methode ist weniger rechenaufwendig (speziell wenn eine größere Anzahl als 4 gewonnene Runden für den Gesamtsieg betrachtet wird).

Wir werden Pascals Ansatz wieder treffen als (einen Spezialfall) der Lösung des Dirichlet-Problems für eine Markovkette.

Kapitel 1

Grundlegendes

In diesem Kapitel geht es um grundlegende Begriffe und Definitionen, die zur mathematischen Modellierung von Zufallssituationen verwendet werden.

1.1 Ereignisse und Wahrscheinlichkeiten

Sei Ω (nicht-leere) Menge („Ergebnisraum“ oder „Stichprobenraum“),

$\omega \in \Omega$ heißt ein „Elementarereignis“

(gewisse) $A \subset \Omega$ nennen wir Ereignisse,
insbesondere $\Omega \ldots$ „sicheres Ereignis“, $\varnothing \ldots$ „unmögliches Ereignis“

Vorstellung: „der Zufall“ wählt ein $\omega \in \Omega$, wir sagen „A tritt ein“, wenn $\omega \in A$.

Wir betrachten $\mathcal{F} \subset 2^\Omega$ ($:= \{B : B \subset \Omega\}$) und lassen $A \in \mathcal{F}$ als Ereignisse zu.

Operationen (Mengeninterpretationen und ihre Interpretation für Ereignisse)**:**

$A^c := \Omega \setminus A$	…	„A tritt nicht ein“ (A^c heißt Gegen- oder Komplementärereignis von A)
$A \cap B$	…	„A und B treten ein“
$A \cup B$	…	„A oder B treten ein“
$A \subset B$	…	„A impliziert B“

Manchmal auch:

$A \Delta B := (A \cap B^c) \cup (A^c \cap B)$ (symmetrische Differenz),
… „genau eines der beiden Ereignisse A, B tritt ein“

Wir fordern, dass $\mathcal{F}$ erfüllt

i) $\varnothing \in \mathcal{F}$,
ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,
iii) $A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Definition 1.1. Sei Ω eine nicht-leere Menge. $\mathcal{F} \subset 2^\Omega$, das *i)–iii)* genügt, heißt eine *σ-Algebra* (über Ω).

Ein Paar $(\Omega, \mathcal{F})$ mit $\mathcal{F}$ σ-Algebra über Ω heißt ein messbarer Raum (auch: Messraum oder Ereignisraum).

Bemerkung. Das „σ" im Namen erinnert an die abzählbare Operation in *iii)*; wenn man stattdessen

$$\textit{iii)'} \quad A_1, A_2 \in \mathcal{F} \;\Rightarrow\; A_1 \cup A_2 \in \mathcal{F}$$

fordert, so heißt $\mathcal{F}$ eine *Algebra*.

Eine σ-Algebra ist insbesondere eine Algebra, denn $A_1 \cup A_2 = A_1 \cup A_2 \cup \varnothing \cup \varnothing \cup \cdots$

Beispiel 1.2. Ω endliche (oder abzählbare) Menge, $\mathcal{F} = 2^\Omega$ („diskreter messbarer Raum")
Beispiel-Instanzen:

- Wurf einer Münze, $\Omega = \{K, Z\}$
- Wurf eines Würfels, $\Omega = \{1, 2, \ldots, 6\}$
- Dreifacher Würfelwurf, $\Omega = \{1, 2, \ldots, 6\}^3$ $(= \{(a_1, a_2, a_3) : a_i \in \{1, \ldots, 6\}\})$
- Wartezeit, bis in einer Münzwurffolge der erste Erfolg kommt, $\Omega = \mathbb{N}_0$

Abgesehen von maßtheoretischen Schwierigkeiten im überabzählbaren Fall, die dann i.A. die Wahl $\mathcal{F} = 2^\Omega$ unmöglich machen – vgl. z.B. [G, Satz 1.5], („Warum so vorsichtig?") –, hilft eine σ-Algebra auch bei der Modellierung unterschiedlicher „Informationsgenauigkeit":

Bemerkung. $\mathcal{F}$ modelliert, welche Ereignisse wir beobachten können, z.B.

$$\Omega = \{1, \ldots, 6\}, \quad \mathcal{F} = \big\{\varnothing, \{2, 4, 6\}, \{1, 3, 5\}, \{1, \ldots, 6\}\big\}$$

entspricht folgendem Zufallsexperiment: Jemand wirft einen 6er-Würfel, verrät uns nur, ob Augenzahl gerade oder ungerade.

Definition 1.3. $(\Omega, \mathcal{F})$ ein messbarer Raum. Eine Abbildung $P : \mathcal{F} \to [0, 1]$ mit

(N) $P(\Omega) = 1$ („Normierung") und

(A) $A_1, A_2, \cdots \in \mathcal{F}$ paarw. disjunkt $\Rightarrow P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n)$ („σ-Additivität")

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

Ein solches Tripel $(\Omega, \mathcal{F}, P)$ heißt ein *Wahrscheinlichkeitsraum*.

Für $A \in \mathcal{F}$ nennen wir $P(A)$ die Wahrscheinlichkeit (des Ereignisses A unter dem Maß P).

Beispiel 1.4 (diskreter Wahrscheinlichkeitsraum)**.** Ω endlich oder abzählbar, $p : \Omega \to [0,1]$ Abbildung mit $\sum_{\omega\in\Omega} p(\omega) = 1$.

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

definiert ein Wahrscheinlichkeitsmaß auf $(\Omega, 2^\Omega)$.

p heißt die (Wahrscheinlichkeits-)Gewichtsfunktion von P, $p(\omega)$ heißt das (Wahrscheinlichkeits-) Gewicht von ω.

Beispiel-Instanzen

1. Uniforme Verteilung auf einer endlichen Menge: $|\Omega| < \infty$, $p(\omega) = \frac{1}{|\Omega|}$ sind die Gewichte der uniformen Verteilung auf Ω. Das zugehörige P heißt oft auch die Laplace-Verteilung[1] (auf Ω).
   (a) (Wurf eines fairen 6er-Würfels) $\Omega = \{1,2,3,4,5,6\}$, $p(\omega) = 1/6$ für $\omega \in \Omega$
   (b) (dreimaliger Wurf eines fairen 6er-Würfels) $\Omega = \{(\omega_1,\omega_2,\omega_3) : \omega_i \in \{1,2,3,4,5,6\}$ für $i = 1,2,3\}$, $p((\omega_1,\omega_2,\omega_3)) = 1/6^3 = 1/216$
   (c) (n verschiedene Objekte / Zahlen in zufälliger Reihenfolge)

   $$\Omega = \{(x_1,x_2,\ldots,x_n) : x_1,\ldots,x_n \in \{1,2,\ldots,n\} \text{ paarweise verschieden}\}$$

   („symmetrische Gruppe der Ordnung n"), $p((x_1,x_2,\ldots,x_n)) = 1/n!$
2. Ein verfälschter Münzwurf: $\Omega = \{\text{Kopf}, \text{Zahl}\}$, $p(\text{Kopf}) = 0.6 = 1 - p(\text{Zahl})$
3. Eine winziges Modell für Spam, das Sprache und Status einer Email betrachtet:

   $$\Omega = \{\texttt{Deutsch}, \texttt{Englisch}\} \times \{\texttt{Spam}, \texttt{keinSpam}\},$$

   $p\big((\texttt{Deutsch},\texttt{Spam})\big) = 0.2$, $p\big((\texttt{Deutsch},\texttt{keinSpam})\big) = 0.1$, $p\big((\texttt{Englisch},\texttt{Spam})\big) = 0.6$, $p\big((\texttt{Englisch},\texttt{keinSpam})\big) = 0.1$
4. Anzahl Würfe, bevor beim wiederholten fairen Münzwurf zum ersten Mal Kopf kommt: $\Omega = \mathbb{N}_0$, $p(n) = (1/2)^n \cdot (1/2) = 2^{-n-1}$ für $n \in \Omega$

Lemma 1.5. Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum, $A, B, A_1, A_2, \cdots \in \mathcal{F}$, so gilt

$$P(\varnothing) = 0 \tag{1.1}$$

$$P(A\cup B) + P(A\cap B) = P(A) + P(B) \qquad \text{(endliche Additivität)}, \tag{1.2}$$
$$\text{insbesondere } P(A) + P(A^c) = 1$$

$$A \subset B \Rightarrow\ P(A) \le P(B) \qquad \text{(Monotonie)} \tag{1.3}$$

$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) \le \sum_{n=1}^{\infty} P(A_n) \qquad (\sigma\text{-Subadditivität}) \tag{1.4}$$

$$A_n \nearrow_{n\to\infty} A \text{ (d.h. } A_1 \subset A_2 \subset \ldots \text{ mit } A = \cup_{n=1}^{\infty} A_n)$$
$$\text{oder } A_n \searrow_{n\to\infty} A \text{ (d.h. } A_1 \supset A_2 \supset \ldots \text{ mit } A = \cap_{n=1}^{\infty} A_n),$$
$$\text{so gilt } P(A) = \lim_{n\to\infty} P(A_n) \qquad (\sigma\text{-Stetigkeit}) \tag{1.5}$$

[1] nach Pierre-Simon Laplace, 1749–1827

Beweis. (1.1):

$$P(\varnothing) = P\big(\varnothing \cup \varnothing \cup \dots\big) = \sum_{n=1}^{\infty} P(\varnothing), \quad \text{also } P(\varnothing) = 0.$$

(1.2): Betrachte zunächst den Fall $A \cap B = \varnothing$:

$$P(A \cup B) = P\big(A \cup B \cup \varnothing \cup \varnothing \cup \dots\big) = P(A) + P(B) + \underbrace{P(\varnothing)}_{=0} + \underbrace{P(\varnothing)}_{=0} + \cdots, \quad \text{d.h. (1.2) gilt.}$$

Allgemeiner Fall: Schreibe

$$A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B) \qquad \text{(paarw. disjunkt)}$$

(mit $A \setminus B := A \cap B^c$), also

$$\begin{aligned} P(A \cup B) + P(A \cap B) &= P(A \setminus B) + P(B \setminus A) + 2P(A \cap B) \\ &= \big(P(A \setminus B) + P(A \cap B)\big) + \big(P(B \setminus A) + P(A \cap B)\big) = P(A) + P(B). \end{aligned}$$

(1.3):

$$P(A) \le P(A) + P(B \setminus A) = P(B)$$

(1.4): Stelle $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} A'_n$ als disjunkte Vereinigung dar mit $A'_n := A_n \setminus \bigcup_{j=1}^{n-1} A_j$ ($\subset A_n$), so ist

$$P\Big(\bigcup_{i=1}^{\infty} A_n\Big) = P\Big(\bigcup_{i=1}^{\infty} A'_n\Big) = \sum_{n=1}^{\infty} P(A'_n) \le \sum_{n=1}^{\infty} P(A'_n)$$

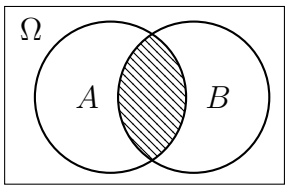
(1.5): Betrachte zunächst den Fall $A_n \nearrow A$: Setze $A'_i := A_i \setminus \bigcup_{j<i} A_j = A_i \setminus A_{i-1}$ ($A_0 := \varnothing$),

$$P(A) = P\Big(\bigcup_{i=1}^{\infty}(A_i \setminus A_{i-1})\Big) = \sum_{i=1}^{\infty} P(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \underbrace{\sum_{i=1}^{n} P(A_i \setminus A_{i-1})}_{=P(A_n)}.$$

Falls $A_n \searrow A$, so beachte, dass $A_n^c \nearrow A$ gilt, dann verwende obiges zusammen mit (1.2). □

Bemerkung 1.6 (Einschluss-Ausschluss-Formel)**.** Formel (1.2) aus Lemma 1.5 kann man auch folgendermaßen schreiben

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Man kann diese Formel anhand eines Venn-Diagramms veranschaulichen: Der schraffierte Bereich $A \cap B$ wird in $P(A) + P(B)$ doppelt gezählt, in $P(A \cap B)$ aber nur einmal.

Allgemein gilt für $A_1, A_2, \dots, A_n \in \mathcal{F}$ (man nennt dies auch die „Siebformel“):

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i) - \sum_{\substack{i,j=1 \\ i\neq j}}^{n} P(A_i \cap A_j) + \sum_{\substack{i,j,k=1 \\ i\neq j, i\neq k, j\neq k}}^{n} P(A_i \cap A_j \cap A_k) \pm \cdots + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

d.h. knapp ausgedrückt

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{\emptyset \neq J \subset \{1,\dots,n\}} (-1)^{(\#J)-1} P\Big(\bigcap_{i\in J} A_i\Big) \tag{1.6}$$

Man kann (1.6) beispielsweise per Induktion beweisen oder indem man die „Atome“ $\bigcap_{j\in J} A_j \cap \bigcap_{j\in I\setminus J} A_j^c$, $J \subset \{1, 2, \dots, n\}$ der von $A_1, \dots, A_n$ erzeugten σ-Algebra betrachtet.

Beispiel. Betrachten wir in Bsp. 1.4, 1 (c) mit $n = 3$ die Ereignisse $A_i = \{(x_1, x_2, x_3) \in \Omega : x_i = i\}$ für $i = 1, 2, 3$. Wenn wir eine uniform verteilte Permutation von $1, 2, 3$ betrachten, so ist $A_i = \{i \text{ ist ein Fixpunkt}\}$. Dann ist

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\ &\quad - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) \\ &= \frac{1}{3} + \frac{1}{3} + \frac{1}{3} - \frac{1}{6} - \frac{1}{6} - \frac{1}{6} + \frac{1}{6} = \frac{2}{3} \end{aligned}$$

die Wahrscheinlichkeit, dass es mindestens einen Fixpunkt gibt.

Bemerkung 1.7. Wahrscheinlichkeitsräume $(\Omega, \mathcal{F}, P)$ gemäß Def. 1.3 werden heute, nicht zuletzt wegen des einflussreichen Werks von A.N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933, oft als „Basisobjekt“ der mathematischen Modellierung von Zufallsvorgängen verwendet. Insoweit ist die explizite Wahl eines geeigneten Wahrscheinlichkeitsraums (oft) ein Teil der Arbeit, wenn ein umgangssprachlich formuliertes Problem in Mathematik übersetzt werden soll. Man sollte sich allerdings bewusst sein, dass es dabei i.A. viele mögliche Wahlen gibt.

(Wir werden bald sogenannte Zufallsvariablen einführen – formal als geeignete Funktionen auf Ω definiert –, die dazu einen für die Intuition und für das Argumentieren sehr angenehmen Zugang ermöglichen.)

Wir lassen hier die „philosophische“ Frage offen, welche Interpretation die Wahrscheinlichkeit $P(A)$ eines Ereignisses A haben soll. Es bieten sich etwa an:

- „naive“ Interpretation der Wahrscheinlichkeit: Die „Natur“ enthält inhärente Unbestimmtheit („ist selbst nicht sicher, was sie tut“), und $P(A)$ beschreibt den Grad der Sicherheit, mit dem sie sich für das Ereignis A entscheidet.

- „frequentistische“ Interpretation der Wahrscheinlichkeit: Wenn man das zufällige Experiment unter exakt denselben Bedingungen sehr oft wiederholte, wäre der relative Anteil der Ausgänge, in denen A eingetreten ist, etwa $P(A)$.

- „subjektive“ Interpretation der Wahrscheinlichkeit: $P(A)$ misst, wie sicher ich mir persönlich bin, dass A eintreten wird.

  (Beispielsweise: Wieviel Wetteinsatz wäre ich bereit zu bezahlen, wenn mir 1 € ausgezahlt würde, sofern A eintritt?)

[KW, S. vi] schreiben dazu: „Es kann nicht darum gehen, eine spezielle Intuition gegenüber den anderen durchzusetzen. Dass sich dies innerhalb der Mathematik auch gar nicht als nötig erweist, ist eine der Stärken mathematischer Wissenschaft.“ Siehe z.B. auch die Diskussion in [G], S. 14 am Ende von Abschn. 1.1.3 oder [P, Sect. 1.2. Interpretations].

Beispiel 1.8 (Kollision von Kennzeichen (oder „Hash-Werten“))**.** n Objekte erhalten „rein zufällig“ ein Kennzeichen aus einer Menge von r möglichen Werten.

Wir formalisieren dies via

$$\Omega = \{1, 2, \dots, r\}^n \ni \omega = (\omega_1, \dots, \omega_n)$$

mit

$$\mathcal{F} = 2^\Omega$$

und

$$P(A) = \frac{|A|}{|\Omega|},\ A \subset \Omega, \quad \text{der uniformen Verteilung auf } \Omega.$$

Betrachte das Ereignis

$$B = \big\{(\omega_1, \dots, \omega_n) : \omega_i \neq \omega_j \text{ für alle } 1 \leq i \neq j \leq n\big\}$$

(„alle Kennzeichen sind verschieden“)

Frage: $P(B) = ?$

Es ist

$$|B| = r(r-1)(r-2)\cdots(r-n+1)$$

(r Wahlmöglichkeiten für ω_1, dann $r-1$ Wahlmöglichkeiten für ω_2, etc.), also

$$P(B) = \frac{|B|}{|\Omega|} = \frac{r(r-1)(r-2)\cdots(r-n+1)}{r^n} = \prod_{i=0}^{n-1} \frac{r-i}{r} = \prod_{i=1}^{n-1} \big(1 - \frac{i}{r}\big).$$

Mit $1 - x \leq e^{-x}$ für $x \in \mathbb{R}$ ergibt sich

$$P(B) \leq \prod_{i=1}^{n-1} e^{-i/r} = \exp\Big(-\frac{1}{r}\sum_{i=1}^{n-1} i\Big) = \exp\Big(-\frac{n(n-1)}{2r}\Big)$$

für eine untere Schranke beachte

$$B^c = \big\{(\omega_1, \dots, \omega_n) : \omega_i = \omega_j \text{ für ein Paar } i \neq j\big\} = \bigcup_{1 \leq i < j \leq n} \{\omega_i = \omega_j\}.$$

Demnach

$$P(B^c) \leq \sum_{1 \leq i < j \leq n} P(\omega_i = \omega_j) = \sum_{1 \leq i < j \leq n} \frac{r^{n-1}}{r^n} = \frac{n(n-1)}{2r},$$

insgesamt

$$\exp\Big(-\frac{n(n-1)}{2r}\Big) \geq P(B) = 1 - P(B^c) \geq 1 - \frac{n(n-1)}{2r}.$$

1.2 Bericht zum Fall $\Omega = \mathbb{R}^d$

In vielen Situationen interessieren wir uns für zufällige Experimente, deren Ausgang eine reelle Zahl bzw. eine (reeller) Vektor [d-Tupel] ist, z.B. physikalische / chemische / biologische Messungen (etwa Masse, Zeit, Länge, Konzentration,...). Um dies mathematisch zu modellieren, möchten wir den Fall $\Omega = \mathbb{R}^d$ (oder auch Ω eine geeignete Teilmenge von $\mathbb{R}^d$) in unserem Rahmen betrachten können.

Aus maßtheoretischen Gründen ist für überabzählbares Ω die Wahl $\mathcal{F} = 2^\Omega$ i.A. nicht möglich[2]. Eine explizite Beschreibung einer σ-Algebra ist i.A. schwierig, man betrachtet daher oft sog. „Erzeugermengen". Dazu:

Beobachtung 1.9. *Sei J eine beliebige Indexmenge, für $j \in J$ sei $\mathcal{F}_j \subset 2^\Omega$ eine σ-Algebra. Dann ist auch*

$$\mathcal{G} := \bigcap_{j \in J} \mathcal{F}_j$$

eine σ-Algebra.

Sei $\mathcal{E} \subset 2^\Omega$. Man schreibt

$$\sigma(\mathcal{E}) := \bigcap \big\{ \mathcal{F} : \mathcal{F} \text{ ist } \sigma\text{-Algebra mit } \mathcal{E} \subset \mathcal{F} \subset 2^\Omega \big\},$$

die „von $\mathcal{E}$ erzeugte σ-Algebra". (Dies ist offenbar die kleinste σ-Algebra über Ω, die $\mathcal{E}$ umfasst.)

Für $\Omega = \mathbb{R}^d$ möchten wir (im Fall $d = 1$) mindestens Intervalle bzw. (im Fall $d > 1$) Quader in $\mathcal{F}$ haben, d.h. wir fordern

$$\forall\, a < b : \ (a, b] \in \mathcal{F} \quad \text{(im Fall } d = 1\text{)}$$
$$\forall\, a_1 < b_1, a_2 < b_2, \ldots, a_d < b_d : \ (a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_d, b_d] \in \mathcal{F} \quad \text{(im Fall } d > 1\text{)}$$

Beobachtung. Aus dieser Forderung folgt, dass jede offene Menge in $\mathcal{F}$ liegen muss, denn für $O \subset \mathbb{R}^1$ offen gilt

$$O = \bigcup_{x,y \in O \cap \mathbb{Q},\, x<y} (x, y]$$

und analog im $\mathbb{R}^d$.

Man verwendet daher die Borel-σ-Algebra[3]

$$\mathcal{B}(\mathbb{R}^d) := \sigma\big(\{O : O \subset \mathbb{R}^d \text{ offen}\}\big).$$

Bericht 1.10. Es gilt

$$\mathcal{E} \subset 2^\Omega \ \cap\text{-stabil} \quad (\text{d.h. } A, B \in \mathcal{E} \Rightarrow A \cap B \in \mathcal{E})$$

so ist ein Wahrscheinlichkeitmaß P auf $\sigma(\mathcal{E})$ bereits durch seine Werte auf $\mathcal{E}$ festgelegt, d.h. sind P, P' W'maße auf $(\Omega, \sigma(\mathcal{E}))$ mit $P(E) = P'(E)$ für alle $E \in \mathcal{E}$, so gilt $P = P'$. Siehe z.B. [G, Satz 1.12].

[2] Siehe z.B. [G, Satz 1.5 und Diskussion dort], wir diskutieren dies (und mehr) erst in der Vorlesung Stochastik I genauer.

[3] nach Émile Borel, 1871–1956

Man verwendet zweckmäßigerweise

$$\mathcal{E} = \big\{(-\infty, x] : x \in \mathbb{R}\big\} \quad \text{(Halbintervalle im Fall } d = 1\text{)},$$
$$\mathcal{E} = \big\{(-\infty, x_1] \times (-\infty, x_2] \times \dots (-\infty, x_d] : x_1, x_2, \dots, x_d \in \mathbb{R}\big\} \quad \text{(Halbquader im Fall } d > 1\text{)}$$

als $\cap$-stabile Erzeugermengen der entsprechenden Borel-σ-Algebra.

Beobachtung und Definition 1.11. Für ein Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ist die Funktion $F_P : \mathbb{R} \to [0,1]$, $F_P(x) := P\big((-\infty, x]\big)$ nicht-fallend und rechtsstetig mit

$$\lim_{x\to\infty} F_P(x) = 1, \qquad \lim_{x\to-\infty} F_P(x) = 0.$$

F_p heißt die *Verteilungsfunktion* von P.

Die Eigenschaften folgen aus Lemma 1.5, (1.5): Für $x < y$ ist $(-\infty, x] \subset (-\infty, y]$ und somit $F_p(x) = P((-\infty, x]) \le P((-\infty, y]) = F_p(y)$; seien $x_n \ge x$ mit $x_n \searrow x$, setze $A_n := (-\infty, x_n]$, $A := (-\infty, x]$, so gilt $A_n \searrow A$ für $n \to \infty$ und daher auch $F_P(x_n) = P(A_n) \searrow P(A) = F_P(A)$, d.h. F_P ist rechtsstetig. Analog gilt $F_P(x_n) = P(A_n) \nearrow P(\mathbb{R}) = 1$ für $x_n \nearrow \infty$ und $F_P(x_n) = P(A_n) \searrow P(\varnothing) = 0$ für $x_n \searrow -\infty$.

Analog betrachtet man im Fall $d > 1$ für $(x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ die Funktion

$$F_P(x_1, x_2, \dots, x_d) := P\big((-\infty, x_1] \times (-\infty, x_2] \times \dots (-\infty, x_d]\big),$$

die entsprechende Eigenschaften besitzt.

Bericht 1.12. Umgekehrt definiert jede Funktion $F : \mathbb{R} \to [0,1]$ (bzw. $F : \mathbb{R}^d \to [0,1]$) mit den Eigenschaften aus Def. 1.11 ein (eindeutiges) Wahrscheinlichkeitsmaß P mit $F = F_P$.

Bemerkung 1.13. 1 (Bezug zum diskreten Fall). Sei eine (höchstens) abzählbare Menge $\Omega' = \{x_1, x_2, \dots\} \subset \mathbb{R}$ (z.B. $\Omega' = \mathbb{N}$, $\Omega' = \mathbb{Z}$, ...) und ein diskretes W'maß P auf $(\Omega', 2^{\Omega'})$ mit Gewichten $p(\cdot)$ (wie in Bsp. 1.4) gegeben. Wir können P auch als W'maß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ auffassen via

$$P(A) := \sum_{n\,:\,x_n \in A} p(x_n),$$

dann ergibt sich als Verteilungsfunktion

$$F_P(x) = \sum_{n\,:\,x_n \le x} p(x_n).$$

(Diese ist stückweise konstant mit (höchstens) abzählbar vielen Sprüngen.)

2. Sei P Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Verteilungsfunktion F_P. Die (verallgemeinerte) inverse Funktion von F_P,

$$F_P^{-1}(t) := \inf\{x \in \mathbb{R} : F(x) \ge t\},$$

heißt auch die *Quantilfunktion* von P.

(Beachte, dass die so definierte Funktion F_P^{-1} linksstetig ist. Mit dieser Definition ergibt sich für $x \in \mathbb{R}$, $t \in [0,1]$ die Beziehung

$$F_P^{-1}(t) \le x \iff t \le F_P(x).$$

In der Literatur gibt es leicht verschiedene Definitionen der „Quantilfunktion“, man prüfe ggfs. jeweils die verwendete Konvention.)

Beispiel (Dirac-Maß)**.**

$$\delta_x(A) = \begin{cases} 1, & x \in A, \\ 0 & x \notin A. \end{cases}$$

(Im Fall $\Omega = \mathbb{R}$ ist $F_{\delta_x}(y) = 1(x \le y)$.)

Beispiel 1.14 (Maße mit Dichten auf $\mathbb{R}$ (bzw. auf $\mathbb{R}^d$))**.** Sei $f_P : \mathbb{R} \to \mathbb{R}_+$ integrierbar[4] mit

$$\int_{\mathbb{R}} f_P(x)\,dx = 1.$$

Dann definiert[5]

$$P(A) := \int_A f_P(x)\,dx$$

ein Wahrscheinlichkeitsmaß, die Funktion f_P heißt die *Dichte* (auch: Wahrscheinlichkeitsdichte) von P.

Die Verteilungsfunktion

$$F_P(x) = P\big((-\infty, x]\big) = \int_{-\infty}^{x} f_P(y)\,dy$$

ist dann (zumindest an Stetigkeitsstellen von f_P) differenzierbar mit

$$\frac{d}{dx} F_P(x) = f_P(x)$$

Analog definiert man für (geeignet) integrierbares $f_P : \mathbb{R}^d \to \mathbb{R}_+$ mit $\int_{\mathbb{R}^d} f_P(x)\,dx = 1$ durch

$$P(A) := \int_A f_P(x)\,dx$$

ein W'maß P auf $\mathbb{R}^d$ mit Dichte f_P. Wir denken an dieser Stelle wiederum an geeignet „gutartige" Funktionen f_P und Teilmengen $A \subset \mathbb{R}^d$, für die das Integral beispielsweise als iteriertes Riemann-Integral wohldefiniert ist, d.h.

$$\int_A f_P(x)\,dx = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{1}_A(x_1, \ldots, x_d) f_P(x_1, \ldots, x_d)\,dx_1 \cdots dx_d.$$

Dies wird für die im Rahmen der Vorlesung betrachteten Beispiele genügen.

Beispiel 1.15 („Klassische" eindimensionale Verteilungen mit Dichte)**.**

1. (uniforme Verteilung) $a, b \in \mathbb{R}$, $a < b$. $\mathrm{Unif}_{[a,b]}$ mit Dichte $\frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$, Verteilungsfunktion $\left(\frac{x-a}{b-a} \wedge 1\right) \vee 0$

[4] In einem mit den Vorkenntnissen der Hörer verträglichen Sinn: Wir werden nur Beispiele betrachten, in denen f_P wenigstens stückweise stetig ist, so dass man hier durchaus an das Riemann-Integral (oder auch ganz salopp an die „Fläche unter der Kurve") denken kann. Für einen Bericht zum Lebesgue-Integral z.B. [G, Tatsache 1.14].

[5] Wiederum hängt es vom verwendeten Integralbegriff ab, für welche Mengen A das Integral $\int_A f_P(x)\,dx$ sinnvoll definiert ist. Man verliert an dieser Stelle wenig, wenn man bei A etwa an eine endliche Vereinigung von Intervallen denkt.

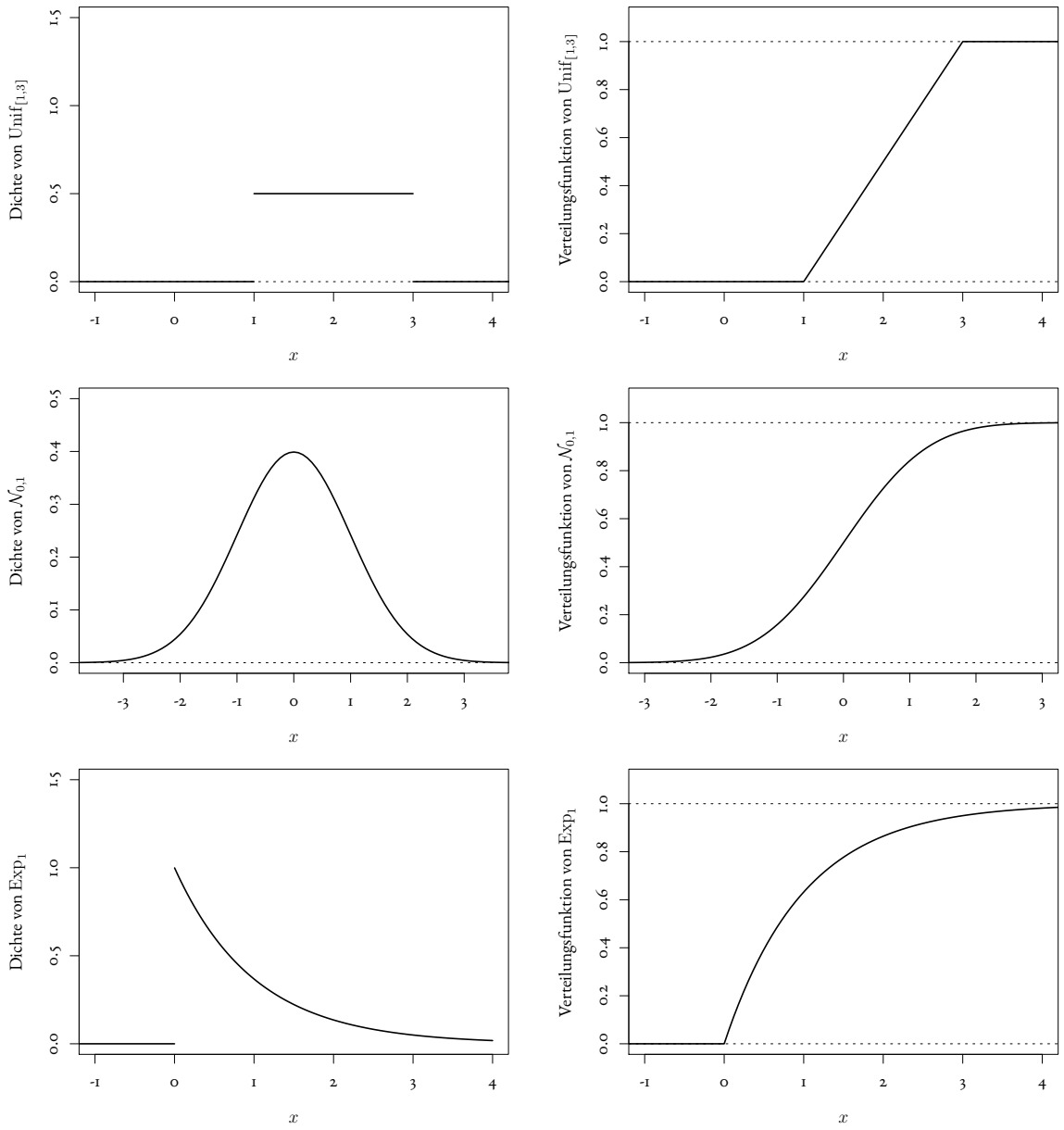
2. (Normalverteilung[en]) $\mu \in \mathbb{R}$, $\sigma > 0$. $\mathcal{N}_{\mu,\sigma^2}$ mit Dichte $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ heißt Normalverteilung mit Mittelwert μ und Varianz σ^2.

   $\mathcal{N}_{0,1}$ heißt die *Standardnormalverteilung*, die Verteilungsfunktion

$$\Phi(x) := F_{\mathcal{N}_{0,1}}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

   ist tabelliert bzw. in vielen Computerprogrammen implementiert und es ist $\mathcal{N}_{\mu,\sigma^2}((-\infty, x]) = \Phi((x-\mu)/\sigma)$ (Übung).

3. (Exponentialverteilung[en]) $\theta > 0$, Exp_θ hat Dichte $\theta e^{-\theta x} \mathbf{1}_{[0,\infty)}(x)$, Verteilungsfunktion $(1 - e^{-\theta x})\mathbf{1}_{[0,\infty)}(x)$



Beispiel 1.16. Laplace-Verteilung auf einem beschränkten Gebiet $\Omega \subset \mathbb{R}^d$: Für $A \subset \Omega$ (geeignet[6]) ist

$$P(A) := \frac{\mathrm{vol}(A)}{\mathrm{vol}(\Omega)} = \frac{\int_A 1\,dx}{\int_\Omega 1\,dx}.$$

1.3 Klassische diskrete Verteilungen

Beispiel 1.17 (Urnenmodelle)**.** Eine Urne enthalte n (nummerierte) Kugel, wir ziehen zufällig k ($\leq n$) heraus.

1. mit Zurücklegen, mit Beachtung der Reihenfolge:

$$\Omega_1 = \big\{(\omega_1, \dots, \omega_k) : \omega_1, \dots, \omega_k \in \{1, 2, \dots, n\}\big\} = \{1, 2, \dots, n\}^k,$$
$$P_1\big(\{(\omega_1, \dots, \omega_k)\}\big) = \frac{1}{n^k} \quad \Big(= \frac{1}{|\Omega_1|}, \text{ vgl. auch Beispiel 1.8}\Big)$$

2. ohne Zurücklegen, mit Beachtung der Reihenfolge:

$$\Omega_2 = \big\{(\omega_1, \dots, \omega_k) : \omega_1, \dots, \omega_k \in \{1, 2, \dots, n\}, \omega_i \neq \omega_j \text{ für } i \neq j\big\},$$
$$P_2\big(\{(\omega_1, \dots, \omega_k)\}\big) = \frac{1}{n \cdot (n-1) \cdots (n-k+1)} = \frac{(n-k)!}{n!} \quad \Big(= \frac{1}{|\Omega_2|}\Big)$$

3. ohne Zurücklegen, ohne Beachtung der Reihenfolge:

$$\Omega_3 = \big\{A \subset \{1, 2, \dots, n\} : |A| = k\big\},$$
$$P_3\big(\{A\}\big) = \frac{1}{\binom{n}{k}} = \frac{k!(n-k)!}{n!} \quad \Big(= \frac{1}{|\Omega_3|}, \text{ denn es gibt } \binom{n}{k} \text{ versch. } k\text{-elementige Teilmengen}\Big)$$

4. mit Zurücklegen, ohne Beachtung der Reihenfolge:

$$\Omega_4 = \big\{(\ell_1, \ell_2, \dots, \ell_n) \in \mathbb{N}_0^n : \ell_1 + \ell_2 + \cdots + \ell_n = k\big\}$$

(ℓ_i gibt an, wie oft Kugel i gezogen wurde)

Für $(\ell_1, \ell_2, \dots, \ell_n) \in \Omega_4$ gibt es

$$\binom{k}{\ell_1, \ell_2, \dots, \ell_n} := \frac{k!}{\ell_1! \cdot \ell_2! \cdot \cdots \cdot \ell_n!} \qquad \text{„Multinomialkoffizient“}$$

verschiedene $\omega = (\omega_1, \dots, \omega_k) \in \Omega_1$ mit

$$\big|\{1 \leq j \leq k : \omega_j = i\}\big| = \ell_i \quad \text{für } i = 1, \dots, n.$$

$$P_4\big(\{(\ell_1, \dots, \ell_n)\}\big) = \binom{k}{\ell_1, \ell_2, \dots, \ell_n}\Big(\frac{1}{n}\Big)^k, \quad (\ell_1, \dots, \ell_n) \in \Omega_4$$

[6] in dem Sinne, dass ein „Volumen“ $\mathrm{vol}(A)$ definierbar ist

Bemerkung 1.18. Es gilt

$$|\Omega_4| = \binom{n+k-1}{k} \quad \left(= \binom{n+k-1}{n-1} \right)$$

Ein "Zähltrick": Lege k Kugeln und $n-1$ "Trennstäbe" – also insgesamt $n+k-1$ Objekte – in eine Reihe:

$$\underbrace{\bigcirc\cdots\bigcirc}_{\ell_1 \text{ Kugeln}} \,|\, \underbrace{\bigcirc\bigcirc\cdots\bigcirc}_{\ell_2 \text{ Kugeln}} \,|\, \underbrace{\;}_{\ell_3 = 0} \,|\cdots|\, \underbrace{\bigcirc\bigcirc\cdots\bigcirc}_{\ell_{n-1} \text{ Kugeln}} \,|\, \underbrace{\bigcirc\cdots\bigcirc}_{\ell_n \text{ Kugeln}}$$

Insbesondere ist die Verteilung auf Ω_4 aus Beispiel 1.17, 4. nicht die uniforme.

Die uniforme Verteilung auf der Menge Ω_4 aus Beispiel 1.17, 4. heißt auch die „Bose-Einstein-Verteilung", die in Beispiel 1.17, 4. tatsächlich betrachtete Verteilung heißt (im Kontext der statistischen Phyisk) auch die „Maxwell-Boltzmann-Verteilung".

Beispiel. Eine Hörsaalreihe habe n Plätze, darauf nehmen m ($\leq n/2$) Männer und $n-m$ Frauen rein zufällig Platz.

Die Wahrscheinlichkeit, dass keine zwei Männer nebeneinander sitzen

$$= \frac{\binom{n-m+1}{m}}{\binom{n}{m}}$$

Beispiel 1.19 (Hypergeometrische Verteilung)**.** Eine Urne enthalte n Kugeln, davon s schwarze und w weiße ($s+w=n$), ziehe k-mal ohne Zurücklegen,

$$\mathrm{Hyp}_{s,w,k}\big(\{\ell\}\big) = \frac{\binom{s}{\ell}\binom{w}{k-\ell}}{\binom{s+w}{k}}, \quad \ell = 0,1,\ldots,k$$

ist die W'keit, genau ℓ schwarze Kugeln zu ziehen.

Beispiel 1.20 (p-Münzwurf)**.** 1. $\Omega = \{0,1\}$, $\mathrm{Ber}_p(\{1\}) = p = 1 - \mathrm{Ber}_p(\{0\})$ mit einem $p \in [0,1]$ („Bernoulli-Verteilung"[7])

2. n-facher p-Münzwurf (mit $p \in [0,1]$): $\Omega = \{0,1\}^n$,

$$\mathrm{Ber}_p^{\otimes n}\Big(\big\{(\omega_1,\ldots,\omega_n)\big\}\Big) = p^{|\{i\leq n\,:\,\omega_i=1\}|}(1-p)^{|\{i\leq n\,:\,\omega_i=0\}|}$$

3. Binomialverteilung (zum Parameter n und p, $n \in \mathbb{N}$, $p \in [0,1]$):

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

(dies ist die W'keit, beim n-fachen Münzwurf genau k Erfolge zu beobachten)

[7] nach Jakob Bernoulli, 1654–1705

Beispiel 1.21 (Geometrische Verteilung). $p \in (0,1)$, $\Omega = \mathbb{N}_0$,

$$\mathrm{Geom}_p\big(\{k\}\big) = p(1-p)^k, \quad k \in \mathbb{N}_0$$

ist die W'keit, bei wiederholtem p-Münzwurf genau k Misserfolge vor dem ersten Erfolg zu beobachten

Beachte: Manche Autoren betrachten die geometrische Verteilung auf $\mathbb{N}$ (statt auf $\mathbb{N}_0$), dann ist das Gewicht $p(1-p)^{k-1}$ und die Interpretation „k Würfe (einschließlich) bis ersten Erfolg."

Beispiel 1.22 (Negative Binomialverteilung, auch Pascal-Verteilung genannt). Für $r \in (0,\infty)$, $p \in (0,1)$ ist

$$\mathrm{NegBin}_{r,p}\big(\{k\}\big) = \binom{-r}{k}(-1)^k p^r (1-p)^k, \quad k \in \mathbb{N}_0$$

wobei $\binom{-r}{k} := \frac{(-r)(-r-1)\cdots(-r-k+1)}{k!}$ $\Big(= (-1)^k \binom{r+k-1}{k}$ für $r \in \mathbb{N}\Big)$.

$\mathrm{NegBin}_{r,p}\big(\{k\}\big)$ ist die W'keit, insgesamt k Misserfolge vor dem r-ten Erfolg in einer p-Münzwurffolge zu sehen ($\mathrm{NegBin}_{1,p} = \mathrm{Geom}_p$).

(Beachte

$$\sum_{k=0}^{\infty} \binom{-r}{k}(-1)^k p^r (1-p)^k = p^r \big(1+p-1\big)^{-r} = 1$$

und allg. für $r > 0$ $\sum_{k=0}^{\infty} \binom{-r}{k} x^k = (1+x)^{-r}$ (für $|x| < 1$) mittels Taylor-Entwicklung in $x = 0$ bzw. allgemeinem Binomischem Lehrsatz.)

Beispiel 1.23 (Multinomialverteilung). $s \in \{2,3,\dots\}$, $p_1,\dots,p_s \in [0,1]$, $p_1 + \cdots + p_s = 1$, $n \in \mathbb{N}$, $\Omega = \big\{(k_1,\dots,k_s) \in \mathbb{N}_0^s : k_1 + \cdots + k_s = n\big\}$,

$$\mathrm{Mult}_{n;p_1,\dots,p_s}\big(\big\{(k_1,\dots,k_s)\big\}\big) = \binom{n}{k_1,k_2,\dots,k_s} p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s}$$

Interpetation: n Züge mit Zurücklegen ohne Beachtung der Reihenfolge aus einer Urne mit s Kugeln (s verschiedene „Farben"), Farbe i wird mit W'keit p_i gezogen), obiges ist die W'keit, genau k_i-mal Farbe i zu ziehen für $i = 1,2,\dots,s$.

Beispiel 1.24 (Poissonverteilung[8]). $\lambda \in (0,\infty)$,

$$\mathrm{Poi}_\lambda\big(\{k\}\big) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{N}_0$$

Proposition 1.25 (Poissonapproximation der Binomialverteilung). *Seien $p_n \in [0,1]$ mit $p_n \to 0$ und $np_n \to \lambda \in (0,\infty)$ für $n \to \infty$, so gilt für jedes $k \in \mathbb{N}_0$*

$$\mathrm{Bin}_{n,p_n}\big(\{k\}\big) \underset{n\to\infty}{\longrightarrow} \mathrm{Poi}_\lambda\big(\{k\}\big).$$

[8] nach Siméon Denis Poisson, 1781–1840

Beweis. Es ist

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} = \underbrace{\frac{n(n-1)\cdots(n-k+1)}{k!\, n^k}}_{\to 1/k!} \Big(\underbrace{np_n}_{\to\lambda} \Big)^k \underbrace{\Big(1 - \frac{np_n}{n}\Big)^n}_{\to e^{-\lambda}} \big(1-p_n\big)^{-k}$$

$$\longrightarrow e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{für } n \to \infty.$$

□

Prop. 1.25 motiviert, warum die Poissonverteilung oft in Anwendungssituationen vorkommt, in denen man viele unabhängige Ereignisse betrachtet, von denen jedes nur mit einer sehr kleinen W'keit eintritt – man denke etwa an Schädensfälle bei Versicherungen, Zerfallsereignisse in einer Probe radioaktiven Materials oder an genetische Mutationen.

Beispiel 1.26. L. von Bortkewitsch[9] berichtete in seinem Buch *Das Gesetz der kleinen Zahlen*, Teubner, 1898 verschiedene Datensätze, die gut zur Poissonverteilung passen.

Speziell in § 12, 4. („Die durch Schlag eines Pferdes im preußischen Heere getöteten") werden für 20 Jahre (1875–1894) und 10 Armeekops der preußischen Kavallerie, also insgesamt $20 \cdot 10 = 200$ „Korpsjahre" berichtet, in wievielen davon sich x Todesfälle durch Schlag eines Pferds ereigneten (Tabelle b) auf S. 25):

Ergebnis x	Anz. „Korpsjahre"
0	109
1	65
2	22
3	3
4	1
≥5	0

Angenommen, die Anzahl durch Schlag eines Pferdes während eines Jahres in einem Korps getöteter Soldaten wäre Poi_λ-verteilt mit $\lambda = 0{,}61$, so würden wir das Resultat x je $200 \times \mathrm{Poi}_{0{,}61}(x)$-mal erwarten:

Ergebnis x	Anz. „Korpsjahre"	$200 \times \mathrm{Poi}_{0{,}61}(x)$
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
≥5	0	0,08

Von Bortkewitsch, a.a.O., S. 25 schreibt: „Die Kongruenz der Theorie mit der Erfahrung lässt [...], wie man sieht, nichts zu wünschen übrig."

Übrigens: Wie ist von Bortkewitsch auf $\lambda = 0{,}61$ gekommen?

Die beobachtete „mittlere Anzahl Todesfälle pro Korpsjahr" in den Daten ist

$$\widehat{\lambda} = \frac{109}{200} \cdot 0 + \frac{65}{200} \cdot 1 + \frac{22}{200} \cdot 2 + \frac{3}{200} \cdot 3 + \frac{1}{200} \cdot 4 + 0 = 0{,}61$$

[9] Ladislaus von Bortkewitsch, 1868–1931

und es ist auch

$$\sum_{x=0}^{\infty} x\mathrm{Poi}_\lambda(x) = \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!}e^{-\lambda} = \lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}e^{-\lambda} = \lambda$$

(„der Erwartungswert von Poi_λ ist λ“) und somit ist obiges der naheliegende „Momentenschätzer“ – wir werden darauf zurückkommen.

1.4 Zufallsvariablen

Zufallsvariablen (oft abgekürzt ZV; manchmal auch Zufallsgrößen gennant) sind (in einem gewissen Sinn sogar: die) „Fundamentalobjekte“ der Stochastik[10], sie sind zudem oft sehr angenehm zum Rechnen/Notieren und zum intuitiven Argumentieren über zufällige Vorgänge.

So kann man die Beispiel-Instanzen aus Beispiel 1.4 auch folgendermaßen aussprechen:

1. (a) Sei W das Ergebnis eines Wurfs eines fairen 6er-Würfels (Wertebereich $\{1,2,3,4,5,6\}$).
   (b) Werfe Würfel dreimal, seien W_1, W_2, W_3 Ergebnisse des 1., 2., 3. Wurfs des Würfels ($W = (W_1, W_2, W_3)$ hat Wertebereich $\{1,2,3,4,5,6\}^3$).
   (c) n verschiedene Objekte in zufälliger Reihenfolge: Sei $X = (X_1, X_2, \ldots, X_n)$ eine zufällige Permutation von $1, 2, \ldots, n$
   (Wertebereich $\{(x_1, x_2, \ldots, x_n) : x_1, \ldots, x_n \in \{1, 2, \ldots, n\}$ paarweise verschieden$\}$).
2. Sei M das Ergebnis eines (möglicherweise verfälschten) Münzwurfs (Wertebereich $\{\mathrm{Kopf}, \mathrm{Zahl}\}$).
3. Wähle uniform eine Email aus meiner Inbox (die deutsche und englische Emails enthält, die jeweils Spam sein können oder nicht), sei X die Sprache und Y der Spam-Status der betrachteten Email ((X, Y) hat Wertebereich $\{\texttt{Deutsch}, \texttt{Englisch}\} \times \{\texttt{Spam}, \texttt{keinSpam}\}$)
4. Wir werfen eine faire Münze, bis zum ersten Mal Kopf kommt. G zählt, wie oft bis dahin Zahl gefallen ist (Wertebereich $\mathbb{N}_0$).

Beispiel 1.27. Wir werfen 2 Würfel (naheliegende Modellierung: $\Omega = \{1,2,\ldots,6\}^2$ [und $\mathcal{F} = 2^\Omega$, P=uniforme Verteilung auf Ω]) und beobachten die Augensumme, nennen wir sie X.

Wir könnten übergehen zu $\Omega' = \{2,3,\ldots,12\}$ [und entsprechendem $P'(\{x\}) = \frac{6-|x-7|}{36}$] oder wir betrachten

$$\begin{array}{rccc} X: & \Omega & \to & \Omega' \\ & \cup & & \\ & (\omega_1,\omega_2) & \mapsto & \omega_1+\omega_2 \end{array}$$

und Ereignisse

$$\{X = x\} := \{(\omega_1,\omega_2) \in \Omega : \omega_1 + \omega_2 = x\} \quad \text{für } x = 2,3,\ldots,12.$$

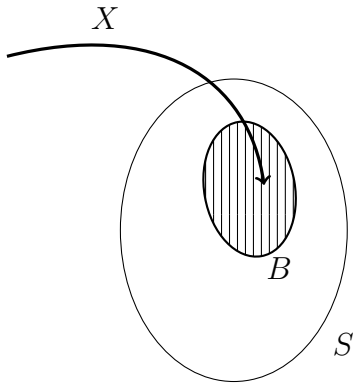
[10] Man kann ein Zufallsexperiment auch stets auffassen als: „Der Zufall wählt aus einer Menge S möglicher Realisierungen der Zufallsvariable X eine aus“ und sich in diesem Sinn den „Umweg“ über die Diskussion von Ereignissen und Wahrscheinlichkeitsräume (zunächst) ersparen, tatsächlich stellt das Buch von Kersting & Wakolbinger [KW] Zufallsvariablen auch gleich an den Anfang der Diskussion; die beiden Zugänge sind logisch äquivalent, unterscheiden sich aber etwas in der „Betonung“. Siehe auch Bemerkung 1.36 unten.

Definition 1.28. $(\Omega, \mathcal{F}, P)$ ein W'raum, $(S, \mathcal{S})$ ein messbarer Raum (der „Wertebereich").

Eine Abbildung $X : \Omega \to S$ heißt eine *Zufallsvariable* (mit Wertebereich S), wenn gilt

$$\forall\, B \in \mathcal{S} : X^{-1}(B) = \big\{\omega \in \Omega : X(\omega) \in B\big\} \in \mathcal{F}.$$

(im Sinne der Maßtheorie ist X eine messbare (strikt: eine $\mathcal{F}$-$\mathcal{S}$)-messbare Abbildung von Ω nach S.)



„Logo-Bild" für eine Zufallsvariable X.

Wir schreiben

$$\{X \in B\} := X^{-1}(B)$$

für das Ereignis „X nimmt einen Wert in B an" und abkürzend oft auch $\{X = x\} := \big\{X \in \{x\}\big\}$; im Fall $S = \mathbb{R}$ oft auch $\{X \le x\} := \big\{X \in (-\infty, x]\big\}$, etc.

Im Fall reellwertiger Zufallsvariablen lassen wir oft auch die Werte $+\infty$ oder $-\infty$ zu (Übergang von $S = \mathbb{R}$ zu $S = \overline{\mathbb{R}}$).

Beachte hier die übliche Notationskonvention: ZV werden meist mit Großbuchstaben benannt, mögliche Werte („Realisierungen") mit Kleinbuchstaben.

Bericht 1.29. 1. Wenn Ω und S abzählbar sind (und wir in kanonischer Weise die jeweiligen Potenzmengen als σ-Algebren verwenden), so ist jede Funktion eine Zufallsvariable.

2. Sei $\mathcal{S} = \sigma(\mathcal{E})$ für ein $\mathcal{E} \subset 2^S$, dann ist X eine Zufallsvariable, sofern gilt

$$\forall\, C \in \mathcal{E} : X^{-1}(C) \in \mathcal{F}$$

(d.h. es genügt Messbarkeit auf einer Erzeugermenge der σ-Algebra zu prüfen, speziell für $S = \mathbb{R}$ [oder $\overline{\mathbb{R}}$] für Mengen der Form $(-\infty, x]$), denn $\{B \subset S : X^{-1}(B) \in \mathcal{F}\}$ *ist* eine σ-Algebra (Übung), umfasst nach Voraussetzung $\mathcal{E}$.

3. Im Fall $\Omega = \mathbb{R}^m$, $S = \mathbb{R}^d$ (jeweils mit der zugehörigen Borel-σ-Algebra versehen) ist jede stetige Abbildung $f : \Omega \to S$ messbar

Beispiel 1.30 (Indikatorvariable)**.** $A \in \mathcal{F}$, $\mathbf{1}_A : \Omega \to \{0, 1\}$,

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \text{falls } \omega \in A, \\ 0, & \text{falls } \omega \notin A \end{cases}$$

$\mathbf{1}_A$ heißt die Indikatorvariable des Ereignisses A.

Zufallsvariablen sind nicht zuletzt deshalb nützlich für die Modellierung zufälliger Vorgänge, weil man mit ihnen gewissermaßen genauso operieren und „rechnen“ kann wie mit anderen Variablen (oder „unbestimmten Größen“) in der Mathematik:

Beobachtung 1.31 (Rechnen mit ZVn)**.** Sind $X_1, X_2, \dots$ reelle ZVn, so sind auch

$$(X_1, X_2), X_1 + X_2, X_1 - X_2, X_1 \cdot X_2, X_1 \wedge X_2, X_1 \vee X_2,$$
$$\sup_n X_n, \inf_n X_n, \limsup_{n\to\infty} X_n \text{ und } \liminf_{n\to\infty} X_n$$

Zufallsvariablen, denn

$$\big\{(X_1, X_2) \in (-\infty, x_1] \times (-\infty, x_2]\big\} = \{X_1 \le x_1\} \cap \{X_2 \le x_2\} \in \mathcal{F},$$
$$\text{die Abbildungen } (x, y) \mapsto x + y, \text{ bzw. } \mapsto x - y, \ \mapsto x \cdot y, \text{ etc. sind stetig,}$$
$$\Big\{\sup_n X_n \le x\Big\} = \bigcap_{n\in\mathbb{N}} \{X_n \le x\} \in \mathcal{F}, \text{ etc.,}$$
$$\limsup_{n\to\infty} X_n = \inf_{n\in\mathbb{N}} \sup_{m\ge n} X_m, \ \liminf_{n\to\infty} X_n = \sup_{n\in\mathbb{N}} \inf_{m\ge n} X_m.$$

Insbesondere können wir das Ereignis

$$\big\{X_n \text{ konvergiert}\big\} = \Big\{\limsup_{n\to\infty} X_n \le \liminf_{n\to\infty} X_n\Big\}$$
$$\left(= \Big\{\Big(\limsup_{n\to\infty} X_n, \liminf_{n\to\infty} X_n\Big) \in \big\{(x, y) \in \overline{\mathbb{R}}^2 : x \le y\big\}\Big\}\right)$$

sinnvoll betrachten.

Beobachtung 1.32 (Verkettung messbarer Abbildungen, Funktionen von Zufallsvariablen)**.** $(S, \mathcal{S})$, $(S', \mathcal{S}')$, messbare Räume, $f : \Omega \to S$, $g : S \to S'$ messbar, dann ist

$$g \circ f : \Omega \to S', \quad g \circ f(\omega) = g\big(f(\omega)\big)$$

($\mathcal{F}$-$\mathcal{S}$-)messbar, denn

$$\big(g \circ f\big)^{-1}(B') = f^{-1}\Big(\underbrace{g^{-1}(B')}_{\in\mathcal{S}}\Big) \in \mathcal{F} \quad \text{für jedes } B' \in \mathcal{S}'.$$

Insbesondere ist für eine S-wertige ZV X auf $(\Omega, \mathcal{F}, P)$

$$g(X) := g \circ X$$

wiederum eine Zufallsvariable.

(Interpretation: Wir werten die Funktion g an der zufälligen Stelle X aus.)

Beobachtung und Definition 1.33. *1. (Verteilung) X ZV (auf einem W'raum $(\Omega, \mathcal{F}, P)$) mit Werten in S,*

$$P_X(B) := P\big(\{X \in B\}\big), \quad B \in \mathcal{S}$$

definiert ein W'maß auf $(S, \mathcal{S})$ (Übung), wir nennen P_X die Verteilung *von X (unter P) und schreiben auch $\mathscr{L}_P(X) := P_X$, oft auch nur $\mathscr{L}(X)$, wenn P fixiert ist oder aus dem Kontext klar.*

(Das $\mathscr{L}$ erinnert an English "law" bzw. Französisch «loi», d.h. „Gesetz".)

2. μ ein W'maß auf S, wir schreiben $X \sim \mu$, wenn $\mathscr{L}_P(X) = \mu$.

3. X und Y ZVn mit Werten in S heißen identisch verteilt, *wenn $\mathscr{L}_P(X) = \mathscr{L}_P(Y)$ (man schreibt dies auch als $X =^d Y$)*

4. $X_1, X_2, \dots, X_d$ (reelle) ZVn, $Y := \big(X_1, X_2, \dots, X_d\big)$, so heißt $\mathscr{L}_P(Y)$ die gemeinsame Verteilung *der $X_1, X_2, \dots, X_d$, $\mathscr{L}_P(X_i)$ heißt die i-te* Randverteilung *(oder* Marginalverteilung*) von Y.*

Schreibweise. Wir kürzen (auch im Folgenden) oft ab $P(X \in B) := P\big(\{X \in B\}\big)$, $P(X = x) := P\big(\{X = x\}\big)$, $P(X_1 \in B_1, X_2 \in B_2) := P\big(\{X_1 \in B_1\} \cap \{X_2 \in B_2\}\big)$, etc.

Beispiel 1.34. 1. Für $A \in \mathcal{F}$ ist $\mathscr{L}_P(\mathbf{1}_A) = \mathrm{Ber}_{P(A)}$.

2. Sei $\Omega = \{0,1\}^n$, $P = \mathrm{Ber}_p^{\otimes n}$ (aus Bsp. 1.20, 2.), $X_i(\omega) = \omega_i$, $Y := X_1 + X_2 + \cdots + X_n$, so ist

$$\mathscr{L}_P(X_i) = \mathrm{Ber}_p,\ \mathscr{L}_P(Y) = \mathrm{Bin}_{n,p}$$

3. (Augensumme beim zweifachen Münzwurf, vgl. auch Bsp. 1.4 1. (b) und Bsp. 1.27) W_1 und W_2 das Ergebnis des ersten bzw. des zweiten Wurfs beim zweimaligen Werfen eines fairen Würfels und $X := W_1 + W_2$ die Augensumme (z.B. formalisiert via $\Omega = \{1,2,\dots,6\}^2$ mit $p\big((\omega_1,\omega_2)\big) = 1/36$ für $(\omega_1,\omega_2) \in \Omega$, $W_i\big((\omega_1,\omega_2)\big) = \omega_i$, $i = 1,2$), so hat X Wertebereich $S_X = \{2,3,\dots,12\}$ und Verteilung $\mathscr{L}(X) =: \mu$ mit

$$\begin{aligned}\mu(\{x\}) = P(X = x) &= \sum_{(w_1,w_2)\,:\,w_1+w_2=x} P(W_1 = w_1, W_2 = w_2)\\ &= \frac{\#\{(w_1,w_2) \in \{1,2,\dots,6\}^2 : w_1 + w_2 = x\}}{36} = \frac{6 - |7-x|}{36}, \quad x \in S_X\end{aligned}$$

Bemerkung 1.35. Die Randverteilungen legen (i.A.) nicht die gemeinsame Verteilung fest, z.B.:

Wir haben eine faire Münze M_1 und zwei gezinkte Münzen M_2, M_3, wobei

$$P(M_2 = K) = \frac{3}{4},\ P(M_3 = K) = \frac{1}{4}.$$

Wir werfen erst M_1, wenn M_1 K (Kopf) zeigt, so werfen wir dann M_2, sonst M_3.

Sei X_i = Resultat des i-ten Wurfs, $i = 1,2$.

Die gemeinsame Verteilung von (X_1, X_2) ist

X_1 \ X_2	K	Z	
K	$\frac{1}{2}\cdot\frac{3}{4} = \frac{3}{8}$	$\frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8}$	$\frac{1}{2}$
Z	$\frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8}$	$\frac{1}{2}\cdot\frac{3}{4} = \frac{3}{8}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	

also $P(X_1 = K) = P(X_2 = K) = \frac{1}{2}$ und dieselben Randverteilungen ergeben sich, wenn man zwei Mal M_1 wirft, aber die gemeinsame Verteilung wäre eine andere.

Bemerkung 1.36 (Kanonisches Modell für eine ZV)**.** Man kann eine Zufallsvariable X mit Wertebereich S und Verteilung μ stets in „kanonischer Weise" mit der Wahl $\Omega = S$ und $P = \mu$ und geeignetem $\mathcal{F} \subset 2^S$ (im diskreten Fall kann man $\mathcal{F} = 2^S$ wählen) als $X = \mathrm{Id}_S$ auf dem W'raum $(S, \mathcal{F}, \mu)$ formulieren.

Man kann daher genausogut die mathematische Modellierung eines Zufallsphänomens mit der Formulierung geeigneter Zufallsvariablen samt Verteilung beginnen (dies ist der in dem Buch von G. Kersting und A. Wakolbinger [KW] beschrittene Weg).

Definition 1.37. X ZV mit Werten in $\mathbb{R}$

$$F_X : \mathbb{R} \to [0,1], \quad F_X(x) = P(X \le x),\ x \in \mathbb{R}$$

heißt die *Verteilungsfunktion* von X. (Strenggenommen: die Verteilungsfunktion von $\mathscr{L}_P(X)$)

Beobachtung 1.38 (Erzeugung reeller ZVn mit vorgegebener Verteilung)**.** *Sei $F : \mathbb{R} \to [0,1]$ eine Verteilungsfunktion,*

$$F^{-1}(t) := \inf\big\{x \in \mathbb{R} : F(x) \ge t\big\}, \quad t \in [0,1]$$

die inverse Verteilungsfunktion oder Quantilfunktion (aus Bem. 1.13), U reelle ZV, $U \sim \mathrm{Unif}_{[0,1]}$, dann hat

$$X := F^{-1}(U)$$

die Verteilungsfunktion $F_X = F$.

Beweis. Es gilt $F^{-1}(t) \le x \iff t \le F(x)$ nach Bem. 1.13, somit ist für $x \in \mathbb{R}$

$$P(X \le x) = P(F^{-1}(U) \le x) = P(U \le F(x)) = P(0 \le U \le F(x)) = F(x) - 0 = F(x).$$

□

Beispiel 1.39. Exp_θ hat Verteilungsfunktion $F_{\mathrm{Exp}_\theta}(x) = (1 - e^{-\theta x})\mathbf{1}_{[0,\infty)}(x)$ mit inverser Funktion $F^{-1}_{\mathrm{Exp}_\theta}(t) = -\frac{1}{\theta}\log(1-t)$,

also ist $-\dfrac{1}{\theta}\log(1-U) \sim \mathrm{Exp}_\theta$ (und natürlich ebenso $-\dfrac{1}{\theta}\log(U)$)

Beobachtung 1.40 (Dichtetransformation im Fall $\mathbb{R}^1$)**.** *X reelle ZV mit Dichte f_X, d.h. $F_X(x) = \int_{-\infty}^{x} f_X(z)\,dz$, $I \subset \mathbb{R}$ (möglicherweise unbeschränktes) offenes Intervall mit $P(X \in I) = 1$, $J \subset \mathbb{R}$, $\varphi : I \to J$ stetig differenzierbar, bijektiv.*

Dann hat $Y := \varphi(X)$ die Dichte

$$f_Y(y) = \begin{cases} \dfrac{f_X\big(\varphi^{-1}(y)\big)}{\big|\varphi'\big(\varphi^{-1}(y)\big)\big|}, & y \in J, \\ 0, & y \notin J. \end{cases}$$

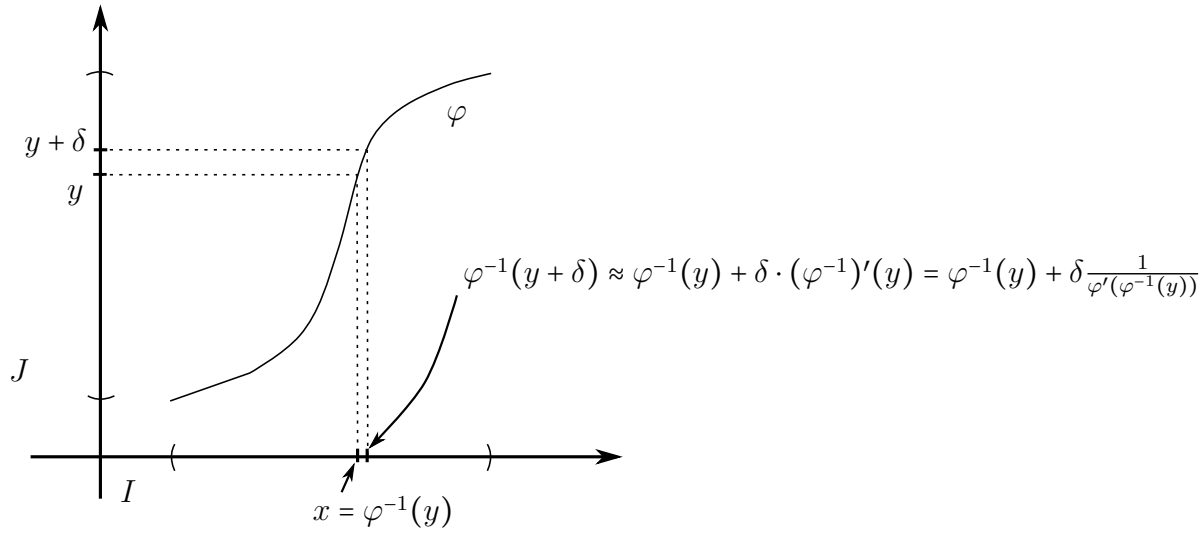
Beweis. φ muss offenbar strikt wachsend oder strikt fallend sein, wir betrachten den wachsenden Fall.

Für $z < \inf J$ ist $P(Y \le z) = 0$, für $z > \sup J$ ist $P(Y \le z) = 1$.

Sei $z \in J$:

$$P(Y \le z) = P\big(\varphi(X) \le z\big) = P\big(X \le \varphi^{-1}(z)\big)$$
$$= \int_{-\infty}^{\varphi^{-1}(z)} f_X(x)\,dx = \int_{-\infty}^{z} f_X\big(\varphi^{-1}(y)\big)\frac{1}{\big|\varphi'\big(\varphi^{-1}(y)\big)\big|}dy,$$

wobei wir $x = \varphi^{-1}(y)$ (und somit $\frac{dx}{dy} = \frac{1}{\varphi'(\varphi^{-1}(y))}$ substituiert haben). Siehe auch die Skizze unten. □



Beispiel 1.41. $X \sim \mathcal{N}_{0,1}$, $\mu \in \mathbb{R}$, $\sigma > 0$, so ist $Y := \sigma X + \mu \sim \mathcal{N}_{\mu,\sigma^2}$ (Übung).

Bericht 1.42 (Dichtetransformation im $\mathbb{R}^d$)**.** X $\mathbb{R}^d$-wertige ZV mit Dichte f_X, $I \subset \mathbb{R}^d$ offen mit $P(X \in I) = 1$, $J \subset \mathbb{R}^d$ offen, $\varphi : I \to J$ bijektiv, stetig differenzierbar mit Ableitung

$$\varphi'(x) = \left(\frac{\partial \varphi_i}{\partial x_j}(x)\right)_{i,j=1}^{d} \qquad (\text{„Jacobi-Matrix“}),$$

dann hat $Y := \varphi(X)$ die Dichte

$$f_Y(y) = \begin{cases} \dfrac{f_X\big(\varphi^{-1}(y)\big)}{\big|\det \varphi'\big(\varphi^{-1}(y)\big)\big|}, & y \in J, \\ 0, & y \notin J. \end{cases}$$

Beweise finden sich in Analysis-Lehrbüchern, z.B. G. Kersting und M. Brokate, *Maß und Integral*, S. 107, H. Heuser, *Analysis, Teil 2*, Satz 205.2 (“Substitutions-Regel”), O. Forster, *Analysis 3*, Kap. 9, Satz 1 (“Transformationsformel”).

Wir betrachten hier nur folgende Heuristik (im Fall $d = 2$): Lokal sieht

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \varphi(x) = \begin{pmatrix} \varphi_1(x) \\ \varphi_2(x) \end{pmatrix}$$

„aus wie“

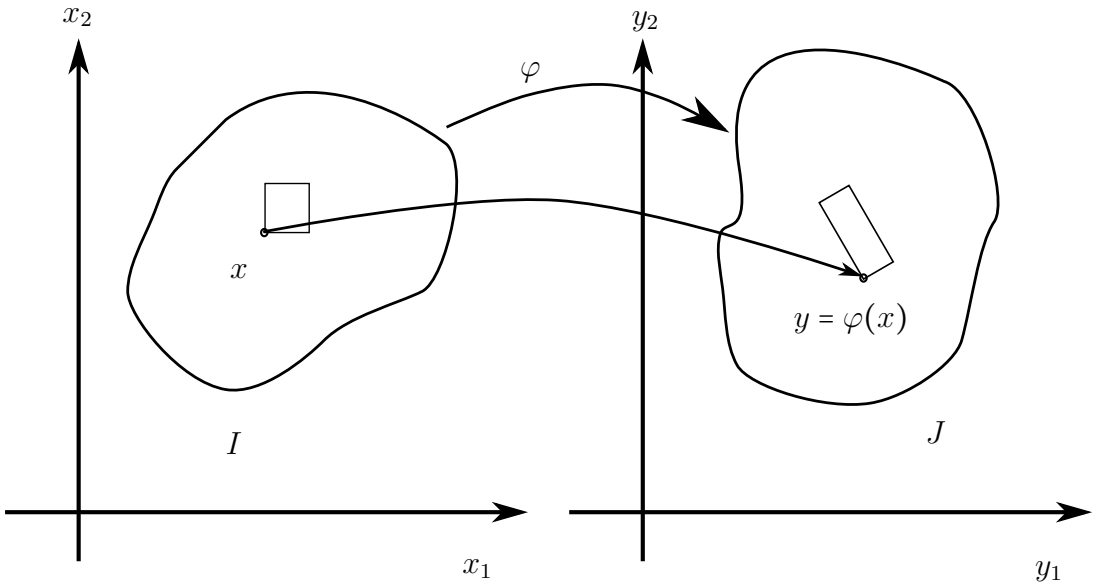
$$\begin{aligned}\varphi(x') &\approx \varphi(x) + \varphi'(x)\cdot(x'-x)\\ &= \varphi(x) + \begin{pmatrix} \frac{\partial}{\partial x_1}\varphi_1(x) & \frac{\partial}{\partial x_2}\varphi_1(x) \\ \frac{\partial}{\partial x_1}\varphi_2(x) & \frac{\partial}{\partial x_2}\varphi_2(x)\end{pmatrix}\cdot\begin{pmatrix} x_1'-x_1 \\ x_2'-x_2\end{pmatrix}\end{aligned}$$

(plus Terme, die $O(\|x'-x\|^2)$ sind), also:

die Fläche der Größe $h_1\cdot h_2$ „rund um x“

wird auf

$\approx$ Fläche $h_1\cdot h_2\cdot|\det\varphi'(x)|$ „rund um y“ abgebildet.



Wenden wir dies auf $Y=\varphi(X)$ an, so bedeutet das anschaulich: Für $y=\varphi(x)\in J$ (und sehr kleines $h>0$) ist

$$\begin{aligned}f_Y(y)h^2 &\approx \mathbb{P}\big(Y \text{ nimmt Wert in einem Quadrat der Fläche } h^2 \text{ mit „Aufpunkt“ } y \text{ an}\big)\\ &\approx \mathbb{P}\big(X \text{ nimmt Wert in einem Quader der Fläche } h^2/|\det\varphi'(x)| \text{ mit „Aufpunkt“ } x \text{ an}\big)\\ &\approx f_X(x)\frac{h^2}{|\det\varphi'(x)|} = \frac{f_X(\varphi^{-1}(y))}{|\det\varphi'(\varphi^{-1}(y))|}h^2.\end{aligned}$$

Kapitel 2

Bedingte Wahrscheinlichkeiten und Unabhängigkeit

2.1 Bedingte Wahrscheinlichkeiten

Beispiel 2.1. Wir ziehen zwei Kugeln ohne Zurücklegen aus einer Urne mit $s > 0$ schwarzen und $w > 0$ weißen Kugeln.

Modell: Nummeriere die Kugeln, $1, \dots, w$ seien weiß, $w+1, \dots, w+s$ schwarz,

$\Omega = \big\{(i,j) : 1 \le i, j \le w+s, i \ne j\big\}$,

P = uniforme Verteilung auf Ω ($|\Omega| = (w+s)(w+s-1)$),

betrache die Ereignisse

$$A = \big\{\text{erste Kugel ist weiß}\big\} = \{(i,j) \in \Omega : i \le w\},$$
$$B = \big\{\text{zweite Kugel ist weiß}\big\} = \{(i,j) \in \Omega : j \le w\}.$$

Ohne weitere Informationen ist

$$P(B) = \frac{|B|}{|\Omega|} = \frac{w(w+s-1)}{(w+s)(w+s-1)} = \frac{w}{w+s}.$$

Nehmen wir an, wir haben den ersten Zug beobachtet und gesehen, dass A eingetreten ist. Mit dieser Information sollte die W'keit von B

$$\frac{w-1}{w+s-1} < \frac{w}{w+s}$$

sein (denn es „wurde schon eine weiße Kugel verbraucht").

Beobachtung und Definition 2.2. *Sei $(\Omega, \mathcal{F}, P)$ W'raum, $A \in \mathcal{F}$ mit $P(A) > 0$.*

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

heißt bedingte Wahrscheinlichkeit von B, gegeben A *(für $B \in \mathcal{F}$).*

$P(\cdot \mid A)$ ist ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{F})$, man prüft leicht per Inspektion, dass die Eigenschaften aus Definition 1.3, Normierung und σ-Additivität, erfüllt sind.

Wir lassen $P(B \mid A)$ undefiniert, wenn $P(A) = 0$.

In Beispiel 2.1 ist $P(A) = \frac{w}{w+s}$, $P(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{w(w-1)}{(w+s)(w+s-1)}$, also ergibt sich tatsächlich $P(B \mid A) = \frac{w-1}{w+s-1}$.

Bemerkung 2.3 („Natürlichkeit von Definition 2.2“)**.** Nehmen wir an, wir möchten angesichts der Information „A ist eingetreten“ das W'maß P revidieren zu einem W'maß $\widetilde{P}$ mit

1. $\widetilde{P}(A) = 1$ (d.h. A ist sicher unter $\widetilde{P}$) und
2. $\widetilde{P}(B) = c_A P(B)$ für $B \subset A$ mit einem $c_A > 0$
   (d.h. Teilereignisse von A erhalten bis auf Normierung ihr altes Gewicht).

Dann gilt

$$\widetilde{P}(C) = \frac{P(A \cap C)}{P(A)} \quad \big(= P(C \mid A) \big) \quad \text{für alle } C \in \mathcal{F}.$$

Beweis. Für $C \in \mathcal{F}$ ist

$$\widetilde{P}(C) = \widetilde{P}(A \cap C) + \underbrace{\widetilde{P}(C \setminus A)}_{\leq \widetilde{P}(A^c) = 0} \overset{2.}{=} c_A P(C),$$

mit Wahl $C = A$ und 1. folgt $1 = \widetilde{P}(A) = c_A P(A)$, also $c_A = 1/P(A)$. □

Bemerkung 2.4. $P(B \mid A) \neq P(B)$ kann nicht notwendigerweise als „Kausalität“ (im Sinne von „A beeinflusst, ob B eintritt“) interpretiert werden:

In Beispiel 2.1 ist auch

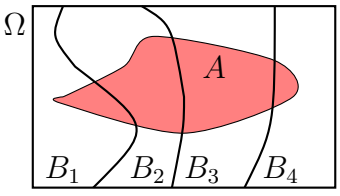
$$P(A \mid B) = \frac{P(B \cap A)}{P(B)} = \frac{w-1}{w+s-1} \neq P(A),$$

aber es passt nicht zu unserer Vorstellung, dass der 2. Zug den 1. Zug beeinflusst.

Satz 2.5. *I abzählbare Indexmenge, $B_i \in \mathcal{F}$ paarweise disjunkt mit $P\big(\bigcup_{i \in I} B_i\big) = 1$ (und $P(B_i) > 0$ für $i \in I$).*

1. *(Formel von der totalen Wahrscheinlichkeit) Für $A \in \mathcal{F}$ gilt*

$$P(A) = \sum_{i \in I} P(B_i) P(A \mid B_i)$$



   Man kann dies anhand eines Diagramms veranschaulichen: $P(A \mid B_i)$ ist der Anteil von $A \cap B_i$ an der „Wahrscheinlichkeitsmasse“ $P(B_i)$ von B_i

2. *(Formel von Bayes[1]) Für $A \in \mathcal{F}$ mit $P(A) > 0$ und jedes $k \in I$ gilt*

$$P(B_k \mid A) = \frac{P(B_k) P(A \mid B_k)}{\sum_{i \in I} P(B_i) P(A \mid B_i)}$$

[1] nach Thomas Bayes, 1702–1761; die Arbeit (die eine Frage von Laplace beantwortet) wurde posthum 1763 publiziert

Beweis. 1. $\sum_{i\in I} P(B_i)P(A \mid B_i) = \sum_{i\in I} P(A \cap B_i) = P\Big(\bigcup_{i\in I}(A \cap B_i)\Big) = P(A)$

(verwende die σ-Additivität von P).

2. Der Nenner ist = $P(A)$ nach 1., der Zähler ist = $P(A \cap B_k)$ nach Definition. □

Beispiel 2.6 (Medizinische Reihenuntersuchung)**.** Eine Krankheit

- komme bei 2% der Bevölkerung vor („Prävalenz 2%"),
- ein Test schlage bei 95% der Kranken an („Sensitivität 95%"),
- aber auch bei 10% der Gesunden („Spezifität 90%").

Eine zufällig gewählte Person wird mit positivem Resultat getestet. Wie wahrscheinlich ist es, dass sie krank ist?

Sei

$$A = \{\text{Test fällt positiv aus}\},$$
$$B = \{\text{getestete Person ist krank}\},$$

also $P(B) = 0{,}02$, $P(A \mid B) = 0{,}95$, $P(A \mid B^c) = 0{,}1$, somit

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(B^c)P(A \mid B^c)} = \frac{0{,}02 \cdot 0{,}95}{0{,}02 \cdot 0{,}95 + 0{,}98 \cdot 0{,}1} \approxeq 0{,}162$$

(wir sehen: geringe „positive Korrektheit"), andererseits

$$P(B \mid A^c) = \frac{P(B)P(A^c \mid B)}{P(B)P(A^c \mid B) + P(B^c)P(A^c \mid B^c)} = \frac{0{,}02 \cdot 0{,}05}{0{,}02 \cdot 0{,}05 + 0{,}98 \cdot 0{,}9} \approxeq 0{,}0011$$

(recht hohe „negative Korrektheit").

Demnach: Ein negatives Testergebnis schließt die Krankheit mit hoher W'keit aus, ein positives Testergebnis sollte eher als „der Fall sollte weiter beobachtet / untersucht werden" interpretiert werden als „die Testperson ist krank".

S.a. Gerd Gigerenzer, *Das Einmaleins der Skepsis*, Berlin Verlag, 2002, der auch einlädt, den Sachverhalt anschaulich anhand einer „Vierfelder-Tafel" bezogen auf eine Gesamtpopulation der Größe 1000 zu betrachten:

	krank	gesund	Σ
pos. getestet	19	98	117
neg. getestet	1	882	883
Σ	20	980	1000

2.2 Mehrstufige Zufallsexperimente

Wir betrachten folgende Situation:

Wir haben ZVn $X_1, X_2, \dots, X_n$ im Sinn und kennen

1. die Verteilung von X_1,
2. für $2 \le k \le n$ die bedingte Verteilung von X_k, wenn $X_1, X_2, \dots, X_{k-1}$ schon beobachtet wurden.

Frage / Aufgabe : Wie modellieren? (und: wie damit rechnen?)

Beispiel 2.7 (Pólyas Urne[2])**.** $s, w \in \mathbb{N}_0$, $c \in \{-1, 0, 1, 2, \dots\}$, eine Urne enthält anfangs s schwarze und w weiße Kugeln. Wir greifen in jedem Zug einmal rein zufällig hinein und legen dann die gezogene Kugel zurück zusammen mit c weiteren Kugeln derselben Farbe (im Fall $c = -1$ legen wir die gezogene Kugel nicht zurück). Wir setzen

$$X_i = \mathbf{1}_{\{i\text{-te gezogene Kugel ist schwarz}\}}.$$

Beobachtung 2.8 (Multiplikationsformel)**.** *$A_1, A_2, \dots, A_n$ Ereignisse mit*

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

so ist

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

(Beweis per Inspektion, das Produkt rechts teleskopiert)

Satz 2.9 (Konstruktion von W'maßen via bedingte Wahrscheinlichkeiten)**.** *Seien $\Omega_1, \Omega_2, \dots, \Omega_n$ ($\neq \varnothing$) abzählbar, $p_1 : \Omega_1 \to [0,1]$ Wahrscheinlichkeitsgewichtsfunktion, für $2 \le k \le n$ und beliebige $\omega_1 \in \Omega_1, \dots, \omega_{k-1} \in \Omega_{k-1}$ sei $p_{k|\omega_1,\dots,\omega_{k-1}} : \Omega_k \to [0,1]$ W'gewichtsfunktion, setze $\Omega := \Omega_1 \times \Omega_2 \times \dots \times \Omega_n$, wir schreiben $X_i : \Omega \to \Omega_i$ mit $X_i\big((\omega_1, \dots, \omega_n)\big) = \omega_i$ für die i-te Koordinatenprojektion.*

Dann ist $p : \Omega \to [0,1]$ mit

$$p\big((\omega_1, \dots, \omega_n)\big) := p_1(\omega_1) \cdot p_{2|\omega_1}(\omega_2) \cdot p_{3|\omega_1,\omega_2}(\omega_3) \cdot \dots \cdot p_{n|\omega_1,\dots,\omega_{n-1}}(\omega_n)$$

eine W'gewichtsfunktion, das W'maß P auf Ω mit Gewichten p erfüllt

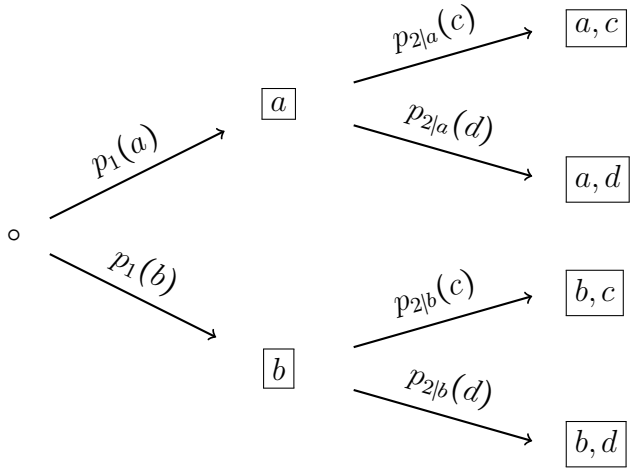
1. *$P(X_1 = \omega_1) = p_1(\omega_1)$ für alle $\omega_1 \in \Omega_1$,*
2. $$P(X_k = \omega_k \mid X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}) = p_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k)$$

für $2 \le k \le n$, $\omega_1 \in \Omega_1, \dots, \omega_{k-1} \in \Omega_{k-1}$, sofern $P(X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}) > 0$.

P ist durch 1. und 2. eindeutig festgelegt.

[2] nach George Pólya, 1887–1985; siehe F. Eggenberger, G. Pólya, Über die Statistik verketteter Vorgänge, Zeitschr. f. Angew. Math. Mech. 3, 279–290, (1923).

Man kann die Situation aus Satz 2.9 in einem Baumdiagramm veranschaulichen (z.B. für $n = 2$, $\Omega_1 = \{a, b\}, \Omega_2 = \{c, d\}$):



Beweis von Satz 2.9. Für $1 \le k \le n, \omega_1 \in \Omega_1, \dots, \omega_k \in \Omega_k$ ist

$$\begin{aligned} P\big(X_1 = \omega_1, \dots, X_k = \omega_k\big) &= \sum_{\omega_{k+1} \in \Omega_{k+1}} \cdots \sum_{\omega_n \in \Omega_n} P\big(\{(\omega_1, \dots, \omega_n)\}\big) \\ &= p_1(\omega_1) p_{2|\omega_1}(\omega_2) \cdots p_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k) \underbrace{\sum_{\omega_{k+1} \in \Omega_{k+1}} p_{k+1|\omega_1,\dots,\omega_k}(\omega_{k+1})}_{=1} \cdot \\ &\qquad \underbrace{\cdots}_{=1} \cdot \underbrace{\sum_{\omega_n \in \Omega_n} p_{n|\omega_1,\dots,\omega_{n-1}}(\omega_n)}_{=1} \\ &= p_1(\omega_1) p_{2|\omega_1}(\omega_2) \cdots p_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k), \end{aligned}$$

insbesondere für $k = 1$ also $P(X_1 = \omega_1) = p_1(\omega_1)$, d.h. 1. gilt und

$$P(\Omega) = \sum_{\omega_1 \in \Omega_1} P(X_1 = \omega_1) = \sum_{\omega_1 \in \Omega_1} p_1(\omega_1) = 1,$$

P ist demnach ein W'maß.

Für $2 \le k \le n$ folgt

$$P\big(X_1 = \omega_1, \dots, X_k = \omega_k\big) = P\big(X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}\big) p_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k),$$

d.h. 2. gilt.

Die Eindeutigkeit von P folgt aus Beob. 2.8 (mit Wahl $A_i = \{X_i = \omega_i\}$). □

Anwendung in Bsp. 2.7 (Pólya-Urne)

Die Urne enthält anfangs s schwarze und w weiße Kugeln, man legt jeweils die gezogene Kugel zurück zusammen mit c weiteren Kugeln derselben Farbe.

Wir betrachten n Züge, $\Omega = \{0,1\}^n \ni \omega = (\omega_1, \dots, \omega_n)$, $X_i = \omega_i$ (mit Interpretation $1 \mathrel{\widehat{=}}$ schwarz), es ist

$$p_{k|\omega_1,\dots,\omega_{k1}}(\omega_k) = \begin{cases} \dfrac{s + c\ell}{s + w + c(k-1)}, & \text{falls } \omega_k = 1, \\ \dfrac{w + c(k-1-\ell)}{s + w + c(k-1)}, & \text{falls } \omega_k = 0 \end{cases}$$

mit $\ell = \omega_1 + \cdots + \omega_{k-1}$.

Satz 2.9 liefert (für $\omega = (\omega_1, \dots, \omega_n)$ mit $\omega_1 + \cdots + \omega_n = m$)

$$\begin{aligned} P\big(\{(\omega_1, \dots, \omega_n)\}\big) &= p_1(\omega_1) \prod_{k=2}^{n} p_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k) \\ &= \frac{\prod_{i=0}^{m-1}(s+ci) \cdot \prod_{j=0}^{n-m-1}(w+cj)}{\prod_{k=0}^{n-1}(s+w+ck)}. \end{aligned}$$

Beachte: dies hängt nicht von der Reihenfolge der Werte ω_i, nur von der Gesamtsumme ab (man sagt dazu auch, dass die X_i „austauschbar“ sind).

Halten wir Spezialfälle fest:

- für $c = 0$ (d.h. Ziehen mit Zurücklegen, also n unabhängige Versuche) ergibt sich $\left(\frac{s}{s+w}\right)^m \left(\frac{w}{s+w}\right)^{n-m}$
- für $c = -1$ (d.h. Ziehen ohne Zurücklegen) ergibt sich

  $$\frac{s(s-1)\cdots(s-m+1) \cdot w(w-1)\cdots(w-n+m+1)}{(s+w)(s+w-1)\cdots(s+w-n+1)} = \frac{\binom{s}{m}\binom{w}{n-m}}{\binom{s+w}{n}} \frac{1}{\binom{n}{m}} = \mathrm{Hyp}_{s,w,n}\big(\{m\}\big) \frac{1}{\binom{n}{m}}$$

  mit der hypergeometrischen Verteilung aus Bsp. 1.19.

  Wir können diese Rechnung folgendermaßen interpretieren: Die Anzahl gezogener schwarzer Kugeln in den n Zügen ist $\mathrm{Hyp}_{s,w,n}$-verteilt, gegeben, dass m schwarze Kugeln gezogen wurden, sind deren Positionen in der Reihenfolge uniform verteilt (es gibt $\binom{n}{m}$ mögliche Wahlen für m Positionen).

- für $c = 1$ ergibt sich

  $$\frac{\frac{(s+m-1)!}{(s-1)!}\frac{(w+n-m-1)!}{(w-1)!}}{\frac{(s+w+n-1)!}{(s+w-1)!}} = \frac{(s+w-1)!}{(s-1)!(w-1)!} \frac{(s+m-1)!(w+n-m-1)!}{(s+w+n-1)!}$$

Sei $S_n := X_1 + \cdots + X_n$ (die Anzahl gezogene schwarze Kugeln unter den ersten n Zügen), für

$0 \le m \le n$ ist (für $c \neq 0$)

$$P(S_n = m) = \sum_{\substack{(w_1,\dots,\omega_n)\in\{0,1\}^n,\\ \omega_1+\cdots+\omega_n=m}} P\big(\{(\omega_1,\dots,\omega_n)\}\big)$$

$$= \binom{n}{m} \frac{\prod_{i=0}^{m-1}(s+ci)\cdot\prod_{j=0}^{n-m-1}(w+cj)}{\prod_{k=0}^{n-1}(s+w+ck)} = \frac{\frac{s}{c}(\frac{s}{c}+1)\cdots(\frac{s}{c}+m-1)}{m!}\cdot\frac{\frac{w}{c}(\frac{w}{c}+1)\cdots(\frac{w}{c}+n-m-1)}{(n-m)!}}{\frac{\frac{s+w}{c}(\frac{s+w}{c}+1)\cdots(\frac{s+w}{c}+n-1)}{n!}}$$

$$= \frac{\binom{-s/c}{m}\cdot\binom{-w/c}{n-m}}{\binom{-(s+w)/c}{n}}$$

(mit $\binom{r}{m} := \dfrac{r(r-1)\cdots(r-m+1)}{m!}$ für $r \in \mathbb{R}$, $m \in \mathbb{N}$), beachte, dass der Faktor $(-1)^n$ sich oben herauskürzt.

Diese Verteilung heißt auch *Pólya-Verteilung*.

(Im Fall $c = 0$ ergibt sich natürlich $P(S_n = m) = \mathrm{Bin}_{n,s/(s+w)}\big(\{m\}\big)$, vgl. Bsp. 1.20.)

In Satz 2.9 hatten wir ein W'maß auf einem Produktraum (mit $n \in \mathbb{N}$ Faktoren) konstruiert. Gelegentlich benötigt man ein analoges Resultat für den Fall (abzählbar) unendlich vieler Faktoren – beispielsweise, um in dem hier betrachteten Kontext eine unendliche Münzwurffolge zu modellieren.

Bericht 2.10 (Konstruktion von W'maßen auf unendlichen Produkträumen)**.** Seien Ω_i, $i \in \mathbb{N}$ jeweils (höchstens) abzählbar, $\neq \varnothing$, $p_1(\cdot)$ W'gewichtsfunktion auf Ω_1, für $k \in \{2,3,\dots\}$ und $\omega_1 \in \Omega_1, \dots, \omega_{k-1} \in \Omega_{k-1}$ sei $p_{k|\omega_1,\dots,\omega_{k-1}}(\cdot)$ W'gewichtsfunktion auf Ω_k, setze

$$\Omega := \bigtimes_{i\in\mathbb{N}} \Omega_i = \big\{(\omega_1,\omega_2,\omega_3,\dots) : \omega_i \in \Omega_i\big\}$$

(versehen mit

$$\mathcal{F} = \sigma\Big(\big\{\{\omega_1\}\times\{\omega_2\}\times\cdots\times\{\omega_n\}\times\Omega_{n+1}\times\Omega_{n+2}\times\cdots : n\in\mathbb{N}, \omega_i\in\Omega_i \text{ für } 1\le i\le n\big\}\Big),$$

der sogenannten Produkt-σ-Algebra), für $i \in \mathbb{N}$ sei

$$X_i : \Omega \to \Omega_i,\ X_i\big((\omega_n)_{n\in\mathbb{N}}\big) = \omega_i \qquad \text{(die Projektion auf die } i\text{-te Koordinate).}$$

Dann gibt es (genau) ein W'maß P auf $(\Omega,\mathcal{F})$ mit

$$P\big(X_1 = \omega_1,\dots,X_k = \omega_k\big) = p_1(\omega_1)p_{2|\omega_1}(\omega_2)\cdots p_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k) \tag{2.1}$$

für alle $k \in \mathbb{N}$ und $\omega_i \in \Omega_i$ für $1 \le i \le k$.

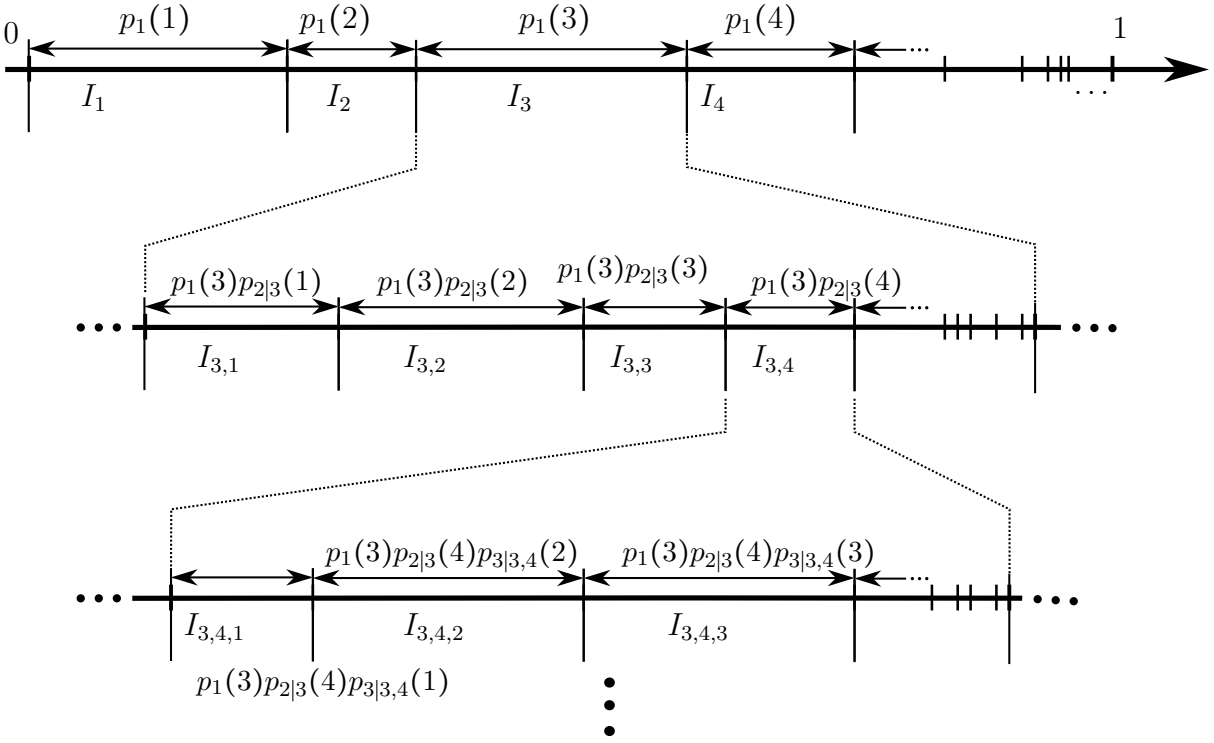
Dieses Resultat, das Satz 2.9 auf die Situation abzählbar unendlich vieler Ω_i, $i \in \mathbb{N}$ erweitert, werden wir in dieser Vorlesung nicht mit mathematischer Strenge beweisen (dafür siehe z.B. Georgii [G, Satz 3.12]).

Grobe Beweisidee. Gedankenexperiment: Stellen wir uns vor, wir haben einen Computer, der genau eine uniform in $[0,1]$ verteilte Zufallsvariable U (mit beliebiger Präzision) simulieren kann. Wir möchten aus dem Wert von U die Werte von $X_1, X_2, \dots$ derart ablesen, dass (2.1) für alle k gilt.

Sei $U \sim \mathrm{unif}_{[0,1]}$.

- Zerlege $I = [0,1)$ in disjunkte, halboffene Intervalle $I_{\omega_1}, \omega_1 \in \Omega_1$ der Längen $p_1(\omega_1)$ (1. Stufe),
- zerlege jedes I_{ω_1} in disjunkte, halboffene Intervalle $I_{\omega_1,\omega_2}, \omega_2 \in \Omega_2$ der Längen $p_1(\omega_1)p_{2|\omega_1}(\omega_2)$ (2. Stufe),

  $\dots$
- wenn $I_{\omega_1,\omega_2,\dots,\omega_{k-1}}$ konstruiert (mit Länge $p_1(\omega_1)p_{2|\omega_1}(\omega_2)\cdots p_{k-1|\omega_1,\dots,k-2}(\omega_{k-1})$), zerlege in disjunkte, halboffene Intervalle $I_{\omega_1,\omega_2,\dots,\omega_{k-1},\omega_k}, \omega_k \in \Omega_k$ der Längen $p_1(\omega_1)p_{2|\omega_1}(\omega_2)\cdots p_{k-1|\omega_1,\dots,k-2}(\omega_{k-1})p_{k|\omega_1,\dots,k-1}(\omega_k)$ (k. Stufe), etc.



Für $u \in [0,1)$ gibt es genau ein

$$X(u) = \big(X_1(u), X_2(u), \dots\big) \in \Omega,$$

so dass gilt (siehe auch die Skizze für den Fall $\Omega_1 = \Omega_2 = \dots = \mathbb{N}$)

$$u \in I_{X_1(u),X_2(u),\dots,X_k(u)} \quad \text{für alle } k \in \mathbb{N}$$

und dies definiert $X : [0,1) \to \Omega$. Damit ist

$$\begin{aligned} P\big(X_1 = \omega_1, \dots, X_k = \omega_k\big) &= \big(\mathrm{unif}_{[0,1]} \circ X^{-1}\big)\big(\{\omega_1\} \times \{\omega_2\} \times \dots \times \{\omega_k\} \times \Omega_{k+1} \times \Omega_{k+2} \times \dots\big) \\ &= \mathrm{unif}_{[0,1]}\big(\{u \in [0,1) : X_1(u) = \omega_1, X_2(u) = \omega_2, \dots, X_k(u) = \omega_k\}\big) \\ &= \mathrm{unif}_{[0,1]}\big(I_{\omega_1,\omega_2,\dots,\omega_k}\big) = \text{Länge von } I_{\omega_1,\omega_2,\dots,\omega_k} \\ &= p_1(\omega_1)p_{2|\omega_1}(\omega_2)\cdots p_{k-1|\omega_1,\dots,k-2}(\omega_{k-1})p_{k|\omega_1,\dots,k-1}(\omega_k), \end{aligned}$$

d.h. (2.1) gilt.

2.3 Unabhängigkeit

Definition 2.11. Wir betrachten einen W'raum $(\Omega, \mathcal{F}, P)$. Ereignisse A und B heißen (stochastisch) *unabhängig*, wenn gilt

$$P(A \cap B) = P(A) \cdot P(B)$$

(strenggenommen: unabhängig bezüglich P). Im Fall $P(A) > 0$ sind also A und B unabhängig g.d.w. $P(B \mid A) = P(B)$.

Beispiel. 1. Ziehen mit Zurücklegen aus einer Urne, $A = \{\text{erste gezogene Kugel ist schwarz}\}$, $B = \{\text{zweite gezogene Kugel ist schwarz}\}$, so sind A und B unabhängig.

(Dies gilt natürlich nicht, wenn ohne Zurücklegen gezogen wird.)

2. Ziehe zwei Karten ohne Zurücklegen aus einem (perfekt gemischten) Skatblatt, sei

$$A = \{\text{erste gezogene Karte ist ein As}\},$$
$$B = \{\text{zweite gezogene Karte hat Farbe Pik}\}.$$

Es ist

$$P(A) = \frac{4}{32} = \frac{1}{8}, \quad P(B) = \frac{8}{32} = \frac{1}{4} \left(= \frac{8 \cdot 7}{32 \cdot 31} + \frac{24 \cdot 8}{32 \cdot 31} \right),$$
$$P(A \cap B) = \frac{1 \cdot 7 + 3 \cdot 8}{32 \cdot 31} = \frac{31}{32 \cdot 31} = \frac{1}{32} = P(A) \cdot P(B),$$

d.h. A und B sind unabhängig.

Bemerkung. Gilt $P(A) \in \{0, 1\}$, so ist A von sich selbst unabhängig.

Definition 2.12. Sei I (beliebige) Indexmenge, A_i, $i \in I$ Ereignisse. Die Familie $(A_i)_{i \in I}$ heißt *unabhängig*, wenn für jedes endliche $J \subset I$, $J \neq \varnothing$ gilt

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

Beispiel 2.13. (Aus paarweiser Unabhängigkeit folgt i.A. nicht Unabhängigkeit)**.** Werfe eine faire Münze zwei Mal, sei

$$A = \{\text{erster Wurf ist } K\}, \; B = \{\text{zweiter Wurf ist } K\},$$
$$C = \{\text{beide Würfe haben das gleiche Ergebnis}\}.$$

Es ist

$$P(A) = P(B) = P(C) = \frac{1}{2}, \quad P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{1}{4},$$

d.h. je zwei der betrachteten Ereignisse sind unabhängig (man sagt: A, B, C sind paarweise unabhängig), aber

$$P(A \cap B \cap C) = \frac{1}{4} \neq P(A) \cdot P(B) \cdot P(C) = \frac{1}{8},$$

d.h. A, B, C sind nicht unabhängig. (Dies ist auch intuitiv klar, denn $(A \cap B) \cup (A^c \cap B^c) = C$.)

Bemerkung 2.14. Sind $(A_i, i \in I)$ unabhängig, so auch $(A_i^c, i \in I)$ und für $J, J' \subset I$ endlich mit $J \cap J' = \varnothing$ gilt

$$P\Big(\bigcap_{j\in J} A_j \cap \bigcap_{j'\in J'} A_{j'}^c\Big) = \prod_{j\in J} P(A_j) \cdot \prod_{j'\in J'} (1 - P(A_{j'})). \tag{2.2}$$

Beweis. Betrachte zunächst den Fall $|J| = |J'| = 1$:

$$P\big(A_j \cap A_{j'}^c\big) = P(A_j) - P(A_j \cap A_{j'}) = P(A_j) - P(A_j)P(A_{j'}) = P(A_j)(1 - P(A_{j'})),$$

sei nun J mit $|J| \geq 1$ allgemein, $J' = \{j'\}$ mit $j' \notin J$, wie oben ist

$$\begin{aligned} P\Big(\bigcap_{j\in J} A_j \cap A_{j'}^c\Big) &= P\Big(\bigcap_{j\in J} A_j\Big) - P\Big(\bigcap_{k\in J\cup\{j'\}} A_k\Big) \\ &= \prod_{j\in J} P(A_j) - \prod_{k\in J\cup\{j'\}} P(A_k) = \Big(\prod_{j\in J} P(A_j)\Big)\big(1 - P(A_{j'})\big). \end{aligned}$$

Nehmen wir nun an, (2.2) gilt für alle J, J' mit $J \cap J' = \varnothing$ und $|J'| \leq m$.

Sei $J'' = J' \cup \{j''\}$ mit einem $j'' \notin J \cap J'$ und $|J'| = m$: Nach Induktionsvoraussetzung ist

$$\begin{aligned} P\Big(\bigcap_{j\in J} A_j \cap \bigcap_{j'\in J''} A_{j'}^c\Big) &= P\Big(\bigcap_{j\in J} A_j \cap \bigcap_{j'\in J'} A_{j'}^c\Big) - P\Big(\bigcap_{j\in J\cup\{j''\}} A_j \cap \bigcap_{j'\in J'} A_{j'}^c\Big) \\ &= \prod_{j\in J} P(A_j) \cdot \prod_{j'\in J'} (1 - P(A_{j'})) \;\; - \prod_{j\in J\cup\{j''\}} P(A_j) \cdot \prod_{j'\in J'} (1 - P(A_{j'})) \\ &= \prod_{j\in J} P(A_j) \cdot \prod_{j'\in J''} (1 - P(A_{j'})), \end{aligned}$$

d.h. (2.2) gilt auch für J und J'' und somit allgemein. □

Diskussion. 1. Stochastische Unabhängigkeit ist eine (gemeinsame) Eigenschaft von Ereignissen und deren Wahrscheinlichkeiten; (Un-)abhängigkeit ist nicht automatisch mit (Nicht-)Existenz eines kausalen Zusammenhangs gleichzusetzen.

Beispiel: Wir befragen eine zufällig an einem Samstagnachmittag auf dem Mainzer Gutenbergplatz ausgewählte Testperson. Die Ereignisse „hat Schuhgröße ≥ 41“ und „hat Führerschein“ sind nicht unabhängig (gegeben $\{\text{hat Schuhgröße} \geq 41\}$ handelt es sich vermutlich eher um einen Erwachsenen, daher ist die Chance, dass die Person auch einen Führerschein hat größer als der Anteil der Führerscheinbesitzer in der Gesamtbevölkerung, die auch viele Kinder umfasst). Trotzdem wäre die Behauptung, dass große Füße Führerscheine hervorbringen, natürlich unsinnig.

2. Nichtsdestoweniger *modelliert* man die erneute Wiederholung eines gewissen zufälligen Experiments unter gleichen Bedingungen (oder auch die Befragung verschiedener Versuchspersonen aus einer großen Grundgesamtheit) zumeist mittels (angenommener) stochastischer Unabhängigkeit. Die Annahme unabhängiger Kopien eines gewissen Zufallsexperiments bildet häufig einen zentralen Ansatzpunkt statistischer Analysen.

Definition 2.15. I (nicht-leere) Indexmenge, X_i, $i \in I$ Zufallsvariablen (auf demselben W'raum), X_i habe Wertebereich $(S_i, \mathscr{A}_i)$. Die Familie $(X_i)_{i\in I}$ heißt *unabhängig*, wenn für jedes endliche $\varnothing \neq J \subset I$ und beliebige $B_j \in \mathscr{A}_j$ gilt

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

Also: Eine Familie von ZVn ist u.a. g.d.w. jede endliche Teilfamilie u.a. ist.

Beobachtung 2.16. 1. $(X_i)_{i\in I}$ unabhängig und $\varnothing \neq I' \subset I$, so ist auch $(X_i)_{i\in I'}$ eine unabhängige Familie. (Dies folgt sofort aus der Definition: wähle $B_j = S_j$ für $j \in I \setminus I'$.)

2. Seien $(S_i', \mathscr{A}_i')$ messbare Räume, $f_i : S_i \to S_i'$ messbar, X_i, $i \in I$ unabhängige ZVn (X_i hat Werte in S_i). Dann sind auch

$$Y_i := f_i(X_i),\ i \in I \ \text{ unabhängig,}$$

denn für endliches $J \subset I$, $B_j' \in \mathscr{A}_j'$ ist

$$\begin{aligned} P\Big(\bigcap_{j\in J}\{Y_j \in B_j'\}\Big) &= P\Big(\bigcap_{j\in J}\{X_j \in f_j^{-1}(B_j')\}\Big) \\ &= \prod_{j\in J} P\big(X_j \in f_j^{-1}(B_j')\big) = \prod_{j\in J} P\big(Y_j \in B_j'\big). \end{aligned}$$

3. Eine Familie von Ereignissen A_i, $i \in I$ ist unabhängig g.d.w. die Familie $\mathbf{1}_{A_i}$, $i \in I$ der zugehörigen Indikatorvariablen u.a. ist. (Dies folgt sofort aus der Definition zusammen mit Bem. 2.14.)

Bericht 2.17. In der Situation von Def. 2.15 sei $\mathcal{E}_i$ ein $\cap$-stabiler Erzeuger von $\mathscr{A}_i$ und es gelte für alle $J \subset I$ mit $|J| < \infty$ und $B_j \in \mathcal{E}_j$

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

Dann sind die $(X_i)_{i\in I}$ unabhängig.

Insbesondere gilt im reellwertigen Fall ($S_i = \mathbb{R}$): $(X_i)_{i\in I}$ sind unabhängig g.d.w.

$$\text{für alle } J \subset I,\ |J| < \infty,\ x_j \in \mathbb{R} \text{ gilt} \quad P\Big(\bigcap_{j\in J}\{X_j \le x_j\}\Big) = \prod_{j\in J} P(X_j \in x_j).$$

Beweisidee. Zeige induktiv über $\#\{j \in J : B_j \in \mathscr{A}_j \setminus \mathcal{E}_j\}$, dass die Bed. aus Def. 2.15 erfüllt ist, verwende dazu Eindeutigkeitssatz für Maße (Bericht 1.10). Für Details siehe z.B. [G, Satz 3.19].

Satz 2.18. *$X_1, X_2, \dots, X_n$ ZVn, X_i habe Werte in S_i, S_i abzählbar für $i = 1, 2, \dots, n$. Dann sind $X_1, X_2, \dots, X_n$ unabhängig g.d.w. gilt*

$$\forall\, x_1 \in S_1, x_2 \in S_2, \dots, x_n \in S_n :\ P\big(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\big) = \prod_{i=1}^{n} P(X_i = x_i).$$

Beweis. „⇒“: Klar (Wähle $A_i = \{x_i\}$ in Def. 2.15.)

„⇐“: Seien $A_1 \subset S_1, \dots A_n \subset S_n$, es ist

$$\begin{aligned} P\big(\{X_1 \in A_1\} \cap \dots \cap \{X_n \in A_n\}\big) &= P\Big(\bigcup_{(x_1,\dots,x_n)\in A_1\times\dots\times A_n} \{X_1 = x_1, \dots, X_n = x_n\}\Big) \\ &= \sum_{x_1\in A_1,\dots,x_n\in A_n} \underbrace{P\big(X_1 = x_1, \dots, X_n = x_n\big)}_{=P(X_1=x_1)\cdots P(X_n=x_n)} \\ &= \sum_{x_1\in A_1} P(X_1 = x_1) \cdots \sum_{x_n\in A_n} P(X_n = x_n) \\ &= P(X_1 \in A_1) \cdots P(X_n \in A_n). \end{aligned}$$

□

Bericht 2.19 (Unabhängigkeit im reellwertigen Fall mit Dichte)**.** Seien $X_1, X_2, \dots, X_n$ reellwertige ZVn, $f_1, \dots, f_n : \mathbb{R} \to \mathbb{R}_+$ Wahrscheinlichkeitsdichten (d.h. $\int_{\mathbb{R}} f_i(x)\,dx = 1$), dann sind äquivalent:

1. $X_1, \dots, X_n$ sind u.a. und X_i hat Dichte f_i für $i = 1, \dots, n$
   (d.h. $P(X_i \in B) = \int_B f_i(x)\,dx$).
2. Die ZV $X = (X_1, \dots, X_n)$ mit Werten in $\mathbb{R}^n$ hat Dichte
   $$f(x) = f_1(x_1) \cdot f_2(x_2) \cdot \dots \cdot f_n(x_n), \quad x = (x_1, \dots, x_n) \in \mathbb{R}^n,$$
   d.h. die gemeinsame Dichte hat Produktgestalt.

Beweisidee. „Naiv" rechnen wir

$$\int_{-\infty}^{x_1} f_1(y_1)\,dy_1 \cdots \int_{-\infty}^{x_n} f_n(y_n)\,dy_n = \int_{(-\infty,x_1]\times\dots\times(-\infty,x_n]} f_1(x_1)\cdots f_n(x_n)\,dy_1 \dots dy_n$$

(verwende den Satz von Fubini um die Integralvertauschungen zu rechtfertigen und dann Bericht 2.17. Siehe z.B. [KW, Satz auf S. 70 in Kap. III.10] und [G, Bsp. 3.30] für Details.

Beispiel 2.20. 1. $X_1, \dots, X_n$ u.a., $X_i \sim \mathcal{N}_{0,1}$, dann hat $X := (X_1, \dots, X_n)$ Dichte

$$f_X(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2}\|x\|^2\big),\ x = (x_1, \dots, x_n) \in \mathbb{R}^n$$

(mit $\|x\| = \sqrt{x_1^2 + \dots + x_n^2}$, der euklidischen Norm).

Sei $M = (m_{ij})_{i,j=1}^n$ orthogonale $n \times n$-Matrix (d.h. $M^T M = I$, die $n \times n$-Identitätsmatrix),

$$Y^T := MX^T \quad \text{d.h. } Y = (Y_1, \dots, Y_n) \text{ mit } Y_i = \sum_{j=1}^n m_{ij} X_j,$$

dann sind $Y_1, \dots, Y_n$ u.a., $Y_i \sim \mathcal{N}_{0,1}$.

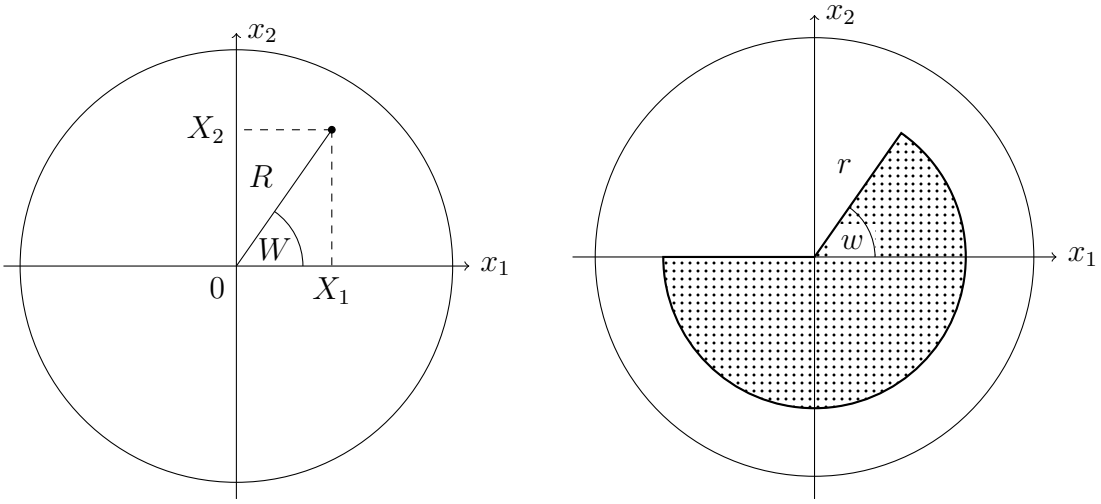
$\varphi : \mathbb{R}^n \to \mathbb{R}^n$, $\varphi(x) = Mx$ ist bijektiv und differenzierbar mit $\varphi^{-1}(y) = M^T y$, $\varphi' = M$, die Dichtetransformationsformel (Bericht 1.42) zeigt: Y hat Dichte

$$\begin{aligned} f_Y(y) &= \frac{f_X(\varphi^{-1}(y))}{|\det \varphi'(\varphi^{-1}(y))|} = \frac{f_X(M^T y)}{|\det M|} = f_X(M^T y) \\ &= (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2} \underbrace{\|M^T y\|^2}_{\|y\|^2}\big) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2}. \end{aligned}$$

2. Sei X ein uniform im Einheitskreis $\{x \in \mathbb{R}^2 : \|x\| \le 1\}$ verteilter Punkt (in kartesischen Koordinaten $X = (X_1, X_2)$), R der Radius, W der Winkel von X (in Polarkoordinaten), also

$$R = \sqrt{X_1^2 + X_2^2}, \quad W = \begin{cases} \arcsin\left(\frac{X_2}{R}\right), & X_1 \ge 0, \\ \pi - \arcsin\left(\frac{X_2}{R}\right), & X_1 < 0, X_2 \ge 0, \\ -\pi - \arcsin\left(\frac{X_2}{R}\right), & X_1 < 0, X_2 < 0, \end{cases}$$

(siehe auch die Skizze unten).



Links: Punkt (X_1, X_2) und seine Polarkoordinaten (R, W). Rechts: Schraffiert ist $B(r, w) := \{\text{Punkte mit Radius} \le r \text{ und Winkel} \le w\}$.

Dann sind R und W unabhängig,

$$R \text{ hat Dichte } f_R(r) = 2r\mathbf{1}_{[0,1]}(r),$$
$$W \text{ hat Dichte } f_W(w) = \frac{1}{2\pi}\mathbf{1}_{[-\pi,\pi)}(w),$$

denn (für $0 \le r \le 1$, $-\pi \le w < \pi$)

$$P(R \le r, W \le w) = P\big(X \in B(r,w)\big)$$
$$= \frac{\pi r^2 \frac{w+\pi}{2\pi}}{\pi 1^2} = r^2\frac{w+\pi}{2\pi} = \int_0^r 2s\,ds \cdot \int_{-\pi}^{w} \frac{1}{2\pi}\,dv.$$

Definition 2.21. Seien $(\Omega_i, \mathcal{F}_i)$, messbare Räume, P_i ein Wahrscheinlichkeitsmaß auf Ω_i, $i = 1, \dots, n$,

$$\Omega := \Omega_1 \times \cdots \times \Omega_n$$

(versehen mit $\mathcal{F} = \sigma\big(A_1 \times A_2 \times \cdots \times A_n, A_i \in \mathcal{F}_i \text{ für } i = 1, \dots, n\big)$, der „Produkt-$\sigma$-Algebra), dann heißt das Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{F})$ mit

$$P\big(A_1 \times A_2 \times \cdots \times A_n\big) = \prod_{i=1}^{n} P_i(A_i) \quad \text{für } A_1 \in \mathcal{F}_1, \dots, A_n \in \mathcal{F}_n$$

das Produkt (oder Produktmaß) der $P_1, \dots, P_n$, man schreibt $P = P_1 \otimes P_2 \otimes \cdots \otimes P_n$ (im Fall $P_1 = P_2 = \cdots = P_n$ auch $P = P_1^{\otimes n}$).

Bemerkung 2.22. ZVn $X_1, \dots, X_n$ (mit Wertebereichen $S_1, \dots, S_n$), die auf demselben W'raum $(\Omega, \mathcal{F}, P)$ definiert sind unabhängig, g.d.w. gilt (mit $X = (X_1 \dots, X_n)$)

$$\mathscr{L}_P(X) = \mathscr{L}_P(X_1) \otimes \mathscr{L}_P(X_2) \otimes \cdots \otimes \mathscr{L}_P(X_n).$$

(D.h. ZVn sind unabhängig g.d.w. die gemeinsame Verteilung (vgl. Beob. und Def. 1.33) ein Produktmaß ist).

Insbesondere: Unabhängigkeit von ZVn ist eine Eigenschaft der (gemeinsamen) Verteilung.

(Dies ergibt sich direkt aus dem Vergleich von Def. 2.15 und Def. 2.21.)

Beobachtung und Bericht 2.23 (Existenz von u.a. ZVn mit vorgegebenen Verteilungen)**.** $(S_i, \mathscr{A}_i)$ *messbare Räume,* μ_i *W'maß auf* $(S_i, \mathscr{A}_i)$ *für* $i = 1, \dots, n$ *(für ein* $n \in \mathbb{N}$*) oder für* $i \in \mathbb{N}$*.*

Dann gibt es (auf einem geeigeten W'raum $(\Omega, \mathcal{F}, P)$*) unabhängige ZVn* $X_1, X_2, \dots$ *mit* $X_i \sim \mu_i$*. Man kann* $\Omega = \bigtimes_{i=1}^n S_i$*,* $P = \bigotimes_{i=1}^n \mu_i$ *(bzw.* $\Omega = \bigtimes_{i=1}^\infty S_i$*,* $P = \bigotimes_{i=1}^\infty \mu_i$*) und* $X_i = i$*-te Koordinatenprojektion wählen.*

Diskussion. Wenn alle S_i abzählbar sind, folgt dies im Fall endlich vieler X_i aus Satz 2.18, im Fall unendlich vieler X_i aus Bericht 2.10.

Im Fall endlich vieler X_i mit Werten in $\mathbb{R}$, die jeweils eine Dichtefunktion f_i besitzen, folgt die Behauptung aus der Existenz des n-dim (Lebesgue-)Maßes („Volumen-Maß"), der Fall endlich oder unendlich vieler X_i mit Werten in $\mathbb{R}$ kann zudem (via Beob. 1.38 und Bericht 2.10) auf den diskreten Fall zurückgespielt werden:

Seien $U_{i,j}$, $i, j \in \mathbb{N}$ u.a., $\mathrm{Ber}_{1/2}$, dann sind $V_i := \sum_{j=1}^\infty 2^{-j} U_{i,j}$ u.a. (vgl. Beob. 2.16) und $V_i \sim \mathrm{Unif}_{[0,1]}$ für $i \in \mathbb{N}$.

Sei F_i die Verteilungsfunktion von μ_i, dann leistet $X_i := F_i^{-1}(V_i)$ das Gewünschte (vgl. Beob. 1.38).

2.4 Faltung

Definition 2.24. X und Y unabhängige reellwertige ZVn, $X \sim \mu$, $Y \sim \nu$ (definiert auf einem W'raum). Die Verteilung von $X + Y$ heißt die *Faltung* von μ und ν, geschrieben $\mu * \nu$:

$$(\mu * \nu)(B) = P(X + Y \in B), \quad B \in \mathcal{B}(\mathbb{R})$$

(alternativ: $\mu * \nu = (\mu \otimes \nu) \circ f^{-1}$ mit $f(x, y) = x + y$).

Bemerkung. $\mu * \nu = \nu * \mu$ (denn $X + Y = Y + X$).

Beobachtung 2.25 (Diskreter Fall)**.** Falls $\mu(\mathbb{Z}) = \nu(\mathbb{Z}) = 1$ (d.h. X und Y haben Werte in $\mathbb{Z}$), so ist

$$(\mu * \nu)(\{k\}) = P(X + Y = k) = \sum_{m \in \mathbb{Z}} P(X = m, Y = k - m) = \sum_{m \in \mathbb{Z}} \mu(\{m\}) \nu(\{k - m\}).$$

Beispiel 2.26.

1. X, Y u.a., $\sim \mathrm{Ber}_p$, so ist $X + Y \sim \mathrm{Bin}_{2,p}$, d.h. $\mathrm{Ber}_p * \mathrm{Ber}_p = \mathrm{Bin}_{2,p}$.
2. (Binomialfamilie) $X_1, X_2, \dots, X_n$ u.a., $\sim \mathrm{Ber}_p$, so ist $X_1 + X_2 + \cdots + X_n \sim \mathrm{Bin}_{n,p}$, d.h.

$$\mathrm{Ber}_p^{*n} = \underbrace{\mathrm{Ber}_p * \mathrm{Ber}_p * \cdots * \mathrm{Ber}_p}_{n\text{-mal}} = \mathrm{Bin}_{n,p}.$$

Insbesondere gilt

$$\mathrm{Bin}_{n_1,p} * \mathrm{Bin}_{n_2,p} = \mathrm{Bin}_{n_1+n_2,p} \quad \text{für } p \in [0,1], n_1, n_2 \in \mathbb{N},$$

die Binomialverteilungen bilden (für festes p) eine *Faltungsfamilie*.

3. (Poissonfamilie) Für $\alpha, \beta > 0$ ist $\mathrm{Poi}_\alpha * \mathrm{Poi}_\beta = \mathrm{Poi}_{\alpha+\beta}$, denn

$$\begin{aligned}\sum_{m=0}^{k} e^{-\alpha}\frac{\alpha^m}{m!}\cdot e^{-\beta}\frac{\beta^{k-m}}{(k-m)!} &= e^{-(\alpha+\beta)}\frac{1}{k!}\sum_{m=0}^{k}\binom{k}{m}\alpha^m\beta^{k-m}\\ &= e^{-(\alpha+\beta)}\frac{(\alpha+\beta)^k}{k!} = \mathrm{Poi}_{\alpha+\beta}(\{k\}), \quad k\in\mathbb{N}_0.\end{aligned}$$

Auch die Poissonverteilungen bilden eine Faltungsfamilie.

Beobachtung 2.27 (Faltung von Dichten)**.** X, Y u.a. reellwertige ZVn mit Dichte f_X bzw. f_Y, so hat $X + Y$ die Dichte

$$(f_X * f_Y)(z) := \int_{\mathbb{R}} f_X(x) f_Y(z-x)\,dx, \quad z\in\mathbb{R}.$$

Es ist nämlich

$$\begin{aligned}P(X+Y\le w) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathbf{1}_{\{x+y\le w\}} f_X(x) f_Y(y)\,dydx\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathbf{1}_{\{x+z-x\le w\}} f_X(x) f_Y(z-x)\,dzdx\\ &= \int_{-\infty}^{\infty}\mathbf{1}_{\{z\le w\}}\int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx\,dz = \int_{-\infty}^{w}(f_X * f_Y)(z)\,dz\end{aligned}$$

wobei wir in der 2. Zeile $y = z - x$ substituiert haben.

Beispiel 2.28 (Die Normalverteilungen bilden eine Faltungsfamilie)**.** Es gilt

$$\mathcal{N}_{\mu_1,\sigma_1^2} * \mathcal{N}_{\mu_2,\sigma_2^2} = \mathcal{N}_{\mu_1+\mu_2,\sigma_1^2+\sigma_2^2} \quad \text{für } \mu_1,\mu_2\in\mathbb{R},\ \sigma_1,\sigma_2>0$$

Beweis. Betrachte o.E. den Fall $\mu_1 = \mu_2 = 0$ (denn $Z\sim\mathcal{N}_{\mu,\sigma^2}$, $a\in\mathbb{R}$, so ist $Z + a\sim\mathcal{N}_{\mu+a,\sigma^2}$):

Für $z\in\mathbb{R}$ ist

$$\begin{aligned}&\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\Big(-\frac{x^2}{2\sigma_1^2}\Big)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\Big(-\frac{(z-x)^2}{2\sigma_2^2}\Big)\,dx\\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\Big(-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}\Big)\int_{\mathbb{R}}\frac{1}{\big(2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\big)^{1/2}}\exp\left(\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}-\frac{x^2}{2\sigma_1^2}-\frac{z^2}{2\sigma_2^2}+\frac{zx}{\sigma_2^2}-\frac{x^2}{2\sigma_2^2}\right)dx\\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\Big(-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}\Big)\int_{\mathbb{R}}\frac{1}{\big(2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\big)^{1/2}}\exp\left(-\frac{\big(x-\frac{z}{1+(\sigma_2/\sigma_1)^2}\big)^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}\right)dx\\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\Big(-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}\Big).\end{aligned}$$

(Nebenrechnung: Das Argument der Exponentialfunktion innerhalb des Integrals in der 2. Zeile ist

$$\begin{aligned}
&\frac{z^2}{2(\sigma_1^2+\sigma_2^2)} - \frac{x^2}{2\sigma_1^2} - \frac{z^2}{2\sigma_2^2} + \frac{zx}{\sigma_2^2} - \frac{x^2}{2\sigma_2^2} \\
&= -\frac{1}{2}\big(\sigma_1^{-2}+\sigma_2^{-2}\big)\Big(x^2 - \frac{2xz}{\sigma_2^2(\sigma_1^{-2}+\sigma_2^{-2})} + \underbrace{\big(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2+\sigma_2^2}\big)\big(\sigma_1^{-2}+\sigma_2^{-2}\big)^{-1}}_{=\frac{\sigma_1^2}{\sigma_2^2(\sigma_1^2+\sigma_2^2)(\sigma_1^{-2}+\sigma_2^{-2})}=\frac{1}{\sigma_1^4}\frac{1}{(\sigma_1^{-2}+\sigma_2^{-2})^2}} z^2\Big) \\
&= -\frac{1}{2}\underbrace{\big(\sigma_1^{-2}+\sigma_2^{-2}\big)}_{=\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}}\big(x - \frac{z}{1+(\sigma_2/\sigma_1)^2}\big)^2,
\end{aligned}$$

das Integral in der 2. Zeile ist $\mathcal{N}_{z/(1+(\sigma_2/\sigma_1)^2),\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2)}(\mathbb{R}) = 1$.) □

Bemerkung. Man kann anstelle obiger expliziter Rechnung auch mit orthogonalen Transformationen und Beispiel 2.20, 1. (Invarianz der multi-dimensionalen Standard-Normalverteilung unter orthogonalen Transformationen) argumentieren, vgl. auch [KW, Bsp. auf S. 71]:

Seien $a, b \in (0,1)$ mit $a^2 + b^2 = 1$, so ist die 2×2-Matrix $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ orthogonal, seien Z_1, Z_2 u.a., $\sim \mathcal{N}_{0,1}$, dann haben

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} a & b \\ -b & a \end{pmatrix}\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} aZ_1 + bZ_2 \\ -bZ_1 + aZ_2 \end{pmatrix}$$

dieselbe Verteilung, d.h. auch aZ_1+bZ_2 und $-bZ_1+aZ_2$ sind u.i.v., $\sim \mathcal{N}_{0,1}$, insbesondere ist aZ_1+bZ_2 standard-normalverteilt.

Setzen wir $a := \frac{\sigma_1}{\sqrt{\sigma_1^2+\sigma_2^2}}$, $b := \frac{\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}}$, so finden wir: $X_1 := \sigma_1 Z_1 \sim \mathcal{N}_{0,\sigma_1^2}$, $X_2 := \sigma_2 Z_2 \sim \mathcal{N}_{0,\sigma_2^2}$ (und X_1, X_2 sind u.a.),

$$\frac{X_1}{\sqrt{\sigma_1^2+\sigma_2^2}} + \frac{X_2}{\sqrt{\sigma_1^2+\sigma_2^2}} = aZ_1 + bZ_2 \sim \mathcal{N}_{0,1},$$

also gilt $X_1 + X_2 \sim \mathcal{N}_{0,\sigma_1^2+\sigma_2^2}$.

2.5 Asymptotische Ereignisse

In diesem Abschnitt geht es um Ereignisse, die gewissermaßen (wenn man bei der Nummerierung an einen Zeitablauf denkt) „unendlich spät“ entschieden werden.

Definition 2.29. Seien $A_1, A_2, \ldots$ Ereignisse,

$$\limsup_{n\to\infty} A_n := \bigcap_{n\geq 1}\bigcup_{m\geq n} A_m$$

(„unendlich viele der A_n treten ein“),

$$\liminf_{n\to\infty} A_n := \bigcup_{n\geq 1} \bigcap_{m\geq n} A_m$$

(„von einem (möglicherweise zufälligen) Index ab treten alle A_n ein“).

Beachte: Es ist $\Big(\liminf_{n\to\infty} A_n\Big)^c = \limsup_{n\to\infty} A_n^c$, mit Indikatorvariablen ausgedrückt gilt

$$\limsup_{n\to\infty} \mathbf{1}_{A_n}(\omega) = \mathbf{1}_{\limsup_n A_n}(\omega), \quad \liminf_{n\to\infty} \mathbf{1}_{A_n}(\omega) = \mathbf{1}_{\liminf_n A_n}(\omega).$$

Satz 2.30 (Lemma von Borel-Cantelli[3])**.** *Seien $A_1, A_2, \dots$ Ereignisse, $A = \limsup_{n\to\infty} A_n$.*

1. $\sum_{n=1}^{\infty} P(A_n) < \infty \quad \Longrightarrow \quad P(A) = 0$

2. $\sum_{n=1}^{\infty} P(A_n) = \infty$ *und die $A_1, A_2, \dots$ seien unabhängig* $\quad \Longrightarrow \quad P(A) = 1$

Beweis. 1. Für jedes n ist

$$0 \leq P(A) \leq P\Big(\bigcup_{m\geq n} A_m\Big) \leq \sum_{m=n}^{\infty} P(A_m) \xrightarrow[n\to\infty]{} 0,$$

also gilt $P(A) = 0$.

2. Es ist

$$P(A^c) = P\Big(\bigcup_{n\geq 1}\bigcap_{m\geq n} A_m^c\Big) \leq \sum_{n=1}^{\infty} P\Big(\bigcap_{m\geq n} A_m^c\Big) = \sum_{n=1}^{\infty} \lim_{k\to\infty} \underbrace{P\Big(\bigcap_{m=n}^{k} A_m^c\Big)}_{=\prod_{m=n}^{k}(1-P(A_m)) \leq \prod_{m=n}^{k} e^{-P(A_m)}}$$

$$\leq \sum_{n=1}^{\infty} \underbrace{\lim_{k\to\infty} \exp\Big(-\sum_{m=n}^{k} P(A_m)\Big)}_{=0} = 0.$$

□

Beispiel 2.31. 1. Betrachte unendlich iterierten (unabhängigen) p-Münzwurf ($0 < p < 1$), sei

$$A_n = \{\text{Erfolg im } n\text{-ten Wurf}\},$$

dann ist $P(\limsup_n A_n) = 1$, d.h. es treten fast sicher ∞ viele Erfolge auf.

2. Seien $X_1, X_2, \dots$ ZVn, $X_n \sim \mathrm{Poi}_{\lambda_n}$ mit $\sup_n \lambda_n =: \Lambda < \infty$, $A_n = \{X_n \geq n\}$. Es ist

$$P(A_n) = e^{-\lambda_n} \sum_{k=n}^{\infty} \frac{(\lambda_n)^k}{k!} \leq e^{-\Lambda} \sum_{k=n}^{\infty} \frac{\Lambda^k}{k!}$$

(verwende z.B. Bsp 2.26, 3. zur Poisson-Faltungseigenschaft, um dies einzusehen: Falls $\lambda_n < \Lambda$, sei Y_n u.a. von X_n, $\sim \mathrm{Poi}_{\Lambda-\lambda_n}$, dann ist $X_n + Y_n \sim \mathrm{Poi}_{\Lambda}$ und somit $P(X_n \geq n) \leq P(X_n + Y_n \geq n)$. Somit

$$\sum_{n=1}^{\infty} P(A_n) \leq e^{-\Lambda} \sum_{k=1}^{\infty} \frac{\Lambda^k}{k!} \sum_{1\leq n\leq k} 1 = e^{-\Lambda} \sum_{k=1}^{\infty} \frac{\Lambda^k}{k!} k < \infty,$$

also $P(\limsup_n A_n) = 0$. (Beachte: die X_n brauchen nicht unabhängig zu sein.)

[3] nach Émile Borel (1871–1956) und Francesco Cantelli (1875–1966)

Kapitel 3

Erwartungswert, Varianz und Kovarianz

Der Erwartungswert ist eine wichtige Kenngröße der Verteilung einer reellwertigen Zufallsvariable X, er gibt eine Antwort auf die – etwas salopp formulierte – Frage „Wie groß ist X typischerweise?“

3.1 Diskreter Fall

Sei X reelle ZV mit abzählbarem Wertebereich (auf einem W'raum $(\Omega, \mathcal{F}, P)$ definiert), d.h. es gibt eine abzählbare Menge $S = S_X \subset \mathbb{R}$ mit $P(X \in S) = 1$ und $\mathscr{L}_P(X)$ hat Gewichte $P(X = x)$, $x \in S$.

Definition 3.1. X besitzt einen Erwartungswert, wenn

$$\sum_{x \in S_X} |x| P(X = x) < \infty$$

gilt, man schreibt dies auch als $X \in \mathcal{L}^1$ (bzw. $X \in \mathcal{L}^1(P)$, wenn das zugrundeliegende W'maß P nicht aus dem Kontext klar ist).

In diesem Fall heißt

$$\mathbb{E}[X] := \sum_{x \in S_X} x P(X = x)$$

der Erwartungswert von X.

Bemerkung. Die Summe $\sum_{x \in S_X} x P(X = x)$ ist dann wohldefiniert (unabhängig von der Summationsreihenfolge) und es gilt $|\mathbb{E}[X]| \leq \sum_{x \in S_X} |x| P(X = x) < \infty$.

Für ein „Gegenbeispiel“ sei $P(X = n) = P(X = -n) = \frac{1}{2n(n-1)}$ für $n = 2, 3, \ldots$ (es ist $\sum_{n=2}^{\infty} 2 \frac{1}{2n(n-1)} = \sum_{n=2}^{\infty} \left(\frac{1}{n-1} - \frac{1}{n} \right) = 1$, d.h. dies sind W'gewichte), wenn man die Werte durchnummerierte mit $x_{2i} = i + 1$, $x_{2i-1} = -i - 1$, $i \in \mathbb{N}$, so wäre

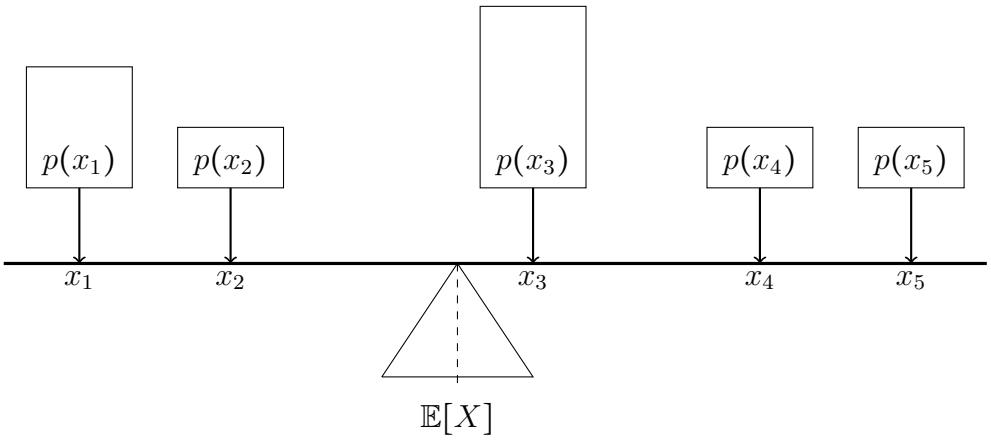
$$\sum_{j=1}^{\infty} x_j P(X = x_j) = \lim_{N \to \infty} \sum_{j=1}^{N} x_j P(X = x_j) = 0,$$

andererseits ist $\sum_{j=1}^{\infty} |x_j| P(X = x_j) = \sum_{n=2}^{\infty} \frac{n}{n(n-1)} = \sum_{k=1}^{\infty} \frac{1}{k} = \infty$.

Bemerkung 3.2. 1. Falls $|\Omega| < \infty$, so besitzt jede ZV eine Erwartungswert und es gilt

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} P(\{\omega\}) X(\omega).$$

2. Wenn X endlich viele mögliche Werte $x_1, \dots, x_n$ (mit Gewichten $p(x_i) = P(X = x_i)$) hat, so besitzt es einen Erwartungswert und man kann $\mathbb{E}[X]$ als den „Massenschwerpunkt“ interpretieren.



Auf einer Balkenwaage (deren Balken Eigengewicht 0 habe) liege an der Position x_i das Gewicht $p(x_i)$. Damit der Balken in Ruhelage ist, muss man ihn an der Stelle $\sum_{i=1}^{n} x_i p(x_i) = \mathbb{E}[X]$ unterstützen, denn dann ist das Gesamtdrehmonent (proportional zu)

$$\sum_{i=1}^{n} p(x_i)(x_i - \mathbb{E}[X]) = \mathbb{E}[X] - \mathbb{E}[X] = 0$$

3. Wenn $X \geq 0$ (bzw. $P(X \geq 0) = 1$), so ist $\sum_x x\, P(X = x)$ stets wohldefiniert (möglicherweise mit Wert $+\infty$, den man dann formal zulässt).

4. Der Erwartungswert von X muss nicht notwendigerweise ein möglicher Wert von X sein: $P(X = \mathbb{E}[X]) = 0$ ist durchaus möglich, beispielsweise gilt für W das Ergebnis eines fairen Würfelwurfs

$$\mathbb{E}[W] = \sum_{w=1}^{6} \frac{1}{6} w = \frac{7}{2} \notin \{1, 2, \dots, 6\}$$

Daher kann man die Interpretation von $\mathbb{E}[X]$ als „typischer Wert von X“ i.A. nicht wörtlich nehmen.

Es gilt aber: Sind $X_1, X_2, \dots$ unabhängig mit derselben Verteilung wie X, so konvergiert

$$M_n := \frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{} \mathbb{E}[X] = \sum_x x P(X = x)$$

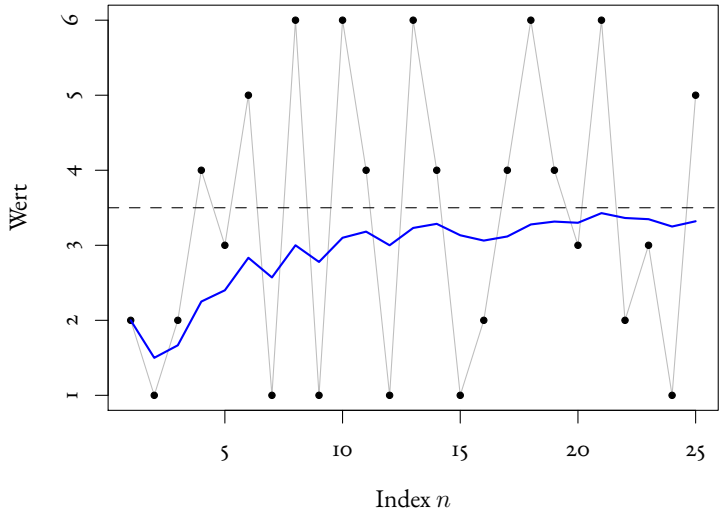
(in geeignetem Sinn), dies ist die Aussage des *Gesetzes der großen Zahlen*, das wir später sehen werden.

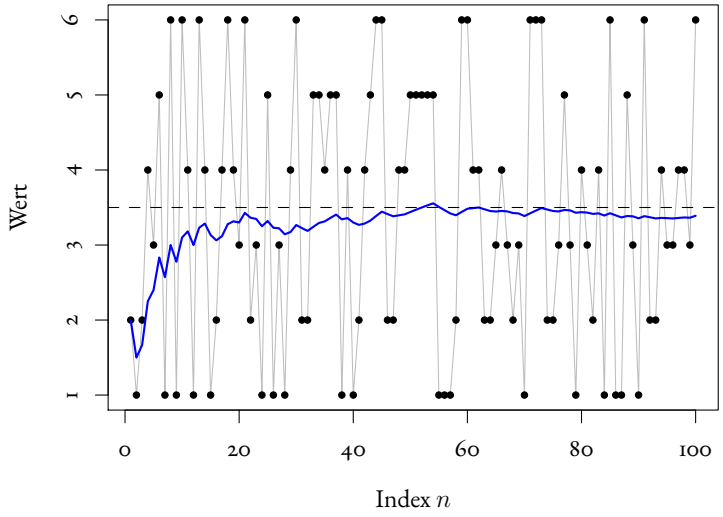
Es ist nämlich

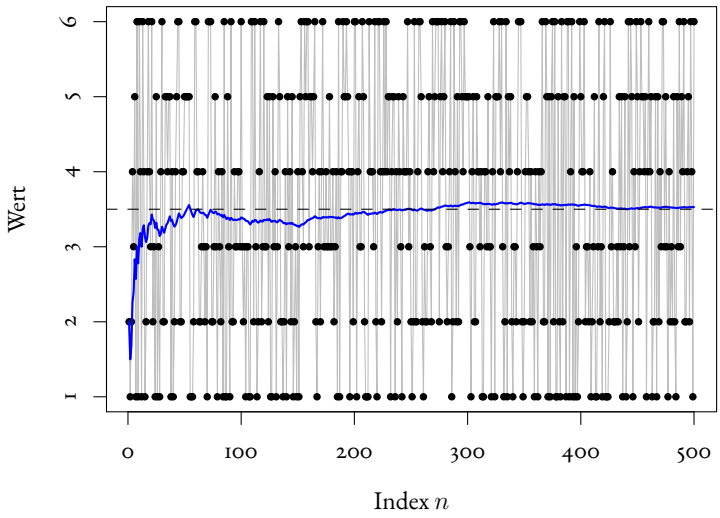
$$M_n = \sum_x x \cdot \frac{\#\{i \leq n : X_i = x\}}{n}$$

und $\#\{i \leq n : X_i = x\}/n \xrightarrow[n\to\infty]{} P(X = x)$.

Illustration: $X_1, X_2, \ldots$ uniform auf $\{1, 2, 3, 4, 5, 6\}$,
X_n sind jeweils die schwarzen Punkte, M_n die blaue Linie







5. Man kann $\mathbb{E}[X]$ als den erforderlichen Einsatz in einem „fairen Spiel“ interpretieren, bei dem man eine zufällige Auszahlung X erhält.

6. Der Erwartungswert ist eine Eigenschaft der Verteilung: $\mathscr{L}(X) = \mathscr{L}(Y)$ impliziert $\mathbb{E}[X] = \mathbb{E}[Y]$. (Klar, da dann $P(X = x) = P(Y = x)$ für alle x gilt.)

Beispiel 3.3. 1. A Ereignis, so ist $\mathbb{E}[\mathbf{1}_A] = 1 \cdot P(\mathbf{1}_A = 1) + 0 \cdot P(\mathbf{1}_A = 0) = P(A)$.

2. Sei $X \sim \mathrm{Bin}_{n,p}$, $n \in \mathbb{N}$, $p \in [0,1]$:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{n} kP(X = k) = \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-1-(k-1)} = np\mathrm{Bin}_{n-1,p}(\{0,1,\dots,n-1\}) = np\end{aligned}$$

3. Sei $X \sim \mathrm{Geom}_p$, $p \in (0,1]$:

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} np(1-p)^n = p(1-p)\sum_{n=1}^{\infty} n(1-p)^{n-1} = \frac{p(1-p)}{p^2} = \frac{1-p}{p}$$

(Wir verwenden, dass $f(t) := \sum_{n=0}^{\infty} t^n = \frac{1}{1-t}$ (für $|t| < 1$) erfüllt $\frac{d}{dt}f(t) = \frac{1}{(1-t)^2} = \sum_{n=1}^{\infty} nt^{n-1}$ (die Potenzreihe darf im Inneren des Konvergenzbereichs gliedweise abgeleitet werden) und setzen dann $t = 1 - p$ ein.)

4. Sei $X \sim \mathrm{Poi}_\alpha$, $\alpha > 0$:

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} ne^{-\alpha}\frac{\alpha^n}{n!} = \alpha\sum_{n=1}^{\infty} e^{-\alpha}\frac{\alpha^{n-1}}{(n-1)!} = \alpha$$

Satz 3.4 (Rechenregeln für Erwartungswerte)**.** *Seien $X, Y, X_1, X_2, \dots, Y_1, Y_2, \dots \in \mathcal{L}^1(P)$.*
1. (Linearität) Für $a, b \in \mathbb{R}$ gilt $aX + bY \in \mathcal{L}^1(P)$ und

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

2. (Monotonie) Wenn $X \geq Y$ (es genügt $P(X \geq Y) = 1$), so gilt $\mathbb{E}[X] \geq \mathbb{E}[Y]$; insbesondere gilt $\mathbb{E}[X] \geq 0$ für $X \geq 0$.

3. $P(X \geq 0) = 1$ und $\mathbb{E}[X] = 0 \Rightarrow P(X = 0) = 1$.

4. (Faktorisierung für unabhängige Produkte) Wenn X und Y unabhängig sind, so ist $XY \in \mathcal{L}^1(P)$ und

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y].$$

5. (Monotone Konvergenz) Sei $X_n \nearrow_{n\to\infty} X$, so gilt

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X];$$

insbesondere für $Y_n \geq 0$, $Y := \sum_{n=1}^{\infty} Y_n$ gilt $\mathbb{E}[Y] = \sum_{n=1}^{\infty} \mathbb{E}[Y_n]$ (möglicherweise als $\infty = \infty$).

Beweis. 1. Beachte, dass $aX + bY$ ebenfalls diskret ist, der Wertebereich $\{ax + by : x \in S_X, y \in S_Y\}$ ist abzählbar. Es ist

$$\sum_z |z| P(aX + bY = z) = \sum_{x,y} \underbrace{|ax + by|}_{\leq |a||x| + |b||y|} P(X = x, Y = y) \leq |a| \sum_x |x| P(X = x) + |b| \sum_y |y| P(Y = y) < \infty,$$

d.h. $aX + bY \in \mathcal{L}^1(P)$. Analog ist

$$\begin{aligned}\mathbb{E}[aX + bY] &= \sum_{x,y} (ax + by) P(X = x, Y = y) \\ &= a \sum_{x,y} x P(X = x, Y = y) + b \sum_{x,y} y P(X = x, Y = y) = a\mathbb{E}[X] + b\mathbb{E}[Y].\end{aligned}$$

2.

$$\begin{aligned}\mathbb{E}[X] = \sum_x x P(X = x) &= \sum_{x,y} x \underbrace{P(X = x, Y = y)}_{=0 \text{ falls } y > x} \\ &\geq \sum_{x,y} y P(X = x, Y = y) = \sum_y y P(Y = y) = \mathbb{E}[Y]\end{aligned}$$

3. $\mathbb{E}[X] = \sum_{x\,(\geq 0)} x P(X = x)$ wäre > 0, wenn $P(X = x) > 0$ für ein $x > 0$ gälte.

4. Beachte, dass XY wiederum diskret ist. Weiter ist

$$\sum_z |z| P(XY = z) = \sum_{x, y \neq 0} |xy| \underbrace{P(X = x, Y = y)}_{= P(X=x)P(Y=y)} = \sum_{x \neq 0} |x| P(X = x) \cdot \sum_{y \neq 0} |y| P(Y = y) = \mathbb{E}[X]\,\mathbb{E}[Y],$$

d.h. $XY \in \mathcal{L}^1(P)$. Analog folgt

$$\begin{aligned}\mathbb{E}[XY] = \sum_z z P(XY = z) &= \sum_{x, y \neq 0} xy P(X = x, Y = y) \\ &= \sum_{x \neq 0} x P(X = x) \cdot \sum_{y \neq 0} y P(Y = y) = \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

5. $Y \geq \sum_{n=1}^N Y_n$, also

$$\mathbb{E}[Y] \overset{2.}{\geq} \mathbb{E}\Big[\sum_{n=1}^N Y_n\Big] \overset{1.}{=} \sum_{n=1}^N \mathbb{E}[Y_n]$$

für jedes $N \in \mathbb{N}$, somit $\mathbb{E}[Y] \geq \sum_{n=1}^\infty \mathbb{E}[Y_n]$.

Sei $\varepsilon \in (0, 1)$,

$$\tau := \inf\Big\{N \in \mathbb{N} : \sum_{n=1}^N Y_n \geq (1 - \varepsilon) Y\Big\},$$

n. Vor. ist (auf $\{Y < \infty\}$) $P(\tau < \infty) = 1$. Setze $S := \sum_{n=1}^{\tau} Y_n$.

$$\begin{aligned}(1-\varepsilon)\mathbb{E}[Y] \le \mathbb{E}[S] &= \sum_{N,s} sP(S=s,\tau=N) = \sum_{N,s} sP\Big(\tau=N, \sum_{n=1}^{N} Y_n = s\Big)\\ &= \sum_{N} \mathbb{E}\Big[\mathbf{1}_{\{\tau=N\}} \sum_{n=1}^{N} Y_n\Big] = \sum_{N\in\mathbb{N}} \sum_{1\le n\le N} \mathbb{E}[Y_n \mathbf{1}_{\{\tau=N\}}]\\ &= \sum_{n\in\mathbb{N}} \sum_{N\ge n} \mathbb{E}[Y_n \mathbf{1}_{\{\tau=N\}}] = \sum_{n\in\mathbb{N}} \sum_{N\ge n} \sum_{y} yP(Y_n=y,\tau=N)\\ &= \sum_{n\in\mathbb{N}} \sum_{y} yP(Y_n=y,\tau\ge n) = \sum_{n\in\mathbb{N}} \mathbb{E}[Y_n\mathbf{1}_{\{\tau\ge n\}}] \le \sum_{n\in\mathbb{N}} \mathbb{E}[Y_n],\end{aligned}$$

mit $\varepsilon \downarrow 0$ folgt der 2. Teil der Beh.

Für den ersten Teil der Beh. schreibe $Y_n := X_n - X_{n-1}$ (mit $X_0 := 0$), dann verwende den 2. Teil. □

Beobachtung 3.5 (Erwartungswerte für Kompositionen)**.** X (diskrete) reelle ZV, $g : \mathbb{R} \to \mathbb{R}$, $Y := g(X)$.

Dann gilt $Y \in \mathcal{L}^1(P)$ g.d.w. $\sum_x |g(x)|P(X=x) < \infty$ und in diesem Fall ist

$$\mathbb{E}[Y] = \sum_x g(x)P(X=x).$$

(Schreibe $\sum_y yP(Y=y) = \sum_y \sum_{x:g(x)=y} g(x)P(X=x) = \sum_x g(x)P(X=x)$.)

Beispiel 3.6. 1. Seien $X_1,\dots,X_n$ u.i.v., $\sim \mathrm{Ber}_p$, so ist $X := X_1 + \cdots + X_n \sim \mathrm{Bin}_{n,p}$ und

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np.$$

(Wir hatten den Erwartungswert einer binomialverteilten ZV bereits in Bsp. 3.3, 2. bestimmt, hier kommen wir allerdings ohne explizite Rechnung aus).

2. Sei $X \sim \mathrm{Hyp}_{s,w,k}$ hypergeometrisch verteilt, vgl. Bsp. 1.19. Denken wir an eine Urne mit s schwarzen und w weißen Kugeln, aus der k mal ohne Zurücklegen gezogen wird, so ist

$$X \overset{d}{=} \mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_k} \quad \text{mit } A_i = \{i\text{-te gezogene Kugel ist schwarz}\}$$

und $P(A_1) = P(A_2) = \cdots = P(A_k) = \dfrac{s}{s+w}$, also $\mathbb{E}[X] = \dfrac{ks}{s+w}$.

Bericht 3.7 (Allgemeiner Fall: Erwartungswerte für nicht notwendig diskrete ZV)**.** Sei X eine reelle ZV, setze

$$X_{(n)} := \frac{1}{n}\lfloor nX \rfloor \quad \text{(Diskretisierung von } X \text{ auf das Gitter } \tfrac{1}{n}\mathbb{Z})$$

mit $\lfloor x \rfloor := \max\{z : z \in \mathbb{Z}, z \le x\}$, offenbar ist $X_{(n)}$ eine diskrete ZV (für jedes $n \in \mathbb{N}$).

Wenn $X_{(n)} \in \mathcal{L}^1$ (für ein, und dann automatisch für alle n) gilt, so existiert

$$\mathbb{E}[X] := \lim_{n\to\infty} \mathbb{E}[X_{(n)}]$$

und die Rechenregeln aus Satz 3.4 gelten ebenso.

Beweisidee. Wir betrachten hier nur eine Skizze, siehe z.B. [G, Kap. 4.1.2] (oder die Vorlesung Stochastik I) für das „volle Argument".

Es ist $X_{(n)} \le X < X_{(n)} + \frac{1}{n}$, also

$$X_{(m)} < X_{(n)} + \frac{1}{n} \text{ und } X_{(n)} < X_{(m)} + \frac{1}{m} \quad \text{für alle } m, n \in \mathbb{N}.$$

Demnach

$$|X_{(n)}| \le |X_{(m)}| + \big(\frac{1}{m} \vee \frac{1}{n}\big),$$

insbesondere gilt $X_{(n)} \in \mathcal{L}^1$ g.w.d. $X_{(m)} \in \mathcal{L}^1$ und

$$\mathbb{E}\big[X_{(n)}\big] \le \mathbb{E}\big[X_{(m)}\big] + \frac{1}{m}, \quad \mathbb{E}\big[X_{(m)}\big] \le \mathbb{E}\big[X_{(m)}\big] + \frac{1}{n},$$

also

$$\Big|\mathbb{E}\big[X_{(n)}\big] - \mathbb{E}\big[X_{(m)}\big]\Big| \le \big(\frac{1}{m} \vee \frac{1}{n}\big),$$

d.h. $\big(\mathbb{E}[X_{(n)}]\big)_{n\in\mathbb{N}}$ ist eine Cauchy-Folge, somit existiert ihr Grenzwert.

Um die Rechenregeln aus Satz 3.4 auf den allgemeinen Fall zu übertragen, muss man jeweils prüfen, dass sie mit der Grenzwertbildung in der diskreten Approximation verträglich sind. □

Satz 3.8 (Jensen'sche Ungleichung[1])**.** *Sei $\varphi : \mathbb{R} \to \mathbb{R}$ konvex, $X \in \mathcal{L}^1$ und $\varphi(X) \in \mathcal{L}^1$, dann gilt*

$$\varphi\big(\mathbb{E}[X]\big) \le \mathbb{E}\big[\varphi(X)\big],$$

insbesondere gilt für $X \in \mathcal{L}^1$

$$\big|\mathbb{E}[X]\big| \le \mathbb{E}\big[|X|\big]$$

und

$$\big(\mathbb{E}[X]\big)^2 \le \mathbb{E}\big[X^2\big]$$

(letzeres möglicherweise im Sinne von $(\mathbb{E}[X])^2 \le \infty$).

Beweis. Erinnerung:

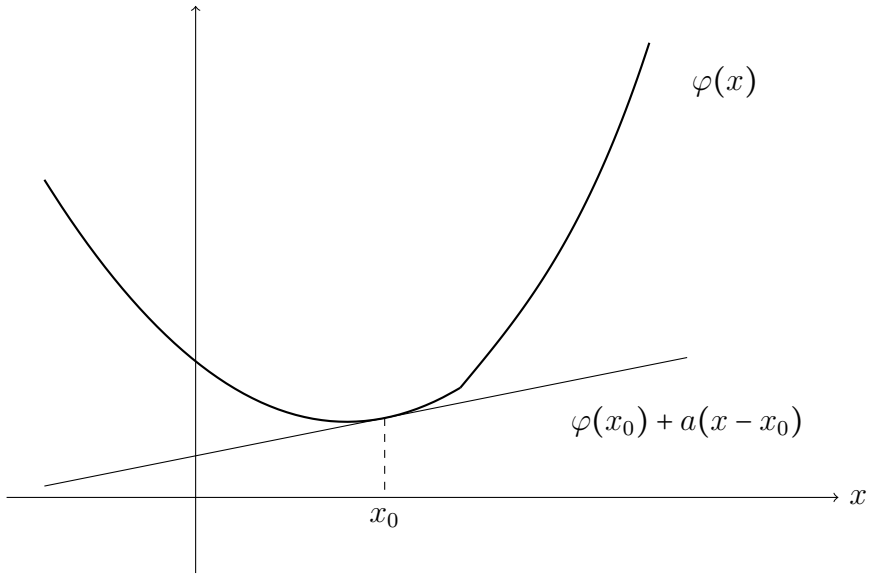
$$\varphi \text{ ist konvex} \quad \Longleftrightarrow \quad \forall\, x, y \in \mathbb{R}, \alpha \in [0,1] : \ \alpha\varphi(x) + (1-\alpha)\varphi(y) \ge \varphi(\alpha x + (1-\alpha)y)$$

(eine konvexe Funktion liegt oberhalb jeder ihrer Tangenten) und zu jedem $x_0 \in \mathbb{R}$ gibt es $a = a(x_0) \in \mathbb{R}$, so dass

$$\forall\, x \in \mathbb{R} : \ \varphi(x_0) + a(x - x_0) \le \varphi(x)$$

gilt.

[1] nach Johan Ludvig Jensen, 1859–1925 benannt



Wegen Monotonie und Linearität des Erwartungswerts (Satz 3.4) gilt also

$$\varphi(x_0) + a\big(\mathbb{E}[X] - x_0\big) \le \mathbb{E}\big[\varphi(X)\big],$$

mit der Wahl $x_0 = \mathbb{E}[X]$ folgt die Behauptung. □

3.2 Der Fall mit Dichte

Definition 3.9. X reelle ZV mit Dichte f_X, dann ist

$$X \in \mathcal{L}^1 \quad :\Leftrightarrow \quad \int_{\mathbb{R}} |x| f_X(x)\, dx < \infty$$

(man sagt dann, X besitzt einen Erwartungswert) und

$$\mathbb{E}[X] := \int_{\mathbb{R}} x f_X(x)\, dx.$$

Beispiel 3.10. 1. $X \sim \mathcal{N}_{0,1}$ hat $\mathbb{E}[X] = 0$, denn

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |x| e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x e^{-x^2/2}\, dx = \sqrt{2/\pi} \Big[-e^{-x^2/2} \Big]_0^{\infty} = \sqrt{2/\pi} < \infty$$

und aus der Symmetrie der Dichte folgt

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} x e^{-x^2/2}\, dx - \frac{1}{\sqrt{2\pi}} \int_0^{\infty} x e^{-x^2/2}\, dx = 0.$$

2. Die Gammaverteilung $\Gamma_{a,\nu}$ $(a, \nu > 0)$ hat Dichte

$$\frac{a^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-ax} \mathbf{1}_{(0,\infty)}(x)$$

(mit $\Gamma(\nu) = \int_0^\infty x^{\nu-1}e^{-x}\,dx$).

Für $X \sim \Gamma_{a,\nu}$ ist

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x\,\frac{a^\nu}{\Gamma(\nu)}x^{\nu-1}e^{-ax}\,dx \\ &= \frac{\Gamma(\nu+1)}{a\Gamma(\nu)}\int_0^\infty \frac{a^{\nu+1}}{\Gamma(\nu+1)}x^{(\nu+1)-1}e^{-ax}\,dx = \frac{\Gamma(\nu+1)}{a\Gamma(\nu)} = \frac{\nu}{a}\end{aligned}$$

(denn $\Gamma(\nu+1) = \nu\Gamma(\nu)$).

3. Die Cauchy-Verteilung mit Dichte $\frac{1}{\pi}\frac{1}{1+x^2}$ besitzt keinen Erwartungswert:

$$\int_{-\infty}^\infty \frac{1}{\pi}\frac{|x|}{1+x^2}\,dx = \frac{2}{\pi}\int_0^\infty \frac{1}{\pi}\frac{x}{1+x^2}\,dx = \Big[\frac{1}{\pi}\log\big(1+x^2\big)\Big]_0^\infty = \infty.$$

Bericht 3.11. 1. (Zur Herleitung des Falls mit Dichte aus der allgemeinen Situation aus Bericht 3.7)

$X_{(n)} = \frac{1}{n}\lfloor nX\rfloor$ nimmt den Wert $\frac{k}{n}$, $k \in \mathbb{Z}$ an mit

$$P\Big(X_{(n)} = \frac{k}{n}\Big) = \int_{k/n}^{(k+1)/n} f_X(x)\,dx,$$

also ist (sofern die Reihe absolut konvergiert, was man analog überprüft)

$$\begin{aligned}\mathbb{E}\big[X_{(n)}\big] &= \sum_{k\in\mathbb{Z}}\frac{k}{n}\int_{k/n}^{(k+1)/n} f_X(x)\,dx \\ &= \int_{\mathbb{R}}\frac{1}{n}\lfloor nx\rfloor f_X(x)\,dx \underset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}} x f_X(x)\,dx\end{aligned}$$

(die Konvergenz folgt, falls $xf_X(x)$ [uneigentlich] Riemann-integrierbar ist, aus der Riemann-Approximation, für den allgemeinen Fall mit messbarem f_X aus dem Konvergenzsatz von Lebesgue).

2. (Analogon zu Beob. 3.5 im Fall mit Dichte)
Sei $X = (X_1, \dots, X_d)$ $\mathbb{R}^d$-wertig mit Dichte $f_X : \mathbb{R}^d \to [0,\infty]$, $g : \mathbb{R}^d \to \mathbb{R}$, $Y := g(X)$. Dann gilt $Y \in \mathcal{L}^1$ g.d.w.

$$\int_{\mathbb{R}^d} |g(x_1,\dots,x_d)| f_X(x_1,\dots,x_d)\,dx_1\dots dx_d < \infty$$

und in diesem Fall

$$\mathbb{E}[Y] = \int_{\mathbb{R}^d} g(x_1,\dots,x_d) f_X(x_1,\dots,x_d)\,dx_1\dots dx_d < \infty.$$

(Siehe z.B. [G, Korollar 4.13])

3.3 Varianz und Kovarianz

Definition 3.12. X reelle ZV, $p > 0$. X besitzt p-tes Moment (auch geschrieben $X \in \mathcal{L}^p$), wenn $\mathbb{E}\big[|X|^p\big] < \infty$.

$\mathbb{E}\big[X^p\big]$ heißt das p-te Moment von X.

Bemerkung 3.13. Für $p \geq p' \geq 1$ ist $\mathcal{L}^p \subset \mathcal{L}^{p'}$ (denn $|x|^{p'} \leq 1 + |x|^p$).

Definition 3.14. Für $X, Y \in \mathcal{L}^2$ heißt

1. $\mathrm{Var}[X] := \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2$ die *Varianz* von X
   (manchmal schreibt man auch $\sigma_X^2 := \mathrm{Var}[X]$),
   $\sqrt{\mathrm{Var}[X]}$ die *Standardabweichung* (oder *Streuung*) von X
   (manchmal auch $\sigma_X = \sqrt{\sigma_X^2}$ geschrieben),
2. $\mathrm{Cov}[X, Y] := \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
   die *Kovarianz* von X und Y.

X und Y heißen *unkorreliert*, wenn $\mathrm{Cov}[X, Y] = 0$.

Beobachtung 3.15.

1. Wegen $|XY| \leq X^2 + Y^2$ ist die Kovarianz wohldefiniert. Es gilt (offensichtlich)
$$\mathrm{Cov}[X, Y] = \mathrm{Cov}[Y, X].$$
2. Es ist
$$\begin{aligned}\mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] &= \mathbb{E}\big[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]\big] \\ &= \mathbb{E}[YX] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$
   (und analog für $\mathrm{Var}[X] = \mathrm{Cov}[X, X]$).
3. $\mathrm{Var}[X] = 0 \iff P\big(X = \mathbb{E}[X]\big) = 1$
   („⇐“ ist klar, für „⇒“ wende Satz 3.4, 3. an auf die ZV $(X - \mathbb{E}[X])^2$)
4. $\mathrm{Var}[X]$ ist eine Eigenschaft der Verteilung von X, $\mathrm{Cov}[X, Y]$ ist eine Eigenschaft der gemeinsamen Verteilung von X und Y.

Beispiel 3.16. 1. $X \sim \mathrm{Ber}_p$, $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p)$.

2. $X \sim \mathrm{Poi}_\alpha$,
$$\mathbb{E}[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)e^{-\alpha}\frac{\alpha^k}{k!} = \alpha^2 \sum_{k=2}^{\infty} e^{-\alpha}\frac{\alpha^{k-2}}{(k-2)!} = \alpha^2,$$
also
$$\mathrm{Var}[X] = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2 = \mathbb{E}\big[X(X-1)\big] + \mathbb{E}[X] - \big(\mathbb{E}[X]\big)^2 = \alpha^2 + \alpha - \alpha^2 = \alpha$$

3. $X \sim \mathrm{Bin}_{n,p}$,
$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\binom{n}{k}p^k(1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}(1-p)^{(n-2)-(k-2)} = n(n-1)p^2\end{aligned}$$

also

$$\mathrm{Var}[X] = \mathbb{E}\big[X(X-1)\big] + \mathbb{E}[X] - \big(\mathbb{E}[X]\big)^2 = n(n-1)p^2 + np - (np)^2 = -np^2 + np = np(1-p)$$

4. $X \sim \mathrm{Geom}_p$, $p \in [0,1]$ (d.h. $P(X=k) = p(1-p)^k$, $k \in \mathbb{N}_0$, vgl. Bsp. 1.21 und wir hatten gesehen, dass $\mathbb{E}[X] = (1-p)/p$, siehe Bsp. 3.3, 3).

Es ist

$$\mathbb{E}[X(X-1)] = \sum_{n=0}^{\infty} n(n-1)p(1-p)^n = p(1-p)^2 \sum_{n=2}^{\infty} n(n-1)(1-p)^{n-2} = 2\frac{(1-p)^2}{p^2}$$

(verwende, dass $f(t) := \sum_{n=0}^{\infty} t^n = \frac{1}{1-t}$ (für $|t| < 1$) erfüllt $\frac{d^2}{dt^2} f(t) = \frac{2}{(1-t)^3} = \sum_{n=2}^{\infty} n(n-1)t^{n-2}$), somit

$$\mathrm{Var}[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X]\big(1 - \mathbb{E}[X]\big) = 2\frac{(1-p)^2}{p^2} - \frac{1-p}{p} \cdot \frac{2p-1}{p} = \frac{1-p}{p^2}.$$

5. $X \sim \mathcal{N}_{\mu,\sigma^2}$, $\mathrm{Var}[X] = \sigma^2$:

$$\begin{aligned}
\mathrm{Var}[X] = \mathbb{E}\big[(X-\mu)^2\big] &= \int_{\mathbb{R}} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)\, dx \\
&= \int_{\mathbb{R}} \sigma^2 z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{z^2 e^{-z^2/2}}_{= z\left(-\frac{d}{dz} e^{-z^2/2}\right)}\, dz \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(\Big[z\big(-e^{-z^2/2}\big) \Big]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-z^2/2}\, dz \right) = \frac{\sigma^2}{\sqrt{2\pi}} \big(0 + \sqrt{2\pi}\big) = \sigma^2
\end{aligned}$$

(Wir haben Bericht 3.11, 2. verwendet, dann im Integral $z = (x-\mu/\sigma)$ substituiert und partiell integriert.)

Satz 3.17 (Rechenregeln für Varianz und Kovarianz)**.** *Seien $X, Y, X_1, X_2, \ldots, X_n \in \mathcal{L}^2$, $a, b, c, d \in \mathbb{R}$.*

1. $aX + b, cY + d \in \mathcal{L}^2$ und

$$\mathrm{Cov}[aX + b, cY + d] = ac\mathrm{Cov}[X, Y],$$

insbesondere

$$\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$$

(die Kovarianz ist eine Bilinearform, die Varianz ein quadratisches Funktional).

2. $\mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + \sum_{\substack{1 \le i,j \le n \\ i \ne j}} \mathrm{Cov}[X_i, X_j],$

insbesondere gilt für paarweise unkorrelierte $X_1, \ldots, X_n$ also $\mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i]$.

3. Sind X und Y unabhängig, so gilt $\mathrm{Cov}[X,Y] = 0$.

4. Es gilt

$$\big|\mathrm{Cov}[X,Y]\big| \le \sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]} \quad \textit{(Cauchy-Schwarz-Ungleichung}^{2}\textit{)}$$

Beweis. 1. Es ist

$$\begin{aligned}\mathrm{Cov}[aX+b,cY+d] &= \mathrm{Cov}[aX,cY] \quad (\text{denn } \mathbb{E}[aX+b] = \mathbb{E}[aX]+b \text{ und } \mathbb{E}[cY+d] = \mathbb{E}[cY]+d)\\ &= \mathbb{E}[aX\,cY] - \mathbb{E}[aX]\,\mathbb{E}[cY] = ac\big(E[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\big)\\ &= ac\mathrm{Cov}[X,Y].\end{aligned}$$

2. Dies folgt etwa per Induktion über n aus 1., oder direkt folgendermaßen:

Sei o.E. $\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] = 0$ (sonst ziehe jeweils die Erwartungswerte ab, verwende 1.), dann ist

$$\begin{aligned}\mathrm{Var}\Big[\sum_{i=1}^n X_i\Big] &= \mathbb{E}\Big[\Big(\sum_{i=1}^n X_i\Big)^2\Big] = \sum_{i,j=1}^n \mathbb{E}\big[X_iX_j\big]\\ &= \sum_{i=1}^n \mathbb{E}\big[X_i^2\big] + \sum_{i\neq j}^n \mathbb{E}\big[X_iX_j\big] = \sum_{i=1}^n \mathrm{Var}[X_i] + \sum_{i\neq j}^n \mathrm{Cov}[X_i,X_j]\end{aligned}$$

3. Klar, denn für X und Y unabhängig ist $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ nach Satz 3.4, 4.

4. Falls $\mathrm{Var}[Y] = 0$, so ist die Ungleichung (als $0 \le 0$) erfüllt
(denn dann ist $P(Y - \mathbb{E}[Y] = 0) = 1$ nach Beob. 3.15, 3. und somit auch $\mathrm{Cov}[X,Y] = 0$).

Falls $\mathrm{Var}[Y] > 0$, setze $\alpha := -\dfrac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[Y]}$, es ist

$$\begin{aligned}0 \le \mathrm{Var}[X+\alpha Y]\,\mathrm{Var}[Y] &\overset{1.}{=} \big(\mathrm{Var}[X] + 2\alpha\mathrm{Cov}[X,Y] + \alpha^2\mathrm{Var}[Y]\big)\mathrm{Var}[Y]\\ &= \mathrm{Var}[X]\,\mathrm{Var}[Y] - \big(\mathrm{Cov}[X,Y]\big)^2.\end{aligned}$$

□

Bemerkung 3.18. Es gilt Gleichheit in der Cauchy-Schwarz-Ungleichung g.d.w.

$$\text{es gibt } a,b,c \in \mathbb{R} \text{ (mit } a \neq 0 \text{ oder } b \neq 0\text{), so dass } P\big(aX+bY+c=0\big) = 1.$$

In diesem Fall heißen X und Y *perfekt korreliert*.

(Denn wir sehen aus dem Beweis von Satz 3.17, 4., dass Gleichheit genau dann eintritt, wenn $\mathrm{Var}[Y] = 0$ oder $\mathrm{Var}[X+\alpha Y] = 0$.)

[2] nach Augustin-Louis Cauchy (1789–1857) und Hermann Amandus Schwarz (1843–1921)

Beispiel 3.19. 1. $X \sim \mathrm{Bin}_{n,p}$, schreibe $X = Y_1 + \cdots + Y_n$ mit Y_i u.i.v. $\sim \mathrm{Ber}_p$, so ist (mit Satz 3.17, 2.)

$$\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[Y_i] = n\mathrm{Var}[Y_1] = np(1-p)$$

(vgl. auch Bsp. 3.16, 3.).

2. $X \sim \mathrm{Hyp}_{s,w,n}$, stelle dar als $X = Y_1 + \cdots + Y_n$ mit $Y_i = \mathbf{1}_{A_i}$, $A_i = \{i\text{-te gezogene Kugel ist schwarz}\}$ (bei n-fachem Ziehen ohne Zurücklagen aus einer Urne mit s schwarzen und w weißen Kugeln).

Es ist $\mathbb{E}[Y_i] = P(A_i) = P(A_1) = \dfrac{s}{s+w} =: p$, $\mathrm{Var}[Y_i] = p(1-p)$; für $i \neq j$ ist

$$\begin{aligned}
\mathbb{E}[Y_i Y_i] &= \mathbb{E}[Y_1 Y_2] = P(A_1 \cap A_2) = \frac{s}{s+w}\frac{s-1}{s+w-1}, \\
\mathrm{Cov}[Y_i, Y_j] &= \mathbb{E}[Y_i Y_i] - \mathbb{E}[Y_i]\,\mathbb{E}[Y_j] = \frac{s}{s+w}\frac{s-1}{s+w-1} - \Big(\frac{s}{s+w}\Big)^2 \\
&= \frac{s}{s+w}\Big(\underbrace{\frac{s-1}{s+w-1} - \frac{s}{s+w}}_{=-\frac{w}{s+w}\,\frac{1}{s+w-1}}\Big) = -p(1-p)\frac{1}{s+w-1},
\end{aligned}$$

also

$$\begin{aligned}
\mathrm{Var}[X] &= \sum_{i=1}^{n} \mathrm{Var}[Y_i] + \sum_{i \neq j}^{n} \mathrm{Cov}[Y_i, Y_j] = np(1-p) - n(n-1)\Big(-p(1-p)\frac{1}{s+w-1}\Big) \\
&= np(1-p)\Big(1 - \frac{n-1}{s+w-1}\Big)
\end{aligned}$$

3. Z reelle ZV mit $\mathbb{E}\big[|Z|^3\big] < \infty$ und *symmetrischer* Verteilung, d.h. es gilt $P(Z > z) = P(Z < -z)$ für alle $z \geq 0$ (z.B. $Z \sim \mathcal{N}_{0,1}$), setze

$$Y := Z^2,$$

dann gilt

$$\mathrm{Cov}[Y, Z] = \mathbb{E}[Z^2 Z] - \mathbb{E}[Z^2]\mathbb{E}[Z] = \mathbb{E}[Z^3] - \mathbb{E}[Z^2]\mathbb{E}[Z] = 0 - \mathbb{E}[Z^2] \cdot 0 = 0.$$

Z und Y sind also unkorreliert, aber i.A. *nicht* unabhängig.

Definition 3.20. Seien $X, Y \in \mathcal{L}^2$.

$$\kappa_{X,Y} := \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} \in [-1, 1]$$

heißt *Korrelationskoeffizient* von X und Y (manche Autoren schreiben auch $\rho_{X,Y}$).

(Die Cauchy-Schwarz-Ungleichung (Satz 3.17, 4.) zeigt, dass $|\kappa_{X,Y}| \leq 1$.)

Beobachtung 3.21 (Interpretation des Korrelationskoeffizienten via „beste lineare Vorhersage“)**.** Es ist

$$\min_{\beta_0,\beta_1\in\mathbb{R}} \mathbb{E}\big[(Y-\beta_1 X-\beta_0)^2\big] = (1-\kappa_{X,Y}^2)\min_{\beta_0\in\mathbb{R}} \mathbb{E}\big[(Y-\beta_0)^2\big] \quad \Big(= (1-\kappa_{X,Y}^2)\mathrm{Var}[Y]\Big),$$

denn der Ausdruck auf der linken Seite ist

$$\begin{aligned}
&\mathrm{Var}[Y-\beta_1 X-\beta_0] + \big(\mathbb{E}[Y]-\beta_1\mathbb{E}[X]-\beta_0\big)^2\\
&= \mathrm{Var}[Y] - 2\beta_1\mathrm{Cov}[X,Y] + \beta_1^2\mathrm{Var}[X] + \big(\mathbb{E}[Y]-\beta_1\mathbb{E}[X]-\beta_0\big)^2\\
&= \sigma_Y^2 - 2\beta_1\sigma_X\sigma_Y\kappa_{X,Y} + \beta_1^2\sigma_X^2 + \big(\mathbb{E}[Y]-\beta_1\mathbb{E}[X]-\beta_0\big)^2\\
&= \sigma_Y^2\big(1-\kappa_{X,Y}^2\big) + \sigma_X^2\Big(\beta_1 - \frac{\sigma_Y}{\sigma_X}\kappa_{X,Y}\Big)^2 + \big(\mathbb{E}[Y]-\beta_1\mathbb{E}[X]-\beta_0\big)^2,
\end{aligned}$$

was offensichtlich minimal wird für die Wahl

$$\beta_1 = \beta_1^* := \frac{\sigma_Y}{\sigma_X}\kappa_{X,Y}, \quad \beta_0 = \beta_0^* := \mathbb{E}[Y] - \beta_1^*\mathbb{E}[X]$$

und dann den Wert $(1-\kappa_{X,Y}^2)\sigma_Y^2$ hat.

(Für den Zusatz beachte analog:

$$\mathbb{E}\big[(Y-\beta_0)^2\big] = \mathbb{E}[Y^2] - 2\beta\mathbb{E}[Y] + \beta^2 = \mathrm{Var}[Y] + \big(\beta - \mathbb{E}[Y]\big)^2$$

ist minimal für die Wahl $\beta = \mathbb{E}[Y]$.)

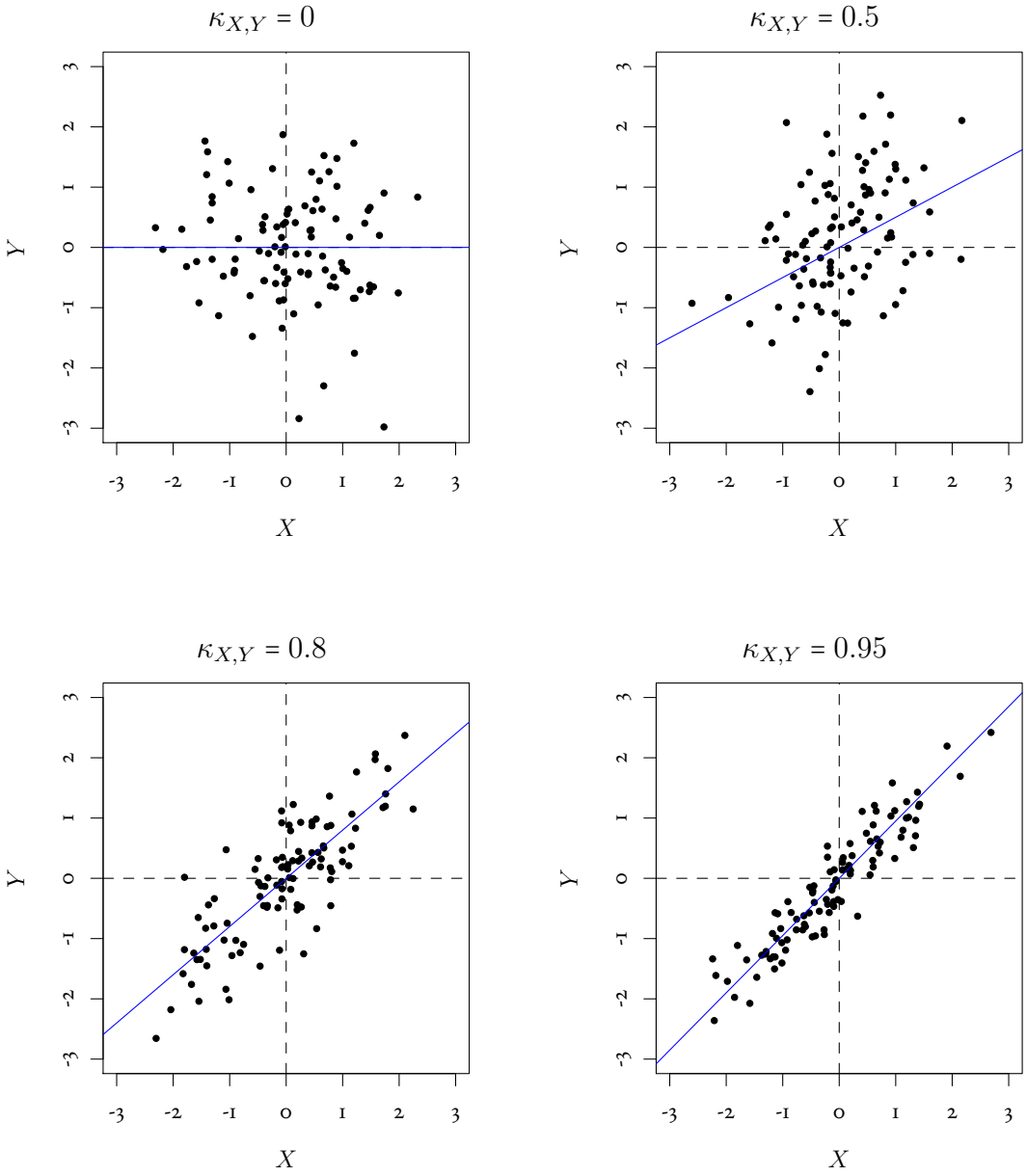
Im Sinne einer möglichst kleinen quadratischen Abweichung ist $\mathbb{E}[Y]$ die beste konstante „Vorhersage“ von Y. Man kann demnach um einen Faktor $(1-\kappa_{X,Y}^2)$ besser vohersagen, wenn man stattdessen eine affin-lineare Funktion von X verwenden darf.

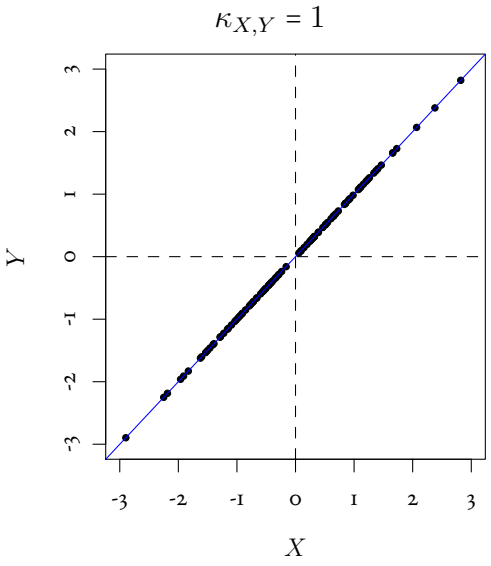
Demnach (vgl. auch Bem. 3.18)

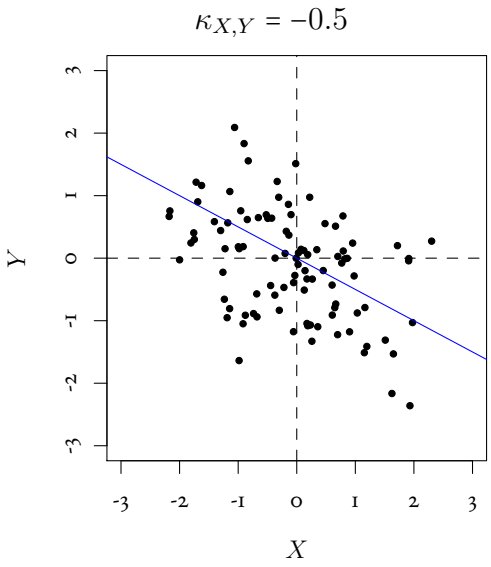
$\lvert\kappa_{X,Y}\rvert = 1$	$\leftrightarrow$	perfekter linearer Zusammenhang zwischen X und Y
$\kappa_{X,Y} = 1$	$\leftrightarrow$	perfekter linearer Zusammenhang zwischen X und Y mit positivem Koeffizienten (X größer als $\mathbb{E}[X] \iff Y$ größer als $\mathbb{E}[Y]$)
$\kappa_{X,Y} = -1$	$\leftrightarrow$	perfekter linearer Zusammenhang zwischen X und Y mit negativem Koeffizienten (X größer als $\mathbb{E}[X] \iff Y$ kleiner als $\mathbb{E}[Y]$)

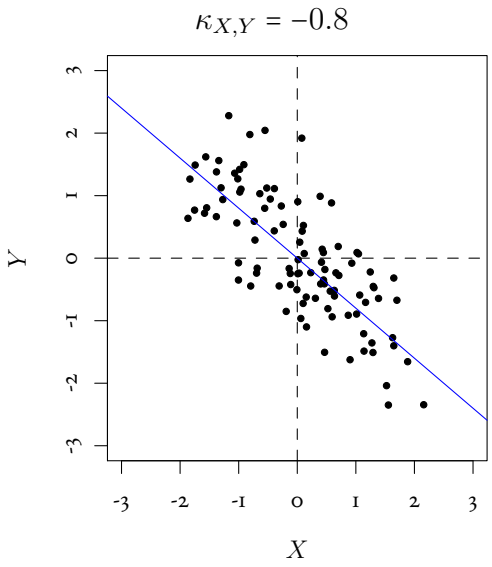
Nicht-lineare Zusammenhänge erfasst der Korrelationskoeffizient möglicherweise nicht korrekt (oder gar nicht), vgl. Bsp. 3.19, 3.

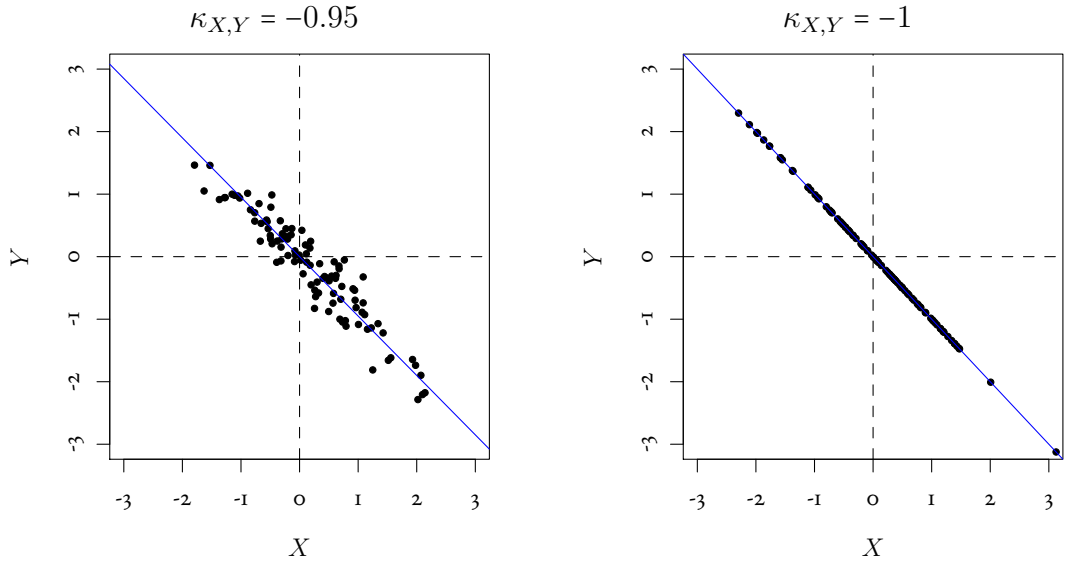
Die folgenden Scatterplots zeigen jeweils 100 simulierte Paare (X,Y), wobei $\sigma_X = \sigma_Y = 1$ und $\kappa_{X,Y}$ den angegebenen Wert hat. (Blau eingezeichnet ist die „Vorhersagegerade“ $x \mapsto \beta_1^* x + \beta_0^*$.)

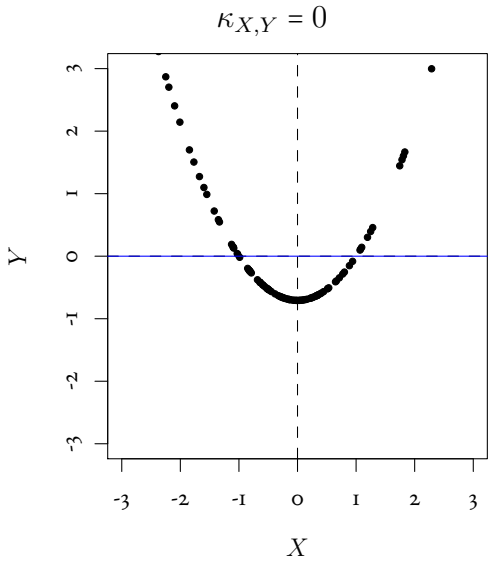

$\kappa_{X,Y} = 0$
$\kappa_{X,Y} = 0.5$
$\kappa_{X,Y} = 0.8$
$\kappa_{X,Y} = 0.95$
X
Y
-3
-2
-1
0
1
2
3


$\kappa_{X,Y} = 1$
Y
X
-3
-2
-1
0
1
2
3


$\kappa_{X,Y} = -0.5$
Y
X
-3
-2
-1
0
1
2
3


$\kappa_{X,Y} = -0.8$
Y
X
-3
-2
-1
0
1
2
3





3.4 Median(e)

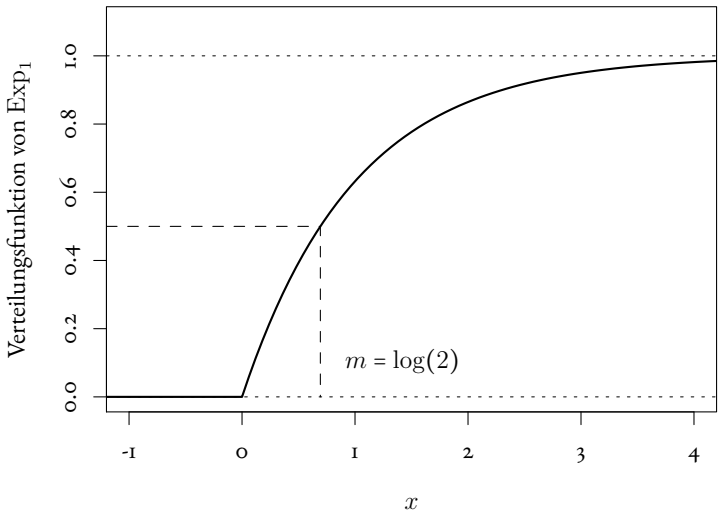
Definition 3.22. X reelle ZV, m heißt (ein) Median von X (auch „Zentralwert“, manchmal auch m_X geschrieben), wenn gilt

$$P(X \geq m) \geq \frac{1}{2} \quad \text{und} \quad P(X \leq m) \geq \frac{1}{2}.$$

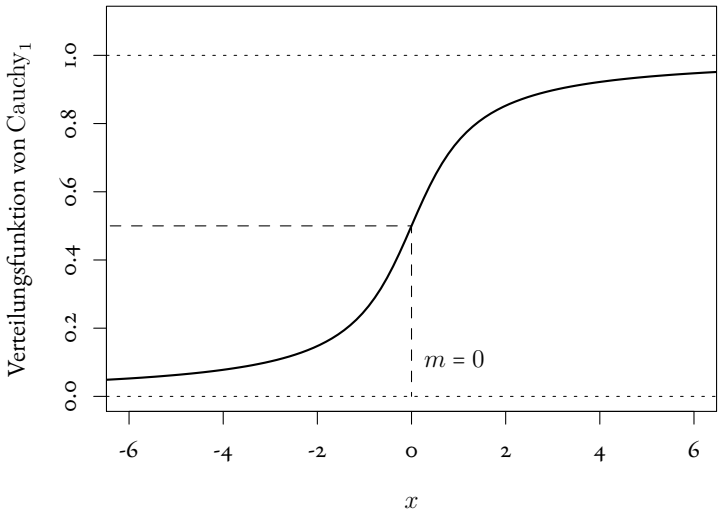
Ein Median existiert stets, auch wenn X keinen Erwartungswert besitzt.

Man kann den Median als eine „robustere“ Antwort auf die Aufgabe, für eine ZV *nur einen* „typischen Wert“ anzugeben, ansehen. Allerdings gibt es für Mediane keine so angenehmen Rechenregeln, wie sie Satz 3.4 für den Erwartungswert liefert.

Beispiel 3.23. 1. $X \sim \mathrm{Exp}_\theta$ hat Dichte $\theta e^{-\theta x}\mathbf{1}_{[0,\infty)}(x)$, Verteilungsfunktion $(1 - e^{-\theta x})\mathbf{1}_{[0,\infty)}(x)$, demnach ist der (eindeutig bestimmte) Median $m = \frac{1}{\theta}\log 2$.

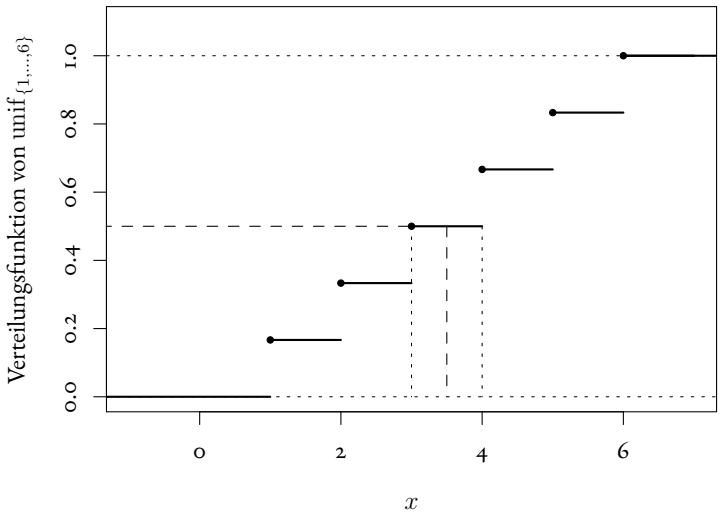


2. X Cauchy-verteilt mit Dichte $\frac{1}{\pi}\frac{1}{1+x^2}$, Verteilungsfunktion $\frac{1}{2}\frac{1}{\pi}\arctan(x)$, der (eindeutig bestimmte) Median ist $m = 0$.



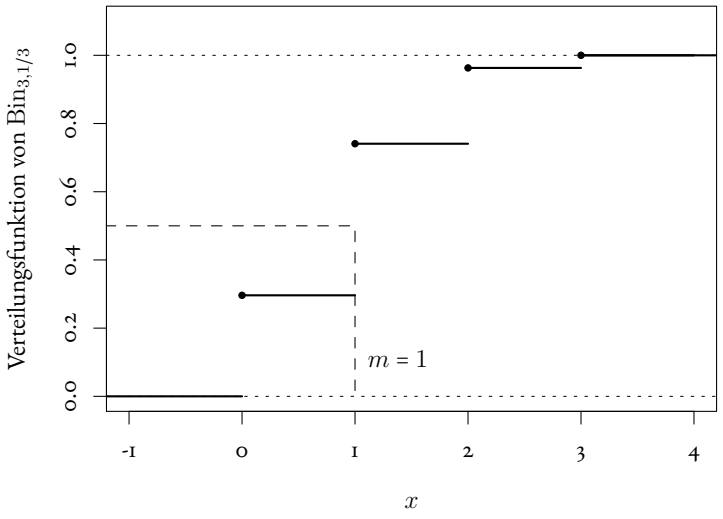
(Wegen der Symmetrie der Dichte, es gibt keinen Erwartungswert, vgl. Bsp. 3.10, 3.)

3. $X \sim \mathrm{unif}_{\{1,2,\ldots,6\}}$



Jeder Wert $m \in [3,4]$ ist ein Median (und die vielleicht „kanonischste“ Wahl wäre $m = 3{,}5$).

4. $X \sim \mathrm{Bin}_{3,1/3}$ hat Median 1



Bemerkung 3.24. Sei $X \in \mathcal{L}^1$.

1. Jeder Median von X ist ein Minimierer von $a \mapsto \mathbb{E}\big[|X-a|\big]$.

2. Für jeden Median m ist $\big|\mathbb{E}[X]-m\big| \le \sqrt{\mathrm{Var}[X]}$.

Beweis. 1. Sei m ein Median. Falls $a > m$:

$$|X-a| - |X-m| \ge (a-m)\mathbf{1}_{\{X\le m\}} - (a-m)\mathbf{1}_{\{X>m\}},$$

also

$$\mathbb{E}\big[|X-a|\big] - \mathbb{E}\big[|X-m|\big] \ge (a-m)\big(\underbrace{P(X\le m)}_{\ge 1/2} - \underbrace{P(X>m)}_{\le 1/2}\big) \ge 0,$$

analog im Fall $a < m$.

2. Es ist

$$\big|\mathbb{E}[X] - m\big| \le \mathbb{E}\big[|X - m|\big] \overset{1.}{\le} \mathbb{E}\big[|X - \mathbb{E}[X]|\big] = \sqrt{\Big(\mathbb{E}\big[|X - \mathbb{E}[X]|\big]\Big)^2} \le \sqrt{\mathbb{E}\big[|X - \mathbb{E}[X]|^2\big]},$$

wobei wir für die erste und die dritte Ungleichung jeweils die Jensensche Ungleichung (Satz 3.8) verwenden. □

3.5 Erzeugende Funktionen*

Definition 3.25. Sei X eine ZV mit Werten in $\mathbb{N}_0$,

$$\varphi_X(s) = \mathbb{E}\big[s^X\big] = \sum_{n=0}^{\infty} s^n P(X = n), \quad s \in [0,1].$$

heißt die *erzeugende Funktion* von X.

Analog ist für ein Wahrscheinlichkeitsmaß μ auf $\mathbb{N}_0$ die erzeugende Funktion

$$\varphi_\mu(s) = \sum_{n=0}^{\infty} s^n \mu(\{n\}).$$

Beobachtung. φ_X (und genauso φ_μ) ist zumindest für $s \in [0,1]$ wohldefiniert mit $0 \le \varphi_X(s) \le 1 = \varphi_X(1)$, ist auf $[0,1)$ glatt, ist konvex (strikt konvex, sofern $P(X > 1) > 0$ bzw. $\mu(\{2,3,\dots\}) > 0$).

Satz 3.26 (Momentenbestimmung mittels der erzeugenden Funktion)**.**

1. $P(X = n) = \frac{\varphi_X^{(n)}(0)}{n!}$ *für* $n \in \mathbb{N}_0$*, insbesondere ist* $\mathscr{L}(X)$ *durch* φ_X *eindeutig bestimmt.*
2. $\mathbb{E}[X] = \varphi_X'(1)$ *(sofern existent)*
3. $\mathbb{E}[X(X-1)] = \varphi_X''(1)$*, insbesondere ist* $\mathrm{Var}[X] = \varphi_X''(1) + \mathbb{E}[X] - \big(\mathbb{E}[X]\big)^2$ *(sofern existent)*

Beweis. 1. Wegen $0 \le P(X = k)$ und $\sum_{k=0}^{\infty} P(X = k) = 1$ hat die Potenzreihe

$$\sum_{k=0}^{\infty} s^k P(X = k) = \varphi_X(s)$$

mindestens Konvergenzradius 1. Aus der Analysis ist bekannt, dass sie somit an jeder Stelle s mit $|s| < 1$ (n-mal) gliedweise differenziert werden kann und es gilt

$$\frac{d^n}{ds^n}\varphi_X(s) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1)s^{k-n}P(X = k),$$

also für $s = 0$

$$\frac{d^n}{ds^n}\varphi_X(0) = n!P(X = n).$$

2. Für $0 \le s < 1$ ist nach obigem

$$\varphi_X'(s) = \sum_{n=1}^{\infty} n s^{n-1} P(X = n) = \mathbb{E}\big[X s^{X-1}\big].$$

Für $s \nearrow 1$ steigt die rechte Seite monoton auf gegen $\sum_{n=1}^{\infty} nP(X = n)$, wenn φ_X in $s = 1$ differenzierbar ist, konvergiert die linke Seite gegen $\varphi'(1-) = \varphi'(1)$.

(Da φ_X konvex ist in $[0, 1)$, existiert

$$\varphi_X'(1-) = \lim_{s \nearrow 1} \varphi_X'(s) \in (-\infty, \infty]$$

stets, und wir sehen aus obigem Argument, dass $\mathbb{E}[X] < \infty \iff \varphi_X'(1-) < \infty$.)

3. Analog ist für $0 \le s < 1$

$$\varphi_X''(s) = \sum_{n=2}^{\infty} n(n-1) s^{n-2} P(X = n) = \mathbb{E}\big[X(X-1)s^{X-2}\big]$$

und sofern φ_X in $s = 1$ zweimal differenzierbar ist, folgt die Behauptung mit $s \nearrow 1$.

(Wie oben ist $\varphi_X''(1-) < \infty \iff \mathbb{E}\big[X(X-1)\big] < \infty$.) □

Satz 3.27. *$X_1, \dots, X_m$ unabhängige, $\mathbb{N}_0$-wertige ZVn, so ist*

$$\varphi_{X_1 + \dots + X_m}(s) = \varphi_{X_1}(s) \cdot \varphi_{X_2}(s) \cdot \dots \cdot \varphi_{X_m}(s)$$

Analog gilt für W'maße $\mu_1, \mu_2, \dots, \mu_n$ auf $\mathbb{N}_0$

$$\varphi_{\mu_1 * \dots * \mu_m}(s) = \varphi_{\mu_1}(s) \cdot \dots \cdot \varphi_{\mu_m}(s)$$

Beweis. Für $s \in [0, 1]$ ist

$$\mathbb{E}\big[s^{X_1 + \dots + X_m}\big] = \mathbb{E}\big[s^{X_1} \cdot s^{X_2} \dots s^{X_m}\big] = \mathbb{E}\big[s^{X_1}\big] \cdot \mathbb{E}\big[s^{X_2}\big] \dots \mathbb{E}\big[s^{X_m}\big]$$

mit Satz 3.4, 4. □

Beispiel 3.28.

1. $X \sim \mathrm{Ber}_p$, $\varphi_X(s) = ps + 1 - p$
2. $X \sim \mathrm{Bin}_{n,p}$, $\varphi_X(s) = (ps + 1 - p)^n = (1 - p(1-s))^n$ (explizite Rechnung oder verwende Satz 3.27)
3. $X \sim \mathrm{Poi}_\lambda$, $\varphi_X(s) = \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} s^n = e^{-\lambda} e^{\lambda s} = e^{-\lambda(1-s)}$
4. $X \sim \mathrm{Geom}_p$, $\varphi_X(s) = \sum_{n=0}^{\infty} p(1-p)^n s^n = \frac{p}{1-(1-p)s}$

Beispiel 3.29 (Eine Skizze zu Galton-Watson-Prozessen)**.** Wir beobachten zunächst: Sei X $\mathbb{N}_0$-wertig, $X_1, X_2, \dots$ u.i.v. Kopien von X, davon unabhängig Y $\mathbb{N}_0$-wertig, $Z := \sum_{i=1}^{Y} X_i$ (mit Interpretation

$Z = 0$ auf $\{Y = 0\}$), dann ist $\varphi_Z(s) = \varphi_Y(\varphi_X(s))$, denn

$$\begin{aligned}
\mathbb{E}\big[s^Z\big] &= \sum_{z=0}^{\infty} s^z P(Z = z) = \sum_{z=0}^{\infty}\sum_{y=0}^{\infty} s^z P(Z = z, Y = y) \\
&= \sum_{z=0}^{\infty}\sum_{y=0}^{\infty} \sum_{\substack{x_1,\dots,x_y\in\mathbb{N}_0 \\ x_1+\cdots+x_y=z}} s^z P(Y = y, X_1 = x_1, \dots, X_y = x_y) \\
&= \sum_{z=0}^{\infty}\sum_{y=0}^{\infty} \sum_{\substack{x_1,\dots,x_y\in\mathbb{N}_0 \\ x_1+\cdots+x_y=z}} s^{x_1+\cdots+x_y} P(Y = y)P(X_1 = x_1)\cdots P(X_y = x_y) \\
&= \sum_{y=0}^{\infty} P(Y = y) \sum_{x_1=0}^{\infty} \cdots \sum_{x_y=0}^{\infty} s^{x_1} P(X = x_1)\cdots s^{x_y} P(X = x_y) \\
&= \sum_{y=0}^{\infty} P(Y = y)\big(\varphi_X(s)\big)^y = \varphi_Y(\varphi_X(s)).
\end{aligned}$$

Man findet daraus (zusammen mit Satz 3.26, 2.) insbesondere die Wald'sche Identität (zumindest im Fall $\mathbb{N}_0$-wertiger Summanden):

$$\mathbb{E}[Z] = \varphi'_Z(1) = \varphi'_Y(\varphi_X(1))\varphi'_X(1) = \varphi'_Y(1)\varphi'_X(1) = \mathbb{E}[Y]\mathbb{E}[X].$$

Seien nun $X_{n,i}, n, i \in \mathbb{N}$ u.i.v. Kopien einer $\mathbb{N}_0$-wertigen ZV X (mit $m := \mathbb{E}[X] = \varphi'_X(1) < \infty$ und $P(X = 1) < 1$), setze

$$Z_0 = 1, \quad Z_n = \sum_{i=1}^{Z_{n-1}} X_{n,i} \quad \text{für } n \in \mathbb{N}.$$

Interpretation: Z_n ist die Größe der n-ten Generation in einer Population, in der jedes Individuum unabhängig eine zufällige, wie X verteilte Anzahl Nachkommen hat. (Man nennt $(Z_n)_{n\in\mathbb{N}_0}$ auch einen Verzweigungsprozess oder Galton-Watson-Prozess[3].)

[3]Diese Art von zufälligen Populationsprozessen wird üblicherweise in der mathematischen Literatur nach Sir Francis Galton (1822–1911) und Henry William Watson (1827–1903) benannt, die die Eigenschaft (3.1) untersuchten: Galton stellte dazu eine Aufgabe, „Problem 4001", in der Ausgabe vom April 1873 der Zeitschrift *Educational times*, in der er nach der Wahrscheinlichkeit des Aussterbens eines Familiennamens fragte (man lese Z_n als die Anzahl der männlichen Nachkommen eines, sagen wir adeligen, Stammvaters in der n-ten Generation, in Galtons gesellschaftlich-historischem Kontext war klar, dass es für die Namensfrage genügt, die Anzahl der Söhne zu untersuchen). Watson sandte einen – fast richtigen – Lösungsvorschlag, den die beiden dann gemeinsam in dem Artikel „On the probability of the extinction of families", *J. Roy. Anthropol. Inst.*, 4 (1874), 138–144 veröffentlichten.

Die Geschichte des Studiums der Verzweigungsprozesse ist nicht allzu geradlinig verlaufen: Wie sich im Nachhinein herausstellte, hatten Galton und Watson zwar den Fall $m \le 1$ richtig behandelt, nicht aber den Fall $m > 1$, obwohl der französische Mathematiker und Statistiker Irénée-Jules Bienaymé (1796–1878) das Problem bereits 1845 korrekt gelöst hatte – sein Artikel darüber war in Vergessenheit geraten. Die Frage, unter welchen Bedingungen $P(Z \text{ stirbt aus}) = 1$ gilt, ist wohl derart natürlich, dass sie im Lauf der Zeit von verschiedenen Autoren unabhängig und mit verschiedenen Interpretationen „entdeckt" und auch beantwortet wurde; siehe dazu David Kendall, The genealogy of genealogy: branching processes before (and after) 1873. With a French appendix containing Bienaymé's paper of 1845, *Bull. London Math. Soc.*, 7 (1975), no. 3, 225–253 und die lesenswerten Vortragsnotizen von Peter Jagers, Some Notes on the History of Branching Processes, from my Perspective. Lecture at the Oberwolfach Symposium on Random Trees, 18–24 January, 2009 `http://www.math.chalmers.se/~jagers/Branching History.pdf`

Aus der Definition (zusammen mit obiger Beobachtung und Satz 3.26) findet man induktiv

$$\varphi_{Z_n}(s) = \big(\underbrace{\varphi_X \circ \varphi_X \circ \cdots \circ \varphi_X}_{n\text{-fache Hintereinanderausführung}}\big)(s) \quad \text{und} \quad \mathbb{E}[Z_n] = \varphi'_{Z_n}(1) = m^n.$$

$Z = (Z_n)_n$ heißt *subkritisch*, wenn $m < 1$, *kritisch*, wenn $m = 1$, *superkritisch*, wenn $m > 1$.

Sei $q_n = P(Z_n = 0) = \varphi_{Z_n}(0)$, offenbar ist $q_n \le q_{n+1}$ (denn $\{Z_n = 0\} \subset \{Z_{n+1} = 0\}$) und

$$q_n \nearrow_{n\to\infty} q = P\Big(\bigcup_{n=1}^{\infty}\{Z_n = 0\}\Big) = P(Z \text{ stirbt aus}).$$

Tatsächlich gilt

$$P(Z \text{ stirbt aus}) = 1 \iff m \le 1. \tag{3.1}$$

Wegen $q_{n+1} = \varphi_X(q_n)$ folgt mit $n \to \infty$ und Stetigkeit von φ_X, dass $q = \varphi_X(q)$, d.h. $q = \lim_{n\to\infty} q_n$ ist ein Fixpunkt von φ_X.

Sei $\tilde{q} = \varphi_X(\tilde{q})$ ein (möglicherweise anderer) Fixpunkt von φ_X, dann ist $0 = q_0 \le \tilde{q}$, n-malige Anwendung von φ_X liefert

$$q_n \le \varphi_X(\tilde{q}) = \tilde{q} \text{ für alle } n, \quad \text{d.h. } q \text{ ist der kleinste Fixpunkt von } \varphi_X$$

Da φ_X konvex ist (und n. Vor. $\varphi_X(s) \not\equiv s$), gilt:

wenn $m = \varphi'_X(1) \le 1$, so ist $q = 1$ der einzige Fixpunkt in $[0, 1]$

wenn $m = \varphi'_X(1) > 1$, so ist $0 \le q < 1$ und die Fixpunkte sind $\{q, 1\}$

(Details als Übung)

Ohne Zweifel hat die Frage nach dem Aussterben (adeliger) Familiennamen Menschen schon lange vor Galton und Watson beschäftigt, so schreibt Jane Austen (1775–1817) in ihrem Roman *Persuasion* über den Vater der Protagonistin (der sich offenbar gerne mit dem in durch (3.1) im Fall $m \le 1$ beschriebenen Phänomen beschäftigte, auch wenn er es sicherlich nicht als mathematische Frage auffasste):

> "Sir Walter Elliot, of Kellynch Hall, in Somersetshire, was a man who, for his own amusement, never took up any book but the Baronetage; ... there his faculties were roused into admiration and respect, by contemplating the limited remnant of the earliest patents ..."

Kapitel 4

Gesetz der großen Zahlen

Satz 4.1. *X reelle ZV, $f : [0,\infty) \to [0,\infty)$ monoton wachsend.*
1. Für $a > 0$ mit $f(a) > 0$ gilt

$$P\big(|X| \geq a\big) \leq \frac{1}{f(a)}\mathbb{E}\big[f(|X|)\big] \qquad \textit{(Markov[1]-Ungleichung).} \tag{4.1}$$

2. Für $X \in \mathcal{L}^2$ gilt

$$P\big(|X - \mathbb{E}[X]| \geq a\big) \leq \frac{\mathrm{Var}[X]}{a^2} \qquad \textit{(Chebyshev[2]-Ungleichung).} \tag{4.2}$$

Beweis. 1. Sei $Y := f(a)\mathbf{1}_{\{|X|\geq a\}}$, so ist $Y \leq f(|X|)$ und

$$\mathbb{E}[Y] = f(a)P\big(|X| \geq a\big) \leq \mathbb{E}\big[f(|X|)\big]$$

nach Satz 3.4, 2.

2. Wende *1.* an auf $\widetilde{X} := X - \mathbb{E}[X]$ und $f(a) = a^2$. □

Korollar 4.2. *1. $X_1, X_2, \cdots \in \mathcal{L}^2$ seien paarweise unkorreliert mit*

$$\sup_n \mathrm{Var}[X_n] \leq \theta < \infty,$$

dann gilt für $\varepsilon > 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big)\right| > \varepsilon\right) \leq \frac{\theta}{\varepsilon^2 n}\ \big(\underset{n\to\infty}{\longrightarrow} 0\big). \tag{4.3}$$

2. Speziell gilt für $X_1, X_2, \cdots \in \mathcal{L}^2$ u.i.v. mit $\mu := \mathbb{E}[X_1]$ und $\sigma^2 := \mathrm{Var}[X_1]$ $(< \infty)$

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2 n} \underset{n\to\infty}{\longrightarrow} 0 \tag{4.4}$$

(schwaches Gesetz der großen Zahlen).

[2] Andrei Andrejewich Markov, 1856–1922.
[2] Pafnuty Lvovich Chebyshev, 1821–1894.

Beweis. Zu (4.3): Wende die Chebyshev-Ungleichung an auf

$$Y := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]),$$

es ist

$$\mathbb{E}[Y] = 0, \quad \operatorname{Var}[Y] = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}[X_i] \le \frac{1}{n^2}n\theta = \frac{\theta}{n}.$$

(4.4) ist ein Spezialfall von (4.3), denn Unabhängigkeit impliziert Unkorreliertheit (Satz 3.17, 3.). □

Definition 4.3. Seien $X_1, X_2, \ldots, X$ relle ZVn (auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, P)$ definiert).

1. $(X_n)_{n\in\mathbb{N}}$ *konvergiert stochastisch* gegen X, auch geschrieben

$$X_n \xrightarrow[n\to\infty]{\text{stoch.}} X,$$

(auch $X_n \xrightarrow[n\to\infty]{} X$ stoch. oder $X_n \xrightarrow[n\to\infty]{P} X$) wenn gilt

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

2. $(X_n)_{n\in\mathbb{N}}$ *konvergiert fast sicher* gegen X, auch geschrieben

$$X_n \xrightarrow[n\to\infty]{\text{f.s.}} X,$$

(oder $X_n \xrightarrow[n\to\infty]{} X$ fast sicher / f.s.) wenn gilt

$$P\Big(X_n \xrightarrow[n\to\infty]{} X\Big) = 1.$$

Beachte:

$$\Big\{X_n \xrightarrow[n\to\infty]{} X\Big\} = \bigcap_{\varepsilon\in\mathbb{Q}\cap(0,\infty)} \bigcup_{n\in\mathbb{N}} \bigcap_{m\ge n} \big\{|X_m - X| \le \varepsilon\big\} \quad \textit{ist} \text{ ein Ereignis.}$$

Wir können Kor. 4.2, 2. aussprechen als

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{\text{stoch.}} \mu$$

Beobachtung 4.4. $X_n \to X$ f.s. $\Rightarrow X_n \to X$ stoch., die Umkehrung gilt i.A. nicht.

Beweis. Sei $P\big(\{X_n \to X\}\big) = 1$. Für $\varepsilon > 0$ gilt

$$P\big(|X_n - X| > \varepsilon\big) \le P\Big(\bigcup_{m\ge n}\big\{|X_m - X| > \varepsilon\big\}\Big) \xrightarrow[n\to\infty]{} P\Big(\big\{|X_m - X| > \varepsilon \text{ für } \infty \text{ viele } m\big\}\Big) = 0$$

d.h. es gilt $X_n \to X$ stochastisch.

Für ein Gegenbeispiel zur Umkehrung seien $X_1, X_2, \dots$ u.a., $X_n \sim \mathrm{Ber}_{1/n}$, dann gilt

$$X_n \xrightarrow[n\to\infty]{\text{stoch.}} 0 \qquad \text{(denn für jedes } 1 > \varepsilon > 0 \text{ ist } P\big(|X_n - 0| > \varepsilon\big) = \frac{1}{n} \to 0\text{)},$$

aber

$$P\Big(\underbrace{\{X_n = 1 \text{ } \infty\text{-oft}\}}_{=\limsup_{n\to\infty}\{X_n=1\}} \Big) = 1$$

□

Man kann mit konvergenten Folgen „wie gewohnt“ rechnen:

Lemma 4.5. *$X_1, X_2, \dots, Y_1, Y_2, \dots, X, Y$ reelle ZVn, $(a_n)_{n\in\mathbb{N}}$ Folge von reellen Zahlen mit $a_n \to a \in \mathbb{R}$, es gelte $X_n \to X$ stochastisch (bzw. fast sicher) und $Y_n \to Y$ stochastisch (bzw. fast sicher). Dann gilt auch*

$$X_n + Y_n \underset{n\to\infty}{\longrightarrow} X + Y \text{ stochastisch (bzw. fast sicher)}, \tag{4.5}$$

$$a_n X_n \underset{n\to\infty}{\longrightarrow} aX \text{ stochastisch (bzw. fast sicher)}. \tag{4.6}$$

Beweis. 1) f.s.-Fall:

$$A := \{X_n \to X\}, \quad B := \{Y_n \to Y\}$$

erfüllen $P(A) = P(B) = 1$, also auch $P(A \cap B) = 1$, und

$$A \cap B = \{X_n \to X, Y_n \to Y\} \subset \{X_n + Y_n \to X + Y\} \quad \text{sowie} \quad A \subset \{a_n X_n \to aX\}$$

2) Der Fall stochastischer Konvergenz:

i) Sei $\varepsilon > 0$, es ist

$$\limsup_{n\to\infty} P\Big(\underbrace{\big|(X_n + Y_n) - (X + Y)\big| > \varepsilon}_{\subset\{|X_n - X| > \varepsilon/2\}\cup\{|Y_n - Y| > \varepsilon/2\}} \Big)$$
$$\le \limsup_{n\to\infty} P\big(|X_n - X| > \varepsilon/2\big) + \limsup_{n\to\infty} P\big(|Y_n - Y| > \varepsilon/2\big) = 0.$$

ii) Sei $\varepsilon > 0$ (und wir nehmen o.E. an, dass $\sup_m |a_m| > 0$), es gilt

$$|a_n X_n - aX| = |a_n(X_n - X) + (a_n - a)X| \le |a_n|\,|X_n - X| + |a_n - a|\,X,$$

also

$$\limsup_{n\to\infty} P\big(|a_n X_n - aX| > \varepsilon\big)$$
$$\le \limsup_{n\to\infty} P\Big(|X_n - X| > \frac{\varepsilon}{2\sup_m |a_m|}\Big) + \limsup_{n\to\infty} P\Big(|X| > \underbrace{\frac{\varepsilon}{2|a_n - a|}}_{\to\infty}\Big) = 0.$$

□

Die Aussage von Korollar 4.2 gilt tatsächlich auch für den stärkeren Begriff der fast sicheren Konvergenz:

Satz 4.6 (Eine Version des starken Gesetzes der großen Zahlen)**.** *In der Situation von Kor. 4.2, i. gilt*

$$\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big) \underset{n\to\infty}{\longrightarrow} 0 \quad \textit{fast sicher}, \tag{4.7}$$

insbesondere gilt für $X_1, X_2, \cdots \in \mathcal{L}^2$ *u.i.v. mit* $\mathbb{E}[X_1] = \mu$

$$\frac{X_1 + \cdots + X_n}{n} \underset{n\to\infty}{\longrightarrow} \mu \quad \textit{fast sicher}. \tag{4.8}$$

Den Beweis betrachten wir als Ergänzung in Abschnitt 4.1.

Beispiel 4.7 (Borels[3] Gesetz über normale Zahlen, 1909)**.** Sei $U \sim \mathrm{Unif}_{[0,1]}$, X_i die i-te Ziffer in der (nicht-abbrechenden) Dezimaldarstellung von U (d.h. $U = \sum_{i=1}^{\infty} X_i\, 10^{-i}$).

Dann gilt für $q = 0, 1, \ldots, 9$:

$$\frac{1}{n}\big|\{1 \le i \le n : X_i = q\}\big| \underset{n\to\infty}{\longrightarrow} \frac{1}{10} \quad \text{fast sicher.}$$

Beweis. $X_1, X_2, \ldots$ sind u.a., $X_i \sim \mathrm{Unif}_{\{0,1,\ldots,9\}}$, also sind $\mathbf{1}_{\{X_i = q\}}$, $i = 1, 2, \ldots$ u.i.v., $\mathrm{Ber}_{1/10}$, die Beh. folgt aus Satz 4.6. □

Bericht 4.8. Das starke Gesetz der großen Zahlen gilt tatsächlich auch ohne $\mathcal{L}^2$-Annahme:

Für $X_1, X_2, \cdots \in \mathcal{L}^1$ u.i.v. mit $\mathbb{E}[X_1] = \mu$ gilt

$$\frac{X_1 + \cdots + X_n}{n} \underset{n\to\infty}{\longrightarrow} \mu \quad \text{fast sicher.}$$

4.1 Beweis von Satz 4.6*

Die Aussage von Satz 4.6 (das starke Gesetz der großen Zahlen für unkorrelierte ZVn) lautete: Sind $X_1, X_2, \cdots \in \mathcal{L}^2$ seien paarweise unkorrelierte ZVn mit

$$\sup_n \mathrm{Var}[X_n] \le \theta < \infty$$

dann gilt

$$\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big) \underset{n\to\infty}{\longrightarrow} 0 \quad \text{fast sicher} \tag{4.9}$$

Wir wollen hier als Ergänzung den Beweis betrachten.

[3] Émile Borel, 1871–1956

Beweis von Satz 4.6. Sei o.E. $\mathbb{E}[X_i] = 0$, sonst betrachte $X_i' := X_i - \mathbb{E}[X_i]$, setze

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

1. Schritt: Zeige

$$S_{n^2} \underset{n\to\infty}{\longrightarrow} 0 \text{ f.s.}$$

Für $\varepsilon > 0$ zeigt Kor. 4.2, 1.

$$P\big(|S_{n^2}| > \varepsilon\big) \le \frac{\theta}{\varepsilon^2 n^2},$$

also

$$P\big(|S_{n^2}| > \varepsilon \text{ für } \infty\text{-viele } n\big) = 0$$

(mit Borel-Cantelli-Lemma (Satz 2.30), da $\sum_n \frac{\theta}{\varepsilon^2 n^2} < \infty$) und somit

$$P\big(\{S_{n^2} \to 0\}^c\big) \le P\Bigg(\bigcup_{\varepsilon\in\mathbb{Q}\cap(0,\infty)} \big\{|S_{n^2}| > \varepsilon \text{ für } \infty\text{-viele } n\big\}\Bigg) = 0.$$

2. Schritt: Zu $m \in \mathbb{N}$ wähle $n = n(m)$ mit $n^2 \le m \le (n+1)^2$:

$$P\Big(\big|\underbrace{mS_m - n^2 S_{n^2}}_{=\sum_{n^2<i\le m} X_i}\big| > \varepsilon m\Big) \le \frac{1}{\varepsilon^2 m^2}\mathrm{Var}\Big[\sum_{n^2<i\le m} X_i\Big] \le \frac{\theta(m-n^2)}{\varepsilon^2 m^2},$$

somit

$$\begin{aligned}
&\sum_{m=1}^{\infty} P\Big(\big|mS_m - n(m)^2 S_{n(m)^2} > \varepsilon m\Big) \\
&\le \frac{\theta}{\varepsilon^2}\sum_{n=1}^{\infty}\sum_{m=n^2}^{(n+1)^2-1}\frac{m-n^2}{m^2} \le \frac{\theta}{\varepsilon^2}\sum_{n=1}^{\infty}\frac{1}{n^4}\underbrace{\sum_{j=1}^{(n+1)^2-1-n^2} j}_{=\frac{2n(2n+1)}{2}} \le \frac{\theta}{\varepsilon^2}\sum_{n=1}^{\infty}\frac{2n(2n+1)}{n^4} < \infty,
\end{aligned}$$

demnach gilt (wie im 1. Schritt)

$$P\Big(S_m - \frac{n(m)^2}{m} S_{n(m)^2} \underset{m\to\infty}{\longrightarrow} 0\Big) = 1,$$

dies zusammen mit dem 1. Schritt und Lemma 4.5 zeigt $S_m \to 0$ f.s. für $m \to \infty$.

Es ist nämlich für $\varepsilon > 0$ wegen $\frac{n(m)^2}{m} \to_{m\to\infty} 1$

$$\begin{aligned}
A_\varepsilon &:= \big\{|S_m| > \varepsilon \text{ für } \infty\text{-viele } m\big\} \\
&\subset \big\{|S_{n^2}| > \frac{\varepsilon}{4} \text{ für } \infty\text{-viele } n\big\} \cup \Big\{\Big|S_m - \frac{n(m)^2}{m} S_{n(m)^2}\Big| > \frac{\varepsilon}{2} \text{ für } \infty\text{-viele } m\Big\},
\end{aligned}$$

also $P(A_\varepsilon) = 0$ und damit auch $P\Big(\bigcap_{\varepsilon\in\mathbb{Q},\varepsilon>0} A_\varepsilon\Big) = 0$. □

Kapitel 5

Zentraler Grenzwertsatz

Vorbemerkung. Seien $X_1, X_2, \dots$ u.i.v., $\mathbb{E}[X_1] = \mu$, $\mathrm{Var}[X_1] = \sigma^2 < \infty$.

Wir haben gesehen, dass $X_1 + \cdots + X_n \approx n\mu$ mit hoher Wahrscheinlichkeit, denn

$$\frac{X_1 + \cdots + X_n}{n} - \mu \underset{n\to\infty}{\longrightarrow} 0 \quad \text{f.s.}$$

gemäß dem starken Gesetz der großen Zahlen (Satz 4.6), aber feiner gefragt:

Wie groß ist $X_1 + \cdots + X_n - n\mu$ typischerweise?

Für $A \gg \sqrt{n}$ ist (mit Chebyshev-Ungleichung, Satz 4.1) zumindest

$$P\big(|X_1 + \cdots + X_n - n\mu| > A\big) \le \frac{n\sigma^2}{A^2} \quad \text{(sehr) klein.}$$

Um einzusehen, dass $\sqrt{n}$ die korrekte Größenordnung der typischen Abweichungen von $X_1+\cdots+X_n$ vom $n\mu$ ist, betrachten wir

$$X_i \sim \mathrm{Ber}_p, \quad \text{mit einem } p \in (0,1) \qquad \text{(der „einfachste Fall“)},$$

dann ist

$$Z_n := X_1 + \cdots + X_n \sim \mathrm{Bin}_{n,p}$$

Erinnerung (Stirling-Approximation[1])**.** Es gilt

$$n! = \sqrt{2\pi} n^{n+1/2} e^{-n}\, e^{\rho(n)} \quad \text{mit } 0 < \rho(n) < \tfrac{1}{12n}$$

Dies findet sich in vielen Analysis-Lehrbüchern oder auch in W. Feller, An introduction to probability and its applications, Vol. I, Wiley, 1968, Kap. II.9.

[1] James Stirling, 1692–1770

Numerische Beispiele zur Güte der Stirling-Approximation

n	$n!$	$\sqrt{2\pi}n^{n+1/2}e^{-n}$	$1-\frac{\sqrt{2\pi}n^{n+1/2}e^{-n}}{n!}$
3	6	5,836	0,027
4	24	23,506	0,021
5	120	118,019	0,016
6	720	710,078	0,014
7	5.040	4980,396	0,012
8	40.320	39902,395	0,011
9	362.880	359.536,873	0,0088
10	3.628.800	3.598.695,619	0,0078
15	1.307.674.368.000	1.300.430.722.199,47	0,0054

Sei $C > 0$ fest, $p \in (0,1)$, betrachte

$$n, k \in \mathbb{N} \text{ mit } |k - np| \le C\sqrt{n}.$$

Es ist

$$\begin{aligned} P(Z_n = k) = \mathrm{Bin}_{n,p}(\{k\}) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n\frac{k}{n}\left(1-\frac{k}{n}\right)}} \left(\left(p\frac{n}{k}\right)^{k/n} \left((1-p)\frac{n}{n-k}\right)^{(n-k)/n} \right)^n \cdot K(n,k) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n\frac{k}{n}\left(1-\frac{k}{n}\right)}} \exp\Big(-nh\Big(\frac{k}{n}\Big)\Big) \cdot K(n,k) \end{aligned}$$

mit

$$h(t) := t\log\left(\tfrac{t}{p}\right) + (1-t)\log\left(\tfrac{1-t}{1-p}\right), \quad t \in [0,1]$$

und Korrekturfaktor

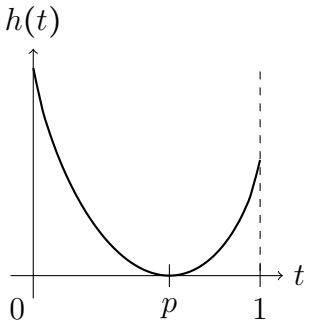
$$K(n,k) := \exp\big(\rho(n) - \rho(k) - \rho(n-k)\big) \quad \Big(= \big(1 + o(1)\big) \text{ für } n, k \to \infty \Big)$$

denn mit Stirling-Approximation ist

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{1}{\sqrt{2\pi}} \Big(\frac{n}{k(n-k)} \Big)^{1/2} \frac{n^n}{k^k (n-k)^{n-k}} \cdot e^{\rho(n)-\rho(k)-\rho(n-k)}$$

Es ist

$$\begin{aligned} h(p) = 0, \quad \text{für } t \in (0,1) \text{ ist } \; h'(t) &= \log\left(\tfrac{t}{p}\right) - \log\left(\tfrac{1-t}{1-p}\right), \\ h''(t) &= \frac{1}{t(1-t)} \; \big(\ge 0 \big), \end{aligned}$$



Taylorentwicklung von h im Punkt p liefert ($h(p) = h'(p) = 0$)

$$h\Big(\tfrac{k}{n}\Big) = \frac{1}{2}\frac{1}{p(1-p)}\Big(\tfrac{k}{n} - p\Big)^2 + O\Big(\big(\tfrac{k-np}{n}\big)^3\Big) = \frac{1}{2}\frac{1}{p(1-p)}\Big(\tfrac{k}{n} - p\Big)^2 + O\Big(\frac{1}{n^{3/2}}\Big),$$

weiter ist

$$\frac{1}{\sqrt{n\frac{k}{n}\big(1-\frac{k}{n}\big)}} = \frac{1}{\sqrt{np(1-p)}}\big(1 + o(1)\big)$$

(jeweils für $k, n \to \infty$ so, dass $|k - np| \le C\sqrt{n}$ gilt). Insgesamt:

$$\mathrm{Bin}_{n,p}(\{k\}) = \frac{1}{\sqrt{2\pi np(1-p)}} \exp\Big(-\frac{n}{2p(1-p)}\Big(\frac{k}{n} - p\Big)^2\Big) \cdot \widetilde{K}(n,k)$$

mit einem Korrekturfaktor $\widetilde{K}(n,k)$, der

$$\lim_{n\to\infty} \max_{k\,:\,|k-np|\le C\sqrt{n}} \big|\widetilde{K}(n,k) - 1\big| = 0$$

erfüllt.

Zusammenfassend haben wir bewiesen:

Satz 5.1 (Satz von de Moivre-Laplace[2], lokale Normalapproximation der Binomialverteilung)**.** *Sei* $C > 0$, $p \in (0,1)$,

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\big(-\tfrac{1}{2}z^2\big)$$

die Dichte der Standard-Normalverteilung. Mit

$$z_n(k) := \frac{k - np}{\sqrt{np(1-p)}}$$

gilt

$$\lim_{n\to\infty} \max_{k\,:\,|k-np|\le C\sqrt{n}} \left| \frac{\mathrm{Bin}_{n,p}(\{k\})}{\frac{1}{\sqrt{np(1-p)}}\varphi\big(z_n(k)\big)} - 1 \right| = 0.$$

[2] Abraham de Moivre, 1667–1754; Pierre-Simon Laplace, 1749–1827

Korollar 5.2 (Normalapproximation der Binomialverteilung)**.** *Sei* $p \in (0,1)$, $Z_n \sim \mathrm{Bin}_{n,p}$, *setze*

$$Z_n^* := \frac{Z_n - np}{\sqrt{np(1-p)}}.$$

Dann gilt für $-\infty \le a < b \le \infty$

$$\lim_{n\to\infty} P(a \le Z_n^* \le b) = \int_a^b \varphi(z)\,dz.$$

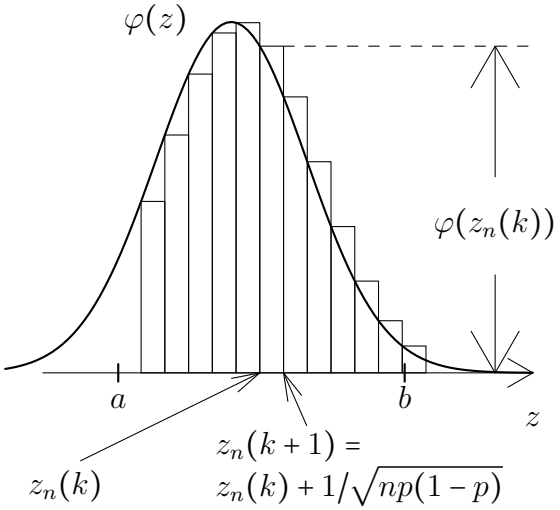
Beweis. Betrachte zunächst $-\infty < a < b < \infty$:

$$P(a \le Z_n^* \le b) = \Bigg(\sum_{k\,:\, z_n(k)\in[a,b]} \frac{1}{\sqrt{np(1-p)}} \varphi\big(z_n(k)\big) \Bigg) \big(1 + o(1)\big)$$

und

$$\sum_{k\,:\, z_n(k)\in[a,b]} \frac{1}{\sqrt{np(1-p)}} \varphi\big(z_n(k)\big) = \sum_{k=\lceil np+a\sqrt{np(1-p)}\rceil}^{k=\lfloor np+b\sqrt{np(1-p)}\rfloor} \frac{1}{\sqrt{np(1-p)}} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\Big(\frac{k-np}{\sqrt{np(1-p)}} \Big)^2 \Big)$$
$$\xrightarrow[n\to\infty]{} \int_a^b \varphi(z)\,dz$$

(die Summe stellt eine Riemann-Approximation des Integrals dar, siehe auch folgende Skizze).



Sei nun $a = -\infty < b < \infty$:
Zu $\varepsilon > 0$ wähle $a' < 0$ so, dass für $n \in \mathbb{N}$

$$P(Z_n^* < a') \le P(|Z_n^*| > |a'|) \le \frac{1}{(a')^2} < \frac{\varepsilon}{4}$$

(mit Chebychev-Ungleichung, Satz 4.1) und

$$\int_{-\infty}^{a'} \varphi(z)\,dz < \frac{\varepsilon}{4}$$

gilt. Dann ist für n genügend groß

$$\left|P(Z_n^* \le b) - \int_{-\infty}^{b} \varphi(z)\,dz\right| \le \left|P(a' \le Z_n^* \le b) - \int_{a'}^{b} \varphi(z)\,dz\right| + \left|P(Z_n^* < a')\right| + \int_{-\infty}^{a'} \varphi(z)\,dz$$
$$\le \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon.$$

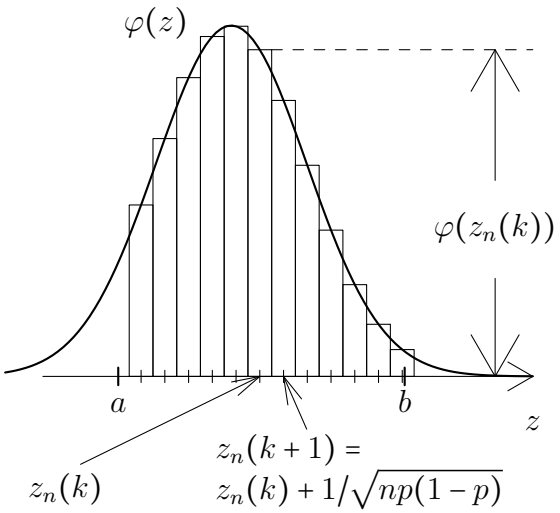
Der Fall $b = \infty$ kann analog behandelt werden (Übung). □

Bemerkung 5.3 (Stetigkeitskorrektur)**.** Für numerische Approximationen betrachtet man oft

$$\mathrm{Bin}_{n,p}\big(\{k, k+1, \dots, \ell\}\big) \approx \Phi\Big(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\Big) - \Phi\Big(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\Big)$$

für $0 \le k \le \ell \le n$ (mit $\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} e^{-z^2/2}\,dz$ der Verteilungsfunktion der Standardnormalverteilung).

Man betrachtet also Histogramm-Balken wie im Beweis von Kor. 5.2, die aber jeweils an $z_n(k)$ „zentriert“ sind (siehe auch Skizze unten).



Die numerische Approximation ist (speziell für eher kleine Werte von n) meist besser mit „Stetigkeitskorrektur“, Beispiele:

n	p	k	ℓ	$Bin_{n,p}(\{k, k+1, \dots, \ell\})$	Approx. ohne	Approx. mit
15	0,3	5	9	0,4809	0,3835	0,4976
50	0,3	9	15	0,5509	0,4680	0,5389
1000	0,3	250	290	0,2567	0,2448	0,2558

Hierbei gibt „Approx. ohne“ jeweils den Wert $\Phi\Big(\frac{\ell - np}{\sqrt{np(1-p)}}\Big) - \Phi\Big(\frac{k - np}{\sqrt{np(1-p)}}\Big)$ und „Approx. mit“ den Wert $\Phi\Big(\frac{\ell + 0{,}5 - np}{\sqrt{np(1-p)}}\Big) - \Phi\Big(\frac{k - 0{,}5 - np}{\sqrt{np(1-p)}}\Big)$ an (jeweils auf 4 Nachkommastellen gerundet).

Bericht und Definition 5.4. Der Sachverhalt aus Korollar 5.2 wird auch ausgesprochen als „Konvergenz in Verteilung“:

$$X_n \underset{n\to\infty}{\longrightarrow} X \text{ in Verteilung} \quad \big(\text{auch } X_n \xrightarrow[n\to\infty]{d} X \text{ oder } X_n \xrightarrow[n\to\infty]{\mathscr{L}} X \text{ geschrieben}\big),$$

wenn gilt

$$\lim_{n\to\infty} P(X_n \le x) = P(X \le x) \ \big(= F_X(x)\big)$$

für jedes $x \in \mathbb{R}$, an dem F_X stetig ist.

Beispiel („Macht entschlossener Minderheiten“)**.** An einer Wahl zwischen Vorschlag A und Vorschlag B nehmen 100.000 Wähler teil. Darunter sind 300, die fest entschlossen sind, für Vorschlag A zu stimmen; die übrigen sind unentschlossen und entscheiden sich (in unserem Modell) unabhängig per fairem Münzwurf zwischen den beiden Vorschlägen. Wie wahrscheinlich ist es, dass Vorschlag A die Mehrheit erhält?

Sei Y die Anzahl unentschlossener Wähler, die für A stimmt, n. Vor. ist $Y \sim \mathrm{Bin}_{n,1/2}$ mit $n = 100.000 - 300 = 99.700$ und Vorschlag A erhält $300 + Y$ Stimmen.

$$\begin{aligned} P(300 + Y \ge 50.001) &= P\Big(\frac{Y - 0{,}5n}{\sqrt{n \cdot 0{,}5 \cdot 0{,}5}} \ge \frac{49.701 - 0{,}5n}{\sqrt{n \cdot 0{,}5 \cdot 0{,}5}}\Big) = P\Big(\frac{Y - 0{,}5n}{\sqrt{n \cdot 0{,}5 \cdot 0{,}5}} \ge -0.9438\Big) \\ &\approx \int_{-0.9438}^{\infty} \varphi(z)\, dz \approx 0{,}827 \end{aligned}$$

Satz 5.5 („Zentraler Grenzwertsatz“)**.** *Seien $X_1, X_2, \ldots$ u.i.v. reelle ZVn $\in \mathscr{L}^2$ mit $\mathrm{Var}[X_1] \in (0, \infty)$, dann gilt für $-\infty \le a < b \le \infty$*

$$\lim_{n\to\infty} P\Big(a \le \frac{X_1 + \cdots + X_n - n\mathbb{E}[X_1]}{\sqrt{n\mathrm{Var}[X_1]}} \le b\Big) = P(a \le Z \le b) \quad \textit{mit } Z \sim \mathcal{N}_{0,1}$$

Bemerkung. Korollar 5.2 ist ein Spezialfall dieses Satzes, indem man $Z_n \sim \mathrm{Bin}_{n,p}$ darstellt als $Z_n = X_1 + X_2 + \cdots + X_n$ mit X_i u.i.v., $X_1 \sim \mathrm{Ber}_p$.

Beobachtung. Die Aussage von Satz 5.5 gilt trivialerweise (sogar für festes $n \in \mathbb{N}$), wenn $X_i \sim \mathcal{N}_{0,1}$, da dann

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \sim \mathcal{N}_{0,1}$$

vgl. Beispiel 2.28.

Die Beweisidee für Satz 5.5 ist den allgemeinen Fall auf diesen Spezialfall zurückzuführen. Wir betrachten als Ergänzung in Abschnitt 5.1 den Beweis.

5.1 Beweis von Satz 5.5*

Seien $X_1, X_2, \dots$ u.i.v. reelle ZVn $\in \mathscr{L}^2$ mit $\mathrm{Var}[X_1] \in (0, \infty)$. Wir möchten zeigen, dass für $-\infty \le a < b \le \infty$ gilt

$$\lim_{n\to\infty} P\Big(a \le \frac{X_1 + \dots + X_n - n\mathbb{E}[X_1]}{\sqrt{n\mathrm{Var}[X_1]}} \le b\Big) = P(a \le Z \le b) \quad \text{mit } Z \sim \mathcal{N}_{0,1}$$

Wir können o.E. $\mathbb{E}[X_1] = 0$, $\mathrm{Var}[X_1] = 1$ annehmen, ansonsten betrachten wir

$$\widetilde{X}_i := \frac{X_i - \mathbb{E}[X_1]}{\sqrt{\mathrm{Var}[X_1]}}.$$

Lemma 5.6. *$X_1, X_2, \dots$ u.i.v. reelle ZVn, $X_i \in \mathcal{L}^2$ mit $\mathbb{E}[X_i] = 0$, $\mathrm{Var}[X_i] = 1$, $f : \mathbb{R} \to \mathbb{R}$ dreimal stetig differenzierbar, die ersten drei Ableitungen seien gleichmäßig beschränkt. Dann gilt*

$$\lim_{n\to\infty} \mathbb{E}\Big[f\Big(\frac{X_1 + \dots + X_n}{\sqrt{n}}\Big)\Big] = \mathbb{E}\big[f(Z)\big]$$

mit $Z \sim \mathcal{N}_{0,1}$.

Beweis. Seien $Z_1, Z_2, \dots$ u.i.v., $\sim \mathcal{N}_{0,1}$, unabhängig von den X_i, schreibe

$$\begin{aligned} &f\Big(\frac{X_1 + \dots + X_n}{\sqrt{n}}\Big) - f\Big(\frac{Z_1 + \dots + Z_n}{\sqrt{n}}\Big) \\ &= \sum_{i=1}^{n} \left(f\Big(W_{i,n} + \frac{X_i}{\sqrt{n}}\Big) - f\Big(W_{i,n} + \frac{Z_i}{\sqrt{n}}\Big)\right) \end{aligned} \tag{5.1}$$

mit

$$W_{i,n} := \frac{1}{\sqrt{n}}\Big(X_1 + X_2 + \dots + X_{i-1} + Z_{i+1} + Z_{i+1} + \dots + Z_n\Big).$$

Taylor-Enwicklung (im Punkt $W_{i,n}$) liefert

$$\begin{aligned} &f\Big(W_{i,n} + \frac{X_i}{\sqrt{n}}\Big) - f\Big(W_{i,n} + \frac{Z_i}{\sqrt{n}}\Big) \\ &= f'\big(W_{i,n}\big)\frac{X_i - Z_i}{\sqrt{n}} + \frac{1}{2} f''\big(W_{i,n}\big)\frac{X_i^2 - Z_i^2}{n} + R_{i,n} \end{aligned}$$

mit

$$\big|R_{i,n}\big| \le \big|f''\big(W_{i,n} + \vartheta_{i,n}\big) - f''\big(W_{i,n}\big)\big|\frac{X_i^2}{2n} + \big|f''\big(W_{i,n} + \widetilde{\vartheta}_{i,n}\big) - f''\big(W_{i,n}\big)\big|\frac{Z_i^2}{2n}$$

wobei $|\vartheta_{i,n}| \le \dfrac{|X_i|}{\sqrt{n}}$, $|\widetilde{\vartheta}_{i,n}| \le \dfrac{|Z_i|}{\sqrt{n}}$.

Mit $C_2 := \sup_{x\in\mathbb{R}} |f''(x)|$, $C_3 := \sup_{x\in\mathbb{R}} |f'''(x)|$ gilt für jedes $K > 0$:

$$|R_{i,n}| \le C_3 \frac{K^3}{2n^{3/2}} \mathbf{1}_{\{|X_i|\le K\}} + C_2 \frac{X_i^2}{n} \mathbf{1}_{\{|X_i|>K\}} + C_3 \frac{|Z_i|^3}{n^{3/2}}.$$

Nehme Erwartungswert in (5.1):

$$\begin{aligned}
&\left| \mathbb{E}\Big[f\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}} \Big) \Big] - \mathbb{E}\big[f(Z) \big] \right| \\
&= \left| \sum_{i=1}^n \Bigg(\underbrace{\mathbb{E}\Big[f'(W_{i,n}) \frac{X_i - Z_i}{\sqrt{n}} \Big]}_{=\mathbb{E}[f'(W_{i,n})]\mathbb{E}[X_i - Z_i]/\sqrt{n} = 0} + \underbrace{\mathbb{E}\Big[\frac{1}{2} f''(W_{i,n}) \frac{X_i^2 - Z_i^2}{n} \Big]}_{=\mathbb{E}[f''(W_{i,n})]\mathbb{E}[X_i^2 - Z_i^2]/(2n) = 0} + \mathbb{E}\big[R_{i,n} \big] \Bigg) \right| \\
&\le \sum_{i=1}^n \mathbb{E}\big[|R_{i,n}| \big] \le n \Big(C_3 \frac{K^3}{2n^{3/2}} + \frac{C_2}{n} \mathbb{E}\big[X_1^2 \mathbf{1}_{\{|X_1|>K\}} \big] + \frac{C_3}{n^{3/2}} \mathbb{E}\big[|Z_1|^3 \big] \Big),
\end{aligned}$$

also

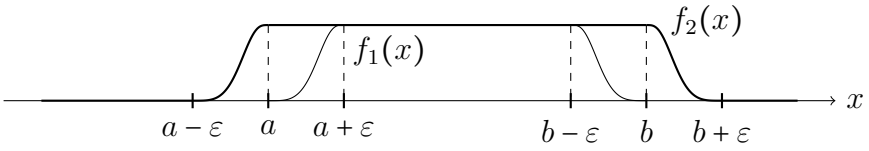
$$\limsup_{n\to\infty} \left| \mathbb{E}\Big[f\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}} \Big) \Big] - \mathbb{E}\big[f(Z) \big] \right| \le C_2 \mathbb{E}\big[X_1^2 \mathbf{1}_{\{|X_1|>K\}} \big] \underset{K\to\infty}{\longrightarrow} 0$$

□

Beweis von Satz 5.5. Seien $-\infty < a < b < \infty$. Zu $0 < \varepsilon < \frac{b-a}{2}$ wähle f_1, f_2, die den Voraussetzungen von Lemma 5.6 genügen und

$$\mathbf{1}_{[a+\varepsilon, b-\varepsilon]} \le f_1 \le \mathbf{1}_{[a,b]} \le f_2 \le \mathbf{1}_{[a-\varepsilon, b+\varepsilon]}$$

erfüllen (siehe Skizze).



Es ist

$$\begin{aligned}
P\big(Z \in [a+\varepsilon, b-\varepsilon] \big) &\le \mathbb{E}\big[f_1(Z) \big] = \lim_{n\to\infty} \mathbb{E}\Big[f_1\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}} \Big) \Big] \\
&\le \liminf_{n\to\infty} P\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}} \in [a,b] \Big) \\
&\le \limsup_{n\to\infty} P\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}} \in [a,b] \Big) \\
&\le \lim_{n\to\infty} \mathbb{E}\Big[f_2\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}} \Big) \Big] = \mathbb{E}\big[f_2(Z) \big] \le P\big(Z \in [a-\varepsilon, b+\varepsilon] \big),
\end{aligned}$$

mit $\varepsilon \downarrow 0$ folgt die Behauptung.

Die Fälle $a = -\infty$ oder $b = +\infty$ kann man analog behandeln. □

Kapitel 6

Ideen und Begriffe aus der Statistik

6.1 Zur deskriptiven Statistik

In diesem Kapitel der Vorlesung werden Ideen und Verfahren der beschreibenden Statistik kurz diskutiert, vergleiche die Folien `Folien_deskriptive_Statistik.pdf` auf der Homepage der Vorlesung; Themen dort: Histogramme, Dichtepolygone, Stripcharts, Boxplots, statistische Kenngrößen: Lageparameter (empirischer Mittelwert, Median, Quartile) und Streuungsparameter (empirische Streuung, Quartilsabstand)

6.2 Grundlegende Begriffe, Schätzen von Parametern

Beispiel 6.1. Bei einer biologischen Expedition wurden $n = 53$ Krebse einer gewissen Art gefangen, davon $k = 23$ Weibchen.

Was sagt uns dies über den Weibchenanteil in der Population?

Gibt die Beobachtung (Weibchenanteil in der Stichprobe $\frac{23}{53} \approxeq 0{,}434$) Anlass, an einem ausgeglichenen Geschlechterverhältnis in dieser Population zu zweifeln?

Vorstellung: Eine sehr große Population von Krebsen mit (uns unbekanntem) Weibchenanteil $\vartheta \in (0,1)$,
der naheliegendste Schätzwert für ϑ (angesichts der Beobachtungen) ist

$$\widehat{\vartheta} = \frac{23}{53} \ (\approxeq 0{,}434).$$

Modell: $X = (X_1, X_2, \ldots, X_n)$,

$$X_i = \begin{cases} 1, & i\text{-ter gefangener Krebs ist Weibchen,} \\ 0, & i\text{-ter gefangener Krebs ist Männchen} \end{cases}$$

X_i sind u.a., $\sim \mathrm{Ber}_\vartheta$ (Wir tun hier so, als ob wir mit Zurücklegen rein zufällig aus der Gesamtpopulation gezogen hätten – diese Approximation ist für große Populationen gerechtfertigt.)

Wir interpretieren $\frac{k}{n} = \frac{23}{53}$ als Realisierung der Zufallsvariable

$$\widehat{\vartheta} := \frac{1}{n}\big(X_1 + \cdots + X_n\big)$$

1. (Punktschätzung) Was auch immer ϑ ist, es gilt

$$\mathbb{E}_\vartheta\big[\widehat{\vartheta}\big] = \vartheta,$$
$$\mathrm{Var}_\vartheta\big[\widehat{\vartheta}\big] = \frac{1}{n^2} n\vartheta(1-\vartheta) = \frac{1}{n}\vartheta(1-\vartheta) = \frac{\sigma^2}{n}$$

mit $\sigma = \sigma(\vartheta) = \sqrt{\vartheta(1-\vartheta)}$. ($\mathbb{E}_\vartheta$, etc. bezieht sich auf Erwartungswerte bezüglich dem Wahrscheinlichkeitsmaß, unter dem $X_i \sim \mathrm{Ber}_\vartheta$ und u.a. sind.)

$$\widehat{\sigma} := \sqrt{\widehat{\vartheta}(1-\widehat{\vartheta})}$$

ist ein naheliegender Schätzer für σ.

Ein Schätzer für die Standardabweichung von $\widehat{\vartheta}$ ist $\dfrac{\widehat{\sigma}}{\sqrt{n}}$.

Im Statistik-Jargon heißt dies auch der „Standardfehler“ (englisch: standard error [of the mean], SEM), dies ist eine naheliegende Maßzahl für die „Genauigkeit der Schätzung“.

(Im Beispiel:

$$\frac{23}{53}\cdot\frac{30}{53} \approxeq 0{,}246, \quad \widehat{\sigma} \approxeq 0{,}496, \quad \frac{\widehat{\sigma}}{\sqrt{53}} \approxeq 0{,}0681$$

man gibt also an: geschätzter Weibchenanteil $0{,}43 \pm 0{,}068$.)

2. („Wie genau ist die Schätzung?“: Konfidenzintervall) Wenn der wahre Weibchenanteil ϑ ist, so ist $X_1 + \cdots + X_n \sim \mathrm{Bin}_{n,\vartheta}$, also

$$\frac{\widehat{\vartheta}-\vartheta}{\widehat{\sigma}/\sqrt{n}} \approx \frac{\widehat{\vartheta}-\vartheta}{\sigma(\vartheta)/\sqrt{n}} \overset{d}{\approx} \mathcal{N}_{0,1}$$

gemäß dem Satz von de Moivre-Laplace (Satz 5.1 und Korollar 5.2), also

$$P_\vartheta\Big(-1{,}96 \le \frac{\widehat{\vartheta}-\vartheta}{\widehat{\sigma}/\sqrt{n}} \le 1{,}96\Big) \approx P(-1{,}96 \le Z \le 1{,}96) \approx 0{,}95$$

mit $Z \sim \mathcal{N}_{0,1}$, d.h. das (zufällige) Intervall

$$I := \Big[\widehat{\vartheta} - 1{,}96\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + 1{,}96\frac{\widehat{\sigma}}{\sqrt{n}}\Big] \quad \text{erfüllt } P_\vartheta(\vartheta \in I) \approx 0{,}95$$

(für jede Wahl von $\vartheta \in [0,1]$).

I heißt ein Konfidenzintervall (für ϑ) zum (approximativen) Niveau $0{,}95$

(Im Beispiel: $I \approx [0{,}30, 0{,}57]$)

3. (Testen von Hypothesen) Passen die Beobachtungen zur Hypothese, dass der wahre Weibchenanteil in der Population $\vartheta_0 = \frac{1}{2}$ ist?

Wir beobachten

$$\Big|\widehat{\vartheta} - \frac{1}{2}\Big| = \Big|\frac{23}{53} - \frac{1}{2}\Big| \approxeq 0{,}07,$$

es ist

$$P_{\vartheta_0}\Big(\big|\widehat{\vartheta} - \vartheta_0\big| \ge 0{,}07\Big) \approx P\big(|Z| \ge \sqrt{4\cdot n}\cdot 0{,}07\big) \approxeq 2\big(1 - \Phi(0{,}96)\big) \approxeq 0{,}336.$$

Demnach: Wenn die Hypothese $\vartheta = \vartheta_0 = \frac{1}{2}$ zutrifft, würden wir in ca. 1/3 der Fälle eine mindestens so große Abweichung $\big|\widehat{\vartheta} - \frac{1}{2}\big|$ wie die tatsächlich anhand der Daten beobachtete finden. Insoweit gibt die Beobachtung keinen Anlass, diese Hypothese anzuzweifeln.

Definition 6.2. Ein *statistisches Modell* ist ein Tripel $(\mathcal{M} =) \big(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta}\big)$, wo $\mathcal{X} \neq \varnothing$ Menge („Beobachtungs- oder Stichprobenraum"), $\mathscr{F} \subset 2^{\mathcal{X}}$ eine σ-Algebra, Θ eine Menge (mit $|\Theta| > 1$) und für jedes $\vartheta \in \Theta$ ist P_ϑ ein W'maß auf $(\mathcal{X}, \mathscr{F})$.

Das Modell $\mathcal{M}$ heißt *parametrisch*, wenn $\Theta \subset \mathbb{R}^d$ für ein $d \in \mathbb{N}$, speziell *einparametrisch*, wenn $d = 1$.

$\mathcal{M}$ heißt *diskret*, wenn $\mathcal{X}$ abzählbar ist, $\mathcal{M}$ heißt *stetig*, wenn $\mathcal{X} \subset \mathbb{R}^n$ und jedes P_ϑ eine Dichte $\rho_\vartheta : \mathcal{X} \to [0, \infty]$ besitzt.

Ein diskretes oder stetiges Modell heißt ein *Standardmodell*.

$\big(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (P_\vartheta^{\otimes n})_{\vartheta\in\Theta}\big)$ heißt das n-fache Produktmodell von $\mathcal{M}$ (für Produktmaße vgl. Def. 2.21).

Definition 6.3. $\big(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta}\big)$ statistisches Modell, $(S, \mathscr{A})$ messbarer Raum.

1. Eine Zufallsvariable X (definiert auf $(\mathcal{X}, \mathscr{F})$ mit Werten in S, d.h. $X : \mathcal{X} \to S$ ist $\mathscr{F}$-$\mathscr{A}$-messbar) heißt eine *Statistik* (manchmal auch: „Stichprobe").

2. Sei $\tau : \Theta \to \mathbb{R}$ eine reelle Kenngröße (oder „Parametermerkmal"), eine Statistik $T : \mathcal{X} \to \mathbb{R}$ heißt ein *Schätzer* (genauer: „Punktschätzer") für τ.

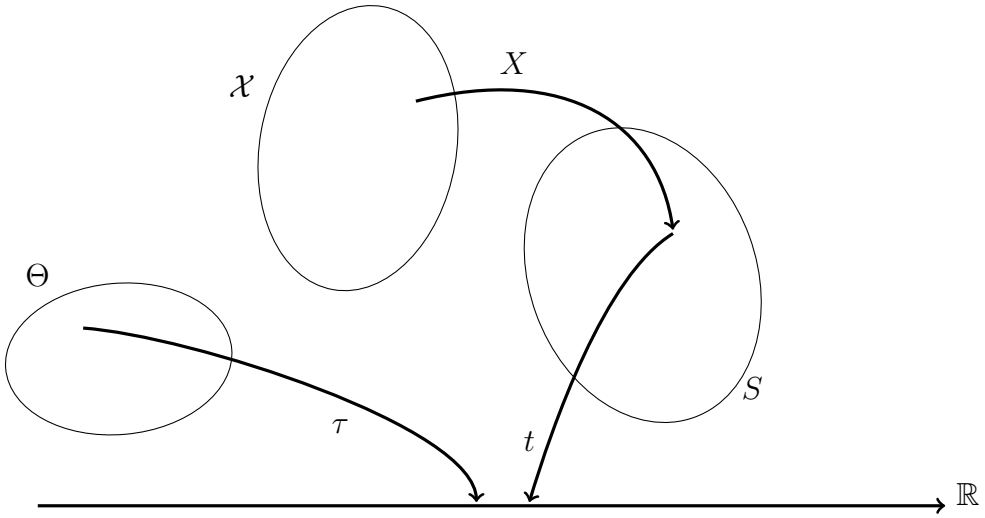
3. Ein Schätzer T für τ heißt *erwartungstreu* (oder „unverzerrt"), wenn gilt

$$\forall\, \vartheta \in \Theta : \quad \mathbb{E}_\vartheta[T] = \tau(\vartheta).$$

$b_\vartheta(T) := \mathbb{E}_\vartheta[T] - \tau(\vartheta)$ heißt die *Verzerrung* (englisch: bias) von T.

Die typische Konstruktion / Situation eines Schätzers ist $T = t(X)$ für eine Funktion $t : S \to \mathbb{R}$.

Man schreibt / benennt einen Schätzer für τ oft $\widehat{\tau}$.



Schematische Darstellung eines Schätzers $T = t(X)$ für τ

Das Startbeispiel 6.1 in der Formalisierung von Definition 6.2 und 6.3 ausgedrückt:

- Statistisches Modell: $\mathcal{X} = \{0, 1\}^n$, $\mathscr{F} = 2^{\mathcal{X}}$, $\Theta = [0, 1]$, $P_\vartheta = \mathrm{Ber}_\vartheta^{\otimes n}$

- Parametermerkmal: $\tau : \Theta \to \mathbb{R}$, $\tau(\vartheta) = \vartheta$
- Statistik: $S = \mathcal{X}$, $X = \mathrm{Id}_{\mathcal{X}}$
- Schätzer: $T : \mathcal{X} \to \mathbb{R}$, $T\big((x_1, \dots, x_n)\big) = \dfrac{x_1 + \cdots + x_n}{n}$

Dies ist ein Standardmodell im Sinne von Def. 6.2, es gilt $\mathbb{E}_\vartheta[T] = \vartheta = \tau(\vartheta)$ für alle $\vartheta \in [0,1]$, d.h. T ist hier ein erwartungstreuer Schätzer für τ.

Beispiel 6.4 (Erwartungstreue Schätzer für Mittelwert und Varianz im Produktmodell)**.** Für $\vartheta \in \Theta$ sei Q_ϑ ein W'maß auf $\mathbb{R}$ mit endlichem Mittelwert

$$m(\vartheta) := \int_{\mathbb{R}} x\, Q_\vartheta(dx)$$

und endlicher Varianz

$$v(\vartheta) := \int_{\mathbb{R}} \big(x - m(\vartheta)\big)^2\, Q_\vartheta(dx).$$

Unter P_ϑ seien $X_1, \dots, X_n$ u.i.v., $X_i \sim Q_\vartheta$.

(In der Formalisierung von Definition 6.2 und 6.3 könnten wir wählen: $\mathcal{M} = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (P_\vartheta)_{\vartheta\in\Theta})$ mit $P_\vartheta = Q_\vartheta^{\otimes n}$ für $\vartheta \in \Theta$, als Statistik betrachten wir $X = (X_1, \dots, X_n)$ mit $X_i : \mathbb{R}^n \to \mathbb{R}$ die Projektion auf die i-te Koordinate.

Bemerke: dies ist u.U. kein parametrisches Modell, man könnte z.B.

$$\Theta := \Big\{Q : Q \text{ ist W'maß auf } \mathbb{R} \text{ mit } \textstyle\int_{\mathbb{R}} x^2\, Q(dx) < \infty\Big\}$$

wählen.)

Dann ist

$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i \quad \text{ein erwartungstreuer Schätzer für } m(\vartheta),$$

$$S^2 := \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}\big)^2 \quad \text{ein erwartungstreuer Schätzer für } v(\vartheta).$$

(In diesem Kontext heißt $\overline{X}$ auch der empirische Mittelwert oder Stichprobenmittelwert, S^2 die korrigierte Stichprobenvarianz, vgl. auch die Diskussion in Kapitel 6.1 über deskriptive Statistik.)

Für $\vartheta \in \Theta$ gilt nämlich

$$\mathbb{E}_\vartheta[\overline{X}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\vartheta[X_i] = \frac{1}{n}\cdot n\, m(\vartheta) = m(\vartheta),$$

$$\begin{aligned}\mathbb{E}_\vartheta\Big[\sum_{i=1}^n \big(X_i - \overline{X}\big)^2\Big] &= n\mathbb{E}_\vartheta\big[(X_i - \overline{X})^2\big] = n\mathrm{Var}_\vartheta\big[X_i - \overline{X}\big] \\ &= n\mathrm{Var}_\vartheta\Big[\frac{n-1}{n}X_1 - \frac{1}{n}\sum_{i=2}^n X_i\Big] = n\Big(\big(\tfrac{n-1}{n}\big)^2\mathrm{Var}_\vartheta[X_1] + \tfrac{n-1}{n^2}\mathrm{Var}_\vartheta[X_1]\Big) \\ &= (n-1)\mathrm{Var}_\vartheta[X_1],\end{aligned}$$

also

$$\mathbb{E}_\vartheta\big[S^2\big] = \frac{1}{n-1}\mathbb{E}_\vartheta\Big[\sum_{i=1}^n \big(X_i - \overline{X}\big)^2\Big] = v(\vartheta).$$

Beobachtung und Definition 6.5. Betrachten wir in der Situation von Beispiel 6.4 die Stichprobengröße n als variabel (formal: wir gehen zum unendlichen Produktmodell $\mathcal{M} = (\mathbb{R}^\infty, \mathcal{B}^{\otimes\infty}, (Q_\vartheta^{\otimes\infty})_{\vartheta\in\Theta})$ über (vgl. Beobachtung und Bericht 2.23), mit $X_i : \mathbb{R}^\infty \to \mathbb{R}$ Projektion auf i-te Koordinate).

Dann gilt für jedes $\vartheta \in \Theta$

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \underset{n\to\infty}{\longrightarrow} m(\vartheta) \quad \text{stochastisch bzgl. } P_\vartheta \quad \text{und}$$

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2 \underset{n\to\infty}{\longrightarrow} v(\vartheta) \quad \text{stochastisch bzgl. } P_\vartheta.$$

Man sagt: Diese (Folgen von) Schätzer(n) sind *konsistent*.

Für $\overline{X}_n$ folgt dies direkt aus dem Gesetz der großen Zahlen (siehe Korollar 4.2), weiterhin ist

$$\frac{n-1}{n}S_n^2 = \frac{1}{n}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2 = \Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) - 2\Big(\frac{1}{n}\sum_{i=1}^n X_i\overline{X}_n\Big) + \big(\overline{X}_n\big)^2 = \Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) - \big(\overline{X}_n\big)^2$$
$$\underset{n\to\infty}{\overset{P_\vartheta}{\longrightarrow}} \int_{\mathbb{R}} x^2\, Q_\vartheta(dx) - \big(m(\vartheta)\big)^2 = v(\vartheta)$$

gemäß dem Gesetz der großen Zahlen (siehe Bericht 4.8) zusammen mit Lemma 4.5 und obigem, wegen $\frac{n-1}{n} \to 1$ folgt mit Lemma 4.5 die Behauptung.

Bericht. Tatsächlich gilt sogar

$$\overline{X}_n \underset{n\to\infty}{\longrightarrow} m(\vartheta) \quad P_\vartheta\text{-f.s.} \qquad \text{und} \qquad S_n^2 \underset{n\to\infty}{\longrightarrow} v(\vartheta) \quad P_\vartheta\text{-f.s.}$$

(diese Schätzer sind auch „stark konsistent“).

Für $\overline{X}_n$ folgt dies unter den Voraussetzungen von Beispiel 6.4 aus der Version des starken Gesetzes der großen Zahlen, die wir in Satz 4.6 bewiesen haben, für S_n^2 folgt dies aus Bericht 4.8.

6.2.1 Maximum-Likelihood-Schätzer

Definition 6.6. Sei $\mathscr{M} = (\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Standardmodell mit

$$\text{Gewichten } \rho_\vartheta(\cdot) \quad \text{bzw.} \quad \text{Dichte } \rho_\vartheta(\cdot) \qquad \text{für } \vartheta \in \Theta.$$

Die Funktion

$$\begin{array}{rccl} \rho: & \mathcal{X}\times\Theta & \to & [0,\infty) \\ & \cup & & \\ & (x,\vartheta) & \mapsto & \rho(x,\vartheta) := \rho_\vartheta(x) \end{array}$$

heißt *Likelihood-Funktion* (manchmal auch „Plausibilitäts-Funktion“), für $x \in \mathcal{X}$ heißt

$$L_x : \Theta \to [0,\infty), \quad L_x(\vartheta) = \rho(x,\vartheta)$$

die Likelihood-Funktion zum Beobachtungswert x.

Ein Schätzer $T : \mathcal{X} \to \Theta$ heißt (ein) Maximum-Likelihood-Schätzer, wenn

$$\rho(x, T(x)) = \max_{\vartheta\in\Theta} \rho(x,\vartheta) \quad \forall x \in \mathcal{X}$$

(auch kurz ML-Schätzer genannt, engl. MLE = maximum likelihood estimator).

Beispiel 6.7. 1. („Rückfangmethode", engl. "capture-recapture") Ein Teich enthalte ϑ Fische (einer gewissen Art, $\vartheta \in \mathbb{N}$ ist der unbekannte Parameter), fange und markiere m, setze wieder aus. Wenn sich die markierten Fische gut verteilt haben, fange erneut n Fische.

Nehmen wir an, wir beobachten unter den erneut gefangenen x markierte Fische. Formalisierung als statistisches Modell:

$\mathcal{X} = \{0, 1, \dots, n\}$, $\Theta = \{(m \vee n), (m \vee n) + 1, (m \vee n) + 2, \dots\}$, $P_\vartheta = \mathrm{Hyp}_{m,\vartheta-m,n}$

Die Likelihood-Funktion ist

$$\rho(x, \vartheta) = \frac{\binom{m}{x}\binom{\vartheta-m}{n-x}}{\binom{\vartheta}{n}},$$

der ML-Schätzer ist

$$\widehat{\vartheta}_{\mathrm{ML}} = T(x) = \Big\lfloor \frac{n}{x} \cdot m \Big\rfloor,$$

denn

$$\begin{aligned}\frac{\rho(x,\vartheta)}{\rho(x,\vartheta-1)} &= \frac{\binom{\vartheta-m}{n-x}\binom{\vartheta-1}{n}}{\binom{\vartheta}{n}\binom{\vartheta-1-m}{n-x}} \\ &= \frac{(\vartheta-m)(\vartheta-n)}{\vartheta(\vartheta-m-n+x)} = 1 - \frac{\vartheta x - mn}{\vartheta(\vartheta-m-n+x)} \begin{cases} > 1, & \vartheta < \frac{mn}{x}, \\ = 1, & \vartheta = \frac{mn}{x}, \\ < 1, & \vartheta > \frac{mn}{x} \end{cases}\end{aligned}$$

(beachte: stets ist $\vartheta - m \geq n - x$, es gibt im Teich mindestens so viele unmarkierte Fische wie in der Rückfang-Stichprobe).

2. (Erfolgsw'keit im Binomialmodell)

$$\rho(x,\vartheta) = L_x(\vartheta) = \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}, \quad \big(P_\vartheta = \mathrm{Bin}_{n,\vartheta} \text{ für } \theta \in \Theta = [0,1],\ x \in \mathcal{X} = \{0,1,\dots,n\}\big),$$

$$\begin{aligned}\frac{d}{d\vartheta}\log L_x(\vartheta) &= \frac{d}{d\vartheta}\Big(\log\binom{n}{x} + x\log\vartheta + (n-x)\log(1-\vartheta)\Big) \\ &= \frac{x}{\vartheta} - \frac{n-x}{1-\vartheta} = 0 \iff \vartheta = \frac{x}{n},\end{aligned}$$

d.h. hier ist $\widehat{\vartheta}_{\mathrm{ML}} = \dfrac{x}{n}$.

(Es ist $\frac{d}{d\vartheta}\log L_x(\vartheta) > 0$ für $\vartheta < x/n$ und $\frac{d}{d\vartheta}\log L_x(\vartheta) < 0$ für $\vartheta > x/n$, d.h. es handelt sich tatsächlich um ein Maximum; Inspektion zeigt, dass auch in den Randfällen $x = 0$ und $x = n$ $\widehat{\vartheta}_{\mathrm{ML}} = \frac{x}{n}$ gilt.)

3. (Normales Modell mit bekannter Varianz) n Beobachtungen seien u.i.v. $\sim \mathcal{N}_{\vartheta,\sigma^2}$, $\sigma^2 > 0$ sei bekannt.

Mit $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ ist

$$\begin{aligned}\rho(x,\vartheta) = L_x(\vartheta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x_i-\vartheta)^2}{2\sigma^2}\Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\vartheta)^2\Big)\end{aligned}$$

d.h.

$$L_x(\vartheta) \stackrel{!}{=} \max \iff \sum_{i=1}^{n}(x_i - \vartheta)^2 \stackrel{!}{=} \min .$$

Mit $m(x) := \frac{1}{n}\sum_{i=1}^{n} x_i$ ist

$$\sum_{i=1}^{n}(x_i - \vartheta)^2 = \sum_{i=1}^{n}\big(x_i - m(x)\big)^2 + \big(m(x) - \vartheta\big)^2,$$

d.h. es ist $\widehat{\vartheta}_{\mathrm{ML}} = m(x)$, das empirische Mittel der Beobachtungen.

4. (Normales Modell, unbekannter Erwartungswert und unbekannte Varianz) n Beobachtungen seien u.i.v. $\sim \mathcal{N}_{\mu,v}$ mit unbekannten $\mu \in \mathbb{R}$ und $v \in (0,\infty)$.

(Formalisierung: $\mathcal{X} = \mathbb{R}^n$, $\Theta = \{\vartheta = (\mu, v) : \mu \in \mathbb{R}, v > 0\}$, $P_{(\mu,v)} = \mathcal{N}_{\mu,v}^{\otimes n}$)

Wie in 3. ist

$$\log L_x\big((\mu,v)\big) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log v - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2,$$

nach obigem ist $\widehat{\mu}_{\mathrm{ML}} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ Maximierer bezüglich μ (für jeden Wert von v), weiter ist

$$\frac{\partial}{\partial v}\log L_x\big((\mu,v)\big)\Big|_{\mu=\widehat{\mu}_{\mathrm{ML}}} = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{\mathrm{ML}})^2,$$

also $\dfrac{\partial}{\partial v}\log L_x\big((\widehat{\mu}_{\mathrm{ML}}, v)\big) = 0 \iff v = \widehat{v}_{\mathrm{ML}} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{\mathrm{ML}})^2.$

(Und man prüft: $\log L_x\big((\widehat{\mu}_{\mathrm{ML}}, v)\big)$ ist wachsend für $v < \widehat{v}_{\mathrm{ML}}$, fallend für $v > \widehat{v}_{\mathrm{ML}}$.)

Beachte: Der ML-Schätzer für die unbekannte Varianz ist hier die (unkorrigierte) Stichprobenvarianz, also ist er nicht erwartungstreu (vgl. Beob. und Def. 6.5).

5. n Beobachtungen seien u.i.v. uniform auf $[0,\vartheta]$ (mit einem unbekannten $\vartheta \in (0,\infty)$).

Es ist

$$\widehat{\vartheta}_{\mathrm{ML}} = \max\{x_1, x_2, \ldots, x_n\},$$

denn

$$L_{(x_1,\ldots,x_n)}(\vartheta) = \begin{cases} \dfrac{1}{\vartheta^n}, & \text{falls } \vartheta \geq x_1, x_2, \ldots, x_n, \\ 0, & \text{sonst.} \end{cases}$$

Bericht 6.8 (Cramér-Rao-Schranke und „beste“ Schätzer)**.** Sei $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Standardmodell, $\rho(x,\vartheta)$ die Likelihoodfunktion.

Ein erwartungstreuer Schätzer T für ein reelles Parametermerkmal $\tau(\vartheta)$ heißt *varianzminimierend* (auch „gleichmäßig bester Schätzer“, engl. UMVU (= uniformly minimum variance unbiased) estimator), falls für jeden anderen erwartungstreuen Schätzer $\widetilde{T}$ für τ gilt

$$\mathrm{Var}_\vartheta[T] \leq \mathrm{Var}_\vartheta[\widetilde{T}] \quad \text{für alle } \vartheta \in \Theta.$$

(In diesem Sinne ist T optimal und beantwortet – so existent – auf diese Weise die Frage „Wie gut kann man $\tau(\vartheta)$ anhand der Beobachtungen überhaupt schätzen?“)

Ein einparametriges Standardmodell (d.h. $\Theta \subset \mathbb{R}$) heißt regulär, falls gilt:

(i) $\Theta \subset \mathbb{R}$ ist ein offenes Intervall.

(ii) Likelihood-Funktion $\rho(x, \vartheta)$ ist strikt positiv auf $\mathcal{X} \times \Theta$ und für jedes x ist $\vartheta \mapsto \rho(x, \vartheta)$ stetig diff'bar.

(iii) $U_\vartheta(x) := \frac{d}{d\vartheta} \log \rho(x, \vartheta)$ erfüllt $I_\vartheta := \mathrm{Var}_\vartheta[U_\vartheta] \in (0, \infty)$
(U_ϑ heißt die „Scorefunktion" und I_ϑ heißt die *Fisher-Information*) und es gilt

$$\int_{\mathcal{X}} \frac{d}{d\vartheta} \rho(x, \vartheta)\, dx = \frac{d}{d\vartheta} \int_{\mathcal{X}} \rho(x, \vartheta)\, dx \quad (= 0).$$

Weiter heißt ein Schätzer T regulär, wenn für jedes $\vartheta \in \Theta$ gilt

$$\frac{d}{d\vartheta} \int_{\mathcal{X}} T(x) \rho(x, \vartheta)\, dx = \int_{\mathcal{X}} T(x) \frac{d}{d\vartheta} \rho(x, \vartheta)\, dx.$$

(Wenn $\mathcal{X}$ diskret ist, so ist jeweils das Integral $\int_{\mathcal{X}} \ldots dx$ durch die Summe $\sum_{x \in \mathcal{X}} \ldots$ zu ersetzen.)

Sei $\tau : \Theta \to \mathbb{R}$ ein stetig differenzierbares Parametermerkmal, T ein regulärer, erwartungstreuer Schätzer für τ in einem regulären Standardmodell. Dann gilt die Cramér-Rao-Schranke[1]:

$$\mathrm{Var}_\vartheta[T] \geq \frac{(\tau'(\vartheta))^2}{I(\vartheta)} \quad \forall\, \vartheta \in \Theta,$$

wobei Gleichheit genau dann gilt, wenn

$$T(x) - \tau(\vartheta) = \frac{\tau'(\vartheta) U_\vartheta(x)}{I(\vartheta)}.$$

Beispiel(-klasse). Exponentielle Familien (bzgl. der Statistik T)
Sei

$$\rho(x, \vartheta) = h(x) \cdot \exp\big(a(\vartheta) \cdot T(x) - b(\vartheta)\big)$$

für gewisse Funktionen $a, b : \Theta \to \mathbb{R}$ und $h : \mathcal{X} \to \mathbb{R}$

Dann ist

$$U_\vartheta(x) = a'(\vartheta) T(x) - b'(\vartheta),$$

also

$$\mathbb{E}_\vartheta[T] = \frac{b'(\vartheta)}{a'(\vartheta)} =: \tau(\vartheta),$$

man kann zeigen, dass

$$\begin{aligned} I(\vartheta) &= \mathrm{Var}_\vartheta[U_\vartheta] \\ &= a'(\vartheta) \cdot \tau'(\vartheta), \end{aligned}$$

d.h. es gilt

$$T(x) = \frac{b'(\vartheta)}{a'(\vartheta)} + \frac{U_\vartheta(x)}{a'(\vartheta)} = \tau(\vartheta) + \frac{\tau'(\vartheta) U_\vartheta(x)}{I(\vartheta)}.$$

T ist demnach in dieser Situation ein varianzminimierender erwartungstreuer Schätzer für τ.

Beispiel-Instanzen.

[1] nach Harald Cramér, 1983–1985 und Calyampudi Radhakrishna Rao, *1920

1. Binomialverteilungen: $P_\vartheta = \mathrm{Bin}_{n,\vartheta}$, $\vartheta \in [0,1]$

$$\rho(x,\vartheta) = \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x} = \binom{n}{x}\exp\Big(\underbrace{\frac{x}{n}}_{T(x)}\underbrace{n\log(\frac{\vartheta}{1-\vartheta})}_{=a(\vartheta)} + \underbrace{n\log(1-\vartheta)}_{=-b(\vartheta)}\Big), \quad x \in \mathbb{N}_0,$$

$T(x) = \frac{x}{n}$ ist varianzminimierender erwartungstreuer Schätzer für $\tau(\vartheta) = b'(\vartheta)/a'(\vartheta) = \vartheta$.

2. Poissonverteilungen: $P_\vartheta = \mathrm{Poi}_\vartheta$, $\vartheta \in (0,\infty)$

$$\rho(x,\vartheta) = e^{-\vartheta}\frac{\vartheta^x}{x!} = \underbrace{\frac{1}{x!}}_{=h(x)} e^{\overbrace{x}^{T(x)}\overbrace{\log\vartheta}^{=a(\vartheta)} - \overbrace{\vartheta}^{b(\vartheta)}}$$

Es ist $\tau(\vartheta) = \frac{1}{1/\vartheta} = \vartheta$, $T(x) = x$ ist varianzminimierender erwartungstreuer Schätzer für ϑ, seine Varianz ist $\frac{(\tau'(\vartheta))^2}{a'(\vartheta)\tau'(\vartheta)} = \frac{1^2}{\frac{1}{\vartheta}\cdot 1} = \vartheta$.

3. Normalverteilungen bei bekannter Varianz: $P_\vartheta = \mathcal{N}_{\vartheta,\sigma^2}$ mit festem $\sigma^2 > 0$

$$\begin{aligned}\rho(x,\vartheta) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{1}{2\sigma^2}(x-\vartheta)^2\Big)\\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/(2\sigma^2)}}_{=h(x)}\cdot\exp\Big(-\underbrace{x}_{=T(x)}\cdot\underbrace{\frac{\vartheta}{\sigma^2}}_{=a(\vartheta)} - \underbrace{\frac{\vartheta^2}{2\sigma^2}}_{=b(\vartheta)}\Big),\end{aligned}$$

also: $T(x) = x$ ist varianzminimierender erwartungstreuer Schätzer für $\vartheta = \tau(\vartheta) = \frac{b'(\vartheta)}{a'(\vartheta)}$, seine Varianz ist $\sigma^2 = \frac{1}{I(\vartheta)} = 1^2/(a'(\vartheta)\tau'(\vartheta))$.

Bemerkung. Das n-fache Produktmodell $\mathscr{M}^{\otimes n}$ eines regulären Modells ist wiederum regulär, seine Fisher-Information erfüllt $I^{(n)}(\vartheta) = n \cdot I(\vartheta)$.

Ist $\mathscr{M}$ exponentielles Modells bzgl. der Statistik T, so ist $\mathscr{M}^{\otimes n}$ ebenfalls ein exponentielles Modell, und die zugrundeliegende Statistik ist

$$T_n(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^n T(x_i),$$

denn

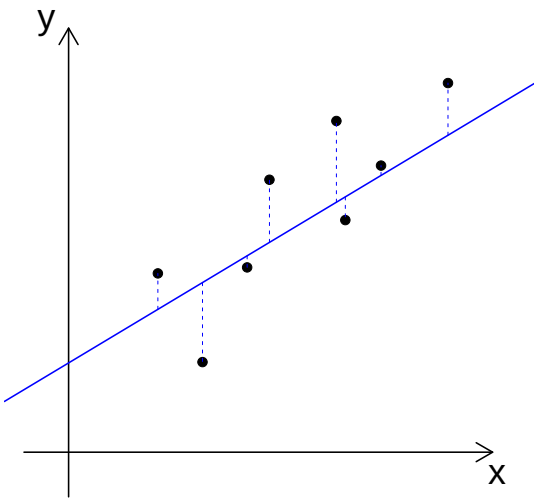
$$\rho^{\otimes n}((x_1,\dots,x_n),\vartheta) = \prod_{i=1}^n \rho(x_i,\vartheta) = \prod_{i=1}^n h(x_i)\exp\left(na(\vartheta)\cdot\frac{1}{n}\cdot\sum_{i=1}^n(T(x_i)) - nb(\vartheta)\right).$$

6.2.2 Eine Anmerkung zu linearer Regression und kleinste-Quadrate-Schätzung

Beobachtung 6.9 (Lineare Regression als kleinste-Quadrate-Schätzer)**.** Nehmen wir an, die Beobachtungen bestehen aus n Messwertpaaren (x_i, y_i), $i = 1,\dots,n$ (Werte in $\mathbb{R}^2$) und wir vermuten aus theoretischen Gründen einen (affin-)linearen Zusammenhang, d.h. bei „perfekter“ Messung gälte $y_i = \beta_0 + \beta_1 x_1$ für gewisse (uns unbekannte) Zahlen β_0 und β_1.

(Ein „Lehrbuchbeispiel“: y_i ist die Länge einer Stahlfeder bei Zugbelastung mit Gewicht x_i innerhalb des Gültigkeitsbereich des Hooke'schen Gesetzes.)

Aufgrund beispielsweise von Messungenauigkeiten (oder womöglich auch weil der lineare Zusammenhang in Wirklichkeit nur approximativ gilt) werden die realen Datenpunkte typischerweise nicht auf einer Geraden liegen.



Formulierung als statistisches Modell: $x_1, \dots, x_n$ seien feste (bekannte) Werte (x ist die „erklärende Variable“), für $\vartheta = (\beta_0, \beta_1) \in \Theta = \mathbb{R}^2$ sei unter P_ϑ

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \dots, n \quad \text{mit } \varepsilon_i \text{ u.i.v. mit } \mathbb{E}[\varepsilon_i] = 0,\ \mathrm{Var}[\varepsilon_i] = \sigma^2$$

und wir fassen die beobachteten y_i-Werte als Realisierungen der Y_i auf (y ist die „abhängige Variable“ oder „Zielgröße“).

Ein naheliegender Ansatz, $\vartheta = (\beta_0, \beta_1)$ zu schätzen, ist der *kleinste-Quadrate-Schätzer*: Finde $\widehat{\beta}_0, \widehat{\beta}_1$ so, dass

$$\sum_{i=1}^n \big(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\big)^2 = \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^n \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2$$

Die Lösung kennen wir schon (vgl. Beob. 3.21, die wir hier gewissermaßen nur „statistisch aussprechen“): Mit

$$\overline{x} := \frac{1}{n}\sum_{i=1}^n x_i, \quad \overline{y} := \frac{1}{n}\sum_{i=1}^n y_i, \quad \sigma_x^2 := \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2, \quad \sigma_y^2 := \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2$$
$$\mathrm{cov}_{x,y} := \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$$

ist

$$\widehat{\beta}_1 = \frac{\mathrm{cov}_{x,y}}{\sigma_x^2}, \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}. \tag{6.1}$$

(Betrachte nämlich eine ZV $(\widetilde{X}, \widetilde{Y})$ mit Werten in $\mathbb{R}^2$, deren Verteilung die empirische Verteilung der Datenpunkte ist, d.h. $\frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$, so ist $\mathbb{E}[\widetilde{X}] = \overline{x}, \mathbb{E}[\widetilde{Y}] = \overline{y}, \mathrm{Var}[\widetilde{X}] = \sigma_x^2, \mathrm{Var}[\widetilde{Y}] = \sigma_y^2, \mathrm{Cov}[\widetilde{X}, \widetilde{Y}] = \mathrm{cov}_{x,y}$ und die Behauptung folgt wörtlich aus Beob. 3.21, man kann natürlich auch die Rechnung dort nochmals hier wiederholen).

Übrigens: Wenn man zusätzlich annimmt, dass die ε_i u.i.v. $\sim \mathcal{N}_{0,\sigma^2}$ sind, so ist der kleinste-Quadrate-Schätzer hier auch zugleich der Maximum-Likelihood-Schätzer (mit einer Rechnung analog zu Bsp. 6.7, 3.).

6.3 Konfidenzintervalle (und Konfidenzbereiche)

Definition 6.10. Sei $\mathscr{M} = (\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\tau(\vartheta)$ reelles Parametermerkmal, L, R Statistiken mit $L \le R$, $\alpha \in (0,1)$.

Das (zufällige) Intervall $I := [L, R]$ heißt ein *Konfidenzintervall* (manchmal auch „Vertrauensintervall") für τ zum (Sicherheits-)Niveau $1-\alpha$ (bzw. Irrtumsniveau α), wenn gilt

$$\forall\, \vartheta \in \Theta \;:\quad P_\vartheta\big(\tau(\vartheta) \in I\big) \ge 1-\alpha.$$

Beachte: I ist zufällig, nicht aber ϑ (zumindest in unserer (sogenannten frequentistischen) Interpretation).

Allgemeiner heißt eine in Abhängigkeit von den Beobachtungen $x \in \mathcal{X}$ konstruierte Menge $C(x) \subset \Theta$ heißt ein *Konfidenzbereich* für τ zum (Sicherheits-)Niveau $1-\alpha$, wenn gilt

$$\forall\, \vartheta \in \Theta \;:\quad P_\vartheta\big(\{x \in \mathcal{X} : C(x) \ni \tau(\vartheta)\}\big) \ge 1-\alpha.$$

Offenbar möchte man i.A. I so kurz wie möglich wählen (soweit verträglich mit dem geforderten Niveau).

Beispiel 6.11 (Konfidenzintervall für den Mittelwert im normalen Modell bei bekannter Varianz)**.** Unter P_ϑ seien $X_1, X_2, \dots, X_n$ u.i.v. $\sim \mathcal{N}_{\vartheta,\sigma^2}$ mit $\vartheta \in \Theta := \mathbb{R}$, $\sigma^2 > 0$ sei bekannt (und fest).

Sei $q := \Phi^{-1}(1-\frac{\alpha}{2})$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung, $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

$$I := \Big[\overline{X} - q\cdot\frac{\sigma}{\sqrt{n}}, \overline{X} + q\cdot\frac{\sigma}{\sqrt{n}}\Big]$$

ist ein Konfidenzintervall für ϑ zum (Sicherheits-)Niveau $1-\alpha$, denn unter P_ϑ ist $\overline{X} \sim \mathcal{N}_{\vartheta,\sigma^2/n}$,

$$\begin{aligned} P_\vartheta\Big(\overline{X} - q\cdot\frac{\sigma}{\sqrt{n}} \le \vartheta \le \overline{X} + q\cdot\frac{\sigma}{\sqrt{n}}\Big) &= P_\vartheta\Big(q \ge \frac{\overline{X}-\vartheta}{\sigma/\sqrt{n}} \ge -q\Big) \\ = P(-q \le Z \le q) = P(Z \le q) - P(Z \ge -q) &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha \end{aligned}$$

(mit $Z \sim \mathcal{N}_{0,1}$).

Beispiel 6.12 (Approximatives Konfidenzintervall im Binomialmodell mittels Normalapproximation, vgl. auch Bsp. 6.1)**.** $X \sim \mathrm{Bin}_{n,\vartheta}$, $\vartheta \in \Theta = [0,1]$,

$\widehat{\vartheta} := \frac{X}{n}$, $\widehat{\sigma} := \sqrt{\widehat{\vartheta}(1-\widehat{\vartheta})}$, $\alpha \in (0,1)$, $q := \Phi^{-1}(1-\frac{\alpha}{2})$, dann ist

$$I := \Big[\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\Big]$$

ein (approximatives) Konfidenzintervall für ϑ zum Sicherheitsniveau $1-\alpha$, denn unter P_ϑ gilt für $n \to \infty$

$$\widehat{\vartheta} \xrightarrow{d} \vartheta, \quad \widehat{\sigma} \xrightarrow{d} \sqrt{\vartheta(1-\vartheta)}, \quad \text{und } \frac{\widehat{\vartheta}-\vartheta}{\widehat{\sigma}/\sqrt{n}} = \frac{X - n\vartheta}{\widehat{\sigma}\sqrt{n}} \xrightarrow{d} \mathcal{N}_{0,1}$$

(mit Satz von de Moivre-Laplace, Satz 5.1 und Korollar 5.2)

$$P_\vartheta\Big(\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}} \le \vartheta \le \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\Big) = P_\vartheta\Big(-q \le \frac{\widehat{\vartheta}-\vartheta}{\widehat{\sigma}/\sqrt{n}} \le q\Big) \approx P(-q \le Z \le q) = 1-\alpha$$

6.3.1 Rund um die multivariate Normalverteilung

Für das weitere Vorgehen benötigen wir einige Eigenschaften der multivariaten Normalverteilung.

Beobachtung und Definition 6.13. *$n \in \mathbb{N}$, $X_1, X_2, \ldots, X_n$ u.i.v. $\sim \mathcal{N}_{0,1}$, so hat*

$$X := X_1^2 + X_2^2 + \cdots + X_n^2 \quad \textit{die Dichte } \tfrac{1}{\Gamma(n/2)} 2^{-n/2} x^{\frac{n}{2}-1} e^{-x/2} \mathbf{1}_{[0,\infty)}(x),$$

$\chi_n^2 := \mathscr{L}(X)$ *heißt* Chiquadrat-Verteilung *mit n Freiheitsgraden.*

Beachte: $\chi_n^2 = \Gamma_{1/2,n/2}$, wo die Gamma-Verteilung $\Gamma_{\alpha,\nu}$ die Dichte $\frac{1}{\Gamma(\nu)} \alpha^\nu x^{\nu-1} e^{-\alpha x} \mathbf{1}_{(0,\infty)}(x)$ besitzt (α = Skalen-, ν = Formparameter), siehe Aufg. 4.1.

Proposition 6.14. *Seien $\alpha, r, s > 0$, $X \sim \Gamma_{\alpha,r}, Y \sim \Gamma_{\alpha,s}$ unabhängig. Dann sind*

$$X + Y \;\; \textit{und} \;\; V := \frac{X}{X+Y} \;\; \textit{unabhängig}$$

und $X + Y \sim \Gamma_{\alpha,r+s}$, $V \sim \beta_{r,s}$, wobei die Beta-Verteilung $\beta_{r,s}$ die Dichte

$$\tfrac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} v^{r-1}(1-v)^{s-1} \mathbf{1}_{(0,1)}(v)$$

besitzt.

Insbesondere bilden die Gamma-Verteilungen eine Faltungsfamilie (bezüglich des zweiten, des sogenannten Formparameters): $\Gamma_{\alpha,r} * \Gamma_{\alpha,s} = \Gamma_{\alpha,r+s}$.

Beweis. (X, Y) hat Dichte

$$f_{(X,Y)}(x,y) = \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} x^{r-1} y^{s-1} e^{-\alpha(x+y)} \text{ auf } (0,\infty)^2.$$

Sei $\varphi(x,y) = \begin{pmatrix} x+y \\ \frac{x}{x+y} \end{pmatrix}$, so ist

$$\varphi^{-1}(z,v) = \begin{pmatrix} zv \\ z(1-v) \end{pmatrix}, \quad D\varphi(x,y) = \begin{pmatrix} 1 & \frac{y}{(x+y)^2} \\ 1 & -\frac{x}{(x+y)^2} \end{pmatrix}, \quad |\det D\varphi(x,y)| = \frac{|x+y|}{(x+y)^2} = \frac{1}{|x+y|}$$

Schreibe $Z := X + Y$, $V := \frac{X}{X+Y}$. Gemäß 2-dimensionaler Dichtetransformation (siehe Bericht 1.42) ist die Dichte von (Z,V):

$$\begin{aligned} f_{(Z,V)}(z,v) &= \frac{f_{(X,Y)}(\varphi^{-1}(z,v))}{|\det D\varphi(\varphi^{-1}(z,v))|} \\ &= z \cdot \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} (zv)^{r-1} (z(1-v))^{s-1} e^{-\alpha z} \\ &= \underbrace{\frac{\alpha^{r+s}}{\Gamma(r+s)} z^{r+s-1} e^{-\alpha z}}_{\text{Dichte von } \Gamma_{\alpha,r+s}} \underbrace{\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} v^{r-1}(1-v)^{s-1}}_{\text{Dichte von } \beta_{r,s}} \end{aligned}$$

□

Beweis von Beob. 6.13. $X \sim \mathcal{N}_{0,1}$, so ist $X^2 \sim \Gamma_{\frac{1}{2},\frac{1}{2}}$ $(= \chi_1^2)$:

$|X|$ hat Dichte $\frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ auf $(0,\infty)$; sei $\varphi : (0,\infty) \to (0,\infty), x \mapsto x^2$, $\varphi^{-1}(y) = \sqrt{y}$, $\frac{d}{dy}\varphi(y) = 2y$, also hat X^2 die Dichte $\frac{1}{2\sqrt{y}} \frac{2}{\sqrt{2\pi}} e^{-\frac{y}{2}} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \frac{1}{\Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} e^{-\frac{1}{2}y}$ (siehe Beob. 1.40).

Dies zeigt die Behauptung für $n = 1$, der allgemeine Fall folgt daraus induktiv unter Verwendung von Proposition 6.14. □

Korollar und Definition 6.15. Seien $m, n \in \mathbb{N}$, $X_1, \dots, X_m, Y_1, \dots, Y_n$ unabhängig, $\sim \mathcal{N}_{0,1}$.

1. $F_{m,n} := \dfrac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2}$ hat Dichte $f_{m,n}(x) = \dfrac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \dfrac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{(m+n)}{2}}} \mathbf{1}_{(0,\infty)}(x)$.

$\mathcal{L}(F_{m,n})$ heißt *Fisher-Verteilung*[2] mit m und n Freiheitsgraden (präziser: mit m Zähler- und n Nenner-Freiheitsgraden).

2. $T_n := \dfrac{X}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} Y_j^2}}$ hat Dichte $t_n(x) = \dfrac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})} \dfrac{1}{\sqrt{n}} \left(1 + \dfrac{x^2}{n}\right)^{-\frac{n+1}{2}}$.

$\mathcal{L}(T_n)$ heißt *Student-Verteilung*[3] mit n Freiheitsgraden (auch Student'sche T-Verteilung genannt).

Bemerke: Die Student-Verteilung mit einem Freiheitsgrad ist die Cauchy-Verteilung.

Bemerkung. Sei T_n Student-verteilt mit n Freiheitsgraden, so ist $T_n \xrightarrow[n\to\infty]{d} \mathcal{N}_{0,1}$.
(denn es gilt $t_n(x) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ lokal gleichmäßig).

Beweis von Korollar 6.15. 1. $X := \sum_{i=1}^{m} X_i^2 \sim \Gamma_{\frac{1}{2},\frac{m}{2}}$, $Y := \sum_{j=1}^{n} Y_j^2 \sim \Gamma_{\frac{1}{2},\frac{n}{2}}$ sind unabhängig, also ist $V := \dfrac{X}{X+Y} \sim \beta_{\frac{m}{2},\frac{n}{2}}$ nach Proposition 6.14.

Dann ist

$$F_{m,n} = \frac{nX}{mY} = \frac{n}{m} \frac{\frac{X}{X+Y}}{\frac{Y}{X+Y}} = \frac{n}{m} \frac{V}{1-V},$$

mit

$$\varphi : (0,1) \to (0,\infty), v \mapsto \frac{n}{m}\frac{v}{1-v}, \quad \text{also } \varphi^{-1}(z) = \frac{mz}{n+mz}, \quad \frac{d}{dv}\varphi(v) = \frac{n}{m}\frac{1}{(1-v)^2}$$

ist $F_{m,n} = \varphi(V)$, hat also die Dichte

$$\begin{aligned} f_{m,n}(z) &= \frac{mnz}{(n+mz)^2} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{mz}{n+mz}\right)^{\frac{m}{2}-1} \left(\frac{n}{n+mz}\right)^{\frac{n}{2}-1} \\ &= \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{z^{\frac{m}{2}-1}}{(n+mz)^{\frac{(m+n)}{2}}} \end{aligned}$$

[2] Nach Ronald Aylmer Fisher, 1890–1962

[3] Nach William Sealy Gosset, 1876–1937, der sie 1908 unter dem Pseudonym "Student" veröffentlichte.

2. T_n^2 hat (nach 1.) Dichte $f_{1,n}$, also hat $|T_n|$ Dichte $2tf_{1,n}(t^2)\mathbf{1}_{[0,\infty)}(t)$.

Da T_n symmetrisch um 0 verteilt ist (klar aus der Symmetrie von X_1), hat T_n die Dichte

$$\begin{aligned}|t|f_{1,n}(t^2) &= |t|\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}n^{n/2}\frac{(t^2)^{\frac{1}{2}-1}}{(n+t^2)^{(n+1)/2}}\\ &= \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}\frac{1}{\sqrt{n}}\frac{1}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}\end{aligned}$$

□

Satz 6.16. *$X_1, \dots, X_n$ u.i.v. $\sim \mathcal{N}_{\mu,\sigma^2}$ mit $\mu \in \mathbb{R}$, $\sigma > 0$,*

$$M := \frac{1}{n}\sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - M)^2.$$

Es gilt

1. M und S^2 sind unabhängig, $M \sim \mathcal{N}_{\mu,\sigma^2/n}$, $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

2. $T := \dfrac{\sqrt{n}(M-\mu)}{\sqrt{S^2}}$ ist Student-verteilt mit $n-1$ Freiheitsgraden.

Beweis. Sei o.E. $\mu = 0$, $\sigma^2 = 1$, sonst betrachte $X_i' := (X_i - \mu)/\sqrt{\sigma^2}$.

1. Sei O orthogonale $n \times n$-Matrix, deren erste Zeile $z_1 = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$ ist, d.h. ergänze z_1 zu einer Orthonormalbasis $z_1, \dots, z_n$ von $\mathbb{R}^n$, setze

$$O = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}.$$

Dann ist

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} := O \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

n-dimensional Standardnormalverteilt (nach Beispiel 2.20, Invarianz der n-dim. Normalverteilung unter orthogonalen Transformationen).

Somit

$$\begin{aligned} Y_1 &= \sum_{i=1}^n \frac{1}{\sqrt{n}}X_i = \sqrt{n}M, \qquad \text{also } M \sim \mathcal{N}_{0,1/n},\\ (n-1)S^2 &= \sum_{i=1}^n (X_i - M)^2 = \sum_{i=1}^n X_i^2 - nM^2\\ &= \|(X_1, \dots, X_n)^T\|^2 - Y_1^2 = \|(Y_1, \dots, Y_n)^T\|^2 - Y_1^2 = \sum_{i=2}^n Y_i^2, \end{aligned}$$

also $(n-1)S^2 \sim \chi^2_{n-1}$ und unabhängig von M.

(Geometrisch ausgedrückt: Zerlege $\mathbb{R}^n = D \oplus D^\perp$ in die Diagonale $D = \{(x, x, \dots, x) : x \in \mathbb{R}\} \subset \mathbb{R}^n$ und ihr orthogonales Komplement $D^\perp = \{(x_1, x_2, \dots, x_n) : x_1 + \cdots + x_n = 0\}$, seien $\mathcal{P}_D : \mathbb{R}^n \to D$, $\mathcal{P}_{D^\perp} = \mathrm{Id}_{\mathbb{R}^n} - \mathcal{P}_D : \mathbb{R}^n \to D^\perp$ die orthogonalen Projektionen auf D bzw. auf $D^\perp$, dann ist $\sqrt{n}M$ die (signierte) Länge von $\mathcal{P}_D X$ und $(n-1)S^2 = \|\mathcal{P}_{D^\perp} X\|^2$.)

2. Dies folgt aus 1. und der Definition (vgl. Korollar und Definition 6.15, 2.) □

Korollar 6.17 (Student-Konfidenzintervall für den Erwartungswert im normalen Modell)**.** *Unter P_ϑ, $\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ seien*

$$X_1, X_2, \dots, X_n \text{ u.i.v. } \sim \mathcal{N}_{\mu,\sigma^2}.$$

Sei $\alpha \in (0,1)$, $q = q_{n-1,1-\alpha/2}$ das $1 - \frac{\alpha}{2}$-Quantil der Student-Verteilung mit $n-1$ Freiheitsgraden,

$$M := \frac{1}{n}\sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - M)^2.$$

Dann ist

$$I := \Big[M - q\sqrt{\frac{S^2}{n}}, M + q\sqrt{\frac{S^2}{n}}\Big]$$

ein Konfidenzintervall für μ zum Irrtumsniveau α.

Beweis. $T := \frac{\sqrt{n}(M-\mu)}{\sqrt{S^2}}$ ist Student-verteilt mit $n-1$ Freiheitsgraden (für jede Wahl von μ und σ^2),

$$\begin{aligned} &P_{(\mu,\sigma^2)}\Big(M - q\sqrt{\tfrac{S^2}{n}} \le \mu \le M + q\sqrt{\tfrac{S^2}{n}}\Big) \\ &= P(-q \le T \le q) = P(T \le q) - P(T \le -q) = 1 - \frac{\alpha}{2}. \end{aligned}$$

□

Beispiel. Zwei Schlafmittel sollen verglichen werden, 10 Patienten erhielten in aufeinanderfolgenden Nächten Medikament A und B.

Die Daten[4] (x_i = Anz. Stunden Schlaf mit Mittel A - Anz. Stunden Schlaf mit Mittel B bei Patient Nr. i):

i	1	2	3	4	5	6	7	8	9	10
x_i	1,2	2,4	1,3	1,3	0,0	1,0	1,8	0,8	4,6	1,4

Es ist

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i \approxeq 1{,}58, \quad s = \Big(\frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2\Big)^{1/2} \approxeq 1{,}23.$$

Nehmen wir an, die Daten stammen aus einer Normalverteilung mit unbekanntem Mittelwert μ und unbekannter Varianz σ^2 (und die Ergebnisse der verschiedenen Patienten sind unabhängig).

[4] Aus Student (= William S. Gosset, 1876–1937), The Probable Error of a Mean, Biometrika 6:1–25 (1908), siehe Illustration I in Section IX dort.

Es ist $q_{9,\,0{,}995} \approx 3{,}25$ (aus einer Quantiltabelle oder beispielsweise mit R berechnet), demnach ist

$$\left[\overline{x} \pm q\frac{s}{\sqrt{n}}\right] \approx [0{,}31,\ 2{,}85]$$

ein Konfidenzintervall für μ (die mittlere zusätzliche Anzahl Stunden Schlaf, die Medikament A mehr bringt als Medikament B) zum Sicherheitsniveau $0{,}99 = 1 - 0{,}01$.

Beachte: (Sinnlos) genaue Werte mit Rechnergenauigkeit sind $\overline{x} - q\frac{s}{\sqrt{n}} \approx 0{,}3159481$, $\overline{x} + q\frac{s}{\sqrt{n}} \approx 2{,}8440519$, man sollte allerdings die Grenzen eines Konfidenzintervalls stets „konservativ", d.h. nach außen, runden.

6.3.2 Ein Konfidenzintervall für den Median (ein kleiner Ausflug in die nicht-parametrische Statistik)

Satz 6.18 (Ein Konfidenzintervall für den Median)**.** *Seien $X_1, \ldots, X_n$ u.i.v. reellwertig, mit (unbekannter) Verteilung Q, die eine stetige Verteilungsfunktion besitzt (d.h. Q hat keine Atome). (Im Formalismus: $\Theta = \{\vartheta : \vartheta$ nicht-atomares W'maß auf $\mathbb{R}\}$, $\mathcal{X} = \mathbb{R}^n$, $P_\vartheta = \vartheta^{\otimes n}$)*

$m(Q)$ sei „der" Median von Q (d.h. $Q\big((-\infty, m(Q)]\big) = \frac{1}{2} = Q\big([m(Q), \infty)\big)$, vgl. Def. 3.22; falls mehrere Werte in Frage kommen, nehmen wir das arithmetische Mittel aus dem kleinsten und dem größten möglichen Wert).

Die zugehörige Ordnungsstatistik ist

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

Zu $\alpha \in (0, 1)$ wähle k maximal, so dass $\mathrm{Bin}_{n,1/2}(\{0, \ldots, k-1\}) \le \frac{\alpha}{2}$, dann ist

$$\left[X_{(k)}, X_{(n-k+1)}\right]$$

ein Konfidenzintervall für den Median $m(Q)$ zum Sicherheitsniveau $1 - \alpha$.

Beweis. Es ist

$$\begin{aligned} Q^{\otimes n}\big(X_{(k)} > m(Q)\big) &= Q^{\otimes n}\big(|\{1 \le i \le n : X_i \le m(Q)\}| \le k-1\big) \\ &= \mathrm{Bin}_{n,1/2}(\{0, \ldots, k-1\}) \le \frac{\alpha}{2}, \end{aligned}$$

analog ist

$$Q^{\otimes n}\big(X_{(n-k+1)} < m(Q)\big) = Q^{\otimes n}\big(|\{1 \le i \le n : X_i \ge m(Q)\}| \le k-1\big) \le \frac{\alpha}{2},$$

somit

$$Q^{\otimes n}\Big(\left[X_{(k)}, X_{(n-k+1)}\right] \not\ni m(Q)\Big) \le Q^{\otimes n}\big(X_{(k)} > m(Q)\big) + Q^{\otimes n}\big(X_{(n-k+1)} < m(Q)\big) \le \alpha.$$

□

Im „Schlafmittel-Vergleich"-Beispiel oben ergäbe sich für $\alpha = 0{,}01$: $n = 10$, man muss in Satz 6.18 $k = 1$ wählen (es ist $\mathrm{Bin}_{10,1/2}(\{0\}) \approx 0{,}001$, aber $\mathrm{Bin}_{10,1/2}(\{0, 1\}) \approx 0{,}012$), d.h. ein Konfidenzintervall für den Median (der Differenz der Schlafdauer unter Mittel A versus Mittel B) zum Sicherheitsniveau 99% ist $[X_{(1)}, X_{(10)}] = [0,\ 4{,}6]$.

(Für $\alpha = 0{,}05$ könnte man $k = 2$ wählen und erhielte $[X_{(2)}, X_{(9)}] = [0{,}8,\ 2{,}4]$ als Konfidenzintervall zum Sicherheitsniveau 95%.)

6.3.3 Exkurs: Exakte Konfidenzintervalle für den Erfolgsparameter in der Binomialverteilung*

Unter n unabhängigen Versuchen seien x Erfolge beobachtet worden, wir fassen x als Realisierung einer $\mathrm{Bin}_{n,\vartheta}$-verteilten ZV auf und wollen anhand der Beobachtung auf ϑ schließen.

Wir hatten in Beispiel 6.12 das auf asymptotischer Normalität fußende (approximative) Konfidenzintervall für ϑ zum Niveau $1-\alpha$ kennen gelernt:

$$\left[\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

mit $\widehat{\vartheta} = \dfrac{x}{n}$, $\widehat{\sigma} = \sqrt{\widehat{\vartheta}(1-\widehat{\vartheta})}$, q das $1-\frac{\alpha}{2}$-Quantil von $\mathcal{N}_{0,1}$

Wie könnten wir vorgehen, wenn wir uns nicht auf die Asymptotik verlassen möchten? Wir beobachte $X \sim P_\vartheta := \mathrm{Bin}_{n,\vartheta}$ und möchten anhand der Beobachtung ein (nur nicht approximativ korrektes) Konfidenzintervall für $\vartheta \in \Theta = [0,1]$ konstruieren.

Idee: Zu $\vartheta \in \Theta := [0,1]$ wähle $c_\vartheta \in (0,1)$, so dass für

$$C_\vartheta := \left\{x \in \{0,1,\dots,n\} : \mathrm{Bin}_{n,\vartheta}(\{x\}) \geq c_\vartheta\right\}$$

gilt $\mathrm{Bin}_{n,\vartheta}(C_\vartheta) \geq 1-\alpha$ (und c_ϑ möglichst groß, so dass C_ϑ möglichst klein).

Setze $C(x) := \{\vartheta \in \Theta : x \in C_\vartheta\}$ für $x \in \mathcal{X} := \{0,1,\dots,n\}$, dann gilt

$$\forall\, \vartheta \in \Theta :\ P_\vartheta(\vartheta \in C(X)) = P_\vartheta(X \in C_\vartheta) \geq 1-\alpha$$

nach Konstruktion.

Es gilt

1. Für $\vartheta \in (0,1)$ ist $\{0,\dots,n\} \ni x \mapsto \mathrm{Bin}_{n,\vartheta}(\{x\})$ strikt wachsend auf $\{0,1,\dots,\lceil (n+1)\vartheta - 1\rceil\}$, strikt fallend auf $\{\lfloor (n+1)\vartheta\rfloor,\dots,n\}$, also maximal auf $x = \lfloor (n+1)\vartheta\rfloor$ (und auf $(n+1)\vartheta - 1$, wenn $(n+1)\vartheta \in \mathbb{Z}$).

2. Für $x \in \{1,\dots,n\}$ ist $[0,1] \ni \vartheta \mapsto Bin_{n,\vartheta}(\{x, x+1,\dots,n\})$ stetig, strikt monoton wachsend mit

   $$\mathrm{Bin}_{n,\vartheta}(\{x, x+1,\dots,n\}) = \beta_{x,n-x+1}([0,\vartheta]),$$

   wo $\beta_{a,b}$ (die Beta-Verteilung) die Dichte

   $$f_{\mathrm{Beta}_{a,b}}(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}$$

   auf $(0,1)$ hat.

Argument:

1.

$$\begin{aligned}\frac{\mathrm{Bin}_{n,\vartheta}(\{x\})}{\mathrm{Bin}_{n,\vartheta}(\{x-1\})} &= \frac{\binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}}{\binom{n}{x-1}\vartheta^{x-1}(1-\vartheta)^{n-x+1}} \\ &= \frac{(n-x+1)\vartheta}{x(1-\vartheta)} > 1 \qquad \Longleftrightarrow\ x < (n+1)\vartheta\end{aligned}$$

2. $U_1, \dots, U_n$ unabhängig und uniform auf $[0,1]$, so ist

$$S_\vartheta := \sum_{i=1}^{n} \mathbf{1}_{[0,\vartheta]}(U_i)$$

ist $\mathrm{Bin}_{n,\vartheta}$-verteilt.

Sei $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ die „Ordnungsstatistik“.

$$\begin{aligned}
\mathrm{Bin}_{n,\vartheta}(\{x,\dots,n\}) &= P(S_\vartheta \geq x) = P(U_{(x)} \leq \vartheta) \\
&= \sum_{k=1}^{n} \sum_{\substack{B \subseteq \{1,\dots,n\} \setminus \{k\} \\ |B| = x-1}} \underbrace{P\Big(\begin{array}{l} U_k \leq \vartheta, U_m \leq U_k \text{ für } m \in B, \\ U_l > U_k \text{ für } l \in \{1,\dots,n\} \setminus (\{k\} \cup B) \end{array} \Big)}_{= \int_0^\vartheta u^{|B|}(1-u)^{n-|B|-1}\,du = \int_0^\vartheta u^{x-1}(1-u)^{n-x}\,du} \\
&= \underbrace{n\binom{n-1}{x-1}}_{= \frac{n!}{(x-1)!(n-x)!} = \frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)}} \int_0^\vartheta u^{x-1}(1-u)^{n-x}\,du
\end{aligned}$$

Wähle $C_\vartheta := \{x_-(\vartheta), x_-(\vartheta)+1, \dots, x_+(\vartheta)\}$ mit $x_-(\vartheta) = \max\{x : \mathrm{Bin}_{n,\vartheta}(\{0,\dots,x-1\}) \leq \frac{\alpha}{2}\}$ und $x_+(\vartheta) = \min\{x : \mathrm{Bin}_{n,\vartheta}(\{x+1,\dots,n\}) \leq \frac{\alpha}{2}\}$.

Es gilt:

- $x \leq x_+(\vartheta) \iff \mathrm{Bin}_{n,\vartheta}(\{x,\dots,n\}) = \beta_{x,n-x+1}([0,\vartheta]) > \frac{\alpha}{2}$
  $\iff \vartheta > p_-(x) := \frac{\alpha}{2}$-Quantil von $\mathrm{Beta}_{x,n-x+1}$.
- $x \geq x_+(\vartheta) \iff \mathrm{Bin}_{n,\vartheta}(\{0,\dots,x\}) = 1 - \mathrm{Bin}(\{x+1,\dots,n\}) = \beta_{x+1,n-x}([\vartheta,1]) \geq \frac{\alpha}{2}$.
  $\iff \vartheta < p_+(x) := 1 - \frac{\alpha}{2}$-Quantil von $\mathrm{Beta}_{x+1,n-x}$.

Somit haben wir bewiesen:

Satz 6.19 (Exaktes Konfidenzintervall im Binomialmodell)**.**

$$\begin{aligned}
p_-(x) &:= \tfrac{\alpha}{2}\text{-}\textit{Quantil von } \mathrm{Beta}_{x,n-x+1}, \\
p_+(x) &:= 1 - \tfrac{\alpha}{2}\text{-}\textit{Quantil von } \mathrm{Beta}_{x+1,n-x}
\end{aligned}$$

$x \mapsto [p_-(x), p_+(x)]$ *ist ein Konfidenzintervall für* ϑ *zum Sicherheitsniveau* $1-\alpha$.

Bemerkung.

- Quantile der Beta-Verteilungen sind tabelliert, gelegentlich kann beim Nachschlagen in Tabellen die Symmetrieeigenschaft

$$\beta_{a,b}([0,x]) = \beta_{b,a}([1-x,1]) = 1 - \beta_{b,a}([0,1-x])$$

nützlich sein.

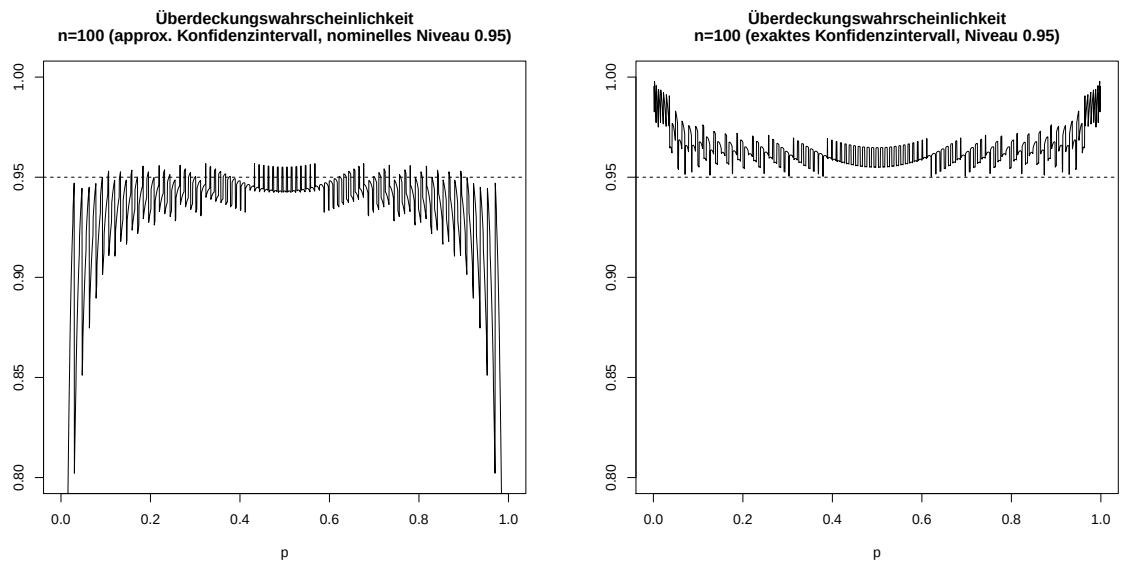


Abbildung 6.1: Überdeckungswahrscheinlichkeit eines (nominellen) 95%-Konfidenzintervalls für den Erfolgsparameter p einer Binomialverteilung mit $n = 100$ als Funktion von p: Links approximatives 95%-Konfidenzintervall (aus Bsp. 6.12), rechts exaktes Konfidenzintervall (aus Satz 6.19). Speziell für p nahe an 0 oder 1 hält das approximative Konfidenzintervall das nominelle Niveau nicht ein.

- R kennt die Beta-Verteilungen, ihre Verteilungsfunktionen `pbeta(x, a, b)` und ihre Quantile `qbeta(p, a, b)`

Für die Daten aus Beispiel 6.1 ($n = 53$, beobachtet $x = 23$) mit $\alpha = 0{,}05$ finden wir $\widehat{\vartheta} = \frac{23}{53} \approx 0{,}434$, $\widehat{\sigma} \approx 0{,}496$, $q_{0{,}975} \approx 1{,}96$, also ist das approximative 95%-Konfidenzintervall für ϑ hier $\left[\widehat{\vartheta} \pm q\frac{\widehat{\sigma}}{\sqrt{53}}\right] \approx [0{,}30, 0{,}57]$.

Es ist $p_-(23) = 0{,}025$-Quantil von $\beta_{23,31} \approx 0{,}30$, $p_+(23) = 0{,}975$-Quantil von $\beta_{24,30} \approx 0{,}57$, d.h. das exakte Konfidenzintervall $[p_-(x), p_+(x)] \approx [0{,}30, 0{,}57]$ (stimmt hier bei Rundung mit obigem überein).

(Die Intervalle sind nicht gleich: Absurd präzise Werte wären: $\left[\widehat{\vartheta} \pm q\frac{\widehat{\sigma}}{\sqrt{53}}\right] \approx [0{,}3005306, 0{,}5673939]$, $[p_-(23), p_+(23)] \approx [0{,}2983921, 0{,}5771742]$.)

Bericht. Manchmal betrachtet man auch den Schätzer

$$\widetilde{\vartheta} = \frac{x+1}{n+2}$$

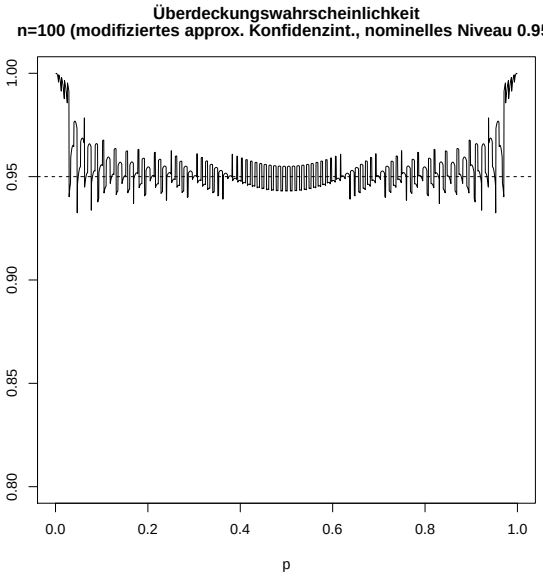
für ϑ und bildet als (approximatives) Konfidenzintervall zum Niveau $1-\alpha$

$$\left[\widetilde{\vartheta} - q\frac{\widetilde{\sigma}}{\sqrt{n}}, \widetilde{\vartheta} + q\frac{\widetilde{\sigma}}{\sqrt{n}}\right]$$

mit $\widetilde{\sigma} = \sqrt{\widetilde{\vartheta}(1-\widetilde{\vartheta})}$, q das $1 - \frac{\alpha}{2}$-Quantil von $\mathcal{N}_{0,1}$.

$\widetilde{\vartheta}$ wirkt vielleicht zunächst „ad hoc“ gewählt. Die Form von $\widetilde{\vartheta}$ ist natürlich im Kontext eines Bayesschen Ansatzes, den wir in dieser Vorlesung nicht weiter diskutieren werden. Die tatsächliche Überdeckungswahrscheinlichkeit dieses approximativen Konfidenzintervalls ist speziell für ϑ nahe an 0

oder 1 besser als die des „klassischen“ approximativen Konfidenzintervalls aus Bsp. 6.12 (vergleiche die Grafik unten mit Abb. 6.1, linke Grafik):



6.4 Statistische Tests

Beispiel 6.20 (für einen einseitiger Binomialtest)**.** Herr A behauptet, (mit W'keit $\vartheta > 1/2$) vorhersagen zu können, ob die oberste Karte eines verdeckten, gut gemischten Skatblatts rot oder schwarz ist.

Frau B ist skeptisch (und verdächtigt, dass A einfach rät, d.h. $\vartheta = 1/2$) und schlägt vor, $n = 20$ Versuche durchzuführen.

Sei

$$X := \text{Anzahl richtige Vorhersagen von A,}$$

wir modellieren $X \sim \mathrm{Bin}_{n,\vartheta}$ (mit uns unbekanntem $\vartheta \in [0,1]$).

B wählt $\alpha = 0{,}05$, sagen wir, und k (möglichst klein) mit (und hier $\vartheta_0 := 1/2$)

$$\mathrm{Bin}_{n,\vartheta_0}\big(\{k, k+1, \ldots, n\}\big) \leq \alpha$$

(hier $k = 15$, denn $\mathrm{Bin}_{n,\vartheta_0}(\{15, 16, \ldots, 20\}) \approx 0{,}021$, $\mathrm{Bin}_{n,\vartheta_0}(\{14, 15, \ldots, 20\}) \approx 0{,}058$) und wird (auf dem Signifikanzniveau α) die

$$\text{Nullhypothese} : \vartheta \leq \frac{1}{2}$$

verwerfen zugunsten der

$$\text{Alternative} : \vartheta > \frac{1}{2},$$

wenn das Ereignis $\{X \geq k\}$ eintritt, ansonsten die Nullhypothese beibehalten.

Demnach: Falls A tatsächlich rät (also in Wirklichkeit $\vartheta = 1/2$ gilt), ist die W'keit, ihm versehentlich hellseherische Fähigkeiten zuzuschreiben (dies wäre dann ein sogenannter „Fehler 1. Art": die Nullhypothese abzulehnen, obwohl sie zutrifft), höchstens α.

Nehmen wir an, die 20 Versuche werden durchgeführt und A erzielt 13 „Treffer". B wird also die Nullhypothese beibehalten (denn $\{X \geq k\}$ tritt nicht ein; quantitativer: der p-Wert ist $P_{1/2}(X \geq 13) \approx 0{,}132 > 0{,}05$; d.h., wenn A einfach nur rät, würde er in ca. 13,2% der Fälle mindestens ebensoviele „Treffer" erzielen wie beobachtet) und etwa sagen:

„Die Beobachtungen zeigen (auf dem Niveau $\alpha = 5\%$) keine keine signifikante Abweichung von der Nullhypothese."

6.4.1 Der formale Rahmen statistischer Tests

Definition 6.21. Sei $(\mathcal{X}, \mathscr{F}, (P_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\Theta = \Theta_0 \dot\cup \Theta_1$ disjunkte Zerlegung in „Nullhypothese" und „Alternative" (auch „Gegenhypothese").

Eine Statistik $\varphi : \mathcal{X} \to [0,1]$ heißt ein *Test* von Θ_0 gegen Θ_1.

Der Test heißt *randomisiert*, wenn $\varphi(X) \notin \{0,1\}$, sonst *nicht-randomisiert*,

für einen nicht randomisierten Test φ heißt

$$\{x : \varphi(x) = 1\} \text{ der Ablehnungs- oder Verwerfungsbereich (von } \Theta_0).$$

$$G_\varphi : \Theta \to [0,1], \quad G_\varphi(\vartheta) = \mathbb{E}_\vartheta[\varphi]$$

heißt die *Gütefunktion* von φ ($1 - G_\varphi$ heißt Operationscharakteristisk), φ heißt ein Test zum Niveau (auch: Signifikanzniveau) $\alpha \in (0,1)$, wenn gilt

$$\sup_{\vartheta\in\Theta_0} G_\varphi(\vartheta) \leq \alpha.$$

$\sup_{\vartheta\in\Theta_0} G_\varphi(\vartheta)$ heißt effektives Niveau (auch: Umfang) von φ.

Für $\vartheta \in \Theta_1$ heißt $G_\varphi(\vartheta)$ die Macht (auch: Schärfe, englisch: power) des Tests φ bei ϑ.

Sei $(\varphi_\alpha)_{\alpha\in(0,1)}$ eine Familie von nicht-randomisierten Tests mit $\varphi_\alpha \leq \varphi_{\alpha'}$ für $\alpha \leq \alpha'$, φ_α habe effektives Niveau α. Dann heißt für $x \in \mathcal{X}$

$$p\,\big(=p(x)\big) = \inf\{\alpha \in (0,1) : \varphi_\alpha(x) = 1\}$$

der *p-Wert* (bei Beobachtung x).

Interpretation.

1. Man interpretiert φ als Entscheidungsregel: Bei gegebener Beobachtung x

$$\begin{aligned}
\varphi(x) = 0 &\quad : \quad \text{behalte Nullhypothese bei} \\
\varphi(x) = 1 &\quad : \quad \text{verwirf Nullhypothese, entscheide für die Alternative} \\
\varphi(x) \in (0,1) &\quad : \quad \text{wirf eine Münze, die mit W'keit } \varphi(x) \\
&\qquad\quad\ \ \text{für die Alternative entscheidet}
\end{aligned}$$

2. Niveau α bedeutet, dass die W'keit für einen „Fehler 1. Art" (die Nullhypothese fälschlicherweise zu verwerfen) $\leq \alpha$ ist (uniform in $\vartheta \in \Theta_0$).

3. Für $\vartheta \in \Theta_1$ ist $1 - G_\varphi(\vartheta)$ die Wahrscheinlichkeit, einen „Fehler 2. Art" zu begehen (die Nullhypothese fälschlicherweise zu akzeptieren).

4. Viele „praktische" Tests haben die folgende Form, z.B. Bspe. 6.23, 6.24, 6.25, 6.32: Berechne eine gewisse (Test-)Statistik Y (aus den Beobachtungen), verwirf die Nullhypothese, wenn $Y > q$ für einen gewissen Wert $q = q(\alpha)$, der in Abhängigkeit von den Parametern des Tests (insbesondere dem gewünschten Niveau α) gewählt wird. In der Sprache von Definition 6.21 also: $\varphi_\alpha(x) = \mathbf{1}(Y(x) > q(\alpha))$ und $\sup_{\vartheta\in\Theta_0} \mathbb{E}_\vartheta[\varphi] = \alpha$.

   Dann kann man den p-Wert des Test(ergebnisses) interpretieren als die Wahrscheinlichkeit, bei Gültigkeit der Nullhypothese einen mindestens so „extremen" Wert der Teststatistik zu finden wie den tatsächlich anhand der Daten beobachteten.

Demnach sind für φ wünschenswert:

$$G_\varphi \text{ sollte auf } \Theta_1 \text{ möglichst groß sein}$$

(solange mit dem gewünschten Signifikanzniveau verträglich), zudem sollte für einen Test zum Niveau α gelten

$$\sup_{\vartheta\in\Theta_0} G_\varphi(\vartheta) \leq \alpha \leq \sup_{\vartheta\in\Theta_1} G_\varphi(\vartheta) \qquad \text{(dann heißt } \varphi \text{ „unverfälscht").}$$

Beispiel 6.22 (Binomialtest)**.** $\Theta = [0,1]$, unter P_ϑ sei die Beobachtung $X \sim \mathrm{Bin}_{n,\vartheta}$ (oder aber Beobachtungen $X_1, \dots, X_n$ sind unter P_ϑ u.i.v. $\sim \mathrm{Ber}_\vartheta$ und wir bilden $X := X_1 + \cdots + X_n$). Wähle $\alpha \in (0, 1/2)$.

1. Zweiseitiger Binomialtest: $\Theta_0 = \{\vartheta_0\}$ für ein $\vartheta_0 \in [0,1]$, $\Theta_1 = \Theta \setminus \Theta_0$. Setze

$$\begin{aligned} c_\ell &:= \max\big\{x \in \{0,1,2,\dots,n\} : \mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\}) \leq \alpha/2\big\}, \\ c_r &:= \min\big\{x \in \{0,1,2,\dots,n\} : \mathrm{Bin}_{n,\vartheta_0}(\{x,x+1,\dots,n\}) \leq \alpha/2\big\}, \\ \varphi(x) &:= \mathbf{1}_{\{0,1,\dots,c_\ell\}}(x) + \mathbf{1}_{\{c_r,c_r+1,\dots,n\}}(x). \end{aligned}$$

Dann gilt

$$\begin{aligned} \mathbb{E}_{\vartheta_0}[\varphi(X)] &= P_{\vartheta_0}(X \leq c_\ell) + P_{\vartheta_0}(X \geq c_r) \\ &= \mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,c_\ell\}) + \mathrm{Bin}_{n,\vartheta_0}(\{c_r, c_r+1,\dots,n\}) \leq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha \end{aligned}$$

nach Konstruktion, d.h. der Test hält Niveau α ein. (Wegen der Diskretheit der möglichen Beobachtungen ist das tatsächliche Niveau i.A. etwas kleiner.)

Bei gegebener Beobachtung x ist der p-Wert dann $2\mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\})$ falls $x < n\vartheta_0$ und $2\mathrm{Bin}_{n,\vartheta_0}(\{x,x+1,\dots,n\})$ falls $x > n\vartheta_0$.

2. a) Einseitiger Binomialtest (linksseitige Alternative): $\Theta_0 = [\vartheta_0, 1]$ für ein $\vartheta_0 \in (0,1]$, $\Theta_1 = [0,\vartheta_0) = \Theta \setminus \Theta_0$. Setze

$$\begin{aligned} c &:= \max\big\{x \in \{0,1,2,\dots,n\} : \mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\}) \leq \alpha\big\}, \\ \varphi(x) &:= \mathbf{1}_{\{0,1,\dots,c\}}(x). \end{aligned}$$

Nach Konstruktion ist $\mathbb{E}_{\vartheta_0}[\varphi(X)] = P_{\vartheta_0}(X \le c) \le \alpha$ und man kann (leicht) zeigen, dass für $\vartheta > \vartheta_0$ gilt $P_\vartheta(X \le c) \le P_{\vartheta_0}(X \le c)$ $(\le \alpha)$, d.h. der Test hält Niveau α ein.

Bei gegebener Beobachtung x ist der p-Wert dann $\mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\})$.

2. b) Einseitiger Binomialtest (rechtssseitige Alternative): $\Theta_0 = [0,\vartheta_0]$ für ein $\vartheta_0 \in [0,1)$, $\Theta_1 = (\vartheta_0,1] = \Theta \setminus \Theta_0$. Analog setze

$$C := \min\big\{x \in \{0,1,2,\dots,n\} : \mathrm{Bin}_{n,\vartheta_0}(\{x,x,+1,\dots,n\}) \le \alpha\big\},$$
$$\varphi(x) := \mathbf{1}_{\{C,C+1,\dots,n\}}(x).$$

Der Test hält Niveau α ein, bei gegebener Beobachtung x ist der p-Wert dann $\mathrm{Bin}_{n,\vartheta_0}(\{x,x,+1,\dots,n\})$.

Bemerkung. Offenbar benötigt man die Verteilungsfunktion der Binomialverteilung, um die „kritischen Werte“ c_ℓ, c_r bzw. c, C für den Binomialtest bei vorgegebenem n und α zu bestimmen. Für kleine Werte von n kann man diese „von Hand“ bestimmen, für größere Werte konsultiert man entweder ein Computerprogramm oder eine entsprechende Tabelle oder man verwendet die Normalapproximation der Binomialverteilung: Mit Satz 5.1 und Korollar 5.2 ist

$$\mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\}) \approxeq \Phi\left(\frac{x - n\vartheta_0}{\sqrt{n\vartheta_0(1-\vartheta_0)}}\right)$$

wobei $\Phi = F_{\mathcal{N}_{0,1}}$ die Verteilungsfunktion der Standard-Normalverteilung ist.

Beispiel 6.23 (z-Test oder Gauß-Test)**.** $\Theta = \mathbb{R}$, unter P_ϑ seien die Beobachtungen $X_1,\dots,X_n$ u.i.v. $\sim \mathcal{N}_{\vartheta,\sigma^2}$ mit bekanntem, festem $\sigma^2 > 0$. Wähle $\alpha \in (0,1)$.

1. Zweiseitiger z-Test: $\Theta_0 = \{\vartheta_0\}$ für ein $\vartheta \in \mathbb{R}$, $\Theta_1 = \Theta \setminus \Theta_0$

$$M := \frac{1}{n}\sum_{i=1}^n X_i, \quad Z := \frac{M - \vartheta_0}{\sqrt{\sigma^2/n}}$$

mit $q := \Phi^{-1}(1-\alpha/2)$ (das $(1-\alpha/2)$-Quantil der Standardnormalverteilung) ist

$$\varphi(X_1,\dots,X_n) := \mathbf{1}_{\{|Z|>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α (denn unter P_{ϑ_0} ist $M \sim \mathcal{N}_{\vartheta_0,\sigma^2/n}$).

Der p-Wert ist dann $2\big(1-\Phi(|Z|)\big)$, wobei Φ die Verteilungsfunktion der Standard-Normalverteilung ist.

2. Einseitiger Test: $\Theta_0 = \{\vartheta : \vartheta \le \vartheta_0\}$ für ein $\vartheta \in \mathbb{R}$, $\Theta_1 = \Theta \setminus \Theta_0 = \{\vartheta : \vartheta > \vartheta_0\}$.

Mit $q := \Phi^{-1}(1-\alpha)$ ist

$$\varphi(X_1,\dots,X_n) := \mathbf{1}_{\{Z>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α. Als p-Wert ergibt sich $1 - \Phi(Z)$

(Je nach Anwendungssituation kann man auch $\Theta_0 = \{\vartheta : \vartheta \ge \vartheta_0\}$ betrachten, dann ist $\varphi(X_1,\dots,X_n) := \mathbf{1}_{\{Z<-q\}}$ zu wählen und der p-Wert wäre $\Phi(Z) = 1 - \Phi(-Z)$).

Beispiel 6.24 ((ein-Stichproben- oder gepaarter) t-Test)**.** $\Theta = \mathbb{R} \times (0,\infty) \ni \vartheta = (\mu, \sigma^2)$, unter P_ϑ seien die Beobachtungen $X_1, \dots, X_n$ u.i.v. $\sim \mathcal{N}_{\mu,\sigma^2}$ (mit unbekanntem $\mu \in \mathbb{R}$ und unbekanntem $\sigma^2 > 0$). Wähle $\alpha \in (0,1)$.

Sei $M := \dfrac{1}{n}\sum_{i=1}^n X_i$, $S^2 := \dfrac{1}{n-1}\sum_{i=1}^n (X_i - M)^2$.

1. Zweiseitiger (ein-Stichproben) t-Test: $\Theta_0 = \{\vartheta = (\mu,\sigma^2) \in \Theta : \mu = \mu_0\}$ für ein $\mu_0 \in \mathbb{R}$ (man schreibt dies oft knapp als „$\Theta_0 : \mu = \mu_0$“), $\Theta_1 = \Theta \setminus \Theta_0$.

$$T := \sqrt{n}\frac{M - \mu_0}{\sqrt{S^2}}$$

Mit $q := q_{n-1,1-\alpha/2} = (1-\alpha/2)$-Quantil der Student-$(n-1)$-Verteilung ist

$$\varphi(X_1, \dots, X_n) := \mathbf{1}_{\{|T|>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α (denn nach Satz 6.16 ist T für jedes $\vartheta \in \Theta_0$ unter P_ϑ Student-$(n-1)$-verteilt).

Der p-Wert ist $2\big(1 - F_{T_{n-1}}(|T|)\big)$ mit $F_{T_{n-1}}$ der Verteilungsfunktion der Student-$(n-1)$-Verteilung.

2. Einseitiger Test: $\Theta_0 = \{\vartheta = (\mu,\sigma^2) \in \Theta : \mu \le \mu_0\}$ für ein $\mu_0 \in \mathbb{R}$ (oft knapp geschrieben als „$\Theta_0 : \mu \le \mu_0$“), $\Theta_1 = \Theta \setminus \Theta_0$. Mit $q := q_{n-1,1-\alpha} = (1-\alpha)$-Quantil der Student-$(n-1)$-Verteilung ist

$$\varphi(X_1, \dots, X_n) := \mathbf{1}_{\{T>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α (und analog $\varphi(X_1, \dots, X_n) := \mathbf{1}_{\{T<-q\}}$ ein Test für $\Theta_0 = \{\mu \ge \mu_0\}$, beachte auch: $-q$ ist das α-Quantil der Student-$(n-1)$-Verteilung).

Der p-Wert ist $1 - F_{T_{n-1}}(T)$.

(Je nach Anwendungssituation kann man auch $\Theta_0 = \{\vartheta = (\mu,\sigma^2) \in \Theta : \mu \ge \mu_0\}$ betrachten, dann ist $\varphi(X_1, \dots, X_n) := \mathbf{1}_{\{T<-q\}}$ zu wählen und der p-Wert wäre $F_{T_{n-1}}(T) = 1 - F_{T_{n-1}}(-T)$).

Anwendungsbeispiel. a) Die Wirksamkeit eines gewissen Schlafmittels soll geprüft werden. 10 Patienten erhalten das Schlafmittel, die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet.

Wir nehmen an, die Beob. sind u.i.v. $\sim \mathcal{N}_{\mu,\sigma^2}$ und wir möchten die Nullhypothese $\mu = 0$, sagen wir, zum Niveau $\alpha = 0{,}05$ testen.

Die Daten[5]:

Patient i	1	2	3	4	5	6	7	8	9	10
zus. Schl.	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0

Es ist $n = 10$, $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 0{,}75$, $s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 \approxeq 1{,}79$, $t = \frac{\overline{x}-0}{s/\sqrt{10}} \approxeq 1{,}326$

Das $0{,}975$-Quantil der Student-9-Verteilung ist $\approxeq 2{,}262$, demnach können wir die Nullhypothese nicht ablehnen.

(Für ein Student-9-verteiltes T ist $P(|T| \ge 1{,}326) \approxeq 0{,}2176$, dies ist der p-Wert des Tests.)

[5] Aus Student (William S. Gosset), The Probable Error of a Mean, Biometrika 6:1–25 (1908)

Man kann diesen Befund folgendermaßen formulieren:

„Die Beobachtungen sind mit der Nullhypothese $\mu = 0$ (im statistischen Sinne) verträglich."

oder

„Die beobachtete Abweichung $\overline{x} = 0{,}75$ ist nicht signifikant von 0 verschieden (t-Test, $\alpha = 0{,}05$)."

b) Die Wirksamkeit eines Schlafmittels soll mit der eines anderen verglichen werden. 10 Patienten erhalten Schlafmittel A, die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet. Dann erhalten dieselben 10 Patienten Schlafmittel B, wieder wird die Anzahl zusätzlicher Stunden Schlaf in einer Nacht beobachtet.

Da dieselben Patienten untersucht werden, können (und sollten) wir die Messungen paaren: Wir interessieren uns bei jedem Patienten für die Differenz des (zusätzlichen) Schlafs bei Mittel 2 und bei Mittel 1.

Wir nehmen an, die beobachteten Differenzen sind Realisierungen von u.i.v. ZVn mit Vert. $\mathcal{N}_{\mu,\sigma^2}$ und wir möchten die Nullhypothese $\mu \leq 0$ gegen die Alternative $\mu > 0$, sagen wir, zum Niveau $\alpha = 0{,}05$ testen.

(Dies wäre beispielsweise in folgender Situation angemessen: Wir möchten darlegen, dass Mittel B wirksamer ist als Mittel A, indem wir die Nullhypothese „$\mu \leq 0$" entkräften.)

Die Daten (wiederum aus Student, a.a.O.) :

Patient i	1	2	3	4	5	6	7	8	9	10
Mittel A	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
Mittel B	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4
Diff.	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

Es ist $n = 10$, $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 1{,}58$, $s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 \approxeq 1{,}23$, $t = \frac{\overline{x}-0}{s/\sqrt{10}} \approxeq 4{,}062$

Das $0{,}95$-Quantil der Student-9-Verteilung ist $\approxeq 1{,}833$, demnach können wir die Nullhypothese ablehnen.

(Für ein Student-9-verteiltes T ist $P(T > 4{,}062) \approxeq 0{,}0014$, dies ist der p-Wert des Tests.)

Mögliche knappe Formulierung dieses Befunds:

„Die beobachtete Differenz $\overline{x} = 1{,}58$ ist signifikant größer als 0 (einseitiger t-Test, $\alpha = 0{,}05$)."

Beispiel 6.25 (Test für die Varianz im normalen Modell)**.** In der Situation von Beispiel 6.24 sei $\Theta_0 = \{\vartheta = (\mu, \sigma^2) \in \Theta : \sigma^2 \leq v_0\}$ für ein $v_0 > 0$, $\Theta_1 = \Theta \setminus \Theta_0$. Wähle $\alpha \in (0, 1)$.

Mit $q :=$ $(1-\alpha)$-Quantil der χ^2_{n-1}-Verteilung ist

$$\varphi(X_1, \ldots, X_n) := \mathbf{1}_{\{S^2 > qv_0/(n-1)\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α (vgl. Satz 6.16).

Beispiel 6.26 (zwei-Stichproben oder ungepaarter t-Test [mit Annahme gleicher Varianzen])**.** $\Theta = \mathbb{R}\times\mathbb{R}\times(0,\infty) \ni \vartheta = (\mu_1, \mu_2, \sigma^2)$, unter P_ϑ sind $X_1, \ldots, X_m$ u.i.v. und davon unabhängig $Y_1, \ldots, Y_n$ u.i.v. $(m, n \in \mathbb{N})$, $X_i \sim \mathcal{N}_{\mu_1,\sigma^2}$, $Y_j \sim \mathcal{N}_{\mu_2,\sigma^2}$. Seien

$$M_X := \frac{1}{m}\sum_{i=1}^{m} X_i,\ M_Y := \frac{1}{n}\sum_{j=1}^{n} Y_j$$

die jeweiligen Stichprobenmittelwerte,

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - M_X)^2,\ S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - M_Y)^2,$$

die (korrigierten) Stichprobenvarianzen,

$$S^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \quad \left(= \frac{1}{m+n-2}\Big(\sum_{i=1}^{m}(X_i - M_X)^2 + \sum_{j=1}^{n}(Y_j - M_Y)^2\Big)\right),$$

(die „gepoolte Stichprobenvarianz"),

$$T = \frac{M_X - M_Y}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

(Beachte: Stets gilt $\mathbb{E}_{(\mu_1,\mu_2,\sigma^2)}[S^2] = \sigma^2$ [S^2 ist ein erwartungstreuer Schätzer für σ] und T ist unter $P_{(\mu_1,\mu_1,\sigma^2)}$ Student-$(n+m-2)$-verteilt, Argument analog zum Beweis von Satz 6.16).

1. Zweiseitiger ungepaarter t-Test : $\Theta_0 = \{(\mu_1,\mu_2,\sigma^2) \in \Theta : \mu_1 = \mu_2\}$ (oft knapp geschrieben als „$\Theta_0 : \mu_1 = \mu_2$"), $\Theta_1 = \Theta \setminus \Theta_0$.

Wähle $\alpha \in (0,1)$, mit $q := q_{m+n-2,1-\alpha/2} = (1-\alpha/2)$-Quantil der Student-$(m+n-2)$-Verteilung ist

$$\varphi(X_1,\dots,X_m,Y_1,\dots,Y_n) := \mathbf{1}_{\{|T|>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α.

1. Einseitiger Test : $\Theta_0 = \{(\mu_1,\mu_2,\sigma^2) \in \Theta : \mu_1 \le \mu_2\}$ (oft knapp geschrieben als „$\Theta_0 : \mu_1 \le \mu_2$"), $\Theta_1 = \Theta \setminus \Theta_0$. Mit $q := q_{m+n-2,1-\alpha} = (1-\alpha)$-Quantil der Student-$(m+n-2)$-Verteilung ist

$$\varphi(X_1,\dots,X_m,Y_1,\dots,Y_n) := \mathbf{1}_{\{T>q\}}$$

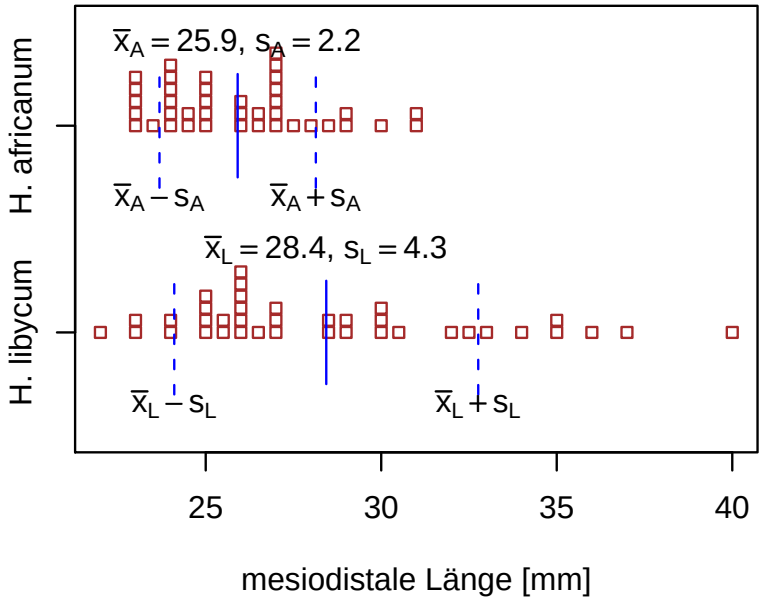
ein Test von Θ_0 gegen Θ_1 zum Niveau α.

(Analog ist $\varphi(X_1,\dots,X_m,Y_1,\dots,Y_n) := \mathbf{1}_{\{T<-q\}}$ ein Test von $\Theta_0 = \{(\mu_1,\mu_2,\sigma^2) \in \Theta : \mu_1 \ge \mu_2\}$.)

(p-Werte werden analog zum ein-Stichproben-Fall (Bsp. 6.24) berechnet, wobei $F_{T_{n-1}}$ durch $F_{T_{m+n-2}}$ ersetzt wird.)

Anwendungsbeispiel. Es wurden fossile Backenzähne gefunden, die zwei Arten von Urpferden zugeordnet wurden, und jeweils die („mesiodistale") Länge bestimmt. Wir möchten die (Null-)Hypothese prüfen, ob die mittlere Zahnlänge bei den beiden Arten gleich ist.

Die Daten:



Für *Hipparion africanum*: $n_A = 39$, $\overline{x}_A = 25{,}9$, $s_A = 2{,}2$

für *Hipparion lybicum*: $n_L = 38$, $\overline{x}_L = 28{,}4$, $s_L = 4{,}3$

Wir verwenden Signifikanzniveau $\alpha = 0{,}01$, das 99,5%-Quantil der Student-Vert. mit 75 Freiheitsgraden ist $\approx 2{,}64$. Es ist

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_L - 1)s_L^2}{n_A + n_L - 1} \approx 11{,}94, \quad t = \frac{\overline{x}_A - \overline{x}_L}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_L}}} \approx -3{,}229,$$

Wir können die Nullhypothese „die mittlere mesiodistale Länge bei *H. lybicum* und bei *H. africanum* sind gleich“ zum Signifikanzniveau 1% ablehnen.

(Für ein Student-75-verteiltes T ist $P(|T| > 3{,}229) \approx 0{,}0018$, dies ist der p-Wert des Tests.)

Mögliche Formulierung dieses Befunds: „Die mittlere mesiodistale Länge war signifikant größer (28,4 mm) bei *H. libycum* als bei *H. africanum* (25,9 mm) (t-Test, $\alpha = 0{,}01$).“

Bericht 6.27. Die Tests aus Bsp. 6.23–6.26 sind (in ihrem jeweiligen statistischen Modell) optimal (sie sind „gleichmäßig beste Tests“) in dem Sinne, dass sie unter allen allen Tests von Θ_0 gegen Θ_1 mit Niveau α die größte Macht haben.

Dies ist das „Test-Analogon“ zu Bericht 6.8, vgl. z.B. die Diskussion in [G, Kap. 10.3 und 10.4].

Bericht 6.28 (zwei-Stichproben-t-Test ohne Annahme gleicher Varianz, Welchs t-Test[6])**.** Es gibt auch eine Version des zwei-Stichproben-t-Tests, der die Annahme gleicher Varianzen nicht trifft (wir werden ihn im Verlauf der Vorlesung allerdings nicht verwenden):

[6] B. L. Welch, The Significance of the Difference between Two Means When the Population Variances Are Unequal, Biometrika 29:350–362, (1938)

Man schätzt die Streuung von $M_X - M_Y$ durch

$$\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} \quad \text{und bildet} \quad T = \frac{M_X - M_Y}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}.$$

Unter $P_{(\mu_0,\mu_0,\sigma_1^2,\sigma_2^2)}$ ist T „approximativ Student-verteilt mit g Freiheitsgraden“, wobei

$$g = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{s_X^4}{n_X^2(n_X-1)} + \frac{s_Y^4}{n_Y^2(n_Y-1)}}$$

aus den Daten geschätzt wird.

Seien die Werte $T = t$ und g beobachtet worden, man verwirft die Nullhypothese „$\mu_1 = \mu_2$“ (zum Niveau α), wenn $1 - F_{T_g}(t) \leq \alpha/2$, wobei F_{T_g} die Verteilungsfunktion der Student-Verteilung mit g Freiheitsgraden, d.h. wenn die Wahrscheinlichkeit, dass eine Student-verteilte Zufallsgröße mit g Freiheitsgraden einen betragsmäßig mindestens so großen Wert wie den beobachteten t-Wert annimmt, $\leq \alpha$ ist.

(Wir hatten in Korollar 6.15 die Student-Verteilung nur für ganzzahlige Werte von n definiert, aber man kann dort allgemeine Werte $n > 0$ zulassen).

Dieser Test hat approximativ Niveau α und wird in der Praxis häufig verwendet. Beispielsweise führt der Befehl `t.test` in dem Statistikprogramm R automatisch diese Version des zwei-Stichproben-t-Tests durch, wenn man zwei Stichproben übergibt und keine weiteren Zusatzparameter setzt.

Bemerkung 6.29 (Zur „reinen Lehre“ des statistischen Testens)**.** Nehmen wir an, wir möchten eine gewisse Aussage anhand experimenteller oder empirischer Daten statistisch prüfen. Das korrekte („lehrbuchmäßige“) Vorgehen sieht folgendermaßen aus:

1. Statistisches Modell formulieren, Nullhypothese und Alternative angeben (was die Nullhypothese ist, hängt von der konkreten Anwendungsfrage ab, oft ernennt man „das Gegenteil dessen, was man erhärten möchte“ zur Nullhypothese).

2. Dann einen Test (einschließlich gewünschtem Niveau) festlegen.

3. Dann erst: Daten erheben (bzw. Daten anschauen), Test-Entscheidung fällen.

Die Kontrolle der Fehlerwahrscheinlichkeiten, die die Theorie des statistischen Testens liefert, bezieht sich auf dieses Vorgehen. Wenn man die Reihenfolge herumdreht, also zuerst die Daten anschaut und dann einen Test wählt, verfälscht man strenggenommen zumindest das Signifikanzniveau, möglicherweise bis ins Unsinnige (Beispiel: zuerst den empirischen Mittelwert bestimmen, dann je nachdem, ob er links oder rechts von ϑ_0 liegt, entscheiden, ob man eine rechts- oder eine linksseitige Alternative wählt, ist offenbar „geschummelt“.)

Man sollte dieselben Daten nicht für explorative Statistik (d.h. Beobachtungen, die zu neuen Hypothesen führen [sollen]) und schließende Statistik (d.h. Beobachtungen, anhand denen eine Hypothese getestet werden soll) zugleich verwenden.

6.4.2 Alternativtests und das Lemma von Neyman-Pearson*

Wir betrachten ein Standardmodell (vgl. Def. 6.2) mit jeweils einpunktiger Nullhypothese und Alternative, d.h. $\Theta = \{0, 1\}$, $\mathcal{X} \subset \mathbb{R}^n$ oder $\mathcal{X}$ diskret und P_i besitzt Dichte bzw. $\rho(x, i)$ auf $\mathcal{X}$ für $i = 0, 1$.

Setze

$$R(x) := \begin{cases} \dfrac{\rho(x,1)}{\rho(x,0)} & \text{wenn } \rho(x,0) > 0, \\ \infty & \text{sonst.} \end{cases}$$

R heißt der *Likelihood-Quotient*.

Ein Test von P_0 gegen P_1 (formal hier wörtlich $\Theta_0 = \{0\}$ gegen $\Theta_1 = \{1\}$) der Form

$$\varphi(x) = \begin{cases} 1 & \text{für } R(x) > c, \\ 0 & \text{für } R(x) < c, \end{cases}$$

für ein $c \geq 0$ heißt ein Neyman-Pearson-Test[7]. (Im Fall $R(x) = c$ kann Randomisierung notwendig sein.)

Satz 6.30 (Neyman-Pearson-Lemma)**.** *Betrachte Standardmodell mit einpunktiger Nullhypothese und einpunktiger Alternative, $\alpha \in (0,1)$.*

1. Es gibt einen Neyman-Pearson-Test φ mit $\mathbb{E}_0[\varphi] = \alpha$.

2. Sei φ ein Neyman-Pearson-Test φ mit $\mathbb{E}_0[\varphi] = \alpha$, $\widetilde{\varphi}$ irgendein Test von P_0 gegen P_1 zum Niveau α. Dann gilt $\mathbb{E}_1[\varphi] \geq \mathbb{E}_1[\widetilde{\varphi}]$, d.h. die Macht von φ ist mindestens so groß wie die von $\widetilde{\varphi}$.

Man sagt: φ ist (in dieser Situation) ein gleichmäßig bester *Test.*

Beweis. 1. Wähle c mit $P_0(R \geq c) \geq \alpha$ und $P_0(R \leq c) \geq 1 - \alpha$.

Falls $P_0(R = c) = 0$, so ist $\varphi(x) := \mathbf{1}_{\{R(x)>c\}}$ ein Neyman-Pearson-Test mit $\mathbb{E}_0[\varphi] = P_0(R > c) = P_0(R \geq c) = \alpha$.

Falls $P_0(R = c) > 0$, so setze $\gamma := \dfrac{\alpha - P_0(R > c)}{P_0(R = c)}$ $(\in [0,1])$ und

$$\varphi(x) := \begin{cases} 1 & \text{wenn } R(x) > c, \\ \gamma & \text{wenn } R(x) = c, \\ 0 & \text{wenn } R(x) < c. \end{cases}$$

Dies ist ein Neyman-Pearson-Test mit $\mathbb{E}_0[\varphi] = 1 \cdot P_0(R > c) + \gamma P_0(R = c) + 0 \cdot P_0(R < c) = \alpha$.

2. Sei φ ein Neyman-Pearson-Test (mit Schwellenwert c) mit $\mathbb{E}_0[\varphi] = \alpha$, $\widetilde{\varphi}$ irgendein Test mit $\mathbb{E}_0[\widetilde{\varphi}] \leq \alpha$. Es gilt

$$\text{für alle } x \in \mathcal{X}: \quad \big(\varphi(x) - \widetilde{\varphi}(x)\big)\big(\rho(x,1) - c\rho(x,0)\big) \geq 0$$

denn die beiden Faktoren haben dasselbe Vorzeichen (sofern der zweite $\neq 0$), da $\varphi(x) \geq \mathbf{1}(\rho(x,1) > c\rho(x,0))$. Somit

$$f_1(x) := \big(\varphi(x) - \widetilde{\varphi}(x)\big)\rho(x,1) \geq c\big(\varphi(x) - \widetilde{\varphi}(x)\big)\rho(x,0) =: c f_0(x) \tag{6.2}$$

und folglich

$$\mathbb{E}_1[\varphi] - \mathbb{E}_1[\widetilde{\varphi}] = \int_{\mathcal{X}} f_1(x)\,dx \geq c \int_{\mathcal{X}} f_0(x)\,dx = c(\alpha - \mathbb{E}_0[\widetilde{\varphi}]) \geq 0. \tag{6.3}$$

(Wenn $\mathcal{X}$ diskret ist, muss das Integral natürlich durch eine Summe ersetzt werden.) □

[7] nach Jerzy Neyman, 1894–1981 und Egon Pearson, 1895–1980

Bemerkung. Wir sehen aus dem Beweis auch: Wenn $\widetilde{\varphi}$ ebenfalls ein gleichmäßig bester Test von P_0 gegen P_1 mit $\mathbb{E}_0[\widetilde{\varphi}] = \alpha$ ist, so gilt Gleichheit in (6.3) und daher auch Gleichheit in (6.2) (möglicherweise mit Ausnahme einer Menge von [Lebesgue-]Maß 0). In diesem Sinne ist also hier ein gleichmäßig bester Test „identisch“ mit einem Neyman-Pearson-Test.

Beispiel. Beobachtungen $X_1, \dots, X_n$ seien unter P_0 u.i.v. mit $X_i \sim \mathcal{N}_{\mu_0,\sigma^2}$, unter P_1 u.i.v. mit $X_i \sim \mathcal{N}_{\mu_1,\sigma^2}$, wobei $\sigma^2 > 0$ und $\mu_0 < \mu_1$ bekannt (und fest) sind.

Mit $x = (x_1, \dots, x_n)$ und $\overline{x} := \frac{1}{n}\sum_{i=1}^n x_i$ ist

$$\begin{aligned} R(x) &= \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2 \Big) \\ &= \exp\Big(-\frac{n}{2\sigma^2}\big(2(\mu_1-\mu_0)\overline{x} + \mu_1^2 - \mu_0^2\big)\Big). \end{aligned}$$

Offenbar ist $R(x)$ eine montone (fallende) Funktion von $\overline{x}$ und unter P_0 ist $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}_{\mu_0,\sigma^2/n}$, also ist

$$\varphi(x) := \mathbf{1}_{\{\overline{x}>c\}}$$

mit der Wahl $c := \mu_0 + \sqrt{\sigma^2/n}\,\Phi^{-1}(1-\alpha)$ ein Neyman-Pearson-Test zum Niveau $\alpha \in (0,1)$ (denn dann ist $\mathbb{E}_0[\varphi] = P_0\big(\overline{X} > c\big) = \alpha$).

6.4.3 Tests für kategorielle Beobachtungen (zum χ^2-Test)*

Ein Experiment mit s möglichen Ausgängen werde n mal (unabhängig) wiederholt, Ausgang i habe die (unbekannte) Wahrscheinlichkeit ϑ_i, $i = 1, \dots, s$.

Angenommen, wir beobachten h_i-mal Ausgang i für $i = 1, \dots, s$. Passt dies zur (Null-)Hypothese, dass

$$\vartheta = (\vartheta_1, \dots, \vartheta_s) = (\rho_1, \dots, \rho_s) = \rho$$

gilt für einen vorgegebenen Vektor ρ von Wahrscheinlichkeitsgewichten (auf $\{1, \dots, s\}$)?

Satz 6.31. *Sei* $\rho \in \Delta_s := \{(\vartheta_1, \dots, \vartheta_s) \in [0,1]^s : \vartheta_1 + \dots + \vartheta_s = 1\}$,

$$(H_1^{(n)}, \dots, H_s^{(n)}) \sim \mathrm{Mult}_{n;\rho_1,\dots,\rho_s},$$

dann gilt

$$\sum_{i=1}^s \frac{(H_i^{(n)} - n\rho_i)^2}{n\rho_i} \xrightarrow[n\to\infty]{d} \chi^2_{s-1}$$

Beweisskizze. Seien $X_1^{(n)}, \dots, X_s^{(n)}$ u.a., $X_i^{(n)} \sim \mathrm{Poi}_{n\rho_i}$, dann ist

$$N_n := X_1^{(n)} + \dots + X_s^{(n)} \sim \mathrm{Poi}_n$$

(siehe Bsp. 2.26, 3.)

Beachte: Für $m \in \mathbb{N}$, $h_1, \dots, h_s \in \mathbb{N}_0$ mit $h_1 + \dots + h_s = m$ ist

$$\begin{aligned} &P\big(X_1^{(n)} = h_1, \dots, X_s^{(n)} = h_s \,\big|\, N_n = m\big) \\ &= \Big(e^{-n}\frac{n^m}{m!}\Big)^{-1} \prod_{i=1}^s e^{-n\rho_i}\frac{(n\rho_i)^{h_i}}{h_i!} = \binom{m}{h_1, h_2, \dots, h_s}\rho_1^{h_1}\cdots\rho_s^{h_s}, \end{aligned}$$

d.h. bedingt auf $\{N_n = m\}$ ist $(X_1^{(n)}, \dots, X_s^{(n)}) \sim \mathrm{Mult}_{m;\rho_1,\dots,\rho_s}$.

Sei

$$\widetilde{X}_i^{(n)} := \frac{X_i^{(n)} - n\rho_i}{\sqrt{n\rho_i}}, \quad \widetilde{H}_i^{(n)} := \frac{H_i^{(n)} - n\rho_i}{\sqrt{n\rho_i}}$$

(beachte: $\mathbb{E}[\widetilde{X}_i^{(n)}] = 0$, $\mathrm{Var}[\widetilde{X}_i^{(n)}] = 1$),

$$\widetilde{N}_n := \frac{N_n - n}{\sqrt{n}} = \sum_{i=1}^{s} \sqrt{\rho_i}\widetilde{X}_i^{(n)}$$

(beachte: $N_n = n \iff \widetilde{N}_n = 0 \iff (\widetilde{X}_1^{(n)}, \dots, \widetilde{X}_s^{(n)})^T \in \mathbb{H}_\rho$)

Der zentrale Grenzwertsatz (Satz 5.5), angewendet auf jede der (unabhängigen) Koordinaten, liefert

$$\widetilde{X}^{(n)} := (\widetilde{X}_1^{(n)}, \dots, \widetilde{X}_s^{(n)})^T \xrightarrow[n\to\infty]{d} \mathcal{N}_{0,1}^{\otimes s}$$

und somit gilt auch

$$\sum_{i=1}^{s} \big(\widetilde{X}_i^{(n)}\big)^2 \xrightarrow[n\to\infty]{d} \chi_s^2.$$

Das macht zumindest plausibel, dass auch

$$\mathscr{L}\Big(\sum_{i=1}^{s} \big(\widetilde{X}_i^{(n)}\big)^2 \,\Big|\, \widetilde{N}_n = 0\Big) \xrightarrow[n\to\infty]{d} \chi_{s-1}^2$$

gilt.

Hier sind etwas mehr Details: Sei O eine orthogonale $s \times s$-Matrix, deren letzte Spalte u_ρ ist (ergänze u_ρ zur ONB),

$$\Pi_\rho = O \cdot \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & & 0 \end{pmatrix} O^T$$

ist die (orthogonale) Projektion(smatrix) auf $\mathbb{H}_\rho$.

Sei $\widetilde{N}_\rho := \mathcal{N}_{0,1}^{\otimes s} \circ (\Pi_\rho)^{-1}$,
$a_1, \dots, a_s \in \mathbb{R}$, $A := (-\infty, a_1] \times \times (-\infty, a_s]$,

$$q_{m,n} := P\big(\widetilde{X}^{(n)} \in A \,\big|\, N_n = m\big)$$

Es gilt

$$q_{m,n} \geq q_{m+1,n} \quad \text{für } m, n \in \mathbb{N}$$

(verwende die „natürliche“ Kopplung der Multinomialverteilungen und die spezielle Form von A), somit für $\varepsilon > 0$

$$P\big(\widetilde{X}^{(n)} \in A \,\big|\, \widetilde{N}_n \in [0, \varepsilon]\big) \leq P\big(\widetilde{H}^{(n)} \in A\big) \leq P\big(\widetilde{X}^{(n)} \in A \,\big|\, \widetilde{N}_n \in [-\varepsilon, 0]\big)$$

(denn

$$q_{n,n} \geq \sum_{m=n}^{\lceil n+\varepsilon\sqrt{n}\rceil} q_{m,n} P\big(N_n = m \,\big|\, \widetilde{N}_n \in [0,\varepsilon]\big) = P\big(\widetilde{X}^{(n)} \in A \,\big|\, \widetilde{N}_n \in [0,\varepsilon]\big)$$

und analog für die andere Schranke).

Setze $\widetilde{Y}^{(n)} := O^T \widetilde{X}^{(n)}$, $U_\varepsilon := \{(x_1,\dots,x_s)^T \in \mathbb{R}^s : 0 \leq x_s \leq \varepsilon\} = \mathbb{R}^{s-1} \times [0,\varepsilon]$

$$\begin{aligned}
&P\big(\widetilde{X}^{(n)} \in A \,\big|\, \widetilde{N}_n \in [0,\varepsilon]\big) = P\big(\widetilde{Y}^{(n)} \in O^T A \,\big|\, \widetilde{Y}_n \in U_\varepsilon\big) \\
&= \frac{P\big(\widetilde{Y}_n \in O^T A \cap U_\varepsilon\big)}{P\big(\widetilde{Y}_n \in U_\varepsilon\big)} \\
&\xrightarrow[n\to\infty]{} \frac{\mathcal{N}_{0,1}^{\otimes s}(O^T A \cap U_\varepsilon)}{\mathcal{N}_{0,1}^{\otimes s}(U_\varepsilon)} \\
&\qquad = \frac{1}{\mathcal{N}_{0,1}([0,\varepsilon])} \int_0^\varepsilon \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathcal{N}_{0,1}^{\otimes(s-1)}\big(\{x \in \mathbb{R}^{s-1} : (x,t) \in O^T A\}\big)\, dt
\end{aligned}$$

denn mit zentralem Grenzwertsatz (Satz 5.5) und Rotationssymmetrie der s-dim. Normalverteilung (Beispiel 2.20) folgt $\widetilde{Y}^{(n)} \xrightarrow[n\to\infty]{d} \mathcal{N}_{0,1}^{\otimes s}$. Mit $\varepsilon \downarrow 0$ konvergiert die rechte Seite gegen $\widetilde{\mathcal{N}}_\rho(A)$. □

Korollar 6.32 (χ^2-Anpassungstest[8])**.** *Sei* $\vartheta \in \Theta = \Delta_s := \{(\vartheta_1,\dots,\vartheta_s) \in [0,1]^s : \vartheta_1 + \dots + \vartheta_s = 1\}$, *unter* P_ϑ *sei* $(H_1,\dots,H_s) \sim \mathrm{Mult}_{n;\vartheta_1,\dots,\vartheta_s}$.

Sei $\rho \in \Delta_s$,

$$D := \sum_{i=1}^s \frac{(H_i - n\rho_i)^2}{n\rho_i},$$

$\alpha \in (0,1)$, q *das* $(1-\alpha)$*-Quantil der* χ^2_{s-1}*-Verteilung.*

Der Test von $H_0 : \{\vartheta = \rho\}$ *gegen* $H_1 : \{\vartheta \neq \rho\}$ *mit Ablehnungsbereich* $\{D > q\}$ *hat (asymptotisches) Niveau* α.

Dies folgt aus Satz 6.31. Satz 6.31 macht allerdings keine Aussage darüber, wie groß n sein sollte, damit die Approximation plausibel ist. Eine oft zitierte Faustregel (für die Gültigkeit der χ^2-Approximation) ist $n\rho_i \geq 5$ für alle i.

Beispiel 6.33 (Mendels Erbsenexperimente[9])**.** Betrachte zwei Merkmale: Farbe: grün (rezessiv) vs. gelb (dominant), Form: rund (dominant) vs. runzlig (rezessiv)

Beim Kreuzen von Doppelhybriden erwarten wir folgende Phänotypwahrscheinlichkeiten unter Mendel'scher Segregation („rund" und „gelb" sind jeweils dominant, $n = 556$ Versuche):

Typ	rund/gelb	rund/grün	kantig/gelb	rund/gelb
Anteil	9/16	3/16	3/16	1/16
Erwartete Anzahl	315	104,25	104,25	34,75
beobachtet	315	108	101	32

[8] von Karl Pearson (1857–1936) im Jahr 1900 vorgeschlagen

[9] Gregor Mendel, 1822–1884; G. Mendel, Versuche über Pflanzenhybriden, Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen: 3–47, (1866).

Wir finden $D \cong 0{,}47$, $\chi^2_3([0, 0.47]) \cong 0{,}075$, ein χ^2-Test zum 1%-Niveau lehnt H_0 nicht ab (und auch zu „nahezu egal welchem Niveau“ nicht). Insoweit passen die Daten sehr gut zu den theoretischen Häufigkeiten.

Beispiel. Wir vermuten, dass ein gegebener sechsseitiger Würfel unfair ist und möchten dies auf dem 5%-Niveau testen. Bei 120-maligem Würfeln finden wir folgende Häufigkeiten:

i	1	2	3	4	5	6
h_i	13	12	20	18	26	31

Es ist $D \cong 13{,}7$, das 95%-Quantil der χ^2_5-Verteilung ist $\cong 11{,}07$, wir können die Nullhypothese „$\vartheta = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$“ also auf dem 5%-Niveau ablehnen.

```
> w <- c(13,12,20,18,26,31)
> chisq.test(w,p=c(1/6,1/6,1/6,1/6,1/6,1/6))

Chi-squared test for given probabilities

data:  w
X-squared = 13.7, df = 5, p-value = 0.01763
```

Exkurs: χ^2-Test auf Homogenität (auch: „auf Unabhängigkeit“)

In einem Experiment werden zwei „Merkmale“ beobachtet, wobei das erste Merkmal a und das zweite Merkmal b viele Ausprägungen besitzt (also insgesamt $s = a \cdot b$ mögliche Ausgänge).

Unter n u.a. Wiederholungen werde h_{ij} mal Ausgang (i, j) beobachtet ($i \in \{1, 2, \dots, a\}$, $j \in \{1, 2, \dots, b\}$), man fasst die Beobachtungen in einer $a \times b$-Kontingenztafel zusammen:

i \ j	1	2	3	
1	h_{11}	h_{12}	h_{13}	$h_{1\cdot}$
2	h_{21}	h_{22}	h_{23}	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot 3}$	$h_{\cdot\cdot} = n$

mit Zeilensummen $h_{i\cdot} = \sum_{j=1}^{b} h_{ij}$, Spaltensummen $h_{\cdot j} = \sum_{i=1}^{a} h_{ij}$ und Gesamtsumme $h_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{b} h_{ij} = n$.

Wir fassen die beobachteten Häufigkeiten als Realisierungen einer

$$\text{multinomial}(n, (\vartheta_{ij})_{i=1,\dots,a;j=1,\dots,b})\text{-verteilten ZV } (H_{ij})_{i=1,\dots,a;j=1,\dots,b}$$

auf, wobei

$(\vartheta_{ij})_{i=1,\dots,a;j=1,\dots,b}$ ein $a \cdot b$-dimensionaler Vektor von Wahrscheinlichkeitsgewichten ist.

Passen die Beobachtungen zur Nullhypothese, dass

$$\vartheta_{ij} = \eta_i \cdot \rho_j, \quad \text{für } i = 1, \dots, a,\ j = 1, \dots, b$$

mit $(\eta_i)_{i=1,\dots,a}$, $(\rho_j)_{j=1,\dots,b}$ gewissen a- bzw. b-dimensionalen Vektoren von Wahrscheinlichkeitsgewichten?

Wir bilden

$$\widehat{\vartheta}_{i\cdot} = \frac{H_{i\cdot}}{n}, \quad \widehat{\vartheta}_{\cdot j} = \frac{H_{\cdot j}}{n}$$

und die Teststatistik

$$D = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(H_{ij} - n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j})^2}{n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j}}$$

Bericht 6.34. Unter H_0 : „$(\vartheta_{ij})_{i=1,\dots,a;j=1,\dots,b}$ hat Produktform“ ist D (approximativ) $\chi^2_{(a-1)(b-1)}$-verteilt.

Wir würden also H_0 zum Niveau α ablehnen, falls der beobachtete Wert größer ist als das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $(a-1)(b-1)$ Freiheitsgraden.

Das Simpson-Paradoxon

Durch Zusammenfassen von Gruppen können sich (scheinbare) statistische Trends in ihr Gegenteil verkehren. Dieses Phänomen heißt Simpson-Paradoxon oder Yule-Simpson-Effekt[10].

Beispiel (Zulassungsstatistik der UC Berkeley 1973)**.** Im Herbst 1973 haben sich an der Universität Berkeley 12763 Kandidaten für ein Studium beworben, davon 8442 Männer und 4321 Frauen. Es kam zu folgenden Zulassungszahlen:

	Aufgenommen	Abgelehnt
Männer	3738	4704
Frauen	1494	2827

Demnach betrug die Zulassungsquote
bei den Männern $\frac{3738}{8442} \approx 44\%$, bei den Frauen nur $\frac{1494}{4321} \approx 35\%$.

Ein χ^2-Test auf Homogenität (z.B. mit R) zeigt, dass eine solche Unverhältnismäßigkeit nur mit verschwindend kleiner Wahrscheinlichkeit durch „reinen Zufall“ entsteht:

```
> berkeley <- matrix(c(3738,1494,4704,2827),nrow=2)
> berkeley
     [,1] [,2]
[1,] 3738 4704
[2,] 1494 2827
> chisq.test(berkeley,correct=FALSE)

Pearson's Chi-squared test

data:  berkeley
X-squared = 111.2497, df = 1, p-value < 2.2e-16
```

Dieser Fall hat einiges Aufsehen erregt, s.a. P.J. Bickel, E.A. Hammel, J.W. O'Connell, Sex Bias in Graduate Admissions: Data from Berkeley, *Science* 187, no. 4175, 398–404, (1975).

Das Ungleichgewicht verschwindet, wenn man die Zulassungszahlen nach Departments aufspaltet:

[10] nach Edward H. Simpson, *1922 und George Udny Yule, 1871–1951

Es stellt sich heraus, dass innerhalb der Departments die Aufnahmewahrscheinlichkeiten nicht signifikant vom Geschlecht abhängen, aber sich Frauen häufiger bei Departments mit (absolut) niedriger Aufnahmequote beworben haben als Männer – dies ist ein Beispiel für das *Simpson-Paradox*.

Die genauen nach Departments aufgeschlüsselten Bewerber- und Zulassungszahlen sind leider nicht öffentlich zugänglich (siehe aber Abb. 1 in Bickel et. al, loc. cit., für eine grafische Aufbereitung der Daten, die den Simpson-Effekt zeigt).

Bickel et. al demonstrieren das Phänomen mittels eines hypothetischen Beispiels:

	Aufgenommen	Abgelehnt
	Department of machismathics	
Männer	200	200
Frauen	100	100
	Department of social warfare	
Männer	50	100
Frauen	150	300
	Gesamt	
Männer	250	300
Frauen	250	400

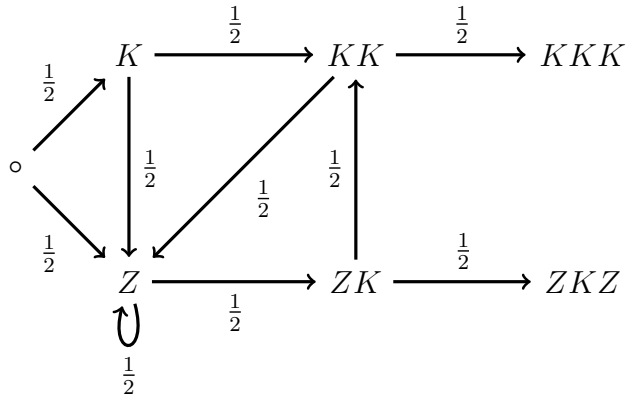
Kapitel 7

Markovketten

Beispiel 7.1. Eine faire Münze wird solange geworfen, bis entweder das Muster KKK ($\widehat{=}$ Sieg von Spieler A) oder das Muster ZKZ ($\widehat{=}$ Sieg von Spieler B) gefallen ist.

$$P(\text{Sieg A}) = ?$$

Aus der Problemstellung ergibt sich, dass es genügt, sich die Ergebnisse der beiden letzten Münzwürfe zu merken, mögliche Zustände:



(Wir stellen uns vor, ein „Spielstein“ springt zufällig auf diesem Graph herum, beginnend im Knoten „$\circ$“, indem er in jedem Schritt zufällig einem der vom aktuellen Knoten ausgehenden Pfeile folgt (mit der daran angegebenen Wahrscheinlichkeit, und die Wahl ist unabhängig von den bis dahin getroffenen Entscheidungen).

Sei $w(x) = P(\text{A gewinnt von Zustand } x \text{ aus})$, zerlege nach dem ersten Schritt:

$$\begin{aligned} w(KK) &= \frac{1}{2} + \frac{1}{2}w(Z), \\ w(ZK) &= \frac{1}{2}w(KK), \\ w(K) &= \frac{1}{2}w(KK) + \frac{1}{2}w(Z), \\ w(Z) &= \frac{1}{2}w(Z) + \frac{1}{2}w(ZK) \end{aligned}$$

mit Lösung $w(KK) = \frac{2}{3}, w(K) = \frac{1}{2}, w(Z) = \frac{1}{3} = w(ZK)$,

$$P(\text{A gewinnt}) = w(\circ) = \frac{1}{2}w(K) + \frac{1}{2}w(Z) = \frac{5}{12}.$$

Definition 7.2. Sei S abzählbare Menge ($S \neq \varnothing$).

1. $A = (a_{x,y})_{x,y\in S}$ heißt eine *stochastische Matrix* (über S), wenn gilt

$$a_{x,y} \geq 0 \text{ für alle } x, y \in S \quad \text{und} \quad \sum_{y\in S} a_{x,y} = 1 \text{ für alle } x \in S.$$

2. Eine Folge $X = (X_n)_{n\in\mathbb{N}_0}$ von Zufallsvariablen mit Werten in S heißt eine *Markovkette*[1] (mit Zustandsraum S und Übergangsmatrix A), wenn

$$P(X_{n+1} = y \,|\, X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x) = P(X_{n+1} = y \,|\, X_n = x) = a_{x,y}$$

für alle $n \in \mathbb{N}, x_0, \ldots, x_{n-1}, x, y \in S$ mit $P(X_0 = x_0, \ldots, X_n = x) > 0$ gilt.

Die Verteilung von X_0 heißt die *Startverteilung* (von X).

Die entscheidende Eigenschaft ist, dass der (bzw. die Verteilung des) „neue" Zustand X_{n+1} nur vom direkt vorhergehenden X_n abhängt, nicht von der „gesamten Vorgeschichte" $X_0, X_1, \ldots, X_n$ – dies nennt man auch die „Gedächtnislosigkeit" einer Markovkette (s.a. Beob. 7.3, 3. unten).

Beispiele.

1. $X_0, X_1, \ldots$ u.i.v. mit $X_i \sim \nu$ (ν ist ein W'maß auf S) sind (trivialerweise) eine Markovkette, mit $a_{x,y} = \nu(\{y\})$.

2. (Irrfahrt) $Y_1, Y_2, \ldots$ u.i.v. mit Werten in $\mathbb{Z}^d$, $Y_i \sim \nu$,

$$X_n := Y_1 + \cdots + Y_n, \ n \in \mathbb{N} \quad (\text{und } X_0 := 0).$$

$(X_n)_{n\in\mathbb{N}_0}$ ist eine Markovkette, $a_{x,y} = \nu(\{y - x\})$:

Für $x_0 = 0, x_1, \ldots, x_n \in \mathbb{Z}^d$ ist

$$\begin{aligned} P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) &= P(Y_1 = x_1, Y_2 = x_2 - x_1, \ldots, Y_n = x_n - x_{n-1}) \\ &= \prod_{i=1}^{n} P(Y_i = x_i - x_{i-1}) = \prod_{i=1}^{n} \nu(\{x_i - x_{i-1}\}) \end{aligned}$$

3. (Pólya-Urne, vgl. Bsp. 2.7) S_n = Anz. schwarze, W_n = Anz. weiße Kugeln in der Pólya-Urne nach n Zügen, $X_n := (S_n, W_n)$ mit Werten in $\mathbb{N}^2$ (und $X_0 = (1, 1)$, sagen wir).

$(X_n)_n$ ist Markovkette, für $x = (x_s, x_w), y = (y_s, y_w) \in \mathbb{N}^2$ ist

$$a_{x,y} = \begin{cases} \frac{x_s}{x_s+x_w}, & \text{wenn } y_s = x_s + 1, y_w = x_w, \\ \frac{x_w}{x_s+x_w}, & \text{wenn } y_s = x_s, y_w = x_w + 1, \\ 0, & \text{sonst.} \end{cases}$$

[1] zu Ehren von Andrei Andreyevich Markov, 1852–1922 benannt

4. Das „Problem der Punkte“ aus Kapitel 0 ist eine Variation über 2. mit $S = \mathbb{Z}^2$, $\nu = \frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(1,0)}$.

Beobachtung 7.3. 1. Sei $X = (X_n)_{n\in\mathbb{N}_0}$ eine Folge von ZVn mit Werten in S. S ist eine Markovkette mit Übergangsmatrix A und Startverteilung μ g.d.w.

$$\forall\, n \in \mathbb{N},\, x_0, \dots, x_n \in S :\ P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \mu(\{x_0\}) \prod_{i=1}^{n} a_{x_{i-1},x_i}.$$

(„⇐“ per Inspektion, für „⇒“ beachte

$$\begin{aligned} &P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) \\ &= \underbrace{P(X_n = x_n \,|\, X_0 = x_0, \dots, X_{n-1} = x_{n-1})}_{= a_{x_{n-1},x_n}} P(X_0 = x_0, \dots, X_{n-1} = x_{n-1}) \\ &= \cdots = a_{x_{n-1},x_n} a_{x_{n-2},x_{n-2}} \cdots a_{x_0,x_1} P(X_0 = x_0) \end{aligned}$$

sofern die linke Seite > 0 ist.)

2. Zu jeder Übergangsmatrix A und Startverteilung μ (auf einer abzählbaren Menge S) gibt es eine Markovkette mit dieser Übergangsmatrix und dieser Startverteilung: Dies folgt aus Bericht 2.10 (Konstruktion von W'maßen auf unendlichen Produkträumen), setze dort $p_1 = \mu$ und $p_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k) = a_{\omega_{k-1},\omega_k}$.

   Man schreibt oft P_μ (und $\mathbb{E}_\mu$ für Erwartungswerte unter P_μ), wenn $\mathscr{L}(X_0) = \mu$, um die Startverteilung einer Markovkette zu betonen, im Fall $\mu = \delta_x$ mit einem $x \in S$ auch P_x und $\mathbb{E}_x$.

3. („Markov-Eigenschaft“) Für jede Markovkette $(X_n)_n$, $n, m \in \mathbb{N}_0$, $B \subset S^{n+1}$, $B' \subset S^{m+1}$ und $x \in S$ gilt

$$\begin{aligned} &P\big((X_n, X_{n+1}, \dots, X_{n+m}) \in B' \,\big|\, (X_0, \dots, X_n) \in B, X_n = x\big) \\ &= P_x\big((X_0, X_1, \dots, X_m) \in B'\big) \end{aligned}$$

   (sofern $P\big((X_0, \dots, X_n) \in B, X_n = x\big) > 0$),

   denn für $x_0, x_1, \dots, x_n, x, y_0, y_1, \dots, y_m \in S$ (mit $x_n = x = y_0$, sonst ergibt sich = 0) ist nach 1.

$$\begin{aligned} &P\big(X_0 = x_0, \dots, X_{n-1} = x_{n-1}, X_n = x_n,\ X_n = x,\ X_n = y_0, X_{n+1} = y_1, \dots, X_{n+m} = y_m\big) \\ &= \cdots = P\big(X_0 = x_0, \dots, X_{n-1} = x_{n-1}, X_n = x_n,\ X_n = x\big) P_x\big(X_0 = y_0, X_1 = y_1, \dots, X_m = y_m\big). \end{aligned}$$

   Summiere dies über alle $(x_0, x_1, \dots, x_n) \in B$, $(y_0, y_1, \dots, y_m) \in B'$, so folgt

$$\begin{aligned} &P\big((X_0, \dots, X_n) \in B, X_n = x, (X_n, X_{n+1}, \dots, X_{n+m} \in B'\big) \\ &= P\big((X_0, \dots, X_n) \in B, X_n = x\big) P_x\big((X_0, X_1, \dots, X_m) \in B'\big). \end{aligned}$$

   Anschaulich gesprochen sind bedingt auf die „Gegenwart“ (den Zustand X_n zur Zeit n) die „Zukunft“ (das Pfadstück $(X_n, X_{n+1}, \dots, X_{n+m})$) und die „Vergangenheit“ (das Pfadstück $(X_0, X_1, \dots, X_n)$) unabhängig.

Korollar 7.4. *Sei $X = (X_n)_n$ Markovkette mit Übergangsmatrix $A = (a_{x,y})_{x,y\in S}$ und*

$$a_{x,y}^{(n)} = P_x(X_n = y),\ x, y \in S$$

die n-Schritt-Übergangswahrscheinlichkeit (für $n \in \mathbb{N}_0$).

Dann gilt

$$a_{x,z}^{(n+m)} = \sum_{y\in S} a_{x,y}^{(n)} a_{y,z}^{(m)}, \quad x, z \in S,\ m, n \in \mathbb{N}_0, \tag{7.1}$$

denn $P_x(X_{n+m} = z) = \sum_{y\in S} P_x(X_n = y, X_{n+m} = z) = \sum_{y\in S} P_x(X_n = y) P_y(X_m = z)$.

Insbesondere ist $A^n = (a_{x,y}^{(n)})_{,y\in S}$ gegeben durch

$$A^n = \underbrace{A \cdot A \cdot \cdots \cdot A}_{n \text{ Faktoren}}, \quad \textit{das } n\textit{-fache Matrixprodukt von } A \textit{ mit sich selbst}.$$

Das System von Gleichungen (7.1) heißt die Chapman-Kolmogorov-Gleichungen[2].

Beispiel 7.5 (Gewöhnliche Irrfahrt auf $\mathbb{Z}$)**.** Sei $p \in [0, 1]$, $S = \mathbb{Z}$,

$$a_{x,y} = \begin{cases} p, & y = x + 1, \\ 1 - p, & y = x - 1, \\ 0, & \text{sonst} \end{cases}$$

Es ist

$$a_{x,y}^{(n)} = a_{0,y-x}^{(n)} = \begin{cases} \binom{n}{\frac{n+y-x}{2}} p^{(n+y-x)/2} (1-p)^{(n-y+x)/2}, & \text{wenn } y - x \equiv n \pmod 2 \text{ und } |y - x| \le n, \\ 0, & \text{sonst} \end{cases}$$

Man kann per Induktion prüfen, dass dies aus den Chapman-Kolmogorov-Gleichungen folgt:

$$a_{0,z}^{(n)} = \sum_y a_{0,y}^{(n-1)} a_{y,z} = a_{0,z-1}^{(n-1)} \cdot p + a_{0,z+1}^{(n-1)} \cdot (1 - p),$$

alternativ (und zum Argumentieren vielleicht angehmer) stelle die Markovkette X mit Übergangsmatrix A dar als

$$X_n = Y_1 + \cdots + Y_n \quad \text{mit } Y_i \text{ u.i.v., } P(Y_1 = +1) = p = 1 - P(Y_1 = -1),$$

um von 0 nach $y - x$ in n Schritten zu gelangen, muss man

$$k = \frac{n + y - x}{2} \text{ Schritte der Größe } +1 \text{ machen und } n - k = \frac{n - y + x}{2} \text{ Schritte der Größe } -1.$$

Es gibt $\binom{n}{k}$ Wahlen, welche k der n Schritte die +1-Schritte sein sollen, jede solche Wahl hat W'keit $p^k(1 - p)^{n-k}$

(Wir sehen hier ein Beispiel für einen Periodeneffekt: Man kann gerade Zustände $x \in 2\mathbb{Z}$ nur zu geraden Zeiten besuchen.)

[2] Andrey Nikolaevich Kolmogorov, 1903–1987 ; Sydney Chapman, 1888–1970

7.1 Treffwahrscheinlichkeiten und erwartete Eintrittszeiten

Sei $X = (X_n)_{n\in\mathbb{N}_0}$ Markovkette mit Zustandsraum S und Übergangsmatrix A, für $B \subset S$ sei

$$T_B := \min\{n \in \mathbb{N}_0 : X_n \in B\}$$

die (erste) Treffzeit von B und X_{T_B} der Ort des ersten Besuchs in B.

Für $z \in B$ sei

$$h_z(x) := P_x(T_B < \infty, X_{T_B} = z)$$

Satz 7.6. *h_z ist die kleinste nicht-negative Lösung von*

$$\begin{cases} f(x) = \mathbf{1}_{\{x=z\}}, & x \in B, \\ f(x) = \sum_y a_{x,y} f(y), & x \in S \setminus B. \end{cases} \tag{7.2}$$

In Matrixschreibweise lautet die zweite Zeile von (7.2) $f(x) = Af(x)$, $x \in S \setminus B$.

Man sagt dazu auch: f ist „harmonisch" (bezgl. A) auf $S \setminus B$.

Beweis. Offenbar gilt $h_z(x) = \mathbf{1}_{\{x=z\}}$ für $x \in B$, für $x \in S \setminus B$ ist

$$\begin{aligned} P_x\big(T_B < \infty, X_{T_B} = z\big) &= \sum_{y\in S} P_x\big(X_1 = y, T_B < \infty, X_{T_B} = z\big) \\ &= \sum_{y\in S} P_x(X_1 = y) P_y\big(T_B < \infty, X_{T_B} = z\big) = \sum_{y\in S} a_{x,y} h_z(y) \end{aligned}$$

(wobei wir in der zweiten Gleichung die Markov-Eigenschaft, Beob. 7.3, 3. verwendet haben), d.h. $h_z(\cdot)$ löst (7.2).

Sei $f : S \to [0, \infty)$ eine Lösung von (7.2), zeige induktiv (über n):

$$\forall\, x \in S \,:\, P_x\big(T_B \le n, X_{T_B} = z\big) \le f(x) \tag{7.3}$$

Für $x \in B$ gilt (7.2) offenbar stets, ebenso für $n = 0$.

Sei (7.2) für n erfüllt, betrachte ein $x \in S \setminus B$:

$$\begin{aligned} P_x\big(T_B \le n+1, X_{T_B} = z\big) &= \sum_{y\in S} P_x\big(X_1 = y, T_B \le n+1, X_{T_B} = z\big) \\ &= \sum_{y\in S} P_x(X_1 = y) \underbrace{P_y\big(T_B \le n, X_{T_B} = z\big)}_{\le f(y) \text{ n. Vor.}} \le \sum_{y\in S} a_{x,y} f(y) = f(x) \end{aligned}$$

wobei wir für die letzte Gleichung verwenden, dass f (7.2) löst. Mit $n \to \infty$ in (7.3) folgt $h_z \le f$. □

Bemerkung. Für die Untersuchung in Satz 7.6 könnten wir die Zustände $x \in B$ absorbierend machen, d.h. übergehen zu

$$\widetilde{A} = (\widetilde{a}_{x,y})_{x,y\in S} \quad \text{mit} \quad \widetilde{a}_{x,y} = \begin{cases} a_{x,y}, & x \notin B \\ \mathbf{1}_{\{y=x\}}, & x \in B \end{cases}$$

Es ergibt sich damit dieselbe Lösung.

Bemerkung 7.7 (Stochastische Lösung eines diskreten Dirichlet-Problems)**.** In der Situation von Satz 7.6 sei $|S| < \infty$, $\varnothing \neq B \subset S$, es gelte

$$\forall\, x \in S :\ P_x(T_B < \infty) > 0$$

Dann ist für jedes $g : B \to \mathbb{R}$ das (lineare) Gleichungssystem

$$\begin{cases} Af(x) = f(x), & x \in S \setminus B \\ f(x) = g(x), & x \in B \end{cases} \tag{7.4}$$

mit $Af(x) := \sum_{y\in S} a_{x,y} f(y)$ eindeutig lösbar (und nach Satz 7.6 bzw. einer leichten Erweiterung davon gegeben durch $f(x) = \mathbb{E}_x\big[g(X_{T_B})\big]$).

Beweis. Sei $\widehat{a}_{x,y} := \mathbf{1}_{S\setminus B}(x) a_{x,y}$, $\widehat{A} = (\widehat{a}_{x,y})_{x,y\in S}$, $\widehat{g}(x) = \mathbf{1}_B(x) g(x)$, so ist (7.4) (in Matrixschreibweise)

$$\big(I_S - \widehat{A}\big) f = \widehat{g}$$

(mit $I_s = (\delta_{x,y})_{x,y\in S}$ der Identitätsmatrix auf S) also ist

$$f = \big(I_S - \widehat{A}\big)^{-1} \widehat{g}$$

Beachte: $\big(I_S - \widehat{A}\big)^{-1} = \sum_{n=0}^{\infty} \widehat{A}^n$, die Reihe konvergiert, denn $\widehat{a}_{x,y} \geq 0$ und nach Voraussetzung ist für ein $n_0 \in \mathbb{N}$ und ein $\delta \in (0,1)$

$$\sup_{x\in S} \sum_{y\in S} \widehat{a}^{(n)}_{x,y} \leq 1 - \delta \quad \text{für } n \geq n_0,$$

demnach sind die Einträge von $\widehat{A}^{n_0+k}$ nicht-negativ und beschränkt durch $(1-\delta)^k$. □

Beispiel 7.8. 1. (Symmetrische gewöhnliche Irrfahrt auf $\mathbb{Z}$, Rückkehr zur 0)
Betrachte (X_n) aus Bsp 7.5 mit $p = 1/2$
Sei $T_0 := \inf\{n \in \mathbb{N}_0 : X_n = 0\}$, so gilt

$$P_x(T_0 < \infty) = 1 \quad \text{für jedes } x \in \mathbb{Z}.$$

Die Funktion $h_0(x) := P_x(T_0 < \infty)$ löst nach Satz 7.6

$$h_0(0) = 1, \quad h_0(x) = \frac{1}{2} h_0(x-1) + \frac{1}{2} h_0(x+1),\ x \neq 0,$$

d.h. die Werte $h_0(x)$ liegen auf einer Geraden mit $h_0(0) = 1$, wegen $0 \leq h_0(x) \leq 1\ \forall\, x \in \mathbb{Z}$ folgt $h_0(x) \equiv 1$.

2. („Ruinproblem des Glücksspielers")
Seien $b, c \in \mathbb{N}$, $S := \{-b-b+1, \ldots, 0, 1, \ldots, c\}$, $p \in (0,1)$,

$$a_{x,y} = \begin{cases} p, & y = x+1 \leq c, \\ 1-p, & y = x-1 \geq -b, \\ 1, & y = x = -b \text{ oder } y = x = c, \\ 0, & \text{sonst} \end{cases}$$

Wir interpretieren die zugehörige Markovkette (X_n) als (kumlierten) Gewinnprozess in einem (iterierten) Münzwurfspiel:

Gewinne 1 € bei Kopf (W'keit p), verliere 1 € bei Zahl (W'keit $1-p$), spiele solange, bis entweder c € gewonnen („Sieg") oder b € verloren („Ruin").

$h_c(x) := P_x(T_{\{c\}} < \infty)$ löst

$$h_c(c) = 1,\ h_c(-b) = 0, \quad h_c(x) = ph_c(x+1) + (1-p)h_c(x-1),\ -b < x < c,$$

die Lösung ist für $p \neq 1/2$

$$h_c(x) = \frac{(q/p)^x - (q/p)^{-b}}{(q/p)^c - (q/p)^{-b}} \quad \text{mit } q := 1-p$$

und für $p = 1/2$

$$h_c(x) = \frac{x+b}{c+b}.$$

Analog findet man

$$P_x(T_{\{-b\}} < \infty) = \frac{(q/p)^c - (q/p)^x}{(q/p)^c - (q/p)^{-b}} = 1 - P_x(T_{\{c\}} < \infty)$$

(bzw. $P_x(T_{\{-b\}} < \infty) = (c-x)/(c+b)$ für $p = 1/2$), es gilt also $P_x(T_{\{-b,c\}} < \infty) = 1$, das Spiel endet mit Sicherheit in endlicher Zeit.

Erwartete Eintrittszeiten

Man kann für eine Markovkette die erwartete Zeit bis zum Besuch eines gewissen Zustands in ähnlicher Weise bestimmen wie die Auftreffwahrscheinlichkeiten.

Bericht 7.9. (In der Situation von Satz 7.6) sei $g(x) = \mathbb{E}_x[T_B]$. g ist die kleinste nicht-negative Lösung von

$$\begin{cases} f(x) = 0, & x \in B, \\ f(x) = 1 + \sum_y a_{x,y} f(y), & x \in S \setminus B \end{cases} \tag{7.5}$$

(wobei die Summe $\sum_y a_{x,y} f(x)$ möglicherweise divergieren kann).

Man kann den Beweis analog zu dem Satz 7.6 führen; Bem. 7.7 zur Eindeutigkeit gilt entsprechend.

Beispiel 7.10. Sei (X_n) symm. gew. Irrfahrt auf $\mathbb{Z}$ (startend in $X_0 = 0$), vgl. Beispiel 7.5 (und Bsp. 7.8). Es ist

$$\mathbb{E}_0[T_{\{1\}}] = 1 + \frac{1}{2} \cdot 0 + \frac{1}{2}\mathbb{E}_{-1}[T_{\{1\}}],$$

andererseits ist

$$\mathbb{E}_{-1}[T_{\{1\}}] = \underbrace{\mathbb{E}_{-1}[T_{\{0\}}]}_{=\mathbb{E}_0[T_{\{1\}}]} + \mathbb{E}_0[T_{\{1\}}] = 2\mathbb{E}_0[T_{\{1\}}].$$

Somit ist $\mathbb{E}_0[T_{\{1\}}] = 1 + \mathbb{E}_0[T_{\{1\}}]$, was $\mathbb{E}_0[T_{\{1\}}] = +\infty$ erzwingt.

7.2 Gleichgewichte

Definition 7.11. Sei $X = (X_n)_n$ Markovkette auf S mit Übergangsmatrix A. Ein W'maß π auf S heißt ein *Gleichgewicht* für X (bzw. für A), wenn gilt

$$P_\pi(X_1 = x) = \pi(x) \text{ für alle } x \in S.$$

(Wir schreiben hier und im Folgenden abkürzend $\pi(x)$ für $\pi(\{x\})$).

Da $P_\pi(X_1 = x) = \sum_{y\in S} \pi(y)a_{y,x}$, ist dies gleichbedeutend mit

$$\pi A = \pi, \quad \text{d.h. } \pi \text{ ist links-Eigenvektor von } A \text{ zum Eigenwert } 1.$$

Es gilt dann auch $\pi = \pi A^n$, $P_\pi(X_n = x) = \pi(x)$ für $n \in \mathbb{N}$.

Definition 7.12. $A = (a_{x,y})_{x,y\in S}$ heißt *irreduzibel*, wenn es

$$\text{für alle } x, y \in S \text{ ein } n = n(x, y) \text{ gibt mit } a^{(n)}_{x,y} > 0.$$

Die zugehörige Markovkette kann also von jedem Startzustand x aus jeden anderen Zustand y mit positiver Wahrscheinlichkeit (irgendwann) erreichen.

Satz 7.13. *Sei $|S| < \infty$, A irreduzibel, dann gibt es genau ein Gleichgewicht π für A. π erfüllt $\pi(x) > 0$ für alle $x \in S$.*

Beweis. Existenz: Wähle $x_0 \in S$.

$$a_n := \Big(\frac{1}{n}\sum_{j=0}^{n-1} a^{(j)}_{x_0,x}\Big)_{x\in S} \quad \text{ist ein Vektor in } [0,1]^S.$$

Wähle mit Satz von Bolzano-Weierstraß ($[0,1]^S$ ist kompakt) eine konvergente Teilfolge $n_k \nearrow_{k\to\infty} \infty$,

$$\pi(x) := \lim_{k\to\infty} \frac{1}{n_k} \sum_{j=0}^{n_k-1} a^{(j)}_{x_0,x}, \quad x \in S$$

ist ein Gleichgewicht:

$$\begin{aligned}
\pi A(x) &= \sum_{y\in S} \Big(\lim_{k\to\infty} \frac{1}{n_k} \sum_{j=0}^{n_k-1} a^{(j)}_{x_0,y}\Big) a_{y,x} \\
&= \lim_{k\to\infty} \frac{1}{n_k} \sum_{j=0}^{n_k-1} \sum_{y\in S} a^{(j)}_{x_0,y} a_{y,x} = \lim_{k\to\infty} \frac{1}{n_k} \sum_{j=0}^{n_k-1} a^{(j+1)}_{x_0,x} \\
&= \lim_{k\to\infty} \frac{1}{n_k} \sum_{j=0}^{n_k-1} a^{(j)}_{x_0,x} + \lim_{k\to\infty} \frac{1}{n_k}\big(a^{(n_k)}_{x_0,x} - a^{(0)}_{x_0,x}\big) = \pi(x).
\end{aligned}$$

Zeige $\pi(x) > 0$ für alle $x \in S$:

Wähle $\widetilde{x} \in S$ mit $\pi(\widetilde{x}) > 0$. Für $x \in S$ gibt es nach Voraussetzung ein n (= $n(\widetilde{x}, x)$) mit $a^{(n)}_{\widetilde{x},x} > 0$, somit

$$\pi(x) = \pi A^n(x) = \sum_{y\in S} \pi(y) a^{(n)}_{y,x} \geq \pi(\widetilde{x}) a^{(n)}_{\widetilde{x},x} > 0.$$

Eindeutigkeit: Sei $\widetilde{\pi}$ ebenfalls ein Gleichgewicht, wähle $x_0 \in S$ mit

$$\frac{\widetilde{\pi}(x_0)}{\pi(x_0)} = \max\left\{\frac{\widetilde{\pi}(x)}{\pi(x)} : x \in S\right\} =: \lambda,$$

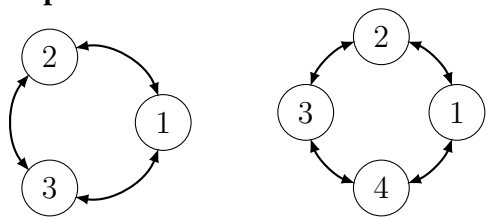
setze $\nu := \lambda\pi - \widetilde{\pi}$. Dann gilt $\nu A = \nu$, $\nu(x_0) = 0$, $\nu(x) \geq 0$ für $x \in S$.

Gäbe es ein $\widetilde{x} \in S$ mit $\nu(\widetilde{x}) > 0$, so wäre (mit Argument wie oben) $\nu(x) > 0$ für alle $x \in S$, im Widerspruch zu $\nu(x_0) = 0$. Folglich gilt $\nu \equiv 0$, d.h. $\pi = \widetilde{\pi}$. □

Definition 7.14. Eine Markovkette X auf S (bzw. eine Übergangsmatrix $A = (a_{x,y})_{x,y\in S}$) heißt *aperiodisch*, wenn es ein $n_0 \in \mathbb{N}$ gibt mit

$$a_{x,x}^{(n)} > 0 \text{ für alle } x \in S, n \geq n_0.$$

Beispiel.



aperiodisch periodisch (mit Periode 2)

Satz 7.15. *$|S| < \infty$, A irreduzibel und aperiodisch, $(X_n)_n$ Markovkette mit Übergangsmatrix A und Startverteilung μ. Dann gilt*

$$\textit{für alle } x \in S: \quad P_\mu(X_n = x) \underset{n\to\infty}{\longrightarrow} \pi(x),$$

wobei π das (eindeutige) Gleichgewicht zu A ist.

Beweis. Wir konstruieren eine Folge von $S \times S$-wertigen ZVn $(X_n, X_n')_{n\in\mathbb{N}_0}$ mit

1. (X_n) ist A-Markovkette mit Startverteilung μ
   (X_n') ist A-Markovkette mit Startverteilung π
2. $P(X_n \neq X_n') \underset{n\to\infty}{\longrightarrow} 0$.

Damit ergibt sich

$$\begin{aligned}\mu A^n(x) = P(X_n = x) &= P(X_n = x, X_n = X_n') + P(X_n = x, X_n \neq X_n') \\ &= \underbrace{P(X_n' = x)}_{=\pi(x)} - \underbrace{P(X_n' = x, X_n \neq X_n')}_{\leq P(X_n\neq X_n')\to 0} + \underbrace{P(X_n = x, X_n \neq X_n')}_{\leq P(X_n\neq X_n')\to 0} \\ &\underset{n\to\infty}{\longrightarrow} \pi(x)\end{aligned}$$

Wir werden $(X_n, X_n')_n$ mit den geforderten Eigenschaften als eine Markovkette auf $S \times S$ konstruieren: Sei $\widetilde{A}$ Übergangsmatrix auf $S \times S$:

$$\widetilde{a}_{(x,x'),(y,y')} = \begin{cases} a_{x,y}a_{x',y'}, & \text{wenn } x \neq x', \\ a_{x,y}, & \text{wenn } x = x' \text{ und } y = y', \\ 0, & \text{wenn } x = x' \text{ und } y \neq y'. \end{cases}$$

Man prüft, dass dies eine Übergangsmatrix ist und dass die Marginalverteilungen stimmen (d.h. jede Koordinate ist für sich eine Markovkette mit Übergangsmatrix A).

Wir konstruieren $(X_n, X'_n)_n$ folgendermaßen:

Sei $Y = (Y_n)_n$ (A, μ)-MK, $Y' = (Y'_n)_n$ (A, π)-MK, Y und Y' seien unabhängig.

Setze $T := \inf\{n \geq 0 : Y_n = Y'_n\}$, $X_n = Y_n$, $X'_n = \begin{cases} Y'_n & \text{für } n < T \\ Y_n & \text{für } n \geq T \end{cases}$

($(X_n, X'_n)_n$ hat tatsächlich Übergangsmatrix $\widetilde{A}$, per Inspektion).

Nach Voraussetzung gibt es $n_0 \in \mathbb{N}$, $y_0 \in S$, $\delta > 0$ so dass

$$\forall\, y \in S : \; a^{(n_0)}_{y,y_0} \geq \delta.$$

Somit

$$P(X_{n_0} \neq X'_{n_0}) \leq P(Y_{n_0} \neq y_0 \text{ oder } Y'_{n_0} \neq y_0) = 1 - P(Y_{n_0} = y_0)P(Y'_{n_0} = y_0) \leq 1 - \delta^2.$$

Analog liefert die Markov-Eigenschaft für $k \in \mathbb{N}$, $x, x' \in S$

$$P(X_{n_0 k} \neq X'_{n_0 k} \mid X_{n_0(k-1)} = x, X'_{n_0(k-1)} = x') \leq 1 - \delta^2,$$

Iteration zeigt

$$P(X_{n_0 k} \neq X'_{n_0 k}) \leq (1 - \delta^2)^k \underset{k\to\infty}{\longrightarrow} 0.$$

Nach Konstruktion ist $n \mapsto P(X_n \neq X'_n)$ nicht-wachsend, daher folgt die Behauptung. □

Beispiel 7.16 (Ehrenfest-Modell[3])**.** d Teilchen sind verteilt auf einen linken und einen rechten Behälter, in jedem Schritt wechselt ein rein zufällig ausgewähltes Teilchen von seinem aktuellen in den anderen Behälter. Sei X_n = Anz. Teilchen im linken Behälter nach n Schritten.

$(X_n)_n$ ist Markovkette auf $\{0, 1, \ldots, d\}$ mit Übergangsmatrix

$$a_{x,y} = \begin{cases} (d-x)/d, & y = x+1 \leq d, \\ x/d, & y = x-1 \geq 0, \\ 0, & \text{sonst} \end{cases}$$

Das Gleichgewicht ist $\pi = \mathrm{Bin}_{d,1/2}$: Es ist

$$\pi(x)a_{x,x+1} = \pi(x+1)a_{x+1,x}$$

(und allgemein $\pi(x)a_{x,y} = \pi(y)a_{y,x}$ und dies impliziert $\pi A = \pi$, denn dann ist $\sum_y \pi(y)a_{y,x} = \sum_y \pi(x)a_{x,y} = \pi(x)\sum_y a_{x,y} = \pi(x)$).

Definition und Beobachtung 7.17. Eine Markovkette X auf S (bzw. ihre Übergangsmatrix $A = (a_{x,y})_{x,y\in S}$) heißt *reversibel* (bezüglich π), wenn gilt

$$\pi(x)a_{x,y} = \pi(y)a_{y,x} \quad \text{für alle } x, y \in S \qquad (\text{„detaillierte Balancegleichung“})$$

Dann ist π ein Gleichgewicht für X und es gilt für $n \in \mathbb{N}$, $x_0, x_1, \ldots, x_n \in S$

$$P_\pi(X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n) = P_\pi(X_0 = x_n, X_1 = x_{n-1}, \ldots, X_{n-1} = x_1, X_n = x_0)$$

[3] Paul Ehrenfest, 1880–1933; Tatjana Ehrenfest(-Afanasyeva), 1876–1946

Satz 7.18. *Sei $X = (X_n)$ Markovkette auf S mit Übergangsmatrix A,*

$$\tau_x := \inf\{n \geq 1 : X_n = x\}$$

der Zeitpunkt des ersten Besuchs bzw. bei Start in x der ersten Rückkehr nach x. Für $x \in S$ sind äquivalent:

1. $\mathbb{E}_x[\tau_x] < \infty$.
2. *Es gibt ein Gleichgewicht π mit $\pi(x) > 0$.*

Beweis. „1. $\Rightarrow$ 2." : Sei

$$\rho(y) := \mathbb{E}_x\Big[\underbrace{\sum_{n=0}^{\tau_x - 1} \mathbf{1}_{\{X_n = y\}}}_{\leq \tau_x}\Big] \ (< \infty) \quad \text{für } y \in S,$$

es gilt

$$\rho(y) = \sum_{z \in S} \rho(z) a_{z,y},$$

denn

$$\begin{aligned}
\sum_{z \in S} \rho(z) a_{z,y} &= \sum_{z \in S} \mathbb{E}_x\Big[\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n = z, \tau_x > n\}}\Big] a_{z,y} \\
&= \sum_{z \in S} \sum_{n=0}^{\infty} \underbrace{P_x\big(X_n = z, \tau_x > n\big) a_{z,y}}_{= P_x(\tau_x > n, X_n = z, X_{n+1} = y)} = \sum_{n=0}^{\infty} P_x(\tau_x > n, X_{n+1} = y) \\
&= \mathbb{E}_x\Big[\sum_{n=1}^{\tau_x} \mathbf{1}_{\{X_n = y\}}\Big] = \rho(y).
\end{aligned}$$

Weiter ist

$$\sum_{y \in S} \rho(y) = \sum_{y \in S} \sum_{n=0}^{\infty} P_x(X_n = y, \tau_x > n) = \sum_{n=0}^{\infty} P_x(\tau_x > n) = \mathbb{E}_x[\tau_x] < \infty,$$

d.h. $\pi(y) := \dfrac{\rho(y)}{\mathbb{E}_x[\tau_x]}$, $y \in S$ leistet das Gewünschte.

„2. $\Rightarrow$ 1." : Für $\ell \in \mathbb{N}$ gilt

$$\mathbb{E}_y\big[\tau_x \wedge (\ell + 1)\big] = 1 + \sum_{z \in S,\, z \neq x} a_{y,z} \mathbb{E}_z\big[\tau_x \wedge \ell\big], \quad y \in S$$

(zerlege gemäß dem ersten Sprung, analog zu obigem), Sei π Gleichgewicht mit $\pi(x) > 0$:

$$\sum_{y \in S} \pi(y) \underbrace{\mathbb{E}_y\big[\tau_x \wedge \ell\big]}_{\leq \mathbb{E}_y[\tau_x \wedge (\ell+1)]} \leq 1 + \sum_{z \neq x} \pi(z) \mathbb{E}_z\big[\tau_x \wedge \ell\big],$$

d.d. $\pi(x)\mathbb{E}_x\big[\tau_x \wedge \ell\big] \leq 1$. Mit $\ell \to \infty$ folgt $\pi(x)\mathbb{E}_x\big[\tau_x\big] \leq 1$, also

$$\mathbb{E}_x\big[\tau_x\big] \leq \frac{1}{\pi(x)} < \infty.$$

□

Korollar 7.19. *Wenn X (in der Situation von Satz 7.18) ein eindeutiges Gleichgewicht besitzt, so gilt*

$$\mathbb{E}_x[\tau_x] = \frac{1}{\pi(x)} \quad \textit{für } x \in S.$$

Für das Ehrenfest-Modell (Bsp. 7.16) mit d Kugeln ist

$$\mathbb{E}_0[\tau_0] = \frac{1}{\mathrm{Bin}_{d,1/2}(\{0\})} = 2^d.$$

Beispiel 7.20 (Erneuerungskette)**.** Sei $m \in \mathbb{N}$, ν W'maß auf $\{1, 2, \dots, m+1\}$ mit $\nu(j) > 0$ für $1 \le j \le m+1$, X Markovkette auf $S = \{0, 1, 2, \dots, m\}$ mit Übergangsmatrix

$$a_{i,j} = \begin{cases} \nu(j+1) & i = 0 \le j \le m, \\ 1, & j = i - 1, \\ 0, & \text{sonst.} \end{cases}$$

Offenbar X ist irreduzibel und aperiodisch. Mit $\mu := \sum_{j=1}^{m+1} j\nu(j)$ ist das (eindeutige) Gleichgewicht gegeben durch

$$\pi(x) = \frac{1}{\mu}\nu\big(\{x+1, x+2, \dots, m+1\}\big), \quad x \in S,$$

denn

$$\pi(j+1)p_{j+1,j} + \pi(0)p_{0,j} = \frac{1}{\mu}\Big(\nu(\{j+2, \dots, m+1\}) \cdot 1 + 1 \cdot \nu(j+1)\Big) = \pi(j).$$

Mit Satz 7.15 folgt

$$P_{x_0}(X_n = x) \xrightarrow[n\to\infty]{} \pi(x)$$

(für alle $x_0, x \in S$).

Man kann X folgendermaßen darstellen: Seien $T_1, T_2, \dots$ u.i.v., $T_i \sim \nu$, $W_n := T_1 + \cdots + T_n$, $n \in \mathbb{N}$ ($W_0 := 0$), dann ist

$$X_n := \inf\big\{W_k - n : k \in \mathbb{N}_0, W_k \ge n\big\}$$

eine Markovkette mit obiger Übergangsmatrix (Übung).

(Interpretation: T_k = Lebensdauer der k-ten „Glühbirne", W_j = Zeitpunkt, zu dem zum j-tem Mal die Glühbirne ausgewechselt wird, dann ist X_n = „Restlebensdauer" der zum Zeitpunkt n brennenden Glühbirne.) Wir finden eine Version des sogenannten Erneuerungssatzes:

$$P\big(\exists\, k \in \mathbb{N} : W_k = n\big) = P_0(X_n = 0) \xrightarrow[n\to\infty]{} \pi(0) = \frac{1}{\mu}.$$

Bericht. Man kann die Voraussetzungen deutlich abschwächen, tatsächlich gilt obiges auch in dem Fall, dass ν ein W'maß auf $\mathbb{N}$ ist mit $\sum_{x\in\mathbb{N}} x\nu(x) < \infty$, die Bedingung, dass ν strikt positiv ist, kann ersetzt werden durch die Forderung $\mathrm{ggT}(\{x : \nu(x) > 0\}) = 1$.

7.3 Rekurrenz und Transienz*

Sei $X = (X_n)_{n\in\mathbb{N}_0}$ Markovkette mit Zustandsraum S und Übergangsmatrix A.

Definition 7.21. Ein Zustand $x \in S$ heißt *rekurrent*, wenn $P_x(\tau_x < \infty) = 1$ gilt, ansonsten *transient*.

Die Kette $(X_n)_n$ heißt rekurrent (transient), falls alle Zustände rekurrent (transient) sind.

Beispiele. 1. Die gewöhnliche Irrfahrt auf $\mathbb{Z}$ mit Drift $2p-1 \neq 0$ ($a_{x,x+1} = p = 1-a_{x,x-1}$) ist transient (vgl. Bsp. 7.8, 2.), für $p = \frac{1}{2}$ ist sie rekurrent (vgl. Bsp. 7.8, 1.).

2. $|S| < \infty$ und A irreduzibel, so ist X rekurrent (dies folgt aus Satz 7.13 zusammen mit Satz 7.18).

Beobachtung 7.22.

1. $x \in S$ rekurrent, so gilt $P_x\big(X_n = x \text{ für unendlich viele } n\big) = 1$.
2. $x \in S$ transient, so gilt $P_\mu\big(X_n = x \text{ für unendlich viele } n\big) = 0$ für jede Startverteilung μ.
3. $x \in S$ ist transient g.d.w. $\sum_{n=1}^{\infty} P_x(X_n = x) < \infty$.

Beweis. Sei

$$B_x := \big|\{n \geq 1 : X_n = x\}\big| \quad \text{die Anzahl Besuche in } x,$$

$m \in \mathbb{N}$, μ beliebige Startverteilung. Es ist

$$\begin{aligned} P_\mu(B_x \geq m) &= \sum_{\ell=1}^{\infty} P_\mu\big(X_1, \ldots, X_{\ell-1} \neq x, X_\ell = x,\ X_n = x \text{ für mind. } m-1 \text{ Werte von } n \geq \ell\big) \\ &= \sum_{\ell=1}^{\infty} P_\mu\big(X_1, \ldots, X_{\ell-1} \neq x, X_\ell = x\big) P_x(B_x \geq m-1) \\ &= P_\mu(\tau_x < \infty) P_x(B_x \geq m-1) \end{aligned}$$

also gilt

$$P_\mu(B_x \geq m) = P_\mu(\tau_x < \infty) P_x(\tau_x < \infty)^{m-1}.$$

Im transienten Fall (Punkt 2.) folgt

$$P_\mu(B_x = \infty) = \lim_{m\to\infty} P_\mu(B_x \geq m) = 0,$$

im rekurrenten Fall (Punkt 1.) folgt

$$P_x(B_x \geq m) = P_x(\tau_x < \infty)^m = 1 \qquad \text{für jedes } m \in \mathbb{N},$$

somit $P_x(B_x = \infty) = 1$.

Zu 3. :
„⇐“ : Reihe endlich $\Rightarrow P_x(B_x < \infty) = 1$ mit Borel-Cantelli-Lemma.
„⇒“ : Nach obigem ist

$$\begin{aligned} P_x(B_x = m) &= P_x(B_x \geq m) - P_x(B_x \geq m+1) \\ &= \big(1 - P_x(\tau_x < \infty)\big) P_x(\tau_x < \infty)^m, \end{aligned}$$

d.h. $B_x \sim \mathrm{Geom}_{1-P_x(\tau_x<\infty)}$ unter P_x, also

$$\sum_{n=1}^{\infty} P_x(X_n = x) = \mathbb{E}_x[B_x] = \frac{1}{1 - P_x(\tau_x < \infty)} - 1 < \infty.$$

□

Beispiel 7.23 (Gewöhnliche symmetrische Irrfahrt auf $\mathbb{Z}^d$)**.** (X_n) Markovkette auf $S = \mathbb{Z}^d$, Übergangsmatrix $a_{x,y} = \begin{cases} \frac{1}{2d}, & \text{falls } \|x-y\| = 1, \\ 0, & \text{sonst} \end{cases}$ (d.h. X springt jeweils zu einem uniform ausgewählten (direkten) Nachbarpunkt auf dem d-dimensionalen Gitter).

X ist rekurrent für $d = 1$ und $d = 2$, transient für $d \geq 3$.

Beweis. Wegen Verschiebungsinvarianz (es gilt $a_{x,y} = a_{x+z,y+z} = a_{0,y-x}$ für alle $x, y, z \in \mathbb{Z}^d$) genügt es den Startpunkt $x = 0 \in \mathbb{Z}^d$ zu betrachten.

Den Fall $d = 1$ haben wir bereits behandelt (Bsp. 7.8, 1.).

Der Fall $d = 2$: Es ist

$$\begin{aligned} P_0(X_{2n} = 0) &= 4^{-2n} \sum_{k=0}^{n} \binom{2n}{k, k, n-k, n-k} \\ &= 4^{-2n} \binom{2n}{n} \sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k} = 4^{-2n}\binom{2n}{n}^2 \sim \frac{1}{\pi n} \end{aligned}$$

mit Stirling-Approximation, wegen $\sum_n 1/n = \infty$ folgt aus Beob. 7.22 Rekurrenz.

Der Fall $d \geq 3$: Es ist

$$\begin{aligned} P_0(X_{2n} = 0) &= (2d)^{-2n} \sum_{\substack{n_1,\dots n_d \in \mathbb{Z}_+ \\ n_1+\cdots+n_d = n}} \binom{2n}{n_1, n_1, n_2, n_2, \dots, n_d, n_d} \\ &= (2d)^{-2n} \binom{2n}{n} \sum_{\substack{n_1,\dots n_d \in \mathbb{Z}_+ \\ n_1+\cdots+n_d = n}} \binom{n}{n_1, n_2, n_3, \dots, n_d}^2 \end{aligned}$$

denn man muss in jeder der d Koordinatenrichtungen gleich viele $+1$ und -1-Schritte ausführen, um nach $2n$ Schritten wieder zurück in der 0 zu sein.

Mit $m := \lceil n/d \rceil$ ist

$$\binom{n}{n_1, n_2, n_3, \dots, n_d} \leq \binom{dm}{m, m, \dots, m}$$

(denn falls $n_i \leq n_j - 2$, so ist $\binom{n}{n_1,\dots,n_i,\dots,n_j,\dots,n_d} \leq \binom{n}{n_1,\dots,n_i+1,\dots,n_j-1,\dots,n_d}$), somit ist

$$\begin{aligned} P_0(X_{2n} = 0) &\leq 2^{-2n}\binom{2n}{n}\frac{(dm)!}{(m!)^d} d^{-n} \sum_{\substack{n_1,\dots n_d \in \mathbb{Z}_+ \\ n_1+\cdots+n_d = n}} \binom{n}{n_1, n_2, n_3, \dots, n_d} d^{-n} \\ &\sim \frac{1}{\sqrt{\pi n}} \frac{(2\pi dm)^{1/2}}{(2\pi m)^{d/2}} d^{dm-n} \end{aligned}$$

für $n \to \infty$ (und damit auch $m = \lceil n/d \rceil \to \infty$) mit Stirling-Approximation.

Insgesamt folgt $P_0(X_{2n} = 0) = O(n^{-d/2})$, für $d \geq 3$ ist daher $\sum_{k=1}^{\infty} P_0(X_k = 0) < \infty$, mit Beob. 7.22 folgt Transienz. □

Literaturverzeichnis

[KW] G. Kersting und A. Wakolbinger, *Elementare Stochastik*, Birkhäuser 2010

[G] H.-O. Georgii, *Stochastik*, De Gruyter 2015

[H] N. Henze, *Stochastik: Eine Einführung mit Grundzügen der Maßtheorie*, Springer 2019

[Pf] J. Pfanzagl, *Elementare Wahrscheinlichkeitsrechnung*, de Gruyter 1991

[Kr] U. Krengel, Einführung in die Wahrscheinlichkeitstheorie und Statistik, Vieweg 2000

[DH] H. Dehling und B. Haupt, Einführung in die Wahrscheinlichkeitstheorie und Statistik, Springer 2003

[GS] C.M. Grinstead und J.L. Snell, *Introduction to Probability*, AMS 1997. `http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html`

[P] J. Pitman, *Probability*, Springer 1997

[F] W. Feller, *An introduction to probability theory and its applications*, Vol I und II, Wiley 1968, 1971

[R] The R Project for Statistical Computing, `http://www.r-project.org/`