Statistik für Informatiker

Notizen zu einer Vorlesung an der
Johannes-Gutenberg-Universität Mainz, Sommer 2017

Matthias Birkner

Vorläufige Version, 21. Juni 2017

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

Inhaltsverzeichnis

Kapitel 0

Auftakt

Sei $Z = (X, Y)$ ein „rein zufällig“ gewählter Punkt im Einheitsquadrat $S = \{(x, y) : 0 \leq x, y \leq 1\}$ (geschrieben in kartesischen Koordinaten).

Für die praktische Implementierung stellen wir uns etwa vor, dass S in (sehr kleine) disjunkte Quadrate („Pixel“) zerlegt wird, und dass wir unter allen möglichen Pixeln eines wählen, wobei jedes dieselbe Chance hat, gezogen zu werden. Eine Möglichkeit, dies in `R` zu implementieren, wäre

```
Z <- c(runif(1),runif(1))
```

Hierbei generiert der Befehl `runif(1)` eine (Pseudo-)Zufallszahl im Einheitsintervall $[0, 1]$.

Wir nennen Z eine *Zufallsvariable*, die möglichen Werte S ihren *Wertebereich*.

Sei $B \subset S$ eine gewisse Teilmenge, dann können wir das *Ereignis*

$$\{Z \in B\}$$

(ausgesprochen als „Z nimmt einen Wert in B an“) betrachten.

Je nach Ausgang des zufälligen Experiments (für das Computerexperiment hängt dies vom internen Zustand des Pseudo-Zufallsgenerators und damit implizit vom gewählten “random seed” ab) wird Z in B liegen oder nicht, d.h. das Ereignis $\{Z \in B\}$ tritt ein oder nicht.

Die *Wahrscheinlichkeit* des Ereignisses $\{Z \in B\}$ ist plausiblerweise

$$P\big(\{Z \in B\}\big) = \frac{\text{Anzahl Pixel in } B}{\text{Anzahl Pixel in } S} = \frac{\text{Fläche von } B}{\text{Fläche von } S} = \frac{\text{Fläche von } B}{1}$$

(das P erinnert an Englisch “probability” oder Französisch ≪probabilité≫, die natürlich beide vom Lateinischen Wort probabilitas abstammen).

Sei $\mathbf{1}_B$ die *Indikatorfunktion* von B, d.h. für $z \in S$ ist

$$\mathbf{1}_B(z) = \begin{cases} 1, & z \in B, \\ 0, & z \notin B. \end{cases}$$

Wir können eine weitere Zufallsvariable $W := \mathbf{1}_B(Z)$ mit Wertebereich $\{0, 1\}$ bilden: W ist gleich 1, wenn der Wert von Z in B liegt, sonst gleich 0 (man nennt eine solche Zufallsvariable auch eine *Indikatorvariable*). Mit Ereignissen ausgesprochen also

$$\{W = 1\} = \{Z \in B\}, \quad \{W = 0\} = \{Z \in B^c\}$$

(hierbei ist $B^c = S \setminus B = \{z \in S : z \notin B\}$ die Komplementmenge von B) und somit auch

$$P\big(\{W = 1\}\big) = P\big(\{Z \in B\}\big) = \frac{\text{Fläche von } B}{1}.$$

Anwendung: eine einfache Monte Carlo-Methode zur Integration

Sei p = Fläche von $B = P\big(\{Z \in B\}\big)$. Wir können den Zufall verwenden, um p (wenigstens approximativ) zu bestimmen:

Stellen wir uns vor, wir wiederholen obiges Zufallsexperiment n-mal, wobei der Zufall „jedes mal neu wirkt“. Seien $Z_1, Z_2, \ldots, Z_n$ die Ergebnisse dieser n Experimente (im Jargon: die Z_i sind *unabhängige Kopien* von Z), setze $W_i := \mathbf{1}_B(Z_i)$.

Die Zufallsvariable

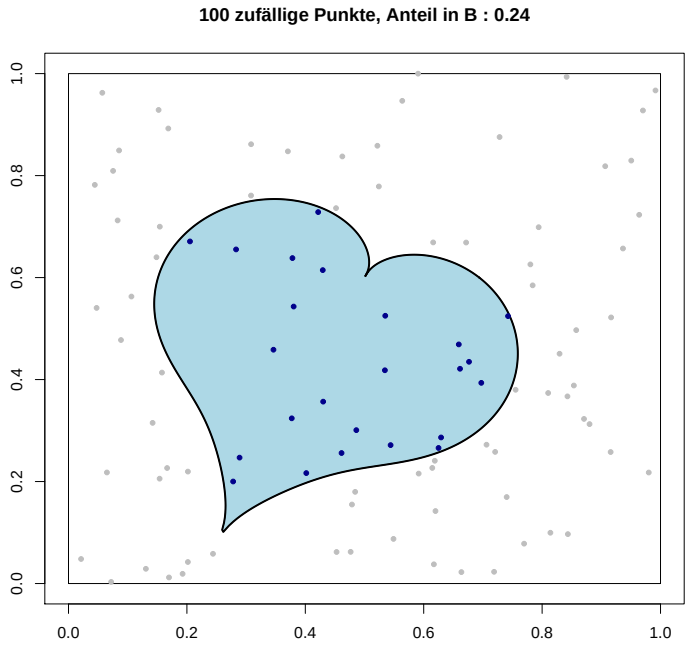
$$\widetilde{M}_n := \sum_{i=1}^{n} W_i$$

gibt an, wieviele der n zufällig gewählten Punkte in B gelandet sind. Der empirische Anteil

$$M_n := \frac{1}{n} \sum_{i=1}^{n} W_i$$

ist ein (plausibler) „Schätzwert“ für p. (Der Wertebereich von M_n ist offenbar $\{0, \frac{1}{n}, \frac{2}{n}, \ldots \frac{n-1}{n}, 1\}$.)

Man kann dies (nämlich die Simulation von M_n und Ausgabe des berechneten Werts) als ein einfaches Beispiel eines sogenannten Monte Carlo-Algorithmus betrachten: Das Verfahren bestimmt zwar nicht genau den Wert von p, wir werden aber quantifizieren können, wie (un-)wahrscheinlich es ist, dass es um mehr als ein vorgegebenes ε „daneben liegt“.



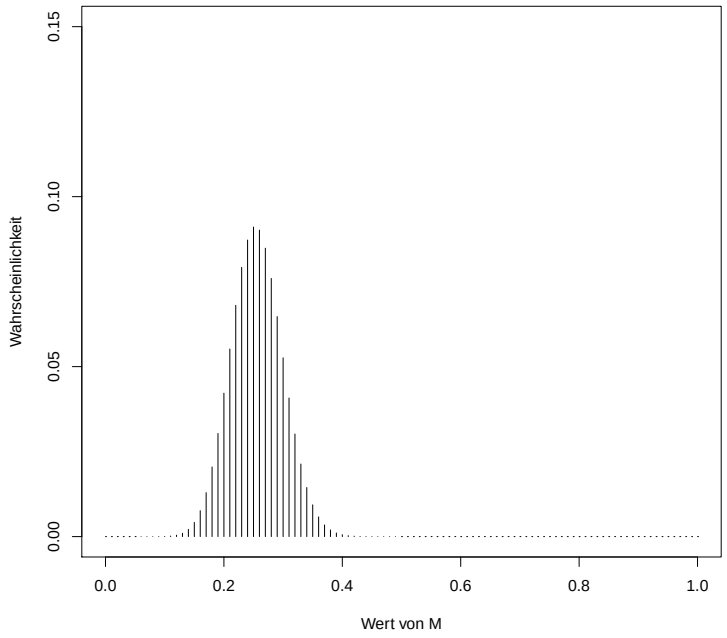
Visualisierung eines solchen Versuchs für $n = 100$.

$\widetilde{M}_n$ ist Binomial-verteilt mit Parametern n und p (abgekürzt $\mathrm{Bin}_{n,p}$-verteilt), d.h.

$$P\big(\{\widetilde{M}_n = k\}\big) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{für } k = 0, 1, \ldots, n.$$

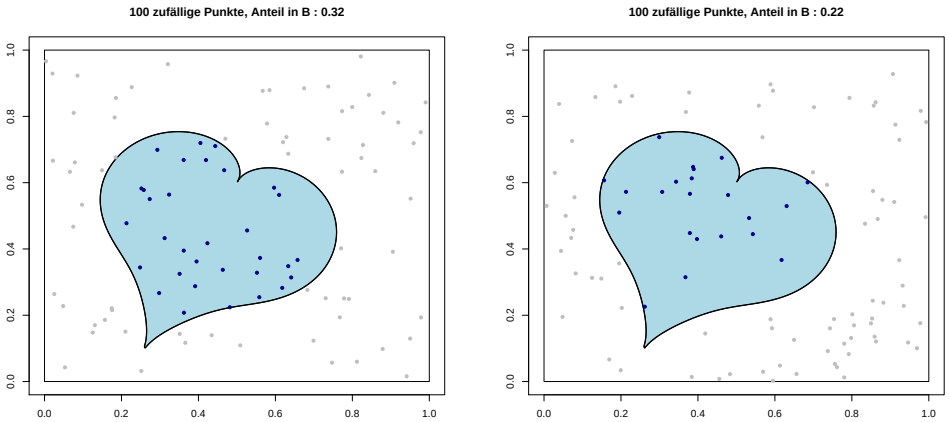
(Wir werden dies später noch genauer betrachten).

Damit ergibt sich für die Verteilungsgewichte von M_{100} folgendes (Balken-)Diagramm (an die Stelle k/n zeichnen wir einen Balken der Höhe $P(\{M_n = \frac{k}{n}\})$):



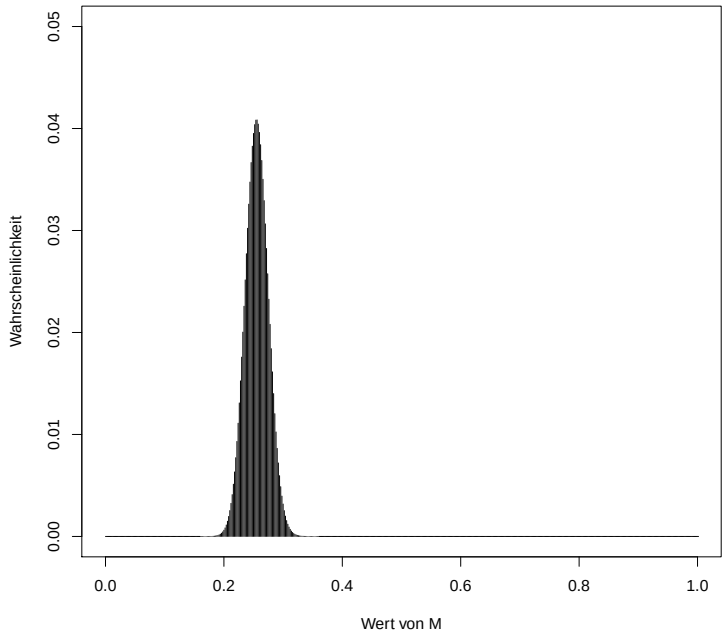
Tatsächlich ist der Wert von p in diesem Beispiel

$$p = \int_0^{2\pi} \frac{1}{98} \left(2 - 2\sin(w) + \frac{\sin(w)\sqrt{|\cos(w)|}}{\sin(w) + 7/5} \right)^2 dw \approxeq 0.25557221...$$



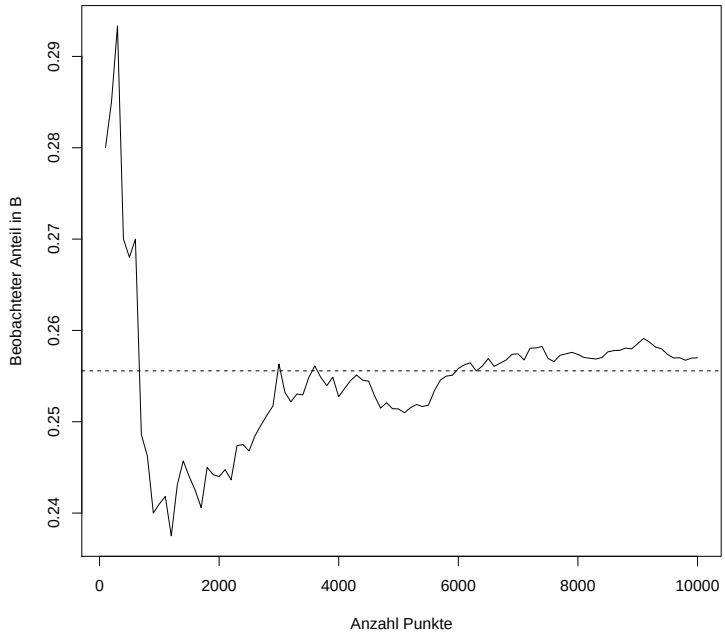
Zur Illustration: Zwei weitere Wiederholungen mit jeweils $n = 100$ zufälligen Punkten in $[0, 1]^2$

Durch Erhöhung von n können wir die Genauigkeit der Approximation erhöhen. Wir werden dies später quantifizieren können, für den Moment betrachten wir das Histogramm der Verteilungsgewichte von M_{500}



(Wir sehen: Die Verteilung von M_{500} ist deutlich stärker um p konzentriert als die von M_{100}).

Betrachten wir zum Abschluss (für eine Folge von $Z_1, Z_2, \ldots, Z_{10000}$) die Folge der M_n als Funktion von n (die gestrichelte Linie ist der Wert p):



Wir werden im Lauf der Vorlesung sehen: Für $n \to \infty$ konvergiert M_n (in geeignetem Sinne) gegen p. Dies folgt aus dem sogenannten Gesetz der großen Zahlen.

Kapitel 1

Grundlagen aus der Wahrscheinlichkeitstheorie

1.1 Zufallsvariablen, Ereignisse und Wahrscheinlichkeiten

1.1.1 Zufallsvariablen

In vielen Situationen tritt „Zufall“ in Erscheinung, in diesem Kapitel geht es uns darum, einen mathematischen Rahmen zu formulieren, um solche Phänomene zu beschreiben. Zufallsvariablen sind (in einem gewissen Sinn sogar: die) „Fundamentalobjekte“ der Stochastik, sie sind zudem oft sehr angenehm zum Notieren und Rechnen sowie zum intuitiven Argumentieren über zufällige Vorgänge.

Beispielsweise:

- Wir werfen einen Würfel, sei X die Augenzahl (für einen 6er-Würfel kommen als Werte die Elemente von $\{1, 2, \ldots, 6\}$ in Frage).
- Wir werfen eine Münze, sei W das Ergebnis (es kommen als Werte $\{\mathrm{Kopf}, \mathrm{Zahl}\}$ in Frage).
- Sei Y die Anzahl Anfragen an einen gewissen Web-Server an einem Tag (es kommen als Werte prinzipiell Elemente von $\mathbb{N}_0$ in Frage)
- Der zufällig gewählte Punkt Z aus Kapitel 0, dort kamen als Werte die Punkte aus $[0, 1]^2$ in Frage.
- Bei einer zufällig ausgewählten Person (etwa für eine wissenschaftliche Studie) werden Gewicht G (in kg) und Blutdruck D (in mmHg) gemessen. Als Werte kommen beispielsweise Zahlen in $\mathbb{R}_+$ in Frage (auch wenn wir praktischerweise Gewicht und Blutdruck nicht mit beliebiger Genauigkeit messen können und auch aus physiologischen Gründen Gewicht und Blutdruck nicht beliebig hoch sein können).

Wir nennen eine gewisse Menge von Zufallsvariablen $\mathscr{X}$ (die wir für die Modellierung eines gewissen Sachverhalts/Phänomens, in dem Unsicherheit oder Zufall vorkommt, verwenden wollen) ein *Zufallsexperiment*. Jede Zufallsvariable (abgekürzt: ZV) X besitzt einen gewissen *Wertebereich* S (wir nehmen in unserem Modell an, dass, egal wie das zufällige Experiment

ausgeht, X irgendeinen Wert in S annimmt; Vorstellung: „der Zufall“ wählt einen der möglichen Werte aus). Man sagt auch: X ist eine S-wertige Zufallsvariable, auch „Zufallsgröße“, S heißt auch der „Zielbereich“ von X.

(Übliche Notationskonvention: ZV werden meist mit Großbuchstaben benannt.)

Lesehinweis Man kann dieses Kapitel auf mehrere Weisen lesen. Die grundlegenden Objekte Zufallsvariablen, Ereignisse und Wahrscheinlichkeiten und ihre Eigenschaften werden uns durch den gesamten Verlauf der Vorlesung begleiten. Einige Beweise und Diskussionen bzw. Berichte zu Hintergrundmaterial sind (wie hier) am Rand mit  gekennzeichnet. Sie wurden der Vollständigkeit halber für interessierte Leser aufgenommen, speziell auch für solche, die nachprüfen möchten, an welchen Stellen der Autor die „volle Wahrheit“ vereinfacht darstellt. Sie können aber übersprungen werden, ohne im weiteren Verlauf der Vorlesung „abgehängt“ zu werden.

Definition 1.1. Ein *Zufallsexperiment*[1] $\mathscr{X}$ genügt folgenden Forderungen:

(Z1) Zu jeder Zufallsvariable X in $\mathscr{X}$ mit Wertebereich S und Funktion $\varphi : S \to S'$ gibt es die Zufallsvariable

$$X' = \varphi(X) \text{ in } \mathscr{X}$$

(und dies bestimmt $\varphi(X)$ eindeutig). $\varphi(X)$ hat dann Wertebereich S'. Für $\varphi = \mathrm{id}_S$, $\mathrm{id}_S(x) = x$ soll $\mathrm{id}_S(X) = X$ gelten.

(Vorstellung: Setze den zufälligen Wert X in die (deterministische) Funktion φ ein, erhalte einen (zufälligen) Funktionswert.)

Wenn $\psi : S' \to S''$ eine weitere Funktion ist, so können wir die Hintereinanderausführung $\psi \circ \varphi : S \to S''$ betrachten (als Abbildung von S nach S'', $(\psi \circ \varphi)(a) = \psi(\varphi(a))$ für $a \in S$). Es soll dann stets gelten

$$(\psi \circ \varphi)(X) = \psi\big(\varphi(X)\big).$$

[1] Wir verwenden hier eine etwas „heruntergekochte“ Version des axiomatischen Zugangs von Götz Kersting [K09], der sich implizit auch in [KW] findet. Ausgelassene Details finden sich in Abschnitt 1.1.4.

Beachte: der Name Zufallsexperiment ist nicht in der Literatur standardisiert. Meist sprechen Stochastik- oder Statistik-(Lehr-)Bücher von Wahrscheinlichkeitsräumen, siehe Bem. 1.2 und Abschnitt 1.1.4 unten.

(Z2) Sind $X_1, X_2, \ldots, X_n$ in $\mathscr{X}$ Zufallsvariablen mit Wertebereichen $S_1, S_2, \ldots, S_n$, so gibt es die Zufallsvariable

$$X = (X_1, X_2, \ldots, X_n) \text{ in } \mathscr{X}$$

mit Wertebereich

$$\begin{aligned} S &= S_1 \times S_2 \times \cdots \times S_n \\ &= \{(x_1, x_2, \ldots, x_n) : x_i \in S_i \text{ für } i = 1, \ldots, n\} \end{aligned}$$

(wir schreiben auch $S = \bigtimes_{i=1}^{n} S_i$).
Sei $\pi_i : S \to S_i$, $\pi_i : (x_1, x_2, \ldots, x_n) \mapsto x_i$ die i-te Koordinatenprojektion auf S, dann soll gelten

$$\pi_i(X) = X_i \quad \text{für } i = 1, \ldots, n$$

und X dadurch eindeutig bestimmt X sein. X heißt die *Produkt(zufalls)variable* von $X_1, \ldots, X_n$.

Wir fordern auch, dass dies auch für eine unendliche Folge $X_1, X_2, \ldots$ von Zufallsvariablen möglich ist, d.h. wir können $X = (X_i)_{i \in \mathbb{N}}$ als Zufallsvariable bilden.

(Z3) Um triviale Fälle (in denen der Zufall immer „nur gleich ausgehen“ kann) auszuschließen, fordern wir von $\mathscr{X}$ eine gewisse Reichhaltigkeit: Es gibt mindestens ein Paar Zufallsvariablen X, Y mit demselben Wertebereich $\{0, 1\}$ und $X \neq Y$.

Bemerkung 1.2 („Klassischer Zugang“ zu Zufallsvariablen)**.** Sehr viele Autoren verwenden im Vergleich zu obiger Definition 1.1 einen etwas anderen Zugang zu Zufallsvariablen und Ereignissen (siehe z.B. [Ge, Kap. 1], [MP90, Kap. 1]): Man beginnt mit einer Menge Ω (dem „Stichprobenraum“), die Punkte $\omega \in \Omega$ nennt man „Elementarereignisse“. Die Vorstellung dabei ist, dass die Wirkung des Zufalls darin besteht, einen Punkt ω auszuwählen (jeweils mit einer gewissen Wahrscheinlichkeit). Wie Ω zu wählen ist, hängt von der konkreten Anwendungsfrage ab, es gibt i.A. viele Wahlmöglichkeiten. Um triviale Fälle zu vermeiden, nehmen wir an, dass Ω mindestens zwei verschiedene Punkte enthält.

Ein Ereignis ist in diesem Zugang eine (geeignete) Teilmenge $A \subset \Omega$ und eine Zufallsvariable X mit Wertebereich S eine (geeignete) Funktion $X : \Omega \to S$, in dieser Formulierung definiert man das Ereignis für (geeignete) Teilmengen $B \subset S$ des Wertebereichs $\{X \in B\} := \{\omega \in \Omega : X(\omega) \in B\} = X^{-1}(B)$ (d.h. das Urbild der Menge B unter der Abbildung X). „Geeignet“ in diesen Formulierungen bezieht sich auf gewisse maßtheoretische Bedingungen, die für den Fortgang dieser Vorlesung keine Rolle spielen und der Vollständigkeit halber in Abschnitt 1.1.4 berichtet werden.

Man kann (recht leicht) nachprüfen, dass die so definierten Zufallsvariablen die Bedingungen (Z1), (Z2), (Z3) erfüllen und die so erklärten Ereignisse die Eigenschaften aus Proposition 1.5 und Lemma 1.6 besitzen.

Diese Formulierung verwendete Andrej N. Kolmogorov in seinem berühmten Buch *Grundbegriffe der Wahrscheinlichkeitstheorie*, Springer, 1933, das gewissermaßen die Wahrscheinlichkeitstheorie als mathematische Disziplin „salonfähig“ machte; sie ist auch die in der heutigen Mathematik „übliche“. Die beiden Zugänge sind logisch äquivalent, unterscheiden sich

aber etwas in der „Betonung“. Siehe dazu auch die Diskussion in dem Buch von G. Kersting und A. Wakolbinger [KW, Vorwort] und in [K09].

Bemerkung 1.3. 1. „Deterministische Identitäten“ übertragen sich auf ZVn: $X_1, X_2, \dots, X_n$ ZVn (X_i habe Wertebereich S_i), dann gibt es gemäß (Z2) die ZV $X = (X_1, \dots, X_n)$ mit Werten in $S = \bigtimes_{i=1}^n S_i$, für eine Abbildung $\varphi : S \to S'$ können wir

$$\varphi(X_1, \dots, X_m) := \varphi(X)$$

nach (Z1) bilden. Für Abbildungen $\varphi_1 : S \to S'_1, \dots, \varphi_m : S \to S'_m$ und $\psi : S'_1 \times \cdots \times S'_m \to S''$ gilt dann

$$\psi\big(\varphi_1(X_1, \dots, X_n), \dots, \varphi_m(X_1, \dots, X_n)\big) = \big(\psi \circ (\varphi_1, \dots, \varphi_m)\big)(X_1, \dots, X_n), \tag{1.1}$$

denn $\pi_i\big((\varphi_1, \dots, \varphi_m)(X)\big) = \varphi_i(X)$ nach (Z1), also $\big(\varphi_1, \dots, \varphi_m\big)(X) = \big(\varphi_1(X), \dots, \varphi_m(X)\big)$ nach (Z2) und (1.1) folgt dann aus (Z1).

Insbesondere gelten im Fall, dass die Wertebereiche (Teilmengen von) $\mathbb{R}$ sind, die üblichen Rechenregeln. Sind z.B. X_1, X_2, X_3 in $\mathscr{X}$ ZVn mit Werten in $\mathbb{R}$, so gilt

$$(X_1 + X_2) + X_3 = X_1 + (X_2 + X_3)$$

als Zufallsvariablen. (Verwende $\varphi(x_1, x_2) := x_1 + x_2$, es ist $(X_1 + X_2) + X_3 = \big(\varphi \circ (\varphi \circ (\pi_1, \pi_2), \pi_3)\big)(X_1, X_2, X_3) = \big(\varphi \circ (\pi_1, \varphi \circ (\pi_2, \pi_3))\big)(X_1, X_2, X_3) = X_1 + (X_2 + X_3)$, dann benutze (1.1).) Analog kann man Kommutativität, Distributivgesetz, etc. prüfen.

2. Speziell gilt auch: Sind X, Y S-wertige ZVn, $\varphi : S \to S'$ injektiv (d.h. für alle $a_1, a_2 \in S$: $\varphi(a_1) = \varphi(a_2) \Rightarrow a_1 = a_2$), so gilt

$$\varphi(X) = \varphi(Y) \quad \Longrightarrow \quad X = Y$$

(Gleichheiten als Zufallsvariablen).

Sei nämlich $\varphi^{-1} : \varphi(S) \to S$ die Umkehrabbildung (d.h. $\varphi^{-1}(\varphi(a)) = \mathrm{id}_S(a) = a$ für alle $a \in S$), so ist

$$\varphi^{-1}(\varphi(X)) = \big(\varphi^{-1} \circ \varphi\big)(X) = \mathrm{id}_S(X) = X$$

gemäß (Z1) und da nach Voraussetzung $\varphi(X) = \varphi(Y)$, folgt mit analoger Rechnung $Y = \varphi^{-1}(\varphi(Y)) = \varphi^{-1}(\varphi(X)) = X$.

3. Konstante können in diesem Rahmen als Spezialfall von ZVn aufgefasst werden: S ein Wertebereich, $c \in S$ ein fester Punkt (eine „Konstante“). Sei X' irgendeine ZV (mit Wertebereich S'), $\varphi : S' \to S$ die konstante Abbildung mit Wert c ($\varphi(a) = c$ für alle $a \in S'$), dann ist $\varphi(X')$ eine Zufallsvariable und man erhält aus (Z1), (Z2) und (1.1), dass sich $\varphi(X')$ in allen „Rechnungen“ wie die Konstante c verhält (und dass es gar nicht darauf ankommt, welches X' wir verwendet haben).

4. Prinzipiell gibt es Uneindeutigkeiten in der Wahl des Wertebereiches S einer ZV X, so kann z.B. eine ganzzahlige (d.h. $\mathbb{Z}$-wertige) ZV auch immer als eine $\mathbb{R}$-wertige ZV interpretiert werden. Was wir als den „angemessenen“ Wertebereich ansehen, ist gewissermaßen Teil der „Modellierungsfrage“; es wird in unseren Beispielen i.A. aus dem Kontext klar sein, was wir als Wertebereich wählen.

1.1.2 Ereignisse

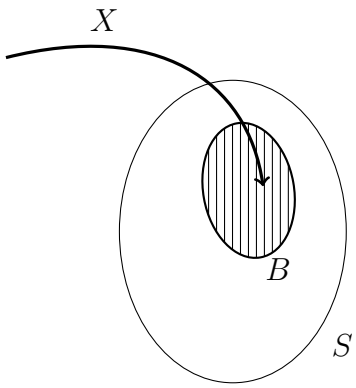
Sehr salopp gesagt übersetzen Ereignisse Zufallsvariablen in „einen Haufen ja-nein-Fragen“: Zu jedem X in $\mathscr{X}$ mit Wertebereich S und jeder[2] Teilmenge $B \subset S$ gibt es das *Ereignis*

$$\{X \in B\}$$

(ausgesprochen als „X nimmt einen Wert in B an“). Wir schreiben auch oft abkürzend für $x \in S$

$$\{X = x\} := \{X \in \{x\}\}$$

(ausgesprochen als „X nimmt den Wert x an“).



„Logo-Bild“ für eine Zufallsvariable X (und das Ereignis $\{X \in B\}$).

Identität von Ereignissen (via Indikatorfunktionen): Für $B \subset S$ sei $\mathbf{1}_B : S \to \{0,1\}$

$$\mathbf{1}_B(a) = \begin{cases} 1, & a \in B, \\ 0, & a \notin B \end{cases}$$

die *Indikatorfunktion*.

Seien X Zufallsvariable mit Wertebereich S, X' ZV mit Wertebereich S', $B \subset S, B' \subset S'$. Wir betrachten die Ereignisse $\{X \in B\}$ und $\{X' \in B'\}$ als gleich, wenn gilt $\mathbf{1}_B(X) = \mathbf{1}_{B'}(X)$ (und schreiben dies auch als $\{X \in B\} = \{X' \in B'\}$).

(In diesem Sinne entspricht ein Ereignis einer Äquivalenzklasse von Paaren (X, B) bestehend aus einer ZV und einer (geeigneten) Teilmenge von deren Wertebereich.)

Beispiel 1.4. 1. X Augenzahl beim Würfelwurf, $X' = X^2$ die quadrierte Augenzahl, so ist $\{X = 6\} = \{X' = 36\}$.

2. X S-wertige ZV, $B \subset S$, $\varphi : S \to S'$, $B' \subset S'$, so ist $\{\varphi(X) \in B'\} = \{X \in \varphi^{-1}(B')\}$ (mit $\varphi^{-1}(B') = \{a \in S : \varphi(a) \in B'\}$ dem Urbild im Sinne der Abbildungen) denn $\mathbf{1}_{B'} \circ \varphi = \mathbf{1}_{\varphi^{-1}(B')}$.
(Mit $S = \{1, 2, \ldots, 6\}, S' = \{1, 4, 9, 16, 25, 36\}$, $\varphi(a) = a^2$ ist 1. ein Spezialfall von 2.)

3. (Ereignisse und Indikatorvariablen) $A = \{X \in B\}$ Ereignis, so ist

$$I_A := \mathbf{1}_B(X)$$

[2] Strenggenommen: Zu jeder „erlaubten“, d.h. im geeigneten Sinne messbaren, Teilmenge B, siehe Abschn. 1.1.4.

die *Indikatorvariable* des Ereignisses A. I_A hat Wertebereich $\{0,1\}$ und es gilt $\{I_A = 1\} = A$ (denn $\mathbf{1}_{\{1\}} = \mathrm{id}_{\{0,1\}}$ auf $\{0,1\}$, somit auch $\mathbf{1}_{\{1\}}(I_A) = \mathrm{id}_{\{0,1\}}(I_A) = I_A = \mathbf{1}_B(X)$ gemäß (Z1).)

In diesem Sinne könnte man Ereignisse also selbst auch als Zufallsvariablen auffassen.

Schreibweise. Wir bezeichnen mit $\mathscr{E}$ die Menge aller so „bildbaren" Ereignisse in einem Zufallsexperiment $\mathscr{X}$.

Man kann mit Ereignissen rechnen wie mit logischen Aussagen (bzw. wie mit Mengen, siehe auch Bemerkung 1.2); die folgende Proposition kann man ausprechen als „$\mathscr{E}$ bildet einen (σ-vollständigen) Boole'schen Verband".

Proposition 1.5. *1. Zu jedem Ereignis $E \in \mathscr{E}$ gibt es ein eindeutiges* Komplementärereignis E^c *(auch Gegenereignis genannt), so dass für jede ZV X mit Wertebereich S und $B \subset S$ gilt*

$$\{X \in B\}^c = \{X \in B^c\}$$

(hierbei ist $B^c = S \setminus B = \{a \in S : a \notin B\}$ das Komplement von B in S). Es gilt $(E^c)^c = E$.

(Vorstellung: E^c tritt genau dann ein, wenn E nicht eintritt.)

2. Zu Ereignissen $A_1, A_2 \in \mathscr{E}$ gibt es eindeutig bestimmte Ereignisse

$$A_1 \cup A_2 \quad \text{(„Vereinigungsereignis", ausgesprochen als „A_1 oder A_2 treten ein")}$$
$$\text{und } A_1 \cap A_2 \quad \text{(„Schnittereignis", ausgesprochen als „A_1 und A_2 treten beide ein")}$$

so dass für X S-wertige ZV, $B_1, B_2 \subset S$ gilt $\{X \in B_1\} \cup \{X \in B_1\} = \{X \in B_1 \cup B_2\}$ und $\{X \in B_1\} \cap \{X \in B_1\} = \{X \in B_1 \cap B_2\}$; stets ist $A_1 \cup A_2 = A_2 \cup A_1$ und $A_1 \cap A_2 = A_2 \cap A_1$ und die Operationen $\cup$ und $\cap$ sind assoziativ.

Ebenso gibt es $\bigcap_{n=1}^{\infty} A_n$ und $\bigcup_{n=1}^{\infty} A_n$ für jede Folge $A_1, A_2, \dots \in \mathscr{E}$ von Ereignissen.

3. Es gibt eindeutig bestimmte Ereignisse E_s (das sichere Ereignis*) und E_u (das* unmögliche Ereignis*), $E_\mathrm{s} \neq E_\mathrm{u}$ mit $E_\mathrm{s}^c = E_\mathrm{u}$, so dass für jede ZV X mit Werten in S gilt*

$$\{X \in S\} = E_\mathrm{s}, \quad \{X \in \varnothing\} = E_\mathrm{u}.$$

(Vorstellung: E_s tritt immer ein, E_u tritt nie ein, egal, wie der Zufall ausgeht.)

Wir führen eine Relation $\subset$ auf den Ereignissen $\mathscr{E}$ ein durch

$$E \subset E' \ :\Leftrightarrow\ E \cap E' = E,$$

ausgesprochen als „E impliziert E'" oder „E zieht E' nach sich" oder „E ist ein Teilereignis von E'".

(Vorstellung: Wenn wir wissen, dass E eingetreten ist, dann sind wir sicher, dass auch E' eingetreten ist.)

Lemma 1.6. *$\subset$ definiert eine partielle Ordnung auf $\mathscr{E}$ (d.h. stets gilt $A \subset A$; $A \subset A'$ und $A' \subset A \Rightarrow A = A'$; $A \subset A'$ und $A' \subset A'' \Rightarrow A \subset A''$).*

Für jedes Ereignis A gilt $E_\mathrm{u} \subset A \subset E_\mathrm{s}$, d.h. E_s ist maximal, E_u ist minimal bezüglich $\subset$.

Sprechweisen. Ereignisse E und E' heißen *disjunkt*, wenn gilt

$$E \cap E' = E_{\mathrm{u}}$$

(d.h. E und E' können nicht gleichzeitig eintreten).

Wir notieren

$$A \smallsetminus B := A \cap B^c$$

(„A tritt ein, aber B nicht“) für Ereignisse $A, B \in \mathscr{E}$.

Beweis von Proposition 1.5. 1. Für $E = \{X \in B\}$ definiere $E^c := \{X \in B^c\}$. Wir müssen die Wohldefiniertheit prüfen: Sei dazu $\eta : \{0,1\} \to \{0,1\}$, $\eta(x) = 1-x$ (η „flippt“ ein Bit), offenbar ist $\mathbf{1}_{B^c} = \eta \circ \mathbf{1}_B$.

Wenn also auch $E = \{X' \in B'\}$ für eine gewisse ZV X' mit Werten in S' und $B' \subset S'$ gilt (d.h. $I_E = \mathbf{1}_B(X) = \mathbf{1}_{B'}(X')$), so gilt

$$\mathbf{1}_{B^c}(X) = \eta\big(\mathbf{1}_B(X)\big) = \eta\big(\mathbf{1}_{B'}(X')\big) = \mathbf{1}_{(B')^c}(X'),$$

gemäß (Z1), d.h. auch $E^c = \{X' \in (B')^c\}$.

2. Sei zunächst $A_1 = \{X \in B_1\}, A_2 = \{X \in B_2\}$ für eine gewisse ZV X. Wir setzen dann

$$\{X \in B_1\} \cup \{X \in B_1\} := \{X \in B_1 \cup B_2\}, \quad \{X \in B_1\} \cap \{X \in B_1\} := \{X \in B_1 \cap B_2\}.$$

Wiederum ist die Wohldefiniertheit zu zeigen: Wenn $\{X \in B_1\} = \{X' \in B_1'\}$, $\{X \in B_2\} = \{X' \in B_2'\}$ für eine gewisse S'-wertige ZV X' und $B_1', B_2' \subset S'$ gilt, so ist $\mathbf{1}_{B_1}(X) = \mathbf{1}_{B_1'}(X')$ und $\mathbf{1}_{B_2}(X) = \mathbf{1}_{B_2'}(X')$ und daher auch

$$\mathbf{1}_{B_1 \cup B_2}(X) = \max(\mathbf{1}_{B_1}(X), \mathbf{1}_{B_2}(X)) = \max(\mathbf{1}_{B_1'}(X'), \mathbf{1}_{B_2}(X')) = \mathbf{1}_{B_1' \cup B_2'}(X')$$

(mit Bem. 1.3, 1.), d.h. $\{X \in B_1 \cup B_2\} = \{X' \in B_1' \cup B_2'\}$ gilt; analog ist

$$\mathbf{1}_{B_1 \cap B_2}(X) = \min(\mathbf{1}_{B_1}(X), \mathbf{1}_{B_2}(X)) = \min(\mathbf{1}_{B_1'}(X'), \mathbf{1}_{B_2}(X')) = \mathbf{1}_{B_1' \cap B_2'}(X'),$$

d.h. $\{X \in B_1 \cap B_2\} = \{X' \in B_1' \cap B_2'\}$.

Kommutativität von $\cup$ und von $\cap$ sowie Assoziativität folgen genauso aus den entsprechenden Eigenschaften der Mengen- bzw. Bit-Operationen.

Zu zeigen bleibt: Beliebige Ereignisse A_1 und A_2 aus $\mathscr{E}$ können in dieser Form (d.h. als $A_1 = \{X \in B_1\}, A_2 = \{X \in B_2\}$ mit derselben ZV X dargestellt werden): Setze dazu $X := (I_{A_1}, I_{A_2})$ mit Werten in $\{0,1\}^2$, wobei I_{A_1}, I_{A_2} die zugehörigen Indikatorvariablen gemäß Bsp. 1.4, 3. sind. Damit ist

$$\big\{X \in \{(1,0),(1,1)\}\big\} = \big\{X \in \pi_1^{-1}(\{1\})\big\} = \{\pi_1(X) = 1\} = \{I_{A_1} = 1\} = A_1,$$

analog ist $\{X \in \{(0,1),(1,1)\}\} = A_2$.

Der allgemeine Fall ($n \in \mathbb{N}$ oder abzählbar unendlich viele Ereignisse) kann mit etwas mehr Notationsaufwand analog behandelt werden (Details in [K09]).

3. Sei X ZV mit Wertebereich S, setze $\{X \in S\} := E_\mathrm{s}$, $\{X \in \varnothing\} = E_\mathrm{u}$. Zur Wohldefiniertheit: Sei X' mit Wertebereich S' eine weitere ZV, $Y := (X, X')$ und π_1, π_2 die Koordinatenprojektionen auf $S \times S'$, dann ist $\mathbf{1}_S(X) = (\mathbf{1}_S \circ \pi_1)(Y) = (\mathbf{1}_{S'} \circ \pi_2)(Y) = \mathbf{1}_{S'}(X')$ und damit $\{X \in S\} = \{X' \in S'\}$. Analog ist $\{X \in \varnothing\} = \{X' \in \varnothing\}$.

Offenbar gilt zusammen mit 1. $E_\mathrm{s}^c = \{X \in S\}^c = \{X \in S^c\} = \{X \in \varnothing\} = E_\mathrm{u}$.

Es bleibt zu zeigen, dass $E_\mathrm{s} \neq E_\mathrm{u}$ gilt. Wäre $E_\mathrm{s} = E_\mathrm{u}$, so wäre für jede ZV X mit Wertebereich S und jedes $B \subset S$

$$\{X \in B\} = \{X \in B \cap S\} = \{X \in B\} \cap \{X \in S\} = \{X \in B\} \cap \{X \in \varnothing\} = \{X \in \varnothing\} = E_\mathrm{u},$$

d.h. dann wären alle Ereignisse $= E_\mathrm{u}$ und alle Indikatorvariablen wären gleich. Dies widerspricht der Forderung (Z3). □

Beweis von Lemma 1.6. Wie im Beweis von Proposition 1.5 können wir $A, A', A'' \in \mathscr{E}$ darstellen als

$$A = \{X \in B\},\ A' = \{X \in B'\},\ A'' = \{X \in B''\}$$

für eine geeignete ZV X mit Werten in einem S und geeigneten Teilmengen $B, B', B'' \subset S$. Nach Definition gilt $A \cap A = \{X \in B \cap B\} = \{X \in B\} = A$, also $A \subset A$; $A \cap E_\mathrm{u} = \{X \in B \cap \varnothing\} = \{X \in \varnothing\} = E_\mathrm{u}$, also $E_\mathrm{u} \subset A$; analog gilt $A \subset E_\mathrm{s}$.

Sei $A \subset A'$ und $A' \subset A$, dann ist aus der Definition (und Kommutativität von $\cap$)

$$A = A \cap A' = A' \cap A = A';$$

sei $A \subset A'$ und $A' \subset A''$, dann ist wiederum aus der Definition (und Assoziativität von $\cap$)

$$A \cap A'' = (A \cap A') \cap A'' = A \cap (A' \cap A'') = A \cap A' = A,$$

also $A \subset A''$. □

1.1.3 Wahrscheinlichkeiten

Wir betrachten ein Zufallsexperiment $\mathscr{X}$ und die zugehörige Menge von Ereignissen $\mathscr{E}$.

Definition 1.7. Für ein Ereignis A nennen wir $P(A) \in [0, 1]$ die *Wahrscheinlichkeit* von A. P erfüllt folgende Forderungen:

(N) $P(E_\mathrm{s}) = 1$ („Normierung“) und

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

Proposition 1.8. *Es gilt*

$$P(E_{\mathrm{u}}) = 0 \tag{1.2}$$

$$P(A \cup B) + P(A \cap B) = P(A) + P(B) \qquad \textit{(endliche Additivität)}, \tag{1.3}$$

$$\textit{insbesondere } P(A) + P(A^c) = 1$$

$$A \subset B \Rightarrow\ P(A) \le P(B) \qquad \textit{(Monotonie)} \tag{1.4}$$

zudem gilt auch

$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) \le \sum_{n=1}^{\infty} P(A_n) \qquad \textit{(σ-Subadditivität)} \tag{1.5}$$

$$A_n \nearrow_{n\to\infty} A \textit{ (d.h. } A_1 \subset A_2 \subset \dots \textit{ mit } A = \textstyle\bigcup_{n=1}^{\infty} A_n\textit{)}$$

$$\textit{oder } A_n \searrow_{n\to\infty} A \textit{ (d.h. } A_1 \supset A_2 \supset \dots \textit{ mit } A = \textstyle\bigcap_{n=1}^{\infty} A_n\textit{)},$$

$$\textit{so gilt } P(A) = \lim_{n\to\infty} P(A_n) \qquad \textit{(σ-Stetigkeit)} \tag{1.6}$$

Beweis. (1.2): $P(E_{\mathrm{u}}) = P\big(E_{\mathrm{u}} \cup E_{\mathrm{u}} \cup \dots\big) = \sum_{n=1}^{\infty} P(E_{\mathrm{u}})$, also $P(E_{\mathrm{u}}) = 0$.

(1.3): Betrachte zunächst den Fall $A \cap B = E_{\mathrm{u}}$:

$$P(A\cup B) = P\big(A\cup B\cup E_{\mathrm{u}}\cup E_{\mathrm{u}}\cup\dots\big) = P(A)+P(B)+\underbrace{P(E_{\mathrm{u}})}_{=0}+\underbrace{P(E_{\mathrm{u}})}_{=0}+\cdots,$$ d.h. (1.3) gilt hier.

Allgemeiner Fall: Schreibe

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

also

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

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

(1.5): 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

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

(die zweite Gleichung verwendet (A), die letzte (Un-)gleichung verwendet (1.4)).

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

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

Falls $A_n \searrow A$, so beachte, dass $A_n^c \nearrow A$ gilt, dann verwende obiges zusammen mit (1.3). □

Bemerkung 1.9. 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 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 [Ge], S. 14 am Ende von Abschn. 1.1.3).

Beispiel 1.10 (**Kollision von Kennzeichen** (oder „Hash-Werten“))**.** n Objekte erhalten „rein zufällig“ ein Kennzeichen aus einer Menge von r möglichen Werten (die wir als $\{1, 2, \ldots, r\}$ notieren). Beispielsweise:

- Wir fragen n „zufällig gewählte“ Personen nach ihrem Geburtstag: Wähle $r = 365$ (wir ignorieren Schaltjahre).
- Wir haben n verschiedene Daten, der MD5-Algorithmus (z.B. im Unix-Kommando `md5sum`) berechnet für jede Datei einen 128 bit-Hashwert (und wir blenden die tatsächlichen Details des Algorithmus und der Dateiinhalte aus und tun so, als ob der Hash-Wert jeweils „zufällig“ wäre). Wähle $r = 2^{128} \approx 3.4 \cdot 10^{38}$.

Wir formalisieren dies im Sinne von Definition 1.1 mit einer Zufallsvariable

$$X = (X_1, \ldots, X_n)$$

(X_i sei das Kennzeichen, das das i-te Objekt erhält) mit Werten in

$$S = \{1, \ldots, r\}^n,$$

wobei jeder mögliche Wert mit der gleichen Wahrscheinlichkeit angenommen werde, d.h.

$$P\big(\{X = x\}\big) = \frac{1}{|S|} \quad \text{für jedes } x = (x_1, \ldots, x_n) \in S.$$

Betrachte das Ereignis

$$A = \{X_i \neq X_j \text{ für } 1 \leq i \neq j \leq n\} = \{X \in B\} \quad (\text{„alle Kennzeichen sind verschieden“})$$

mit

$$B = \big\{(x_1, \ldots, x_n) : x_i \neq x_j \text{ für alle } 1 \leq i \neq j \leq n\big\}$$

Frage:

$$P(A) = ?$$

(Man nennt dies auch manchmal das „Geburtstagsproblem“.)

Es ist

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

(r Wahlmöglichkeiten für x_1, dann $r-1$ Wahlmöglichkeiten für x_2, etc.), also

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

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

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

für eine untere Schranke beachte

$$A^c = \{X_i = X_j \text{ für ein Paar } i \neq j\big\} = \bigcup_{1\leq i<j\leq n} \{X_i = X_j\}.$$

Demnach

$$P(A^c) \leq \sum_{1\leq i<j\leq n} P(X_i = X_j) = \sum_{1\leq i<j\leq n} \frac{r^{n-1}}{r^n} = \frac{n(n-1)}{2r},$$

insgesamt

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

Wir sehen: Kollision hat substantielle W’keit, wenn $n^2 \gtrsim r$ (d.h. $n \gtrsim \sqrt{r}$) gilt.

1.1.4 „Kleingedrucktes“: Bericht zur Maßtheorie

Hinweis Dieser Abschnitt dient nur der Vollständigkeit, zum Vergleich mit der (Lehrbuch-) Literatur und zur Information besonders interessierter (und ggfs. schon anderweitig vorgebildeter) Leser. Alle in dieser Vorlesung in Beispielen vorkommenden Mengen und Funktionen werden geeignete Messbarkeitseigenschaften haben, auch wenn wir dies nicht explizit erwähnen.

In der Situation, dass man überabzählbar große Wertebereiche S für Zufallsvariablen betrachten möchte (was beispielsweise schon den sehr naheliegenden Fall $S = \mathbb{R}$ betrifft), muss man – damit es am Ende tatsächlich eine Wahrscheinlichkeit auf den Ereignissen mit den Eigenschaften aus Def. 1.7 gibt – in Def. 1.1 gewisse Einschränkungen an die Menge der erlaubten Wertebereiche S und Abbildungen φ zwischen Wertebereichen machen.

Im Jargon der Maßtheorie muss jedes solche S zunächst ein *messbarer Raum* sein, d.h. es ist eine σ-Algebra, d.h. eine Teilmenge $\mathscr{S} \subset 2^S := \{B : B \subset S\}$ der Potenzmenge von S ausgezeichnet, die erfüllt:

$$\varnothing \in \mathscr{S}, \;\; A \in \mathscr{S} \;\Rightarrow\; A^c \in \mathscr{S}, \;\; A_1, A_2, \cdots \in \mathscr{S} \;\Rightarrow\; \bigcup_{n=1}^{\infty} A_n \in \mathscr{S}$$

(man sieht leicht, dass dann $\mathscr{S}$ bezüglich denselben Operationen abgeschlossen ist wie die Ereignisse in Prop. 1.5). Falls S abzählbar ist, so kann man einfach $\mathscr{S} = 2^S$ wählen, was offenbar allen obigen Bedingungen genügt. Im überabzählbaren Fall geht dies i.A. nicht, siehe z.B. die Diskussion in [Ge, Abschnitt 1.1.2] („Warum so vorsichtig?“).

Zudem muss für unseren Zugang $\mathscr{S}$ eine abzählbare, Punkte trennende Erzeugermenge besitzen (“measurable space with denumerable separation”, siehe [K09, Section 1.3]) – dies ist für alle „interessanten“ Wertebereiche der Fall (es gilt beispielweise für jeden separablen metrischen Raum, der mit seiner Borel-σ-Algebra ausgestattet ist, insbesondere für $\mathbb{R}^d$).

Sind weiter $(S, \mathscr{S})$ und $(S', \mathscr{S}')$ messbare Räume (die wir als Wertebereiche für die Zufallsvariablen aus einem Zufallsexperiment $\mathscr{X}$ ins Auge fassen), so lassen wir nicht alle Abbildungen $\varphi : S \to S'$ zu, sondern nur *messbare* (streng: $\mathscr{S}$-$\mathscr{S}'$-messbare) Abbildungen φ, d.h. es muss gelten

$$\forall\, B' \in \mathscr{S}' : \; \varphi^{-1}(S') = \big\{a \in S : \varphi(a) \in B'\big\} \in \mathscr{S}.$$

Weiterhin sind für eine S-wertige ZV X als Ereignisse in Abschnitt 1.1.2 nur $\{X \in B\}$ mit $B \in \mathscr{S}$ zugelassen (für $B \subset S$ mit $B \notin \mathscr{S}$ wird die Frage, ob X in B liegt, nicht „erlaubt“).

Ebenso muss man im „klassischen“ Zugang wie in Bem. 1.2 den Stichprobenraum Ω mit einer σ-Algebra $\mathscr{F} \subset 2^\Omega$ ausstatten, die Elemente $A \in \mathscr{F}$ übernehmen dann die Rolle der Ereignisse und die Wahrscheinlichkeit P ist (nur) für $A \in \mathscr{F}$ definiert. Das Tripel $(\Omega, \mathscr{F}, P)$ heißt dann typischerweise ein *Wahrscheinlichkeitsraum*. Als Zufallsvariablen kann man wiederum nicht alle Funktionen $X : \Omega \to S$ zulassen, sondern nur $\mathscr{F}$-$\mathscr{S}$-messbare.

1.1.5 Verteilung von Zufallsvariablen (diskreter Fall)

Schreibweise. Wir kürzen 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.

Definition 1.11. Eine ZV X heißt *diskret*, wenn ihr Wertebereich S (endlich oder) abzählbar ist oder zumindest eine (endliche oder) abzählbare Teilmenge S enthält mit $P(X \in S) = 1$.

$$\rho_X : A \mapsto P(X \in A), \quad A \subset S$$

heißt die *Verteilung* von X. Die Zahlen

$$\rho_X(\{a\}) := P(X = a), \quad a \in S$$

heißen die *Verteilungsgewichte* von X (oft auch nur: die *Gewichte*). Wir kürzen oft ab (mit einem kleinen „Notationsmissbrauch“) $\rho_X(a) := \rho_X(\{a\})$.

Offenbar gilt (mit Eigenschaft (A) aus Def. 1.7) $\rho_X(a) \geq 0$,

$$\sum_{a \in A} \rho_X(a) = P(X \in A) \;\text{ für } A \subset S, \qquad \text{insbesondere } \sum_{a \in S} \rho_X(a) = P(X \in S) = 1$$

und ρ_X ist ein *Wahrscheinlichkeitsmaß* (auch *Wahrscheinlichkeitsverteilung*) auf S (d.h. ρ_X : $\{\text{Teilmengen von } S\} \to [0,1]$ und die (analogen) Eigenschaften aus Def. 1.7 und aus Prop. 1.8 gelten, mit Lesung $E_\text{s} = S$, $E_\text{u} = \varnothing$).

Man schreibt für die Verteilung einer ZV X oft auch $\mathscr{L}(X)$ (das $\mathscr{L}$ erinnert an English "law" bzw. Französisch «loi», d.h. „Gesetz"), man schreibt auch $X \sim \rho$, wenn $\mathscr{L}(X) = \rho$ gilt (für eine gewisse Wahrscheinlichkeitsverteilung).

Beispiel 1.12 (Uniforme verteilung auf endlicher Menge)**.** $\#S < \infty$, eine ZV X (mit Werten in S) ist *uniform verteilt auf* S (auch $\mathscr{L}(X) = \text{unif}_S$ oder $S \sim \text{unif}_S$) geschrieben, wenn gilt

$$\rho_X(x) = P(X = x) = \frac{1}{\#S} \quad \text{für } x \in S.$$

Man nennt eine uniforme Verteilung auch eine *Laplace-Verteilung*[3].

Beispiel-Instanzen:

- $S = \{\text{Kopf}, \text{Zahl}\}$, (fairer) Münzwurf
- $S = \{1, 2, \ldots, 20\}$, Wurf eines (fairen) 20-seitigen Würfels (d.h. eines symm. Ikosaeders)
- $S = \{1, 2, \ldots, r\}^n$ aus Bsp. 1.10
- $S = \{0, 1, \ldots, \ell_x - 1\} \times \{0, 1, \ldots, \ell_y - 1\}$, wenn wir im „Auftakt"-Beispiel aus Kap. 0 das Quadrat in $\ell_x \times \ell_y$ viele Pixel unterteilt wählen (mit $\ell_x, \ell_y \in \mathbb{N}$)

Definition 1.13 (Gemeinsame Verteilung und Marginalverteilungen)**.** Seien $X_1, X_2, \ldots, X_d$ ZVn (in einem gewissen Zufallsexperiment $\mathscr{X}$, X_i habe Werte in S_i),

$$X := \big(X_1, X_2, \ldots, X_d\big)$$

(gemäß (Z2) aus Def. 1.1) die Produkt-Zufallsvariable (mit Werten in $S_1 \times \cdots \times S_d$), so heißt $\mathscr{L}(X)$ die *gemeinsame Verteilung* der $X_1, X_2, \ldots, X_d$, $\mathscr{L}(X_i)$ heißt die i-te *Randverteilung* (oder *Marginalverteilung*) von X.

Speziell heißen $X_1, X_2, \ldots, X_d$ (stochastisch) *unabhängig*, wenn für alle Ereignisse $\{X_i \in B_i\}$ gilt

$$P\big(X_1 \in B_1, X_2 \in B_2, \ldots, X_d \in B_d\big) = \prod_{i=1}^{d} P(X_i \in B_i) \tag{1.7}$$

Bemerkung. 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

[3] nach Pierre-Simon Laplace, 1749–1927

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 ergäben sich, wenn man zwei Mal M_1 wirft, aber die gemeinsame Verteilung wäre eine andere).

Beobachtung 1.14. *$X_1, X_2, \ldots, X_n$ ZVn, X_i habe Werte in S_i, S_i abzählbar für $i = 1, 2, \ldots, n$. Dann sind $X_1, X_2, \ldots, X_n$ unabhängig g.d.w. gilt*

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

Beweis. „⇒“: Klar (Wähle $B_i = \{x_i\}$ in (1.7)).

„⇐“: Seien $B_1 \subset S_1, \ldots B_n \subset S_n$, es ist

$$\begin{aligned} P\big(\{X_1 \in B_1\} \cap \cdots \cap \{X_n \in B_n\}\big) &= P\Big(\bigcup_{(x_1,\ldots,x_n)\in B_1\times\cdots\times B_n} \{X_1 = x_1, \ldots, X_n = x_n\}\Big) \\ &= \sum_{x_1\in B_1,\ldots,x_n\in B_n} \underbrace{P\big(X_1 = x_1, \ldots, X_n = x_n\big)}_{=P(X_1=x_1)\cdots P(X_n=x_n)} \\ &= \sum_{x_1\in B_1} P(X_1 = x_1)\cdots \sum_{x_n\in B_n} P(X_n = x_n) \\ &= P(X_1 \in B_1)\cdots P(X_n \in B_n). \end{aligned}$$

□

„Klassische“ diskrete Verteilungen

Beispiel 1.15 (Urnenmodelle)**.** Eine Urne enthalte n (mit $1, 2, \ldots, n$ nummerierte) Kugeln, wir ziehen zufällig k ($\leq n$) heraus. Sei

$$X_i \text{ die Nummer der Kugel im } i\text{-ten Zug}, \quad i = 1, \ldots, k$$

und $X = (X_1, X_2, \ldots, X_k)$.

Wir betrachten 4 mögliche Situationen:

1. Ziehen mit Zurücklegen, mit Beachtung der Reihenfolge:

   X hat Werte in $W_1 = \big\{(x_1, \ldots, x_k) : x_1, \ldots, x_k \in \{1, 2, \ldots, n\}\big\} = \{1, 2, \ldots, n\}^k$,

   $P\big(X = (x_1, \ldots, x_k)\big) = \dfrac{1}{n^k}$

   (d.h. X ist uniform auf $\{1, \ldots, n\}^k$ verteilt, vgl. auch Beispiel 1.10).

2. Ziehen ohne Zurücklegen, mit Beachtung der Reihenfolge:

$$X \text{ hat Werte in } W_2 = \big\{(x_1,\dots,x_k) : x_1,\dots,x_k \in \{1,2,\dots,n\}, x_i \neq x_j \text{ für } i \neq j\big\},$$
$$P\big(X = (x_1,\dots,x_k)\big) = \frac{1}{n\cdot(n-1)\cdots(n-k+1)} = \frac{(n-k)!}{n!} \quad \Big(= \frac{1}{|W_2|}\Big)$$

3. Ziehen ohne Zurücklegen, ohne Beachtung der (Zug-)Reihenfolge: Wir beobachten nicht X (das hier wie in 2. verteilt wäre), sondern mit

$\varphi\big((x_1,\dots,x_k)\big) = \{x_1,\dots,x_k\}$ (verwandle Vektor in Menge, d.h. vergiss die Reihenfolge)

(nur) $Y = \varphi(X)$ mit Werten in $W_3 = \big\{A \subset \{1,2,\dots,n\} : \#A = k\big\}$.
Es ist

$$P(Y = A) = \frac{1}{\binom{n}{k}} = \frac{k!(n-k)!}{n!} \quad \Big(= \frac{1}{|W_3|},\ \text{es gibt } \binom{n}{k} \text{ versch. } k\text{-elementige Teilmengen}\Big)$$

denn

$$\begin{aligned} P(Y = A) &= P\big(\varphi(X) = A\big) = P\big(X \in \varphi^{-1}(A)\big) \\ &= \sum_{x \in W_2 : \varphi(x) = A} P(X = x) = \sum_{x \in W_2 : \varphi(x) = A} \frac{(n-k)!}{n!} = k!\frac{(n-k)!}{n!} \end{aligned}$$

(es gibt $k!$ viele verschiedene Elemente x von W_2 mit $\varphi(x) = A$, nämlich alle verschiedenen Anordnungen der k Elemente von A).

4. Ziehen mit Zurücklegen, ohne Beachtung der (Zug-)Reihenfolge: Sei $X = (X_1,\dots,X_k)$ wie in 1., wir beobachten aber (nur) $Y = (Y_1,\dots,Y_n)$, wobei

$$Y_j = \#\{1 \le i \le k : X_i = j\} \quad \text{für } j = 1,\dots,n.$$

(Y_j gibt an, wie oft Kugel j gezogen wurde). Y hat Werte in

$$W_4 = \big\{(\ell_1,\ell_2,\dots,\ell_n) \in \mathbb{N}_0^n : \ell_1 + \ell_2 + \cdots + \ell_n = k\big\}$$

Für $(\ell_1,\ell_2,\dots,\ell_n) \in W_4$ gibt es

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

verschiedene $x = (x_1,\dots,x_k) \in W_1$ mit

$$\big|\{1 \le i \le k : x_i = j\}\big| = \ell_j \quad \text{für } j = 1,\dots,n.$$

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

Bemerkung 1.16. Es gilt

$$|W_4| = \binom{n+k-1}{k} \quad \left(= \binom{n+k-1}{n-1} \right)$$

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

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

Insbesondere ist die Verteilung auf W_4 aus Beispiel 1.15, 4. nicht die uniforme.

Die uniforme Verteilung auf dem W_4 aus Beispiel 1.15, 4. heißt auch die „Bose-Einstein-Verteilung", die in Beispiel 1.15, 4. betrachtete Verteilung heißt 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 1.17 (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 1.18 (p-Münzwurf)**.** 1. $S = \{0,1\}$, $\mathrm{Ber}_p(\{1\}) = p = 1 - \mathrm{Ber}_p(\{0\})$ mit einem $p \in [0,1]$ („Bernoulli-Verteilung"[4])

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

$$\mathrm{Ber}_p^{\otimes n}\big(\{(x_1, \ldots, x_n)\}\big) = p^{|\{i \leq n \,:\, x_i = 1\}|}(1-p)^{|\{i \leq n \,:\, x_i = 0\}|}$$

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

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

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

[4] nach Jakob Bernoulli, 1654–1705

Beispiel 1.19 (Geometrische Verteilung)**.** $p \in (0,1)$, $S = \mathbb{N}_0$,

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

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

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

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 1.21 (Poissonverteilung[5])**.** $\lambda \in (0,\infty)$,

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

Proposition 1.22 (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 $k \in \mathbb{N}_0$*

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

Beweis. Es ist

$$\begin{aligned}\binom{n}{k}p_n^k(1-p_n)^{n-k} &= \underbrace{\frac{n(n-1)\cdots(n-k+1)}{k!\,n^k}}_{\to 1/k!}\Big(\underbrace{np_n}_{\to\lambda}\Big)^k\underbrace{\Big(1-\frac{np_n}{n}\Big)^n}_{\to e^{-\lambda}}\big(1-p_n\big)^{-k}\\ &\longrightarrow e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{für } n\to\infty.\end{aligned}$$

□

Prop. 1.22 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 1.23. L. von von Bortkewitsch[6] 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):

[5] nach Siméon Denis Poisson, 1781–1840

[6] Ladislaus von Bortkewitsch, 1868–1931

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

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

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.

```
# Daten aus
# Ladislaus von Bortkewitsch,
# Das Gesetz der kleinen Zahlen, Teubner, 1898.
# Kap. 12.4 Beispiel: Die durch Schlag eines Pferdes
#           im preussischen Heer getoeteten
#
# Fuer 20 Jahre (1875-1894) und 10 Armeekops der
# preussischen Kavallerie wird berichtet, in wievielen
# "Korpsjahren" x Soldaten des Korps in diesem Jahr durch
# Schlag eines Pferdes starben (x=0,1,2,3,4 oder >=5)
beob <- c(109,65,22,3,1,0)
label <- c("0","1","2","3","4",">=5")
```

```
# angenommen, die Anzahl durch Schlag eines Pferdes
# waehrend eines Jahres in einem Korps
# getoeter Soldaten waere Poisson(0.61)-verteilt,
# dann sollten wir finden:
lambda <- 0.61
angepasst <- round(200*c(dpois(0:4,lambda), 1-ppois(4,lambda)), digits=2)

# die Daten und die theoretischen Werte passen gut zusammen:
cat('Ergebn.\t Daten \t Theorie \n');
for (i in 1:6) cat(paste(label[i],'\t',beob[i],'\t',angepasst[i],'\n'))
#   Ergebn.      Daten   Theorie
#   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 laesst [...], wie man sieht,
#  nichts zu wuenschen uebrig."

# (Uebrigens: 0.61 ist der Momentenschaetzer)
sum(beob[1:5]*(0:4))/200
```

1.1.6 Verteilungen mit Dichte

Zufallsvariablen mit Dichten sind ein kontinuierliches Analogon zu Zufallsvariablen mit Gewichten. In vielen Situationen ist eine Modellierung eines zufälligen Werts X als „allgemeine“ reelle Zahl angemessen, d.h. die Annahme, dass der Wertebereich S diskret ist, ist zu „eng“.
(Auch wenn man argumentieren könnte, dass die Menge der im Rechner mit gegebener Genauigkeit darstellbaren Werte prinzipiell diskret ist, ist es oft „unpraktisch“, sich immer auf eine konkrete Diskretisierung festlegen zu müssen.)

Beispiel 1.24 (Approximation der Exponentialverteilung durch reskalierte geometrisch verteilte ZVn)**.** Sei $W \sim \mathrm{Geom}_p$ mit $p \ll 1$, $X := pW$ (hat Werte in $p\mathbb{Z}_+ \subset \mathbb{R}$). Für jedes feste $K \in \mathbb{N}$ ist

$$P(W \geq K) = \sum_{j=K}^{\infty} P(W = j) = \sum_{j=K}^{\infty} p(1-p)^j = (1-p)^K p \sum_{\ell=0}^{\infty} (1-p)^\ell = (1-p)^K p \frac{1}{1-(1-p)} = (1-p)^K$$

(≈ 1, wenn p sehr klein), andererseits ist für (festes) $x > 0$

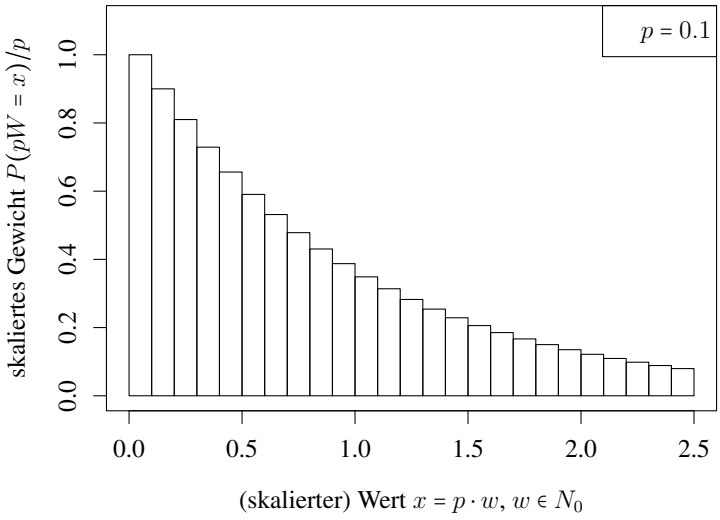
$$P(X \geq x) = P(pW \geq x) = P\big(W \geq \lceil x/p \rceil\big) = (1-p)^{\lceil x/p \rceil} \approx (1-p)^{x/p} \approx e^{-x}$$

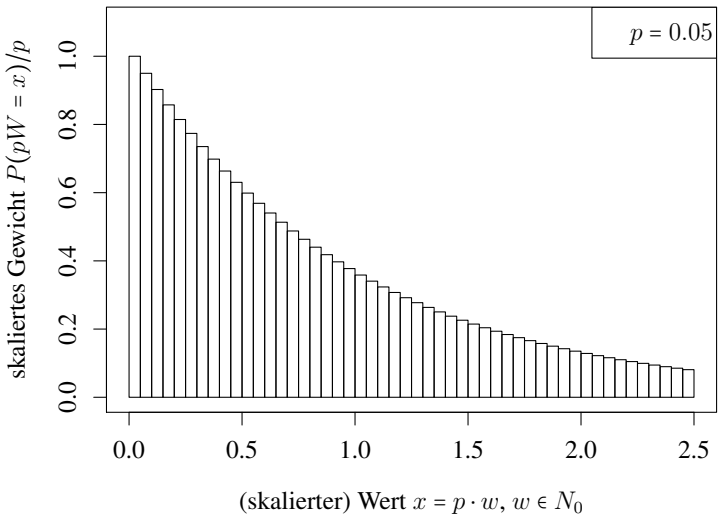
(und zwar “egal, wie klein” p ist).

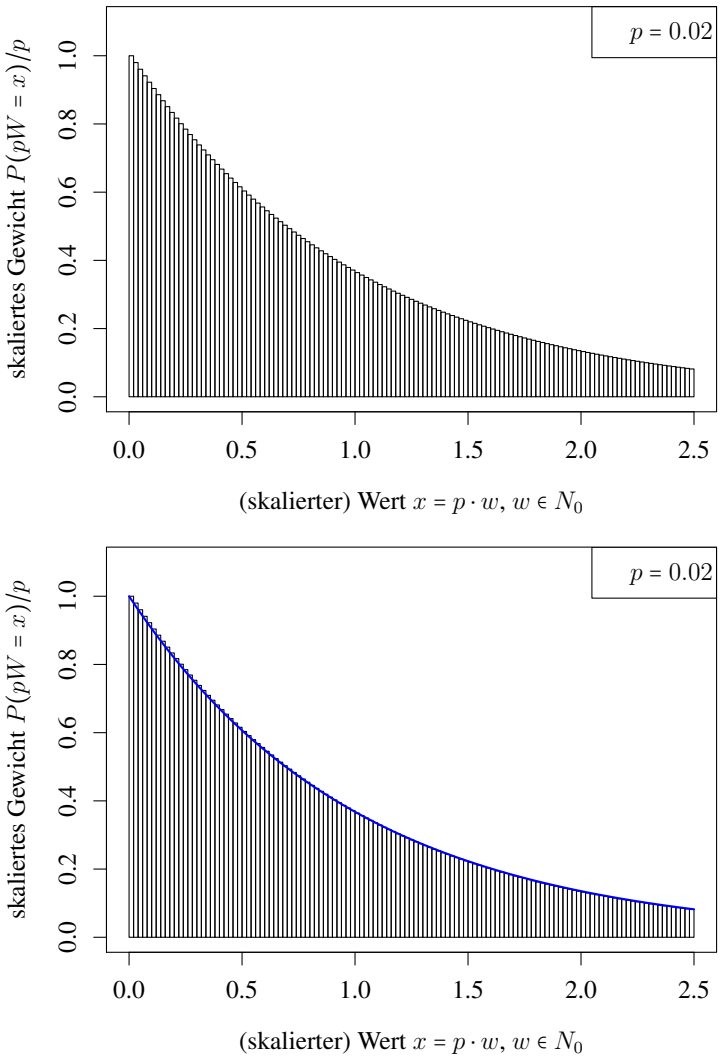
Interpretation z.B.: via Wartezeiten auf sehr fein zeitdiskretisiertem Gitter

Frage also: Gibt es eine reellwertige ZV X, für die obiges ($P(X \geq x) = e^{-x}$) als Identität gilt? Ja, wie wir sehen werden.

(In den folgenden Bildern skalieren wir die Höhe der Balken so, dass die Fläche des Balkens bei $x = pw$ gerade $P(pW = x) = p \cdot P(pW = x)/p$ entspricht.)







(Die blaue Kurve ist e^{-x})

Beispiel 1.25 (Approximation einer Normalverteilung durch reskalierte binomialverteilte ZVn)**.** Betrachten wir für $S_n \sim \mathrm{Bin}_{n,1/2}$ (und verschiedene n) die Verteilungsgewichte, so beobachten wir, dass sich die „Form stabilisiert“.

Zentrieren und stauchen wir:

$$X_n := \frac{S_n - n/2}{\sqrt{n}}$$

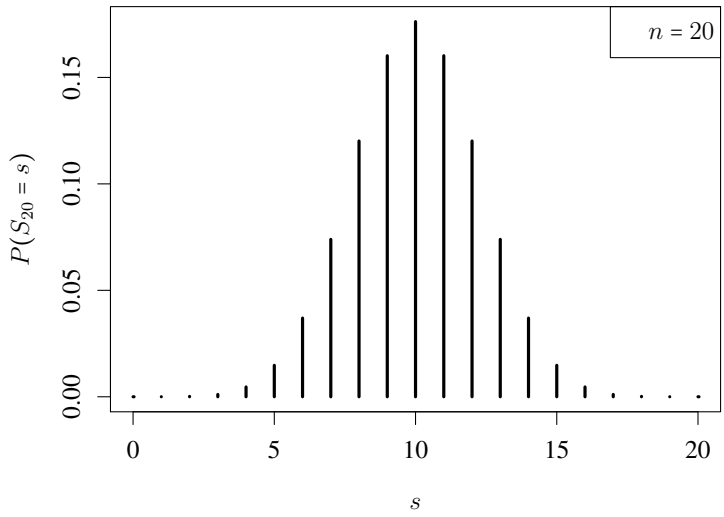
und skalieren die „Balken“ so, dass die Fläche des Balkens bei $(s - n/2)/\sqrt{n}$ gerade

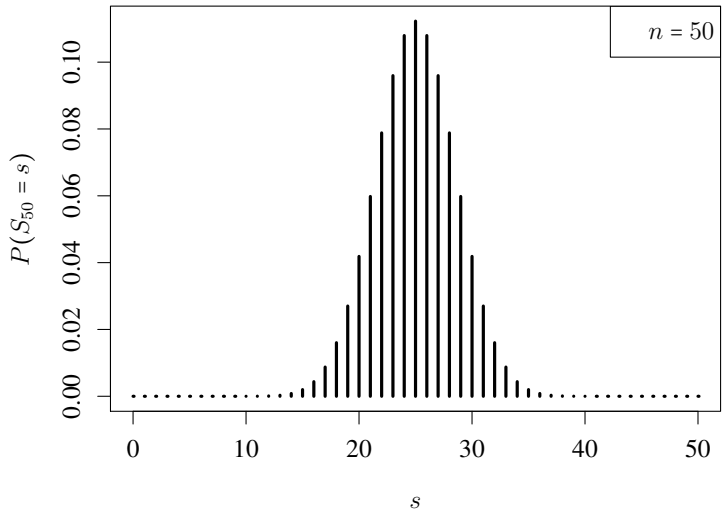
$$P(S_n = s) = \frac{1}{\sqrt{n}} \times \sqrt{n} P(S_n = s)$$

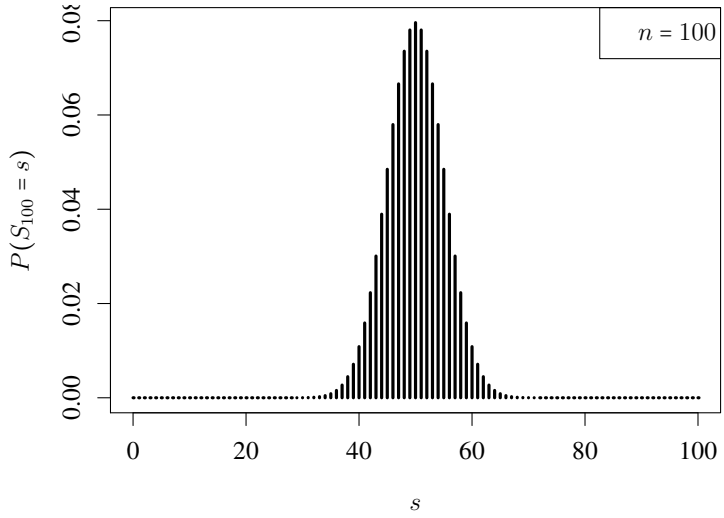
entspricht.

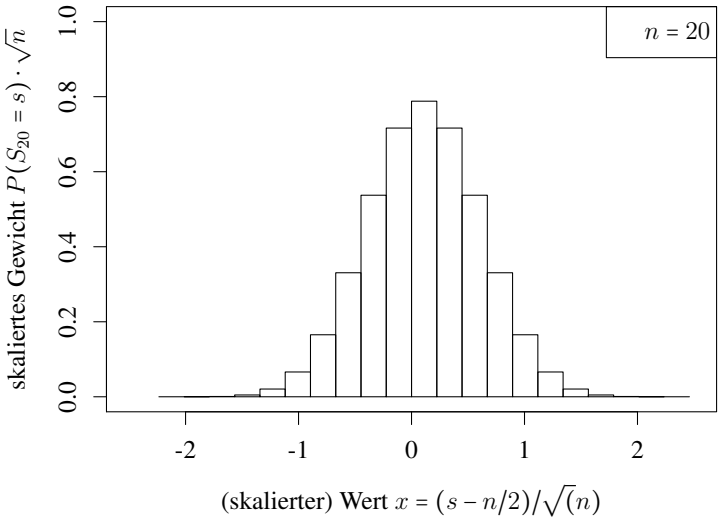
Es zeigt sich: Die Balken werden gut durch die Funktion $x \mapsto \frac{1}{\sqrt{\pi/2}} \exp(-2x^2)$ beschrieben.

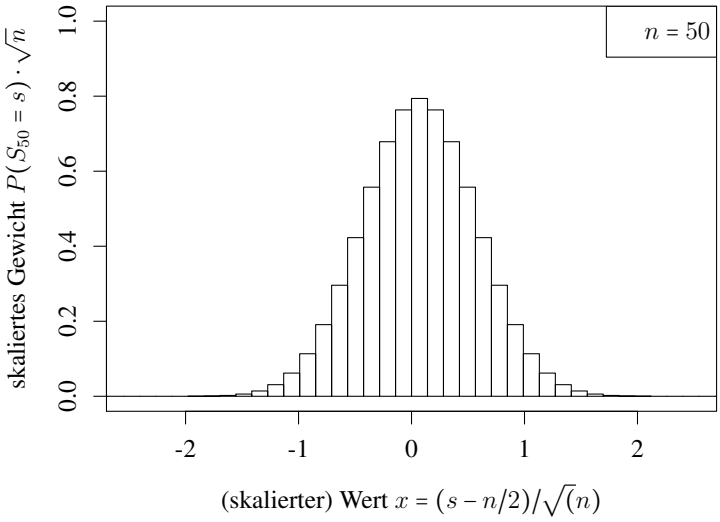
(dies ist eine „Vorschau“ auf den zentralen Grenzwertsatz)

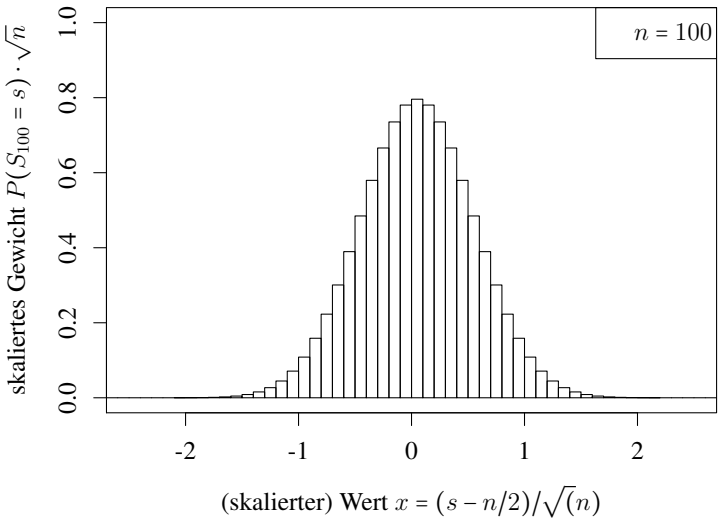

n = 20
$P(S_{20} = s)$
0.00
0.05
0.10
0.15
0
5
10
15
20
s

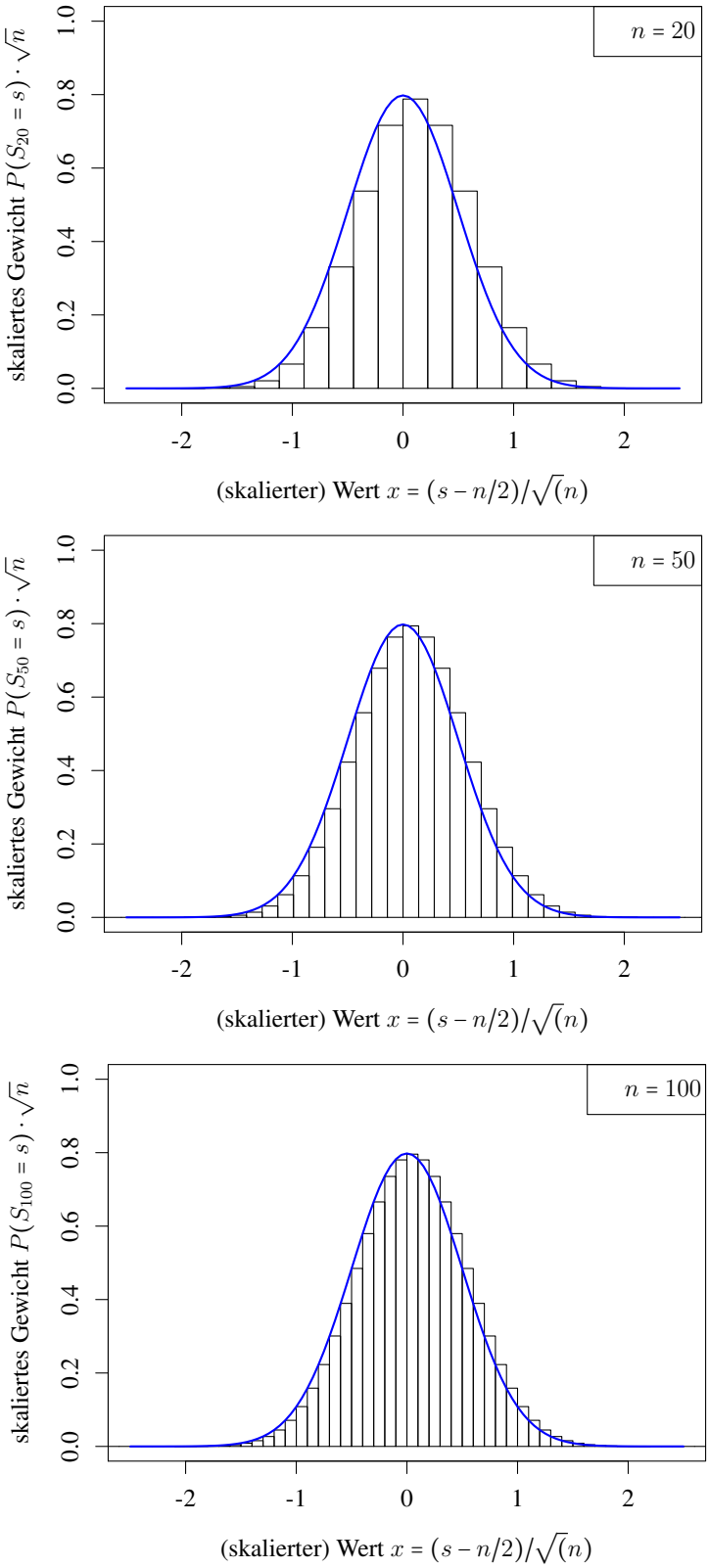

n = 50
$P(S_{50} = s)$
0.00
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
50
s


n = 100
$P(S_{100} = s)$
0.00
0.02
0.04
0.06
0
20
40
60
80
100
s


n = 20
skaliertes Gewicht $P(S_{20} = s) \cdot \sqrt{n}$
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
(skalierter) Wert $x = (s - n/2)/\sqrt{(n)}$


n = 50
skaliertes Gewicht $P(S_{50} = s) \cdot \sqrt{n}$
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
(skalierter) Wert $x = (s - n/2)/\sqrt{(n)}$


n = 100
skaliertes Gewicht $P(S_{100} = s) \cdot \sqrt{n}$
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
(skalierter) Wert $x = (s - n/2)/\sqrt{(n)}$



Definition 1.26. Sei X eine Zufallsvariable mit Werten in einem Intervall $S = [a, b] \cap \mathbb{R} \subset \mathbb{R}$ mit $-\infty \leq a < b \leq \infty$ (im Fall $a > -\infty, b = \infty$ meinen wir $S = [a, \infty)$, etc.) und sei $f : S \to [0, \infty)$

eine integrierbare[7] Funktion mit

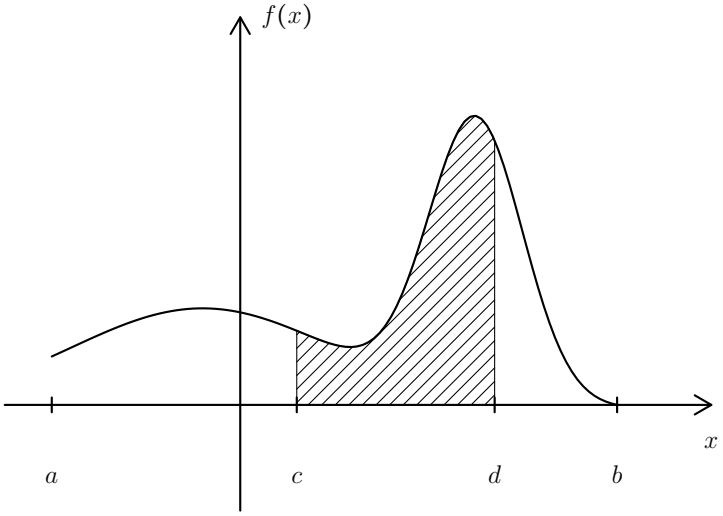
$$\int_a^b f(x)\,dx = 1.$$

X besitzt die *Dichte* (auch: *Wahrscheinlichkeitsdichte*) f, wenn gilt

$$P\big(X \in [c,d]\big) = \int_c^d f(x)\,dx \quad \text{für jedes Teilintervall } [c,d] \subset S. \tag{1.8}$$

Wir notieren oft auch f_X für die Dichte einer ZV X (um den Bezug zu X zu betonen, speziell wenn wir mehrere ZVn zugleich ins Auge fassen).

X hat Dichte f, so ist $P\big(X \in [c,d]\big) = \int_c^d f(x)\,dx$:



Interpretation der Dichte: X ZV mit Dichte f_X, für $x \in \mathbb{R}$ und kleines $\delta > 0$ ist

$$P\big(X \in [x, x+\delta]\big) = \int_x^{x+\delta} f_X(a)\,da \approx \delta f_X(x)$$

(wörtlich zumindest für Stetigkeitsstellen x von f_X), also

$$f_X(x) = \lim_{\delta \downarrow 0} \frac{1}{\delta} P\big(X \in [x, x+\delta]\big)$$

Man formuliert dies gelegentlich auch mit „infinitesimalen Größen“ als

$$P(X \in dx) = f_X(x)\,dx$$

(Dieser suggestive Ausdruck erhält einen Sinn im Sinne der „Standard-Analysis“, wenn man auf beiden Seiten x über ein Intervall $[c,d]$ integriert, dann erhält man (1.8).)

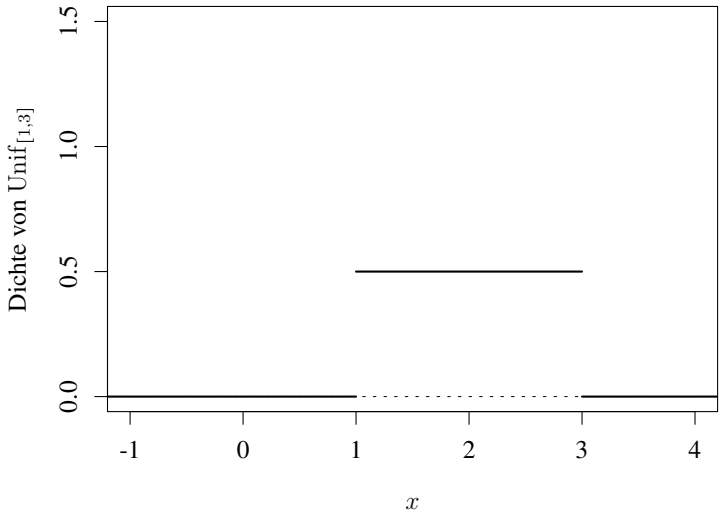
[7] In einem mit den Vorkenntnissen der Hörer verträglichen Sinn: Wir werden nur Beispiele betrachten, in denen f wenigstens stückweise stetig ist, so dass man hier durchaus an das Riemann-Integral (oder auch ganz salopp an die „Fläche unter der Kurve“) denken kann. Für einen Bericht zum (allgemeineren)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

Bemerkung. Für eine ZV X mit Dichte f_X ist es – im Gegensatz zum Fall mit Gewichten – nicht besonders sinnvoll, nach der Wahrscheinlichkeit von Ereignissen $\{X = x\}$ für feste Punkte $x \in \mathbb{R}$ zu fragen, es gilt dann nämlich immer

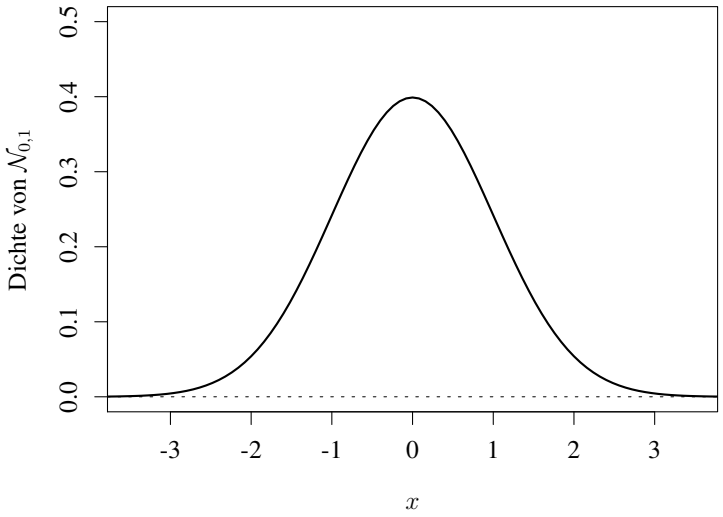
$$P(X = x) = \lim_{\delta \downarrow 0} P\big(X \in [x, x+\delta]\big) = \lim_{\delta \downarrow 0} \int_x^{x+\delta} f_X(a)\, da = \int_x^x f_X(a)\, da = 0.$$

Beispiel 1.27 (Einige „klassische“ eindimensionale Verteilungen mit Dichte)**.**

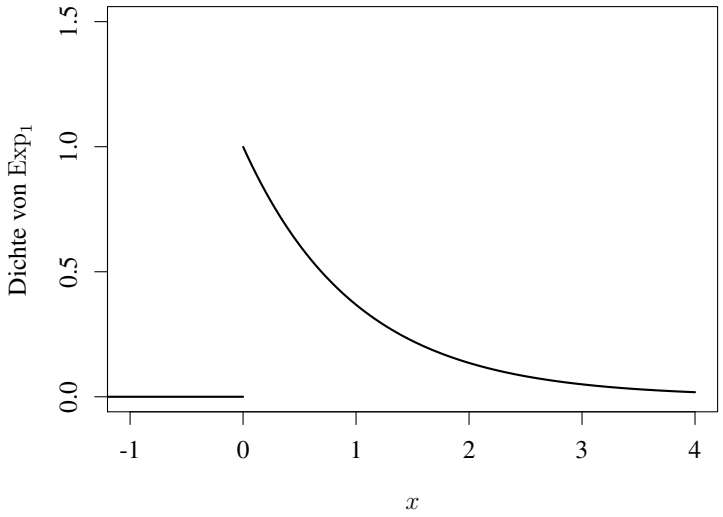
1. (uniforme Verteilung) $a, b \in \mathbb{R}$, $a < b$. $\mathrm{Unif}_{[a,b]}$ mit Dichte $\frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$



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



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



Definition 1.28 (Verteilungsfunktion)**.** Für eine Zufallsvariable X mit Werten in $\mathbb{R}$ (bzw. in einer Teilmenge $S \subset \mathbb{R}$) heißt die Funktion

$$F_X(x) := P(X \leq x), \quad x \in \mathbb{R} \tag{1.9}$$

die *Verteilungsfunktion* von X.

Wenn X mit Wertebereich $S \subset \mathbb{R}$ die Dichte f_X besitzt, so gilt offenbar

$$F_X(x) = \int_{-\infty}^{x} f_X(a)\, da \tag{1.10}$$

(mit Setzung $f_X(a) = 0$ für $a \notin S$, dem Wertebereich von X).

Beispiel 1.27 (Fortsetzung)**.**

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 $\max\left\{\min\left\{\frac{x-a}{b-a}, 1\right\}, 0\right\}$

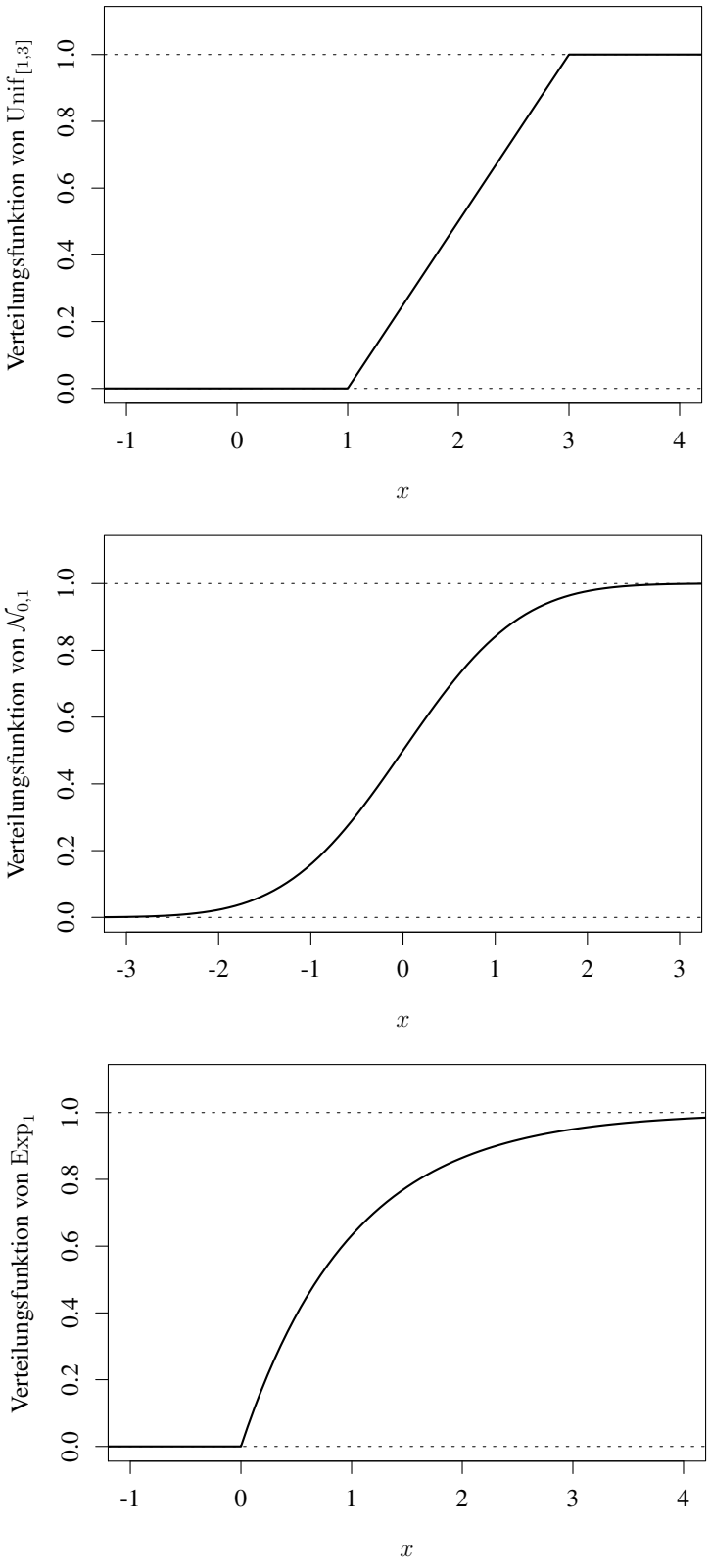
2. (Normalverteilung[en]) $\mu \in \mathbb{R}$, $\sigma > 0$. $\mathcal{N}_{\mu,\sigma^2}$ mit Dichte $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ heißt Normalverteilung mit Mittelwert μ und Varianz σ^2.

   $\mathcal{N}_{0,1}$ 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

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



Bemerkung 1.29. 1. Die Dichte / Verteilungsfunktion von X hängt nur von der Verteilung von X ab:

wenn $Y =^d X$ („Gleichheit in Verteilung“), also $P(X \in B) = P(Y \in B)$ für alle B gilt, so hat Y dieselbe (offenbar).

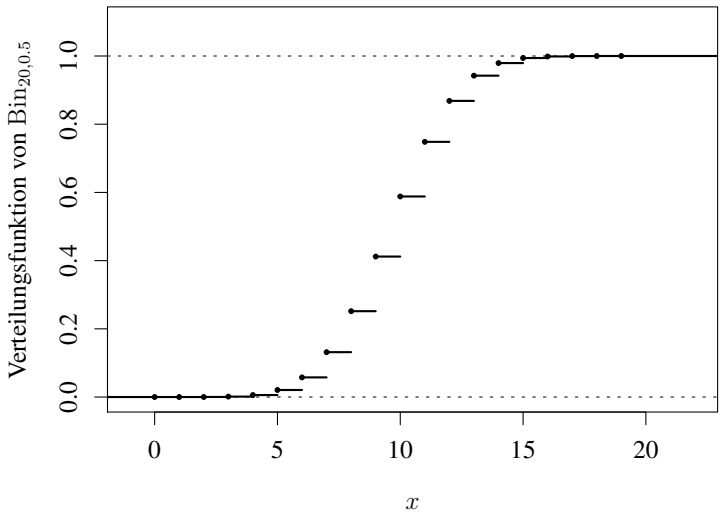
Wir sprechen daher oft auch kurz von der Dichte bzw. Verteilungsfunktion einer Wahrscheinlichkeitsverteilung auf $\mathbb{R}$, ohne die zugehörige ZV explizit zu machen.

2. Wenn X Dichte f_X und Verteilungsfunktion F_X besitzt, so ist
$$\frac{d}{dx} F_X(x) = f_X(x)$$
(zumindest an Stetigkeitspunkten von f_X, leite (1.10) nach x ab)

3. X ZV mit Werten in $S \subset \mathbb{R}$ mit Verteilungsfunktion F_X, $c < d$, so ist
$$P\big(X \in (c,d]\big) = P(X \le d) - P(X \le c) = F(d) - F(c)$$
(und falls $P(X = c) = 0$, z.B. weil X eine Dichte besitzt, so ist natürlich auch $P\big(X \in [c,d]\big) = P(X = c) + P\big(X \in (c,d]\big) = F(d) - F(c)$).

   Für $B = \bigcup_{i=1}^n (c_i, d_i]$ mit $c_1 < d_1 < c_2 < d_2 < \cdots < c_{n-1} < d_n$ ist (mit Eigenschaft (A) aus Def. 1.7)
$$P(X \in B) = \sum_{i=1}^n P\big(X \in (c_i, d_i]\big) = \sum_{i=1}^n \big(F_X(d_i) - F_X(c_i)\big)$$
(und „allgemeine“ Mengen $B \subset \mathbb{R}$ können auf diese Weise approximiert werden). In diesem Sinne „weiß F_X alles“ über die Verteilung von X.

4. (Bezug zum diskreten Fall). Sei X ZV mit (höchstens) abzählbarem Wertebereich $S = \{x_1, x_2, \dots\} \subset \mathbb{R}$ und Gewichten $\rho_X(x_n)$ wie in in Def. 1.11, d.h.
$$P(X \in B) = \sum_{n\,:\,x_n \in B} \rho_X(x_n),$$
dann ergibt sich als Verteilungsfunktion
$$F_X(x) = \sum_{n\,:\,x_n \le x} \rho_X(x_n).$$
(Diese ist stückweise konstant mit (höchstens) abzählbar vielen Sprüngen.)



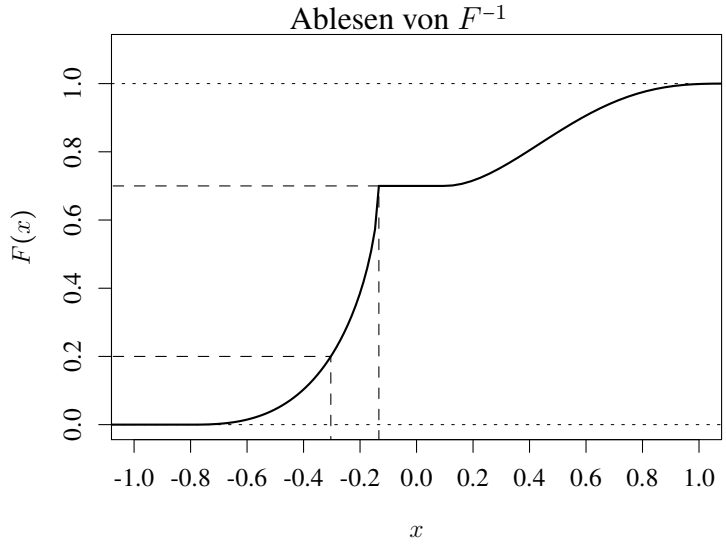
5. Stets ist F_X nicht-fallend und rechtsstetig (wenn X eine Dichte besitzt, so ist F_X stetig) mit
$$\lim_{x\to\infty} F_X(x) = 1, \quad \lim_{x\to-\infty} F_X(x) = 0.$$

6. Umgekehrt gibt es zu jeder Funktion $F : \mathbb{R} \to [0,1]$ mit den Eigenschaften aus Bem. 1.29, 5. eine ZV X mit $F_X = F$.

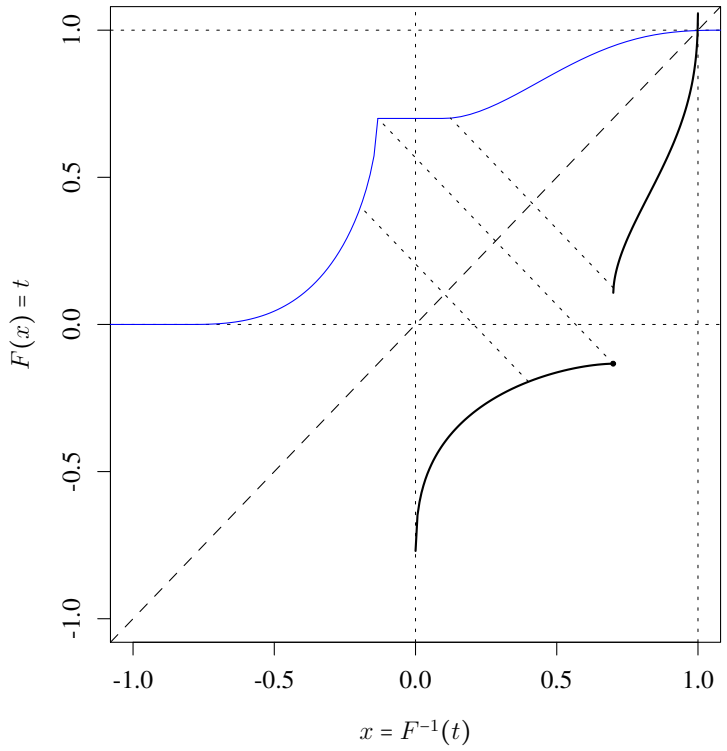
   (Wir kommen darauf zurück, siehe Beob. 1.31 unten.)

Definition 1.30. Die (verallgemeinerte) inverse Funktion von F_X,
$$F_X^{-1}(t) := \inf\{x \in \mathbb{R} : F(x) \geq t\}, \quad t \in [0,1]$$
(mit Setzung $\inf \varnothing = +\infty$) heißt auch die *Quantilfunktion* von X.

(Beachte, dass die so definierte Funktion F_X^{-1} linksstetig ist. Mit dieser Definition ergibt sich für $x \in \mathbb{R}$, $t \in [0,1]$ die Beziehung
$$F_X^{-1}(t) \leq x \iff t \leq F_X(x).$$
In der Literatur gibt es leicht verschiedene Definitionen der „Quantilfunktion“, man prüfe ggfs. jeweils die verwendete Konvention.)



Erinnerung: Inverse via Spiegelung an der Diagonalen



Beobachtung 1.31 (Erzeugung reeller ZVn mit vorgegebener Verteilung)**.** *Sei* $F : \mathbb{R} \to [0,1]$ *eine Verteilungsfunktion ,*

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

die inverse Verteilungsfunktion oder Quantilfunktion (aus Def. 1.30), U *reelle ZV,* $U \sim \mathrm{Unif}_{[0,1]}$*, dann hat*

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

die Verteilungsfunktion $F_X = F$*.*

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

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

□

Beispiel 1.32. 1. 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 $-\frac{1}{\theta}\log(1-U) \sim \mathrm{Exp}_\theta$ (und natürlich ebenso $-\frac{1}{\theta}\log(U)$)

2. $p(k), k \in \mathbb{N}_0$ Wahrscheinlichkeitsgewichte, $F(x) = \sum_{k:k\leq x} p(k)$ zugehörige Verteilungsfunktion (vgl. Bem. 1.29, 4.), so hat

$$X := \min\{n \in \mathbb{N}_0 : \sum_{k=0}^{n} p(k) \geq U\}$$

die Gewichte $(p(k))_{k\in\mathbb{N}_0}$.

(Dies ist etwa eine Möglichkeit, eine Poisson-verteilte ZV zu simulieren.)

Bericht: Verteilungen in R. R verwendet folgende Namenskonvention: Wenn `name` für eine Verteilung steht, so ist

- `dname` die Dichte- bzw. Gewichtsfunktion von `name`
- `pname` die Verteilungsfunktion von `name` ("`p`" steht für "probability distribution function")
- `qname` die Quantilfunktion von `name`
- `rname` simuliert gemäß `name` ("`r`" steht für "random")

Beispiele:

- Uniforme Verteilung $\mathrm{Unif}_{[a,b]}$: `[d|p|q|r]unif(...,min=`a`,max=`b`)`
- Normalverteilung $\mathcal{N}_{\mu,\sigma^2}$: `[d|p|q|r]norm(...,mean=`$\mu$`,sd=`$\sigma$`)` (beachte: R parametrisiert mit σ, nicht mit σ^2)
- Exponentialverteilung Exp_λ: `[d|p|q|r]exp(...,rate=`λ`)`
- Poissonverteilung Poi_λ: `[d|p|q|r]pois(...,lambda=`λ`)`
- Binomialverteilung $\mathrm{Bin}_{n,p}$: `[d|p|q|r]binom(...,size=`n`,prob=`p`)`

(Siehe auch die Online-Hilfe in R.)

Transformation von Dichten

Beobachtung 1.33. X habe Dichte f_X, sei $a > 0$, $b \in \mathbb{R}$, so hat $Y := aX + b$ die Dichte

$$f_Y(y) = \frac{1}{a} f_X\big((y-b)/a\big)$$

denn

$$P(Y \le y) = P\big(X \le \tfrac{y-b}{a}\big) = \int_{-\infty}^{(y-b)/a} f_X(x)\,dx = \int_{-\infty}^{y} f_X\big(\tfrac{z-b}{a}\big)\,\frac{1}{a}\,dz$$

wir substituieren $x = (z-b)/a$ ($\Leftrightarrow z = ax + b$), $\frac{dx}{dz} = 1/a$.

Beachte: $Y \in [y, y+\delta] \iff X \in \big[\tfrac{y-b}{a}, \tfrac{y-b}{a} + \tfrac{\delta}{a}\big]$.

Beispiel 1.34. 1. $X \sim \mathcal{N}_{0,1}$, $\mu \in \mathbb{R}$, $\sigma > 0$, so ist $Y := \sigma X + \mu \sim \mathcal{N}_{\mu,\sigma^2}$

Insbesondere gilt $\mathcal{N}_{\mu,\sigma^2}((-\infty, x]) = \Phi((x-\mu)/\sigma)$

2. $X \sim \mathrm{Exp}_1$, $a > 0$, so hat $Y := aX$ die Dichte $\frac{1}{a} e^{-x/a}$ (d.h. $Y \sim \mathrm{Exp}_{1/a}$)

Proposition 1.35 (Allgemeine 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 $P(X \in I) = 1$, $J \subset I$, $\varphi : I \to J$ stetig differenzierbar, bijektiv.*

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

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

Beachte: Wenn φ nicht bijektiv ist, so besitzt $\varphi(X)$ i.A. keine Dichte. Sei z.B. $X \sim \mathcal{N}_{0,1}$, $\varphi(x) = \mathbf{1}_{(0,\infty)}(x)$, so ist $\varphi(X) \sim \mathrm{Ber}_{1/2}$, d.h. $\varphi(X)$ ist hier diskret.

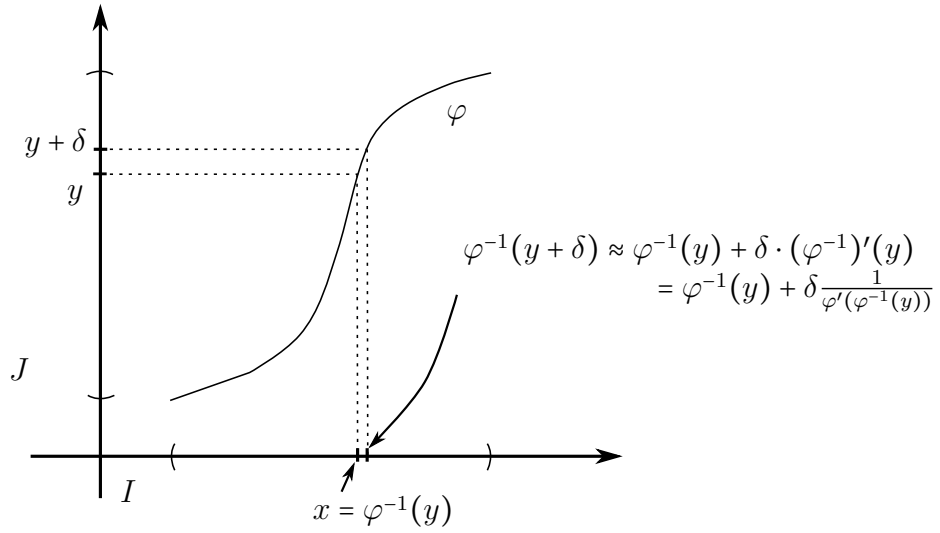
Beweis. φ muss offenbar strikt wachsend oder strikt fallend sein, wir betrachten den wachsenden Fall.

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

Sei $z \in J$:

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

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 1.36. 1. $U \sim \mathrm{Unif}_{[0,1]}$, so hat $X := -\log(U)$ Dichte e^{-x} für $x \ge 0$ (wie wir in Bsp. 1.32 gesehen haben)

2. $U \sim \mathrm{Unif}_{[0,1]}$, $n \in \mathbb{N}$, so hat $Y := U^n$ Dichte $f_Y(y) = n^{-1}y^{1/n-1}$ (für $0 \le y \le 1$)

Dichten im mehrdimensionalen Fall

Definition 1.37. Sei $f : \mathbb{R}^d \to \mathbb{R}_+$ eine (geeignet) integrierbare[8] Funktion mit

$$\int_{\mathbb{R}^d} f(x)\,dx = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\dots,x_d)\,dx_1\cdots dx_d = 1.$$

Eine Zufallsvariable X mit Werten in $\mathbb{R}^d$ besitzt die *Dichte* f, wenn für (geeignete[9]) Teilmengen $A \subset \mathbb{R}^d$ gilt

$$P(X \in A) = \int_A f(x)\,dx := \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \mathbf{1}_A(x_1,\dots,x_d) f(x_1,\dots,x_d)\,dx_1\cdots dx_d.$$

[8] Für die Zwecke dieser Vorlesung genügt es, an ein d-fach iteriertes Riemann-Integral zu denken.

[9] Damit das multivariate Integral wohldefiniert ist, muss A strenggenommen gewisse sogenannte Messbarkeitseigenschaften erfüllen. Dies wird für die von uns betrachteten Beispiele, z.B. A ein verallgemeinerter Quader oder A eine Menge mit stückweise differenzierbarem Rand, immer erfüllt sein.

Analog zum 1-dimensionalen Fall besitzt die Dichte f_X einer d-dimensionalen ZV X die Interpretation

$$P\big(X \in [x_1, x_1+\delta_1] \times [x_2, x_2+\delta_2] \times \cdots \times [x_d+\delta_d]\big) \approx \delta_1\delta_2\cdots\delta_d \cdot f_X\big((x_1,\ldots,x_d)\big)$$

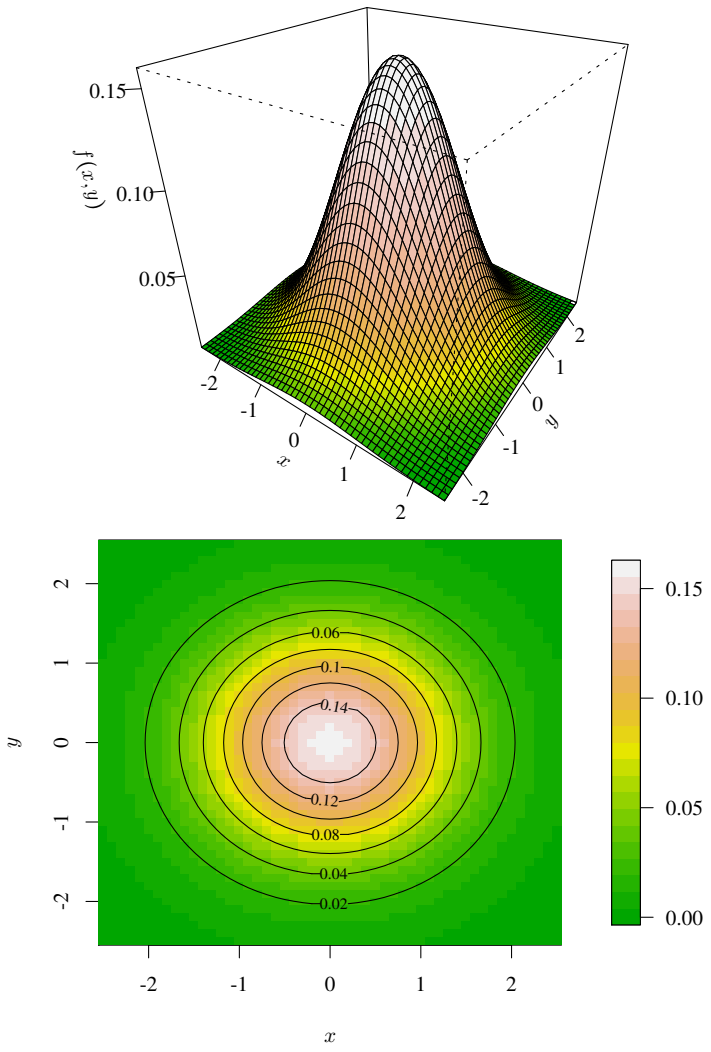
für $(x_1,\ldots,x_d) \in \mathbb{R}^d$ und $0 < \delta_1, \delta_2, \ldots, \delta_d \ll 1$.

Beispiel 1.38. 1. Uniforme oder Laplace-Verteilung auf einem beschränkten Gebiet $S \subset \mathbb{R}^d$: X mit Dichte $f_X(x) = \frac{1}{\mathrm{vol}(S)}\mathbf{1}_S(x)$ erfüllt für $A \subset S$ (geeignet[10])

$$P(X \in A) = \int_A \frac{1}{\mathrm{vol}(S)}\mathbf{1}_S(x)\,dx = \frac{\int_A 1\,dx}{\int_S 1\,dx} = \frac{\mathrm{vol}(A)}{\mathrm{vol}(S)}.$$

(Z.B. der zufällig gewählte Punkt Z aus Kapitel 0 war uniform auf $S = [0,1]^2$ verteilt.)

2. Die 2-dimensionale Standard-Normalverteilung hat Dichte $f(x,y) = \dfrac{1}{2\pi}\exp\big(-\frac{1}{2}(x^2+y^2)\big)$



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

3. Allgemeiner: Die d-dimensionale Standard-Normalverteilung hat Dichte

$$f(x_1,\dots,x_d) = \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{1}{2}(x_1^2+\cdots+x_d^2)\Big)$$

Beobachtung 1.39 (Marginaldichten)**.** *Die Zufallsvariable $X = (X_1, X_2)$ mit Werten in ($S \subset$) $\mathbb{R}^2$ habe (gemeinsame) Dichte $f_X(x_1, x_2)$, so hat X_1 die Dichte $f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_X(x_1, x_2)\, dx_2$ ($x_1 \in \mathbb{R}$) und analog hat X_2 die Dichte $f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_X(x_1, x_2)\, dx_1$ ($x_2 \in \mathbb{R}$). f_{X_1} und f_{X_2} heißen die* Marginal- *oder* Randdichten *von f_X.*

Es ist nämlich

$$\begin{aligned} P\big(X_1 \in [a,b]\big) = P\big(X_1 \in [a,b], X_2 \in \mathbb{R}\big) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_{[a,b]}(x_1)\mathbf{1}_{\mathbb{R}}(x_2) f(x_1,x_2)\, dx_1 dx_2 \\ &= \int_a^b \Big(\int_{-\infty}^{\infty} f(x_1,x_2)\, dx_2\Big) dx_1. \end{aligned}$$

Allgemein: Für $X = (X_1,\dots,X_d)$ mit Werten in $\mathbb{R}^n$ und Dichte f_X ist

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_X(x_1,\dots,x_n)\, dx_1\cdots dx_{i-1} dx_{i+1}\cdots dx_n$$

die i-te Marginaldichte.

Das kontinuierliche Analogon zu Prop. 1.14 lautet:

Bericht 1.40 (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. $\int_{-\infty}^{\infty} f_i(x)\, dx = 1$), 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. $P(X_i \in B) = \int_B f_i(x)\, dx$ und $P\big(X_1 \in B_1,\dots,X_n \in B_n\big) = \prod_{i=1}^n P(X_i \in B_i)$ für $B_1,\dots,B_n \subset \mathbb{R}$).

2. Die ZV $X = (X_1,\dots,X_n)$ mit Werten in $\mathbb{R}^n$ hat Dichte

   $$f(x) = f_1(x_1)\cdot f_2(x_2)\cdot\cdots\cdot f_n(x_n), \quad x = (x_1,\dots,x_n) \in \mathbb{R}^n,$$

   d.h. die gemeinsame Dichte hat Produktgestalt.

   Vergleicht man dies mit Beob. 1.39, so folgt: In diesem Fall ist die gemeinsame Dichtefunktion das Produkt der Marginaldichten.

Beweisidee. „Naiv“ rechnen wir

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

□

Beispiel 1.41. 1. Wähle $A = [a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$ in Bsp. 1.38, 1., d.h. $X = (X_1, X_2)$ ist uniform verteilt auf einem (achsenparallelen) Rechteck (mit Fläche $\mathrm{vol}(A) = (b_1 - a_1)(b_2 - a_2)$). X hat Dichte

$$f_X(x_1, x_2) = \frac{1}{(b_1 - a_1)(b_2 - a_2)} \mathbf{1}_{[a_1,b_1]\times[a_2,b_2]}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

mit Marginaldichten $f_{X_i}(x_i) = \frac{1}{(b_i - a_i)} \mathbf{1}_{[a_i,b_i]}(x_i)$ $(i = 1, 2)$, d.h. die Koordinaten X_1 und X_2 sind unabhängig (und jeweils uniform auf $[a_i, b_i]$ verteilt).

2. Wähle $A = \left\{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\right\}$ in Bsp. 1.38, d.h. $X = (X_1, X_2)$ ist uniform verteilt auf dem Einheitskreis (mit Fläche $\mathrm{vol}(A) = \pi$) mit Dichte

$$f_X(x_1, x_2) = \frac{1}{\pi} \mathbf{1}_{[0,1]}(x_1^2 + x_2^2).$$

Die Marginaldichte ist

$$\begin{aligned} f_{X_1}(x_1) &= \int_{-\infty}^{\infty} \frac{1}{\pi} \mathbf{1}_{[0,1]}(x_1^2 + x_2^2)\, dx_2 = \frac{1}{\pi} \mathbf{1}_{[-1,1]}(x_1) \int_{-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}} 1\, dx_2 \\ &= \mathbf{1}_{[-1,1]}(x_1) \frac{2}{\pi} \sqrt{1 - x_1^2} \end{aligned}$$

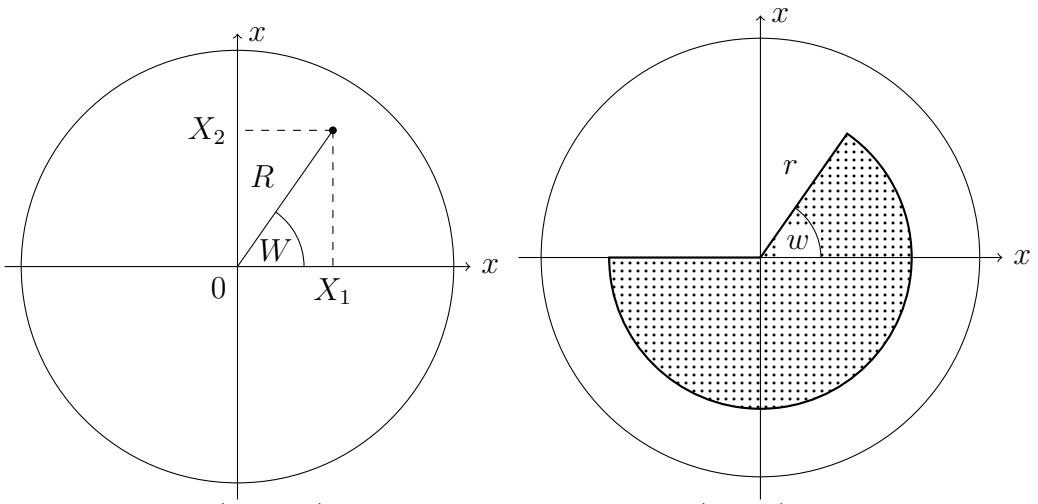
(und analog bzw. offensichtlich aus Symmetrie ist $f_{X_2} = f_{X_1}$).

Insbesondere sind (im Gegensatz zu 1.) X_1 und X_2 sind (natürlich) *nicht* unabhängig, denn $f_X(x_1, x_2) \neq f_{X_1}(x_1) f_{X_2}(x_2)$.

3. Betrachten wir $X = (X_1, X_2)$ aus 2. in Polarkoordinaten: Sei R der Radius, W der Winkel von X (in Polarkoordinaten), also

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

bzw. $X_1 = R\cos(W), X_2 = R\sin(W)$ (siehe auch die Skizze unten).



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

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

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

4. $X = (X_1, \dots, X_d)$ d-dimensional Standard-normalverteilt, so sind $X_1, \dots, X_d$ unabhängig und jeweils $\sim \mathcal{N}_{0,1}$ (d.h. die X_i sind [1-dimensional] Standard-normalverteilt), denn

$$\frac{1}{(2\pi)^{d/2}}\exp\big(-\frac{1}{2}(x_1^2 + \cdots + x_d^2)\big) = \prod_{i=1}^d \Big(\frac{1}{(2\pi)^{1/2}}e^{-x_i^2/2}\Big)$$

Die Beobachtung zur Rotationssymmetrie aus Bsp. 1.41, 3. verallgemeinert sich folgendermaßen:

Beobachtung 1.42. *$X = (X_1, X_2)^T$ habe eine rotationssymmetrische Dichte f_X, d.h. $f_X(x_1, x_2)$ hängt nur von $r = \sqrt{x_1^2 + x_2^2}$ ab (also $f_X(x_1, x_2) = g(r)$ für eine gewisse Funktion $g \ge 0$), $X' = (X_1', X_2')^T$ entstehe aus X durch Drehung (um Winkel α um den Ursprung), d.h.*

$$\begin{pmatrix} X_1' \\ X_2' \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \cdot \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \cos(\alpha)X_1 - \sin(\alpha)X_2 \\ \sin(\alpha)X_1 + \cos(\alpha)X_2 \end{pmatrix}.$$

Dann hat X' dieselbe Dichte (und somit dieselbe Verteilung) wie X. (Dies ist anschaulich sehr plausibel, man kann es z.B. mit Bericht 1.43 beweisen.)

Sei (R, W) die Polarkoordinatendarstellung von X wie in Bsp. 1.41, 3. (insbes. $R = \sqrt{X_1^2 + X_2^2}$), so sind R und W unabhängig, W ist uniform verteilt auf $[-\pi, \pi)$ und R hat Dichte $2\pi r g(r)\mathbf{1}_{[0,\infty)}(r)$:

$$P(R \le u) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_{[0,u]}\big(\sqrt{x_1^2 + x_2^2}\big)g\big(\sqrt{x_1^2 + x_2^2}\big)\,dx_1 dx_2$$
$$= \int_{-\pi}^{\pi}\int_0^u g(r)\,r\,dr\,dw = \int_0^u 2\pi r g(r)\,dr$$

(wir verwenden hier etwas salopp die „Polarkoordinatenform des Flächenelements“ $dx_1 dx_2 = r\,dr\,dw$, man kann dies wiederum mit Bericht 1.43 beweisen.)

Speziell für X_1, X_2 unabhängig, $\sim \mathcal{N}_{0,1}$, somit $X = (X_1, X_2)$ 2-dim. Standard-normalverteilt, ergibt sich (für $u > 0$):

$$P(R^2 \le u) = P(R \le \sqrt{u}) = \int_0^{\sqrt{u}} 2\pi r \tfrac{1}{2\pi}e^{-r^2/2}\,dr = \big[e^{-r^2/2}\big]_{r=0}^{r=\sqrt{u}} = 1 - e^{-u/2},$$

d.h. $R^2 \sim \mathrm{Exp}_{1/2}$. (Dies ist der Hintergrund der Box-Muller-Methode[11] *zur Simulation normalverteilter ZVn, vgl. auch Übungsblatt 5.)*

Bericht 1.43 (Allgemeine 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

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

(wobei $\varphi(x) = \big(\varphi_1(x), \ldots, \varphi_d(x)\big)^T$, d.h. φ_i ist die i-te Koordinatenfunktion von φ), 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“

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

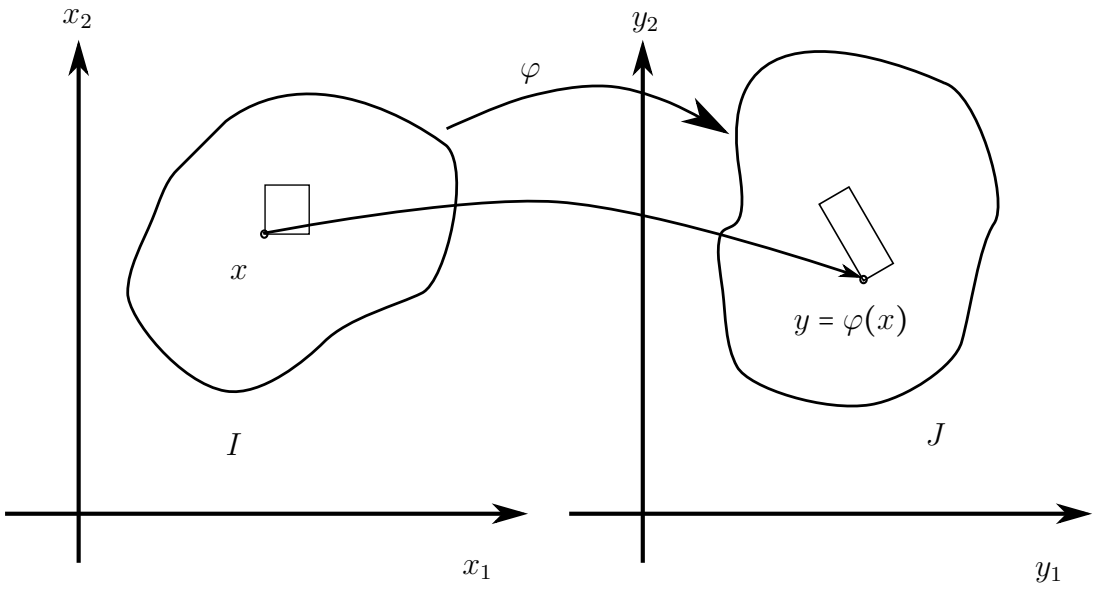
(plus Terme, die $O(\|x' - x\|^2)$ sind), also:

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.

[11] George E.P. Box, und Mervin E. Muller, A Note on the Generation of Random Normal Deviates. Ann. Math. Stat. 29, 610-611, 1958.



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

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

Nur der Vollständigkeit halber, wir werden dies im Verlauf der Vorlesung nicht benötigen:

Bericht 1.44. Analog betrachtet zu Def. 1.28 betrachtet man im Fall $d > 1$ für $(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ gelegentlich die d-dimensionale Verteilungsfunktion

$$F_X(x_1, x_2, \ldots, x_d) := P\big(X \in (-\infty, x_1] \times (-\infty, x_2] \times \ldots (-\infty, x_d]\big),$$

die allerdings etwas weniger „handlich“ ist als im 1-dimensionalen Fall.

Analog zu Bem. 1.29, 3. „weiß“ die d-dimensionale Verteilungsfunktion von X „alles“ über die Verteilung von X. Die d-dimensionale Verallgemeinerung der Eigenschaften aus Bem. 1.29, 5. besteht in folgenden Bedingungen:

1. F_X rechtsstetig, d.h. $x_n \searrow x$ (koordinatenweise) $\Rightarrow F_X(x_n) \to F(x)$
2. $F_X(x_n) \to 1$ wenn $x_n \to (+\infty, \ldots, +\infty)$
3. $F_X(x_n) \to 0$ wenn $\min_{i=1,\ldots,d} x_{n,i} \to -\infty$
4. Für $x = (x_1, \ldots, x_d) < y = (y_1, \ldots, y_d)$ (koordinatenweise), parametrisiere die Ecken des d-Quaders $(x, y]$ mit $\{1, 2\}^d$ via $\big\{u = (z_1^{(i_1)}, \ldots, z_d^{(i_d)}) : i_1, \ldots, i_d \in \{1, 2\}\big\}$ wo $z_j^{(1)} = x_j$, $z_j^{(2)} = y_j$, es muss gelten

$$\sum_{u \text{ Ecken}} (-1)^{\#\{1 \le m \le d : i_m = 1\}} F_X(u) \Big(= \mu_F((x, y]) \Big) \ge 0$$

1.2 Bedingte Wahrscheinlichkeiten und mehrstufige Zufallsexperimente

Beispiel 1.45. Wir ziehen zwei Kugeln ohne Zurücklegen aus einer Urne mit $s > 0$ schwarzen und $w > 0$ weißen Kugeln. Wie in Bsp. 1.15, 2. stellen wir uns die Kugeln nummeriert vor, Nr. $1, \dots, w$ seien weiß, $w+1, \dots, w+s$ schwarz. Sei X_i die Nr. der Kugel im i-ten Zug (= $1, 2$).

Also: $X = (X_1, X_2)$ ist uniform verteilt auf $S = \big\{(i,j) : 1 \le i,j \le w+s, i \ne j\big\}$ (vgl. Bsp. 1.15, 2.).

Betrachte die Ereignisse

$$A = \big\{\text{erste Kugel ist weiß}\big\} = \{X_1 \le w\},$$
$$B = \big\{\text{zweite Kugel ist weiß}\big\} = \{X_2 \le w\} \Big(= \big\{X \in \{(i,j) \in S : j \le w\}\big\}\Big)$$

Ohne weitere Informationen ist

$$P(B) = P\big(X \in \{(i,j) \in S : j \le w\}\big) = \frac{\#\{(i,j) \in S : j \le w\}}{\#S} = \frac{w(w+s-1)}{(w+s)(w+s-1)} = \frac{w}{w+s}.$$

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 1.46. *Sei $\mathscr{E}$ die Menge der Ereignisse in einem gewissen Zufallsexperiment, $A \in \mathscr{E}$ mit $P(A) > 0$. Für $B \in \mathscr{E}$*

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

heißt bedingte Wahrscheinlichkeit von B, gegeben A.

$P(\cdot \mid A)$ ist ein Wahrscheinlichkeitsmaß auf $\mathscr{E}$, man prüft leicht per Inspektion, dass die Eigenschaften aus Definition 1.7, Normierung und σ-Additivität, erfüllt sind.

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

In Beispiel 1.45 ist $P(A) = \frac{w}{w+s}$,

$$P(A \cap B) = P(X_1 \le w, X_2 \le w) = \frac{w(w-1)}{(w+s)(w+s-1)},$$

also ergibt sich tatsächlich $P(B \mid A) = \frac{w-1}{w+s-1}$.

Bemerkung 1.47 („Natürlichkeit von Definition 1.46")**.** 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

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

Beweis. Für $C \in \mathscr{E}$ 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 1.48. $P(B \mid A) \neq P(B)$ kann nicht notwendigerweise als „Kausalität“ (im Sinne von „A beeinflusst, ob B eintritt“) interpretiert werden:
In Beispiel 1.45 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 1.49. *Sei $\mathscr{E}$ die Menge der Ereignisse in einem gewissen Zufallsexperiment, I eine (höchstens abzählbare) Indexmenge, seien $B_i \in \mathscr{E}$ 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 $A \in \mathscr{E}$ gilt*

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

2. *(Formel von Bayes[12]) Für $A \in \mathscr{E}$ mit $P(A) > 0$ und jedes $k \in I$ gilt*

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

Insbesondere gilt für Ereignisse A, B mit $P(A) > 0$

$$P(B \mid A) = \frac{P(B) P(A \mid B)}{P(B) P(A \mid B) + P(B^c) P(A \mid B^c)}$$

(verwende Zerlegung $B \cup B^c = E_s$)

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. □

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

Beispiel 1.50 (Verrauschter Übertragungskanal)**.** Ein (einzelnes) Bit X (aus $\{0,1\}$, es gelte $P(X=1)=a\in(0,1)$) wird über einen fehleranfälligen Kanal gesendet, der jede 1 mit W'keit f_1 und jede 0 mit W'keit f_0 flippt, sei Y das empfangene Bit.

Dann ist

$$\begin{aligned} P(Y=0\mid X=1)&=f_1, \quad P(Y=1\mid X=1)=1-f_1,\\ P(Y=1\mid X=0)&=f_0, \quad P(Y=0\mid X=0)=1-f_0, \end{aligned}$$

also

$$\begin{aligned} P(X=1\mid Y=1)&=\frac{P(X=1)P(Y=1\mid X=1)}{P(X=1)P(Y=1\mid X=1)+P(X=0)P(Y=1\mid X=0)}\\ &=\frac{a(1-f_1)}{a(1-f_1)+(1-a)f_0},\\ P(X=0\mid Y=0)&=\frac{P(X=0)P(Y=0\mid X=0)}{P(X=0)P(Y=0\mid X=0)+P(X=1)P(Y=0\mid X=1)}\\ &=\frac{(1-a)(1-f_0)}{(1-a)(1-f_0)+af_1} \end{aligned}$$

Z.B. für die konkreten Werte $a=0.3$, $f_1=0.05$, $f_0=0.1$ ergibt sich $P(Y=1)=0.355$, $P(X=1\mid Y=1)\approxeq 0.803$, $P(X=0\mid Y=0)\approxeq 0.977$.

Beobachtung 1.51 (Multiplikationsformel)**.** *$A_1,A_2,\dots,A_n$ Ereignisse (aus der Menge $\mathscr{E}$ der Ereignisse in einem gewissen Zufallsexperiment) 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)

Oft betrachtet man 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.

Dann kann man die gemeinsame Verteilung (zumindest im diskreten Fall, d.h. die gemeinsamen Gewichte) des Vektors $(X_1,X_2,\dots,X_n)$ mittels der Multiplikationsformel (Beob. 1.51, lese dort $A_i=\{X_i=x_i\}$) bestimmen („Pfadregel"):

$$\begin{aligned} &P(X_1=x_1,X_2=X_2,\dots,X_n=x_n)\\ &=P(X_1=x_1)\cdot P(X_2=x_2\mid X_1=x_1)\cdot P(X_3=x_3\mid X_1=x_1,X_2=x_2)\\ &\qquad\cdot\dots\cdot P(X_n=x_n\mid X_1=x_1,X_2=x_2,\dots,X_{n-1}=x_{n-1}) \end{aligned}\tag{1.11}$$

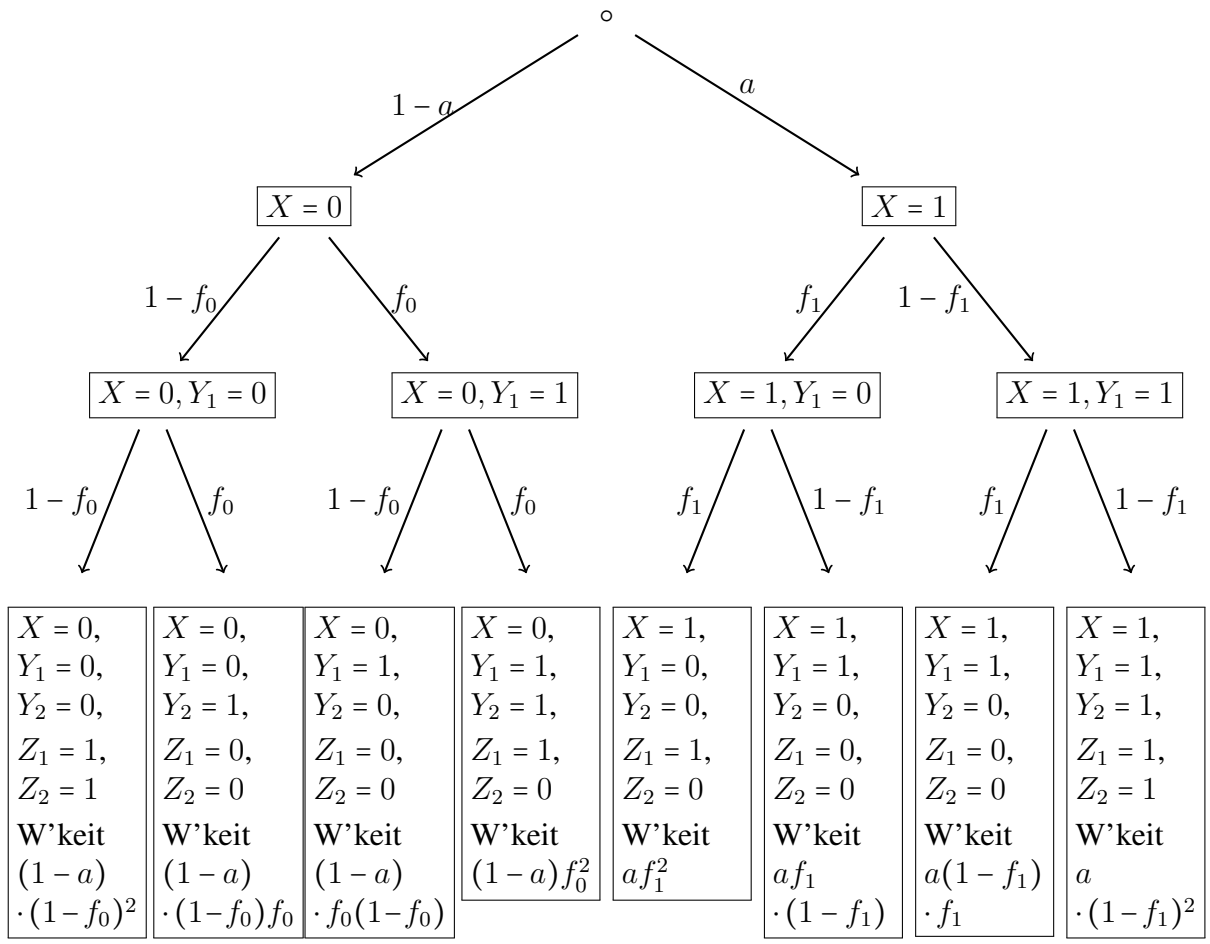
Man stellt Rechnungen, die die verschiedenen Fälle in dieser Weise aufzählen, oft mittels eines Baumdiagramms dar, wie in dem folgenden Beispiel.

Beispiel 1.52. Nehmen wir an, in der Situation von Bsp. 1.50 wird das zufällige Bit X sicherheitshalber zweimal gesendet (wobei jedesmal unabhängig mit den genannten W'keiten ein Übertragungsfehler auftritt), seien Y_1 und Y_2 die beiden empfangenen Bits, $Z_1 = I_{\{Y_1=Y_2\}}$, $Z_2 = I_{\{Y_1=Y_2=X\}}$ (beachte, dass der Empfänger Z_1 beobachten kann, nicht aber Z_2). Dann ist wegen $\{Z_2 = 1\} \subset \{Z_1 = 1\}$

$$P(Z_2 = 1 \mid Z_1) = \frac{P(Z_2 = 1)}{P(Z_1 = 1)}$$

(dies ist die Wahrscheinlichkeit, mit der ein Empfänger, der den gesendeten Bits „vertraut", sofern er zweimal dasselbe empfangen hat, „richtig liegt").

Wir fassen die möglichen Ausgänge von (X, Y_1, Y_2) (und die sich daraus ergebenden Werte von Z_1, Z_2) in einem Baumdiagramm zusammen:



Wir sehen:

$$P(Z_1 = 1) = (1-a)(1-f_0)^2 + (1-a)f_0^2 + af_1^2 + a(1-f_1)^2,$$
$$P(Z_2 = 1) = (1-a)(1-f_0)^2 + a(1-f_1)^2,$$
$$P(Z_2 = 1 \mid Z_1) = \frac{(1-a)(1-f_0)^2 + a(1-f_1)^2}{(1-a)(1-f_0)^2 + (1-a)f_0^2 + af_1^2 + a(1-f_1)^2}.$$

Für die konkreten Zahlenwerte aus Beispiel 1.50 ($a = 0.3, f_1 = 0.05, f_0 = 0.1$) ergibt sich $P(Z_2 = 1 \mid Z_1) \approxeq 0.991$.

1.2.1 Nochmal zur Unabhängigkeit

Erinnerung. Zufallsvariablen $X_1, \dots, X_n$ (X_i habe Wertebereich S_i) heißen (stochastisch) *unabhängig*, wenn für alle Ereignisse $\{X_i \in B_i\}$ gilt

$$P\big(X_1 \in B_1, X_2 \in B_2, \dots, X_n \in B_n\big) = \prod_{i=1}^{n} P(X_i \in B_i) \tag{1.12}$$

(vgl. Def. 1.13).

Dann ist die Pfadformel (vgl. Beob. 1.51) besonders „angenehm".

Definition 1.53. Ereignisse $A_1, \dots, A_n$ heißen *unabhängig*, wenn dies für ihre Indikatorvariablen $I_{A_1}, \dots, I_{A_n}$ gilt.

Speziell: A und B unabhängige Ereignisse (mit $P(B) > 0$), so ist $P(A \cap B) = P(A)P(B)$ und somit $P(A \mid B) = P(A)$.

Bemerkung 1.54. Sind ZVn $X_1, \dots, X_n$ unabhängig, so auch

1. jede Teilfamilie $X_{i_1}, \dots, X_{i_k}$ (für $1 \le i_1 < \cdots < i_k \le n$) (wähle $B_i = S_i$ in (1.7) für $i \notin \{i_1, \dots, i_k\}$) ;

2. $f_1(X_1), f_2(X_2), \dots, f_n(X_n)$ für Funktionen $f_i : S_i \to S_i'$ (beachte $\{f_i(X_i) \in B_i'\} = \{X_i \in f_i^{-1}(B_i')\}$ in (1.7), vgl. Bsp. 1.4, 2.)

3. In Def. 1.53 genügt es i.A. nicht, jeweils nur Paare auf Unabhängigkeit zu prüfen:

   Beispiel: Seien X_1, X_2, X_3 unabhängige faire Münzwürfe $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$, $Y_1 = I_{\{X_1 = X_2\}}, Y_2 = I_{\{X_1 = X_3\}}, Y_3 = I_{\{X_2 = X_3\}}$. Dann sind jeweils Y_1 und Y_2, Y_1 und Y_3, Y_2 und Y_3 unabhängig, aber Y_1, Y_2, Y_3 zusammen *nicht*.

   (Es ist $P(Y_i = 1) = \frac{1}{2}$, z.B. $P(Y_1 = 1, Y_2 = 1) = P\big(\{X_1 = X_2 = X_3 = 1\} \cup \{X_1 = X_2 = X_3 = 0\}\big) = (1/2)^3 + (1/2)^3 = 1/4 = P(Y_1 = 1)P(Y_2 = 1)$, $P(Y_1 = 1, Y_2 = 0) = P\big(\{X_1 = X_2 = 1, X_3 = 0\} \cup \{X_1 = X_2 = 0, X_3 = 1\}\big) = (1/2)^3 + (1/2)^3 = 1/4 = P(Y_1 = 1)P(Y_2 = 0)$, etc.)

   Man sagt dazu auch: Y_1, Y_2, Y_3 sind *paarweise unabhängig*, aber eben nicht unabhängig.

1.2.2 Faltung

Definition 1.55. X und Y unabhängige reellwertige ZVn, $X \sim \mu$, $Y \sim \nu$ (in einem gewissen Zufallsexperiment $\mathscr{X}$). Die Verteilung von $X + Y$ heißt die *Faltung* von μ und ν, geschrieben $\mu * \nu$:

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

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

Beobachtung 1.56 (Diskreter Fall)**.** Falls $\mu(\mathbb{Z}) = \nu(\mathbb{Z}) = 1$ (d.h. X und Y haben Werte in $\mathbb{Z}$), so ist

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

(Im allg. diskreten Fall $P(X \in \{x_i, i \in \mathbb{N}\}, Y \in \{y_j, j \in \mathbb{N}]\}) = 1$ muss man die „Doppelsumme" betrachten: $P(X + Y = z) = \sum_{i,j\,:\,x_i+y_j=z} P(X = x_i)P(Y = y_j)$.)

Beispiel 1.57. 1. W_1, W_2 unabhängige 6-er Würfelwürfe, dann ist

$$S := W_1 + W_2 \sim \mathrm{Unif}_{\{1,2,\dots,6\}} * \mathrm{Unif}_{\{1,2,\dots,6\}}$$

mit

$$\begin{aligned} P(S = k) &= \sum_{m=\max\{k-6,1\}}^{\min\{k-1,6\}} P(W_1 = m)P(W_2 = k - m) \\ &= \frac{1}{36}\big(\min\{k - 1, 6\} - \max\{k - 6, 1\} + 1\big) = \frac{6 - |7 - k|}{36} \end{aligned}$$

für $k \in \{2, 3, \dots, 12\}$

2. X, Y u.a., $\sim \mathrm{Ber}_p$, so ist $X + Y \sim \mathrm{Bin}_{2,p}$, d.h. $\mathrm{Ber}_p * \mathrm{Ber}_p = \mathrm{Bin}_{2,p}$.

3. (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*.

(Schreibe $S_1 := X_1 + \cdots + X_{n_1} \sim \mathrm{Bin}_{n_1,p}$, $S_2 := X_{n_1+1} + X_{n_1+2} + \cdots + X_{n_1+n_2} \sim \mathrm{Bin}_{n_2,p}$, so ist $S_1 + S_2 = X_1 + \cdots + X_{n_1+n_2} \sim \mathrm{Bin}_{n_1+n_2,p}$.)

4. (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 1.58 (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

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

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

Beispiel 1.59 (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 $Z+a\sim\mathcal{N}_{\mu+a,\sigma^2}$).

Die Behauptung folgt aus der Invarianz der multi-dimensionalen Standard-Normalverteilung unter orthogonalen Transformationen (vgl. Beob. 1.42):

Seien $a, b \in (0,1)$ mit $a^2 + b^2 = 1$, so ist die 2×2-Matrix

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} \quad \text{orthogonal}$$

(Dies ist eine Drehmatrix, wir könnten $a = \cos(\varphi), b = \sin(\varphi)$ für ein geeign. $\varphi\in[-\pi,\pi)$ schreiben, und

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}\begin{pmatrix} a & -b \\ b & a \end{pmatrix} = \begin{pmatrix} a^2+b^2 & -ab+ba \\ -ba+ab & a^2+b^2 \end{pmatrix}$$

Seien Z_1, Z_2 u.a., $\sim\mathcal{N}_{0,1}$, dann haben nach Beob. 1.42

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

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

Bemerkung. Man kann – anstelle von Beob. 1.42 – in diesem Fall auch das Faltungsintegral explizit ausrechnen:

Betrachte wieder o.E. den Fall $\mu_1 = \mu_2$.

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 $\mathcal{N}_{z/(1+(\sigma_2/\sigma_1)^2),\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2)}(\mathbb{R}) = 1$.)

1.3 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?“

1.3.1 Diskreter Fall

Sei X reelle ZV mit abzählbarem Wertebereich (in einem gewissen Zufallsexperiment $\mathscr{X}$) 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 1.60. Der *Erwartungswert* von X ist definiert als

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

sofern die Reihe absolut konvergiert (d.h. sofern $\sum_{x\in S_X} |x| P(X = x) < \infty$ gilt, dann kann die Summation in beliebiger Reihenfolge erfolgen). Manchmal schreibt man auch $\mu_X := \mathbb{E}[X]$.

Man sagt dann „X besitzt einen Erwartungswert“ und schreibt dies auch als $X \in \mathcal{L}^1$ (bzw. $X \in \mathcal{L}^1(P)$, wenn die zugrundeliegende Wahrscheinlichkeiten $P(\cdot)$ nicht aus dem Kontext klar sind).

Beispiel 1.61. 1. A ein Ereignis, so ist $\mathbb{E}[I_A] = 1 \cdot P(I_A = 1) + 0 \cdot P(I_A = 0) = P(A)$.

2. W Augenzahl bei einem fairen Würfelwurf (W ist uniform auf $\{1,2,3,4,5,6\}$), so ist

$$\mathbb{E}[W] = \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \frac{1}{6}\cdot 3 + \frac{1}{6}\cdot 4 + \frac{1}{6}\cdot 5 + \frac{1}{6}\cdot 6 = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3{,}5$$

(allgemein: X uniform auf $\{1,2,\dots,s\}$ mit $s \in \mathbb{N}$, so ist
$\mathbb{E}[X] = \sum_{i=1}^{s} \frac{1}{s}\cdot i = \frac{1}{s}\frac{s(s+1)}{2} = \frac{s+1}{2}$.)

3. X habe Werte in $S := \{2,3,4,\dots\} \cup \{-2,-3,-4,\dots\}$ mit Gewichten $P(X = n) = P(X = -n) = \frac{1}{2n(n-1)}$ für $n = 2,3,\dots$ (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), dann ist

$$\sum_{x\in S} |x| P(X = x) = \sum_{n=2}^{\infty} \frac{n}{n(n-1)} = \sum_{k=1}^{\infty} \frac{1}{k} = \infty,$$

d.h. X besitzt keinen Erwartungswert.

Wenn man S durchnummerierte mit $x_{2i} = i+1$, $x_{2i-1} = -i-1$, $i \in \mathbb{N}$, so wäre

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

(denn $\sum_{j=1}^{2N} x_j P(X = x_j) = 0$); wenn man andererseits S durchnummerierte mit $x_{3i} = -i-1$, $x_{3i-2} = 2i$, $x_{3i-1} = 2i+1$, $i \in \mathbb{N}$, so wäre

$$\sum_{j=1}^{\infty} x_j P(X = x_j) = \lim_{N\to\infty} \sum_{j=1}^{N} x_j P(X = x_j) = \frac{1}{2}\log(2) \neq 0$$

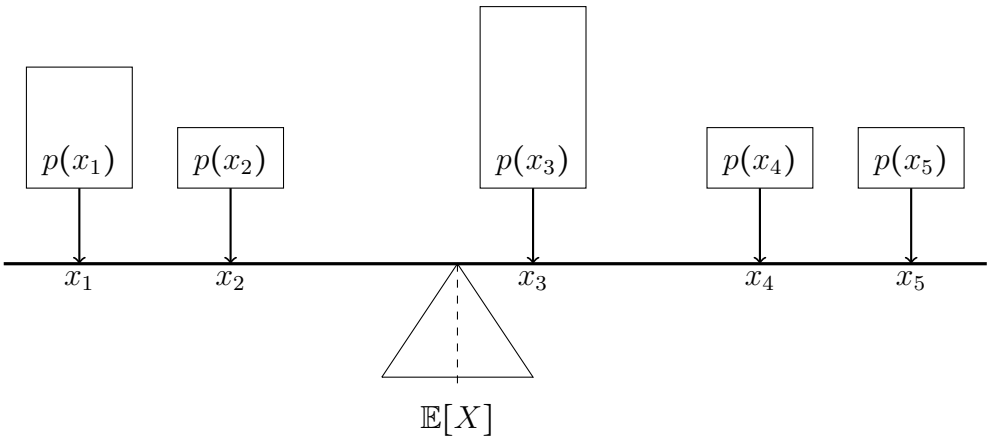
(denn $\sum_{j=1}^{3N} x_j P(X = x_j) = \sum_{i=2}^{N} \frac{-i}{2i(i-1)} + \sum_{i=2}^{2N} \frac{i}{2i(i-1)} = -\frac{1}{2}\sum_{k=1}^{N-1}\frac{1}{k} + \frac{1}{2}\sum_{k=1}^{2N-1}\frac{1}{k} \sim -\frac{1}{2}\log(N) + \frac{1}{2}\log(2N) = \frac{1}{2}\log(2)$).

(Wir sehen hier ein Beispiel für die Tatsache aus der Analysis, dass der Wert einer bedingt konvergenten Reihe von der Summationsreihenfolge abhängt.)

Bemerkung 1.62. 1. Eine beschränkte reellwertige ZV X (d.h. es gibt ein $M < \infty$ mit $P(-M \le X \le M) = 1$) besitzt stets einen Erwartungswert.

(Dies gilt insbesondere, wenn X nur endlich viele mögliche Werte hat.)

2. Wenn X endlich viele mögliche Werte $x_1,\dots,x_n$ (mit Gewichten $p(x_i) = P(X = x_i)$) hat, so kann man $\mathbb{E}[X]$ 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 in 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. Der Erwartungswert von X muss nicht notwendigerweise ein möglicher Wert von X sein ($P(X = \mathbb{E}[X]) = 0$ ist durchaus möglich, siehe Bsp. 1.61), daher kann man die Interpretation von $\mathbb{E}[X]$ als „typischer Wert von X“ i.A. nicht wörtlich nehmen.

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

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

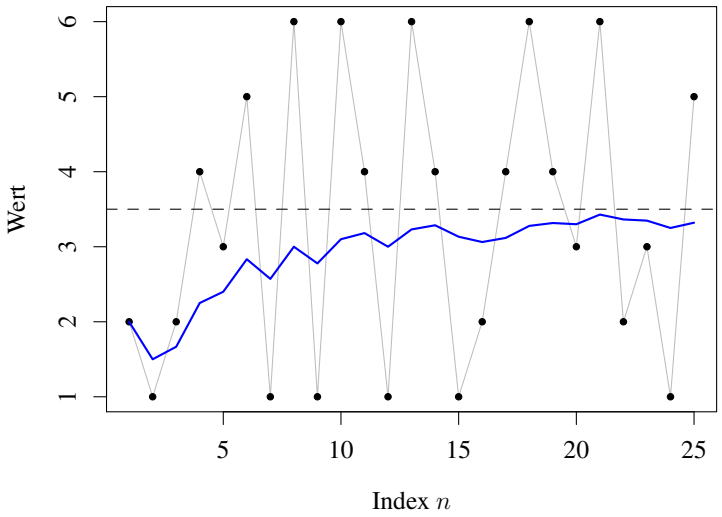
   (in geeignetem Sinn), dies ist die Aussage des *Gesetzes der großen Zahlen*, das wir später sehen werden.

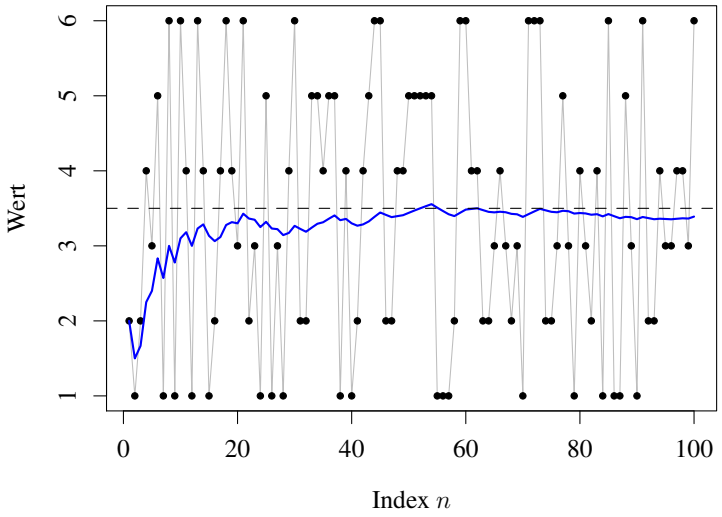
   Es ist nämlich

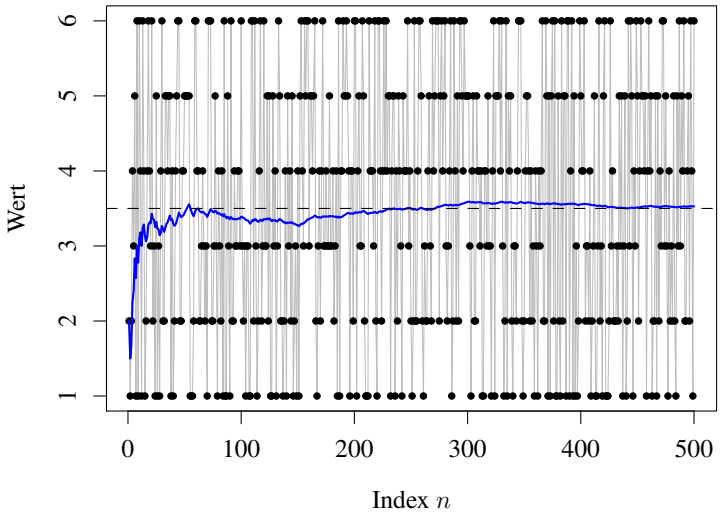
$$M_n = \sum_x x \cdot \frac{\#\{i \le n : X_i = x\}}{n}$$

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

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







(Beachte: Wir haben dies bereits in der Anwendung „eine Monte-Carlo-Methode zur Integration“ in Kapitel 0 verwendet.)

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

5. 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 1.63. 1. Sei $X \sim \mathrm{Bin}_{n,p}$, $n \in \mathbb{N}$, $p \in [0,1]$:

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

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

$$\begin{aligned}\mathbb{E}[X] &= \sum_{n=0}^{\infty} np(1-p)^n = p\sum_{n=1}^{\infty}\sum_{k=1}^{n} 1\cdot(1-p)^n = p\sum_{k=1}^{\infty}\sum_{n=k}^{\infty}(1-p)^n \\ &= p\sum_{k=1}^{\infty}(1-p)^k\sum_{j=0}^{\infty}(1-p)^j = \sum_{k=1}^{\infty}(1-p)^k = \frac{1}{p} - 1\end{aligned}$$

3. 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 1.64 (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

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

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

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

2.

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

3. $\mathbb{E}[X] = \sum_{x\,(\geq 0)} xP(X = x)$ 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

$$\mathbb{E}[XY] = \sum_z zP(XY = z) = \sum_{x,y\neq 0} xyP(X = x, Y = y)$$
$$= \sum_{x\neq 0} xP(X = x) \cdot \sum_{y\neq 0} yP(Y = y) = \mathbb{E}[X]\,\mathbb{E}[Y].$$

□

Beobachtung 1.65 (Erwartungswerte für Kompositionen)**.** X (diskrete) reelle ZV, $g : \mathbb{R} \to \mathbb{R}$, $Y := g(X)$.

Dann besitzt Y einen Erwartungswert 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 1.66. 1. Seien $X_1, \ldots, X_n$ 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. 1.63, 2. bestimmt, hier kommen wir allerdings ohne explizite Rechnung aus).

2. Sei $X \sim \mathrm{Hyp}_{s,w,k}$ hypergeometrisch verteilt, vgl. Bsp. 1.17. Denken wir an eine Urne mit s schwarzen und w weißen Kugeln, aus der k mal ohne Zurücklegen gezogen wird, so ist

$$X \stackrel{d}{=} I_{A_1} + \cdots + I_{A_k} \quad \text{mit } A_i = \{i\text{-te gezogene Kugel ist schwarz}\}$$

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

1.3.2 Der Fall mit Dichte

Definition 1.67. Sei X reellwertige ZV mit Dichte f_X, dann besitzt X einen Erwartungswert (auch $X \in \mathcal{L}^1$ geschrieben), wenn gilt $\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty$ und man setzt dann

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

Beispiel 1.68. 1. $X \sim \mathcal{N}_{0,1}$ hat $\mathbb{E}[X] = 0$, denn aus der Symmetrie der Dichte folgt

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

(strenggenommen muss man auch prüfen, dass $(2\pi)^{-1/2} \int_{-\infty}^{\infty} |x| e^{-x^2/2}\,dx = (2/\pi)^{1/2} \int_0^{\infty} x e^{-x^2/2}\,dx = (2/\pi)^{1/2}\big[-e^{-x^2/2}\big]_0^{\infty} = (\pi/2)^{-1/2} < \infty$)

Somit gilt auch: $Y \sim \mathcal{N}_{\mu,\sigma^2}$ hat $\mathbb{E}[Y] = \mu$, denn $\sigma X + \mu =^d Y$ nach Bsp. 1.34).

2. $X \sim \mathrm{Exp}_\lambda$ hat $\mathbb{E}[X] = 1/\lambda$, denn

$$\begin{aligned}\mathbb{E}[X] &= \int_0^{\infty} x\,\lambda e^{-\lambda x}\,dx = \big[x(-e^{-\lambda x})\big]_0^{\infty} - \int_0^{\infty} 1 \cdot (-e^{-\lambda x})\,dx \\ &= \int_0^{\infty} e^{-\lambda x}\,dx = \frac{1}{\lambda} \int_0^{\infty} \lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}\end{aligned}$$

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

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

Bericht 1.69. 1. Man kann prinzipiell den Fall mit Dichte aus dem diskreten Fall herleiten: X habe Dichte f_X, so nimmt $X_{(n)} = \frac{1}{n}\lfloor nX \rfloor$ den Wert $\frac{k}{n}$, $k \in \mathbb{N}$ 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)

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

2. (Analogon zu Beob. 1.65 im Fall mit Dichte)
Sei $X = (X_1, \ldots, X_d)$ $\mathbb{R}^d$-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.

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

und in diesem Fall

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

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

3. Die Rechenregeln aus Satz 1.64 gelten auch im Fall mit Dichte.

1.3.3 Varianz und Kovarianz

Für eine reellwertige Zufallsvariable X heißt $\mathbb{E}[X^2]$ das *2. Moment von* X (allgemein heißt $\mathbb{E}[X^p]$ das p-te Moment).

Man sagt, dass X ein 2. Moment besitzt, wenn $\mathbb{E}[X^2] < \infty$ gilt und schreibt dies auch als $X \in \mathcal{L}^2$ (bzw. $X \in \mathcal{L}^2(P)$, wenn die zugrundeliegende Wahrscheinlichkeiten $P(\cdot)$ nicht aus dem Kontext klar sind).

Definition 1.70. 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 $\mathrm{Cov}[X, Y] = 0$.

Die Standardabweichung σ_X ist – neben dem Erwartungswert $\mathbb{E}[X]$ – eine weitere wichtige Kenngröße der Verteilung einer Zufallsvariable X, sie gibt eine Antwort auf die – etwas salopp formulierte – Frage „Wie sehr weicht X typischerweise von $\mathbb{E}[X]$ ab?“

Einen Hinweis dazu gibt die Chebyshev-Ungleichung:

Satz 1.71. *Sei X reelle ZV, $f : [0, \infty) \to [0, \infty)$ monoton wachsend.*
1. Für $a > 0$ mit $f(a) > 0$ gilt

$$P\big(|X| \geq a\big) \leq \frac{1}{f(a)} \mathbb{E}\big[f(|X|)\big] \qquad \textit{(Markov}^{13}\textit{-Ungleichung).} \tag{1.13}$$

2. Für $X \in \mathcal{L}^2$ gilt

$$P\big(|X - \mathbb{E}[X]| \geq a\big) \leq \frac{\mathrm{Var}[X]}{a^2} \qquad \textit{(Chebyshev}^{14}\textit{-Ungleichung).} \tag{1.14}$$

Beweis. 1. Sei $Y := f(a) I_{\{|X| \geq a\}}$, 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 1.64, 2.

2. Wende *1.* an auf $\widetilde{X} := X - \mathbb{E}[X]$ und $f(a) = a^2$. □

Insbesondere (wähle $a = b\sigma_X$ in (1.14)): Die W'keit, dass X von $\mathbb{E}[X]$ um mehr als das b-fache von σ_X abweicht, ist höchstens $1/b^2$.

[14] Andrei Andrejewich Markov, 1856–1922.
[14] Pafnuty Lvovich Chebyshev, 1821–1894.

Beobachtung 1.72. 1. Wegen $|XY| \le X^2 + Y^2$ 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. $\mathrm{Var}[X] = 0 \iff P\big(X = \mathbb{E}[X]\big) = 1$

(„$\Leftarrow$“ ist klar, für „$\Rightarrow$“ wende Satz 1.64, 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 1.73. 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. 1.19 und wir hatten gesehen, dass $\mathbb{E}[X] = (1-p)/p$, siehe Bsp. 1.63, 2).

Es ist

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

(verwende, dass $f(t) \coloneqq \sum_{n=0}^{\infty} t^n = \frac{1}{1-t}$ (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. $X \sim \mathcal{N}_{\mu,\sigma^2}$ hat $\mathrm{Var}[X] = \sigma^2$ (wir hatten in Bsp. 1.68 bereits gesehen, dass $\mathbb{E}[X] = \mu$):

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

(Wir haben Bericht 1.69, 2. verwendet, dann im Integral $z = (x - \mu/\sigma)$ substituiert und partiell integriert.)

Satz 1.74 (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

$$\big|\mathrm{Cov}[X, Y]\big| \le \sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]} \quad \textit{(Cauchy-Schwarz-Ungleichung}^{15}\textit{)}$$

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

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 $\mathbb{E}[X\,Y] = \mathbb{E}[X]\,\mathbb{E}[Y]$ nach Satz 1.64, 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. 1.72, 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 1.75. 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, dass Gleichheit genau dann eintritt, wenn $\mathrm{Var}[Y] = 0$ oder $\mathrm{Var}[X+\alpha Y] = 0$.)

Beispiel 1.76.

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 1.74, 2.)
$$\mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[Y_i] = n\mathrm{Var}[Y_1] = np(1-p)$$
(vgl. auch Bsp. 1.73, 3.).

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

(Erinnerung: Für $y_1, \dots, y_n \in \{0,1\}$ mit $y_1 + \cdots + y_n = k$ gilt

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \frac{s(s-1)(s-2)\cdots(s-k+1)\cdot w(w-1)\cdots(w-n+k+1)}{(s+w)(s+w-1)(s+w-2)\cdots(s+w-n+1)},$$

was nicht von der Reihenfolge abhängt – die Y_i sind „austauschbar".)

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

$$\mathbb{E}[Y_i Y_i] = \mathbb{E}[Y_1 Y_2] = P(A_1 \cap A_2) = \frac{s}{s+w}\frac{s-1}{s+w-1},$$

$$\begin{aligned}\operatorname{Cov}[Y_i, Y_j] &= \mathbb{E}[Y_i Y_i] - \mathbb{E}[Y_i]\,\mathbb{E}[Y_j] = \frac{s}{s+w}\frac{s-1}{s+w-1} - \Big(\frac{s}{s+w}\Big)^2 \\ &= \frac{s}{s+w}\Big(\underbrace{\frac{s-1}{s+w-1} - \frac{s}{s+w}}_{= -\frac{w}{s+w}\frac{1}{s+w-1}}\Big) = -p(1-p)\frac{1}{s+w-1},\end{aligned}$$

also

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

(Wir sehen: Die Varianz ist kleiner im Fall ohne Zurücklegen als im Fall mit Zurücklegen – insbes. ist sie natürlich $= 0$ im Fall $n = s + w$.)

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

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

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

Definition 1.77. Seien $X, Y \in \mathcal{L}^2$.

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

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

(Die Cauchy-Schwarz-Ungleichung (Satz 1.74, 4.) zeigt, dass $|\kappa_{X,Y}| \leq 1$.)

Beobachtung 1.78 (Interpretation des Korrelationskoeffizienten via „beste lineare Vorhersage“)**.** Es ist

$$\min_{\beta_0,\beta_1\in\mathbb{R}} \mathbb{E}\big[(Y-\beta_1X-\beta_0)^2\big] = (1-\kappa_{X,Y}^2)\min_{\beta_0\in\mathbb{R}} \mathbb{E}\big[(Y-\beta_0)^2\big] \quad \Big(= (1-\kappa_{X,Y}^2)\mathrm{Var}[Y]\Big),$$

denn der Ausdruck auf der linken Seite ist

$$\begin{aligned}
&\mathrm{Var}[Y-\beta_1X-\beta_0] + \big(\mathbb{E}[Y]-\beta_1\mathbb{E}[X]-\beta_0\big)^2\\
&= \mathrm{Var}[Y] - 2\beta_1\mathrm{Cov}[X,Y] + \beta_1^2\mathrm{Var}[X] + \big(\mathbb{E}[Y]-\beta_1\mathbb{E}[X]-\beta_0\big)^2\\
&= \sigma_Y^2 - 2\beta_1\sigma_X\sigma_Y\kappa_{X,Y} + \beta_1^2\sigma_X^2 + \big(\mathbb{E}[Y]-\beta_1\mathbb{E}[X]-\beta_0\big)^2\\
&= \sigma_Y^2\big(1-\kappa_{X,Y}^2\big) + \sigma_X^2\Big(\beta_1 - \frac{\sigma_Y}{\sigma_X}\kappa_{X,Y}\Big)^2 + \big(\mathbb{E}[Y]-\beta_1\mathbb{E}[X]-\beta_0\big)^2,
\end{aligned}$$

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 $\beta = \mathbb{E}[Y]$.)

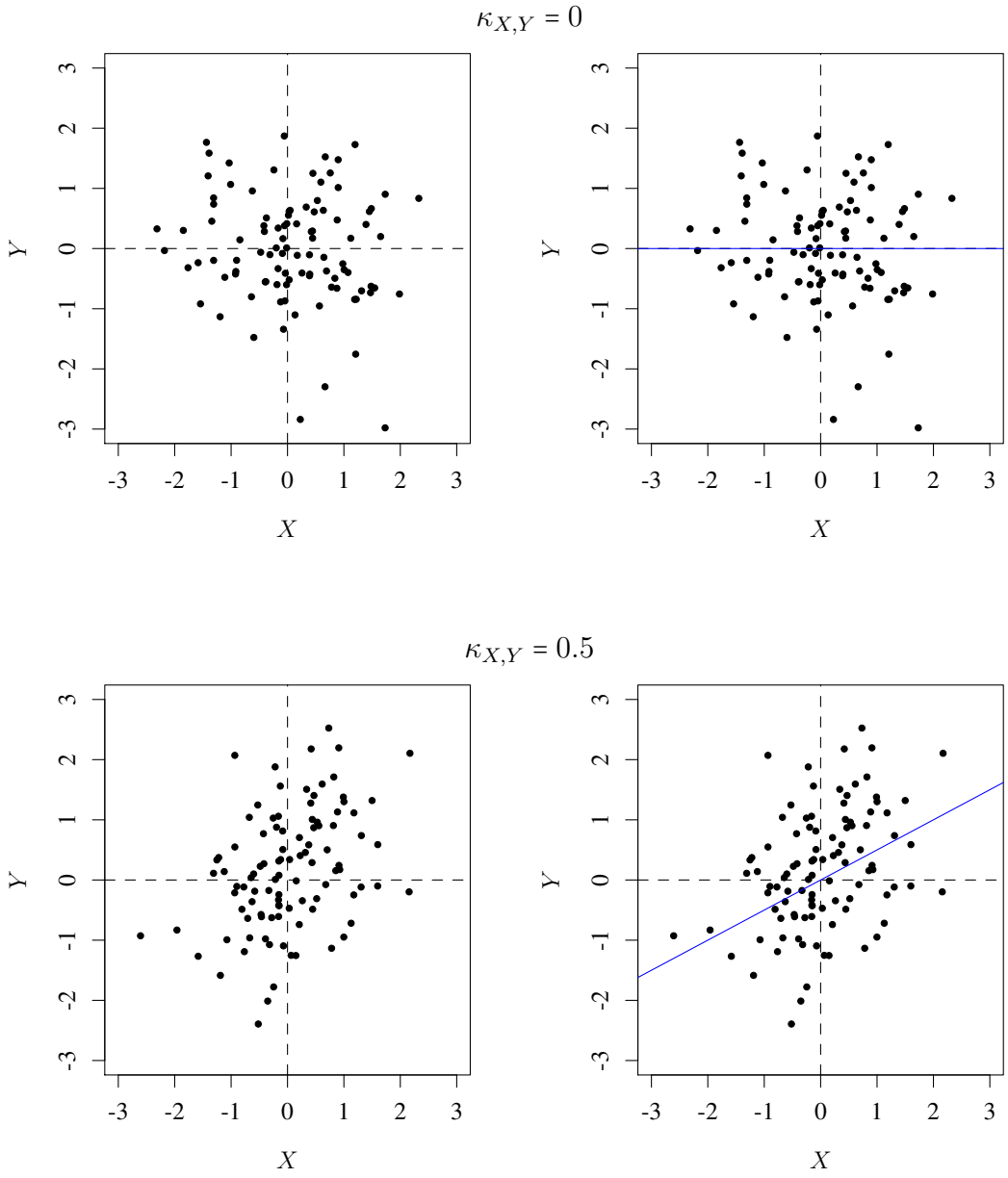
Im Sinne einer möglichst kleinen quadratischen Abweichung ist $\mathbb{E}[Y]$ die beste konstante „Vorhersage“ von Y. Man kann demnach um einen Faktor $(1-\kappa_{X,Y}^2)$ besser vorhersagen, wenn man stattdessen eine affin-lineare Funktion von X verwenden darf.

Demnach (vgl. auch Bem. 1.75)

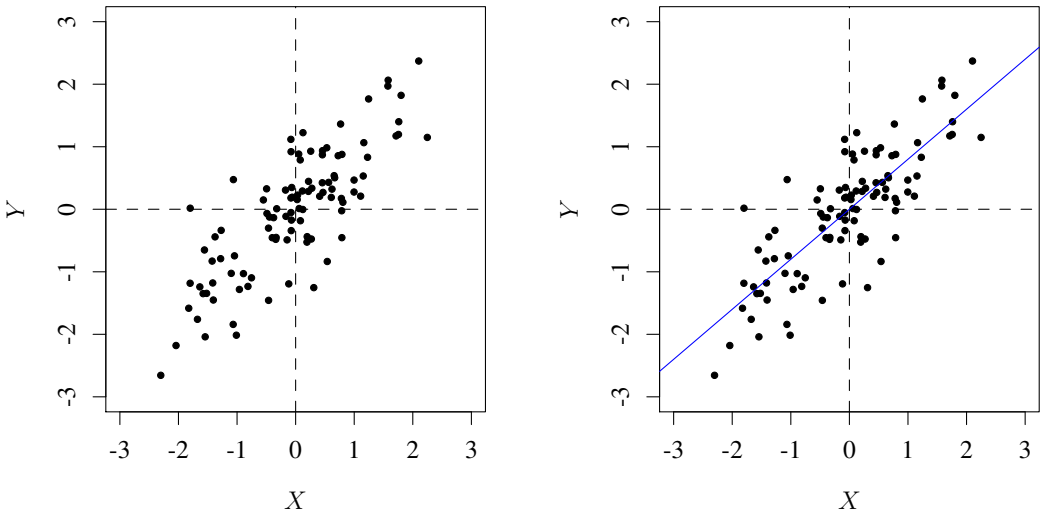
$\lvert\kappa_{X,Y}\rvert = 1$	$\leftrightarrow$	perfekter linearer Zusammenhang zwischen X und Y
$\kappa_{X,Y} = 1$	$\leftrightarrow$	perfekter linearer Zusammenhang zwischen X und Y mit positivem Koeffizienten (X größer als $\mathbb{E}[X] \iff Y$ größer als $\mathbb{E}[Y]$)
$\kappa_{X,Y} = -1$	$\leftrightarrow$	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. 1.76, 3.

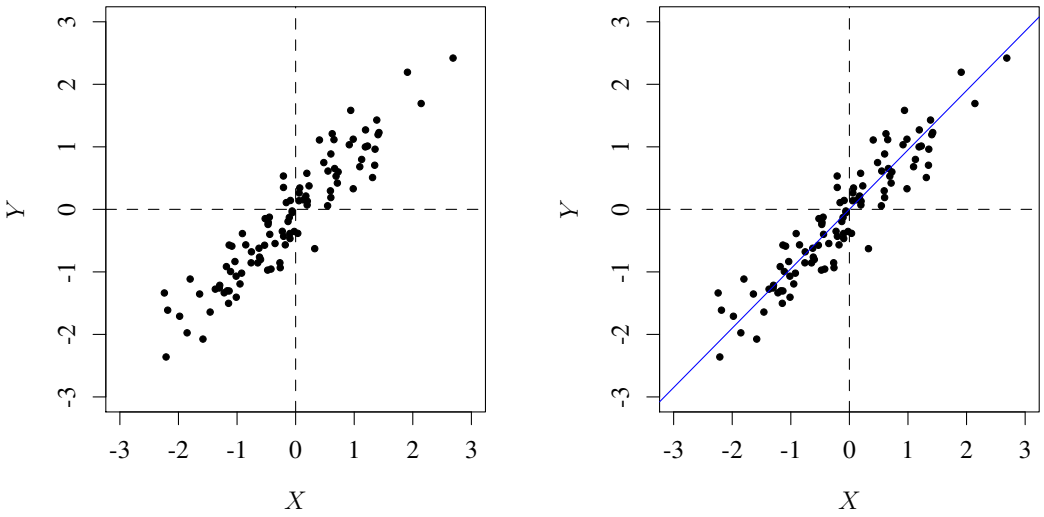
Die folgenden Scatterplots zeigen jeweils 100 simulierte Paare (X,Y), wobei $\sigma_X = \sigma_Y = 1$ und $\kappa_{X,Y}$ den angegebenen Wert hat. (Blau eingezeichnet ist die „Vorhersagegerade“ $x \mapsto \beta_1^*x + \beta_0^*$.)


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

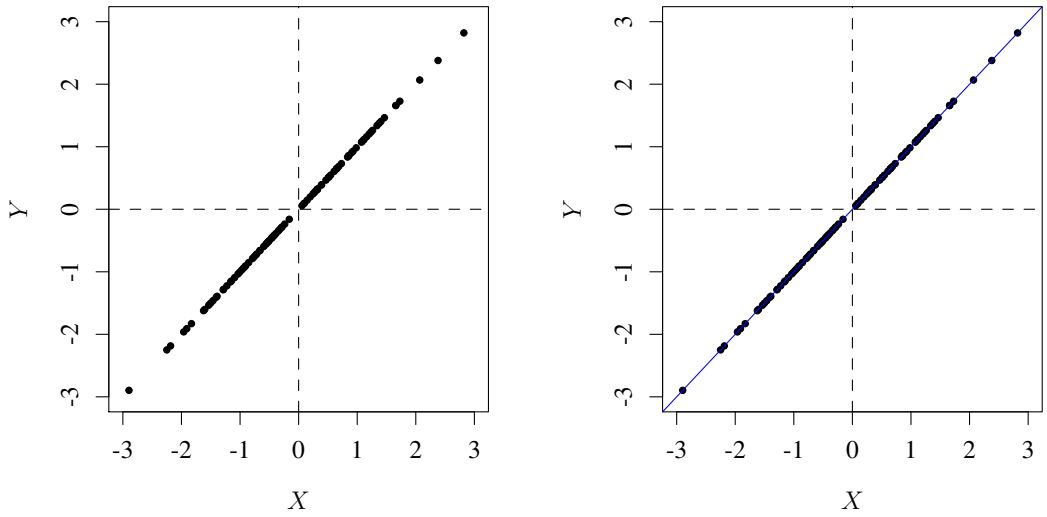
$\kappa_{X,Y} = 0.8$



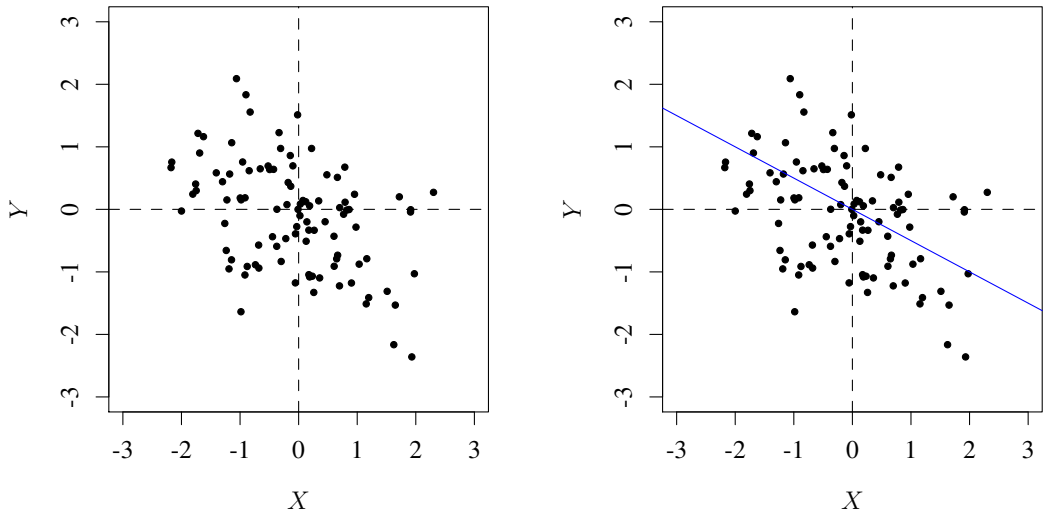
$\kappa_{X,Y} = 0.95$



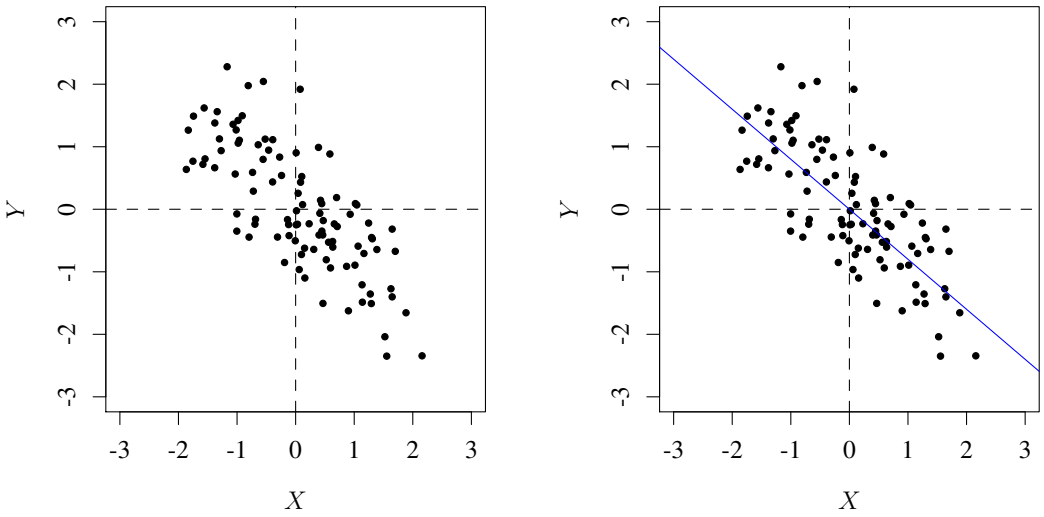
$\kappa_{X,Y} = 1$



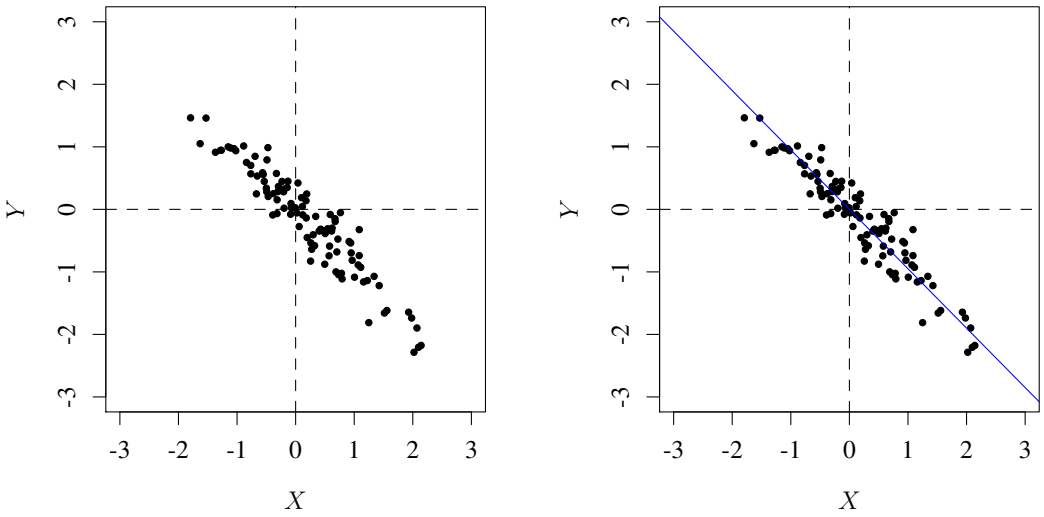
$\kappa_{X,Y} = -0.5$

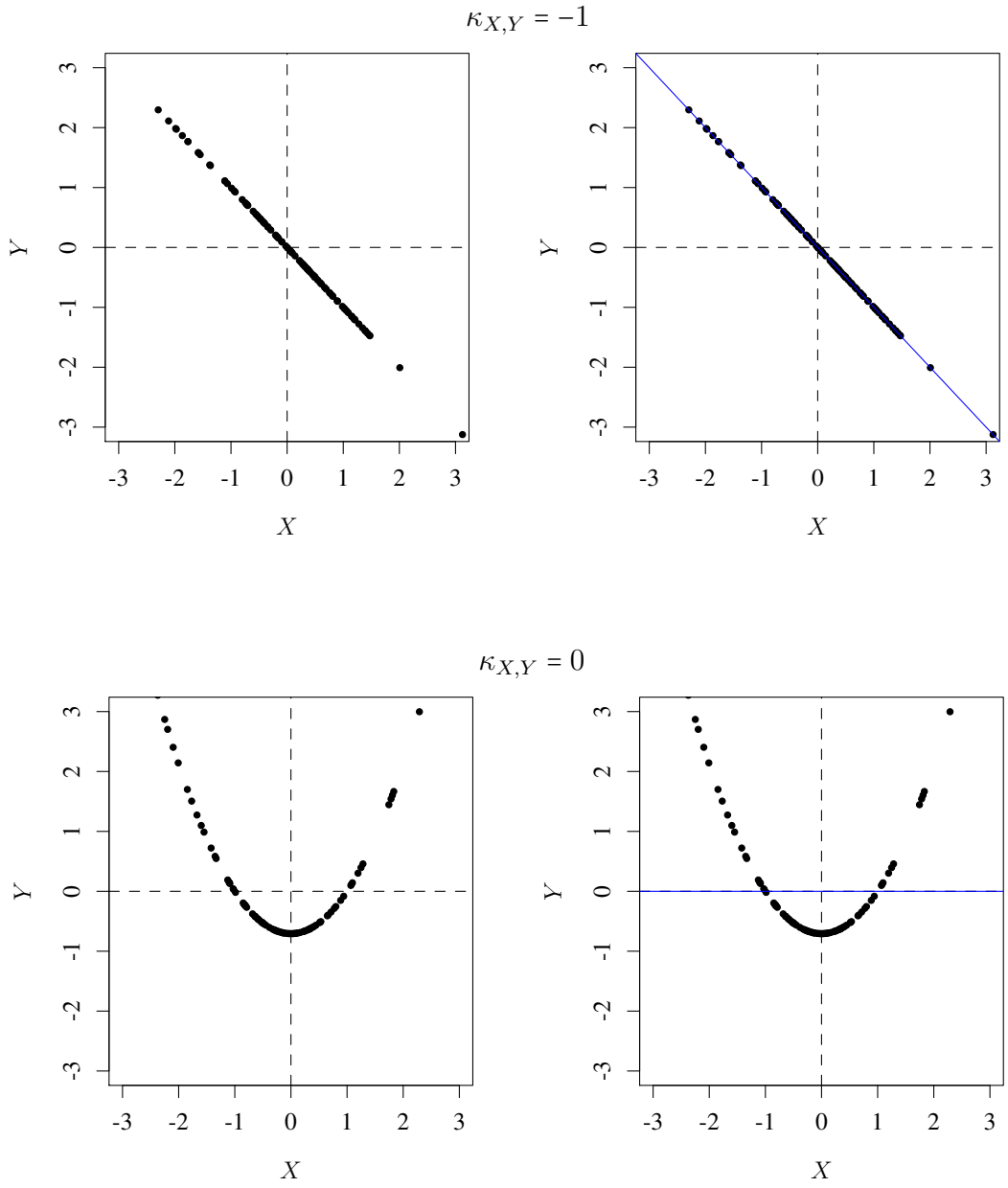


$\kappa_{X,Y} = -0.8$



$\kappa_{X,Y} = -0.95$





1.3.4 Median(e)

Anschaulich ist der Median einer reellen Zufallsvariable X der Wert m, so dass

$$„P(X \leq m) = \frac{1}{2} = P(X \geq m)“$$

gilt.

Da man diese Gleichheit (zumal im diskreten Fall) nicht immer genau einstellen kann, definiert man formal folgendermaßen:

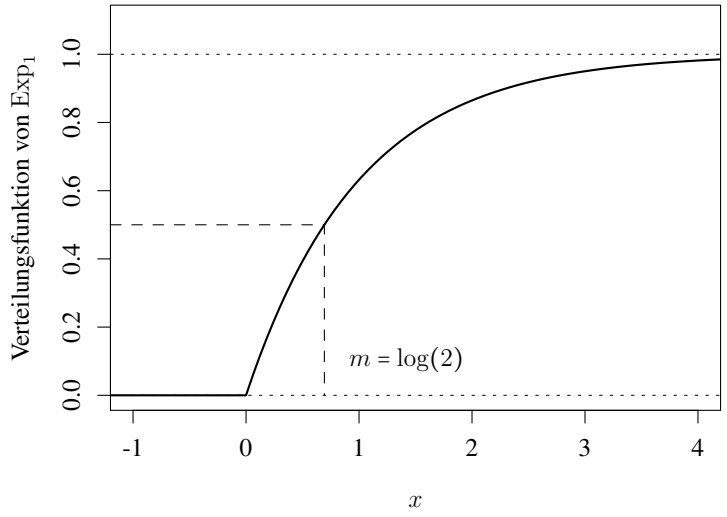
Definition 1.79. X reelle ZV, m heißt (ein) Median von X (auch „Zentralwert“, manchmal auch m_X geschrieben), wenn gilt

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

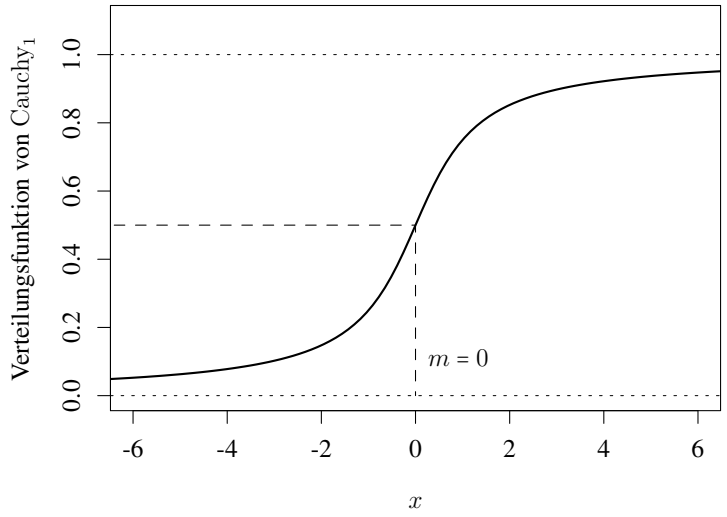
Ein Median existiert stets, auch wenn X keinen Erwartungswert besitzt.

Man kann den Median als eine „robustere“ Antwort auf die Aufgabe, für eine ZV *nur einen* „typischen Wert“ anzugeben, ansehen (im Gegensatz zum Erwartungswert besitzt ja jede Verteilung einen Median). Allerdings gibt es für Mediane keine so angenehmen Rechenregeln, wie sie Satz 1.64 für den Erwartungswert liefert.

Beispiel 1.80. 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 $m = \frac{1}{\theta}\log 2$.

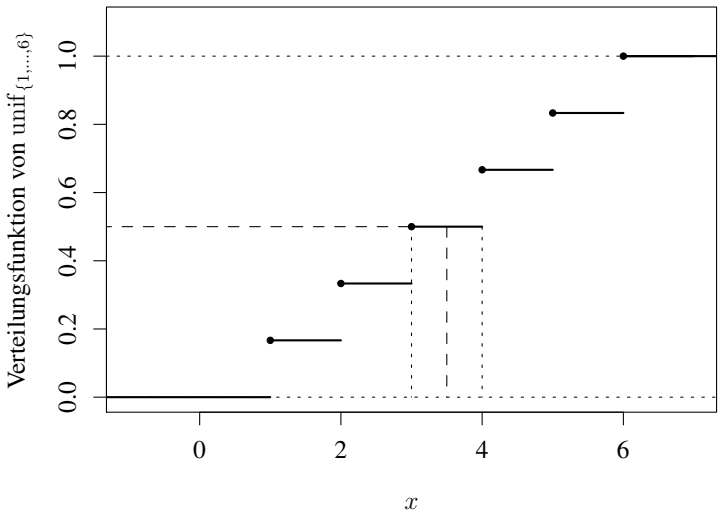


2. X Cauchy-verteilt mit Dichte $\frac{1}{\pi}\frac{1}{1+x^2}$, Verteilungsfunktion $\frac{1}{2}\frac{1}{\pi}\arctan(x)$, der (eindeutig bestimmte) Median ist $m = 0$.



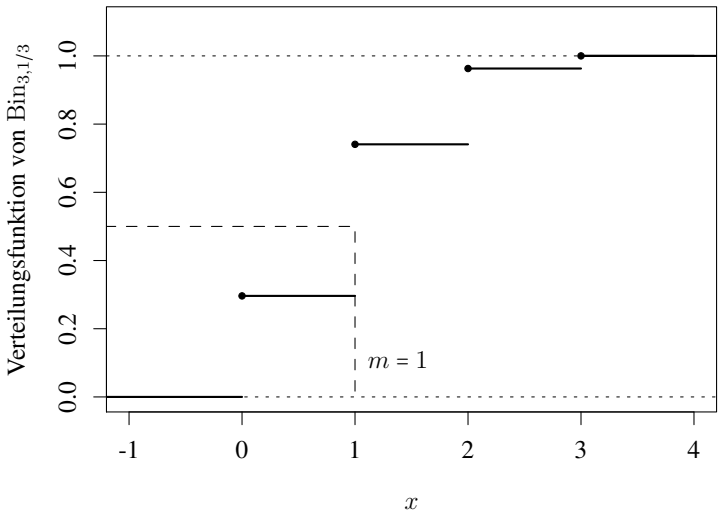
(Wegen der Symmetrie der Dichte, es gibt keinen Erwartungswert, vgl. Bsp. 1.68, 3.)

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



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 1.81. 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

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

wobei für die erste Ungleichung die Monotonie des Erwartungswerts (Satz 1.64, 2., beachte: $X - m \le |X - m|$ und $-(X - m) \le |X - m|$ und für die letzte Ungleichung die Cauchy-Schwarz-Ungleichung (Satz 1.74, 4.) verwenden. □

1.4 Gesetz der großen Zahlen und zentraler Grenzwertsatz

1.4.1 Gesetz der großen Zahlen

Satz 1.82 ((Schwaches) Gesetz der großen Zahlen)**.** *Seien $X_1, X_2, \dots$ unabhängige und identisch verteilte (u.i.v.) reellwertige ZVn mit $\mathbb{E}[X_1] = \mu$ und $\mathrm{Var}[X_1] < \infty$, dann gilt für jedes $\varepsilon > 0$*

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \le \frac{\mathrm{Var}[X_1]}{\varepsilon^2 n} \underset{n\to\infty}{\longrightarrow} 0 \tag{1.15}$$

Beweis. Sei $Y_n := \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$, es ist

$$\begin{aligned}\mathrm{Var}[Y_n] &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}[X_i - \mu] + \sum_{\substack{1\le i,j\le n\\ i\neq j}}\mathrm{Cov}[X_i - \mu, X_j - \mu]\right) \\ &= \frac{1}{n^2}\mathrm{Var}[X_i] = \frac{1}{n}\mathrm{Var}[X_1]\end{aligned}$$

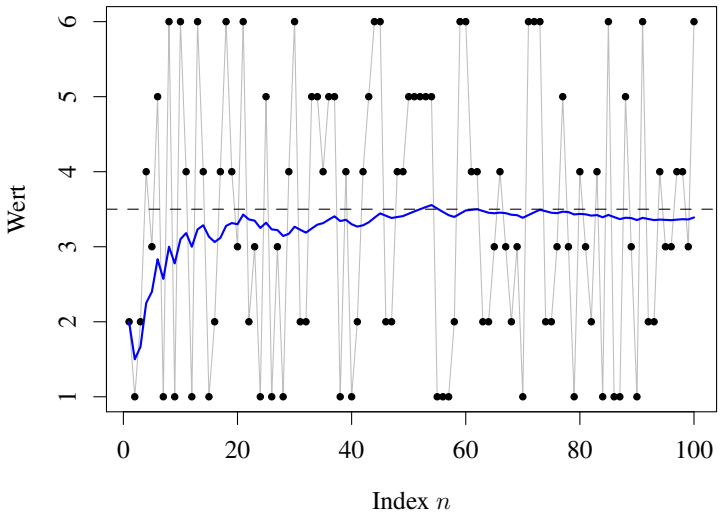
mit Satz 1.74, somit

$$P\big(|Y_n| \ge \varepsilon\big) \le \frac{\mathrm{Var}[Y_n]}{\varepsilon^2} = \frac{\mathrm{Var}[X_1]}{\varepsilon^2 n}$$

gemäß Chebyshev-Ungleichung (Satz 1.71). □

Erinnerung. Wir hatten bereits in Bem. 1.62, 3. das Gesetz der großen Zahlen illustriert:

$X_1, X_2, \dots$ unabhängig, uniform auf $\{1, 2, 3, 4, 5, 6\}$, X_n sind die schwarzen Punkte, $(X_1 + \cdots + X_n)/n$ die blaue Linie



(und es auch schon im „Auftakt“-Beispiel in Kap 0 verwendet)

Bemerkung 1.83. 1. Wir entnehmen dem Beweis von Satz 1.82 folgende kleine Verallgemeinerung:

Sind $X_1, X_2, \cdots \in \mathcal{L}^2$ seien paarweise unkorreliert mit

$$\sup_n \operatorname{Var}[X_n] \leq \theta < \infty,$$

dann gilt für $\varepsilon > 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big)\right| > \varepsilon\right) \leq \frac{\theta}{\varepsilon^2 n} \left(\underset{n\to\infty}{\longrightarrow} 0\right). \tag{1.16}$$

(Das Argument geht genauso wie im Beweis von Satz 1.82, wenn wir $Y_n := \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big)$ setzen.)

2. Seien $Y_n, n \in \mathbb{N}$ und Y reellwertigen ZVn in einem gewissen Zufallsexperiment $\mathscr{X}$.

Man sagt die Folge $(Y_n)_{n\in\mathbb{N}}$ *konvergiert stochastisch* gegen Y, auch geschrieben

$$Y_n \underset{n\to\infty}{\overset{\text{stoch.}}{\longrightarrow}} Y,$$

(auch $Y_n \underset{n\to\infty}{\longrightarrow} Y$ stoch. oder $Y_n \underset{n\to\infty}{\overset{P}{\longrightarrow}} Y$), wenn gilt

$$\forall\, \varepsilon > 0 : \lim_{n\to\infty} P\big(|Y_n - Y| > \varepsilon\big) = 0.$$

Man spricht damit Satz 1.82 oft folgendermaßen aus:

$$\frac{X_1 + \cdots + X_n}{n} \underset{n\to\infty}{\overset{\text{stoch.}}{\longrightarrow}} \mu$$

Bericht 1.84 (Nur der Vollständigkeit halber)**.** Die Konvergenzaussage (1.15) in Satz 1.82 sieht (zumindest mit Blick auf die in der Analysis übliche Definition der Konvergenz) vielleicht etwas merkwürdig aus.

Tatsächlich gilt für $X_1, X_2, \dots$ u.i.v. mit $\mathbb{E}[X_1] = \mu$ auch:

$$\text{Für jedes } \varepsilon > 0 \text{ gibt es ein (vom Zufall abhängiges) } N_0 \text{ mit}$$
$$\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \le \varepsilon \quad \text{für alle } n \ge N_0.$$

In der Literatur heißt dies manchmal das *starke Gesetz der großen Zahlen*, man sagt auch $(X_1 + \cdots + X_n)/n$ *konvergiert fast sicher* gegen μ.

Wir werden dies im weiteren Verlauf der Vorlesung nicht verwenden.

1.4.2 Zum zentralen Grenzwertsatz

Vorbemerkung. Seien $X_1, X_2, \dots$ unabhängig und identisch verteilt (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

$$\frac{X_1 + \cdots + X_n}{n} - \mu \xrightarrow[n\to\infty]{} 0 \quad \text{stochastisch}$$

gemäß dem Gesetz der großen Zahlen (Satz 1.82), aber feiner gefragt:

Wie groß ist $X_1 + \cdots + X_n - n\mu$ typischerweise?

Für $A \gg \sqrt{n}$ ist (mit Chebyshev-Ungleichung, Satz 1.71) zumindest

$$P\big(|X_1 + \cdots + X_n - n\mu| > A\big) \le \frac{n\sigma^2}{A^2} \quad \text{(sehr) klein.}$$

Tatsächlich ist $\sqrt{n}$ die korrekte Größenordnung der typischen Abweichungen von $X_1 + \cdots + X_n$ von $n\mu$, beachte dazu

$$\mathbb{E}\big[(X_1 + \cdots + X_n) - n\mu\big] = 0$$

und $\mathrm{Var}\big[(X_1 + \cdots + X_n) - n\mu\big] = n\mathrm{Var}[X_1] = n\sigma^2$, also

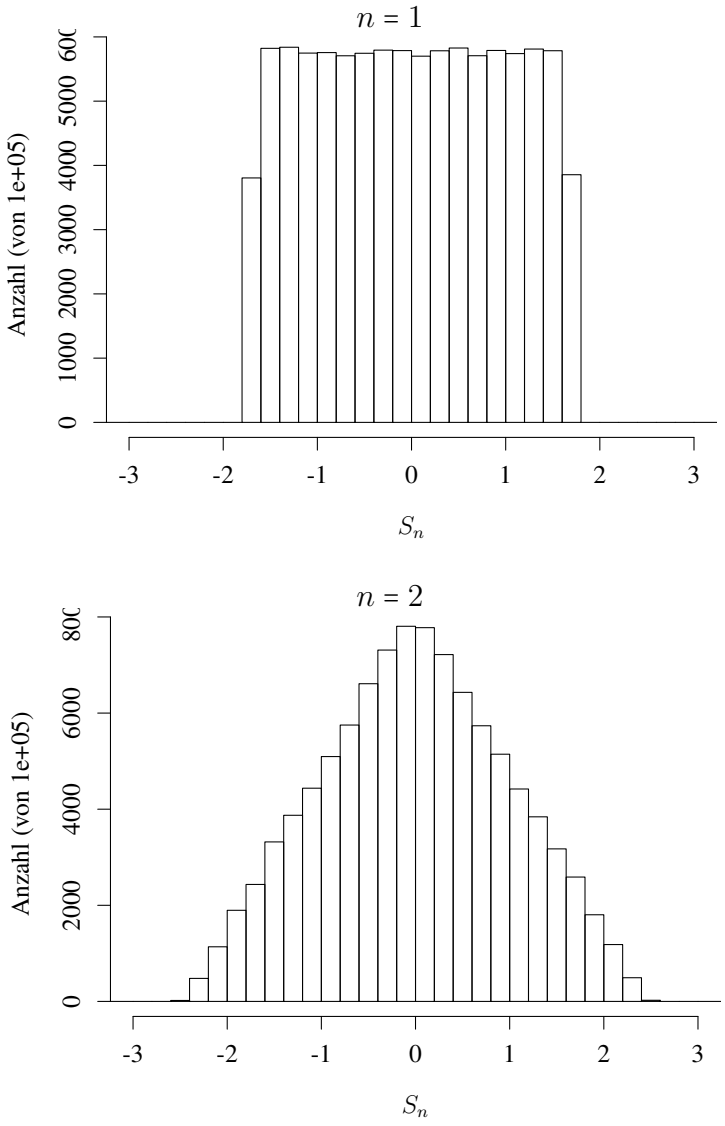
$$\mathrm{Var}\Big[\frac{(X_1 + \cdots + X_n) - n\mu}{\sqrt{n\sigma^2}}\Big] = 1$$

Demnach: Mit dieser Skalierung hängen zumindest Erwartungswert und Varianz nicht mehr von n ab.

Wie sieht es aber mit der „ganzen“ Verteilung aus? Wir betrachten dazu Simulationen:

$X_i \sim \text{unif}_{[0,1]}$, $\mathbb{E}[X_1] = 1/2$, $\text{Var}[X_1] = \frac{1}{12}$

Histogramme jeweils basierend auf 10^5 Simulationen von S_n



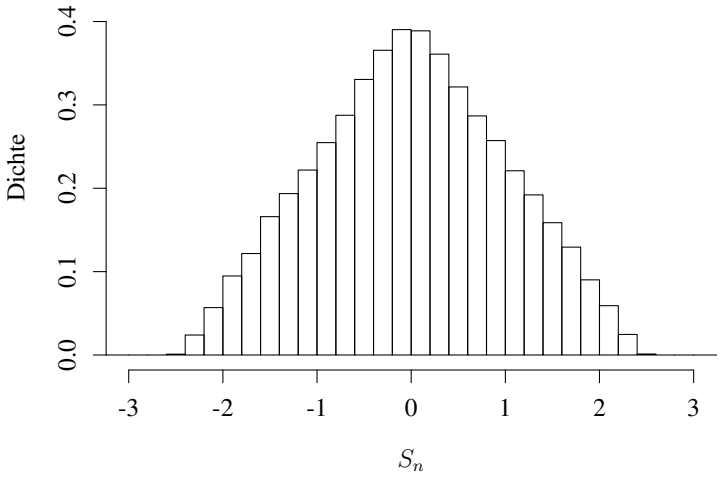
Zur Vergleichbarkeit gehen wir von den absoluten Anzahlen als Balkenhöhen zur sogenannten Dichte über, d.h. die Balkenhöhe ist nun jeweils

$$\frac{\text{Anzahl}}{100.000 \times \text{Balkenbreite}}.$$

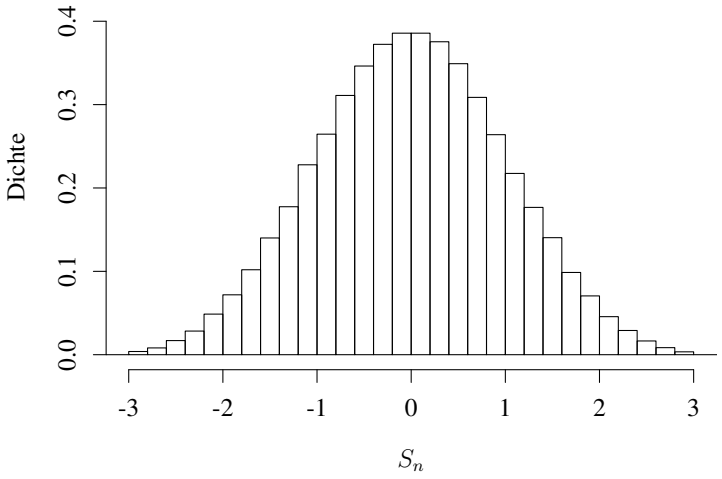
Damit wird die Gesamtfläche der Balken = 1 (wie bei einer Wahrscheinlichkeitsdichte).

(Da wir gleich breite Balken verwendet haben, entspricht dies einfach der Wahl einer anderen Skala auf der y-Achse)

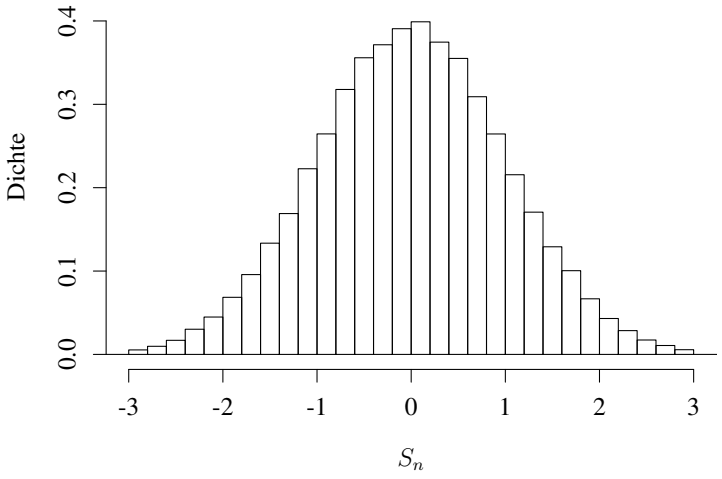
$n = 2$


Dichte
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
S_n

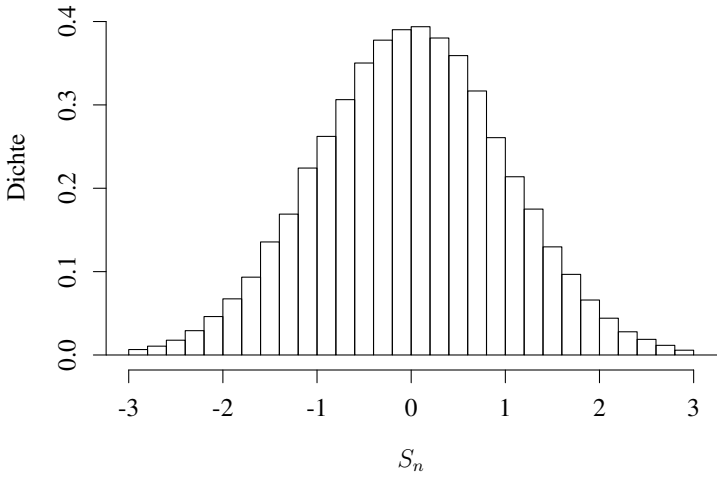
$n = 5$


Dichte
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
S_n

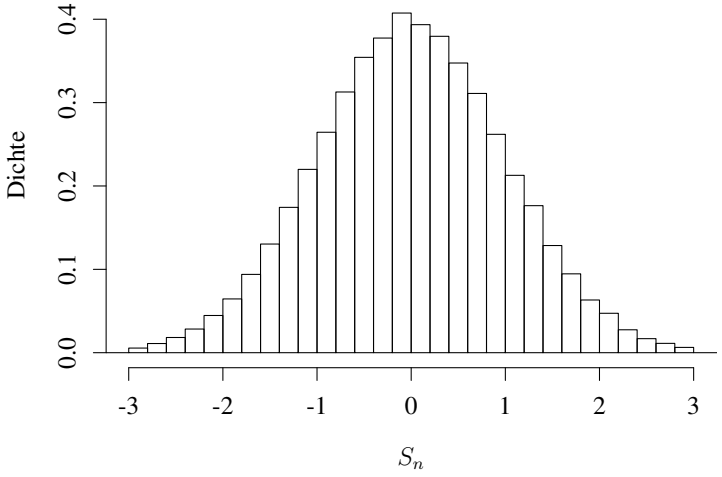
$n = 20$


Dichte
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
S_n

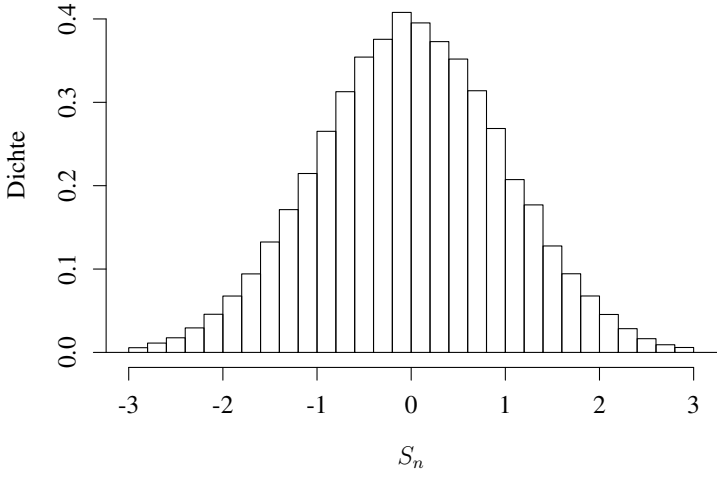
n = 100

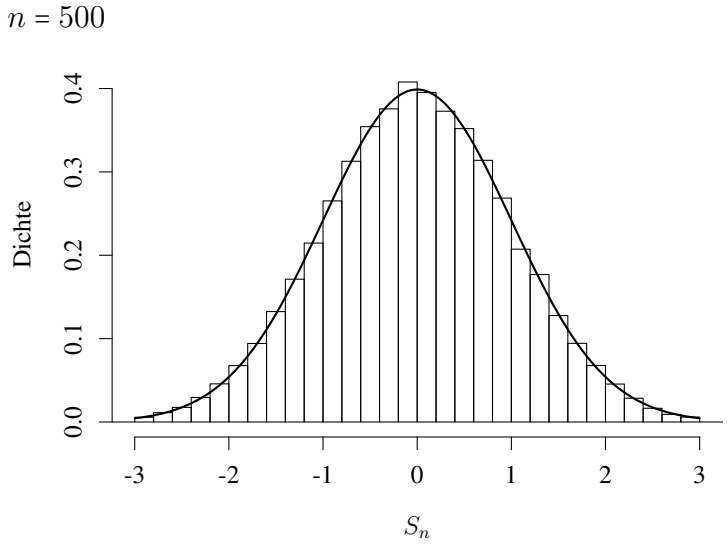


n = 200



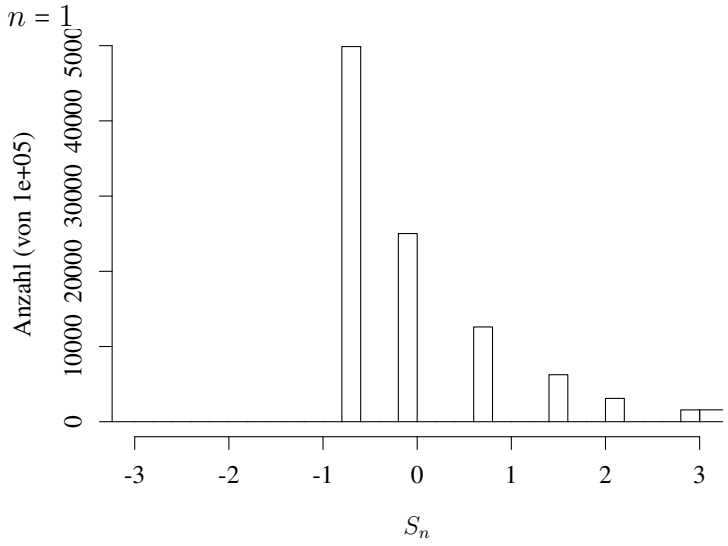
n = 500

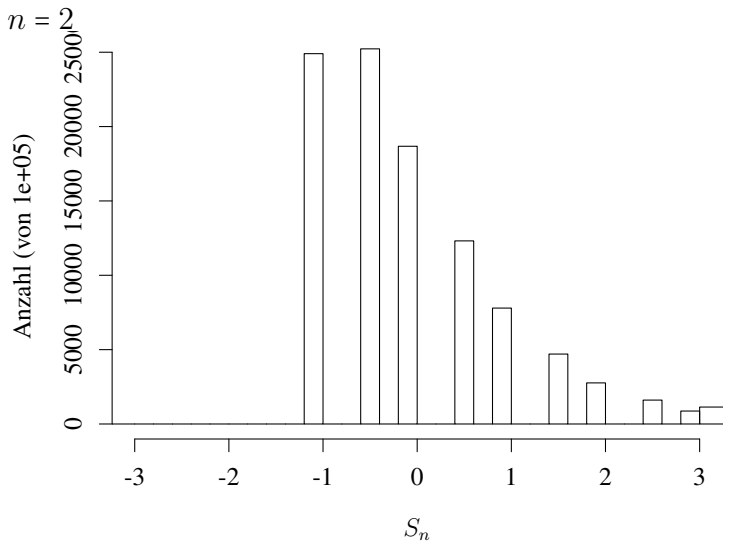




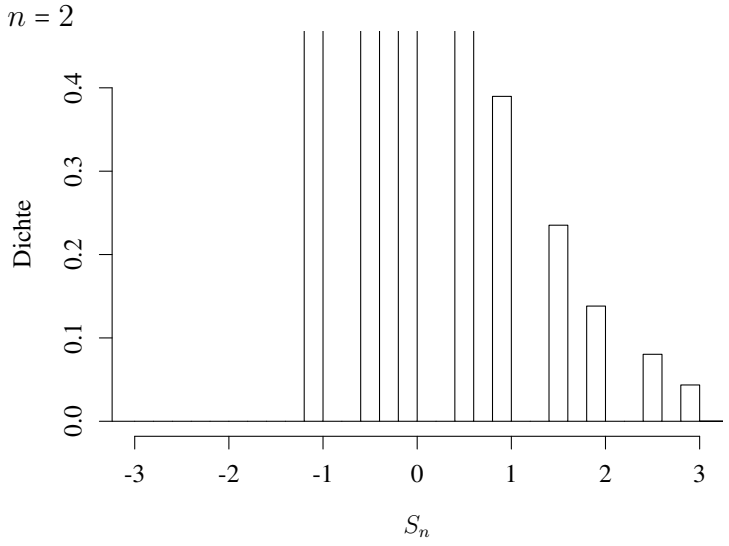
(Die schwarze Kurve ist die Dichte der Standard-Normalverteilung, $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$.)

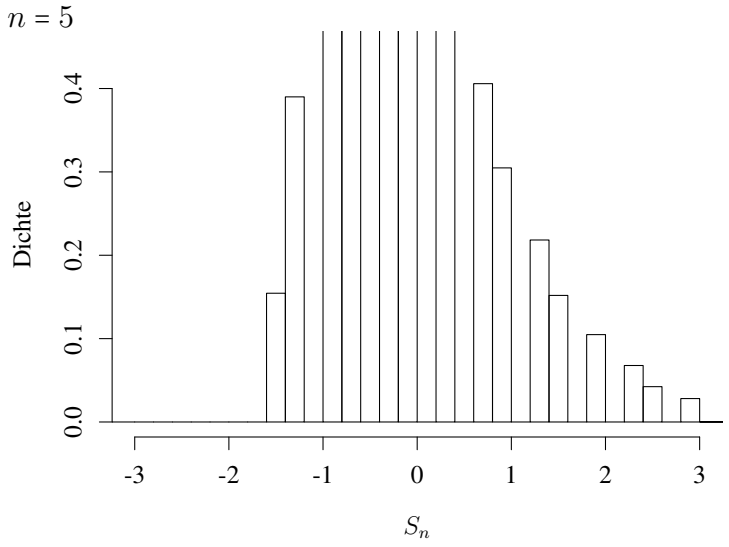
Nun dasselbe nochmal mit $X_i \sim \mathrm{geom}_{1/2}$ (mit $\mathbb{E}[X_1] = 1$, $\mathrm{Var}[X_1] = 2$):
Histogramme jeweils basierend auf 10^5 Simulationen von S_n

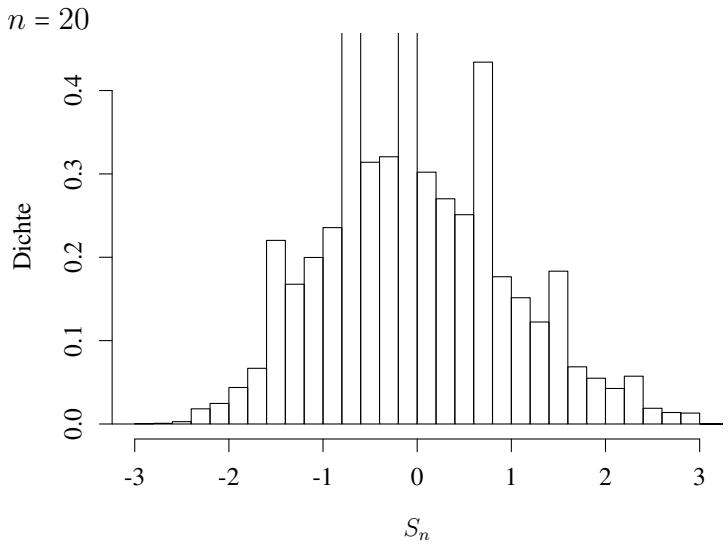


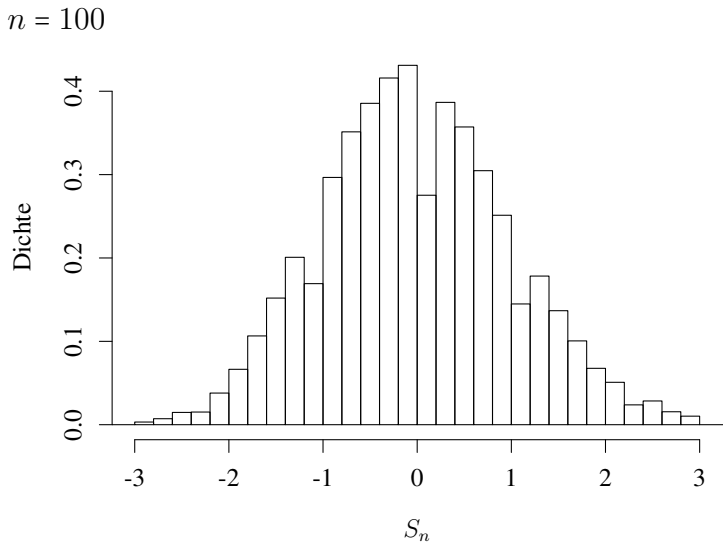


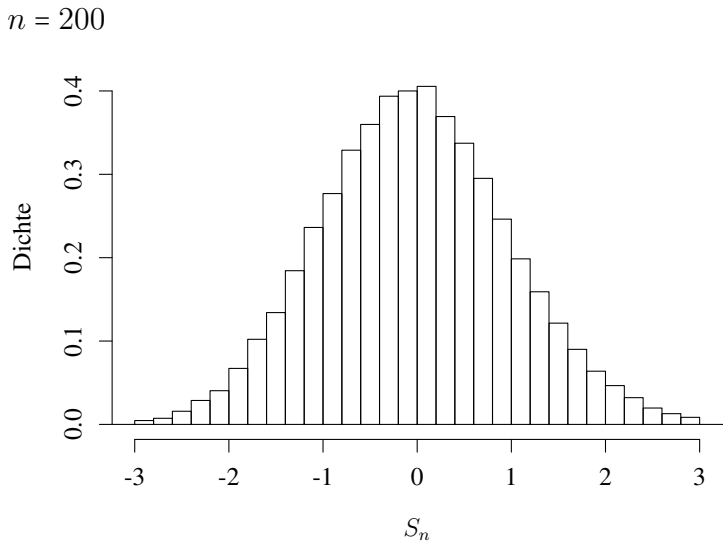
Zur Vergleichbarkeit gehen wir wieder von den absoluten Anzahlen als Balkenhöhen zur sogenannten Dichte über.

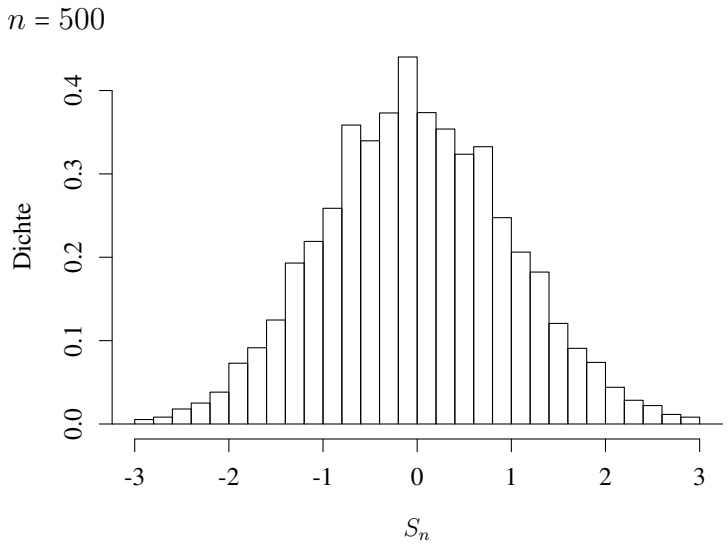


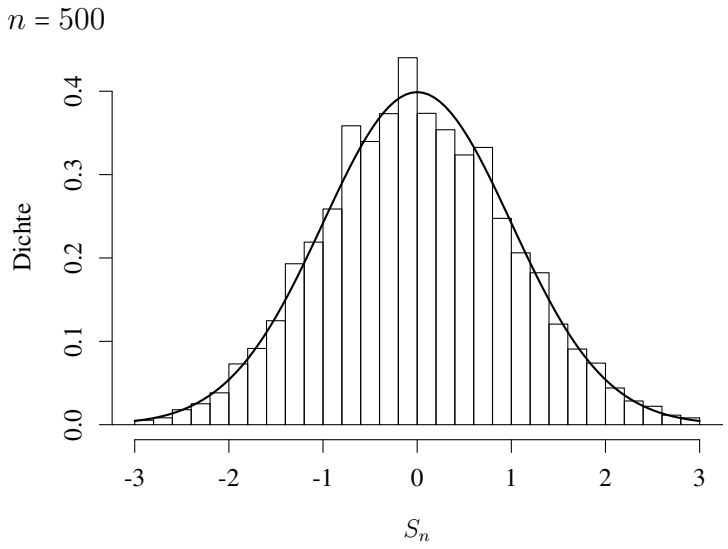



n = 20
Dichte
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
S_n


n = 100
Dichte
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
S_n


n = 200
Dichte
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
S_n





(Die schwarze Kurve ist die Dichte der Standard-Normalverteilung, $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$.)

Wir sehen: Für genügend großes n ist die Verteilung von

$$\frac{(X_1 + \cdots + X_n) - n\mu_{X_1}}{\sqrt{n\sigma^2_{X_1}}} \overset{d}{\approx} Z \quad \text{mit } Z \sim \mathcal{N}_{0,1}. \tag{1.17}$$

Übrigens: Da die Summe unabhängiger, normalverteilter ZVn wieder normalverteilt ist, gilt für $X_i \sim \mathcal{N}_{\mu,\sigma^2}$

$$\frac{(X_1 + \cdots + X_n) - n\mu}{\sqrt{n\sigma^2}} \sim \mathcal{N}_{0,1},$$

d.h. dann gilt (1.17) exakt (dies folgt aus Bsp. 1.59 und Bsp. 1.34, 1.)

Satz 1.85 („Zentraler Grenzwertsatz“). *Seien $X_1, X_2, \dots$ u.i.v. reelle ZVn $\in \mathscr{L}^2$ mit $\mathrm{Var}[X_1] \in (0,\infty)$, dann gilt für $-\infty \le a < b \le \infty$*

$$\lim_{n\to\infty} P\Big(a \le \frac{X_1 + \cdots + X_n - n\mathbb{E}[X_1]}{\sqrt{n\mathrm{Var}[X_1]}} \le b\Big) = P(a \le Z \le b) \quad \textit{mit } Z \sim \mathcal{N}_{0,1}. \tag{1.18}$$

Die wichtige Botschaft von Satz 1.85 lautet: Eine Summe von vielen unabhängigen und identisch verteilten zufälligen Summanden ist (approximativ) normalverteilt.

Bemerkung 1.86. Die Eigenschaft (1.18) wird auch ausgesprochen als „Konvergenz in Verteilung“: X, X_n reellwertige ZVn, so sagt man

$$X_n \underset{n\to\infty}{\longrightarrow} X \text{ in Verteilung} \quad \big(\text{auch } X_n \xrightarrow[n\to\infty]{d} X \text{ oder } X_n \xrightarrow[n\to\infty]{\mathscr{L}} X \text{ geschrieben}\big),$$

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.

Satz 1.85 besagt also: $\dfrac{X_1 + \cdots + X_n - n\mathbb{E}[X_1]}{\sqrt{n\mathrm{Var}[X_1]}} \xrightarrow[n\to\infty]{d} Z$

Bericht 1.87. Es gibt viele Verallgemeinerungen von Satz 1.85, die die Annahme, dass die X_i u.i.v. sind, (stark) abschwächen.

1.4.3 Eine Heuristik zum zentralen Grenzwertsatz

Beweise des zentralen Grenzwertsatzes finden sich in der Lehrbuch-Literatur, z.B. sehr schön in [KW, Kap. III.12], in [Ge, Kap. 5.3] oder in [MP90, Satz 2.3.7]; wir betrachten hier nur ein heuristisches Argument:

„Warum taucht im zentralen Grenzwertsatz die Normalverteilung auf?“

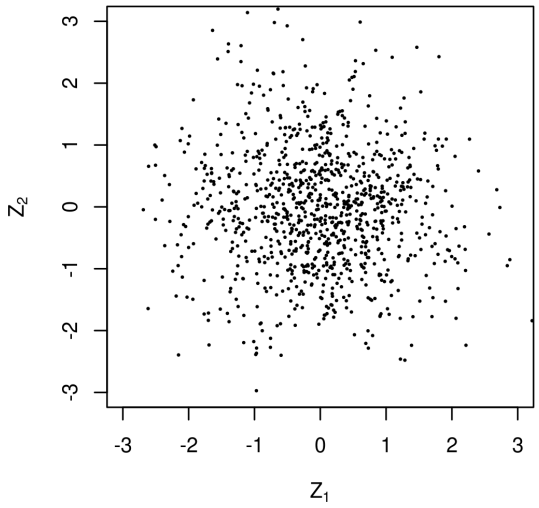
Beobachtung. In der Situation des zentralen Grenzwertsatzes sei

$$(Z_1, Z_2) := \Big(\frac{X_1 + X_2 + \cdots + X_n - n\mu_{X_1}}{\sqrt{n\sigma^2_{X_1}}}, \frac{X_{n+1} + X_{n+2} + \cdots + X_{2n} - n\mu_{X_1}}{\sqrt{n\sigma^2_{X_1}}}\Big),$$

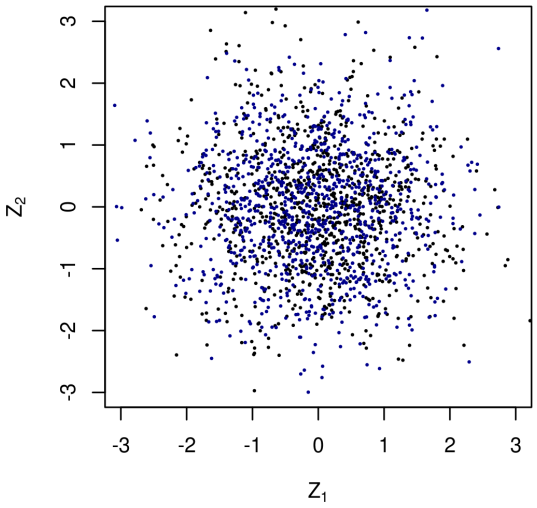
offenbar sind Z_1 und Z_2 unabhängig und identisch verteilt.

Betrachten wir die gemeinsame Verteilung von Z_1 und Z_2 (Simulationen mit $n = 200$, $X_i \sim \mathrm{unif}_{[0,1]}$):

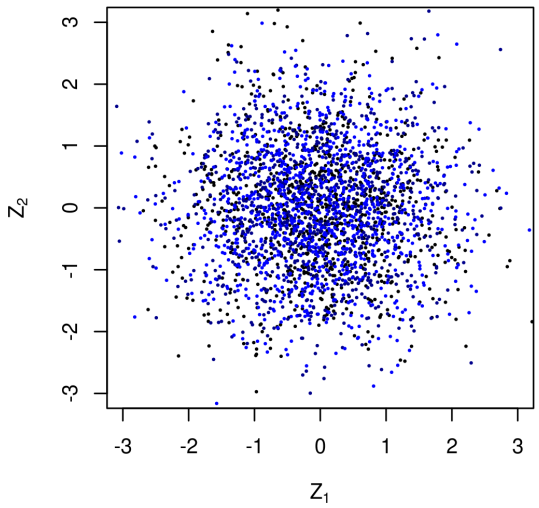
1000 simulierte Werte von (Z_1, Z_2)



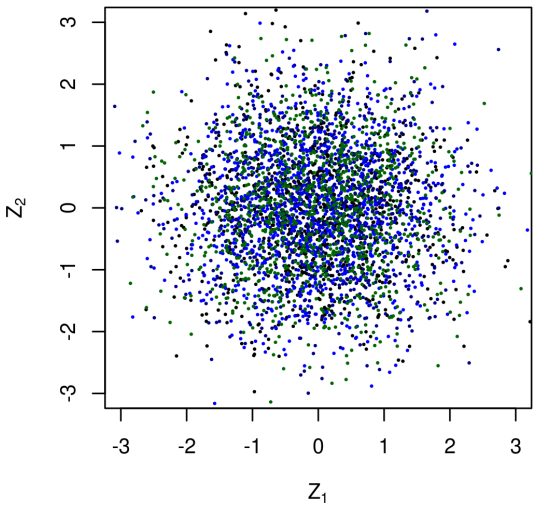
2000 simulierte Werte von (Z_1, Z_2)

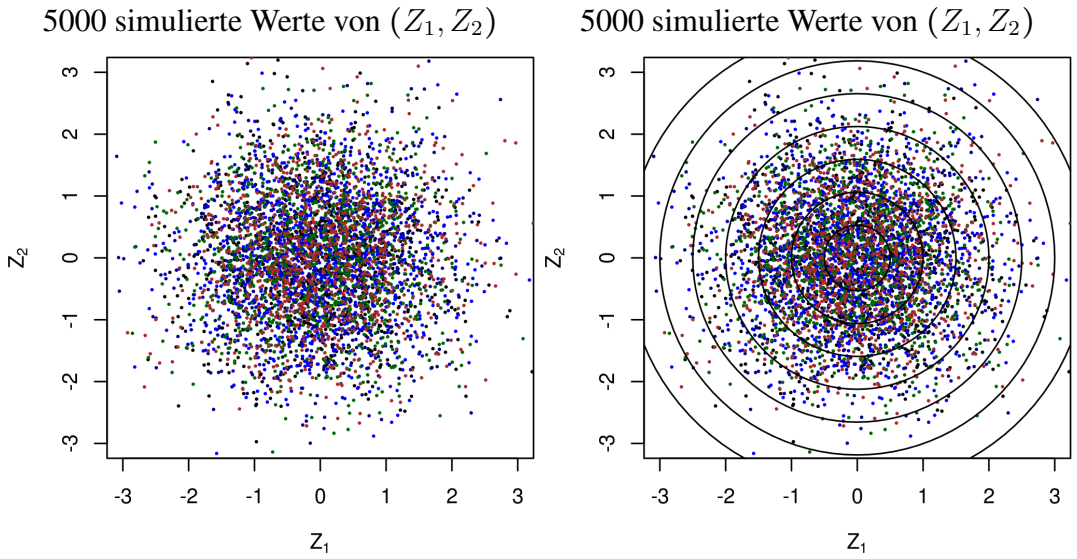


3000 simulierte Werte von (Z_1, Z_2)



4000 simulierte Werte von (Z_1, Z_2)





Die Simulationen legen nahe: (Z_1, Z_2) ist (approximativ) rotationssymmetrisch verteilt.

Beobachtung 1.88 (Eine Charakterisierung der Normalverteilung)**.** *Seien Z_1, Z_2 unabhängige, reellwertige ZVn mit derselben Dichte f, so dass (Z_1, Z_2) rotationssymmetrisch verteilt ist.*

Dann muss f eine (zentrierte) Normaldichte sein, d.h.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{x^2}{2\sigma^2}\Big)$$

für ein $\sigma^2 \in (0, \infty)$.

Beweis. Die (gemeinsame) Dichte $f_{(Z_1,Z_2)}$ ist rotationssymmetrisch, also gilt

$$f_{(Z_1,Z_2)}(z_1, z_2) = g\big(\sqrt{z_1^2 + z_2^2}\big)$$

für eine gewisse Funktion g (vgl. Beob. 1.42), andererseits gilt wegen Unabhängigkeit (vgl. Bericht 1.40)

$$f_{(Z_1,Z_2)}(z_1, z_2) = f(z_1) f(z_2)$$

Mit der Wahl $z_1 = r \geq 0, z_2 = 0$ folgt

$$f(r) f(0) = g\big(\sqrt{r^2}\big) = g(r)$$

(insbesondere muss f symmetrisch sein: $f(-r) = f(r)$).

Somit erfüllt f folgende Funktionalgleichung (setze oben $r = \sqrt{z_1^2 + z_2^2}$):

$$f(z_1) f(z_2) = f(0) f\big(\sqrt{z_1^2 + z_2^2}\big), \quad z_1, z_2 \in \mathbb{R}.$$

(Eine mögliche Lösung ist $f(z) = e^{-z^2}$, denn $e^{-z_1^2} \cdot e^{-z_1^2} = e^{-(z_1^2+z_2^2)} = 1 \cdot e^{-(\sqrt{z_1^2+z_2^2})^2}$.)

Zur allgemeinen Lösung: $w(x) := f\big(\sqrt{|x|}\big)$ erfüllt

$$w(a^2)w(b^2) = w(0)w\big(a^2+b^2\big), \quad a,b \in \mathbb{R}$$

also gilt

$$w(u)w(v) = w(0)w(u+v), \quad u,v \geq 0 \tag{1.19}$$

Die allgemeine Lösung von (1.19) hat die Form

$$w(u) = c_1 e^{-c_2 u}$$

also ist

$$f(z) = w(z^2) = c_1 \cdot \exp(-c_2 z^2)$$

für gewisse $c_1, c_2 > 0$; wegen Normierung $\int_{-\infty}^{\infty} f(x)\,dx = 1$ muss dann $c_1 = (2\pi\sigma^2)^{-1/2}$, $c_2 = 1/(2\sigma^2)$ für ein $\sigma^2 \in (0,\infty)$ gelten. □

1.4.4 Ergänzung: Hoeffding- und McDiarmid-Ungleichung

Seien $X_1, X_2, \ldots, X_n$ u.a., X_i habe Werte in $[a_i, b_i]$ (für gewisse Konstanten $-\infty < a_i < b_i < \infty$), setze

$$S_n := X_1 + \cdots + X_n.$$

Offenbar ist $\mathrm{Var}[X_i] \leq (b_i - a_i)^2$ (denn $|X_i - \mathbb{E}[X_i]| \leq b_i - a_i$) und somit $\mathrm{Var}[S_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] \leq \sum_{i=1}^n (b_i - a_i)^2$, die Chebyshev-Ungleichung (Formel (1.14) aus Satz 1.71) liefert daher für $t > 0$

$$P\big(|S_n - \mathbb{E}[S_n]| \geq t\big) \leq \frac{\sum_{i=1}^n \mathrm{Var}[X_i]}{t^2} \leq \frac{\sum_{i=1}^n (b_i-a_i)^2}{t^2}.$$

Das folgende Resultat stellt oft eine deutliche Verschärfung der Chebyshev-Ungleichung dar (zumindest für beschränkte Summanden):

Bericht 1.89 (Hoeffding-Ungleichung(en)[16])**.** Unter obigen Voraussetzungen gilt für $t \geq 0$

$$P\big(S_n - \mathbb{E}[S_n] \geq t\big) \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\Big), \tag{1.20}$$

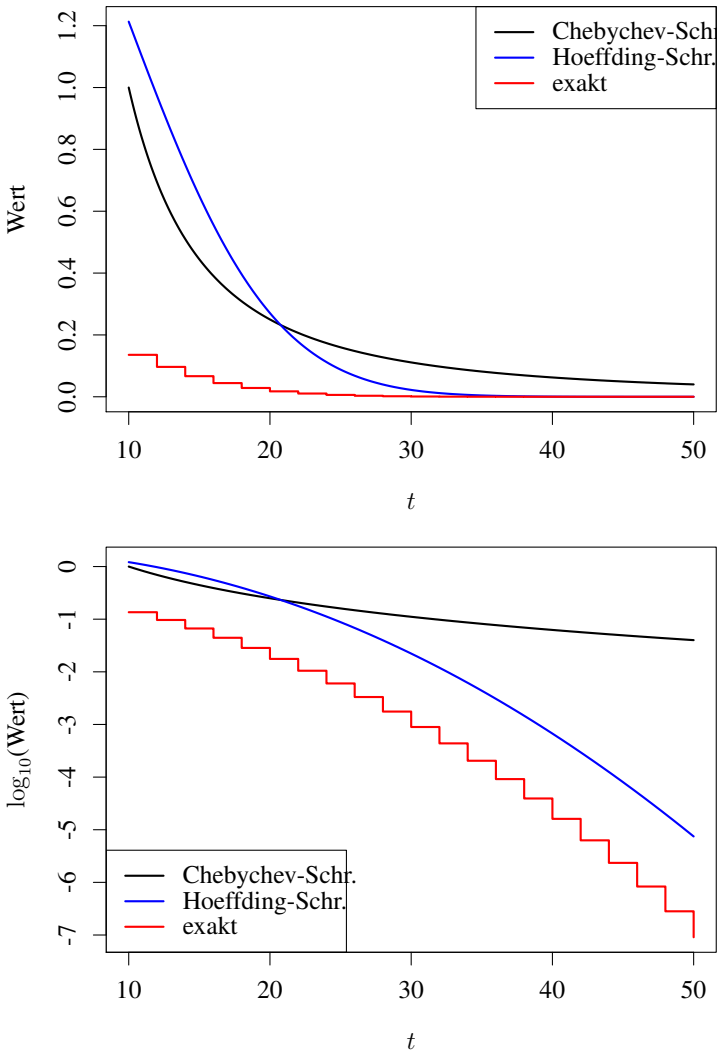
$$P\big(S_n - \mathbb{E}[S_n] \leq -t\big) \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\Big), \tag{1.21}$$

insbesondere

$$P\big(|S_n - \mathbb{E}[S_n]| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\Big). \tag{1.22}$$

Beispiel. Seien X_i u.i.v. mit $P(X_i = +1) = P(X_i = -1) = \frac{1}{2}$, $n = 100$

[16] nach Wassily Hoeffding, 1914–1991

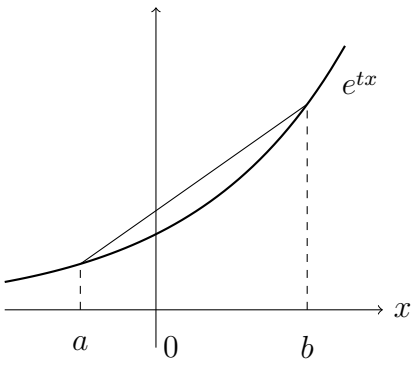


Zum Beweis der Aussagen in Bericht 1.89 verwendet man folgendes Lemma:

Lemma 1.90 (Hoeffdings Lemma)**.** *$a < 0 < b$, X ZV mit Werten in $[a,b]$ und $\mathbb{E}[X] = 0$, dann gilt für $t \in \mathbb{R}$*

$$\mathbb{E}\left[e^{tX}\right] \leq \exp\left(\tfrac{1}{8}t^2(b-a)^2\right)$$

Beweis. $x \mapsto e^{tx}$ ist konvex,



daher gilt für jedes $x \in [a,b]$

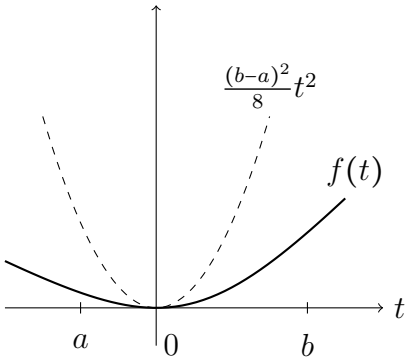
$$e^{tx} \le \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}$$

(beachte $x = \frac{b-x}{b-a}a + \frac{x-a}{b-a}b$) insbesondere

$$\begin{aligned}\mathbb{E}\big[e^{tX}\big] \le \mathbb{E}\Big[\frac{b-X}{b-a}e^{ta} + \frac{X-a}{b-a}e^{tb}\Big] &= \frac{b-\mathbb{E}[X]}{b-a}e^{ta} + \frac{\mathbb{E}[X]-a}{b-a}e^{tb} \\ &= \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb}.\end{aligned}$$

Die Funktion

$$f(t) := \log\Big(\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb}\Big), \quad t \in \mathbb{R}$$



erfüllt

$$f(0) = f'(0) = 0,\ f''(t) \le (b-a)^2/4,$$

(es ist $f'(t) = \dfrac{\frac{ab}{b-a}\big(e^{ta}-e^{tb}\big)}{\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb}} = ab\dfrac{e^{ta}-e^{tb}}{be^{ta}-ae^{tb}}$

$$\begin{aligned}f''(t) &= ab\frac{(ae^{ta}-be^{tb})(be^{ta}-ae^{tb}) - (e^{ta}-e^{tb})(abe^{ta}-abe^{tb})}{(be^{ta}-ae^{tb})^2} \\ &= ab\frac{abe^{2ta} - a^2e^{t(a+b)} - b^2e^{t(a+b)} + abe^{2tb} - abe^{2ta} + abe^{t(a+b)} + abe^{t(a+b)} - abe^{2tb}}{(be^{ta}-ae^{tb})^2} \\ &= \frac{abe^{t(a+b)}\big(-a^2-b^2+2ab\big)}{(be^{ta}-ae^{tb})^2} = -(b-a)^2\frac{1}{(e^{-t(a+b)/2})^2}\frac{ab}{(be^{ta}-ae^{tb})^2} \\ &= (b-a)^2\frac{(-ab)}{(be^{t(a-b)/2} - ae^{t(b-a)/2})^2} \le \frac{1}{4}(b-a)^2\end{aligned}$$

denn

$$\begin{aligned}(be^{t(a-b)/2} - ae^{t(b-a)/2})^2 &= b^2e^{t(a-b)} - 2ab + a^2e^{t(b-a)} \\ &= -4ab + \big(b^2e^{t(a-b)} + 2ab + a^2e^{t(b-a)}\big) = -4ab + \big(be^{t(a-b)/2} + ae^{t(b-a)/2}\big)^2 \\ &\ge -4ab,\end{aligned}$$

beachte: $-a > 0$).

Also ist

$$f(t) = f(0) + \int_0^t f'(s)\,ds = \int_0^t \Big(f'(0) + \int_0^s f''(u)\,du\Big)\,ds \\ \leq \frac{(b-a)^2}{4}\int_0^t\int_0^s du\,ds = \frac{(b-a)^2}{8}t^2$$

und somit

$$\mathbb{E}\big[e^{tX}\big] \leq \exp\big(f(t)\big) \leq \exp\Big(\frac{1}{8}t^2(b-a)^2\Big)$$

wie behauptet. □

Beweis von Bericht 1.89. Für $u > 0$ ist

$$\begin{aligned} P\big(S_n - \mathbb{E}[S_n] \geq t\big) &= P\Big(\exp\big(u(S_n - \mathbb{E}[S_n])\big) \geq e^{ut}\Big) \\ &\leq e^{-ut}\mathbb{E}\Big[\exp\big(u\big(S_n - \mathbb{E}[S_n]\big)\Big] = e^{-ut}\mathbb{E}\Big[\prod_{i=1}^n e^{u(X_i - \mathbb{E}[X_i])}\Big] \\ &= e^{-ut}\prod_{i=1}^n \mathbb{E}\big[e^{u(X_i-\mathbb{E}[X_i])}\big] \leq e^{-ut}\prod_{i=1}^n \exp\big(\tfrac{1}{8}u^2(b_i-a_i)^2\big) \\ &= \exp\Big(-ut + u^2\tfrac{1}{8}\sum_{i=1}^n (b_i-a_i)^2\Big). \end{aligned}$$

Mit der (optimalen) Wahl $u = 4t/\sum_{i=1}^n (b_i - a_i)^2$ folgt (1.20). (1.21) kann analog bewiesen werden (oder ersetze S_n durch $-S_n$ in (1.20)); (1.22) folgt aus (1.20) und (1.21). □

Bericht 1.91 (McDiarmid-Ungleichung)**.** Seien $X_1, X_2, \ldots, X_n$ unabhängige Zufallsvariablen mit Werten in S, $f : S^n \to \mathbb{R}$. Es gebe Konstanten $c_1, \ldots, c_n < \infty$, so dass für $i \in \{1, \ldots, n\}$, $x_1, \ldots, x_n \in S$, $x_i' \in S$ gilt

$$\big|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)\big| \leq c_i. \tag{1.23}$$

Dann gilt für $t \geq 0$

$$P\Big(f(X_1, \ldots, X_n) - \mathbb{E}\big[f(X_1, \ldots, X_n)\big] \geq t\Big) \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big)$$

und

$$P\Big(\big|f(X_1, \ldots, X_n) - \mathbb{E}\big[f(X_1, \ldots, X_n)\big]\big| \geq t\Big) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big).$$

Die Hoeffding-Ungleichung folgt aus der McDiarmid-Ungleichung (mit der Wahl $f(x_1, \ldots, x_n) = x_1 + \cdots + x_n$), letztere ist aber in allgemeineren Situationen anwendbar. Wir werden sie hier nicht beweisen. (Für einen Beweis siehe z.B. Kapitel 6.1 in Stéphane Boucheron, Gábor Lugosi und Pascal Massart, *Concentration inequalities : a nonasymptotic theory of independence*, Oxford University Press, 2013)

Beispiel. Seien $X_1, X_2, \dots, X_n$ u.i.v. $\sim \mathrm{Ber}_p$, $p \in (0,1)$ (d.h. $P(X_i = 1) = p = 1 - P(X_i = 0)$),

$$W := \sum_{i=2}^{n} I_{\{X_i \neq X_{i-1}\}}$$

die Anzahl der „Wechsel" in der Folge $X_1, X_2, \dots, X_n$ (z.B. enthält $(0,0,1,1,0,1,0)$ 4 Wechsel). Beachte: Die Summanden in W sind nicht unabhängig (wir können also nicht die Hoeffding-Ungleichung verwenden).

Es ist $\mathbb{E}[W] = \sum_{i=2}^{n} \mathbb{E}[I_{\{X_i \neq X_{i-1}\}}] = (n-1) \cdot 2p(1-p)$ und wir können schreiben $W = f(X_1, X_2, \dots, X_n)$ mit $f : \{0,1\}^n \to \mathbb{N}_0$,

$$f(x_1, \dots, x_n) = \sum_{i=2}^{n} \mathbf{1}_{\{x_i \neq x_{i-1}\}}.$$

f erfüllt die Voraussetzungen der McDiarmid-Ungleichungen mit $c_1 = 1, c_2 = c_3 = \cdots = c_{n-1} = 2, c_n = 1$, somit gilt

$$P\Big(\big|W - 2p(1-p)(n-1)\big| \geq t\Big) \leq 2\exp\Big(-\frac{2t^2}{4n-6}\Big)$$

Wir sehen: Abweichungen um $t \gg \sqrt{n}$ sind sehr unwahrscheinlich.

Literaturverzeichnis

[KW] G. Kersting und A. Wakolbinger, Elementare Stochastik, 2. Aufl., Birkhäuser, 2010.

[Ge] H.-O. Georgii, Einführung in die Wahrscheinlichkeitstheorie und Statistik, 5. Aufl., de Gruyter, 2015.

[MP90] R. Mathar, D. Pfeifer, Stochastik für Informatiker, Teubner, 1990.

[N95] P. Naeve, Stochastik für Informatik, Oldenbourg, 1995.

[GT96] M. Greiner, G. Tinhofer, Stochastik für Studienanfänger der Informatik, Hanser, 1996.

[K09] G. Kersting, Random variables –without basic space. p. 13–34 in Trends in stochastic analysis. Festschrift in honour of Heinrich von Weizsäcker. Edited by Jochen Blath, Peter Mörters and Michael Scheutzow. London Math. Soc. Lecture Note Ser., 353, Cambridge Univ. Press, 2009.