# Statistik für Informatiker

## Notizen zu einer Vorlesung an der Johannes-Gutenberg-Universität Mainz, Sommer 2018

Matthias Birkner

Vorläufige Version, 4. Juni 2018

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

# Inhaltsverzeichnis

# Kapitel 0

# Auftakt

## 0.1 Organisatorisches

### Orte, Zeiten, Zulassungskriterien, . . .

- Mo. 16–18h, Raum 03-428

- Homepage: `http://www.staff.uni-mainz.de/birkner/StatInfo18/`
  (dort werden die Folien/Notizen, benutzter R-Code/Datensätze, etc. stehen)

- Übungsbetrieb:

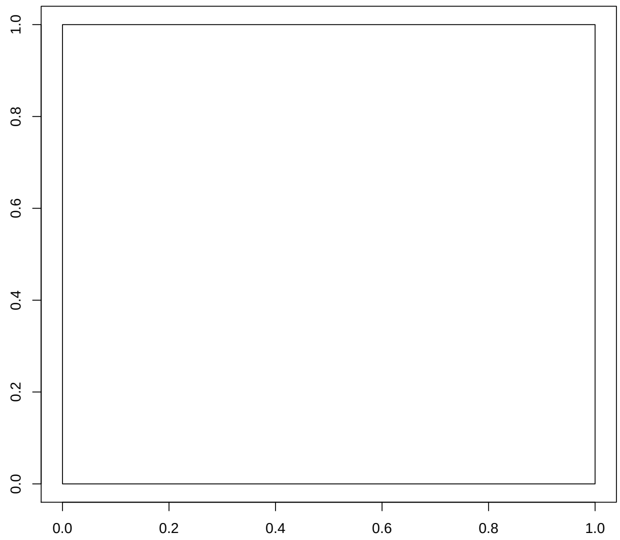
  Organisiert von Timo Schlüter, Informationen unter

  `https://www.stochastik.mathematik.uni-mainz.de/statistik-fuer-informatiker-sose-18/`

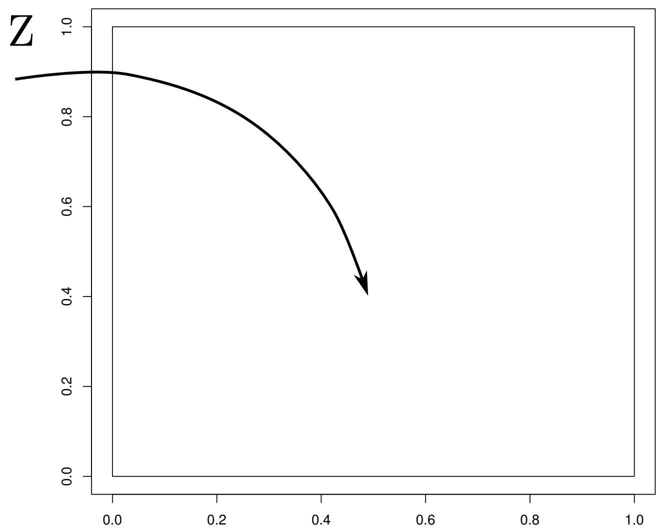
  (verwenden [auch] R `http://www.r-project.org/`, dazu später mehr ...)

## 0.2 Ein Startbeispiel

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




Einheitsquadrat $S$




Einheitsquadrat $S$

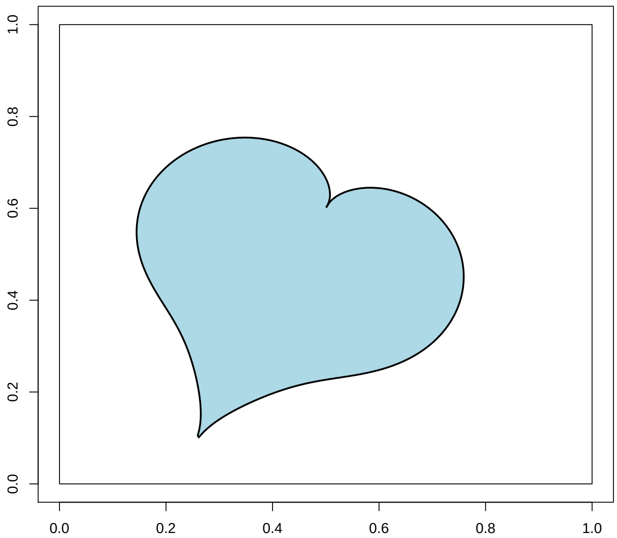
Für die praktische Implementierung stellen wir uns etwa vor, dass $S$ in (sehr kleine) disjunkte Quadrate („Pixel“) zerlegt wird, und dass wir unter allen möglichen Pixeln eines wählen, wobei jedes dieselbe Chance hat, gezogen zu werden. Eine Möglichkeit, dies in `R` zu implementieren, wäre

```
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.




Einheitsquadrat und Teilmenge $B$

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 $p = P(\{Z \in B\})$ = Fläche von $B$. 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 $Z_1, Z_2, \dots, Z_n$ 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

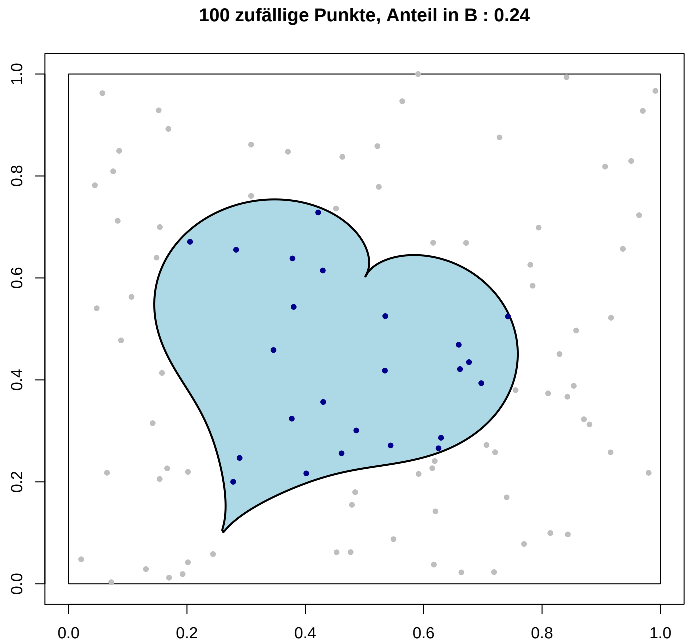
$$\widetilde{M}_n := \sum_{i=1}^n W_i$$

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

$$M_n := \frac{1}{n}\sum_{i=1}^n W_i \quad \Big( = \frac{1}{n}\widetilde{M}_n \Big)$$

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

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 $\varepsilon$ „daneben liegt“.



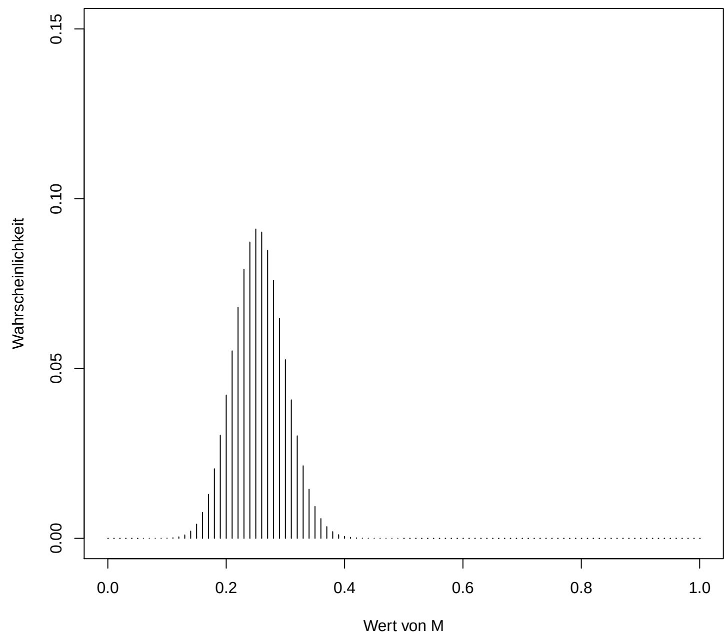

Visualisierung eines solchen Versuchs für $n = 100$.

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

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

(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 $P(\{M_n = \frac{k}{n}\})$):




**Übrigens:** Tatsächlich ist der Wert von $p$ in diesem Beispiel

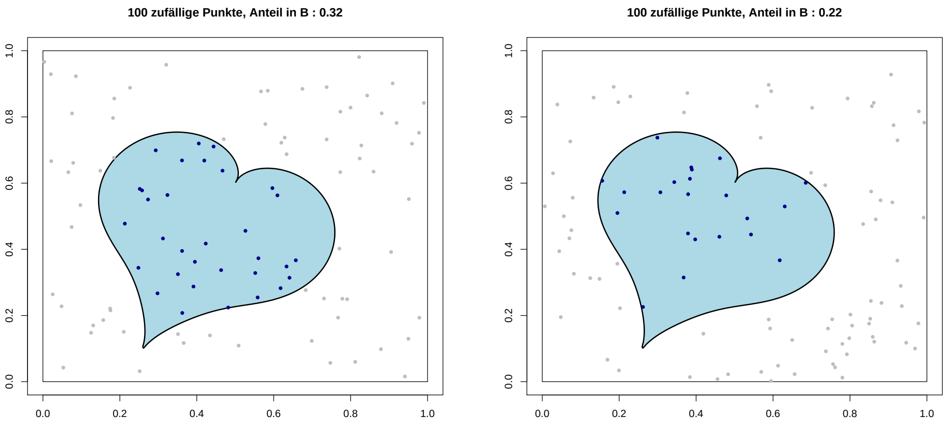
$$p = \int_{-\pi}^{\pi} \frac{1}{98} \left( 2 - 2\sin(w) + \frac{\sin(w)\sqrt{|\cos(w)|}}{\sin(w) + 7/5} \right)^2 dw \approxeq 0.25557221...$$

Denn für

$$B = \left\{(x,y) \in \mathbb{R}^2 : (x^2+y^2)^{1/2} \le f(\arg(\mathrm{x},\mathrm{y}))\right\}$$

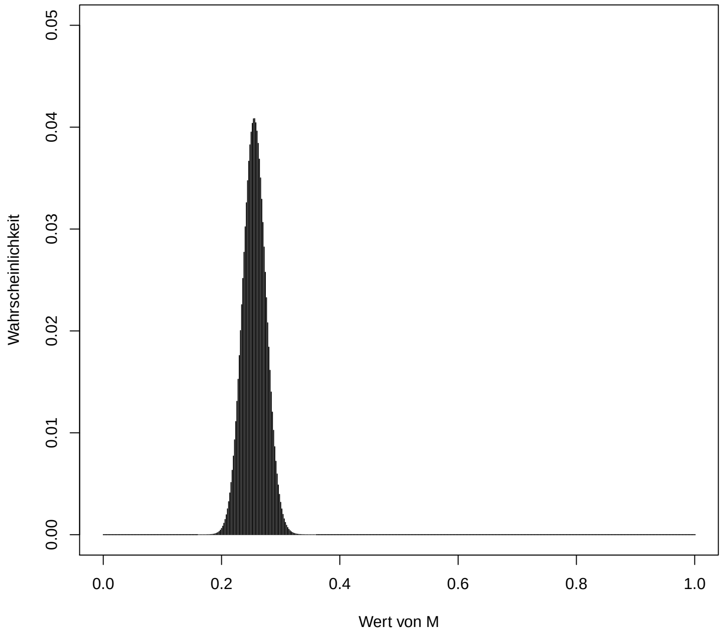
wo $f : [-\pi,\pi] \to \mathbb{R}_+$ und $\arg(x,y) = \mathrm{sign}(x)\arcsin(y/\sqrt{x^2+y^2}) + \pi\mathrm{sign}(y)\mathbf{1}_{(-\infty,0)}(x)$ (dies ist der Winkel, den der Strahl vom Ursprung durch den Punkt $(x,y)$ mit der $x$-Achse einschließt) gilt

$$\text{Fläche von } B = \int_{-\pi}^{\pi} \tfrac{1}{2} f(w)^2 \, dw$$



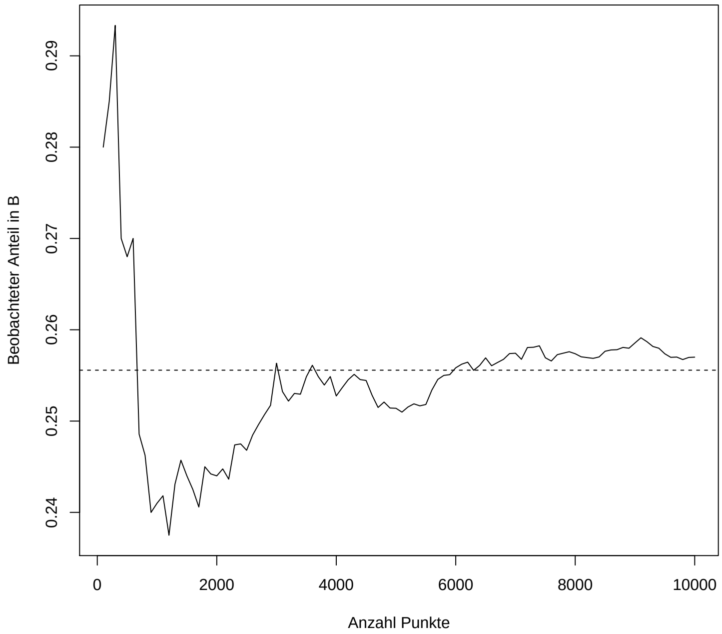

Zur Illustration: Zwei weitere Wiederholungen mit
jeweils $n = 100$ zufälligen Punkten in $[0, 1]^2$

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.

## 0.3 Erste Schritte mit `R`

### Was ist `R`?

- `R` ist eine für die Statistik und für stochastische Simulation entwickelte Programmiersprache, zudem sind viele statistische Standardverfahren bereits in `R` implementiert oder als Zusatzpaket verfügbar.

- `R` hat eine sehr aktive Benutzer- und Entwicklergemeinde, die nahezu alle Bereiche der Statistik und viele Anwendungsbereiche (z.B. Populationsgenetik, Finanzmathematik) überdeckt.

- `R` ist frei verfügbar unter der GNU general public license, für (nahezu) alle Rechnerarchitekturen erhältlich:
  `http://www.r-project.org/`

### Literatur zu `R`

„Standardreferenz“:

W.N. Venables et al, *An Introduction to R*,
`http://cran.r-project.org/manuals.html`

Zur ersten Einführung auch die Hinweise zu R, „nullte Schritte“ auf der Vorlesungs-Homepage.

Weiterführend (z.T. wesentlich über den Stoff der Vorlesung hinaus):

- Günther Sawitzki, *Einführung in R*, `http://sintro.r-forge.r-project.org/`
- William N. Venables, Brian D. Ripley, *Modern applied statistics with S* („Standardwerk“, UB Lehrbuchsammlung)
- Lothar Sachs and Jürgen Hedderich, *Angewandte Statistik – Methodensammlung mit R* (E-Book, UB)
- Christine Duller, *Einführung in die nichtparametrische Statistik mit SAS und R : ein anwendungsorientiertes Lehr- und Arbeitsbuch* (E-Book, UB)
- Helge Toutenburg, Christian Heumann, *Deskriptive Statistik : Eine Einführung in Methoden und Anwendungen mit R und SPSS* (E-Book, UB)
- Uwe Ligges, *Programmieren mit R* (E-Book, UB)

## `R` installieren, starten, anhalten

Installation: Windows, Mac OS: Binaries von `http://www.r-project.org/` (siehe Link `Download, Packages`, CRAN dort)

Linux: Für die meisten Distributionen gibt es fertige Pakete (entweder bei CRAN oder im Repository der Distribution)

Bei Fragen oder Problemen: Übungsgruppen

R starten: Windows, Mac OS: Icon (ggf. aus Menu) anklicken, Linux/Unix: `> R` auf einer Konsole (oder z.B. mit ESS aus Emacs heraus).

R beenden: q() (fragt, ob Daten gespeichert werden sollen)
laufende Rechnungen unterbrechen: CTRL-C

Hinweis: `RStudio` `https://www.rstudio.com/` bietet eine integrierte Entwicklungsumgebung für R [für Privatanwender frei verfügbar].

## Einige einfache (und wichtige) `R`-Befehle

Mathematische Operatoren und Funktionen:
`+, -, *, /, ^, exp, sin, log,` etc.

Hilfe aufrufen:
`help(`*`Befehl`*`)` oder `?`*`Befehl`*

Online-Hilfe starten:
`help.start()`

Variable einen Wert zuweisen:

```
x <- 5
y <- "Zeichenkette"
z <- TRUE
w <- -2+7.5i
```

Vektor erzeugen, Elementzugriff:
`v <- c(1,2,3.1415,-17); v[3]`

Rezyklierungs-Regel bei Vektoren:
`v+2` liefert `3,4,5.1415,-15`

Liste erzeugen, Elementzugriff:
`l <- list(1.2,"text",-5,FALSE); l[[2]]`

(Pseudo-)Zufallszahl generieren:
`runif`

Demos starten (und bestaunen):
`demo()`

Einlesen eines Skripts:
`source`

Ausgabe in Datei umlenken:
`sink`

Grafikausgaben:
`plot,...`

Grafik„geräte“ öffnen:
`x11, postscript, pdf, dev.copy2eps, dev.copy2pdf, ...`

Grafikparameter setzen: `par`

Siehe auch die Befehle und Hinweise in `Auftakt.R`, das wir in der ersten Sitzung diskutieren.

# Kapitel 1

# Grundlagen aus der Wahrscheinlichkeitstheorie

## 1.1 Ereignisse, Zufallsvariablen und Wahrscheinlichkeiten

**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.

**Worum geht es?** In vielen Situationen tritt „Zufall“ oder „Ungewissheit“ auf, in diesem Kapitel geht es uns darum, den üblichen mathematischen Rahmen kennen zu lernen, um solche Phänomene zu beschreiben.

**Beispiele.**

- Im Auftakt-Beispiel in Kapitel 0 haben wir ein zufälliges Pixel $Z$ aus (einer diskretisierten Version von) $[0,1]^2$ gewählt.
- „Glücksspiel-Situationen“: Wir werfen eine Münze oder einen Würfel, spielen Lotto, schauen ein gut gemischtes Pokerblatt an, ...
- Wir befragen einen zufällig ausgewählten Studenten in der Mensa nach seinem Studiengang und danach, wieviel er heute in der Mensa ausgegeben hat.
- Wir übergeben $n$ Zahlen in rein zufälliger Reihenfolge an einen Sortieralgorithmus;

  wir überprüfen zwei Polynome $f(x)$ und $g(x)$ vom Grad $d \gg 1$ auf Gleichheit, indem wir an $k$ zufällig ausgewählten Stellen auf Gleichheit testen;

  wir betrachten einen Router, der Datenpakete an einen zufällig ausgewählten Nachbarknoten im Netzwerk schickt; ...

**Mathematische Modellierung des Zufalls** („Zutaten“)**:** Sei $\Omega$ (nicht-leere) Menge („Ergebnisraum“ oder „Stichprobenraum“ genannt), und sei $\mathcal{F} \subset 2^\Omega$ ($:= \{B : B \subset \Omega\}$).

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

$A \in \mathcal{F}$ nennen wir ein „Ereignis“,

insbesondere $\Omega \dots$ „sicheres Ereignis“, $\varnothing \dots$ „unmögliches Ereignis“

Vorstellung/Interpretation:

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

**Operationen** (Mengenoperationen und ihre Interpretation für Ereignisse)**:** Für $A, B \in \mathcal{F}$

| | | |
|---|---|---|
| $A^c := \Omega \setminus A$ | $\dots$ | „$A$ tritt nicht ein“<br>($A^c$ heißt Gegen- oder Komplementärereignis von $A$) |
| $A \cap B$ | $\dots$ | „$A$ und $B$ treten ein“ |
| $A \cup B$ | $\dots$ | „$A$ oder $B$ treten ein“ |
| $A \subset B$ | $\dots$ | „$A$ impliziert $B$“ (falls $A$ eintritt, so auch $B$) |

$A$ und $B$ heißen *disjunkt*, wenn $A \cap B = \varnothing$.

Wir notieren auch $A \setminus B := A \cap B^c = \{\omega : \omega \in A, \omega \notin B\}$ („$A$ tritt ein, aber $B$ nicht“).

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

$$\begin{aligned} &i) \quad \varnothing \in \mathcal{F}, \\ &ii) \quad A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}, \\ &iii) \quad A_1, A_2, \dots \in \mathcal{F} \Rightarrow \cup_{n=1}^\infty A_n \in \mathcal{F}. \end{aligned} \tag{1.1}$$

$\mathcal{F} \subset 2^\Omega$, das (1.1), *i)–iii)* genügt, heißt eine *$\sigma$-Algebra* (über $\Omega$), die de Morgansche Regeln ergeben dann auch $\Omega = \varnothing^c \in \mathcal{F}$, $\cap_{n=1}^\infty A_n = \left(\cup_{n=1}^\infty A_n^c\right)^c \in \mathcal{F}$.

Eine Abbildung $P : \mathcal{F} \to [0,1]$ mit

$$\begin{aligned} &\text{(N)} \quad P(\Omega) = 1 \quad (\text{„Normierung“}) \quad \text{und} \\ &\text{(A)} \quad A_1, A_2, \dots \in \mathcal{F} \text{ paarweise disjunkt} \\ &\qquad \Rightarrow P\Big(\bigcup_{n=1}^\infty A_n\Big) = \sum_{n=1}^\infty P(A_n) \qquad (\text{„}\sigma\text{-Additivität“}) \end{aligned} \tag{1.2}$$

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

$P(A)$ heißt / nennen wir die *Wahrscheinlichkeit des Ereignisses $A$*.

**Definition 1.1.** Ein Tripel $(\Omega, \mathcal{F}, P)$, wobei $\mathcal{F} \subset 2^\Omega$ die Forderungen (1.1), *i)–iii)* und $P$ die Forderungen (1.2), (N) und (A) erfüllt, heißt ein *Wahrscheinlichkeitsraum*.

**Beispiel und Definition 1.2** (Diskreter Wahrscheinlichkeitsraum)**.** $\Omega$ endlich oder abzählbar, $p : \Omega \to [0,1]$ Abbildung mit $\sum_{\omega \in \Omega} p(\omega) = 1$.

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

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

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

Dies erfüllt „offensichtlich" Forderung (1.2).

**Beispiele 1.3.** 1. (uniforme Wahl aus endlicher Menge, „Laplace-Experimente") $\#\Omega < \infty$ und $p(\omega) = \frac{1}{\#\Omega}$ in Bsp. 1.2, z.B.

(a) (Wurf eines fairen 6er-Würfels) $\Omega = \{1,2,3,4,5,6\}$, $p(\omega) = 1/6$ für $\omega \in \Omega$

(b) (dreimaliger Wurf eines fairen 6er-Würfels) $\Omega = \{(\omega_1,\omega_2,\omega_3) : \omega_i \in \{1,2,3,4,5,6\}$ für $i = 1,2,3\}$, $p((\omega_1,\omega_2,\omega_3)) = 1/6^3 = 1/216$

(c) (Auftakt-Beispiel aus Kapitel 0 mit Diskretisierung auf 32 Bit Genauigkeit) $\Omega = \{(x,y) : x = k/2^{32}, y = \ell/2^{32}$ mit $k, \ell \in \{0,1,2,\dots,2^{32}-1\}\}$, $p((x,y)) = 1/(2^{32})^2 = 2^{-64}$

(d) ($n$ verschiedene Eingaben in zufälliger Reihenfolge)

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

(„symmetrische Gruppe der Ordnung $n$"), $p((x_1,x_2,\dots,x_n)) = 1/n!$

2. (ein verfälschter Münzwurf) $\Omega = \{\text{Kopf}, \text{Zahl}\}$, $p(\text{Kopf}) = 0.6 = 1 - p(\text{Zahl})$

3. (eine winziges Modell für Spam, das Sprache und Status einer Email betrachtet)

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

$p\big((\mathtt{Deutsch}, \mathtt{Spam})\big) = 0.2$, $p\big((\mathtt{Deutsch}, \mathtt{keinSpam})\big) = 0.1$, $p\big((\mathtt{Englisch}, \mathtt{Spam})\big) = 0.6$, $p\big((\mathtt{Englisch}, \mathtt{keinSpam})\big) = 0.1$

4. (Anzahl Würfe, bevor beim wiederholten fairen Münzwurf zum ersten Mal Kopf kommt) $\Omega = \mathbb{N}_0$, $p(n) = (1/2)^n \cdot (1/2) = 2^{-n-1}$ für $n \in \Omega$

**Bemerkung 1.4.** Wahrscheinlichkeitsräume $(\Omega, \mathcal{F}, P)$ gemäß Def. 1.1 werden heute, nicht zuletzt wegen des einflussreichen Werks von A.N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933, oft als „Basisobjekt" der mathematischen Modellierung von Zufallsvorgängen verwendet. Insoweit ist die explizite Wahl eines geeigneten Wahrscheinlichkeitsraums (oft) ein Teil der Arbeit, wenn ein umgangssprachlich formuliertes Problem in Mathematik übersetzt werden soll.

Man sollte sich bewusst sein, dass es dabei i.A. viele mögliche Wahlen gibt und dass die Formulierung mit Zufallsvariablen oftmals einen für die Intuition und für das Argumentieren sehr angenehmen Zugang ergibt (siehe Abschnitt 1.1.1 unten, wo wir sie als Funktionen auf $\Omega$ formal definieren) Siehe auch Abschnitt 1.2 für Hintergrund und weitere Diskussion.

**Lemma 1.5.** *Für $A, B, A_1, A_2, \dots \in \mathcal{F}$ gilt*

$$P(\varnothing) = 0 \tag{1.3}$$

$$P(A \cup B) + P(A \cap B) = P(A) + P(B) \qquad \textit{(endliche Additivität)}, \tag{1.4}$$

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

$$A \subset B \Rightarrow\ P(A) \le P(B) \qquad \textit{(Monotonie)} \tag{1.5}$$

*zudem gilt auch*

$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) \le \sum_{n=1}^{\infty} P(A_n) \qquad \textit{($\sigma$-Subadditivität)} \tag{1.6}$$

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

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

$$\textit{so gilt } P(A) = \lim_{n\to\infty} P(A_n) \qquad \textit{($\sigma$-Stetigkeit)} \tag{1.7}$$

*Beweis.* (1.3): $P(\varnothing) = P\big(\varnothing \cup \varnothing \cup \dots\big) = \sum_{n=1}^{\infty} P(\varnothing)$, also $P(\varnothing) = 0$.

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

$P(A\cup B) = P\big(A\cup B\cup\varnothing\cup\varnothing\cup\dots\big) = P(A)+P(B)+\underbrace{P(\varnothing)}_{=0}+\underbrace{P(\varnothing)}_{=0}+\cdots$, d.h. (1.4) gilt hier.

Allgemeiner Fall: Schreibe

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

also

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

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

(1.6): 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 (1.5)).

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

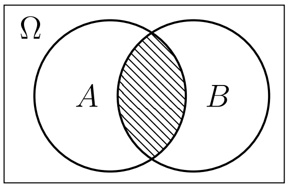
$$P(A) = P\Big(\bigcup_{i=1}^{\infty}(A_i \setminus A_{i-1})\Big) = \sum_{i=1}^{\infty} P(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \underbrace{\sum_{i=1}^{n} P(A_i \setminus A_{i-1})}_{=P(A_n)}.$$

Falls $A_n \searrow A$, so beachte, dass $A_n^c \nearrow A$ gilt, dann verwende obiges zusammen mit (1.4). $\square$

**Bemerkung 1.6** (Einschluss-Ausschluss-Formel)**.** Formel (1.4) aus Lemma 1.5 kann man auch folgendermaßen schreiben

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

Man kann diese Formel anhand eines Venn-Diagramms veranschaulichen:




Der schraffierte Bereich $A \cap B$ wird in $P(A) + P(B)$ doppelt gezählt, in $P(A \cap B)$ aber nur einmal.

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

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

d.h. knapp ausgedrückt

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{\emptyset \neq J \subset \{1,\dots,n\}} (-1)^{(\#J)-1} P\Big(\bigcap_{i \in J} A_i\Big) \tag{1.8}$$

Man kann (1.8) per Induktion beweisen, oder siehe Abschnitt 1.2.3 für einen alternativen Ansatz.

**Beispiel.** Betrachten wir in Bsp. 1.3, 1 (d) mit $n = 3$ die Ereignisse $A_i = \{(x_1, x_2, x_3) \in \Omega : x_i = i\}$ für $i = 1, 2, 3$. Wenn wir eine uniform verteilte Permutation von $1, 2, 3$ betrachten, so ist $A_i = \{i \text{ ist ein Fixpunkt}\}$. Dann ist

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

die Wahrscheinlichkeit, dass es mindestens einen Fixpunkt gibt.

**Bemerkung 1.7.** Wir lassen hier die „philosophische" Frage offen, welche Interpretation die Wahrscheinlichkeit $P(A)$ eines Ereignisses $A$ haben soll. Es bieten sich etwa an:

- „naive" Interpretation der Wahrscheinlichkeit: Die „Natur" enthält inhärente Unbestimmtheit („ist selbst nicht sicher, was sie tut"), und $P(A)$ beschreibt den Grad der Sicherheit, mit dem sie sich für das Ereignis $A$ entscheidet.
- „frequentistische" Interpretation der Wahrscheinlichkeit: Wenn man das zufällige Experiment unter exakt denselben Bedingungen sehr oft wiederholte, wäre der relative Anteil der Ausgänge, in denen $A$ eingetreten ist, etwa $P(A)$.

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

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

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

### 1.1.1 Zufallsvariablen

Oft stellt man sich zumindest intuitiv Zufallsvorgänge so vor, dass ein „zufälliger Wert" beobachtet oder gemessen wurde, mit dem Verständnis, dass sich bei Wiederholung des Experiments möglicherweise ein anderer Wert aus einer gewissen Menge möglicher Werte ergäbe.

So kann man die Beispiele 1.3 auch folgendermaßen aussprechen:

1. (a) Sei $W$ das Ergebnis eines Wurfs eines fairen 6er-Würfels (Wertebereich $\{1,2,3,4,5,6\}$).
   (b) Werfe Würfel dreimal, seien $W_1, W_2, W_3$ Ergebnisse des 1., 2., 3. Wurfs des Würfels ($W = (W_1, W_2, W_3)$ hat Wertebereich $\{1,2,3,4,5,6\}^3$).
   (c) $Z = (X,Y)$ ein zufällig gewähltes Pixel aus $[0,1]$ (aus Auftakt-Beispiel aus Kapitel 0 mit Diskretisierung auf 32 Bit Genauigkeit)
   (d) $n$ verschiedene Eingaben in zufälliger Reihenfolge: Sei $X = (X_1, X_2, \dots, X_n)$ eine zufällige Permutation von $1,2,\dots,n$ (Wertebereich $\{(x_1,x_2,\dots,x_n) : x_1,\dots,x_n \in \{1,2,\dots,n\}$ paarweise verschieden$\}$)
2. Sei $M$ das Ergebnis eines (möglicherweise verfälschten) Münzwurfs (Wertebereich $\{\mathrm{Kopf}, \mathrm{Zahl}\}$)
3. Wähle uniform eine Email aus meiner Inbox (die deutsche und englische Emails enthält, die jeweils Spam sein können oder nicht), sei $X$ die Sprache und $Y$ der Spam-Status der betrachteten Email ($(X,Y)$ hat Wertebereich $\{\texttt{Deutsch}, \texttt{Englisch}\} \times \{\texttt{Spam}, \texttt{keinSpam}\}$)
4. Wir werfen eine faire Münze, bis zum ersten Mal $\mathrm{Kopf}$ kommt. $G$ zählt, wie oft bis dahin $\mathrm{Zahl}$ gefallen ist.

**Definition 1.8.** $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum, $S$ eine Menge, eine Abbildung[1] $X : \Omega \to S$ nennen wir eine *Zufallsvariable*, oft abgekürzt ZV (auch: *Zufallsgröße*) mit Wertebereich $S$.

ZVn und Ereignisse: Für $A \subset S$ schreibt man[2]

$$\{X \in A\} := \{\omega \in \Omega : X(\omega) \in A\} = X^{-1}(A)$$

[1] Im Fall, dass $S$ überabzählbar ist, muss man hier strenggenommen fordern, dass es sich um eine sogenannte *messbare* Abbildung handelt. Für die Zwecke dieser Vorlesung ist es ungefährlich, davon auszugehen, dass uns nur solche Abbildung begegnen werden, wir verzichten daher hier stillschweigend auf die Präzisierung. Siehe auch die Diskussion in Abschnitt 1.2.1.

[2] Wiederum darf man strenggenommen im Fall, dass $S$ überabzählbar ist, hier nur sogenannte *messbare* Teilmengen $A \subset S$ betrachten. Wir werden im Fall $S = \mathbb{R}$ oder $S = \mathbb{R}^n$ nur solche Mengen betrachten und verzichten daher auch hier stillschweigend auf die Präzisierung.

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

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

**Beispiel 1.9** (Indikatorvariable)**.** $A \in \mathcal{F}$, $\mathbf{1}_A : \Omega \to \{0,1\}$,

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

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

**Definition 1.10.** Für eine Zufallsvariable $X$ mit Wertebereich $S$ heißt[3]

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

die *Verteilung* von $X$.

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

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

heißen dann 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 Forderung (1.2)): $\rho_X(a) \ge 0$,

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

und $\rho_X$ ist ein *Wahrscheinlichkeitsmaß* (auch *Wahrscheinlichkeitsverteilung*) auf $S$ (d.h. $\rho_X : \{\text{Teilmengen von } S\} \to [0,1]$ und die (analogen) Eigenschaften aus Forderung (1.2) und aus Lemma 1.5 gelten, mit Lesung $\Omega = S$).

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).

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

**Beobachtung 1.11** (Operationen mit Zufallsvariablen)**.** Zufallsvariablen sind nicht zuletzt deshalb nützlich für die Modellierung zufälliger Vorgänge, weil man mit ihnen gewissermaßen operieren und „rechnen“ kann wie mit den Variablen in einer Programmiersprache:

[3] ☡ Siehe Fußnote 2.

Seien $X$ und $Y$ mit Wertebereichen $S_X$ bzw. $S_Y$ ZVn auf *demselben Wahrscheinlichkeitsraum*, $f : S_X \to S$ eine Abbildung, so können wir die $S$-wertige ZV $f(X) := f \circ X$ bilden und die $S_X \times S_Y$-wertige ZV $(X, Y)$.

Speziell können wir im Fall $S_X, S_Y \subset \mathbb{R}$ die ZVn $X^2$, $X + Y$, $X - Y$, $XY$, etc. bilden.

Mit $f$ wie oben ergibt sich (im diskreten Fall) für die Gewichte von $f(X)$ dann

$$\begin{aligned} \rho_{f(X)}(s) &= P\big(f(X) = s\big) = P\big(X \in \{x \in S_X : f(x) = s\}\big) \\ &= \sum_{x : f(x)=s} P(X = x) = \sum_{x : f(x)=s} \rho_X(x) \qquad \text{für } s \in S \end{aligned}$$

**Beispiel** (Verteilung der Augensumme zweier fairer Würfel)**.** Sei $(W_1, W_2)$ das Ergebnis des 1. und 2. Wurf eines fairen Würfels, $f : \{1, 2, \dots, 6\}^2 \to \{2, 3, \dots, 12\}, (w_1, w_2) \mapsto w_1 + w_2$, so ist

$$\begin{aligned} P(W_1 + W_2 = s) = P(f(W_1, W_2) = s) &= \sum_{(w_1,w_2) : w_1+w_2=s} P(W_1 = w_1, W_2 = w2) \\ &= \frac{\#\{(w_1, w_2) \in \{1, 2, \dots, 6\}^2 : w_1 + w_2 = s\}}{36} = \frac{6 - |7 - s|}{36}, \quad s \in \{2, 3, \dots, 12\} \end{aligned}$$

**Bemerkung 1.12** (Kanonisches Modell für eine ZV)**.** Man kann eine Zufallsvariable $X$ mit Wertebereich $S$ und Verteilung $\rho$ stets in „kanonischer Weise" mit der Wahl $\Omega = S$ und $P = \rho$ und geeignetem $\mathcal{F} \subset 2^S$ (im diskreten Fall kann man $\mathcal{F} = 2^S$ wählen) als $X = \mathrm{Id}_S$ auf dem W'raum $(S, \mathcal{F}, \rho)$ formulieren.

Man kann daher genausogut die mathematische Modellierung eines Zufallsphänomens mit der Formulierung geeigneter Zufallsvariablen samt Verteilung beginnen, siehe auch Diskussion und Referenzen in Abschnitt 1.2.2

### 1.1.2 Einige „klassische" diskrete Verteilungen

**Beispiel 1.13** (Ziehen mit und ohne Zurücklegen: Urnenmodelle)**.** Eine Urne enthalte $n$ (mit $1, 2, \dots, n$ nummerierte) Kugeln, wir ziehen zufällig $k$ $(\le n)$ heraus. Sei

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

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

Wir betrachten 4 mögliche Situationen:

1. Ziehen mit Zurücklegen, mit Beachtung der Reihenfolge:

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

   (d.h. $X$ ist uniform auf $\{1, \dots, n\}^k$ verteilt).

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

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

(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

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

Man nennt $Y$ auch einen (zufälligen) „Besetzungsvektor“. 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

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

Begründung: Wir müssen $\ell_1$ 1-er, $\ell_2$ 2-er, ..., $\ell_n$ $n$-er auf $k$ (angeordnete) Plätze verteilen, dazu gibt es

$$\underbrace{\frac{k\cdot(k-1)\cdots(k-\ell_1-1)}{\ell_1!}}_{\text{Anz. Mögl'keiten, die 1-er zu platzieren}}\cdot\underbrace{\frac{(k-\ell_1)\cdot(k-\ell_1-1)\cdots(k-\ell_1-\ell_2-1)}{\ell_2!}}_{\text{Anz. Mögl'keiten, die 2-er zu platzieren}}\cdot\cdots$$

$$\cdots\cdot\underbrace{\frac{(k-\ell_1-\cdots-\ell_{n-1})\cdot(k-\ell_1-\cdots-\ell_{n-1}-1)\cdots(k-\ell_1-\ell_{n-1}-\ell_n+1)}{\ell_n!}}_{\text{Anz. Mögl'keiten, die } n\text{-er zu platzieren}}$$

$$=\frac{k!}{\ell_1!\cdot\ell_2!\cdot\cdots\cdot\ell_n!}=\binom{k}{\ell_1,\ell_2,\ldots,\ell_n}$$

Möglichkeiten, somit ist

$$P\big(Y=(\ell_1,\ldots,\ell_n)\big)=\binom{k}{\ell_1,\ell_2,\ldots,\ell_n}\Big(\frac{1}{n}\Big)^k,\quad (\ell_1,\ldots,\ell_n)\in W_4$$

d.h. eine Instanz der Multinomialverteilung, siehe Bsp. 1.18 unten.

**Bemerkung 1.14.** 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:

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

Insbesondere ist die Verteilung auf $W_4$ aus Beispiel 1.13, 4. nicht die uniforme.

Die uniforme Verteilung auf dem $W_4$ aus Beispiel 1.13, 4. heißt auch die „Bose-Einstein-Verteilung", die in Beispiel 1.13, 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 1.15** (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 $\ell$ schwarze Kugeln zu ziehen.

Z.B. ist die W'keit, dass der Geber beim Skat genau 3 Asse bekommt $= \frac{\binom{4}{3}\binom{28}{7}}{\binom{32}{10}} \approx 0{,}073$.

**Beispiel 1.16** ($p$-Münzwurf und Binomialverteilung)**.** 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 $p \in [0,1]$): $S = \{0,1\}^n$,

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

Für $X = (X_1,\dots,X_n) \sim \mathrm{Ber}_p^{\otimes n}$ sagt man: $X_1,\dots,X_n$ unabhängig und $X_i \sim \mathrm{Ber}_p$ für $i = 1,2,\dots,n$

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

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

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

Dieser Zusammenhang mit ZVn ausgesprochen:

$$X = (X_1,\dots,X_n) \sim \mathrm{Ber}_p^{\otimes n}, \quad Y := X_1 + X_2 + \dots + X_n,$$

so ist $Y \sim \mathrm{Bin}_{n,p}$, denn für $k \in \{0,1,\dots,n\}$ und für jede Wahl von $(x_1,\dots,x_n) \in \{0,1\}^n$ mit $x_1 + \dots + x_n = k$ ist $P(X_1 = x_1,\dots,X_n = x_n) = p^k(1-p)^{n-k}$, somit mit $B_k = \big\{(x_1,\dots,x_n) \in \{0,1,\}^n : x_1 + \dots + x_n = k\big\}$

$$\begin{aligned} P(Y = k) &= P\Big((X_1,\dots,X_n) \in B_k\Big) = \sum_{(x_1,\dots,x_n)\in B_k} P\big((X_1,\dots,X_n) = (x_1,\dots,x_n)\big) \\ &= \sum_{(x_1,\dots,x_n)\in B_k} p^k(1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

da $\#B_k = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \binom{n}{k}$. Insoweit ist dies eine Beispiel-Instanz zu Beobachtung 1.11.

**Beispiel 1.17** (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 zum ersten Erfolg.“

**Beispiel 1.18** (Multinomialverteilung)**.** $s \in \{2,3,\dots\}$, $p_1,\dots,p_s \in [0,1]$, $p_1 + \dots + p_s = 1$, $n \in \mathbb{N}$, $S = \big\{(k_1,\dots,k_s) \in \mathbb{N}_0^s : k_1 + \dots + k_s = n\big\}$,

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

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 $i = 1,2,\dots,s$.

[4]nach Jakob Bernoulli, 1654–1705

**Beispiel 1.19** (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 1.20** (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

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

□

Prop. 1.20 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 1.21.** L. 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):

| 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 $\mathrm{Poi}_\lambda$-verteilt mit $\lambda = 0{,}61$, so würden wir das Resultat $x$ je $200 \times \mathrm{Poi}_{0,61}(x)$-mal erwarten:

[5] nach Siméon Denis Poisson, 1781–1840
[6] Ladislaus von Bortkewitsch, 1868–1931

| 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 $\mathrm{Poi}_\lambda$ ist $\lambda$") und somit ist obiges der naheliegende „Momentenschätzer" – wir werden darauf zurückkommen.

**Bsp. 1.21: Das „making of" mit R:**

```
# 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
```

### 1.1.3 Der Fall 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 (man denke z.B. an Längen, Gewichte, Konzentrationen, ...), 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 1.22** (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

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

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

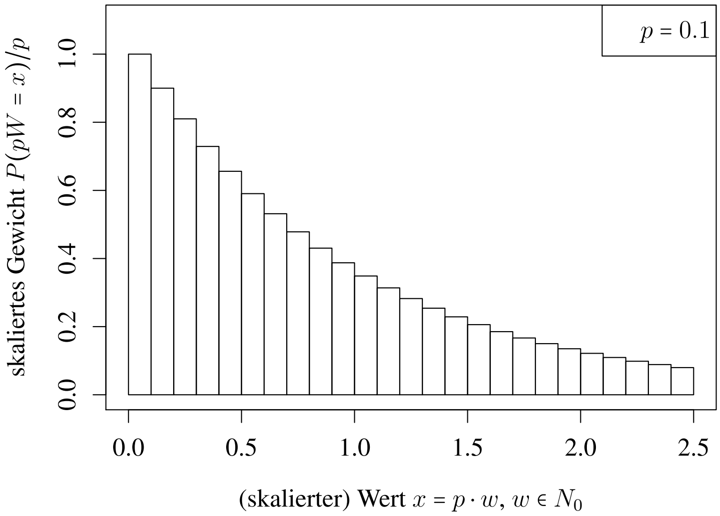
$$P(X \geq x) = P(pW \geq x) = P\big(W \geq \lceil x/p \rceil\big) = (1-p)^{\lceil x/p \rceil} \approx (1-p)^{x/p} \approx e^{-x}$$

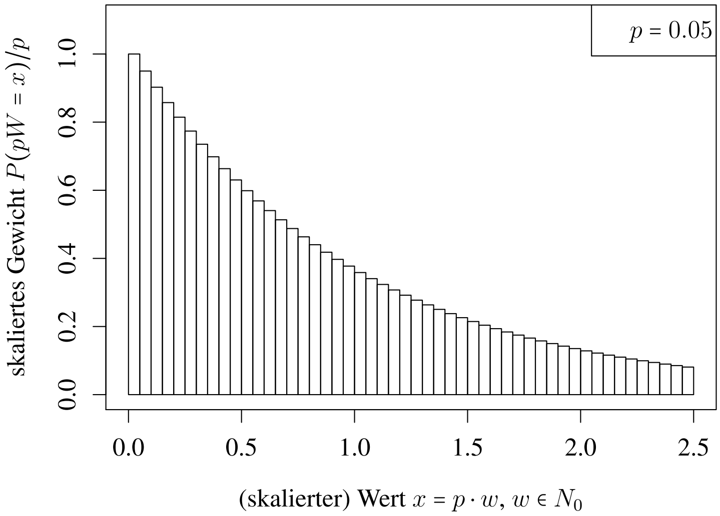
(und zwar "egal, wie klein" $p$ ist).

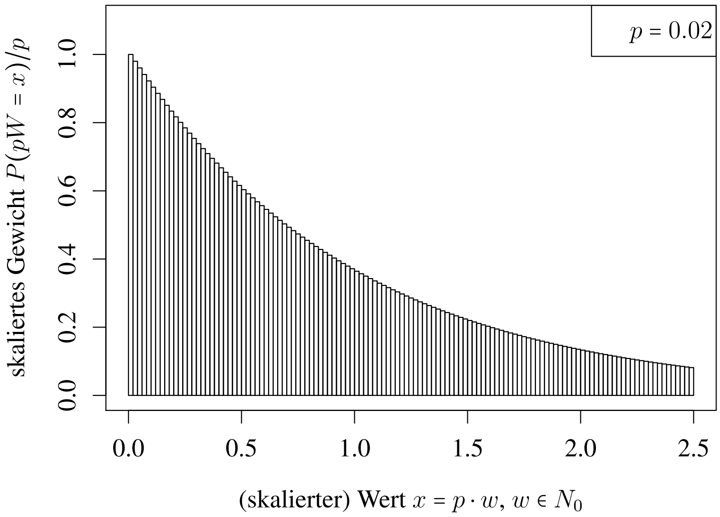
Interpretation z.B.: via Wartezeiten auf sehr fein zeitdiskretisiertem Gitter

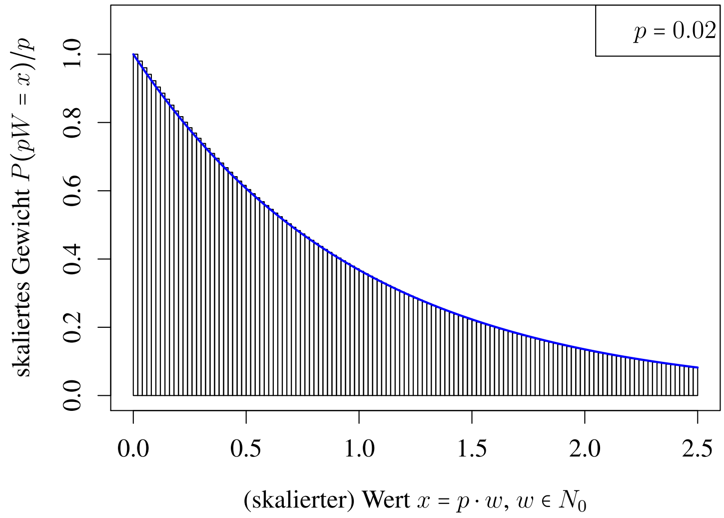
Frage also: Gibt es eine reellwertige ZV $X$, für die obiges ($P(X \geq x) = e^{-x}$) als Identität gilt? Ja, wie wir sehen werden.

(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.)


p = 0.1
skaliertes Gewicht P(pW = x)/p
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
(skalierter) Wert x = p · w, w ∈ N0


p = 0.05
skaliertes Gewicht P(pW = x)/p
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
(skalierter) Wert x = p · w, w ∈ N0


p = 0.02
skaliertes Gewicht P(pW = x)/p
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
(skalierter) Wert x = p · w, w ∈ N0




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

**Beispiel 1.23** (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:

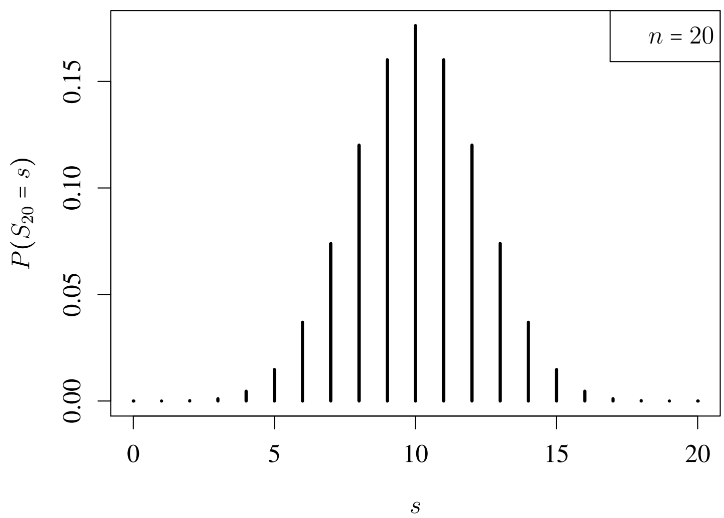
$$X_n := \frac{S_n - n/2}{\sqrt{n}}$$

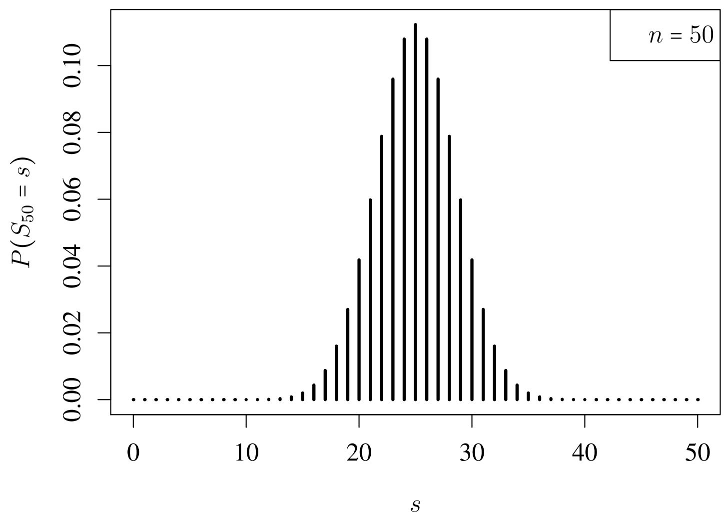
und skalieren die „Balken“ so, dass die Fläche des Balkens bei $(s - n/2)/\sqrt{n}$ gerade

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

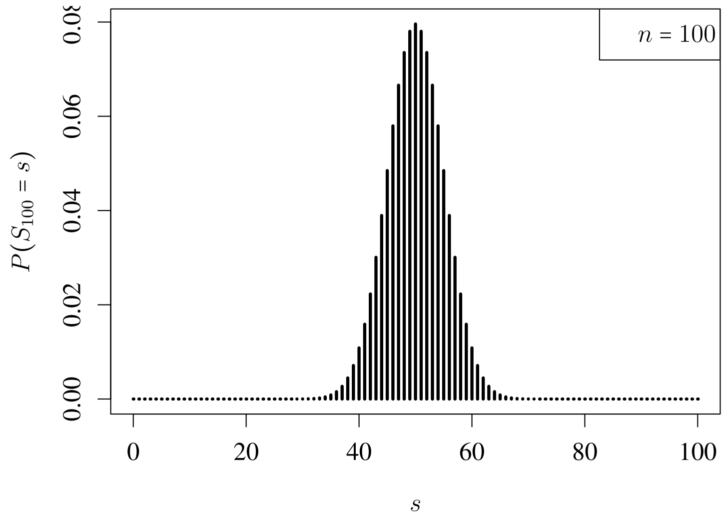
entspricht.

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)

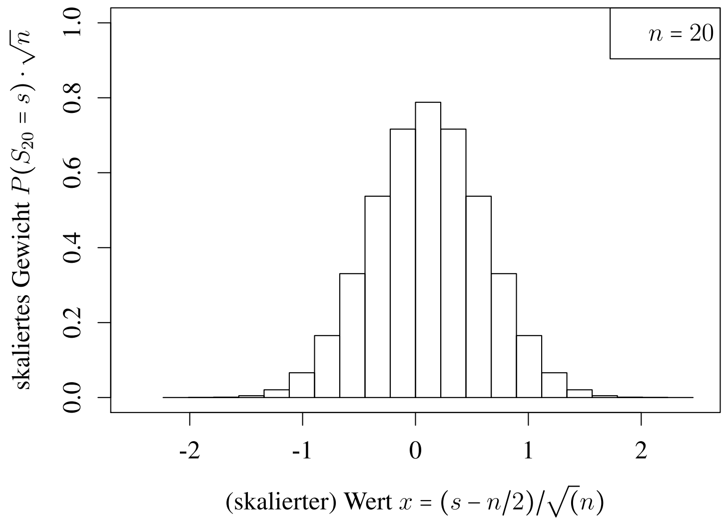




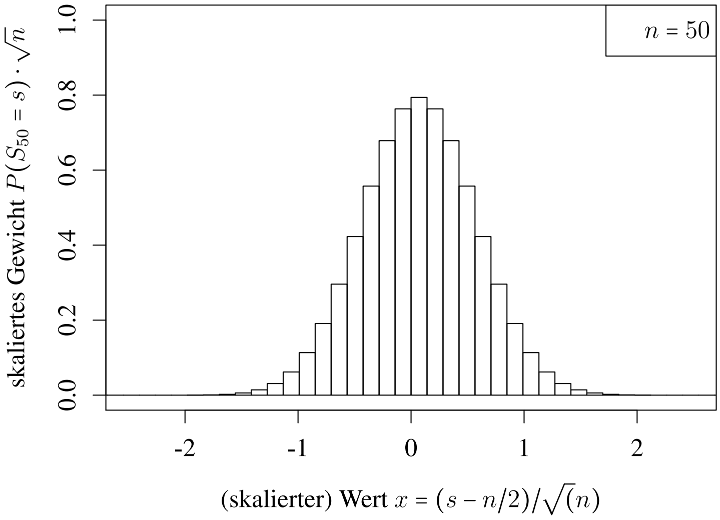
$n = 50$
$P(S_{50} = s)$
$s$

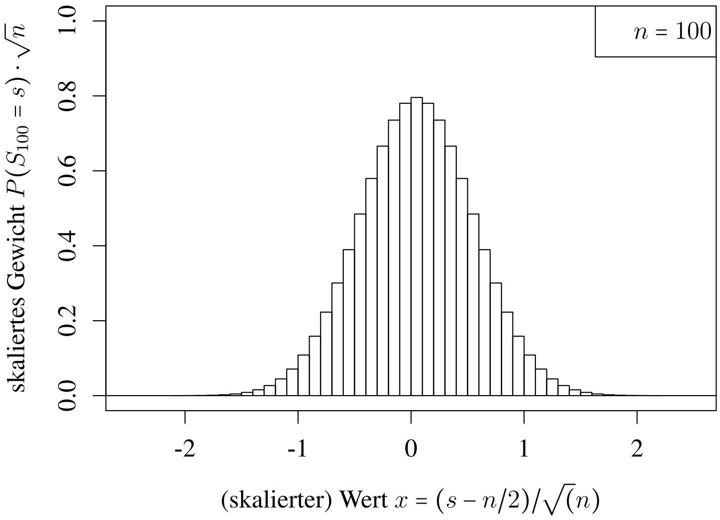



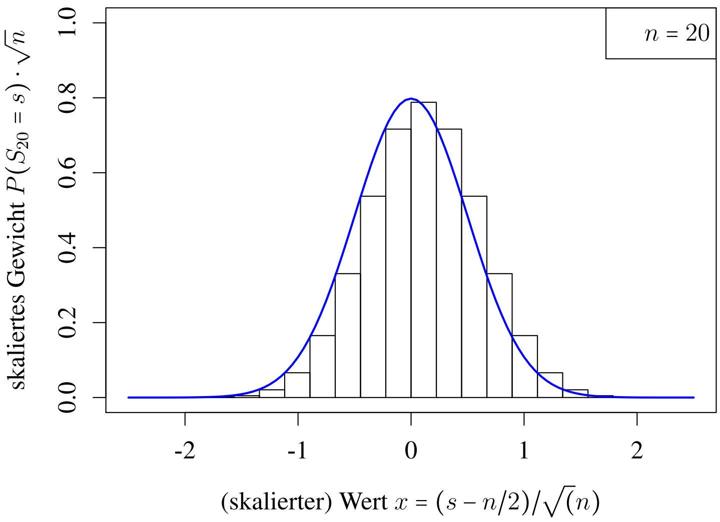
$n = 100$
$P(S_{100} = s)$
$s$

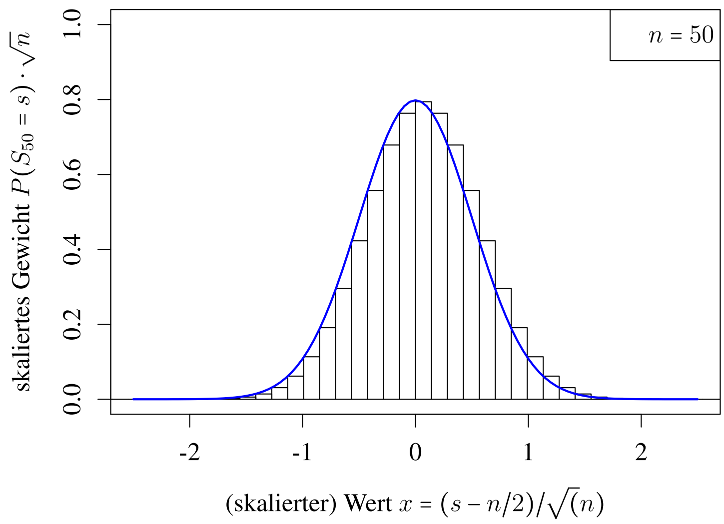



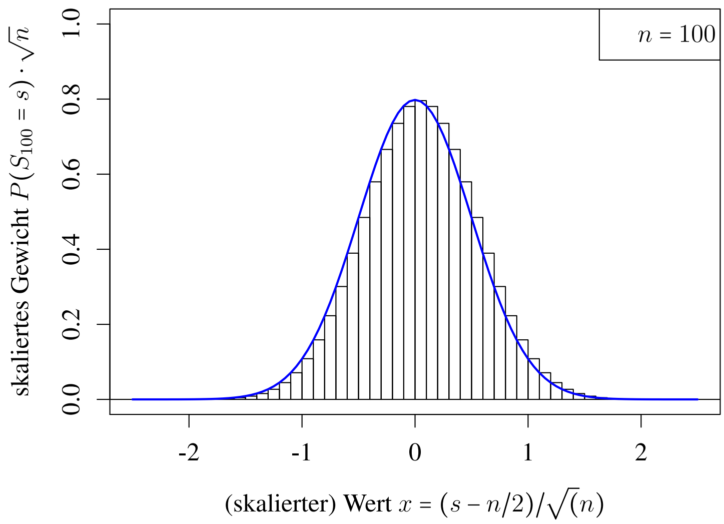
$n = 20$
skaliertes Gewicht $P(S_{20} = s) \cdot \sqrt{n}$
(skalierter) Wert $x = (s - n/2)/\sqrt{(n)}$


$n = 50$
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 $x = (s - n/2)/\sqrt{(n)}$


$n = 20$
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 $x = (s - n/2)/\sqrt{(n)}$




**Definition 1.24.** Sei $X$ eine Zufallsvariable mit Werten in einem Intervall $S = [a, b] \subset \mathbb{R}$ mit $-\infty \le a < b \le \infty$ (im Fall $a > -\infty, b = \infty$ meinen wir $S = [a, \infty)$, etc.) und sei $f : S \to [0, \infty)$ eine integrierbare[7] Funktion mit

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

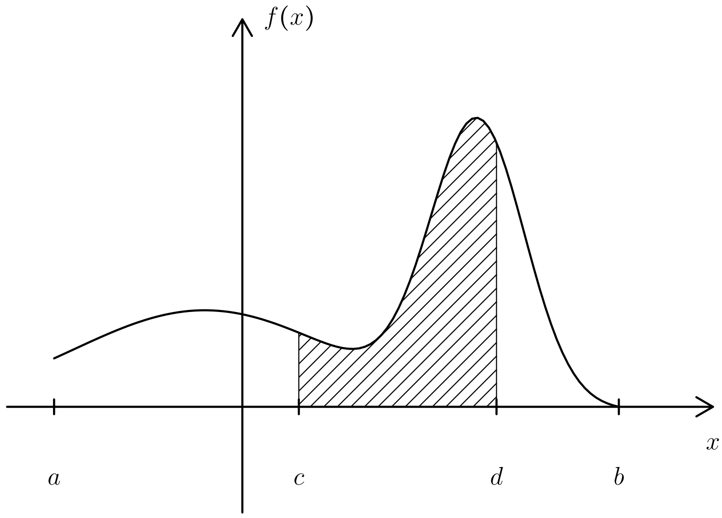
$X$ besitzt die *Dichte* (auch: *Wahrscheinlichkeitsdichte*) $f$, wenn gilt

$$P\big(X \in [c, d]\big) = \int_c^d f(x)\,dx \quad \text{für jedes Teilintervall } [c, d] \subset S. \tag{1.9}$$

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

[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].




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

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

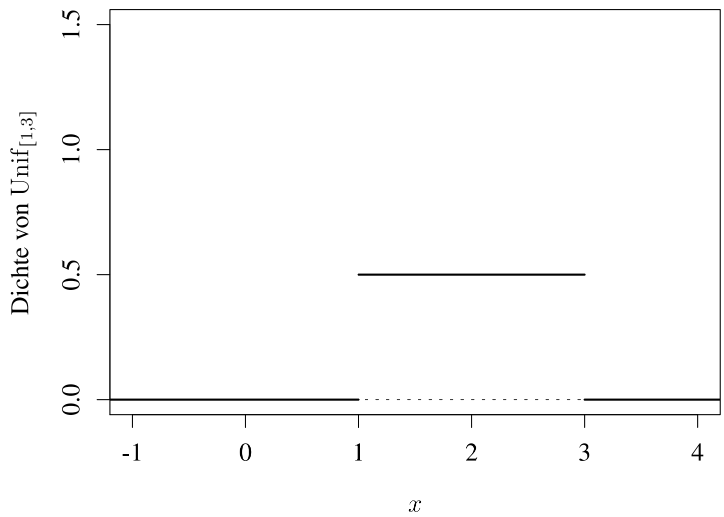
(Dieser suggestive Ausdruck erhält einen Sinn im Sinne der „Standard-Analysis“, wenn man auf beiden Seiten $x$ über ein Intervall $[c, d]$ integriert, dann erhält man (1.9).)

**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

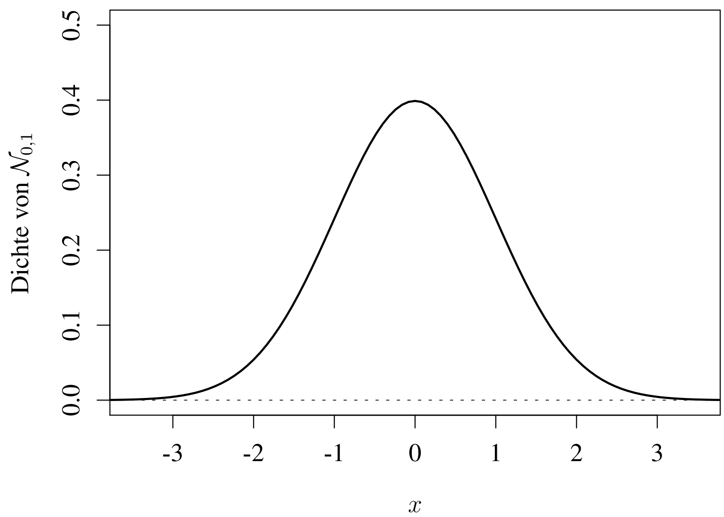
$$P(X = x) = \lim_{\delta \downarrow 0} P\big(X \in [x, x+\delta]\big) = \lim_{\delta \downarrow 0} \int_x^{x+\delta} f_X(a)\, da = \int_x^x f_X(a)\, da = 0.$$

**Beispiel 1.25** (Einige „klassische“ eindimensionale Verteilungen mit Dichte)**.**

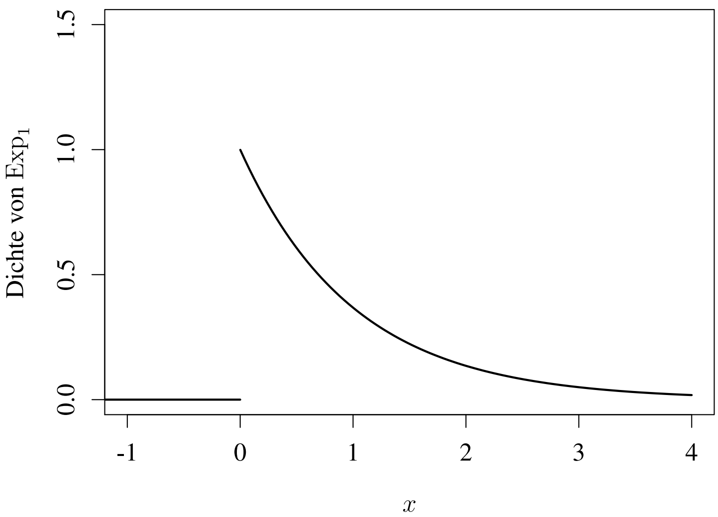
1. (uniforme Verteilung) $a, b \in \mathbb{R}$, $a < b$. $\mathrm{Unif}_{[a,b]}$ mit Dichte $\frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$




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 $\mu$ und Varianz $\sigma^2$.
   $\mathcal{N}_{0,1}$ heißt die *Standardnormalverteilung*.




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




**Definition 1.26** (Verteilungsfunktion)**.** Für eine Zufallsvariable $X$ mit Werten in $\mathbb{R}$ (bzw. in einer Teilmenge $S \subset \mathbb{R}$) heißt die Funktion

$$F_X(x) := P(X \leq x), \quad x \in \mathbb{R} \tag{1.10}$$

die *Verteilungsfunktion* von $X$.

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

$$F_X(x) = \int_{-\infty}^{x} f_X(a)\, da \tag{1.11}$$

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

**Beispiel 1.25** (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 $\max\big\{\min\big\{\frac{x-a}{b-a}, 1\big\}, 0\big\}$

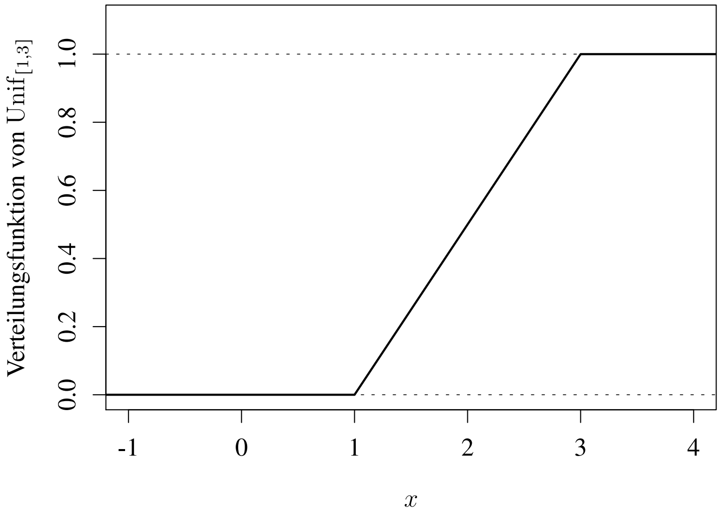
2. (Normalverteilung[en]) $\mu \in \mathbb{R}$, $\sigma > 0$. $\mathcal{N}_{\mu,\sigma^2}$ mit Dichte $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$ heißt Normalverteilung mit Mittelwert $\mu$ und Varianz $\sigma^2$.

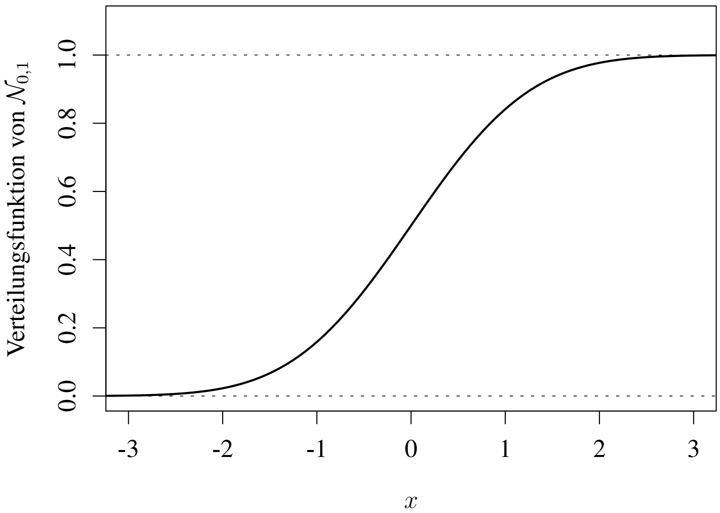
   $\mathcal{N}_{0,1}$ heißt die *Standardnormalverteilung*, die Verteilungsfunktion

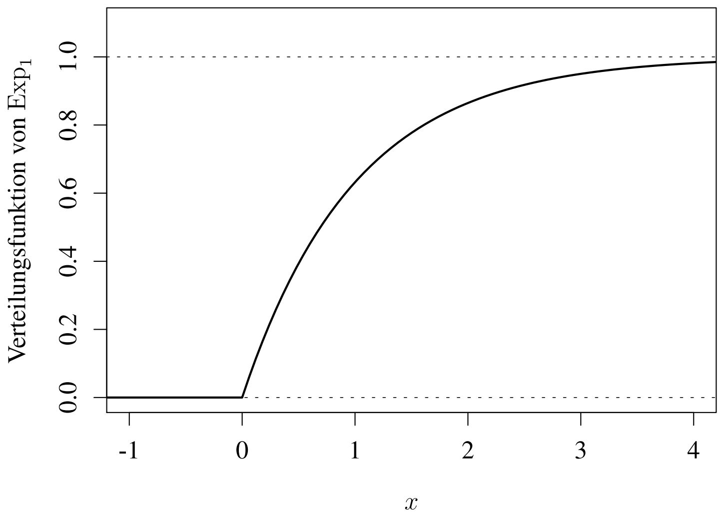
   $$\Phi(x) := F_{\mathcal{N}_{0,1}}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

   ist tabelliert bzw. in vielen Computerprogrammen implementiert

3. (Exponentialverteilung[en]) $\theta > 0$, $\mathrm{Exp}_\theta$ 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 1.27.** 1. Die Dichte / Verteilungsfunktion von $X$ hängt nur von der Verteilung von $X$ ab:

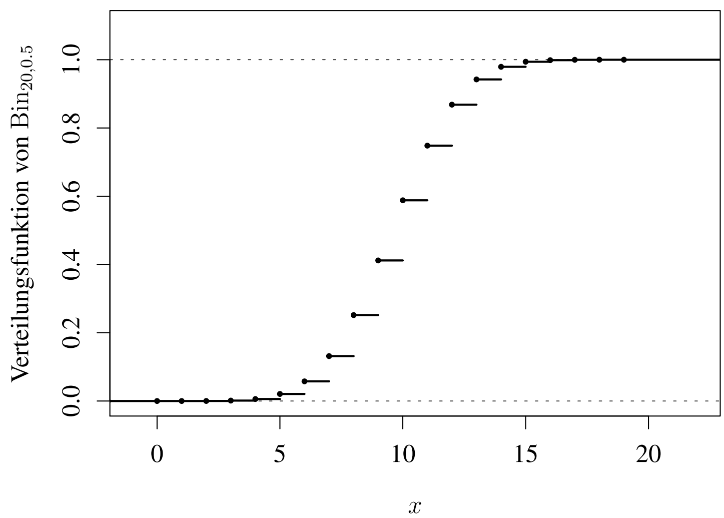
wenn $Y =^d X$ („Gleichheit in Verteilung“), also $P(X \in B) = P(Y \in B)$ für alle $B$ gilt, so hat $Y$ dieselbe (offenbar).

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 (1.11) 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 Forderung 1.2)
$$P(X \in B) = \sum_{i=1}^n P\big(X \in (c_i, d_i]\big) = \sum_{i=1}^n \big(F_X(d_i) - F_X(c_i)\big)$$
(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. 1.10, 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. 1.27, 5. eine ZV $X$ mit $F_X = F$.

   (Wir kommen darauf zurück, siehe Beob. 1.29 unten.)

**Definition 1.28.** Die (verallgemeinerte) inverse Funktion von $F_X$,

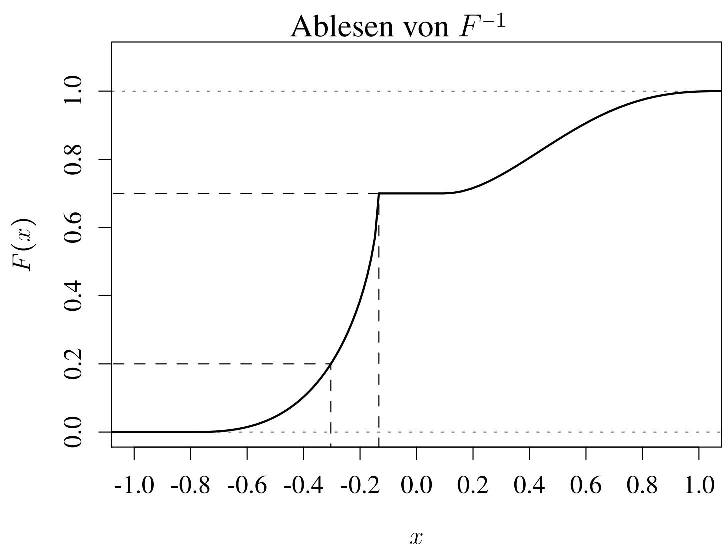
$$F_X^{-1}(t) := \inf\{x \in \mathbb{R} : F(x) \geq t\}, \quad t \in [0,1]$$

(mit Setzung $\inf \varnothing = +\infty$) heißt auch die *Quantilfunktion* von $X$.

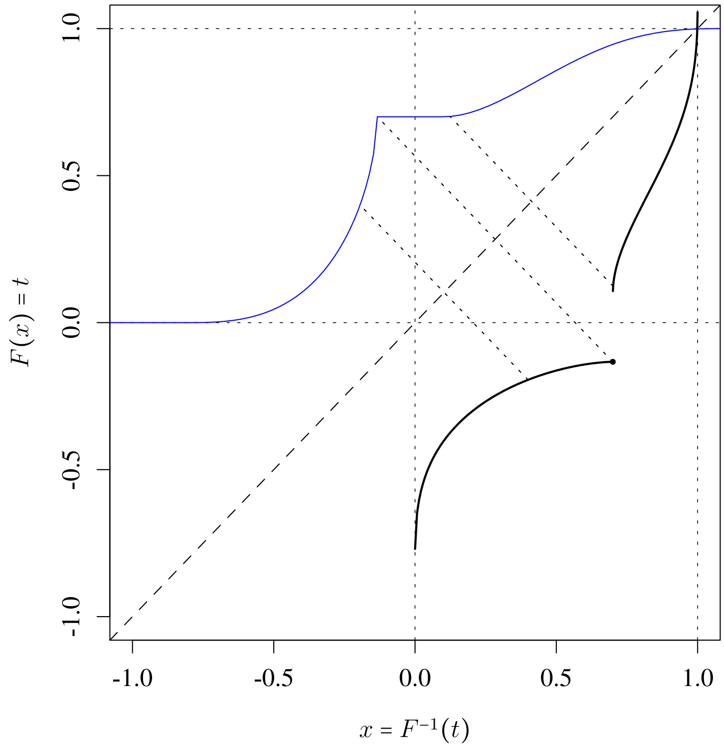
(Beachte, dass die so definierte Funktion $F_X^{-1}$ linksstetig ist. Mit dieser Definition ergibt sich für $x \in \mathbb{R}$, $t \in [0,1]$ die Beziehung

$$F_X^{-1}(t) \leq x \iff t \leq F_X(x).$$

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 1.29** (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. 1.28),* $U$ *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 1.30.** 1. $\mathrm{Exp}_\theta$ hat Verteilungsfunktion $F_{\mathrm{Exp}_\theta}(x) = (1 - e^{-\theta x})\mathbf{1}_{[0,\infty)}(x)$ mit inverser Funktion $F^{-1}_{\mathrm{Exp}_\theta}(t) = -\frac{1}{\theta}\log(1-t)$, also ist $-\dfrac{1}{\theta}\log(1-U) \sim \mathrm{Exp}_\theta$ (und natürlich ebenso $-\dfrac{1}{\theta}\log(U)$)

2. $p(k), k \in \mathbb{N}_0$ Wahrscheinlichkeitsgewichte, $F(x) = \sum_{k:k\leq x} p(k)$ zugehörige Verteilungsfunktion (vgl. Bem. 1.27, 4.), so hat

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

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 $\sigma$, nicht mit $\sigma^2$)
- Exponentialverteilung $\mathrm{Exp}_\lambda$: `[d|p|q|r]exp(...,rate=`$\lambda$`)`
- Poissonverteilung $\mathrm{Poi}_\lambda$: `[d|p|q|r]pois(...,lambda=`$\lambda$`)`
- Binomialverteilung $\mathrm{Bin}_{n,p}$: `[d|p|q|r]binom(...,size=`$n$`,prob=`$p$`)`

(Siehe auch die Online-Hilfe in R.)

### Transformation von Dichten

In Beob. 1.11 hatten wir gesehen, wie Gewichte einer diskreten ZV $f(X)$ aus denen von $X$ (mit Wertebereich $S_X$) berechnet werden, wenn $f : S_X \to S$ eine Funktion ist. Die analoge Frage im Fall mit Dichte beantwortet die folgende Beobachtung.

**Beobachtung 1.31.** $X$ 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: $Y \in [y, y+\delta] \iff X \in \big[\frac{y-b}{a}, \frac{y-b}{a} + \frac{\delta}{a}\big]$.

**Beispiel 1.32.** 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 $\mathcal{N}_{\mu,\sigma^2}\big((-\infty, x]\big) = \Phi\big((x-\mu)/\sigma\big)$

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 1.33** (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 $\varphi$ 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.

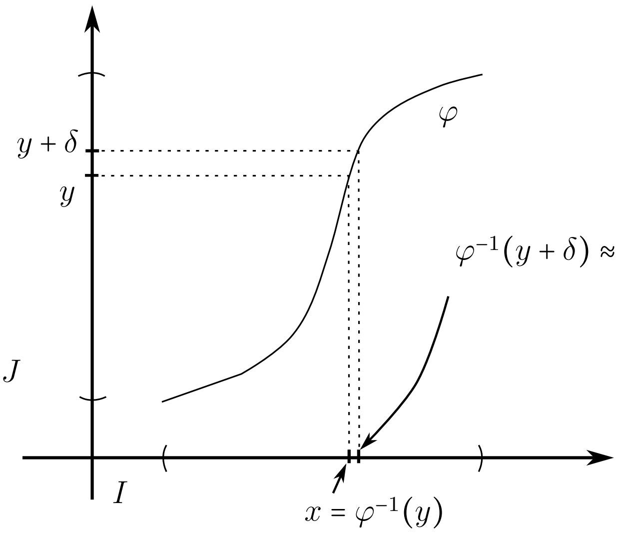
*Beweis.* $\varphi$ muss offenbar strikt wachsend oder strikt fallend sein, wir betrachten den wachsenden Fall.

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. $\square$




**Beispiel 1.34.** 1. $U \sim \mathrm{Unif}_{[0,1]}$, so hat $X := -\log(U)$ Dichte $e^{-x}$ für $x \ge 0$ (wie wir in Bsp. 1.30 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

Wir werden im weiteren Verlauf Gelegenheit haben, mehrdimensionale Zufallsvariablen mit Dichte zu betrachten, hier zunächst die grundlegenden Begriffe und Definitionen dazu.

Das multivariate Analogon zu Def. 1.24 ist folgende

**Definition 1.35.** 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

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

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

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

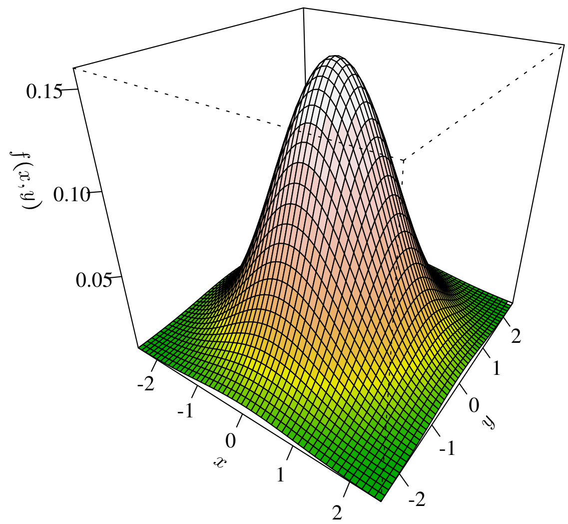
für $(x_1,\dots,x_d) \in \mathbb{R}^d$ und $0 < \delta_1, \delta_2, \dots, \delta_d \ll 1$.

**Beispiel 1.36.** 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.)

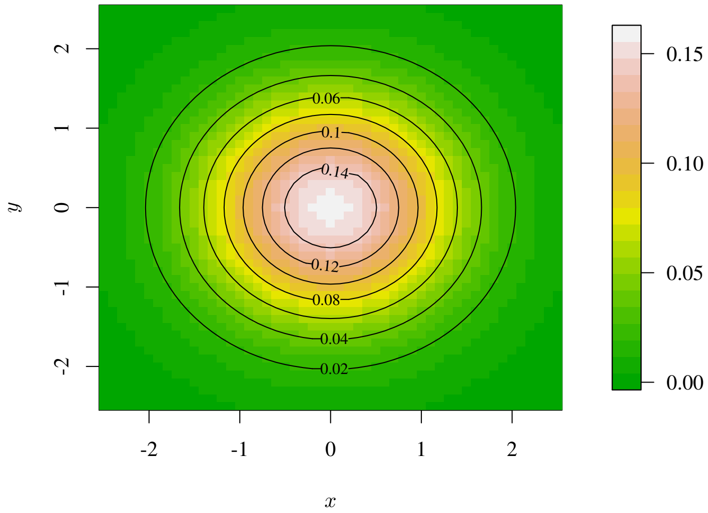
2. Die 2-dimensionale Standard-Normalverteilung hat Dichte $f(x,y) = \dfrac{1}{2\pi} \exp\big(-\frac{1}{2}(x^2+y^2)\big)$




[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.

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




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

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

## 1.2 „Kleingedrucktes“

Dieses Unterkapitel sei allen interessierten Lesern an Herz gelegt, es kann aber übersprungen werden, ohne im weiteren Verlauf der Vorlesung „abgehängt“ zu werden.

### 1.2.1 Eine Anmerkung zur Maßtheorie

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 Forderung 1.2 gibt – in Def. 1.8 gewisse Einschränkungen an die Menge der erlaubten Wertebereiche $S$ und Abbildungen $X$ sowie an die Abbildungen $f$ zwischen Wertebereichen in Beob. 1.11 machen.

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

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

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?“).

Im Fall $S = \mathbb{R}$ oder $S = \mathbb{R}^d$ wählt man für $\mathscr{S}$ meist die sogenannte Borel-$\sigma$-Algebra, die kleinste $\sigma$-Algebra, die alle offenen Mengen enthält.

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

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

Für einen gegebenen Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, P)$, in dem offenbar $(\Omega, \mathscr{F})$ einen messbaren Raum bilden, vgl. Def. 1.1, und Wertebereich $(S, \mathscr{S})$ sind nicht alle Funktionen $X : \Omega \to S$ zugelassen, sondern nur $\mathscr{F}$-$\mathscr{S}$-messbare.

Weiterhin sind für eine $S$-wertige ZV $X$ als Ereignisse in Def. 1.8 nur $\{X \in A\}$ mit $A \in \mathscr{S}$ zugelassen (für $A \subset S$ mit $A \notin \mathscr{S}$ wird die Frage, ob $X$ in $A$ liegt, nicht „erlaubt“).

### 1.2.2 Alternativer Zugang: Zufallsvariablen als „Fundamentalobjekte“ der Wahrscheinlichkeitstheorie

Wir haben in diesem Kapitel, dem „traditionellen Weg“ folgend, Wahrscheinlichkeitsräume und Ereignisse an den Anfang der Diskussion gestellt (vgl. Def. 1.1) und dann Zufallsvariablen (vgl. Def. 1.8) sozusagen als „abgeleitete Objekte“ erhalten. Man kann – logisch äquivalent – die Zufallsvariablen an den Anfang stellen und dann daraus Ereignisse via Indikatorvariable (ähnlich zu Bsp. 1.9) ableiten. Diesen Zugang findet man im Detail in [K09] und in [KW], dieser Weg wurde auch in der Version der Vorlesungsnotizen aus SS2017 [SfI17] begangen. Dieses Mal bleibt es hauptsächlich aus Zeitgründen – das SS 2018 ist an der JGU besonders kurz – nur beim üblichen Zugang.

### 1.2.3 Zur Einschluss-Ausschluss-Formel

Hier der Vollständigkeit halber ein Beweis der Einschluss-Ausschluss-Formel (1.8): Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum und $A_i \in \mathcal{F}$, wobei $i \in \{1, ..., n\} =: I$, so gilt

$$P\Big(\bigcup_{i\in I} A_i\Big) = \sum_{\varnothing \neq K \subset I} (-1)^{|K|-1} P\Big(\bigcap_{k\in K} A_k\Big).$$

Wir gehen dafür in mehreren Schritten vor. Für $J \subset I$ sei

$$B_J = \bigcap_{j\in J} A_j \cap \bigcap_{j\in I\setminus J} A_j^c,$$

dabei sei ein Durchschnitt mit leerer Indexmenge definiert als $\Omega$, d.h. $\bigcap_{j\in\varnothing} A_j := \Omega$.

Wir zeigen die folgende Aussagen:

(a) Für jedes $K \subset I$ gilt
$$P\Big(\bigcap_{k\in K} A_k\Big) = \sum_{K\subset J\subset I} P(B_J).$$

(b) Für jedes $J \subset I$ gilt
$$P(B_J) = \sum_{J\subset K\subset I} (-1)^{|K\setminus J|} P\Big(\bigcap_{k\in K} A_k\Big).$$

(c) Die oben angegebene Einschluss-Ausschluss-Formel folgt, indem wir $J = \varnothing$ setzen und Aussage (b) mit
$$P\Big(\bigcap_{i\in I} A_i^c\Big) = 1 - P\Big(\bigcup_{i\in I} A_i\Big)$$
verwenden.

Zu (a): Wir zeigen zunächst, dass
$$B_J \cap B_{\widetilde{J}} = \varnothing \text{ für } J \neq \widetilde{J} \quad \text{und} \bigcup_{\varnothing\subset J\subset I} B_J = \Omega.$$

Sei $J \neq \widetilde{J}$, d.h. es gibt ein $j \in (J \setminus \widetilde{J}) \cup (\widetilde{J} \setminus J)$, dann gilt $B_J \cap B_{\widetilde{J}} \subset A_j \cap A_j^c = \varnothing$. Für $x \in \Omega$ sei $J(x) := \{j \in I : x \in A_j\}$, so ist nach Definition $x \in B_{J(x)}$. Da zu jedem $x \in \Omega$ ein solches $J(x) \subset I$ gehört, ist $\Omega \subset \bigcup_{J:\varnothing\subset J\subset I} B_J \subset \Omega$, d.h. es gilt $\bigcup_{J:\varnothing\subset J\subset I} B_J = \Omega$.

Sei nun $K \subset I$ gegeben. Nach obigem ist
$$\bigcap_{k\in K} A_k = \Big\{x \in \Omega : J(x) \supset K\Big\} = \bigcup_{J:K\subset J\subset I} B_J, \tag{1.12}$$
und die Zerlegung ist disjunkt. (a) folgt somit aus der Additivitätseigenschaft des Wahrscheinlichkeitsmaßes $P$.

Zu (b): Sei $J \subset I$ fest gewählt. Es ist nach (a)
$$\begin{aligned}\sum_{K:J\subset K\subset I} (-1)^{|K\setminus J|} P\Big(\bigcap_{k\in K} A_k\Big) &= \sum_{K:J\subset K\subset I} (-1)^{|K\setminus J|} \sum_{J':K\subset J'\subset I} P\big(B_{J'}\big)\\ &= \sum_{J':J\subset J'\subset I} \sum_{K:J\subset K\subset J'} (-1)^{|K\setminus J|} P\big(B_{J'}\big)\\ &= \sum_{J':J\subset J'\subset I} P\big(B_{J'}\big) \underbrace{\sum_{K:J\subset K\subset J'} (-1)^{|K\setminus J|}}_{=\begin{cases}1, & J'=J\\ 0, & J'\neq J\end{cases}} = P\big(B_J\big).\end{aligned}$$

Für die letzte Gleichung beachten wir beispielsweise folgendes: Sei $J' \supsetneq J$, d.h. es gibt ein $j' \in J' \setminus J$, dann ist
$$\sum_{K:J\subset K\subset J'} (-1)^{|K\setminus J|} = \sum_{K:J\subset K\subset J'\setminus\{j'\}} \Big((-1)^{|K\setminus J|} + (-1)^{|(K\cup\{j'\})\setminus J|}\Big) = \sum_{K:J\subset K\subset J'\setminus\{j'\}} 0 = 0.$$

(Bemerkung: Teil (b) ist ein Spezialfall der sogenannten Möbius-Inversion.)

Zu (c): Es ist
$$\begin{aligned}P\Big(\bigcup_{i\in I} A_i\Big) &= 1 - P\bigg(\Big(\bigcup_{i\in I} A_i\Big)^c\bigg) = 1 - P\Big(\bigcap_{i\in I} A_i^c\Big) = 1 - P\big(B_\varnothing\big)\\ &= 1 - \sum_{K:\varnothing\subset K\subset I} (-1)^{|K|} P\Big(\bigcap_{k\in K} A_k\Big) = 1 - (-1)^0 P\big(\Omega\big) + \sum_{K:\varnothing\neq K\subset I} (-1)^{|K|-1} P\Big(\bigcap_{k\in K} A_k\Big)\\ &= \sum_{K:\varnothing\neq K\subset I} (-1)^{|K|-1} P\Big(\bigcap_{k\in K} A_k\Big).\end{aligned}$$

## 1.3 Bedingte Wahrscheinlichkeiten, Unabhängigkeit, gemeinsame Verteilung

### 1.3.1 Bedingte Wahrscheinlichkeit

**Beispiel 1.37.** Wir ziehen zwei Kugeln ohne Zurücklegen aus einer Urne mit $s > 0$ schwarzen und $w > 0$ weißen Kugeln. Wie in Bsp. 1.13, 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. 1.13, 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

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

(und übrigens $= P(A)$)

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 1.38.** *Sei* $(\Omega, \mathcal{F}, P)$ *W'raum,* $A \in \mathcal{F}$ *mit* $P(A) > 0$*. Für* $B \in \mathcal{F}$

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

*heißt* bedingte Wahrscheinlichkeit von $B$, gegeben $A$.

$P(\,\cdot \mid A)$ *ist ein Wahrscheinlichkeitsmaß auf* $\Omega$*, d.h. die Eigenschaften aus Forderung* (1.2)*, Normierung und* $\sigma$*-Additivität sind erfüllt. (Prüfung per Inspektion)*

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

In Beispiel 1.37 ist

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

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

**Bemerkung 1.39** („Natürlichkeit von Definition 1.38"). Nehmen wir an, wir möchten angesichts der Information „$A$ ist eingetreten" das W'maß $P$ revidieren zu einem W'maß $\widetilde{P}$ mit

1. $\widetilde{P}(A) = 1$ (d.h. $A$ ist sicher unter $\widetilde{P}$) und
2. $\widetilde{P}(B) = c_A P(B)$ für $B \subset A$ mit einem $c_A > 0$

   (d.h. Teilereignisse von $A$ erhalten bis auf Normierung ihr altes Gewicht).

Dann gilt

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

*Beweis.* Für $C \in \mathcal{F}$ ist

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

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

**Bemerkung 1.40.** $P(B \mid A) \neq P(B)$ kann nicht notwendigerweise als „Kausalität" (im Sinne von „$A$ beeinflusst, ob $B$ eintritt") interpretiert werden:

So ist in Beispiel 1.37 ist auch

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

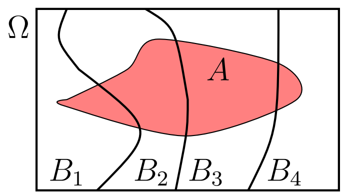
aber es passt nicht zu unserer Vorstellung, dass der 2. Zug den 1. Zug beeinflusst.

**Satz 1.41.** *Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum, $I$ eine (höchstens abzählbare) Indexmenge, $B_i \in \mathcal{F}$ paarweise disjunkt mit $P\big(\bigcup_{i\in I} B_i\big) = 1$ und $P(B_i) > 0$ für $i \in I$.*

1. *(Formel von der totalen Wahrscheinlichkeit)*

   *Für $A \in \mathcal{F}$ gilt*

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




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

2. *(Formel von Bayes[11])*

   *Für $A \in \mathcal{F}$ mit $P(A) > 0$ und jedes $k \in I$ gilt*

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

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

*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 $B \cup B^c = \Omega$)*

*Beweis.* 1. Mit Definition 1.38 ist

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

(verwende die ($\sigma$-)Additivität von $P$).

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

**Beispiel 1.42** (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

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

und

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

Z.B. für die konkreten Werte $a = 0.3$, $f_1 = 0.05$, $f_0 = 0.1$ ergibt sich $P(Y = 1) = 0.355$

$$P(X = 1 \mid Y = 1) \approxeq 0.803, \quad P(X = 0 \mid Y = 0) \approxeq 0.977.$$

**Beispiel 1.43** (Medizinische Reihenuntersuchung)**.** Eine Krankheit

- komme bei 2% der Bevölkerung vor („Prävalenz 2%"),
- ein Test schlage bei 95% der Kranken an („Sensitivität 95%"),

- aber auch bei 10% der Gesunden („Spezifität 90%").

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

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

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

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

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

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

(recht hohe „negative Korrektheit").

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

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

| | krank | gesund | $\Sigma$ |
|---|---|---|---|
| pos. getestet | 19 | 98 | 117 |
| neg. getestet | 1 | 882 | 883 |
| $\Sigma$ | 20 | 980 | 1000 |

**Beobachtung 1.44** (Multiplikationsformel). *$A_1, A_2, \dots, A_n \in \mathcal{F}$ Ereignisse (auf einem W'raum $(\Omega, \mathcal{F}, P)$ erklärt) 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)

Betrachten wir z.B. in der Situation von Bsp. 1.37 (Ziehen ohne Zurücklegen aus Urne mit $s$ schwarzen und $w$ weißen Kugeln) drei Züge, sei $A_1 = \{\text{erste Kugel ist weiß}\}$, $A_2 = \{\text{zweite Kugel ist weiß}\}$, $A_3 = \{\text{dritte Kugel ist schwarz}\}$, so ist

$$P(A_1) = \frac{w}{s+w}, \; P(A_2 \mid A_1) = \frac{w-1}{s+w-1}, \; P(A_3 \mid A_1 \cap A_2) = \frac{s}{s+w-2},$$

also $P(A_1 \cap A_2 \cap A_3) = \dfrac{w}{s+w} \cdot \dfrac{w-1}{s+w-1} \cdot \dfrac{s}{s+w-2} = \dfrac{w(w-1)s}{(s+w)(s+w-1)(s+w-2)}.$

### 1.3.2 Gemeinsame Verteilung und Randverteilungen

**Definition 1.45** (Gemeinsame Verteilung und Marginalverteilungen)**.** Seien $X_1, X_2, \dots, X_d$ ZVn (die auf einem gemeinsamen W'raum $(\Omega, \mathcal{F}, P)$ definiert sind), $X_i$ habe Werte in $S_i$,

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

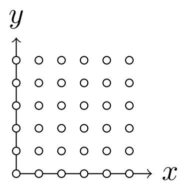
die Produkt-Zufallsvariable (mit Werten in $S_1 \times \cdots \times S_d$), so heißt $\mathscr{L}(X)$ die *gemeinsame Verteilung* der $X_1, X_2, \dots, X_d$ ($\mathscr{L}(X)$ ist ein Wahrscheinlichkeitsmaß auf dem Wertebereich von $X$, d.h. auf dem Produktraum $S_1 \times \cdots \times S_d$, vgl. Definition 1.10). $\mathscr{L}(X_i)$ heißt die $i$-te *Randverteilung* (oder *Marginalverteilung*) von $X$.

**Beispiel.** Fassen wir die Häufigkeitstabelle (normiert auf Gesamtsumme = 1) aus Bsp. 1.43 als Verteilungstafel der gemeinsamen Verteilung von $X_1$ („Testergebnis" mit Werten in $S_1 = \{\text{positiv}, \text{negativ}\}$) und $X_2$ („Gesundheitszustand" mit Werten in $S_2 = \{\text{krank}, \text{gesund}\}$) auf:

| $X_1$ \ $X_2$ | krank | gesund | |
|---|---|---|---|
| positiv | 0.019 | 0.098 | 0.117 |
| negativ | 0.001 | 0.882 | 0.883 |
| | 0.02 | 0.98 | 1 |

**Beispiel 1.46.** 1. Sei $X = (X_1, Y_1)$ uniform verteilt auf dem (diskreten) Quadrat $Q_L := \{0, 1, \dots, L-1\}^2 = \{(x, y) \in \mathbb{N}_0^2 : \max\{x, y\} < L\}$ (mit $L \in \mathbb{N}$), d.h.

$$P(X_1 = x, Y_1 = y) = \frac{1}{\#Q_L} = \frac{1}{L^2} \quad \text{für } x, y \in \{0, 1, \dots, L-1\}.$$

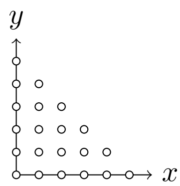



Somit ist $P(X_1 = x) = \sum_{y=0}^{L-1} P(X_1 = x, Y_1 = y) = \frac{1}{L}$ für $0 \le x < L$ und ebenso $P(Y_1 = y) = \frac{1}{L}$ für $0 \le y < L$.

Wir sehen in diesem Fall: $P(X_1 = x, Y_1 = y) = P(X_1 = x) \cdot P(Y_1 = y)$, d.h. die gemeinsamen Gewichte erhält man als *Produkt* der Randverteilungs-Gewichte. Man sagt dazu auch: $X_1$ und $Y_1$ sind unabhängig (siehe Def. 1.51 unten für die allgemeine Definition).

2. Sei $X = (X_2, Y_2)$ uniform verteilt auf dem (diskreten) Dreieck $D_L := \{(x, y) \in \mathbb{N}_0^2 : x + y < L\}$, d.h.

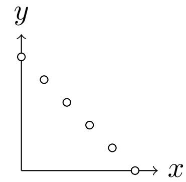
$$P(X_2 = x, Y_2 = y) = \frac{1}{\#D_L} = \frac{2}{L(L+1)} \quad \text{für } (x, y) \in D_L$$




Somit ist $P(X_2 = x) = \sum_y P(X_2 = x, Y_2 = y) = \frac{L-x}{\#D_L} = \frac{2(L-x)}{L(L+1)}$ für $0 \le x < L$ und analog $P(Y_2 = y) = \frac{2(L-y)}{L(L+1)}$ für $0 \le y < L$.

Wir sehen in diesem Fall: $P(X_2 = x, Y_2 = y) \ne P(X_2 = x) \cdot P(Y_2 = y)$, die Koordinaten $X_2$ und $Y_2$ sind nicht unabhängig.

3. Sei $X = (X_3, Y_3)$ uniform verteilt auf der (diskreten, verschobenen) Nebendiagonale $N_L := \{(x,y) \in \mathbb{N}_0^2 : x + y = L\}$, d.h.




$$P(X_3 = x, Y_3 = y) = \frac{1}{\# N_L} = \frac{1}{L} \quad \text{für } (x,y) \in N_L$$

Somit ist $P(X_3 = x) = \sum_y P(X_3 = x, Y_3 = y) = \frac{1}{L}$ für $0 \le x < L$ und analog $P(Y_3 = y) = \frac{1}{L}$ für $0 \le y < L$.

Wir sehen in diesem Fall: $P(X_3 = x, Y_3 = y) \neq P(X_3 = x) \cdot P(Y_3 = y)$, die Koordinaten $X_3$ und $Y_3$ sind nicht unabhängig. (Es gilt hier sogar $Y_3 = L - X_3$, d.h. es gibt einen deterministischen Zusammenhang.)

Zudem sehen wir an 1. und 3. zusammen: Die Randverteilungen legen die gemeinsame Verteilung nicht fest. So gilt $\mathscr{L}(X_1) = \mathscr{L}(X_3)$ und $\mathscr{L}(Y_1) = \mathscr{L}(Y_3)$ (man schreibt dies auch als $X_1 =^d X_3$, $Y_1 =^d Y_3$, „Gleichheit in Verteilung“), aber $\mathscr{L}\big((X_1, Y_1)\big) \neq \mathscr{L}\big((X_3, Y_3)\big)$.

**Bedingte Verteilung und mehrstufige Experimente** Oft betrachtet man folgende Situation:

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

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

   (d.h. für beliebige beobachtete Werte $x_1, x_2, \ldots, x_{k-1}$ kennen wir $P(X_k \in \cdot \mid X_1 = x_1, \ldots, X_{k-1} = x_{k-1})$)

Dann kann man die gemeinsame Verteilung (zumindest im diskreten Fall, d.h. die gemeinsamen Gewichte) des Vektors $(X_1, X_2, \ldots, X_n)$ mittels der Multiplikationsformel (Beob. 1.44, lese dort $A_i = \{X_i = x_i\}$) bestimmen („Pfadregel“):

$$\begin{aligned} &P(X_1 = x_1, X_2 = X_2, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \cdot P(X_2 = x_2 \mid X_1 = x_1) \cdot P(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2) \\ &\quad \cdot \cdots \cdot P(X_n = x_n \mid X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \end{aligned} \tag{1.13}$$

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

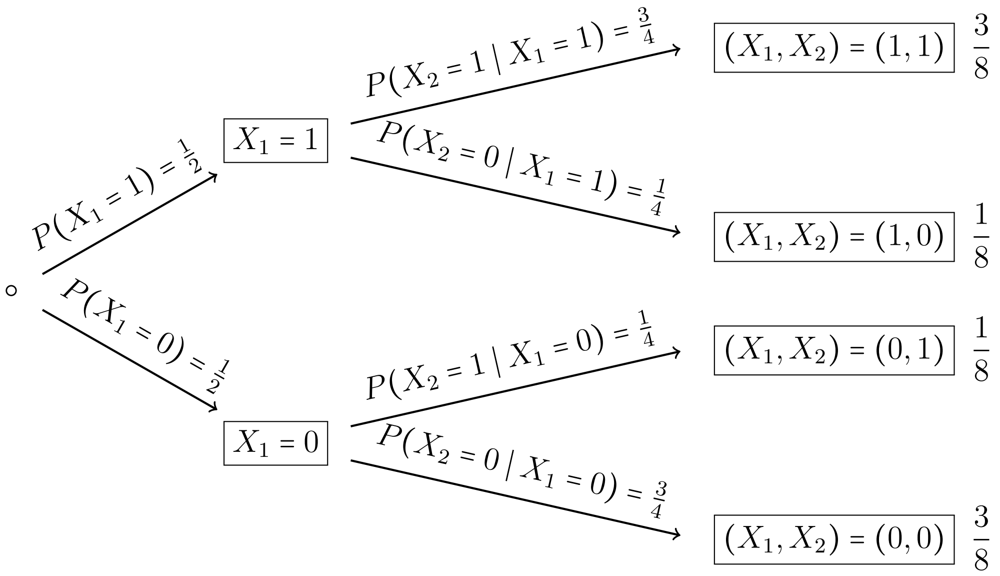
**Beispiel 1.47.** Wir haben eine faire Münze $M_1$ und zwei gezinkte Münzen $M_2$, $M_3$ (deren Seiten jeweils mit 0 und 1 beschriftet seien), wobei

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

Wir werfen erst $M_1$, wenn $M_1$ den Wert 1 zeigt, so werfen wir dann $M_2$, sonst $M_3$. Sei

$$X_i = \text{Resultat des } i\text{-ten Wurfs}, \ i = 1, 2.$$

Somit $S_1 = S_2 = \{0,1\}$, $P(X_1 = 0) = P(X_1 = 1) = \frac{1}{2}$, $P(X_2 = 1 \mid X_1 = 1) = \frac{3}{4} = 1 - P(X_2 = 0 \mid X_1 = 1)$, $P(X_2 = 1 \mid X_1 = 0) = \frac{1}{4} = 1 - P(X_2 = 0 \mid X_1 = 0)$




Die gemeinsame Verteilung von $(X_1, X_2)$ ist damit

| $X_1 \backslash X_2$ | 1 | 0 | |
|---|---|---|---|
| 1 | $\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$ | $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$ | $\frac{1}{2}$ |
| 0 | $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$ | $\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$ | $\frac{1}{2}$ |
| | $\frac{1}{2}$ | $\frac{1}{2}$ | |

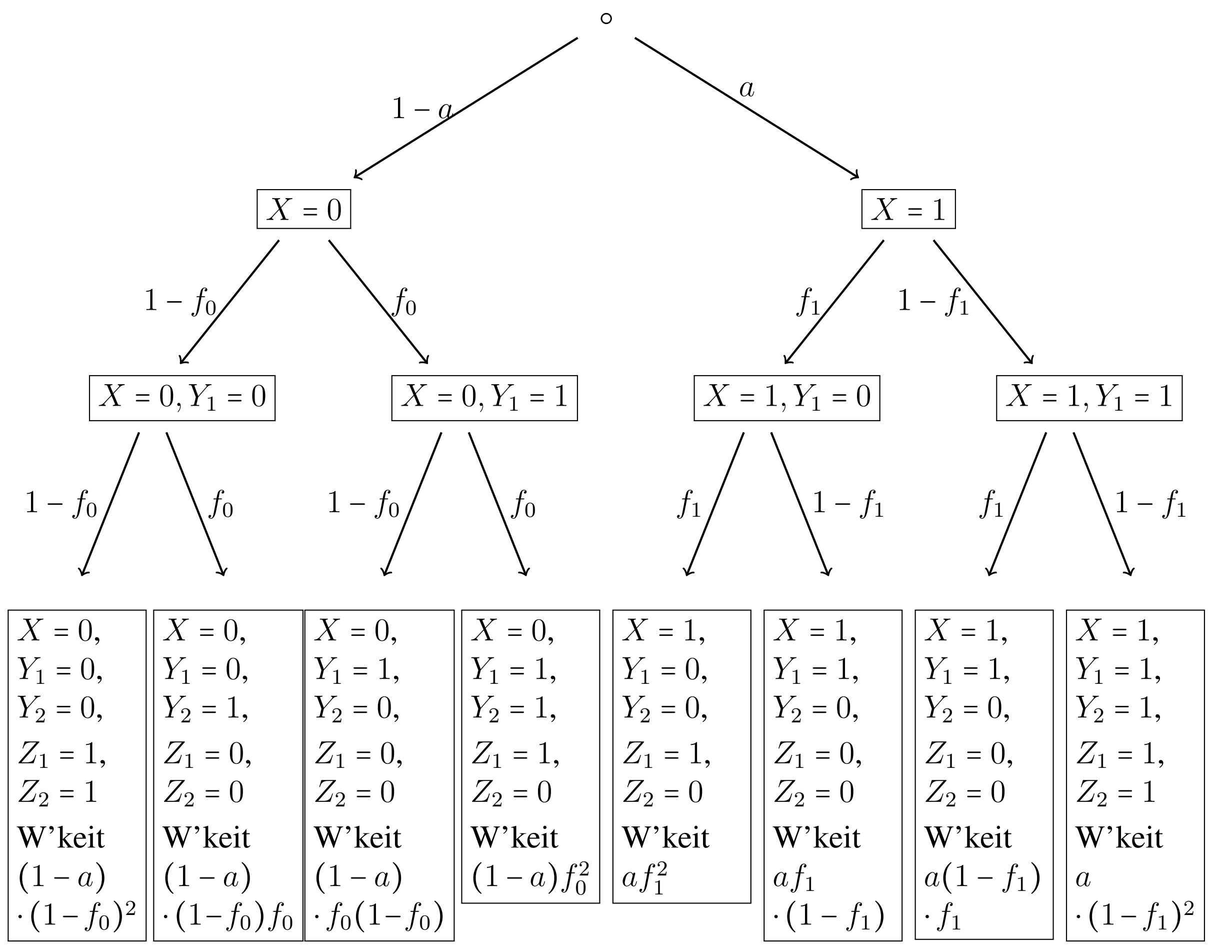
Wir sehen hier wieder: Die Randverteilungen legen (i.A.) nicht die gemeinsame Verteilung fest. Es ist $P(X_1 = 1) = P(X_2 = 1) = \frac{1}{2}$ (und dieselben Randverteilungen ergäben sich, wenn man zwei Mal $M_1$ wirft, aber die gemeinsame Verteilung wäre eine andere).

**Beispiel 1.48.** Nehmen wir an, in der Situation von Bsp. 1.42 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, $Z_1 = \mathbf{1}_{\{Y_1 = Y_2\}}$, $Z_2 = \mathbf{1}_{\{Y_1 = Y_2 = X\}}$ (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:

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

Für die konkreten Zahlenwerte aus Beispiel 1.42 ($a = 0.3, f_1 = 0.05, f_0 = 0.1$) ergibt sich $P(Z_2 = 1 \mid Z_1 = 1) \approxeq 0.991$.

### 1.3.3 Unabhängigkeit

**Definition 1.49.** Ereignisse $A$ und $B$ (auf demselben W'raum) heißen (stochastisch) *unabhängig*, falls gilt

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

(Sofern $P(A) > 0$, so gilt dann auch $P(B \mid A) = P(B)$, d.h. die bedingte und die unbedingte W'keit von $B$ stimmen überein, analog mit vertauschten Rollen.)

Allgemein: Ereignisse $A_1, A_2, \dots, A_n$ heißen (stochastisch) *unabhängig*, falls gilt

$$P\Big(\bigcap_{i\in J} A_i\Big) = \prod_{i\in J} P(A_i) \quad \text{für alle } \varnothing \neq J \subset \{1,2,\dots,n\} \tag{1.14}$$

(für unendlich viele Ereignisse $A_1, A_2, \dots$ : obiges für jedes endliche $J \subset \mathbb{N}$).

**Beispiel.** Wir ziehen eine Karte (nennen wir sie $K$) aus einem gut gemischtem Skatblatt, sei $A = \{K \text{ ist ein As}\}$, $B = \{K \text{ ist kreuz}\}$.

Es ist $P(A) = \frac{4}{32} = \frac{1}{8}$, $P(B) = \frac{8}{32} = \frac{1}{4}$, $P(A \cap B) = \frac{1}{32} = \frac{1}{8} \cdot \frac{1}{4} = P(A) \cdot P(B)$, d.h. $A$ und $B$ sind unabhängig.

Aber auch: Ziehe eine zweite Karte $K_2$ (ohne $K$ zurückzustecken), sei $C = \{K_2 \text{ ist kreuz}\}$. Es ist $P(C) = \frac{8 \cdot 7 + 24 \cdot 8}{32 \cdot 31} = \frac{1}{4}$, $P(A \cap C) = \frac{3 \cdot 8 + 1 \cdot 7}{32 \cdot 31} = \frac{1}{32} = P(A) \cdot P(C)$, d.h. auch $A$ und $C$ sind unabhängig (was vielleicht zumindest auf den ersten Blick nicht zur Intuition passt).

Die Ereignisse $B$ und $C$ sind *nicht* unabhängig.

**Bemerkung 1.50.** 1. Sind Ereignisse $A_1, A_2, \dots, A_n$ unabhängig, so gilt dies offenbar auch für jede Teilfamilie $A_{i_1}, \dots, A_{i_k}$ (mit $1 \le i_1 < \cdots < i_k \le n$).

2. Es genügt i.A. nicht, in (1.14) nur Paare zu prüfen.

Beispiel: Wir werfen dreimal eine faire Münze, mit Ergebnissen $W_1, W_2, W_3$, sei $A_1 = \{W_1 = W_2\}$, $A_2 = \{W_1 = W_3\}$, $A_3 = \{W_2 = W_3\}$.

Es ist $P(A_i) = \frac{1}{2}$, $P(A_i \cap A_j) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4} = P(A_i)P(A_j)$ für $i \neq j$, $i, j \in \{1, 2, 3\}$, aber $P(A_1 \cap A_2 \cap A_3) = \frac{1}{4} \neq P(A_1)P(A_2)P(A_3)$.

Man sagt dazu, dass die Ereignisse $A_i$ hier *paarweise unabhängig* sind.

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

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

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

### Unabhängige Zufallsvariablen

**Definition 1.51.** Zufallsvariablen $X_1, X_2, \dots, X_d$ (die auf einem gemeinsamen W'raum definiert sind) heißen (stochastisch) *unabhängig*, wenn für alle Ereignisse $\{X_i \in B_i\}$ gilt

$$P\big(X_1 \in B_1, X_2 \in B_2, \dots, X_d \in B_d\big) = \prod_{i=1}^{d} P(X_i \in B_i) \tag{1.15}$$

**Beobachtung 1.52.** *$X_1, X_2, \dots, X_n$ ZVn, $X_i$ habe Werte in $S_i$, $S_i$ abzählbar für $i = 1, 2, \dots, n$. Dann sind $X_1, X_2, \dots, X_n$ unabhängig g.d.w. gilt*

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

*(d.h. die gemeinsamen Gewichte haben Produktgestalt).*

*Beweis.* „$\Rightarrow$“: Klar (wähle $B_i = \{x_i\}$ in (1.15)).

„$\Leftarrow$“: Seien $B_1 \subset S_1, \dots B_n \subset S_n$, es ist

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

□

**Beobachtung und Definition 1.53.** *Für unabhängige Zufallsvariablen $X_1, \dots, X_n$ ist also die gemeinsame Verteilung, d.h. die Verteilung von $X := \big(X_1, \dots, X_n\big)$ durch die Randverteilungen festgelegt (siehe Def. 1.45).*

*Man nennt dann $\mu := \mathscr{L}(X)$ das* Produkt *der $\mu_1 := \mathscr{L}(X_1), \dots, \mu_n := \mathscr{L}(X_n)$ und schreibt*

$$\mu = \mu_1 \otimes \mu_2 \otimes \dots \otimes \mu_n$$

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

1. jede Teilfamilie $X_{i_1}, \dots, X_{i_k}$ (mit $1 \le i_1 < \dots < i_k \le n$)
   (wähle $B_i = S_i$ in (1.15) 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 (1.15))

3. In Def. 1.51 genügt es i.A. nicht, jeweils nur Paare auf Unabhängigkeit zu prüfen:

   Beispiel (wir sprechen Bsp. 1.50, 2. mit ZVn aus): Seien $X_1, X_2, X_3$ unabhängige faire Münzwürfe $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$, $Y_1 = \mathbf{1}_{\{X_1=X_2\}}, Y_2 = \mathbf{1}_{\{X_1=X_3\}}, Y_3 = \mathbf{1}_{\{X_2=X_3\}}$. 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.

4. Vergleichen wir Def. 1.49 und Def. 1.51, so sehen wir: Ereignisse $A_1, A_2, \dots, A_n$ sind genau dann unabhängig, wenn dies für ihre Indikatorvariable $\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, \dots, \mathbf{1}_{A_N}$ gilt.

### 1.3.4 Die Situation im Fall mit Dichten

**Beobachtung 1.55** (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. 1.52 lautet:

**Bericht 1.56** (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

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

   d.h. die gemeinsame Dichte hat Produktgestalt.

   Vergleicht man dies mit Beob. 1.55, 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 1.57.** 1. Wähle $A = [a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$ in Bsp. 1.36, 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 $A = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \le 1\}$ in Bsp. 1.36, 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

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

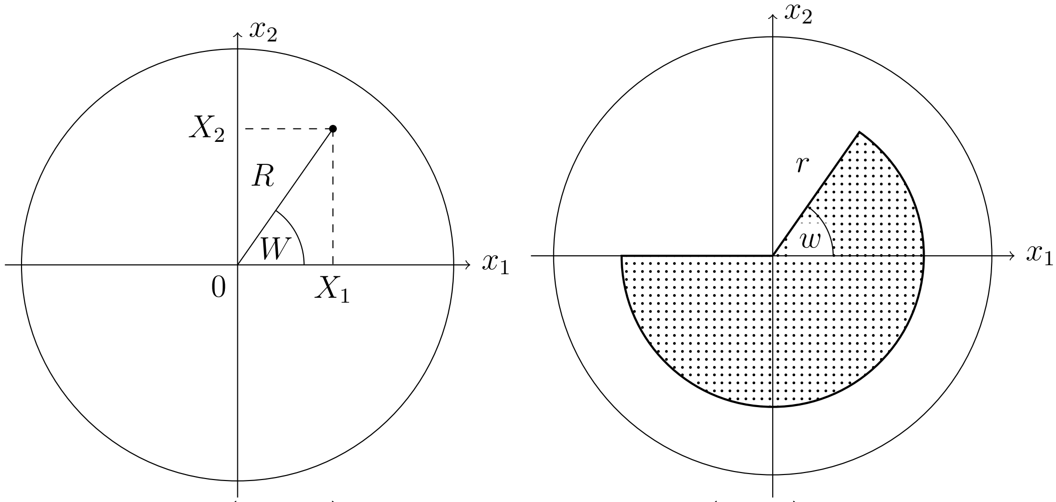
(und analog bzw. offensichtlich aus Symmetrie ist $f_{X_2} = f_{X_1}$).

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

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

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 $B(r, w) := \{\text{Punkte mit Radius} \le r \text{ und Winkel} \le w\}$.

Dann sind $R$ und $W$ unabhängig,

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

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. 1.57, 3. verallgemeinert sich folgendermaßen:

**Beobachtung 1.58.** *$X = (X_1, X_2)^T$ 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 $\alpha$ 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 1.59 beweisen.)*

*Sei $(R, W)$ die Polarkoordinatendarstellung von $X$ wie in Bsp. 1.57, 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 $2\pi r g(r) \mathbf{1}_{[0,\infty)}(r)$:*

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

*(wir verwenden hier etwas salopp die „Polarkoordinatenform des Flächenelements" $dx_1 dx_2 = r\,dr dw$, man kann dies wiederum mit Bericht 1.59 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. $R^2 \sim \mathrm{Exp}_{1/2}$.*

**Bericht 1.59** (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 $\varphi(x) = \big(\varphi_1(x), \dots, \varphi_d(x)\big)^T$, d.h. $\varphi_i$ ist die $i$-te Koordinatenfunktion von $\varphi$), dann hat $Y := \varphi(X)$ die Dichte

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

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

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

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

„aus wie“

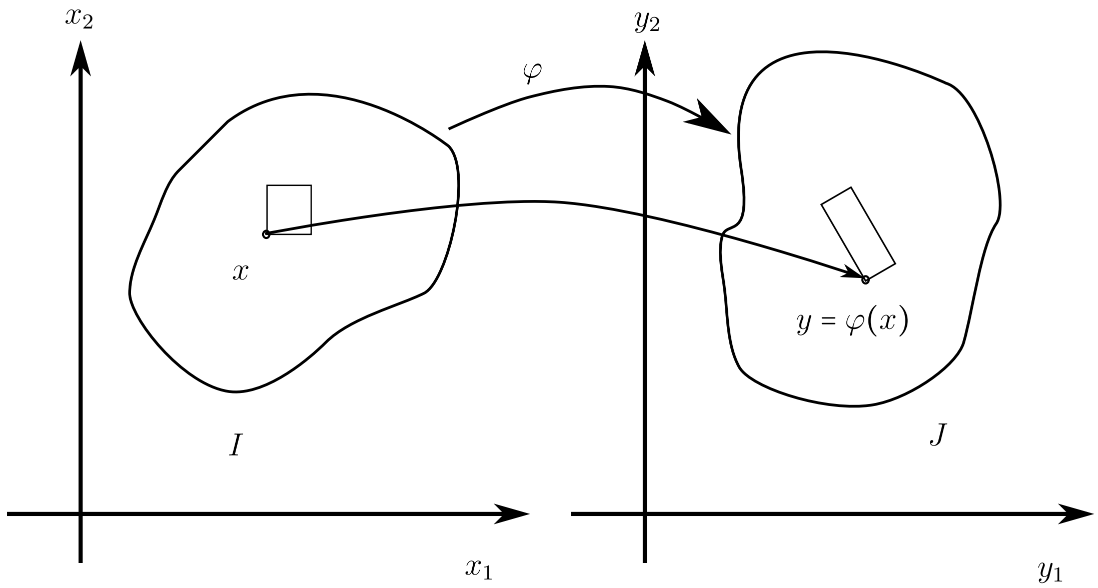
$$\begin{aligned} \varphi(x') &\approx \varphi(x) + \varphi'(x) \cdot (x' - x) \\ &= \varphi(x) + \begin{pmatrix} \frac{\partial}{\partial x_1}\varphi_1(x) & \frac{\partial}{\partial x_2}\varphi_1(x) \\ \frac{\partial}{\partial x_1}\varphi_2(x) & \frac{\partial}{\partial x_2}\varphi_2(x) \end{pmatrix}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

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

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

wird auf

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



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

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

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

**Bericht 1.60.** Analog betrachtet zu Def. 1.26 betrachtet man im Fall $d > 1$ für $(x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ gelegentlich die $d$-dimensionale Verteilungsfunktion

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

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

Analog zu Bem. 1.27, 3. „weiß“ die $d$-dimensionale Verteilungsfunktion von $X$ „alles“ über die Verteilung von $X$. Die $d$-dimensionale Verallgemeinerung der Eigenschaften aus Bem. 1.27, 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 $x_n \to (+\infty, \dots, +\infty)$
3. $F_X(x_n) \to 0$ wenn $\min_{i=1,\dots,d} x_{n,i} \to -\infty$
4. Für $x = (x_1, \dots, x_d) < y = (y_1, \dots, y_d)$ (koordinatenweise), parametrisiere die Ecken des $d$-Quaders $(x, y]$ mit $\{1,2\}^d$ via $\big\{u = (z_1^{(i_1)}, \dots, z_d^{(i_d)}) : i_1, \dots, i_d \in \{1,2\}\big\}$ wo $z_j^{(1)} = x_j$, $z_j^{(2)} = y_j$, es muss gelten

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

### 1.3.5 Faltung

**Definition 1.61.** $X$ und $Y$ unabhängige reellwertige ZVn, $X \sim \mu$, $Y \sim \nu$ (definiert auf demselben W'raum). Die Verteilung von $X + Y$ heißt die *Faltung* von $\mu$ und $\nu$, 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 1.62** (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 1.63.** 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 $k \in \{2, 3, \dots, 12\}$

(Wir hatten dies breits in dem Beispiel direkt nach Beob. 1.11 betrachtet.)

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 $X_1 + X_2 + \dots + X_n \sim \mathrm{Bin}_{n,p}$, d.h.

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

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

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 1.64** (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 1.65** (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. 1.58):

Seien $a, b \in (0,1)$ mit $a^2 + b^2 = 1$, so ist die $2 \times 2$-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. 1.58

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

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. 1.58 – in diesem Fall auch das Faltungsintegral explizit ausrechnen:

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

Für $z \in \mathbb{R}$ ist

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

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

## 1.4 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?"

### 1.4.1 Diskreter Fall

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

**Definition 1.66.** 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 1.67.** 1. $A$ ein Ereignis, so ist $\mathbb{E}[\mathbf{1}_A] = 1 \cdot P(\mathbf{1}_A = 1) + 0 \cdot P(\mathbf{1}_A = 0) = P(A)$.

2. $W$ 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

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

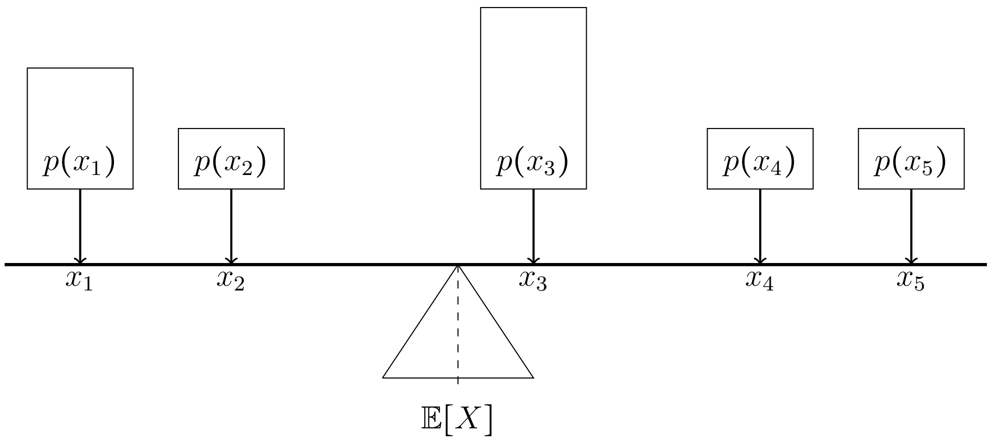
(denn $\sum_{j=1}^{3N} x_j P(X = x_j) = \sum_{i=2}^{N} \frac{-i}{2i(i-1)} + \sum_{i=2}^{2N} \frac{i}{2i(i-1)} = -\frac{1}{2} \sum_{k=1}^{N-1} \frac{1}{k} + \frac{1}{2} \sum_{k=1}^{2N-1} \frac{1}{k} \sim -\frac{1}{2} \log(N) + \frac{1}{2} \log(2N) = \frac{1}{2} \log(2)$).

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

**Bemerkung 1.68.** 1. Eine beschränkte reellwertige ZV $X$ (d.h. es gibt eine Konstante $M < \infty$ mit $P(-M \leq X \leq M) = 1$) 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. 1.67), daher kann man die Interpretation von $\mathbb{E}[X]$ als „typischer Wert von $X$" i.A. nicht wörtlich nehmen.

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

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

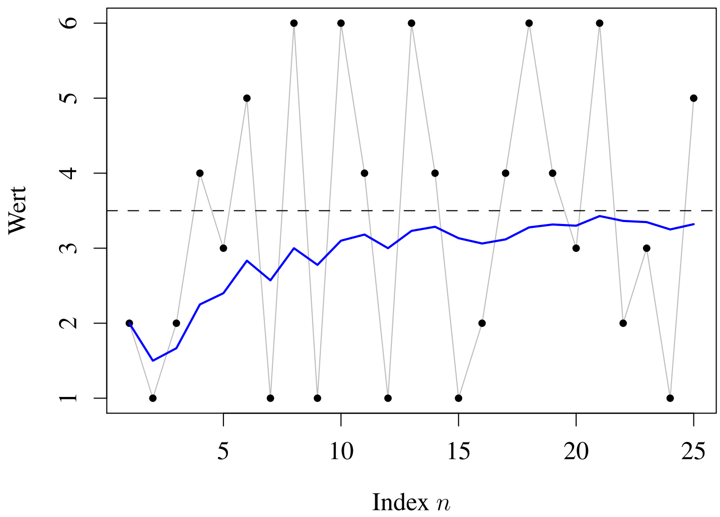
   (in geeignetem Sinn), dies ist die Aussage des *Gesetzes der großen Zahlen*, das wir später sehen werden.

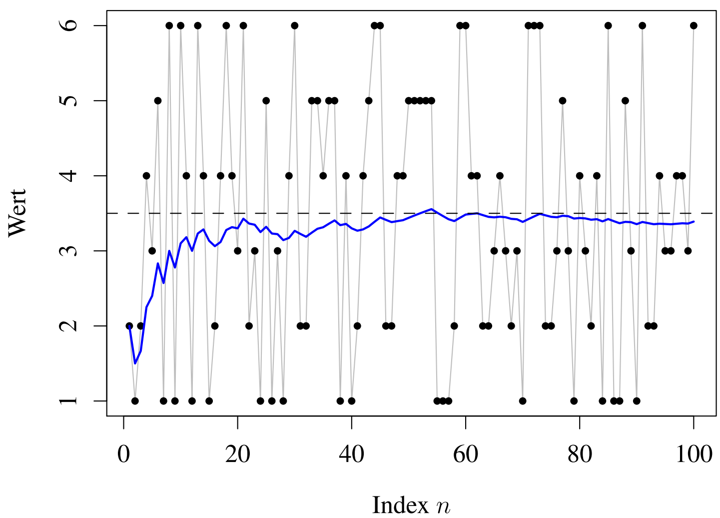
   Es ist nämlich

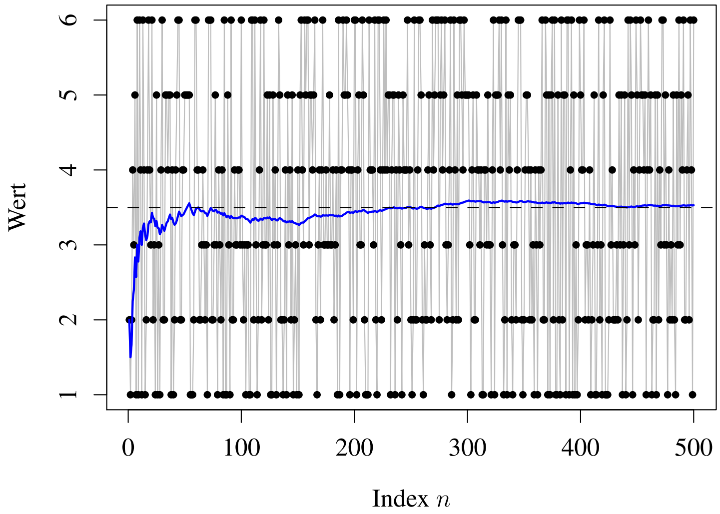
   $$M_n = \sum_x x \cdot \frac{\#\{i \le n : X_i = x\}}{n}$$

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

Illustration: $X_1, X_2, \dots$ 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 1.69.** 1. Sei $X \sim \mathrm{Bin}_{n,p}$, $n \in \mathbb{N}$, $p \in [0,1]$ :

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

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

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

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 1.70** (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

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

2.

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

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

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

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

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

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

□

**Beobachtung 1.71** (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 $\sum_y y P(Y = y) = \sum_y \sum_{x : g(x) = y} g(x) P(X = x) = \sum_x g(x) P(X = x)$.)

**Beispiel 1.72.** 1. Seien $X_1, \dots, X_n$ u.i.v., $\sim \mathrm{Ber}_p$, so ist $X := X_1 + \dots + X_n \sim \mathrm{Bin}_{n,p}$ und

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

(Wir hatten den Erwartungswert einer binomialverteilten ZV bereits in Bsp. 1.69, 2. bestimmt, hier kommen wir allerdings ohne explizite Rechnung aus.)

2. Sei $X \sim \mathrm{Hyp}_{s,w,k}$ hypergeometrisch verteilt, vgl. Bsp. 1.15. Denken wir an eine Urne mit $s$ schwarzen und $w$ weißen Kugeln, aus der $k$ mal ohne Zurücklegen gezogen wird, so ist

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

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

### 1.4.2 Der Fall mit Dichte

**Definition 1.73.** 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 1.74.** 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. 1.32).

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 1.75.** 1. Man kann prinzipiell den Fall mit Dichte aus dem diskreten Fall herleiten: $X$ habe Dichte $f_X$, so nimmt $X_{(n)} = \dfrac{1}{n}\lfloor nX \rfloor$ den Wert $\dfrac{k}{n}$, $k \in \mathbb{N}$ 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. 1.71 im Fall mit Dichte)
Sei $X = (X_1,\dots,X_d)$ $\mathbb{R}^d$-wertig mit Dichte $f_X : \mathbb{R}^d \to [0,\infty]$, $g : \mathbb{R}^d \to \mathbb{R}$, $Y := g(X)$. Dann gilt $Y \in \mathcal{L}^1$ g.d.w.

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

und in diesem Fall

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

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

3. Die Rechenregeln aus Satz 1.70 gelten auch im Fall mit Dichte.

### 1.4.3 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 1.76.** 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 $\sigma_X$ 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 1.77.** *Sei $X$ reelle ZV, $f : [0,\infty) \to [0,\infty)$ monoton wachsend.*
*1. Für $a > 0$ mit $f(a) > 0$ gilt*

$$P\big(|X| \geq a\big) \leq \frac{1}{f(a)}\mathbb{E}\big[f(|X|)\big] \qquad \textit{(Markov}^{12}\textit{-Ungleichung).} \tag{1.16}$$

*2. Für $X \in \mathcal{L}^2$ gilt*

$$P\big(|X - \mathbb{E}[X]| \geq a\big) \leq \frac{\mathrm{Var}[X]}{a^2} \qquad \textit{(Chebyshev}^{13}\textit{-Ungleichung).} \tag{1.17}$$

*Beweis. 1.* Sei $Y := f(a)\mathbf{1}_{\{|X|\geq a\}}$, so ist $Y \leq f(|X|)$ und

$$\mathbb{E}[Y] = f(a)P\big(|X| \geq a\big) \leq \mathbb{E}\big[f(|X|)\big]$$

nach Satz 1.70, 2.

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

Insbesondere (wähle $a = b\sigma_X$ in (1.17)): Die W'keit $P\big(|X - \mathbb{E}[X]| \geq b\,\sigma_X\big)$, dass $X$ von $\mathbb{E}[X]$ um mehr als das $b$-fache von $\sigma_X$ abweicht, ist höchstens $1/b^2$.

**Beobachtung 1.78.**

1. Wegen $|XY| \leq X^2 + Y^2$ ist die Kovarianz wohldefiniert. Es gilt (offensichtlich)
$$\mathrm{Cov}[X,Y] = \mathrm{Cov}[Y,X].$$
2. Es ist
$$\begin{aligned}\mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big] &= \mathbb{E}\big[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]\big]\\ &= \mathbb{E}[YX] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]\\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$
(und analog für $\mathrm{Var}[X] = \mathrm{Cov}[X,X]$).
3. $\mathrm{Var}[X] = 0 \iff P\big(X = \mathbb{E}[X]\big) = 1$
(„$\Leftarrow$“ ist klar, für „$\Rightarrow$“ wende Satz 1.70, 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 1.79.** 1. $X \sim \mathrm{Ber}_p$, $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p)$.

2. $X \sim \mathrm{Poi}_\alpha$,

$$\mathbb{E}[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)e^{-\alpha}\frac{\alpha^k}{k!} = \alpha^2 \sum_{k=2}^{\infty} e^{-\alpha}\frac{\alpha^{k-2}}{(k-2)!} = \alpha^2,$$

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

also

$$\operatorname{Var}[X] = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2 = \mathbb{E}\big[X(X-1)\big] + \mathbb{E}[X] - \big(\mathbb{E}[X]\big)^2 = \alpha^2 + \alpha - \alpha^2 = \alpha$$

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

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

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. 1.17 und wir hatten gesehen, dass $\mathbb{E}[X] = (1-p)/p$, siehe Bsp. 1.69, 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 $f(t) := \sum_{n=0}^{\infty} t^n = \frac{1}{1-t}$ (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

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

5. $X \sim \mathcal{N}_{\mu,\sigma^2}$ hat $\operatorname{Var}[X] = \sigma^2$ (wir hatten in Bsp. 1.74 bereits gesehen, dass $\mathbb{E}[X] = \mu$):

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

(Wir haben Bericht 1.75, 2. verwendet, dann im Integral $z = (x - \mu/\sigma)$ substituiert und partiell integriert.)

**Satz 1.80** (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*

$$\big|\mathrm{Cov}[X, Y]\big| \le \sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]} \quad \textit{(Cauchy-Schwarz-Ungleichung}^{14}\textit{)}$$

*Beweis.* 1. Es ist

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

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

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

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

3. Klar, denn für $X$ und $Y$ unabhängig ist $\mathbb{E}[X\,Y] = \mathbb{E}[X]\,\mathbb{E}[Y]$ nach Satz 1.70, 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. 1.78, 3. und somit auch $\mathrm{Cov}[X, Y] = 0$).

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

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 1.81.** 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 1.82.** 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 1.80, 2.)

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

(vgl. auch Bsp. 1.79, 3.).

2. $X \sim \mathrm{Hyp}_{s,w,n}$, stelle dar als $X = Y_1 + \cdots + Y_n$ mit $Y_i = \mathbf{1}_{A_i}$, $A_i = \{i\text{-te gezogene Kugel ist schwarz}\}$ (bei $n$-fachem Ziehen ohne 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 $\mathbb{E}[Y_i] = P(A_i) = P(A_1) = \dfrac{s}{s+w} =: p$, $\mathrm{Var}[Y_i] = p(1-p)$; für $i \neq j$ ist

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

also

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

(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

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

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

**Definition 1.83.** Seien $X, Y \in \mathcal{L}^2$.

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

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

(Die Cauchy-Schwarz-Ungleichung (Satz 1.80, 4.) zeigt, dass $|\kappa_{X,Y}| \leq 1$.)

**Beobachtung 1.84** (Interpretation des Korrelationskoeffizienten via „beste lineare Vorhersage“)**.** Es ist

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

denn der Ausdruck auf der linken Seite ist

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

was offensichtlich minimal wird für die Wahl

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

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

(Für den Zusatz beachte analog:

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

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

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

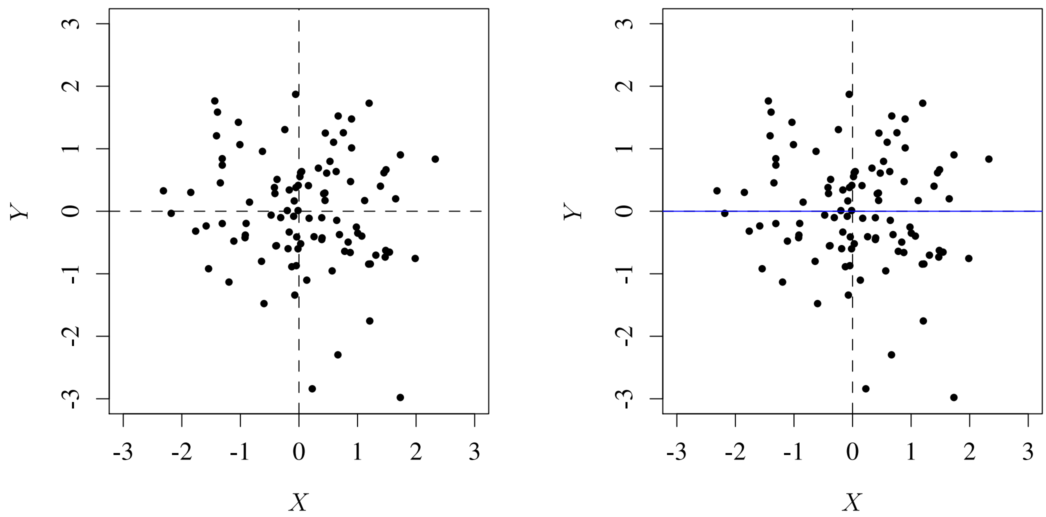
Demnach (vgl. auch Bem. 1.81)

| | | |
|---|---|---|
| $\lvert\kappa_{X,Y}\rvert = 1$ | $\leftrightarrow$ | 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 $\mathbb{E}[Y]$) |

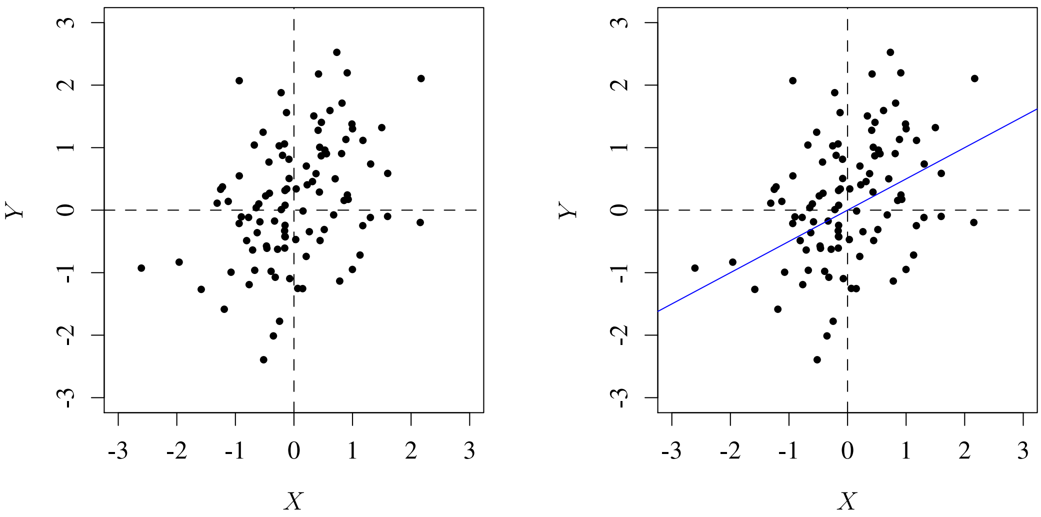
Nicht-lineare Zusammenhänge erfasst der Korrelationskoeffizient möglicherweise nicht korrekt (oder gar nicht), vgl. Bsp. 1.82, 3.

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

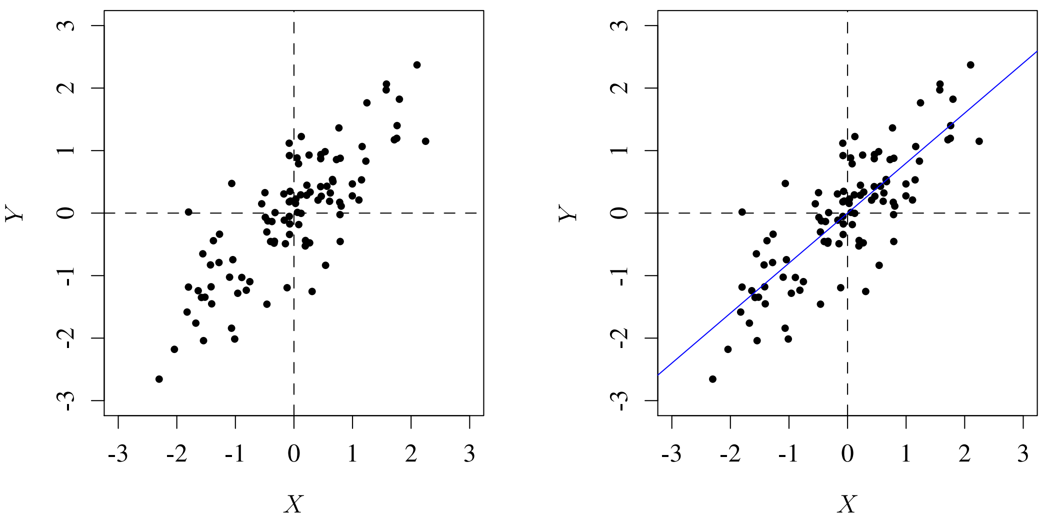
$\kappa_{X,Y} = 0$



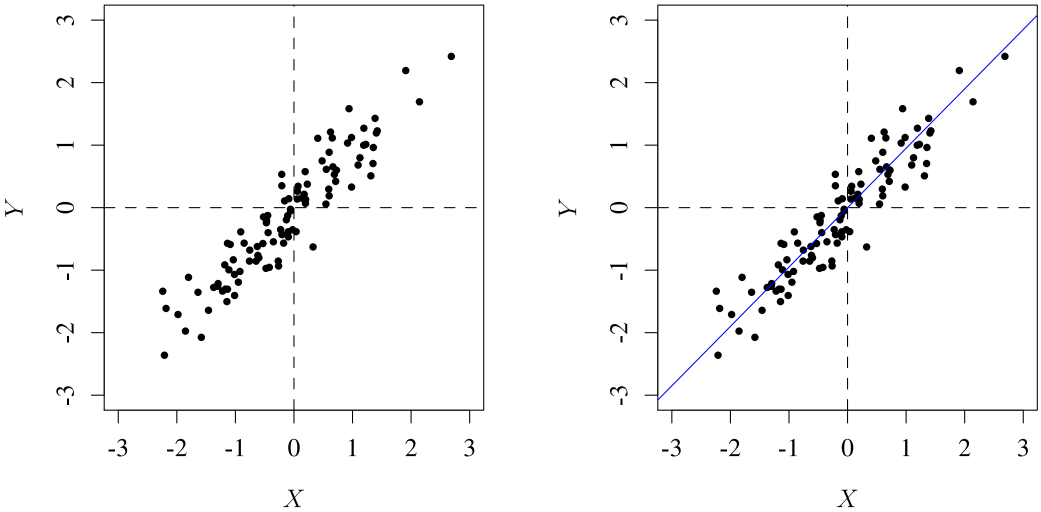
$\kappa_{X,Y} = 0.5$



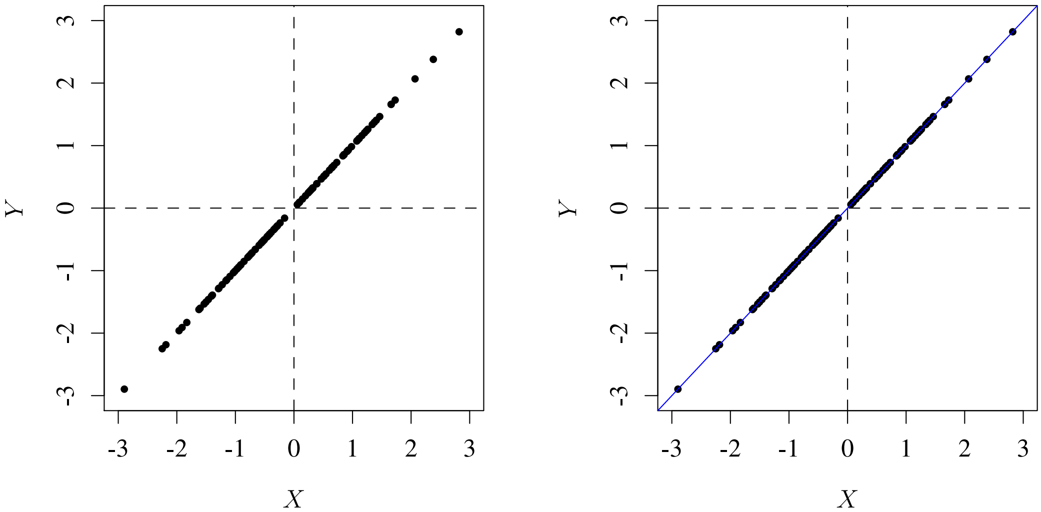

$\kappa_{X,Y} = 0.8$



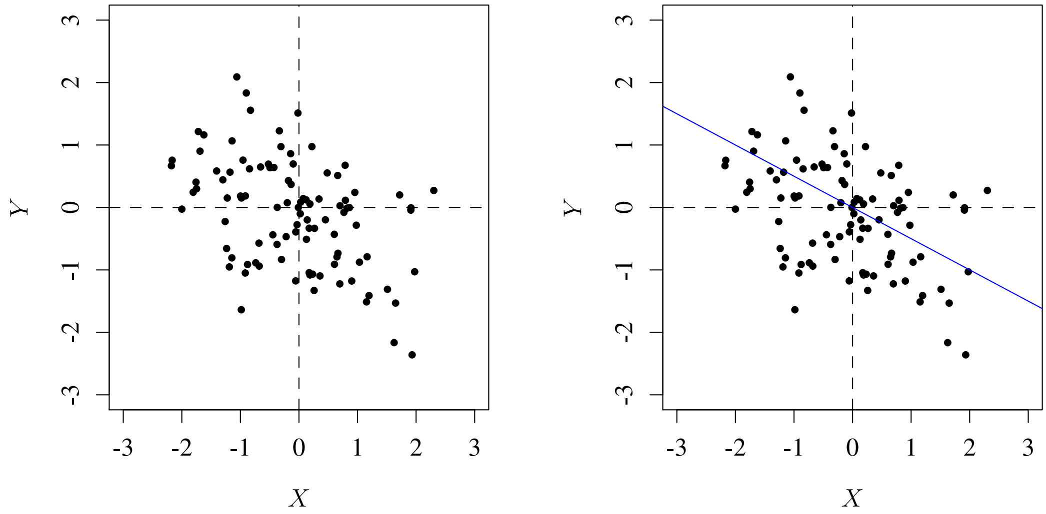
$\kappa_{X,Y} = 0.95$



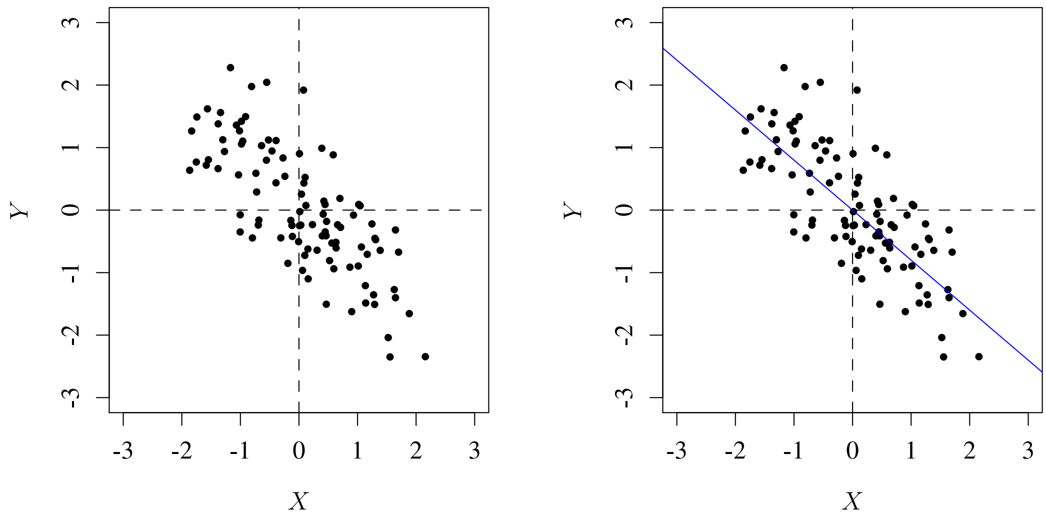

$\kappa_{X,Y} = 1$



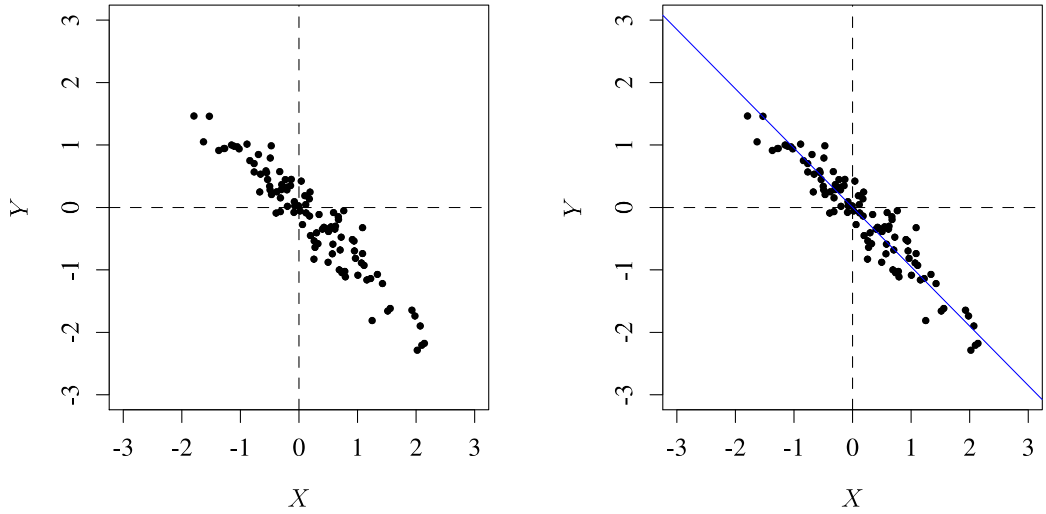
$\kappa_{X,Y} = -0.5$



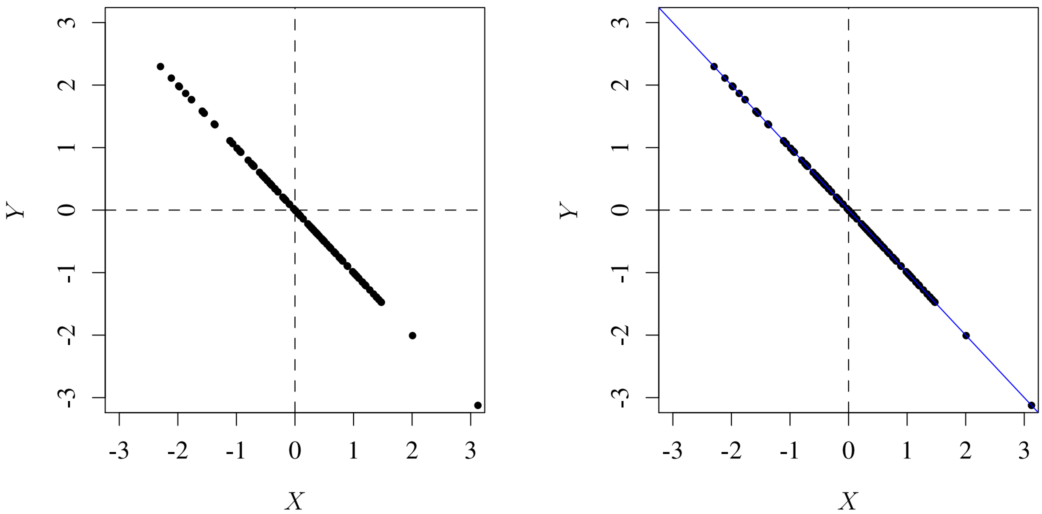

$\kappa_{X,Y} = -0.8$

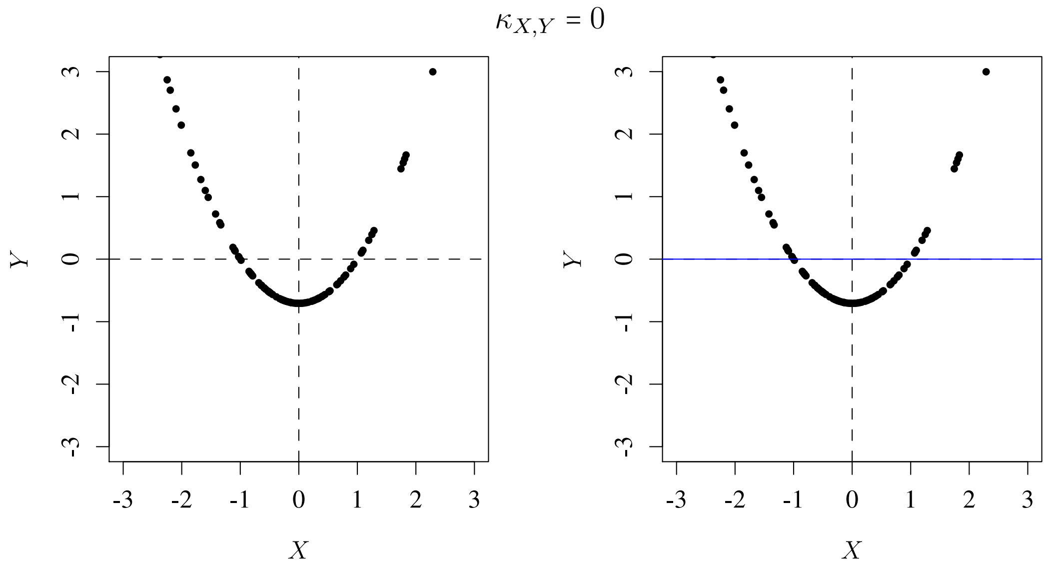


$\kappa_{X,Y} = -0.95$




$\kappa_{X,Y} = -1$






### 1.4.4 Median(e)

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

$$„P(X \le m) = \frac{1}{2} = P(X \ge m)“$$

gilt.

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

**Definition 1.85.** $X$ reelle ZV, $m$ heißt (ein) Median von $X$ (auch „Zentralwert“, manchmal auch $m_X$ geschrieben), wenn gilt

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

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

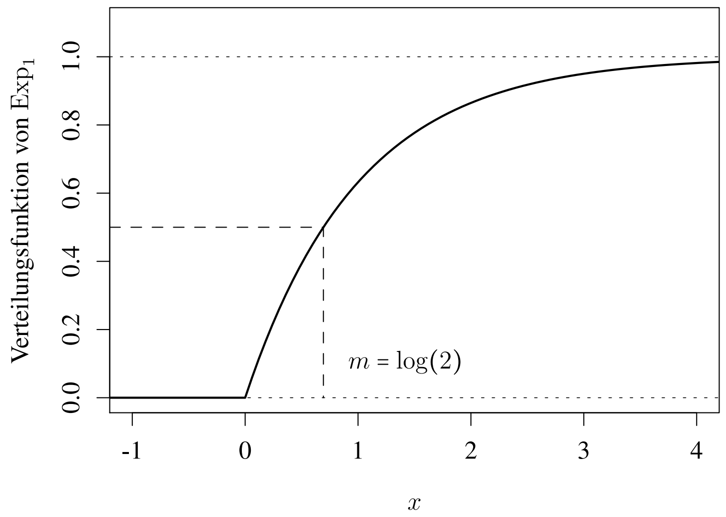
Man kann den Median als eine „robustere“ Antwort auf die Aufgabe, für eine ZV *nur einen* „typischen Wert“ anzugeben, ansehen (im Gegensatz zum Erwartungswert besitzt ja jede Verteilung einen Median). Allerdings gibt es für Mediane keine so angenehmen Rechenregeln, wie sie Satz 1.70 für den Erwartungswert liefert.

**Bemerkung.** Falls der Median von $X$ (im Sinne von Def. 1.85) uneindeutig ist, d.h. wenn die Verteilungsfunktion $F_X$ von $X$ den Wert $\frac{1}{2}$ auf einem nicht-trivialen Intervall annimmt, so betrachtet man gelegentlich „pragmatisch“

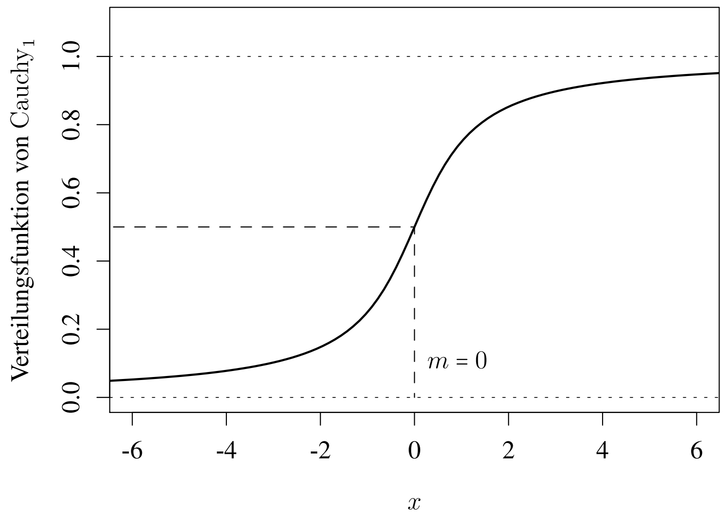
$$\frac{1}{2}\Big( \inf\{x \in \mathbb{R} : F_X(x) = \tfrac{1}{2}\} + \sup\{x \in \mathbb{R} : F_X(x) = \tfrac{1}{2}\}\Big),$$

das arithmetische Mittel des kleinst- und des größtmöglichen Medians, als „den“ Median. (So berechnet es beispielsweise `R`.)

**Beispiel 1.86.** 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 $m = \frac{1}{\theta}\log 2$.

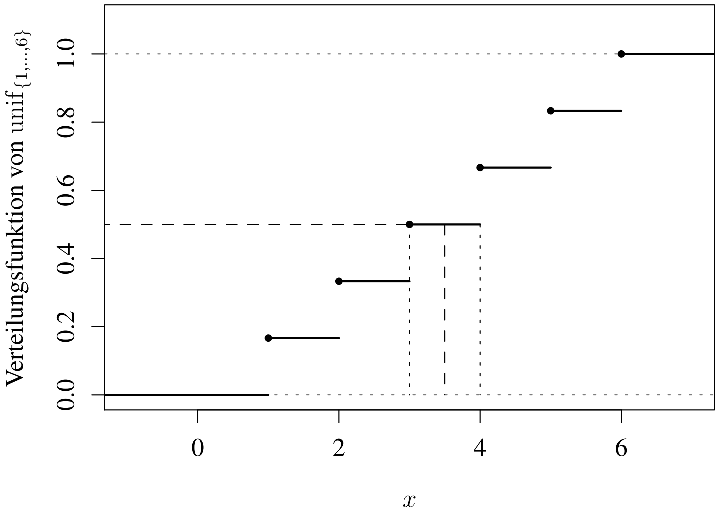



2. $X$ Cauchy-verteilt mit Dichte $\frac{1}{\pi}\frac{1}{1+x^2}$, Verteilungsfunktion $\frac{1}{2}\frac{1}{\pi}\arctan(x)$, der (eindeutig bestimmte) Median ist $m = 0$.



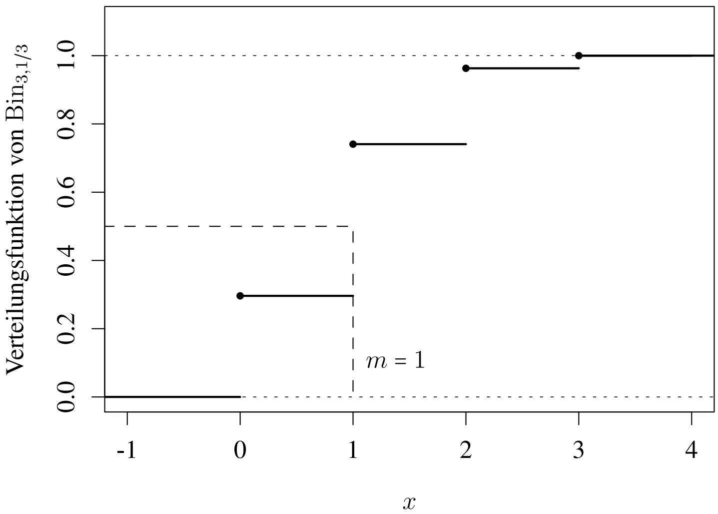

(Wegen der Symmetrie der Dichte, es gibt keinen Erwartungswert, vgl. Bsp. 1.74, 3.)

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




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

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




**Bemerkung 1.87.** 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 1.70, 2., beachte: $X-m \le |X-m|$ und $-(X-m) \le |X-m|$ und für die letzte Ungleichung die Cauchy-Schwarz-Ungleichung (Satz 1.80, 4.) verwenden. □

## 1.5 Gesetz der großen Zahlen und zentraler Grenzwertsatz

### 1.5.1 Gesetz der großen Zahlen

**Satz 1.88** ((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 $\varepsilon > 0$*

$$P\left(\left|\frac{X_1+\cdots+X_n}{n} - \mu\right| > \varepsilon\right) \le \frac{\mathrm{Var}[X_1]}{\varepsilon^2 n} \underset{n\to\infty}{\longrightarrow} 0 \tag{1.18}$$

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

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

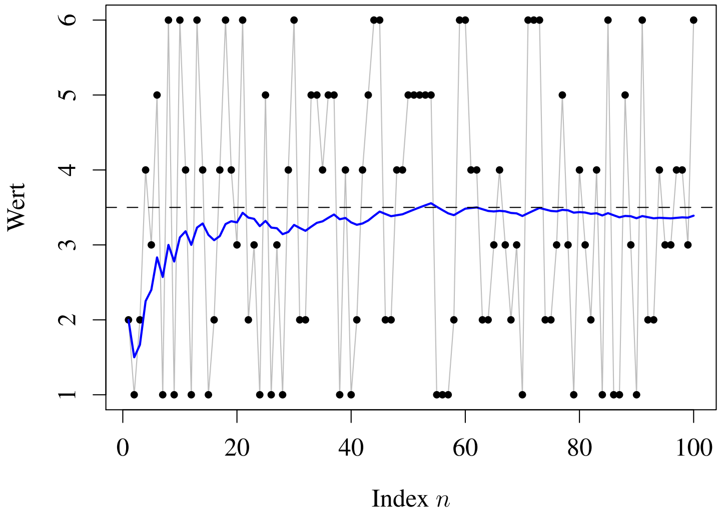
mit Satz 1.80, somit

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

gemäß Chebyshev-Ungleichung (Satz 1.77). □

**Erinnerung.** Wir hatten bereits in Bem. 1.68, 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 1.89.** 1. Wir entnehmen dem Beweis von Satz 1.88 folgende kleine Verallgemeinerung:

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

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

dann gilt für $\varepsilon > 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathbb{E}[X_i]\right)\right| > \varepsilon\right) \le \frac{\theta}{\varepsilon^2 n} \left(\xrightarrow[n\to\infty]{} 0\right). \tag{1.19}$$

(Das Argument geht genauso wie im Beweis von Satz 1.88, wenn wir $Y_n := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathbb{E}[X_i]\right)$ setzen.)

2. Seien $Y_n, n \in \mathbb{N}$ und $Y$ reellwertigen ZVn, die auf einem gemeinsamen Wahrscheinlichkeitsraum definiert sind.

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

$$Y_n \xrightarrow[n\to\infty]{\text{stoch.}} Y,$$

(auch $Y_n \xrightarrow[n\to\infty]{} Y$ stoch. oder $Y_n \xrightarrow[n\to\infty]{P} Y$), wenn gilt

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

Man spricht damit Satz 1.88 oft folgendermaßen aus:

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow[n\to\infty]{\text{stoch.}} \mu$$

**Bericht 1.90** (Nur der Vollständigkeit halber)**.** Die Konvergenzaussage (1.18) in Satz 1.88 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:

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

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

Wir werden dies im weiteren Verlauf der Vorlesung nicht verwenden.

### 1.5.2 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 $X_1 + \dots + X_n \approx n\mu$ mit hoher Wahrscheinlichkeit, denn

$$\frac{X_1 + \dots + X_n}{n} - \mu \underset{n\to\infty}{\longrightarrow} 0 \quad \text{stochastisch}$$

gemäß dem Gesetz der großen Zahlen (Satz 1.88), aber feiner gefragt:

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

Für $A \gg \sqrt{n}$ ist (mit Chebyshev-Ungleichung, Satz 1.77) zumindest

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

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

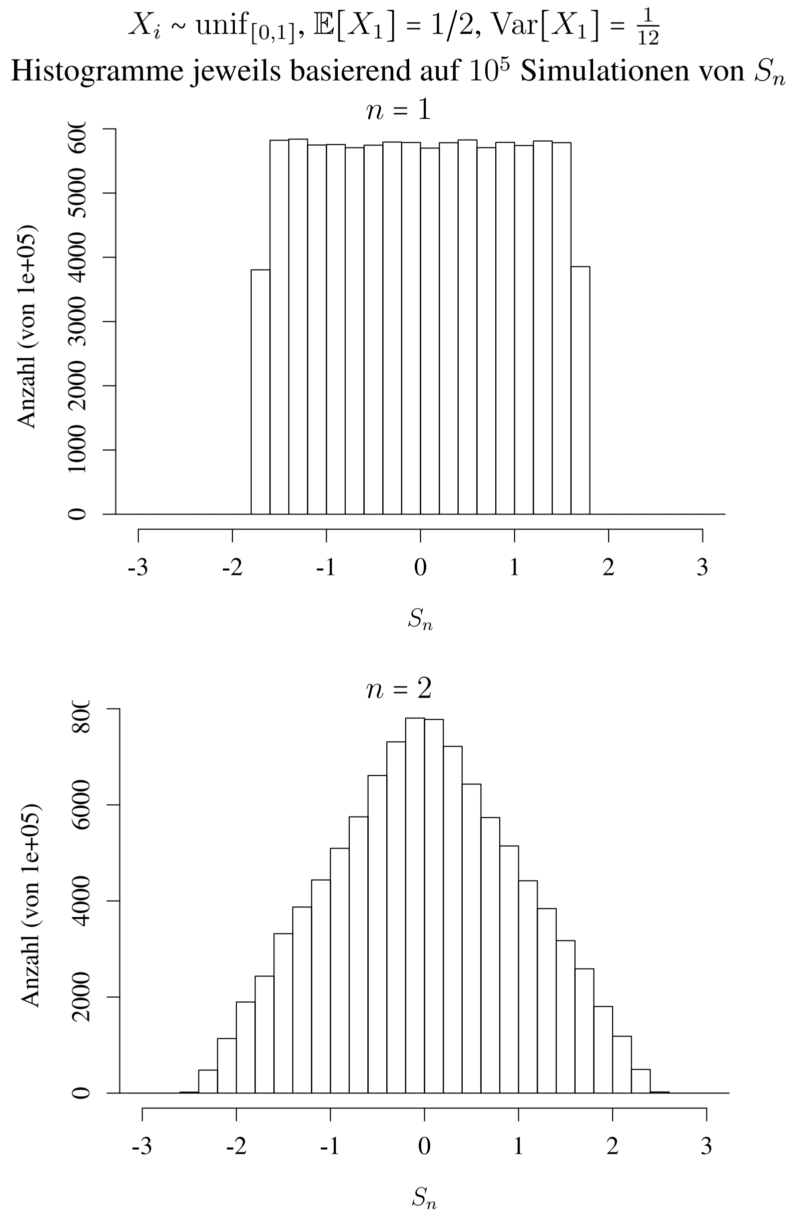
$$\mathbb{E}\big[(X_1 + \dots + X_n) - n\mu\big] = 0$$

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

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

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:



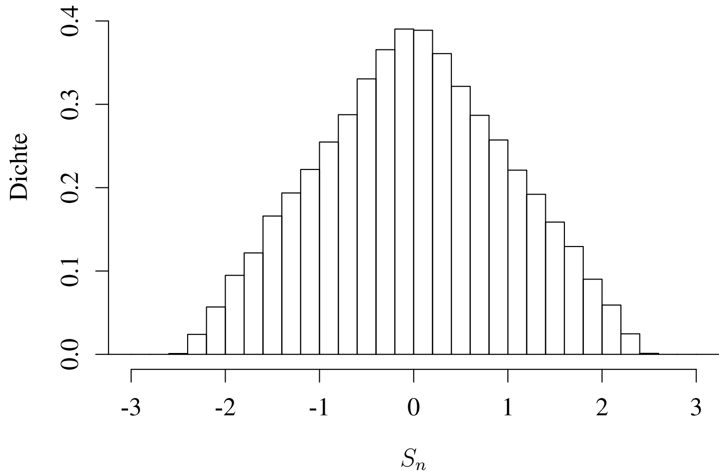

Zur Vergleichbarkeit gehen wir von den absoluten Anzahlen als Balkenhöhen zur sogenannten Dichte über, d.h. die Balkenhöhe ist nun jeweils

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

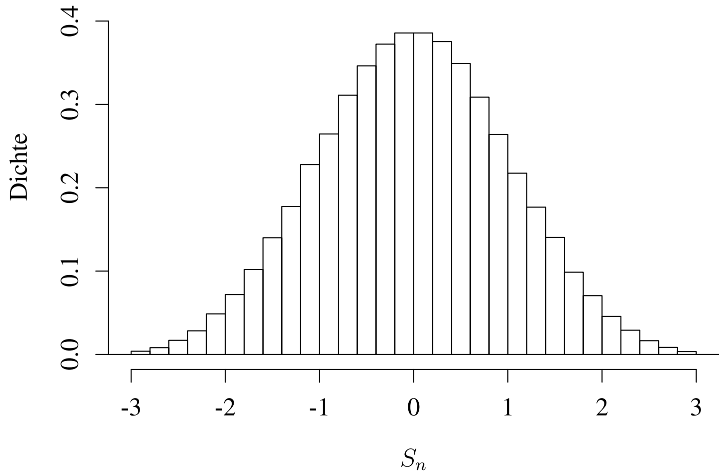
Damit wird die Gesamtfläche der Balken = 1 (wie bei einer Wahrscheinlichkeitsdichte).

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

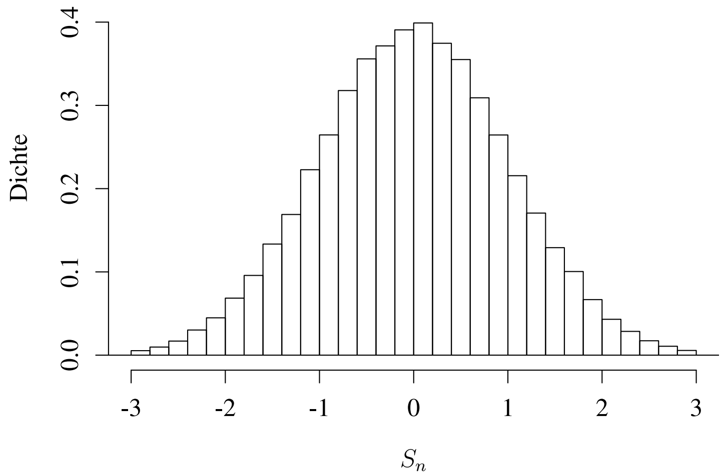
$n = 2$


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

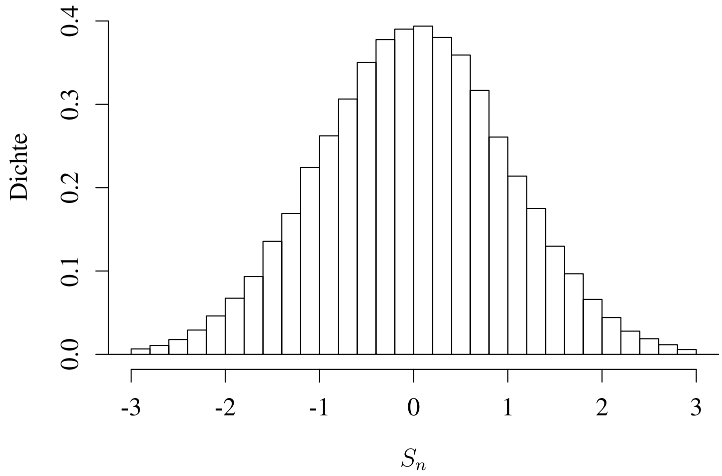
$n = 5$


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

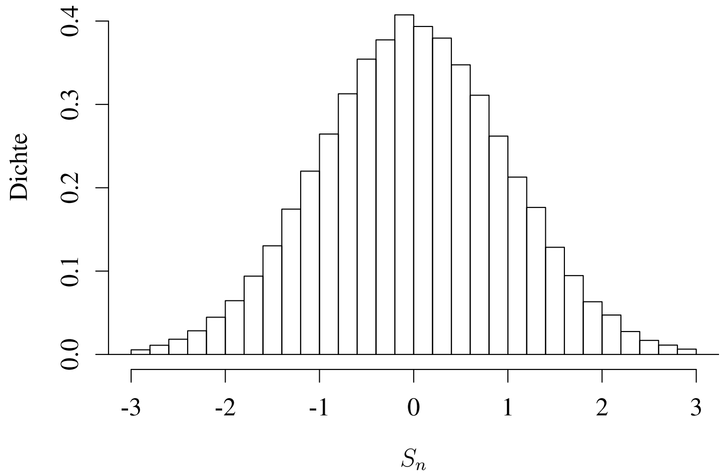
$n = 20$


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

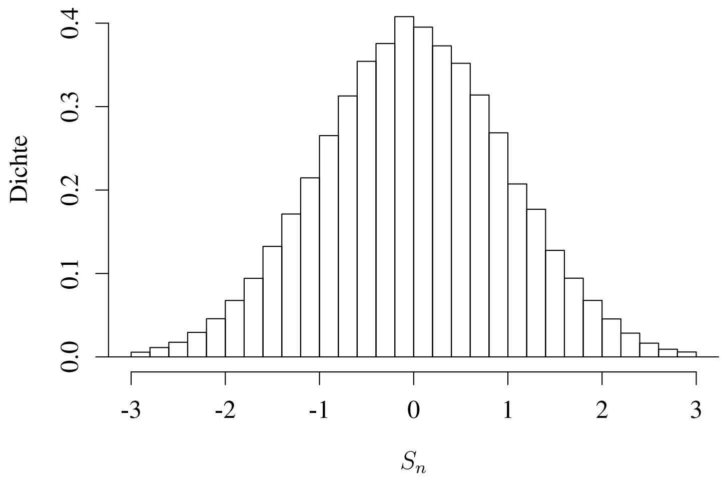
$n = 100$



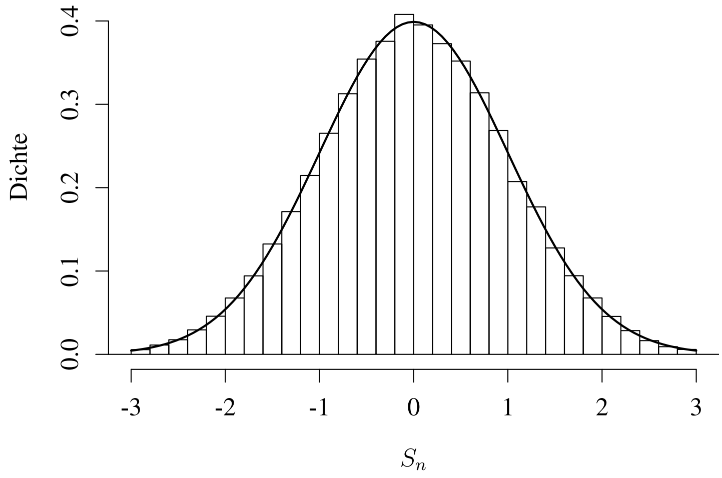

$n = 200$




$n = 500$



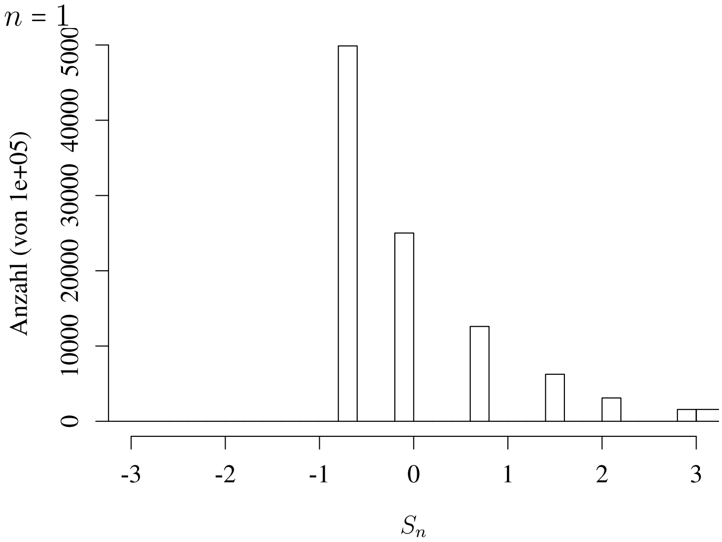
$n = 500$

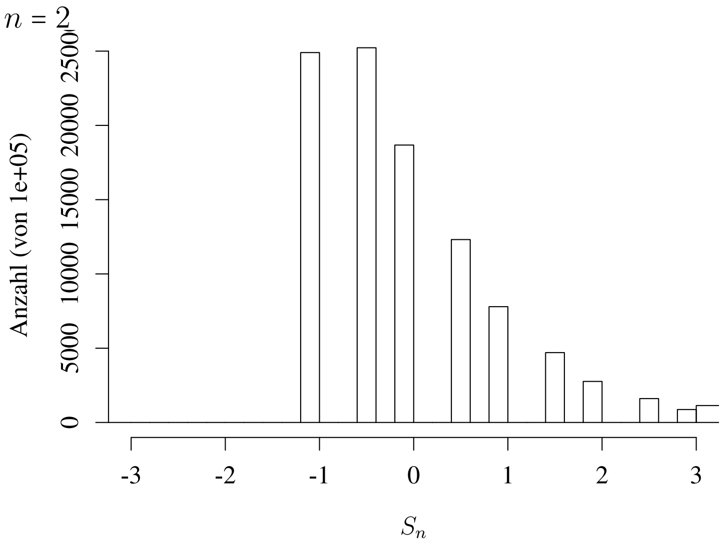



(Die schwarze Kurve ist die Dichte der Standard-Normalverteilung, $f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$.)

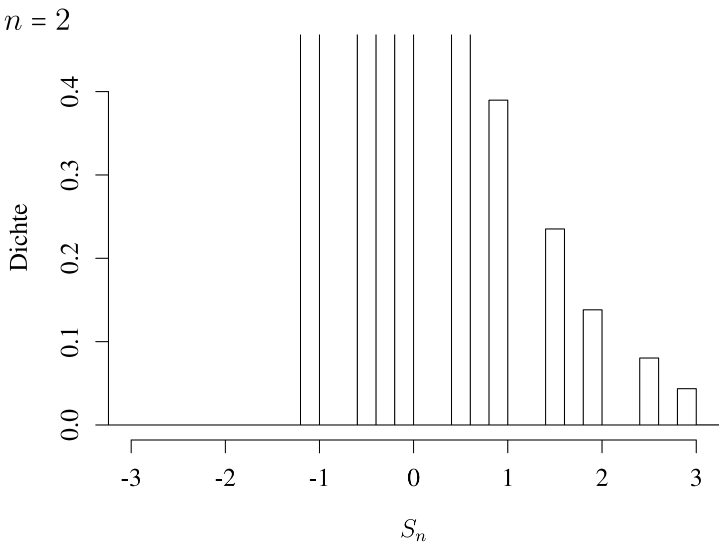
Nun dasselbe nochmal mit $X_i \sim \mathrm{geom}_{1/2}$ (mit $\mathbb{E}[X_1] = 1$, $\mathrm{Var}[X_1] = 2$):
Histogramme jeweils basierend auf $10^5$ Simulationen von $S_n$

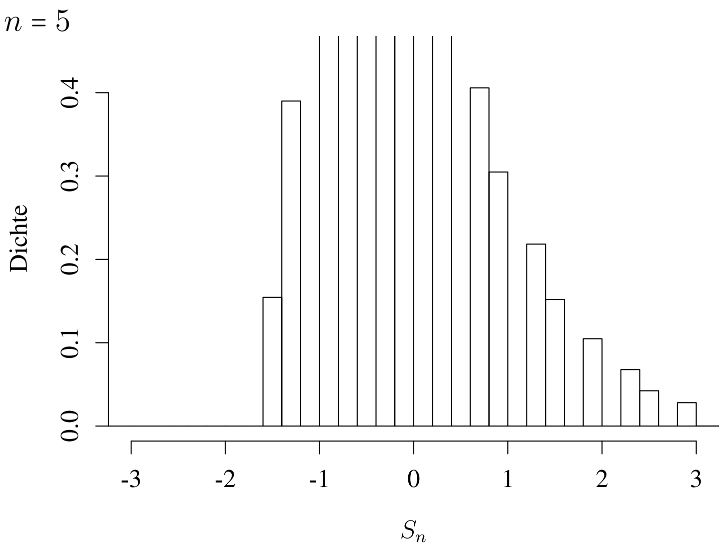
$n = 1$



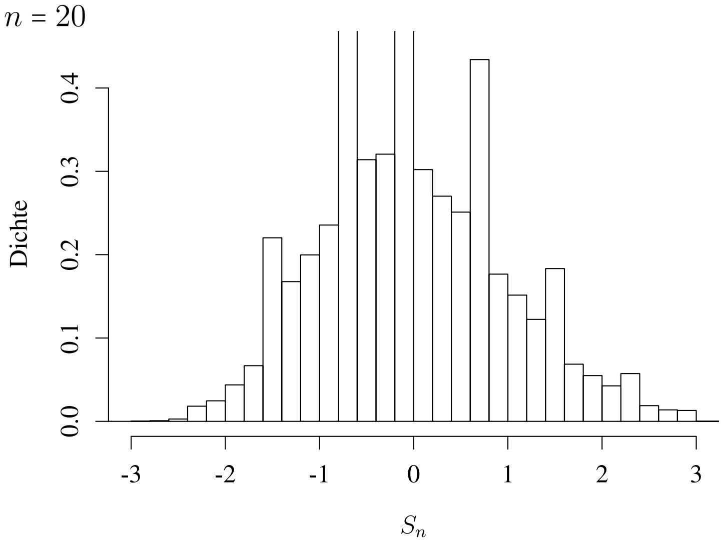



Zur Vergleichbarkeit gehen wir wieder von den absoluten Anzahlen als Balkenhöhen zur sogenannten Dichte über.

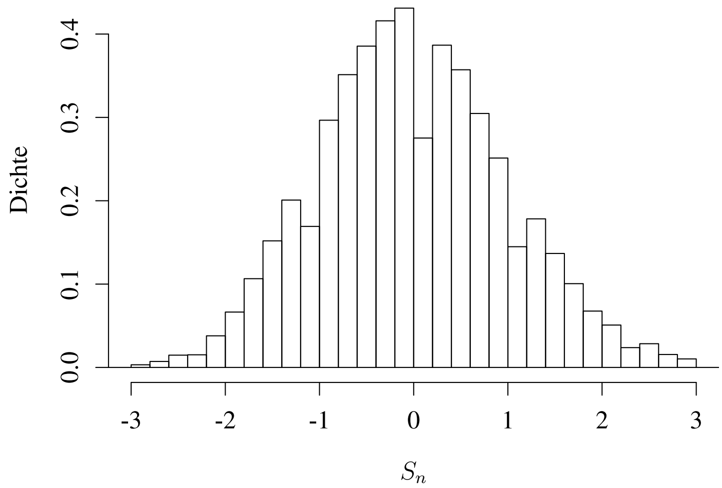


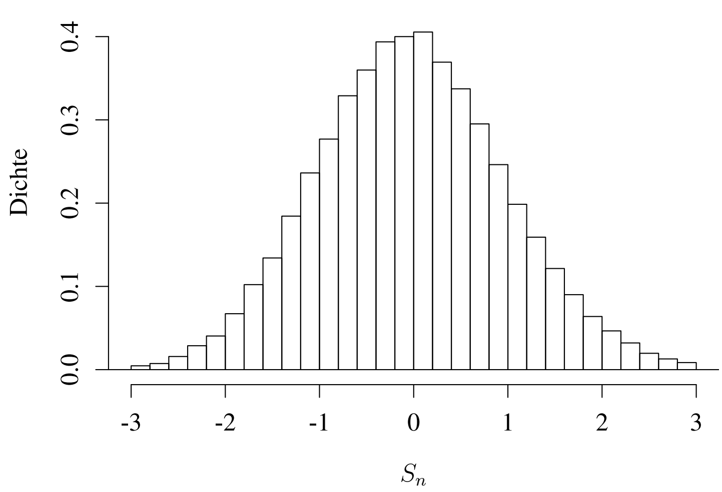
$n = 20$

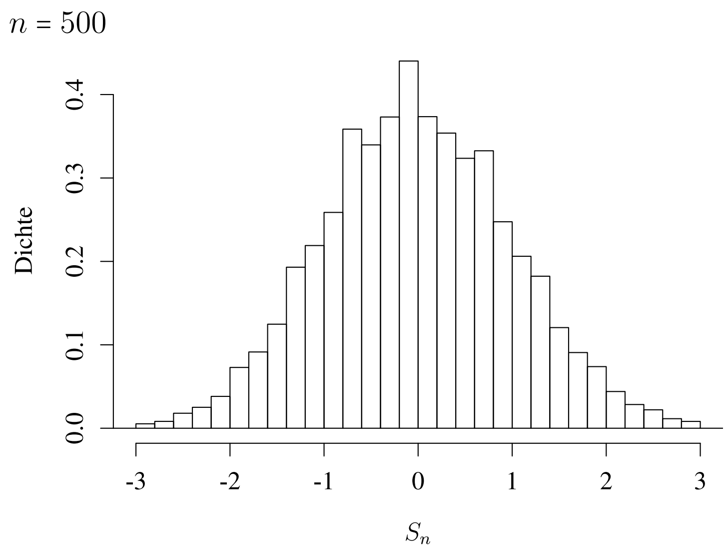


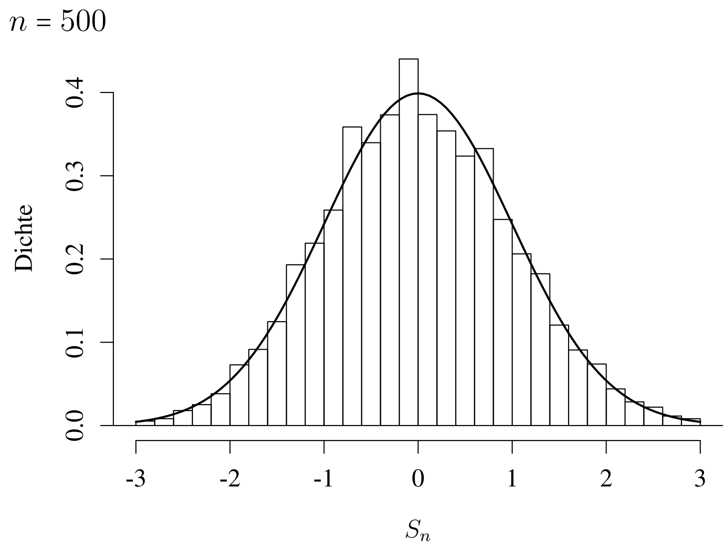

$n = 100$




$n = 200$






(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

$$\frac{(X_1 + \cdots + X_n) - n\mu_{X_1}}{\sqrt{n\sigma^2_{X_1}}} \overset{d}{\approx} Z \quad \text{mit } Z \sim \mathcal{N}_{0,1}. \tag{1.20}$$

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 (1.20) exakt (dies folgt aus Bsp. 1.65 und Bsp. 1.32, 1.)

**Satz 1.91** („Zentraler Grenzwertsatz"). *Seien $X_1, X_2, \ldots$ u.i.v. reelle ZVn $\in \mathscr{L}^2$ mit* $\mathrm{Var}[X_1] \in (0,\infty)$*, dann gilt für* $-\infty \le a < b \le \infty$

$$\lim_{n\to\infty} P\Big(a \le \frac{X_1 + \cdots + X_n - n\mathbb{E}[X_1]}{\sqrt{n\mathrm{Var}[X_1]}} \le b\Big) = P(a \le Z \le b) \quad \textit{mit } Z \sim \mathcal{N}_{0,1}. \tag{1.21}$$

Die wichtige Botschaft von Satz 1.91 lautet: Eine Summe von vielen unabhängigen und identisch verteilten zufälligen Summanden ist (approximativ) normalverteilt.

**Bemerkung 1.92.** Die Eigenschaft (1.21) wird auch ausgesprochen als „Konvergenz in Verteilung": $X, X_n$ reellwertige ZVn, so sagt man

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

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 1.91 besagt also: $\dfrac{X_1 + \cdots + X_n - n\mathbb{E}[X_1]}{\sqrt{n\mathrm{Var}[X_1]}} \overset{d}{\underset{n\to\infty}{\longrightarrow}} Z$

**Bericht 1.93.** Es gibt viele Verallgemeinerungen von Satz 1.91, die die Annahme, dass die $X_i$ u.i.v. sind, (stark) abschwächen.

### 1.5.3 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:

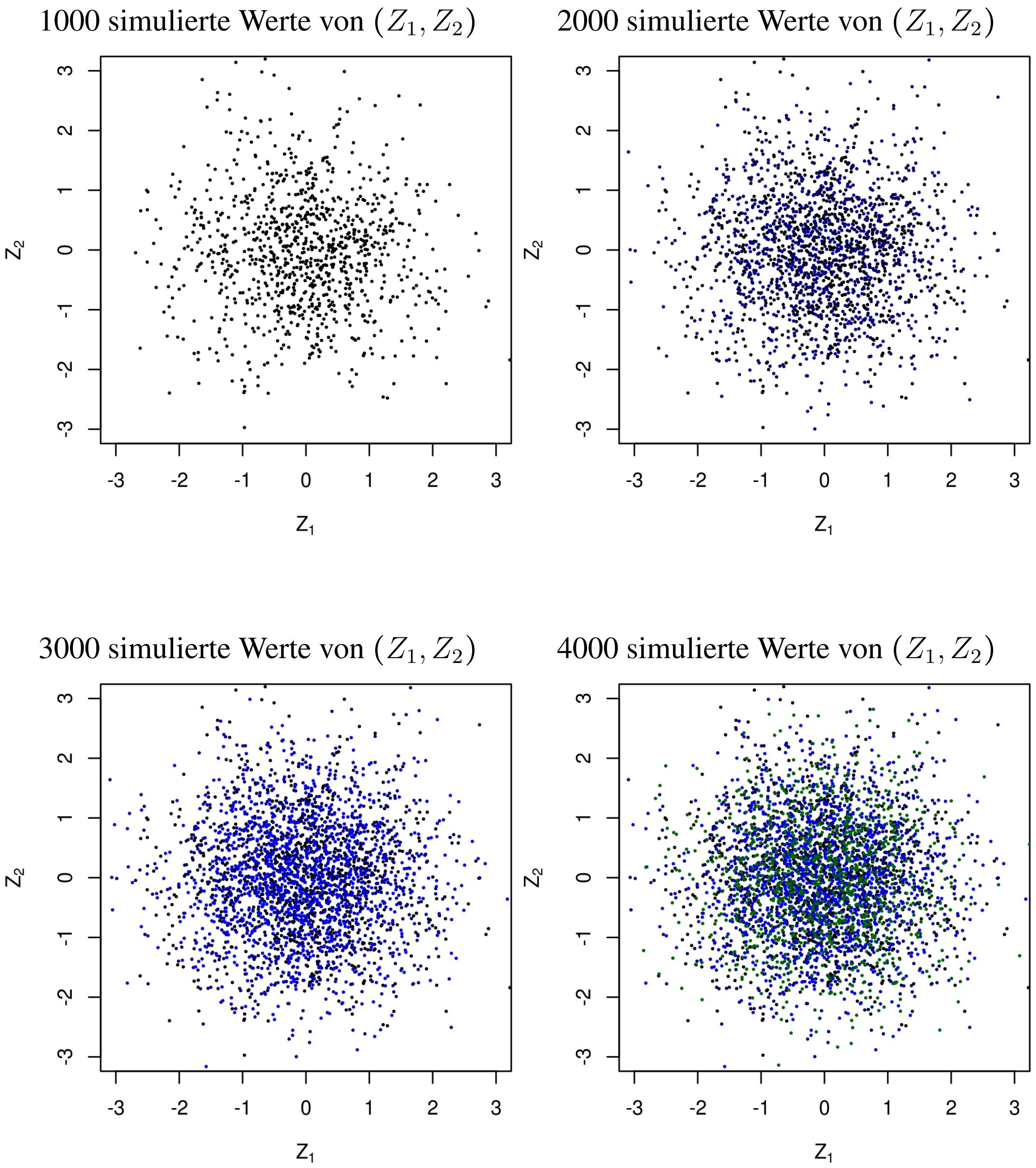
„Warum taucht im zentralen Grenzwertsatz die Normalverteilung auf?"

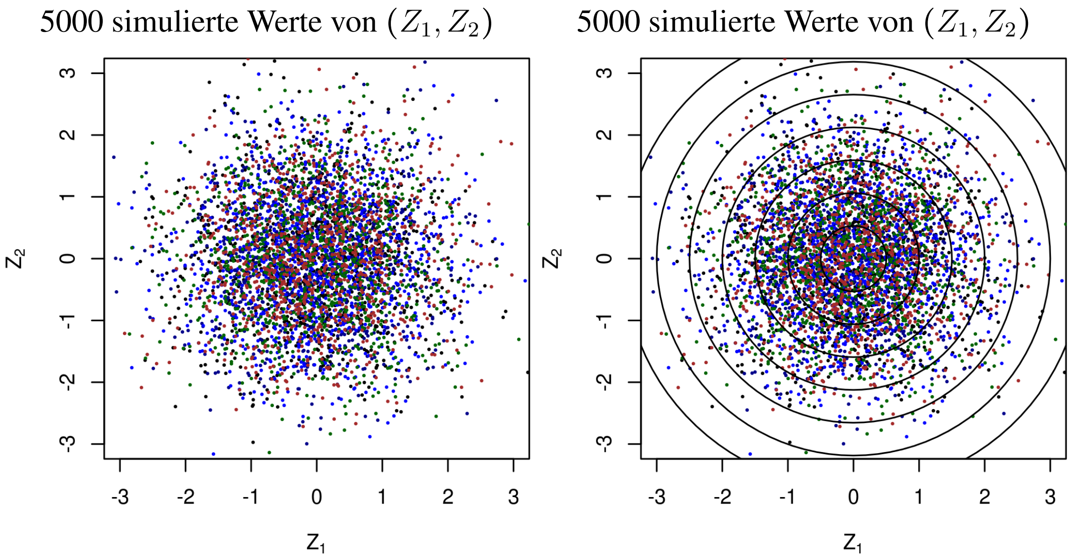
**Beobachtung.** In der Situation des zentralen Grenzwertsatzes sei

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

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 $(Z_1, Z_2)$
$Z_1$
$Z_2$
-3 -2 -1 0 1 2 3




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

**Beobachtung 1.94** (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. 1.58), andererseits gilt wegen Unabhängigkeit (vgl. Bericht 1.56)

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

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

$$w(u)w(v) = w(0)w(u+v), \quad u,v \ge 0 \tag{1.22}$$

Die allgemeine Lösung von (1.22) 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. □

### 1.5.4 Ergänzung: Hoeffding- und McDiarmid-Ungleichung

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

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

Offenbar ist $\mathrm{Var}[X_i] \le (b_i - a_i)^2$ (denn $|X_i - \mathbb{E}[X_i]| \le b_i - a_i$) und somit $\mathrm{Var}[S_n] = \mathrm{Var}[X_1] + \dots + \mathrm{Var}[X_n] \le \sum_{i=1}^n (b_i - a_i)^2$, die Chebyshev-Ungleichung (Formel (1.17) aus Satz 1.77) liefert daher für $t > 0$

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

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

**Bericht 1.95** (Hoeffding-Ungleichung(en)[15])**.** Unter obigen Voraussetzungen gilt für $t \ge 0$

$$P\big(S_n - \mathbb{E}[S_n] \ge t\big) \le \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big), \tag{1.23}$$

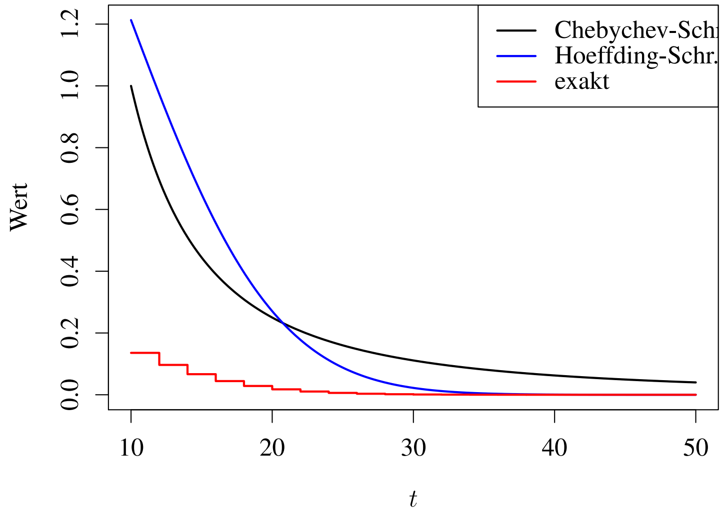
$$P\big(S_n - \mathbb{E}[S_n] \le -t\big) \le \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big), \tag{1.24}$$

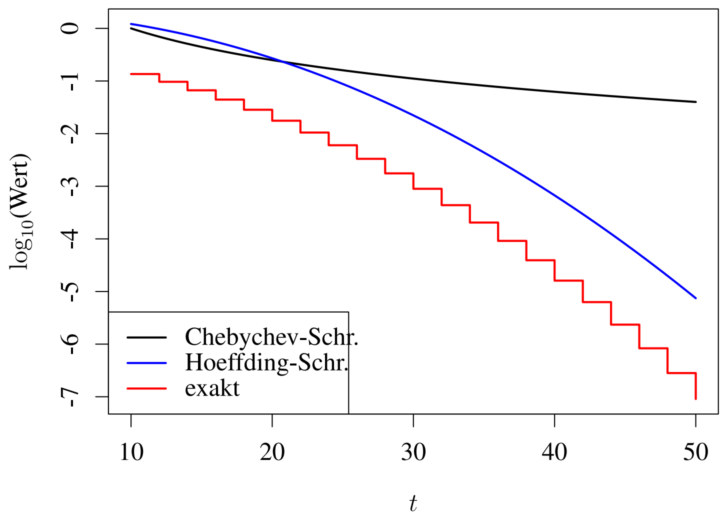
insbesondere

$$P\big(|S_n - \mathbb{E}[S_n]| \ge t\big) \le 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big). \tag{1.25}$$

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

[15] nach Wassily Hoeffding, 1914–1991



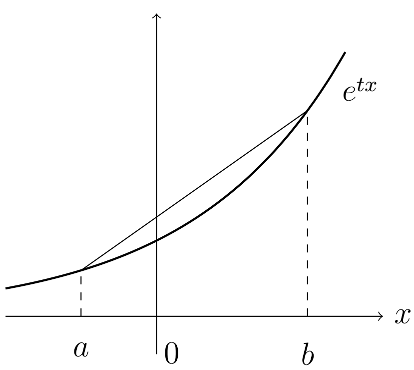




Zum Beweis der Aussagen in Bericht 1.95 verwendet man folgendes Lemma:

**Lemma 1.96** (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 $x \in [a,b]$

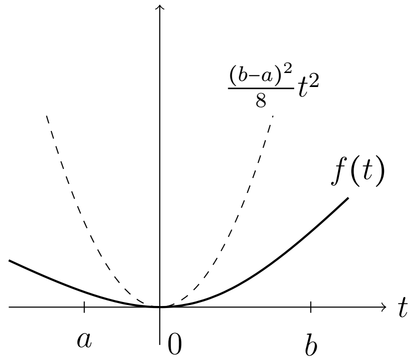
$$e^{tx} \le \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}$$

(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

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

und somit

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

wie behauptet. □

*Beweis von Bericht 1.95.* Für $u > 0$ ist

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

Mit der (optimalen) Wahl $u = 4t / \sum_{i=1}^n (b_i - a_i)^2$ folgt (1.23). (1.24) kann analog bewiesen werden (oder ersetze $S_n$ durch $-S_n$ in (1.23)); (1.25) folgt aus (1.23) und (1.24). □

**Bericht 1.97** (McDiarmid-Ungleichung)**.** Seien $X_1, X_2, \dots, X_n$ unabhängige Zufallsvariablen mit Werten in $S$, $f : S^n \to \mathbb{R}$. Es gebe Konstanten $c_1, \dots, c_n < \infty$, so dass für $i \in \{1, \dots, n\}$, $x_1, \dots, x_n \in S$, $x_i' \in S$ gilt

$$\big|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)\big| \le c_i. \tag{1.26}$$

Dann gilt für $t \ge 0$

$$P\Big(f(X_1, \dots, X_n) - \mathbb{E}\big[f(X_1, \dots, X_n)\big] \ge t\Big) \le \exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big) \tag{1.27}$$

und

$$P\Big(\big|f(X_1, \dots, X_n) - \mathbb{E}\big[f(X_1, \dots, X_n)\big]\big| \ge t\Big) \le 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big). \tag{1.28}$$

Die Hoeffding-Ungleichung folgt aus der McDiarmid-Ungleichung (mit der Wahl $f(x_1, \dots, x_n) = x_1 + \dots + x_n$), 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 $X_1, X_2, \ldots, X_n$ u.i.v. ~ $\mathrm{Ber}_p$, $p \in (0,1)$ (d.h. $P(X_i = 1) = p = 1 - P(X_i = 0)$),

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

die Anzahl der „Wechsel" in der Folge $X_1, X_2, \ldots, X_n$ (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 $\mathbb{E}[W] = \sum_{i=2}^{n} \mathbb{E}[\mathbf{1}_{\{X_i \neq X_{i-1}\}}] = (n-1) \cdot 2p(1-p)$ und wir können schreiben $W = f(X_1, X_2, \ldots, X_n)$ mit $f : \{0,1\}^n \to \mathbb{N}_0$,

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

$f$ 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.

[SfI17] M. Birkner, Notizen zur Vorlesung Statistik für Informatiker, JGU Mainz, SS 2017. `http://www.staff.uni-mainz.de/birkner/StatInfo17/Statistik_f_Informatiker_SS17.pdf`