

Statistik für Informatiker

Notizen zu einer Vorlesung an der
Johannes-Gutenberg-Universität Mainz, Sommer 2022

Matthias Birkner

Vorläufige Version, 17. Juni 2022

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

Inhaltsverzeichnis

0	Auftakt	3
0.1	Erste Schritte mit \mathbb{R}	9
1	Grundlagen aus der Wahrscheinlichkeitstheorie	12
1.1	Ereignisse, Zufallsvariablen und Wahrscheinlichkeiten	12
1.1.1	Zufallsvariablen	17
1.1.2	Einige „klassische“ diskrete Verteilungen	19
1.1.3	Der Fall mit Dichte	25
1.2	„Kleingedrucktes“	41
1.2.1	Eine Anmerkung zur Maßtheorie	41
1.2.2	Alternativer Zugang: Zufallsvariablen als „Fundamentalobjekte“ der Wahrscheinlichkeitstheorie	42
1.2.3	Zur Einschluss-Ausschluss-Formel	42
1.3	Bedingte Wahrscheinlichkeiten, Unabhängigkeit, gemeinsame Verteilung	44
1.3.1	Bedingte Wahrscheinlichkeit	44
1.3.2	Gemeinsame Verteilung und Randverteilungen	48
1.3.3	Unabhängigkeit	51
1.3.4	Die Situation im Fall mit Dichten	54
1.3.5	Faltung	58
1.4	Erwartungswert, Varianz und Kovarianz	61
1.4.1	Diskreter Fall	61
1.4.2	Der Fall mit Dichte	66
1.4.3	Varianz und Kovarianz	67
1.4.4	Median(e)	78
1.5	Gesetz der großen Zahlen und zentraler Grenzwertsatz	81
1.5.1	Gesetz der großen Zahlen	81
1.5.2	Zum zentralen Grenzwertsatz	83
1.5.3	Eine Heuristik zum zentralen Grenzwertsatz	91
1.5.4	Ergänzung: Hoeffding- und McDiarmid-Ungleichung*	94
2	Ideen aus der Statistik	99
2.1	Zur deskriptiven Statistik	99
2.2	Schätzer und Tests für (Populations-)Mittelwerte	99
2.2.1	Schätzer und Konfidenzintervall	100
2.2.2	Konfidenzintervall, bekannte Varianz	100
2.2.3	Konfidenzintervall, unbekannte Varianz	102

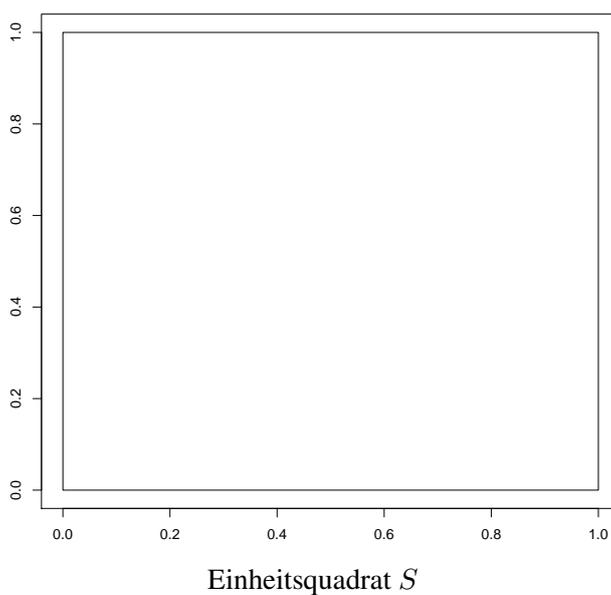
2.2.4	Statistische Tests	104
2.2.5	Ein-Stichproben- oder gepaarter t -Test	106
2.2.6	Zwei-Stichproben- t -Test oder ungepaarter t -Test mit Annahme gleicher Varianzen	108
2.2.7	Bericht: Welchs zwei Stichproben- t -Test (t -Statistik ohne Annahme gleicher Varianz)	110
2.3	Schätzprinzipien	111
2.3.1	ML-Schätzer	111
2.3.2	Bayes-Statistik	116
2.3.3	Kleinste-Quadrate-Schätzer und lineare Regression	121

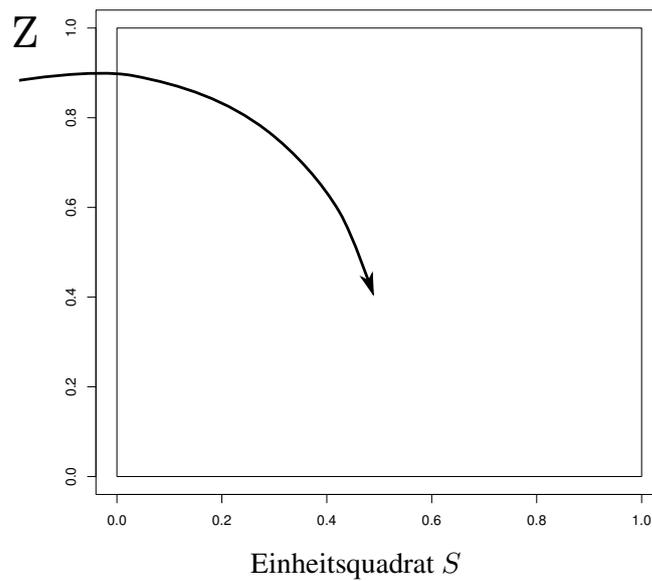
Kapitel 0

Auftakt

Ein Startbeispiel

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



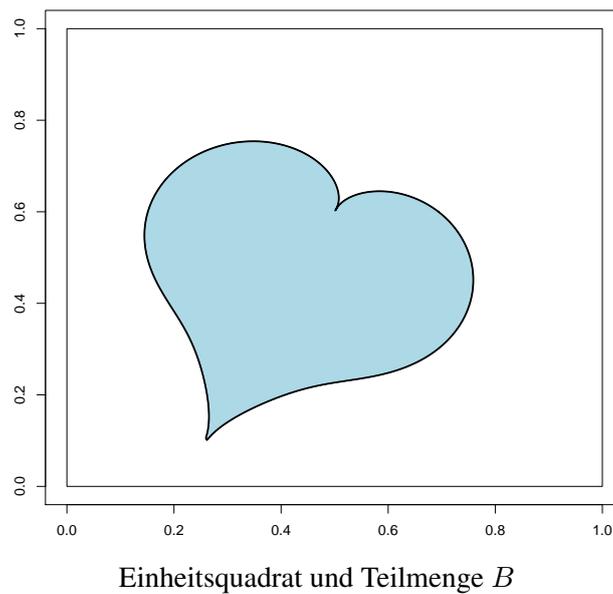


Für die praktische Implementierung stellen wir uns etwa vor, dass S in (sehr kleine) disjunkte Quadrate („Pixel“) zerlegt wird, und dass wir unter allen möglichen Pixeln eines wählen, wobei jedes dieselbe Chance hat, gezogen zu werden. Eine Möglichkeit, dies in R zu implementieren, wäre

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

Hierbei generiert der Befehl `runif(1)` eine (Pseudo-)Zufallszahl im Einheitsintervall $[0, 1]$. Wir nennen Z eine *Zufallsvariable*, die möglichen Werte S ihren Wertebereich.

Sei $B \subset S$ eine gewisse Teilmenge.



Dann können wir das *Ereignis*

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

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

Je nach Ausgang des zufälligen Experiments (für das Computereperiment 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

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

(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

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

Anwendung: eine einfache Monte Carlo-Methode zur Integration

Sei $p = P(\{Z \in B\}) = \text{Fläche von } B$. 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

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

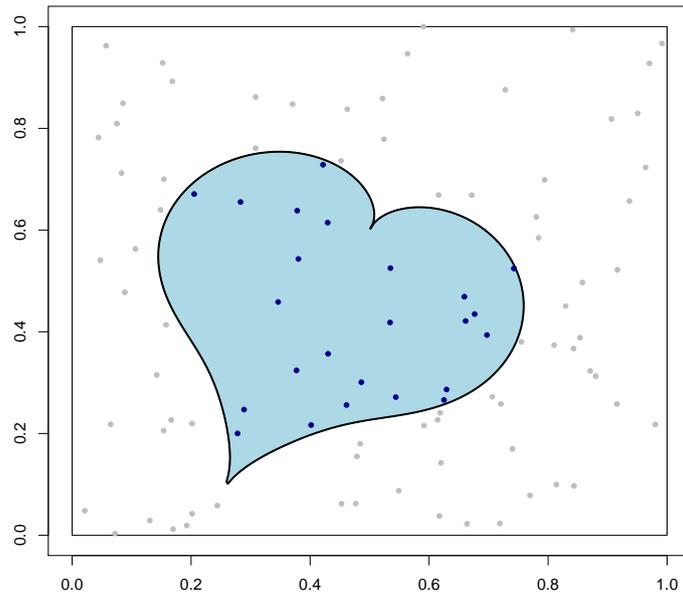
gibt an, wieviele der n zufällig gewählten Punkte in B gelandet sind. Der empirische Anteil

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

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 ε „daneben liegt“.

100 zufällige Punkte, Anteil in B : 0.24



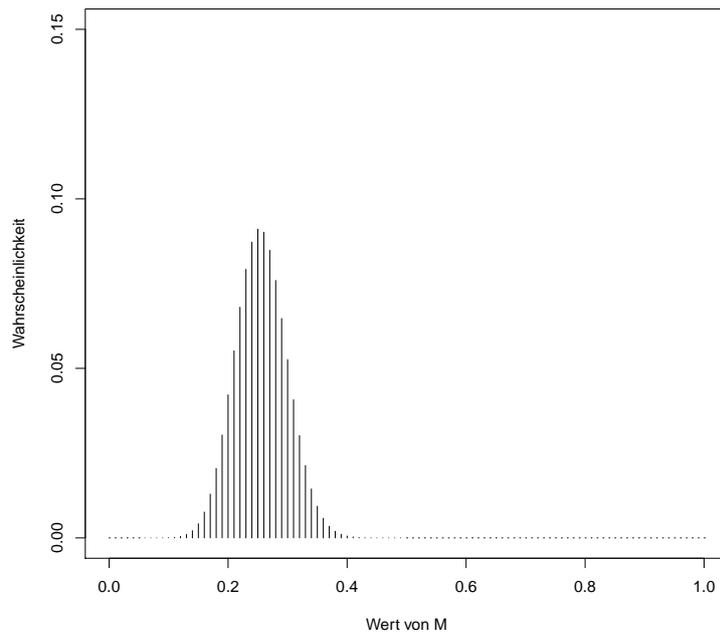
Visualisierung eines solchen Versuchs für $n = 100$.

\widetilde{M}_n ist Binomial-verteilt mit Parametern n und p (abgekürzt $\text{Bin}_{n,p}$ -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

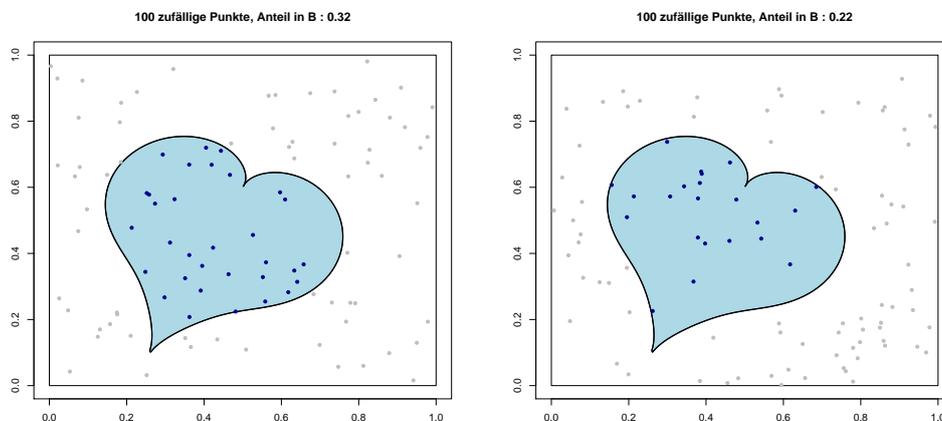
$$p = \int_{-\pi}^{\pi} \frac{1}{98} \left(2 - 2 \sin(w) + \frac{\sin(w) \sqrt{|\cos(w)|}}{\sin(w) + 7/5} \right)^2 dw \approx 0.25557221\dots$$

Denn für

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

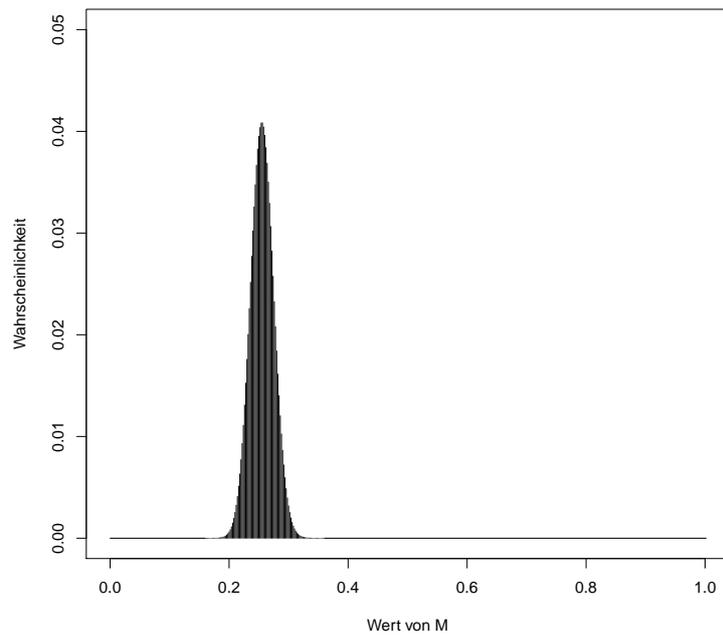
wo $f : [-\pi, \pi] \rightarrow \mathbb{R}_+$ und $\arg(x, y) = \text{sign}(x) \arcsin(y/\sqrt{x^2 + y^2}) + \pi \text{sign}(y) \mathbf{1}_{(-\infty, 0)}(x)$ (dies ist der Winkel, den der Strahl vom Ursprung durch den Punkt (x, y) mit der x -Achse einschließt) gilt

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



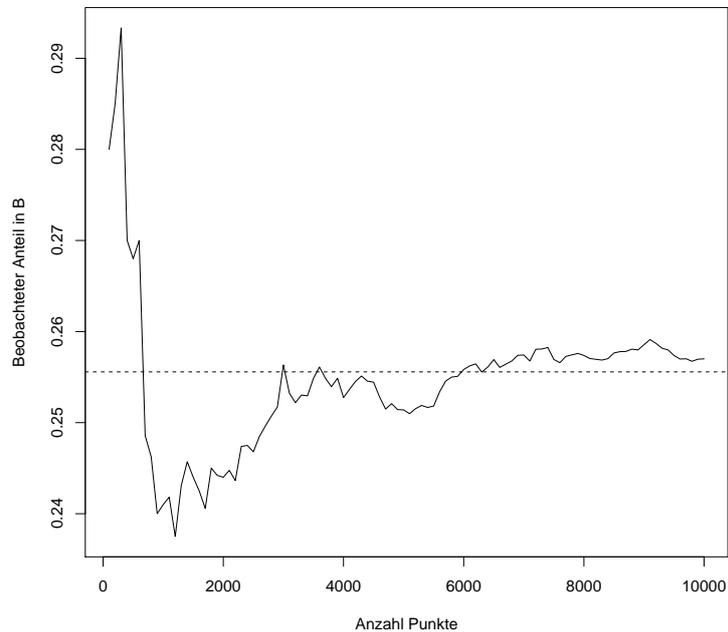
Zur Illustration: Zwei weitere Wiederholungen mit jeweils $n = 100$ zufälligen Punkten in $[0, 1]^2$

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 $Z_1, Z_2, \dots, Z_{10000}$) die Folge der M_n als Funktion von n (die gestrichelte Linie ist der Wert p):



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

0.1 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/) <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 Ω (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“, $\emptyset \dots$ „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“

(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 $A \cap B = \emptyset$.

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

- i) $\emptyset \in \mathcal{F}$,
 - ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,
 - iii) $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \cup_{n=1}^\infty A_n \in \mathcal{F}$.
- (1.1)

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

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

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

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

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

Definition 1.1. Ein Tripel (Ω, \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). Ω endlich oder abzählbar, $p : \Omega \rightarrow [0, 1]$ Abbildung mit $\sum_{\omega \in \Omega} p(\omega) = 1$.

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

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

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

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) $\Omega = \{(\omega_1, \omega_2, \omega_3) : \omega_i \in \{1, 2, 3, 4, 5, 6\} \text{ für } i = 1, 2, 3\}$, $p((\omega_1, \omega_2, \omega_3)) = 1/6^3 = 1/216$

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

(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) $\Omega = \{\text{Kopf, Zahl}\}$, $p(\text{Kopf}) = 0.6 = 1 - p(\text{Zahl})$

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

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

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

4. (Anzahl Würfe, bevor beim wiederholten fairen Münzwurf zum ersten Mal Kopf kommt)

$$\Omega = \mathbb{N}_0, p(n) = (1/2)^n \cdot (1/2) = 2^{-n-1} \text{ für } n \in \Omega$$

Bemerkung 1.4. Wahrscheinlichkeitsräume (Ω, \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 Ω 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

$$P(\emptyset) = 0 \quad (1.3)$$

$$P(A \cup B) + P(A \cap B) = P(A) + P(B) \quad (\text{endliche Additivität}), \quad (1.4)$$

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

$$A \subset B \Rightarrow P(A) \leq P(B) \quad (\text{Monotonie}) \quad (1.5)$$

zudem gilt auch

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n) \quad (\sigma\text{-Subadditivität}) \quad (1.6)$$

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

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

$$\text{so gilt } P(A) = \lim_{n \rightarrow \infty} P(A_n) \quad (\sigma\text{-Stetigkeit}) \quad (1.7)$$

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

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

$$P(A \cup B) = P(A \cup B \cup \emptyset \cup \emptyset \cup \dots) = P(A) + P(B) + \underbrace{P(\emptyset)}_{=0} + \underbrace{P(\emptyset)}_{=0} + \dots, \quad \text{d.h. (1.4) gilt hier.}$$

Allgemeiner Fall: Schreibe

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

also

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

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

(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 

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} A'_n\right) = \sum_{n=1}^{\infty} P(A'_n) \leq \sum_{n=1}^{\infty} P(A_n)$$

(die zweite Gleichung verwendet (A), die letzte (Un-)gleichung verwendet (1.5)).

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

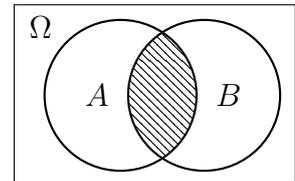
$$P(A) = P\left(\bigcup_{i=1}^{\infty} (A_i \setminus A_{i-1})\right) = \sum_{i=1}^{\infty} P(A_i \setminus A_{i-1}) = \lim_{n \rightarrow \infty} \underbrace{\sum_{i=1}^n P(A_i \setminus A_{i-1})}_{=P(A_n)}.$$

Falls $A_n \searrow A$, so beachte, dass $A_n^c \nearrow A^c$ 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“):

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

d.h. knapp ausgedrückt

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

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

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

die Wahrscheinlichkeit, dass es mindestens einen Fixpunkt gibt.

Bemerkung 1.7. 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 € ausbezahlt 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 $\{(x_1, x_2, \dots, x_n) : x_1, \dots, x_n \in \{1, 2, \dots, n\} \text{ paarweise verschieden}\}$)
2. Sei M das Ergebnis eines (möglicherweise verfälschten) Münzwurfs (Wertebereich $\{\text{Kopf}, \text{Zahl}\}$)
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 $\{\text{Deutsch}, \text{Englisch}\} \times \{\text{Spam}, \text{keinSpam}\}$)
4. Wir werfen eine faire Münze, bis zum ersten Mal Kopf kommt. G zählt, wie oft bis dahin Zahl gefallen ist.

Definition 1.8. (Ω, \mathcal{F}, P) ein Wahrscheinlichkeitsraum, S eine Menge, eine Abbildung¹ $X : \Omega \rightarrow S$ nennen wir eine *Zufallsvariable*, oft abgekürzt ZV (auch: *Zufallsgröße*) mit Wertebereich S .

ZVn und Ereignisse: Für $A \subset S$ schreibt man²

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

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 $\{X \leq x\} := \{X \in (-\infty, x]\}$, etc.

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

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

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

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

Definition 1.10. Für eine Zufallsvariable X mit Wertebereich S heißt³

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

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)): $\rho_X(a) \geq 0$,

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

und ρ_X ist ein *Wahrscheinlichkeitsmaß* (auch *Wahrscheinlichkeitsverteilung*) auf S (d.h. $\rho_X : \{\text{Teilmengen von } S\} \rightarrow [0, 1]$ 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 $\mathcal{L}(X)$ (das \mathcal{L} erinnert an English “law” bzw. Französisch «loi», d.h. „Gesetz“), man schreibt auch $X \sim \rho$, wenn $\mathcal{L}(X) = \rho$ gilt (für eine gewisse Wahrscheinlichkeitsverteilung).

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

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

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

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

Bemerkung 1.12 (Kanonisches Modell für eine ZV). Man kann eine Zufallsvariable X mit Wertebereich S und Verteilung ρ 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 $X = \text{Id}_S$ auf dem W -raum (S, \mathcal{F}, ρ) formulieren.

Man kann daher genauso gut 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 ($\leq n$) heraus. Sei

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

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

Wir betrachten 4 mögliche Situationen:

1. Ziehen mit Zurücklegen, mit Beachtung der Reihenfolge:

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

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

2. Ziehen ohne Zurücklegen, mit Beachtung der Reihenfolge:

$$X \text{ hat Werte in } W_2 = \{(x_1, \dots, x_k) : x_1, \dots, x_k \in \{1, 2, \dots, n\}, x_i \neq x_j \text{ für } i \neq j\},$$

$$P(X = (x_1, \dots, x_k)) = \frac{1}{n \cdot (n-1) \cdots (n-k+1)} = \frac{(n-k)!}{n!} \quad \left(= \frac{1}{|W_2|} \right)$$

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

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

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

Es ist

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

denn

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

(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

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

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

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

Man nennt Y auch einen (zufälligen) „Besetzungsvektor“. Für $(\ell_1, \ell_2, \dots, \ell_n) \in W_4$ gibt es

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

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

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

Begründung: Wir müssen ℓ_1 1-er, ℓ_2 2-er, \dots , ℓ_n n -er auf k (angeordnete) Plätze verteilen, dazu gibt es

$$\frac{k \cdot (k-1) \cdots (k-\ell_1-1)}{\ell_1!} \cdot \frac{(k-\ell_1) \cdot (k-\ell_1-1) \cdots (k-\ell_1-\ell_2-1)}{\ell_2!} \cdots$$

Anz. Mögl'keiten, die 1-er zu platzieren
Anz. Mögl'keiten, die 2-er zu platzieren

$$\cdots \cdot \frac{(k-\ell_1-\cdots-\ell_{n-1}) \cdot (k-\ell_1-\cdots-\ell_{n-1}-1) \cdots (k-\ell_1-\ell_{n-1}-\ell_n+1)}{\ell_n!}$$

Anz. Mögl'keiten, die n -er zu platzieren

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

Möglichkeiten, somit ist

$$P(Y = (\ell_1, \dots, \ell_n)) = \binom{k}{\ell_1, \ell_2, \dots, \ell_n} \left(\frac{1}{n}\right)^k, \quad (\ell_1, \dots, \ell_n) \in W_4$$

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

Bemerkung 1.14. Es gilt

$$|W_4| = \binom{n+k-1}{k} = \binom{n+k-1}{n-1}$$

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

$$\underbrace{\circ \cdots \circ}_{\ell_1 \text{ Kugeln}} \mid \underbrace{\circ \circ \cdots \circ}_{\ell_2 \text{ Kugeln}} \mid \underbrace{\quad}_{\ell_3 = 0} \mid \cdots \mid \underbrace{\circ \circ \cdots \circ}_{\ell_{n-1} \text{ Kugeln}} \mid \underbrace{\circ \cdots \circ}_{\ell_n \text{ Kugeln}}$$

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,

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

ist die W'keit, genau ℓ 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. $S = \{0, 1\}$, $\text{Ber}_p(\{1\}) = p = 1 - \text{Ber}_p(\{0\})$ mit einem $p \in [0, 1]$ („Bernoulli-Verteilung“⁴)

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

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

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

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

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

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

Dieser Zusammenhang mit ZVn ausgesprochen:

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

so ist $Y \sim \text{Bin}_{n,p}$, 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 $B_k = \{(x_1, \dots, x_n) \in \{0, 1, \dots\}^n : x_1 + \dots + x_n = k\}$

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

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

Beispiel 1.17 (Geometrische Verteilung). $p \in (0, 1)$, $S = \mathbb{N}_0$,

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

ist die W'keit, bei wiederholtem p -Münzwurf genau j 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). $s \in \{2, 3, \dots\}$, $p_1, \dots, p_s \in [0, 1]$, $p_1 + \dots + p_s = 1$, $n \in \mathbb{N}$, $S = \{(k_1, \dots, k_s) \in \mathbb{N}_0^s : k_1 + \dots + k_s = n\}$,

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

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

⁴nach Jakob Bernoulli, 1654–1705

Beispiel 1.19 (Poissonverteilung⁵). $\lambda \in (0, \infty)$,

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

Proposition 1.20 (Poissonapproximation der Binomialverteilung). Seien $p_n \in [0, 1]$ mit $p_n \rightarrow 0$ und $np_n \rightarrow \lambda \in (0, \infty)$ für $n \rightarrow \infty$, so gilt für jedes $k \in \mathbb{N}_0$

$$\text{Bin}_{n,p_n}(\{k\}) \xrightarrow{n \rightarrow \infty} \text{Poi}_\lambda(\{k\}).$$

Beweis. Es ist

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

□

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 Schadensfälle bei Versicherungen, Zerfallsereignisse in einer Probe radioaktiven Materials oder an genetische Mutationen.

Beispiel 1.21. L. von Bortkewitsch⁶ berichtete in seinem Buch *Das Gesetz der kleinen Zahlen*, Teubner, 1898 verschiedene Datensätze, die gut zur Poissonverteilung passen.

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

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

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

⁵nach Siméon Denis Poisson, 1781–1840

⁶Ladislav von Bortkewitsch, 1868–1931

Ergebnis x	Anz. „Korpsjahre“	$200 \times \text{Poi}_{0,61}(x)$
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

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

und es ist auch

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

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

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 $W \sim \text{Geom}_p$ mit $p \ll 1$, $X := pW$ (hat Werte in $p\mathbb{Z}_+ \subset \mathbb{R}$). Für jedes feste $K \in \mathbb{N}$ ist

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

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

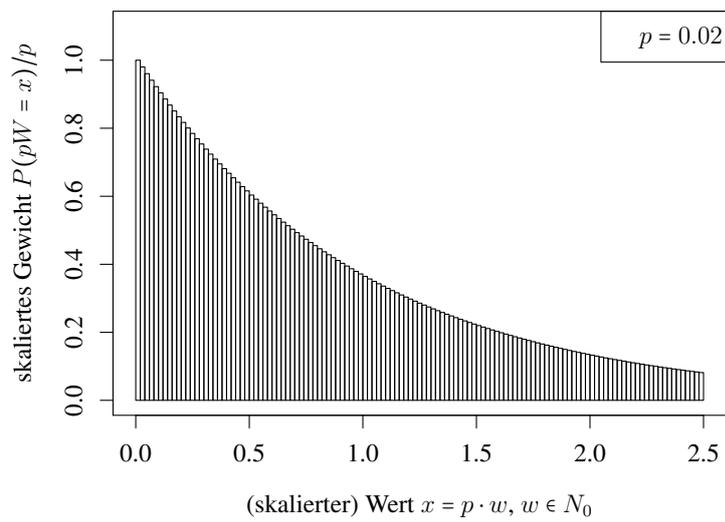
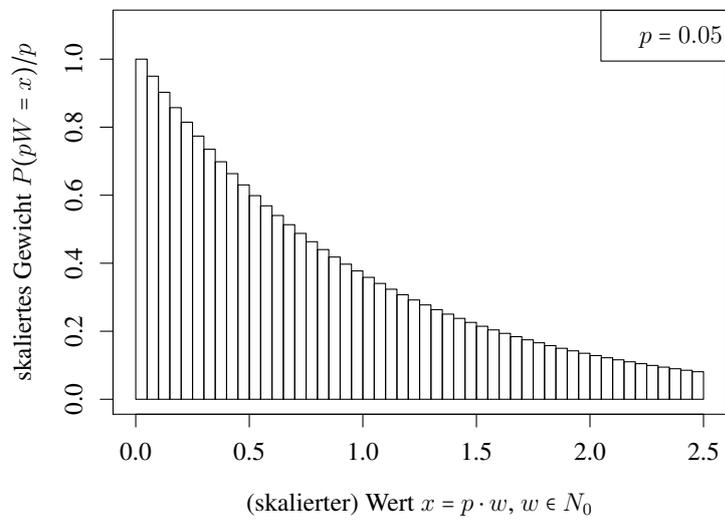
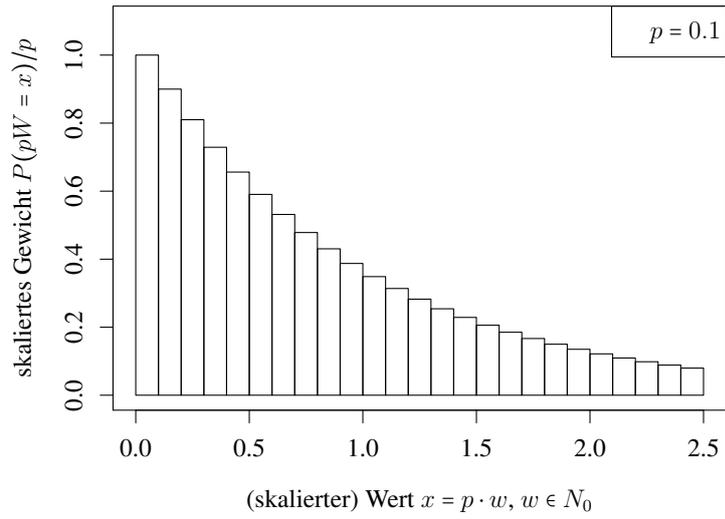
$$P(X \geq x) = P(pW \geq x) = P(W \geq \lceil x/p \rceil) = (1-p)^{\lceil x/p \rceil} \approx (1-p)^{x/p} \approx e^{-x}$$

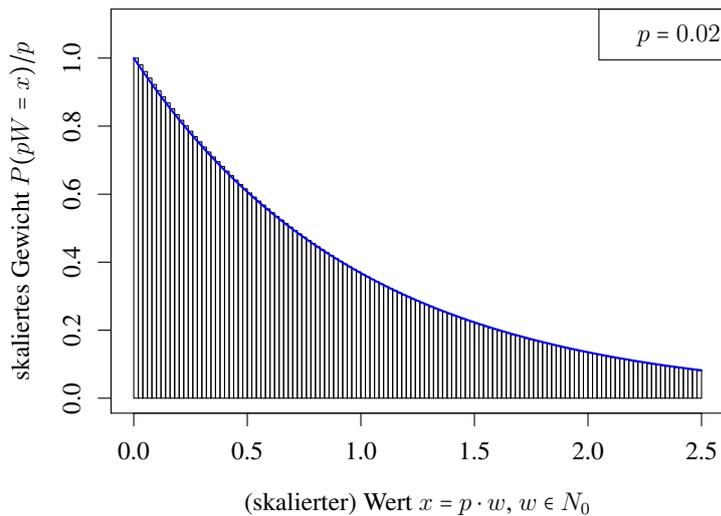
(und zwar „egal, wie klein“ p ist).

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

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

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





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

Beispiel 1.23 (Approximation einer Normalverteilung durch reskalierte binomialverteilte ZVn). Betrachten wir für $S_n \sim \text{Bin}_{n,1/2}$ (und verschiedene n) die Verteilungsgewichte, so beobachten wir, dass sich die „Form stabilisiert“.

Zentrieren und stauchen wir:

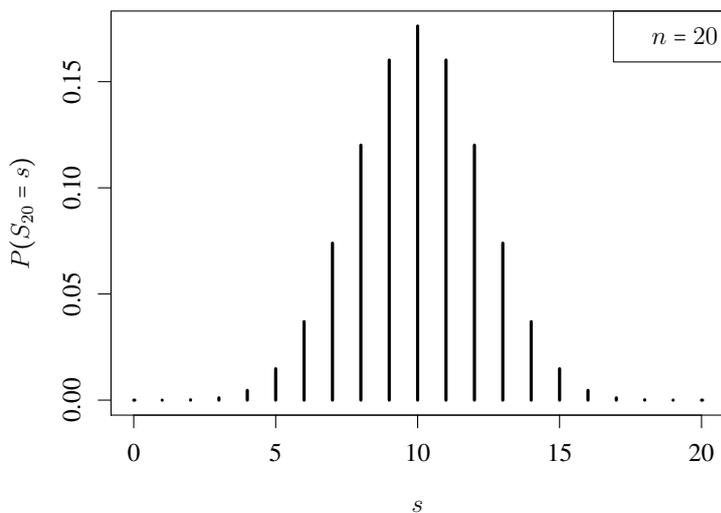
$$X_n := \frac{S_n - n/2}{\sqrt{n}}$$

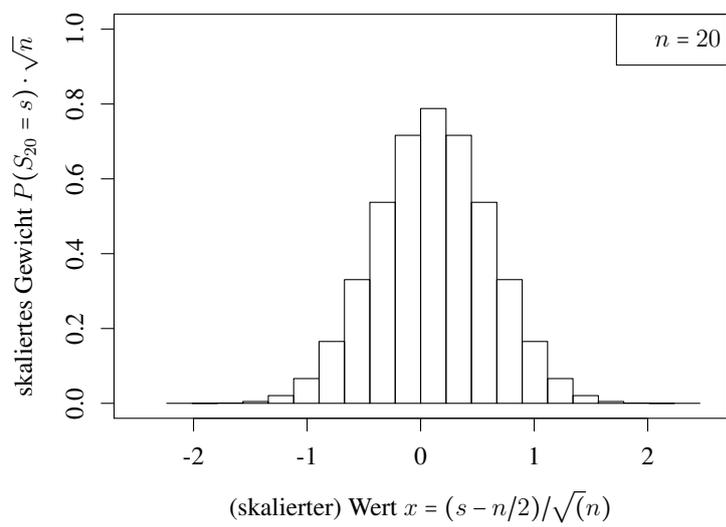
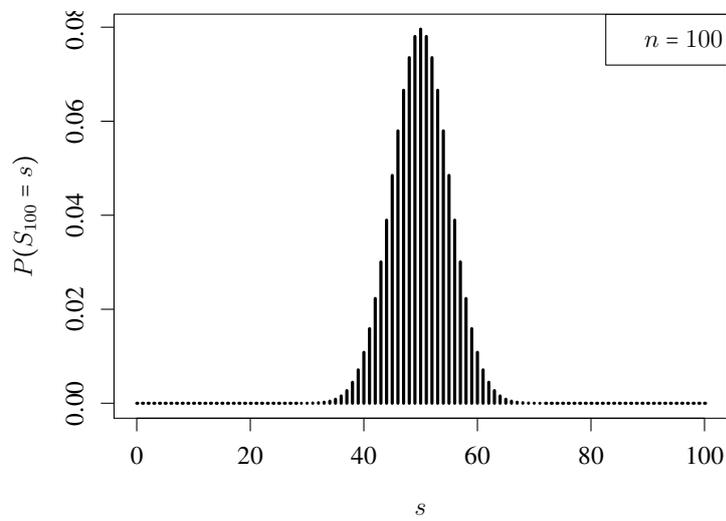
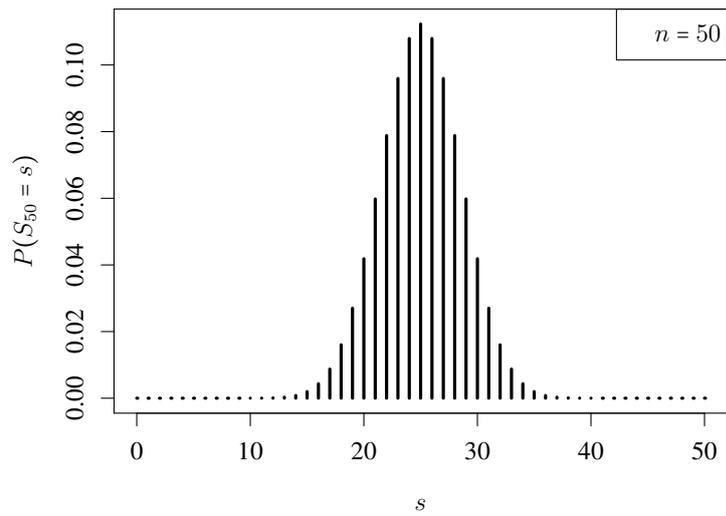
und skalieren die „Balken“ so, dass die Fläche des Balkens bei $(s - n/2)/\sqrt{n}$ gerade

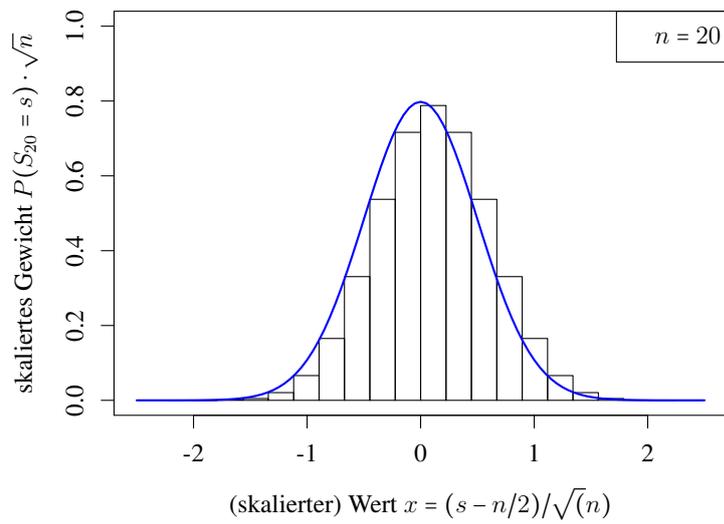
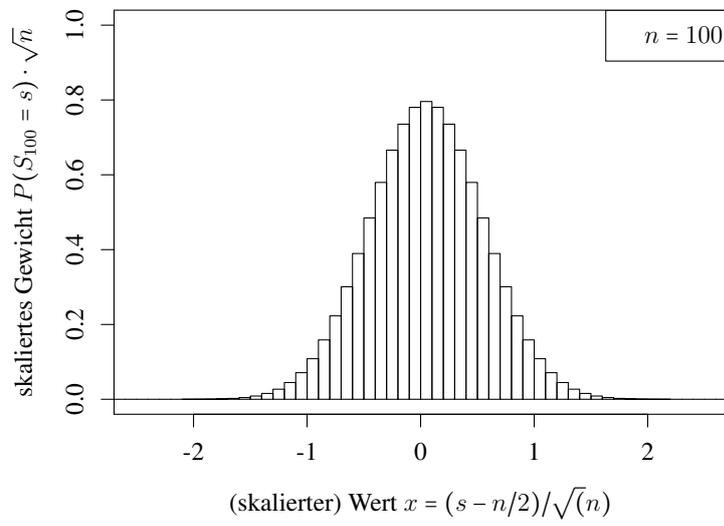
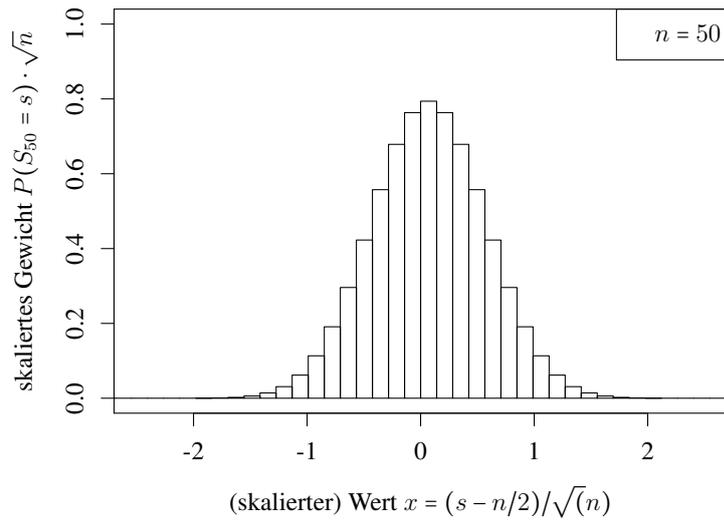
$$P(S_n = s) = \frac{1}{\sqrt{n}} \times \sqrt{n} P(S_n = s)$$

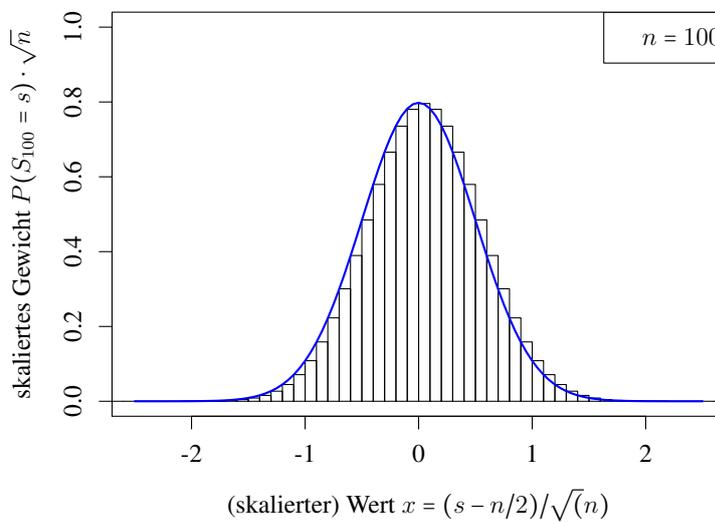
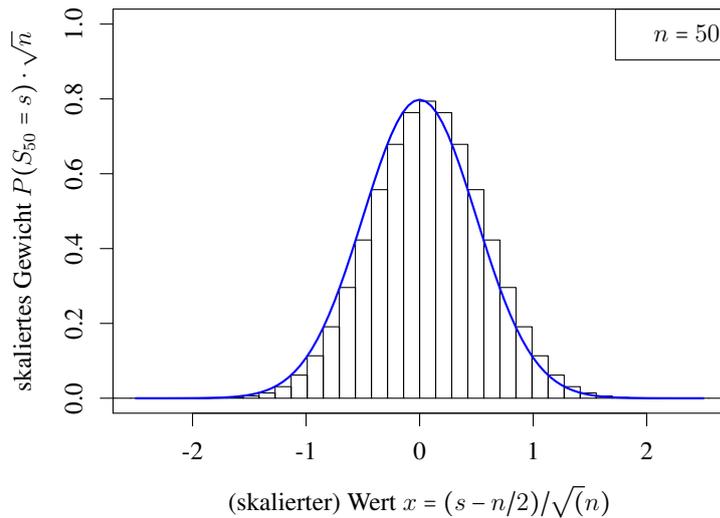
entspricht.

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)









Definition 1.24. Sei X eine Zufallsvariable mit Werten in einem Intervall $S = [a, b] \subset \mathbb{R}$ mit $-\infty \leq a < b \leq \infty$ (im Fall $a > -\infty, b = \infty$ meinen wir $S = [a, \infty)$, etc.) und sei $f : S \rightarrow [0, \infty)$ eine integrierbare⁷ Funktion mit

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

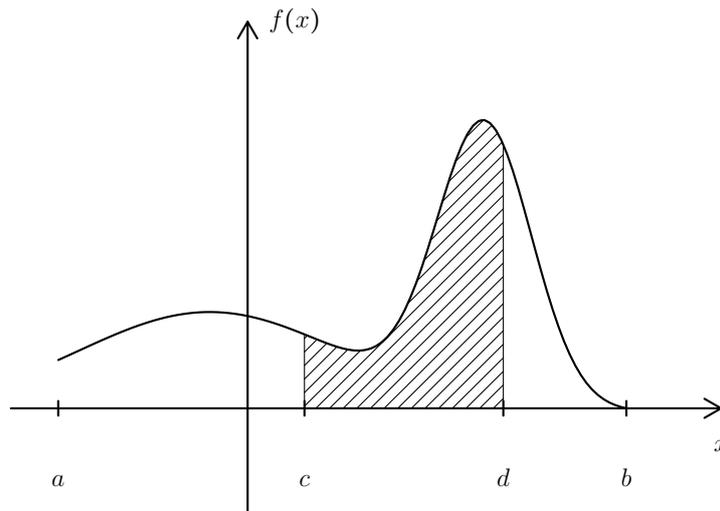
X besitzt die *Dichte* (auch: *Wahrscheinlichkeitsdichte*) f , wenn gilt

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

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).

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

⁷In 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

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

(wörtlich zumindest für Stetigkeitsstellen x von f_X), also

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

Man formuliert dies gelegentlich auch mit „infinitesimalen Größen“ als

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

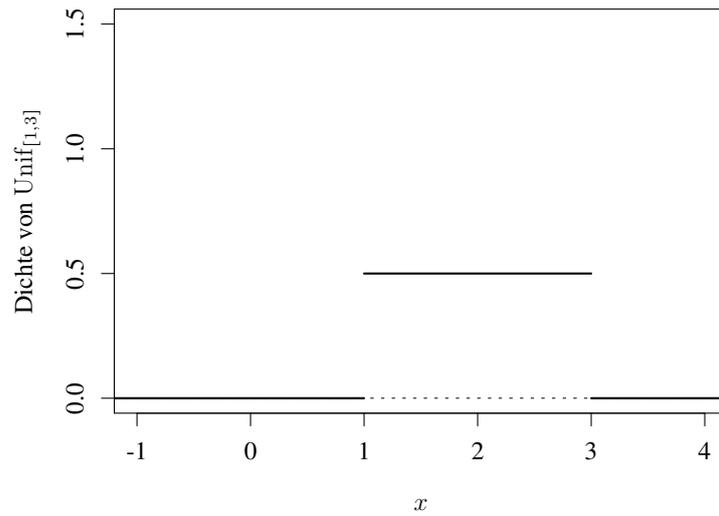
(Dieser suggestive Ausdruck erhält einen Sinn im Sinne der „Standard-Analysis“, wenn man auf beiden Seiten x über ein Intervall $[c, d]$ integriert, dann erhält man (1.9).)

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

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

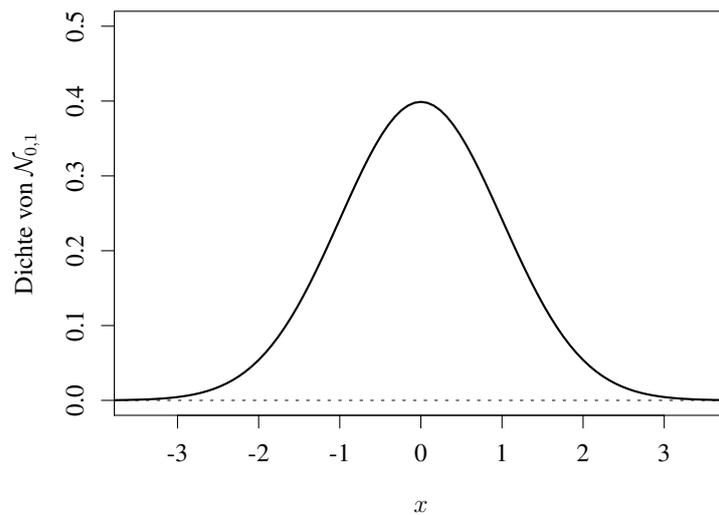
Beispiel 1.25 (Einige „klassische“ eindimensionale Verteilungen mit Dichte).

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

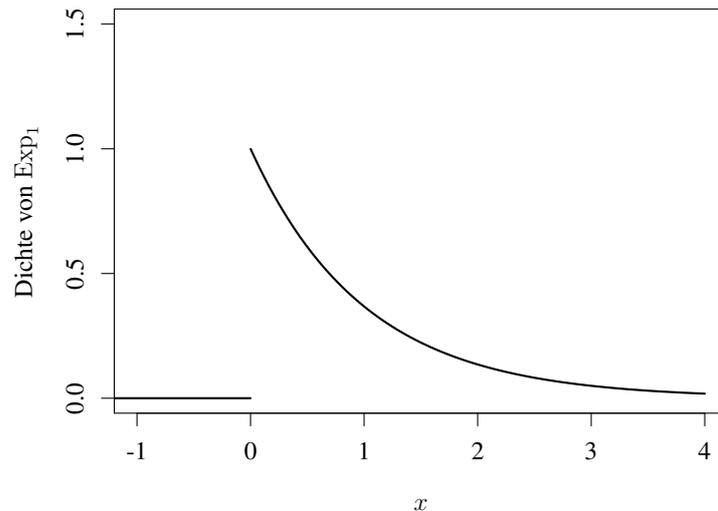


2. (Normalverteilung[en]) $\mu \in \mathbb{R}$, $\sigma > 0$. $\mathcal{N}_{\mu, \sigma^2}$ mit Dichte $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ heißt Normalverteilung mit Mittelwert μ und Varianz σ^2 .

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



3. (Exponentialverteilung[en]) $\theta > 0$, Exp_θ 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

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

die *Verteilungsfunktion* von X .

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

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

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

Beispiel 1.25 (Fortsetzung).

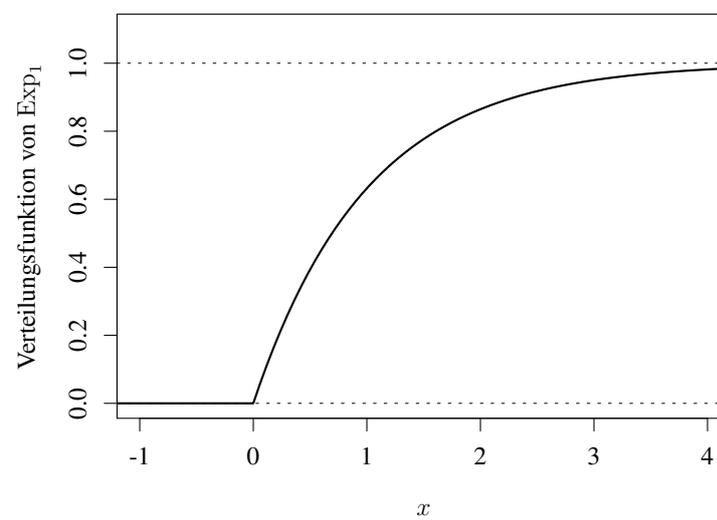
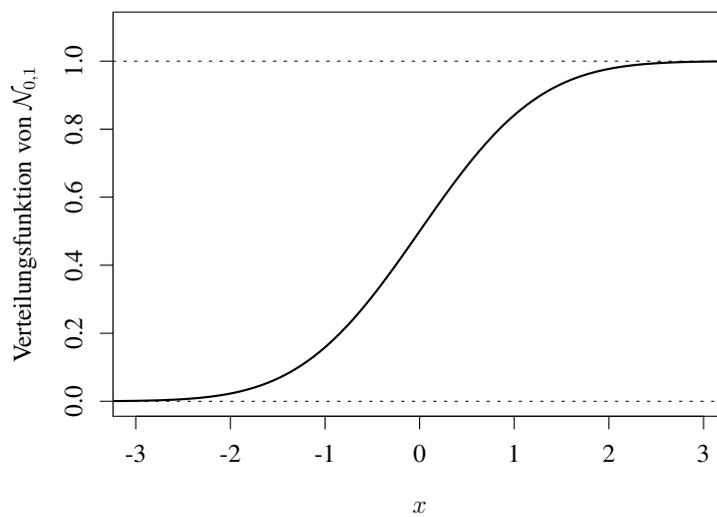
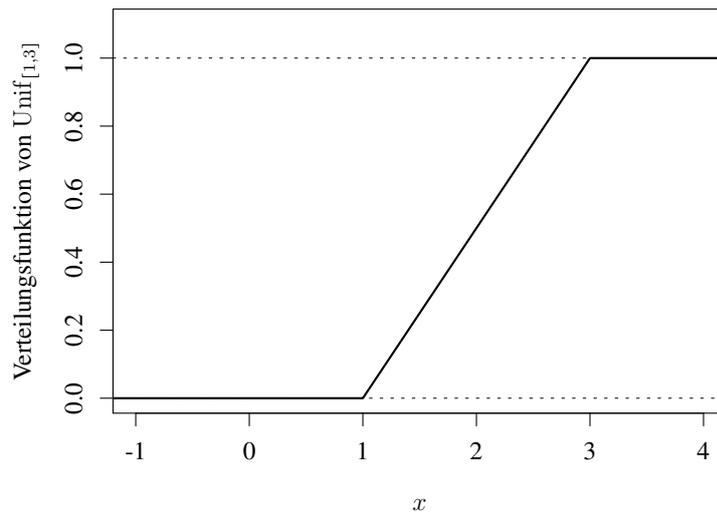
1. (uniforme Verteilung) $a, b \in \mathbb{R}$, $a < b$. $\text{Unif}_{[a,b]}$ mit Dichte $\frac{1}{b-a} \mathbf{1}_{[a,b]}(x)$, Verteilungsfunktion $\max \left\{ \min \left\{ \frac{x-a}{b-a}, 1 \right\}, 0 \right\}$
2. (Normalverteilung[en]) $\mu \in \mathbb{R}$, $\sigma > 0$. $\mathcal{N}_{\mu, \sigma^2}$ mit Dichte $\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$ heißt Normalverteilung mit Mittelwert μ und Varianz σ^2 .

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

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

ist tabelliert bzw. in vielen Computerprogrammen implementiert

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



Bemerkung 1.27. 1. Die Dichte / Verteilungsfunktion von X hängt nur von der Verteilung von X ab:

wenn $Y \stackrel{d}{=} X$ („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

$$P(X \in (c, d]) = P(X \leq d) - P(X \leq c) = F_X(d) - F_X(c)$$

(und falls $P(X = c) = 0$, z.B. weil X eine Dichte besitzt, so ist natürlich auch $P(X \in [c, d]) = P(X = c) + P(X \in (c, d]) = F_X(d) - F_X(c)$).

Für $B = \cup_{i=1}^n (c_i, d_i]$ mit $c_1 < d_1 < c_2 < d_2 < \dots < c_{n-1} < d_n$ ist (mit Eigenschaft (A) aus Forderung 1.2)

$$P(X \in B) = \sum_{i=1}^n P(X \in (c_i, d_i]) = \sum_{i=1}^n (F_X(d_i) - F_X(c_i))$$

(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 .

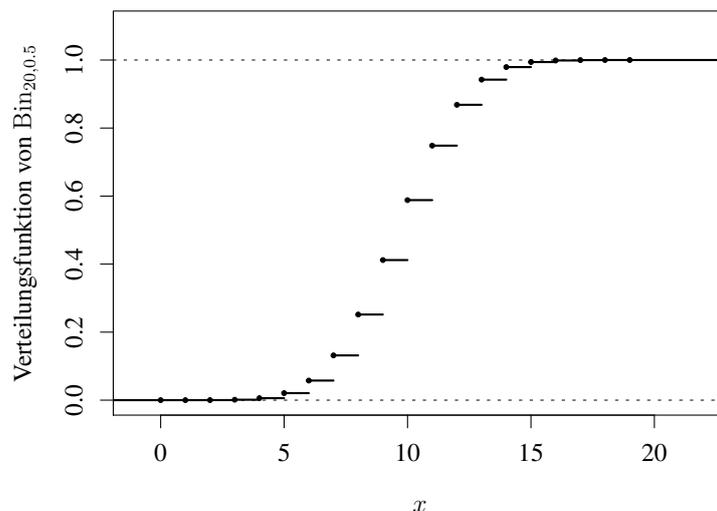
4. (Bezug zum diskreten Fall). Sei X ZV mit (höchstens) abzählbarem Wertebereich $S = \{x_1, x_2, \dots\} \subset \mathbb{R}$ und Gewichten $\rho_X(x_n)$ wie in in Def. 1.10, d.h.

$$P(X \in B) = \sum_{n: x_n \in B} \rho_X(x_n),$$

dann ergibt sich als Verteilungsfunktion

$$F_X(x) = \sum_{n: x_n \leq x} \rho_X(x_n).$$

(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

$$\lim_{x \rightarrow \infty} F_X(x) = 1, \quad \lim_{x \rightarrow -\infty} F_X(x) = 0.$$

6. Umgekehrt gibt es zu jeder Funktion $F : \mathbb{R} \rightarrow [0, 1]$ 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 ,

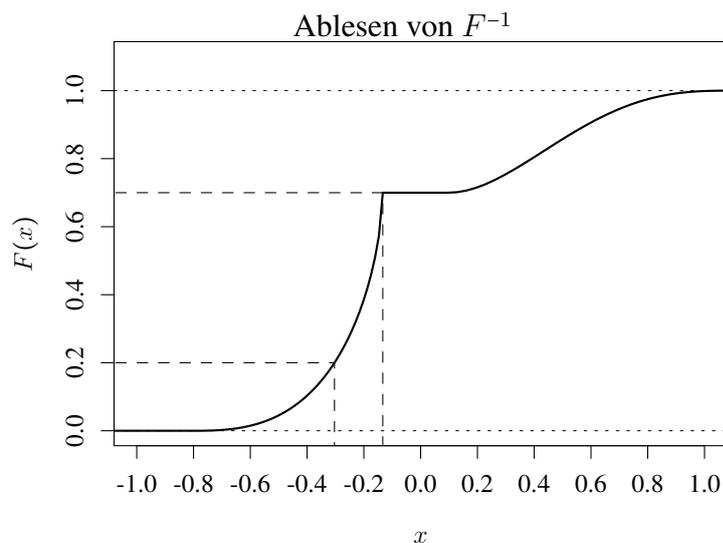
$$F_X^{-1}(t) := \inf\{x \in \mathbb{R} : F_X(x) \geq t\}, \quad t \in [0, 1]$$

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

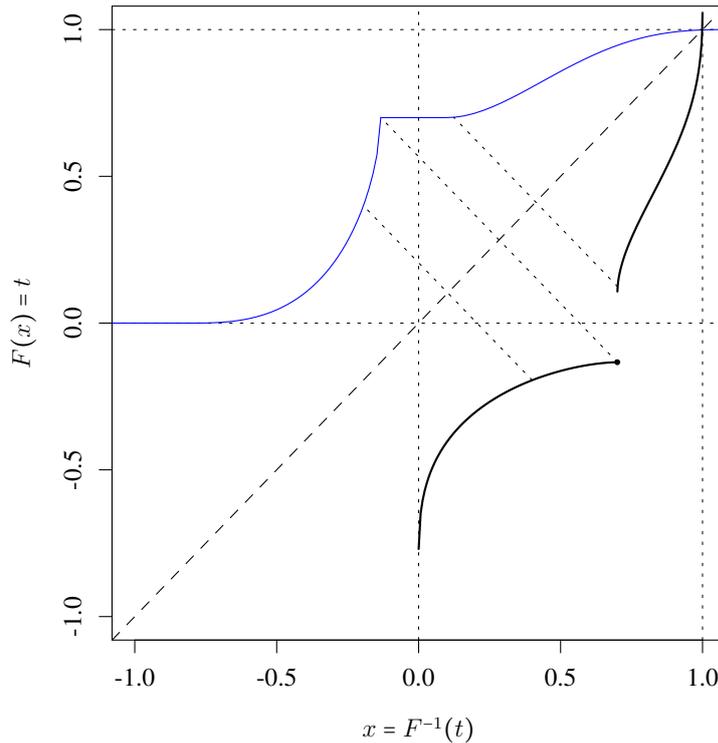
(Beachte, dass die so definierte Funktion F_X^{-1} linksstetig ist. Mit dieser Definition ergibt sich für $x \in \mathbb{R}, t \in [0, 1]$ die Beziehung

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

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 $F : \mathbb{R} \rightarrow [0, 1]$ 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, $U \sim \text{Unif}_{[0,1]}$, 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. Exp_θ hat Verteilungsfunktion $F_{\text{Exp}_\theta}(x) = (1 - e^{-\theta x})\mathbf{1}_{[0, \infty)}(x)$ mit inverser Funktion $F_{\text{Exp}_\theta}^{-1}(t) = -\frac{1}{\theta} \log(1 - t)$, also ist $-\frac{1}{\theta} \log(1 - U) \sim \text{Exp}_\theta$ (und natürlich ebenso $-\frac{1}{\theta} \log(U)$)

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

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

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

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

denn

$$P(Y \leq y) = P\left(X \leq \frac{y-b}{a}\right) = \int_{-\infty}^{(y-b)/a} f_X(x) dx = \int_{-\infty}^y f_X\left(\frac{z-b}{a}\right) \frac{1}{a} dz$$

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

$$\text{Beachte: } Y \in [y, y + \delta] \iff X \in \left[\frac{y-b}{a}, \frac{y-b}{a} + \frac{\delta}{a}\right].$$

Beispiel 1.32. 1. $X \sim \mathcal{N}_{0,1}$, $\mu \in \mathbb{R}$, $\sigma > 0$, so ist $Y := \sigma X + \mu \sim \mathcal{N}_{\mu,\sigma^2}$.

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

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

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 $P(X \in I) = 1$, $J \subset \mathbb{R}$, $\varphi: I \rightarrow J$ stetig differenzierbar, bijektiv.

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

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

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

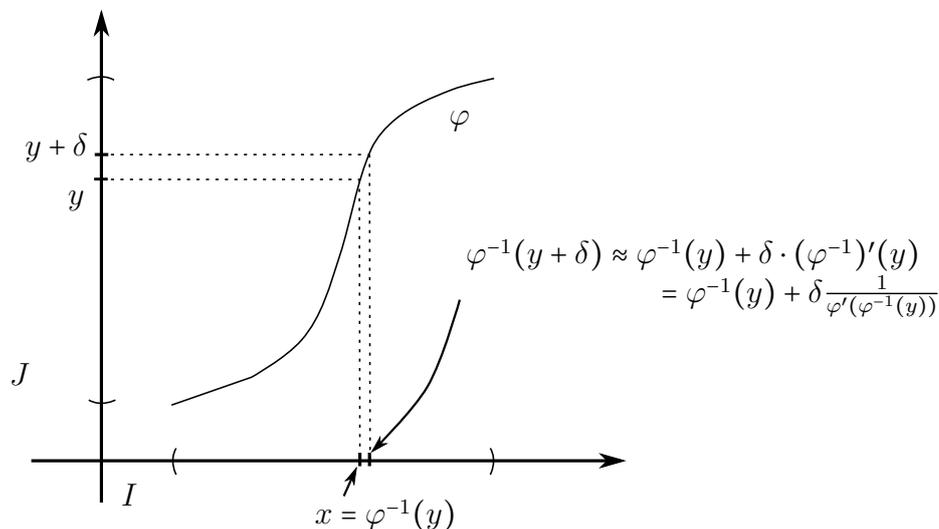
Beweis. φ muss offenbar strikt wachsend oder strikt fallend sein, wir betrachten den wachsenden Fall.

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

Sei $z \in J$:

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

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



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

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

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}_+$ eine (geeignet) integrierbare⁸ 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 (geeignete⁹) 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

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

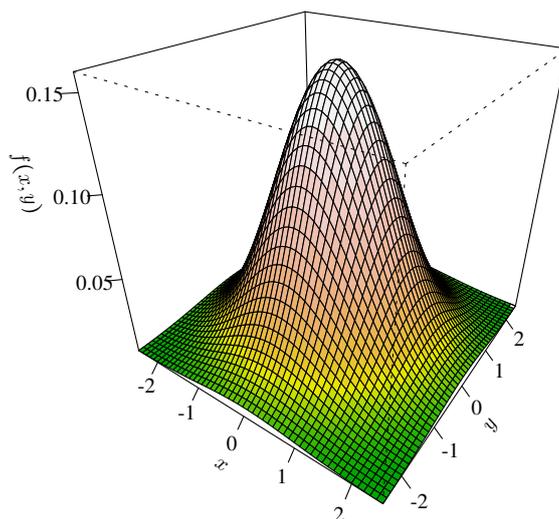
für $(x_1, \dots, x_d) \in \mathbb{R}^d$ und $0 < \delta_1, \delta_2, \dots, \delta_d \ll 1$.

Beispiel 1.36. 1. Uniforme oder Laplace-Verteilung auf einem beschränkten Gebiet $S \subset \mathbb{R}^d$: X mit Dichte $f_X(x) = \frac{1}{\text{vol}(S)} \mathbf{1}_S(x)$ erfüllt für $A \subset S$ (geeignet¹⁰)

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

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

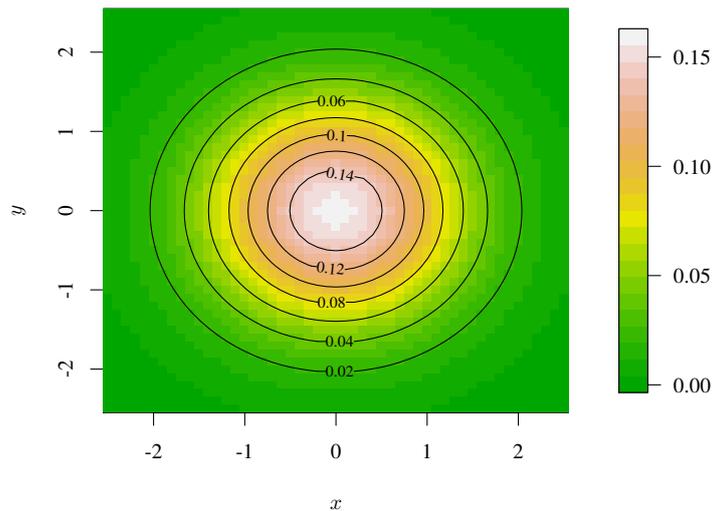
2. Die 2-dimensionale Standard-Normalverteilung hat Dichte $f(x, y) = \frac{1}{2\pi} \exp(-\frac{1}{2}(x^2 + y^2))$



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

⁹Damit das multivariate Integral wohldefiniert ist, muss A strenggenommen gewisse sogenannte Messbarkeits-eigenschaften 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.

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



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

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

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 σ -Algebra, d.h. eine Teilmenge $\mathcal{S} \subset 2^S := \{B : B \subset S\}$ der Potenzmenge von S ausgezeichnet, die erfüllt:

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

Falls S abzählbar ist, so kann man einfach $\mathcal{S} = 2^S$ 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 \mathcal{S} meist die sogenannte Borel- σ -Algebra, die kleinste σ -Algebra, die alle offenen Mengen enthält.

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

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

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

Weiterhin sind für eine S -wertige ZV X als Ereignisse in Def. 1.8 nur $\{X \in A\}$ mit $A \in \mathcal{S}$ zugelassen (für $A \subset S$ mit $A \notin \mathcal{S}$ 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.

1.2.3 Zur Einschluss-Ausschluss-Formel

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

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

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 Ω , d.h. $\bigcap_{j \in \emptyset} A_j := \Omega$.

Wir zeigen die folgende Aussagen:

(a) Für jedes $K \subset I$ gilt

$$P\left(\bigcap_{k \in K} A_k\right) = \sum_{K \subset J \subset I} P(B_J).$$

(b) Für jedes $J \subset I$ gilt

$$P(B_J) = \sum_{J \subset K \subset I} (-1)^{|K \setminus J|} P\left(\bigcap_{k \in K} A_k\right).$$

(c) Die oben angegebene Einschluss-Ausschluss-Formel folgt, indem wir $J = \emptyset$ setzen und Aussage (b) mit

$$P\left(\bigcap_{i \in I} A_i^c\right) = 1 - P\left(\bigcup_{i \in I} A_i\right)$$

verwenden.

Zu (a): Wir zeigen zunächst, dass

$$B_J \cap B_{\tilde{J}} = \emptyset \text{ für } J \neq \tilde{J} \quad \text{und} \quad \bigcup_{\emptyset \subset J \subset I} B_J = \Omega.$$

Sei $J \neq \tilde{J}$, d.h. es gibt ein $j \in (J \setminus \tilde{J}) \cup (\tilde{J} \setminus J)$, dann gilt $B_J \cap B_{\tilde{J}} \subset A_j \cap A_j^c = \emptyset$. 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 $\Omega \subset \bigcup_{J: \emptyset \subset J \subset I} B_J \subset \Omega$, d.h. es gilt $\bigcup_{J: \emptyset \subset J \subset I} B_J = \Omega$.

Sei nun $K \subset I$ gegeben. Nach obigem ist

$$\bigcap_{k \in K} A_k = \{x \in \Omega : J(x) \supset K\} = \bigcup_{J: K \subset J \subset I} B_J, \quad (1.12)$$

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)

$$\begin{aligned} \sum_{K: J \subset K \subset I} (-1)^{|K \setminus J|} P\left(\bigcap_{k \in K} A_k\right) &= \sum_{K: J \subset K \subset I} (-1)^{|K \setminus J|} \sum_{J': K \subset J' \subset I} P(B_{J'}) \\ &= \sum_{J': J \subset J' \subset I} \sum_{K: J \subset K \subset J'} (-1)^{|K \setminus J|} P(B_{J'}) \\ &= \sum_{J': J \subset J' \subset I} P(B_{J'}) \underbrace{\sum_{K: J \subset K \subset J'} (-1)^{|K \setminus J|}}_{= \begin{cases} 1, & J' = J \\ 0, & J' \neq J \end{cases}} = P(B_J). \end{aligned}$$

Für die letzte Gleichung beachten wir beispielsweise folgendes: Sei $J' \neq J$, d.h. es gibt ein $j' \in J' \setminus J$, dann ist

$$\sum_{K: J \subset K \subset J'} (-1)^{|K \setminus J|} = \sum_{K: J \subset K \subset J' \setminus \{j'\}} \left((-1)^{|K \setminus J|} + (-1)^{|(K \cup \{j'\}) \setminus J|} \right) = \sum_{K: J \subset K \subset J' \setminus \{j'\}} 0 = 0.$$

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

Zu (c): Es ist

$$\begin{aligned} P\left(\bigcup_{i \in I} A_i\right) &= 1 - P\left(\left(\bigcup_{i \in I} A_i\right)^c\right) = 1 - P\left(\bigcap_{i \in I} A_i^c\right) = 1 - P(B_{\emptyset}) \\ &= 1 - \sum_{K: \emptyset \subset K \subset I} (-1)^{|K|} P\left(\bigcap_{k \in K} A_k\right) = 1 - (-1)^0 P(\Omega) + \sum_{K: \emptyset \neq K \subset I} (-1)^{|K|-1} P\left(\bigcap_{k \in K} A_k\right) \\ &= \sum_{K: \emptyset \neq K \subset I} (-1)^{|K|-1} P\left(\bigcap_{k \in K} A_k\right). \end{aligned}$$

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 $S = \{(i, j) : 1 \leq i, j \leq w + s, i \neq j\}$ (vgl. Bsp. 1.13, 2.).

Betrachte die Ereignisse

$$A = \{\text{erste Kugel ist weiß}\} = \{X_1 \leq w\},$$

$$B = \{\text{zweite Kugel ist weiß}\} = \{X_2 \leq w\} \left(= \{X \in \{(i, j) \in S : j \leq w\}\} \right)$$

Ohne weitere Informationen ist

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

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

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

heißt bedingte Wahrscheinlichkeit von B , gegeben A .

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

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

In Beispiel 1.37 ist

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

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

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 \tilde{P} mit

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

Dann gilt

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

Beweis. Für $C \in \mathcal{F}$ ist

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

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

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

So ist in Beispiel 1.37 ist auch

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

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

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

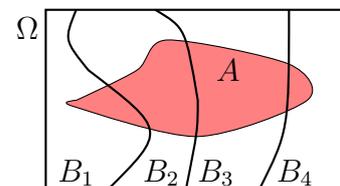
1. (Formel von der totalen Wahrscheinlichkeit)

Für $A \in \mathcal{F}$ gilt

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

Man kann dies anhand eines Diagramms veranschaulichen:

$P(A | B_i)$ ist der Anteil von $A \cap B_i$ an der „Wahrscheinlichkeitsmasse“ $P(B_i)$ von B_i



2. (Formel von Bayes¹¹)

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

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

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

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

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

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

Beweis. 1. Mit Definition 1.38 ist

$$\sum_{i \in I} P(B_i)P(A | B_i) = \sum_{i \in I} P(A \cap B_i) = P\left(\bigcup_{i \in I} (A \cap B_i)\right) = P(A)$$

(verwende die (σ) -Additivität von P).

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

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

also

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

und

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

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

$$P(X = 1 | Y = 1) \approx 0.803, \quad P(X = 0 | Y = 0) \approx 0.977.$$

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

also $P(B) = 0,02$, $P(A | B) = 0,95$, $P(A | B^c) = 0,1$, somit

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

(wir sehen: geringe „positive Korrektheit“, andererseits

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

(recht hohe „negative Korrektheit“).

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

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

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

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

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

so ist

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

(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

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

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

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 (Ω, \mathcal{F}, P) definiert sind), X_i habe Werte in S_i ,

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

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

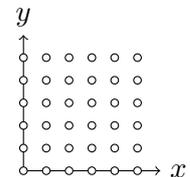
Beispiel. Fassen wir die Häufigkeitstabelle (normiert auf Gesamtsumme = 1) aus Bsp. 1.43 als Verteilungstafel der gemeinsamen Verteilung von X_1 („Testergebnis“ mit Werten in $S_1 = \{\text{positiv, negativ}\}$) und X_2 („Gesundheitszustand“ mit Werten in $S_2 = \{\text{krank, gesund}\}$) auf:

$X_1 \backslash X_2$	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 $0 \leq x < L$ und ebenso $P(Y_1 = y) = \frac{1}{L}$ für $0 \leq y < L$.

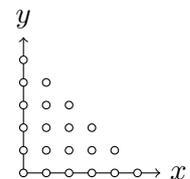


Wir sehen in diesem Fall: $P(X_1 = x, Y_1 = y) = P(X_1 = x) \cdot P(Y_1 = y)$, d.h. die gemeinsamen Gewichte erhält man als *Produkt* der Randverteilungs-Gewichte. Man sagt dazu auch: X_1 und Y_1 sind *unabhängig* (siehe Def. 1.51 unten für die allgemeine Definition).

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 $0 \leq x < L$ und analog $P(Y_2 = y) = \frac{2(L-y)}{L(L+1)}$ für $0 \leq y < L$.

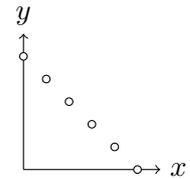


Wir sehen in diesem Fall: $P(X_2 = x, Y_2 = y) \neq P(X_2 = x) \cdot P(Y_2 = y)$, die Koordinaten X_2 und Y_2 sind *nicht unabhängig*.

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 $0 \leq x < L$ und analog $P(Y_3 = y) = \frac{1}{L}$ für $0 \leq y < L$.



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 $\mathcal{L}(X_1) = \mathcal{L}(X_3)$ und $\mathcal{L}(Y_1) = \mathcal{L}(Y_3)$ (man schreibt dies auch als $X_1 \stackrel{d}{=} X_3, Y_1 \stackrel{d}{=} Y_3$, „Gleichheit in Verteilung“, aber $\mathcal{L}((X_1, Y_1)) \neq \mathcal{L}((X_3, Y_3))$).

Bedingte Verteilung und mehrstufige Experimente

Oft betrachtet man folgende Situation:

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

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

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

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

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

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

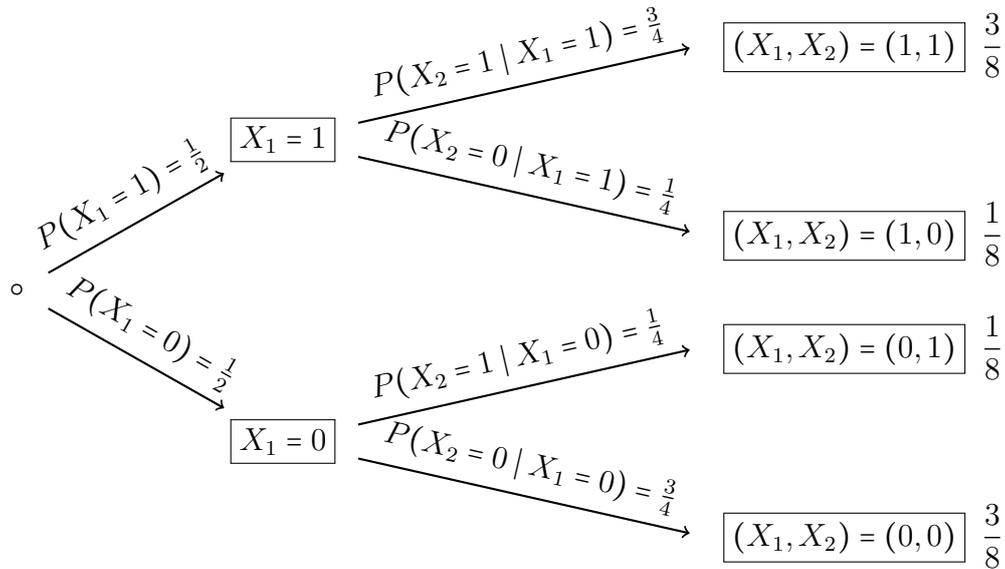
Beispiel 1.47. Wir haben eine faire Münze M_1 und zwei gezinkte Münzen M_2, M_3 (deren Seiten jeweils mit 0 und 1 beschriftet seien), wobei

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

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



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	$\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$	$\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	

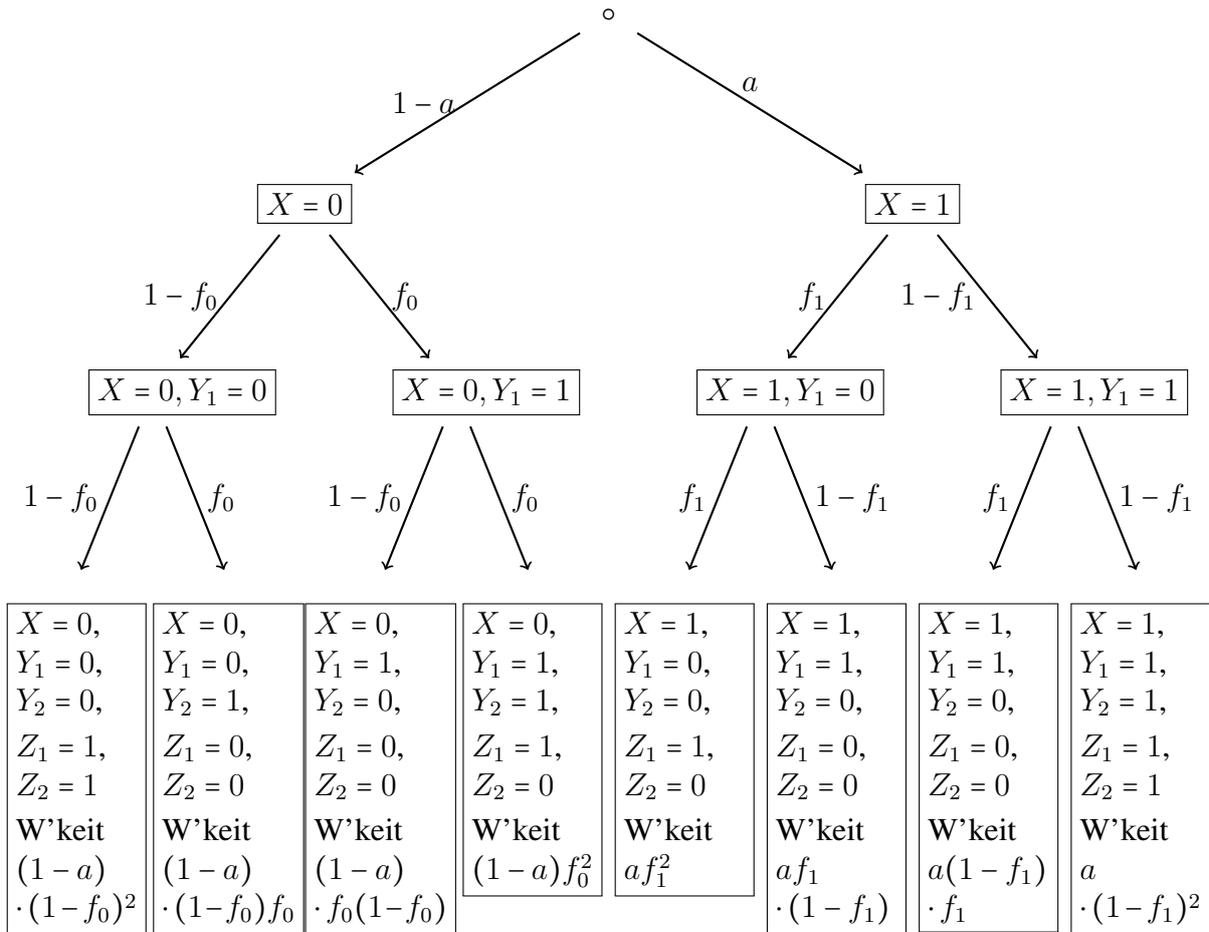
Wir sehen hier wieder: Die Randverteilungen legen (i.A.) nicht die gemeinsame Verteilung fest. Es ist $P(X_1 = 1) = P(X_2 = 1) = \frac{1}{2}$ (und dieselben Randverteilungen ergäben sich, wenn man zwei Mal M_1 wirft, aber die gemeinsame Verteilung wäre eine andere).

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 $\{Z_2 = 1\} \subset \{Z_1 = 1\}$

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

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

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



Wir sehen:

$$P(Z_1 = 1) = (1-a)(1-f_0)^2 + (1-a)f_0^2 + af_1^2 + a(1-f_1)^2,$$

$$P(Z_2 = 1) = (1-a)(1-f_0)^2 + a(1-f_1)^2,$$

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

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

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 $P(B | A) = P(B)$, 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

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

(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 $1 \leq i_1 < \dots < i_k \leq n$).

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 ≥ 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

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

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

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

(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

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

□

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

Man nennt dann $\mu := \mathcal{L}(X)$ das Produkt der $\mu_1 := \mathcal{L}(X_1), \dots, \mu_n := \mathcal{L}(X_n)$ 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 $1 \leq i_1 < \dots < i_k \leq n$)
(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 $f_i : S_i \rightarrow S'_i$
(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. $P(Y_1 = 1, Y_2 = 1) = P(\{X_1 = X_2 = X_3 = 1\} \cup \{X_1 = X_2 = X_3 = 0\}) = (1/2)^3 + (1/2)^3 = 1/4 = P(Y_1 = 1)P(Y_2 = 1)$, $P(Y_1 = 1, Y_2 = 0) = P(\{X_1 = X_2 = 1, X_3 = 0\} \cup \{X_1 = X_2 = 0, X_3 = 1\}) = (1/2)^3 + (1/2)^3 = 1/4 = P(Y_1 = 1)P(Y_2 = 0)$, 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

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

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

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

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, $f_1, \dots, f_n : \mathbb{R} \rightarrow \mathbb{R}_+$ Wahrscheinlichkeitsdichten (d.h. $\int_{-\infty}^{\infty} f_i(x) dx = 1$), dann sind äquivalent:

- 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 $P(X_1 \in B_1, \dots, X_n \in B_n) = \prod_{i=1}^n P(X_i \in B_i)$ für $B_1, \dots, B_n \subset \mathbb{R}$).
- 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

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

□

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 $\text{vol}(A) = (b_1 - a_1)(b_2 - a_2)$). 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 $A = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}$ in Bsp. 1.36, d.h. $X = (X_1, X_2)$ ist uniform verteilt auf dem Einheitskreis (mit Fläche $\text{vol}(A) = \pi$) mit Dichte

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

Die Marginaldichte ist

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

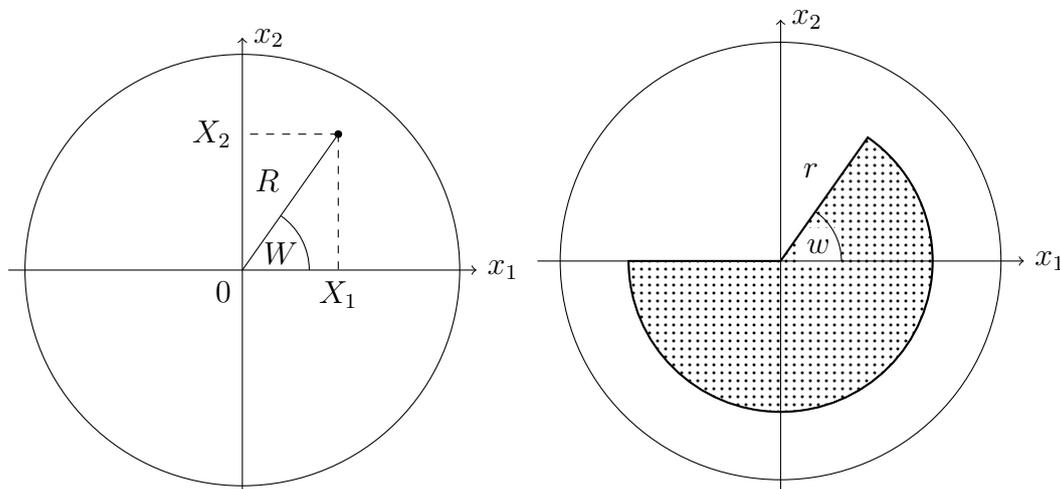
(und analog bzw. offensichtlich aus Symmetrie ist $f_{X_2} = f_{X_1}$).

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

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

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

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



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

Dann sind R und W unabhängig,

$$R \text{ hat Dichte } f_R(r) = 2r \mathbf{1}_{[0,1]}(r),$$

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

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

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

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

$$\frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x_1^2 + \dots + x_d^2)\right) = \prod_{i=1}^d \left(\frac{1}{(2\pi)^{1/2}} e^{-x_i^2/2}\right)$$

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 $g \geq 0$), $X' = (X'_1, X'_2)^T$ entstehe aus X durch Drehung (um Winkel α um den Ursprung), d.h.

$$\begin{pmatrix} X'_1 \\ X'_2 \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \cdot \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \cos(\alpha)X_1 - \sin(\alpha)X_2 \\ \sin(\alpha)X_1 + \cos(\alpha)X_2 \end{pmatrix}.$$

Dann hat X' dieselbe Dichte (und somit dieselbe Verteilung) wie X . (Dies ist anschaulich sehr plausibel, man kann es z.B. mit Bericht 1.59 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 $2\pi r g(r) \mathbf{1}_{[0, \infty)}(r)$:

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

(wir verwenden hier etwas salopp die „Polarkoordinatenform des Flächenelements“ $dx_1 dx_2 = r dr dw$, man kann dies wiederum mit Bericht 1.59 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 $u > 0$):

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

d.h. $R^2 \sim \text{Exp}_{1/2}$.

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, $\varphi : I \rightarrow J$ bijektiv, stetig differenzierbar mit Ableitung

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



(wobei $\varphi(x) = (\varphi_1(x), \dots, \varphi_d(x))^T$, d.h. φ_i ist die i -te Koordinatenfunktion von φ), dann hat $Y := \varphi(X)$ die Dichte

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

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“

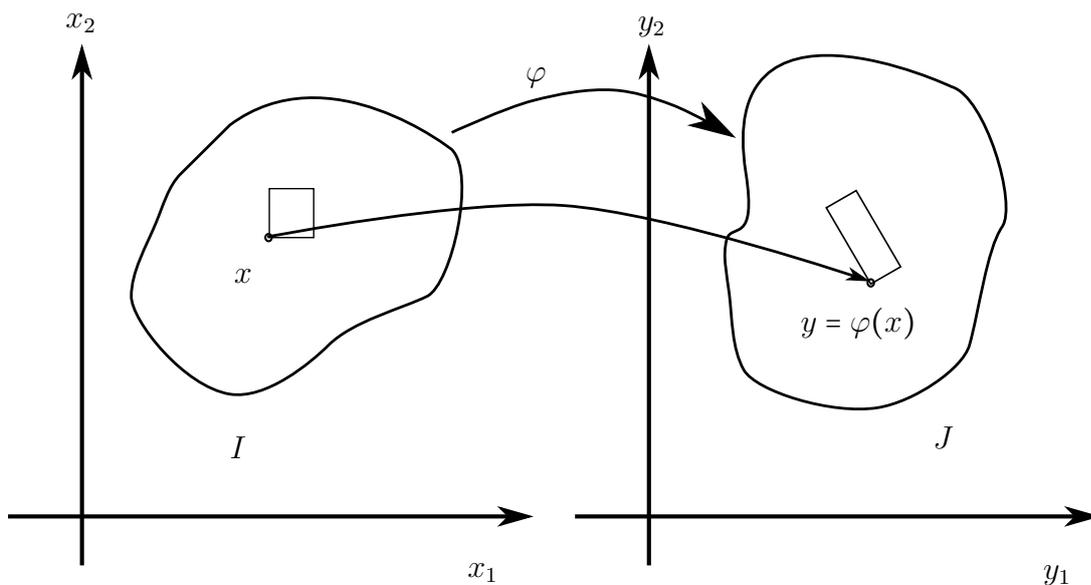
$$\begin{aligned} \varphi(x') &\approx \varphi(x) + \varphi'(x) \cdot (x' - x) \\ &= \varphi(x) + \begin{pmatrix} \frac{\partial}{\partial x_1} \varphi_1(x) & \frac{\partial}{\partial x_2} \varphi_1(x) \\ \frac{\partial}{\partial x_1} \varphi_2(x) & \frac{\partial}{\partial x_2} \varphi_2(x) \end{pmatrix} \cdot \begin{pmatrix} x'_1 - x_1 \\ x'_2 - x_2 \end{pmatrix} \end{aligned}$$

(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

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

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

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

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) $\Rightarrow F_X(x_n) \rightarrow F(x)$
2. $F_X(x_n) \rightarrow 1$ wenn $x_n \rightarrow (+\infty, \dots, +\infty)$
3. $F_X(x_n) \rightarrow 0$ wenn $\min_{i=1, \dots, d} x_{n,i} \rightarrow -\infty$
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 $\{u = (z_1^{(i_1)}, \dots, z_d^{(i_d)}) : i_1, \dots, i_d \in \{1, 2\}\}$ wo $z_j^{(1)} = x_j$, $z_j^{(2)} = y_j$, es muss gelten

$$\sum_{u \text{ Ecken}} (-1)^{\#\{1 \leq m \leq d : i_m = 1\}} F_X(u) \left(= \mu_F((x, y]) \right) \geq 0$$

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 μ und ν , geschrieben $\mu * \nu$:

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

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

Beobachtung 1.62 (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

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

mit

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

für $k \in \{2, 3, \dots, 12\}$

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

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

3. (Binomialfamilie) X_1, X_2, \dots, X_n u.a., $\sim \text{Ber}_p$, so ist $X_1 + X_2 + \dots + X_n \sim \text{Bin}_{n,p}$, d.h.

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

Insbesondere gilt

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

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

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

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

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

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

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

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 2×2 -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

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

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

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

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 (Ω, \mathcal{F}, P) definiert), d.h. es gibt eine abzählbare Menge $S = S_X \subset \mathbb{R}$ mit $P(X \in S) = 1$ und $\mathcal{L}_P(X)$ 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

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

(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

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

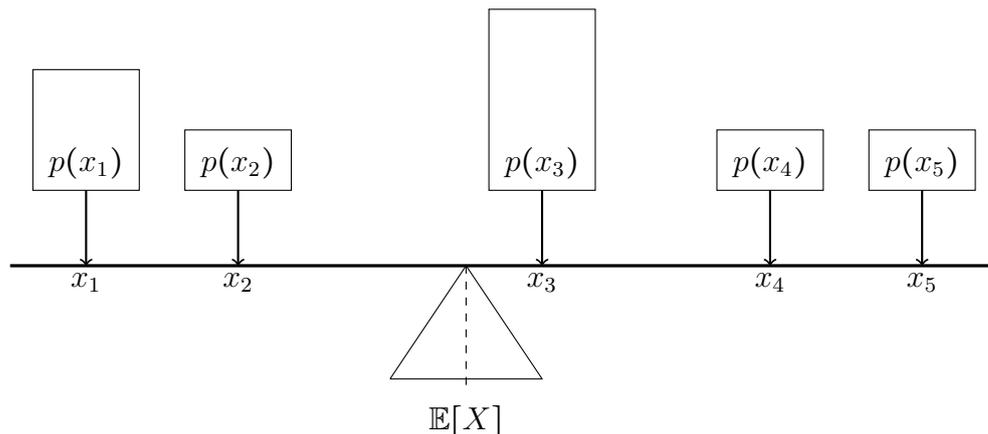
(denn $\sum_{j=1}^{3N} x_j P(X = x_j) = \sum_{i=2}^N \frac{-i}{2i(i-1)} + \sum_{i=2}^{2N} \frac{i}{2i(i-1)} = -\frac{1}{2} \sum_{k=1}^{N-1} \frac{1}{k} + \frac{1}{2} \sum_{k=1}^{2N-1} \frac{1}{k} \sim -\frac{1}{2} \log(N) + \frac{1}{2} \log(2N) = \frac{1}{2} \log(2)$).

(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 Gesamtdrehmoment (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

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

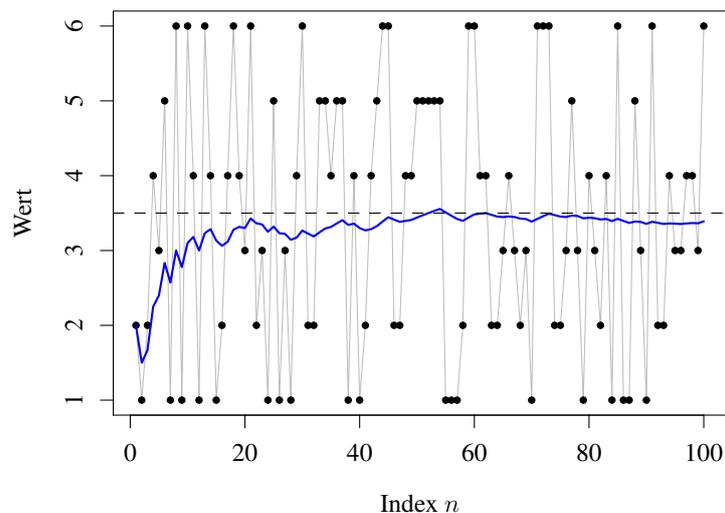
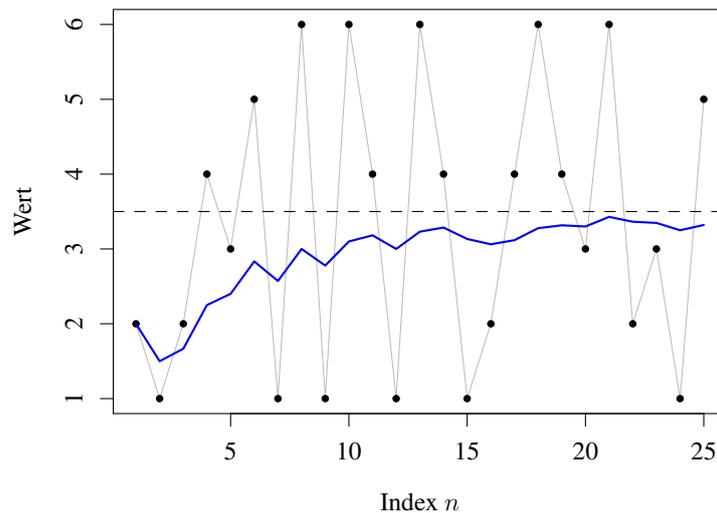
(in geeignetem Sinn), dies ist die Aussage des *Gesetzes der großen Zahlen*, das wir später sehen werden.

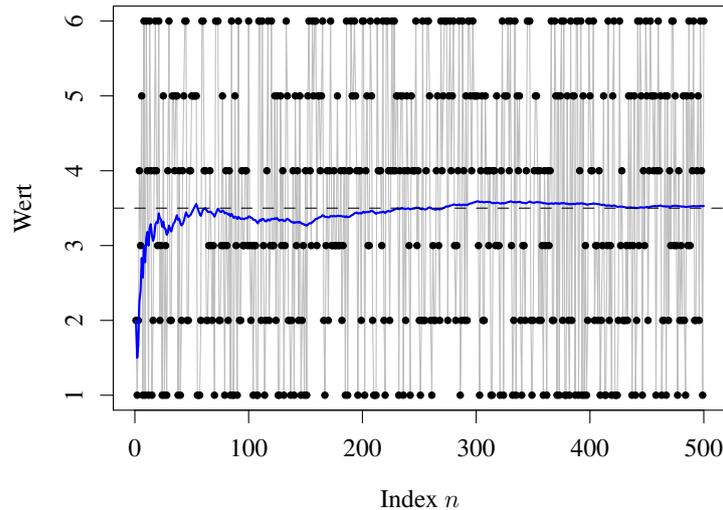
Es ist nämlich

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

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

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: $\mathcal{L}(X) = \mathcal{L}(Y)$ 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 $X \sim \text{Bin}_{n,p}$, $n \in \mathbb{N}$, $p \in [0, 1]$:

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

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

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

3. Sei $X \sim \text{Poi}_{\alpha}$, $\alpha > 0$:

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

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, $g : \mathbb{R} \rightarrow \mathbb{R}$, $Y := g(X)$.

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., $\sim \text{Ber}_p$, so ist $X := X_1 + \dots + X_n \sim \text{Bin}_{n,p}$ 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 $X \sim \text{Hyp}_{s,w,k}$ 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 $P(A_1) = P(A_2) = \dots = P(A_k) = \frac{s}{s+w}$, also $\mathbb{E}[X] = k \cdot \frac{s}{s+w}$.

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 $(2\pi)^{-1/2} \int_{-\infty}^{\infty} |x| e^{-x^2/2} dx = (2/\pi)^{1/2} \int_0^{\infty} x e^{-x^2/2} dx = (2/\pi)^{1/2} [-e^{-x^2/2}]_0^{\infty} = (\pi/2)^{-1/2} < \infty$)

Somit gilt auch: $Y \sim \mathcal{N}_{\mu, \sigma^2}$ hat $\mathbb{E}[Y] = \mu$, denn $\sigma X + \mu \stackrel{d}{=} Y$ nach Bsp. 1.32).

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

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

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

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

Bericht 1.75. 1. Man kann prinzipiell den Fall mit Dichte aus dem diskreten Fall herleiten: X habe Dichte f_X , so nimmt $X_{(n)} = \frac{1}{n} \lfloor nX \rfloor$ den Wert $\frac{k}{n}$, $k \in \mathbb{N}$ an mit

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

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

$$\begin{aligned}\mathbb{E}[X_{(n)}] &= \sum_{k \in \mathbb{Z}} \frac{k}{n} \int_{k/n}^{(k+1)/n} f_X(x) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{n} [nx] f_X(x) dx \xrightarrow{n \rightarrow \infty} \int_{-\infty}^{\infty} x f_X(x) dx\end{aligned}$$

2. (Analogon zu Beob. 1.71 im Fall mit Dichte)

Sei $X = (X_1, \dots, X_d) \mathbb{R}^d$ -wertig mit Dichte $f_X : \mathbb{R}^d \rightarrow [0, \infty]$, $g : \mathbb{R}^d \rightarrow \mathbb{R}$, $Y := g(X)$. 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. $\text{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ die *Varianz* von X
(manchmal schreibt man auch $\sigma_X^2 := \text{Var}[X]$),
 $\sqrt{\text{Var}[X]}$ die *Standardabweichung* (oder *Streuung*) von X
(manchmal auch $\sigma_X = \sqrt{\sigma_X^2}$ geschrieben),
2. $\text{Cov}[X, Y] := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
die *Kovarianz* von X und Y .

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

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

Einen Hinweis dazu gibt die Chebyshev-Ungleichung:

Satz 1.77. Sei X reelle ZV, $f : [0, \infty) \rightarrow [0, \infty)$ monoton wachsend.

1. Für $a > 0$ mit $f(a) > 0$ gilt

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

2. Für $X \in \mathcal{L}^2$ gilt

$$P(|X - \mathbb{E}[X]| \geq a) \leq \frac{\text{Var}[X]}{a^2} \quad (\text{Chebyshev}^{13}\text{-Ungleichung}). \quad (1.17)$$

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

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

nach Satz 1.70, 2.

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

Insbesondere (wähle $a = b\sigma_X$ in (1.17)): Die W'keit $P(|X - \mathbb{E}[X]| \geq b\sigma_X)$, dass X von $\mathbb{E}[X]$ um mehr als das b -fache von σ_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)

$$\text{Cov}[X, Y] = \text{Cov}[Y, X].$$

2. Es ist

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[YX] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

(und analog für $\text{Var}[X] = \text{Cov}[X, X]$).

3. $\text{Var}[X] = 0 \iff P(X = \mathbb{E}[X]) = 1$

(„ \Leftarrow “ ist klar, für „ \Rightarrow “ wende Satz 1.70, 3. an auf die ZV $(X - \mathbb{E}[X])^2$)

4. $\text{Var}[X]$ ist eine Eigenschaft der Verteilung von X , $\text{Cov}[X, Y]$ ist eine Eigenschaft der gemeinsamen Verteilung von X und Y .

Beispiel 1.79. 1. $X \sim \text{Ber}_p$, $\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p)$.

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

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

¹³Andrei Andrejewich Markov, 1856–1922.

¹³Pafnuty Lvovich Chebyshev, 1821–1894.

also

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

3. $X \sim \text{Bin}_{n,p}$,

$$\begin{aligned} \mathbb{E}[X(X-1)] &= \sum_{k=0}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^n \binom{n-2}{k-2} p^{k-2} (1-p)^{(n-2)-(k-2)} = n(n-1)p^2 \end{aligned}$$

also

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

4. $X \sim \text{Geom}_p$, $p \in [0, 1]$ (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

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

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

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

(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 $X, Y, X_1, X_2, \dots, X_n \in \mathcal{L}^2$, $a, b, c, d \in \mathbb{R}$.

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

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

insbesondere

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

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

$$2. \text{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \text{Var}[X_i] + \sum_{\substack{1 \leq i, j \leq n \\ i \neq j}} \text{Cov}[X_i, X_j],$$

insbesondere gilt für paarweise unkorrelierte X_1, \dots, X_n also $\text{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \text{Var}[X_i]$.

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

4. Es gilt

$$|\text{Cov}[X, Y]| \leq \sqrt{\text{Var}[X]} \sqrt{\text{Var}[Y]} \quad (\text{Cauchy-Schwarz-Ungleichung}^{14})$$

Beweis. 1. Es ist

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

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

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

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

3. Klar, denn für X und Y unabhängig ist $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$ nach Satz 1.70, 4.

4. Falls $\text{Var}[Y] = 0$, so ist die Ungleichung (als $0 \leq 0$) erfüllt

(denn dann ist $P(Y - \mathbb{E}[Y] = 0) = 1$ nach Beob. 1.78, 3. und somit auch $\text{Cov}[X, Y] = 0$).

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

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

$$\begin{aligned} 0 \leq \text{Var}[X + \alpha Y] \text{Var}[Y] &\stackrel{!}{=} (\text{Var}[X] + 2\alpha \text{Cov}[X, Y] + \alpha^2 \text{Var}[Y]) \text{Var}[Y] \\ &= \text{Var}[X] \text{Var}[Y] - (\text{Cov}[X, Y])^2. \end{aligned}$$

□

Bemerkung 1.81. Es gilt Gleichheit in der Cauchy-Schwarz-Ungleichung g.d.w.

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

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

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

Beispiel 1.82. 1. $X \sim \text{Bin}_{n,p}$, schreibe $X = Y_1 + \dots + Y_n$ mit Y_i u.i.v. $\sim \text{Ber}_p$, so ist (mit Satz 1.80, 2.)

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

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

2. $X \sim \text{Hyp}_{s,w,n}$, stelle dar als $X = Y_1 + \dots + Y_n$ 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 $y_1 + \dots + y_n = k$ gilt

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

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

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

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

also

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

(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 $\mathbb{E}[|Z|^3] < \infty$ 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

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

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

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

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

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

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

denn der Ausdruck auf der linken Seite ist

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

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:

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

ist minimal für die Wahl $\beta_0 = \mathbb{E}[Y]$.)

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)

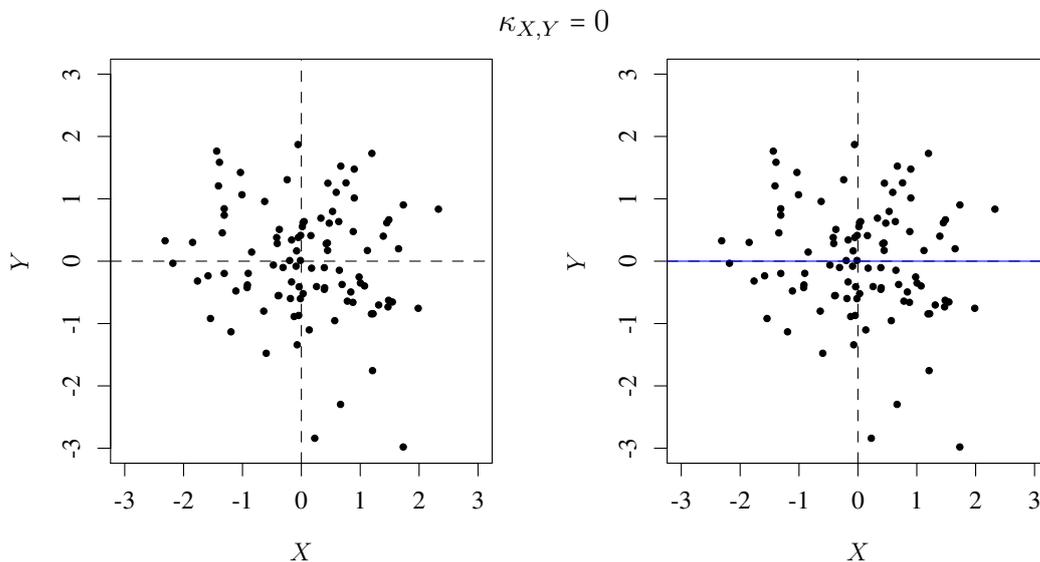
$|\kappa_{X,Y}| = 1 \leftrightarrow$ perfekter linearer Zusammenhang zwischen X und Y

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

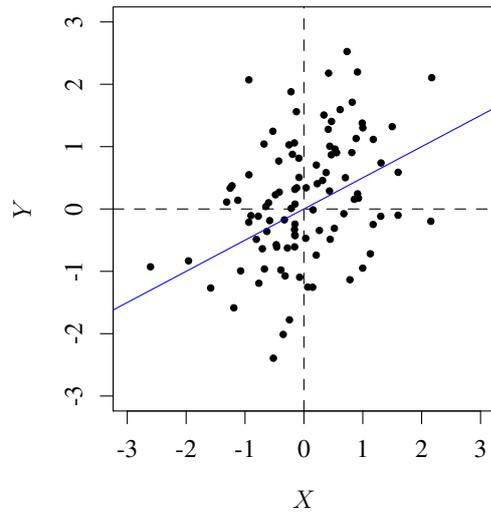
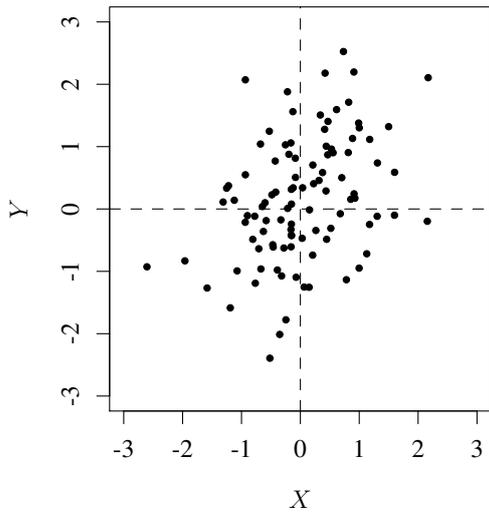
$\kappa_{X,Y} = -1 \leftrightarrow$ perfekter linearer Zusammenhang zwischen X und Y
mit negativem Koeffizienten
(X größer als $\mathbb{E}[X] \iff Y$ kleiner als $\mathbb{E}[Y]$)

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

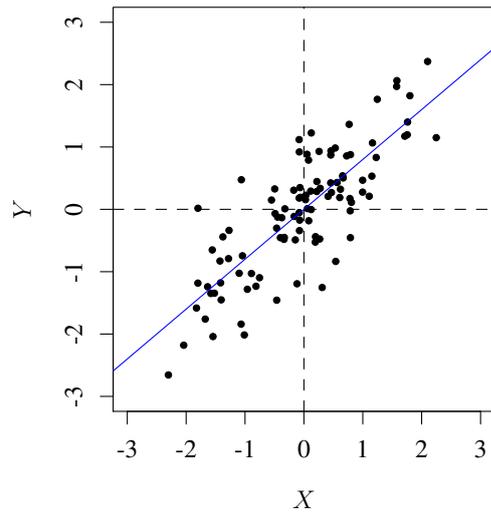
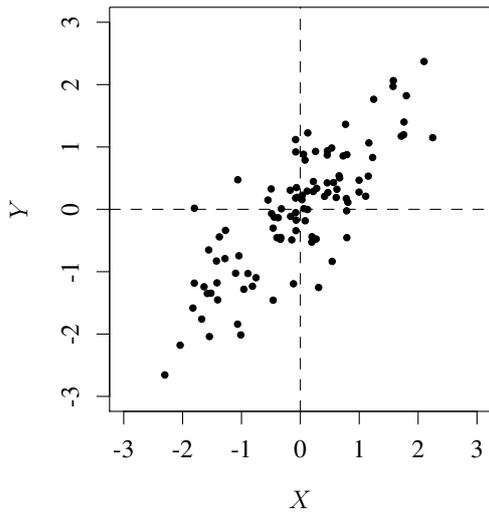
Die folgenden Scatterplots zeigen jeweils 100 simulierte Paare (X, Y) , wobei $\sigma_X = \sigma_Y = 1$ und $\kappa_{X,Y}$ den angegebenen Wert hat. (Blau eingezeichnet ist die „Vorhersagegerade“ $x \mapsto \beta_1^* x + \beta_0^*$.)



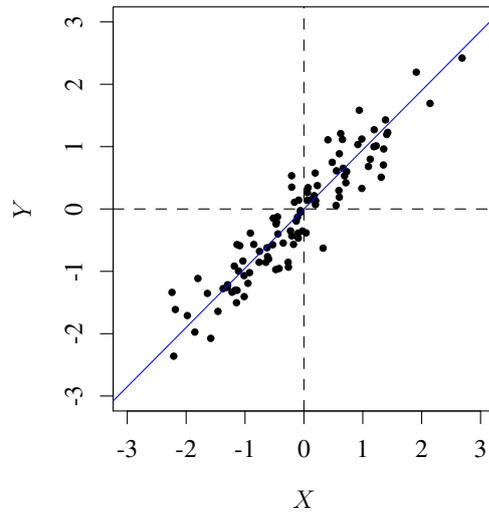
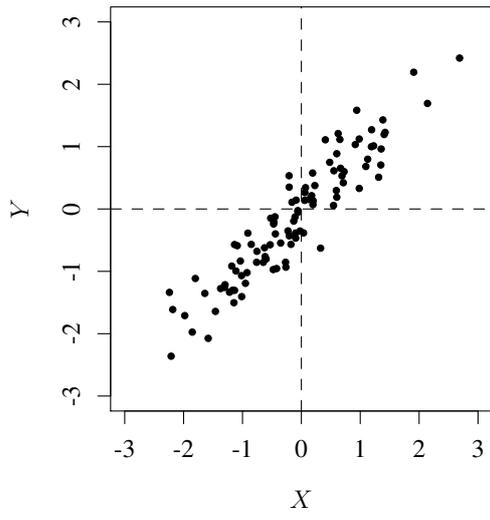
$\kappa_{X,Y} = 0.5$



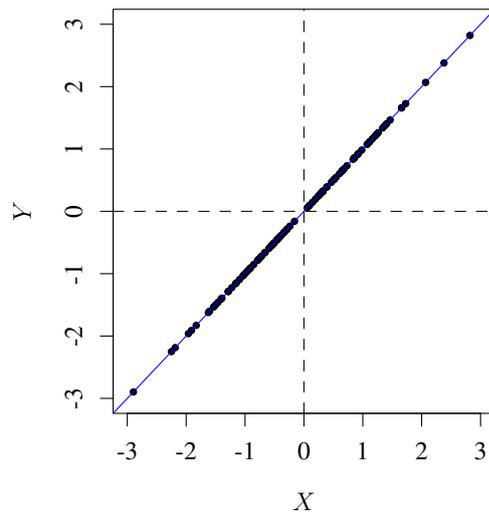
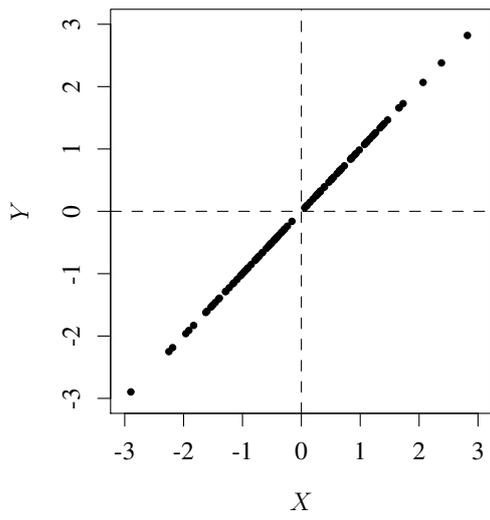
$\kappa_{X,Y} = 0.8$



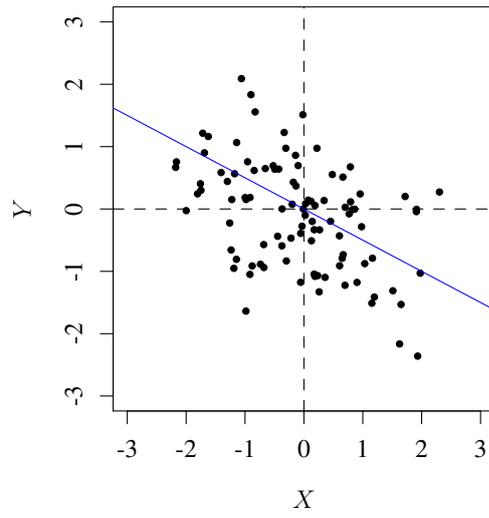
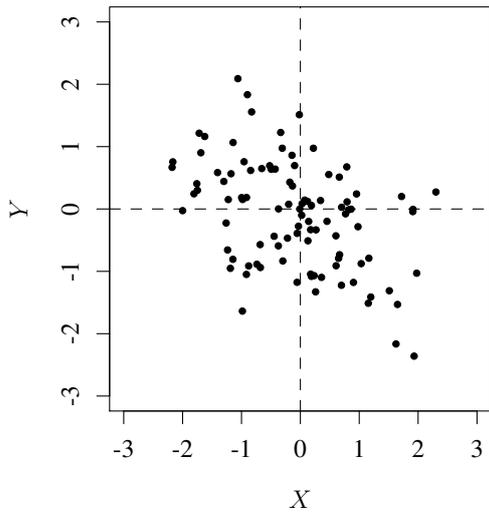
$$\kappa_{X,Y} = 0.95$$



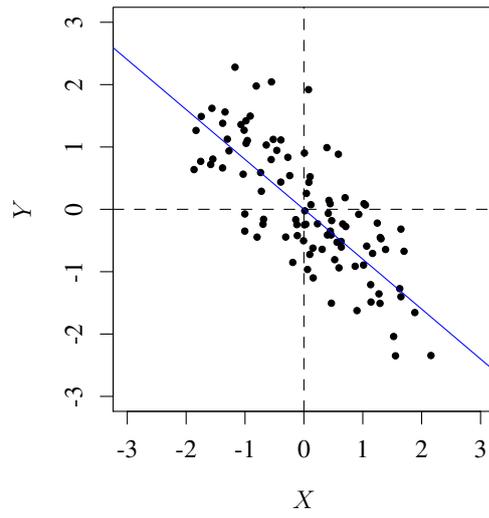
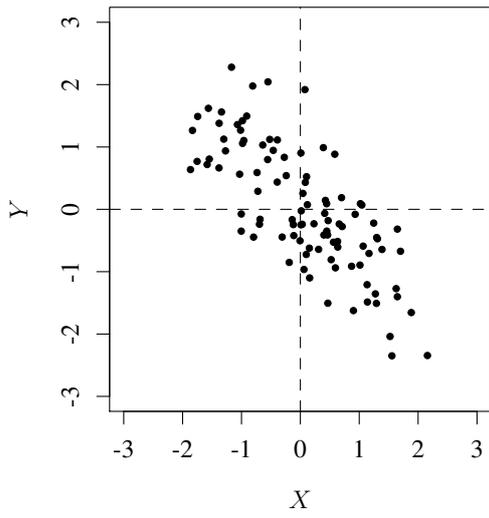
$$\kappa_{X,Y} = 1$$



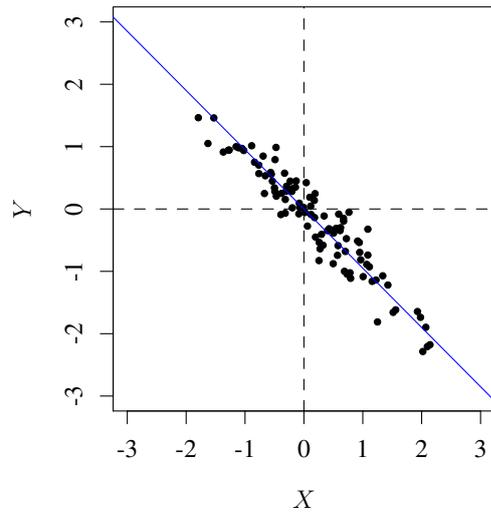
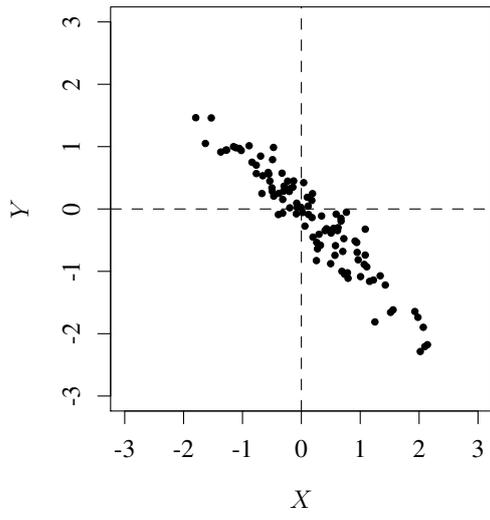
$$\kappa_{X,Y} = -0.5$$



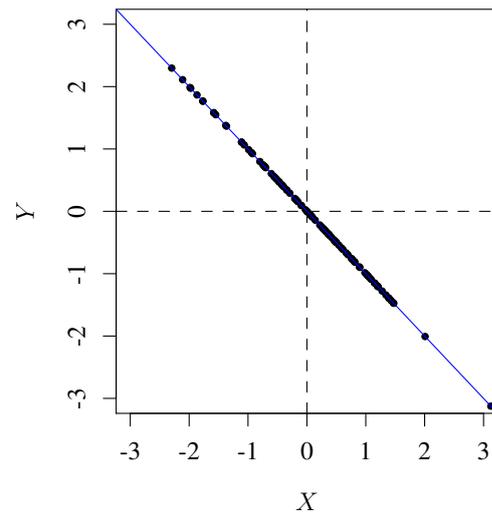
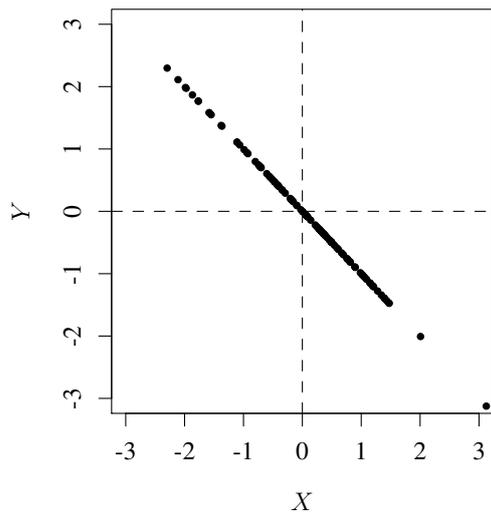
$$\kappa_{X,Y} = -0.8$$

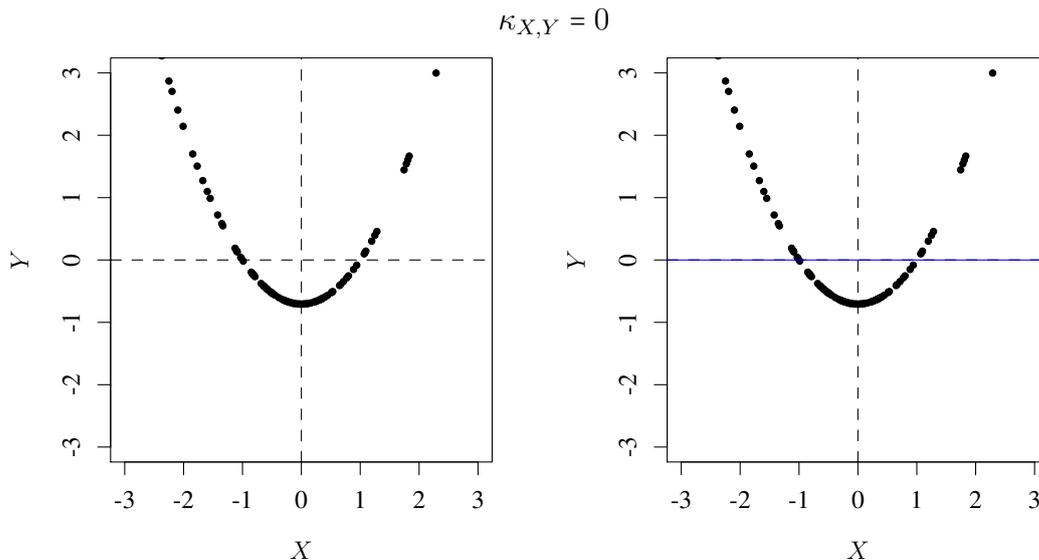


$$\kappa_{X,Y} = -0.95$$



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





1.4.4 Median(e)

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

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

gilt.

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

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

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

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

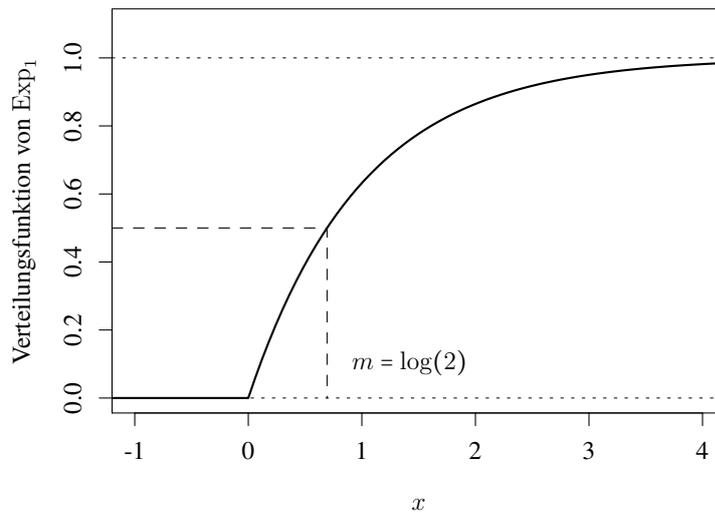
Man kann den Median als eine „robustere“ Antwort auf die Aufgabe, für eine ZV *nur einen* „typischen Wert“ anzugeben, ansehen (im Gegensatz zum Erwartungswert besitzt ja jede Verteilung einen Median). Allerdings gibt es für Mediane keine so angenehmen Rechenregeln, wie sie Satz 1.70 für den Erwartungswert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

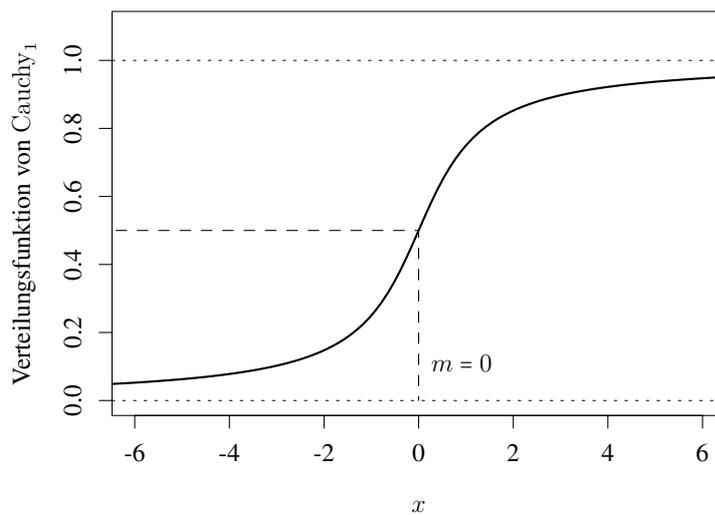
$$\frac{1}{2} \left(\inf \{x \in \mathbb{R} : F_X(x) = \frac{1}{2}\} + \sup \{x \in \mathbb{R} : F_X(x) = \frac{1}{2}\} \right),$$

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

Beispiel 1.86. 1. $X \sim \text{Exp}_\theta$ 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 $m = \frac{1}{\theta} \log 2$.

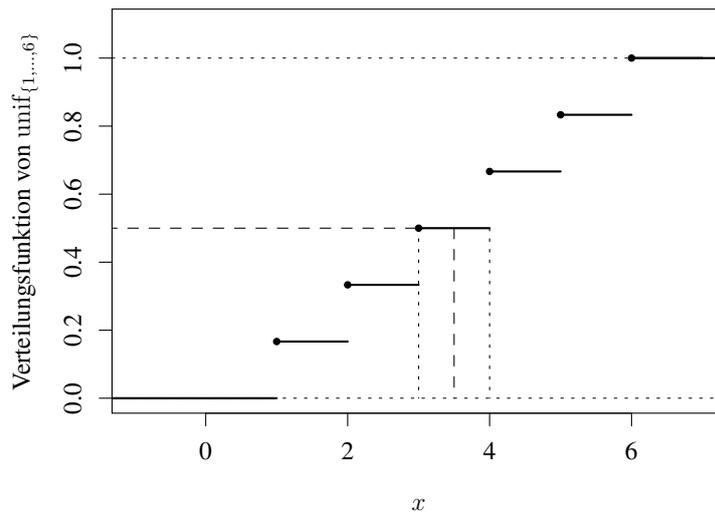


2. X Cauchy-verteilt mit Dichte $\frac{1}{\pi} \frac{1}{1+x^2}$, Verteilungsfunktion $\frac{1}{2} \frac{1}{\pi} \arctan(x)$, der (eindeutig bestimmte) Median ist $m = 0$.



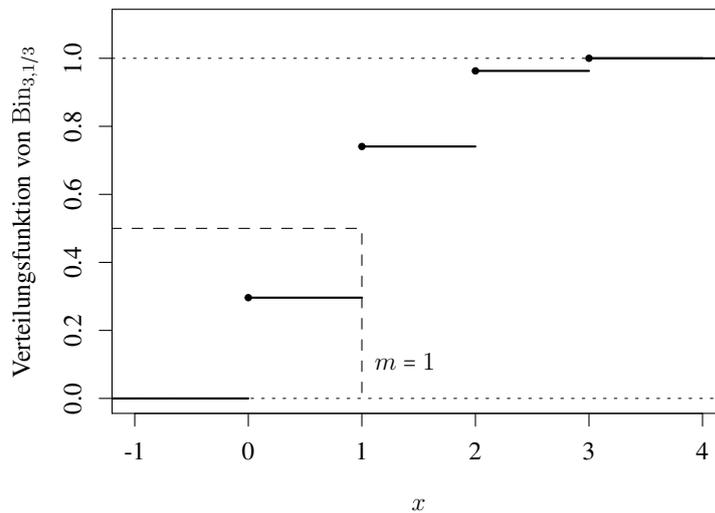
(Wegen der Symmetrie der Dichte, es gibt keinen Erwartungswert, vgl. Bsp. 1.74, 3.)

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



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

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



Bemerkung 1.87. Sei $X \in \mathcal{L}^1$.

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

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

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

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

also

$$\mathbb{E}[|X - a|] - \mathbb{E}[|X - m|] \geq (a - m) \left(\underbrace{P(X \leq m)}_{\geq 1/2} - \underbrace{P(X > m)}_{\leq 1/2} \right) \geq 0,$$

analog im Fall $a < m$.

2. Es ist

$$\begin{aligned} |\mathbb{E}[X] - m| &= |\mathbb{E}[X - m]| \leq \mathbb{E}[|X - m|] \\ &\stackrel{1.}{\leq} \mathbb{E}[|X - \mathbb{E}[X]|] = \sqrt{\left(\mathbb{E}[|X - \mathbb{E}[X]|]\right)^2} \leq \sqrt{\mathbb{E}[|X - \mathbb{E}[X]|^2]}, \end{aligned}$$

wobei für die erste Ungleichung die Monotonie des Erwartungswerts (Satz 1.70, 2., beachte: $X - m \leq |X - m|$ und $-(X - m) \leq |X - m|$) und für die letzte Ungleichung die Cauchy-Schwarz-Ungleichung (Satz 1.80, 4.) verwenden. \square

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 $\text{Var}[X_1] < \infty$, dann gilt für jedes $\varepsilon > 0$*

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\text{Var}[X_1]}{\varepsilon^2 n} \xrightarrow{n \rightarrow \infty} 0 \quad (1.18)$$

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

$$\begin{aligned} \text{Var}[Y_n] &= \frac{1}{n^2} \left(\sum_{i=1}^n \text{Var}[X_i - \mu] + \sum_{\substack{1 \leq i, j \leq n \\ i \neq j}} \text{Cov}[X_i - \mu, X_j - \mu] \right) \\ &= \frac{1}{n^2} (n \cdot \text{Var}[X_1] + 0) = \frac{1}{n} \text{Var}[X_1] \end{aligned}$$

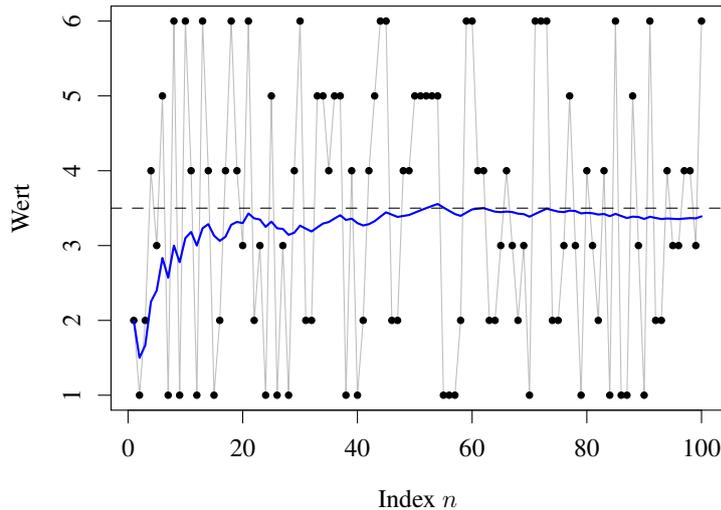
mit Satz 1.80, somit

$$P(|Y_n| \geq \varepsilon) \leq \frac{\text{Var}[Y_n]}{\varepsilon^2} = \frac{\text{Var}[X_1]}{\varepsilon^2 n}$$

gemäß Chebyshev-Ungleichung (Satz 1.77). \square

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, $(X_1 + \dots + X_n)/n$ 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

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

dann gilt für $\varepsilon > 0$

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

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

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

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

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

Man spricht damit Satz 1.88 oft folgendermaßen aus:

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

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:

Für jedes $\varepsilon > 0$ gibt es ein (vom Zufall abhängiges) N_0 mit

$$\left| \frac{X_1 + \dots + X_n}{n} - \mu \right| \leq \varepsilon \quad \text{für alle } n \geq N_0.$$

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 μ .

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.), $\mathbb{E}[X_1] = \mu$, $\text{Var}[X_1] = \sigma^2 < \infty$.

Wir haben gesehen, dass $X_1 + \dots + X_n \approx n\mu$ mit hoher Wahrscheinlichkeit, denn

$$\frac{X_1 + \dots + X_n}{n} - \mu \xrightarrow[n \rightarrow \infty]{\text{stochastisch}} 0$$

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

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

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

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

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

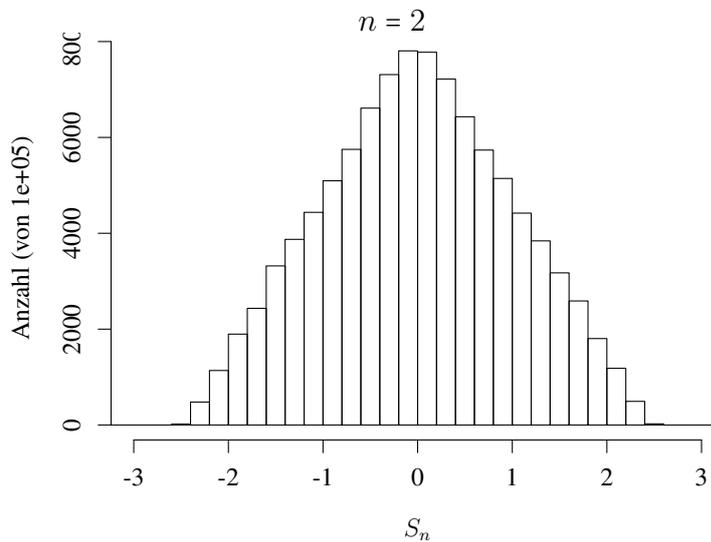
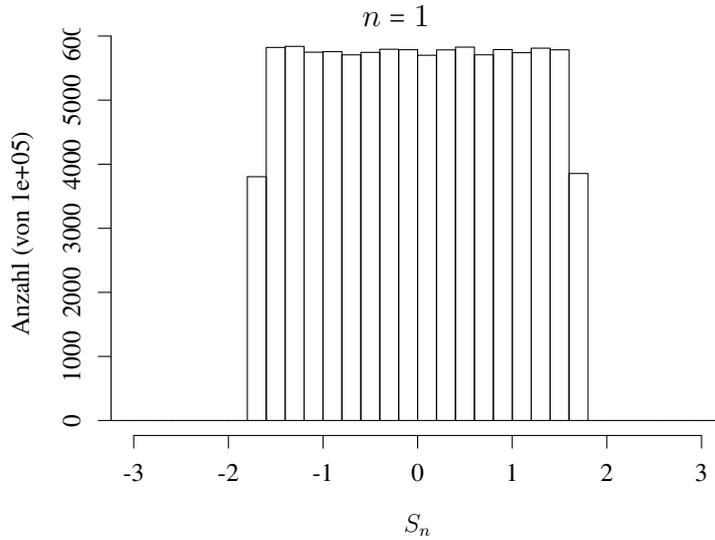
$$\text{Var}\left[\frac{(X_1 + \dots + X_n) - n\mu}{\sqrt{n\sigma^2}}\right] = 1$$

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

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

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

Histogramme jeweils basierend auf 10^5 Simulationen von S_n



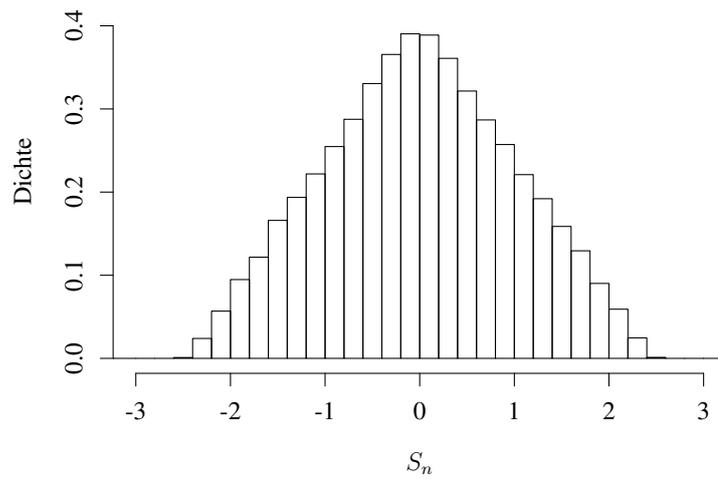
Zur Vergleichbarkeit gehen wir von den absoluten Anzahlen als Balkenhöhen zur sogenannten Dichte über, d.h. die Balkenhöhe ist nun jeweils

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

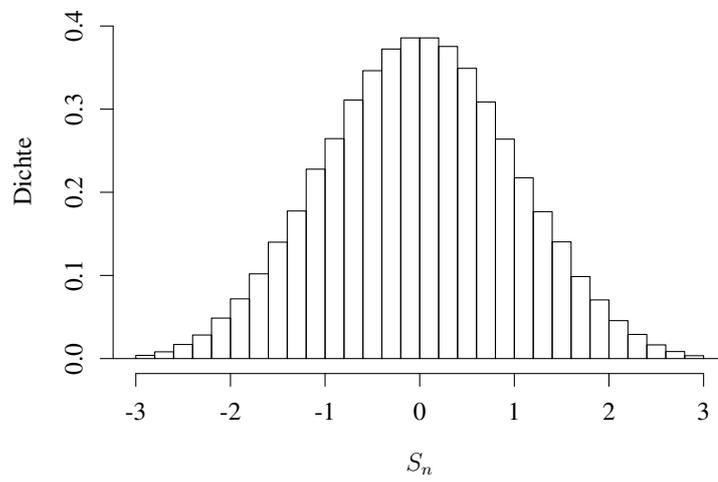
Damit wird die Gesamtfläche der Balken = 1 (wie bei einer Wahrscheinlichkeitsdichte).

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

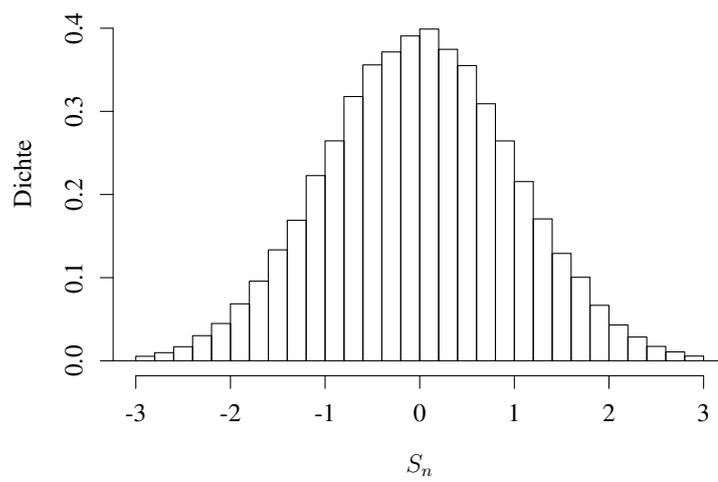
$n = 2$



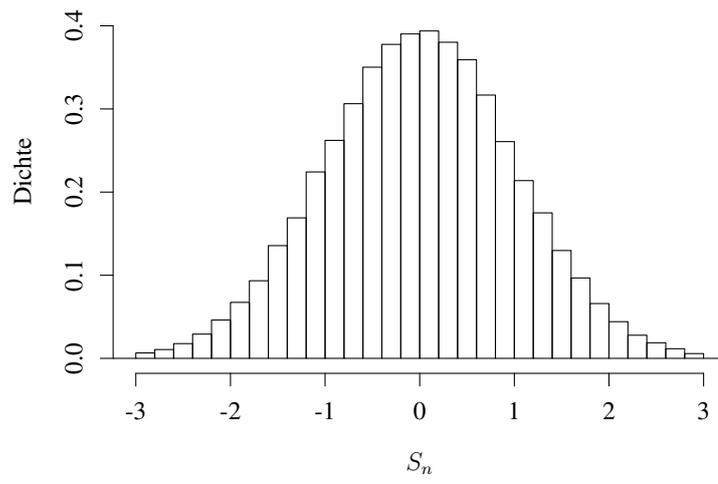
$n = 5$



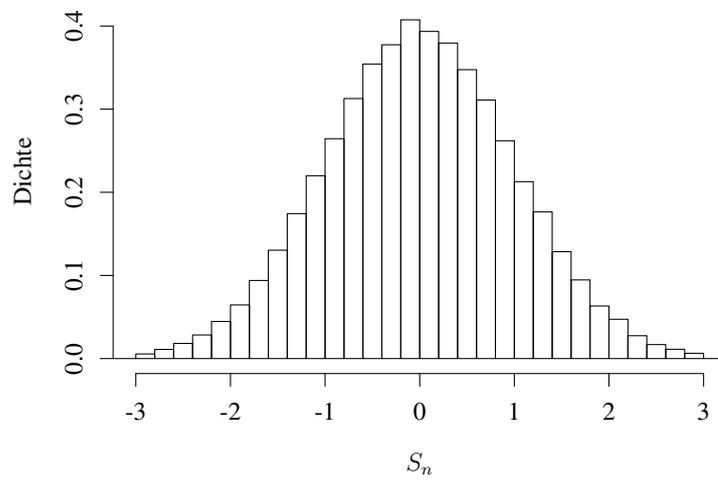
$n = 20$



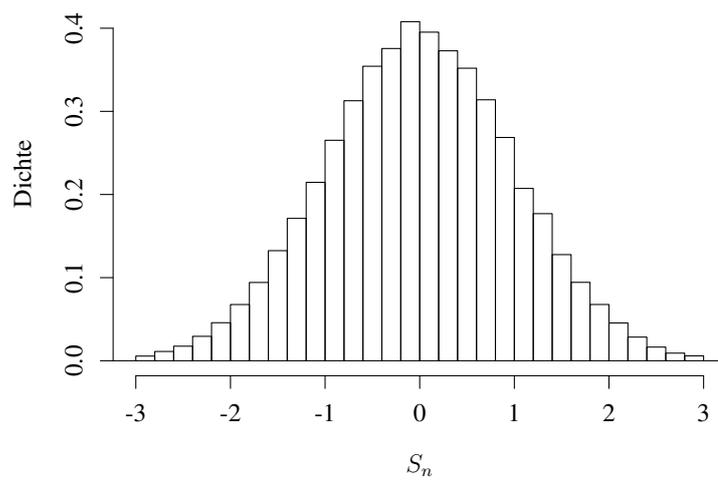
$n = 100$



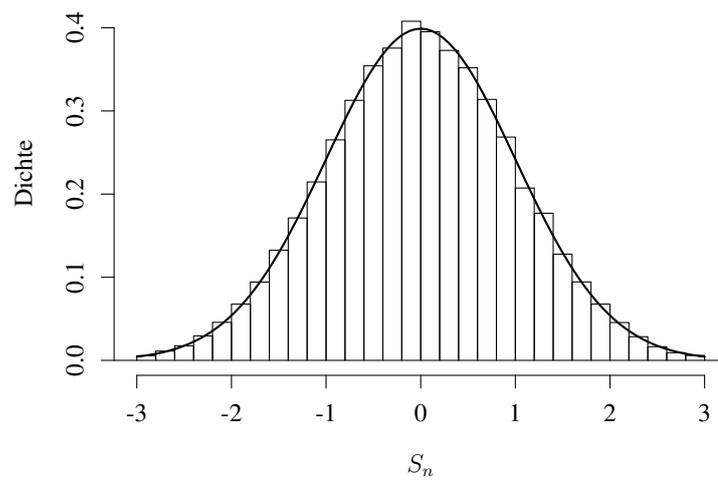
$n = 200$



$n = 500$

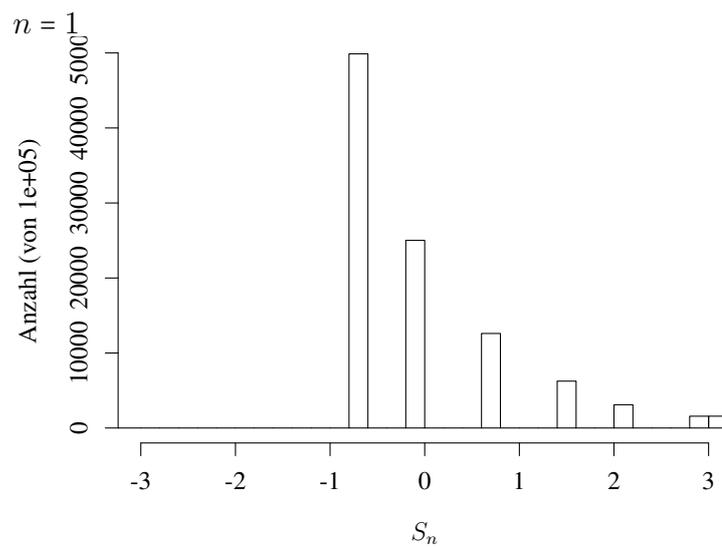


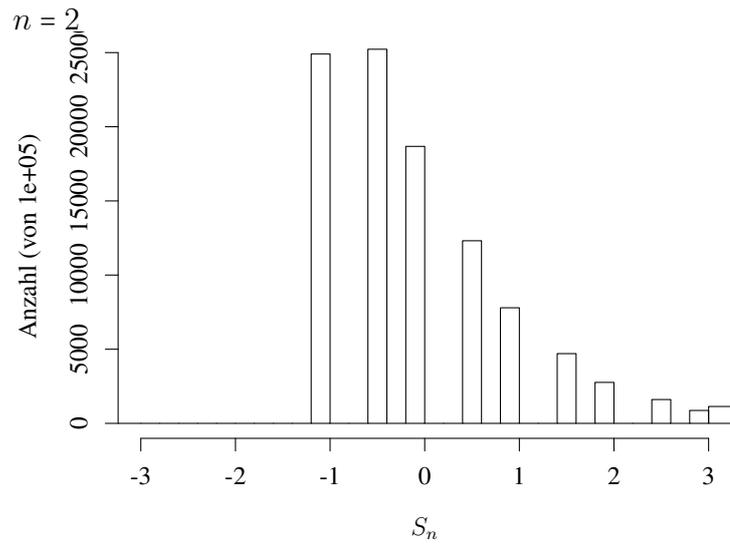
$n = 500$



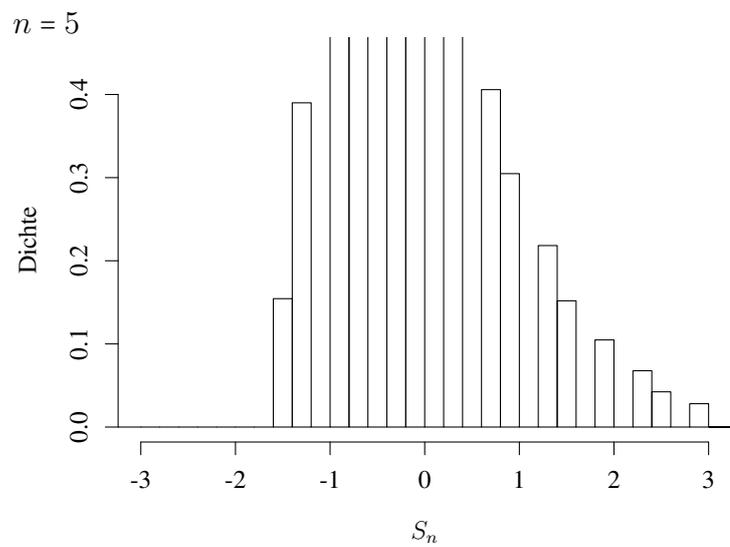
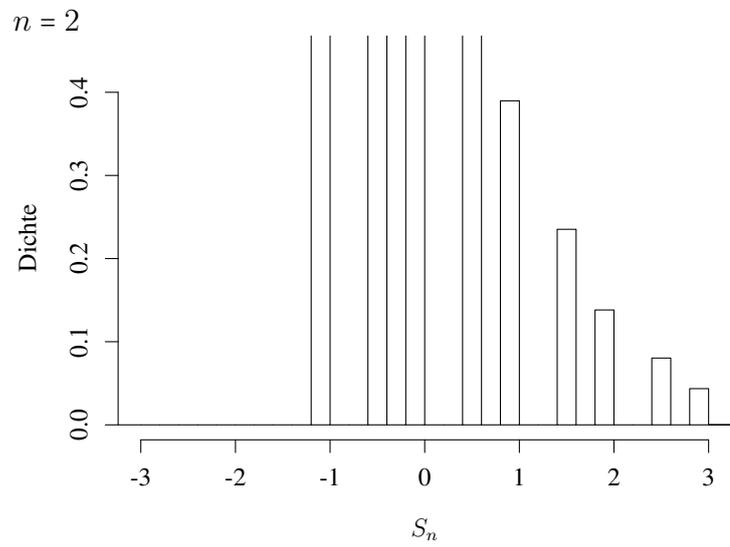
(Die schwarze Kurve ist die Dichte der Standard-Normalverteilung, $f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$.)

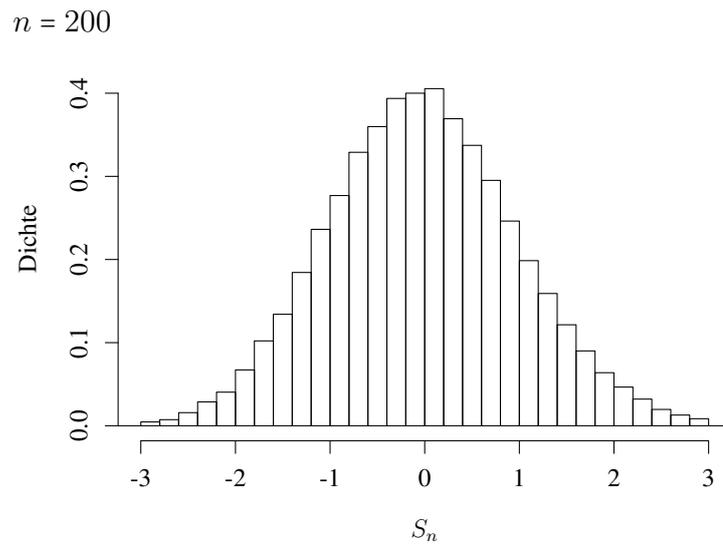
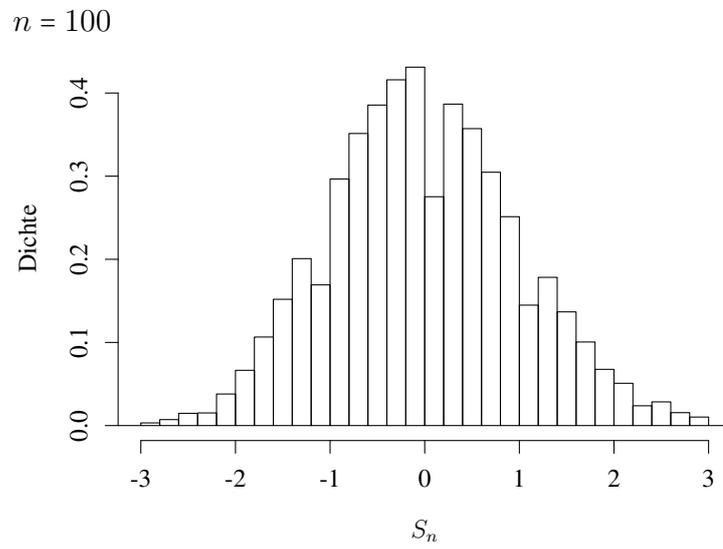
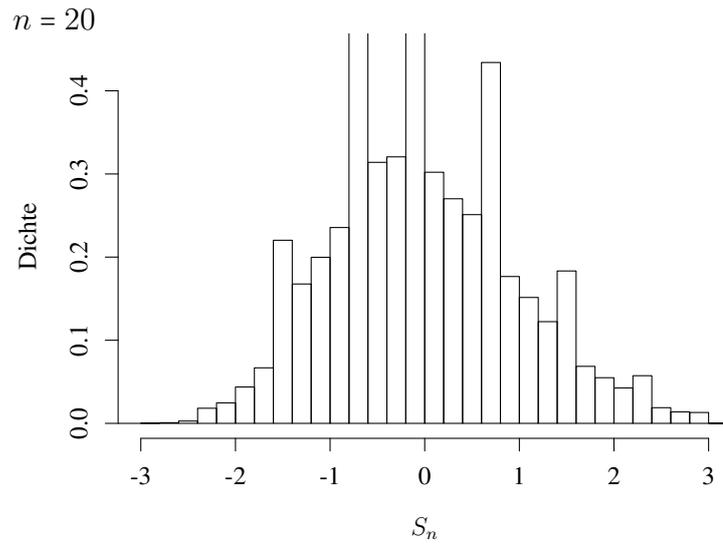
Nun dasselbe nochmal mit $X_i \sim \text{geom}_{1/2}$ (mit $\mathbb{E}[X_1] = 1$, $\text{Var}[X_1] = 2$):
Histogramme jeweils basierend auf 10^5 Simulationen von S_n

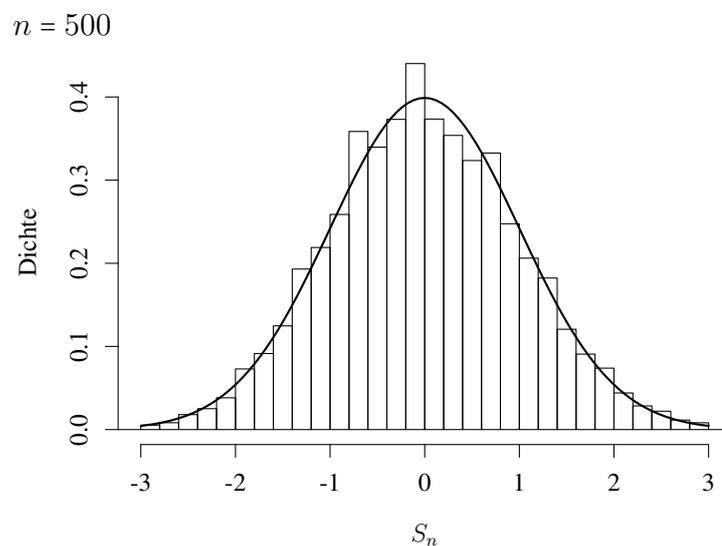
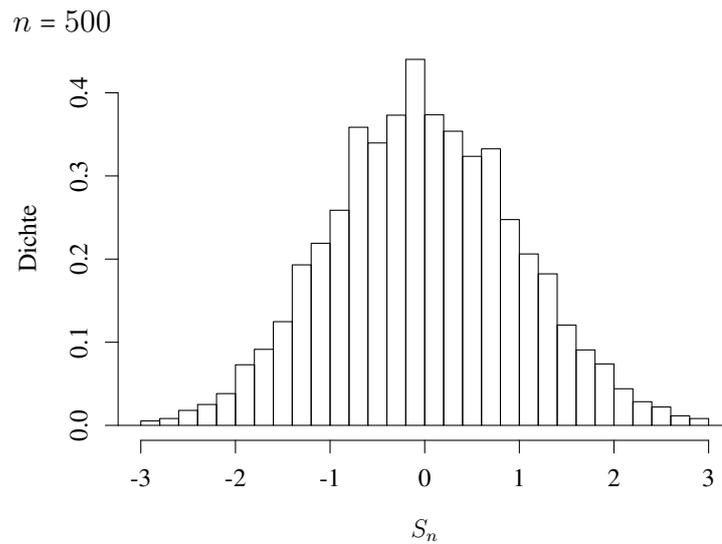




Zur Vergleichbarkeit gehen wir wieder von den absoluten Anzahlen als Balkenhöhen zur sogenannten Dichte über.







(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

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

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

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

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, \dots u.i.v. reelle ZVn $\in \mathcal{L}^2$ mit $\text{Var}[X_1] \in (0, \infty)$, dann gilt für $-\infty \leq a < b \leq \infty$

$$\lim_{n \rightarrow \infty} P\left(a \leq \frac{X_1 + \dots + X_n - n\mathbb{E}[X_1]}{\sqrt{n\text{Var}[X_1]}} \leq b\right) = P(a \leq Z \leq b) \quad \text{mit } Z \sim \mathcal{N}_{0,1}. \quad (1.21)$$

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

$$X_n \xrightarrow[n \rightarrow \infty]{} X \text{ in Verteilung} \quad \left(\text{auch } X_n \xrightarrow[n \rightarrow \infty]{d} X \text{ oder } X_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} X \text{ geschrieben}\right),$$

wenn gilt

$$\lim_{n \rightarrow \infty} P(X_n \leq x) = P(X \leq x) \quad (= F_X(x))$$

für jedes $x \in \mathbb{R}$, an dem F_X stetig ist.

Satz 1.91 besagt also: $\frac{X_1 + \dots + X_n - n\mathbb{E}[X_1]}{\sqrt{n\text{Var}[X_1]}} \xrightarrow[n \rightarrow \infty]{d} Z$

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?“

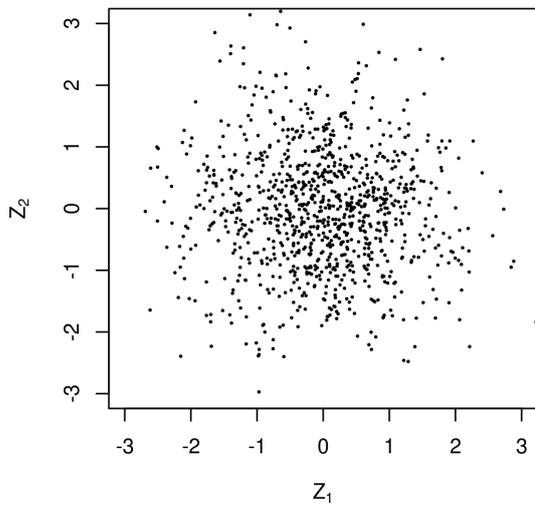
Beobachtung. In der Situation des zentralen Grenzwertsatzes sei

$$(Z_1, Z_2) := \left(\frac{X_1 + X_2 + \dots + X_n - n\mu_{X_1}}{\sqrt{n\sigma_{X_1}^2}}, \frac{X_{n+1} + X_{n+2} + \dots + X_{2n} - n\mu_{X_1}}{\sqrt{n\sigma_{X_1}^2}} \right),$$

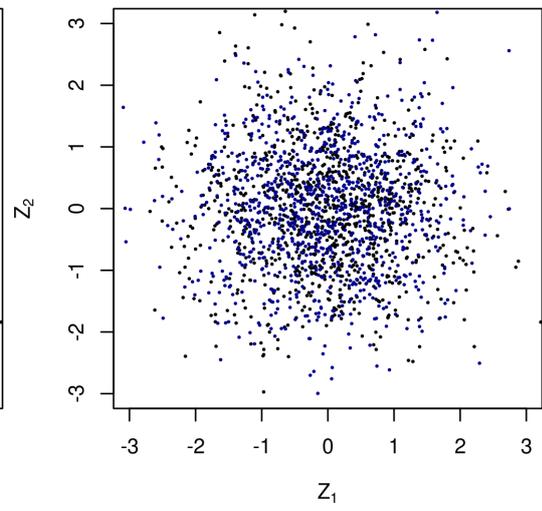
offenbar sind Z_1 und Z_2 unabhängig und identisch verteilt.

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

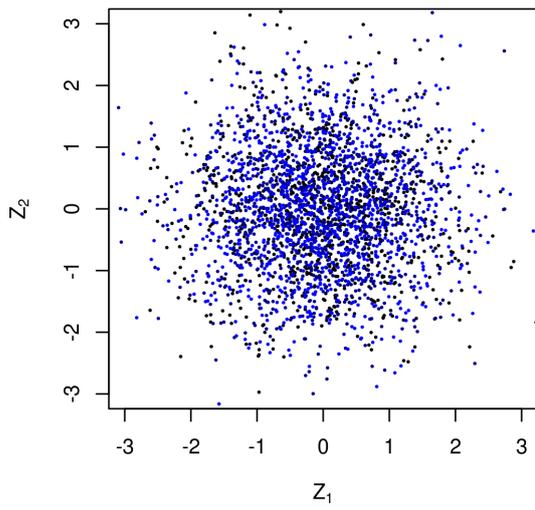
1000 simulierte Werte von (Z_1, Z_2)



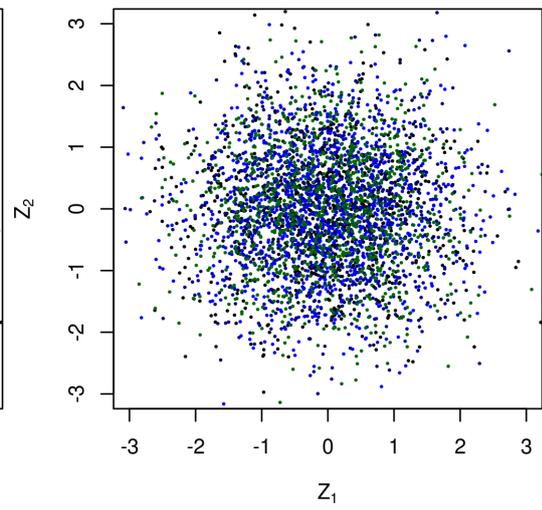
2000 simulierte Werte von (Z_1, Z_2)

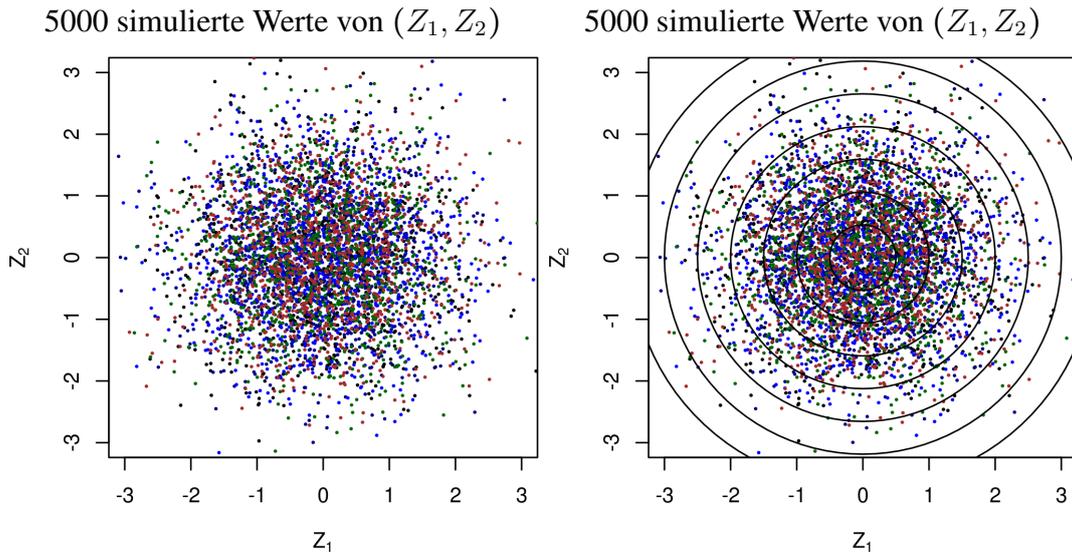


3000 simulierte Werte von (Z_1, Z_2)



4000 simulierte Werte von (Z_1, Z_2)





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

Beobachtung 1.94 (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.

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

für ein $\sigma^2 \in (0, \infty)$.

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

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

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

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

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

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

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

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

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

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

also gilt

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

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. \square

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

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

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

Bericht 1.95 (Hoeffding-Ungleichung(en)¹⁵). Unter obigen Voraussetzungen gilt für $t \geq 0$

$$P(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \tag{1.23}$$

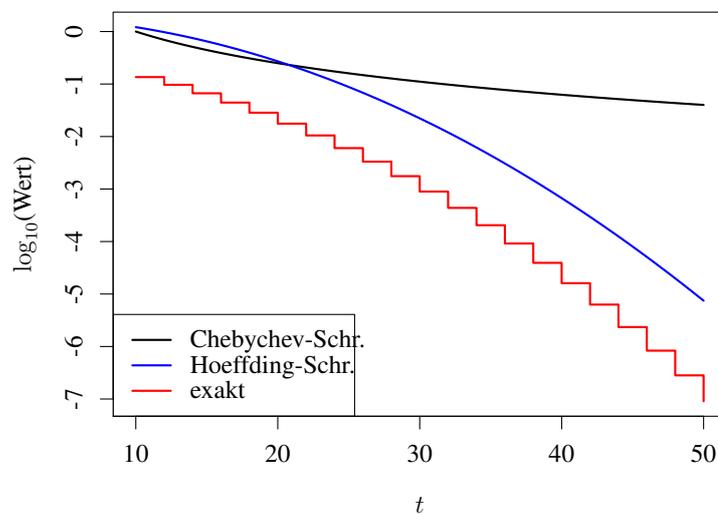
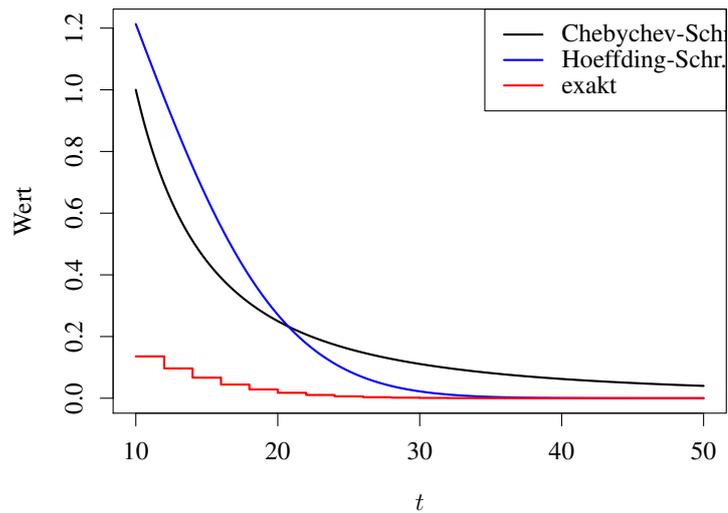
$$P(S_n - \mathbb{E}[S_n] \leq -t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \tag{1.24}$$

insbesondere

$$P(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{1.25}$$

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

¹⁵nach Wassily Hoeffding, 1914–1991

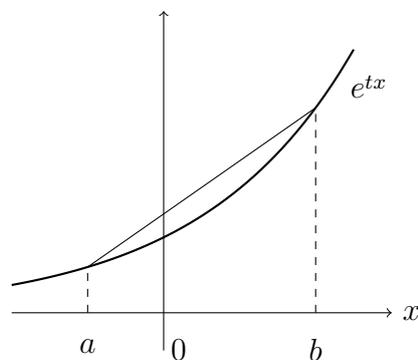


Zum Beweis der Aussagen in Bericht 1.95 verwendet man folgendes Lemma:

Lemma 1.96 (Hoeffdings Lemma). $a < 0 < b$, X ZV mit Werten in $[a, b]$ und $\mathbb{E}[X] = 0$, dann gilt für $t \in \mathbb{R}$

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

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



daher gilt für jedes $x \in [a, b]$

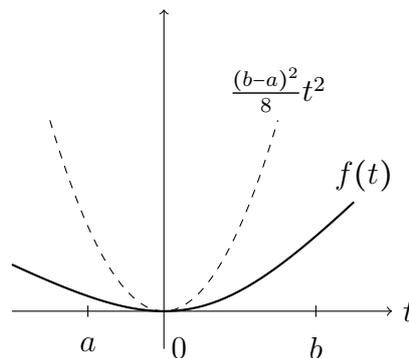
$$e^{tx} \leq \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}$$

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

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

Die Funktion

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



erfüllt

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

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

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

denn

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

beachte: $-a > 0$).

Also ist

$$\begin{aligned} f(t) &= f(0) + \int_0^t f'(s) ds = \int_0^t \left(f'(0) + \int_0^s f''(u) du \right) ds \\ &\leq \frac{(b-a)^2}{4} \int_0^t \int_0^s du ds = \frac{(b-a)^2}{8} t^2 \end{aligned}$$

und somit

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

wie behauptet. □

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

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

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 , $f : S^n \rightarrow \mathbb{R}$. 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

$$\left|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)\right| \leq c_i. \quad (1.26)$$

Dann gilt für $t \geq 0$

$$P\left(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \quad (1.27)$$

und

$$P\left(\left|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]\right| \geq t\right) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \quad (1.28)$$

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, \dots, X_n u.i.v. $\sim \text{Ber}_p, p \in (0, 1)$ (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, \dots, X_n (z.B. enthält $(0, 0, 1, 1, 0, 1, 0)$ 4 Wechsel). Beachte: Die Summanden in W sind nicht unabhängig (wir können also nicht die Hoeffding-Ungleichung verwenden).

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

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

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

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

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

Kapitel 2

Ideen aus der Statistik

2.1 Zur deskriptiven Statistik

In diesem Kapitel der Vorlesung werden Ideen und Verfahren der beschreibenden Statistik kurz diskutiert, vergleiche die Folien `Folien_deskriptive_Statistik.pdf` auf der Homepage der Vorlesung.

2.2 Schätzer und Tests für (Populations-)Mittelwerte

Eine „Standardsituation“ in der Statistik: Wir haben n Beobachtungswerte x_1, \dots, x_n (z.B. die Carapaxlängen der gefangenen Springkrebse aus dem Einstiegsbeispiel zur deskriptiven Statistik) und möchten anhand dessen etwas über die „Gesamtpopulation“, aus der die Daten stammen, aussagen.

Die grundlegende Idee: Die Daten werden als Realisierungen von Zufallsvariablen aufgefasst, die in einem stochastischen Modell spezifiziert werden. Man versucht dann, anhand der Daten Rückschlüsse auf Parameter des Modells zu ziehen, und so systematische Effekte von Zufälligem zu trennen.

Wir formulieren dies folgendermaßen („abstrakt“):

Definition 2.1. Ein *statistisches Modell* ist ein Tripel $(\Omega, \mathcal{F}, (P_\vartheta)_{\vartheta \in \Theta})$, wobei Ω der „Stichprobenraum“, $\mathcal{F} \subset 2^\Omega$ eine σ -Algebra über Ω , Θ eine (möglicherweise recht allgemeine) Indexmenge und jedes P_ϑ ein Wahrscheinlichkeitsmaß (auf (Ω, \mathcal{F})).

(Vergleiche auch Def. 1.1.) Die verschiedenen P_ϑ beschreiben unterschiedliche Arten, „wie der Zufall wirken kann“ (d.h. unterschiedliche Verteilungen der Zufallsvariablen auf Ω). Im Kontext des Modells nehmen wir an, dass die Daten gemäß einem gewissen P_ϑ generiert wurden, wir kennen aber das „wahre“ ϑ nicht. Die Aufgabe der (schließenden) Statistik besteht grob gesprochen darin, anhand der Beobachtungen Informationen über das „zugrundeliegende“ ϑ zu gewinnen.

2.2.1 Schätzer und Konfidenzintervall

Punktschätzer (für die Lageschätzung)

Modell: X_1, \dots, X_n unabhängige ZVn mit derselben Verteilung ϑ (die wir möglicherweise nicht kennen), mit Mittelwert $\mu (= \mu(\vartheta))$ und Varianz $\sigma^2 (= \sigma^2(\vartheta))$

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i$$

\bar{X} ist ein *Schätzer* für μ .

\bar{X} ist *erwartungstreu*:

$$\mathbb{E}_{\vartheta}[\bar{X}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\vartheta}[X_i] = \mathbb{E}_{\vartheta}[X_1] = \mu = \mu(\vartheta) \quad \text{unter } P_{\vartheta},$$

egal, was der tatsächliche Wert von μ ist

Seien X_1, \dots, X_n, \dots unabhängige ZVn mit derselben Verteilung, mit Mittelwert $\mu (= \mu(\vartheta))$ und Varianz $\sigma^2 (= \sigma^2(\vartheta))$

\bar{X} ist *konsistent*:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{\text{stoch.}} \mu = \mu(\vartheta) \quad \text{unter } P_{\vartheta}$$

(gemäß dem Gesetz der großen Zahlen, Satz 1.88)

Die allgemeine Situation:

Definition 2.2. Wir interessieren uns für eine gewisse (numerische) Eigenschaft $h(\vartheta)$ der Verteilung ϑ der X_i , eine Funktion Y der X_1, \dots, X_n ist ein erwartungstreuer Schätzer für $h(\vartheta)$, wenn stets gilt

$$\mathbb{E}_{\vartheta}[Y] = h(\vartheta) \quad \text{für } \vartheta \in \Theta$$

Demnach ist

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i$$

ein „Punktschätzer“ für $\mu = \mathbb{E}_{\vartheta}[X_i]$ mit „guten“ Eigenschaften.

2.2.2 Konfidenzintervall, bekannte Varianz

Soweit, so schön, aber: wie genau ist die Schätzung?

$$\text{Es ist } \text{Var}_{\vartheta}[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}_{\vartheta}[X_i] = \frac{\sigma^2}{n}.$$

Wir machen zunächst „unser Leben leicht“: Nehmen wir an, dass wir σ kennen, und dass die Daten aus einer Normalverteilung stammen (im Modell: X_i u.i.v., $\sim \mathcal{N}_{\mu, \sigma^2}$).

Konfidenzintervall für den Mittelwert im normalen Modell bei bekannter Varianz

Beobachtung 2.3. X_1, X_2, \dots, X_n u.i.v. $\sim \vartheta = \mathcal{N}_{\mu, \sigma^2}$ mit (uns unbekanntem) $\mu \in \mathbb{R}$, $\sigma^2 > 0$ sei bekannt (und fest). Sei $\alpha \in (0, 1)$ [z.B. $\alpha = 0.05$ oder $\alpha = 0.01$], $q := \Phi^{-1}(1 - \frac{\alpha}{2})$ das $(1 - \alpha/2)$ -Quantil der Standardnormalverteilung [mit R: `qnorm(1 - alpha/2)`].

Das (anhand der Daten konstruierte) Intervall

$$I := \left[\bar{X} - q \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + q \cdot \frac{\sigma}{\sqrt{n}} \right]$$

hat die Eigenschaft

$$P_{\vartheta}(\mu \in I) = 1 - \alpha$$

(egal, was μ ist), denn

$$\begin{aligned} \bar{X} - q \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + q \cdot \frac{\sigma}{\sqrt{n}} \\ \iff \\ -q \leq \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \leq q \end{aligned}$$

und (egal was μ und σ^2 sind) stets ist unter P_{ϑ}

$$\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}_{0,1}$$

Definition 2.4. Allgemein heißt ein (anhand der Beobachtungswerte zu konstruierendes) Intervall I ein *Konfidenzintervall* (manchmal auch „Vertrauensintervall“) für μ zum (Sicherheits-)Niveau $1 - \alpha$ (bzw. Irrtumsniveau α), wenn gilt

$$\forall \vartheta \in \Theta : P_{\vartheta}(\mu \in I) \geq 1 - \alpha.$$

(mit Notation $\mu = \mu(\vartheta)$).

Beachte: I ist zufällig, nicht aber ϑ (zumindest in unserer (sogenannten frequentistischen) Interpretation).

Offenbar möchte man i.A. I so kurz wie möglich wählen (soweit verträglich mit dem geforderten Niveau).

Bemerkung. Auch wenn die X_i nicht wörtlich normalverteilt sind (sondern unabhängig und identisch verteilt mit einer gewissen Verteilung ϑ , die Mittelwert μ und Varianz σ^2 hat), so ist

$$\sqrt{n} \frac{\bar{X} - \mu}{\sigma} = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \stackrel{d}{\approx} \mathcal{N}_{0,1}$$

gemäß dem zentralen Grenzwertsatz.

Daher ist auch für allgemeine Verteilung das auf Normalität fußende Konfidenzintervall zumindest für große n „approximativ korrekt“ (man muss aber σ kennen, um es zu konstruieren).

2.2.3 Konfidenzintervall, unbekannte Varianz

Wenn man die Varianz nicht kennt, kann man sie immerhin schätzen:

$$S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

[im Sinne der deskriptiven Statistik wäre dies die „korrigierte Stichprobenvarianz“] ist erwartungstreuer Schätzer für σ^2 :

$$\begin{aligned} \mathbb{E}_\vartheta \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] &= n \mathbb{E}_\vartheta [(X_i - \bar{X})^2] = n \text{Var}_\vartheta [X_i - \bar{X}] \\ &= n \text{Var}_\vartheta \left[\frac{n-1}{n} X_1 - \frac{1}{n} \sum_{i=2}^n X_i \right] \\ &= n \left(\left(\frac{n-1}{n} \right)^2 \text{Var}_\vartheta [X_1] + \frac{n-1}{n^2} \text{Var}_\vartheta [X_1] \right) = (n-1) \text{Var}_\vartheta [X_1] = (n-1) \sigma^2 \end{aligned}$$

S^2 ist konsistent (als Folge in n betrachtet):

$$\begin{aligned} \frac{n-1}{n} S^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) - 2 \left(\frac{1}{n} \sum_{i=1}^n X_i \bar{X} \right) + (\bar{X})^2 \\ &= \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) - (\bar{X})^2 \xrightarrow[n \rightarrow \infty]{\text{stoch.}} \mathbb{E}_\vartheta [X_1^2] - (\mathbb{E}_\vartheta [X_1])^2 = \sigma^2 \end{aligned}$$

(verwende jeweils das Gesetz der großen Zahlen).

Wir wissen bereits: Die Standardabweichung von \bar{X} ist $\frac{\sigma}{\sqrt{n}}$. Somit ist

$$\frac{S}{\sqrt{n}}$$

ein (naheliegender) Schätzer für die (unbekannte) Standardabweichung von \bar{X} , man nennt es auch den *Standardfehler*.

Satz 2.5 (Student (=William Gosset), 1908). X_1, \dots, X_n u.i.v. $\sim \mathcal{N}_{\mu, \sigma^2}$ mit $\mu \in \mathbb{R}$, $\sigma > 0$,

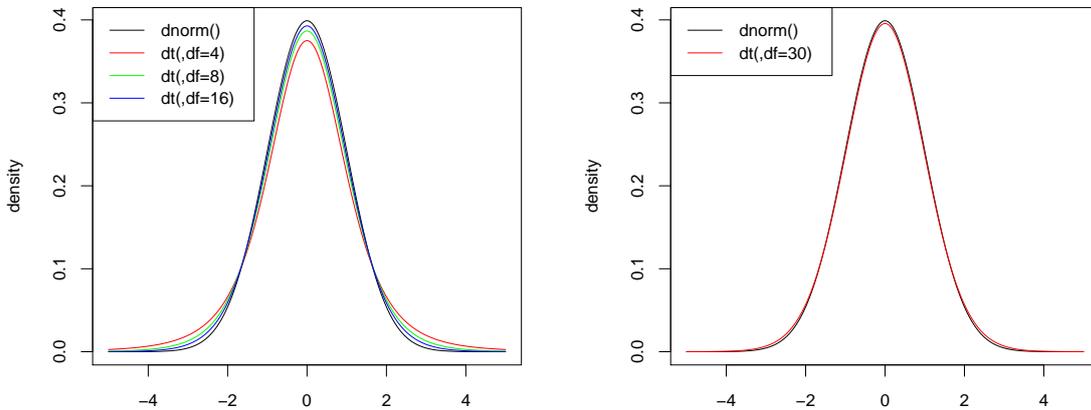
$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Dann besitzt $T := \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S^2}}$ die Dichte $t_{n-1}(x)$, wobei

$$t_m(x) = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{m}{2})} \frac{1}{\sqrt{m}} \left(1 + \frac{x^2}{m} \right)^{-\frac{m+1}{2}}$$

t_m ist die Dichte der Student- t -Verteilung mit m Freiheitsgraden (manchmal auch Student-Verteilung oder t -Verteilung genannt).

Dichte der t -Verteilungen



Bemerkung. Die t -Verteilung wurde von William Gosset erforscht und von ihm 1908 veröffentlicht, während er in einer Guinness-Brauerei arbeitete. Da sein Arbeitgeber die Veröffentlichung nicht gestattete, veröffentlichte Gosset sie unter dem Pseudonym *Student*.

Beweise von Satz 2.5 finden sich in der Literatur, siehe z.B. Kersting & Wakolbinger [KW], Kap. 22 oder Georgii [Ge], Satz 9.17. (Es gibt dazu ein schönes geometrisches Argument, das die Rotationsinvarianz der multivariaten Normalverteilung ausnutzt.)

Zum Begriff der Freiheitsgrade zunächst ein Beispiel: Es gibt 5 Freiheitsgrade im Vektor

$$x = (x_1, x_2, x_3, x_4, x_5)$$

da 5 Werte frei wählbar sind. Der Vektor

$$v := x - \bar{x} := (x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, x_4 - \bar{x}, x_5 - \bar{x})$$

hat 4 Freiheitsgrade, denn nach Wahl von v_1, v_2, v_3, v_4 ist v_5 festgelegt wegen $v_1 + \dots + v_4 + v_5 = 0$.

Die Bezeichnung „Student-Verteilung mit $n - 1$ Freiheitsgraden“ ist motiviert durch die Tatsache, dass $t = (\bar{x} - \mu)/(s/\sqrt{n})$, wo $s = \sqrt{s^2} = \frac{1}{\sqrt{n-1}} \left((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right)^{1/2}$ (bis auf Normierung) die Länge eines „Vektors mit $n - 1$ Freiheitsgraden“ ist.

Student-Konfidenzintervall für den Erwartungswert im normalen Modell

Korollar 2.6. Unter P_ϑ , $\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ seien

$$X_1, X_2, \dots, X_n \text{ u.i.v. } \sim \mathcal{N}_{\mu, \sigma^2}.$$

Sei $\alpha \in (0, 1)$, $q = q_{n-1, 1-\alpha/2}$ das $1 - \frac{\alpha}{2}$ -Quantil der Student-Verteilung mit $n - 1$ Freiheitsgraden [mit R : $\text{qt}(1 - \alpha/2, \text{df}=n - 1)$],

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Dann ist

$$I := \left[\bar{X} - q\sqrt{\frac{S^2}{n}}, \bar{X} + q\sqrt{\frac{S^2}{n}} \right]$$

ein Konfidenzintervall für μ zum Irrtumsniveau α .

Beweis.

$$I := \left[\bar{X} - q\sqrt{\frac{S^2}{n}}, \bar{X} + q\sqrt{\frac{S^2}{n}} \right]$$

ist ein Konfidenzintervall für μ zum Irrtumsniveau α , denn $T := \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{S^2}}$ ist Student-verteilt mit $n - 1$ Freiheitsgraden (für jede Wahl von μ und σ^2), also

$$\begin{aligned} P_{(\mu, \sigma^2)} \left(\bar{X} - q\sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{X} + q\sqrt{\frac{S^2}{n}} \right) \\ = P(-q \leq T \leq q) = P(T \leq q) - P(T \leq -q) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha. \end{aligned}$$

□

Beispiel. 10 Patienten erhielten in aufeinanderfolgenden Nächten Schlafmittel A, dann Mittel B.

Die Daten¹ ($x_i = \text{Anz. Stunden Schlaf mit Mittel A} - \text{Anz. Stunden Schlaf mit Mittel B}$ bei Patient Nr. i):

i	1	2	3	4	5	6	7	8	9	10
x_i	1,2	2,4	1,3	1,3	0,0	1,0	1,8	0,8	4,6	1,4

Es ist

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i \approx 1,58, \quad s = \left(\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2 \right)^{1/2} \approx 1,23.$$

Es ist $q_{9, 0,995} \approx 3,25$ (aus einer Quantiltabelle oder beispielsweise mit R berechnet), demnach ist

$$\left[\bar{x} \pm q \frac{s}{\sqrt{n}} \right] \approx [0,31, 2,85]$$

ein Konfidenzintervall für μ (die mittlere zusätzliche Anzahl Stunden Schlaf, die Medikament A mehr bringt als Medikament B) zum Sicherheitsniveau $0,99 = 1 - 0,01$.

Beachte: (Sinnlos) genaue Werte mit Rechnergenauigkeit sind $\bar{x} - q \frac{s}{\sqrt{n}} \approx 0,3159481$, $\bar{x} + q \frac{s}{\sqrt{n}} \approx 2,8440519$, man sollte allerdings die Grenzen eines Konfidenzintervalls stets „konservativ“, d.h. nach außen, runden.

2.2.4 Statistische Tests

Abstrakter Rahmen

Die Fragestellung statistischer Tests: Sei $(\Omega, \mathcal{F}, (P_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell, $\Theta = \Theta_0 \dot{\cup} \Theta_1$ disjunkte Zerlegung in „Nullhypothese“ und „Alternative“ (auch „Gegenhypothese“).

¹Aus Student (William S. Gosset), The Probable Error of a Mean, Biometrika 6:1–25 (1908)

Wir suchen ein Verfahren, um anhand von Beobachtungswerten x_1, \dots, x_n zu entscheiden, ob wir

$$H_0 : \vartheta \in \Theta_0 \quad (\text{die Nullhypothese}) \text{ oder}$$

$$H_1 : \vartheta \in \Theta_1 \quad (\text{die Alternative})$$

für plausibler halten (und zugleich eine geeignete Art, „plausibler“ zu quantifizieren).

Abstrakt ist ein statistischer Test von Θ_0 gegen Θ_1 eine Funktion $\varphi(x_1, \dots, x_n)$ der Beobachtungen mit Werten $\{0, 1\}$ und der Entscheidungsregel:

$$\varphi(x_1, \dots, x_n) = 0 \text{ dann } \begin{array}{l} \text{entscheide für } H_0 \\ \text{ („behalte die Nullhypothese bei“)} \end{array}$$

$$\varphi(x_1, \dots, x_n) = 1 \text{ dann } \begin{array}{l} \text{entscheide für } H_1 \\ \text{ („verwirf die Nullhypothese zugunsten der Alternative“)} \end{array}$$

Sei $\alpha \in (0, 1)$. Der Test φ hat *Niveau* α , wenn gilt

$$\sup_{\vartheta \in \Theta_0} P_{\vartheta}(\varphi(X_1, \dots, X_n) = 1) \leq \alpha$$

(d.h. die W'keit für einen „Fehler 1. Art“ (die Nullhypothese fälschlicherweise zu verwerfen) ist $\leq \alpha$).

Zugleich ist wünschenswert:

$$P_{\vartheta}(\varphi(X_1, \dots, X_n) = 0) \stackrel{!}{=} \text{möglichst klein} \quad \text{für } \vartheta \in \Theta_1,$$

die W'keit für einen „Fehler 2. Art“ (die Nullhypothese fälschlicherweise zu akzeptieren) sollte klein sein. Dann hat φ große „Schärfe“ oder „Macht“.

Statistische Tests: Niveau vs. p -Wert In der Praxis haben Tests meist die folgende Form:

Berechne eine gewisse (Test-)Statistik $Y = Y(x_1, \dots, x_n)$, setze

$$\varphi(x_1, \dots, x_n) = \mathbf{1}(Y(x_1, \dots, x_n) > q)$$

für einen gewissen Wert $q = q(\alpha)$, der in Abhängigkeit von den Parametern des Tests gewählt wird.

Sei $y = Y(x_1, \dots, x_n)$ beobachtet worden. Dann nennt man (speziell wenn $\Theta_0 = \{\vartheta_0\}$)

$$P_{\vartheta_0}(Y(X_1, \dots, X_n) > y)$$

den p -Wert des Test(ergebnisses). Dies ist die W'keit, bei Gültigkeit der Nullhypothese einen mindestens so „extremen“ Wert der Teststatistik zu finden wie den tatsächlich anhand der Daten beobachteten. (Es gilt: Test lehnt H_0 zum Niveau α ab g.d.w. p -Wert $\leq \alpha$.)

2.2.5 Ein-Stichproben- oder gepaarter t -Test

Modell: n u.i.v. Beobachtungen X_1, \dots, X_n , $X_i \sim \mathcal{N}_{\mu, \sigma^2}$, $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ unbekannt.

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Wir wissen: Für jedes $\mu_0 \in \mathbb{R}$, $\sigma^2 > 0$ ist unter $P_{(\mu_0, \sigma^2)}$

$$T := \sqrt{n} \frac{\bar{X} - \mu_0}{\sqrt{S^2}} \sim \text{Student}-(n-1)$$

Signifikanzniveau $\alpha \in (0, 1)$:

Zweiseitiger Test von $H_0 : \{\mu = \mu_0\}$ gegen $H_1 : \{\mu \neq \mu_0\}$ zum Niveau α :

Lehne H_0 ab, wenn $|T| > q_{n-1, 1-\alpha/2}$, wobei $q_{n-1, 1-\alpha/2} = (1 - \alpha/2)$ -Quantil der Student- $(n-1)$ -Vert.

Einseitiger Test von $H_0 : \{\mu \leq \mu_0\}$ gegen $H_1 : \{\mu > \mu_0\}$:

Lehne H_0 ab, wenn $T > q_{n-1, 1-\alpha}$, wobei $q_{n-1, 1-\alpha} = (1 - \alpha)$ -Quantil der Student- $(n-1)$ -Vert.

Beispiel. Die Wirksamkeit eines gewissen Schlafmittels soll geprüft werden. 10 Patienten erhalten das Schlafmittel, die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet. Wir nehmen an, die Beob. sind u.i.v. $\sim \mathcal{N}_{\mu, \sigma^2}$ und wir möchten die Nullhypothese $\mu = 0$, sagen wir, zum Niveau $\alpha = 0.05$ testen.

Die Daten²:

Patient i	1	2	3	4	5	6	7	8	9	10
zus. Schl.	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0

Es ist $n = 10$, $\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 0.75$, $s^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2 \approx 1.79$, $t = \frac{\bar{x} - 0}{s/\sqrt{10}} \approx 1.326$. Das 0.975-Quantil der Student-9-Verteilung ist ≈ 2.262 , demnach können wir die Nullhypothese nicht ablehnen. (Für ein Student-9-verteiltes T ist $P(|T| \geq 1.326) \approx 0.2176$, dies ist der p -Wert des Tests.)

Man kann unseren Befund folgendermaßen formulieren: „Die Beobachtungen sind mit der Nullhypothese $\mu = 0$ (im statistischen Sinne) verträglich.“ oder „Die beobachtete Abweichung $\bar{x} = 0.75$ ist nicht signifikant von 0 verschieden (t -Test, $\alpha = 0.05$).“

Das Beispiel in R:

```
> schlaf <- c(0.7, -1.6, -0.2, -1.2, -0.1, 3.4,
             3.7, 0.8, 0.0, 2.0)
> t.test(schlaf)
```

One Sample t-test

```
data: schlaf
t = 1.3257, df = 9, p-value = 0.2176
alternative hypothesis: true mean is not equal to 0
```

²Aus Student (William S. Gosset), The Probable Error of a Mean, Biometrika 6:1–25 (1908)

95 percent confidence interval:

-0.5297804 2.0297804

sample estimates:

mean of x

0.75

Beispiel. Die Wirksamkeit eines Schlafmittels soll mit der eines anderen verglichen werden. 10 Patienten erhalten Schlafmittel *A*, die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet. Dann erhalten dieselben 10 Patienten Schlafmittel *B*, wieder wird die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet.

Da dieselben Patienten untersucht werden, können (und sollten) wir die Messungen paaren: Wir interessieren uns bei jedem Patienten für die Differenz des (zusätzlichen) Schlafs bei Mittel *B* und bei Mittel *A*.

Wir nehmen an, die beobachteten Differenzen sind Realisierungen von u.i.v. ZVn mit Vert. $\mathcal{N}_{\mu, \sigma^2}$ und wir möchten die Nullhypothese $\mu \leq 0$ gegen die Alternative $\mu > 0$, sagen wir, zum Niveau $\alpha = 0.05$ testen.

Dies wäre beispielsweise in folgender Situation angemessen: Wir möchten darlegen, dass Mittel *B* wirksamer ist als Mittel *A*, indem wir die Nullhypothese „ $\mu \leq 0$ “ entkräften.

Die Daten³:

Patient <i>i</i>	1	2	3	4	5	6	7	8	9	10
Mittel <i>A</i>	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
Mittel <i>B</i>	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4
Diff.	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

Patient <i>i</i>	1	2	3	4	5	6	7	8	9	10
Diff.	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

Es ist $n = 10$, $\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 1.58$, $s^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2 \approx 1.23$, $t = \frac{\bar{x} - 0}{s/\sqrt{10}} \approx 4.062$

Das 0.95-Quantil der Student-9-Verteilung ist ≈ 1.833 , demnach können wir die Nullhypothese ablehnen. (Für ein Student-9-verteiltetes T ist $P(T > 4.062) \approx 0.0014$, dies ist der p -Wert des Tests.)

Mögliche knappe Formulierung: „Die beobachtete Differenz $\bar{x} = 1.58$ ist signifikant größer als 0 (einseitiger t -Test, $\alpha = 0.05$).“

Das Beispiel in R:

```
> diff <- c(1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8,
            0.8, 4.6, 1.4)
> t.test(diff, alternative="greater")
```

One Sample t-test

```
data: diff
t = 4.0621, df = 9, p-value = 0.001416
alternative hypothesis: true mean is greater than 0
```

³Ebenfalls aus Student (William S. Gosset), The Probable Error of a Mean, Biometrika 6:1–25 (1908)

95 percent confidence interval:

0.8669947 Inf

sample estimates:

mean of x

1.58

2.2.6 Zwei-Stichproben- t -Test oder ungepaarter t -Test mit Annahme gleicher Varianzen

Modell: m u.i.v. Beobachtungen X_1, \dots, X_m und davon unabhängig n u.i.v. Beobachtungen Y_1, \dots, Y_n , unter P_ϑ

$X_i \sim \mathcal{N}_{\mu_1, \sigma^2}, Y_j \sim \mathcal{N}_{\mu_2, \sigma^2}$ mit $\vartheta = (\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$

Seien

$$\bar{X} := \frac{1}{m} \sum_{i=1}^m X_i, \quad \bar{Y} := \frac{1}{n} \sum_{j=1}^n Y_j$$

die jeweiligen Stichprobenmittelwerte,

$$S_X^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y})^2,$$

die (korrigierten) Stichprobenvarianzen.

$$S^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \left(= \frac{1}{m+n-2} \left(\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2 \right) \right),$$

(bem.: $\mathbb{E}_{(\mu_1, \mu_2, \sigma^2)}[S^2] = \sigma^2$)

$$T = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Für $\mu_0 \in \mathbb{R}, \sigma^2 > 0$ ist unter $P_{(\mu_0, \mu_0, \sigma^2)}$

$$T = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \text{Student-}(m+n-2)$$

Signifikanzniveau $\alpha \in (0, 1)$:

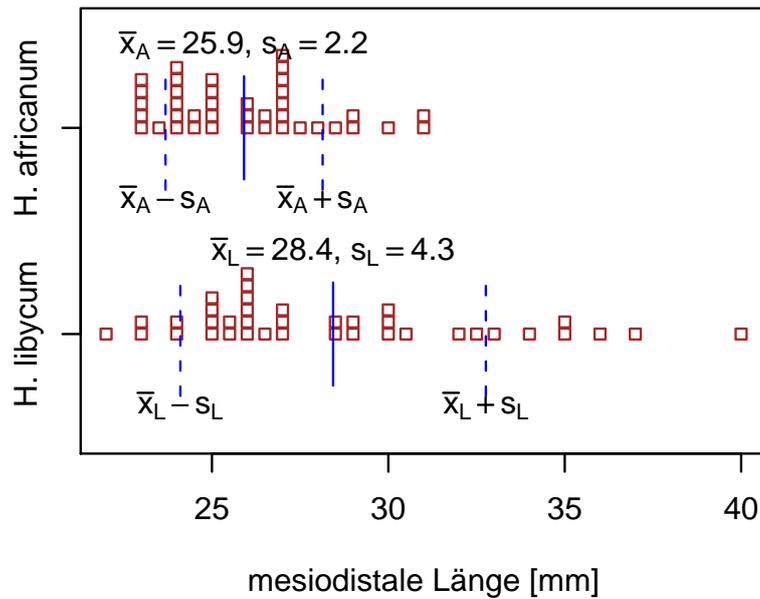
Zweiseitiger Test von $H_0 : \{\mu_1 = \mu_2\}$ gegen $H_1 : \{\mu_1 \neq \mu_2\}$:

Lehne H_0 ab, wenn $|T| > q_{m+n-2, 1-\alpha/2}$, wobei $q_{m+n-2, 1-\alpha/2} = (1 - \alpha/2)$ -Quantil der Student- $(m+n-2)$ -Vert.

Einseitiger Test von $H_0 : \{\mu_1 \leq \mu_2\}$ gegen $H_1 : \{\mu_1 > \mu_2\}$:

Lehne H_0 ab, wenn $T > q_{m+n-2, 1-\alpha}$, wobei $q_{m+n-2, 1-\alpha} = (1 - \alpha)$ -Quantil der Student- $(m+n-2)$ -Vert.

Beispiel. Es wurden fossile Backenzähne gefunden, die zwei Arten von Urpferden zugeordnet wurden, und jeweils die („mesiodistale“) Länge bestimmt. Wir möchten die (Null-)Hypothese prüfen, ob die mittlere Zahnlänge bei den beiden Arten gleich ist.



Die Daten:	Hipparion africanum	Hipparion lybicum
	$n_A = 39$	$n_L = 38$
	$\bar{x}_A = 25.9$	$\bar{x}_L = 28.4$
	$s_A = 2.2$	$s_L = 4.3$

Wir verwenden Signifikanzniveau $\alpha = 0.01$, das 99,5%-Quantil der Student-Vert. mit 75 Freiheitsgraden ist ≈ 2.64 .

Es ist

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_L - 1)s_L^2}{n_A + n_L - 1} \approx 11.94, \quad t = \frac{\bar{x}_A - \bar{x}_L}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_L}}} \approx -3.229,$$

Wir können die Nullhypothese „die mittlere mesiodistale Länge bei *H. lybicum* und bei *H. africanum* sind gleich“ zum Signifikanzniveau 1% ablehnen.

(Für ein Student-75-verteiltes T ist $P(|T| > 3.229) \approx 0.0018$, dies ist der p -Wert des Tests.)

Mögliche Formulierung unseres Befunds: „Die mittlere mesiodistale Länge war signifikant größer (28.4 mm) bei *H. lybicum* als bei *H. africanum* (25.9 mm) (t -Test, $\alpha = 0,01$).“

Das Beispiel in R:

```
> t.test(md[Art=="africanum"], md[Art=="lybicum"],
         var.equal=TRUE)
```

Two Sample t-test

```
data: md[Art == "africanum"] and md[Art == "lybicum"]
t = -3.2289, df = 75, p-value = 0.001845
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
-4.0811448 -0.9667634
sample estimates:
mean of x mean of y
25.91026 28.43421
```

Betrachten wir (spaßeshalber) nochmal die Schlafmittel-Daten, diesmal ungepaart:

```
> medA <- c(0.7, -1.6, -0.2, -1.2, -0.1, 3.4, 3.7, 0.8, 0.0, 2.0)
> medB <- c(1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4)
> t.test(medB, medA, var.equal=TRUE)
```

Two Sample t-test

```
data: medB and medA
t = 1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.203874  3.363874
sample estimates:
mean of x mean of y
  2.33      0.75
```

2.2.7 Bericht: Welchs zwei Stichproben-*t*-Test (*t*-Statistik ohne Annahme gleicher Varianz)

Es gibt auch eine Version des zwei-Stichproben-*t*-Tests, der die Annahme gleicher Varianzen nicht trifft (wir werden ihn im Verlauf der Vorlesung allerdings nicht verwenden):

Wäre eine beobachtete Abweichung $\bar{x} - \bar{y}$ mit der Nullhypothese verträglich, dass $\mu_X = \mu_Y$?

Wir schätzen die Streuung von $\bar{X} - \bar{Y}$ durch

$$\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} \quad \text{und bilden} \quad T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}.$$

Unter $P_{(\mu_0, \mu_0, \sigma_1^2, \sigma_2^2)}$ ist T „approximativ Student-verteilt mit g Freiheitsgraden“, wobei g aus den Daten

$$\text{geschätzt wird, } g = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{s_X^4}{n_X^2(n_X-1)} + \frac{s_Y^4}{n_Y^2(n_Y-1)}}$$

Welchs⁴ zwei Stichproben-*t*-Test

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}, \quad g = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{s_X^4}{n_X^2(n_X-1)} + \frac{s_Y^4}{n_Y^2(n_Y-1)}}$$

verwirft die Nullhypothese „ $\mu_1 = \mu_2$ “ (zum Niveau α), wenn

$$\text{pt}(|t|, \text{df}=g, \text{lower.tail}=\text{FALSE}) \leq \alpha/2$$

ist, d.h. wenn die Wahrscheinlichkeit, dass eine Student-verteilte Zufallsgröße mit g Freiheitsgraden einen betragsmäßig mindestens so großen Wert wie den beobachteten *t*-Wert annimmt, $\leq \alpha$ ist (und analoges Vorgehen für einseitige Tests).

⁴B. L. Welch, The Significance of the Difference between Two Means When the Population Variances Are Unequal, *Biometrika* 29:350–362, (1938)

Zwei-Stichproben-t-Test mit R:

```
> A <- md[Art=="africanum"]
> L <- md[Art=="libycum"]
> t.test(L,A)
```

Welch Two Sample t-test

```
data: L and A
t = 3.2043, df = 54.975, p-value = 0.002255
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
 0.9453745 4.1025338
sample estimates:
mean of x mean of y
 28.43421 25.91026
```

2.3 Schätzprinzipien

Wir haben bisher (nur) den empirischen Mittelwert

$$\frac{1}{n} \sum_{i=1}^n x_i$$

als Schätzer für den Erwartungswert einer (unbekannten) Verteilung verwendet.

Das ist sehr naheliegend und diese Schätzer haben auch „gute“ Eigenschaften (Erwartungstreue, Konsistenz, asymptotische Normalität), wie wir in Kapitel 2.2 gesehen haben. Andererseits trifft man auch Situationen, in denen zumindest auf den ersten Blick kein „offensichtlicher“ Schätzer auf der Hand liegt; (spätestens) dann lohnen sich systematischere Ansätze, von denen wir einige in diesem Kapitel betrachten.

2.3.1 ML-Schätzer

Schätzer, allgemein

Wir betrachten eine Zufallsvariable X mit Wertebereich S in einem statistischen Modell $(\Omega, \mathcal{F}, (P_\vartheta)_{\vartheta \in \Theta})$. Wir verwenden X , um die Beobachtungen zu modellieren (oft schreiben wir $X = (X_1, X_2, \dots, X_n)$, wenn das Experiment aus n Wiederholungen oder „Bauteilen“ besteht).

Abstrakt gesehen ist ein Schätzer für ϑ eine Funktion $T : S \rightarrow \Theta$ (mit der Interpretation, dass wir anhand von Beobachtungen $x \in S$ im Rahmen des Modells „tippen“ würden, dass der Wert von ϑ wohl $T(x)$ ist).

Definition 2.7. Man nennt diese Situation ein *statistisches Standardmodell*, wenn entweder S diskret ist oder $S \subset \mathbb{R}^n$ gilt und X unter P_ϑ

$$\text{Gewichte } \rho_\vartheta(\cdot) \quad \text{bzw.} \quad \text{Dichte } \rho_\vartheta(\cdot) \quad \text{für } \vartheta \in \Theta$$

besitzt. Die Funktion

$$\begin{aligned} \rho : S \times \Theta &\rightarrow [0, \infty) \\ \downarrow \\ (x, \vartheta) &\mapsto \rho(x, \vartheta) := \rho_\vartheta(x) \end{aligned}$$

heißt die *Likelihood-Funktion* (manchmal auch „Plausibilitäts-Funktion“), für $x \in S$ heißt

$$L_x : \Theta \rightarrow [0, \infty), \quad L_x(\vartheta) = \rho(x, \vartheta)$$

die Likelihood-Funktion zum Beobachtungswert x .

ML-Schätzer

Definition 2.8. Ein Schätzer $T : S \rightarrow \Theta$ heißt (ein) *Maximum-Likelihood-Schätzer*, wenn

$$\rho(x, T(x)) = \max_{\vartheta \in \Theta} \rho(x, \vartheta) \quad \forall x \in S$$

(auch kurz ML-Schätzer genannt, engl. MLE = maximum likelihood estimator). Man schreibt oft auch $\widehat{\vartheta}_{\text{ML}}$ für den ML-Schätzer.

Beispiel 2.9. 1. („Rückfangmethode“, engl. “capture-recapture”) Ein Teich enthalte ϑ Fische (einer gewissen Art, $\vartheta \in \mathbb{N}$ ist der unbekannte Parameter), fange und markiere m , setze wieder aus. Wenn sich die markierten Fische gut verteilt haben, fange erneut n Fische. Nehmen wir an, wir beobachten unter den erneut gefangenen x markierte Fische.

Formalisierung als statistisches Modell: $S = \{0, 1, \dots, n\}$, $\Theta = \{(m \vee n), (m \vee n) + 1, (m \vee n) + 2, \dots\}$, unter P_ϑ ist die beobachtete Anzahl $X \sim \text{Hyp}_{m, \vartheta-m, n}$ (hypergeometrische Verteilung, s.a. Bsp. 1.15).

Die Likelihood-Funktion ist demnach

$$\rho(x, \vartheta) = \frac{\binom{m}{x} \binom{\vartheta-m}{n-x}}{\binom{\vartheta}{n}},$$

der ML-Schätzer ist

$$\widehat{\vartheta}_{\text{ML}} = \left\lfloor \frac{n}{x} \cdot m \right\rfloor,$$

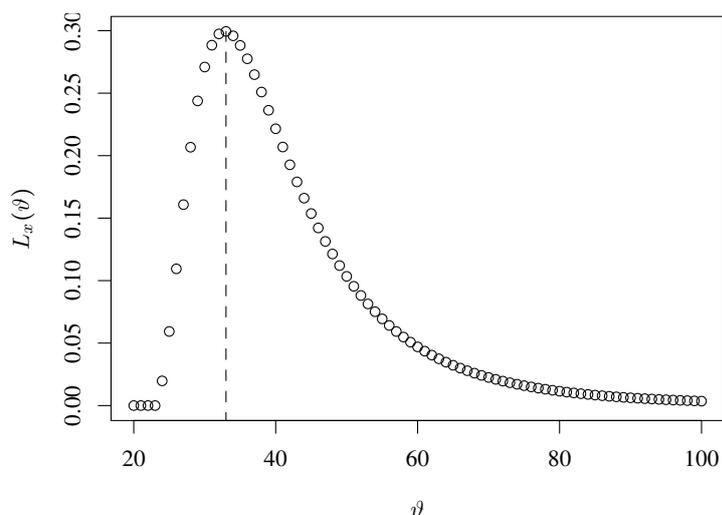
(mit $\lfloor a \rfloor = \max\{z \in \mathbb{Z} : z \leq a\}$),
es ist nämlich

$$\begin{aligned} \frac{\rho(x, \vartheta)}{\rho(x, \vartheta-1)} &= \frac{\binom{\vartheta-m}{n-x} \binom{\vartheta-1}{n}}{\binom{\vartheta}{n} \binom{\vartheta-1-m}{n-x}} \\ &= \frac{(\vartheta-m)(\vartheta-n)}{\vartheta(\vartheta-m-n+x)} = 1 - \frac{\vartheta x - mn}{\vartheta(\vartheta-m-n+x)} \begin{cases} > 1, & \vartheta < \frac{mn}{x}, \\ = 1, & \vartheta = \frac{mn}{x}, \\ < 1, & \vartheta > \frac{mn}{x} \end{cases} \end{aligned}$$

(beachte: stets ist $\vartheta - m \geq n - x$, es gibt im Teich mindestens so viele unmarkierte Fische wie in der Rückfang-Stichprobe).

Der ML-Schätzer ist hier auch anschaulich plausibel, denn $\frac{m}{\widehat{\vartheta}_{\text{ML}}} \approx \frac{x}{n}$.

Likelihood-Funktion $L_x(\vartheta)$, hier $m = 10, n = 20, x = 6$ ($\hat{\vartheta}_{ML} = 33$)



Bemerkung: Falls $\frac{mn}{x} \in \mathbb{N}$, so ist der ML-Schätzer hier nicht eindeutig: $\frac{mn}{x} - 1$ und $\frac{mn}{x}$ maximieren die Likelihood.

2. (Erfolgsw'keit im Binomialmodell per ML schätzen) Ein zufälliges Experiment mit zwei möglichen Ausgängen (Idealisierung: Werfen einer Münze) werde unabhängig (unter identischen Bedingungen) n -fach wiederholt, wir zählen die Anzahl X der „Erfolge“.

$S = \{0, 1, \dots, n\}$, unter $P_\vartheta, \vartheta \in \Theta = [0, 1]$ ist $X \sim \text{Bin}_{n,\vartheta}$, also

$$\rho(x, \vartheta) = L_x(\vartheta) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x}$$

$$\begin{aligned} \frac{d}{d\vartheta} \log L_x(\vartheta) &= \frac{d}{d\vartheta} \left(\log \binom{n}{x} + x \log \vartheta + (n-x) \log(1-\vartheta) \right) \\ &= \frac{x}{\vartheta} - \frac{n-x}{1-\vartheta} = 0 \iff \vartheta = \frac{x}{n} \end{aligned}$$

(Es ist $\frac{d}{d\vartheta} \log L_x(\vartheta) > 0$ für $\vartheta < x/n$ und $\frac{d}{d\vartheta} \log L_x(\vartheta) < 0$ für $\vartheta > x/n$, d.h. es handelt sich tatsächlich um ein Maximum; Inspektion zeigt, dass auch in den Randfällen $x = 0$ und $x = n$ $\hat{\vartheta}_{ML} = \frac{x}{n}$ gilt.)

Demnach ist $\hat{\vartheta}_{ML} = \frac{x}{n}$.

$\hat{\vartheta}_{ML}$ ist hier auch ein sehr „naheliegender“ Schätzer: Wir schätzen die (unbekannte) Erfolgswahrscheinlichkeit durch die relative Anzahl der beobachteten Erfolge.

3. (Normales Modell mit bekannter Varianz) n Beobachtungen seien u.i.v. $\sim \mathcal{N}_{\vartheta, \sigma^2}$, $\sigma^2 > 0$ sei bekannt und $\vartheta \in \Theta = \mathbb{R}$ soll geschätzt werden.

Mit $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ ist

$$\begin{aligned} \rho(x, \vartheta) = L_x(\vartheta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \vartheta)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \vartheta)^2\right) \end{aligned}$$

d.h.

$$L_x(\vartheta) \stackrel{!}{=} \max \iff \sum_{i=1}^n (x_i - \vartheta)^2 \stackrel{!}{=} \min.$$

$$L_x(\vartheta) \stackrel{!}{=} \max \iff \sum_{i=1}^n (x_i - \vartheta)^2 \stackrel{!}{=} \min.$$

Mit $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$ ist

$$\frac{1}{n} \sum_{i=1}^n (x_i - \vartheta)^2 = (\bar{x} - \vartheta)^2 + \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

d.h. es ist $\widehat{\vartheta}_{\text{ML}} = \bar{x}$, das empirische Mittel der Beobachtungen.

(Für die Gleichung beachte $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2$.)

Die ML-Schätzer $\widehat{\vartheta}_{\text{ML}} = \bar{x}$ stimmen hier mit dem naheliegenden Lageschätzer aus Kap. 2.2 überein, sie sind insbesondere konsistent und (hier sogar exakt) normalverteilt. (Dies gilt asymptotisch für $n \rightarrow \infty$ recht allgemein für ML-Schätzer [siehe Bericht 2.10 unten].)

4. (Normales Modell, unbekannter Erwartungswert und unbekannte Varianz) n Beobachtungen seien u.i.v. $\sim \mathcal{N}_{\mu,v}$ mit unbekanntem $\mu \in \mathbb{R}$ und $v \in (0, \infty)$.

(Formalisierung: $S = \mathbb{R}^n$, $\Theta = \{(\mu, v) : \mu \in \mathbb{R}, v > 0\}$, unter $P_{(\mu,v)}$ ist $X = (X_1, \dots, X_n) \sim \mathcal{N}_{\mu,v}^{\otimes n}$)

Wie in 3. ist

$$\log L_x((\mu, v)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log v - \frac{1}{2v} \sum_{i=1}^n (x_i - \mu)^2,$$

nach obigem ist $\widehat{\mu}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n x_i$ Maximierer bezüglich μ (für jeden Wert von v).

$$\log L_x((\mu, v)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log v - \frac{1}{2v} \sum_{i=1}^n (x_i - \mu)^2,$$

somit ist

$$\left. \frac{\partial}{\partial v} \log L_x((\mu, v)) \right|_{\mu=\widehat{\mu}_{\text{ML}}} = -\frac{n}{2v} + \frac{1}{2v^2} \sum_{i=1}^n (x_i - \widehat{\mu}_{\text{ML}})^2$$

also $\frac{\partial}{\partial v} \log L_x((\widehat{\mu}_{\text{ML}}, v)) = 0 \iff v = \widehat{v}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n (x_i - \widehat{\mu}_{\text{ML}})^2$.

(Und man prüft: $\log L_x((\widehat{\mu}_{\text{ML}}, v))$ ist wachsend für $v < \widehat{v}_{\text{ML}}$, fallend für $v > \widehat{v}_{\text{ML}}$.)

Beachte: Der ML-Schätzer für die unbekannte Varianz ist hier die (unkorrigierte) Stichprobenvarianz, also ist er nicht erwartungstreu (vgl. Kap. 2.2; allerdings ist er für halbwegs großes n auch nicht weit weg von Erwartungstreue).

5. n Beobachtungen seien u.i.v. uniform auf $\{1, 2, \dots, \vartheta\}$ (mit einem unbekanntem $\vartheta \in \mathbb{N}$).

Mögliche Interpretation: In einer Stadt gibt es ϑ Taxis, die mit $1, 2, \dots, \vartheta$ durchnummeriert sind. Wir beobachten die Nummern von n zufällig gewählten Taxis und schätzen, wieviele Taxis es insgesamt gibt.

Formalisierung: $X = (X_1, \dots, X_n)$ mit Werten in $S = \mathbb{N}^n$, $\rho((x_1, \dots, x_n), \vartheta) = \vartheta^{-n} \mathbf{1}_{x_1, \dots, x_n \leq \vartheta}$ für $\vartheta \in \Theta = \mathbb{N}$.

Es ist

$$\widehat{\vartheta}_{\text{ML}} = \max\{x_1, x_2, \dots, x_n\},$$

denn

$$L_{(x_1, \dots, x_n)}(\vartheta) = \begin{cases} \frac{1}{\vartheta^n}, & \text{falls } \vartheta \geq x_1, x_2, \dots, x_n, \\ 0, & \text{sonst.} \end{cases}$$

6. n Beobachtungen seien u.i.v. $\sim \text{Poi}_\vartheta$ (mit einem unbekanntem $\vartheta \in \Theta := (0, \infty)$), d.h.

$$\rho((x_1, \dots, x_n), \vartheta) = \prod_{i=1}^n e^{-\vartheta} \frac{\vartheta^{x_i}}{x_i!}$$

Für $x = (x_1, \dots, x_n)$ ist

$$L_x(\vartheta) = \frac{\exp(-n\vartheta + (x_1 + \dots + x_n) \log \vartheta)}{x_1! x_2! \dots x_n!}$$

also

$$\frac{\partial}{\partial \vartheta} \log L_x(\vartheta) = -n + \frac{x_1 + \dots + x_n}{\vartheta} \stackrel{!}{=} 0 \iff \vartheta = \widehat{\vartheta}_{\text{ML}}$$

mit $\widehat{\vartheta}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n x_i$, d.h. der ML-Schätzer ist hier wiederum der empirische Mittelwert. (Wegen $\mathbb{E}[X] = \vartheta$ für $X \sim \text{Poi}_\vartheta$ ist es auch zugleich der sogenannte „Momentenschätzer“.)

Betrachten wir Beispiel 1.21 nochmals in diesem Licht: Ladislaus von Bortkewitsch, *Das Gesetz der kleinen Zahlen*, Teubner, 1898, § 12, 4. („Die durch Schlag eines Pferdes im preußischen Heere getöteten“) berichtet 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“	$200 \times \text{Poi}_{0,61}(x)$
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
≥ 5	0	0,08

Sei $X_i =$ Anzahl der im i -ten Korpsjahr Getöteten $\sim \text{Poi}_\vartheta$, u.a. für verschiedene i .

Die Daten geben uns zwar nicht die Beobachtungen x_i selbst, enthalten aber genug Informationen, um

$$\widehat{\vartheta}_{\text{ML}} = \frac{1}{200} \sum_{i=1}^{200} x_i = \frac{109}{200} \cdot 0 + \frac{65}{200} \cdot 1 + \frac{22}{200} \cdot 2 + \frac{3}{200} \cdot 3 + \frac{1}{200} \cdot 4 + 0 = 0,61$$

zu berechnen.

Bericht 2.10 („gute“ Eigenschaften der ML-Schätzer). Betrachte ein Produktmodell, d.h. $X = (X_1, \dots, X_n)$ mit Werten in $S = S_1^n$ und

$$\rho((x_1, \dots, x_n), \vartheta) = \prod_{i=1}^n \rho^{(1)}(x_i, \vartheta)$$

für eine Likelihood-Funktion $\rho^{(1)}$ auf $S_1 \times \Theta$ (unter P_ϑ sind die Beobachtungen X_1, \dots, X_n u.i.v. mit Gewichten / Dichte $\rho^{(1)}(\cdot, \vartheta)$).

Sei $\widehat{\vartheta}_{ML,n} = \widehat{\vartheta}_{ML,n}(X_1, \dots, X_n)$ der ML-Schätzer basierend auf n unabhängigen Beobachtungen, dann gilt (unter gewissen Bedingungen an $\rho^{(1)}(\cdot, \cdot)$):

- Die ML-Schätzer sind konsistent, d.h.

$$P_\vartheta(|\widehat{\vartheta}_{ML,n} - \vartheta| \geq \varepsilon) \xrightarrow{n \rightarrow \infty} 0 \text{ für jedes } \varepsilon > 0 \text{ und jedes } \vartheta \in \Theta.$$

- Die ML-Schätzer sind asymptotisch normal mit (asymptotisch optimaler) Varianz $1/(nI(\vartheta))$, d.h.

$$\sqrt{n}(\widehat{\vartheta}_{ML,n} - \vartheta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}_{0,1/I(\vartheta)} \text{ unter } P_\vartheta \text{ für jedes } \vartheta \in \Theta,$$

wobei $I(\vartheta) := \text{Var}_\vartheta \left[\frac{d}{d\vartheta} \log \rho(X, \vartheta) \right]$ die sogenannte Fisher-Information ist.

(Bemerk.: Im Fall des ML-Schätzers für den Erwartungswert im normalen Modell haben wir dies schon gesehen; die Botschaft hier ist: es gilt ziemlich allgemein.)

2.3.2 Bayes-Statistik

Wir folgten bisher (und werden dies auch in den späteren Teilen wieder tun) dem klassischen, sogenannten frequentistischen Ansatz der Statistik:

Wir fassen eine Menge Θ von „Parametern“ ins Auge, für $\Theta \ni \vartheta$ beschreibt (in einem statistischen Modell) P_ϑ die Verteilung der Beobachtungen, wenn dieses ϑ der tatsächlich zutreffende (sozusagen der „wahre“) Parameter ist. In der konkreten Anwendungssituation kennen wir dieses „wahre“ ϑ natürlich i.A. nicht, wir fassen es als zwar unbekannt, aber prinzipiell feste Größe auf. Wahrscheinlichkeitsausagen beziehen sich *nicht* auf ϑ , sondern auf zufällige Beobachtungen unter P_ϑ .

Ansatz der Bayes-Statistik Dies ist anders in der *Bayes-Statistik*: Man wählt eine Wahrscheinlichkeitsverteilung α auf Θ , die *a priori*-Verteilung (auch *Vorbewertung*) und stellt sich vor, dass die Daten einem zweistufigen Experiment entstammen:

- Zunächst wird der Parameter ϑ gemäß der *a priori*-Verteilung α erzeugt (insbesondere ist ϑ jetzt selbst eine Zufallsvariable),
- dann werden die Beobachtungen X zufällig erzeugt mit einer Verteilung, die vom gewählten ϑ abhängt.

Insbesondere besitzt in dieser Formulierung das Paar (X, ϑ) eine *gemeinsame Verteilung*.

Es gelte: Die *a priori*-Verteilung auf Θ hat Dichte bzw. Gewichte $\alpha(\vartheta)$ (je nachdem, ob ϑ kontinuierlich oder diskret verteilt ist; wir betrachten im Folgenden nur den Fall, dass $\Theta \subset \mathbb{R}$ ein Intervall ist und ϑ eine Dichte besitzt).

Interpretation: Ohne Kenntnis der Beobachtungen nehmen wir an, dass ϑ die *a priori*-Verteilung besitzt (z.B. aus „Erfahrung“ oder aus „Expertenwissen“). Wir interpretieren die Likelihood-Funktion $\rho : S \times \Theta \rightarrow [0, \infty)$ als

$$\rho(x, p) = P(X = x \mid \vartheta = p)$$

(bzw. $P(X \in dx \mid \vartheta = p) = \rho(x, p) dx$ falls X eine Dichte besitzt). Mit Formel von der totalen Wahrscheinlichkeit (Satz 1.41, 1. ist

$$P(X = x) = \int_{\Theta} \rho(x, t) \alpha(t) dt$$

(bzw. $P(X \in dx) = \int_{\Theta} \rho(x, t) \alpha(t) dt dx$, d.h. $P(X \leq x) = \int_{\Theta} \int_{-\infty}^x \rho(y, t) dy \alpha(t) dt$, wenn X eine Dichte besitzt), mit der Formel von Bayes (Satz 1.41, 2. ist

$$P(\vartheta \in dp \mid X = x) = \frac{\rho(x, p) \alpha(p) dp}{\int_{\Theta} \alpha(\vartheta') \rho(x, \vartheta') d\vartheta'}$$

Die *a posteriori-Dichte* (bzw. a posteriori-Gewicht, wenn ϑ diskret) bei Beobachtung x ,

$$\pi_x(\vartheta) = \frac{\alpha(\vartheta) \rho(x, \vartheta)}{\int_{\Theta} \alpha(\vartheta') \rho(x, \vartheta') d\vartheta'},$$

ist die Dichte von ϑ , bedingt auf Beobachtung $X = x$, d.h.

$$P(\vartheta \leq u \mid X = x) = \int_{\Theta \cap (-\infty, u]} \pi_x(p) dp$$

Definition 2.11. Der *Bayes-Schätzer* (zur a priori-Verteilung α) ist

$$\widehat{\vartheta}_B = \widehat{\vartheta}_B(x) := \mathbb{E}_{\pi_x}[\vartheta] = \int_{\Theta} \vartheta \pi_x(\vartheta) d\vartheta$$

d.h. der Erwartungswert von ϑ bedingt auf $X = x$.

(Wir betrachten hier nur den Fall, dass Θ ein Intervall ist.)

Definition 2.12. Für einen Schätzer $Y = Y(X)$ (für ϑ) ist

$$F_{\alpha}(Y) := \int_{\Theta} \mathbb{E}[(Y - \vartheta)^2 \mid \vartheta = p] \alpha(p) dp$$

der erwartete quadratische Fehler (zur Vorbewertung α).

Der Bayes-Schätzer minimiert den erwarteten quadratischen Fehler (zur Vorbewertung α):

Satz 2.13. *Stets gilt $F_{\alpha}(Y) \geq F_{\alpha}(\widehat{\vartheta}_B(X))$.*

Beweis.

$$\begin{aligned} & F_{\alpha}(Y(X)) - F_{\alpha}(\widehat{\vartheta}_B(X)) \\ &= \int_{\Theta} \mathbb{E}[(Y(X) - \vartheta)^2 - (\widehat{\vartheta}_B(X) - \vartheta)^2 \mid \vartheta = p] \alpha(p) dp \\ &= \int_{\Theta} \mathbb{E}[Y(X)^2 - 2Y(X)\vartheta - \widehat{\vartheta}_B(X)^2 + 2\vartheta\widehat{\vartheta}_B(X) \mid \vartheta = p] \alpha(p) dp \\ &= \mathbb{E}[Y(X)^2 - 2Y(X)\vartheta - \widehat{\vartheta}_B(X)^2 + 2\vartheta\widehat{\vartheta}_B(X)] \\ &= \mathbb{E}[Y(X)^2 - \widehat{\vartheta}_B(X)^2] - 2\mathbb{E}[Y(X)\vartheta] + 2\mathbb{E}[\vartheta\widehat{\vartheta}_B(X)] \end{aligned}$$

Weiter ist

$$\begin{aligned} & \mathbb{E}[\vartheta\widehat{\vartheta}_B(X)] \\ &= \sum_{x \in S} \mathbb{E}[\vartheta\widehat{\vartheta}_B(X) I_{\{X=x\}}] = \sum_{x \in S} P(X=x) \widehat{\vartheta}_B(x) \mathbb{E}[\vartheta \mid X=x] \\ &= \sum_{x \in S} P(X=x) \widehat{\vartheta}_B(x) \mathbb{E}_{\pi_x}[\vartheta] = \sum_{x \in S} P(X=x) (\widehat{\vartheta}_B(x))^2 \\ &= \mathbb{E}[(\widehat{\vartheta}_B(X))^2] \end{aligned}$$

und analog

$$\mathbb{E}[\vartheta Y(X)] = \mathbb{E}[Y(X)\widehat{\vartheta}_B(X)].$$

Insgesamt:

$$\begin{aligned} & F_\alpha(Y(X)) - F_\alpha(\widehat{\vartheta}_B(X)) \\ &= \mathbb{E}[Y(X)^2 - \widehat{\vartheta}_B(X)^2 - 2Y(X)\widehat{\vartheta}_B(X) + 2\widehat{\vartheta}_B(X)^2] \\ &= \mathbb{E}[(Y(X) - \widehat{\vartheta}_B(X))^2] \geq 0. \end{aligned}$$

□

Beispiel: Münzwurf mit zufälliger Erfolgswahrscheinlichkeit

Beispiel 2.14 (Münzwurf mit zufälliger Erfolgswahrscheinlichkeit). ϑ wird gemäß einer Verteilung α auf $\Theta := [0, 1]$ „ausgewürfelt“, dann:

$n \in \mathbb{N}$, gegeben $\vartheta = u \in [0, 1]$ seien X_1, X_2, \dots, X_n unabhängig und jeweils $\sim \text{Ber}_u$ (d.h. $P(X_i = 1 \mid \vartheta = u) = u = 1 - P(X_i = 0 \mid \vartheta = u)$).

Somit: Für $x = (x_1, \dots, x_n) \in \{0, 1\}^n$ ist

$$\rho(x, \vartheta) = \vartheta^{\#\{i \leq n: x_i=1\}} (1 - \vartheta)^{\#\{i \leq n: x_i=0\}}.$$

Eine Situation, in der dieses Modell sinnvoll ist, könnte folgende sein: Nehmen wir an, ein Versicherungsnehmer hat jedes Jahr mit einer gewissen (zu ihm „gehörigen“) Wahrscheinlichkeit ϑ einen Schadensfall (unabhängig über die Jahre), und $\alpha(\vartheta) d\vartheta$ beschreibt die Verteilung der Schadenswahrscheinlichkeiten aller Kunden dieser Versicherung (diese Verteilung sei der Versicherung aus Erfahrungswerten bekannt).

Mit $\rho(x, \vartheta) = \text{Bin}_{n,\vartheta}(x)$ ist dann die Wahrscheinlichkeit, dass ein „typischer Kunde“ in n Jahren k Schadensfälle verursacht $\int_0^1 \alpha(\vartheta) \rho(k, \vartheta) d\vartheta$, und $\pi_k(\vartheta)$ ist die Verteilung der Schadenswahrscheinlichkeit pro Jahr eines Kunden, bedingt darauf, dass er in den letzten n Jahren k Schäden hatte. Diese Information kann die Versicherung beispielsweise für Vertragsverlängerung, Tarifierung, etc. benutzen.

Wir betrachten hier (nur) den Fall $\alpha = \text{unif}_{[0,1]}$.

Gehe über zu $Y = X_1 + \dots + X_n$, dann ist gegeben $\vartheta = u$, $Y \sim \text{Bin}_{n,u}$ und für $k \in \{0, 1, \dots, n\}$

$$\begin{aligned} P(Y = k) &= \int_0^1 \binom{n}{k} u^k (1-u)^{n-k} du \\ &= \frac{n!}{k!(n-k)!} \frac{1}{(n+1)!} = \frac{1}{n+1} \end{aligned}$$

(d.h. Y ist uniform auf $\{0, 1, \dots, n\}$).

Definition und Lemma 2.15 (Beta-Verteilungen). Für $a, b \in (0, \infty)$ ist die Beta-Funktion gegeben durch

$$B(a, b) = \int_0^1 u^{a-1} (1-u)^{b-1} du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

wobei $\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} dy$ die Gamma-Funktion ist (beachte: $\Gamma(a+1) = a\Gamma(a)$, speziell für $a \in \mathbb{N}$ ist $\Gamma(a) = (a-1)!$, wie man mit partieller Integration nachrechnen kann).

Für $a, b \in \mathbb{N}$ kann man explizit rechnen: Es ist dann

$$B(a, b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

denn

$$B(a, 1) = \int_0^1 u^{a-1} du = \left[\frac{1}{a} u^a \right]_{u=0}^{u=1} = \frac{1}{a}$$

und für $b \in \{2, 3, \dots\}$ ist (mit partieller Integration)

$$\begin{aligned} \int_0^1 u^{a-1} (1-u)^{b-1} du &= \left[\frac{1}{a} u^a (1-u)^{b-1} \right]_{u=0}^{u=1} - \int_0^1 \frac{1}{a} u^a \cdot (b-1)(1-u)^{b-2} (-1) du \\ &= \frac{b-1}{a} \int_0^1 u^a (1-u)^{b-2} du \end{aligned}$$

also

$$\begin{aligned} B(a, b) &= \frac{b-1}{a} B(a+1, b-1) = \frac{(b-1) \cdot (b-2) \cdots 2 \cdot 1 \cdot B(a+b-1, 1)}{a \cdot (a+1) \cdots (a+b-3) \cdots (a+b-2)} \\ &= \frac{(b-1) \cdot (b-2) \cdots 2 \cdot 1}{a \cdot (a+1) \cdots (a+b-2) \cdots (a+b-1)} = \frac{(a-1)!(b-1)!}{(a+b-1)!} \end{aligned}$$

Seien $r, s > 0$. Eine ZV V mit Werten in $[0, 1]$ ist *Beta-verteilt* mit Parametern $r, s > 0$, in Formeln $V \sim \beta_{r,s}$, wenn V die Dichte

$$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} v^{r-1} (1-v)^{s-1} \mathbf{1}_{(0,1)}(v)$$

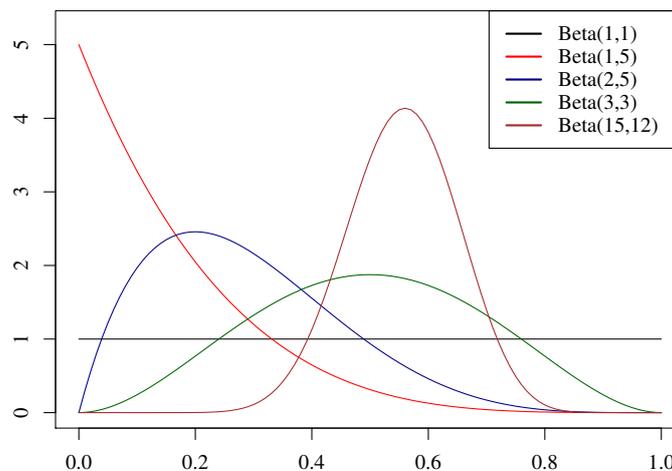
besitzt. Es gilt dann

$$\mathbb{E}[V] = \frac{r}{r+s}$$

denn

$$\begin{aligned} \mathbb{E}[V] &= \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \int_0^1 v \cdot v^{r-1} (1-v)^{s-1} dv \\ &= \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \frac{\Gamma(r+1)\Gamma(s)}{\Gamma(r+s+1)} = \frac{\Gamma(r+s)}{\Gamma(r+s+1)} \frac{\Gamma(r+1)}{\Gamma(r)} = \frac{r}{r+s} \end{aligned}$$

Einige Beta-Dichten:



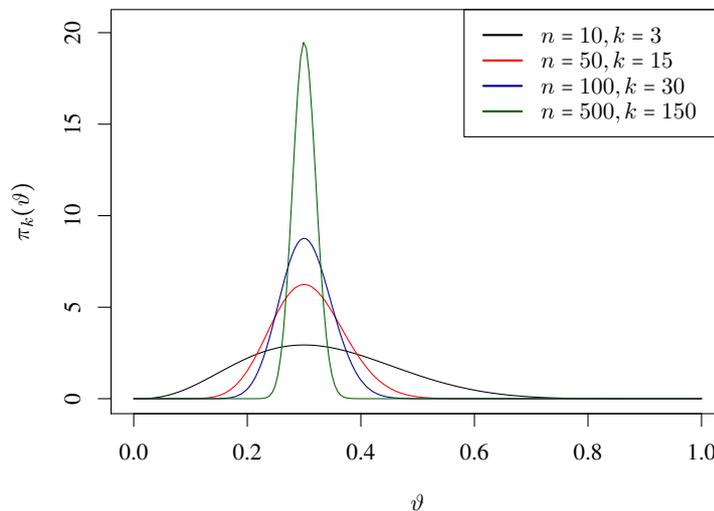
Zurück zum Beispiel 2.14 („Münzwürfe mit zufälliger Erfolgsw'keit“): Die a posteriori-Verteilung ist $\mathcal{L}(\vartheta | Y = k) = \beta_{k+1, n-k+1}$:

$$\begin{aligned} P(\vartheta \in dp | Y = k) &= \frac{P(\vartheta \in dp, Y = k)}{P(Y = k)} \\ &= \frac{1}{1/(n+1) \binom{n}{k}} p^k (1-p)^{n-k} dp \\ &= \frac{(n+1)!}{k!(n-k)!} p^k (1-p)^{n-k} dp \end{aligned}$$

Demnach (mit obigem Lemma 2.15 zur Beta-Verteilung) ist

$$\widehat{\vartheta}_B = \widehat{\vartheta}_B(Y) = \frac{Y+1}{n+2}$$

A posteriori-Dichte $\pi_k(\vartheta)$ für verschiedene n und k mit $k/n = 0.3$:



Wir sehen, dass für großes n die a posteriori-Verteilung recht eng um $\frac{Y+1}{n+2} \approx \frac{Y}{n}$ konzentriert ist. Zudem: die frequentistische und die Bayes'sche „Antwort“ stimmen für große n „nahezu“ überein.

Bemerkung. Laplace⁵ antwortete auf die die von ihm (vielleicht mit einem Augenzwinkern) gestellte Frage:

„Angenommen die Sonne ist bis heute n -mal aufgegangen. Mit welcher Wahrscheinlichkeit geht sie morgen auf?“

$$\frac{n+1}{n+2}$$

Dies passt zur Antwort des Bayes-Schätzers in obigem Beispiel.

⁵Pierre-Simon Laplace, 1749–1827; zitiert nach Kersting & Wakolbinger, *Elementare Stochastik*, 2. Aufl., Birkhäuser 2010, S. 127

2.3.3 Kleinste-Quadrate-Schätzer und lineare Regression

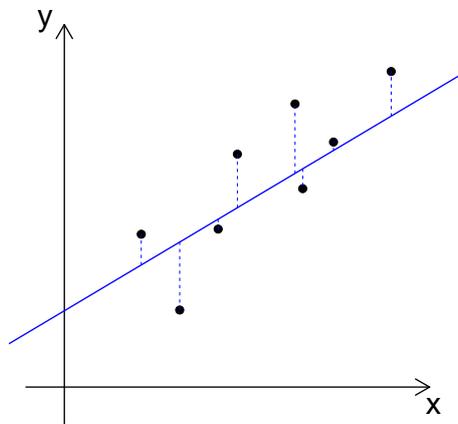
Der sogenannte kleinste-Quadrate-Ansatz ist ebenfalls ein recht allgemeines Prinzip zur Konstruktion von Schätzern, wir betrachten ihn hier am Beispiel des linearen Regressionsmodells:

Nehmen wir an, die Beobachtungen bestehen aus n Messwertpaaren (x_i, y_i) , $i = 1, \dots, n$ (Werte in \mathbb{R}^2) und wir vermuten aus theoretischen Gründen einen zumindest „ungefähren“ (affin-)linearen Zusammenhang, d.h. bei „perfekter“ Messung und „perfektem“ Zusammenhang gälte

$$y_i = \beta_0 + \beta_1 x_i$$

für gewisse (uns unbekannte) Zahlen β_0 und β_1 . (Ein „Lehrbuchbeispiel“: y_i ist die Länge einer Stahlfeder bei Zugbelastung mit Gewicht x_i innerhalb des Gültigkeitsbereich des Hooke'schen Gesetzes.)

Aufgrund beispielsweise von Messungenauigkeiten (oder womöglich auch weil der lineare Zusammenhang in Wirklichkeit nur approximativ gilt) werden die realen Datenpunkte typischerweise nicht auf einer Geraden liegen.



Formulierung als statistisches Modell: x_1, \dots, x_n seien feste (bekannte) Werte (x ist die „erklärende Variable“), für $\vartheta = (\beta_0, \beta_1) \in \Theta = \mathbb{R}^2$ sei unter P_ϑ

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

$$\text{mit } \varepsilon_i \text{ u.i.v. mit } \mathbb{E}[\varepsilon_i] = 0, \quad \text{Var}[\varepsilon_i] = \sigma^2$$

und wir fassen die beobachteten y_i -Werte als Realisierungen der Y_i auf (y ist die „abhängige Variable“ oder „Zielgröße“).

Ein naheliegender Ansatz, $\vartheta = (\beta_0, \beta_1)$ zu schätzen, ist der *kleinste-Quadrate-Schätzer*: Finde $\widehat{\beta}_0, \widehat{\beta}_1$ so, dass

$$\sum_{i=1}^n (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2 = \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

Die Lösung kennen wir schon (vgl. Beob. 1.84, die wir hier gewissermaßen nur „statistisch aussprechen“): Mit

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} := \frac{1}{n} \sum_{i=1}^n y_i,$$

$$\sigma_x^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \sigma_y^2 := \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\text{cov}_{x,y} := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

($\text{cov}_{x,y}$ ist die „empirische Kovarianz“ der x - und der y -Werte) ist

$$\widehat{\beta}_1 = \frac{\text{cov}_{x,y}}{\sigma_x^2}, \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

Die Gerade $x \mapsto \widehat{\beta}_0 + \widehat{\beta}_1 x$ heißt auch die *Ausgleichsgerade*, der Wert $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ der anhand von x_i „vorhergesagte Wert“ oder „Ausgleichswert“.

Man nennt weiter

$$r_i := y_i - \widehat{y}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$$

das *Residuum* zum i -ten Beobachtungswert (der „Rest“ der Abweichung, die von dem Modell (nur) durch den „Rauschterm“ erklärt wird).

$$\kappa_{x,y} := \frac{\text{cov}_{x,y}}{\sigma_x \sigma_y}$$

ist der (empirische) Korrelationskoeffizient, auch *Pearsons Korrelationskoeffizient*⁶ (stets ist $-1 \leq \kappa_{x,y} \leq 1$, nach Cauchy-Schwarz-Ungleichung).

$R = \kappa_{x,y}^2$ nennt man auch das *Bestimmtheitsmaß*. Je näher R an 1 liegt, um so besser passt die lineare Approximation der y -Werte durch die x -Werte. Das sieht man auch gut an der alternativen Formel

$$R = \frac{\sum_{i=1}^n (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n r_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Zu den Formeln: Betrachte eine ZV $(\widetilde{X}, \widetilde{Y})$ mit Werten in \mathbb{R}^2 , deren Verteilung die empirische Verteilung der Datenpunkte $\frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$ ist, so ist

$$\begin{aligned} \mathbb{E}[\widetilde{X}] &= \bar{x}, & \mathbb{E}[\widetilde{Y}] &= \bar{y}, \\ \text{Var}[\widetilde{X}] &= \sigma_x^2, & \text{Var}[\widetilde{Y}] &= \sigma_y^2, \\ \text{Cov}[\widetilde{X}, \widetilde{Y}] &= \text{cov}_{x,y}, & \kappa_{\widetilde{X}, \widetilde{Y}} &= \kappa_{x,y} \end{aligned}$$

und die Behauptung folgt wörtlich aus Beob. 1.84, dort hatten wir gerechnet:

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

denn der Ausdruck auf der linken Seite ist

$$\begin{aligned} & \text{Var}[\widetilde{Y} - \beta_1 \widetilde{X} - \beta_0] + (\mathbb{E}[\widetilde{Y}] - \beta_1 \mathbb{E}[\widetilde{X}] - \beta_0)^2 \\ &= \text{Var}[\widetilde{Y}] - 2\beta_1 \text{Cov}[\widetilde{X}, \widetilde{Y}] + \beta_1^2 \text{Var}[\widetilde{X}] + (\mathbb{E}[\widetilde{Y}] - \beta_1 \mathbb{E}[\widetilde{X}] - \beta_0)^2 \\ &= \sigma_{\widetilde{Y}}^2 - 2\beta_1 \sigma_{\widetilde{X}} \sigma_{\widetilde{Y}} \kappa_{\widetilde{X}, \widetilde{Y}} + \beta_1^2 \sigma_{\widetilde{X}}^2 + (\mathbb{E}[\widetilde{Y}] - \beta_1 \mathbb{E}[\widetilde{X}] - \beta_0)^2 \\ &= \sigma_{\widetilde{Y}}^2 (1 - \kappa_{\widetilde{X}, \widetilde{Y}}^2) + \sigma_{\widetilde{X}}^2 \left(\beta_1 - \frac{\sigma_{\widetilde{Y}}}{\sigma_{\widetilde{X}}} \kappa_{\widetilde{X}, \widetilde{Y}} \right)^2 + (\mathbb{E}[\widetilde{Y}] - \beta_1 \mathbb{E}[\widetilde{X}] - \beta_0)^2, \end{aligned}$$

was offensichtlich minimal wird für die Wahl

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

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

⁶nach Karl Pearson, 1858–1936

Zudem ist

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

dazu beachte analog:

$$\mathbb{E}[(\tilde{Y} - \beta_0)^2] = \mathbb{E}[\tilde{Y}^2] - 2\beta_0 \mathbb{E}[\tilde{Y}] + \beta_0^2 = \text{Var}[\tilde{Y}] + (\beta_0 - \mathbb{E}[\tilde{Y}])^2$$

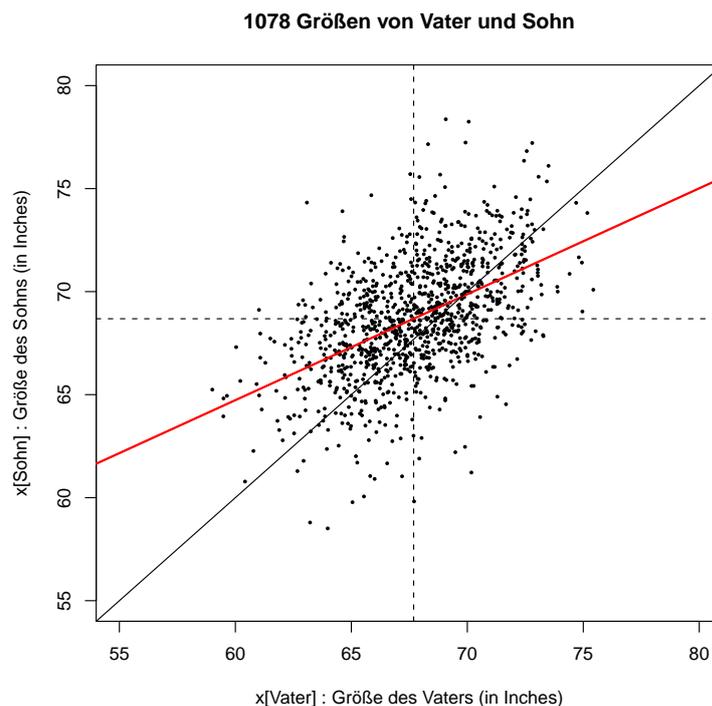
ist minimal für die Wahl $\beta_0 = \mathbb{E}[\tilde{Y}]$.

Übrigens: Wenn man zusätzlich annimmt, dass die ε_i u.i.v. $\sim \mathcal{N}_{0, \sigma^2}$ sind, so ist der kleinste-Quadrate-Schätzer hier auch zugleich der Maximum-Likelihood-Schätzer (mit einer Rechnung analog zum Beispiel für den Erwartungswert-ML-Schätzer, siehe Bsp. 2.9, 3.)

Beispiel: Größen von Vätern und Söhnen

Woher kommt der Name „Regression“ (nach lat. regressio, Zurückkommen)? Francis Galton (1822–1911, engl. Wissenschaftler) hat angesichts biometrischer Beobachtungen den Ausdruck “regression towards the mean” geprägt.

Ein (relativ berühmter) Datensatz von Karl Pearson (1858-1936):



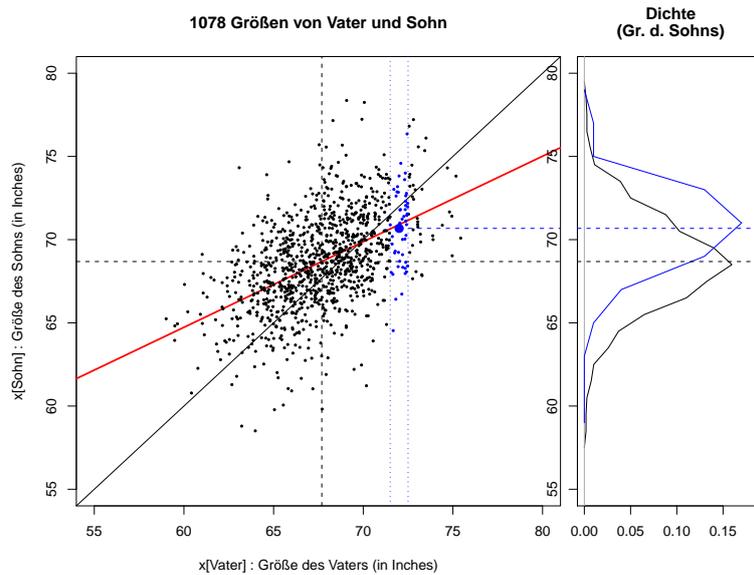
$$\bar{x}_{\text{Vater}} = 67.7, \bar{x}_{\text{Sohn}} = 68.7, \sigma_{\text{Vater}}^2 = 7.52 \quad (\sigma_{\text{Vater}} = 2.74, \sigma_{\text{Sohn}} = 2.81), \text{cov}(x_{\text{Vater}}, x_{\text{Sohn}}) = 3.87$$

(Korrelationskoeffizient $\kappa = \text{cov}(x_{\text{Vater}}, x_{\text{Sohn}}) / (\sigma_{\text{Vater}} \sigma_{\text{Sohn}}) = 0.50$)

Regressionsgerade: $x_{\text{Sohn}} = 33.89 + 0.514x_{\text{Vater}}$ (rot).

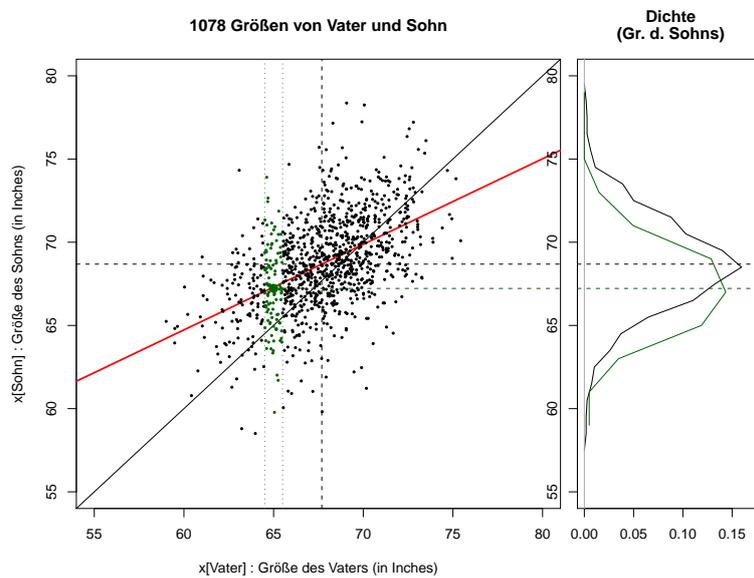
Schwarz gestrichelt: die jeweiligen Mittelwerte \bar{x}_{Vater} und \bar{x}_{Sohn} , schwarz durchgezogen: die Diagonale.

Betrachten wir die Söhne von überdurchschnittlich großen Vätern
(z.B. Väter, die ca. 72 Inches groß sind):



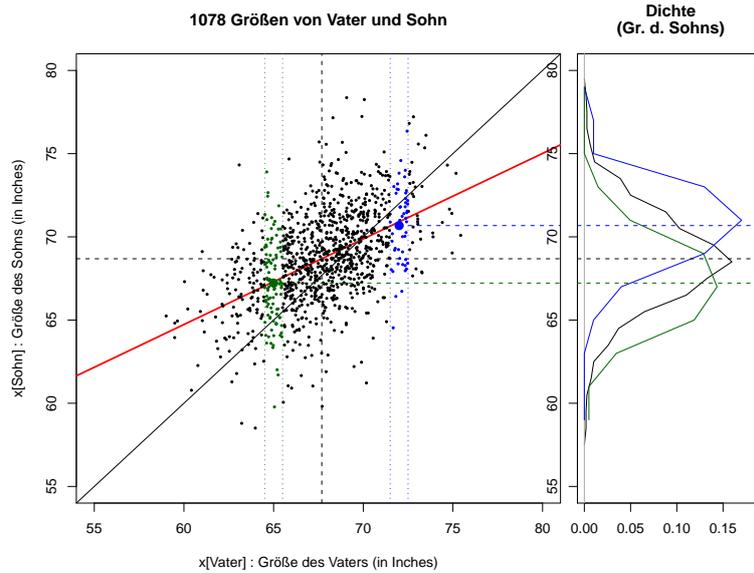
Diese Söhne sind überdurchschnittlich groß (im Vergleich zu allen Söhnen), aber im Mittel kleiner als ihr Vater.

Betrachten wir andererseits die Söhne von unterdurchschnittlich großen Vätern
(z.B. Väter, die ca. 65 Inches groß sind):



Diese Söhne sind unterdurchschnittlich groß (im Vergleich zu allen Söhnen), aber im Mittel größer als ihr Vater.

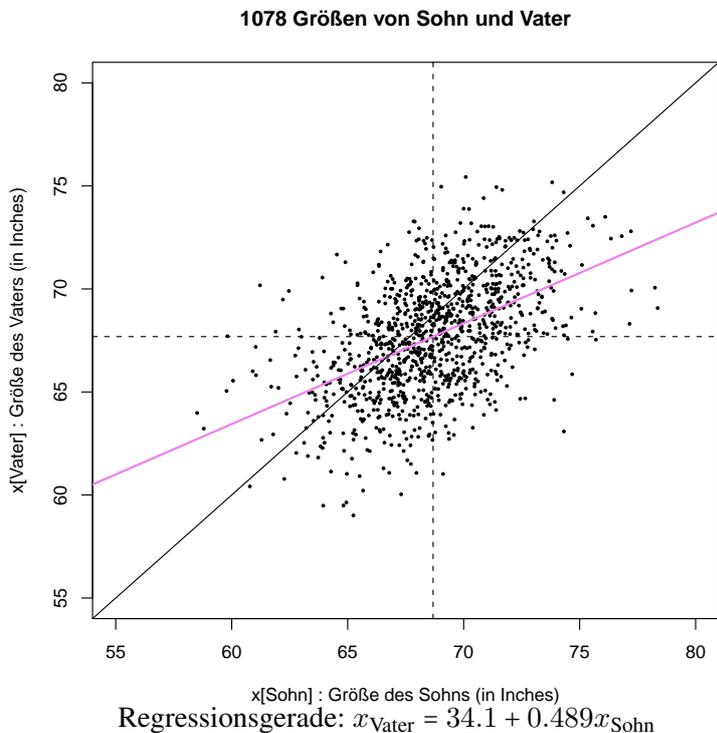
“Regression towards the mean”:



Wir sehen: Söhne überdurchschnittlich großer Väter sind im Mittel kleiner als ihr Vater (aber immer noch größer als der Populationsdurchschnitt), für Söhne unterdurchschnittlich großer Väter ist es umgekehrt: „Rückkehr zum Mittelwert“.

Bemerkung. Das beobachtete Phänomen der „Rückkehr zum Mittelwert“ bedeutet nicht notwendigerweise einen tieferen kausalen Zusammenhang, es tritt stets im Zusammenhang mit natürlicher Variabilität auf (technisch gesehen stets, wenn für den Korrelationskoeffizient κ gilt $|\kappa| < 1$).

Bestimmen wir (spaßeshalber) im Größen-Beispiel die Regressionsgerade für die Größe des Vaters als Funktion der Größe des Sohns: Wir hatten $\bar{x}_{\text{Vater}} = 67.7$, $\bar{x}_{\text{Sohn}} = 68.7$, $\text{cov}(x_{\text{Vater}}, x_{\text{Sohn}}) = 3.87$, $\sigma_{\text{Sohn}}^2 = 7.92$ und finden die Regressionsgerade $x_{\text{Vater}} = 34.1 + 0.489x_{\text{Sohn}}$



Literaturverzeichnis

- [K] N. Kurt, Stochastik für Informatiker: Eine Einführung in einheitlich strukturierten Lerneinheiten, Springer, 2020.
- [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.
- [Sf17] 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