Stochastik und Statistik

Prof. Dr. Leonhard Held
mit Ergänzungen von Prof. Dr. Torsten Hothorn
Ludwig–Maximilians–Universität München
Institut für Statistik

12. April 2016

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Als Literatur zur Vorlesung sind folgende Bücher geeignet:

- Held (2008): "Methoden der statistischen Inferenz. Likelihood und Bayes", Spektrum, 29,95 EUR (304 Seiten)
- Dümbgen (2003): "Stochastik für Informatiker", Springer Verlag, 30,79 EUR (268 Seiten)
- Fahrmeir, Künstler, Pigeot, Tutz (2010): "Statistik: Der Weg zur Datenanalyse", 7. Auflage, Springer Verlag, 30,79 EUR (610 Seiten)
- Georgii (2009): "Stochastik", 4. Auflage, deGruyter, 29,95 EUR (404 Seiten)
- Grimmett, Stirzaker (2001): "Probability and Random Processes", 3rd Edition, Oxford University Press, ca. 40 EUR (608 Seiten)
- Ligges (2008): "Programmieren mit R", 3. Auflage, Springer Verlag, 33,87 EUR (251 Seiten)

Das Gebiet **"Stochastik"** beinhaltet zum einen die **Wahrscheinlichkeitstheorie** und zum anderen die **Statistik**:

- Wahrscheinlichkeitstheorie: Mathematische Beschreibung von zufälligen Phänomenen
- Statistik: Erhebung, Auswertung und Interpretation von Daten sowie Quantifizierung von Unsicherheit

Wieso ist die Stochastik wichtig in der (Bio)-Informatik? Diese Frage kann beantwortet werden, wenn man die vielfältigen Anwendungsgebiete betrachtet:

- Simulation von zufälligen Phänomenen/Prozessen am Computer
  Bsp.: Verbreitung von Epidemien

- Analyse von statistischen/randomisierten Algorithmen
  Bsp.: Quicksort

- Statistische Analyse von Daten aus der Biologie und Genetik
  Bsp.: Genexpression und Überlebenszeit

- Statistische Modelle für das Auftreten von Daten
  Bsp.: Hardy-Weinberg-Gesetz

Ein Grundbegriff der Stochastik ist die **Wahrscheinlichkeit**, wie zum Beispiel die Wahrscheinlichkeit $P(A)$ für das Auftreten eines bestimmten Ereignisses A

$$\begin{array}{lcll} P(A) & = & 1 & : \quad A \text{ tritt mit Sicherheit ein} \\ P(A) & = & 0 & : \quad A \text{ tritt mit Sicherheit } \textit{nicht} \text{ ein} \\ P(A) & = & p \in (0,1) & : \quad \text{Ereignis } A \text{ tritt mit Wahrscheinlichkeit } p \text{ ein} \end{array}$$

Was gibt es für Möglichkeiten, eine Wahrscheinlichkeit zu interpretieren?

Einerseits ist die **subjektivistische Interpretation** möglich. Man fragt sich etwa:

- "Wie sicher bist **Du**, dass das Ereignis A eintreten wird?"

- "Wie groß schätzt **Du** die Wahrscheinlichkeit $P(A)$ des Ereignisses A ein?"

Die Wahrscheinlichkeit ist also ein Maß für die subjektive Unsicherheit. Unterschiedliche Individuen können den gleichen Ereignissen unterschiedliche Wahrscheinlichkeiten zuordnen. Um die Wahrscheinlichkeit $P(A)$ einer bestimmten Person zu quantifizieren, kann man der Person folgende Wette anbieten:

- "Wie viel Einsatz E wirst **Du** maximal auf das Eintreffen von A setzen, wenn beim Eintreten von A ein Gewinn G ausgezahlt wird?"

Dann ergibt sich die subjektive Wahrscheinlichkeit der betrachteten Person zu $P(A) = E/G$. Hierbei sieht man die Wahrscheinlichkeit als (entsprechend normierten) Wetteinsatz an.

Beispiel 1.0.1 (Spiel mit drei Bechern)
Unter einen von drei gleichartigen Bechern wird eine weiche Kugel gelegt. Nun beginnt der Spielanbieter, die Becher vor den Augen des Spielers zu vertauschen.
Der Spieler muss nach einer gewissen Zeit sagen, unter welchem Becher die Kugel liegt. Wenn er die Kugel findet, gewinnt er den doppelten Einsatz.
Der Spieler glaubt, dass er mit Wahrscheinlichkeit größer als 1/2 den richtigen Becher findet, denn ansonsten würde er im Mittel Geld verlieren. Der Spielanbieter hingegen glaubt, dass er die drei Becher so schnell vertauscht, dass der Spieler nicht mehr verfolgen kann, unter welchem Becher sich die Kugel befindet. Somit muss der Spieler raten und wird mit Wahrscheinlichkeit 1/3 den richtigen Becher finden. Selbst wenn diese Wahrscheinlichkeit größer als 1/3, aber zumindest kleiner als 1/2 ist, wird der Spielanbieter im Mittel Geld gewinnen. Die Wahrscheinlichkeit des gleichen Ereignisses wird also von unterschiedlichen Person unterschiedlich angesetzt; ansonsten würde das Spiel wohl nicht zustandekommen.

Andererseits kann man die Wahrscheinlichkeit auch **frequentistisch** interpretieren (auch **Häufigkeitsinterpretation**). Angenommen das betrachtete Zufallsexperiment kann beliebig oft unter gleichen Bedingungen wiederholt werden. Dann definiert man die Wahrscheinlichkeit $P(A)$ eines bestimmten Ereignisses A als den Grenzwert, gegen den die relative Häufigkeit des Eintretens des Ereignisses A konvergiert. Festzuhalten bleibt, dass der mathematische Calculus für Wahrscheinlichkeiten *unabhängig* von der jeweiligen Interpretation ist.
Die Graphiken und Beispiele im Skript und in der Vorlesung wurden mit R erstellt. Das Analysesystem R ist für alle gängigen Betriebssysteme frei erhältlich unter

http://www.R-Project.org

Eine geeignete deutschsprachige Einführung ist das Buch von Uwe Ligges "Programmieren mit R" (siehe Literturliste). Auf der obigen Webseite sind elektronische Handbücher und diverse Einführungen in diese Programmiersprache erhältlich. Am Institut für Statistik werden Vorlesungen und Kurse zu R angeboten.

Kapitel 2

Laplace-Verteilungen und diskrete Modelle

2.1 Laplace Wahrscheinlichkeiten

Man betrachtet die endliche Grundgesamtheit $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Nun kann man für ein Ereignis $A \subset \Omega$ die Laplace-Wahrscheinlichkeit definieren:

Definition 2.1.1
*Die **Laplace-Wahrscheinlichkeit** ist definiert als die Zahl*

$$P(A) := \frac{|A|}{|\Omega|} = \frac{|A|}{n}$$

wobei $|A|$ die Anzahl der Elemente in A ist.
*Jedes **Elementarereignis** $\{\omega_i\}$ mit $i = 1, \ldots, n$ hat also die Wahrscheinlichkeit $P(\{\omega_i\}) = \frac{1}{n}$. Die entsprechende Abbildung $\mathcal{P}(\Omega) \to [0, 1]$ nennt man **Laplace-Verteilung** oder auch **diskrete Gleichverteilung** auf Ω, wobei $\mathcal{P}(\Omega)$ die **Potenzmenge**, also die Menge aller Teilmengen von Ω, ist.*

Beispiel 2.1.1 (Augensumme von zwei Würfeln)
Für die Augensumme von zwei Würfeln hat die Grundgesamtheit die Form $\Omega = \{(1, 1), (1, 2), \ldots, (6, 5), (6, 6)\}$. Die Anzahl der Elemente in Ω beträgt dann $|\Omega| = n = 6^2 = 36$. Sei A_k das Ereignis "Augensumme ist k".

Es gilt:

$$\begin{aligned} P(A_1) &= 0 \\ P(A_2) &= \frac{1}{36} \\ P(A_3) &= \frac{2}{36} \\ &\vdots \quad\quad \vdots \\ P(A_7) &= \frac{6}{36} \\ P(A_8) &= \frac{5}{36} \\ &\vdots \quad\quad \vdots \\ P(A_{12}) &= \frac{1}{36} \end{aligned}$$

Allgemein:

$$P(A_k) = \frac{6 - |k-7|}{36} \quad \textit{für} \quad k = 2, \ldots, 12$$

Beispiel 2.1.2 (Skat)
Beim Skatspiel werden 32 *verschiedene Karten, darunter* 4 *Buben an* 3 *Spieler verteilt. Jeder Spieler erhält* 10 *Karten. Zwei Karten kommen in den Skat. Wie groß ist nun die Wahrscheinlichkeit folgender Ereignisse?*

$$\begin{aligned} A_1 &:= \textit{“Spieler 1 erhält alle Buben”} \\ A_2 &:= \textit{“Jeder Spieler erhält genau einen Buben”} \end{aligned}$$

Dazu wird zunächst die Grundgesamtheit Ω *geeignet definiert, welche die Elementarereignisse* ω *enthält:*

$$\omega = (\ \underbrace{\omega_1, \omega_2, \ldots, \omega_{10}}_{\textit{Karten für Sp. 1}}\ ,\ \underbrace{\omega_{11}, \omega_{12}, \ldots, \omega_{20}}_{\textit{Karten für Sp. 2}}\ ,\ \underbrace{\omega_{21}, \omega_{22}, \ldots, \omega_{30}}_{\textit{Karten für Sp. 3}}\ ,\ \underbrace{\omega_{31}, \omega_{32}}_{\textit{Skat}}\ )$$

Da es $32!$ *Permutationen von* $\omega_1, \ldots, \omega_{32}$ *gibt, hat* Ω *die Größe* $|\Omega| = 32!$ *und jedes Elementarereignis tritt mit der Wahrscheinlichkeit*

$$P(\{\omega\}) \ = \ \frac{1}{32!} \ = \ 3.8 \cdot e^{-36}$$

auf.

Wieviele Elemente enthält A_1?

$$\begin{aligned} |A_1| = \quad & 10 && \textit{(wo ist der ♣-Bube)} \\ & \cdot 9 && \textit{(wo ist der ♠-Bube)} \\ & \cdot 8 && \textit{(wo ist der ♡-Bube)} \\ & \cdot 7 && \textit{(wo ist der ♢-Bube)} \\ & \cdot 28! && \textit{(wo sind die übrigen 28 Karten)} \end{aligned}$$

Damit erhält man für die Wahrscheinlichkeit von A_1:

$$P(A_1) \;=\; \tfrac{10\cdot 9\cdot 8\cdot 7\cdot 28!}{32!} \;=\; \tfrac{10\cdot 9\cdot 8\cdot 7}{32\cdot 31\cdot 30\cdot 29} \;\approx\; 0.0058$$

Für das Ereignis A_2 geht man analog vor:

$$\begin{aligned} |A_2| = \quad & 10 \cdot 10 \cdot 10 \cdot 2 && \textit{(wo ist irgendein Bube)} \\ & \cdot 4! && \textit{(verteile die Buben auf 4 feste Positionen)} \\ & \cdot 28! && \textit{(wo sind die übrigen 28 Karten)} \end{aligned}$$

$$P(A_2) \;=\; \frac{|A_2|}{|\Omega|} \;=\; \frac{10^3 \cdot 2 \cdot 4! \cdot 28!}{32!} \;\approx\; 0.0556$$

Die Annahme gleichwahrscheinlicher Elementarereignisse bei Laplace-Verteilungen ist häufig zu restriktiv. Weiterhin wollen wir auch den Fall, dass Ω unendlich (aber zumindest abzählbar) ist, zulassen. Wir verallgemeinern daher im Folgenden Laplace-Wahrscheinlichkeitsräume zu diskreten Wahrscheinlichkeitsräumen.

2.2 Diskrete Wahrscheinlichkeitsräume

Definition 2.2.1
*Ein **diskreter Wahrscheinlichkeitsraum** ist ein Paar (Ω, P), wobei Ω eine abzählbare Grundgesamtheit ist und P ein diskretes Wahrscheinlichkeitsmaß, das jeder Teilmenge $A \subset \Omega$ eine Wahrscheinlichkeit $P(A)$ zuordnet. Diese definiert man wieder über die Wahrscheinlichkeiten der Elementarereignisse $\omega \in A$:*

$$P(A) = \sum_{\omega \in A} P(\{\omega\}),$$

wobei für die $P(\{\omega\})$ gelten muss

$$0 \leq P(\{\omega\}) \leq 1 \quad \text{für alle } \omega$$

$$\text{und} \quad \sum_{\omega \in \Omega} P(\{\omega\}) = 1$$

Beispiel 2.2.1 (Unsymmetrischer/ Unfairer Würfel)
Angenommen, man betrachtet wieder einen Würfelwurf mit Grundgesamtheit $\Omega = \{1, 2, 3, 4, 5, 6\}$. Nun würfelt der Würfel aber nicht jede Augenzahl mit der Wahrscheinlichkeit $1/6$, sondern

$$P(\{\omega\}) = \begin{cases} 0.2 & : \quad \omega \in \{1, 2\} \\ 0.15 & : \quad \omega \in \{3, 4, 5, 6\} \end{cases}$$

Offensichtlich gilt: $\sum_{\omega \in \Omega} P(\{\omega\}) = 1$

Beispiel 2.2.2 (Laplace-Verteilung)
Hier ist Ω endlich und $P(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{n}$

Beispiel 2.2.3
Man interessiere sich für die Anzahl i der Würfe einer fairen Münze, bis zum ersten Mal Zahl eintritt. Sei ω_i das Elementarereignis, dass beim i-ten Wurf zum ersten Mal Zahl eintrifft. Dann ist Ω offensichtlich unendlich mit $\Omega = \{\omega_1, \omega_2, \ldots\}$.
Für die Wahrscheinlichkeiten der Elementarereignisse ω_i gilt:

$$P(\{\omega_1\}) = \frac{1}{2}, P(\{\omega_2\}) = \frac{1}{4}, P(\{\omega_3\}) = \frac{1}{8}, \ldots$$

also

$$P(\{\omega_i\}) \; = \; \tfrac{1}{2^i} \text{ für } i = 1, 2, 3, \ldots$$

Man beachte, dass sich auch hier die Wahrscheinlichkeiten der Elementarereignisse zu eins addieren. Dies folgt aus der geometrischen Reihenformel

$$\sum_{i=1}^{\infty} b^i = \frac{b}{1-b} \quad \textit{(für } b < 1\textit{)},$$

angewandt auf $P(\{\omega_i\})$*:*

$$\sum_{i=1}^{\infty} P(\{\omega_i\}) = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = \frac{1/2}{1-1/2} = 1$$

2.3 Axiome von Kolmogorov für Wahrscheinlichkeiten

Andrei Kolmogorov [1903-1987] gilt als einer der Pioniere der modernen Wahrscheinlichkeitstheorie. Unter anderem stellte er folgende Axiome für Wahrscheinlichkeitsverteilungen auf.
Wir betrachten nun eine beliebige abzählbare Grundgesamtheit Ω und eine Funktion P auf der Potenzmenge $\mathcal{P}(\Omega)$, die jedem Ereignis $A \subset \Omega$ eine Wahrscheinlichkeit zuordnet.
Wir nennen P eine **Wahrscheinlichkeitsverteilung** auf Ω, wenn für sie die **Axiome von Kolmogorow** gelten:

A1) $P(A) \geq 0$ für beliebige $A \subset \Omega$

A2) $P(\Omega) = 1$

A3) $P(A \cup B) = P(A) + P(B)$ für disjunkte Ereignisse $A, B \subset \Omega$

Folgerungen aus den Axiomen:

- $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$
  für paarweise disjunkte Ereignisse $A_1, A_2, \ldots, A_n \subset \Omega$

- $P(A) \leq P(B)$ falls $A \subset B$

  Beweis:

  Zunächst definieren wir $B \setminus A = B \cap \bar{A}$ (in Worten: "B ohne A"). Dann gilt $B = (B \setminus A) \cup A$
  und daher $P(B) \overset{\text{A3}}{=} P(B \setminus A) + P(A) \geq P(A)$

- Das **Komplement** wird definiert als $\bar{A} = \Omega \setminus A$.
  Dann gilt $P(\bar{A}) = 1 - P(A)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ für beliebige $A, B \subset \Omega$

Die letzte Formel kann wie folgt verallgemeinert werden:

Satz 2.3.1 (Siebformel von Sylvester-Poincaré)
Von James Sylvester [1814-1897] und Jules Henri Poincaré [1854-1912]:

Für beliebiges $n \in \mathbb{N}$ und Ereignisse $A_1, A_2, \ldots, A_n \subset \Omega$ ist

$$\begin{aligned} P(A_1 \cup A_2 \cup \ldots \cup A_n) &= \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) \\ &\quad + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \\ &\quad \pm \ldots + (-1)^{n+1} \cdot P(A_1 \cap A_2 \cap \ldots \cap A_n). \end{aligned}$$

Insbesondere erhält man für drei beliebige Mengen $A, B, C \subset \Omega$

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ &\quad - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

Anstelle der doch recht komplexen Siebformel verwendet man häufig auch Ungleichungen zur Abschätzung von $P(A_1 \cup A_2 \cup \ldots \cup A_n)$:

Satz 2.3.2 (Bonferroni-Ungleichungen)
Für beliebige Ereignisse $A_1, A_2, \ldots A_n$ gilt

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) \begin{cases} \leq \sum_i P(A_i) \\ \geq \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) \end{cases}$$

Kapitel 3

Bedingte Wahrscheinlichkeiten und stochastische Unabhängigkeit

3.1 Bedingte Wahrscheinlichkeiten

Definition 3.1.1
Für Ereignisse $A, B \subset \Omega$ mit $P(B) > 0$ definiert man die ***bedingte Wahrscheinlichkeit*** *von A gegeben B als die Zahl*

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

Beispiel 3.1.1 (Spiel mit drei Bechern)
Wir nehmen im Folgenden an, dass der Spielanbieter die Becher so schnell vertauschen kann, dass der Spieler raten muss, wo sich die Kugel befindet. Allerdings hat er einen der drei Becher heimlich im Vorfeld markiert.
Betrachte nun folgende Ereignisse:

$A :=$ "Spieler findet den richtigen Becher"
$B :=$ "Spielanbieter legt die Kugel unter den markierten Becher"

Dann gilt für die bedingten Wahrscheinlichkeiten:

$$P(A|B) = 1 \quad \textit{und} \quad P(A|\bar{B}) = 1/2$$

Somit ergibt sich

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) = 1/3 + 1/2 \cdot 2/3 = 2/3.$$

Für bedingte Wahrscheinlichkeiten gilt offensichtlich:

$$\begin{aligned}
P(B|B) &= 1 \\
P(\bar{B}|B) &= 0 \\
P(\Omega|B) &= 1 \\
P(A|B) &= \frac{P(A \cap B)}{P(B)} \geq 0 \\
P((A_1 \cup A_2)|P(B)) &= \frac{P((A_1 \cup A_2) \cap B)}{P(B)} \\
&= \frac{P((A_1 \cap B) \cup (A_2 \cap B))}{P(B)} \\
&= \frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)} \\
&= P(A_1|B) + P(A_2|B),
\end{aligned}$$

hierbei folgt die vorletzte Zeile unter der Annahme, dass A_1 und A_2 disjunkt sind. Als Funktion von $A \subset \Omega$ ist $P(A|B)$ (bei festem B) also offensichtlich eine Wahrscheinlichkeitsverteilung, die den Axiomen von Kolmogorov genügt.

Beispiel 3.1.2 (Skat)
Die Ereignisse sind:

$A :=$ *"Mindestens eine der acht Karokarten liegt im Skat"*
$B :=$ *"Spieler 1 erhält beim Austeilen keine der acht Karokarten"*

Daher gilt:

$$\begin{aligned}
P(A|B) &= 1 - P(\bar{A}|B) \\
&= 1 - \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot 15 \cdot 14 \cdot 13 \cdot 24!}{24 \cdot 23 \cdot 22 \cdot 21 \cdot 20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot 15 \cdot 22!} \\
&= 1 - \frac{14 \cdot 13}{22 \cdot 21} \approx 0.606
\end{aligned}$$

$$\begin{aligned}
P(A) &= 1 - P(\bar{A}) \\
&= 1 - \frac{30 \cdot 29 \cdot 28 \cdot 27 \cdot 26 \cdot 25 \cdot 24 \cdot 23 \cdot 24!}{32!} \\
&= 1 - \frac{24 \cdot 23}{32 \cdot 31} \approx 0.444
\end{aligned}$$

Bevor die Karten ausgeteilt sind, ist die Wahrscheinlichkeit, dass mindestens eine der acht Karokarten im Skat liegt gleich 0.444. Wenn Spieler 1 bei sich

keine Karokarte findet, steigt die Wahrscheinlichkeit dafür, dass mindestens eine der acht Karokarten im Skat liegt auf 0.606.

Satz 3.1.1 (Multiplikationssatz)
Für beliebige Ereignisse $A_1, A_2, \ldots, A_n$ mit $P(A_1 \cap A_2 \cap \ldots \cap A_n) > 0$ gilt:

$$\begin{aligned} & P(A_1 \cap A_2 \cap \ldots \cap A_n) \\ = \; & P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n|A_1 \cap \ldots \cap A_{n-1}) \end{aligned}$$

wobei man die rechte Seite offensichtlich auch in jeder anderen möglichen Reihenfolge faktorisieren kann.

Wir schreiben im folgenden auch gerne $P(A_1, A_2) := P(A_1 \cap A_2)$ etc. Insbesondere gilt also

$$\begin{aligned} P(A_1, A_2) &= P(A_1) \cdot P(A_2|A_1) \\ P(A_1, A_2, A_3) &= P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \end{aligned}$$

Definition 3.1.2
*Man nennt Ereignisse $B_1, B_2, \ldots, B_n$, $B_i \in \Omega$ eine **disjunkte Zerlegung** von Ω, falls $B_1, B_2, \ldots, B_n$ paarweise disjunkt mit $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$.*

Satz 3.1.2 (Satz der totalen Wahrscheinlichkeit)
Sei $B_1, B_2, \ldots, B_n$ eine disjunkte Zerlegung von Ω. Dann gilt für jedes $A \subset \Omega$:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

Beweis:

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_n) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n) \end{aligned}$$

Bemerkung: Insbesondere gilt:

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}),$$

da $B, \bar{B}$ eine disjunkte Zerlegung von Ω ist.

Beispiel 3.1.3 (Spiel mit drei Bechern, Fortsetzung)
Da

$$P(A|B) = 1 \quad \textit{und} \quad P(A|\bar{B}) = 1/2$$

ergibt sich

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) = 1/3 + 1/2 \cdot 2/3 = 2/3.$$

3.2 Der Satz von Bayes

Satz 3.2.1 (Satz von Bayes)
Von Thomas Bayes [1701-1761]:
Dieser Satz beruht auf der Asymmetrie der Definition von bedingten Wahrscheinlichkeiten:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \quad \Rightarrow \quad P(A \cap B) = P(A|B)P(B) \\ P(B|A) &= \frac{P(A \cap B)}{P(A)} \quad \Rightarrow \quad P(A \cap B) = P(B|A)P(A) \end{aligned}$$

Damit folgt:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad \overset{\text{Totale W'keit}}{=} \quad \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}$$

Allgemeiner gilt für eine disjunkte Zerlegung $B_1, \ldots, B_n$ von Ω:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Bezeichnung:
$P(B_i)$ heißen "*a-priori*-Wahrscheinlichkeiten"
$P(B_i|A)$ heißen "*a-posteriori*-Wahrscheinlichkeiten"
Nach der Beobachtung von A ändert sich die Wahrscheinlichkeit für B_i von $P(B_i)$ zu $P(B_i|A)$.

Beispiel 3.2.1 (Diagnostische Tests)
Bei diagnostischen Tests betrachtet man die Ereignisse

$K :=$ *"Person ist krank"*
$T :=$ *"Test auf Krankheit ist positiv"*

Man kennt die:

$$\begin{aligned} \textbf{\textit{Sensitivität}}\ P(T|K) &\Rightarrow P(\bar{T}|K) = 1 - P(T|K) \\ \textbf{\textit{Spezifität}}\ P(\bar{T}|\bar{K}) &\Rightarrow P(T|\bar{K}) = 1 - P(\bar{T}|\bar{K}) \\ \textbf{\textit{Prävalenz}}\ P(K) &\Rightarrow P(\bar{K}) = 1 - P(K) \end{aligned}$$

Zahlenbeispiel:

$$P(T|K) = 0.222 \quad P(\bar{T}|\bar{K}) = 0.993 \quad P(K) = 0.0264$$

$$
\begin{aligned}
P(K|T) &= \frac{P(T|K)\cdot P(K)}{P(T|K)\cdot P(K) + P(T|\bar{K})\cdot P(\bar{K})} \\
&\quad \frac{0.222 \cdot 0.0264}{0.222 \cdot 0.0264 + 0.007(1-0.0264)} \\
&\approx 0.46 \\
P(\bar{K}|\bar{T}) &\approx 0.98
\end{aligned}
$$

Beispiel 3.2.2 (Diagnose Creutzfeldt-Jakob)
Bei der Creutzfeldt-Jakob-Erkrankung (CJD) handelt es sich um eine spongiforme Encephalopathie. Die Inzidenz (Anzahl der Neuerkrankungen in einem Jahr bezogen auf die Anzahl der Gesunden am Beginn des Jahres) beträgt 1 Fall pro 1 Million pro Jahr. Das große Problem dieser Krankheit ist eine Diagnose zu Lebzeiten. Die definitive Diagnose von CJD findet durch den Nachweis des PRP^{Sc} (gestörte Isoform des Prion-Proteins PRP) im Gehirn bei der Autopsie statt. Nur so erhält man einen gesicherten CJD-Fall.
Seit 1993 existiert die CJD Surveillance Studie, die alle CJD-Verdachtsfälle bundesweit registriert, dokumentiert und weiterhin überwacht. Falls möglich werden zusätzlich die Gehirne der Verstorbenen gesammelt. Bis Dezember 1996 wurden 289 Patienten Liquor (CSF-Proben) entnommen und analysiert. Davon war bei 238 Patienten eine klinische Diagnosestellung möglich.
Es wird ein diagnostischer Test gesucht, der genau die Situation abdeckt, in der der Test auch später bei lebenden (!) dementen Patienten angewandt werden kann, um die CJD-Erkrankung von anderen Demenzen zu trennen. Es soll untersucht werden, ob der Nachweis des 14-3-3-Proteins im Liquor ein "valider" Hinweis auf einen CJD-Fall ist.

Tabelle 3.1: Diagnostischer Test für Creutzfeldt-Jakob.

	CJD		
14-3-3	+	-	Total
+	126	7	133
-	8	97	105
Total	134	104	238

Die Sensitivität ist $P(\text{14-3-3} = +|CJD = +) = 126/134 = 0.94$ *und die Spezifität ist* $P(\text{14-3-3} = -|CJD = -) = 97/104 = 0.93$.

Etwas störend bei der Anwendung des Satzes von Bayes ist der unhandliche Nenner auf der rechten Seite. Durch den Übergang von Wahrscheinlichkeiten

π zu den zugehörigen Chancen (engl. "Odds") $\gamma = \pi/(1-\pi)$ können wir den Nenner loswerden und eine einfachere Version des Satzes von Bayes formulieren. Zunächst halten wir fest, dass zwischen Wahrscheinlichkeiten π und den zugehörigen Chancen γ offensichtlich folgende Zusammenhänge bestehen:

$$\gamma = \frac{\pi}{1-\pi} \tag{3.1}$$

$$\pi = \frac{\gamma}{1+\gamma} \tag{3.2}$$

Der Satz von Bayes lässt sich nun in folgender Variante darstellen:

Satz 3.2.2 (Satz von Bayes für Chancen)

$$\frac{P(B|A)}{P(\bar{B}|A)} = \frac{P(B)}{P(\bar{B})} \cdot \frac{P(A|B)}{P(A|\bar{B})}$$

$$\text{Posteriori Chance} = \text{Priori Chance} \cdot \text{Likelihood Quotient}$$

Man erhält also die Posteriori-Chance, indem man die Priori-Chance mit dem sogenannten Likelihood-Quotienten multipliziert.

Beispiel 3.2.3
In obigem Zahlenbeispiel zum diagnostischen Test ist die priori-Chance gleich $0.0264/(1-0.0264) \approx 0.0271$. *Der Likelihood-Quotient ist das Verhältnis von Sensitivität zu* $1-$*Spezifität, also gleich* $0.222/(1-0.993) = 0.222/0.007 \approx 31.7$. *Nach einem positiven Testergebnis erhöht sich also die Chance für Krankheit um den Faktor 31.7. Die Posteriori-Chance für Krankheit ist somit gleich* $0.0271 \cdot 31.7 \approx 0.86$, *was wiederum einer Posteriori-Wahrscheinlichkeit von* $0.86/(1+0.86) = 0.46$ *entspricht. Natürlich muss hier das Ergebnis, das mit der ersten Version des Satzes von Bayes berechnet wurde, bestätigt werden.*

Bemerkung:
Bedingte Wahrscheinlichkeiten $P(A|B)$ verhalten sich für festes B (angenommen $P(B) > 0$) wie gewöhnliche Wahrscheinlichkeiten. Somit kann man Rechenregeln, die für gewöhnliche Wahrscheinlichkeiten allgemeine Gültigkeit haben, verallgemeinern, indem man alle darin auftretenden Terme bzgl. eines weiteren Ereignisses bedingt.

Zum Beispiel kann man die **Siebformel** (Satz 2.3.1) wie folgt verallgemeinern:

$$P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C) \quad \text{für } P(C) > 0.$$

Andere Formeln und Rechenregeln können analog verallgemeinert werden. Im Folgenden wenden wir dieses Rezept auf den Satz von Bayes in der zweiten Variante an:

Satz 3.2.3 (Satz von Bayes bei zwei bedingenden Ereignissen)
Unter der Voraussetzung, dass die beteiligten Wahrscheinlichkeiten definiert sind, gilt:

$$\frac{P(A|B,C)}{P(\bar{A}|B,C)} = \frac{P(A|C)}{P(\bar{A}|C)} \cdot \frac{P(B|A,C)}{P(B|\bar{A},C)}$$

oder auch

$$\frac{P(A|B,C)}{P(\bar{A}|B,C)} = \frac{P(A|B)}{P(\bar{A}|B)} \cdot \frac{P(C|A,B)}{P(C|\bar{A},B)}$$

Beispiel 3.2.4
Der Amerikaner O.J. Simpson, berühmter Footballspieler und später auch Schauspieler, wurde 1994 wegen Mordes an seiner Ex-Frau und deren Liebhaber angeklagt. Im Prozess wurde Simpson vorgeworfen, seine Frau früher geschlagen und vergewaltigt zu haben. Simpsons Verteidiger, Alan Dershowitz, wies diese Vorwürfe als irrelevant zurück, da nur jeder 1000. Mann, der seine Frau schlägt, sie schließlich auch umbringen würde. Der emeritierte Statistikprofessor I.J. Good hielt dagegen, dass es hier nicht um die Wahrscheinlichkeit gehe, dass ein Mann seine Frau umbringe, wenn er sie zuvor geschlagen habe. Gesucht sei vielmehr die Wahrscheinlichkeit, dass ein Mann seine Frau umgebracht hat, wenn er sie zuvor geschlagen hat und wenn diese Frau dann tatsächlich von jemandem umgebracht worden ist. Auf der Grundlage eigener Schätzungen und von der Verteidigung gelieferter Zahlen berechnete Good diese Wahrscheinlichkeit als keineswegs verschwindend gering. Er schickte seinen Artikel sowohl an die Zeitschrift Nature *als auch Simpsons Verteidigung und die Polizei von Los Angeles. Es ist jedoch davon auszugehen, dass nicht alle Empfänger die wahrscheinlichkeitstheoretischen Überlegungen gleichermaßen verstanden haben.*
Simpson wurde im Strafprozess von den Geschworenen freigesprochen, von einem Zivilgericht jedoch zu Schadensersatzzahlungen an die Hinterbliebenen der Opfer verurteilt.

Im Folgenden wollen wir uns mit den Berechnungen eines Artikels von J.F. Merz und J.P. Caulkins[1] *und denen von Good*[2] *befassen. Dazu werden drei Ereignisse definiert:*

$$\begin{array}{rcl} A(Abuse) & : & \text{"Mann hat seine Frau geschlagen"} \\ M(Murder) & : & \text{"Frau wurde umgebracht"} \\ G(Guilty) & : & \text{"Mann hat seine Frau umgebracht"} \end{array}$$

Folgende bedingte Wahrscheinlichkeiten werden von Merz & Caulkins verwendet:

$$\begin{aligned} P(G|M) &= \frac{1430}{4936} = 0.29 \\ \Rightarrow P(\bar{G}|M) &= 0.71 \\ P(A|G,\, M) &= 0.5 \\ P(A|\bar{G},\, M) &= 0.05. \end{aligned}$$

Diese Wahrscheinlichkeiten basieren auf folgenden empirischen Zahlen: Von den 4936 Frauen, die 1992 ermordet wurden, wurden 1430 von Ihren Ehemännern ermordert, daher $P(G|M) = 0.29$. Aus einem Zeitungsartikel zitieren die Autoren den Verteidiger Dershowitz wie folgt: "It is, of course, true that, among the small number of men who do kill their present or former mates, a considerable number did first assault them." Merz & Caulkins interpretieren "a considerable number" als 50%, so dass $P(A|G,\, M) = 0.5$ folgt. Schließlich nehmen sie an, dass $P(A|\bar{G},\, M)$ gleich der Wahrscheinlichkeit ist, dass eine zufällig ausgewählte Frau geschlagen wird. Empirische Daten schätzen den Anteil von Frauen, die geschlagen werden, auf 5%, also $P(A|\bar{G},\, M) = 0.05$. Unter Anwendung von Satz 3.2.3 ergibt sich somit

$$\begin{aligned} \frac{P(G|A,\, M)}{P(\bar{G}|A,\, M)} &= \frac{P(G|M)}{P(\bar{G}|M)} \cdot \frac{P(A|G,\, M)}{P(A|\bar{G},\, M)} \\ &= \frac{0.29}{0.71} \cdot \frac{0.5}{0.05} \\ &\approx 4. \end{aligned}$$

Die Chance von $G|A,\, M$ ist also ungefähr gleich 4, mit (3.2) folgt, dass die zugehörige Wahrscheinlichkeit $P(G|A,\, M) \approx 4/(1+4) = 0.8$ ist. Das heißt mit einer Wahrscheinlichkeit von 0.8 hat ein Mann seine Frau umgebracht, wenn er sie zuvor geschlagen hat und wenn diese Frau dann tatsächlich von

[1] J.F. Merz and J.P. Caulkins (1995): "Propensity to abuse–propensity to murder?", *Chance*, 8(2), 14.

[2] I.J. Good (1995): "When batterer turns murderer", *Nature*, 375(6532), 541.

jemandem umgebracht worden ist.
Überraschend ist das Ergebnis, wenn man die Methode von Good verwendet, die ebenfalls auf Satz 3.2.3 beruht: Er verwendet die Zerlegung

$$\frac{P(G|A,\, M)}{P(\bar{G}|A,\, M)} = \frac{P(G|A)}{P(\bar{G}|A)} \cdot \frac{P(M|G,\, A)}{P(M|\bar{G},\, A)} \tag{3.3}$$

und schätzt $P(G|A) = 1/10000$. *Hintergrund dieser Schätzung ist die Aussage von Dershowitz, dass nur jeder 1000. Mann, der seine Frau schlägt, sie schließlich auch umbringen würde. Good nimmt also an, dass das (mindestens) mit Wahrscheinlichkeit 1/10 in dem fraglichen Jahr passieren wird. Da* $P(M|G,\, A) = 1$, *bleibt nur noch* $P(M|\bar{G},\, A)$ *zu quantifizieren. Offensichtlich gilt* $P(M|\bar{G},\, A) = P(M|\bar{G}) \approx P(M)$. *Da es jährlich in den Vereinigten Staaten ca. 25,000 Morde gibt, folgt bei 250,000,000 Einwohnern, dass*

$$P(M|\bar{G},\, A) \;=\; \frac{25000}{250000000} = \frac{1}{10000}.$$

Setzt man diese Zahlen in (3.3) ein, so erhält man

$$\frac{P(G|A,\, M)}{P(\bar{G}|A,\, M)} = \frac{\frac{1}{10000}}{\frac{9999}{10000}} \cdot \frac{1}{\frac{1}{10000}} \approx 1,$$

also $P(G|A,\, M) = 0.5$.
Beide Ansätze führen, bei völlig unterschiedlicher empirischer Grundlage, zu deutlich höheren Wahrscheinlichkeiten dafür, dass O.J. Simpson seine Frau ermordet hat.

3.3 Stochastische Unabhängigkeit

Wann bezeichnet man zwei Ereignisse A, B als (stochastisch) **unabhängig**? Immer dann, wenn sich die Wahrscheinlichkeit für eines der beiden Ereignisse nicht ändert, wenn man mit dem anderen Ereignis bedingt, d.h. wenn

$$P(A|B) = P(A) \text{ bzw. } P(B|A) = P(B)$$

gilt. Dies führt wegen $P(A|B) = P(A \cap B)/P(B)$ zu folgender Definition:

Definition 3.3.1
*Zwei Ereignisse A, B sind genau dann **unabhängig**, wenn*

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

gilt.

Man zeigt leicht, dass dann auch $\bar{A}$ und B, A und $\bar{B}$ oder auch $\bar{A}$ und $\bar{B}$ unabhängig sind. Beispielsweise gilt

$$\begin{aligned} P(A \cap \bar{B}) &= P(A) - P(A \cap B) \\ &= P(A) - P(A) \cdot P(B) \\ &= P(A)[1 - P(B)] \\ &= P(A) \cdot P(\bar{B}) \end{aligned}$$

Beispiel 3.3.1 (Zweimaliges Würfeln)
Ein fairer Würfel wird zweimal hintereinander geworfen. Die Ereignisse sind

A :="Beim 1. Würfelwurf eine Sechs"
B :="Beim 2. Würfelwurf eine Sechs"

Bei jedem einzelnen Würfelwurf ist die Grundgesamtheit $\Omega = \{1, 2, 3, 4, 5, 6\}$.
Nach Laplace gilt $P(A) = P(B) = 1/6$.
Bei "unabhängigem" Werfen gilt somit

$$P(A \cap B) = P(A) \cdot P(B) = 1/36$$

Angenommen der Würfelwerfer ist besonders am Werfen von einem Pasch interessiert. Er kann den zweiten Wurf ein wenig steuern und würfelt mit Wahrscheinlichkeit 1/2 das gleiche wie beim ersten Wurf. Die anderen Ergebnisse seien dann gleichverteilt mit Wahrscheinlichkeit 0.1.
Dann ist zwar $P(A) = 1/6$ und auch $P(B) = 1/6$, aber

$$P(A \cap B) = 1/12 > 1/36$$

Die Ereignisse A und B sind also abhängig, da

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

Die Unabhängigkeit von mehr als zwei Ereignissen lässt sich folgendermaßen definieren:

Definition 3.3.2
$A_1, A_2, \ldots, A_n$ *sind* ***(stochastisch) unabhängig****, wenn für alle Teilmengen* $I \subset \{1, 2, \ldots, n\}$ *mit* $I = \{i_1, i_2, \ldots, i_n)\}$ *gilt:*

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \ldots \cdot P(A_{i_k})$$

Bemerkung: Aus der paarweisen Unabhängigkeit folgt *nicht* die Unabhängigkeit von mehr als zwei Ereignissen.

Beispiel 3.3.2
Seien der Laplace-Wahrscheinlichkeitsraum (Ω, P) *(etwa einmaliges Ziehen aus einer Urne) und die Ereignisse* A_i *folgendermaßen definiert:*

$$\begin{aligned} &\Omega = \{0, 1, 2, 3\} \\ &A_i = \{0\} \cup \{i\} \quad \textit{mit } i = 1, 2, 3 \end{aligned}$$

Dann gilt $P(A_i) = P(\{0\} \cup \{i\}) = P(\{0\}) + P(\{i\}) = \frac{1}{2}$ *(für alle* $i = 1, 2, 3$ *nach A3) und*

$$\begin{aligned} P(A_i \cap A_j) &= P((\{0\} \cup \{i\})) \cap ((\{0\} \cap \{j\}))) \\ &= P((\{0\} \cup \{i\}) \cap \{0\} \cup (\{0\} \cup \{i\}) \cap \{j\}) \\ &= P(\{0\}) = \frac{1}{4} = P(A_i) \cdot P(A_j) \quad \forall i \neq j. \end{aligned}$$

Damit sind die Ereignisse paarweise unabhängig.
Aber:

$$P(A_1 \cap A_2 \cap A_3) = P(\{0\}) = \tfrac{1}{4}$$
und
$$P(A_1) \cdot P(A_2) \cdot P(A_3) = \tfrac{1}{8}$$

A_1, A_2, A_3 *sind also* nicht *unabhängig.*

Abschließend lernen wir noch den Begriff der bedingten Unabhängigkeit kennen, den wir im Kapitel über Markov-Ketten benötigen.

Definition 3.3.3
Sei C *ein beliebiges Ereignis mit* $P(C) > 0$*. Zwei Ereignisse* A *und* B *nennt man* ***bedingt unabhängig gegeben*** C*, wenn*

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

gilt.

Die folgenden Beispiele illustrieren, dass weder aus (unbedingter) Unabhängigkeit bedingte Unabhängigkeit (bzgl. einem Ereignis C), noch aus bedingter Unabhängigkeit bzgl. einem Ereignis C unbedingte Unabhängigkeit folgt.

Beispiel 3.3.3
Eine faire Münze wird zweimal unabhängig voneinander geworfen. Seien A und B die Ereignisse, dass die erste bzw. die zweite Münze Zahl zeigt. Sei C das Ereignis, dass beide Münzen das gleiche Symbol (entweder Kopf oder Zahl) zeigen. Dann sind A und B unabhängig, da

$$P(A) = P(B) = \frac{1}{2} \text{ und } P(A \cap B) = \frac{1}{4} = P(A) \cdot P(B),$$

aber nicht bedingt unabhängig gegeben C, da

$$P(A|C) = P(B|C) = \frac{1}{2} \text{ aber } P(A \cap B|C) = \frac{1}{2} \neq P(A|C) \cdot P(B|C).$$

Beispiel 3.3.4
Ein fairer Würfel wird zweimal unabhängig voneinander geworfen. Sei A das Ereignis, dass die kleinere Zahl gleich 1 ist und B das Ereignis, dass die größere Zahl gleich 4 ist. Sei C das Ereignis, dass die kleinere Zahl kleiner oder gleich drei ist und die größere Zahl größer oder gleich vier ist. Dann sind A und B nicht unabhängig, da

$$P(A) = \frac{11}{36},\ P(B) = \frac{7}{36} \text{ aber } P(A \cap B) = \frac{2}{36} \neq P(A) \cdot P(B),$$

A und B sind aber bedingt unabhängig gegegen C:

$$P(A|C) = P(B|C) = \frac{1}{3} \text{ und } P(A \cap B|C) = \frac{1}{9} = P(A|C) \cdot P(B|C).$$

3.4 Das Hardy-Weinberg-Gesetz

Gegeben sind eine Population von diploiden Organismen sowie zwei Allele a und b. Für ein bestimmtes Gen gibt folgende drei Genotypen:

$$\begin{array}{lll} aa & & P(aa) = p_{aa} \\ ab & \text{mit Wahrscheinlichkeiten} & P(ab) = p_{ab} \\ bb & & P(bb) = p_{bb} \end{array}$$

Folgende Annahmen werden vereinfachend getroffen:

- Wahrscheinlichkeiten sind unabhängig vom Geschlecht
- die Paarungen sind "rein zufällig"
- es tritt keine Mutation, Selektion, ... auf

Betrachte die Ereignisse:

$$\left.\begin{array}{l} M_x = \text{"Mutter ist von Typ } x\text{"} \\ V_x = \text{"Vater ist von Typ } x\text{"} \\ K_x = \text{"Kind ist von Typ } x\text{"} \end{array}\right\} x \in \{aa, ab, bb\}$$

Nun wird die bedingte Wahrscheinlichkeit berechnet, dass das Kind K den Genotyp x hat, gegeben die Eltern haben die Genotypen y und z, also $P(K_x | M_y \cap V_z)$. Die Einträge in den einzelnen Zellen folgender Tabelle geben diese bedingten Wahrscheinlichkeiten, die sich unter den gemachten Annahmen ergeben, in der Reihenfolge $x = (aa, ab, bb)$ an.

	V_{aa}	V_{ab}	V_{bb}
M_{aa}	$1, 0, 0$	$\frac{1}{2}, \frac{1}{2}, 0$	$0, 1, 0$
M_{ab}	$\frac{1}{2}, \frac{1}{2}, 0$	$\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$	$0, \frac{1}{2}, \frac{1}{2}$
M_{bb}	$0, 1, 0$	$0, \frac{1}{2}, \frac{1}{2}$	$0, 0, 1$

Tabelle 3.2: Bedingte Wahrscheinlichkeiten, dass das Kind Genotyp (aa, ab, bb) hat, gegeben den Genotyp der Eltern.

Wir interessieren uns nun für die unbedingten Wahrscheinlichkeiten $P(K_{aa})$, $P(K_{bb})$ und $P(K_{ab})$

$$\begin{aligned}
P(K_{aa}) &= \sum_{y,z} P(K_{aa}|M_y \cap V_z) P(M_y \cap V_z) \\
&= \sum_{y,z} P(K_{aa}|M_y \cap V_z) P(M_y) P(V_z) \\
&= \sum_{y,z} P(K_{aa}|M_y \cap V_z) \cdot p_y \cdot p_z \\
&= 1 \cdot p_{aa} \cdot p_{aa} + \frac{1}{2} \cdot p_{aa} \cdot p_{ab} + 0 + \frac{1}{2} \cdot p_{ab} \cdot p_{aa} + \frac{1}{4} \cdot p_{ab} \cdot p_{ab} + 0 + 0 \\
&= p_{aa}^2 + p_{aa} p_{ab} + \frac{1}{4} p_{ab}^2 \\
&= \left(\underbrace{p_{aa} + \frac{p_{ab}}{2}}_{=: q} \right)^2 = q^2
\end{aligned}$$

Aus Symmetriegründen gilt:

$$P(K_{bb}) \quad = \left(p_{bb} + \frac{p_{ab}}{2} \right)^2 \;=\; (1-q)^2$$

und somit

$$P(K_{ab}) \;=\; 1 - q^2 - (1-q)^2 \;=\; 2q(1-q)$$

In der nächsten Generation erhält ein Individuum den Genotyp $x \in \{aa, ab, bb\}$ mit der Wahrscheinlichkeit $P(x)$:

$$P(x) = \begin{cases} q^2 & \text{für } x = aa \\ 2q(1-q) & \text{für } x = ab \\ (1-q)^2 & \text{für } x = bb \end{cases}$$

Dies ist die **Hardy-Weinberg-Verteilung**, die nur noch von einem Parameter q abhängt. Dieser beschreibt die Häufigkeit des Allels a (welches in den Genotypen aa und ab vorkommt). Interessanterweise stellt diese Verteilung sogar einen **Gleichgewichtszustand** dar. Um dies zu sehen nehmen wir an, dass die Wahrscheinlichkeiten der einzelnen Genotypen durch eine Hardy-Weinberg-Verteilung mit Parameter q beschrieben wird. In der folgenden Generation erhält man die Wahrscheinlichkeiten $\bar{q}^2$, $2\bar{q}(1-\bar{q})$, $(1-\bar{q})^2$ der Genotypen aa, ab und bb, wobei aber offensichtlich $\bar{q} = q^2 + \frac{2q(1-q)}{2} = q$

gilt. Folglich ist die Verteilung in der betrachteten Generation identisch mit der Verteilung der vorhergehenden Generation.
Das Hardy-Weinberg-Gleichgewicht impliziert, dass man bei Kenntnis der Häufigkeit des Genotyps *aa* oder *bb* auch die Häufigkeiten der anderen beiden Genotypen kennt. Bei Kenntnis der Häufigkeit des Genotyps *ab* ist dies allerdings nicht ganz so, wie folgende Abbildung illustriert ($P(ab)$ ist nicht umkehrbar).

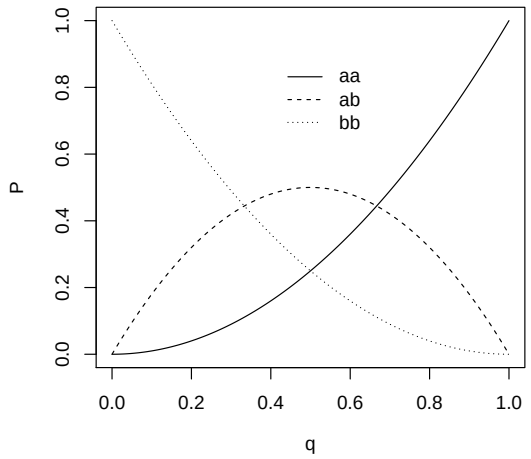

Abbildung 3.1: Das Hardy-Weinberg-Gleichgewicht für die Genotypen *aa*, *bb* und *ab*.

Beispiel 3.4.1
Angenommen nur der Genotyp bb verursacht eine Krankheit ("Phänotyp"). Die Genotypen aa und ab seien rein äußerlich nicht unterscheidbar. Wenn der sichtbare Genotyp bb mit einer relativen Häufigkeit p_{bb} auftritt, dann kann man unter der Annahme, dass die Population im H-W-Gleichgewicht ist, die anderen relativen Häufigkeiten berechnen. Sei q die Häufigkeit des Allels a. Dann gilt:

$$\begin{aligned} p_{bb} &= (1-q)^2 \quad \Rightarrow \quad \sqrt{p_{bb}} = 1-q \quad \Rightarrow \quad q = 1-\sqrt{p_{bb}} \\ p_{aa} &= q^2 = (1-\sqrt{p_{bb}})^2 \\ p_{ab} &= 2q(1-q) = 2(1-\sqrt{p_{bb}})\,[1-(1-\sqrt{p_{bb}})] = 2\,(1-\sqrt{p_{bb}})\,\sqrt{p_{bb}} \end{aligned}$$

Zahlenbeispiel:

$$p_{bb} = 0.0001 \qquad \Rightarrow \qquad \begin{aligned} p_{aa} &= 0.9801 \\ p_{ab} &= 0.0198 \end{aligned}$$

Es kann noch das "Hardy-Weinberg-**Un**gleichgewicht" formuliert werden: Die Werte hängen dabei nicht nur von q ab, sondern auch noch von einem weiteren Parameter, dem **Ungleichgewichtskoeffizienten** d:

$$P(x) = \begin{cases} q^2+d & \text{für } x = aa \\ 2q(1-q)-2d & \text{für } x = ab \\ (1-q)^2+d & \text{für } x = bb \end{cases}$$

Für $d = 0$ erhält man wieder das Hardy-Weinberg-Gleichgewicht. Problematisch ist unter Umständen, dass der Wertebereich für q und d beschränkt ist, wie folgende Abbildung illustriert:

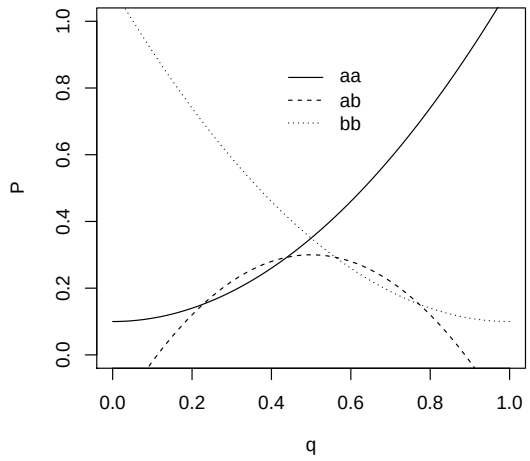

Abbildung 3.2: Das Hardy-Weinberg-Ungleichgewicht mit $d = 0.1$ für die Genotypen aa, bb und ab.

Später werden wir statistische Verfahren kennenlernen, wie man bei gegebenen empirischen Daten die Parameter q und d schätzen kann und auch untersuchen kann, ob die Daten statistisch signifikant gegen die Hardy-Weinberg-Annahme ($d = 0$) sprechen.

Kapitel 4

Diskrete Zufallsvariablen

4.1 Einleitung

Die Ergebnisse von Zufallsvorgängen sind nicht notwendigerweise Zahlen. Oft ist es aber hilfreich, diese durch Zahlen zu repräsentieren, um mit ihnen rechnen zu können.

Beispiel 4.1.1 (4-maliger Wurf einer Münze)
$\Omega = \{Wappen, Zahl\} = \{W, Z\}^4$
Beispiel für ein Ereignis: $\omega = \{W, Z, Z, W\}$
Angenommen man interessiert sich für

$$X := \text{“Anzahl von Wappen”}$$

Dann nennt man X eine ***Zufallsvariable*** *(ZV) mit reellen* ***Ausprägungen*** $X = x \in \mathbb{R}$.
X ist eine Abbildung von Ω nach $\mathbb{R}$.

Vorteile:

1. man kann mit X “rechnen”.

$$\text{z. B. } P(X \leq x) \quad \text{oder} \quad P(X^2 > y)$$

2. der Wahrscheinlichkeitsraum Ω wird meist nicht mehr (explizit) benötigt

Definition 4.1.1
Eine Zufallsvariable X heißt **diskret**, *falls sie nur endlich oder abzählbar unendlich viele Werte $x_1, x_2, \ldots$ annehmen kann. Die Menge $\{x_1, x_2, \ldots\}$ der möglichen Ausprägungen von X, also alle $x_i \in \mathbb{R}$ mit $f(x_i) > 0$ heißt*

Träger $\mathcal{T}$ *der Zufallsvariable* X*.*

Die **Wahrscheinlichkeitsfunktion** *von* X *ist durch*

$$f(x_i) = P(X = x_i)$$

für $x_i \in \mathcal{T}$ *gegeben. Dabei steht* $P(X = x_i)$ *für* $P(\{\omega \in \Omega : X(\omega) = x_i\})$*.*

Offensichtlich muss für jede Wahrscheinlichkeitsfunktion $f(x)$ *gelten, dass*

$$\sum_{x_i \in \mathcal{T}} f(x_i) = 1 \quad \textit{und} \quad f(x) \geq 0 \quad \textit{für alle } x \in \mathbb{R}$$

Die Wahrscheinlichkeitsfunktion $f(x_i)$ *heißt auch* **Wahrscheinlichkeitsdichte**. *Die* **Verteilungsfunktion** *einer diskreten Zufallsvariable ist definiert als*

$$F(x) = P(X \leq x) = \sum_{i:x_i \leq x} f(x_i)$$

Kennt man also für alle Werte von x den Wert $f(x)$, so kennt man auch $F(x)$ (dies gilt auch umgekehrt).

Eigenschaften der Verteilungsfunktion:

- $F(x)$ ist monoton wachsend ("Treppenfunktion")
- $F(x)$ ist stückweise konstant mit Sprungstellen an allen Elementen $x_i \in \mathcal{T}$ (d.h. $f(x_i) > 0$)
- $\lim\limits_{x \to \infty} F(x) = 1$
- $\lim\limits_{x \to -\infty} F(x) = 0$

Beispiel 4.1.2 (4-maliger unabh. Münzwurf mit einer fairen Münze) *Sei* X *die Zufallsvariable "Anzahl Kopf". Der Träger ist dann* $\mathcal{T} = \{0, 1, 2, 3, 4\}$*. Für die Wahrscheinlichkeitsfunktion ergibt sich:*

$$\begin{array}{rcl} f(0) & = & \left(\frac{1}{2}\right)^4 = 1/16 \\ f(1) & = & 4/16 \\ f(2) & = & 6/16 \\ f(3) & = & 4/16 \\ f(4) & = & 1/16 \end{array} \qquad \Rightarrow \qquad F(x) = \begin{cases} 0 & : & x < 0 \\ 1/16 & : & 0 \leq x < 1 \\ 5/16 & : & 1 \leq x < 2 \\ 11/16 & : & 2 \leq x < 3 \\ 15/16 & : & 3 \leq x < 4 \\ 1 & : & x \geq 4 \end{cases}$$

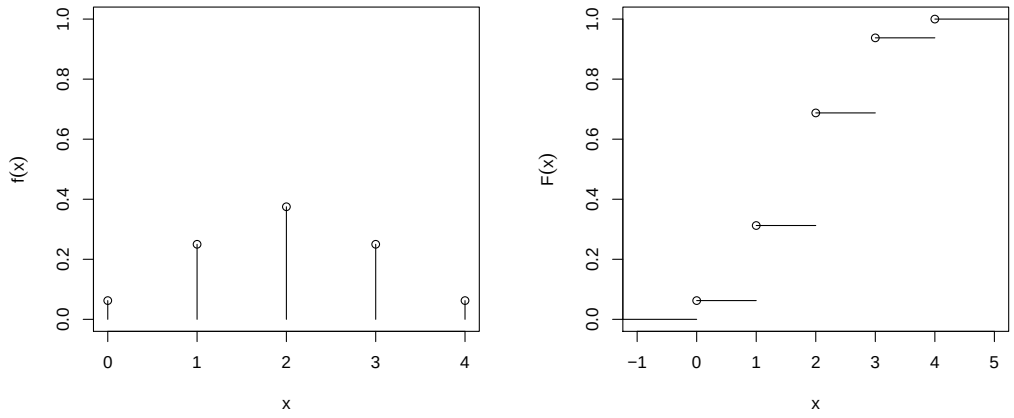

Abbildung 4.1: Wahrscheinlichkeitsfunktion (links) und Verteilungsfunktion (rechts) für den viermaligen Münzwurf

Man unterscheidet nun bestimmte "gängige" Verteilungen, um verschiedene Zufallsprozesse zu modellieren. Diese Verteilungen hängen von Parametern ab. Das einfachste Beispiel ist die **Bernoulli-Verteilung**. Eine Bernoulli-verteilte Zufallsvariable kann nur die Werte 0 und 1 annehmen:

$$\begin{array}{rcrcr} P(X=1) & = & f(1) & = & \pi \\ P(X=0) & = & f(0) & = & 1-\pi \end{array}$$

$\pi \in [0,1]$ ist der Parameter der Bernoulli-Verteilung (symbolisch $X \sim \mathcal{B}(\pi)$). Die Verteilungsfunktion lautet:

$$F(x) = \begin{cases} 0 & : & x < 0 \\ 1-\pi & : & 0 \leq x < 1 \\ 1 & : & x \geq 1 \end{cases}$$

Die **diskrete Gleichverteilung** hat den endlichen Träger $\mathcal{T} = \{x_1, x_2, \ldots, x_k\}$, wobei für $i = 1, \ldots, k$ gilt:

$$P(X = x_i) = f(x_i) = \frac{1}{k}$$

Häufig beinhaltet der Träger $\mathcal{T}$ alle natürlichen Zahlen zwischen a, $b \in \mathbb{N}$. Die Grenzen a und b sind dann die Parameter der Verteilung.

Interessanter ist die **geometrische Verteilung**:
Ein Zufallsvorgang, bei dem mit einer Wahrscheinlichkeit $\pi \in [0,1]$ ein Ereignis A eintritt, wird unabhängig voneinander sooft wiederholt, bis zum ersten

Mal A eintritt.
Sei X die Zufallsvariable "Anzahl der Versuche bis zum ersten Mal A eintritt". Dann ist der Träger von X gleich $\mathbb{N}$ und die Wahrscheinlichkeitsfunktion lautet:

$$P(X = x) = f(x) = \underbrace{(1-\pi)^{x-1}}_{(x-1)\text{-mal } \bar{A}} \cdot \underbrace{\pi}_{1\text{-mal } A} \qquad x = 1, 2, \ldots$$

π ist der Parameter der geometrischen Verteilung (symbolisch $X \sim \mathcal{G}(\pi)$).
Die Verteilungsfunktion von X lautet:

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= 1 - P(X > x) \\ &= 1 - (1-\pi)^x \end{aligned}$$

$\rightarrow$ Modell für (diskrete) Lebenszeiten und Wartezeiten.

Eine Variante der geometrischen Verteilung ist es, die Zufallsvariable $Y :=$ "Anzahl der Versuche *bevor* das erste mal A eintritt" zu betrachten. Offensichtlich gilt $Y = X - 1$, sodass Y Träger und Wahrscheinlichkeitsfunktion

$$\begin{aligned} \mathcal{T} &= \{0, 1, 2, \ldots\} \\ f(y) &= (1-\pi)^y \cdot \pi \end{aligned}$$

besitzt.
Die Wahrscheinlichkeitsfunktionen (Dichten) und Verteilungsfunktionen gängiger Verteilungen sind in R implementiert. Für die geometrische Verteilung liefert etwa

- `dgeom()` die Wahrscheinlichkeitsfunktion/-dichte
- `pgeom()` die Verteilungsfunktion und
- `rgeom()` Zufallszahlen aus der geometrischen Verteilung.

Definition 4.1.2
*Sei X eine diskrete Zufallsvariable mit Verteilungsfunktion $F(x)$, $x \in \mathbb{R}$. Sei $p \in [0, 1]$. Das p-**Quantil** x_p der Verteilung von X ist definiert als der* kleinste *Wert x für den gilt:*

$$F(x) \geq p$$

Somit gilt $P(X \leq x_p) = F(x_p) \geq p$ und daher "$x_p = F^{-1}(p)$" ("Inversion der Verteilungsfunktion").
*Das 0.5-Quantil der Verteilung wird **Median** x_{med} genannt.*

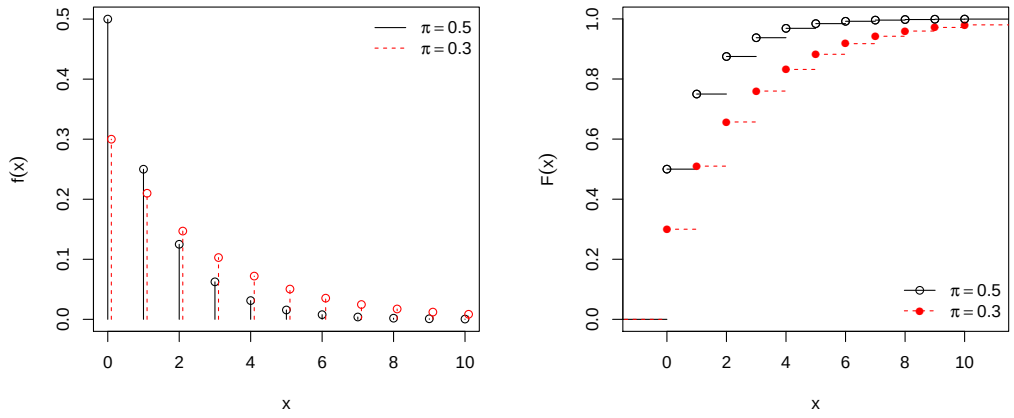

Abbildung 4.2: Vergleich der geometrischen Wahrscheinlichkeitsfunktionen (links) und Verteilungsfunktionen (rechts) für die beiden Parameter $\pi = 0.3$ und $\pi = 0.5$

Beispiel 4.1.3 (Quantile der geometrischen Verteilung)

- *p-Quantil x_p: In einem Anteil p aller Fälle muss man maximal x_p Versuche unternehmen,* bevor *zum ersten mal A eintritt.*

- *In R berechnet die Funktion* `qgeom()` *die Quantilfunktion der geometrischen Verteilung. Etwa das* 0.95*-Quantil und den Median für* $\pi = 0.01$ *und* $\pi = 0.9$*:*

```
> qgeom(c(0.5,0.95), prob = 0.01)

[1]  68 298

> qgeom(c(0.5,0.95), prob = 0.9)

[1] 0 1
```

4.2 Unabhängigkeit von diskreten Zufallsvariablen

Definition 4.2.1
Seien X und Y zwei Zufallsvariablen auf dem Wahrscheinlichkeitsraum (Ω,P) mit den Trägern $\mathcal{T}_X = \{x_1, x_2, \ldots\}$ und $\mathcal{T}_Y = \{y_1, y_2, \ldots\}$ und Wahrscheinlichkeitsfunktionen $f_X(x)$ und $f_Y(y)$.
Die Funktion

$$f_{X,Y}(x,y) = P(X = x \text{ und } Y = y) = P(X = x, Y = y)$$

heißt **gemeinsame Wahrscheinlichkeitsfunktion** *von X und Y.*
X und Y heißen **unabhängig**, *falls für alle $x \in \mathcal{T}_\mathcal{X}$ und $y \in \mathcal{T}_\mathcal{Y}$ gilt:*

$$\begin{aligned} f_{X,Y}(x,y) &= P(X = x) \cdot P(Y = y) \\ &= f_X(x) \cdot f_Y(y) \end{aligned}$$

Allgemeiner kann man die Unabhängigkeit von n Zufallsvariablen $X_1, X_2, \ldots, X_n$ wie folgt definieren:

Definition 4.2.2
$X_1, X_2, \ldots, X_n$ heißen **unabhängig**, *falls*

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

für alle $x_1, x_2, \ldots, x_n$ aus den entsprechenden Trägern gilt.

Definition 4.2.3
Sind die Zufallsvariablen $X_1, X_2, \ldots, X_n$ unabhängig und Bernoulli-verteilt mit Parameter π, so heißt $X = X_1, X_2, \ldots, X_n$ **Bernoulli-Folge**.

Beispiel 4.2.1 (Bernoulli-Folge)
Betrachten wir eine Bernoulli-Folge der Länge $n = 3$ mit dem Parameter $\pi = \frac{1}{6}$. Wegen der Unabhängigkeit gilt z. B.

$$P(X_1 = 1, X_2 = 0, X_3 = 0) = \frac{1}{6} \cdot \left(\frac{5}{6}\right)^2 = \frac{25}{216}$$

Meist interessiert aber nur die Anzahl $X = \sum_{i=1}^{n} X_i$, wie oft in der Bernoulli-Kette $X_i = 1$ aufgetreten ist.

Diese Zufallsvariable X heißt **binomialverteilt** (symbolisch $X \sim \mathcal{B}(n, \pi)$) mit Parameter $n \in \mathbb{N}$, $\pi \in [0, 1]$ und hat den Träger $\mathcal{T} = \{0, 1, \ldots, n\}$ sowie die Wahrscheinlichkeitsfunktion:

$$P(X = x) = f(x) = \binom{n}{x} \cdot \pi^x (1 - \pi)^{n-x} \text{ für } x \in \mathcal{T}$$

Es gilt $\mathcal{B}(1, \pi) = \mathcal{B}(\pi)$.

Funktionen in R: `dbinom()`, `pbinom()`, `rbinom()` und `qbinom()`.

Beispiel 4.2.2 (Würfeln)
Sei X: "Anzahl 6-er bei n-maligem Würfeln". Dann ist $X \sim \mathcal{B}(n, 1/6)$. Berechne

- ***Modus*** *(wahrscheinlichster Wert) und*
- *Median*

in Abhängigkeit von n.

Berechnung in R:

```
> pi <- 1/6
> modmed <- matrix(0, nrow = 10, ncol = 2)
> colnames(modmed) <- c("Modus", "Median")
> rownames(modmed) <- 1:nrow(modmed)
> for (n in 1:nrow(modmed)) {
+     traeger <- c(0:n)
+     dichte <- dbinom(traeger, prob = pi, size = n)
+     modmed[n, "Modus"] <- traeger[which.max(dichte)]
+     modmed[n, "Median"] <- qbinom(0.5, prob = pi, size = n)
+ }
> modmed

   Modus Median
1      0      0
2      0      0
3      0      0
4      0      1
5      0      1
6      1      1
```

```
7       1       1
8       1       1
9       1       1
10      1       2
```

Im **Urnenmodell** befinden sich N Kugeln in einer Urne, davon M markierte. Nun wird eine Stichprobe aus n Kugeln *mit Zurücklegen* gezogen. Sei X die Anzahl der markierten Kugeln in der Stichprobe. Dann gilt: $X \sim \mathcal{B}(n, M/N)$. Häufig wird jedoch *ohne Zurücklegen* gezogen, d. h. die Wahrscheinlichkeiten ändern sich von Ziehung zu Ziehung.
Die Verteilung von X nennt man dann **hypergeometrisch** (symbolisch $X \sim \mathcal{H}(n, N, M)$). Sie hat den Träger

$$\mathcal{T} = \{\max(0, n - (N - M)), \ldots, \min(n, M)\}$$

und die Wahrscheinlichkeitsfunktion

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad \text{für } x \in \mathcal{T}.$$

Funktionen in R: `{dpqr}hyper()`.

Ist N relativ "groß" und n "klein", so kann man die hypergeometrische Verteilung gut durch die Binomialverteilung approximieren (siehe Abbildung 4.5):

$$\mathcal{H}(n, N, M) \quad \approx \quad \mathcal{B}\left(n, \pi = \frac{M}{N}\right)$$

Beispiel 4.2.3 (Capture-Recapture-Experiment)
Das Ziel ist hierbei die Schätzung der Anzahl N von Individuen in einer Population. Dazu werden zunächst M Individuen markiert und mit der Gesamtpopulation zufällig vermischt. Anschließend wird eine Stichprobe ohne Zurücklegen vom Umfang n gezogen und die Anzahl $X = x$ an markierten Individuen beobachtet. Somit ist die Zufallsvariable X hypergeometrisch mit Parametern N, M und n.

$$X \sim \mathcal{H}(n, N, M)$$

*Die Aufgabe der Statistik ist es nun, einen **Schätzer** $\hat{N}$ für die Anzahl N der Individuen in der Population zu konstruieren. Ein naiver Ansatz zur Konstruktion eines **Schätzers** $\hat{N}$ für N lautet:*

$$\frac{N}{M} \quad \approx \quad \frac{n}{x} \qquad \Rightarrow \qquad \hat{N} \quad = \quad \frac{n}{x} \cdot M$$

Dieser Ansatz ist aber aus verschiedenen Gründen problematisch. Zunächst ist $\hat{N} = \infty$ für $x = 0$. Weiterhin ist im Allgemeinen $\hat{N} \notin \mathbb{N}$. Außerdem kann keine Angabe zur Genauigkeit der Schätzung gegeben werden.
In einem späteren Abschnitt werden wir statistische Verfahren kennenlernen, die diese Probleme lösen.

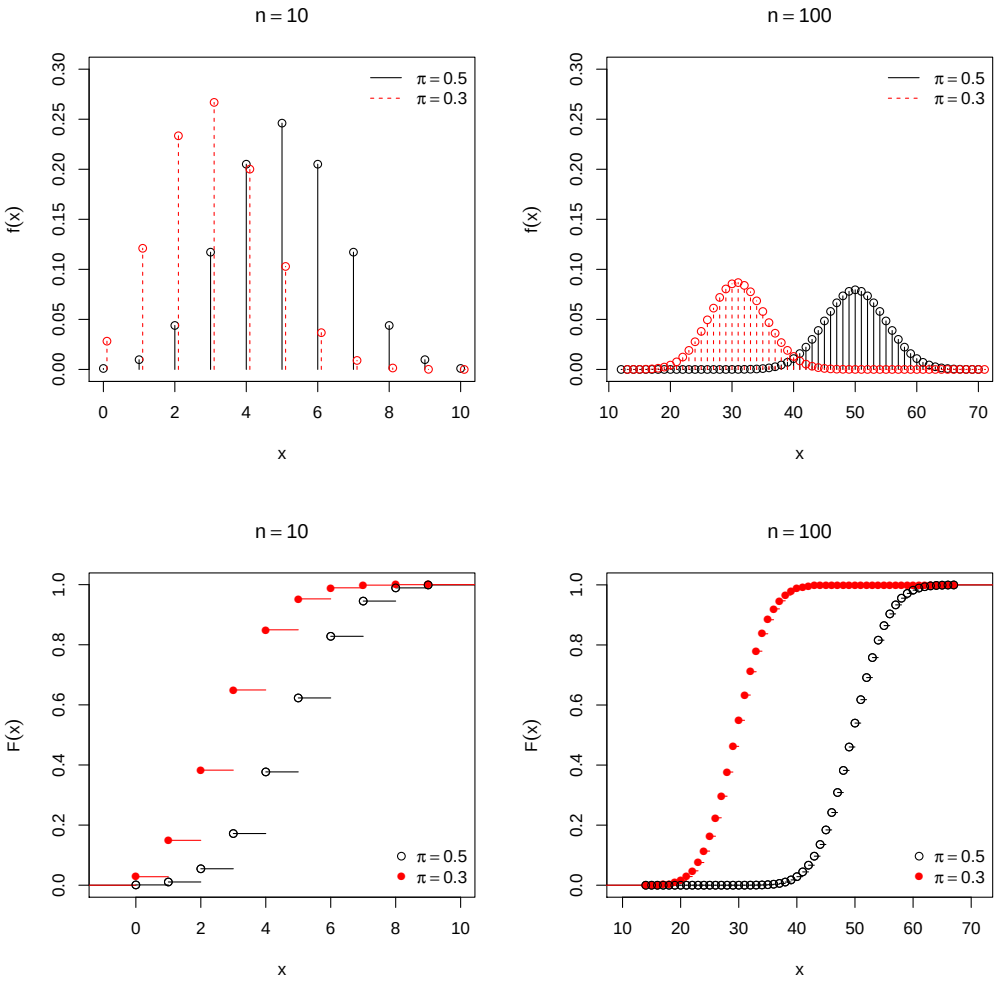

Abbildung 4.3: Vergleich der binomialen Wahrscheinlichkeitsfunktionen (oben) und Verteilungsfunktionen (unten) für $X \sim \mathcal{B}(n, \pi)$ mit $n = 10$ und $n = 100$, jeweils für $\pi = 0.5$ und $\pi = 0.3$

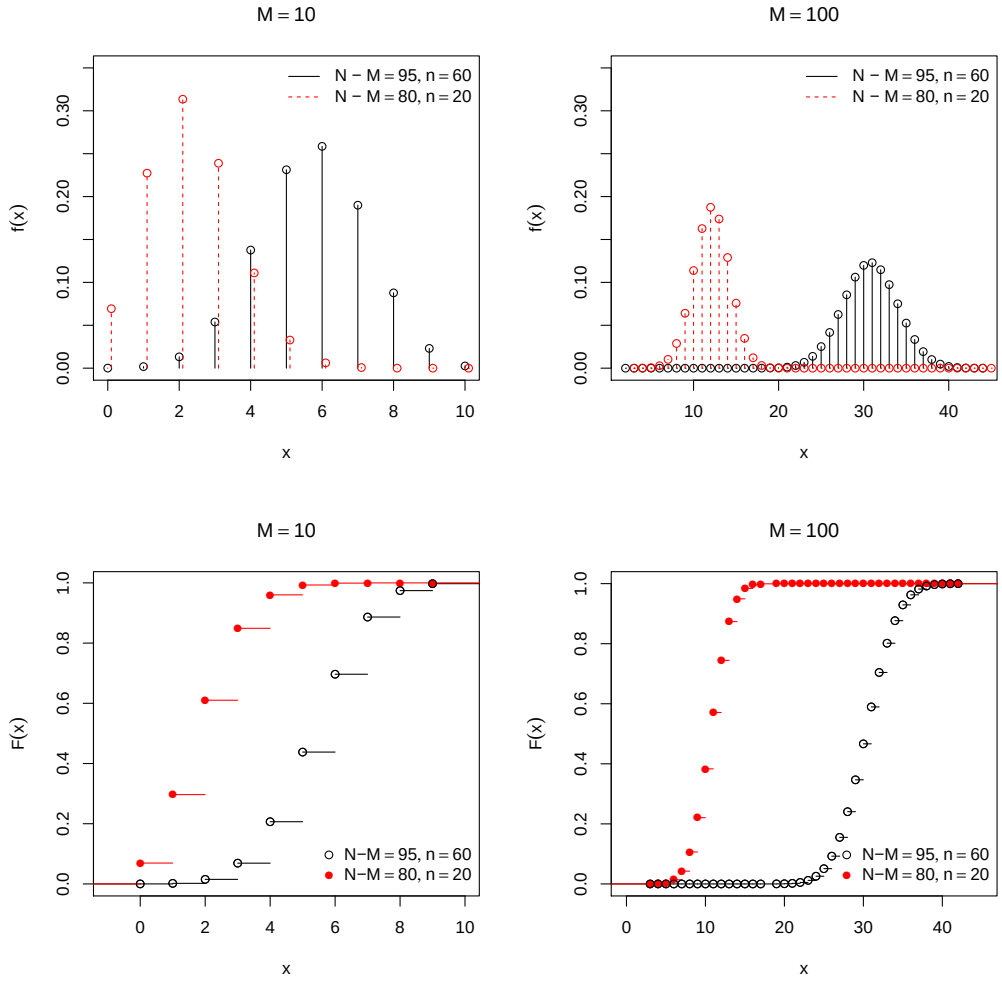

Abbildung 4.4: Vergleich der hypergeometrischen Wahrscheinlichkeitsfunktionen (oben) und der Verteilungsfunktionen (unten) für $X \sim \mathcal{H}(n, N, M)$ mit $M = 10$ (links) und $M = 100$ (rechts) und jeweils $N - M = 80$, $n = 20$ und $N - M = 95$, $n = 60$

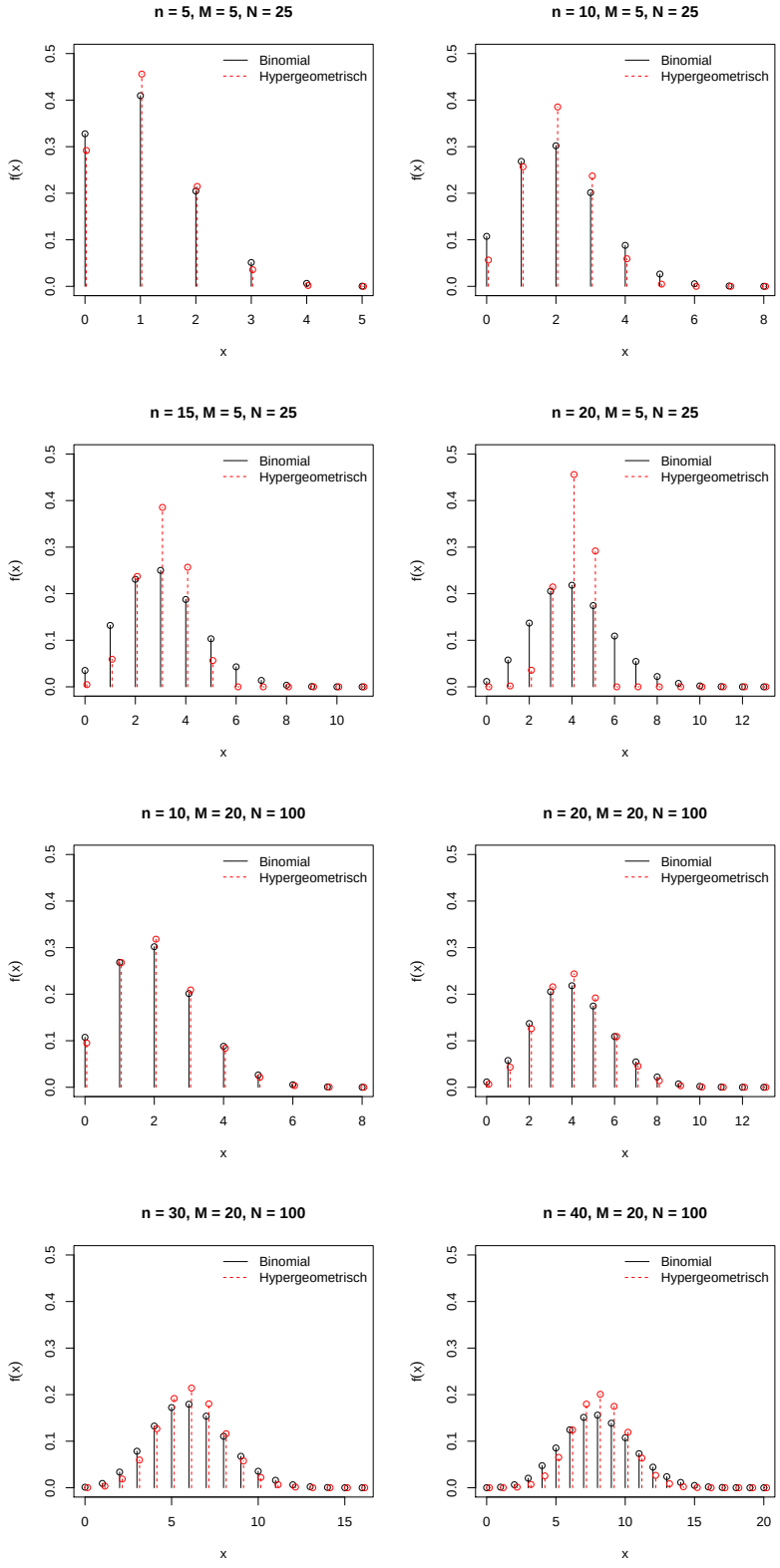

Abbildung 4.5: Vergleich der hypergeometrischen und binomialen Wahrscheinlichkeitsfunktionen

4.3 Die Poisson-Verteilung

Häufig gibt es zufällige Ereignisse, bei denen es keine natürliche obere Grenze für die Anzahl an Ereignissen gibt, z.B. die Anzahl an Telefonanrufen in einem "Call-Center" pro Stunde. Klassisches Beispiel ist eine Untersuchung zur Anzahl von Todesfällen durch Hufschlag in der Preußischen Armee (L. von Bortkiewicz, 1893). Die einfachste Verteilung für solche Phänomene ist die **Poisson-Verteilung** (symbolisch $X \sim \mathcal{P}(\lambda)$) mit Träger $\mathcal{T} = \mathbb{N}_0$ und Wahrscheinlichkeitsfunktion

$$f(x) = \frac{\lambda^x}{x!} \cdot \exp(-\lambda)$$

Der Parameter $\lambda \in \mathbb{R}^+$ der Verteilung reflektiert die **Rate** oder **Intensität**, mit der die Ereignisse in dem zugrundeliegenden Zeitintervall eintreffen.

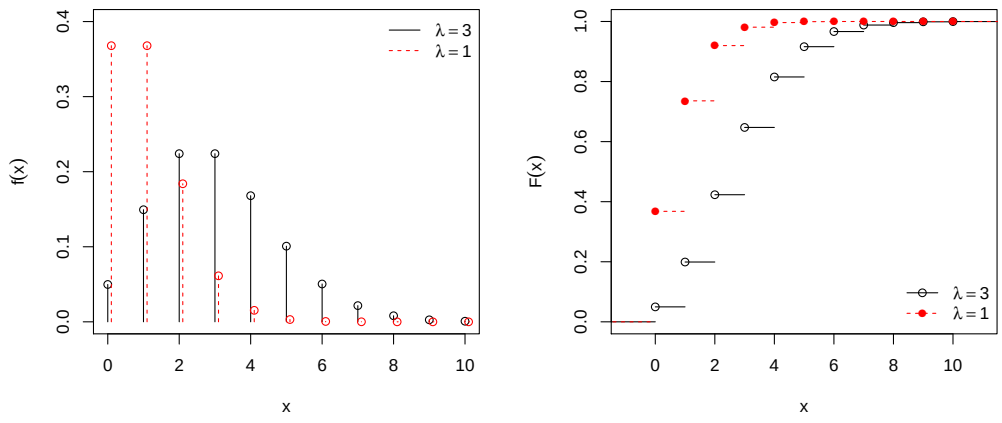

Abbildung 4.6: Vergleich der Wahrscheinlichkeitsfunktionen (links) und Verteilungsfunktionen (rechts) für eine Poissonverteilte Zufallsvariable mit dem Parameter $\lambda = 1$ bzw. $\lambda = 3$

Die Binomialverteilung $\mathcal{B}(n, \pi)$ kann für "großes n" und "kleines π" mit $n \cdot \pi = const.$ gut durch die Poisson-Verteilung mit $\lambda = n \cdot \pi$ approximiert werden.

$$\mathcal{B}(n, \pi) \quad \approx \quad \mathcal{P}(\lambda = n \cdot \pi)$$

Dies ist auch in den Abbildungen 4.7 und 4.8 zu sehen, welche Binomialverteilung und Poissonverteilung für unterschiedliche Werte von n und π zeigen. Je größer n ist und je kleiner π, desto besser ist die Approximation.

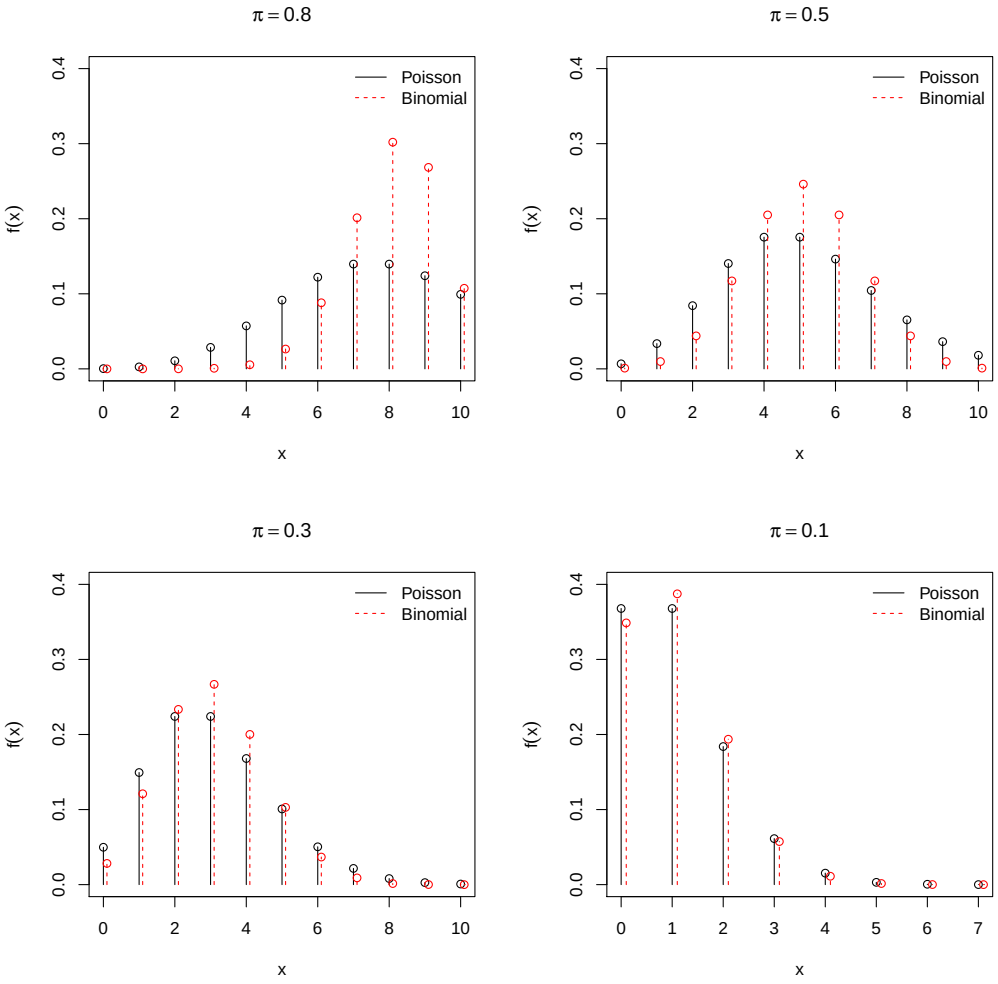

Abbildung 4.7: Vergleich der Wahrscheinlichkeitsfunktionen von Binomialverteilung und Poissonverteilung für $n = 10$ und für $\pi = (0.1, 0.3, 0.5, 0.8)$

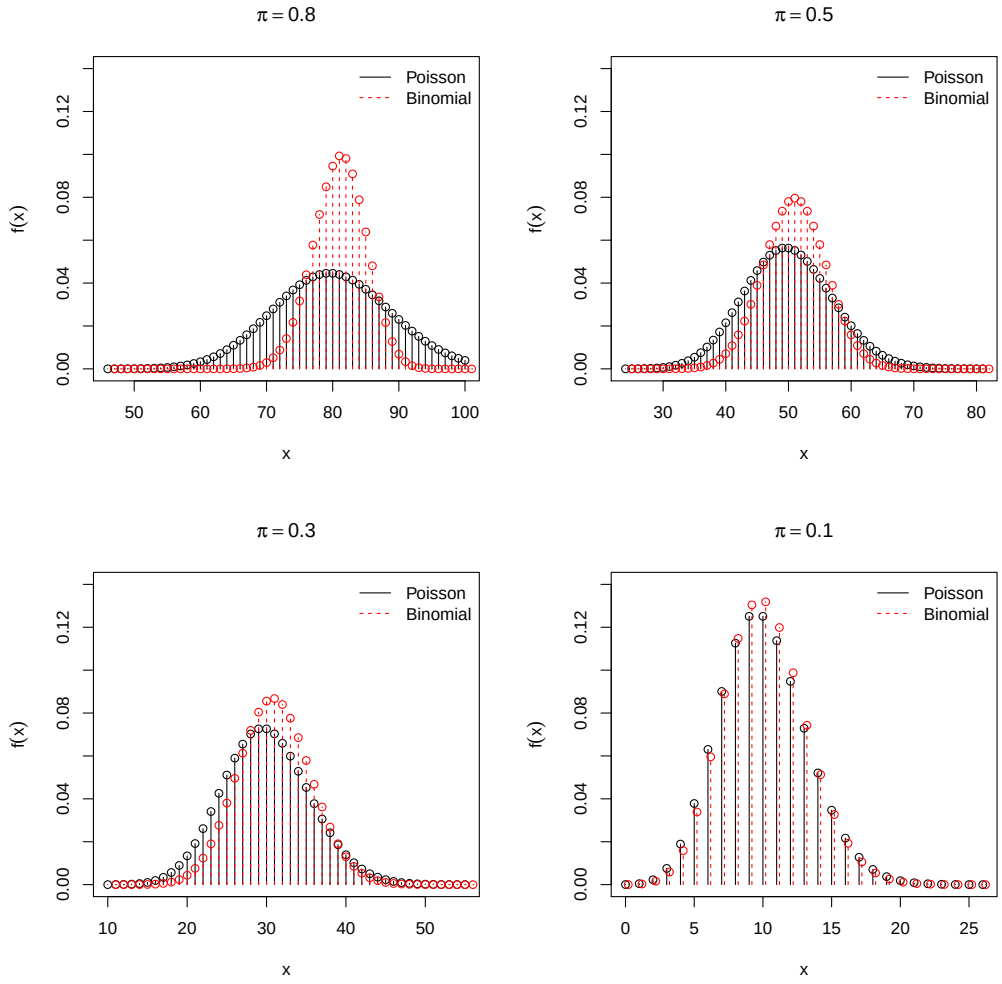

Abbildung 4.8: Vergleich der Wahrscheinlichkeitsfunktionen von Binomialverteilung und Poissonverteilung für $n = 100$ und für $\pi = (0.1, 0.3, 0.5, 0.8)$

4.4 Faltungen

Definition 4.4.1
Seien X und Y unabhängige Zufallsvariablen mit Wahrscheinlichkeitsfunktionen $f_X(x)$ und $f_Y(y)$. Sei $Z = X + Y$. Dann nennt man die Wahrscheinlichkeitsfunktion f_Z von Z die ***Faltung*** *von f_X und f_Y:*

$$\begin{aligned} f_Z(z) = P(X+Y=z) &= \sum_x P(X=x, X+Y=z) \\ &= \sum_x P(X=x, Y=z-X) \\ &= \sum_x P(X=x) \cdot P(Y=z-x|X=x) \\ &\overset{unabh.}{=} \sum_x P(X=x) \cdot P(Y=z-x) \\ &= \sum_x f_X(x) \cdot f_Y(z-x) \\ &= \sum_y f_X(z-y) \cdot f_Y(y) \end{aligned}$$

Beispiel 4.4.1 (Faltung von poissonverteilten Zufallsvariablen)
Seien X und Y zwei unabhängige poissonverteilte Zufallsvariablen mit $X \sim \mathcal{P}(\lambda_1)$ und $Y \sim \mathcal{P}(\lambda_2)$. Dann hat $Z = X + Y$ Wahrscheinlichkeitsfunktion

$$\begin{aligned} f_Z(z) &= \sum_{x=0}^{z} f_X(x) \cdot f_Y(z-x) \\ &= \sum_{x=0}^{z} \left[\frac{\lambda_1^x}{x!} \cdot \exp(-\lambda_1) \cdot \frac{\lambda_2^{z-x}}{(z-x)!} \cdot \exp(-\lambda_2) \right] \\ &= \underbrace{\sum_{x=0}^{z} \left[\frac{\lambda_1^x \cdot \lambda_2^{z-x}}{x!\,(z-x)!} \right]}_{\overset{!}{=} \frac{(\lambda_1+\lambda_2)^z}{z!}} \cdot \exp\left(-(\lambda_1+\lambda_2)\right) \end{aligned}$$

dies gilt weil:

$$\begin{aligned} \sum_{x=0}^{z} \frac{\lambda_1^x \cdot \lambda_2^{z-x}}{x!\,(z-x)!} &= \frac{1}{z!} \cdot \sum_{x=0}^{z} \binom{z}{x} \lambda_1^x \cdot \lambda_2^{z-x} \\ &= \frac{(\lambda_1+\lambda_2)^z}{z!} \quad \underbrace{\cdot \sum_{x=0}^{z} \binom{z}{x} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x \cdot \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{z-x}}_{= 1,\ \text{da Summe über W'keitsfkt der Bin.Vtlg}} \end{aligned}$$

Somit erhält man:

$$f_Z(z) = \frac{(\lambda_1 + \lambda_2)^z}{z!} \cdot \exp\left(-(\lambda_1 + \lambda_2);\right.$$

$Z = X + Y$ ist also ebenfalls poissonverteilt mit Parameter $(\lambda_1 + \lambda_2)$: $Z \sim \mathcal{P}(\lambda_1 + \lambda_2)$.

Definition 4.4.2
Sei X die Summe von n unabhängigen geometrischen Zufallsvariablen $X_1, \ldots, X_n$:

$$X = X_1 + \ldots + X_n$$

Dann hat X eine **negative Binomialverteilung** *mit Parametern $n \in \mathbb{N}$ und $\pi \in (0, 1)$ und Wahrscheinlichkeitsfunktion*

$$f(x) = \binom{x-1}{n-1} \pi^n (1-\pi)^{x-n} \text{ für } x = n, n+1, \ldots$$

Funktionen in R: `{dpqr}nbinom`.
Beachte: Unterschiedliche Definition in R! Träger immer gleich $\mathbb{N}_0$

Beispiel 4.4.2 (Faltung von zwei geometrischen Zufallsvariablen)
Seien X und Y zwei unabhängige geometrische Zufallsvariablen mit $X \sim \mathcal{G}(\pi)$ und $Y \sim \mathcal{G}(\pi)$. Dann hat $Z = X + Y$ Wahrscheinlichkeitsfunktion

$$\begin{aligned}
f_Z(z) &= \sum_{x=1}^{z-1} f_X(x) \cdot f_Y(z-x) \\
&= \sum_{x=1}^{z-1} \pi(1-\pi)^{x-1} \cdot \pi(1-\pi)^{z-x-1} \\
&= \pi^2 \sum_{x=1}^{z-1} (1-\pi)^{[(x-1)+(z-x-1)]} \\
&= \pi^2 \sum_{x=1}^{z-1} (1-\pi)^{z-2} \\
&= \pi^2 (z-1)(1-\pi)^{z-2}.
\end{aligned}$$

Z ist also ***negativ binomialverteilt*** *mit Parametern $n = 2$ und π.*

4.5 Die Verteilung von Zufallsvektoren

Seien X und Y zwei diskrete Zufallsvariablen, welche auf dem Wahrscheinlichkeitsraum (Ω, P) definiert sind. Es stellt sich die Frage, wie man Information über deren gemeinsames stochastisches Verhalten quantifizieren könnte. Dazu betrachtet man den **Zufallsvektor** (X, Y) als Abbildung von $\mathbb{R}^2$ nach $[0, 1]$.
Wie bereits zuvor definiert, lautet die **gemeinsame Wahrscheinlichkeitsfunktion** von X und Y folgendermaßen:

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

Definition 4.5.1
Die **gemeinsame Verteilungsfunktion** *zweier Zufallsvariablen X und Y ist*

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

Die **gemeinsame Verteilung** von X und Y enthält i. A. mehr Information, als in den **Randverteilungen** $f_X(x)$ und $f_Y(y)$ steckt. Diese lassen sich leicht berechnen, wenn die gemeinsame Wahrscheinlichkeitsfunktion bekannt ist:

$$\begin{aligned} f_X(x) &= \sum_y f_{X,Y}(x, y) \\ f_Y(y) &= \sum_x f_{X,Y}(x, y) \end{aligned}$$

Beispiel 4.5.1 (Münzwurf)
Ein Lehrer bittet seine Schüler, eine (faire) Münze zweimal zu werfen, und das Ergebnis ("Kopf" = 0, "Zahl" = 1) für jeden Wurf zu notieren. Sei X das Ergebnis des ersten Wurfes und Y das Ergebnis des zweiten Wurfes.
Ein gewissenhafter Schüler folgt genau den Anweisungen des Lehrers und notiert das Ergebnis X_G und Y_G. Ein fauler Schüler wirft nur eine Münze und notiert das erzielte Ergebnis zweimal: X_F und Y_F.
Berechne die gemeinsame Wahrscheinlichkeitsfunktion von (X_G, Y_G) und von (X_F, Y_F):

Gewissenhafter Schüler:

$f_{X_G,Y_G}(X,Y)$	$Y_G = 0$	$Y_G = 1$	$f_{X_G}(x)$
$X_G = 0$	1/4	1/4	1/2
$X_G = 1$	1/4	1/4	1/2
$f_{Y_G}(y)$	1/2	1/2	

Fauler Schüler:

$f_{X_F,Y_F}(X,Y)$	$Y_F=0$	$Y_F=1$	$f_{X_F}(x)$
$X_F=0$	1/2	0	1/2
$X_F=1$	0	1/2	1/2
$f_{Y_F}(y)$	1/2	1/2	

Festzuhalten bleibt, dass die Randverteilungen von X_G und X_F und auch die Randverteilungen von Y_G und Y_F identisch sind, nicht aber die gemeinsame Verteilung von (X_G, Y_G) und von (X_F, Y_F).

Allgemeiner kann man einen Zufallsvektor $\mathbf{X} = (X_1, \ldots, X_n)$ der Dimension n betrachten. Dieser hat dann die Wahrscheinlichkeitsfunktion

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

und die folgenden Randverteilungen

$$f_{X_i}(x_i) = \sum_{x_j : j \neq i} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

Definition 4.5.2
Für zwei Zufallsvariablen X und Y ist die **bedingte Wahrscheinlichkeitsfunktion** *von X, gegeben $Y = y$, definiert als:*

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Die **bedingte Verteilungsfunktion** *von X, gegeben $Y = y$, lautet:*

$$F_{X|Y}(x|y) = P(X \leq x|Y = y) = \frac{P(X \leq x, Y = y)}{P(Y = y)}$$

Bedingte Wahrscheinlichkeitsfunktion und bedingte Verteilungsfunktion sind definiert für alle y mit $P(Y = y) > 0$.

Aus dieser Definition folgt, dass man die gemeinsame Wahrscheinlichkeitsfunktion zweier Zufallsvariablen X und Y durch ein Produkt aus bedingter Wahrscheinlichkeitsfunktion und Randverteilung ausdrücken kann:

$$\begin{aligned} f_{X,Y}(x, y) &= f_{X|Y}(x|y) \cdot f_Y(y) \\ &= f_{Y|X}(y|x) \cdot f_X(x) \end{aligned}$$

Daher kann man die **Unabhängigkeit** von Zufallsvariablen auch anders definieren:
X und Y sind genau dann unabhängig wenn $\forall x, y$ gilt:

$$f_{X|Y}(x|y) = f_X(x) \text{ oder } f_{Y|X}(y|x) = f_Y(y)$$

Bedingte Verteilungen haben dieselben Eigenschaften wie gewöhnliche (unbedingte) Verteilungen, wie z. B. :

$$\sum_{X \in \mathcal{T}_{\mathcal{X}}} f_{X|Y}(x|y) = 1 \qquad \forall y \in \mathcal{T}_{\mathcal{Y}}$$

Beispiel 4.5.2
Gegeben sei folgende gemeinsame Wahrscheinlichkeitsfunktion:

$f_{X,Y}(x,y)$	$Y=-1$	$Y=0$	$Y=2$
$X=1$	$1/18$	$3/18$	$2/18$
$X=2$	$2/18$	0	$3/18$
$X=3$	0	$4/18$	$3/18$

Es sollen die Randverteilungen $f_X(x)$ und $f_Y(y)$ sowie die bedingten Verteilungen $f_{X|Y}(x|y)$ und $f_{Y|X}(y|x)$ berechnet werden. Anschließend werden X und Y auf Unabhängigkeit untersucht.

$f_{X,Y}(x,y)$	$Y=-1$	$Y=0$	$Y=2$	$f_X(x)$
$X=1$	$1/18$	$3/18$	$2/18$	$6/18$
$X=2$	$2/18$	0	$3/18$	$5/18$
$X=3$	0	$4/18$	$3/18$	$7/18$
$f_Y(y)$	$3/18$	$7/18$	$8/18$	

Zum Beispiel sind

$$f_{X|Y}(x|y=-1) = \begin{cases} 1/3 & \text{für } x=1 \\ 2/3 & \text{für } x=2 \\ 0 & \text{für } x=3 \end{cases} \qquad f_{Y|X}(y|x=1) = \begin{cases} 1/6 & \text{für } y=-1 \\ 1/2 & \text{für } y=0 \\ 1/3 & \text{für } y=2 \end{cases}$$

X und Y sind nicht *unabhängig, da $\forall x \in \mathcal{T}_x$ und $\forall y \in \mathcal{T}_y$ gelten müsste, dass $f_{X,Y}(x,y) = f_Y(y) \cdot f_X(x)$. Dies gilt aber nicht, da z. B. :*

$$f_{X,Y}(X=3, Y=-1) = 0 \neq \frac{21}{18^2} = f_X(3) \cdot f_Y(-1)$$

Beispiel 4.5.3
Betrachte zwei unabhängige Zufallsvariablen X und Y, die beide poissonverteilt sind mit Parameter λ bzw. μ. Sei $Z = X + Y$.
Man zeige: Die bedingte Verteilung von $X|Z = z$ ist binomialverteilt mit Parametern $n = z$ und $\pi = \lambda/(\lambda + \mu)$:

$$X|Z = z \quad \sim \quad \mathcal{B}(z, \pi = \lambda/(\lambda + \mu))$$

Gesucht wird also:

$$f_{X|Z}(x|z) = \frac{f_{X,Z}(x,z)}{f_Z(z)}$$

Da Z die Faltung zweier poissonverteilter Zufallsvariablen ist, gilt $Z \sim \mathcal{P}(\lambda+\mu)$.
Die Wahrscheinlichkeitsfunktion von Z ist:

$$f_Z(z) = \frac{(\lambda+\mu)^z}{z!}\exp(-(\lambda+\mu))$$

Für die gemeinsame Verteilung von X und Z erhält man:

$$\begin{aligned} f_{X,Z}(x,z) &= f_{X,Y}(x,z-x) \overset{X,Y \text{ unabh.}}{=} f_X(x) f_Y(z-x) \\ &= \frac{\lambda^x}{x!}\exp(-\lambda)\cdot\frac{\mu^{(z-x)}}{(z-x)!}\exp(-\mu) \end{aligned}$$

Wie ist $Z|X$ verteilt? Wir wissen, dass $Z - X|X \sim Y|X \overset{X,Y \text{ unabh.}}{\sim} Y \sim \mathcal{P}(\mu)$. Daher ist $Z|X = x \ \sim \ \mathcal{P}(\mu) + x$. Somit gilt:

$$\begin{aligned} f_{X|Z}(x|z) &= \frac{f_{X,Z}(x,z)}{f_Z(z)} \\ &= \frac{\frac{\mu^{(z-x)}}{(z-x)!}\exp(-\mu)\cdot\frac{\lambda^x}{x!}\exp(-\lambda)}{\frac{(\lambda+\mu)^z}{z!}\exp(-(\lambda+\mu))} \\ &= \binom{z}{x}\frac{\lambda^x\mu^{(z-x)}}{(\lambda+\mu)^z} \\ &= \binom{z}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x \underbrace{\left(\frac{\mu}{\lambda+\mu}\right)}_{1-\frac{\lambda}{\lambda+\mu}}^{z-x} \end{aligned}$$

$$\Rightarrow \quad X|Z = z \ \sim \mathcal{B}(z, \lambda/(\lambda+\mu))$$

Beispiel 4.5.4 (Gedächtnislosigkeit der geometrischen Verteilung)
Sei X eine geometrisch verteilte Zufallsvariable, d.h. $X \sim \mathcal{G}(\pi)$. X beschreibt also die Anzahl der Versuche, die ausgeführt werden müssen, bis *ein Ereignis A eintritt. Nun wird für $k = 1, 2, \ldots$ die Zufallsvariable Y_k definiert als die Anzahl der noch auszuführenden Versuche für X bis A eintritt, wenn bereits bekannt ist, dass in den letzten k Versuchen das Ereignis nicht eingetreten ist. Es gilt also:*

$$Y_k = X - k | X > k$$

Wie lautet die Verteilung von Y_k?

$$\begin{aligned} P(Y_k = y) &= P(X - k = y | X > k) \\ &= P(X = y + k | X > k) \\ &= \frac{P(X = y + k, X > k)}{P(X > k)} \\ &\overset{y \geq 0}{=} \frac{P(X = y + k)}{P(X > k)} \\ &= \frac{\pi(1-\pi)^{(y+k)-1}}{P(X > k)} \end{aligned}$$

Wir benötigen noch die Wahrscheinlichkeit $P(X > k)$:

$$\begin{aligned} P(X > k) &= 1 - P(X \leq k) \\ &= 1 - F(k) \\ &= 1 - [1 - (1-\pi)^k] \\ &= (1-\pi)^k \end{aligned}$$

Dieses Ergebnis wird in obige Formel eingesetzt:

$$\begin{aligned} P(Y_k = y) &= \frac{\pi(1-\pi)^{(y+k)-1}}{(1-\pi)^k} \\ &= \pi(1-\pi)^{y-1} \end{aligned}$$

Somit ergibt sich, dass auch Y_k geometrisch verteilt ist mit Parameter π, d.h. X und Y_k besitzen die gleiche Verteilung, unabhängig von k. Die geometrische Verteilung kann sich also nicht "merken", wie oft schon ein Misserfolg aufgetreten ist. Diese Eigenschaft der geometrischen Verteilung nennt man ***Gedächtnislosigkeit****.*

Abschließend wollen wir nun eine Verallgemeinerung der Binomialverteilung kennenlernen, die **Multinomialverteilung**. Ein Experiment, bei dem eins von drei möglichen Ereignissen mit Wahrscheinlichkeit π_1, $\pi_2, \ldots, \pi_k$ ($\pi_1 + \pi_2 + \cdots + \pi_k = 1$) auftritt, wird unabhängig voneinander n-mal wiederholt. Sei $\mathbf{X}$ ein drei-dimensionaler Zufallsvektor, dessen i-te Komponente angibt, wie oft das i-te Ereignis eingetreten ist. Dann nennt man $\mathbf{X}$ multinomialverteilt.

Definition 4.5.3
Ein k-dimensionaler diskreter Zufallsvektor $\mathbf{X}$ heißt **multinomialverteilt**, *falls er Träger $\mathcal{T} = \{\boldsymbol{x} = (x_1, x_2, \ldots, x_k) : x_i \in \{0, 1, \ldots, n\}$ und $x_1 + x_2 +$*

$\cdots + x_k = n\}$ *und Wahrscheinlichkeitsfunktion*

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_k) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \pi_i^{x_i}$$

besitzt.
Man schreibt kurz: $\mathbf{X} \sim \mathcal{M}_k(n, \pi = (\pi_1, \pi_2, \ldots, \pi_k))$*; hierbei steht* $\mathcal{M}_3$ *für* **Multinomialverteilung** *der Dimension 3. Man kann zeigen, dass für die Randverteilungen von* $\mathbf{X}$ *gilt:* $X_i \sim \mathcal{B}(n, \pi_i), i = 1, \ldots, k$.

Beispiel 4.5.5
In einer Population mit Häufigkeiten π_1, π_2 *und* π_3 *der Genotypen aa, ab und bb wird eine Stichprobe vom Umfang* n *gezogen. Die Anzahlen* X_1, X_2 *und* X_3 *der drei Genotypen ist dann* **trinomialverteilt**.

Kapitel 5

Erwartungswerte, Varianzen und Kovarianzen

Zur Charakterisierung von Verteilungen unterscheidet man **Lageparameter**, wie z. B.

- **Erwartungswert** ("mittlerer Wert")
- **Modus** (Maximum der Wahrscheinlichkeitsfunktion, d. h. , wahrscheinlichster Wert)
- **Median** (0.5-Quantil : mindestens 50 % der Wahrscheinlichkeitsmasse liegt über diesem Punkt, und 50 % darunter; u. U. kompliziert zu berechnen) schwer zu berechnen und nicht eindeutig)

und **Streuungsparameter**:

- **Varianz** (mittlere quadratische Abweichung)
- **Standardabweichung** (Wurzel aus der Varianz)
- mittlere absolute Abweichung

5.1 Der Erwartungswert einer diskreten Zufallsvariable

Definition 5.1.1
*Der **Erwartungswert** $E(X) = EX$ einer diskreten Zufallsvariable X mit Träger $\mathcal{T}$ ist definiert als*

$$E(X) = \sum_{x \in \mathcal{T}} x \cdot P(X = x) = \sum_{x \in \mathcal{T}} x \cdot f(x)$$

wenn diese Summe absolut konvergent ist. Den Erwartungswert eines Zufallsvektors definiert man als den Vektor der Erwartungswerte der einzelnen Komponenten.

Man könnte die Summe in obiger Definition auch über alle $x \in \mathbb{R}$ laufen lassen, da für $x \notin \mathcal{T}$ gilt: $f(x) = 0$, also letztlich nur abzählbar viele Summanden ungleich Null sind.

Beispiel 5.1.1 (Erwartungswert der Bernoulliverteilung)
Für $X \sim \mathcal{B}(\pi)$ gilt:

$$f(x) = \begin{cases} 1 - \pi & \text{für } x = 0 \\ \pi & \text{für } x = 1 \end{cases}$$

Daher gilt für den Erwartungswert:

$$E(X) = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi$$

Beispiel 5.1.2 (Erwartungswert der Poissonverteilung)
Für $X \sim \mathcal{P}(\lambda)$ gilt:

$$f(x) = \frac{\lambda^x}{x!} \exp(-\lambda) \qquad \text{für } x \in \mathbb{N}_0$$

Der Erwartungswert lautet daher

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \cdot \underbrace{\frac{\lambda^x}{x!} \exp(-\lambda)}_{f(x)} = \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \exp(-\lambda) \\ &= \sum_{x=0}^{\infty} \frac{\lambda^{(x+1)}}{x!} \exp(-\lambda) = \lambda \underbrace{\left(\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \right)}_{\exp(\lambda)} \exp(-\lambda) \\ &= \lambda \end{aligned}$$

Eigenschaften des Erwartungswertes:

1. Sei $X = a$ mit Wahrscheinlichkeit 1, d.h. $P(X = a) = 1$. Dann heißt X **deterministische Zufallsvariable**. Es gilt:
$$\mathrm{E}X = a$$

2. *Linearität des Erwartungswertes*
Seien $a, b \in \mathbb{R}$ und X, Y beliebige Zufallsvariablen. Dann gilt:
$$\mathrm{E}(a \cdot X + b \cdot Y) = a \cdot \mathrm{E}X + b \cdot \mathrm{E}Y$$
Allgemeiner gilt dann auch:
$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mathrm{E}(X_i) \qquad \text{für } a_i \in \mathbb{R} \text{ und beliebige Zufallsvariablen } X_i$$
Natürlich gilt auch $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$, denn man kann Y als deterministische Zufallsvariable mit Träger $\mathcal{T}_Y = \{1\}$ auffassen.

Anwendung für die Linearität des Erwartungswertes:

Der Erwartungswert einer binomialverteilten Zufallsvariable X, d.h. $X \sim \mathcal{B}(n, \pi)$, muss $\mathrm{E}X = n \cdot \pi$ sein, da X darstellbar ist als Summe von n unabhängigen bernoulliverteilten Zufallsvariable $X_i \sim \mathcal{B}(\pi)$ (mit $i = 1, \ldots, n$), wobei jede Zufallsvariable den Erwartungswert $\mathrm{E}(X_i) = \pi$ hat:
$$\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) \;=\; \sum_{i=1}^{n} \mathrm{E}(X_i) \;=\; \sum_{i=1}^{n} \pi \;=\; n \cdot \pi$$

Satz 5.1.1 (Erwartungswert von Zufallsvariablen mit Träger $\mathbb{N}$)
Hat X den Träger $\mathcal{T} = \mathbb{N}$, so gilt:
$$E(X) = \sum_{k=1}^{\infty} P(X \geq k)$$

Beweis:
$$\begin{aligned}
\sum_{k=1}^{\infty} P(X \geq k) &= \sum_{k=1}^{\infty} \sum_{t=k}^{\infty} P(X = t) \\
&= \sum_{t=1}^{\infty} \sum_{k=1}^{t} P(X = t) \\
&= \sum_{t=1}^{\infty} t \cdot P(X = t) \\
&= \mathrm{E}X
\end{aligned}$$

Beispiel 5.1.3 (Erwartungswert der geometrischen Verteilung)
Sei $X \sim \mathcal{G}(\pi)$, dann gilt

$$f(x) = \pi \cdot (1-\pi)^{x-1} \text{ für } x \in \mathbb{N}$$

und damit

$$P(X \geq k) = (1-\pi)^{k-1}$$

Unter Anwendung von Satz 5.1.1 kann man den Erwartungswert ausrechnen:

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} P(X \geq k) \\ &= \sum_{k=1}^{\infty} (1-\pi)^{k-1} \\ &= \sum_{k=0}^{\infty} (1-\pi)^{k} \\ &\overset{geom.Reihe}{=} \frac{1}{1-(1-\pi)} \\ &= \frac{1}{\pi} \end{aligned}$$

Satz 5.1.2 (Transformationsregel für Erwartungswerte)
Sei X eine diskrete Zufallsvariable und $g(X)$ eine reelle Funktion. Dann gilt für $Y = g(X)$:

$$E(Y) \;=\; E(g(X)) \;=\; \sum_{x \in \mathcal{T}} g(x)\, f(x)$$

Sei (X, Y) ein Zufallsvektor *aus den Zufallsvariablen X und Y mit gemeinsamer Wahrscheinlichkeitsfunktion $f_{X,Y}(x, y)$ und sei $g(x, y)$ eine reellwertige Funktion. Dann gilt für $Z = g(X, Y)$:*

$$E(Z) = E(g(X,Y)) = \sum_x \sum_y g(x,y)\; f_{X,Y}(x,y)$$

Speziell gilt daher:

$$E(X \cdot Y) = \sum_x \sum_y x \cdot y \cdot f_{X,Y}(x,y)$$

Man beachte, dass im Allgemeinen nur bei linearen Funktionen g gilt: $\mathrm{E}(g(X)) = g(\mathrm{E}(X))$.

Beispiel 5.1.4
Sei X eine Zufallsvariable mit Wahrscheinlichkeitsfunktion

$$f(x) = \begin{cases} 1/4 & \text{für} \quad x = -2 \\ 1/8 & \text{für} \quad x = -1 \\ 1/4 & \text{für} \quad x = 1 \\ 3/8 & \text{für} \quad x = 3 \end{cases}$$

Dann ergibt sich der Erwartungswert von $E(X^2)$ zu:

$$\begin{aligned} E(X^2) &= \sum_{x \in \mathcal{T}_x} x^2 \cdot f(x) \\ &= (-2)^2 \cdot \frac{1}{4} + (-1)^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{3}{8} \\ &= 4\frac{3}{4} \end{aligned}$$

Alternativ könnte man zunächst die Wahrscheinlichkeitsfunktion von $Y = X^2$ berechnen,

$$f(y) = \begin{cases} (1/4 + 1/8) = 3/8 & \text{für} \quad y = 1 \\ 1/4 & \text{für} \quad y = 4 \\ 3/8 & \text{für} \quad y = 9 \end{cases}$$

und dann den Erwartungswert von Y direkt bestimmen:

$$\begin{aligned} E(Y) &= \sum_{y \in \mathcal{T}_y} y \cdot f(y) \\ &= 1 \cdot \frac{3}{8} + 4 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} \\ &= 4\frac{3}{4} \end{aligned}$$

Man beachte, dass

$$4\frac{3}{4} = E(X^2) \neq E(X)^2 = \left(\frac{3}{4}\right)^2 = \frac{9}{16}$$

.

Definition 5.1.2
*Das **arithmetische Mittel** $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ mit X_i $(i = 1, \ldots, n)$ unabhängig und identisch verteilte Zufallsvariablen aus einer Verteilung mit Erwartungswert μ und Varianz σ^2 besitzt folgende Eigenschaften:*

$$E(\bar{X}_n) = \mu$$

und

$$Var(\bar{X}_n) = \frac{1}{n}\sigma^2$$

Daher folgt sofort für $n \to \infty$*:* $\bar{X}_n \to \mu$ *und* $Var(\bar{X}_n) \to 0$
$\Rightarrow$ *Das arithmetische Mittel konvergiert gegen den Erwartungswert. Mehr dazu in Kapitel 8.4.*

5.2 Varianz und Standardabweichung

Definition 5.2.1
Die ***Varianz*** $V(X)$*, auch* $Var(X)$*, einer diskreten Zufallsvariable ist definiert als:*

$$Var(X) = E[(X - EX)^2]$$

"Erwartete quadratische Abweichung vom Erwartungswert"

Satz 5.2.1 (Verschiebungssatz)
Zur einfacheren Berechnung der Varianz kann der Verschiebungssatz angewendet werden:

$$Var(X) = E(X^2) - [E(X)]^2$$

Beweis:

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[(X - \mathrm{E}X)^2] \\ &= \mathrm{E}[X^2 - 2X\mathrm{E}X + (\mathrm{E}X)^2] \\ &= \mathrm{E}X^2 - \mathrm{E}(2X\mathrm{E}X) + \mathrm{E}((\mathrm{E}X)^2) \\ &= \mathrm{E}X^2 - 2\mathrm{E}X\mathrm{E}X + (\mathrm{E}X)^2 \\ &= \mathrm{E}X^2 - (\mathrm{E}X)^2 \end{aligned}$$

Eigenschaften von Varianzen:

1. $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$ für alle $a, b \in \mathbb{R}$
   Dies gilt weil:

$$\begin{aligned} \mathrm{Var}(aX + b) &= \mathrm{E}\left[((aX + b) - \mathrm{E}(aX + b))^2\right] \\ &= \mathrm{E}\left[(a(X - \mathrm{E}X))^2\right] \\ &= \mathrm{E}\left[a^2(X - \mathrm{E}X)^2\right] \\ &= a^2\mathrm{E}\left[(X - \mathrm{E}X)^2\right] \\ &= a^2 \cdot \mathrm{Var}(X) \end{aligned}$$

2. Sind X und Y unabhängig, so gilt:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \tag{5.1}$$

   "Die Varianz der Summe entspricht der Summe der Varianzen"

Als Streuungsparameter sind Varianzen noch nicht auf der richtigen Skala, denn sie geben ja die mittlere *quadratische* Abweichung wieder! Daher definiert man die **Standardabweichung**:

Definition 5.2.2
*Die **Standardabweichung** einer diskreten Zufallsvariable X ist definiert als die Wurzel aus der Varianz:*

$$\sigma = \sigma(X) = \sqrt{Var(X)}$$

Im Gegensatz zur Varianz gilt für die Standardabweichung:

$$\sigma(aX + b) = |a|\sigma(X) \qquad \text{für alle } a, b \in \mathbb{R}$$

Bemerkung:
Die mittlere absolute Abweichung $\mathrm{E}(|X - \mathrm{E}X|)$ erscheint intuitiver, ist aber deutlich schwerer mathematisch zu handhaben.

Beispiel 5.2.1 (Varianz der Bernoulli-Verteilung)
Sei $X \sim \mathcal{B}(\pi)$. Wir wissen $EX = \pi$. Ferner ist

$$\begin{aligned} E(X^2) &= 0^2 \cdot f(0) + 1^2 \cdot f(1) \\ &= 0 \cdot (1 - \pi) + 1 \cdot \pi \\ &= \pi \end{aligned}$$

Daher:

$$\begin{aligned} Var(X) &= EX^2 - (EX)^2 \\ &= \pi - \pi^2 \\ &= \pi(1 - \pi) \end{aligned}$$

Daher lautet die Varianz für eine binomialverteilte Zufallsvariable $Y \sim \mathcal{B}(n, \pi)$:

$$\mathrm{Var}(Y) = n \cdot \pi \cdot (1 - \pi),$$

da sie ja die Summe von n unabhängigen Bernoulliverteilten Zufallsvariablen $Y_1, \ldots, Y_n$, jeweils mit $\mathrm{Var}(Y_i) = \pi(1 - \pi)$, ist. Verwende dazu die Gleichung (5.1).

Als Maß für die Streuung einer Verteilung ist die Varianz bzw. Standardabweichung einer Zufallsvariable X schwer direkt zu interpretieren. Es gilt aber folgender Satz:

Satz 5.2.2 (Ungleichung von Tschebyscheff)

$$P(|X - E(X)| \geq c) \leq \frac{Var(X)}{c^2}$$

Beispiel 5.2.2
Sei $E(X)$ beliebig und $Var(X) = 1$. Dann ist

$$\begin{aligned} P(|X - E(X)| \geq 1) &\leq 1 \\ P(|X - E(X)| \geq 2) &\leq \frac{1}{4} \\ P(|X - E(X)| \geq 3) &\leq \frac{1}{9} \end{aligned}$$

Definition 5.2.3
Stichprobenvarianz: *Seien X_i $(i = 1, \ldots, n)$ unabhängig und identisch verteilte Zufallsvariablen aus einer Verteilung mit Erwartungswert μ und Varianz σ^2. Dann gilt für*

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2,$$

dass $E(S^2) = \sigma^2$ und für $n \to \infty$: $S^2 \to \sigma^2$
Wenn μ unbekannt ist und durch $\bar{X}_n$ ersetzt werden muss, so gilt allerdings $E(S^2) \neq \sigma^2$.
Für $S^{2*} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ gilt die Gleichheit wieder.

5.3 Kovarianz und Korrelation

Definition 5.3.1
Als Maß für die lineare stochastische Abhängigkeit von zwei Zufallsvariablen X und Y definiert man die ***Kovarianz****:*

$$Cov(X,Y) = E[(X - EX)(Y - EY)]$$

und die ***Korrelation***

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{VarX} \cdot \sqrt{VarY}}$$

unter der Voraussetzung, dass $Var(X) > 0$ und $Var(Y) > 0$ gilt.

Für die einfachere Berechnung der Kovarianz kann man auch den Verschiebungssatz anwenden:

Satz 5.3.1 (Verschiebungssatz für die Kovarianz)

$$Cov(X,Y) = E(XY) - EX \cdot EY$$

Beweis:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)] \\ &= \mathrm{E}[XY - X\mathrm{E}Y - Y\mathrm{E}X + \mathrm{E}X\mathrm{E}Y] \\ &= \mathrm{E}(XY) - \mathrm{E}X \cdot \mathrm{E}Y \end{aligned}$$

Beachte: $\mathrm{E}(XY)$ kann mit dem Transformationssatz für Erwartungswerte leicht über die gemeinsame Wahrscheinlichkeitsfunktion $f_{XY}(x,y)$ von X und Y berechnet werden.

Beispiel 5.3.1
Die vorliegende Wahrscheinlichkeitsfunktion ist die gleiche wie im Beispiel 4.5.2. Es wurde festgestellt, dass die Zufallsvariablen X und Y stochastisch abhängig sind. Nun soll die Kovarianz $Cov(X,Y)$ und die Korrelation ρ berechnet werden.

$f_{X,Y}(x,y)$	$Y=-1$	$Y=0$	$Y=2$	$f_X(x)$
$X=1$	$1/18$	$3/18$	$2/18$	$6/18$
$X=2$	$2/18$	0	$3/18$	$5/18$
$X=3$	0	$4/18$	$3/18$	$7/18$
$f_Y(y)$	$3/18$	$7/18$	$8/18$	

Für die Berechnung der Kovarianz werden folgende Werte benötigt:

$$\begin{aligned} E(X \cdot Y) &= \left(1 \cdot (-1) \cdot \frac{1}{18}\right) + \left(1 \cdot 0 \cdot \frac{3}{18}\right) + \ldots = \frac{29}{18} \\ E(X) &= 1 \cdot \frac{6}{18} + 2 \cdot \frac{5}{18} + 3 \cdot \frac{7}{18} = \frac{37}{18} \\ E(Y) &= (-1) \cdot \frac{3}{18} + 0 \cdot \frac{7}{18} + 2 \cdot \frac{8}{18} = \frac{13}{18} \end{aligned}$$

Es ergibt sich:

$$Cov(X,Y) = \frac{29}{18} - \frac{37}{18} \cdot \frac{13}{18} = \frac{41}{324}$$

Für die Korrelation benötigt man noch

$$\begin{aligned} E(X^2) &= 1 \cdot \frac{6}{18} + 4 \cdot \frac{5}{18} + 9 \cdot \frac{7}{18} = \frac{89}{18} \\ E(Y^2) &= 1 \cdot \frac{3}{18} + 0 \cdot \frac{7}{18} + 4 \cdot \frac{8}{18} = \frac{35}{18} \end{aligned}$$

woraus sich

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 = \frac{233}{324} \\ Var(Y) &= E(Y^2) - E(Y)^2 = \frac{461}{324} \end{aligned}$$

und schließlich

$$\rho = \frac{41}{\sqrt{233 \cdot 461}} \approx 0.125$$

ergibt.

Definition 5.3.2
Zwei Zufallsvariablen X und Y heißen ***unkorreliert****, wenn*

$$Cov(X,Y) = 0 \quad bzw. \;\; \rho(X,Y) = 0$$

d.h. wenn gilt:

$$E(X \cdot Y) = EX \cdot EY$$

X und Y sind ***positiv/negativ korreliert*** *, falls gilt:*

$$\rho(X,Y) > 0 \; bzw. \;\; \rho(X,Y) < 0$$

Das heißt für größere Werte von X erhält man eher größere/kleinere Werte von Y.

Aus der Unabhängigkeit zweier Zufallsvariablen folgt deren Unkorreliertheit aber der Umkehrschluss gilt i. A. nicht!

Beweis:
“ $\Rightarrow$”

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_x \sum_y x \cdot y \, f_{X,Y}(x,y) \\
&\overset{\text{unabh.}}{=} \sum_x \sum_y x \cdot y \, f_X(x) \, f_Y(y) \\
&= \sum_x x \, f_X(x) \sum_y y \, f_Y(y) \\
&= \mathrm{E}X \cdot \mathrm{E}Y
\end{aligned}$$

“ $\not\Leftarrow$”
Gegenbeispiel:
Seien $X \sim \mathcal{B}(\pi = \frac{1}{2})$ und $Y \sim \mathcal{B}(\pi = \frac{1}{2})$ unabhängig.
Betrachte:

$$Z_1 = X + Y = \begin{cases} 0 & \text{mit W'keit } \frac{1}{4} \\ 1 & \text{mit W'keit } \frac{1}{2} \\ 2 & \text{mit W'keit } \frac{1}{4} \end{cases}$$

$$Z_2 = X - Y = \begin{cases} -1 & \text{mit W'keit } \frac{1}{4} \\ 0 & \text{mit W'keit } \frac{1}{2} \\ 1 & \text{mit W'keit } \frac{1}{4} \end{cases}$$

Die gemeinsame Wahrscheinlichkeitsfunktion von Z_1 und Z_2 lautet:

$f_{Z_1,Z_2}(z_1,z_2)$	$z_1 = 0$	$z_1 = 1$	$z_1 = 2$	$f_{Z_2}(z_2)$
$z_2 = -1$	0	1/4	0	1/4
$z_2 = 0$	1/4	0	1/4	1/2
$z_2 = 1$	0	1/4	0	1/4
$f_{Z_1}(z_1)$	1/4	1/2	1/4	

Klarerweise gilt nicht:

$$f(z_1, z_2) = f(z_1) \cdot f(z_2)$$

für alle z_1 und z_2, d. h. Z_1, Z_2 sind nicht unabhängig.

Aber

$$\left.\begin{array}{lcl} \mathrm{E}(Z_1) & = & 1 \\ \mathrm{E}(Z_2) & = & 0 \\ \mathrm{E}(Z_1 Z_2) & = & 0 \end{array}\right\} \Rightarrow \quad \mathrm{Cov}(Z_1, Z_2) = 0 = \mathrm{E}(Z_1) \cdot \mathrm{E}(Z_2)$$

Also sind Z_1 und Z_2 unkorreliert.

□

Während die Kovarianz nicht leicht zu interpretieren ist, ist dies leichter für die Korrelation, da für alle Zufallsvariablenn X, Y gilt:

$$-1 \leq \rho(X, Y) \leq 1$$

Dies folgt aus der **Cauchy-Schwarzschen Ungleichung**:

$$[\mathrm{E}(XY)]^2 \leq \mathrm{E}(X^2) \cdot \mathrm{E}(Y^2)$$

Das heißt, dass die Korrelation in gewisser Weise normiert ist.
$|\rho(X, Y)| = 1$ gilt genau dann, wenn perfekte lineare Abhängigkeit zwischen X und Y besteht:

$$Y = a + b \cdot X$$

für bestimmte a und $b \in \mathbb{R}$ mit $b \neq 0$. Ist $b > 0$ so ist $\rho(X, Y) = 1$. Ist $b < 0$ so ist $\rho(X, Y) = -1$.
Eigenschaften von Kovarianzen:
Seien X, Y beliebige Zufallsvariablen und $a, b, c, d \in \mathbb{R}$, wobei $b \cdot d > 0$. Dann gilt:

1. $\mathrm{Cov}(a + bX, c + dY) = b \cdot d \cdot \mathrm{Cov}(X, Y)$
   Daher:

$$\begin{array}{rcl} \rho(a + bX, c + dY) & = & \dfrac{b\, d\, \mathrm{Cov}(X, Y)}{\sqrt{b^2\, \mathrm{Var}X}\sqrt{d^2\, \mathrm{Var}Y}} \\ & = & \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}X}\sqrt{\mathrm{Var}Y}} \\ & = & \rho(X, Y) \end{array}$$

   d.h. die Korrelation ist **invariant bzgl. linearer Transformationen**

2. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

3. Schließlich gilt für $X + Y$:

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{E}((X+Y) - \mathrm{E}(X+Y))^2 \\ &\overset{\text{lin. EWert}}{=} \mathrm{E}(X - \mathrm{E}X + Y - \mathrm{E}Y)^2 \\ &= \underbrace{\mathrm{E}(X-\mathrm{E}X)^2}_{\mathrm{Var}(X)} + \underbrace{\mathrm{E}(Y-\mathrm{E}Y)^2}_{\mathrm{Var}(Y)} + 2\,\underbrace{\mathrm{E}\left[(X-\mathrm{E}X)(Y-\mathrm{E}Y)\right]}_{\mathrm{Cov}(X,Y)}\end{aligned}$$

Insgesamt :

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X,Y)$$

Wie bereits erwähnt, gilt für unabhängige X und Y:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Definition 5.3.3
Seien (X_i, Y_i) $(i = 1, \ldots, n)$ unabhängig und identisch verteilte Zufallsvariablen aus Verteilungen mit Erwartungswerten μ_X, μ_Y und existierenden Varianzen. Dann gilt für

$$S^2_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y),$$

dass $E(S^2_{XY}) = Cov(X,Y)$ und für $n \to \infty$: $S^2_{XY} \to Cov(X,Y)$.
S^2_{XY} wird **empirische Kovarianz** *genannt.*

Kapitel 6

Elemente statistischer Inferenz

Ziel der Statistik ist es, unter bestimmten Modellannahmen Aussagen über unbekannte Parameter $\theta \in \Theta$ zu machen, nachdem Beobachtungen X gemacht wurden. Dabei unterscheidet man:

- **Punktschätzungen**: Was ist der "beste" Schätzwert $\hat{\theta}$ für den unbekannten Parameter θ?
- **Intervallschätzung**:

  $\rightarrow$ Angabe eines Vertrauensintervalls:
  - **Konfidenzintervalle** überdecken mit gewisser Sicherheit den unbekannten Parameter θ (bei hypothetischer vielfacher Wiederholung des Zufallsexperiments)
    θ fest, X zufällig
    $\rightarrow$ **frequentistischer** Wahrscheinlichkeitsbegriff
  - in einem **Kredibilitätsintervall** liegt der unbekannte Parameter mit einer gewissen Wahrscheinlichkeit
    θ zufällig, X fest
    $\rightarrow$ **subjektivistischer** Wahrscheinlichkeitsbegriff

Beispiel 6.0.2 (Binomialverteilung)
Sei $X \sim \mathcal{B}(n, \pi)$. Man beobachtet $X = 7$ bei $n = 10$ Versuchen. Was ist der "beste" Schätzer $\hat{\theta}$ für den unbekannten Parameter $\theta = \pi$? Hier gilt, dass $\theta \in [0, 1]$ stetig *ist.*

Beispiel 6.0.3 (Capture-Recapture Experiment)
Angenommen $M = 100$, $n = 50$ und $X = 33$. Was ist der "beste" Schätzer $\hat{N}$ für den unbekannten Parameter $\theta = N$? Hier gilt, dass
$\theta \in \Theta = \{N_{min}, N_{min} + 1, \ldots\}$ diskret *ist.*

6.1 Likelihood-Inferenz

Wir haben die Wahrscheinlichkeitsfunktion $f(x)$ einer Zufallsvariable X in Abhängigkeit von einem Parameter θ kennengelernt.

Beispielsweise ist $\theta = \pi$ im Binomialexperiment:

$$X \sim \mathcal{B}(n, \pi) \quad \Rightarrow \quad f(x) = f(x, \theta = \pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

Nun wird $f(x, \theta)$ als Funktion von $\theta \in \Theta$ für festes $X = x$ betrachtet. Die möglichen Werte des Parameters θ liegen im sogenannten Parameterraum Θ.

Definition 6.1.1

$$L(\theta) := f(x, \theta) \qquad \textit{heißt } \textbf{Likelihoodfunktion}$$

und

$$l(\theta) = \log L(\theta) \qquad \textit{heißt } \textbf{Loglikelihoodfunktion}$$

Je größer die (Log-)Likelihoodfunktion $L(\theta)$ bzw. $l(\theta)$ als Funktion von θ bei gegebenen Daten $X = x$ ist, desto "plausibler" ist der entsprechende Wert von θ. Daher definiert man:

Definition 6.1.2
Ein Schätzer $\hat{\theta}_{ML}$ für einen Parameter $\theta \in \Theta$ heißt **Maximum-Likelihood (ML)-Schätzer**, *wenn er die (Log-)Likelihoodfunktion maximiert:*

$$L(\hat{\theta}_{ML}) = \max L(\theta) \quad \theta \in \Theta$$

bzw.

$$l(\hat{\theta}_{ML}) = \max l(\theta) \quad \theta \in \Theta$$

Die Optimierung der (Log-)Likelihoodfunktion in Θ kann nun mit einem geeigneten Optimierungsverfahren durchgeführt werden. Eine Variante zur Berechnung des ML-Schätzers (bei Θ stetig!) ist, falls dies möglich ist, die Ableitung der Loglikelihoodfunktion gleich Null zu setzen.

Beispiel 6.1.1 (ML-Schätzer bei Binomialverteilung)

$$\begin{aligned} L(\theta = \pi) &= \binom{n}{x}\pi^x(1-\pi)^{n-x} \\ l(\pi) &= x\log\pi + (n-x)\log(1-\pi) \\ l'(\pi) &= \frac{x}{\pi} - \frac{n-x}{1-\pi} \stackrel{!}{=} 0 \\ \Rightarrow \quad & x(1-\pi) = (n-x)\pi \\ \Rightarrow \quad & \hat{\pi}_{ML} = \frac{x}{n} \end{aligned}$$

Der ML-Schätzer ist also gleich der relativen Häufigkeit.

Beispiel 6.1.2 (ML-Inferenz im Capture-Recapture-Experiment)
Im Capture-Recapture-Beispiel lautet die Wahrscheinlichkeitsfunktion:

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Dabei sind M und n bekannt. Bei beobachteter Stichprobe $X = x$ lautet die Likelihoodfunktion für den unbekannten Parameter N

$$L(\theta = N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

unter der Restriktion, dass $N \geq \max(M + n - x, n)$.
Im Unterschied zum vorhergehenden Abschnitt ist nun der unbekannte Parameter N ganzzahlig.
Man kann zeigen, dass für $x > 0$ und $\hat{N}_{naive} = \frac{Mn}{x} \notin \mathbb{N}$ gilt:

$$\hat{N}_{ML} = trunc\left\{\frac{Mn}{x}\right\} = trunc\left\{\hat{N}_{naive}\right\}$$

Im Fall $\hat{N}_{naive} \in \mathbb{N}$ ist der ML-Schätzer i. A. nicht eindeutig: dann wird die Likelihood sowohl durch $\hat{N}_{naive}$ als auch durch $\hat{N}_{naive} - 1$ maximiert, was auch beim folgenden Zahlenbeispielen beobachtet werden kann:

M	n	x	$\hat{N}_{ML}$	$\hat{N}_{naive}$
100	50	33	151	151.51
100	50	41	121	121.95
7	23	4	40	40.25
25	30	10	74 und 75	75
13	10	5	25 und 26	26
100	50	0	∞	∞
100	50	50	100	100

Den "naiven" Schätzer $\hat{N}_{naive}$ kann man als ML-Schätzer unter Verwendung der Binomialapproximation zur hypergeometrischen Verteilung herleiten.

Eine andere Möglichkeit zur Berechnung des ML-Schätzers bieten die Optimierungsfunktionen `optim()`(mehrdimensional) und `optimize()`(nur eindimensional) in R oder die Anwendung des **EM-Algorithmus**.
Der ML-Schätzer hat eine wichtige und nützliche Eigenschaft:

Satz 6.1.1 (Invarianz des ML-Schätzers)
Sei $\hat{\theta}_{ML}$ der ML-Schätzer für θ und sei $\varphi = \varphi(\theta)$ eine beliebige (eineindeutige) Funktion von θ. Dann ist der ML-Schätzer von φ:

$$\hat{\varphi}_{ML} = \varphi(\hat{\theta}_{ML})$$

Beispiel 6.1.3 (ML-Schätzer für Chance im Binomialexperiment)
Man kann Satz 6.1.1 dazu verwenden, um für die Chance $\gamma = \frac{\pi}{1-\pi}$ im Binomialexperiment einen ML-Schätzer zu berechnen:

$$\hat{\gamma}_{ML} = \frac{\hat{\pi}_{ML}}{1 - \hat{\pi}_{ML}} = \frac{\frac{x}{n}}{1 - \frac{x}{n}} = \frac{x}{n - x}$$

Dieses Ergebnis kann auch direkt (ohne Ausnutzung der Invarianzeigenschaft) gezeigt werden, jedoch ist die Rechnung etwas aufwendiger:
Sei $X \sim \mathcal{B}(n, \pi)$. Damit gilt $\hat{\pi}_{ML} = x/n$.

$$\begin{aligned} l(\pi) &= x \cdot \log(\pi) + (n - x) \log(1 - \pi) \\ l(\gamma) &= x \cdot \log\left(\frac{\gamma}{1+\gamma}\right) + (n - x) \log\left(\frac{1}{1+\gamma}\right) \\ &= x \log(\gamma) - n \log(1 + \gamma) \end{aligned}$$

Ableiten der Loglikelihood liefert das Ergebnis:

$$l'(\gamma) = \frac{x}{\gamma} - \frac{n}{1+\gamma} \stackrel{!}{=} 0 \Leftrightarrow \hat{\gamma}_{ML} = \frac{x}{n - x}$$

Wir sind nun daran interessiert, Aussagen darüber zu treffen, in welchen Teilmengen des Parameterraumes Θ sich der feste, aber unbekannte, Parameter $\theta \in \Theta$ mutmaßlich befindet.

Beispiel 6.1.4 (Ist eine Münze fair?)
Wir möchten feststellen, ob eine Münze unfair ist. Dazu wird die Münze n-mal geworfen und wir schätzen mittels $\hat{\pi} = x/n$ (mit x gleich der Anzahl "Kopf") per ML-Schätzung die Wahrscheinlichkeit π. Die Behauptung, das feste, wahre aber leider unbekannte π ist ungleich 0.5 läßt sich jetzt schreiben als $\Theta_1 = \Theta \setminus \{1/2\}$.

Wir unterteilen den Parameterraum Θ in zwei disjunkte Teilmengen Θ_0 und Θ_1 mit $\Theta = \Theta_0 \cup \Theta_1$, wobei $\Theta_0 \cap \Theta_1 = \emptyset$. Es ist nun eine Entscheidungsregel gesucht, für welche der beiden Zustände $\theta \in \Theta_0$ oder $\theta \in \Theta_1$ wir uns basierend auf einem Experiment (also Daten $\boldsymbol{X} = (X_1, \ldots, X_n)$) entscheiden sollen.

Definition 6.1.3
Θ_0 heißt Hypothese oder Null-Hypothese. Θ_1 heißt Alternative. Die Formulierung

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1$$

heißt Testproblem.

Eine Entscheidungsregel, um uns zwischen H_0 bzw. H_1 zu entscheiden, nennt man einen statistischen Test.

Definition 6.1.4
Eine Funktion $\psi : \mathbb{R}^n \to \{0, 1\}$ heißt Test für H_0 gegen H_1. Wenn $E(\psi) = P(\psi = 1) \leq \alpha, \alpha \in [0, 1]$, für alle $\theta \in \Theta_0$, heißt ψ Niveau-α-Test.

Es können jetzt folgende Situationen auftreten. Das wahre θ ist Element der Hypothese und $\psi(\boldsymbol{X} = (X_1, \ldots, X_n)) = 0$. Dann hat der Test die richtige Entscheidung getroffen. Umgekehrt, wenn θ Element der Alternative ist und $\psi(\boldsymbol{X}) = 1$, ist dies ebenfalls richtig. Wenn nun $\theta \in \Theta_0$ und $\psi(X) = 1$, ist dies eine fälschliche Entscheidung gegen H_0. Wenn ψ ein Niveau-α-Test ist, passiert dies aber nur mit Wahrscheinlichkeit kleiner gleich α. Dieser Fehler wird Fehler 1. Art genannt. Letztendlich, wenn $\theta \in \Theta_1$ und $\psi = 0$, heißt diese Fehlentscheidung für H_0 Fehler 2. Art. Die Wahrscheinlichkeit für dessen Auftreten ist *nicht* begrenzt.

Beispiel 6.1.5
Binomialverteilung $\boldsymbol{X} = (X_1, X_2, X_3, X_4), X_i \sim \mathcal{B}(\pi)$ mit Parameter $\pi \in \Theta$ unbekannt. Zu Testen sei das Testproblem

$$H_0 : \pi \leq \frac{1}{2} \text{ vs. } H_1 : \pi > \frac{1}{2}$$

Vorüberlegung: falls $X = \sum_i X_i$ zu groß, so spricht das gegen H_0. Was heißt "zu groß"? Sagen wir mal 4. Definiere den Test

$$\psi(\boldsymbol{X}) = \psi(X_1, \ldots, X_4) = I\left(\sum_i X_i = 4\right)$$

Für $\pi \in \Theta$ gilt $E(\psi(\boldsymbol{X}) = 1) = P(\sum_i X_i = 4) = \binom{4}{4}\pi^4(1-\pi)^0 = \pi^4$. Somit gilt für alle Parameter $\pi_0 \in \Theta_0$ aus der Hypothese $E_0(\psi(\boldsymbol{X}) = 1) = \pi_0^4 \leq 0.5^4 = 0.0625$. Somit ist ψ ein Niveau-α-Test für alle $\alpha \geq 0.0625$.

Achtung: Für $\pi_1 \in \Theta_1(\pi_1 > 1/2)$ gilt $E_1(\psi(\boldsymbol{X}) = 1) = \pi_1^4 \in [0.0625, 1]$ und damit ist der Fehler 2. Art $E_1(\psi(\boldsymbol{X}) = 0) = 1 - E_1(\psi(\boldsymbol{X}) = 1) \in [0, 0.9375]$ und ist somit nicht kontrollierbar.

Definition 6.1.5
Ein Test ψ ist von der Form

$$\psi(\boldsymbol{X}) = I(T(\boldsymbol{X}) \in \mathcal{C})$$

Die Funktion $T : \mathbb{R}^n \to \mathbb{R}$ heißt Teststatistik und die Menge $\mathcal{C}$ heißt kritische Region (Ablehnungsbereich), in der die Teststatistik liegen muss, um sich gegen die Null-Hypothese zu entscheiden.

Bleiben zwei Fragen zu klären. Welche Teststatistik T ist in einem speziellen Fall sinnvoll und was ist die kritische Region $\mathcal{C}$? Erstere Frage ist nicht allgemein zu beantworten und es gibt eine Vielzahl von Vorschlägen. Die kritische Region kann so bestimmt werden, dass ψ ein Niveau-α-Test ist. Es gilt:

$$E_0(\psi(\boldsymbol{X}) = 1) = P_0(\psi(\boldsymbol{X}) = 1) = P_0(T(\boldsymbol{X}) \in \mathcal{C}) \leq \alpha.$$

In der Regel lautet der Test ψ:

$\psi(\boldsymbol{X}) = I(T(\boldsymbol{X}) > c_{1-\alpha})$ falls große Werte von $T(\boldsymbol{X})$ gegen H_0 sprechen.

$\psi(\boldsymbol{X}) = I(T(\boldsymbol{X}) < c_{\alpha})$ falls kleine Werte von $T(\boldsymbol{X})$ gegen H_0 sprechen.

$\psi(\boldsymbol{X}) = I(T(\boldsymbol{X}) < c_{\alpha/2}$ oder $T(\boldsymbol{X}) > c_{1-\alpha/2})$
falls große und kleine Werte von $T(\boldsymbol{X})$ gegen H_0 sprechen.

Die Schranken sind die Quantile der Verteilung von $T(\boldsymbol{X})$ wenn $\theta \in \Theta_0$, z.B. ist $c_{1-\alpha}$ das $1-\alpha$-Quantil.
Die sogenannte Prüfverteilung (Verteilung von $T(\boldsymbol{X})$ unter H_0) ist in einigen Fällen (d.h., Teststatistiken) leicht, aber u. U. auch sehr schwer zu bestimmen.

Beispiel 6.1.6 (Fortsetzung: Ist eine Münze fair?)
Das Testproblem lautet

$$H_0 : \pi = \frac{1}{2} \text{ vs. } H_1 : \pi \neq \frac{1}{2}.$$

Als Teststatistik T eignet sich $T(X) = X$ und wir lehnen die Null-Hypothese ab, wenn x zu groß oder zu klein ist (man nennt dies einen zweiseitigen Test, im Unterschied zu obigem einseitigen Test). Der Test lautet also

$$\psi(X) = I(T(X) < c_{\alpha/2} \text{ oder } T(X) > c_{1-\alpha/2}).$$

Der Ausgang unseres Experimentes sei $x = 7$ bei $n = 10$. Wir bestimmen jetzt das $\alpha/2$-Quantil und das $1-\alpha/2$-Quantil der Verteilung von $T(x)$ unter H_0, also der Verteilung $\mathcal{B}(n, 0.5)$ für $\alpha = 0.05$:

```
> qbinom(0.05/2, size = 10, prob = 0.5)

[1] 2

> qbinom(1 - 0.05/2, size = 10, prob = 0.5)

[1] 8
```

Damit können wir uns nicht gegen H_0 (Fairness der Münze) entscheiden, da $x = 7$ weder zu klein (kleiner 2) noch zu groß (größer 8) ist.
Wie groß ist α für den Test $\psi(X) = I(X < 2$ oder $X > 8)$?

Eine formale Möglichkeit, Teststatistiken zu konstruieren, ist die Anwendung des Likelihood-Quotienten-Prinzips, wobei der Quotient der maximierten Likelihood unter der Null-Hypothese und des ML-Schätzers als Teststatistik verwendet wird.

Definition 6.1.6
Der Quotient

$$\frac{\max\limits_{\Theta_0} L(\theta)}{\max\limits_{\Theta} L(\theta)} = \frac{\max\limits_{\Theta_0} L(\theta)}{L(\hat{\theta}_{ML})}$$

heißt Likelihood-Quotient.

Der Likelihood-Quotient steht in engem Zusammenhang zur normierten (Log)-Likelihoodfunktion:

Definition 6.1.7

$$\tilde{L}(\theta) := \frac{L(\theta)}{L(\hat{\theta}_{ML})} \quad \textit{heißt } \textbf{normierte Likelihoodfunktion}$$

und

$$\tilde{l}(\theta) = l(\theta) - l(\hat{\theta}_{ML}) \quad \textit{heißt } \textbf{normierte Loglikelihoodfunktion}$$

Es gilt:

$$0 \leq \tilde{L}(\theta) \leq 1$$
$$-\infty < \tilde{l}(\theta) \leq 0$$

Beispiel 6.1.7 (Likelihood-Quotient der Binomialverteilung)
Für das Testproblem

$$H_0 : \pi = \frac{1}{2} \text{ vs. } H_1 : \pi \neq \frac{1}{2}$$

lautet die Likelihood-Quotienten-Teststatistik

$$T(x) = \frac{0.5^x 0.5^{n-x}}{(x/n)^x (1 - x/n)^{n-x}}$$

und für $\pi = 1/2$ gilt (ohne Beweis) $-2\log(T(x)) \sim \chi^2_1$, d.h., die Prüfverteilung ist eine quadrierte Normalverteilung (später in Vorlesung).

Zur Bestimmung von likelihood-basierten Konfidenzintervallen muss man nun den Test 'invertieren' um zu entscheiden, welche Werte von $\tilde{l}(\theta)$ zu "unplausibel" sind. Es wird nun nicht mehr zu einem festen Θ_0 ein Niveau-α-Test durchgeführt, sondern es wird eine Menge Θ_0 gesucht, für die der Niveau-α-Test ψ gerade *nicht* ablehnt.

Definition 6.1.8
Die Menge

$$\{\theta \in \Theta : \psi(X) = 0\} = \{\theta \in \Theta : T(X) < c_{1-\alpha}\}$$

wird Konfidenzintervall oder Konfidenzbereich zum Konfidenzniveau $1 - \alpha$ genannt.

Das Konfidenzniveau $1 - \alpha$ lässt sich wie folgt frequentistisch interpretieren: Bei hypothetischer Wiederholung des zugrundeliegenden Zufallsexperiments überdecken die so konstruierten Likelihood-Intervalle in ungefähr $(1 - \alpha) \cdot 100\%$ aller Fälle den wahren (aber als unbekannt angesehenen) Parameter θ.

Beispiel 6.1.8 (Likelihood-Konfidenzintervall für π)
Wir wissen, dass -2 mal die normierte Loglikelihood evaluiert an der Null-Hypothese $\theta \in \Theta_0$ verteilt ist nach χ^2_1. Also ist unser Konfidenzintervall

$$\left\{\pi \in [0,1] : \ -2\log\left(\frac{\pi^x(1-\pi)^{n-x}}{(x/n)^x(1-x/n)^{n-x}}\right) < c_{1-\alpha}\right\}$$

bzw.

$$\left\{\pi \in [0,1] : \ \tilde{l}(\pi) \geq -c_{1-\alpha}/2\right\}$$

und wir brauchen nur die Nullstellen der Funktion $\tilde{l}(\pi) + c_{1-\alpha}/2$ zu suchen:

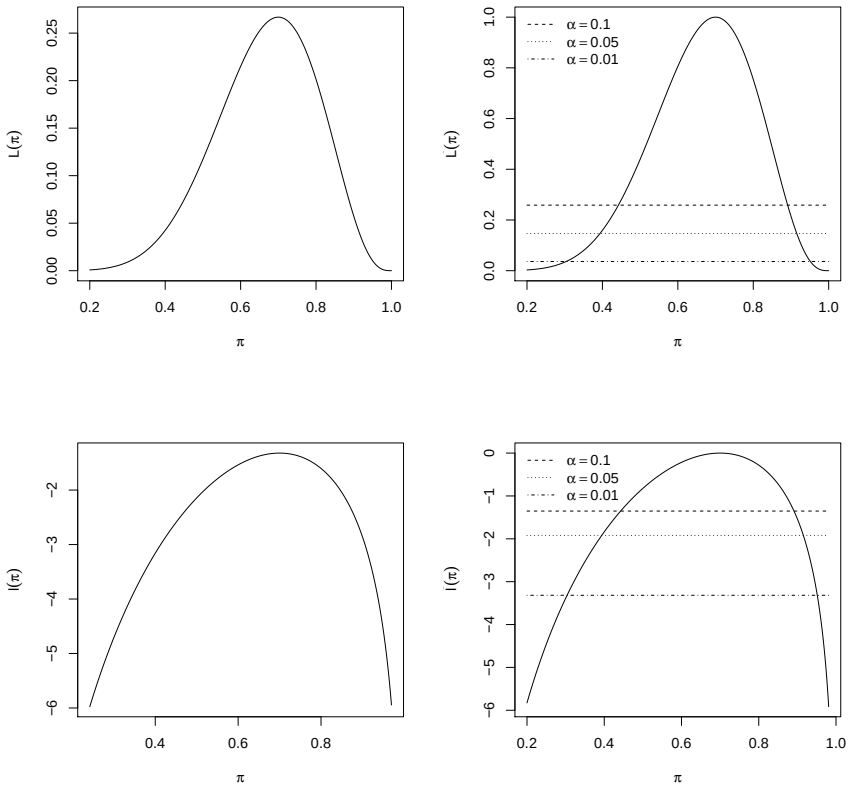

Abbildung 6.1: Für $n = 10, x = 7$ in der Binomialverteilung: Likelihood (oben links), normierte Likelihood (oben rechts), Loglikelihood (unten links) und normierte Loglikelihood (unten rechts). Linien verdeutlichen Konfidenzintervalle.

```
> konfint <- function(x, n, conf.level = 0.95) {
+     lrstat <- function(pi)
+         dbinom(x, n, pi, log = TRUE) -
+         dbinom(x, n, x/n, log = TRUE)
+     foo <- function(pi)
+         lrstat(pi) + qchisq(conf.level, df = 1)/2
+     c(uniroot(foo, c(1e-5, x / n))$root,
+       uniroot(foo, c(x / n, 1 - 1e-5))$root)
+ }
> konfint(7, 10, conf.level = 0.9)

[1] 0.4423946 0.8912495
```

```
> konfint(7, 10, conf.level = 0.95)
[1] 0.3934616 0.9154452
> konfint(7, 10, conf.level = 0.99)
[1] 0.3037870 0.9514883
```

In Abb. 6.1 sind die Werte $c_{1-\alpha}/2$ (unten rechts) bzw. $\exp(c_{1-\alpha}/2)$ (oben rechts) für verschiedene Werte von $\alpha = (0.01, 0.05, 0.1)$ aufgetragen.
Beachte: *Likelihood-Intervalle sind meist unsymmetrisch um $\hat{\pi}_{ML}$!*

Die Bestimmung von Likelihood-Intervallen ist im Allgemeinen nur numerisch möglich, dann aber relativ trivial, wie wir eben gesehen haben.
Wie der ML-Schätzer sind Likelihood-Intervalle invariant bzgl. monotonen Transformationen.

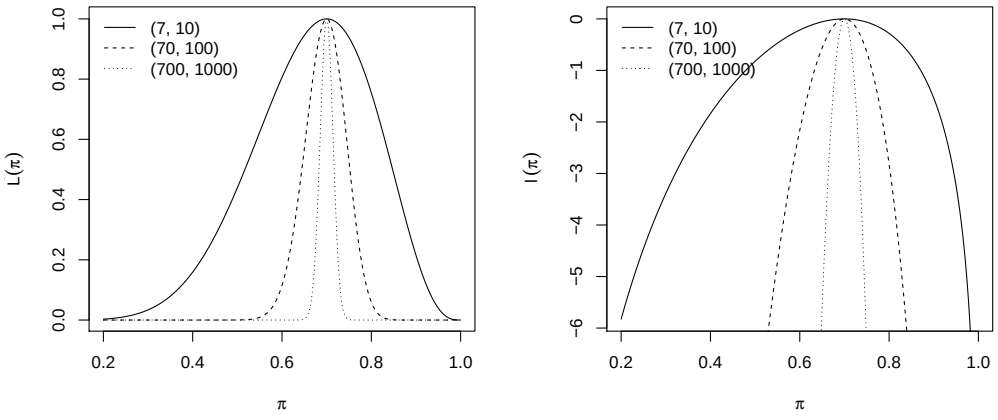

Abbildung 6.2: Vergleich der normierten Likelihoodfunktionen (links) bzw. Loglikelihoodfunktionen (rechts) der Binomialverteilung für $n = (10, 100, 1000)$ und $x = (7, 70, 700)$.

Quadratische Approximation der Log-Likelihood:
Man kann mit Hilfe einer Taylorreihendarstellung um den Entwicklungspunkt $\hat{\theta}_{ML}$ zeigen, dass die normierte Loglikelihoodfunktion $\tilde{l}(\theta)$ approximativ eine quadratische Funktion ist:

$$\begin{aligned}\tilde{l}(\theta) &= \sum_{k=0}^{2} \frac{\tilde{l}^{\,k}(\hat{\theta}_{ML})}{k!}(\theta - \hat{\theta}_{ML})^k + \text{Restglied} \\ &\approx \tilde{l}(\hat{\theta}_{ML}) + \tilde{l}'(\hat{\theta}_{ML}) \cdot (\theta - \hat{\theta}_{ML}) + \frac{1}{2} \cdot \tilde{l}''(\hat{\theta}_{ML}) \cdot (\theta - \hat{\theta}_{ML})^2\end{aligned}$$

Hierbei ist:

$$\begin{aligned} \tilde{l}(\hat{\theta}_{ML}) &= 0 \\ \tilde{l}'(\hat{\theta}_{ML}) &= l'(\hat{\theta}_{ML}) = 0 \\ \tilde{l}''(\hat{\theta}_{ML}) &= l''(\hat{\theta}_{ML}) \end{aligned}$$

Damit erhält man insgesamt:

$$\tilde{l}(\theta) \approx \frac{1}{2} \cdot l''(\hat{\theta}_{ML}) \cdot (\theta - \hat{\theta}_{ML})^2$$

Hier ist $l''(\hat{\theta}_{ML})$ die zweite Ableitung (die Krümmung) von $l(\theta)$, ausgewertet am ML-Schätzer.
Beachte: Die quadratische Approximation wird umso besser, je mehr Daten in die Likelihood eingehen (siehe Abbildung 6.3).

Durch Einsetzen der quadratischen Approximation für $\tilde{l}(\theta)$ in

$$\{\theta \,\Big|\, \tilde{l}(\theta) \geq c\}$$

erhält man

$$\hat{\theta}_{ML} \pm \underbrace{\sqrt{-2c}}_{d} \cdot \sqrt{\left[-l''(\hat{\theta}_{ML})\right]^{-1}}$$

als **approximatives Konfidenzintervall** (nach Abraham Wald (1902-1950) auch **Wald-Intervall** genannt) zum Niveau $1 - \alpha$.
Daher definiert man den **Standardfehler** ("Standard Error")

$$SE(\hat{\theta}_{ML}) := \sqrt{\left[-l''(\hat{\theta}_{ML})\right]^{-1}}$$

Aus der folgenden Tabelle können die Werte für d (Quantile der Normalverteilung, weil $\hat{\theta}_{\text{ML}}$ asymptotisch normalverteilt ist, näheres später) für unterschiedliche Niveaus $1 - \alpha$ entnommen werden:

$1 - \alpha/2$	d
0.95	1.6449
0.975	1.96
0.995	2.5758

Eigenschaften von Wald-Intervallen:

- Sie sind immer symmetrisch um den ML-Schätzer und damit geht die Invarianzeigenschaft verloren.

- Sie sind einfacher zu berechnen, haben aber typischerweise (leicht) schlechtere Eigenschaften.
- Manchmal sind die Grenzen von Wald-Intervallen nicht Element des Parameterraums Θ, siehe z.B. den Fall $1-\alpha = 0.99$ in Beispiel 6.1.9.
- Für große n werden Wald-Intervalle Likelihood-Intervallen immer ähnlicher.

Beispiel 6.1.9 (Wald-Intervalle für Binomialverteilung)

$$
\begin{aligned}
l'(\pi) &= \frac{x}{\pi} - \frac{n-x}{1-\pi} \\
l''(\pi) &= -\frac{x}{\pi^2} - \frac{n-x}{(1-\pi)^2} \\
\Rightarrow -l''(\hat{\pi}_{ML} = \tfrac{x}{n}) &= \frac{x}{\left(\frac{x}{n}\right)^2} + \frac{n-x}{\left(\frac{n-x}{n}\right)^2} = \frac{n^2}{x} + \frac{n^2}{n-x} \\
&= \frac{n}{\hat{\pi}_{ML}(1-\hat{\pi}_{ML})} \\
\Rightarrow SE(\hat{\pi}_{ML}) &= \sqrt{\frac{\hat{\pi}_{ML}(1-\hat{\pi}_{ML})}{n}}
\end{aligned}
$$

Für das Zahlenbeispiel von $x = 7, n = 10$ ergibt sich folgender SE:

$$\Rightarrow SE(\hat{\pi}_{ML}) = \sqrt{\frac{0.7 \cdot 0.3}{10}} = 0.145$$

$\rightsquigarrow$ *Tabelle mit Wald-Intervallen:*

$1-\alpha$	*Wald-Konfidenzintervall*
0.9	$[0.461; 0.939]$
0.95	$[0.416; 0.984]$
0.99	$[0.327; 1.073]$

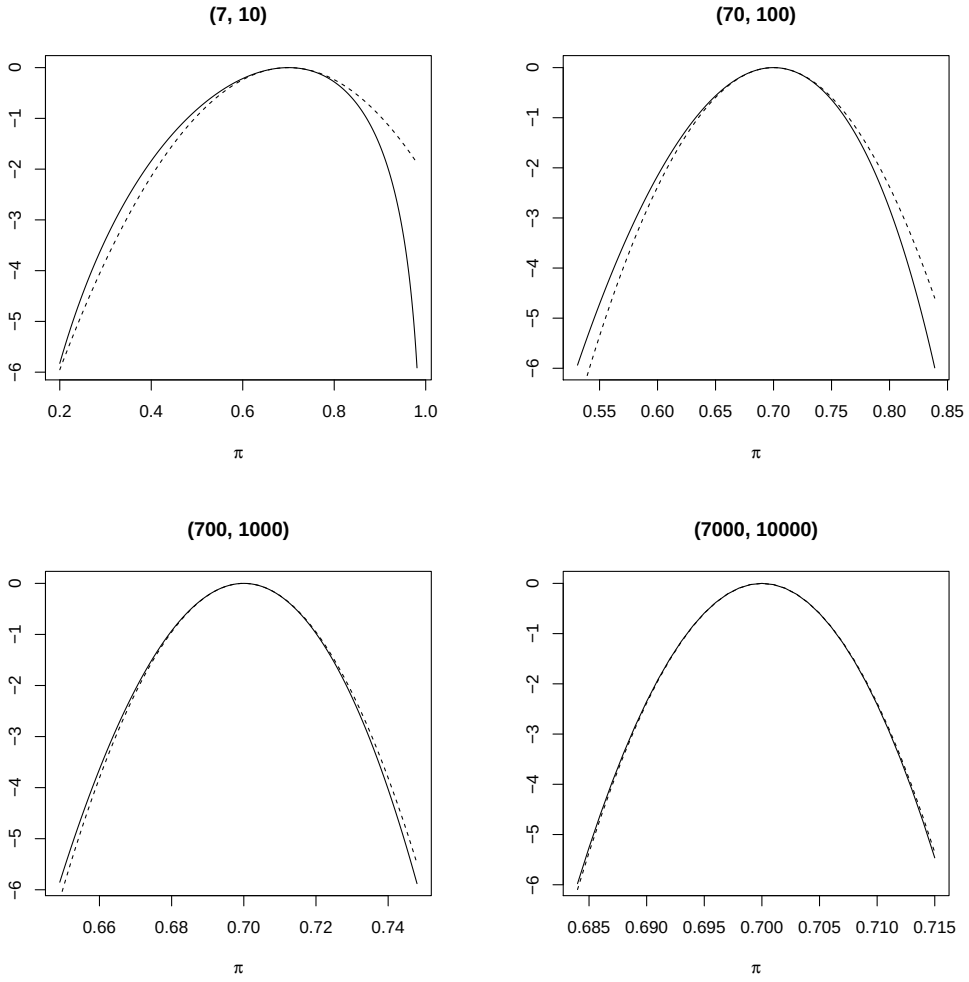

Abbildung 6.3: Vergleich der normierten Loglikelihood mit der quadratischen Approximation für die Binomialverteilung mit $n = (10, 1000, 1000, 10000)$ und $x = (7, 70, 700, 7000)$.

6.2 Erwartungstreue

Eine wünschenswerte Eigenschaft von Punktschätzern (im frequentistischen Sinne) ist die **Erwartungstreue:**

Definition 6.2.1
Ein Schätzer $\hat{\theta}$ (als Funktion der zufälligen Stichprobe X) heißt **erwartungstreu** *oder* **unverzerrt** *für einen unbekannten Parameter θ, falls gilt:*

$$E(\hat{\theta}) = \theta$$

Beispiel 6.2.1 (Erwartungstreuer Schätzer im Binomialexperiment)
Sei $X \sim \mathcal{B}(n, \pi)$. Dann gilt $\hat{\pi} = X/n$, also die relative Häufigkeit. Dies ist ein erwartungstreuer Schätzer der Wahrscheinlichkeit π im Binomialexperiment:

$$E(\hat{\pi}) = E\left(\frac{X}{n}\right) = \frac{1}{n} \cdot E(X) = \frac{1}{n} \cdot n \cdot \pi = \pi$$

Bemerkungen zur Erwartungstreue:

- Erwartungstreue ist nicht invariant bzgl. monotoner Transformationen! Das heißt, falls $\hat{\theta}$ erwartungstreu für θ ist, so gilt im Allgemeinen nicht, dass $g(\hat{\theta})$ erwartungstreu für $g(\theta)$ ist (nur wenn g linear).
- Die Existenz von erwartungstreuen Schätzern ist nicht gesichert.
- Erwartungstreue Schätzer sind nicht notwendigerweise Element des Parameterraums Θ.
- ML-Schätzer sind nicht immer erwartungstreu, zumindest aber asymptotisch erwartungstreu (für wachsenden Stichprobenumfang sind sie im Grenzwert erwartungstreu)

Beispiel 6.2.2 (Taxis in Lübeck)
Die Taxis in Lübeck seien von $1, \ldots, N$ durchnummeriert. Ein Besucher sieht an einem Taxistand $n = 3$ Taxis und fragt nach deren Nummern: $Y = \{Y_1, \ldots, Y_n\}$.
Wie kann er daraus einen erwartungstreuen Schätzer für N berechnen?
Klarerweise ist $X = \max(Y)$ eine untere Schranke für N.
Wie lautet die Wahrscheinlichkeitsfunktion von X beim Ziehen ohne Zurücklegen einer n-elementigen Teilmenge?

$$P(X = x) = \frac{\binom{x-1}{n-1}}{\binom{N}{n}} \quad \text{für } x = \{n, n+1, \ldots, N\}$$

Insbesondere gilt:

$$\sum_{x=n}^{N} P(X = x) \;=\; 1$$

und daher:

$$\sum_{x=n}^{N} \binom{x-1}{n-1} = \binom{N}{n} \tag{6.1}$$

Berechne nun den Erwartungswert von X:

$$\begin{aligned}
EX &= \sum_{x=n}^{N} x \cdot \frac{\binom{x-1}{n-1}}{\binom{N}{n}} \\
&= \frac{1}{\binom{N}{n}} \sum_{x=n}^{N} \underbrace{\frac{x \cdot (x-1)!}{(n-1)! \cdot (x-n)!}}_{n \cdot \frac{x!}{n! \cdot (x-n)!}} \\
&= \frac{1}{\binom{N}{n}} \sum_{x=n}^{N} \binom{x}{n} \cdot n \\
&= \frac{n}{\binom{N}{n}} \sum_{x=n+1}^{N+1} \binom{x-1}{n} \\
&= \frac{n}{\binom{N}{n}} \underbrace{\sum_{x=n+1}^{N+1} \binom{x-1}{(n+1)-1}}_{=\binom{N+1}{n+1} \text{ (Gleichung 6.1)}} \\
&= \frac{n(N+1)}{n+1}
\end{aligned}$$

Wegen der Linearität des Erwartungswertes folgt, dass

$$\hat{N} = \frac{n+1}{n} \cdot X - 1$$

ein erwartungstreuer Schätzer für N ist, denn

$$\begin{aligned}
E(\hat{N}) &= \frac{n+1}{n} \cdot EX - 1 \\
&= \frac{n+1}{n} \cdot \frac{n(N+1)}{n+1} - 1 \\
&= N
\end{aligned}$$

Zahlenbeispiel:
$n = 3, \quad X = \max(Y) = 722$
$\Rightarrow \quad \hat{N} = 961.7$

Die Likelihoodfunktion ist

$$L(N) = const \cdot \frac{(N-n)!}{N!} \text{ für } N = x, x+1, \ldots$$

Diese wird maximiert für $N = x$ *(*$L(N)$ *ist monoton fallend), d.h. der ML-Schätzer für* N *ist* $\max(Y)$*:* $\hat{N}_{ML} = \max(Y)$

6.3 Bayes-Inferenz

Die **Bayes-Inferenz** ist ein weiterer Ansatz zur statistischen Inferenz und basiert auf dem subjektivistischen Wahrscheinlichkeitskonzept:
Der unbekannte Parameter θ ist nun eine Zufallsvariable, welche mit einer Wahrscheinlichkeitsfunktion $f(\theta)$ (**Priori-Verteilung**) versehen wird.
Zunächst können wir nur diskrete Parameter behandeln.
Wir benötigen noch folgende Definition:

Definition 6.3.1 (Satz von Bayes für Wahrscheinlichkeitsfunktionen)
Seien X und Y zwei Zufallsvariablen mit gemeinsamer Wahrscheinlichkeitsfunktion $f_{X,Y}(x,y)$ und daraus abgeleiteten Wahrscheinlichkeitsfunktionen $f_X(x)$, $f_Y(y)$, $f_{X|Y}(x|y)$ und $f_{Y|X}(y|x)$.
Dann gilt für alle x und alle y mit $f(y) > 0$:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\sum_x f_{Y|X}(y|x)f_X(x)}$$

Dies folgt direkt aus der Definition der bedingten Wahrscheinlichkeitsfunktion.

Sei $X = x$ eine Beobachtung eines Zufallsexperiments, das von einem unbekannten Parameter $\theta \in \Theta$ abhängt, wobei Θ abzählbar sei.
Mit dem Satz von Bayes für Wahrscheinlichkeitsfunktionen kann man eine **Posteriori-Verteilung** berechnen:

$$\underbrace{f(\theta|x)}_{\text{Posteriori}} = \frac{\overbrace{f(x|\theta)}^{\text{Likelihood-Fkt}} \cdot \overbrace{f(\theta)}^{\text{Priori}}}{f(x)} = \frac{f(x|\theta)f(\theta)}{\sum_\theta f(x|\theta)f(\theta)}$$

Da der Nenner $f(x)$ nicht von θ abhängt, schreibt man auch kürzer:

$$f(\theta|x) \propto f(x|\theta)f(\theta)$$

Beachte: Da $X = x$ beobachtet wurde, muss automatisch $f(x) = P(X = x) > 0$ gelten, der Nenner der Posteriori-Verteilung ist somit echt größer Null.

Beispiel 6.3.1 (Posteriori-Verteilung im Binomialexperiment)
Angenommen wir interessieren uns für den Parameter $\theta = \pi$, nehmen aber an dass nur Werte in $\Pi = \{0.00, 0.02, 0.04, \ldots, 0.98, 1.00\}$ erlaubt sind und damit der Parameterraum abzählbar ist.
Als **Priori-Verteilung** *$f(\pi)$ können wir z.B. eine Gleichverteilung auf den 51 Elementen von Π wählen.*

Nach der Beobachtung $X = x$ aus einer $\mathcal{B}(n, \pi)$-Verteilung ergibt sich die **Posteriori-Verteilung**

$$f(\pi|x) = \frac{f(x|\pi)f(\pi)}{\sum\limits_{\pi} f(x|\pi)f(\pi)} = \frac{f(x|\pi)}{\sum\limits_{\pi} f(x|\pi)}$$

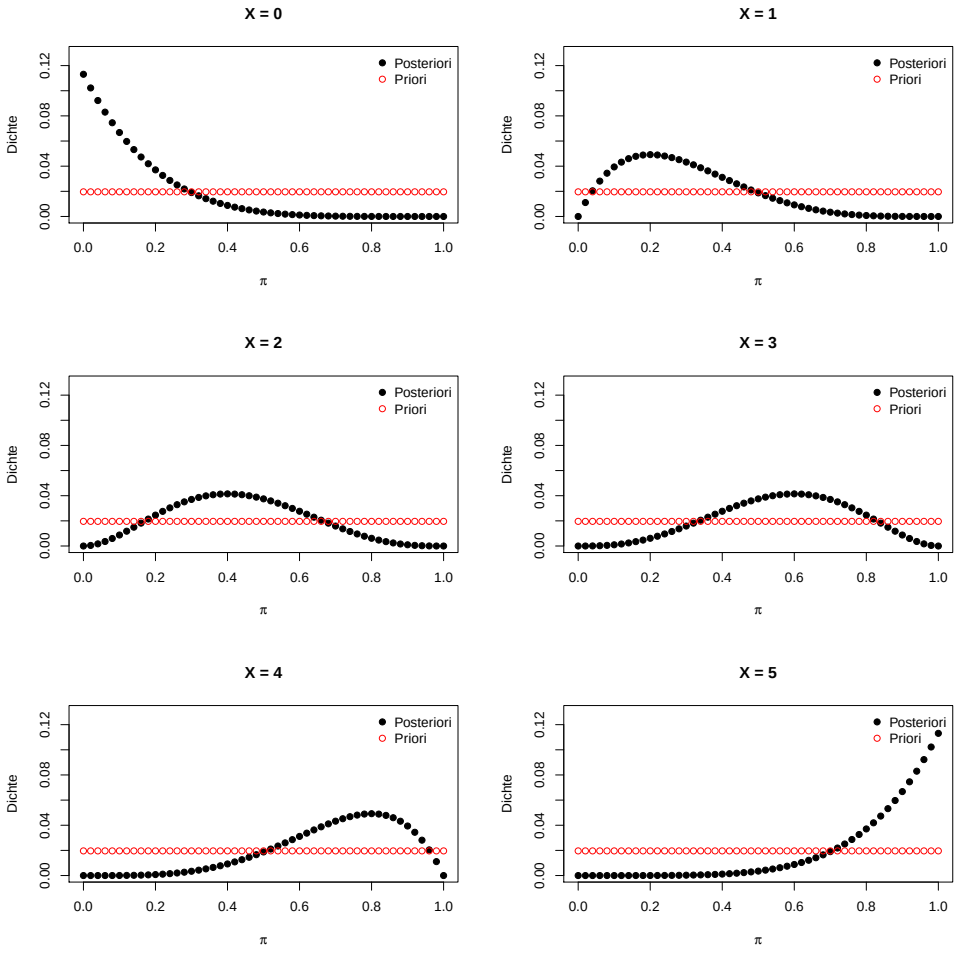

Abbildung 6.4: Posteriori-Verteilung im Binomialexperiment für $X \sim \mathcal{B}(5, \pi)$ bei Priori-Gleichverteilung.

Eine weitere Möglichkeit wäre es, eine Dreiecksverteilung (favorisiere Werte von π nahe bei 0.5) als ***Priori-Verteilung*** *anzunehmen, die z.B. folgende Form haben könnte:*

$$f(\pi) = \frac{1}{C}\{26 - 25 \cdot |2 \cdot \pi - 1|\}$$

für $\pi \in \{0.00, 0.02, 0.04, \ldots, 0.98, 1.00\}$ *und* $C = 1/676$. *Dabei ist* C *so gewählt, dass* $\sum_\pi f(\pi) = 1$ *gilt.*

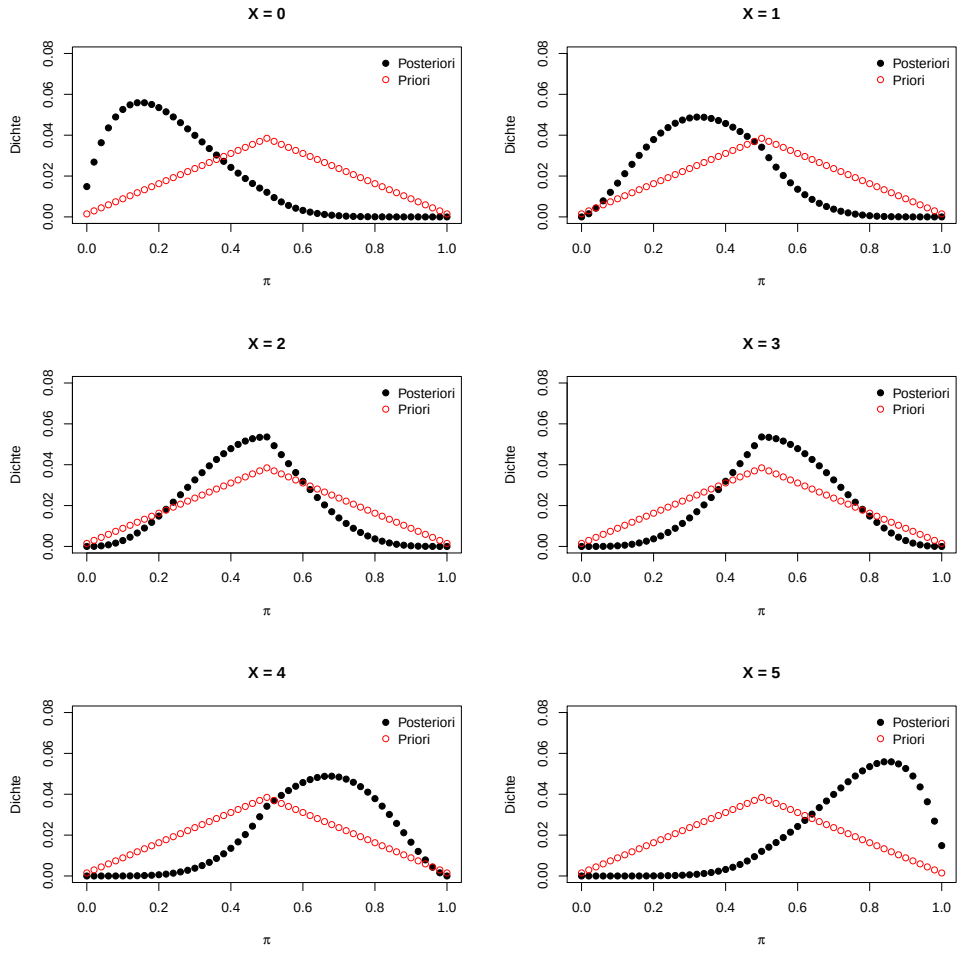

Abbildung 6.5: Posteriori-Verteilung im Binomialexperiment für $X \sim \mathcal{B}(5, \pi)$ bei Priori-Dreiecksverteilung.

Alle weiteren Schlüsse über $\theta \in \Theta$ werden nun aus der Posteriori-Verteilung gezogen. Dafür verwendet man **Bayesianische Punktschätzer** und **Bayesianische Intervallschätzer**. Als Punktschätzer bieten sich Lageparameter der Posteriori-Verteilung an:

- **Posteriori-Erwartungswert** $\hat{\theta}_{\text{Erw}} = E(\theta|x) = \sum_{\theta \in \Theta} \theta f(\theta|x)$
- **Posteriori-Modus** $\hat{\theta}_{\text{Mod}} = \arg\max_{\theta \in \Theta} f(\theta|x)$

- **Posteriori-Median** $\hat{\theta}_{\text{Med}} = \min\{\theta \in \Theta : F(\theta|x) \geq 0.5\}$

wobei $f(\theta|x)$ die Wahrscheinlichkeitsfunktion und $F(\theta|x)$ die Verteilungsfunktion der Posteriori-Verteilung ist.
Wird als Priori-Verteilung die Gleichverteilung verwendet, d.h. $f(\theta) = \text{const}$, so gilt:

$$\text{Posteriori-Modus } \hat{\theta}_{\text{Mod}} = \text{ML-Schätzer } \hat{\theta}_{\text{ML}}$$

da $f(\theta)$ in

$$f(\theta|x) \propto f(x|\theta)f(\theta)$$

nicht von θ abhängt, folgt dass

$$\hat{\theta}_{\text{Mod}} = \arg\max_{\theta\in\Theta} f(\theta|x) = \arg\max_{\theta\in\Theta} f(x|\theta) = \hat{\theta}_{\text{ML}}.$$

Für Beispiel 6.3.1 ergeben sich bei Priori-Gleichverteilung (G) bzw. Dreiecksverteilung (D) folgende Punktschätzer bei Beobachtung von $X = x$ Erfolgen bei $n = 5$ Versuchen:

	Erwartungswert		Modus		Median	
x	G	D	G	D	G	D
0.00	0.13	0.23	0.00	0.16	0.10	0.22
1.00	0.29	0.35	0.20	0.32	0.26	0.34
2.00	0.43	0.45	0.40	0.50	0.42	0.46
3.00	0.57	0.55	0.60	0.50	0.58	0.54
4.00	0.71	0.65	0.80	0.68	0.74	0.66
5.00	0.87	0.77	1.00	0.84	0.90	0.78

Als **Bayesianische Intervallschätzer** (auch Kredibilitätsregion oder -intervall) zum Niveau $1-\alpha$ kommt in Prinzip jede Teilmenge A des Trägers Θ in Frage, für die folgendes gilt:

$$\sum_{\theta\in A} f(\theta|x) \geq 1 - \alpha$$

Wenn für A noch zusätzlich gefordert wird, dass $\forall\,\theta_1 \in A$ und $\forall\,\theta_2 \in \Theta \setminus A$ gilt

$$f(\theta_1|x) \geq f(\theta_2|x)$$

dann wird A eine **"highest posterior density region" (HPD-Region)** genannt.
Eine HPD-Region kann wie folgt aufgestellt werden:

1. sortiere die Werte der Posteriori-Verteilung $f(\theta|x)$ der Größe nach (absteigend)

2. summiere die Werte kumulativ auf (z. B. in R mit der Funktion `cumsum()`), bis die Summe größer als das Niveau $1 - \alpha$ ist

3. die entsprechenden Werte von θ definieren dann eine HPD-Region

Für Beispiel 6.3.1 ergeben sich bei Priori-Gleichverteilung (G) bzw. Dreiecksverteilung (D) folgende 95% HPD-Regionen bei Beobachtung von $X = x$ Erfolgen bei $n = 5$ Versuchen:

	95% HPD-Intervall für π	
x	G	D
0	[0.00;0.36]	[0.00;0.46]
1	[0.02;0.56]	[0.08;0.60]
2	[0.12;0.74]	[0.18;0.72]
3	[0.26;0.88]	[0.28;0.82]
4	[0.44;0.98]	[0.40;0.92]
5	[0.64;1.00]	[0.54;1.00]

Beispiel 6.3.2 (Bayes-Inferenz im Capture-Recapture-Experiment) *Der unbekannte Parameter ist hier die Anzahl N der Fische in einem See, wobei θ als diskrete Zufallsvariable mit Träger*

$$\mathcal{T}_{Priori} = \{M, M+1, \ldots, N_{\max}\}$$

angesehen wird. Im Unterschied zu Abschnitt 6.1 ist damit der unbekannte Parameter N ganzzahlig. **Vor** *der Beobachtung $X = x$, d.h. in der Stichprobe der Größe n befanden sich x markierte Fische, weiß man lediglich, dass mindestens M Fische im See schwimmen, denn genau so viele wurden ja früher markiert und wieder in den See geworfen.*
Als Priori-Verteilung wird z. B. eine Gleichverteilung gewählt:

$$f(\theta = N) \text{ für } N \in \mathcal{T}_{Priori}$$

Über den Satz von Bayes kann nun die Posteriori-Wahrscheinlichkeitsfunktion berechnet werden:

$$f(N|x) = \frac{f(x|N)f(N)}{f(x)} = \frac{f(x|N)f(N)}{\sum_N f(x|N)f(N)}$$

Dabei ist die Zufallsvariable X hypergeometrisch verteilt und hat folgende Wahrscheinlichkeitsfunktion:

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Die Likelihoodfunktion bei beobachteter Stichprobe $X = x$ und unbekanntem Parameter N ist:

$$L(\theta = N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Erst **nach** *der Beobachtung $X = x$ weiß man, dass mindestens $M + n - x$ Fische im See schwimmen, da die Likelihood $f(x|N)$ für $x < M + n - N$ den Wert 0 annimmt, also für N gelten muss: $N < M+n-x$. Außerdem gilt, dass $n \leq N$ ist. Damit kann man den Träger $\mathcal{T}_{Posteriori}$ der Posteriori-Verteilung $f(\theta|x)$ angeben:*

$$\mathcal{T}_{Posteriori} = \{\max(M + n - x, n), \max(M + n - x, n) + 1, \ldots, Y_{\max}\}$$

Als Punktschätzer bietet sich nun der Posteriori-Modus an, also der Wert, der die Posteriori-Verteilung maximiert. Dieser ist in diesem Fall gleich dem ML-Schätzer, da eine Priori-Gleichverteilung gewählt wurde. Außerdem können als weitere Punktschätzer der Posteriori-Erwartungswert und der Posteriori-Modus berechnet werden.

In Abbildung 6.6 wird für die Priori-Verteilung eine Gleichverteilung verwendet, in Abbildung 6.7 hingegen eine gestutzte geometrische Verteilung mit Parameter $\gamma = 0.01$, d.h.

$$f(N) \propto (1 - \gamma)^N \text{ für } N \in \mathcal{T}_{Priori}$$

Diese favorisiert a priori kleinere Werte von N, wie man in der oberen Graphik in Abbildung 6.7 erkennen kann. Entsprechend ist die Posteriori-Verteilung im Vergleich zur Priori-Gleichverteilung (Abbildung 6.6) leicht nach links verschoben.

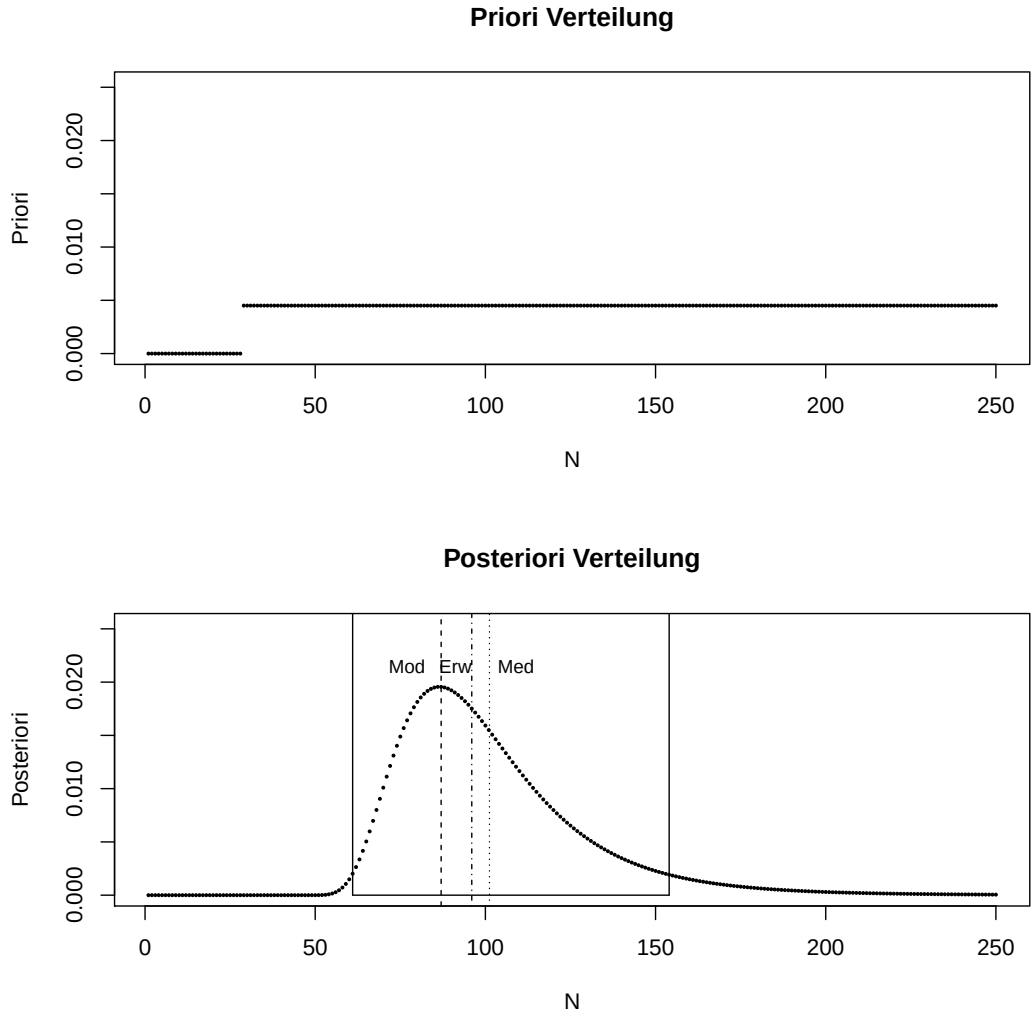

Abbildung 6.6: Posteriori-Verteilung von N bei Priori-Gleichverteilung bei $M = 29$ und $\gamma = 0$ (für $x = 10$ und $n = 30$). Die verschiedenen Punktschätzer werden durch Linien verdeutlicht. Die 95%-HPD-Region ist in durchgezogenen Linien dargestellt.

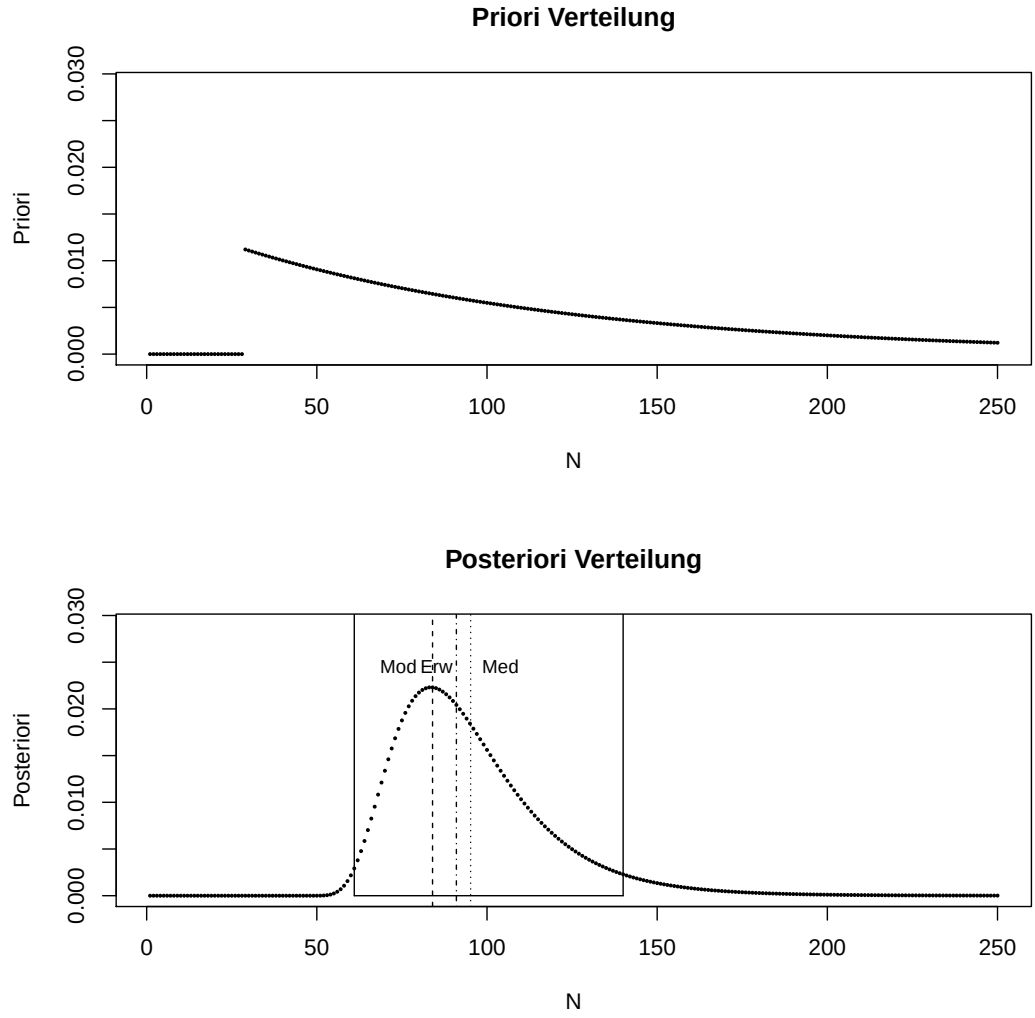

Abbildung 6.7: Posteriori-Verteilung von N bei Verwendung einer gestutzten geometrischen Verteilung mit Parameter $\gamma = 0.01$ als Priori-Verteilung.

Kapitel 7

Markov-Ketten

7.1 Definition und Eigenschaften von Markov-Ketten

Sei $\boldsymbol{X} = (X_0, X_1, X_2, \ldots)$ eine Folge von diskreten Zufallsvariablen, die alle Ausprägungen in einer endlichen bzw. abzählbaren Menge S haben. Dabei heißt S der **Zustandsraum** und $s \in S$ ein **Zustand**. Man nennt X einen **stochastischen Prozess**.

Definition 7.1.1
*$\boldsymbol{X}$ heißt **Markov-Kette**, falls für alle $n \geq 1$ und für alle $s, x_0, x_1, \ldots, x_{n-1} \in S$ gilt:*

$$P(X_n = s | X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = P(X_n = s | X_{n-1} = x_{n-1})$$

In Worten bedeutet die obige Formel, dass, bedingt auf die gesamte Vergangenheit $X_0, X_1, \ldots, X_{n-1}$ des Prozesses $\boldsymbol{X}$, die Verteilung von X_n nur vom letzten Wert $X_{n-1} = x_{n-1}$ abhängt.
Die Abhängigkeit der Werte lässt sich in einem graphischen Modell gut durch Pfeile darstellen:

$$X_0 \to X_1 \to X_2 \to \ldots \to X_{n-1} \to X_n$$

Anders ausgedrückt:
***Bedingt** auf die Gegenwart (X_{n-1}) ist die Zukunft des Prozesses $(X_n, X_{n-1}, \ldots)$ **unabhängig** von seiner Vergangenheit $(X_0, X_1, \ldots, X_{n-2})$.*

Hier wird implizit der Begriff der **bedingten Unabhängigkeit** von Zufallsvariablen verwendet:

Definition 7.1.2 (Bedingt unabhängige Zufallsvariablen)
Zwei diskrete Zufallsvariablen X und Y heißen **bedingt unabhängig** *gegeben Z, wenn für die entsprechend definierten Wahrscheinlichkeitsfunktionen gilt*

$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z) \cdot f_{Y|Z}(y|z)$$

für alle x, y und z aus den entsprechenden Trägern.

Die Entwicklung einer Markov-Kette $\boldsymbol{X}$ (d.h. die Abfolge der Werte, welche $\boldsymbol{X}$ annimmt) ist gekennzeichnet durch die (Ein-Schritt) **Übergangswahrscheinlichkeit** von einem Zustand i in den Zustand j

$$P(X_{n+1} = j|X_n = i) \quad \text{für alle } i,j \in S$$

Man nennt eine Markov-Kette **homogen**, wenn diese Wahrscheinlichkeit nicht von n abhängt und definiert:

$$p_{ij} = P(X_{n+1} = j|X_n = i) = P(X_1 = j|X_0 = i)$$

Die Werte p_{ij} werden in einer $|S| \times |S|$-Matrix $\boldsymbol{P}$ zusammengefasst, der sogenannten **Übergangsmatrix**. $\boldsymbol{P}$ ist eine **stochastische Matrix**, d.h. sie hat folgende Eigenschaften:

1. $p_{ij} \geq 0 \quad$ für alle $i,j \in S$
2. $\sum_j p_{ij} = 1 \quad$ für alle $i \in S$, d.h. die Zeilen addieren sich zu eins.

Beispiel 7.1.1
Betrachte den Zustand eines Telephons (frei, besetzt). Innerhalb eines Zeitabschnittes sei p die Wahrscheinlichkeit, dass angerufen wird (nur ein Anruf pro Zeitabschnitt sei erlaubt, anklopfen geht nicht). Wenn telephoniert wird, sei die Wahrscheinlichkeit, innerhalb eines Zeitabschnittes aufzulegen gleich q. Dies führt zu einer Markov-Kette mit Zustandsraum $S = \{0,1\}$ und Übergangsmatrix

$$\boldsymbol{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

$$\text{etwa } \boldsymbol{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$$

Beispiel 7.1.2
Der Zustandsraum sind die vier Basen (Adenin, Cytosin, Guanin, Thymin) der DNA, d.h. $S = \{A,C,G,T\}$. Die folgende geschätzte Übergangsmatrix[1]

[1] Durbin *et al.* (1998) "Biological sequence analysis", p. 50

kann erhalten werden, wenn man beim "Ablaufen" der DNA die relativen Häufigkeiten der einzelnen Übergänge zählt.

$$\boldsymbol{P} = \begin{pmatrix} 0.300 & 0.205 & 0.285 & 0.210 \\ 0.322 & 0.298 & 0.078 & 0.302 \\ 0.248 & 0.246 & 0.298 & 0.208 \\ 0.177 & 0.239 & 0.292 & 0.292 \end{pmatrix}$$

$\boldsymbol{P}$ beschreibt also die **Kurzzeitentwicklung** (Ein-Schritt-Übergangswahrscheinlichkeiten) einer homogenen Markov-Kette $\boldsymbol{X}$. Die **Langzeitentwicklung** (n-Schritt-Übergangswahrscheinlichkeiten) ist durch **die n-Schritt-Übergangsmatrix** $\boldsymbol{P}(m, m+n)$ mit Elementen

$$p_{ij}(m, m+n) \quad = \quad P(X_{m+n} = j | X_m = i)$$

gekennzeichnet. Für eine homogene Markov-Kette vereinfacht sich diese zu

$$p_{ij}(n) \quad = \quad P(X_n = j | X_0 = i).$$

Dann gilt offensichtlich

$$\boldsymbol{P}(m, m+1) = \boldsymbol{P}$$

und wir schreiben $\boldsymbol{P}_n = \boldsymbol{P}(m, m+n)$.

Wie aber berechnet man $\boldsymbol{P}_n$? Für eine homogene Markov-Kette $\boldsymbol{X}$ gilt:

$p_{ij}(m, m+n+r)$

$$\begin{aligned}
&= && \boldsymbol{P}(X_{m+n+r} = j | X_m = i) \\
&= && \sum_k \boldsymbol{P}(X_{m+n+r} = j, X_{m+n} = k | X_m = i) \\
&= && \sum_k \boldsymbol{P}(X_{m+n+r} = j | X_{m+n} = k, X_m = i) \cdot \boldsymbol{P}(X_{m+n} = k | X_m = i) \\
&\overset{Markov}{=} && \sum_k \boldsymbol{P}(X_{m+n+r} = j | X_{m+n} = k) \cdot \boldsymbol{P}(X_{m+n} = k | X_m = i) \\
&= && \sum_k p_{ik}(m, m+n) p_{kj}(m+n, m+n+r)
\end{aligned}$$

Schreibt man diese Beziehung in Matrizenform so erhält man die **Chapman-Kolmogorov-Gleichung**:

$$\boldsymbol{P}(m, m+n+r) = \boldsymbol{P}(m, m+n) \cdot \boldsymbol{P}(m+n, m+n+r)$$

Nun kann man auch $\boldsymbol{P}_n$ berechnen:

$$\begin{aligned}\boldsymbol{P}_n &= \boldsymbol{P}(m, m+n)\\ &= \boldsymbol{P}(m, m+1+(n-1))\\ &= \boldsymbol{P}(m, m+1)\cdot\boldsymbol{P}(m+1, m+n)\\ &= \boldsymbol{P}\cdot\boldsymbol{P}(m+1, m+n)\\ &= \ldots = \boldsymbol{P}^n\end{aligned}$$

Hierbei ist $\boldsymbol{P}^n$ die n-te Potenz von $\boldsymbol{P}$. Das heißt die n-Schritt-Übergangsmatrix $\boldsymbol{P}_n$ erhält man durch n-faches Potenzieren der Matrix $\boldsymbol{P}$.

Beispiel 7.1.3 (2-Schritt-Übergangsmatrix für Beispiel 7.1.1)
Die 2-Schritt-Übergangsmatrix $\boldsymbol{P}_2$ entspricht der 2-ten Potenz von $\boldsymbol{P}$:

$$\boldsymbol{P}_2 = \boldsymbol{P}^2 = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix} \cdot \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.85 & 0.15 \\ 0.6 & 0.4 \end{pmatrix}$$

Dabei berechnet sich z.B. der Eintrag $p_{21}(2)$ in P_2 folgendermaßen:

$$\begin{aligned}&P(X_{n+2}=1|X_n=2)\\ &= P(X_{n+1}=2|X_n=2)\cdot P(X_{n+2}=1|X_{n+1}=2)\\ &\quad + P(X_{n+1}=1|X_n=2)\cdot P(X_{n+2}=1|X_{n+1}=1)\\ &= 0.6\cdot 0.4 + 0.4\cdot 0.9 = 0.6\end{aligned}$$

Sei nun $\boldsymbol{X}$ eine Markov-Kette mit Übergangsmatrix $\boldsymbol{P}$, welche sich zum Zeitpunkt t im Zustand $i \in S$ befindet. Die **Dauer bis zum nächsten Zustandswechsel** T_i, d.h. die Dauer bis die Markov-Kette in einen anderen Zustand als i übergeht, ist offensichtlich *geometrisch verteilt* mit Parameter $1 - p_{ii}$:

$$T_i \sim \mathcal{G}(1 - p_{ii}).$$

Später werden wir diese nützliche Eigenschaft für einen Modellanpassungstest verwenden, bei dem wir die theoretischen Dauern bis zum nächsten Zustandswechsel (berechnet auf der Basis des geschätzten Markov-Modells) mit den beobachteten Dauern vergleicht.

Um die Entwicklung einer Markov-Kette vollständig zu spezifizieren, muss neben der Übergangsmatrix $\boldsymbol{P}$ noch die sogenannte **Anfangsverteilung** für X_0 angegeben werden, in der festgelegt wird, mit welcher Wahrscheinlichkeit die Markov-Kette in dem jeweiligen Zustand startet. Die Wahrscheinlichkeitsfunktion für X_0 bezeichnet man mit dem Zeilenvektor $\boldsymbol{\mu^{(0)}}$ mit den Elementen $\mu_i^{(0)} = P(X_0 = i)$ für $i \in S$.

Die unbedingte Wahrscheinlichkeitsfunktion von X_n fasst man entsprechend in dem Zeilenvektor $\boldsymbol{\mu}^{(n)}$ zusammen und es gilt offensichtlich:

$$\boldsymbol{\mu}^{(m+n)} = \boldsymbol{\mu}^{(m)} \cdot \boldsymbol{P}_n$$

und somit:

$$\boldsymbol{\mu}^{(n)} = \boldsymbol{\mu}^{(0)} \cdot \boldsymbol{P}^n$$

Die **gemeinsame** Wahrscheinlichkeitsverteilung von $X_0, \ldots, X_n$ ist gegeben durch

$$\begin{aligned} P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) & \\ &= P(X_0 = x_0) \prod_{t=1}^{n} P(X_t = x_t | X_{t-1} = x_{t-1}) \\ &= \mu_{x_0}^{(0)} \prod_{t=1}^{n} p_{x_{t-1}, x_t} \end{aligned}$$

Diese ist wichtig für die Schätzung der Übergangsmatrix bei Beobachtung einer Realisation aus einer Markov-Kette.
Im Folgenden werden nun drei Beispiele für Markov-Ketten behandelt:

Beispiel 7.1.4 (Inzucht)
"Inzucht" bedeutet hier, dass eine Pflanze, welche einen der drei Genotypen $\{aa, ab, bb\}$ besitzen kann, mit sich selbst gekreuzt wird. Die Zufallsvariable X_n gibt den Genotyp in der n-ten Generation an.
Die Markov-Kette hat damit den Zustandsraum:

$$S = \{aa, ab, bb\}$$

und die folgende Übergangsmatrix

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}$$

Die Übergänge von ab zu den anderen Zuständen lassen sich beispielsweise so berechnen:

Man kreuzt $ab \times ab$ und erhält für die verschiedenen Keimzellkombinationen folgende Genotypen für die Tochtergeneration:

$\times$	**a**	**b**
a	aa	ab
b	ab	bb

Die Nachkommen haben also mit einer Wahrscheinlichkeit von

- 1/4 *den Genotyp* aa
- 1/2 *den Genotyp* ab
- 1/4 *den Genotyp* bb

Das entspricht genau den Übergängen in $\boldsymbol{P}$.
Man kann zeigen (im Anschluß numerisch), dass

$$\boldsymbol{P}^n = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} - \left(\frac{1}{2}\right)^{n+1} & \left(\frac{1}{2}\right)^n & \frac{1}{2} - \left(\frac{1}{2}\right)^{n+1} \\ 0 & 0 & 1 \end{pmatrix} \quad \stackrel{n\to\infty}{\longrightarrow} \quad \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}$$

d.h. letztendlich bleiben nur die Genotypen aa *und* bb *übrig, die sich dann deterministisch reproduzieren. Um dies zu sehen, berechnen wir zunächst die Eigenwerte und Eigenvektoren von* $\boldsymbol{P}$*:*

```
> P <- matrix(c(1, 0.25, 0, 0, 1/2, 0, 0, 0.25, 1), nc = 3)
> eP <- eigen(P)
> Q <- eP$vectors
> D <- diag(eP$values)
```

Dann läßt sich $\boldsymbol{P}$ *schreiben als* $\boldsymbol{P} = \boldsymbol{QDQ}^{-1}$

```
> Q %*% D %*% solve(Q)
```

```
     [,1] [,2] [,3]
[1,] 1.00  0.0 0.00
[2,] 0.25  0.5 0.25
[3,] 0.00  0.0 1.00
```

und $\boldsymbol{P}^n = \boldsymbol{QD}^n\boldsymbol{Q}^{-1}$*, die für* $n \to \infty$ *gegen die oben gezeigte Matrix konvergiert:*

```
> round(Q %*% D^100 %*% solve(Q), 2)
```

```
     [,1] [,2] [,3]
[1,]  1.0    0  0.0
[2,]  0.5    0  0.5
[3,]  0.0    0  1.0
```

Beispiel 7.1.5 (Genhäufigkeit in Population konstanter Größe N)
Die Zufallsvariable $X_n = i$ gibt die Anzahl der Individuen einer Population mit bestimmten Genotyp zum Zeitpunkt n an.
Ein einfaches Modell nimmt an, dass zu jedem Zeitpunkt ein zufälliges Mitglied der Population stirbt. Das "nachrückende" Mitglied hat den betrachteten Genotyp mit Wahrscheinlichkeit $\frac{i}{N}$.
Damit ergibt sich für die Einträge p_{ij} in $\boldsymbol{P}$:

$$p_{ij} = \begin{cases} \frac{i}{N} \cdot \frac{(N-i)}{N} = \frac{i(N-i)}{N^2} & \text{für } j = i \pm 1 \\ 1 - 2\frac{i(N-i)}{N^2} & \text{für } j = i \\ 0 & \text{sonst} \end{cases}$$

$\boldsymbol{P}$ hat eine "tri-diagonale" Struktur, d.h. die Hauptdiagonale und zwei Nebendiagonalen enthalten Werte ≥ 0 und sonst enthält die Matrix nur Nullen. Die obige Formulierung beinhaltet auch die beiden Grenzfälle $X_n = 0$ und $X_n = N$.

Beispiel 7.1.6 (Modelle für Epidemien $\rightarrow$ Verzweigungsprozesse)
Die Zufallsvariable X_n gibt die Anzahl der Infizierten in einer Generation n an. Der Zustandsraum S ist

$$S = \{0, 1, \ldots\}$$

Die Idee des Modells beinhaltet, dass jeder Infizierte (unabhängig von den anderen) eine zufällige Anzahl Infizierter in der nächsten Generation "erzeugt" und zwar mit Erwartungswert λ, z.B. könnte die Anzahl Poissonverteilt sein. Dann ist $X_n | X_{n-1} \sim \mathcal{P}(\lambda \cdot X_{n-1})$.
Ein wichtiger Satz ist das Schwellenwerttheorem, das wir hier nur verkürzt wiedergeben:
Für $\lambda < 1$ wird jede Epidemie mit Wahrscheinlichkeit 1 irgendwann aussterben, d.h. X_n wird für großes n gleich Null sein. Die Wahrscheinlichkeit dass der Prozess "explodiert" ist also gleich Null. Für $\lambda > 1$ hingegen ist die Wahrscheinlichkeit, dass die Epidemie explodiert, echt größer Null.

7.2 Klassifikation von Zuständen und Markov-Ketten

Sei $T_i = \min\{n \in \mathbb{N} : X_n = i | X_0 = i\}$ die Rekurrenzzeit für i, also die Zeit, die benötigt wird, um zum Zustand i zurückzukehren.

Definition 7.2.1 (Rekurrenz)
Ein Zustand $i \in S$ einer Markov-Kette heißt **rekurrent** *oder auch* **persistent**, *falls*

$$P(T_i < \infty) = 1$$

Ansonsten (also bei $P(T_i < \infty) < 1$) heißt der Zustand **transient**. *Eine Markov-Kette $\boldsymbol{X}$ kehrt also in einen rekurrenten Zustand mit Wahrscheinlichkeit 1 zurück. Ein Zustand i heißt* ***absorbierend***, *falls die Markov-Kette sobald sie diesen Zustand eingenommen hat, ihn nicht mehr verlassen kann, d.h. wenn $p_{ii} = 1$ und $p_{ij} = 0$ für alle $j \neq i$.*

Beispiel 7.2.1
Die beiden Zustände der Markov-Kette mit folgender Übergangsmatrix sind rekurrent:

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}$$

Dagegen sind die Zustände 1 und 2 bei der Übergangsmatrix der nächsten Markov-Kette rekurrent und die anderen beiden transient, denn falls z.B. die Markov-Kette vom Zustand 3 in den Zustand 2 übergeht, kann sie nicht mehr in den Zustand 3 zurückkehren und wenn wir von 4 nach 3 und dann nach 2 wechseln, ist 4 ebenfalls nicht mehr erreichbar.

$$P = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 \\ 0 & 0.3 & 0.5 & 0.2 \\ 0 & 0 & 0.9 & 0.1 \end{pmatrix}$$

In Beispiel 7.1.6 (Poisson-Verzweigungsprozess) ist der Zustand 0 **rekurrent**, ja sogar **absorbierend**, da $\boldsymbol{X}$ diesen Zustand nie verlässt. Alle anderen Zustände sind **transient**.
Hat man einen Zustand als rekurrent klassifiziert, so interessiert auch noch die Frage, wie lange es dauert, bis die Markov-Kette wieder in diesen Zustand zurückkehrt.

Definition 7.2.2 (Erwartete Rekurrenzzeit)
Die **erwartete Rekurrenzzeit** *eines Zustands i ist:*

$$\mu_i = E(T_i) = \begin{cases} \sum_n n f_i(n) & \text{falls } i \text{ rekurrent ist} \\ \infty & \text{falls } i \text{ transient ist} \end{cases}$$
$$\text{mit } f_i(n) = P(X_1 \neq i, X_2 \neq i, \ldots, X_{n-1} \neq i, X_n = i | X_0 = i)$$

Ein rekurrenter Zustand i heißt **nicht-leer**, *falls seine erwartete Rekurrenzzeit endlich ist. Ansonsten heißt er* **leer**.

Zwischen den Übergangswahrscheinlichkeiten $p_{ij}(n)$ und der Leerheit eines rekurrenten Zustands besteht folgender Zusammenhang: Ein rekurrenter Zustand i ist genau dann leer, wenn

$$p_{ii}(n) \to 0 \text{ für } n \to \infty$$

Dann gilt sogar

$$p_{ji}(n) \to 0 \text{ für } n \to \infty \text{ für alle } j \in S$$

Definition 7.2.3 (Periode)
Die **Periode** ξ *eines Zustandes i ist der größte gemeinsame Teiler der Menge*

$$\{n : p_{ii}(n) > 0\}$$

Man nennt den Zustand i **periodisch**, *falls $\xi > 1$ und* **aperiodisch** *falls $\xi = 1$.*
Haben alle Zustände einer Markov-Kette Periode 1, so heißt sie aperiodisch. Ein Zustand i heißt **ergodisch**, *falls er rekurrent nicht-leer und aperiodisch ist. Sind alle Zustände von $\boldsymbol{X}$ ergodisch, so heißt $\boldsymbol{X}$ ergodisch.*

Beispiel 7.2.2
Die drei Zustände der Markov-Kette mit der folgenden Übergangsmatrix haben Periode 3, da die Markov-Kette nach genau drei Schritten wieder in den ursprünglichen Zustand zurückkehrt. Die Markov-Kette ist damit nicht aperiodisch.

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Bei der Markov-Kette mit der nächsten Übergangsmatrix sind beide Zustände aperiodisch und damit ist auch die Markov-Kette aperiodisch.

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 \\ 0.3 & 0.7 \end{pmatrix}$$

Definition 7.2.4 (Irreduzible Zustände)
Zwei Zustände $i \neq j$ einer Markov-Kette $\boldsymbol{X}$ **kommunizieren** *miteinander, falls $f_{ij} > 0$ und $f_{ji} > 0$. Schreibweise: $i \leftrightarrow j$*
Dabei ist $f_{ij} = \sum_{n=1}^{\infty} f_{ij}(n) = \sum_{n=1}^{\infty} P(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j | X_0 = i)$.
Ein Zustand i kommuniziert per definitionem immer mit sich selber: $i \leftrightarrow i$

Eine Menge $C \subset S$ heißt **irreduzibel**, *falls $i \leftrightarrow j$ für alle $i, j \in C$.*
Eine Menge $C \subset S$ heißt **geschlossen**, *falls $p_{ij} = 0$ für alle $i \in C$ und $j \in \bar{C}$.*

Das bedeutet, dass eine geschlossene Teilmenge des Zustandsraumes nicht verlassen werden kann.

Beispiel 7.2.3
Folgende Markov-Ketten haben einen reduziblen Zustandsraum:

$$\boldsymbol{P_1} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{P_2} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 \end{pmatrix}$$

In $\boldsymbol{P_2}$ zum Beispiel "kommunizieren" nicht alle Zustände miteinander: Die Wahrscheinlichkeit vom Zustand 3 in Zustand 2 oder 1 zu kommen ist 0.
Dagegen ist der Zustandsraum S der Markov-Kette $\boldsymbol{P_3}$ irreduzibel, da die Wahrscheinlichkeit von einem Zustand i irgendwann nach Zustand j zu kommen echt positiv ist für alle $i, j \in S$:

$$\boldsymbol{P_3} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

Satz 7.2.1 (Zerlegungssatz)
Der Zustandraum S einer Markov-Kette $\boldsymbol{X}$ lässt sich zerlegen in

1. *eine Menge T mit transienten Zuständen und*
2. *mehrere Mengen C_k, die irreduzibel und geschlossen sind.*

Also

$$S = T \cup C_1 \cup C_2 \cup \ldots$$

Ferner gilt folgendes ***Lemma****:*
Wenn S endlich ist, dann ist mindestens ein Zustand rekurrent, und alle rekurrenten Zustände sind nicht-leer.

Beispiel 7.2.4
Eine Markov-Kette habe den Zustandsraum $S = \{1, 2, 3, 4, 5, 6\}$ *und die Übergangsmatrix*

$$\boldsymbol{P} = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0.25 & 0.75 & 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 & 0 & 0 \\ 0.25 & 0 & 0.25 & 0.25 & 0 & 0.25 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix}$$

Bestimme nun die Periode jedes Zustands, transiente Zustände und ergodische Zustände, sowie die Zerlegung des Zustandsraumes:

1. *alle Zustände haben die Periode 1 (die Hauptdiagonale ist besetzt), also ist die Markov-Kette aperiodisch*

2. *die Zustände 3 und 4 sind transient (4 nach 6 ist eine Einbahnstraße und 3 nach 4 nach 6 auch) und die anderen sind rekurrent*

3. *die Zustände 1,2,5 und 6 sind ergodisch (aperiodisch und* $E(T_i) < \infty$*). Man zeigt, dass* $E(T_i) = \sum n f_i(n)$ *konvergent ist mittels des Quotientenkriteriums (Reihe* $a_1 + a_2 + \ldots$ *ist konvergent wenn Grenzwert von* $a_{n+1}/a_n < 1$*). Zunächst* T_6*:*

$$\begin{aligned}
E(T_6) &= \sum_{n=1}^{\infty} nP(X_1 = 5, X_2 = 5, \ldots, X_n = 6 | X_0 = 6) \\
&= \sum_{n=1}^{\infty} nP(X_0 = 6, X_1 = 5, X_2 = 5, \ldots, X_n = 6)/P(X_0 = 6) \\
&= P(X_1 = 6|X_0 = 6) + 2P(X_1 = 5|X_0 = 6)P(X_2 = 6|X_1 = 5) + \\
&\quad \sum_{n=3}^{\infty} nP(X_1 = 5|X_0 = 6) \left[\prod_{t=2}^{n-1} P(X_t = 5|X_{t-1} = 5) \right] \cdot \\
&\quad P(X_n = 6|X_{n-1} = 5) \qquad \textit{siehe Seite 97} \\
&= \sum_{n=1}^{\infty} n \prod_{t=1}^{n} \frac{1}{2} \\
&= \sum_{n=1}^{\infty} n \frac{1}{2^n}
\end{aligned}$$

oder

$$
\begin{aligned}
E(T_1) &= \sum_{n=1}^{\infty} nP(X_1 = 2, X_2 = 2, \ldots, X_n = 1 | X_0 = 1) \\
&= \sum_{n=1}^{\infty} nP(X_0 = 1, X_1 = 2, X_2 = 2, \ldots, X_n = 1)/P(X_0 = 1) \\
&= P(X_1 = 1|X_0 = 1) + 2P(X_1 = 2|X_0 = 1)P(X_2 = 1|X_1 = 2) + \\
&\quad \sum_{n=3}^{\infty} nP(X_1 = 2|X_0 = 1) \left[\prod_{t=2}^{n-1} P(X_t = 2|X_{t-1} = 2) \right] \cdot \\
&\quad P(X_n = 1|X_{n-1} = 2) \\
&= \frac{1}{2} + \sum_{n=2}^{\infty} n \frac{1}{2} \left(\frac{3}{4} \right)^{n-2} \frac{1}{4} \\
&= \frac{1}{2} + \frac{1}{8} \sum_{n=0}^{\infty} (n+2) \left(\frac{3}{4} \right)^{n}
\end{aligned}
$$

Damit für T_6:

$$\lim_{n\to\infty} \left(\frac{n+1}{2^{n+1}} \frac{2^n}{n} \right) = \lim_{n\to\infty} \frac{1 + n^{-1}}{2} = \frac{1}{2} < 1 \to E(T_6) < \infty$$

4. *Zerlegung des Zustandsraumes $S = T \cup C_1 \cup C_2$ mit*
   $T = \{3, 4\}$
   $C_1 = \{1, 2\}$
   $C_2 = \{5, 6\}$

7.3 Die stationäre Verteilung und das Grenzwerttheorem

Definition 7.3.1
Eine Wahrscheinlichkeitsverteilung $\boldsymbol{\pi}$ *(Zeilenvektor) mit Einträgen* $(\pi_j : j \in S)$ *heißt* ***stationäre Verteilung*** *einer Markov-Kette* $\boldsymbol{X}$ *mit Übergangsmatrix* $\boldsymbol{P}$*, falls gilt:*

$$\pi_j = \sum_i \pi_i p_{ij}$$

oder in Matrixnotation:

$$\boldsymbol{\pi} = \boldsymbol{\pi} \cdot \boldsymbol{P} \tag{7.1}$$

Interpretation der stationären Verteilung: Hat die Markov-Kette $\boldsymbol{X}$ die Verteilung $\boldsymbol{\pi}$ zu einem gewissen Zeitpunkt n, dann auch im nächsten Zeitpunkt $n+1$ und sogar in allen nachfolgenden Zeitpunkten $i = n+2, n+3, \ldots$
Denn es gilt z.B. für $i = n+2$:

$$\boldsymbol{\pi} \cdot \boldsymbol{P^2} \;=\; (\boldsymbol{\pi P})\boldsymbol{P} \overset{(7.1)}{=} \boldsymbol{\pi P} \overset{(7.1)}{=} \boldsymbol{\pi}$$

Oft wählt man für die Anfangsverteilung $\boldsymbol{\mu}_0$ die stationäre Verteilung, d.h. $\boldsymbol{\mu}_0 = \boldsymbol{\pi}$.
Im folgenden betrachten wir ausschließlich irreduzible Markov-Kette (Zustandsraum S ist irreduzibel).

Satz 7.3.1 (Satz über die stationäre Verteilung)
Eine irreduzible Markov-Kette hat eine **stationäre Verteilung** $\boldsymbol{\pi}$ *genau dann, wenn alle Zustände nicht-leer rekurrent sind. In diesem Fall ist* $\boldsymbol{\pi}$ *eindeutig und gegeben durch*

$$\pi_i = 1/\mu_i$$

wobei μ_i *die erwartete Rekurrenzzeit des Zustands* i *ist.*
Bei endlichem Zustandsraum gilt dann:

$$\boldsymbol{\pi} = \mathbf{1}(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{U})^{-1}$$

wobei $\boldsymbol{I}$ *die Einheitsmatrix und* $\mathbf{1}$ *ein Zeilenvektor mit Einsen ist und die Matrix* $\boldsymbol{U}$ *nur Elemente gleich eins hat.*

Für eine Markov-Kette mit 2 Zuständen lässt sich die stationäre Verteilung leicht durch eine Formel ermitteln:
Sei die noch unbekannte stationäre Verteilung gegeben durch

$$\pi = (\pi_1, \pi_2) = (\pi_1, 1 - \pi_1)$$

und die Übergangsmatrix ist:

$$\boldsymbol{P} = \begin{pmatrix} 1-p_{12} & p_{12} \\ p_{21} & 1-p_{21} \end{pmatrix}$$

Um $\boldsymbol{\pi}$ zu berechnen wendet man die Formel $\boldsymbol{\pi} = \boldsymbol{\pi} \cdot \boldsymbol{P}$ an:

$$\boldsymbol{\pi} = \boldsymbol{\pi} \cdot \boldsymbol{P} \Leftrightarrow (\pi_1, 1-\pi_1) = (\pi_1, 1-\pi_1) \cdot \begin{pmatrix} 1-p_{12} & p_{12} \\ p_{21} & 1-p_{21} \end{pmatrix}$$

In der ersten Spalte der Gleichung ergibt sich für π_1:

$$\begin{aligned} \pi_1 &= \pi_1(1-p_{12}) + (1-\pi_1)p_{21} \\ &= \pi_1 - \pi_1 p_{12} + p_{21} - \pi_1 p_{21} \\ &= \frac{p_{21}}{p_{12}+p_{21}} \end{aligned}$$

Damit lässt sich auch π_2 berechnen:

$$\begin{aligned} \pi_2 &= 1-\pi_1 \\ &= 1 - \frac{p_{21}}{p_{12}+p_{21}} \\ &= \frac{p_{12}}{p_{12}+p_{21}} \end{aligned}$$

Die zweite Spalte ergibt die gleiche Lösung:

$$\boldsymbol{\pi} = (\pi_1, \pi_2) = \left(\frac{p_{21}}{p_{12}+p_{21}}, \frac{p_{12}}{p_{12}+p_{21}} \right) \tag{7.2}$$

Beispiel 7.3.1 (Berechnung einer stationären Verteilung)
Für die Markov-Kette mit folgender Übergangsmatrix

$$\boldsymbol{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$$

soll die zugehörige stationäre Verteilung berechnet werden.
Auf einfachstem Wege lässt sich dieses Ziel erreichen, wenn man die Werte in die eben ausgerechnete Formel (7.2) einsetzt:

$$\boldsymbol{\pi} = (\pi_1, \pi_2) = \left(\frac{0.4}{0.5}, \frac{0.1}{0.5} \right) = (0.8, 0.2)$$

Die erwarteten Rekurrenzzeiten sind demnach $\mu_1 = 5/4$ für Zustand 1 und $\mu_2 = 5$ für Zustand 2.
Eine weitere Möglichkeit zur Berechnung wurde im Satz 7.3.1 beschrieben:

$$\begin{aligned} \boldsymbol{\pi} &= \mathbf{1}(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{U})^{-1} \\ &= (1,1) \cdot \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right]^{-1} \end{aligned}$$

Weitere Rechnung in R:

```
> I <- diag(2)
> P <- matrix(c(0.9, 0.4, 0.1, 0.6), nrow = 2)
> U <- matrix(1, nrow = 2, ncol = 2)
> rep(1, 2) %*% solve(I - P + U)

     [,1] [,2]
[1,]  0.8  0.2
```

Wie sieht die stationäre Verteilung für die Markov-Kette mit dieser Übergangsmatrix aus:

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Alle Zustände sind periodisch mit Periode 3:

```
> P <- diag(3)[,c(3,1,2)]
> P %*% P

     [,1] [,2] [,3]
[1,]    0    0    1
[2,]    1    0    0
[3,]    0    1    0

> P %*% P %*% P

     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

die erwartete Rekurrenzzeit beträgt damit

$$E(T_i) = \sum_n n f_i(n) = 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 1 + 4 \cdot 0 + \cdots = 3.$$

Daher ergibt sich für $\boldsymbol{\pi} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

Reversible Markov-Ketten

Sei $\boldsymbol{X} = (X_0, \ldots, X_N)$ eine reguläre Markov-Kette mit Übergangsmatrix $\boldsymbol{P}$ und stationärer Verteilung $\boldsymbol{\pi}$, die $\boldsymbol{X}$ auch zu jedem Zeitpunkt $n = 0, \ldots, N$ besitze.

Definiere nun $\boldsymbol{Y} = (X_N, \ldots, X_0)$ mit $Y_n = X_{N-n}$. Dann ist $\boldsymbol{Y}$ auch eine Markov-Kette mit den Übergangswahrscheinlichkeiten

$$P(Y_{n+1} = j | Y_n = i) = (\pi_j / \pi_i) p_{ji} \tag{7.3}$$

Beweis der Markov-Eigenschaft für Y:
Sei also $Y_n = X_{N-n}$. Dann gilt:

$$\begin{aligned}
&P(Y_{n+1} = i_{n+1} | Y_n = i_n, \ldots, Y_0 = i_0) \\
&= \frac{P(Y_k = i_k, 0 \leq k \leq n+1)}{P(Y_k = i_k, 0 \leq k \leq n)} \\
&= \frac{P(X_{N-(n+1)} = i_{n+1}, X_{N-n} = i_n, \ldots, X_N = i_0)}{P(X_{N-n} = i_n, \ldots, X_n = i_0)} \\
&= \frac{P(X_{N-(n+1)} = i_{n+1}, X_{N-n} = i_n, \ldots, X_N = i_0)}{P(X_{N-n} = i_n) P(X_{N-(n-1)} = i_{n-1} | X_{N-n} = i_n) \ldots P(X_N = i_0 | X_{N-1} = i_1)} \\
&= \frac{\pi_{i_{n+1}} \cdot p_{i_{n+1} i_n} \cdot \ldots \cdot p_{i_1 i_0}}{\pi_{i_n} \cdot p_{i_n i_{n-1}} \cdot \ldots \cdot p_{i_1 i_0}} \\
&= \frac{\pi_{i_{n+1}}}{\pi_{i_n}} \cdot p_{i_{n+1} i_n}
\end{aligned}$$

$\Rightarrow Y$ ist Markov-Kette

Definition 7.3.2
Man sagt nun $\boldsymbol{X}$ ist **reversibel**, *falls X und Y identische Übergangswahrscheinlichkeiten haben, d.h. falls $\forall\, i, j \in S$ gilt:*

$$\pi_i p_{ij} = \pi_j p_{ji}$$

Beispiel 7.3.2 (Reversible Markov-Kette)
Sei die Übergangsmatrix der Markov-Kette $\boldsymbol{X}$ gegeben als

$$\boldsymbol{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$$

von der bereits die stationäre Verteilung bekannt ist:

$$\boldsymbol{\pi} = (0.8, 0.2)$$

Berechnet man die Übergangsmatrix $\boldsymbol{P'}$ von der Markov-Kette $\boldsymbol{Y}$ (nach 7.3)

$$\begin{aligned}
\boldsymbol{P'} &= \begin{pmatrix} \frac{0.8}{0.8} \cdot 0.9 & \frac{0.2}{0.8} \cdot 0.4 \\ \frac{0.8}{0.2} \cdot 0.1 & \frac{0.2}{0.2} \cdot 0.6 \end{pmatrix} \\
&= \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}
\end{aligned}$$

so stellt man fest, dass $\boldsymbol{P}$ *und* $\boldsymbol{P'}$ *übereinstimmen. Daher ist die Markov-Kette* $\boldsymbol{X}$ *reversibel.*

Satz 7.3.2
Alle irreduziblen Markov-Ketten mit zwei Zuständen sind reversibel.

Beweis:
z.z. $\pi_i \cdot p_{ij} = \pi_j \cdot p_{ji} \qquad \forall i, j \in \{1, 2\}$
Wie bereits bewiesen, haben π_1 und π_2 folgende Werte:

$$\pi_1 = \frac{p_{21}}{p_{12} + p_{21}} \text{ und } \pi_2 = \frac{p_{12}}{p_{12} + p_{21}}$$

- Fall (a) $i = j = 1$ ist trivial
- Fall (b) $i = j = 2$ ebenfalls trivial
- Fall (c) $i = 1, j = 2$: $p_{12} \cdot \pi_1 = p_{12} \cdot \frac{p_{21}}{p_{12}+p_{21}} = p_{21} \cdot \frac{p_{12}}{p_{12}+p_{21}} = p_{21} \cdot \pi_2$
- Fall (d) $i = 2, j = 1$ ist analog Fall (c)

Außerdem sind Markov-Ketten mit tri-diagonaler (nur die Haupt- und zwei Nebendiagonalen sind besetzt, alle anderen Werte sind 0) Übergangsmatrix $\boldsymbol{P}$ reversibel, z.B. der **random walk** auf endlichem Zustandsraum $S = \{0, 1, \ldots, b\}$ oder der Prozess aus Beispiel 7.1.5 (Stichwort: Genhäufigkeit in Population konstanter Größe).

Satz 7.3.3
Sei $\boldsymbol{X}$ *eine irreduzible Markov-Kette mit Übergangsmatrix* $\boldsymbol{P}$*. Ferner gebe es eine Verteilung* $\boldsymbol{\pi}$ *mit* $\pi_i p_{ij} = \pi_j p_{ji}$ *für alle* $i, j \in S$*.*
Dann ist $\boldsymbol{\pi}$ *die stationäre Verteilung und* $\boldsymbol{X}$ *ist bzgl.* $\boldsymbol{\pi}$ *reversibel.*

Beweis:

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$$

Satz 7.3.4 (Grenzwerttheorem)
Eine irreduzible und aperiodische *Markov-Kette konvergiert gegen ihre stationäre Verteilung* $\boldsymbol{\pi}$ *für* $n \to \infty$ *und alle* i*:*

$$p_{ij}(n) \longrightarrow \pi_j = \mu_j^{-1}$$

bzw.

$$\boldsymbol{P}_n = \boldsymbol{P}^n \longrightarrow \begin{pmatrix} \cdots & \boldsymbol{\pi} & \cdots \\ \cdots & \boldsymbol{\pi} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \boldsymbol{\pi} & \cdots \end{pmatrix}$$

und daher $\boldsymbol{\mu}^{(0)}\boldsymbol{P}_n \longrightarrow \boldsymbol{\pi}$ *für alle* $\boldsymbol{\mu}^{(0)}$*, da gilt:*

$$\begin{aligned} \boldsymbol{\mu}^{(0)}\boldsymbol{P}_n &= (\mu_1, \ldots, \mu_m)\boldsymbol{P}_n \\ &\overset{n\to\infty}{=} \left(\left(\sum_i \boldsymbol{\mu}_i\right)\pi_1, \ldots, \left(\sum_i \boldsymbol{\mu}_i\right)\pi_m\right) \\ &= (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_m) \\ &= \boldsymbol{\pi} \end{aligned}$$

Wichtig ist es zu betonen, dass Satz 7.3.4 nur für *irreduzible und aperiodische* Markov-Kette gilt. Denn betrachtet man die folgende Markov-Kette $\boldsymbol{X}$ mit Übergangsmatrix

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

dann stellt man fest, dass diese Kette periodisch (Periode 3) ist und damit die stationäre Verteilung $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$ hat. Für $n \to \infty$ konvergiert $\boldsymbol{P}$ aber nicht gegen $\boldsymbol{\pi}$.

7.4 Inferenz für Markov-Ketten

Mit Hilfe der Inferenz können bei einer Markov-Kette $\boldsymbol{X}$ z.B. die Übergangswahrscheinlichkeiten zwischen den einzelnen Zuständen (Einträge in der $S \times S$ Übergangsmatrix $\boldsymbol{P}$) basierend auf einer (oder mehrerer) Realisationen von $\boldsymbol{X}$ geschätzt werden. Dabei scheint es plausibel zu sein, die Übergangswahrscheinlichkeiten p_{ij} durch die entsprechenden **Übergangshäufigkeiten**

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i}$$

zu schätzen, wobei n_{ij} die Anzahl der beobachteten Übergänge von Zustand i nach j $(i, j \in S)$ ist und $n_i = \sum_j n_{ij}$.

Diese intuitiven Schätzer sind auch die ML-Schätzer, was im folgenden hergeleitet wird. Grundlage ist die Realisation $X_0 = x_0, X_1 = x_1, \ldots, X_N = x_N$ von $\boldsymbol{X}$. Die Likelihood berechnet sich somit als

$$L(\boldsymbol{P}) = \mu_{x_0}^{(0)} \prod_{t=1}^{n} p_{x_{t-1},x_t} = \mu_{x_0}^{(0)} \prod_{i,j} p_{ij}^{n_{ij}}$$

Die Log-Likelihood lautet dann

$$l(\boldsymbol{P}) = \log(\mu_{x_0}^{(0)}) + \sum_{i,j} n_{ij} \log(p_{ij})$$

Es tritt nun das Problem auf, dass *mehrere* Parameter in $\boldsymbol{\theta} = \boldsymbol{P}$ durch Maximierung der Log-Likelihood $l(\boldsymbol{P})$ geschätzt werden müssen und noch zusätzlich die Restriktion

$$\sum_j p_{ij} = 1$$

für alle i zu berücksichtigen ist.
Daher wird zur Maximierung unter Nebenbedingungen die **Lagrangesche Multiplikatorenmethode** angewendet. Dabei wird

$$l^*(\boldsymbol{P}) = \log(\mu_{x_0}^{(0)}) + \sum_{i,j} n_{ij} \log(p_{ij}) - \sum_i \lambda_i \left(\sum_j p_{ij} - 1 \right)$$

maximiert, indem die partiellen Ableitungen nach p_{ij}

$$\frac{dl^*(\boldsymbol{P})}{p_{ij}} = \frac{n_{ij}}{p_{ij}} - \lambda_i$$

gebildet werden. Werden diese Ableitungen gleich Null gesetzt, so erhält man

$$n_{ij} = \lambda_i p_{ij}$$

Durch Summation über j folgt

$$\lambda_i = \sum_j n_{ij} = n_i$$

und schließlich

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i}.$$

Damit hat man nun gezeigt, dass die ML-Schätzer der Übergangswahrscheinlichkeiten die relativen Häufigkeiten der Übergänge in der vorliegenden Realisation sind.

7.5 Hidden Markov Modell

In diesem Kapitel werden wir eine Verallgemeinerung von Markov-Ketten, die so genannten **Hidden Markov Modelle** kennenlernen. Hierbei trennt man die im Modell vorhandenen Zustände und die nach außen sichtbaren Ereignisse voneinander (bei Markov-Ketten waren diese identisch).
Anwendungen finden Hidden Markov Modelle in verschiedenen Bereichen wie beispielsweise in der Ökonometrie, der Genetik (DNA-Sequenzierung, Stammbaumanalyse), der Spracherkennung etc.
Die latenten Parameter $\boldsymbol{X} = (X_1, \ldots, X_N)$ folgen einer (homogenen) **Markovkette** mit diskretem Zustandsraum S mit

- Übergangsmatrix $\boldsymbol{P}$ mit $p_{ij} = P(X_t = j | X_{t-1} = i)$
- Anfangsverteilung $\boldsymbol{\pi}$ mit $\pi_i = P(X_0 = i)$ (oft impliziert durch $\boldsymbol{\pi} = \boldsymbol{\pi P}$)

Die Beobachtungen $y_t | x_t = s$ sind **bedingt unabhängig** aus einer Verteilung mit Wahrscheinlichkeitsfunktion/Dichte $f_s(y_t)$ mit Parametern $\boldsymbol{\theta}_s$, welche z.B.:

- eine **diskrete Verteilung** mit Missklassifizierungswahrscheinlichkeiten p_s sein kann, oder auch
- eine **Poissonverteilung** mit Raten λ_s

Bemerkung: Hidden Markov Modelle liefern eine realistischere Modellierung als eine direkte Modellierung der Beobachtung als homogene Markov-Kette.

Beispiel 7.5.1 (Verrauschtes binäres Signal)
Das empfangene Signal entspricht den zur Verfügung stehenden Daten:

$$\boldsymbol{y} = (2\ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 2)$$

Die Größe des Zustandsraumes $|S| = 2$, *da* $S = \{1, 2\}$. *Weiter gilt* $N = 20$. *Außerdem ist die Matrix mit den Übergangswahrscheinlichkeiten für* $\boldsymbol{X}$ *gegeben:*

$$\boldsymbol{P_X} = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}$$

Daraus ergibt sich als stationäre Verteilung:

$$\boldsymbol{\pi} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

Weiter ist die Verteilung von $\boldsymbol{Y}$ *gegeben* $\boldsymbol{X}$ *bekannt:*

$$\begin{aligned} f(y_t = 1|x_t = 1) &= 0.8 \\ f(y_t = 1|x_t = 2) &= 0.2 \quad \textit{(Verrauschung)} \\ f(y_t = 2|x_t = 1) &= 0.2 \quad \textit{(Verrauschung)} \\ f(y_t = 2|x_t = 2) &= 0.8 \end{aligned}$$

Ziel der statistischen Inferenz ist nun die **Restauration** *der Sequenz* $\boldsymbol{X}$*:*

$$\boldsymbol{x} = (? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ?)$$

Bei der **Inferenz mit festen Hyperparametern** gilt, dass $\boldsymbol{P}$, $\boldsymbol{\pi}$ und $\boldsymbol{\theta}$ (Verteilung von $\boldsymbol{Y}$ gegeben $\boldsymbol{X}$) fest sind.
Ziel ist es, eine möglichst genaue **Schätzung** der latenten Zustände $\boldsymbol{x} = (x_1, \ldots, x_N)$ anzugeben. Die **Posteriori-Verteilung** $f(\boldsymbol{x}|\boldsymbol{y})$ berechnet sich wie folgt:

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{y}) &= f(\boldsymbol{x}, \boldsymbol{y})/f(\boldsymbol{y}) \\ &\propto f(\boldsymbol{x}, \boldsymbol{y}) \\ &= \underbrace{\pi_{x_1} \prod_{t=2}^{N} p_{x_{t-1}x_t}}_{f(\boldsymbol{x})} \cdot \underbrace{\prod_{t=1}^{N} f_{x_t}(y_t)}_{f(\boldsymbol{y}|\boldsymbol{x})} \end{aligned}$$

Das Problem, das sich hierbei ergibt, ist, dass es S^N (!) unterschiedliche Sequenzen $\boldsymbol{x}$ gibt. Folgende Methoden liefern einen Posteriori-Modus-Schätzer (MAP-Schätzer):

- **Viterbi-Algorithmus** von *Viterbi* (1967):
  Dies ist ein rekursiver Algorithmus zur Maximierung von $f(\boldsymbol{x}, \boldsymbol{y})$ bzgl. $\boldsymbol{x}$, welcher numerisch sehr effizient ist: $O(|S|^2 \cdot N)$

- **Simulated annealing** von *Kirkpatrick, Gelatt & Vecchi* (1983)

Beispiel 7.5.2 (Fortsetzung von Beispiel 7.5.1)
Für die Daten

$$\boldsymbol{y} = (2 \; 2 \; 2 \; 1 \; 2 \; 2 \; 1 \; 1 \; 1 \; 1 \; 1 \; 2 \; 1 \; 1 \; 1 \; 2 \; 1 \; 2 \; 2 \; 2)$$

ergeben sich folgende Schätzer für die ursprüngliche Nachricht $\boldsymbol{x}$*:*

Schätzer von $\boldsymbol{x}$		*post. Wkeit* $P(\boldsymbol{x}\|\boldsymbol{y})$
$\hat{\boldsymbol{x}}_{MAP_1} =$	(2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 2 2 2)	*0.0304*
$\hat{\boldsymbol{x}}_{MAP_2} =$	(2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 2 2 2 2 2)	*0.0304*
$\hat{\boldsymbol{x}}_{MPM} =$	(2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 2 1 2 2 2)	*0.0135*
$\boldsymbol{y} =$	(2 2 2 1 2 2 1 1 1 1 1 2 1 1 1 2 1 2 2 2)	*0.0027*

Hierbei bezeichnet $\hat{\boldsymbol{x}}_{MPM}$ *den marginalen Posteriori Modus, d.h. jedes* x_i, $i = 1, \ldots, 20$, *in* $\hat{\boldsymbol{x}}_{MPM}$ *hat marginale Posteriori-Wahrscheinlichkeit* $P(x_i|\boldsymbol{y}) > 0.5$.

Seinen nun bestimmte Hyperparameter $\boldsymbol{\theta}$ unbekannt, wie beispielsweise die Übergangsmatrix $\boldsymbol{P}$ oder die Verteilung $f(y_i|x_i)$.
Zum Schätzen der unbekannten Hyperparameter kommen folgende Ansätze zum Tragen:

- **Klassische** Likelihood-Ansätze maximieren die (marginale) Likelihood $f(\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$ bezüglich $\boldsymbol{\theta}$, z.B. mit dem Baum-Welch-Algorithmus (iterativer (EM-)Algorithmus zur Maximierung von $f(\boldsymbol{y}|\boldsymbol{\theta})$) von *Baum et al.* (1970) und *Dempster, Laird & Rubin* (1977).
  Problematisch ist hierbei die Berechnung von $f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$, da die Anzahl der Summanden im Allgemeinen sehr gross ist.

- **Bayesianische** Ansätze verwenden zusätzliche priori-Verteilungen $f(\boldsymbol{\theta})$ und simulieren aus der Posteriori-Verteilung

  $$f(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})$$

  mit Markov-Ketten Monte Carlo (MCMC) Verfahren

Beispiel 7.5.3 (Fortsetzung von den Beispielen 7.5.1 und 7.5.2)
Die Daten waren:

$$\boldsymbol{y} = \quad (2\ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 2)$$

Bisher war $\boldsymbol{P_X}$ *bekannt:*

$$\boldsymbol{P_X} = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix} \text{ bekannt}$$

Seinen nun die Einträge in $\boldsymbol{P_X}$ *unbekannt:*

$$\boldsymbol{P_X} = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{21} & p_{22} \end{pmatrix} \text{ unbekannt}$$

Die marginale Likelihood $L(p_{11}, p_{22})$ *der Diagonalelemente* p_{11} *und* p_{22} *ist in folgender Graphik dargestellt. Die ML-Schätzungen sind* $\hat{p}_{11} = 0.85$ *und* $\hat{p}_{22} = 0.78$.

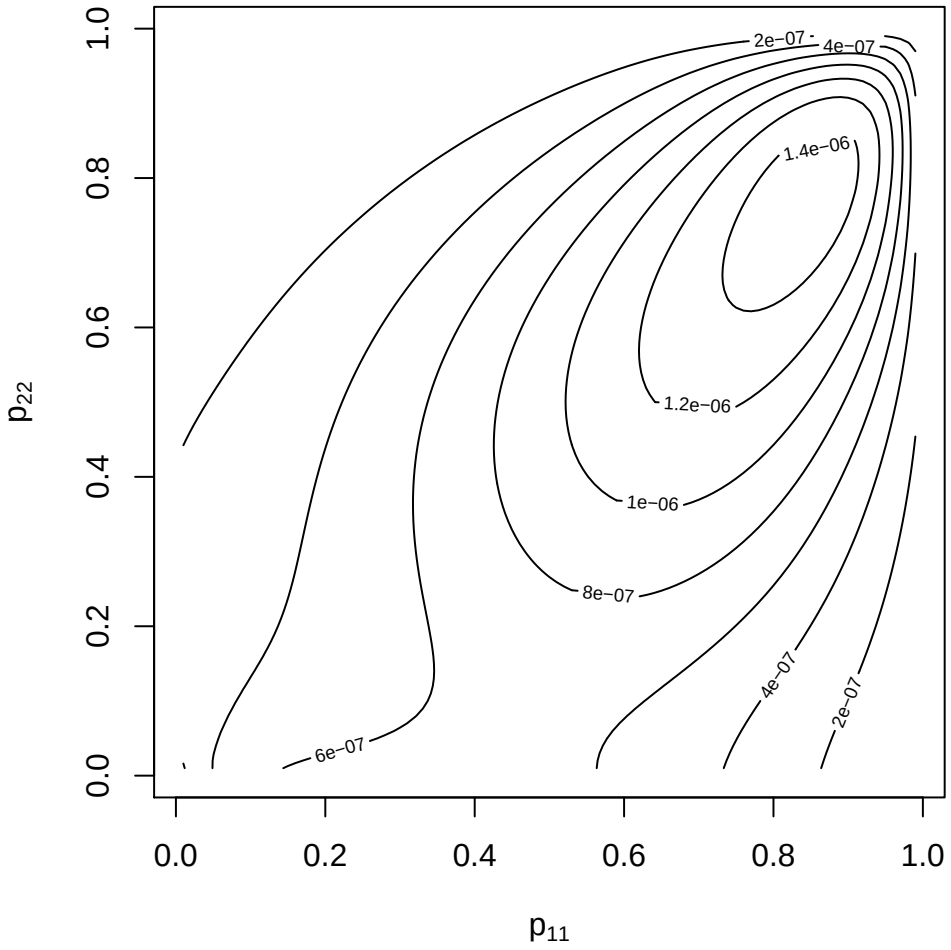

Abbildung 7.1: Marginale Likelihood $L(p_{11}, p_{22})$ dargestellt als Contour-Plot.

Kapitel 8

Stetige Zufallsvariablen

8.1 Definition von stetigen Zufallsvariablen

Idee: Eine Zufallsvariable X heißt **stetig**, falls zu beliebigen Werten $a < b$ aus dem Träger von X auch jeder Zwischenwert in dem Intervall $[a, b]$ möglich ist.
Problem: Wie kann man $P(a \leq X \leq b)$ berechnen, falls alle (also überabzählbar viele) Punkte im Intervall $[a, b]$ möglich sind?

Beispiel 8.1.1 (Glücksrad)
Betrachte ein Glücksrad mit stetigem Wertebereich $[0, 2\pi]$. *Von Interesse ist die Zufallsvariable, die den* exakten *Winkel angibt, an dem das Glücksrad stehen bleibt.*
Aufteilung in 10 *Sektoren, der gleichen Breite. Damit hat jeder Sektor die Wahrscheinlichkeit* $\frac{1}{10}$.

$$P(X \in [0, \pi]) = \frac{5}{10} = \frac{1}{2}$$

Eine feinere Aufteilung in 100 Sektoren der gleichen Breite liefert: jeder Sektor hat Wahrscheinlichkeit $\frac{1}{100}$, *aber*

$$P(X \in [0, \pi]) = \frac{50}{100} = \frac{1}{2}$$

ist konstant.
Im Grenzprozess $n \to \infty$ *erhält man: jeder Sektor hat Wahrscheinlichkeit* 0, *aber*

$$\lim_{n\to\infty} P(X \in [0, \pi]) = \lim_{n\to\infty} \frac{\frac{1}{2}n}{n} = \frac{1}{2}$$

Definition 8.1.1
Eine Zufallsvariable X heißt **stetig**, *wenn es eine Funktion $f(x) \geq 0$ gibt, so dass sich die* ***Verteilungsfunktion*** *$F(x)$ von X wie folgt darstellen lässt:*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du.$$

Die Funktion $f(x)$ heißt **Wahrscheinlichkeitsdichte** *(kurz* **Dichte** *oder* **Dichtefunktion***) von X. Der Träger $\mathcal{T}$ von X ist die Menge aller Elemente $x \in \mathbb{R}$ für die $f(x) > 0$ gilt.*
Beachte den Unterschied zu diskreten Zufallsvariablen! Hier gilt:

$$F(x) = \sum_{i:x_i \leq x} f(x_i)$$

Einige Folgerungen:

1. $P(X = x) = 0 \quad \forall x \in \mathbb{R}$
2. $$\begin{aligned} P(X \in [a,b]) &= P(X \in]a,b]) = P(X \in [a,b[) = P(X \in]a,b[) \\ &= \int_a^b f(x)\, dx \end{aligned}$$
3. $\int_{-\infty}^{+\infty} f(x)\, dx = 1$ "Normierungseigenschaft"

Eigenschaften der Verteilungsfunktionen $F(x)$ von stetigen Zufallsvariablen:

1. $\lim_{x \to -\infty} F(x) = 0$
2. $\lim_{x \to \infty} F(x) = 1$
3. An allen Stetigkeitsstellen von $f(x)$ gilt: $F'(x) = f(x)$
4. $P(a \leq X \leq b) = F(b) - F(a)$
5. $P(X \geq a) = 1 - F(a)$
   etc.

Definition 8.1.2
Als ***Normalisierungskonstante*** *c bezeichnet man multiplikative Terme in der Dichtefunktion $f(x)$, die nicht vom Argument x abhängen (aber im Allgemeinen von den Parametern), der übrige Teil heißt* ***Kern****:*

$$f(x) = c \cdot \underbrace{g(x)}_{Kern}$$

Man schreibt oft $f(x) \propto g(x)$.

Allgemeine Definition von stetigen Zufallsvariablen:
Frage: Für welche Mengen B ist die Aussage

$$P(X \in B) = \int_B f(x)dx$$

überhaupt sinnvoll? Sei $\mathcal{F}$ die Mengenfamilie aller offenen Intervalle in $\mathbb{R}$. Dann gibt es eine sogenannte ***σ*-Algebra** (eine spezielle Mengenfamilie) $\sigma(\mathcal{F})$, die $\mathcal{F}$ enthält.

Für eine σ-Algebra $\sigma(\mathcal{F})$ muss gelten:

1. $\emptyset$ und $\Omega \in \sigma(\mathcal{F})$
2. Für $A, B \in \sigma(\mathcal{F})$ ist auch $B \setminus A \in \sigma(\mathcal{F})$
3. Für $A_1, A_2, \ldots \in \sigma(\mathcal{F})$ ist auch $\bigcup_{n=1}^{\infty} A_n \in \sigma(\mathcal{F})$ und $\bigcap_{n=1}^{\infty} A_n \in \sigma(\mathcal{F})$

Ein Wahrscheinlichkeitsmaß P auf Ω wird nun mittels $\sigma(\mathcal{F})$ definiert: Für alle paarweise disjunkten Mengen $A_1, A_2, \ldots \in \sigma(\mathcal{F})$ soll gelten (vgl. Axiom A3 von Kolmogorow):

$$P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$$

Ferner müssen natürlich auch die Axiome A1 und A2 erfüllt sein:

$$P(\emptyset) = 0$$
$$P(\Omega) = 1$$

Stetige Zufallsvariablen sind also Abbildungen von Ω nach $\mathbb{R}$.

8.2 Wichtige stetige Verteilungen

Im Folgenden werden wir nun wichtige stetige Verteilungen kennenlernen. Stetige Verteilungen hängen wie diskrete Verteilungen von einem oder mehreren **Parametern** ab.
Zur Charakterisierung werden wir meist die **Dichtefunktion** und den **Träger** angeben.

Die einfachste stetige Verteilung ist die **stetige Gleichverteilung**:
Eine Zufallsvariable X heißt **stetig gleichverteilt** auf dem Intervall $[a, b]$ $(a, b \in \mathbb{R})$, kurz $X \sim \mathcal{U}(a, b)$, falls ihre Dichtefunktion die Form

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

hat. Der Träger von X ist also $\mathcal{T} = [a, b]$.
Die Verteilungsfunktion $F(x)$ von X ergibt sich zu

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b \end{cases}$$

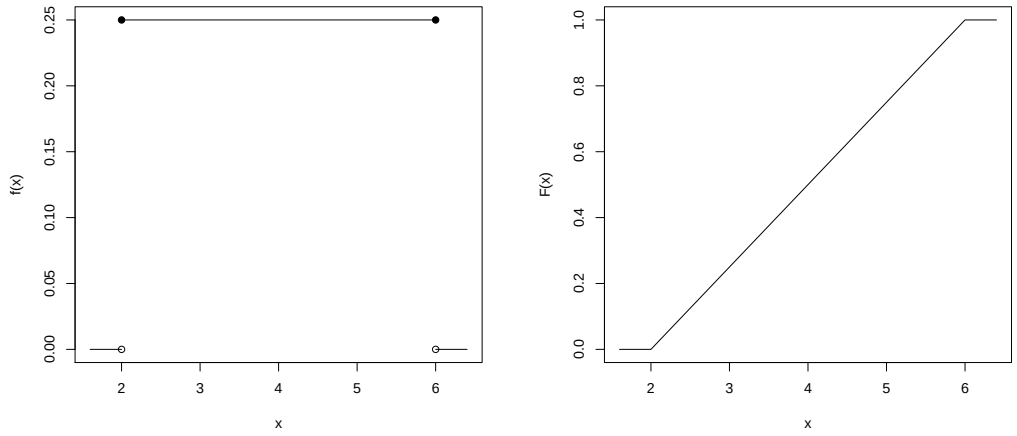

Abbildung 8.1: Dichtefunktion (links) und Verteilungsfunktion (rechts) der stetigen Gleichverteilung für $a = 2$ und $b = 6$

Funktionen in R:

- `dunif(x, min = a, max = b)` liefert Dichtefunktion
- `punif()` liefert Verteilungsfunktion
- `qunif()` liefert Quantile
- `runif()` liefert Zufallszahlen aus der Gleichverteilung

Die Exponentialverteilung
Eine stetige Zufallsvariable X mit positivem Träger $\mathbb{R}_+$, heißt **exponentialverteilt** mit Parameter $\lambda \in \mathbb{R}_+$ (kurz $X \sim \mathcal{E}(\lambda)$), wenn sie die Dichte

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

besitzt. Die Verteilungsfunktion ergibt sich zu

$$F(x) = \begin{cases} 1 - \exp(-\lambda x) & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases}$$

Funktionen in R:

- `dexp(x, rate = λ)`liefert Dichtefunktion
- `pexp()` liefert Verteilungsfunktion
- `qexp()` liefert Quantile
- `rexp()` liefert Zufallszahlen aus der Exponentialverteilung

Es bleibt zu zeigen, dass $\int_0^\infty f(x)\,dx = 1$ gilt:

$$\begin{aligned} \int_0^\infty f(x)\,dx &= \lambda \int_0^\infty \exp(-\lambda x)\,dx \\ &= \lambda \cdot \left[-\frac{1}{\lambda} \exp(-\lambda x) \right]_0^\infty \\ &= \lambda \cdot \left[-0 + \frac{1}{\lambda} \right] \\ &= 1 \end{aligned}$$

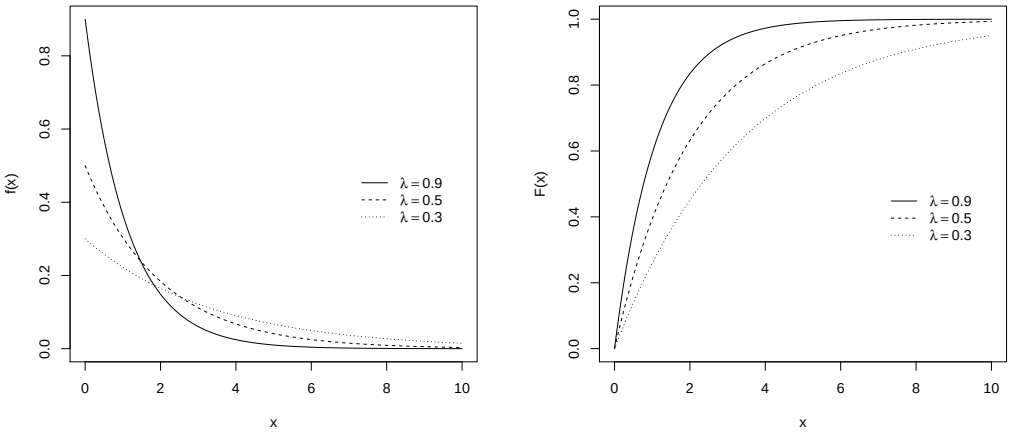

Abbildung 8.2: Dichtefunktion (links) und Verteilungsfunktion (rechts) der Exponentialverteilung für verschiedene Raten.

Beispiel 8.2.1 (Kern der Exponentialverteilung)
Der ***Kern*** *der Exponentialverteilung ist* $\exp(-\lambda x)$*, da dieser Teil der Dichtefunktion* $f(x)$ *von* x *abhängt.*
Die ***Normalisierungskonstante*** *ist* λ*.*

Die Exponentialverteilung steht in engem Zusammenhang zur Poissonverteilung. Die Anzahl der Ereignisse in einem Intervall ist genau dann $\mathcal{P}(\lambda)$-verteilt, wenn die Zeitdauern zwischen aufeinander folgenden Ereignissen unabhängig und exponential verteilt mit Parameter λ sind.

Beispiel 8.2.2
Ebenso wie die geometrische Verteilung besitzt die Exponentialverteilung die Eigenschaft der ***Gedächtnislosigkeit****, d.h.* $P(X > s + x|X > s) = P(X >$

x), *wie man leicht sieht:*

$$\begin{aligned} P(X > s+x|X > s) &= \frac{P(X > s+x, X > s)}{P(X > s)} \\ &\overset{x\geq 0}{=} \frac{P(X > s+x)}{P(X > s)} \\ &= \frac{1 - P(X \leq s+x)}{1 - P(X \leq s)} \\ &= \frac{\exp(-\lambda(s+x))}{\exp(-\lambda s)} \\ &= \exp(-\lambda x) \\ &= P(X > x) \end{aligned}$$

Die Gammaverteilung
Die Gammaverteilung ist eine Verallgemeinerung der Exponentialverteilung. Wie diese hat sie einen positiven Träger $\mathcal{T} = \mathbb{R}_+$, aber einen Parameter mehr: Eine stetige Zufallsvariable X heißt **gammaverteilt** mit Parametern $\alpha \in \mathbb{R}_+$ und $\beta \in \mathbb{R}_+$ (kurz $X \sim \mathcal{G}(\alpha, \beta)$), falls sie die Dichte

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

besitzt.
Hier bezeichnet $\Gamma(\alpha)$ die **Gammafunktion**

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x)\, dx$$

Die Gammafunktion kann als Verallgemeinerung der Fakultät betrachtet werden, da gilt:

$$\begin{aligned} \Gamma(x+1) &= x! \quad \text{für } x \in \mathbb{N}_0 \\ \Gamma(x+1) &= x\Gamma(x) \quad \text{für } x \in \mathbb{R}_+ \end{aligned}$$

Eigenschaften der Gammaverteilung:

- für $\alpha = 1$ entspricht die Gammaverteilung einer Exponentialverteilung mit Parameter $\lambda = \beta$

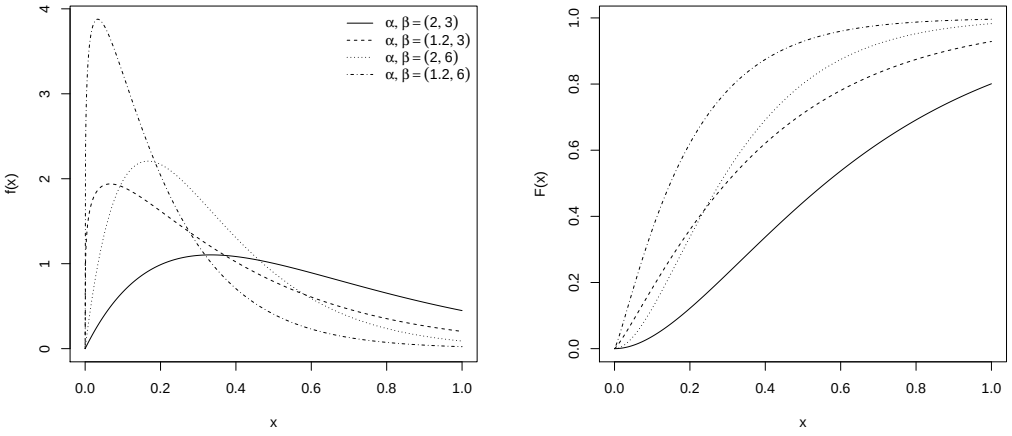

Abbildung 8.3: Dichtefunktion (links) und Verteilungsfunktion (rechts) der Gammaverteilung mit verschiedenen Werten für α und β

- für $\alpha = \frac{d}{2}$ mit $d \in \mathbb{N}$ und $\beta = \frac{1}{2}$ entspricht die Gammaverteilung der sogenannten **Chi-Quadrat(χ^2) -Verteilung** mit d Freiheitsgraden (kurz: $X \sim \mathcal{G}(\frac{d}{2}, \frac{1}{2}) \Rightarrow X \sim \chi^2(d)$)

Funktionen in R:

- `dgamma(x, shape =` α`, rate =` β`)` liefert Dichtefunktion
- `pgamma()` liefert Verteilungsfunktion
- `qgamma()` liefert Quantile
- `rgamma()` liefert Zufallszahlen aus der Gammaverteilung

und

- `dchisq(x, df =` d`, rate =` β`)` liefert Dichtefunktion
- `dchisq()` liefert Verteilungsfunktion
- `dchisq()` liefert Quantile
- `rchisq()` liefert Zufallszahlen aus der χ^2_d-Verteilung

Man kann mit Hilfe der Substitutionsregel

$$\int \tilde{f}(g(x)) \cdot g'(x)\, dx = \int \tilde{f}(z)\, dz$$

zeigen, dass $\int f(x)\,dx = 1$ ist:

$$\begin{aligned}\int f(x)\,dx &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)\,dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}\exp(-\beta x)\,dx\end{aligned}$$

Als Substitution verwendet man $g(x) = \beta \cdot x$. Dann erhält man

$$\begin{aligned}\int f(x)\,dx &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty g(x)^{\alpha-1}\cdot\frac{1}{\beta^{\alpha-1}}\cdot\exp(-g(x))\,dx \\ &= \frac{\beta}{\Gamma(\alpha)}\cdot\frac{1}{\beta}\int_0^\infty \overbrace{g(x)^{\alpha-1}\exp(-g(x))}^{\tilde{f}(g(x))}\cdot\overbrace{\beta}^{g'(x)}\,dx \\ &= \frac{1}{\Gamma(\alpha)}\underbrace{\int_0^\infty \tilde{f}(z)dz}_{=\Gamma(\alpha)} \\ &= 1\end{aligned}$$

Die Normalverteilung
Eine Zufallsvariable X mit Träger $\mathcal{T} = \mathbb{R}$ und Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}_+$ heißt **normalverteilt** (kurz $X \sim \mathcal{N}(\mu, \sigma^2)$), falls sie die Dichtefunktion

$$f(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \quad \text{für } x \in \mathbb{R}$$

hat. Diese wird auch *"Gaußsche Glockenkurve"* genannt. Für $\mu = 0$ und $\sigma^2 = 1$ nennt man die Verteilung **Standardnormalverteilung**.
Beachte:

$$F(x) = \int_{-\infty}^x f(u)\,du$$

ist nicht analytisch zugänglich (d.h. man findet keine Stammfunktion und braucht numerische Integration).

Weshalb gilt für die Dichtefunktion der Normalverteilung $\int_{-\infty}^\infty f(x)\,dx = 1$?
Aus der Analysis ist bekannt, dass für $a > 0$ gilt:

$$\int_{-\infty}^\infty \exp(-a^2x^2)\,dx = \frac{\sqrt{\pi}}{a} \tag{8.1}$$

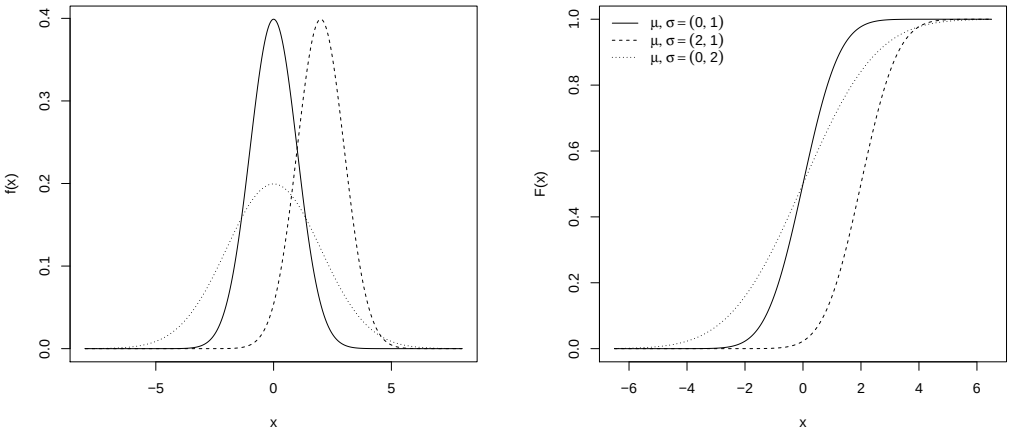

Abbildung 8.4: Dichtefunktion (links) und Verteilungsfunktion (rechts) der Normalverteilung mit verschiedenen Werten für μ und σ

Außerdem stimmen die folgenden beiden Integrale $\forall \mu \in \mathbb{R}$ überein

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx = \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \tag{8.2}$$

da die beiden Integralsfunktionen bis auf eine Verschiebung entlang der x-Achse identisch sind. Daher erhält man:

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\,dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}\cdot\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx \\
&\overset{(8.2)}{=} \frac{1}{\sqrt{2\pi}\cdot\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\
&\overset{(8.1)}{=} \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot \sqrt{\pi} \cdot \sqrt{2\sigma^2} \\
&= 1
\end{aligned}$$

Funktionen in R:

- `dnorm(x, mean =` μ`, sd =` σ`)` liefert Dichtefunktion
- `pnorm()` liefert Verteilungsfunktion
- `qnorm()` liefert Quantile

- `rnorm()` liefert Zufallszahlen aus der Normalverteilung

Die Betaverteilung
Eine Zufallsvariable X mit Träger $\mathcal{T} = (0, 1)$ und Parametern $\alpha \in \mathbb{R}_+$ und $\beta \in \mathbb{R}_+$ heißt **betaverteilt** (kurz $X \sim \mathcal{B}e(\alpha, \beta)$), falls sie die Dichtefunktion

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{für } 0 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

besitzt, wobei die **Betafunktion** $B(\alpha, \beta)$ gerade so definiert ist, dass $\int_0^1 f(x)\,dx = 1$ gilt:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

An dieser Formel erkennt man auch den Zusammenhang zwischen der Beta- und der Gammafunktion.

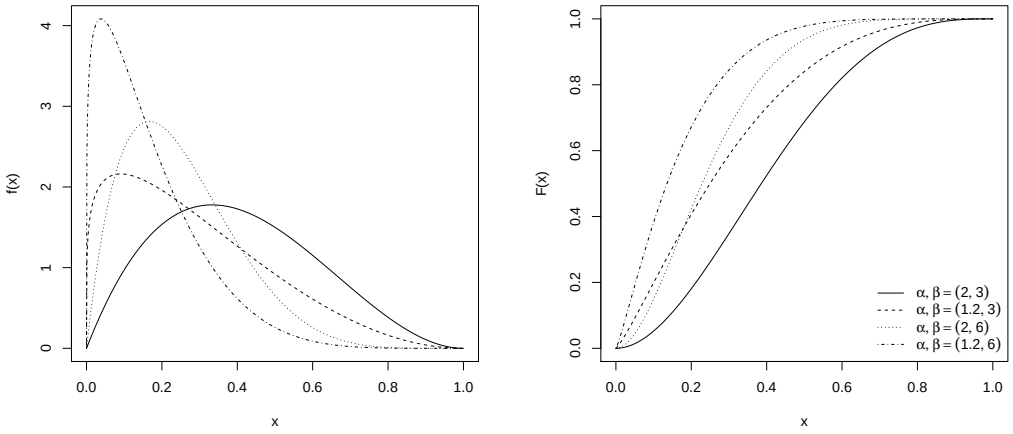

Abbildung 8.5: Dichtefunktion (links) und Verteilungsfunktion (rechts) der Betaverteilung mit verschiedenen Werten für α und β

Beachte: Für $\alpha = \beta = 1$ entspricht die Gammaverteilung der Gleichverteilung auf dem Intervall $[0, 1]$.

Funktionen in R:

- `dbeta(x, shape1 =` α`, shape2 =` β`)` liefert Dichtefunktion

- `pbeta()` liefert Verteilungsfunktion
- `qbeta()` liefert Quantile
- `rbeta()` liefert Zufallszahlen aus der Betaverteilung

8.3 Lageparameter von stetigen Zufallsvariablen

Lageparameter von stetigen Zufallsvariablen sind (ebenso, wie bei diskreten Zufallsvariablen) die folgenden:

- **Erwartungswert**: existiert meistens, ist dann auch eindeutig
- **Median** (0.5-Quantil): existiert immer, ist immer eindeutig, solange der Träger von X ein Intervall ist
- **Modus** (Maximum der Dichtefunktion): existiert nicht immer, ist auch nicht immer eindeutig

Die Definitionen dieser Parameter lauten aber anders.

Definition 8.3.1
Den **Erwartungswert** *einer stetigen Zufallsvariable X ist definiert als*

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

unter der Voraussetzung, dass die Funktion $xf(x)$ **absolut integrierbar** *ist, d.h es muss gelten:*

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

Andernfalls sagt man, der Erwartungswert von X existiert nicht bzw. ist unendlich.

Zur Erinnerung ist hier noch einmal die Definition des Erwartungswertes für stetige Zufallsvariablen aufgeführt:

$$\mathrm{E}(X) = \sum_{x \in \mathcal{T}} x \underbrace{P(X = x)}_{f(x)}$$

Der Erwartungswert für stetige Zufallsvariablen hat sehr ähnliche Eigenschaften wie im diskreten Fall (die Existenz aller auftretenden Erwartungswerte sei im folgenden vorausgesetzt):

1. $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$
   für eine beliebige Funktion $g : \mathbb{R} \to \mathbb{R}$

2. "Linearität des Erwartungswertes":

$$\mathrm{E}(a \cdot X + b) = a\mathrm{E}(X) + b$$

3. "Additivität":

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

4. "Symmetrie":
   Ist $f(x)$ symmetrisch um einen Punkt c, d.h.
   $f(c - x) = f(c + x) \quad \forall x \in \mathbb{R}$, dann ist $\mathrm{E}(X) = c$.

Beispiel 8.3.1 (Erwartungswert der stetigen Gleichverteilung)
Die Dichtefunktion ist

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \textit{sonst} \end{cases}$$

Daher lautet der Erwartungswert

$$E(X) \;=\; \int_a^b x \tfrac{1}{b-a}\, dx \;=\; \tfrac{1}{b-a}\left[\tfrac{x^2}{2}\right]_a^b \;=\; \tfrac{1}{b-a} \cdot \tfrac{1}{2}(b^2 - a^2) \;=\; \tfrac{a+b}{2}$$

Dies ist einfacher über die Symmetrieregel für den Erwartungswert zu zeigen, denn die Dichtefunktion $f(x)$ ist symmetrisch um den Punkt $c = \frac{a+b}{2}$.

Beispiel 8.3.2 (Erwartungswert der Normalverteilung)
Der Erwartungswert der Normalverteilung ist $E(X) = \mu$, da die Dichtefunktion

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right) \quad \textit{für } x \in \mathbb{R}$$

symmetrisch um den Punkt $c = \mu$ ist.

Beispiel 8.3.3 (Erwartungswert der Betaverteilung)

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \textit{für } 0 < x < 1 \\ 0 & \textit{sonst} \end{cases}$$

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f(x)\,dx \\
&= \int_0^1 x \cdot \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \frac{1}{B(\alpha,\beta)} \cdot B(\alpha+1,\beta) \qquad \underbrace{\int_0^1 \frac{1}{B(\alpha+1,\beta)} x^{\alpha}(1-x)^{\beta-1}\,dx}_{=1,\ \textit{Int. über Dichtefkt. von } \mathcal{B}e(\alpha+1,\beta)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\
&\quad \textit{es gilt } \Gamma(x+1) = x\cdot\Gamma(x) \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

Beispiel 8.3.4 (Erwartungswert der Exponentialverteilung)

$$
f(x) = \begin{cases} \lambda \exp(-\lambda x) & \textit{für} \quad x \geq 0 \\ 0 & \textit{sonst} \end{cases}
$$

Mit Hilfe von partieller Integration

$$
\int u(x)v'(x)\,dx = u(x)v(x) - \int u'(x)v(x)\,dx
$$

gilt für den Erwartungswert

$$
\begin{aligned}
E(X) &= \int_0^{\infty} \underbrace{x\lambda}_{u(x)} \underbrace{\exp(-\lambda x)}_{v'(x)}\,dx \\
&= \left[x\lambda(-1)\frac{1}{\lambda}\exp(-\lambda x)\right]_0^{\infty} - \int_0^{\infty} \lambda(-1)\frac{1}{\lambda}\exp(-\lambda x)\,dx \\
&= 0 + \int_0^{\infty} \exp(-\lambda x)\,dx \\
&= \frac{1}{\lambda}
\end{aligned}
$$

Satz 8.3.1
Es gilt für stetige Zufallsvariablen mit positivem Träger $\mathbb{R}_+$:

$$
E(X) = \int_0^{\infty} \underbrace{[1-F(x)]}_{P(X>x)}\,dx
$$

vgl. dazu Satz (5.1.1) für diskrete Zufallsvariablen mit Träger $\mathbb{N}$*:*

$$E(X) = \sum_{k=1}^{\infty} P(X \geq k) = \sum_{k=0}^{\infty} P(X > k)$$

Diese Formel liefert eine einfachere Variante, den Erwartungswert der Exponentialverteilung zu berechnen:

$$\mathrm{E}(X) = \int_0^\infty 1 - [1 - \exp(-\lambda x)]\, dx = \int_0^\infty \exp(-\lambda x)\, dx = -\frac{1}{\lambda} \exp(-\lambda x) \Big|_0^\infty = \frac{1}{\lambda}$$

Bemerkung: Für beliebige Zufallsvariablen X muss zwar immer $\int f(x)\, dx = 1$ gelten, es kann aber durchaus der Fall $\mathrm{E}(X) = \infty$ eintreten, da

$$\mathrm{E}(|X|) = \int |x| f(x)\, dx = \infty$$

Dies sieht man an folgendem Beispiel:

Beispiel 8.3.5 (Erwartungswert der Cauchy-Verteilung)
Die ***Cauchy-Verteilung*** *mit der Dichtefunktion*

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2} \quad \text{für } x \in \mathbb{R}$$

hat **keinen** *(endlichen) Erwartungswert.*
Für die Cauchy-Verteilung gilt, dass $f(x)$ *symmetrisch um den Punkt* 0 *ist, und somit würde man denken, dass* $E(X) = 0$ *ist, was aber nicht der Fall ist. Betrachte dazu zunächst*

$$\begin{aligned} E(|X|) &= 2 \int_0^\infty x f(x)\, dx \\ &= \frac{2}{\pi} \lim_{c \to \infty} \int_0^c \frac{x}{1 + x^2}\, dx \\ &= \frac{2}{\pi} \lim_{c \to \infty} \left[\frac{1}{2} \log(1 + x^2) \right]_0^c \\ &= \frac{1}{\pi} \lim_{c \to \infty} \log(1 + c^2) \\ &= \infty. \end{aligned}$$

Der Erwartungswert der Cauchy-Verteilung existiert somit nicht.

Definition 8.3.2 (Quantile von stetigen Zufallsvariablen)
Wir nehmen an, dass der Träger der stetigen Zufallsvariable X ein Intervall ist und somit die Umkehrfunktion $F^{-1}(p)$ der Verteilungsfunktion $F(x)$ von X eindeutig definiert ist.
*Das p-**Quantil** der Verteilung von X ist definiert als der Wert x_p für den $F(x) = p$ gilt. Somit gilt $x_p = F^{-1}(p)$. Speziell erhält man für $p = 0.5$ den* **Median** x_{Med}.

Ist $f(x)$ symmetrisch um einen Punkt c, so ist $x_{Med} = c$. Beispielsweise ist der Median $x_{Med} = \mu$ bei einer normalverteilten Zufallsvariablen $X \sim \mathcal{N}(\mu, \sigma^2)$.

Definition 8.3.3 (Der Modus von stetigen Zufallsvariablen)
Ein **Modus** *einer stetigen Zufallsvariable X ist ein Wert x_{Mod}, für den für alle $x \in \mathbb{R}$ gilt:*

$$f(x_{Mod}) \geq f(x)$$

Der Modus ist nicht notwendigerweise eindeutig, noch muss er existieren.

Beispiel 8.3.6 (Modi von verschiedenen stetigen Verteilungen)

1. *Modus der Betaverteilung:*

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{für } 0 < x < 1 \\ 0 & \text{sonst} \end{cases}$$

Um das Maximum der Dichtefunktion zu erhalten, wird die erste Ableitung gleich Null gesetzt:

$$\begin{aligned} f'(x) &= \frac{1}{B(\alpha,\beta)} \left[(\alpha-1)x^{\alpha-2}(1-x)^{\beta-1} + x^{\alpha-1}(\beta-1)(1-x)^{\beta-2}(-1)\right] \\ &= \frac{1}{B(\alpha,\beta)} x^{\alpha-2}(1-x)^{\beta-2} \underbrace{\left[(\alpha-1)(1-x) - (\beta-1)x\right]}_{\stackrel{!}{=}0} \\ &\stackrel{!}{=} 0 \\ &\Leftrightarrow \alpha - \alpha x - 1 + x - x\beta + x = 0 \end{aligned}$$

$$x_{Mod} = \frac{\alpha-1}{\alpha+\beta-2} \text{ nur für } \alpha > 1 \text{ und } \beta > 1 \text{ eindeutig!!}$$

2. *Der Modus der Normalverteilung ist μ.*

3. *Der Modus der Gammaverteilung:*
Für $\alpha > 1$ ist der Modus eindeutig gleich $x_{Mod} = (\alpha-1)/\beta$. Für $\alpha < 1$ existieren keine Modi.

Definition 8.3.4
Die **Varianz einer stetigen Zufallsvariablen** *definiert man analog zum diskreten Fall:*

$$VarX = E[X - E(X)]^2 = E[X - \mu]^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$$

mit $\mu = E(X)$. *Die* **Standardabweichung** $\sigma = \sqrt{Var(X)}$ *ist wie im diskreten Fall definiert.*
Beachte: *Auch die Varianz kann nicht existieren, d.h. unendlich sein. Existiert der Erwartungswert nicht, so existiert auch die Varianz nicht.*

Für die Varianz für stetige Zufallsvariablen gelten nun im wesentlichen dieselben Eigenschaften wie im diskreten Fall.

- Verschiebungssatz:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

- Lineare Transformationen: Für $Y = a \cdot X + b$ gilt:

$$\mathrm{Var}(Y) = a^2 \cdot \mathrm{Var}(X)$$

- Sind X und Y unabhängig, so gilt:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Beispiel 8.3.7 (Varianz der stetigen Gleichverteilung)
Wir wissen:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & sonst \end{cases}$$

und

$$E(X) = \frac{a+b}{2}$$

Zunächst folgt für $E(X^2)$*:*

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx \;=\; \int_a^b x^2 \frac{1}{b-a}\, dx \;=\; \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b \;=\; \frac{1}{3} \cdot \frac{b^3 - a^3}{b-a}$$

Mit dem Verschiebungssatz ergibt sich:

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \;=\; \frac{1}{3} \cdot \frac{b^3 - a^3}{b-a} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{3} \cdot (b^2 + ab + a^2) - \frac{1}{4}(b^2 + 2ab + a^2) \\ &= \frac{1}{12}(b^2 - 2ab + a^2) = \frac{(b-a)^2}{12} \end{aligned}$$

Die Varianz wächst also quadratisch mit der Länge des Intervalls, die Standardabweichung somit linear *mit der Länge des Intervalls.*

Im Folgenden nun zusammenfassend die Erwartungswerte und Varianzen der gängigsten stetigen Verteilungen:

Name	Symbol	$\mathrm{E}(X)$	$\mathrm{Var}(X)$
Gleichverteilung	$X \sim \mathcal{U}(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentialverteilung	$X \sim \mathcal{E}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gammaverteilung	$X \sim \mathcal{G}(\alpha,\beta)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Normalverteilung	$X \sim \mathcal{N}(\mu,\sigma^2)$	μ	σ^2
Betaverteilung	$X \sim \mathcal{B}e(\alpha,\beta)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\cdot\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

8.4 Das Gesetz der großen Zahlen

Das **Gesetz der großen Zahlen** ist eine Aussage über das *arithmetische Mittel* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ für $n \to \infty$, wobei X_i, $i = 1, \ldots, n$ unabhängig und identisch verteilte Zufallsvariablen aus einer Verteilung mit Erwartungswert μ und Varianz σ^2 seien.
Klarerweise gilt:

$$E(\bar{X}_n) = \mu$$

und

$$Var(\bar{X}_n) = \frac{1}{n}\,\sigma^2$$

da

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\cdot\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu$$

$$Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{1}{n}\,\sigma^2.$$

Daher folgt sofort, dass für das arithmetische Mittel und seine Varianz im Grenzfall ($n \to \infty$) Folgendes gilt:

$$\bar{X}_n \to \mu \quad \text{und} \quad Var(\bar{X}_n) \to 0$$

In Abbildung 8.6 sieht man anschaulich, dass das arithmetische Mittel von 10000 standardnormalverteilten Zufallsvariablen gegen den Erwartungswert 0 konvergiert.

Dagegen konvergiert das arithmetische Mittel von 10000 Cauchyverteilten Zufallsvariablen nicht (siehe Abb. 8.7), da der Erwartungswert der Cauchy-Verteilung nicht existiert.

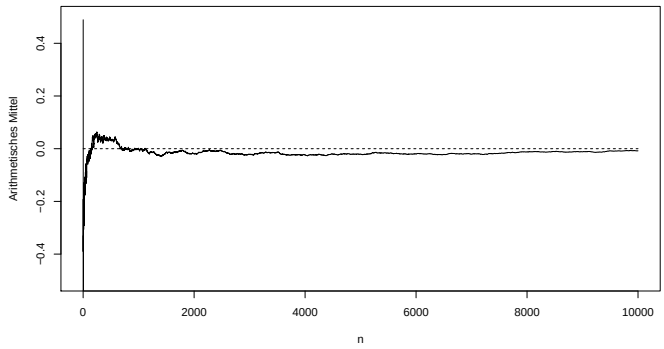

Abbildung 8.6: Arithmetisches Mittel für 10000 standardnormalverteilte Zufallsvariable

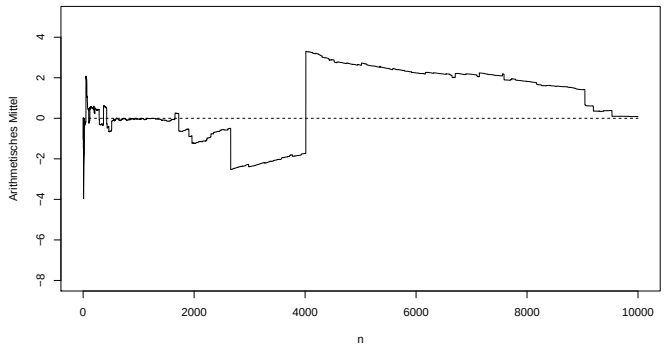

Abbildung 8.7: Arithmetisches Mittel für 10000 Cauchyverteilte Zufallsvariablen

8.5 Der Transformationssatz für Dichten

Sei X eine stetige Zufallsvariable mit Dichte $f_X(x)$. Betrachte nun $Y = g(X)$, wobei z.B. $Y = \exp(X), Y = X^2, \ldots$
Frage: Wie lautet die Dichte $f_Y(y)$ der Zufallsvariable Y?

In dem folgenden Satz wird beschrieben, wie man auf einfache Weise die Dichtefunktion von $Y = g(X)$ berechnen kann:

Satz 8.5.1 (Transformationssatz für Dichten)
Sei g streng monoton und differenzierbar. Dann kann man die Dichte $f_Y(y)$ mit Hilfe des Transformationssatzes berechnen:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \underbrace{\left|\frac{dg^{-1}(y)}{dy}\right|}_{g^{-1\prime}(y)}$$

Beweis (über die Verteilungsfunktion $F_Y(y)$ von Y):
Sei g zunächst streng monoton *wachsend* und differenzierbar:

$$F_Y(y) \;=\; P(g(X) \le y) \;=\; P(X \le g^{-1}(y)) \;=\; F_X(g^{-1}(y))$$

Differenzieren ergibt:

$$\begin{aligned} f_Y(y) = F_Y'(y) &= F_X'(g^{-1}(y)) \cdot \frac{dg^{-1}(y)}{dy} \\ &= f_X(g^{-1}(y)) \cdot \underbrace{\frac{dg^{-1}(y)}{dy}}_{\substack{\text{positiv, da } g^{-1} \text{ streng} \\ \text{monoton wachsend}}} \end{aligned}$$

Sei nun g streng monoton *fallend* und differenzierbar:

$$\begin{aligned} F_Y(y) &= P(g(X) \le y) \;=\; P(X \ge g^{-1}(y)) \;=\; 1 - P(X < g^{-1}(y)) \\ &= 1 - P(X \le g^{-1}(y)) \;=\; 1 - F_X(g^{-1}(y)) \\ \Rightarrow \qquad f_Y(y) &= -f_X(g^{-1}(y)) \cdot \underbrace{\frac{dg^{-1}(y)}{dy}}_{\substack{\text{negativ, da g streng} \\ \text{monoton fallend}}} \end{aligned}$$

Insgesamt ergibt sich also:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{dg^{-1}(y)}{dy}\right|$$

Beispiel 8.5.1 (Erzeugung exponentialverteilter Zufallsvariablen)
Betrachte $X \sim U[0,1]$ und $Y = g(X)$, mit $g(x) = -\log(X)$. Die Umkehrfunktion von $g(x)$ ist damit $g^{-1}(y) = \exp(-y)$. Die Ableitung der Umkehrfunktion lautet dann

$$\frac{dg^{-1}(y)}{dy} = -\exp(-y)$$

Damit ergibt sich für die Dichtefunktion von Y:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |-\exp(-y)| = \exp(-y)$$

Daher folgt, dass Y exponentialverteilt ist mit Parameter $\lambda = 1$, also $Y \sim \mathcal{E}(\lambda = 1)$. Allgemeiner liefert $Y = -\frac{1}{\lambda}\log(x)$ Zufallszahlen aus einer Exponentialverteilung mit Parameter λ : $\quad Y \sim \mathcal{E}(\lambda)$

Beispiel 8.5.2 (Quadrat einer Standardnormalverteilung)
Wie lautet die Dichte von $Y = X^2$, falls $X \sim N(0,1)$, also standardnormalverteilt ist? Die Dichte von X ist

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \quad \text{für } x \in \mathbb{R}$$

Ein Problem ist, dass man für die Verwendung des "Transformationssatzes für Dichten" eine streng monotone Funktion g benötigt, $g(x) = x^2$ aber nicht monoton ist. Daher betrachtet man zunächst $Z = |X|$. Z hat offensichtlich das Doppelte der Dichte der Standardnormalverteilung auf $\mathbb{R}_+$:

$$f(z) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \quad \text{für } z > 0 \text{ und } 0 \text{ sonst}$$

Nun ist $X^2 = Y = Z^2 = g(Z)$ und g monoton wachsend auf dem Wertebereich $\mathbb{R}_+$. Damit ergibt sich $y = z^2 \Leftrightarrow z = \sqrt{y}$ und die Ableitung der Umkehrfunktion von g lautet

$$\frac{dg^{-1}(y)}{dy} = \frac{1}{2}y^{-\frac{1}{2}}$$

Mit dem "Transformationssatz für Dichten" erhält man die Dichte von Y:

$$f(y) \;=\; \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{y})^2\right) \cdot \frac{1}{2}\, y^{-\frac{1}{2}} \;=\; \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y\right) \cdot y^{-\frac{1}{2}}$$

Y ist also gammaverteilt mit den Parametern $\alpha = \beta = \frac{1}{2}$, $Y \sim \mathcal{G}(.5, .5)$. Vergleiche hierzu die Dichte der Gammaverteilung:

$$f(y) = \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)}}_{\frac{\sqrt{\frac{1}{2}}}{\sqrt{\pi}} = \frac{1}{\sqrt{2\pi}}} y^{\alpha-1} \exp(-\beta y)$$

Diese Dichte entspricht auch der Dichte einer χ^2-Verteilung mit 1 Freiheitsgrad: $Y = X^2 \sim \chi^2_1(1)$. Allgemeiner gilt: Für $X_i \sim N(0,1) \quad i = 1, \ldots, d$ und unabhängig ist $Y = X_1^2 + X_2^2 + \ldots + X_d^2$ χ^2-verteilt mit d Freiheitsgraden.

Allgemeiner kann man auch die **Inversions-Methode** zur Erzeugung von n Zufallszahlen aus einer beliebigen stetigen Verteilung mit Dichte $f(x)$ und Verteilungsfunktionen $F(x)$ verwenden. Erzeuge dazu n gleichverteilte Zufallsvariablen $U_1, \ldots, U_n$ auf dem Intervall $[0,1]$. Dann sind

$$X_i = F^{-1}(U_i), \quad i = 1, \ldots, n$$

die gesuchten Zufallszahlen aus der gewünschten Verteilung mit Verteilungsfunktionen $F(x)$.

Beweis:
Die Dichte von X_i ergibt sich mit der Anwendung des Transformationssatzes für Dichten:

$$f_X(x) = \underbrace{f_U(F(x))}_{=1} \cdot \underbrace{F'(x)}_{f(x)} = f(x)$$

Beispiel 8.5.3 (Erzeugung von Cauchyverteilter Zufallsvariablen)
Die Dichtefunktion $f(x)$ von Cauchyverteilten Zufallsvariablen ist

$$f(x) \;=\; \frac{1}{\pi} \cdot \frac{1}{1+x^2}$$

und die Verteilungsfunktion $F(x)$ lautet

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \frac{1}{\pi} \cdot \frac{1}{1+u^2}\, du \;=\; \frac{1}{\pi} \left[\arctan(u)\right]_{\infty}^{x} \;=\; \frac{1}{\pi} \left[\arctan(x) + \frac{\pi}{2}\right] \\ &= \frac{\arctan(x)}{\pi} + \frac{1}{2} \end{aligned}$$

Die inverse Verteilungsfunktion ist somit:

$$F^{-1}(y) = \tan\left[\pi\left(y - \frac{1}{2}\right)\right]$$

Zufallszahlen aus der Cauchy-Verteilung lassen sich also leicht erzeugen, indem man $U_1, \ldots, U_N$ aus $\sim U[0,1]$ erzeugt und $X_i = \tan(\pi(U_i - \frac{1}{2}))$ berechnet.

Beispiel 8.5.4 (log-Normalverteilung)
Anwendung des Transformationssatzes für Dichten:
Betrachte $X \sim N(\mu, \sigma^2)$. Dann heißt $Y = \exp(X)$ ***log-normalverteilt*** *mit Parameter μ und σ^2. Y hat Dichte*

$$f_Y(y) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{(\log(y)-\mu)^2}{\sigma^2})}_{f_X(g^{-1}(y))} \cdot \underbrace{\frac{1}{y}}_{\frac{dg^{-1}(y)}{dy}}$$

für $y > 0$ und 0 sonst.
Es gilt:

$$\begin{aligned} E(Y) &= \exp(\mu + \frac{1}{2}\sigma^2) \\ Var(Y) &= \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1] \end{aligned}$$

8.6 Der zentrale Grenzwertsatz

Der **zentrale Grenzwertsatz(ZGWS)** beinhaltet die Aussage, dass das arithmetische Mittel, geeignet standardisiert, von *beliebigen* unabhängig und identisch verteilten (engl.: *iid*: "independent, identically distributed") Zufallsvariablen gegen die Standardnormalverteilung konvergiert. Diese Tatsache begründet die zentrale Rolle der Normalverteilung in der Stochastik. Doch zunächst müssen wir dazu standardisierte Zufallsvariablen definieren.

Definition 8.6.1
*Eine Zufallsvariable X heißt **standardisiert**, falls sie*

- *Erwartungswert $E(X) = \mu = 0$ und*
- *Varianz $Var(X) = \sigma^2 = 1$*

besitzt.

Jede Zufallsvariable X mit endlichem Erwartungswert $\mathrm{E}(X)$ und endlicher Varianz $\mathrm{Var}(X)$ kann man durch lineare Transformation standardisieren.
Definiere dazu die Zufallsvariable $\tilde{X}$ als

$$\tilde{X} = \frac{X - \mu}{\sigma}.$$

Dann gilt offensichtlich:

$$\begin{aligned} \mathrm{E}(\tilde{X}) &= \frac{1}{\sigma}\left(\mathrm{E}(X) - \mu\right) = 0 \\ \mathrm{Var}(\tilde{X}) &= \frac{1}{\sigma^2}\,\mathrm{Var}(X) = 1 \end{aligned}$$

Auch die Summe von unabhängig und identisch verteilte Zufallsvariablen $X_1, X_2, \ldots, X_n$ mit endlichem Erwartungswert $\mu = \mathrm{E}(X_i)$ und endlicher Varianz $\sigma^2 = \mathrm{Var}(X_i)$ kann standardisiert werden.
Zunächst gilt für die Summe $Y_n = X_1 + X_2 + \ldots + X_n$:

$$\begin{aligned} \mathrm{E}(Y_n) &= n \cdot \mu \\ \mathrm{Var}(Y_n) &= n \cdot \sigma^2 \end{aligned}$$

Somit hat

$$Z_n = \frac{Y_n - n\mu}{\sqrt{n} \cdot \sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

Erwartungswert

$$\begin{aligned} \mathrm{E}(Z_n) &= \mathrm{E}\left(\frac{Y_n - n\mu}{\sqrt{n}\cdot\sigma}\right) = \frac{\mathrm{E}(Y_n) - n\mu}{\sqrt{n}\cdot\sigma} = \frac{n\mu - n\mu}{\sqrt{n}\cdot\sigma} \\ &= 0 \end{aligned}$$

und Varianz

$$\begin{aligned} \mathrm{Var}(Z_n) &= \mathrm{Var}\left(\frac{Y_n - n\mu}{\sqrt{n}\cdot\sigma}\right) = \mathrm{Var}\left(\frac{Y_n - n\cdot 0}{\sqrt{n}\cdot\sigma}\right) \\ &= \left(\frac{1}{\sqrt{n}\cdot\sigma}\right)^2 \mathrm{Var}(Y_n) = \frac{n\cdot\sigma^2}{n\cdot\sigma^2} \\ &= 1. \end{aligned}$$

Die Zufallsvariable Z_n ist also standardisiert.
Die exakte Verteilung von Z_n ist zunächst noch unbekannt. Für $n \to \infty$ kann man jedoch den zentralen Grenzwertsatz anwenden.

Satz 8.6.1 (Zentraler Grenzwertsatz)
Die Verteilungsfunktion $F_n(z)$ von Z_n konvergiert für $n \to \infty$ an jeder Stelle $z \in \mathbb{R}$ gegen die Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung.

Man schreibt:
$F_n(z) \to \Phi(z)$ für $n \to \infty$ und alle $z \in \mathbb{R}$ bzw. kurz $Z_n \overset{a}{\sim} \mathcal{N}(0,1)$ ("asymptotisch standardnormalverteilt")

In der Praxis kann man also die Verteilung von Z_n für n groß gut durch eine Standardnormalverteilung approximieren.

Bemerkungen:

- Satz 8.6.1 gilt sowohl für stetige als auch für diskrete Zufallsvariablen X_i, wenn deren Erwartungswert und Varianz existieren (für Standardisierung nötig)

- X_i kann beliebig "schiefe" (nicht symmetrische) Verteilungen haben, z.B.
$$X_i \sim \mathcal{E}(\lambda)$$
Trotzdem konvergiert Z_n gegen die (symmetrische) $\mathcal{N}(0,1)$-Verteilung.

- Die Standardisierung ist nicht notwendig zur Formulierung des ZGWS. Alternativ kann man auch direkt $Y_n = X_1 + \ldots + X_n$ betrachten. Dann gilt

$$Y_n \overset{a}{\sim} \mathcal{N}(n \cdot \mu, n \cdot \sigma^2)$$

denn

$$\begin{aligned} & Z_n && \overset{a}{\sim} && \mathcal{N}(0,1) \\ \Rightarrow & \underbrace{\sqrt{n}\sigma \cdot Z_n}_{Y_n - n\cdot\mu} && \overset{a}{\sim} && \mathcal{N}(0, n \cdot \sigma^2) \\ \Rightarrow & Y_n && \overset{a}{\sim} && \mathcal{N}(n \cdot \mu, n \cdot \sigma^2) \end{aligned}$$

Beispiel 8.6.1 (Summe von iid Bernoulliverteilten Zufallsvariablen) *Seien X_i Bernoulliverteilte, unabhängige Zufallsvariablen:*

$$X_i \sim \mathcal{B}(\pi), i = 1, \ldots, n$$

Dann ist $Y_n = \sum_{i=1}^{n} X_i$ binomialverteilt mit $Y_n \sim \mathcal{B}(n, \pi)$. Asymptotisch gilt:

$$\frac{Y_n - n \cdot \pi}{\sqrt{n \cdot \pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)$$

bzw.

$$Y_n \overset{a}{\sim} \mathcal{N}(n \cdot \pi, n \cdot \pi(1-\pi))$$

8.7 Die gemeinsame Verteilung von zwei stetigen Zufallsvariablen

Definition 8.7.1
*Die **gemeinsame Verteilungsfunktion** zweier stetiger Zufallsvariablen X und Y ist die Funktion*

$$F(x,y) \;=\; P(X \leq x \text{ und } Y \leq y)$$

*Alternativ kann man die gemeinsame Verteilung von X und Y auch über deren **gemeinsame Dichtefunktion** $f(x,y)$ definieren, wobei*

$$F(x,y) = \int_{v=-\infty}^{y} \int_{u=-\infty}^{x} f(u,v)\,du\,dv$$

für alle $x, y \in \mathbb{R}$ gelten muss.

Falls $f(x,y)$ stetig ist, so gilt:

$$\frac{d^2F(x,y)}{dx\,dy} = f(x,y)$$

Außerdem muss die gemeinsame Dichtefunktion auch normiert sein:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)\,dx\,dy = 1$$

Die **Dichten der Randverteilungen** lassen sich durch Integration (im diskreten Fall war es die Summation) erhalten:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f(x,y)\,dy \\ f_Y(y) &= \int_{-\infty}^{+\infty} f(x,y)\,dx \end{aligned}$$

Der **Erwartungswert einer gemeinsamen Verteilung** lässt sich berechnen durch

$$\mathrm{E}(g(X,Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) \cdot f(x,y)\,dx\,dy$$

für $g : \mathbb{R}^2 \to \mathbb{R}$.

X, Y heißen **unabhängig**, genau dann wenn

$$\begin{array}{rrcll} & F_{X,Y}(x,y) & = & F_X(x)\,F_Y(y) & \\ \text{bzw.} & f_{X,Y}(x,y) & = & f_X(x)\,f_Y(y) & \forall x,y \in \mathbb{R} \end{array}$$

Allgemeiner gilt:
$X_1, X_2, \ldots, X_n$ sind unabhängig $\Leftrightarrow f(x_1, x_2, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n)$.

Weiterhin definiert man analog zum diskreten Fall:

$$\begin{array}{lrcl} \text{die } \textbf{Kovarianz} & \mathrm{Cov}(X,Y) & = & \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))] \\ \text{die } \textbf{Korrelation} & \rho(X,Y) & = & \frac{\mathrm{Cov}_{(X,Y)}}{\sqrt{\mathrm{Var}_{(X)}}\sqrt{\mathrm{Var}_{(Y)}}} \end{array}$$

Es gilt wieder:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(X \cdot Y) - \mathrm{E}(X) \cdot \mathrm{E}(Y)$$

Beispiel 8.7.1
Betrachte

$$f(x,y) = \begin{cases} \frac{1}{x} & \textit{für} \quad 0 \le y \le x \le 1 \\ 0 & \textit{sonst} \end{cases}$$

Die Randverteilung von X ergibt sich zu

$$f_X(x) \;=\; \int_0^x \frac{1}{x}\,dy \;=\; \frac{1}{x}\,[x - 0] \;=\; 1 \qquad \textit{für} \quad 0 \le x \le 1,$$

also einfach eine Gleichverteilung auf $[0,1]$.
Die Randverteilung von Y ist

$$f_Y(y) \;=\; \int_y^1 \frac{1}{x}\,dx \;=\; [\log(x)\,]_y^1 \;=\; \log\left(\frac{1}{y}\right) \quad \textit{für} \quad 0 \le y \le 1.$$

Man überprüft leicht, dass

$$\int_0^1 f(x)\,dx = 1$$

und

$$\int_0^1 f(y)\,dy \;=\; \int_0^1 \log\left(\frac{1}{y}\right)\,dy \;=\; \left[\log\left(\frac{1}{y}\right) \cdot y + y\right]_0^1 \;=\; 1$$

gilt.
Folglich gilt also auch:

$$\int\int f(x,y)\,dy\,dx \;=\; \int f(x)\,dx \;=\; 1$$

bzw.

$$\int\int f(x,y)\,dx\,dy \;=\; \int f(y)\,dy \;=\; 1$$

Weiter erhält man (z. B. mit `MAPLE`*), dass:*

$$\left.\begin{array}{rcl} E(Y) & = & \frac{1}{4} \\ E(Y^2) & = & \frac{1}{9} \end{array}\right\} \Leftrightarrow \quad \begin{array}{l} Var(Y) = E(Y^2) - [E(Y)]^2 \\ \qquad\quad = \frac{1}{9} - \frac{1}{16} \\ \qquad\quad = \frac{7}{144} \end{array}$$

Da $X \sim \mathcal{U}(0,1)$*, gilt* $E(X) = \frac{1}{2}$ *und* $Var(X) = \frac{1}{12}$.
Ferner ergibt sich für

$$\begin{aligned} E(X\cdot Y) &= \int_0^1\int_0^x x\cdot y\cdot\frac{1}{x}\,dy\,dx \;=\; \int_0^1\int_0^x y\,dy\,dx \\ &= \int_0^1 \left[\frac{y^2}{2}\right]_0^x dx \;=\; \int_0^1 \frac{x^2}{2}\,dx \\ &= \left[\frac{x^3}{6}\right]_0^1 \;=\; \frac{1}{6} \end{aligned}$$

Damit erhält man folgende Werte für die Kovarianz

$$Cov(X,Y) \;=\; E(X\cdot Y) - E(X)E(Y) \;=\; \frac{1}{6} - \frac{1}{2}\cdot\frac{1}{4} \;=\; \frac{1}{24}$$

und die Korrelation

$$\rho(X,Y) \;=\; \frac{Cov(X,Y)}{\sqrt{Var(X)}\,\sqrt{Var(Y)}} \;=\; \frac{\frac{1}{24}}{\sqrt{\frac{1}{12}}\,\sqrt{\frac{7}{144}}} \;\approx 0.65$$

Definition 8.7.2
Die ***bivariate ("zweidimensionale") Standardnormalverteilung*** *mit Parameter* ρ *mit* $|\rho| < 1$ *hat die Dichtefunktion*

$$f(x,y) \;=\; \frac{1}{2\pi\sqrt{1-\rho^2}}\,\exp\left(-\frac{1}{2\,(1-\rho^2)}\,(x^2 - 2\rho xy + y^2)\right)$$

Es gilt:

- Die Randverteilungen von X und Y sind (unabhängig von ρ) standardnormalverteilt.
- Die Korrelation zwischen X und Y ist gleich ρ (daher hat $|\rho|$ auch einen Wert < 1).
- Aus Unkorreliertheit von X und Y folgt hier auch die Unabhängigkeit von X und Y: Für $\rho = 0$ ist nämlich die gemeinsame Dichtefunktion das Produkt der Dichten der Randverteilungen:

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2+y^2)\right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)}_{\text{Dichte der } \mathcal{N}(0,1)\text{-Vtlg.}} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)}_{\text{Dichte der } N(0,1)\text{-Vtlg.}} \\ &= f_X(x) \cdot f_Y(y) \end{aligned}$$

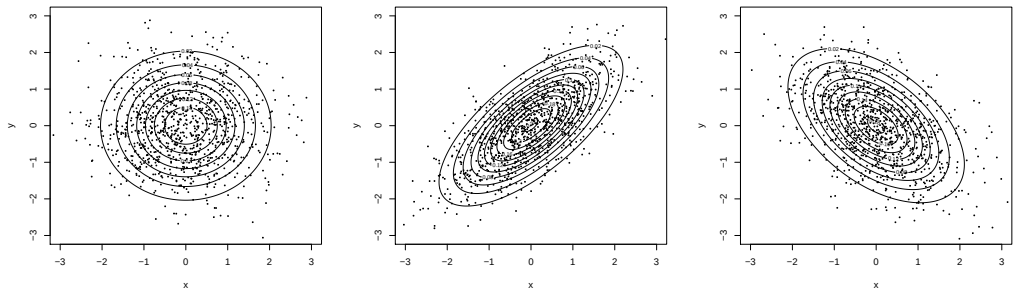

Abbildung 8.8: Die bivariate Standardnormalverteilung für $\rho = 0$ (links), $\rho = 0.7$ (Mitte) und $\rho = -0.5$ (rechts)

Bemerkung:
Die **allgemeine bivariate Normalverteilung** mit insgesamt fünf Parametern $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ erhält man durch folgende lineare Transformationen einer bivariaten Standardnormalverteilung:

$$\begin{aligned} X &\to \mu_X + \sigma_X \cdot X \\ Y &\to \mu_Y + \sigma_Y \cdot Y \end{aligned}$$

8.8 Bedingte Verteilungen von stetigen Zufallsvariablen

Seien X und Y Zufallsvariablen mit gemeinsamer Dichte $f_{X,Y}(x,y)$. Wir interessieren uns für die bedingte Verteilung von X gegeben $Y = y$.
Das Problem bei der Berechnung der Verteilung besteht darin, dass
$P(Y = y) = 0$ und damit

$$P(X \leq x | Y = y) \;=\; \frac{P(X \leq x \quad \text{und} \quad Y = y)}{P(Y = y)}$$

nicht definiert ist. Deshalb geht man nun anders vor und betrachtet

$$\begin{aligned} P(X \leq x | y \leq Y \leq y + dy) &= \frac{P(X \leq x \quad \text{und} \quad y \leq Y \leq y + dy)}{P(y \leq Y \leq y + dy)} \\ &\approx \frac{\int_{-\infty}^{x} f_{X,Y}(u,y)\, dy\, du}{f_Y(y)\, dy} \\ &= \int_{-\infty}^{x} \underbrace{\frac{f_{X,Y}(u,y)}{f_Y(y)}}_{\substack{\text{Dichtefkt. der bed. Vtlg.} \\ \text{von } X \text{ geg. } Y = y}} du \end{aligned}$$

Daher erhält man folgende Definition:

Definition 8.8.1
*Die **bedingte Verteilungsfunktion** von X, gegeben $Y = y$ ist definiert als*

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} \frac{f_{X,Y}(u,y)}{f_Y(y)}\, du$$

*für alle y mit $f_Y(y) > 0$. Die **bedingte Dichte** von X, gegeben $Y = y$ ist somit*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Beispiel 8.8.1
Betrachten wir wieder die gemeinsame Verteilungsfunktion $f(x,y)$ von X und Y aus Beispiel 8.7.1 mit

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{x} & \text{für} \quad 0 \leq y \leq x \leq 1 \\ 0 & \text{sonst} \end{cases}$$

Für die bedingte Dichte von Y, gegeben $X = x$ ergibt sich:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{\frac{1}{x}}{1} \quad \textit{für} \quad 0 \le y \le x \\ &= \begin{cases} \frac{1}{x} & \textit{für} \quad 0 \le y \le x \\ 0 & \textit{sonst} \end{cases} \end{aligned}$$

d.h. $Y|X = x$ ist gleichverteilt auf $[0, x]$ ($Y|X \sim \mathcal{U}(1, x)$).

Für die Dichte von X, gegeben $Y = y$ erhält man:

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{\frac{1}{x}}{\log(\frac{1}{y})} \quad \textit{für} \quad y \le x \le 1 \\ &= \begin{cases} -1/(x\log(y)) & \textit{für} \quad y \le x \le 1 \\ 0 & \textit{sonst} \end{cases} \end{aligned}$$

Bemerkung:
Bedingte Verteilungen sind sehr nützlich zum Simulieren aus gemeinsamen Verteilungen. Da

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) \cdot f_Y(y)$$

gilt, kann man zunächst eine Zufallsvariable $Y = y$ aus der Randverteilung $f_Y(y)$ ziehen, und dann bedingt auf $Y = y$ eine Zufallszahl aus der bedingten Verteilung $f_{X|Y}(x|y)$ ziehen.
Oder andersherum:

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) \cdot f_X(x) \tag{8.3}$$

Im Beispiel 8.8.1 wäre Version (8.3) einfacher zu implementieren.

In R:

```
> x <- runif(1000)
> y <- runif(1000,0,x)
> plot(x,y)
```

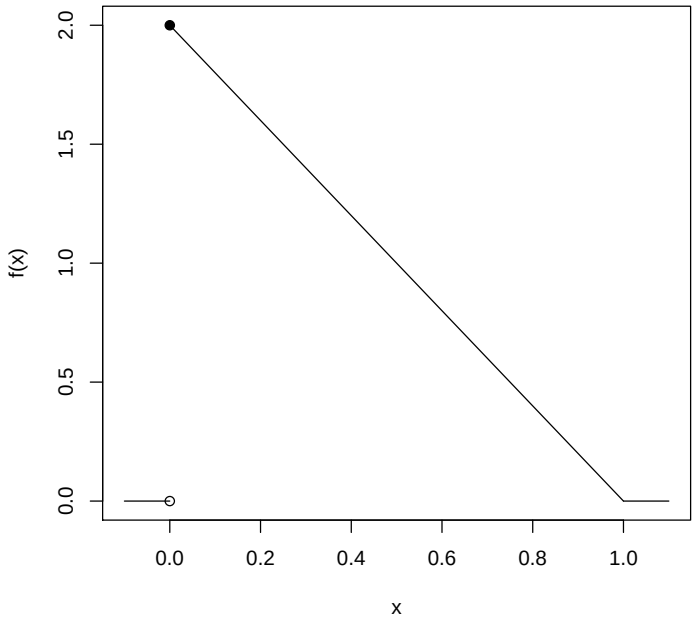

Abbildung 8.9: Die gemeinsame Dichte aus Beispiel 8.8.1

Beispiel 8.8.2
Seien X und Y bivariat standardnormalverteilt. Dann ist die bedingte Dichte von X, gegeben Y

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{\frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\frac{1}{(1-\rho^2)}(x^2-2\rho xy+y^2)\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)} \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\frac{(x-\rho y)^2}{(1-\rho^2)}\right) \end{aligned}$$

Daher ergibt sich: $X|Y=y \sim \mathcal{N}(\rho\cdot y, 1-\rho^2)$
Analog erhält man die Dichte von Y, gegeben X: $Y|X=x \sim \mathcal{N}(\rho\cdot x, 1-\rho^2)$

Nun kann man aus der bivariaten Standardnormalverteilung simulieren.

In R:

```
> x <- rnorm(1000)
> rho <- 0.5
```

```
> y <- rnorm(1000, mean = rho*x, sd = sqrt(1-rho^2))
> plot(x,y)
```

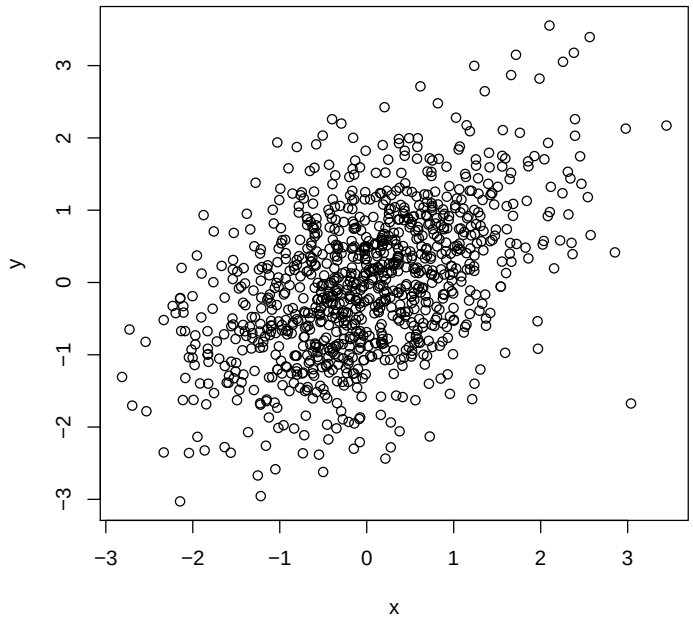

Abbildung 8.10: Die bivariate Standardnormalverteilung aus Beispiel 8.8.2

Kapitel 9

Elemente statistischer Inferenz II

In diesem Kapitel werden wir schon bekannte Grundlagen der statistischen Inferenz auf stetige Zufallsvariablen übertragen und außerdem noch die Tests auf Modellanpassung ("goodness of fit tests") behandeln.

9.1 Likelihood-Inferenz für stetige Zufallsvariablen

Die Likelihood-Inferenz, die wir schon für diskrete Zufallsvariablen kennengelernt haben, lässt sich analog auch auf stetige Zufallsvariablen anwenden. Dies wird an folgendem Beispiel deutlich:

Beispiel 9.1.1
Seien $X_1, X_2, \ldots, X_n$ unabhängige Beobachtungen aus einer $\mathcal{E}(\lambda)$-Verteilung. Wie lautet der ML-Schätzer von λ und dessen Standardfehler?
Die Likelihood-Funktion lautet:

$$L(\theta = \lambda) \;=\; \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) \;=\; \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

Durch Logarithmieren der Likelihood erhält man die Loglikelihoodfunktion:

$$l(\theta = \lambda) \;=\; n \cdot \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

Der ML-Schätzer war definiert als das Maximum der (Log-)Likelihood-Funktion.

Daher bilden wir die erste Ableitung nach λ. Dabei gilt $\bar{x} = \sum_{i=1}^{n} x_i/n$.

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \stackrel{!}{=} 0$$
$$\Rightarrow \quad \hat{\lambda}_{ML} = 1/\bar{x}$$

Ferner erhält man die zweite Ableitung der Loglikelihood: $l''(\lambda) = -\frac{n}{\lambda^2}$
Daher folgt für den Standardfehler:

$$SE(\hat{\lambda}_{ML}) \;=\; \sqrt{[-l''(\hat{\lambda}_{ML})]^{-1}} \;=\; \sqrt{\frac{\hat{\lambda}^2_{ML}}{n}} \;=\; \sqrt{\frac{(1/\bar{x})^2}{n}} \;=\; \frac{1}{\sqrt{n}\cdot\bar{x}}$$

Zahlenbeispiel: Basierend auf $n = 10$ simulierten Beobachtungen aus einer Exponentialverteilung mit wahrem Parameter $\lambda = 2$ ergab sich:

$$\sum_{i=1}^{n} x_i = 6.494 \quad \rightsquigarrow \hat{\lambda}_{ML} = \frac{10}{6.494} = 1.540$$
$$SE(\hat{\lambda}_{ML}) = \frac{\sqrt{10}}{6.494} = 0.4869$$

Ein 95%-Wald-Intervall ist dann:

$$\hat{\lambda}_{ML} \pm 1.96 \cdot SE(\hat{\lambda}_{ML}) \;=\; 1.540 \pm 1.96 \cdot 0.4869 \;=\; [0.585\,,\, 2.495]$$

Um zu entscheiden, wie gut die Approximation des Konfidenzintervalls ist, kann man eine Simulationsstudie durchführen. Dabei werden, mit dem wahren Wert $\theta = \lambda$ und variierendem n, m Konfidenzintervalle berechnet, die jeweils auf n exponentialverteilten Zufallsvariablen basieren.
Anschließend wird die empirische **Verteilung** des ML-Schätzers und die **Überdeckungshäufigkeit**, d.h. der Anteil der Konfidenzintervalle, die den wahren Wert beinhalten, berechnet. Dabei stellt man fest, dass die Überdeckungshäufigkeit nahe beim nominalen Konfidenzniveau von hier 95% liegt. Weiterhin liegt die empirische Verteilung der berechneten ML-Schätzer zentriert um den wahren Wert λ und ähnelt der Form nach einer Normalverteilung. Weshalb ist dies der Fall?
Man kann zeigen, dass (unter Regularitätsbedingungen) asymptotisch, d.h. für großen Stichprobenumfang, für den ML-Schätzer Folgendes gilt:

$$\hat{\theta}_{ML} \stackrel{a}{\sim} \mathcal{N}(\mu = \theta, \sigma^2 = SE(\hat{\theta}_{ML})^2)$$

Nach der Standardisierung ($\mathrm{E}(X) = 0, \mathrm{Var}(X) = 1$) erhält man

$$\tilde{\theta}_{ML} = \frac{\hat{\theta}_{ML} - \theta}{SE(\hat{\theta}_{ML})} \stackrel{a}{\sim} \mathcal{N}(0, 1)$$

Das heißt, der ML-Schätzer ist asymptotisch **unverzerrt** und **normalverteilt** mit einer Standardabweichung, welche gleich dem Standardfehler ist. Genauer müsste man sagen, dass der Standardfehler ein geeigneter Schätzer der Standardabweichung des ML-Schätzers ist.
Mit dieser Kenntnis ist eine Motivation für die Formel der Wald-Intervalle möglich:
Sei dazu $x_\alpha = \Phi^{-1}(\alpha)$ das α-Quantil der Standardnormalverteilung. Dann ist $\tilde{\theta}_{ML}$ asymptotisch im Intervall $[x_{\alpha/2}, x_{1-\alpha/2}]$ mit der Wahrscheinlichkeit $\beta = 1 - \alpha$. Wegen der Symmetrie der Normalverteilung ist $x_{\alpha/2} = -x_{1-\alpha/2}$. Nun ist klar, welchen Wert der Faktor d in der Formel

$$\hat{\theta}_{ML} \pm d \cdot SE(\hat{\theta}_{ML})$$

für die Berechnung des Konfidenzintervalles zum Niveau β für θ hat, nämlich $d = x_{1-\alpha/2}$.
Hier noch einmal die wichtigsten Werte von d für unterschiedliche Niveaus β:

β	d
0.9	1.645
0.95	1.960
0.99	2.576

Beispiel 9.1.2
Nun werden wieder exemplarisch n unabhängige exponentialverteilte Zufallsvariablen $X_1, X_2, \ldots, X_n \sim \mathcal{E}(\lambda)$ betrachtet.
Wie bereits bekannt erhält man für Erwartungswert und Varianz die Werte $\mu = E(X_i) = 1/\lambda$ und $Var(X_i) = 1/\lambda^2$.
Wir wollen einen ML-Schätzer für $\mu = 1/\lambda$ finden. Es ergibt sich

$$\begin{aligned} \hat{\mu}_{ML} &= \frac{1}{\hat{\lambda}_{ML}} \text{ (wegen Invarianzeigenschaft)} = \bar{x} \\ SE(\hat{\mu}_{ML}) &= \bar{x}/\sqrt{n} = \hat{\mu}_{ML}/\sqrt{n} \text{ (ohne Beweis)} \end{aligned}$$

Als 95%-Wald-Intervall für μ erhält man damit:

$$\bar{x} \pm 1.96 \cdot \bar{x}/\sqrt{n}$$

Nun wird der ML-Schätzer als Zufallsvariable betrachtet, wobei $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ mit unabhängigen $X_i \sim \mathcal{E}(\lambda)$. Es folgt für Erwartungswert und Varianz:

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu \\ Var(\bar{X}) &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\frac{1}{\lambda^2}}{n} = \mu^2/n = (\mu/\sqrt{n})^2 \end{aligned}$$

Das heißt, dass der ML-Schätzer $\bar{X}$ ***erwartungstreu*** *mit Standardabweichung $\sigma = \mu/\sqrt{n}$ ist.*
Wegen dem zentralen Grenzwertsatz gilt:

$$\bar{X} \overset{a}{\sim} \mathcal{N}(\mu, \sigma^2 = \mu^2/n)$$

Basierend auf dieser Eigenschaft kann man auch noch ein 95%-Konfidenzintervall für μ angeben. Es lautet

$$\bar{x} \pm 1.96 \cdot \mu/\sqrt{n}$$

Weil aber μ unbekannt ist (λ ist ja nicht bekannt) muss dieser Wert geschätzt werden. Dies geschieht beispielsweise durch eine "plug-in"-Schätzung mit $\hat{\mu}_{ML} = \bar{x}$. Damit sieht man, dass das Intervall basierend auf dem zentralen Grenzwertsatz und das Wald-Intervall (basierend auf dem Standardfehler) identisch ist:

$$\bar{x} \pm 1.96 \cdot \bar{x}/\sqrt{n}$$

Dieses Beispiel zeigt, dass der **Standardfehler** ein empirischer Schätzer der **Standardabweichung** des ML-Schätzers ist. Dabei muss beachtet werden, dass eine exakte Analogie nicht immer funktioniert, zumindest aber asymptotisch gilt.
Beispielsweise gilt dies nicht für $\hat{\lambda}_{ML} = 1/\bar{x}$ mit $SE(\hat{\lambda}_{ML}) = \hat{\lambda}_{ML}/\sqrt{n}$ mit

$$\begin{aligned} \mathrm{E}(1/\bar{X}) &= \lambda \frac{n}{n-1} \\ \mathrm{Var}(1/\bar{X}) &= \frac{\lambda^2}{(n-2)} \frac{n^2}{(n-1)^2} \qquad \text{(ohne Beweis)} \end{aligned}$$

In den folgenden beiden Beispielen gilt die Analogie exakt:

Beispiel 9.1.3 (Schätzung von π im Binomialexperiment)
Sei $X \sim \mathcal{B}(n, \pi)$. Wie bereits berechnet (siehe Beispiel 6.1.1), ist der ML-Schätzer gleich der relativen Häufigkeit: $\hat{\pi}_{ML} = x/n = \bar{x}$. Der Standardfehler lautet (siehe Beispiel 6.1.9):

$$SE(\hat{\pi}_{ML}) = \sqrt{\frac{\hat{\pi}_{ML}(1 - \hat{\pi}_{ML})}{n}}$$

Nun berechnen wir noch Erwartungswert und Varianz des ML-Schätzers $\hat{\pi}_{ML}$:

$$\begin{aligned} E(\hat{\pi}_{ML}) &= E\left(\frac{x}{n}\right) = \frac{1}{n} \cdot E(x) = \pi \\ Var(\hat{\pi}_{ML}) &= Var\left(\frac{x}{n}\right) = \frac{1}{n^2} \cdot Var(x) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n} \end{aligned}$$

Da die Standardabweichung gleich der Wurzel aus der Varianz ist, hat man die Analogie gezeigt.

Beispiel 9.1.4 (Schätzung von q im H-W-Gleichgewicht)
Für die drei möglichen Genotypen gelten unter Annahme einer Trinomialverteilung folgende Wahrscheinlichkeiten:

$$\begin{array}{lll} P(a,a) & = q^2 & = \pi_1 \\ P(a,b) & = 2q(1-q) & = \pi_2 \\ P(b,b) & = (1-q)^2 & = \pi_3 \end{array}$$

Angenommen wir beobachten $\boldsymbol{x} = (x_1, x_2, x_3) = (600, 320, 80)$ bei $n = 1000$ und wollen wissen, wie groß $q = \pi_1 + \pi_2/2$ (Wahrscheinlichkeit für Allel a) ist. Der direkte Weg zur Berechnung des ML-Schätzers verläuft über die Likelihood-Funktion (mit x_i =Anzahl der jeweils aufgetretenen Genotypen), welche lautet:

$$L(q) = \pi_1^{x_1} \cdot \pi_2^{x_2} \cdot \pi_3^{x_3} = (q^2)^{x_1} \cdot [2q(1-q)]^{x_2} \cdot \left[(1-q)^2\right]^{x_3}$$

Eine Maximierung der Likelihood ergibt den ML-Schätzer

$$\hat{q}_{ML} = \frac{x_1 + x_2/2}{n}$$

Die andere Möglichkeit zur Berechnung des ML-Schätzers nutzt die Invarianzeigenschaft aus. Wir wissen, die ML-Schätzer von π_1 und π_2 sind $\hat{\pi}_1 = x_1/n$ und $\hat{\pi}_2 = x_2/n$, daher erhält als ML-Schätzer für q

$$\hat{q}_{ML} = \hat{\pi}_1 + \hat{\pi}_2/2 = \frac{x_1 + x_2/2}{n}$$

Ohne es genau herzuleiten, gilt, dass die zugehörige Zufallsvariable $(X_1 + X_2/2)/n$ die Varianz $\frac{1}{2}\, q\,(1-q)$ hat.
Andererseits kann man zeigen, dass der Standardfehler von $\hat{q}_{ML}$ gleich

$$SE(\hat{q}_{ML}) = \sqrt{\frac{1}{2}\, \hat{q}_{ML}\,(1 - \hat{q}_{ML})}$$

ist.
$\Rightarrow$ Auch hier erhält man den Standardfehler durch "plug-in" des ML-Schätzers in der Formel für die Varianz des ML-Schätzers.

Für die oben angegebenen beobachteten Werte für die drei Genotypen berechnet sich der ML-Schätzer für q dann zu $\hat{q}_{ML} = (600 + 320/2)/1000 = 0.76$.

Daraus lassen sich die Anzahlen berechnen, die erwartet werden, wenn q den Wert seines ML-Schätzers annimmt:

$$E(\boldsymbol{x}) = n \cdot (\hat{q}_{ML}^2, 2\hat{q}_{ML}(1-\hat{q}_{ML}), (1-\hat{q}_{ML})^2) = (577.6\,,\, 364.8\,,\, 57.6)$$

Es stellt sich nun die Frage, ob der Unterschied zwischen erwarteten und beobachteten Anzahlen "zufällig" ist oder darauf hindeutet, dass sich die Population nicht im H-W-Gleichgewicht befindet. Diese Frage wird in Abschnitt 9.2 beantwortet.

Im Folgenden werden nun die Likelihood-Intervalle motiviert. Zunächst erinnern wir uns, dass in Beispiel 8.5.2 festgestellt wurde, dass das Quadrat der Standardnormalverteilung $Y = X^2$, falls $X \sim \mathcal{N}(0,1)$, gammaverteilt ist mit $\mathcal{G}(.5,.5)$. Dies entspricht auch einer χ^2-Verteilung mit 1 Freiheitsgrad: $Y = X^2 \sim \chi_1^2$.

Wie schon bekannt, kann man eine Taylor-Approximation der Loglikelihood durchführen:

$$\begin{aligned} l(\theta) &\approx l(\hat{\theta}_{ML}) + \underbrace{l'(\hat{\theta}_{ML})}_{=0}(\theta - \hat{\theta}_{ML}) + \frac{1}{2}\underbrace{l''(\hat{\theta}_{ML})}_{-\left[SE(\hat{\theta}_{ML})^2\right]^{-1}}(\theta - \hat{\theta}_{ML})^2 \\ &= l(\hat{\theta}_{ML}) - \frac{1}{2}\frac{(\hat{\theta}_{ML} - \theta)^2}{SE(\hat{\theta}_{ML})^2} \\ &= l(\hat{\theta}_{ML}) - \frac{1}{2}\tilde{\theta}_{ML}^2 \end{aligned}$$

Da gilt, dass $\tilde{\theta}_{ML} \overset{a}{\sim} \mathcal{N}(0,1)$ folgt:

$$\begin{aligned} -2\log\frac{L(\theta)}{L(\hat{\theta}_{ML})} &= -2\tilde{l}(\theta) \\ &= -2\left(l(\theta) - l(\hat{\theta}_{ML})\right) \\ &\approx -2\left(l(\hat{\theta}_{ML}) - \frac{1}{2}\tilde{\theta}_{ML}^2 - l(\hat{\theta}_{ML})\right) \\ &= \tilde{\theta}_{ML}^2 \overset{a}{\sim} \chi_1^2 \\ \Leftrightarrow\ \tilde{l}(\theta) &\overset{a}{\sim} -1/2 \cdot \chi_1^2 \end{aligned}$$

Für die Berechnung der Konfidenzintervalle fehlen nur noch die kritischen Werte c in

$$\begin{aligned} \{\theta : \tilde{l}(\theta) &\geq c\} \\ \text{bzw. } \{\theta : \tilde{L}(\theta) &\geq \exp(c)\} \end{aligned}$$

Diese ergeben sich einfach durch Transformation der entsprechenden Quantile $x_\alpha = F^{-1}(\alpha)$ der χ^2_1-Verteilung:

$$c = -x_{1-\alpha}/2$$

Es folgen noch einmal die wichtigsten Werte für c für die gängigsten Niveaus β:

β	c
0.9	-1.33
0.95	-1.92
0.99	-3.32

9.2 Modellanpassung

Häufig ist es von Interesse, die Anpassung eines bestimmten stochastischen Modells an vorliegende Daten zu studieren. Dies ist insbesondere bei kategorialen Daten der Fall.

Beispiel 9.2.1
In Beispiel 9.1.4 hatten wir uns gefragt, ob sich die Population, von der man N Individuen mit $X = (aa, ab, bb) = (600, 320, 80)$ untersucht hatte, im Hardy-Weinberg-Gleichgewicht befindet. Dazu werden die empirischen Häufigkeiten berechnet und mit den beobachteten verglichen.

Beispiel 9.2.2
Gegeben sind Daten einer Umfrage zum Promotionsinteresse von Männern und Frauen. Es stellt sich die Frage, ob die beiden Variablen "Geschlecht" und "Promotionsinteresse" unabhängig sind.

	Interesse		
	Ja	*Nein*	
♂	*5*	*12*	*17*
♀	*6*	*5*	*11*
	11	*17*	*28*

Beispiel 9.2.3
Im 19. Jahrhundert wurden Daten über die Häufigkeit von männlichen Nachkommen bei 6115 Familien mit jeweils (!) 12 Kindern erhoben.

# Jungen	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
# Fam.	*3*	*24*	*104*	*286*	*670*	*1033*	*1343*	*1112*	*829*	*478*	*181*	*45*	*7*

Die Frage ist, ob diese Verteilung einer Binomialverteilung folgt.

In allen drei Beispielen möchte man ein bestimmtes Modell (das sogenannte **"Null-Modell"**, **"Null-Hypothese"**) mit dem allgemeinen Modell (das **"saturierte" Modell**) unter der Annahme einer Multinomialverteilung $M_K(n, \pi)$ (K ist die Anzahl der Kategorien) vergleichen.

Beispiel	**Nullmodell**	**# Parameter p**	**# Kategorien K**
9.2.1	Population ist im H-W Gleichgewicht	1 (Parameter q)	3
9.2.2	Variablen sind unabhängig	2 ($\pi_{♂}$, $\pi_{\text{Interesse}}$)	4
9.2.3	Daten sind binomial verteilt	1 (π in Binomialverteilung)	13

Zum statistischen Testen der "Null-Hypothese" (H_0) geht man nach folgendem Schema vor:

1. *ML-Schätzung der unbekannten Parameter im Null-Modell*
   Beispiel 9.2.1:
   $$\hat{q} = \frac{600 + \frac{300}{2}}{1000} = 0.76$$

   Beispiel 9.2.2:
   $$\hat{\pi}_{\male} = \frac{17}{28} \quad \hat{\pi}_{\text{Interesse}} = \frac{11}{28}$$

   Beispiel 9.2.3:
   $$\hat{\pi} = \frac{0 \cdot 3 + 1 \cdot 24 + \ldots + 12 \cdot 7}{6115 \cdot 12} = 0.519215$$

2. *Berechnung der erwarteten Anzahl E_i an Fällen in Kategorie i unter Annahme des Null-Modells*
   Bsp. 9.2.1:
   $$\begin{aligned} \hat{\pi}_1 = \hat{q}^2 &= 0.5776 \\ \hat{\pi}_2 = 2\hat{q}(1-\hat{q}) &= 0.3648 \\ \hat{\pi}_3 = (1-\hat{q})^2 &= 0.0576 \end{aligned}$$

   $\rightsquigarrow$ Erwartete Anzahl bei $n = 1000$ Individuen:

   $$\begin{aligned} \mathrm{E}_1 = n \cdot \hat{\pi}_1 &= 577.6 \\ \mathrm{E}_2 = n \cdot \hat{\pi}_2 &= 364.8 \\ \mathrm{E}_3 = n \cdot \hat{\pi}_3 &= 57.6 \end{aligned}$$

   Bsp. 9.2.2:
   Unter Unabhängigkeit gilt z.B.:
   $$\hat{\pi}_1 = P(\male \,\&\, \text{Interesse}) = \hat{\pi}_{\male} \cdot \hat{\pi}_{\text{Interesse}} = \frac{17}{28} \cdot \frac{11}{28}$$

   und für die erwarteten Fälle folgt:
   $$\mathrm{E}_1 = n \cdot \hat{\pi}_1 = 28 \cdot \frac{17}{28} \cdot \frac{11}{28} = \frac{17 \cdot 11}{28} \approx 6.68$$

   Verfährt man für die anderen Felder analog, so erhält man insgesamt (erwartete Anzahlen stehen in Klammern)

	Ja	Nein	
m.	5 (6.68)	12 (10.32)	17
w.	6 (4.32)	5 (6.68)	11
	11	17	

Bsp. 9.2.3:
Hier ergeben sich unter Verwendung der binomialen Wahrscheinlichkeitsfunktion $P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$, $x = 0, \ldots, 12$ mit $n = 12$ und $\pi = 0.519215$ folgende erwartete Häufigkeiten E_i:

	0	1	2	3	...	11	12
X_i	3	24	104	286	...	45	7
E_i	0.9	12.1	71.8	258.5	...	26.1	2.3

3. *Berechnung eines Gesamtmaßes für die Abweichung der beobachteten Werte von den erwarteten Werten*
Dazu verwendet man beispielsweise das **Pearson'sche χ^2-Maß**:

$$\chi^2 = \sum_{i=1}^{K} r_i^2 = \sum_{i=1}^{K} \frac{(X_i - \mathrm{E}_i)^2}{\mathrm{E}_i},$$

wobei X_i bzw. E_i die tatsächlich beobachteten bzw. erwarteten Anzahlen in Kategorie i sind.
Unter der Annahme, dass H_0 wahr ist, hat χ^2 eine (asymptotische) χ^2-Verteilung mit $k = K - 1 - p$ Freiheitsgraden (wobei K die Anzahl der Kategorien und p die Anzahl der Modellparameter ist).

Damit erhält man eine Möglichkeit, um einen ***p*-Wert** (***p*-value**) zu berechnen. Dieser gibt die Wahrscheinlichkeit an, unter der Annahme von H_0 ein solches oder ein noch extremeres Resultat zu beobachten. Die Berechnung geschieht über Quantile der χ^2-Verteilung mit k Freiheitsgraden.
Ist der p-Wert klein, so kann die Nullhypothese verworfen werden.

In den 3 Beispielen:

Bsp.	χ^2	k	p-Wert	D	p-Wert
9.2.1	15.08	1	0.0001	14.36	0.00015
9.2.2	1.769	1	0.18	1.765	0.18
9.2.3	110.5	11	0	97.0	$6.66 \cdot 10^{-16}$

Schlussfolgerung aus dieser Tabelle ist:

- Beispiel 9.2.1: Die zugrundeliegende Population befindet sich offensichtlich nicht im Hardy-Weinberg-Gleichgewicht.
- Beispiel 9.2.2: Die Nullhypothese, dass die Variablen unabhängig sind, kann auf den üblichen Signifikanzniveaus nicht abgelehnt werden.
- Beispiel 9.2.3: Die Anzahl der männlichen Nachkommen ist offenbar nicht binomialverteilt.

Bemerkungen:

1. Die χ^2-Verteilung gilt nur asymptotisch, d.h. für $n \to \infty$:
   Faustregel: Alle $\mathrm{E}_i > 1$ und mindestens 80% der $E_i's > 5$.

2. Alternativ bietet sich auch die Berechnung der **Devianz** als Maß für die Abweichung in den Kategorien an:

$$D = 2 \cdot \sum_{i=1}^{K} X_i \log(\frac{X_i}{\mathrm{E}_i})$$

   Diese besitzt unter H_0 die gleiche asymptotische Verteilung wie χ^2. (diese Formel kann über die Likelihood-Funktion der Multinomialverteilung motiviert werden)

3. Man kann in jeder Kategorie noch sogenannte **"Residuen"** berechnen, die angeben, wie weit die beobachteten Fälle X_i und die erwarteten Fälle E_i auseinanderliegen, und in welche Richtung der Unterschied geht. Das i-te Residuum ist definiert als

$$r_i = \frac{X_i - \mathrm{E}_i}{\sqrt{\mathrm{E}_i}}.$$

   Unter H_0 sind die χ^2-Residuen approximativ und asymptotisch standardnormalverteilt, d.h. $r_i \sim N(0,1)$. Residuen mit $|r_i| > 2$ deuten auf schlechte Modellanpassung hin.

Zum Abschluss dieses Abschnitts wird noch ein Beispiel für die Modellanpassung bei Markov Ketten behandelt.

Beispiel 9.2.4 (Regen bei den Snoqualmie Wasserfällen)
Über eine Zeitreihe der Länge $N = 13149$ Tage wurde an den Snoqualmie Wasserfällen registriert, ob es am jeweiligen Tag geregnet hat oder nicht. Nun wird ein Markov-Modell gesucht, welches sich den Daten möglichst genau anpasst.
Der Zustandsraum hat zwei Zustände $S = \{0, 1\}$ mit

- 0: *Kein Regen am Tag* $t = 1, \ldots, N$
- 1: *Regen am Tag* $t = 1, \ldots, N$

Insgesamt war es an $n_0 = 6229$ *Tagen trocken und an* $n_1 = 6920$ *Tagen hat es geregnet. Die Übergangswahrscheinlichkeiten in der Übergangsmatrix* $\boldsymbol{P}$ *werden durch die jeweiligen relativen Übergangshäufigkeiten (ML-Schätzer) geschätzt. Es ergibt sich*

$$\hat{\boldsymbol{P}} = \begin{pmatrix} 0.713 & 0.287 \\ 0.258 & 0.742 \end{pmatrix}$$

Die stationäre Verteilung $\boldsymbol{\pi}$ *ist* $\boldsymbol{\pi} = (0.474\,,\, 0.526)$.
Das Markov-Modell wäre somit spezifiziert, es stellt sich lediglich die Frage, ob das Modell den Daten auch gut angepasst ist.
Dazu führt man einen χ^2*-Anpassungstest durch, wobei man die "Verweildauern", das heißt die Wartezeiten bis zum nächsten Zustandswechsel, betrachtet. In beiden Gruppen (kein Regen, Regen) wird jeweils* $p = 1$ *Parameter geschätzt. Man unterteilt die Verweildauern in* $K = 10$ *Kategorien* $\{0, 1, \ldots, 8, > 8\}$ *Tage. Damit hat der Test* $k = K - 1 - p = 8$ *Freiheitsgrade.*

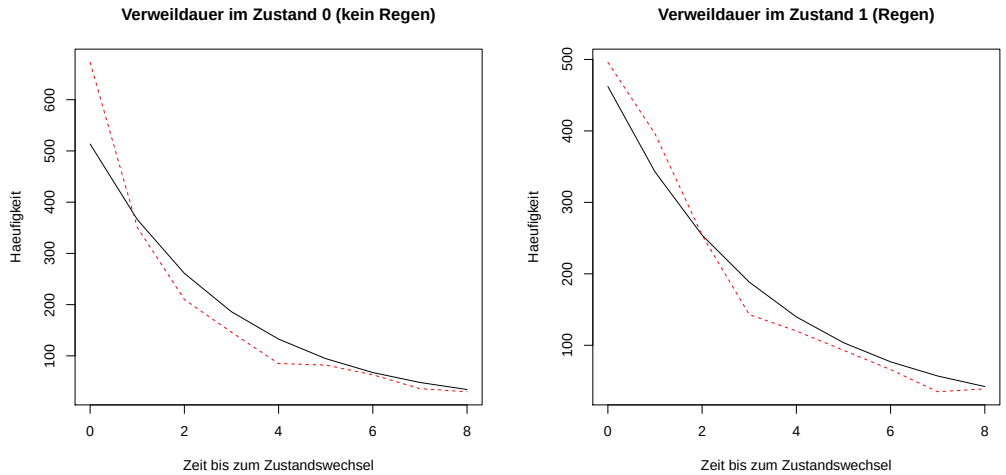

Abbildung 9.1: Verweildauern in den Zuständen "kein Regen" und "Regen" bei den Snoqualmie Wasserfällen

Für die beiden Gruppen ergeben sich folgende Werte für die χ^2-Anpassungsstatistik und die p-Werte:

$KeinRegen$		$Regen$	
χ^2	p-$Wert$	χ^2	p-$Wert$
99.9	0	41.4	$1.8e-6$

Diese Werte (p-Werte sind effektiv gleich Null) drücken eine große Diskrepanz zwischen dem Modell und den Daten aus! Wir benötigen also einen besseren Modellierungsansatz.
Wir könnten eine Markov-Kette höherer Ordnung verwenden. Beispielsweise eine ***Markov-Kette zweiter Ordnung****, bei welcher die Wahrscheinlichkeit, ob es an einem Tag regnet oder nicht, davon abhängt, welches Wetter an den* ***beiden*** *vorangegangenen Tagen geherrscht hat. Die Übergangswahrscheinlicheiten haben die Form*

$$P(X_n = s|X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2})$$

Eine Markov-Kette zweiter Ordnung kann als Markov-Kette (erster Ordnung) dargestellt werden, wenn man den Zustandsraum zu $S = \{00, 01, 10, 11\}$ erweitert, wobei der erste Eintrag x_{n-2} und der zweite Eintrag x_{n-1} darstellt. In der Übergangsmatrix $\boldsymbol{P}$ ergeben sich dann **strukturelle Nullen** *(unmögliche Übergänge):*

$$\boldsymbol{P} = \left(\begin{array}{c||cccc} & 00 & 01 & 10 & 11 \\ \hline\hline 00 & & & 0 & 0 \\ 01 & 0 & 0 & & \\ 10 & & & 0 & 0 \\ 11 & 0 & 0 & & \end{array}\right)$$

Zum Beispiel kann auf $X_{n-2} = 0$ und $X_{n-1} = 0$ nicht $X_{n-1} = 1$ und $X_n = 0$ folgen, ... Insgesamt kann also die Anzahl der Spalten reduziert werden. Man erhält als Werte:

$$\hat{\boldsymbol{P}} = \left(\begin{array}{c||cc} & 0 & 1 \\ \hline\hline 00 & 0.749 & 0.251 \\ 01 & 0.277 & 0.723 \\ 10 & 0.624 & 0.376 \\ 11 & 0.252 & 0.748 \end{array}\right)$$

Ein alternativer Ansatz zur Modellierung wäre die Wahl eines Hidden Markov Modells.

Kapitel 10

Lineare Regression

10.1 Kleinste-Quadrate-Schätzung

Es ist eine Gerade $y = \beta_1 + \beta_2 x$ gesucht, welche die Punktwolke in Abb. 10.1 'bestmöglichst' approximiert.
Dazu betrachten wir eine bivariate Zufallsvariable (Y, X) mit Beobachtungen $(y_i, x_i), i = 1, \ldots, n$ und definieren den Schätzer für die Parameter der Gerade als

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{\beta_1, \beta_2}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Dieser Schätzer $(\hat{\beta}_1, \hat{\beta}_2)$ heißt (aus offensichtlichen Gründen) Kleinster-Quadrate-Schätzer und repräsentiert diejenige Gerade durch die Punktwolke, welche den quadratischen vertikalen Abstand jeder Beobachtung zur Geraden minimiert. Andere Kriterien sind denkbar, wie etwa

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{\beta_1, \beta_2}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - \beta_1 - \beta_2 x_i|.$$

Um nicht nur bivariate Probleme behandeln zu können, betrachten wir einen Zufallsvektor Y, welcher alle n Beobachtungen der Zielgröße zusammenfasst:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n$$

```
> set.seed(290875)
> x <- runif(10)
> y <- 3 + 2 * x + rnorm(length(x), sd = 0.1)
> plot(x, y, xlim = c(0, 1))
```

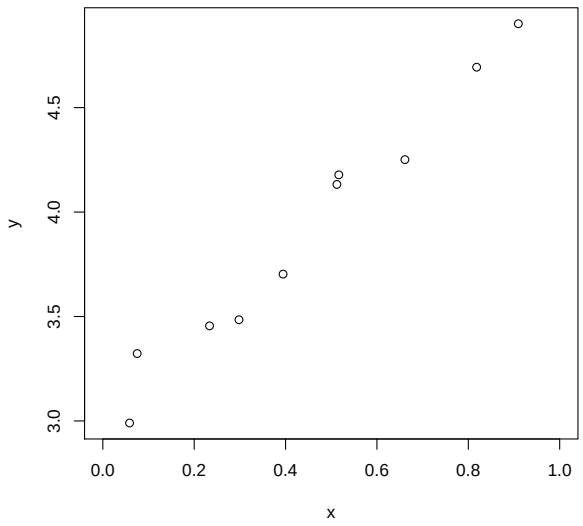

Abbildung 10.1: Streudiagramm für zwei Zufallsvariablen

sowie k Zufallsvektoren

$$X^j = \begin{pmatrix} X_1^j \\ X_2^j \\ \vdots \\ X_n^j \end{pmatrix}, j = 1, \ldots, k$$

welche wir in einer Matrix $\mathbf{X} = (X^1, X^2, \ldots, X^k) \in \mathbb{R}^{n,k}$ aggregieren. Die Parameter der Geraden ($k = 1$) bzw. Hyperebenen ($k > 2$) sind durch

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \in \mathbb{R}^k$$

gegeben und wir betrachten das Modell $Y = \mathbf{X}\beta$. Gesucht ist nach dem Kriterium der Kleinsten Quadrate ein Schätzer $\hat{\beta}$, sodass $||Y - \mathbf{X}\hat{\beta}||_2 \leq ||Y - \mathbf{X}\beta||_2 \forall \beta \in \mathbb{R}^k$.

Satz 10.1.1 (Kleinste-Quadrate-Schätzung)
Sei der Rang von $\mathbf{X}$ *gleich* k*. Dann gilt:*

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||Y - \mathbf{X}\beta|| = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$$

Beweis: Sei $h(\beta) = ||Y - \mathbf{X}\beta||$ mit

$$\begin{aligned} h(\beta) &= (Y - \mathbf{X}\beta)^\top (Y - \mathbf{X}\beta) \\ &= Y^\top Y - \beta^\top \mathbf{X}^\top Y - Y^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta \end{aligned}$$

Betrachte den Gradienten

$$\frac{\partial h(\beta)}{\partial \beta}$$

und bestimme so ein Minimum von $h(\beta)$.

Satz 10.1.2 (Lemma)
Sei $\mathbf{A} \in \mathbb{R}^{q,p}$ *und* $\mathbf{B} = \mathbf{B}^\top \in \mathbb{R}^{q,q}$ *sowie* $z \in \mathbb{R}^q$*. Dann gilt:*

$$\frac{\partial z^\top \mathbf{A}}{\partial z} = \mathbf{A}$$

und

$$\frac{\partial z^\top \mathbf{B} z}{\partial z} = 2\mathbf{B}z$$

```
> plot(x, y)
> X <- cbind(1, x)
> hatbeta <- tcrossprod(solve(crossprod(X, X)), X) %*% y
> abline(a = hatbeta[1], b = hatbeta[2])
```

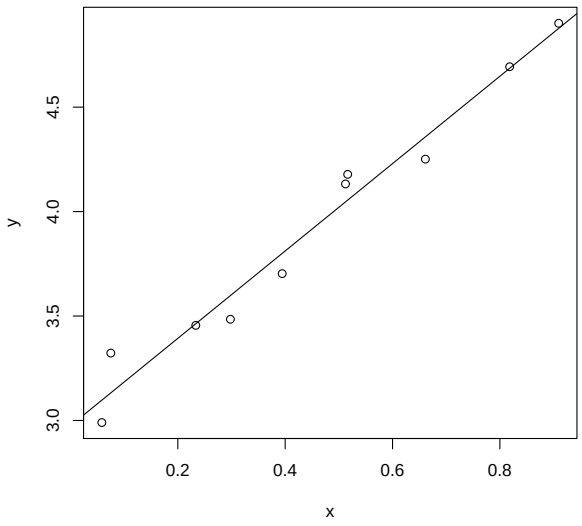

Abbildung 10.2: Streudiagramm mit KQ-Gerade

Damit ist also

$$\frac{\partial h(\beta)}{\partial \beta} = -2\mathbf{X}^\top Y + 2\mathbf{X}^\top \mathbf{X}\beta$$

und $\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top Y$ ist eine notwendige Bedingung für das Minimum. Wegen $\text{Rang}(\mathbf{X}) = k$ ist $\mathbf{X}^\top \mathbf{X}$ invertierbar und somit $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top Y$.

10.2 Das Lineare Regressionsmodell

Das Modell

$$Y = \mathbf{X}\beta + U$$

heißt lineares Regressionsmodell. Dabei ist

$$U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix} \in \mathbb{R}^n$$

ein n-dimensionaler Zufallsvektor. Das lineare Regressionsmodell hat folgende Eigenschaften

- gegeben sind mehrere stetige Merkmale $Y, X^1, \ldots, X^k$
- $X^1, \ldots, X^k$ verursachen Y und *nicht* umgekehrt
- der Zusammenhang ist linear, also $Y_i = \sum_{j=1}^{k} \beta_j X_i^j + U_i$
- die X-Variablen heißen unabhängige Variable, Regressoren, exogene Variable oder Design-Variable
- die Y-Variable heißt abhängige Variable, Regressant, endogene Variable oder Response-Variable
- U sind nicht beobachtbare Störgrößen.

Zudem treffen wir treffen drei Annahmen:

A1) $\mathbf{X}$ ist eine feste (nicht zufällige) $n \times k$ Matrix mit vollem Spaltenrang, also $\text{Rang}(\mathbf{X}) = k$.

A2) U ist ein Zufallsvektor mit $E(U) = (E(U_1), E(U_2), \ldots, E(U_n))^\top = 0$.

A3) Die Komponenten von U sind paarweise unkorreliert und haben alle die gleiche Varianz σ^2, formal: $\text{Cov}(U) = \sigma^2 \text{diag}(n)$.

Beispiel 10.2.1 (Körperfettmessung)
Garcia et al. (2005, Obesity Research) untersuchten $n = 71$ Frauen und erhoben (unter anderem) $k = 5$ Einflussgrößen (Alter, Bauchumfang, Hüftumfang, Ellenbogenbreite und Kniebreite), um deren Einfluss auf die Zielgröße, den Körperfettanteil gemessen mittels Dual Energy X-Ray Absorptiometry

(DXA), zu untersuchen. Die bivariaten Verteilungen sind mittels Streudiagrammen in Abb. 10.3 dargestellt, die Verteilung der Zielgröße mittels eines sogenannten Box-Plots (als Linien sind Minimum / Maximum sowie die $25\%, 50\%$ *und* 75% *Quantile eingetragen). Es stellen sich folgende Fragen: Welche der unabhängigen Variablen haben tatsächlich einen Einfluss auf den Körperfettanteil? Welche haben einen positiven und welche einen negativen Einfluss? Kann man aus den unabhängigen Variablen auf den Körperfettanteil schließen? Diese Fragen können mittels eines linearen Regressionsmodells beantwortet werden.*
Die Koeffizienten β *können wie folgt geschätzt werden:*

```
> bodyfat_lm <- lm(DEXfat ~ age + waistcirc + hipcirc +
+                  elbowbreadth + kneebreadth, data = bodyfat)
> coef(bodyfat_lm)

 (Intercept)          age    waistcirc      hipcirc
-59.57319910   0.06381438   0.32043969   0.43395490
elbowbreadth  kneebreadth
 -0.30117257   1.65380650
```

was äquivalent (aber numerisch stabiler ist) zu

```
> X <- bodyfat[,c("age", "waistcirc", "hipcirc",
+                 "elbowbreadth", "kneebreadth")]
> X <- cbind(1, as.matrix(X))
> Y <- bodyfat$DEXfat
> drop(tcrossprod(solve(crossprod(X, X)), X) %*% Y)

                     age    waistcirc      hipcirc
-59.57319910   0.06381438   0.32043969   0.43395490
elbowbreadth  kneebreadth
 -0.30117257   1.65380650
```

10.2.1 Eigenschaften der KQ-Methode

Offensichtlich gilt $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$. Dabei ist $\hat{\beta}$ ein Zufallsvektor und heißt KQ-Schätzer oder OLS-Schätzer (ordinary least squares).

Satz 10.2.1
Unter A1, A2 und A3 ist $\hat{\beta}$ *ein erwartungstreuer Schätzer für* β *mit Kovarianzmatrix* $Cov(\hat{\beta}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

```
> data("bodyfat", package = "TH.data")
> bodyfat <- bodyfat[,1:6]
> layout(matrix(1:6, nc = 3))
> boxplot(bodyfat$DEXfat, ylab = "DEXfat")
> plot(DEXfat ~ age, data = bodyfat)
> plot(DEXfat ~ waistcirc, data = bodyfat)
> plot(DEXfat ~ hipcirc, data = bodyfat)
> plot(DEXfat ~ elbowbreadth, data = bodyfat)
> plot(DEXfat ~ kneebreadth, data = bodyfat)
```

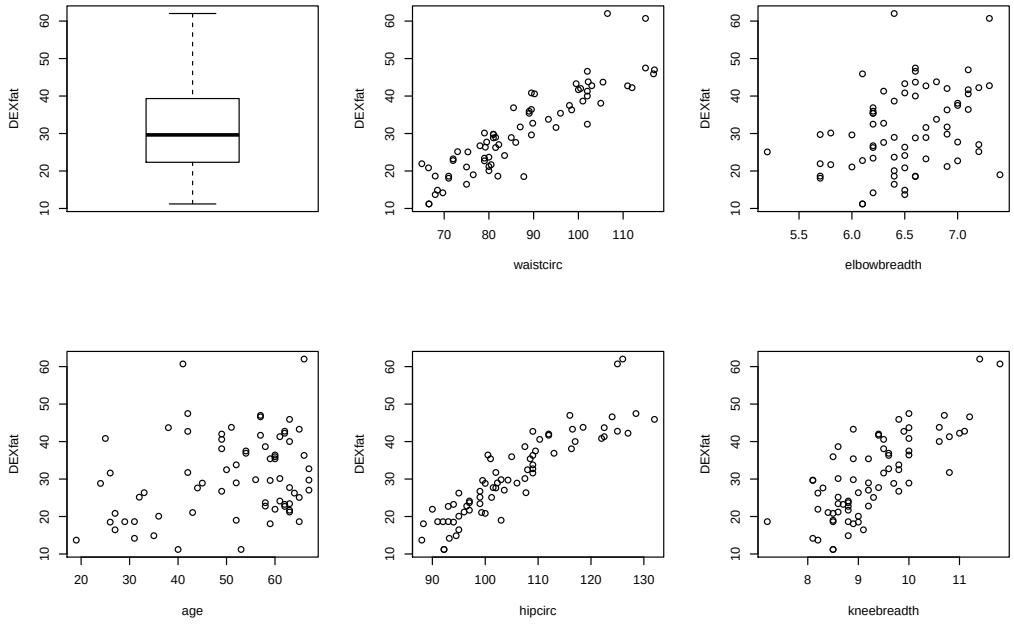

Abbildung 10.3: Bivariate Streudiagramme für Körperfettdaten

Sei $Y \in \mathbb{R}^n$ ein beliebiger Zufallsvektor mit $E(Y) = (E(Y_1), \ldots, E(Y_n))^\top$ und

$$\mathrm{Cov}(Y) = \begin{pmatrix} \mathrm{Var}(Y_1) & \mathrm{Cov}(Y_1, Y_2) & & \\ \mathrm{Cov}(Y_2, Y_1) & \mathrm{Var}(Y_2) & \mathrm{Cov}(Y_2, Y_3) & \\ \vdots & \ddots & \ddots & \\ & & & \mathrm{Var}(Y_n) \end{pmatrix}$$

mit $\mathrm{Cov}(Y) = \mathrm{Cov}(Y)^\top = E((Y - E(Y))(Y - E(Y))^\top)$.

Satz 10.2.2
Es gilt

1. *Cov*(Y) *is positiv semidefinit*
2. $E(\mathbf{A}Y) = \mathbf{A}E(Y)$
3. *Cov*$(\mathbf{A}Y) = \mathbf{A}$*Cov*$(Y)\mathbf{A}^\top$

Beweis: Nr. 2) folgt aus der Linearität des Erwartungswertes. Nr. 3)

$$\begin{aligned} \mathrm{Cov}(\mathbf{A}Y) &= E((\mathbf{A}Y - \mathbf{A}E(Y))(\mathbf{A}Y - \mathbf{A}E(Y))^\top) \\ &= E(\mathbf{A}(Y - E(Y)(Y - E(Y))^\top\mathbf{A}^\top) \\ &= \mathbf{A}E((Y - E(Y)(Y - E(Y))^\top)\mathbf{A}^\top = \mathbf{A}\mathrm{Cov}(Y)\mathbf{A}^\top \end{aligned}$$

Nr. 1) Hier ist für $x \in \mathbb{R}^n$ zu zeigen, dass $x^\top \mathrm{Cov}(Y)x \geq 0 \quad \forall x \in \mathbb{R}^n$.

$$\begin{aligned} x^\top \mathrm{Cov}(Y)x &= E(((x^\top Y - x^\top E(Y))(x^\top Y - x^\top E(Y))^\top) \\ &= E((x^\top Y - x^\top E(Y))^2) \geq 0 \quad \forall x \in \mathbb{R}^n \end{aligned}$$

und damit ist Cov positiv semidefinit.
Beweis zu Satz 10.2.1:

$$\begin{aligned} E(\hat{\beta}) &= E((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(Y) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(\mathbf{X}\beta + U) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\beta + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(U) \\ &= \beta + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top 0 = \beta \end{aligned}$$

Damit ist die Erwartungstreue von $\hat{\beta}$ gezeigt. Desweiteren gilt für die Kovarianz

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}) &= \mathrm{Cov}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{Cov}(Y)((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathrm{diag}(n)((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \end{aligned}$$

Unter Umständen sind wir an Linearkombinationen des Parametervektors β interessiert (welche auch 'Kontraste' genannt werden). Sei $c \in \mathbb{R}^k$ ein Vektor von Konstanten. Dann ist $c^\top \hat{\beta}$ eine erwartungstreue Schätzung von $c^\top \beta$ mit Kovarianzmatrix $\sigma^2 c^\top (\mathbf{X}^\top \mathbf{X})^{-1} c$.

10.2.2 Optimalität der KQ-Methode

Ein Schätzer $\tilde{\beta}$ heißt linear, wenn eine Matrix $\mathbf{C} \in \mathbb{R}^{k,n}$ existiert, sodass $\tilde{\beta} = \mathbf{C}Y$.

Satz 10.2.3 (Gauß-Markov-Theorem)
Unter A1-A3 gilt:

1. *$\hat{\beta}$ ist der beste lineare erwartungstreue Schätzer (BLUE) für β, d.h. $Cov(\hat{\beta}) \leq Cov(\tilde{\beta})$ im Sinne der Löwner-Halbordnung (d.h. $Cov(\tilde{\beta}) - Cov(\hat{\beta})$ psd).*
2. *BLUE ist eindeutig.*

Beweis:
Sei $\tilde{\beta}$ ein beliebiger linearer erwartungstreuer Schätzer für β. Dann folgt

$$E\tilde{\beta} = E(\mathbf{C}Y) = E(\mathbf{C}(\mathbf{X}\beta + U)) = \mathbf{C}\mathbf{X}\beta \stackrel{!}{=} \beta \quad \forall \beta \in \mathbb{R}^k$$

und also $\mathbf{C}\mathbf{X} = \mathrm{diag}(k)$. Betrachte $\Delta = \mathbf{C} - (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$. Wie groß ist die Differenz Δ?

$$\Delta\mathbf{X} = (\mathbf{C} - (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{X} = \mathbf{C}\mathbf{X} - \mathrm{diag}(k) = \mathrm{diag}(k) - \mathrm{diag}(k) = 0$$

$$\Rightarrow \Delta\mathbf{X} = 0.$$

$$\begin{aligned}
\mathrm{Cov}(\tilde{\beta}) &= \mathrm{Cov}(\mathbf{C}Y) \\
&= \mathbf{C}\mathrm{Cov}(Y)\mathbf{C}^\top \\
&= \sigma^2 \mathbf{C}\mathbf{C}^\top \\
&= \sigma^2((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top + \Delta)(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} + \Delta^\top) \\
&= \sigma^2((\mathbf{X}^\top \mathbf{X})^{-1} + \Delta\Delta^\top)
\end{aligned}$$

und also $\mathrm{Cov}(\tilde{\beta}) - \mathrm{Cov}(\hat{\beta}) \geq 0$.
Eindeutigkeit:

$$\begin{aligned}
\tilde{\beta} \text{ ist BLUE} &\iff \mathrm{Cov}(\tilde{\beta}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1} \\
&\iff \Delta\Delta^\top = 0 \\
&\iff \Delta = 0 \\
&\iff \tilde{\beta} = \hat{\beta}
\end{aligned}$$

Satz 10.2.4
Unter A1-A3 ist $c^\top\hat{\beta}$ der BLUE für $c^\top\beta$.

10.2.3 Prognose mit der KQ-Methode

Gegeben sei Y und $X^1, \ldots, X^k$ mit Beobachtungen $(y_i, x_i = (x_i^1, \ldots, x_i^k)), i = 1, \ldots, n$ sowie x_{n+1}. Gesucht sei y_{n+1}. Bekannt ist, dass $Y_{n+1} = x_{n+1}^\top\beta + U_{n+1}$. Da die Störgrößen U nicht beobachtbar sind, jedoch per Annahme einen Erwartungswert gleich 0 haben, schätzen wir $\hat{Y}_{n+1} = x_{n+1}^\top\hat{\beta}$.

Satz 10.2.5
Unter A1-A3 gilt $E(\hat{Y}_{n+1} - Y_{n+1}) = 0$.

Beweis:

$$\begin{aligned} E(\hat{Y}_{n+1} - Y_{n+1}) &= E(x_{n+1}^\top\hat{\beta} - x_{n+1}^\top\beta - U_{n+1}) \\ &= x_{n+1}^\top E(\hat{\beta}) - x_{n+1}^\top\beta + 0 \\ &= x_{n+1}^\top\beta + x_{n+1}^\top\beta = 0 \end{aligned}$$

Beispiel 10.2.2 (Körperfettprognose)
Für die 45jährige Emma mit Baumumfang 90cm, Hüftumfang 110cm, Ellenbogenbreite 7cm und Kniebreite 10cm ist der vorhergesagte Körperfettanteil

```
> emma <- c(intercept = 1, age = 45, waistcirn = 90,
+           hipcirc = 110,  ellbowbreadth = 7,
+           kneebreadth = 10)
> emma %*% coef(bodyfat_lm)

         [,1]
[1,] 34.30292
```

10.2.4 Schätzung von Varianz und Kovarianz

Es fehlen noch Schätzer für σ^2 und $\mathrm{Cov}(\hat{\beta})$. Dazu betrachten wir die Residuen

$$\hat{U} = Y - \mathbf{X}\hat{\beta}$$

als Ersatz für die nicht beobachtbaren Störgrößen U.

Satz 10.2.6

1. *$\hat{U} = \mathbf{M}Y = \mathbf{M}U$ mit $\mathbf{M} = diag(n) - \mathbf{H}$ wobei die sogenannte Hat-Matrix $\mathbf{H}$ gegeben ist durch $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ und $\hat{Y} = \mathbf{H}Y$ ($\mathbf{H}$ setzt dem Y den Hut auf).*

2. $\mathbf{M}$ *ist orthogonaler Projektor mit Rang (gleich Spur)* $n-k$.

Beweis:

$$\begin{aligned}\hat{U} &= Y - \mathbf{X}\hat{\beta} \\ &= Y - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y \\ &= (\mathrm{diag}(n) - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)Y \\ &= \mathbf{M}Y \\ &= \mathbf{M}(\mathbf{X}\beta + U) \\ &= \mathbf{M}U\end{aligned}$$

$\mathbf{M}$ orthogonaler Projektor $\iff$ $\mathbf{M} = \mathbf{M}^\top$ und $\mathbf{M}^2 = \mathbf{M}$.

$$\begin{aligned}\mathbf{M}^2 &= (\mathrm{diag}(n) - \mathbf{H})(\mathrm{diag}(n) - \mathbf{H}) \\ &= \mathrm{diag}(n)^2 - \mathbf{H} - \mathbf{H} + \mathbf{H}^2 \\ &= \mathrm{diag}(n) - 2\mathbf{H} + \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \\ &= \mathrm{diag}(n) - 2\mathbf{H} + \mathbf{H} \\ &= \mathrm{diag}(n) - \mathbf{H} = \mathbf{M}\end{aligned}$$

$$\begin{aligned}\mathbf{M}^\top &= \mathrm{diag}(n)^\top - \mathbf{H}^\top \\ &= \mathrm{diag}(n) - (\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top \\ &= \mathrm{diag}(n) - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{M}\end{aligned}$$

Die Spur von $\mathbf{M}$ ist nun

$$\begin{aligned}\mathrm{Spur}(\mathbf{M}) &= \mathrm{Spur}(\mathrm{diag}(n) - \mathbf{H}) \\ &= \mathrm{Spur}(\mathrm{diag}(n)) - \mathrm{Spur}(\mathbf{H}) \\ &= n - \mathrm{Spur}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) \\ &= n - \mathrm{Spur}(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}) \\ &= n - \mathrm{Spur}(\mathrm{diag}(k)) \\ &= n - k\end{aligned}$$

Satz 10.2.7
Unter A1-A3 gilt

$$\hat{\sigma}^2 = \frac{\hat{U}^\top\hat{U}}{n-k}$$

ist ein erwartungstreuer Schätzer für σ^2.

Beweis:

$$\begin{aligned}
E(\hat{U}^\top \hat{U}) &= E((MU)^\top MU) \\
&= E(U^\top \mathbf{M} U) \\
&= E(\text{Spur}(U^\top \mathbf{M} U)) \\
&= E(\text{Spur}(\mathbf{M} U U^\top)) \\
&= \text{Spur}(E(\mathbf{M} U U^\top)) \\
&= \text{Spur}(\mathbf{M} E(U U^\top)) \\
&= \text{Spur}(\mathbf{M}\text{Cov}(U)) \\
&= \text{Spur}(\mathbf{M}\sigma^2 \text{diag}(n)) \\
&= \text{Spur}(\mathbf{M}\sigma^2 \text{diag}(n)) \\
&= \sigma^2 \text{Spur}(\mathbf{M}) \\
&= \sigma^2 (n-k)
\end{aligned}$$

Damit können wir also auch die Kovarianzmatrix $\text{Cov}(\hat{\beta})$ schätzen, und zwar als

$$\hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Desweiteren ist es möglich, die geschätzten Koeffizienten zu standardisieren, um sie miteinander vergleichen zu können:

$$\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\text{diag}((\mathbf{X}^\top \mathbf{X})^{-1})}}$$

Beispiel 10.2.3 (Körperfett)
Hier sind die standardisierten Regressionskoeffizienten gegeben durch

```
> U <- bodyfat$DEXfat - X %*% coef(bodyfat_lm)
> n <- nrow(bodyfat)
> k <- length(coef(bodyfat_lm))
> sigma2 <- crossprod(U) / (n - k)
> sdbeta <- sqrt(sigma2) * sqrt(diag(solve(crossprod(X))))
> coef(bodyfat_lm) / sdbeta

 (Intercept)          age    waistcirc      hipcirc
  -7.0470856    1.7061408    4.3468577    4.5364912
elbowbreadth  kneebreadth
  -0.2474076    1.9177922
```

oder einfacher

```
> coef(bodyfat_lm) / sqrt(diag(vcov(bodyfat_lm)))

 (Intercept)          age    waistcirc      hipcirc
  -7.0470856    1.7061408    4.3468577    4.5364912
elbowbreadth  kneebreadth
  -0.2474076    1.9177922
```

10.3 Das Lineare Regressionsmodell unter Normalverteilung

Bisher haben wir außer dem Erwartungswert (A2) und der Kovarianzmatrix (A3) nichts über die Verteilung der Störgrößen U angenommen. In diesem Abschnitt betrachten wir zusätzlich

A4) $U_i \sim \mathcal{N}(0, \sigma^2)$.

10.3.1 Eigenschaften der Normalverteilung

Eine n-dimensionale Zufallsvariable Z folgt einer multivariaten Normalverteilung mit Erwartungswertvektor $\mu \in \mathbb{R}^n$ und Kovarianzmatrix $\Sigma \in \mathbb{R}^{n,n}$ (symmetrisch und pd), symbolisch

$$Z \sim \mathcal{N}(\mu, \Sigma).$$

Satz 10.3.1
Es gilt

1. $Z \sim \mathcal{N}(\mu, \Sigma) \Rightarrow E(Z) = \mu, Cov(Z) = \Sigma$ *und* $Z_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$.
2. *Sei* $\mathbf{A} \in \mathbb{R}^{p,n}$ *mit Rang gleich* p *und* $b \in \mathbb{R}^p$, *dann gilt* $\mathbf{A}Z + b \sim \mathcal{N}(\mathbf{A}\mu + b, \mathbf{A}\Sigma\mathbf{A}^\top)$.
3. *Die Komponenten von* Z *sind stochastisch unabhängig* $\iff$ $\Sigma = diag(\sigma_{ii}^2)$.
4. $\mathbf{A} \in \mathbb{R}^{p,n}, \mathbf{B} \in \mathbb{R}^{q,n}, \mathbf{A}\Sigma\mathbf{B}^\top = 0 \Rightarrow \mathbf{A}Z, \mathbf{B}Z$ *sind stochastisch unabhängig.*

Beweis: 1-3 sind klar aus den Rechenregeln für Erwartungswerte und Kovarianzmatrizen. Zu 4:

$$\begin{aligned} \mathrm{Cov}(\mathbf{A}Z, \mathbf{B}Z) &= \mathbf{A}\mathrm{Cov}(Z, \mathbf{B}Z) \\ &= \mathbf{A}\mathrm{Cov}(Z)\mathbf{B}^\top = \mathbf{A}\Sigma\mathbf{B}^\top = 0 \end{aligned}$$

und die Unabhängigkeit folgt.

10.3.2 Konsequenzen für die KQ- und Varianzschätzung

Es gilt

- $Y = \mathbf{X}\beta + U \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \text{diag}(n))$
- $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top Y \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$

Satz 10.3.2
Unter A1 - A4 sind $\hat{\beta}$ und $\hat{\sigma}^2$ stochastisch unabhängig.

Beweis: Mit Satz 10.3.1 4) gilt zunächst, dass $\hat{\beta}$ und $\hat{U} = \mathbf{M}Y$ stochastisch unabhängig sind:

$$\begin{aligned} \text{Cov}(\hat{\beta}, \hat{U}) &= \text{Cov}((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top Y, \mathbf{M}Y) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \text{Cov}(Y)\mathbf{M}^\top \\ &= \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{M}\mathbf{X})^\top = 0 \end{aligned}$$

da $\mathbf{M}\mathbf{X} = (\text{diag}(n) - (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{X} = 0$.
Somit ist $\hat{\sigma}^2 = \hat{U}^\top \hat{U}/(n-k)$ auch von $\hat{\beta}$ stochastisch unabhängig.

Satz 10.3.3
Unter A1 - A4 ist $\hat{\beta}$ die ML-Schätzung für β.

Beweis: Die Likelihoodfunktion für Y ist

$$L(Y, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - \mathbf{X}\beta)^\top (Y - \mathbf{X}\beta)\right)$$

und diese wird maximal, wenn $(Y - \mathbf{X}\beta)^\top (Y - \mathbf{X}\beta)$ minimal wird und dies ist für den KQ-Schätzer $\hat{\beta}$ der Fall.

Satz 10.3.4
Unter A1 - A4 ist $\hat{\sigma}^2_{ML} = \hat{U}^\top \hat{U}/n$ die ML-Schätzung für σ^2.

Beweis: Zu maximieren bleibt bei gegebenem β die Likelihoodfunktion

$$L^\star(\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\hat{U}^\top \hat{U}\right)$$

Dann ist

$$\begin{aligned} \frac{\partial \log(L^\star(\sigma^2))}{\partial \sigma^2} &= \frac{\partial(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\hat{U}^\top \hat{U}}{2\sigma^2})}{\partial \sigma^2} \\ &= -\frac{n}{2\sigma^2} + \frac{\hat{U}^\top \hat{U}}{2\sigma^4} = 0 \\ \iff \hat{\sigma^2} &= \frac{\hat{U}^\top \hat{U}}{n} \end{aligned}$$

10.4 Konfidenzintervalle und Tests für β

Wir möchten nun Hypothesen der Form

$$H_0 : d^\top \beta = 0 \text{ vs. } H_1 : d^\top \beta \neq 0$$

testen oder Konfidenzintervalle für den Parameter $d^\top \beta$ herleiten. Dabei ist $d \in \mathbb{R}^k$ beliebig.

Satz 10.4.1
Unter A1 - A4 gilt

$$\frac{d^\top \hat{\beta} - d^\top \beta}{\sqrt{\hat{\sigma}^2 d^\top (\mathbf{X}^\top \mathbf{X})^{-1} d}} \sim t_{n-k},$$

wobei t_{n-k} die t-Verteilung mit $n-k$ Freiheitsgraden bezeichnet.

Einschub: Sei $V \sim \mathcal{N}(0,1)$ und $W \sim \chi^2_n$, stochastisch unabhängig. Dann folgt die Zufallsvariable $V/\sqrt{W/n}$ einer t-Verteilung mit n Freiheitsgraden. Für $n \to \infty$ ist die t-Verteilung gleich einer Standardnormalverteilung.
Beweis: Nach den Rechenregeln für Erwartungswerte und Kovarianzmatrizen folgt $d^\top \hat{\beta} \sim \mathcal{N}(d^\top \beta, \sigma^2 d^\top (\mathbf{X}^\top \mathbf{X})^{-1} d)$ und damit

$$\frac{d^\top \hat{\beta} - d^\top \beta}{\sqrt{\sigma^2 d^\top (\mathbf{X}^\top \mathbf{X})^{-1} d}} \sim \mathcal{N}(0,1).$$

Betrachte nun die Zufallsvariable

$$\frac{\frac{d^\top \hat{\beta} - d^\top \beta}{\sqrt{\sigma^2 d^\top (\mathbf{X}^\top \mathbf{X})^{-1} d}} \sim \mathcal{N}(0,1)}{\sqrt{\frac{\frac{n-k}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-k} \text{ (o. Beweis)}}{n-k}}} = \frac{d^\top \hat{\beta} - d^\top \beta}{\sqrt{\hat{\sigma}^2 d^\top (\mathbf{X}^\top \mathbf{X})^{-1} d}} \sim t_{n-k}.$$

Damit lautet die Testentscheidung: Lehne H_0 ab, wenn

$$T = \frac{|d^\top \hat{\beta}|}{\sqrt{\hat{\sigma}^2 d^\top (\mathbf{X}^\top \mathbf{X})^{-1} d}} > t_{n-k,1-\alpha/2}$$

und ein $(1-\alpha) \times 100\%$ Konfidenzintervall für $d^\top \beta$ ist

$$d^\top \hat{\beta} \pm t_{n-k,1-\alpha/2} \sqrt{\hat{\sigma}^2 d^\top (\mathbf{X}^\top \mathbf{X})^{-1} d}.$$

Beispiel 10.4.1 (Körperfettmessung)
Jetzt können wir für jede der Einflussgrößen die Teststatistik ausrechnen, dabei ist d der Einheitsvektor, sodass $d^\top \beta = \beta_j$:

```
> T <- coef(bodyfat_lm) / sdbeta
> T

 (Intercept)          age    waistcirc      hipcirc
  -7.0470856    1.7061408    4.3468577    4.5364912
elbowbreadth  kneebreadth
  -0.2474076    1.9177922
```

und die zweiseitigen P-Werte aus der t-Verteilung ablesen

```
> round((1 - pt(abs(T), df = nrow(bodyfat) - length(T))) * 2, 6)

 (Intercept)          age    waistcirc      hipcirc
    0.000000     0.092756     0.000050     0.000025
elbowbreadth  kneebreadth
    0.805373     0.059533
```

Das gleiche Ergebnis erhält man mit

```
> summary(bodyfat_lm)

Call:
lm(formula = DEXfat ~ age + waistcirc + hipcirc + elbowbreadth +
    kneebreadth, data = bodyfat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1782 -2.4973  0.2089  2.5496 11.6504

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -59.57320    8.45359  -7.047 1.43e-09 ***
age            0.06381    0.03740   1.706   0.0928 .
waistcirc      0.32044    0.07372   4.347 4.96e-05 ***
hipcirc        0.43395    0.09566   4.536 2.53e-05 ***
elbowbreadth  -0.30117    1.21731  -0.247   0.8054
kneebreadth    1.65381    0.86235   1.918   0.0595 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.988 on 65 degrees of freedom
Multiple R-squared:  0.8789,        Adjusted R-squared:  0.8696
F-statistic: 94.34 on 5 and 65 DF,  p-value: < 2.2e-16
```

Und als Abschluss noch die Konfidenzintervalle

```
> confint(bodyfat_lm)
```

```
                    2.5 %      97.5 %
(Intercept)  -76.45619185 -42.6902064
age           -0.01088410   0.1385129
waistcirc      0.17321558   0.4676638
hipcirc        0.24291126   0.6249985
elbowbreadth  -2.73231557   2.1299704
kneebreadth   -0.06842371   3.3760367
```

Wir sehen also, dass hauptsächlich der Bauch- und Hüftumfang informativ für den Körperfettanteil sind.