Stochastik 1 und 2

Wintersemester 2017/18 und Sommersemester 2018

Universität Freiburg

von Peter Pfaffelhuber

Version: 20. August 2018

Inhaltsverzeichnis

Teil I.

Stochastik

1. Grundlegendes

1.1. Vorbemerkung

Die *Stochastik* (oder Wahrscheinlichkeitstheorie) hat zwei Gesichter: Einerseits fasziniert es schon seit Jahrhunderten, intuitiv und analytisch Gewinnchancen etwa bei Glücksspielen einzuschätzen und zu berechnen. Dies verleiht der Stochastik einen angewandten und intuitiven Zugang. Andererseits geht es in der Stochastik als Teilgebiet der Mathematik auch um Mengen, Funktionen und deren Zusammenhänge. Letzteres bietet – wie in jedem Teilgebiet der Mathematik – das formale Grundgerüst.

In dieser Vorlesung werden wir von beiden Aspekten Gebrauch machen, wobei die intuitiven Ansätze zumindest in dieser grundlegenden Vorlesung überwiegen werden. In Nachfolgeveranstaltungen wie etwa der Vorlesung *Wahrscheinlichkeitstheorie*, kann dies anders aussehen. Der Grund hierfür ist, dass Wahrscheinlichkeiten durch Maße beschrieben werden, deren Theorie in der Vorlesung *Analysis III* behandelt wird.

Als zu diesem Skript ergänzende Literatur gibt es einige Bücher zu nennen, etwa:

- K. Bosch. Elementare Einführung in die Wahrscheinlichkeitsrechnung. Vieweg+Teubner Verlag, 2011
- H. O. Georgii. Stochastik: Einführung In Die Wahrscheinlichkeitstheorie Und Statistik. de Gruyter, 2009
- N. Henze. Stochastik für Einsteiger: Eine Einführung in die faszinierende Welt des Zufalls. Springer, 2016
- G. Kersting und A. Wakolbinger. Elementare Stochastik. Birkhäuser, 2008

1.2. Wahrscheinlichkeitsräume und Zufallsvariablen

Wir fangen rein formal an und definieren zunächst, worum es in dieser Vorlesung ausschließlich geht.

Definition 1.1 (Wahrscheinlichkeitsräume)**.** *1. Ein Wahrscheinlichkeitsraum ist ein Tripel $(\Omega, \mathbf{P}, \mathcal{F})$, bestehend aus einer Menge Ω, einer σ-Algebra[1] $\mathcal{F}$ und einem Wahrscheinlichkeitsmaß (oder Wahrscheinlichkeitsverteilung) $\mathbf{P} : \mathcal{F} \to [0,1]$, für das*

$$\begin{aligned} \mathbf{P}(\Omega) &= 1, \\ \mathbf{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) &= \sum_{i=1}^{\infty} \mathbf{P}(A_i) \text{ falls } A_i \cap A_j = \emptyset \text{ für alle } i \neq j. \end{aligned} \tag{1.1}$$

gilt.

[1] Mehr dazu in Analysis III. Eine σ-Algebra – $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ – die Potenzmenge von Ω – ist ein Mengensystem mit (i) $\emptyset \in \mathcal{F}$, (ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$, (iii) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

2. *Ein Wahrscheinlichkeitsraum* $(\Omega, \mathbf{P}, \mathcal{F})$ *heißt* diskret*, falls* Ω *höchstens abzählbar ist und* $\mathcal{F} = \mathcal{P}(\Omega)$*. In diesem Fall sagen wir auch, dass* $(\Omega, \mathbf{P})$ *ein diskreter Wahrscheinlichkeitsraum ist. Die Funktion*

$$p : \begin{cases} \Omega & \to [0,1] \\ \omega & \mapsto \mathbf{P}(\{\omega\}) \end{cases}$$

heißt auch Zähldichte*.*

3. *Das Wahrscheinlichkeitsmaß* $\mathbf{P}$ *hat eine Dichte, falls* $\Omega = \mathbb{R}^d$ *für ein* $d = 1, 2, ...$ *und es ein* $f : \mathbb{R}^d \to \mathbb{R}_+$ *gibt mit*

$$\mathbb{P}(B_1 \times \cdots \times B_d) = \int_{B_1} \cdots \int_{B_d} f(x_1, ..., x_d) dx_d \cdots dx_1$$

für alle Intervalle $B_k = [a_k, b_k]$*, oder* $B_k = (a_k, b_k)$*, oder* $B_k = (a_k, b_k]$*, oder* $B_k = [a_k, b_k)$*,* $k = 1, ..., d$*. Wir nennen* $B_1 \times \cdots \times B_d$ *auch einen* Quader*. Wir sagen dann auch, dass* $\mathbf{P}$ *kontinuierlich ist.*

4. *Die Menge* Ω *heißt* Grundraum *oder* Ergebnisraum*,* $\mathcal{F}$ *heißt auch Ereignisraum.*

5. *Jede Abbildung* $X : \Omega \to S$ *heißt Zufallsvariable mit Zielbereich* S *oder* S*-wertige Zufallsvariable.*[2]

Wir benennen zunächst elementare Eigenschaften von Wahrscheinlichkeitsmaßen.

Lemma 1.2 (Grundlegende Eigenschaften von Wahrscheinlichkeitsmaßen)**.** *Sei* $(\Omega, \mathcal{F}, \mathbf{P})$ *ein Wahrscheinlichkeitsraum. Weiter sei* $A, A_1, A_2, ... \in \mathcal{F}$*.*

1. *Es ist* $\mathbf{P}(\emptyset) = 0, \mathbf{P}(\Omega) = 1$*.*

2. *Es ist* $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$*.*

3. *Ist* $A_1 \subseteq A_2$*, so ist* $\mathbf{P}(A_1) \leq \mathbf{P}(A_2)$*.*

4. *Es gilt*

$$\mathbf{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) \leq \sum_{n=1}^{\infty} \mathbf{P}(A_n).$$

Beweis. 1. Da $\emptyset = \emptyset \uplus \emptyset$ und die rechte Seite disjunkte Mengen sind, gilt

$$\mathbf{P}(\emptyset) = \mathbf{P}(\emptyset \uplus \emptyset) = 2\mathbf{P}(\emptyset).$$

Da $\mathbf{P}$ nur Werte in $[0, 1]$ annimmt, muss $\mathbf{P}(\emptyset) = 0$ gelten.
2. Es ist $\mathbf{P}(A^c) + \mathbf{P}(A) = \mathbf{P}(\Omega) = 1$, da $\Omega = A \uplus A^c$.
3. Es gilt $A_2 = A_1 \uplus (A_2 \setminus A_1)$. Also ist

$$\mathbf{P}(A_2) = \mathbf{P}(A_1) + \mathbf{P}(A_2 \setminus A_1) \geq \mathbf{P}(A_1).$$

[2] Hier sind wir etwas schlampig: eigentlich sind nur *messbare* Abbildungen Zufallsvariable. (Das bedeutet: Ist $\mathcal{F}' \subseteq \mathcal{P}(S)$ eine weitere σ-Algebra, so ist $X^{-1}(B) \in \mathcal{F}$ für alle $B \in \mathcal{F}'$.) Da es aber sehr schwierig ist, sich nicht-messbare Abbildungen auszudenken, verzichten wir auf diese Feinheit. Insbesondere ist für höchstens abzählbares Ω jede Abbildung messbar. Mehr dazu erfahren Sie in Analysis III und in der Wahrscheinlichkeitstheorie.

4. Wir schreiben

$$\mathbf{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \mathbf{P}\Big(\biguplus_{n=1}^{\infty} A_n \setminus \big(A_1 \cup \cdots \cup A_{n-1}\big)\Big) = \sum_{n=1}^{\infty} \mathbf{P}\big(A_n \setminus \big(A_1 \cup \cdots \cup A_{n-1}\big)\big)$$
$$\leq \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

nach 2. □

Bemerkung 1.3 (Monotonie von Wahrscheinlichkeitsmaßen)**.** Es gibt weitere Eigenschaften von Wahrscheinlichkeitsmaßen, auf die wir hier ohne maßtheoretische Grundlagen nicht weiter eingehen können. Etwa gilt für $A_1 \subseteq A_2 \subseteq \cdots$

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big)$$

und für $A_1, A_2, \ldots$ mit $\bigcap_{n=1}^{\infty} A_n = \emptyset$ ist

$$\lim_{n\to\infty} \mathbf{P}(A_n) = 0.$$

In der Vorlesung Stochastik werden wir uns ausschließlich mit diskreten Wahrscheinlichkeitsräumen und mit Wahrscheinlichkeitsmaßen mit Dichten beschäftigen. In der Wahrscheinlichkeitstheorie werden dann Wahrscheinlichkeitsräume in ihrer vollen Allgemeinheit behandelt.
Wir beginnen mir einem grundlegenden Fall diskreter Wahrscheinlichkeitsräume. Das nächste Resultat besagt, dass es bei diskreten Wahrscheinlichkeitsräumen genügt, das Maß auf *Elementarereignissen* $\{\omega\}$ für alle $\omega \in \Omega$ zu kennen.

Lemma 1.4 (Zähldichte)**.** *Sei Ω höchstens abzählbar. Eine Funktion $p : \Omega \to [0, 1]$ ist genau dann eine Zähldichte, wenn*

$$\sum_{\omega\in\Omega} p(\omega) = 1.$$

In diesem Fall ist

$$\mathbf{P} : A \mapsto \sum_{\omega\in A} p(\omega)$$

das einzige Wahrscheinlichkeitsmaß mit Zähldichte p.

Beweis. Jede Teilmenge $A \subseteq \Omega$ ist höchstens abzählbar und $A = \bigcup_{\omega\in A}\{\omega\}$ eine Zerlegung von A in disjunkte Teilmengen. Das Resultat folgt nun aus den Eigenschaften für Wahrscheinlichkeitsmaße. □

Wir führen nun noch den wichtigen Begriff der Verteilung einer Zufallsvariablen ein.

Definition 1.5. *Sei $(\Omega, \mathcal{F}, \mathbf{P})$ ein Wahrscheinlichkeitsraum und X eine S-wertige Zufallsvariable (mit σ-Algebra $\mathcal{F}'$). Dann heißt die Abbildung*

$$X_*\mathbf{P} : \begin{cases} \mathcal{F}' & \to [0,1] \\ A & \mapsto \mathbf{P}(\{\omega : X(\omega) \in A\}) \end{cases}$$

Bildmaß von X unter $\mathbf{P}$. Äquivalent schreiben wir

$$\mathbf{P}(X \in A) := \mathbf{P}(\{X \in A\}) := \mathbf{P}(X^{-1}(A)) := \mathbf{P}(\{\omega : X(\omega) \in A\}) = X_*\mathbf{P}(A).$$

Ist S höchstens abzählbar, so ist auch

$$\mathbf{P}(X \in A) = \sum_{s \in A} \mathbf{P}(X = s).$$

Ist $S \subseteq \mathbb{R}^d$ und besitzt $X_*\mathbf{P}$ eine Dichte f_X, so gilt

$$\mathbf{P}(X \in A) = \int_A f_X(x)dx$$

für jeden Quader A.

Beispiel 1.6 (Einfache Beispiele)**.**

1. Die einfachste Zufallsvariable ist $X = \mathrm{id}$. Hier ist $X_*\mathbf{P} = \mathbf{P}$, denn
$$\mathbf{P}(X \in A) = \mathbf{P}(X^{-1}(A) = \mathbf{P}(A).$$
2. Auch sehr einfach sind konstante Zufallsvariable. Sei also $X = c$ für ein $c \in S$. Dann ist
$$\mathbf{P}(X \in A) = \mathbf{P}(c \in A) = 1_{c \in A},$$
also $X_*\mathbf{P} = 1_c(.)$.

1.3. Häufige Modelle: Münzwurf, Würfeln, Urnen

Wir behandeln zunächst einige häufige Beispiele.

Beispiel 1.7 (Münzwurf)**.** Wenn man eine Münze wirft, zeigt sie bekanntlich *Kopf* oder *Zahl*. Bei einem p-Münzwurf ist die Wahrscheinlichkeit für *Kopf* gerade p. Damit ist die Wahrscheinlichkeit für *Zahl* gerade $q := 1 - p$. Bei üblichen Münzen denkt man oft an den Fall $p = 1/2$, weil Kopf und Zahl mit derselben Wahrscheinlichkeit oben zu liegen kommen. Wirft man jedoch eine Reißzwecke (was eine sehr spezielle Münze ist), die entweder mit der spitzen Seite oben oder unten zu liegen kommt, ist klar, dass auch $p \neq 1/2$ sein kann. Wirft man zweimal mit derselben Münze, kann man nach der Wahrscheinlichkeit fragen, zweimal *Kopf* zu werfen. Dies ist p^2.

Um das Beispiel eines n-fach ausgeführten p-Münzwurfes in den Rahmen eines Wahrscheinlichkeitsraumes zu bringen, sei $\Omega = \{0,1\}^n$, wobei $0 \equiv$ Zahl und $1 \equiv$ Kopf. Für das Wahrscheinlichkeitsmaß geben wir die Zähldichte an, nämlich

$$\mathbf{P}((\omega_1, ..., \omega_n)) = \prod_{i=1}^{n} p^{\omega_i}(1-p)^{1-\omega_i} = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}. \tag{1.2}$$

Als Beispiel einer Zufallsvariable nehmen wir

$$X_i(\omega_1, ..., \omega_n) = \omega_i, \qquad X = X_1 + \cdots + X_n.$$

Somit ist X_i das Ergebnis des i-ten Münzwurfes und X ist die Gesamtzahl an Münzwürfen mit Ausgang *Kopf*. Für die Verteilung von X_i und X gilt

$$\begin{aligned}\mathbf{P}(X_i = 1) &= 1 - \mathbf{P}(X_i = 0) = p,\\ \mathbf{P}(X = k) &= \mathbf{P}((\omega_1, ..., \omega_n) : \omega_1 + \cdots + \omega_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.\end{aligned}$$

Die letzte Gleichheit erklärt man dabei wie folgt: Es gibt insgesamt $\binom{n}{k}$ Möglichkeiten, k Ausgänge *Kopf* auf die n Versuche zu verteilen – später mehr zu dieser Anzahl. Jede der Möglichkeiten, bei denen k Köpfe fallen, hat Wahrscheinlichkeit $p^k(1-p)^{n-k}$.
Wir werden ab Abschnitt 4.1 sagen, X habe eine Binomialverteilung mit Parametern n und p.

Beispiel 1.8 (Würfeln)**.** Eine bekannte Herausforderung ist es (etwa bei *Mensch ärgere Dich nicht*), beim Würfeln mit einem Würfel möglichst schnell eine 6 zu werfen. Dies ist ganz ähnlich wie beim (1/6)-Münzwurf, schließlich gibt es nur zwei Möglichkeiten *6* oder *nicht 6*. Damit ist die Wahrscheinlichkeit, k mal keine 6 zu werfen, gerade $(5/6)^k$. Wir berechnen noch die Wahrscheinlichkeit, dass die erste 6 genau im n-ten Wurf kommt. Da wir hierzu unbeschränkt oft den Würfel werfen müssen, sei $\Omega = \{1, ..., 6\}^{\mathbb{N}}$. Weiter sei $X_i \in \{1, ..., 6\}$ das Ergebnis des i-ten Wurfes und

$$T = \min\{n : X_n = 6\}$$

die Nummer des Versuches, bei dem zum ersten mal eine 6 fällt. Die Verteilung von T ist

$$\mathbf{P}(T = n) = \left(\frac{5}{6}\right)^{n-1} \frac{1}{6},$$

denn die einzige Möglichkeit, wie $T = n$ zustande kommt, ist es, zunächst $(n-1)$-mal keine Sechs zu werfen und anschließend eine 6.

Beispiel 1.9 (Urne)**.** In einer Urne befinden sich N Kugeln, und zwar K_1 Kugeln der Farbe 1, ..., K_n Kugeln der Farbe n. (Also ist $K_1 + \cdots + K_n = N$.) Zieht man (mit verschlossenen Augen) aus der Urne, zieht man eine Kugel der Farbe i mit Wahrscheinlichkeit K_i/N. Zieht man anschließend nochmal aus der Urne (wobei die erste Kugel nicht zurückgelegt wurde), ist die Wahrscheinlichkeit nochmal eine Kugel der Farbe i zu ziehen $(K_i - 1)/(N - 1)$. Damit ist die Wahrscheinlichkeit, zwei Kugeln derselben Farbe zu ziehen, gerade

$$\sum_{i=1}^{n} \frac{K_i}{N} \frac{K_i - 1}{N - 1}.$$

Um dies zu formalisieren, sei $\Omega = \{1, ..., n\}^2$ und $X_i \in \{1, ..., n\}$ sei die Farbe des i-ten Zuges, $i = 1, 2$. Wir schreiben damit

$$\mathbf{P}(X_1 = X_2) = \sum_{i=1}^{n} \mathbf{P}(X_1 = X_2 = i) = \sum_{i=1}^{n} \frac{K_i}{N} \frac{K_i - 1}{N - 1}.$$

Im ersten Gleichheitszeichen haben wir verwendet, dass die Ereignisse $\{X_1 = X_2 = i\}, i = 1, ..., n$ disjunkt sind.

Beispiel 1.10 (Geburtstagsproblem)**.** In einem Raum befinden sich 23 Personen. Wie groß ist die Wahrscheinlichkeit, dass es mindestens zwei Personen gibt, die am selben Tag Geburtstag haben?

Um diese Wahrscheinlichkeit zu berechnen, ist es hilfreich, das Gegenteil *Alle Personen haben an unterschiedlichen Tagen Geburtstag* zu betrachten. Wir stellen die Personen (in Gedanken) in einer Reihe auf. Die Wahrscheinlichkeit dass die zweite Person an einem anderen Tag als die erste Person Geburtstag hat ist 364/365. (Von Schaltjahren und dem 29.2. sehen wir einmal ab.) Weiter ist die Wahrscheinlichkeit, dass die dritte Person an einem anderen Tag als die Personen eins und zwei Geburtstag hat, dann 363/365. Überlegen wir das weiter, so ist die Wahrscheinlichkeit, dass alle Personen an unterschiedlichen Tagen Geburtstag haben, gerade

$$\frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-22}{365} \approx 0.493.$$

Damit ist die Wahrscheinlichkeit, dass es zwei Personen gibt, die am gleichen Tag Geburtstag haben, etwa $1 - 0.493 = 0.507$.

2. Kombinatorik und uniform verteilte Zufallsvariablen

Der Grund, warum in der Stochastik zunächst das geschickte Abzählen (also Kombinatorik) so wichtig ist, sind uniforme Verteilungen oder Laplace-Experimente.

2.1. Laplace-Experimente

Unter einem Laplace-Experiment versteht man die Situation, wenn alle möglichen Ausgänge dieselbe Wahrscheinlichkeit besitzen. Beispiele sind der $\frac{1}{2}$-Münzwurf oder das Würfeln mit einem fairen Würfel. Die zugehörigen Wahrscheinlichkeitsmaße heißen uniforme Verteilungen.

Definition 2.1 (Uniforme Verteilung)**.** *Sei X eine S-wertige Zufallsvariable.*

1. *Ist S endlich und gilt $\mathbf{P}(X = a) = \frac{1}{|S|}$ für alle $a \in S$, so nennen wir X (diskret) uniform verteilt auf S. Wir schreiben dann auch $X \sim U(S)$.*
2. *Ist $S = \mathbb{R}^d$ und hat X die Dichte $f_X(x) = 1_S(x)\frac{1}{|S|}$ (wobei $|S|$ das Volumen von S im $\mathbb{R}^d$ ist), so nennen wir X (stetig) uniform verteilt auf S. Wir schreiben dann ebenfalls $X \sim U(S)$.*

Bemerkung 2.2. In beiden Fällen ist $\mathbf{P}(X \in A) = \frac{|A|}{|S|}$.

Beispiel 2.3 (Kollision von Kennzeichen)**.** In Verallgemeinerung zu Beispiel 1.10 werde eine Menge von n Individuen mit $r \geq n$ verschiedenen Kennzeichen gekennzeichnet. Wir denken dabei an Kennzeichen wie r verschiedene Namen, Geburtstage, PINs etc. Wir gehen davon aus, dass alle Individuen jedes Kennzeichen mit gleicher Wahrscheinlichkeit erhalten. Wie groß ist die Wahrscheinlichkeit, dass keine zwei Individuen gleich gekennzeichnet sind?

Hierzu sei $X = (X_1, ..., X_n)$ mit Werten in $S = \{1, ..., r\}^{\{1,...,n\}}$, so dass $1 \leq X_i \leq r$ das Kennzeichen des i-ten Individuums benennt. Wir suchen nun die Wahrscheinlichkeit von

$$\mathbf{P}(X_i \neq X_j \text{ für alle } i \neq j\},$$

oder äquivalent

$$\mathbf{P}(X \in A) \text{ für } A = \{(x_1, ..., x_n) : x_i \neq x_j \text{ für alle } i \neq j\}.$$

Um diese zu bestimmen, müssen wir $|A|$ berechnen. Da das erste Individuum r Kennzeichen zur Auswahl hat, das zweite Individuum (falls A zutrifft) dann noch $r - 1$ usw, ist

$$|A| = r(r-1)\cdots(r-n+1).$$

Damit ist

$$\mathbf{P}(X \in A) = \frac{r(r-1)\cdots(r-n+1)}{r^n} = \prod_{i=1}^{n-1}\left(1 - \frac{i}{r}\right)$$

Wir versuchen nun noch, diese Wahrscheinlichkeit für große n abzuschätzen. Verwenden wir $1 - t \leq e^{-t}$, so folgt

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

Da

$$A^c = \bigcup_{i<j} B_{ij}, \qquad B_{ij} := \{(a_1, ..., a_n) : a_i = a_j\}$$

und

$$\mathbf{P}(X \in B_{ij}) = \frac{1}{r},$$

und da[1] $\{(i,j) \in [1:n]^2 : i < j\} = \frac{n(n-1)}{2}$, folgt, dass

$$\mathbf{P}(X \in A) = 1 - \mathbf{P}(X \in A^c) \geq 1 - \sum_{i<j} \mathbf{P}(X \in B_{ij}) = 1 - \frac{n(n-1)}{2r}$$

Es gilt also

$$1 - \frac{n(n-1)}{2r} \leq \mathbf{P}(X \in A) \leq \exp\Big(-\frac{n(n-1)}{2r}\Big).$$

Für große r muss also n mindestens so groß wie $\sqrt{r}$ sein, damit es merklich zu Kollisionen kommt. (In Beispiel 1.10 war $r = 365$ und $n = 23$, also $n^2 = 529$.)

2.2. Zufällige Permutationen

Aus der linearen Algebra sind Permutationen bekannt. Wir beschäftigen uns nun mit uniform verteilten (also zufälligen) Permutationen. Wir erinnern zunächst an die Bedeutung und Schreibweise von Permutationen.

Bemerkung 2.4 (Permutation)**.** Eine Permutation σ der Menge $[1:n]$ ist eine bijektive Abbildung $\sigma : [1:n] \to [1:n]$. Um σ zu notieren, gibt es zwei verschiedene Möglichkeiten. Entweder man schreibt in zwei Zeilen jeweils i über $\sigma(i)$, was dann für $n = 7$ etwa so aussieht:

$$\sigma = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7 \\ 5\,2\,7\,3\,1\,4\,6 \end{pmatrix}, \tag{2.1}$$

oder man verwendet Zykel von σ. Hierbei ist eine Zykel von σ der Länge j eine Folge $k_1, ..., k_j$ mit $\sigma(k_1) = k_2, ..., \sigma(k_{j-1}) = k_j, \sigma(k_j) = k_1$. In obigem Beispiel ergibt sich

$$\sigma = (15)(2)(3764).$$

Hierbei werden in Klammern jeweils Zykel angegeben. Man beachte, dass die Zykel-Schreibweise nicht eindeutig ist, etwa gilt auch

$$\sigma = (51)(2)(6437).$$

[1] Wir schreiben $[1:n] := \{1, ..., n\}$.

Bekanntermaßen gibt es $n! = 1 \cdots n$ Permutationen von $[1 : n]$. Dies überlegt man sich am besten an der Schreibweise (2.1). Für $\sigma(1)$ gibt es nämlich genau n Möglichkeiten. Anschließend gibt es für $\sigma(2)$ noch $n-1$ Möglichkeiten usw. Zählt man also alle Permutationen durch, so kommt man auf $|[1 : n]| = n!$. Wir bezeichnen die Menge aller Permutationen von $[1 : n]$ mit $\mathbb{S}(n)$.

Beispiel 2.5 (Länge eines Zykel)**.** Sei Σ eine zufällige Permutation von $[1 : n]$. (Das bedeutet: Auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbf{P})$ ist eine Zufallsvariable $\Sigma : \Omega \to \mathbb{S}(n)$ definiert, so dass die Verteilung von Σ uniform auf allen Permutationen ist.) Weiter sei

$$h(\sigma) := \text{Länge des Zykel, der 1 enthält.}$$

Wir wollen nun zeigen, dass

$$\mathbf{P}(h(\Sigma) = b) = \frac{1}{n}.$$

Hierzu setzen wir

$$A := \{\sigma \in \mathbb{S}(n) : h(\sigma) = b\} = \{\sigma \in \mathbb{S}(n) : \sigma(1) \neq 1, \sigma^2(1) \neq 1, ..., \sigma^{b-1}(1) \neq 1, \sigma^b(1) = 1\}.$$

Es gilt (mit derselben Überlegung, mit der wir $\mathbb{S}(n)$ durchgezählt haben)

$$|A| = (n-1)(n-2)\cdots(n-b+1) \cdot 1 \cdot (n-b)\cdots 1,$$

und damit

$$\mathbf{P}(h(\Sigma) = b) = \frac{|A|}{|S|} = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

2.3. Zufällige Teilmengen

Während Permutationen Bijektionen beschreiben, beschäftigen wir uns nun mit zufälligen Teilmengen einer gegebenen Menge $[1 : n]$.

Bemerkung 2.6. Wir bezeichnen mit

$$S_k = \{A : A \subseteq [1 : n], |A| = k\}$$

die Menge der k-elementigen Teilmengen von $[1 : n]$. Sei X uniform auf S_k verteilt. Um nun die Verteilung von X zu bestimmen, benötigen wir $|S_k|$. Wir zeigen

$$|S_k| = \binom{n}{k} := \frac{n!}{k!(n-k)!} \qquad \text{oder} \qquad \mathbf{P}(X = \{1, ..., k\}) = \frac{1}{\binom{n}{k}}.$$

Klar ist, dass aus der zweiten Aussage die erste folgt. Um $\{X = \{1, ..., k\}\}$ zu erhalten, müssen wir als erstes Element eines in $[1 : k]$ wählen. Bei der zweiten Ziehung müssen wir eines der übrigen $k-1$ aus $[1 : k]$ ziehen usw. Daraus folgt

$$\mathbf{P}(X = \{1, ..., k\}) = \frac{k}{n}\frac{k-1}{n-1}\cdots\frac{1}{n-k+1} = \frac{k!(n-k)!}{n!}.$$

Eine alternative Erklärung ist die Folgende: Um ein Element aus $|S_k|$ zu erhalten, wählen wir zunächst ein Element aus n aus, dann eines aus den restlichen $n-1$ usw. Die Einträge eines so erhaltenen k-Tupels bildet eine k-elementige Teilmenge von $[1 : n]$. Allerdings gibt es $k!$ verschiedene Reihenfolgen dieser Elemente, so dass $|S_k| = \binom{n}{k}$ folgt.

2.4. Zufällige Besetzungen

Bemerkung 2.7. Unter einer Besetzung verstehen wir eine Belegung von r Plätzen (Urnen) mit n Objekten (Kugeln). Formal setzen wir

$$S_{n,r} = \{k = (k_1, ..., k_r) : k_j \in \mathbb{N}_0, k_1 + \cdots + k_r = n\}$$

und betrachten nun wieder ein rein zufälliges Element X aus $S_{n,r}$. Wir zeigen nun

$$\mathbf{P}(X = (k_1, ..., k_r)) = \frac{1}{\binom{n+r-1}{n}}$$

für jedes $(k_1, ..., k_r) \in S_{n,r}$.

Um dies einzusehen, verwenden wir am besten eine Hilfskonstruktion: Sei

$$S' = \{\text{01-Folgen der Länge } n + r - 1 \text{ mit genau } n \text{ Einsen}\}$$

$$h : \begin{cases} S_{n,r} & \to S' \\ (k_1, ..., k_r) & \mapsto \underbrace{1 \dots 1}_{k_1-\text{mal}} 0 \underbrace{1 \dots 1}_{k_2-\text{mal}} 0 \dots \underbrace{1 \dots 1}_{k_r-\text{mal}} \end{cases}$$

also z.B. $h(2, 0, 3, 0) = 11001110$. Es ist klar, dass h eine Bijektion ist. Da S' alle n-elementigen Teilmengen von $[1 : (n + r - 1)]$ beschreibt, ist

$$|S_{n,r}| = |S'| = \binom{n + r - 1}{n}.$$

2.5. Kontinuierlich uniform verteilte Zufallsvariable

Wir erinnern zunächst an Definitionen 1.1.3 und 2.1. Danach ist eine Zufallsvariable X genau dann uniform auf $[a, b] \subseteq \mathbb{R}$ verteilt, wenn

$$\mathbf{P}(X \in [w, x]) = \frac{x - w}{b - a}$$

für alle $a \le w \le x \le b$. Dies ist offenbar genau dann der Fall, wenn für $0 \le w \le x \le b$

$$\mathbf{P}(X \le x) = \frac{x - a}{b - a}.$$

(Denn: Gilt dies, so ist $\mathbf{P}(X \in [w, x]) = \mathbf{P}(X \le x) - \mathbf{P}(X \le w)$.)

Für das nächste Resultat benutzen wir folgende Schreibweise für $x \in \mathbb{R}$: Der ganzzahlige Anteil von x ist $\lfloor x \rfloor$, und $\langle x \rangle = x - \lfloor x \rfloor$ ist der gebrochene Anteil von x.

Lemma 2.8 (Verschiebungsinvarianz der uniformen Verteilung)**.** *Sei X uniform auf $[0, r]$ verteilt für $r \in \mathbb{N}$ und $v \in \mathbb{R}$. Dann ist $\langle X + v \rangle \sim U([0, 1])$.*

Beweis. OBdA ist $r = 1$, der allgemeine Fall folgt wegen $\langle X \rangle \sim U([0, 1])$ und $\langle X + v \rangle = \langle \langle X \rangle + v \rangle$. OBdA ist $0 \le v \le 1$, ansonsten ersetzen wir v mit $\langle v \rangle$. Wir berechnen für $0 \le x \le 1$

$$\begin{aligned} \mathbf{P}(\langle X + v \rangle \le x) &= \mathbf{P}(X + v \le x) + \mathbf{P}(1 \le X + v \le 1 + x) \\ &= \mathbf{P}(X \le x - v) + \mathbf{P}(1 - v \le X \le 1 + x - v) = \begin{cases} x - v + v, & x \ge v, \\ 0 + x, & x \le v \end{cases} = x. \end{aligned}$$

Damit ist $\langle X + v \rangle \sim U([0, 1])$. □

Wir kommen nun zu einer interessanten Beobachtung: die führende Ziffer einer auf einem großen Intervall gezogenen uniformen Zufallsvariable ist mit größter Wahrscheinlichkeit eine 1:

Proposition 2.9 (Benford's Gesetz)**.** *Sei X eine Zufallsvariable, so dass $\langle \log_{10} X \rangle \sim U([0,1])$. Weiter sei*

$$h : \begin{cases} \mathbb{R}_+ & \to \{1, ..., 9\} \\ x & \mapsto b \text{ falls } \log_{10} b \leq \langle \log_{10} x \rangle < \log_{10}(b+1). \end{cases}$$

(Mit anderen Worten ist $h(x)$ die führende Ziffer der Dezimaldarstellung von x.) Dann ist

$$\mathbf{P}(h(X) = b) = \log_{10}\left(1 + \frac{1}{b}\right).$$

Bemerkung 2.10. Erstaunlich viele Verteilungen für X sind so, dass $\langle \log_{10} X \rangle \sim U([0,1])$ zumindest approximativ gilt. Um ein einfaches Beispiel zu nennen, sei $Y \sim U([0,r])$ für ein $r \in \mathbb{N}$ und $X = 10^Y$. Dann ist

$$\mathbf{P}(\langle \log_{10} X \rangle \leq x) = \mathbf{P}(\langle Y \rangle \leq x) = x$$

nach Lemma 2.8.

Beweis. Wir schreiben direkt

$$\mathbf{P}(h(X) = b) = \mathbf{P}(\langle \log_{10} X \rangle \in [\log_{10} b, \log_{10}(b+1))) = \log_{10}(b+1) - \log_{10}(b) = \log_{10}\left(1 + \frac{1}{b}\right).$$

□

3. Verteilungen und deren Eigenschaften

Der Begriff des Wahrscheinlichkeitsmaßes oder der Wahrscheinlichkeitsverteilung (siehe Definition 1.1) ist für unsere Zwecke zentral. In diesem Kapitel beschäftigen wir uns meist mit reellwertigen Zufallsvariablen und deren Bildmaßen. Wir gehen dabei nicht weiter auf das Wahrscheinlichkeitsmaß auf Ω ein, sondern interessieren uns ausschließlich für das Bildmaß.

3.1. Die Verteilungsfunktion

Wahrscheinlichkeitsverteilungen sind Abbildungen, die eine σ-Algebra als Definitionsmenge haben. Um solche Verteilungen veranschaulichen zu können, benötigen wir eine andere Darstellung, die wir nun mit den Verteilungsfunktionen kennen lernen werden.

Definition 3.1 (Verteilungsfunktion)**.** *Sei X eine S-wertige Zufallsvariable und $S \subseteq \mathbb{R}$. Die Funktion*

$$F_X : \begin{cases} \mathbb{R} & \to [0,1] \\ x & \mapsto \mathbf{P}(X \leq x) \end{cases}$$

heißt Verteilungsfunktion (des Bildmaßes) von X.

Lemma 3.2 (Verteilungsfunktion von diskreten und kontinuierlichen Zufallsvariablen)**.** *Sei X eine S-wertige Zufallsvariable.*

1. *Ist S diskret, so ist die Verteilungsfunktion*

$$F_X(x) = \sum_{\substack{y \leq x \\ y \in S}} \mathbf{P}(X = y).$$

2. *Besitzt X die Dichte f_X, so ist die Verteilungsfunktion*

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx.$$

   Ist f_X stückweise stetig, so ist $F_X' = f_X$ an allen Stetigkeitsstellen von f_X.

In beiden Fällen wird die Verteilung von X eindeutig durch F_X beschrieben.

Beweis. Die Aussagen 1. und 2. erhält man durch direktes Einsetzen von Zähldichte und Dichte in die Definition für die Verteilungsfunktion. Für die Eindeutigkeit der Verteilungsfunktion stellen wir im Falle einer diskreten Zufallsvariablen X und $x \in S$ fest, dass

$$\begin{aligned} \mathbf{P}(X = x) &= \lim_{\varepsilon \to 0} \mathbf{P}(x \leq X \leq x + \varepsilon) = \lim_{\varepsilon \to 0} \mathbf{P}(X \leq x + \varepsilon) - \mathbf{P}(X \leq x) \\ &= \lim_{\varepsilon \to 0} F_X(x + \varepsilon) - F_X(x). \end{aligned}$$

Für kontinuierliche Zufallsvariablen schreiben wir

$$\mathbf{P}(X \in (a,b]) = \mathbf{P}(X \leq b) - \mathbf{P}(X \leq a) = F_X(b) - F_X(a).$$

Also bestimmt F_X die Werte von $X_*\mathbf{P}$ auf allen Quadern und damit nach Definition 1.1 $X_*\mathbf{P}$ selbst. □

Lemma 3.3 (Eigenschaften von Verteilungsfunktionen)**.** *Sei F_X die Verteilungsfunktion einer Zufallsvariable X. Dann gilt:*

1. *F_X ist monoton wachsend mit*

$$F_X(x) \xrightarrow{x\to-\infty} 0, \qquad F_X(x) \xrightarrow{x\to\infty} 1.$$

2. *F_X ist rechtsstetig und hat linksseitige Grenzwerte. Hat X eine Dichte, dann ist F_X sogar stetig.*

Beweis. 1. Wegen Lemma 1.2.3 gilt für $x \leq y$

$$F_X(x) = \mathbf{P}(X \leq x) \leq \mathbf{P}(X \leq y) = F_X(y)$$

und damit ist F_X monoton wachsend.
2. Der Fall mit Dichte ist klar wegen des Hauptsatzes der Differential- und Integralrechnung. Im Allgemeinen folgt mit Bemerkung 1.3

$$\lim_{\varepsilon\to 0} F_X(x+\varepsilon) = \lim_{\varepsilon\to 0} \mathbf{P}(X \leq x+\varepsilon) = \mathbf{P}(X \leq x) + \lim_{\varepsilon\to 0} \mathbf{P}(x < X \leq x+\varepsilon) = \mathbf{P}(X \leq x),$$
$$\lim_{\varepsilon\to 0} F_X(x-\varepsilon) = \lim_{\varepsilon\to 0} \mathbf{P}(X \leq x-\varepsilon) = \mathbf{P}(X < x).$$

und 2. ist ebenfalls gezeigt. □

Beispiel 3.4 (Verteilungsfunktion der kontinuierlich uniformen Verteilung)**.** Sei $X \in U([a,b])$. Dann ist die Verteilungsfunktion von X

$$F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & x \geq b \end{cases}$$

Denn: Für $x \leq a$ und $x \geq b$ ist die Aussage klar, da $\mathbf{P}(X \leq a) = \mathbf{P}(X > b) = 0$. Die Zufallsvariable X hat nach Definition 2.1 die Dichte $f_X(x) = 1_{[a,b]}(x)\frac{1}{b-a}$. Für $a \leq x \leq b$ ist deshalb

$$\mathbf{P}(X \leq x) = \int_{-\infty}^{x} 1_{[a,b]}(y)\frac{1}{b-a}dy = \int_a^x \frac{1}{b-a} = \frac{x-a}{b-a}.$$

3.2. Bilder von Zufallsvariablen

Da eine S-wertige Zufallsvariable eine Abbildung $X : \Omega \to S$ ist, kann man ebenso Funktionen von Zufallsvariablen betrachten.

Definition 3.5 (Bild einer Zufallsvariablen)**.** *Sei X eine S-wertige Zufallsvariable und $h : S \to S'$ (messbar). Dann heißt $Y := h \circ X$ das Bild von X unter h und ist eine S'-wertige Zufallsvariable.*
Die Verteilung von $Y := h \circ X$ heißt durch h transformierte Verteilung von X. Ist diese Verteilung identisch mit der einer Zufallsvariablen Z, so schreiben wir $Y \sim Z$.

Wir haben bereits Bilder von Zufallsvariablen betrachtet und wiederholen dies kurz.

Beispiel 3.6. 1. Sei Σ eine uniform auf $\mathbb{S}(n)$ verteilte Zufallsvariable (genau wie in Beispiel 2.5), sowie

$$h : \begin{cases} \mathbb{S}(n) & \to \mathbb{N} \\ \sigma & \mapsto \min\{k : \sigma^k(1) = 1\} \end{cases}$$

die Länge des Zykels, der 1 enthält. Dann besagt Beispiel 2.5, dass die Zufallsvariable $Y := h(\Sigma)$ uniform auf $[1 : n]$ verteilt ist.

2. Genau wie in Lemma 2.8 sei $X \sim U([0, r])$ und $h : [0, r] \to [0, 1]$ durch $x \mapsto \langle x + v \rangle$ gegeben. Dann ist $Y := h \circ X$ nach $U([0, 1])$ verteilt, also $\langle X + v \rangle \sim U([0, 1])$.

3. Für eine $\mathbb{R}_+$-wertige Zufallsvariable wird in Kapitel 5.3 das Bild von X unter der Abbildung $h_s : x \mapsto s^x$ für $s \in [0, 1]$, also die Zufallsvariable s^X, eine Rolle spielen.

Lemma 3.7 (Transformation von Verteilungen)**.** *Sei X eine $S \subseteq \mathbb{R}$-wertige Zufallsvariable und $h : S \to \mathbb{R}$ monoton wachsend, sowie $Y := h \circ X$. Dann gilt für $x \in \mathbb{R}$*

$$F_X(x) = F_Y(h(x)).$$

Ist insbesondere h streng monoton und differenzierbar mit differenzierbarer Umkehrabbildung, und hat X eine Dichte f_X, so hat Y die Dichte $x \mapsto f_Y(x) := f_X(h^{-1}(x))|(h^{-1})'|(x)$.

Beweis. Wir schreiben direkt

$$F_X(x) = \mathbf{P}(X \le x) = \mathbf{P}(h(X) \le h(x)) = F_Y(h(x)).$$

Damit gilt wegen $F_X(h^{-1}(x)) = F_Y(x)$

$$f_Y(x) = \frac{d}{dx} F_Y(x) = \frac{d}{dx} F_X(h^{-1}(x)) = f_X(h^{-1}(x))(h^{-1})'(x).$$

□

Beispiel 3.8. Sei $U \sim U([0, 1])$ und $h(u) = -\log u$. Dann hat $Y := h \circ X$ die Verteilung mit Dichte $f(x) = 1_{x \ge 0} e^{-x}$.
Denn: Es ist $h^{-1}(x) = e^{-x}$ und damit

$$f_Y(x) = 1_{e^{-x} \in [0,1]} e^{-x} = 1_{x \ge 0} e^{-x}.$$

Lemma 3.9 (Simulationslemma)**.** *Sei X eine Zufallsvariable mit Verteilungsfunktion F und F^{-1} die verallgemeinerte Inverse, gegeben durch*

$$F^{-1}(u) := \sup\{y : F(y) \le u\}.$$

Ist $U \sim U([0, 1])$, so ist $F^{-1}(U) \sim X$.

Beweis. Es gilt (wegen $F^{-1}(u) < x \iff u < F(x)$)

$$\mathbf{P}(F^{-1}(U) < x) = \mathbf{P}(U < F(x)) = F(x),$$

und die Behauptung ist gezeigt. □

3.3. Die Ein-Ausschlussformel

Für Wahrscheinlichkeitsmaße wissen wir, dass sie additiv sind, d.h. (1.1) gilt. Dies verallgemeinern wir nun auf nicht notwendigerweise disjunkte Mengen.

Proposition 3.10 (Ein- Ausschlussformel)**.** *Für eine S-wertige Zufallsvariable X und (messbare) Mengen $A_1, ..., A_n \subseteq S$ gilt*

$$\begin{aligned}&\mathbf{P}(X \in A_1 \cup \cdots \cup A_n)\\ &= \sum_{i=1}^{n} \mathbf{P}(X \in A_i) - \sum_{1\le i<j\le n} \mathbf{P}(X \in A_i \cap A_j) + \cdots \pm \mathbf{P}(X \in A_1 \cap \cdots \cap A_n).\end{aligned}$$

Beweis. Wir zeigen die Behauptung mittels Induktion nach n. Für $n = 1$ ist die Behauptung klar, für $n = 2$ ist

$$\begin{aligned}\mathbf{P}(X \in A_1 \cup A_2) + \mathbf{P}(X \in A_1 \cap A_2) &= \mathbf{P}(X \in A_1) + \mathbf{P}(X \in A_2 \setminus A_1) + \mathbf{P}(X \in A_1 \cap A_2)\\ &= \mathbf{P}(X \in A_1) + \mathbf{P}(X \in A_2).\end{aligned}$$

Gilt die Behauptung für ein n, so schreiben wir

$$A_1 \cup \cdots \cup A_{n+1} = A_1 \cup \cdots \cup (A_n \cup A_{n+1}).$$

Damit ist (mit $A_i \cap (A_n \cup A_{n+1}) = (A_i \cap A_n) \cup (A_i \cap A_{n+1})$)

$$\begin{aligned}\mathbf{P}(X \in A_1 \cup \cdots \cup A_{n+1}) &= \sum_{i=1}^{n-1} \mathbf{P}(X \in A_i) + \mathbf{P}(X \in A_n \cup A_{n+1})\\ &\quad - \sum_{1\le i<j\le n-1} \mathbf{P}(X \in A_i \cap A_j) - \sum_{1\le i\le n-1} \mathbf{P}(X \in A_i \cap (A_n \cup A_{n+1}))\\ &\qquad + \cdots \pm \mathbf{P}(X \in A_1 \cap \cdots \cap A_{n-1} \cap (A_n \cup A_{n+1}))\\ &= \sum_{i=1}^{n+1} \mathbf{P}(X \in A_i) - \mathbf{P}(X \in A_n \cap A_{n+1})\\ &\quad - \sum_{1\le i<j\le n-1} \mathbf{P}(X \in A_i \cap A_j) - \sum_{i=1}^{n-1} (\mathbf{P}(X \in A_i \cap A_n) + \mathbf{P}(X \in A_i \cap A_{n+1})\\ &\qquad - \mathbf{P}(X \in A_i \cap A_n \cap A_{n+1}))\\ &\quad \pm \mathbf{P}(X \in A_1 \cap \cdots \cap A_{n-1} \cap A_n) \pm \mathbf{P}(X \in A_1 \cap \cdots \cap A_{n-1} \cap A_{n+1})\\ &\qquad \mp \mathbf{P}(X \in A_1 \cap \cdots \cap A_{n-1} \cap A_n \cap A_{n+1}).\end{aligned}$$

Daraus folgt die Behauptung. □

Beispiel 3.11 (Fixpunkte in Permutationen)**.** Sei Σ uniform verteilt auf $\mathbb{S}(n)$, d.h. Σ ist eine zufällige Permutation von $[1 : n]$. Wir wollen die Wahrscheinlichkeit bestimmen, mit der Σ mindestens einen Fixpunkt besitzt. Wir setzen $A_i := \{\sigma \in \mathbb{S}(n) : \sigma(i) = i\}$. Damit gilt für $i_1, ...i_k$ verschieden

$$\mathbf{P}(\Sigma \in A_{i_1} \cap \cdots \cap A_{i_k}) = \frac{(n-k)!}{n!}.$$

Damit gilt

$$\begin{aligned}\mathbf{P}(\Sigma \in A_1 \cup \cdots \cup A_n) &= n \cdot \frac{1}{n} - \binom{n}{2}\frac{1}{n(n-1)} + \binom{n}{3}\frac{1}{n(n-1)(n-2)} \cdots \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} \cdots \pm \frac{1}{n!}.\end{aligned}$$

Anders ausgedrückt haben genau

$$n!\Big(1 - \frac{1}{2!} - + \frac{1}{3!} \cdots \pm \frac{1}{n!}\Big)$$

Permutationen mindestens einen Fixpunkt.

3.4. Die gemeinsame Verteilung und Unabhängigkeit

Betrachtet man mehrere Zufallsvariablen gleichzeitig, so hat man es mit anderen Worten mit einem Vektor von Zufallsvariablen zu tun. Die Verteilung dieses Vektors wird auch gemeinsame Verteilung der Zufallsvariablen genannt. Eine besondere Situation tritt auf, wenn sich Wahrscheinlichkeiten für verschiedene Zufallsvariablen multiplizieren. Dann heißen diese (stochastisch) unabhängig.

Definition 3.12 (Gemeinsame Verteilung)**.** *Ist $X = (X_1, ..., X_n)$ mit Zielbereich $S := S_1 \times \ldots \times S_n$, so heißt $X_*\mathbf{P}$ auch die gemeinsame Verteilung von $X_1, ..., X_n$. Für die Projektion $\pi_i : S \to S_i$ heißt das Bild von X unter π_i auch die i-te Marginal- oder Randverteilung von X.*

Ist die gemeinsame Verteilung von $X_1, ..., X_n$ gerade das Produkt der Marginalverteilungen, d.h. gilt für alle (messbaren) $A_1, ..., A_n$

$$\mathbf{P}(X_1 \in A_1, ..., X_n \in A_n) = \mathbf{P}(X_1 \in A_1) \cdots \mathbf{P}(X_n \in A_n),$$

so heißt die Familie $(X_1, ..., X_n)$ stochastisch unabhängig.

Lemma 3.13 (Berechnung der Marginalverteilungen)**.** *Ist $X = (X_1, ..., X_n)$ mit Zielbereich $S := S_1 \times \ldots \times S_n$. Ist X diskret, so ist die i-te Marginalverteilung gegeben durch*

$$\mathbf{P}(X_i = a_i) = \sum_{a_1 \in S_1} \cdots \sum_{a_{i-1} \in S_{i-1}} \sum_{a_{i+1} \in S_{i+1}} \cdots \sum_{a_n \in S_n} \mathbf{P}(X_1 = a_1, ..., X_n = a_n).$$

Ist X kontinuierlich mit Dichte f_X, so ist die i-te Marginalverteilung gegeben durch die Dichte

$$f_i(x_i) = \int \cdots \int f(x_1, ..., x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

Beweis. Beide Aussagen folgen direkt aus der Definition der Marginalverteilung. □

Beispiel 3.14 (p-Münzwurf)**.** Wir betrachten den p-Münzwurf aus Beispiel 1.7. Es gilt: Ein Zufallsvektor $X = (X_1, ..., X_n)$ ist genau dann ein p-Münzwurf, wenn $X_1, ..., X_n$ unabhängig sind und

$$\mathbf{P}(X_i = 1) = 1 - \mathbf{P}(X_i = 0) = p.$$

Denn dann können wir die Verteilung von X angeben mittels

$$\begin{aligned}\mathbf{P}(X = (\omega_1, ..., \omega_n)) &= \mathbf{P}(X_1 = \omega_1, ..., X_n = \omega_n) = \mathbf{P}(X_1 = \omega_1) \cdots \mathbf{P}(X_n = \omega_n) \\ &= \prod_{i=1}^{n} p^{\omega_i}(1-p)^{1-\omega_i},\end{aligned}$$

was (1.2) entspricht.

Beispiel 3.15 (Zufällige Permutation)**.** Sei Σ uniform auf $\mathbb{S}(n)$ verteilt. Schreiben wir $\Sigma = (\Sigma(1), ..., \Sigma(n))$, so erhalten wir einen Vektor von Zufallsvariablen. Die Marginalverteilungen sind

$$\mathbf{P}(\Sigma(i) = k) = \frac{1}{n}, \qquad i, k = 1, ..., n,$$

d.h. $\Sigma(i)$ ist $U([1:n])$-verteilt, $i = 1, ..., n$. Anders ausgedrückt ist jede Marginalverteilung eine $U([1:n])$-verteilt. Allerdings ist die Familie $\Sigma(1), ..., \Sigma(n)$ nicht unabhängig, etwa ist nämlich für $i \neq j$

$$\mathbf{P}(\Sigma(i) = 1, \Sigma(j) = 1) = 0 \neq \frac{1}{n^2} = \mathbf{P}(\Sigma(i) = 1) \cdot \mathbf{P}(\Sigma(j) = 1).$$

Lemma 3.16 (Eigenschaften unabhängiger Zufallsvariable)**.** *Sei $(X_1, ..., X_n)$ eine unabhängige Familie von Zufallsvariablen.*

1. *Jede Teilfamilie $X_{i_1}, ..., X_{i_k}$ für $i_1, ..., i_k$ paarweise verschieden ist stochastisch unabhängig.*
2. *Seien $h_1, ..., h_n$ Funktionen auf den Zielbereichen $S_1, ..., S_n$. Dann sind $h_1(X_1), ..., h_n(X_n)$ unabhängig.*

Beweis. Für 1. wähle man $A_j = S_j$ für $j \notin \{i_1, ..., i_k\}$ in der Produktformel. Für 2. schreiben wir

$$\begin{aligned}\mathbf{P}(h_1(X_1) \in A_1, \cdots h_n(X_n) \in A_n) &= \mathbf{P}(X_1 \in h_1^{-1}(A_1), \cdots X_n \in h_n^{-1}(A_n)) \\ &= \mathbf{P}(X_1 \in h_1^{-1}(A_1)) \cdots \mathbf{P}(X_n \in h_n^{-1}(A_n)) \\ &= \mathbf{P}(h_1(X_1) \in A_1) \cdots \mathbf{P}(h_n(X_n) \in A_n).\end{aligned}$$

□

Proposition 3.17 (Unabhängigkeit diskreter und stetiger Zufallsvariable)**.** *Seien $X_1, ..., X_n$ Zufallsvariable mit Werten in $S_1, ..., S_n$.*

1. *Im Falle diskreter Zufallsvariablen sind folgende Aussagen äquivalent:*
   (i) $X_1, ..., X_n$ sind unabhängig.
   (ii) Es gilt
   $$\mathbf{P}(X_1 = a_1, ..., X_n = a_n) = \mathbf{P}(X_1 = a_1) \cdots \mathbf{P}(X_n = a_n).$$
2. *Im Falle kontinuierlicher Zufallsvariablen sei f_X die Dichte von X und $f_1, ..., f_n$ die Dichten von $X_1, ..., X_n$. Folgende Aussagen sind äquivalent:*
   (i) $X_1, ..., X_n$ sind unabhängig.

(ii) Es gilt

$$f_X(x_1, ..., x_n) = f_1(x_1) \cdots f_n(x_n).$$

Beweis. 1. '⇒' folgt direkt aus der Definition von Unabhängigkeit, wenn man $A_1 = \{a_1\}, ..., A_n = \{a_n\}$ setzt. '⇐' Es gilt

$$\begin{aligned}\mathbf{P}(X_1 \in A_1, ..., X_n \in A_n) &= \sum_{(a_1,...,a_n)\in A_1\times...\times A_n} \mathbf{P}(X_1 = a_1, ..., X_n = a_n) \\ &= \sum_{(a_1,...,a_n)\in A_1\times...\times A_n} \mathbf{P}(X_1 = a_1) \cdots \mathbf{P}(X_n = a_n) \\ &= \sum_{a_1\in A_1} \mathbf{P}(X_1 = a_1) \cdots \sum_{a_n\in A_n} \mathbf{P}(X_n = a_n) = \mathbf{P}(X_1 \in A_1) \cdots \mathbf{P}(X_n \in A_n).\end{aligned}$$

Eine analoge Rechnung liefert 2. □

4. Weitere wichtige Verteilungen

In diesem Kapitel lernen wir die wichtigsten Verteilungen kennen. Wir gehen wieder davon aus, dass Zufallsvariablen definiert sind, und geben diese wichtigen Verteilungen als Bildmaße an.

4.1. Die Binomial- und die Multinomialverteilung

Definition 4.1. *Sei $n \in \mathbb{N}$ und $p \in [0,1]$. Eine Zufallsvariable X mit Zielbereich $\{0,...,n\}$ heißt* binomialverteilt *zu den Parametern n und p, falls (mit $q := 1-p$)*

$$\mathbf{P}(X=k) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0,...,n.$$

Wir schreiben dann $X \sim B(n,p)$.

Bemerkung 4.2.

1. Wir haben in Beispiel 1.7 gesehen: Ist $X = (X_1,...,X_n)$ ein p-Münzwurf, so ist $X_1 + \cdots + X_n$ binomialverteilt zu den Parametern n und p.
   Generell erhält man die Binomialverteilung immer dann, wenn man ein Experiment hintereinander unabhängig ausführt, und es in jedem Experiment mit derselben Wahrscheinlichkeit p zu einem *Erfolg* und mit Wahrscheinlichkeit $q := 1-p$ zu einem *Misserfolg* kommen kann. Zählt man die Anzahl der *Erfolge*, so ist diese $B(n,p)$-verteilt. Der Grund ist, dass es $\binom{n}{k}$ verschiedene Möglichkeiten gibt, die k Erfolge auf die n Versuche zu verteilen und jede einzelne Abfolge von Erfolgen und Misserfolgen dieselbe Wahrscheinlichkeit (nämlich $p^k q^{n-k}$) hat.

2. Der bekannte binomische Lehrsatz ergibt, dass
   $$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = 1,$$
   so dass es sich bei der Binomialverteilung in der Tat um ein Wahrscheinlichkeitsmaß handelt.

Unterscheidet man bei einem Experiment nicht nur zwischen Erfolgen und Misserfolgen, sondern zwischen ℓ verschiedenen Ausgängen, so erhält man die Multinomialverteilung.

Definition 4.3 (Multinomialverteilung)**.** *Sei $n \in \mathbb{N}$ und $p_1,...,p_\ell \in [0,1]$ mit $p_1 + \cdots + p_\ell = 1$. Eine $\Delta_\ell := \{(k_1,...,k_\ell) : k_1 + \cdots + k_\ell = n\}$-wertige Zufallsvariable $X = (X_1,...,X_\ell)$ heißt* multinomialverteilt *zu den Parametern n und $p_1,...,p_\ell$, falls*

$$\mathbf{P}(X = (k_1,...,k_\ell)) = \binom{n}{k_1 \ \cdots \ k_\ell} p_1^{k_1} \cdots p_\ell^{k_\ell}.$$

Hierbei ist

$$\binom{n}{k_1 \ \cdots \ k_\ell} := \frac{n!}{k_1! \cdots k_\ell!}$$

ein Multinomialkoeffizient.

4.2. Die hypergeometrische Verteilung

Gegeben seien N unterscheidbare Objekte, von denen K markiert sind. Wie viele Möglichkeiten gibt es, n Objekte (ohne Beachtung der Reihenfolge) so auszuwählen, dass genau k Markierte dabei sind? Die Antwort erhält man wie folgt: von den K Markierten muss man k auswählen, das sind schon mal $\binom{K}{k}$ Möglichkeiten. Bei den restlichen $n-k$ auszuwählenden Objekten darf man dann nur noch nicht-Markierte nehmen. Dies sind dann $\binom{N-K}{n-k}$ Möglichkeiten. Insgesamt ergeben sich also

$$\binom{K}{k}\binom{N-K}{n-k}$$

Möglichkeiten. Diese Überlegung führt schon auf ein erstes Resultat.

Lemma 4.4 (Eine Identität mit Binomialkoeffizienten)**.** *Es gilt für $N \geq n, K \geq 0$*

$$\sum_{k=0}^{n} \binom{K}{k}\binom{N-K}{n-k} = \binom{N}{n}.$$

Beweis. Einen anschaulichen Beweis erhält man, wenn man sich die Bedeutung der linken Seite ansieht. Nach obiger Erklärung wird hier über die Anzahl der Markierten Objekte summiert, die man beim Ziehen von n der N Objekte erhält. Insgesamt steht dort also die Anzahl der Möglichkeiten, (ohne Beachtung der Reihenfolge) aus den N Objekten gerade n auszuwählen, also $\binom{N}{n}$. Einen anderen Beweis erhält man durch Induktion. □

Definition 4.5 (Hypergeometrische Verteilung)**.** *Sei $n, N \in \mathbb{N}$ und $K \in \mathbb{N}_0$ mit $n, K \leq N$. Eine $[0 : n \wedge K]$-wertige Zufallsvariable X heißt* hypergeometrisch verteilt *mit n, N, K, wenn*

$$\mathbf{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \qquad k = 0, 1, ..., n \wedge K.$$

Wir schreiben dann auch $X \sim Hyp(n, N, K)$.

Bemerkung 4.6 (Urnen in der Stochastik)**.** Betrachte eine Urne mit insgesamt N Kugeln, wovon K weiß sind. Wir ziehen n Kugeln heraus und betrachten

$$X_i := 1_{i\text{-te Kugel ist weiß}}, \qquad i = 1, ..., n,$$
$$X := X_1 + \cdots + X_n.$$

Ziehen wir mit Zurücklegen, so haben wir in jedem Zug die gleiche Chance von $p = \frac{K}{N}$, eine weiße Kugel zu ziehen. In diesem Fall ist $X \sim B(n, p)$.

Ziehen wir ohne Zurücklegen, so sind die Zufallsvariablen $X_1, ..., X_n$ nicht unabhängig, da jeder Zug die Wahrscheinlichkeiten, im nächsten Zug eine weiße Kugel zu ziehen, ändert. In diesem Fall ist $X \sim \mathrm{Hyp}(n, N, K)$.

4.3. Die Poisson-Verteilung und die Poisson-Approximation

Wie wir in Theorem 4.8 sehen werden, entsteht aus der Binomialverteilung bei sehr kleinen Erfolgswahrscheinlichkeiten, jedoch häufigen Versuchen, die Poisson-Verteilung.

Definition 4.7. *Sei $\lambda \in \mathbb{R}_+$. Eine $\mathbb{N}_0$-wertige Zufallsvariable X heißt Poisson-verteilt mit Parameter λ, falls*

$$\mathbf{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, ...$$

In diesem Fall schreiben wir $X \sim Poi(\lambda)$.

Theorem 4.8 (Gesetz der kleinen Zahlen)**.** *Sei $X_n \sim B(n, p_n)$ für $n = 1, 2, ...$, so dass*

$$n \cdot p_n \xrightarrow{n\to\infty} \lambda > 0.$$

Sei weiter $Z \sim Poi(\lambda)$. Dann gilt

$$\mathbf{P}(X_n = k) \xrightarrow{n\to\infty} \mathbf{P}(Z = k).$$

Beweis. Wir schreiben direkt

$$\begin{aligned}\mathbf{P}(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\to 1} \cdot \frac{1}{k!} \underbrace{(np_n)^k}_{\to \lambda^k} \underbrace{\left(1 - \frac{np_n}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{(1-p_n)^{-k}}_{\to 1} \xrightarrow{n\to\infty} = \mathbf{P}(Z = k).\end{aligned}$$

□

4.4. Die geometrische und die Exponentialverteilung

In Beispiel 1.8 haben wir die Verteilung der Wartezeit bis zur ersten 6 beim Würfelwurf berechnet. Solche Wartezeiten sind oftmals geometrisch verteilt.

Definition 4.9 (Geometrische Verteilung)**.** *Sei $p \in (0, 1]$. Eine $\mathbb{N}$-wertige Zufallsvariable X heißt* geometrisch *verteilt mit Parameter p, falls für $q = 1 - p$*

$$\mathbf{P}(X > i) = q^i, \qquad i = 1, 2, ...$$

Wir schreiben dann auch $X \sim geo(p)$. Es gilt dann also

$$\mathbf{P}(X = i) = \mathbf{P}(X > i - 1) - \mathbf{P}(X > i) = q^{i-1} - q^i = q^{i-1}p.$$

Bemerkung 4.10 (Interpretation beim p-Münzwurf)**.**

1. Sei $X = (X_1, X_2, ...)$ ein unendlicher (!) p-Münzwurf und
$$T := \min\{i : X_i = 1\}.$$
Dann ist $T \sim \text{geo}(p)$.

2. In manchen Büchern nimmt eine geometrisch verteilte Zufallsvariable Y Werte in $\mathbb{N}_0$ an. In diesem Fall ist bei einem unendlichen p-Münzwurf Y nicht der Versuch des ersten Erfolges, sondern der letzte in der Serie von Misserfolgen, d.h. $Y = T - 1$ mit T wie in 1.

3. Für den unendlichen p-Münzwurf $X = (X_1, X_2, ...)$ bemerken wir kurz, dass der Wahrscheinlichkeitsraum, auf dem X definiert ist, nicht diskret sein kann, da es überabzählbar viele $\{0,1\}$-wertige Folgen gibt. Immerhin können wir den unendlichen Münzwurf im Rahmen unserer Möglichkeiten für $p = 1/2$ dennoch definieren. Hierfür sei $\Omega = [0,1]$ und $\mathbf{P}$ eine $U([0,1])$-Verteilung. Für $x \in \Omega$ sei weiter $x = \sum_{i=1}^{\infty} x_i 2^{-i}$ die Binärdarstellung von x, also $x_i = [2^i x - \lfloor 2^i x \rfloor]$, $i = 1, 2, ...$, wobei $[.]$ die Rundung und $\lfloor . \rfloor$ die Abrundung bedeutet. Dann ist mit $X_i(x) = x_i$ der Vektor $X = (X_1, X_2, ...)$ der unendliche $1/2$-Münzwurf.

Wartezeiten sind nicht immer diskrete Anzahlen von Versuchen, sondern können auch kontinuierliche Größen sein. Dies wird zumeist durch Exponentialverteilungen dargestellt.

Definition 4.11 (Exponentialverteilung)**.** *Sei $\lambda > 0$. Eine $\mathbb{R}$-wertige Zufallsvariable X heißt* exponentialverteilt *mit Rate λ, falls sie die Dichte*

$$f_X(x) = \lambda e^{-\lambda x} 1_{x \geq 0} \tag{4.1}$$

und die Verteilungsfunktion

$$F_X(x) = (1 - e^{-\lambda x}) 1_{x \geq 0} \tag{4.2}$$

besitzt. Dies ist genau dann der Fall, wenn

$$\mathbf{P}(X > x) = e^{-\lambda x}, \qquad x \geq 0.$$

Wir schreiben dann $X \sim Exp(\lambda)$.

Proposition 4.12 (Eigenschaften der Exponentialverteilung)**.** *Sei $\lambda, \mu > 0$.*

1. *Sei $X \sim Exp(\lambda)$. Dann ist $\mu X \sim Exp(\lambda/\mu)$*
2. *Sei $U \sim U([0,1])$. Dann ist $-(\log U)/\lambda \sim Exp(\lambda)$.*

Beweis. Sei $x \geq 0$. 1. Für $Y := \mu X$ gilt für $x \geq 0$

$$\mathbf{P}(Y > x) = \mathbf{P}(X > x/\mu) = e^{-\lambda x/\mu}.$$

Daraus folgt bereits die Behauptung. 2. folgt einerseits mit Lemma 3.9. Alternativ schreiben wir für $Z := -\frac{\log U}{\lambda}$

$$\mathbf{P}(Z > x) = \mathbf{P}(-(\log U)/\lambda > x) = \mathbf{P}(U < e^{-\lambda x}) = e^{-\lambda x}.$$

Wieder folgt die Behauptung. □

Proposition 4.13 (Exponentialapproximation)**.** *Sei $Z \sim Exp(1)$ und $X_1, X_2, ...$ eine Folge von geometrisch verteilten Zufallsvariablen zu den Parametern $p_1, p_2, ...$ mit $p_n \downarrow 0$. Dann gilt für $0 \leq c < d \leq \infty$*

$$\mathbf{P}\Big(c \leq p_n X_n \leq d\Big) \xrightarrow{n \to \infty} \mathbf{P}(c \leq Z \leq d).$$

Beweis. Sei

$$c_n := (\lceil c/p_n \rceil - 1)p_n, \qquad d_n := (\lfloor d/p_n \rfloor - 1)p_n,$$

also $c_n \xrightarrow{n\to\infty} c, d_n \xrightarrow{n\to\infty} d$. Dann

$$\begin{aligned}
\mathbf{P}\Big(c \le p_n X_n \le d\Big) &= \mathbf{P}(X_n \ge c/p_n) - \mathbf{P}(X_n > d/p_n) \\
&= (1-p_n)^{\lceil c/p_n \rceil - 1} - (1-p_n)^{\lfloor d/p_n \rfloor - 1} \\
&= \Big((1-p_n)^{1/p_n}\Big)^{c_n} - \Big((1-p_n)^{1/p_n}\Big)^{d_n} \\
&\xrightarrow{n\to\infty} e^{-c} - e^{-d} = \int_c^d e^{-x}dx.
\end{aligned}$$

□

4.5. Die Normalverteilung

In der Statistik werden wir sehen, dass die Normalverteilung grundlegend für viele Anwendungen ist. Sie wird oft mit der Gauß'schen Glockenkurve in Verbindung gebracht. Dies ist der Graph der Funktion

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Aus der Analysis bekannt ist die Tatsache[1]

$$\int e^{-x^2} dx = \sqrt{\pi}.$$

Daraus folgt, dass $\int f(x)dx = 1$ und damit f eine Dichte ist.

[1] Vermutlich wird dies erst am Ende der Vorlesung Analysis 3 bewiesen. Ein elementarer Beweis ist der folgende, basierend auf der Gamma-Funktion

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

Wir schreiben

$$\begin{aligned}
\Gamma\Big(\frac{1}{2}\Big)^2 &= \int_0^\infty \int_0^\infty s^{-1/2} t^{-1/2} e^{-(s+t)} ds dt \\
&\overset{r=t+s}{=} \int_0^\infty \int_0^r s^{-1/2}(r-s)^{-1/2} e^{-r} ds dr \\
&\overset{u=s/r}{=} \int_0^\infty \int_0^1 r u^{-1/2} r^{-1/2} r^{-1/2} (1-u)^{-1/2} e^{-r} du dr \\
&= \int_0^\infty e^{-r} dr \cdot \int_0^1 \frac{1}{\sqrt{u(1-u)}} du \\
&\overset{s=2u-1}{=} \int_{-1}^1 \frac{ds}{\sqrt{1-s^2}} = \arcsin(s)|_{-1}^1 = \pi.
\end{aligned}$$

Daraus folgt

$$\int_{-\infty}^\infty e^{-x^2} dx = 2\int_0^\infty e^{-x^2} dx \overset{y=x^2}{=} \int_0^\infty e^{-y} y^{-\frac{1}{2}} dy = \Gamma(1/2) = \sqrt{\pi}.$$

Definition 4.14. *Sei $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$. Eine $\mathbb{R}$-wertige Zufallsvariable heißt normalverteilt mit Erwartungswert μ und Varianz σ^2, falls sie die Dichte*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big) \tag{4.3}$$

besitzt. Wir schreiben dann auch $X \sim N(\mu, \sigma^2)$. Im Fall $\mu = 0, \sigma^2 = 1$ sagen wir auch, X sei standardnormalverteilt.

Lemma 4.15 (Eigenschaften der Normalverteilung)**.** *Sei $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$.*

1. *Ist $Z \sim N(0,1)$, so ist $\sigma Z + \mu \sim N(\mu, \sigma^2)$.*
2. *Ist $X \sim N(\mu, \sigma^2)$, so ist $\frac{X-\mu}{\sigma} \sim N(0,1)$.*

Beweis. 1. Für $x \in \mathbb{R}$ schreiben wir

$$\begin{aligned}\mathbf{P}(\sigma Z + \mu \le x) = \mathbf{P}\Big(Z \le \frac{x-\mu}{\sigma}\Big) &= \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &\overset{y=\sigma z+\mu}{=} \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y-\mu)^2}{2\sigma^2}\Big) dy.\end{aligned}$$

2. folgt analog. □

5. Kenngrößen von Zufallsvariablen

Um sich etwas unter einer Verteilung vorstellen zu können, ist es oftmals hilfreich, diese durch Kenngrößen zu beschreiben.

5.1. Der Erwartungswert

Wenn man eine Verteilung einer Zufallsvariable durch eine einzige Zahl beschreiben müsste, würde man wohl einen Lageparameter der Verteilung wählen. Diese beschreiben grob, wo Werte von X am ehesten liegen. Der häufigste Lageparameter ist der Erwartungswert.

Definition 5.1 (Erwartungswert)**.** *Sei X eine $S \subseteq \mathbb{R}$-wertige Zufallsvariable.*

1. *Ist X diskret, so ist der* Erwartungswert *von (der Verteilung von) X*

$$\mathbf{E}[X] := \sum_{x \in S} x \cdot \mathbf{P}(X = x),$$

   falls die Summe wohldefiniert ist.

2. *Ist X kontinuierlich mit Dichte f, so ist* Erwartungswert *von (der Verteilung von) X*

$$\mathbf{E}[X] := \int x f(x) dx,$$

   falls das Integral wohldefiniert ist.

Bemerkung 5.2 (k-tes Moment; weitere Lageparameter)**.** 1. Allgemeiner als der Erwartungswert ist das k-te Moment einer Zufallsvariable X, gegeben als $\mathbf{E}[X^k], k = 1, 2, ...$

2. Neben dem Erwartungswert gibt es noch weitere Lageparameter einer Verteilung, nämlich Median und Modalwert. Der Median der Verteilung von X ist dabei jede Zahl m mit

$$\mathbf{P}(X \geq m) \geq \tfrac{1}{2}, \qquad \mathbf{P}(X \leq m) \geq \tfrac{1}{2}.$$

   Der Modalwert ist argmax $\mathbf{P}(X = x)$ im Falle diskreter und argmax $f_X(x)$ im Falle einer Zufallsvariablen mit Dichte.

   Der Erwartungswert ist von allen Lageparametern oft am besten handhabbar, was vor allem an seiner Linearität liegt (Proposition 5.5).

Lemma 5.3 (Transformation von Erwartungswerten)**.** *Sei X eine S-wertige Zufallsvariable und $h : S \to \mathbb{R}$.*

1. *Ist X diskret, so gilt*

$$\mathbf{E}[h(X)] = \sum_{x \in S} h(x) \mathbf{P}(X = x),$$

   falls die Summe existiert.

2. *Ist X kontinuierlich mit Dichte f, und ist h streng monoton und differenzierbar mit differenzierbarer Umkehrabbildung, so gilt*

$$\mathbf{E}[h(X)] = \int h(x)f(x)dx.$$

Beweis. 1. Wir schreiben

$$\begin{aligned}\mathbf{E}[h(X)] &= \sum_{y\in h(S)} y\mathbf{P}(h(X)=y) = \sum_{y\in h(S)} y \sum_{x\in h^{-1}(y)} \mathbf{P}(X=a)\\ &= \sum_{y\in h(S)}\sum_{x\in h^{-1}(y)} h(x)\mathbf{P}(X=a)\\ &= \sum_{x\in S} h(x)\mathbf{P}(X=x)\end{aligned}$$

2. OBdA sei h streng monoton wachsend. (Ansonsten wechseln wir zu $-h$.) Nach Lemma3.7 hat $h(X)$ die Dichte $y \mapsto f(h^{-1}(y)|(h^{-1})'|(y)$. Deshalb gilt

$$\mathbf{E}[h(X)] = \int yf(h^{-1}(y)(h^{-1})'(y)dy \overset{z=h^{-1}(y)}{=} \int h(z)f(z)dz.$$

□

Bemerkung 5.4. In 2. mussten wir starke Annahmen an h machen, um $\mathbf{E}[h(X)]$ berechnen zu können. Richtig ist allerdings, dass die Aussage auch für viel allgemeinere Funktionen h gilt, nämlich für alle, für die $\int f(x)h(x)dx$ existiert. Dies kann man allerdings erst mit Maßtheorie einsehen. Wir werden zur Vereinfachung diese Aussage auch für allgemeine Funktionen h verwenden.

Wir kommen nun zu Eigenschaften des Erwartungswertes.

Proposition 5.5 (Linearität des Erwartungswertes)**.** *1. Sei X eine S-wertige Zufallsvariable und $A \subseteq S$ (messbar). Dann gilt*

$$\mathbf{E}[1_A(X)] = \mathbf{P}(X \in A).$$

2. *Seien $c, d \in \mathbb{R}$ und X, Y Zufallsvariablen mit Zielbereichen $S, T \subseteq \mathbb{R}$ und wohldefiniertem Erwartungswert. Dann gilt*

$$\mathbf{E}[cX + dY] = c\mathbf{E}[X] + d\mathbf{E}[Y].$$

Beweis. 1. Es gilt

$$\mathbf{E}[1_A(X)] = 0 \cdot \mathbf{P}(X \notin A) + 1 \cdot \mathbf{P}(X \in A) = \mathbf{P}(X \in A).$$

2. Sind X, Y diskret, so gilt mit Lemma 5.3

$$\begin{aligned}\mathbf{E}[cX + dY] &= \sum_{X\in S}\sum_{y\in T}(cx + dy)\mathbf{P}(X=x, Y=y)\\ &= c\sum_{x\in S} x\mathbf{P}(X=x) + d\sum_{y\in T} y\mathbf{P}(Y=y)\\ &= c\mathbf{E}[X] + d\mathbf{E}[Y].\end{aligned}$$

Haben X, Y die gemeinsame Dichte f, und haben die Marginalverteilungen f_X und f_Y, so ist

$$\begin{aligned}\mathbf{E}[cX + dY] &= \int (cx + dy) f(x,y) dxdy \\ &= c \int x \int f(x,y) dy dx + d \int y \int f(x,y) dx dy \\ &= c \int x f_X(x) dx + d \int y f_Y(y) dy \\ &= c\mathbf{E}[X] + d\mathbf{E}[Y].\end{aligned}$$

□

Beispiel 5.6 (Fixpunkte in Permutationen)**.** Sei Σ eine zufällige Permutation von $[1 : n]$. Weiter sei

$$X = \sum_{i=1}^{n} X_i, \qquad X_i = 1_{\Sigma(i)=i}$$

die Anzahl an Fixpunkten in Σ. Wir zeigen nun, dass

$$\mathbf{E}[X] = 1.$$

Denn mit Linearität des Erwartungswertes gilt

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} \mathbf{P}(\Sigma(i) = i) = n \cdot \frac{1}{n} = 1.$$

Beispiel 5.7. In einer Kugel befinden sich $1 \leq w \leq g$ markierte und $g - w$ unmarkierte Kugeln. Wir ziehen die Kugeln ohne Zurücklegen aus der Urne, und schreiben $X_i = 1$, falls die i-te Kugel markiert ist und $X_i = 0$ sonst. Wir wollen $\mathbf{E}[T]$ für $T := \min\{i : T_i = 1\}$, also den ersten Zug einer markierten Kugel, berechnen. Wir behaupten

$$\mathbf{E}[T] = \frac{g+1}{w+1}.$$

Eine erste Möglichkeit ist die folgende: Die Verteilung von T lässt sich angeben als

$$\mathbf{P}(T = k+1) = \frac{(g-w)\cdots(g-w-k+1)}{g\cdots(g-k+1)} \frac{w}{g-k} = \frac{\binom{g-w}{k}}{\binom{g-1}{k}} \frac{w}{g}, \qquad k = 0, ..., g-w.$$

Wir zeigen die Behauptung nun mittels Induktion nach g. Für $g = 1$ muss $w = 1$ sein und die Behauptung ist klar. Gilt die Behauptung für $g - 1$, so ist für g

$$\begin{aligned}\mathbf{E}[T] &= \sum_{k=0}^{g-w} (k+1)\mathbf{P}(T = k+1) = 1 + \sum_{k=0}^{g-w} k\mathbf{P}(T = k+1) \\ &= 1 + \frac{w}{g} \sum_{k=0}^{g-w} k \frac{\binom{g-w}{k}}{\binom{g-1}{k}} = 1 + \frac{w}{g} \sum_{k=0}^{g-w-1} (k+1) \frac{\binom{g-w}{k+1}}{\binom{g-1}{k+1}} \\ &= 1 + \frac{g-w}{g} \underbrace{\sum_{k=0}^{g-w-1} (k+1) \frac{\binom{g-1-w}{k}}{\binom{g-2}{k}} \frac{w}{g-1}}_{= \frac{g}{w+1} \text{ nach Induktionsannahme}} \\ &= 1 + \frac{g-w}{w+1} = \frac{g+1}{w+1}.\end{aligned}$$

Eine zweite Möglichkeit ist die folgende: Wir definieren induktiv

$$T_0 := 0, \qquad T_{\ell+1} := \min\{i > T_i : X : i = 1\}, \quad \ell = 0, ..., w-1$$

sowie $T_{w+1} = g + 1$. Dann sind $T_{\ell+1} - T_\ell, \ell = 0, ..., w$ identisch verteilt mit

$$g + 1 = T_{w+1} - T_0 = \sum_{\ell=0}^{w} T_{\ell+1} - T_\ell.$$

Deshalb gilt

$$g + 1 = \sum_{\ell=0}^{w} \mathbf{E}[T_{\ell+1} - T_\ell] = (w+1)\mathbf{E}[T_1] = \mathbf{E}[T]$$

und die Behauptung folgt.

Proposition 5.8 (Positivität und Monotonie des Erwartungswertes)**.** *1. Sei $X \geq 0$ eine $\mathbb{R}$-wertige Zufallsvariable. Dann gilt*

$$\mathbf{E}[X] \geq 0$$

und

$$\mathbf{E}[X] = 0 \quad \textit{genau dann, wenn} \quad \mathbf{P}(X = 0) = 1.$$

2. Für $\mathbb{R}$-wertige Zufallsvariable X_1, X_2 mit $X_1 \leq X_2$ mit wohldefinierten Erwartungswerten gilt $\mathbf{E}[X_1] \leq \mathbf{E}[X_2]$.

Beweis. 1. Ist X diskret, so ergeben sich die ersten beiden Aussagen aus der Definition

$$\mathbf{E}[X] = \sum_{x \in \mathbb{R}} x\mathbf{P}(X = x).$$

Hat X die Dichte f, so muss $f(x) = 0$ für $x \leq 0$ gelten, da sonst $\mathbf{P}(X < 0) > 0$ wäre. Dann ist $\mathbf{E}[X] = \int_0^\infty xf(x)dx > 0$ wegen der Positivität des Integranden. Insbesondere kann die zweite Aussage dann nicht zutreffen. Die letzte Aussage folgt aus der ersten, da

$$0 \leq \mathbf{E}[X_2 - X_1] = \mathbf{E}[X_2] - \mathbf{E}[X_1].$$

□

Proposition 5.9 (Jensen'sche Ungleichung)**.** *Sei $\varphi : \mathbb{R} \to \mathbb{R}$ konvex und X eine $\mathbb{R}$-wertige Zufallsvariable mit endlichem Erwartungswert. Dann ist*

$$\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi X)].$$

Insbesondere gilt

$$|\mathbf{E}[X]| \leq \mathbf{E}[|X|], \qquad \mathbf{E}[X]^2 \leq \mathbf{E}[X^2].$$

Beweis. Da φ konvex ist, gibt es für jedes $d \in \mathbb{R}$ ein $c \in \mathbb{R}$, so dass $\varphi(d) + c(x - d) \leq \varphi(x)$ für alle $x \in \mathbb{R}$. Für $d = \mathbf{E}[X]$ ist also

$$\varphi(\mathbf{E}[X]) = \mathbf{E}\big[\varphi(\mathbf{E}[X]) + c(X - \mathbf{E}[X])\big] \leq \mathbf{E}[\varphi(X)].$$

□

Proposition 5.10 (Cauchy-Schwartz Ungleichung)**.** *Seien X, Y reellwertige Zufallsvariable mit $\mathbf{E}[X^2], \mathbf{E}[Y^2] < \infty$. Dann gilt*

$$\mathbf{E}[XY]^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2].$$

Beweis. Im Fall $\mathbf{P}(X = 0) = 1$ oder $\mathbf{P}(Y = 0) = 1$ ist die Ungleichung trivial. Andernfalls betrachte

$$U := \frac{X}{\sqrt{\mathbf{E}[X^2]}}, \qquad V := \frac{Y}{\sqrt{\mathbf{E}[Y^2]}}.$$

Wegen $(U - V)^2, (U + V)^2 \geq 0$ ist klar, dass

$$\pm 2UV \leq U^2 + V^2, \text{ also } \pm \mathbf{E}[UV] \leq 1.$$

Multiplizieren der letzten Gleichung mit $\sqrt{\mathbf{E}[X^2]} \cdot \sqrt{\mathbf{E}[Y^2]}$ und Quadrieren liefert die Behauptung. □

Proposition 5.11 (Multiplikation von Erwartungswerten bei Unabhängigkeit)**.** *Seien X, Y reellwertige Zufallsvariable mit endlichen Erwartungswerten. Sind X, Y unabhängig, so gilt*

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

Beweis. Falls X, Y diskret sind, schreiben wir

$$\begin{aligned}\mathbf{E}[XY] &= \sum_{x,y} xy\mathbf{P}(X = x, Y = y) = \sum_{x,y} xy\mathbf{P}(X = x)\mathbf{P}(Y = y) \\ &= \sum_x x\mathbf{P}(X = x) \cdot \sum_y y\mathbf{P}(Y = y) = \mathbf{E}[X]\mathbf{E}[Y].\end{aligned}$$

Haben X, Y die gemeinsame Dichte f und $f(x, y) = f_X(x)f_Y(y)$ wegen der Unabhängigkeit (siehe Proposition 5.3), so folgt

$$\begin{aligned}\mathbf{E}[XY] &= \int xyf(x, y)dxdy = \int xf_X(x)yf_Y(y)dxdy = \int xf_X(x)dx \cdot \int yf_Y(y)dy \\ &= \mathbf{E}[X]\mathbf{E}[Y].\end{aligned}$$

□

Beispiel 5.12 (Runs in Münzwürfen)**.** Sei $X = (X_1, ..., X_n)$ ein p-Münzwurf. Ein *Run* ist ein maximaler Block aus 0ern oder 1ern, also enthält beispielsweise 1100010 genau vier Runs. Bezeichnen wir mit Y die Anzahl der Runs, so können wir das formalisieren mittels

$$Y_n = 1 + \sum_{i=2}^{n} 1_{E_i}, \qquad E_i = \{X_i \neq X_{i-1}\}.$$

Mit dieser Formel lässt sich nun ganz einfach die erwartete Anzahl von Runs berechnen. Es gilt nämlich

$$\mathbf{E}[1_{E_i}] = \mathbf{P}(X_i \neq X_{i-1}) = 2pq$$

und damit

$$\mathbf{E}[Y] = 1 + (n - 1)\mathbf{E}[1_{E_1}] = 2(n - 1)pq + 1.$$

5.2. Die Varianz und die Covarianz

Will man die Verteilung einer Zufallsvariable mit zwei Parametern beschreiben, so kommt neben einem Lageparameter meist noch ein Streuungsparameter (oder Skalierungsparameter) hinzu. Dieser gibt an, wie eng X um den Lageparameter herum verteilt ist. Der wichtigste Streuungsparameter ist dabei die Varianz einer Zufallsvariable.

Definition 5.13 (Varianz und Covarianz)**.** *Seien X, Y eine $S \subseteq \mathbb{R}$-wertige Zufallsvariable mit endlichem Erwartungswert.*

1. *Die* Varianz *von X ist*

$$\sigma^2 := \mathbf{Var}[X] := \mathbf{E}[(X - \mathbf{E}[X])^2].$$

Weiter ist die Standardabweichung von X gegeben durch

$$\sigma := \sqrt{\mathbf{Var}[X]}.$$

2. *Die* Covarianz *von X, Y ist*

$$\mathbf{Cov}[X, Y] := \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])].$$

Ist $\mathbf{Cov}[X, Y] = 0$, so heißen X, Y unkorreliert.

Bemerkung 5.14 (k-tes zentriertes Moment)**.** Allgemeiner als die Varianz ist das k-te zentrierte Moment einer Zufallsvariable X, gegeben als $\mathbf{E}[(X - \mathbf{E}[X])^k], k = 2, 3, ...$

Lemma 5.15 (Rechenregeln für die Varianz und die Covarianz)**.** *Seien $X, Y, Z, X_1, ..., X_n$ reellwertige Zufallsvariable mit endlichem Erwartungswert. Es gilt*

1. $$\mathbf{Var}[X] = \mathbf{Cov}[X, X] = \mathbf{E}[X(X-1)] + \mathbf{E}[X] - \mathbf{E}[X]^2 = \mathbf{E}[X^2] - \mathbf{E}[X]^2.$$

2. $$\mathbf{Cov}[X, Y] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

3. $$\mathbf{Cov}[aX + bY, Z] = a\mathbf{Cov}[X, Z] + b\mathbf{Cov}[Y, Z].$$

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

5. *Sind X, Y unabhängig, so ist $\mathbf{Cov}[X, Y] = 0$.*

Beispiel 5.16 (Unkorrelierte Zufallsvariable)**.** Wir betrachten noch ein Beispiel zu Lemma 5.15.5. Seien (X, Y) Zufallsvariable, die jeweils mit Wahrscheinlichkeit $\frac{1}{4}$ die Zustände $(1, 0), (0, 1), (-1, 0)$ und $(0, -1)$ annehmen. Dann ist $\mathbf{P}(XY = \pm 1)\frac{1}{2}$, also $\mathbf{E}[XY] = \mathbf{E}[X] = \mathbf{E}[Y] = 0$. Insbesondere sind X, Y unkorreliert.

Wir haben zwar gerade gezeigt, dass unabhängige Zufallsvariable immer unkorreliert sind. In obigem Beispiel ist allerdings

$$\mathbf{P}(X = 1, Y = 1) = 0 \neq \frac{1}{16} = \mathbf{P}(X = 1) \cdot \mathbf{P}(Y = 1),$$

also sind X, Y zwar unkorreliert, jedoch nicht unabhängig.

Beweis von Lemma 5.15. 2. Es gilt wegen der Linearität des Erwartungswertes

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

1. Die erste Gleichheit liest man direkt aus der Definition ab. Die dritte Gleichheit folgt aus 1. und der ersten Gleichheit. Die zweite Gleichheit folgt aus der dritten Gleichheit und der Linearität des Erwartungswertes.
3. ist eine einfache Anwendung von 2.
4. Wir schreiben mit 3.

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

und die Behauptung folgt mit 1.
5. folgt aus Proposition 5.11. □

Beispiel 5.17 (Runs in Münzwürfen)**.** Wir betrachten nochmal das Beispiel 5.12, wollen nun jedoch die Varianz der Anzahl von Runs mit Hilfe von Lemma (5.15).4 berechnen. Ganz formal schreiben wir

$$\mathbf{Var}[Y] = \mathbf{Var}\Big[\sum_{i=2}^n 1_{E_i}\Big] = \sum_{i=2}^n \mathbf{Var}[1_{E_i}] + \sum_{i\ne j} \mathbf{Cov}[1_{E_i}, 1_{E_j}].$$

Zunächst ist

$$\mathbf{V}[1_{E_i}] = \mathbf{E}[1_{E_i}^2] - (2pq)^2 = \mathbf{E}[1_{E_i}] - (2pq)^2 = 2pq(1-2pq).$$

Weiter sind $1_{E_i}, 1_{E_j}$ für $|i-j|>1$ unabhängig, und für $2\le i\le n-1$ ist

$$\begin{aligned}\mathbf{Cov}[1_{E_i}, 1_{E_i+1}] &= \mathbf{E}[1_{E_i}\cdot 1_{E_{i+1}}] - (2pq)^2 = \mathbf{P}(X_{i+1}\ne X_i \ne X_{i-1}) - (2pq)^2 \\ &= pqp + qpq - (2pq)^2 = pq(1-4pq).\end{aligned}$$

Setzen wir dies zusammen, so ist

$$\begin{aligned}\mathbf{Var}[Y] &= \sum_{i=2}^n \mathbf{Var}[1_{E_i}] + 2\sum_{i=2}^{n-1} \mathbf{Cov}[1_{E_i}, 1_{E_{i+1}}] \\ &= 2pq(1-2pq)(n-1) + 2pq(1-4pq)(n-2).\end{aligned}$$

Beispiel 5.18 (Fixpunkte in Permutationen)**.** Wir betrachten die Situation aus Beispiel 5.6, sei also $\Sigma\sim U(\mathbb{S}(n))$ und

$$X = \sum_{i=1}^n X_i, \qquad X_i = 1_{\Sigma(i)=i}.$$

Ziel ist es, $\mathbf{Var}[X]$ zu bestimmen. Wir berechnen

$$\begin{aligned}\mathbf{P}(X_i=1) &= \frac{1}{n}, \\ \mathbf{P}(X_i=X_j=1) &= \frac{1}{n(n-1)}.\end{aligned}$$

Daraus ergibt sich für $i \neq j$

$$\begin{aligned}\mathbf{Var}[X_i] &= \frac{1}{n} - \frac{1}{n^2} = \frac{n-1}{n^2},\\ \mathbf{Cov}[X_i, X_j] &= \frac{1}{n(n-1)} - \frac{1}{n^2} = \frac{1}{n^2(n-1)},\end{aligned}$$

und insgesamt

$$\mathbf{Var}[X] = n\frac{n-1}{n^2} + n(n-1)\frac{1}{n^2(n-1)} = \frac{n-1}{n} + \frac{1}{n} = 1.$$

Wir formulieren noch die Cauchy-Schwartz-Ungleichung im Kontext von Varianzen und Covarianzen.

Korollar 5.19 (Cauchy-Schwartz-Ungleichung)**.** *Seien X, Y reellwertige Zufallsvariable mit endlichen Varianzen. Dann gilt*

$$\mathbf{Cov}[X, Y]^2 \leq \mathbf{Var}[X]\mathbf{Var}[Y].$$

Insbesondere ist

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

für den Korrelationskoeffizienten

$$\rho(X, Y) := \frac{\mathbf{Cov}[X, Y]}{\sqrt{\mathbf{Var}[X]}\sqrt{\mathbf{Var}[Y]}}.$$

Beweis. Die Behauptungen folgen direkt aus Proposition 5.10, wenn wir dieses Lemma auf $X - \mathbf{E}[X]$ und $Y - \mathbf{E}[Y]$ anwenden. □

5.3. Die erzeugende Funktion und die Laplace-Transformierte

Während Erwartungswert und Varianz nur zwei Zahlen sind, die die Verteilung einer Zufallsvariablen X beschreiben, geben die erzeugende Funktion (oder die Laplace-Transformierte) gleich eine ganze Funktion an, die die Verteilung von X beschreiben. Wie sich herausstellt, sind diese Funktionen so informativ, dass sie die Verteilung von X eindeutig beschreiben; siehe Proposition 5.22 und Bemerkung 5.23.

Definition 5.20 (Erzeugende Funktion und Laplace-Transformierte)**.** *Sei X eine reellwertige Zufallsvariable. Die Abbildungen*

$$\chi_X : \begin{cases}[0,1] & \to [0,\infty),\\ s & \mapsto \mathbf{E}[s^X]\end{cases}, \qquad \mathcal{L}_X : \begin{cases}[0,\infty) & \to [0,\infty),\\ t & \mapsto \mathbf{E}[e^{-tX}]\end{cases}$$

heißen (Wahrscheinlichkeits-)Erzeugende Funktion und Laplace-Transformierte von X.

Bemerkung 5.21 (Zusammenhang zwischen erzeugenden Funktionen und Laplace-Transformierten)**.** Klar ist, dass

$$\mathcal{L}_X(t) = \chi_X(e^{-t}) \qquad \text{sowie} \qquad \chi_X(s) = \mathcal{L}_X(-\log s),$$

d.h. dass sich beide Funktionen schnell ineinander umrechnen lassen.

Wir beginnen mit einer Aussage im Fall diskreter Zufallsvariablen, die den Zusammenhang zwischen erzeugender Funktion und Verteilung einer Zufallsvariablen beschreibt.

Proposition 5.22 (Berechnung der Verteilung aus der erzeugenden Funktion)**.** *Sei X eine $\mathbb{N}_0$-wertige, diskrete Zufallsvariable mit erzeugender Funktion χ. Dann gilt für $k = 0, 1, 2, ...$*

$$\mathbf{P}(X = k) = \frac{1}{k!}\frac{d^k}{ds^k}\chi(s)\Big|_{s=0}.$$

Insbesondere bestimmt χ die Verteilung von X eindeutig.

Beweis. Wir berechnen, da alle Summanden positiv sind und die Reihe damit absolut konvergiert,

$$\frac{d^k}{ds^k}\chi(s)\Big|_{s=0} = \frac{d^k}{ds^k}\sum_{i=0}^{\infty} s^i \mathbf{P}(X = i)\Big|_{s=0} = \sum_{i=k}^{\infty} \frac{i!}{(i-k)!} s^{i-k}\mathbf{P}(X = i)\Big|_{s=0} = k!\mathbf{P}(X = k).$$

□

Bemerkung 5.23 (Laplace-Transformierte bestimmt die Verteilung eindeutig)**.** Das letzte Resultat lässt sich mit maßtheoretischen Methoden auch auf stetige Zufallsvariable übertragen. Es gilt: Sei X eine $\mathbb{R}$-wertige, kontinuierliche Zufallsvariable mit Laplace-Transformierten $\mathcal{L}$, die zumindest auf $\mathbb{R}_+$ existiert. Dann ist die Dichte f_X von X eindeutig durch $\mathcal{L}$ bestimmt.

Erzeugende Funktionen und Laplace-Transformierte sind äußerst praktisch, um Erwartungswerte und Varianzen zu berechnen.

Proposition 5.24 (Momente und erzeugende Funktionen)**.** *Sei X eine reellwertige, diskrete Zufallsvariable mit erzeugender Funktion χ. Dann gilt*

$$\mathbb{E}[X] = \frac{d}{ds}\chi(s)\Big|_{s=1}, \qquad \mathbb{E}[X(X-1)] = \frac{d^2}{ds^2}\chi(s)\Big|_{s=1},$$

falls die Ableitungen existieren. Außerdem gilt für eine reellwertige Zufallsvariable Y mit Laplace-Transformierter $\mathcal{L}$

$$\mathbb{E}[X] = -\frac{d}{dt}\mathcal{L}(t)\Big|_{t=0}, \qquad \mathbb{E}[X^2] = \frac{d^2}{dt^2}\mathcal{L}(t)\Big|_{t=0}$$

falls die Ableitungen existieren.

Bemerkung 5.25 (Berechnung der Varianz)**.** Für die Varianz einer Zufallsvariable X gilt insbesondere (siehe Lemma 5.15)

$$\mathbf{Var}[X] = \chi_X''(1) + \chi_X'(1) - (\chi_X'(1))^2,$$
$$\mathbf{Var}[X] = \mathcal{L}_X''(0) - (\mathcal{L}_X'(0))^2.$$

Beweis von Proposition 5.24. Sei $S = \{k_0, k_1, ...\}$ der Zielbereich von S. Wir schreiben für die erzeugende Funktion

$$\frac{d}{ds}\chi(s)\Big|_{s=1} = \frac{d}{ds}\sum_{i=0}^{\infty} s^{k_i}\mathbf{P}(X = k_i)\Big|_{s=1} = \sum_{i=0}^{\infty} k_i s^{k_i-1}\mathbf{P}(X = k_i)\Big|_{s=1} = \sum_{i=0}^{\infty} k_i \mathbf{P}(X = k_i) = \mathbf{E}[X],$$
$$\frac{d^2}{ds^2}\chi(s)\Big|_{s=1} = \sum_{i=0}^{\infty} k_i(k_i-1)s^{k_i-2}\mathbf{P}(X = k_i)\Big|_{s=1} = \sum_{i=0}^{\infty} k_i(k_i-1)\mathbf{P}(X = k_i) = \mathbf{E}[X(X-1)].$$

Bei der Laplace-Transformierten erhalten wir für diskrete Zufallsvariable mit Bemerkung 5.21

$$-\frac{d}{dt}\mathcal{L}(t)\Big|_{t=0} = \chi'(1),$$
$$\frac{d^2}{dt^2}\mathcal{L}(t)\Big|_{t=0} = -\frac{d}{dt}\chi'(e^{-t})e^{-t}\Big|_{t=0}\chi''(1) + \chi'(1) = \mathbf{E}[X(X-1)] + \mathbf{E}[X] = \mathbf{E}[X^2].$$

Für Verteilungen mit Dichte erinnern wir an parameterabhängige Integrale aus Analysis 2 und berechnen, falls X die Dichte f besitzt,

$$-\frac{d}{dt}\mathcal{L}(t)\Big|_{t=0} = -\frac{d}{dt}\int e^{-tx}f(x)dx\Big|_{t=0} = -\int \frac{d}{dt}e^{-tx}\Big|_{t=0}f(x)dx = \int xf(x)dx = \mathbf{E}[X],$$
$$\frac{d^2}{dt^2}\mathcal{L}(t)\Big|_{t=0} = -\int \frac{d^2}{dt^2}e^{-tx}\Big|_{t=0}f(x)dx = \int x^2f(x)dx = \mathbf{E}[X^2].$$

□

5.4. Kenngrößen wichtiger Verteilungen

Wir sammeln nun Erwartungswerte, Varianzen und erzeugende Funktionen der uns bekannten diskreten Verteilungen. Einige davon sind bereits aus den Übungen bekannt.

Theorem 5.26 (Kenngrößen wichtiger diskreter Verteilungen)**.** *Für X diskret gilt:*

	$\mathbf{P}(X=k)$	$\mathbf{E}[X]$	$\mathbf{Var}[X]$	$\chi_X(s)$
$X \sim U([1:n])$	$\frac{1}{n}1_{1\le k\le n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{1}{n}\frac{s(1-s^n)}{1-s}$
$X \sim B(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$(1+ps-p)^n$
$X \sim Hyp(n,N,K)$	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$	$n\frac{K}{N}$	$n\frac{K}{N}\frac{N-K}{N}\left(1-\frac{n-1}{N-1}\right)$	$\sum_{k=0}^{n}\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}s^k$
$X \sim Poi(\lambda)$	$e^{-\lambda}\frac{\lambda^k}{k!}$	λ	λ	$e^{-(1-s)\lambda}$
$X \sim geo(p)$	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{ps}{1-(1-p)s}$

Beweis. Für $X \sim U([1:n])$ erinnern wir an die (vermutlich) aus der Analysis 1 bekannten Identitäten

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \qquad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

Wir berechnen

$$\begin{aligned}
\mathbf{E}[X] &= \frac{1}{n}\sum_{k=1}^{n} k = \frac{n(n+1)}{2n} = \frac{n+1}{2},\\
\mathbf{Var}[X] &= \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \frac{1}{n}\Big(\sum_{k=1}^{n} k^2\Big) - \frac{(n+1)^2}{4} = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}\\
&= \frac{(n+1)(4n+2-3n-3)}{12} = \frac{n^2-1}{12},\\
\chi_X(s) &= \frac{1}{n}\sum_{k=1}^{n} s^k = \frac{1}{n}s\sum_{k=0}^{n-1} s^k = \frac{1}{n}\frac{s(1-s^n)}{1-s}.
\end{aligned}$$

Für $X \sim B(n,p)$ berechnen wir zunächst die erzeugende Funktion mit Hilfe des binomischen Lehrsatzes

$$\begin{aligned}
\chi_X(s) &= \mathbf{E}[s^X] = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}s^k = \sum_{k=0}^{n}\binom{n}{k}(ps)^k(1-p)^{n-k} = (1+ps-p)^n,\\
\chi_X'(s) &= np(1+ps-p)^{n-1},\\
\chi_X''(s) &= n(n-1)p^2(1+ps-p)^{n-2}.
\end{aligned}$$

Daraus ergeben sich mittels Proposition 5.24 und Bemerkung 5.25

$$\begin{aligned}
\mathbf{E}[X] &= \chi'(1) = np,\\
\mathbf{Var}[X] &= \chi''(1) + \chi'(1) - (\chi'(1))^2 = n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p).
\end{aligned}$$

Für $X \sim \mathrm{Hyp}(n,N,K)$ verwenden wir $X = Z_1 + \cdots + Z_n$ für eine Urne mit den N Kugeln, wovon K weiß sind und

$$Z_i := 1_{\{i\text{-te gezogene Kugel ist weiß}\}}.$$

Klar ist, dass $\mathbf{P}[Z_i = 1] = \frac{K}{N}$, $i = 1, ..., n$ und damit

$$\mathbf{E}[X] = \sum_{i=1}^{n}\mathbf{E}[Z_i] = n\frac{K}{N}.$$

Weiter gilt für $i \neq j$

$$\mathbf{Cov}[Z_i, Z_j] = \mathbf{P}[Z_i = Z_j = 1] - \Big(\frac{K}{N}\Big)^2 = \frac{K}{N}\Big(\frac{K-1}{N-1} - \frac{K}{N}\Big) = -\frac{K}{N}\frac{N-K}{N(N-1)},$$

also nach Lemma 5.15.4 und $\mathbf{Var}[Z_i] = \frac{K}{N}\frac{N-K}{N}$

$$\begin{aligned}
\mathbf{Var}[X] &= \sum_{i=1}^{n}\mathbf{Var}[Z_i] + 2\sum_{1\le i<j\le n}\mathbf{Cov}[Z_i, Z_j]\\
&= n\frac{K}{N}\frac{N-K}{N} - n(n-1)\frac{K}{N}\frac{N-K}{N(N-1)} = n\frac{K}{N}\frac{N-K}{N}\Big(1 - \frac{n-1}{N-1}\Big).
\end{aligned}$$

Für $X \sim \text{Poi}(\lambda)$ ist

$$\begin{aligned}\chi_X(s) &= \mathbf{E}[s^X] = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}s^k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(s\lambda)^k}{k!} = e^{-(1-s)\lambda},\\ \chi_X'(s) &= \lambda e^{-(1-s)\lambda},\\ \chi_X''(s) &= \lambda^2 e^{-(1-s)\lambda}.\end{aligned}$$

Daraus ergeben sich mittels Proposition 5.24 und Bemerkung 5.25

$$\begin{aligned}\mathbf{E}[X] &= \chi'(1) = \lambda,\\ \mathbf{Var}[X] &= \chi''(1) + \chi'(1) - (\chi'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.\end{aligned}$$

Für $X \sim \text{geo}(p)$ ist

$$\begin{aligned}\chi_X(s) &= \mathbf{E}[s^X] = p\sum_{k=1}^{\infty}(1-p)^{k-1}s^k = ps\sum_{k=0}^{\infty}((1-p)s)^k = \frac{ps}{1-(1-p)s},\\ \chi_X'(s) &= \frac{p(1-(1-p)s) + sp(1-p)}{(1-(1-p)s)^2} = \frac{p}{(1-(1-p)s)^2},\\ \chi_X''(s) &= \frac{2p(1-p)}{(1-(1-p)s)^3}.\end{aligned}$$

Daraus ergeben sich mittels Proposition 5.24 und Bemerkung 5.25

$$\begin{aligned}\mathbf{E}[X] &= \chi'(1) = \frac{1}{p},\\ \mathbf{Var}[X] &= \chi''(1) + \chi'(1) - (\chi'(1))^2 = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{(1-p)}{p^2}.\end{aligned}$$

□

Theorem 5.27 (Kenngrößen wichtiger kontinuierlicher Verteilungen)**.** *Für kontinuierliche Zufallsvariablen X gilt:*

	$f_X(x)$	$\mathbf{E}[X]$	$\mathbf{Var}[X]$	$\mathcal{L}_X(t)$
$X \sim U([a,b])$	$\frac{1}{b-a}1_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{-ta}-e^{-tb}}{t(b-a)}$
$X \sim Exp(\lambda)$	$\lambda e^{-\lambda x}1_{x\geq 0}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda+t}$
$X \sim N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$e^{\frac{t^2}{2}\sigma^2 - t\mu}$

Beweis. Für $X \sim U([a,b])$ ist

$$\begin{aligned}
\mathbf{E}[X] &= \frac{1}{b-1}\int_a^b x dx = \frac{1}{b-a}\frac{1}{2}(b^2-a^2) = \frac{b+a}{2},\\
\mathbf{Var}[X] &= \frac{1}{b-a}\int_a^b x^2 dx - \Big(\frac{b+a}{2}\Big)^2 = \frac{a^2+ab+b^2}{3} - \frac{(b+a)^2}{4}\\
&= \frac{4a^2+4ab+4b^2-3b^2-6ab-3a^2}{12} = \frac{(b-a)^2}{12},\\
\mathcal{L}_X(t) &= \mathbf{E}[e^{-tX}] = \frac{1}{b-a}\int_a^b e^{-tx}dx = \frac{e^{-tb}-e^{-ta}}{t(b-a)}.
\end{aligned}$$

Für $X \sim \mathrm{Exp}(\lambda)$ ist

$$\begin{aligned}
\mathcal{L}_X(t) &= \mathbf{E}[e^{-tX}] = \int_0^\infty \lambda e^{-\lambda x}e^{-tx}dx = \lambda\int_0^\infty e^{-(\lambda+t)x}dx = \frac{\lambda}{\lambda+t},\\
\mathcal{L}'_X(t) &= -\frac{\lambda}{(\lambda+t)^2},\\
\mathcal{L}''_X(t) &= \frac{2\lambda}{(\lambda+t)^3}.
\end{aligned}$$

Daraus ergeben sich mittels Proposition 5.24 und Bemerkung 5.25

$$\begin{aligned}
\mathbf{E}[X] &= -\mathcal{L}'_X(0) = \frac{1}{\lambda},\\
\mathbf{Var}[X] &= \mathcal{L}''_X(0) - (\mathcal{L}'_X(0))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.
\end{aligned}$$

Sei zunächst $Z \sim N(0,1)$. Dann ist

$$\mathcal{L}_Z(t) = \frac{1}{\sqrt{2\pi}}\int e^{-\frac{z^2}{2}}e^{-tz}dz = e^{\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int e^{-\frac{1}{2}(z^2-2tz+t^2)}dz = e^{\frac{t^2}{2}}.$$

Damit ist wegen $X := \mu + \sigma Z \sim N(\mu,\sigma^2)$

$$\begin{aligned}
\mathcal{L}_X(t) &= \mathbf{E}[e^{-t(\mu+\sigma Z)}] = e^{\frac{t^2\sigma^2}{2}-t\mu},\\
\mathcal{L}'_X(t) &= (t\sigma^2-\mu)e^{\frac{t^2\sigma^2}{2}-t\mu},\\
\mathcal{L}''_X(t) &= \big((t\sigma^2-\mu)^2+\sigma^2\big)e^{\frac{t^2\sigma^2}{2}-t\mu}.
\end{aligned}$$

Wieder ergeben sich mittels Proposition 5.24 und Bemerkung 5.25

$$\begin{aligned}
\mathbf{E}[X] &= -\mathcal{L}'_X(0) = \mu,\\
\mathbf{Var}[X] &= \mathcal{L}''_X(0) - (\mathcal{L}'_X(0))^2 = (\mu^2+\sigma^2) - \mu^2 = \sigma^2.
\end{aligned}$$

□

6. Approximationssätze

Wie üblich ist ein Approximationssatz ein Grenzwertresultat. Da wir es in der Stochastik mit Zufallsvariablen zu tun haben, werden wir also Grenzwertsätze über Zufallsvariablen (was ja spezielle Funktionen sind) betrachten. Speziell ist hier, dass wir nicht nur über die Konvergenz dieser Funktionen reden können (wie etwa im Gesetz der Großen Zahlen, Theorem 6.6), sondern auch über die Konvergenz ihrer Verteilungen (wie etwa im Zentralen Grenzwertsatz, Theorem 6.8).

6.1. Summen unabhängiger Zufallsvariablen

Wir beschäftigen uns nun mit (der Verteilung von)

$$S := X_1 + \cdots + X_n,$$

wobei $X_1, ..., X_n$ unabhängige (und oft auch identisch verteilte) Zufallsvariable sind. Wir betrachten zunächst zwei spezielle Fälle normalverteilter und Poisson-verteilter Zufallsvariable.

Proposition 6.1 (Summen unabhängiger normal- und Poisson-verteiler Zufallsvariablen)**.**

1. *Sind $X_1, ..., X_n$ unabhängig und $X_i \sim Poi(\lambda_i)$, $i = 1, ..., n$. Dann ist $S = X_1 + \cdots + X_n \sim Poi\Big(\sum_{i=1}^n \lambda_i\Big)$.*
2. *Sind $X_1, ..., X_n$ unabhängig und $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, ..., n$. Dann ist $S = X_1 + \cdots + X_n \sim N\Big(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\Big)$.*

Beweis. 1. Da $X_1, ..., X_n$ unabhängig sind, berechnen wir die erzeugende Funktion von S, nämlich

$$\chi_S(s) = \mathbf{E}[s^{X_1+\cdots+X_n}] = \mathbf{E}[s^{X_1}] \cdots \mathbf{E}[s^{X_n}] = e^{-(1-s)\lambda_1} \cdots e^{-(1-s)\lambda_n} = e^{-(1-s)(\lambda_1+\cdots+\lambda_n)},$$

und da die erzeugende Funktion die Verteilung von S eindeutig festlegt, ist $S \sim \text{Poi}\Big(\sum_{i=1}^n \lambda_i\Big)$.
2. Auch im Falle normalverteilter Zufallsvariablen berechnen wir mit Hilfe der Laplace-Transformierten

$$\mathcal{L}_S(t) = \mathbf{E}[e^{-tX_1}] \cdots \mathbf{E}[e^{-tX_n}] = e^{\frac{t^2}{2}\sigma_1^2 - t\mu_1} \cdots e^{\frac{t^2}{2}\sigma_n^2 - t\mu_n} = e^{\frac{t^2}{2}(\sigma_1^2+\cdots+\sigma_n^2) - t(\mu_1+\cdots+\mu_n)},$$

wobei die rechte Seite die Laplace-Transformierte einer $N\Big(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\Big)$-Verteilung ist. Da diese die Verteilung eindeutig bestimmt (siehe Bemerkung 5.23) folgt die Behauptung. □

Definition 6.2 (Faltung)**.** *Seien X, Y unabhängige Zufallsvariablen. Dann heißt die Verteilung von $X + Y$ die Faltung der Verteilungen von X und Y.*

Proposition 6.3 (Faltung diskreter und kontinuierlicher Zufallsvariable)**.** *Seien X, Y unabhängige Zufallsvariablen.*

1. *Sind X, Y diskret mit Zielbereich $\mathbb{Z}$, so ist die Faltung von X, Y gegeben durch*

$$\mathbf{P}(X+Y=k) = \sum_{\ell=-\infty}^{\infty} \mathbf{P}(X=\ell)\mathbf{P}(Y=k-\ell).$$

   Sind χ_X und χ_Y die erzeugenden Funktionen von X und Y, so ist $\chi_X\chi_Y$ die erzeugende Funktion von $X+Y$.

2. *Sind X, Y kontinuierlich mit Zielbereich $\mathbb{R}$ und Dichten f_X und f_Y, so besitzt die Faltung X, Y ebenfalls eine Dichte, nämlich*

$$f_{X+Y}(z) = \int f_X(x) f_Y(z-x)dx.$$

   Sind $\mathcal{L}_X$ und $\mathcal{L}_Y$ die Laplace-Transformierten von X und Y, so ist $\mathcal{L}_X\mathcal{L}_Y$ die Laplace-Transformierte von $X+Y$.

Beweis. Die Formeln für die Zähldichte und Dichte der Faltung sind klar. Die Multiplikativität von erzeugender Funktion und Laplace-Transformierter haben wir bereits im Beweis von Proposition 6.1 gesehen und verwendet. □

Bemerkung 6.4 ($\sqrt{n}$-Gesetz)**.** Sind $X_1, ..., X_n$ identisch verteilt und $S = X_1 + \cdots + X_n$, so ist (siehe Proposition 5.5)

$$\mathbf{E}[S] = n\mathbf{E}[X_1].$$

Sind weiterhin $X_1, ..., X_n$ unabhängig, so gilt (siehe Lemma 5.15)

$$\mathbf{Var}[S] = n\mathbf{Var}[X_1].$$

Da die Varianz die mittlere *quadratische* Abweichung misst, typische Abweichungen von S um den Erwartungswert also im Bereich $\sqrt{\mathbf{Var}[S]}$ liegen sollten, kann man bereits folgendes für große n ablesen: eine Zufallsvariable, die die Summe von n unabhängigen Zufallsvariablen ist, streut ihre Werte typischerweise in einem Bereich der Größenordnung $\sqrt{n}$ um ihren Erwartungswert. Dies werden wir mit dem Zentralen Grenzwertsatz, Theorem 6.8 (der auf dem Gesetz der großen Zahlen, Theorem Theorem 6.6 aufbaut), präzisieren.

6.2. Das schwache Gesetz der großen Zahlen

Bevor wir zum Hauptresultat dieses Abschnittes kommen, benötigen wir zwei Ungleichungen.

Proposition 6.5 (Die Markov- und Chebyshev-Ungleichung)**.**

1. *Markov-Ungleichung: Sei X eine $\mathbb{R}_+$-wertige Zufallsvariable. Dann gilt für alle $\varepsilon > 0$*

$$\mathbf{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbf{E}[X].$$

2. *Chebyshev-Ungleichung: Sei X eine $\mathbb{R}$-wertige Zufallsvariable mit endlichem Erwartungswert. Dann gilt für alle $\varepsilon > 0$*

$$\mathbf{P}(|X - \mathbf{E}[X]| \geq \varepsilon) \leq \frac{\mathbf{Var}[X]}{\varepsilon^2}.$$

Beweis. 1. Es gilt $X \geq \varepsilon \cdot 1_{X \geq \varepsilon}$ und damit

$$\varepsilon \cdot \mathbf{P}(X \geq \varepsilon) = \mathbf{E}[\varepsilon \cdot 1_{X \geq \varepsilon}] \leq \mathbf{E}[X].$$

2. Wir wenden die Markov-Ungleichung auf die Zufallsvariable $(X - \mathbf{E}[X])^2$ an. Dann gilt

$$\mathbf{P}(|X - \mathbf{E}[X]| \geq \varepsilon) = \mathbf{P}((X - \mathbf{E}[X])^2 \geq \varepsilon^2) \leq \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{\varepsilon^2}$$

und die Behauptung folgt. □

Theorem 6.6 (Schwaches Gesetz der großen Zahlen (von Jacob Bernoulli))**.** *Seien $X_1, X_2, ...$ unabhängige und identisch verteilte, $\mathbb{R}$-wertige Zufallsvariable mit endlichem Erwartungswert und endlicher Varianz. Dann gilt für alle $\varepsilon > 0$*

$$\lim_{n \to \infty} \mathbf{P}\Big(\Big|\frac{X_1 + \cdots X_n}{n} - \mathbf{E}[X_1]\Big| \geq \varepsilon\Big) = 0. \tag{6.1}$$

Beweis. Es gilt

$$\mathbf{E}\Big[\frac{X_1 + \cdots + X_n}{n}\Big] = \frac{\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]}{n} = \mathbf{E}[X_1],$$
$$\mathbf{Var}\Big[\frac{X_1 + \cdots + X_n}{n}\Big] = \frac{\mathbf{Var}[X_1] + \cdots + \mathbf{Var}[X_n]}{n^2} = \frac{\mathbf{Var}[X_1]}{n}.$$

Wendet man nun die Chebyshev-Ungleichung auf $(X_1 + \cdots + X_n)/n$ an, so folgt die Behauptung aus

$$\mathbf{P}\Big(\Big|\frac{X_1 + \cdots X_n}{n} - \mathbf{E}[X_1]\Big| \geq \varepsilon\Big) \leq \frac{\mathbf{Var}[X_1]}{\varepsilon^2 n} \xrightarrow{n \to \infty} 0.$$

□

Bemerkung 6.7 (Starkes Gesetz der großen Zahlen)**.** Es gilt auch folgende, stärkere Version des Gesetzes der großen Zahlen, die wir ohne Beweis angeben:
Seien $X_1, X_2, ...$ unabhängige und identisch verteilte, $\mathbb{R}$-wertige Zufallsvariable mit endlichem Erwartungswert. Dann gilt

$$\mathbf{P}\Big(\lim_{n \to \infty} \frac{X_1 + \cdots + X_n}{n} = \mathbf{E}[X_1]\Big) = 1. \tag{6.2}$$

Auf zwei Dinge möchten wir jedoch hinweisen:

1. Im starken Gesetz der großen Zahlen kommt das Ereignis $\{\lim_{n \to \infty} \frac{X_1 + \cdots + X_n}{n} = \mathbf{E}[X_1]\}$ vor. Um diesem Ereignis sinnvoll eine Wahrscheinlichkeit zuordnen zu können, bedarf es der Maßtheorie.

2. Die Tatsache, dass das starke Gesetz wirklich stärker als das schwache Gesetz ist, liegt daran, dass (6.1) aus (6.2) folgt. Insbesondere gilt das schwache Gesetz der großen Zahlen auch dann, wenn die Varianz von X_1 unendlich ist. (Dies ist ja im oben formulierten schwachen Gesetz gefordert, im starken Gesetz jedoch nicht.)

6.3. Der zentrale Grenzwertsatz

Der Zentrale Grenzwertsatz gibt die Fluktuationen im Gesetz der großen Zahlen (im Falle endlicher Varianzen) an.

Theorem 6.8 (Zentraler Grenzwertsatz)**.** *Seien $X_1, X_2, ...$ unabhängige und identisch verteilte, $\mathbb{R}$-wertige Zufallsvariable mit endlichem Erwartungswert μ und endlicher Varianz σ^2. Dann gilt für alle $-\infty \leq c < d \leq \infty$*

$$\lim_{n\to\infty} \mathbf{P}\Big(c \leq \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} \leq d\Big) = \mathbf{P}(c \leq Z \leq d),$$

wobei $Z \sim N(0,1)$.

Bemerkung 6.9 (Einfache Fälle)**.** 1. Klar ist, dass

$$\mathbf{E}\Big[\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}\Big] = 0,$$
$$\mathbf{Var}\Big[\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}\Big] = 1.$$

Deshalb stimmen schon mal Erwartungswert und Varianz von $\frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}}$ und Z überein. Der Zentrale Grenzwertsatz erweitert dies auf die ganze Verteilung.

2. Generell spricht man für eine Zufallsvariable X von ihrer Standardisierung $\frac{X-\mathbf{E}[X]}{\sqrt{\mathbf{Var}[X]}}$.

3. Klar ist der Zentrale Grenzwertsatz in dem Fall, wenn $X_1, X_2, ... \sim N(\mu, \sigma^2)$. Dann nämlich ist $X_1 + \cdots + X_n \sim N(n\mu, n\sigma^2)$, also

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

und der Satz gilt sogar ohne die Grenzwertbildung.

4. Sei $X_1, X_2, ... \sim B(1,p)$, also $S_n := X_1 + \cdots + X_n \sim B(n,p)$. Dann gilt nach dem Zentralen Grenzwertsatz für $c < d$

$$\mathbf{P}\Big(c \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq d\Big) \xrightarrow{n\to\infty} \mathbf{P}(c \leq Z \leq d).$$

Dies kann man bereits anhand der Laplace-Transformierten erahnen. Es gilt nämlich mit $S_n^* := \frac{S_n - np}{\sqrt{np(1-p)}}$

$$\begin{aligned}
\log \mathcal{L}_{S_n^*}(t) &= \log \mathbf{E}[e^{-tS_n^*}] = \log \mathbf{E}[e^{t(np-S_n)/\sqrt{np(1-p)}}] \\
&= t\sqrt{np/(1-p)} + \log \mathbf{E}[e^{-\frac{t}{\sqrt{np(1-p)}}S_n}] = t\sqrt{np/(1-p)} + \log \chi_{S_n}(e^{-\frac{t}{\sqrt{np(1-p)}}}) \\
&= t\sqrt{np/(1-p)} + n\log(1 - p(1 - e^{-\frac{t}{\sqrt{np(1-p)}}})) \\
&\approx t\sqrt{np/(1-p)} - np(1 - e^{-\frac{t}{\sqrt{np(1-p)}}}) - np^2(1 - e^{-\frac{t}{\sqrt{np(1-p)}}})^2/2 \\
&\approx \frac{t^2}{2(1-p)} - \frac{t^2 p}{2(1-p)} = \frac{t^2}{2} = \log \mathcal{L}_Z(t).
\end{aligned}$$

In diesem Fall (wenn $X_i \sim B(1,p)$) heißt der zentrale Grenzwertsatz auch Satz von deMoivre-Laplace.

5. Eine Anwendung des Satzes von deMoivre Laplace kann etwa so aussehen: Sei X eine $B(n,p)$ verteilte Zufallsvariable mit n groß, npq groß, sowie $c, d \in \mathbb{Z}$. Dann gilt approximativ

$$\begin{aligned}\mathbf{P}(c \le X \le d) &\approx \sum_{k=c}^{d} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z_{k,n}^2}{2}\Big)(z_{k+\frac{1}{2},n} - z_{k-\frac{1}{2},n}) \\ &\approx \frac{1}{\sqrt{2\pi}} \int_{z_{c-1/2,n}}^{z_{d+1/2,n}} \exp\Big(-\frac{z^2}{2}\Big)dz = \Phi\Big(\frac{d+\frac{1}{2}-np}{\sqrt{npq}}\Big) - \Phi\Big(\frac{c-\frac{1}{2}-np}{\sqrt{npq}}\Big).\end{aligned}$$

Bevor wir zum Beweis des Zentralen Grenzwertsatzes kommen, benötigen wir ein Lemma.

Lemma 6.10. *Seien $X_1, X_2, ...$ wie im Zentralen Grenzwertsatz mit $\mu = 0$ und $\sigma^2 = 1$. Weiter sei $h : \mathbb{R} \to \mathbb{R}$ eine 3-mal stetig differenzierbare Funktion mit beschränkten ersten drei Ableitungen und $Z \sim N(0,1)$ verteilt. Dann gilt*

$$\mathbf{E}\Big[h\Big(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\Big)\Big] \xrightarrow{n\to\infty} \mathbf{E}[h(Z)].$$

Beweis. Seien $Z_1, Z_2, ... \sim N(0,1)$ unabhängig und unabhängig von $X_1, X_2, ...$ Wir schreiben

$$U_i := \frac{X_1 + \cdots + X_{i-1} + Z_{i+1} + \cdots + Z_n}{\sqrt{n}}$$

und

$$\begin{aligned}h\Big(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\Big) - h\Big(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}}\Big) \\ = \sum_{i=1}^{n} h\Big(U_i + \frac{X_i}{\sqrt{n}}\Big) - h\Big(U_i + \frac{Z_i}{\sqrt{n}}\Big)\end{aligned} \tag{6.3}$$

Weiter ist mit der Taylor-Formel

$$\begin{aligned}&h\Big(U_i + \frac{X_i}{\sqrt{n}}\Big) - h\Big(U_i + \frac{Z_i}{\sqrt{n}}\Big) \\ &= h'(U_i)\frac{X_i - Z_i}{\sqrt{n}} + h''(U_i)\frac{X_i^2 - Z_i^2}{2n} \\ &\quad + \underbrace{\big(h''(V_i) - h''(U_i)\big)\frac{X_i^2}{2n} + \big(h''(U_i) - h''(W_i)\big)\frac{Z_i^2}{2n}}_{=:R_{in}}\end{aligned}$$

mit Zufallsvariablen V_i, W_i, so dass

$$|V_i - U_i| \le \frac{|X_i|}{\sqrt{n}}, \qquad |W_i - U_i| \le \frac{|Z_i|}{\sqrt{n}}.$$

Weiter ist

$$\begin{aligned}|R_{in}| &= |R_{in}|1_{|X_i|\le k} + |R_{in}|1_{|X_i|>k} \\ &\le c'''\frac{k^3}{n^{3/2}}1_{|X_i|\le k} + c''\frac{X_i^2}{n}1_{|X_i|>k} + c'''\frac{|Z_i|^3}{n^{3/2}}\end{aligned}$$

mit $c'' = \sup_{x\in\mathbb{R}} h''(x), c'' := \sup_{x\in\mathbb{R}} h'''(x)$ und beliebigem k. Da U_i, X_i, Z_i unabhängig sind, folgt

$$\begin{aligned}
&\Big|\mathbf{E}\Big[h\Big(U_i + \frac{X_i}{\sqrt{n}}\Big) - h\Big(U_i + \frac{Z_i}{\sqrt{n}}\Big)\Big]\Big| \\
&\quad = \Big|\mathbf{E}\Big[h'(U_i)\frac{X_i - Z_i}{\sqrt{n}}\Big] + \mathbf{E}\Big[h''(U_i)\frac{X_i^2 - Z_i^2}{2n}\Big] + \mathbf{E}[R_{in}]\Big| \\
&\quad \le \mathbf{E}[|R_{in}|] \le c'''\frac{k^3 + \mathbf{E}[|Z_i|^3]}{n^{3/2}} + c''\frac{\mathbf{E}[X_1^2 1_{|X_1|>k}]}{n}
\end{aligned}$$

Wegen (6.3) und da $\frac{Z_1+\cdots+Z_n}{\sqrt{n}} \sim Z$, ist nun

$$\begin{aligned}
\Big|h\Big(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\Big) - h(Z)\Big| &\le c'''\frac{k^3 + \mathbf{E}[|Z_i|^3]}{n^{1/2}} + c''\mathbf{E}[X_1^2 1_{|X_1|>k}] \\
&\xrightarrow{n\to\infty} c''\mathbf{E}[X_1^2 1_{|X_1|>k}] \xrightarrow{k\to\infty} 0.
\end{aligned}$$

□

Beweis von Theorem 6.8. OBdA ist $\mu = 0$ und $\sigma^2 = 1$, ansonsten gehen wir zu $\frac{X_i-\mu}{\sqrt{\sigma^2}}$ über. Wir setzen

$$Y_n^* := \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

Wähle h_1, h_2 dreimal stetig differenzierbar mit beschränkten ersten drei Ableitungen und für $\varepsilon > 0$

$$1_{c+\varepsilon\le x\le d-\varepsilon} \le h_1(x) \le 1_{c\le x\le d} \le h_2(x) \le 1_{c-\varepsilon\le x\le d+\varepsilon}.$$

Dann gilt

$$\begin{aligned}
\mathbf{P}(c+\varepsilon \le Z \le d-\varepsilon) &\le \mathbf{E}[h_1(Z)] = \liminf_{n\to\infty} \mathbf{E}[h_1(Y_n^*)] \\
&\le \liminf_{n\to\infty} \mathbf{P}(c \le Y_n^* \le d) \le \limsup_{n\to\infty} \mathbf{P}(c \le Y_n^* \le d) \\
&\le \limsup_{n\to\infty} \mathbf{E}[h_2(Y_n^*)] = \mathbf{E}[h_2(Z)] \le \mathbf{P}(c-\varepsilon \le Z \le d+\varepsilon).
\end{aligned}$$

Die Behauptung folgt nun, da $|\mathbf{P}(c \pm \varepsilon \le Z \le d \pm \varepsilon) - \mathbf{P}(c \le Z \le d)| \xrightarrow{\varepsilon\to 0} 0$. □

Ähnlich wie im Zentralen Grenzwertsatz kann man auch eine Approximation von großen Fakultäten erhalten.

Korollar 6.11 (Stirling Formel)**.** *Es gilt*

$$\lim_{n\to\infty} \frac{n!}{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}} = 1.$$

Beweis. Sei $t^+ := \max(t, 0)$. Für $\varepsilon > 0$ sei h eine dreimal stetig differenzierbare Funktion mit beschränkten Ableitungen und

$$t^+ \le h(t) \le t^+ + \varepsilon.$$

Seien $X_1, X_2, ... \sim \text{Poi}(1)$ unabhängig. Nach Proposition 6.1 ist damit $X_1 + \cdots + X_n \sim \text{Poi}(n)$. Sei weiter $Z \sim N(0,1)$. Mit Hilfe von Lemma 6.10 folgt (mit einem ähnlichen Approximationsargument wie im Beweis des Zentralen Grenzwertsatzes)

$$\mathbf{E}\Big[\Big(\frac{X_1 + \cdots + X_n - n}{\sqrt{n}}\Big)^+\Big] \xrightarrow{n\to\infty} \mathbf{E}[Z^+] = \frac{1}{\sqrt{2\pi}} \int_0^\infty z e^{-z^2/2} dz \quad = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}|_0^\infty = \frac{1}{\sqrt{2\pi}} \tag{6.4}$$

Andererseits ist auch

$$\begin{aligned}
&\mathbf{E}\Big[\Big(\frac{X_1 + \cdots + X_n - n}{\sqrt{n}}\Big)^+\Big] \\
&\quad = \frac{1}{\sqrt{n}} e^{-n} \sum_{k=n+1}^{\infty} (k-n) \frac{n^k}{k!} \\
&\quad = \frac{e^{-n}}{\sqrt{n}} \sum_{k=n+1}^{\infty} \Big(\frac{n^k}{(k-1)!} - \frac{n^{k+1}}{k!}\Big) = \frac{e^{-n}}{\sqrt{n}} \frac{n^{n+1}}{n!}.
\end{aligned}$$

Zusammen mit (6.4) folgt die Behauptung. □

Bemerkung 6.12 (Ein einfacher Beweis)**.** Man kann die Stirling-Formel auch fast mit ganz einfachen Methoden beweisen, nämlich

$$\log(n!) = \sum_{k=1}^{n} \log k = \int_1^n \log x + o(n) = x \log x - x|_1^n + o(n) = n \log n - n + o(n).$$

Allerdings zeigt dies nur, dass

$$\frac{\log(n!)}{\log\big(\big(\frac{n}{e}\big)^n\big)} \xrightarrow{n\to\infty} 1,$$

jedoch nicht $\log(n!) - \log\big(\big(\frac{n}{e}\big)^n\big) \xrightarrow{n\to\infty} 0$ (was nach Korollar 6.11 auch gar nicht stimmt).

7. Abhängige Zufallsvariable

Stochastisch unabhängige Zufallsvariablen haben wir bereits in Abschnitt 3.4 kennen gelernt. Nun geht es um das Gegenteil, nämlich (voneinander) abhängige Zufallsvariablen.

7.1. Bedingte Wahrscheinlichkeiten

Zufallsvariablen können auf viele verschiedene Arten voneinander abhängig sein. Um diese Vielfalt in einer uns bekannten Form beschreiben zu können, benötigen wir bedingte Wahrscheinlichkeiten.

Definition 7.1 (Bedingte Wahrscheinlichkeit und bedingte Verteilung)**.** *Seien X und Y Zufallsvariable mit Zielbereichen S und T.*

1. *Die* bedingte Wahrscheinlichkeit von $\{Y \in B\}$, gegeben $\{X \in A\}$ *ist gegeben durch*

$$\mathbf{P}(Y \in B | X \in A) := \frac{\mathbf{P}(X \in A, Y \in B)}{\mathbf{P}(X \in A)}.$$

   Ist $\mathbf{P}(X \in A) = 0$, so definieren wir die rechte Seite als 0.

2. *Die Abbildung $B \mapsto \mathbf{P}(Y \in B | X \in A)$ heißt die* bedingte Verteilung von Y, gegeben $\{X \in A\}$.

3. *Ist X diskret, so ist die bedingte Verteilung von Y gegeben X die Abbildung*

$$\mathbf{P}(Y \in B | X) : \begin{cases} S & \to [0;1] \\ x & \mapsto \mathbf{P}(Y \in B | X = x). \end{cases}$$

4. *Haben X und Y die gemeinsame Dichte f, so ist*

$$\mathbf{P}(Y \in B | X) : \begin{cases} S & \to [0;1] \\ x & \mapsto \dfrac{\displaystyle\int_B f(x,y)dy}{\displaystyle\int f(x,y)dy}. \end{cases}$$

   Die Abbildung

$$y \mapsto f_Y(y | X = x) := \frac{f(x,y)}{\displaystyle\int f(x,y)dy}$$

   heißt auch bedingte Dichte von Y unter X.

Bemerkung 7.2 (Rechenregeln für bedingte Wahrscheinlichkeiten)**.** Mit bedingten Verteilungen kann man genauso rechnen wie gewohnt, z.B. in Falle von diskreten Verteilungen

$$\mathbf{P}(Y \in B|X \in A) = \sum_{y \in B} \mathbf{P}(Y = y|X \in A).$$

Lemma 7.3 (Bedingte Wahrscheinlichkeiten und Unabhängigkeit)**.** *Die Zufallsvariablen X und Y sind genau dann unabhängig, wenn*

$$\mathbf{P}(Y \in B|X \in A) = \mathbf{P}(Y \in B)$$

für alle (messbaren) A, B gilt. Dies ist dies genau dann der Fall, wenn

$$\mathbf{P}(Y \in B|X) = \mathbf{P}(Y \in B)$$

für alle B gilt.

Beweis. Alle Aussagen folgen direkt aus der Definition der bedingten Wahrscheinlichkeit. □

Beispiel 7.4 (Gedächtnislosigkeit der geometrischen Verteilung und der Exponentialverteilung)**.**

1. Sei $X \sim \text{geo}(p)$. Dann gilt für $i, j = 0, 1, 2, ...$

$$\mathbf{P}(X > i + j|X > i) = \mathbf{P}(X > j).$$

   Die Interpretation ist die folgende: wenn i Versuche erfolglos verliefen, ist die Wahrscheinlichkeit für mindestens weitere j erfolglose Versuche genauso groß wie am Anfang. Deshalb spricht man auch von der Gedächtnislosigkeit der geometrischen Verteilung. Der Beweis ist ganz einfach:

$$\begin{aligned}\mathbf{P}(X > i + j|X > i) &= \frac{\mathbf{P}(X > i + j, X > i)}{\mathbf{P}(X > i)} = \frac{(1-p)^{i+j}}{(1-p)^i} = (1-p)^j \\ &= \mathbf{P}(X > j).\end{aligned}$$

2. Analog ist im Fall $X \sim \text{Exp}(\lambda)$ für $s, t \geq 0$

$$\begin{aligned}\mathbf{P}(X > s + t|X > s) &= \frac{\mathbf{P}(X > s + t, X > s)}{\mathbf{P}(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} \\ &= \mathbf{P}(X > t).\end{aligned}$$

   Also ist auch die Exponentialverteilung gedächtnislos.

Beispiel 7.5 (Summen zweier Zufallsvariablen)**.** Seien X, Y zwei unabhängige Zufallsvariable mit Zielbereich $\mathbb{Z}$. Es gilt

$$\mathbf{P}(X + Y \in A|X = x) = \frac{\mathbf{P}(X + Y \in A, X = x)}{\mathbf{P}(X = x)} = \mathbf{P}(x + Y \in A_2),$$

d.h. die bedingte Verteilung von $X+Y$, gegeben $X = x$ ist dieselbe wie die der Zufallsvariable $Y + x$. Andersherum berechnen wir

$$\mathbf{P}(X = x|X + Y = z) = \frac{\mathbf{P}(X = x) \cdot \mathbf{P}(Y = z - x)}{\mathbf{P}(X + Y = z)}.$$

Ist etwa $X, Y \sim \text{geo}(p)$ unabhängig, so ist

$$\mathbf{P}(X+Y=z) = (z-1)p^2(1-p)^{z-2},$$

da genau im z-ten Versuch der zweite Erfolg stattfinden muss. Damit ist

$$\mathbf{P}(X=x|X+Y=z) = \frac{(1-p)^{x-1}p(1-p)^{z-x-1}p}{(z-1)p^2(1-p)^{z-2}} = \frac{1}{z-1}.$$

Also ist X gegeben $X+Y$ gerade $U([1:(X+Y-1)])$-verteilt.

Beispiel 7.6 (Summen zweier Exponentialverteilungen)**.** Sei $X, Y \sim \text{Exp}(\lambda)$. Analog zum letzten Beispiel betrachten wir die gemeinsame Verteilung von X und $X+Y$, also die Verteilung mit Dichte

$$f(x,z) = \lambda^2 e^{-\lambda x} e^{-\lambda(z-x)} 1_{0\le x\le z} = \lambda^2 e^{-\lambda z} 1_{0\le x\le z}.$$

Damit ist die bedingte Dichte von X gegeben $X+Y$

$$\frac{f(x,z)}{\int f(x,z)dx} = \frac{\lambda^2 e^{-\lambda z} 1_{0\le x\le z}}{\int \lambda^2 e^{-\lambda z} 1_{0\le x\le z} dx} = \frac{1}{z} 1_{0\le x\le z},$$

also ist die bedingte Verteilung uniform auf $[0, X+Y]$.

Proposition 7.7 (Formel für die totale Wahrscheinlichkeit)**.** *Seien X, Y Zufallsvariable mit Zielbereichen S und T.*

1. *Sind X, Y diskret, so gilt*

$$\mathbf{P}(Y \in B) = \sum_{x\in S} \mathbf{P}(Y \in B|X=x) \cdot \mathbf{P}(X=x).$$

2. *Haben X, Y die gemeinsame Dichte f, so gilt*

$$f_Y(y) = \int f_Y(y|X=x) \cdot f_X(x)dx.$$

Beweis. 1. Die Formel ergibt sich durch Einsetzen der Definition von $\mathbf{P}(Y \in B|X=x)$ in

$$\mathbf{P}(Y \in B) = \sum_{x\in S} \mathbf{P}(Y \in B, X=x) = \sum_{x\in S} \mathbf{P}(Y \in B|X=x) \cdot \mathbf{P}(X=x).$$

2. Ebenfalls sieht man hier direkt

$$f_Y(y) = \int_S f(x,y)dx = \int_S \frac{f(x,y)}{\int f(x,z)dz} \Big(\int f(x,z)dz \Big) dx = \int f(y|X=x) \cdot f_X(x)dx.$$

□

Beispiel 7.8 (Münzwurf mit zufälliger Erfolgswahrscheinlichkeit)**.** Sei $U \sim U([0,1])$. Gegeben $\{U=u\}$ sei $Z=(Z_1,...,Z_n)$ ein u-Münzwurf und $X = Z_1 + \cdots + Z_n$ die Anzahl der Erfolge. Damit hat X gegeben $\{U=u\}$ eine $B(n,u)$-Verteilung. Wir berechnen nun

$$\mathbf{P}(X=k) = \frac{1}{n+1}, \qquad k=0,...,n.$$

Denn für $0 < k < n$ ist

$$\begin{aligned}\mathbf{P}(X=k) &= \int \mathbf{P}(X=k|U=u) f_U(u) du = \int \underbrace{\binom{n}{k}}_{=f'} u^k \underbrace{(1-u)^{n-k}}_{=g} du \\ &= \int_0^1 \frac{n-k}{k+1}\binom{n}{k} u^{k+1}(1-u)^{n-k-1} du \\ &= \int_0^1 \binom{n}{k+1} u^{k+1}(1-u)^{n-(k+1)} du = \mathbf{P}(X=k+1).\end{aligned}$$

Damit ist die Behauptung gezeigt.

Theorem 7.9 (Formel von Bayes)**.** *Seien X, Y diskrete Zufallsvariable und A so, dass $\mathbf{P}(X \in A) > 0$. Dann gilt*

$$\mathbf{P}(X \in A|Y \in B) = \frac{\mathbf{P}(Y \in B|X \in A) \cdot \mathbf{P}(X \in A)}{\sum_x \mathbf{P}(Y \in B|X=x) \cdot \mathbf{P}(X=x)}.$$

Haben X, Y eine gemeinsame Dichte f, so gilt analog

$$f_X(x|Y=y) = \frac{f_Y(y|X=x) f_X(x)}{\int f_Y(y|X=x) f_X(x) dx}.$$

Beweis. 1. Der Zähler ist gleich $\mathbf{P}(X \in A, Y \in B)$ nach Definition der bedingten Wahrscheinlichkeit, und der Nenner ist

$$\sum_x \mathbf{P}(Y \in B|X=x) \cdot \mathbf{P}(X=x) = \sum_x \mathbf{P}(Y \in B, X=x) = \mathbf{P}(Y \in B).$$

2. folgt analog. □

Beispiel 7.10 (Reihenuntersuchungen)**.** Bei einer Reihenuntersuchung werden nicht nur kranke, sondern manchmal auch gesunde Personen positiv getestet. In einer Population sind insgesamt 0.8% der Personen erkrankt. Eine kranke Person wird in 90% der Fälle positiv getestet, eine gesunde Person mit 7%. Wie groß ist die Wahrscheinlichkeit, dass eine positiv getestete Person wirklich krank ist?
Sei hierzu X eine $\{e, \bar{e}\}$wertige Zufallsvariable, die angibt, ob die zufällig gezogene Person erkrankt ist und Y eine $\{p, n\}$-wertige Zufallsvariable, die angibt, ob die Person positiv oder negativ getestet wird. Gesucht ist also $\mathbf{P}(X=e|Y=p)$. Gegeben ist

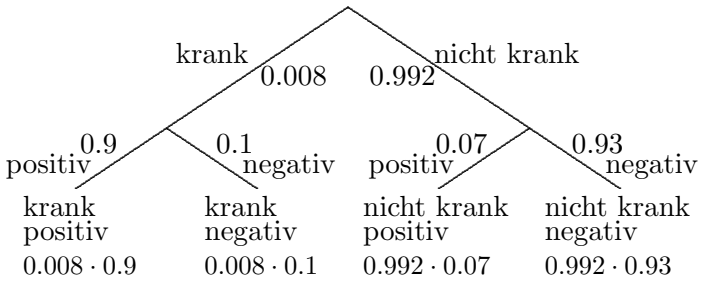
$$\mathbf{P}(X=e) = 0.008, \qquad \mathbf{P}(Y=p|X=e) = 0.9, \qquad \mathbf{P}(Y=p|X=\bar{e}) = 0.07.$$

Wir berechnen mit der Formel von Bayes

$$\begin{aligned}\mathbf{P}(X=e|Y=p) &= \frac{\mathbf{P}(Y=p|X=e) \cdot \mathbf{P}(X=e)}{\mathbf{P}(Y=p|X=e) \cdot \mathbf{P}(X=e) + \mathbf{P}(Y=p|X=\bar{e}) \cdot \mathbf{P}(X=\bar{e})} \\ &= \frac{0.9 \cdot 0.008}{0.9 \cdot 0.008 + 0.07 \cdot 0.992} \approx 0.0939\end{aligned}$$

In der Tat bedeutet also ein positives Testergebnis in nicht einmal einem von zehn Fällen wirklich eine Erkrankung.

Dieses Beispiel lässt sich auch gut anhand eines Wahrscheinlichkeitsbaumes illustrieren:



7.2. Bedingte Erwartungen

Wir kommen nun zu einem Thema, das in der Wahrscheinlichkeitstheorie eine große Rolle spielen wird.

Definition 7.11 (Bedingte Erwartung)**.** *Seien X, Y Zufallsvariable mit Zielbereichen S und T. Weiter sei $h : S \times T \to \mathbb{R}$, so dass $\mathbf{E}[h(X,Y)]$ existiert.*

1. *Sind X, Y diskret, so definieren wir die bedingte Erwartung von $h(X,Y)$, gegeben X, durch*
$$\mathbf{E}[h(X,Y)|X] = \sum_{y \in T} h(X,y) \cdot \mathbf{P}(Y = y|X).$$
2. *Besitzen X, Y die gemeinsame Dichte f, so definieren wir die bedingte Erwartung von $h(X,Y)$, gegeben X, durch*
$$\mathbf{E}[h(X,Y)|X] = \int_T h(X,y) \cdot f_Y(y|X) dy.$$

Proposition 7.12 (Turmeigenschaft)**.** *Seien X, Y Zufallsvariable mit Zielbereichen S und T, sowie $h : S \times T \to \mathbb{R}$ so dass $\mathbf{E}[h(X,Y)]$ existiert. Dann gilt*
$$\mathbf{E}[h(X,Y)] = \mathbf{E}\big[\mathbf{E}[h(X,Y)|X]\big].$$

Beweis. Im diskreten Fall berechnen wir direkt
$$\begin{aligned}\mathbf{E}\big[\mathbf{E}[h(X,Y)|X]\big] &= \sum_{x \in S, y \in T} h(x,y) \cdot \mathbf{P}(Y = y|X = x) \cdot \mathbf{P}(X = x) \\ &= \sum_{x \in S, y \in T} h(x,y) \cdot \mathbf{P}(Y = y, X = x) = \mathbf{E}[h(X,Y)].\end{aligned}$$

Für Zufallsvariable mit Dichte f ist
$$\begin{aligned}\mathbf{E}\big[\mathbf{E}[h(X,Y)|X]\big] &= \int_S \int_T h(x,y) \cdot f_Y(y|X = x) f_X(x) dy dx \\ &= \int_S \int_T h(x,y) \cdot f(x,y) dy dx = \mathbf{E}[h(X,Y)].\end{aligned}$$

□

Beispiel 7.13 (Münzwurf mit zufälliger Erfolgswahrscheinlichkeit)**.** Wir betrachten dieselbe Situation wie in Beispiel 7.8. Wir wissen, dass $Z \sim U([0:n])$, insbesondere ist $\mathbf{E}[Z] = \frac{n}{2}$. dies lässt sich auch ohne Kenntnis der Verteilung von U berechnen, da nämlich Z gegeben U eine $B(n,U)$-Verteilung besitzt, also

$$\mathbf{P}(Z = k|U) = \binom{n}{k} U^k (1-U)^{n-k},$$

und damit

$$\begin{aligned}\mathbf{E}[Z] &= \mathbf{E}\big[\mathbf{E}[Z|U]\big] = \mathbf{E}\Big[\sum_{k=0}^{n} k \cdot \mathbf{P}(Z=k|U)\Big] = \mathbf{E}\Big[\sum_{k=0}^{n} k \cdot \binom{n}{k} U^k (1-U)^{n-k}\Big] \\ &= \mathbf{E}[nU] = \tfrac{1}{2}n.\end{aligned}$$

Analog zur bedingten Erwartung gibt es auch die bedingte Varianz.

Definition 7.14 (Bedingte Varianz)**.** *Seien X, Y Zufallsvariable Zielbereichen S und T, sowie $h : S \times T \to \mathbb{R}$. Wir setzen, falls $\mathbf{Var}[h(X,Y)] < \infty$*

$$\mathbf{Var}[h(X,Y)|X] := \mathbf{E}\big[\big(h(X,Y) - \mathbf{E}[h(X,Y)|X]\big)^2|X\big].$$

Analog zu Lemma 5.15.1 gilt dann

$$\mathbf{Var}[h(X,Y)|X] := \mathbf{E}\big[h(X,Y)^2|X\big] - \mathbf{E}[h(X,Y)|X]^2.$$

Proposition 7.15 (Varianzzerlegung)**.** *Seien X, Y Zufallsvariable mit Zielbereichen $S, T \subseteq \mathbb{R}$, sowie $h : S \to \mathbb{R}$. Sind $\mathbf{Var}[Y], \mathbf{Var}[h(X)] < \infty$, so gilt*

$$\mathbf{E}\big[(Y - h(X))^2\big] = \mathbf{E}\big[\mathbf{Var}[Y|X]\big] + \mathbf{E}\big[\big(\mathbf{E}[Y|X] - h(X)\big)^2\big].$$

Insbesondere ist für $h(X) = \mathbf{E}[Y]$

$$\mathbf{Var}[Y] = \mathbf{E}\big[\mathbf{Var}[Y|X]\big] + \mathbf{Var}\big[\mathbf{E}[Y|X]\big].$$

Beweis. Da man mit bedingten Verteilungen wie üblich rechnen kann, ist

$$\begin{aligned}\mathbf{E}\big[(Y-h(X))^2|X\big] &= \mathbf{Var}[Y - h(X)|X] + \big(\mathbf{E}[Y - h(X)|X]\big)^2 \\ &= \mathbf{Var}[Y|X] + \big(\mathbf{E}[Y|X] - h(X)\big)^2\end{aligned}$$

Bilden des Erwartungswertes ergibt

$$\mathbf{E}\big[(Y - h(X))^2\big] = \mathbf{E}\big[\mathbf{Var}[Y|X]\big] + \mathbf{E}\big[\big(\mathbf{E}[Y|X] - h(X)\big)^2\big].$$

□

Beispiel 7.16 (Zufällige Anzahl unabhängiger Summanden)**.** Sei N eine Zufallsvariable mit Zielbereich $\mathbb{N}$ und $Z_1, Z_2, \ldots$ unabhängig und identisch verteilte Zufallsvariable mit Zielbereich $S \subseteq \mathbb{R}$ und $\mu := \mathbf{E}[Z_1], \sigma^2 = \mathbf{Var}[Z_1]$. Wir untersuchen

$$Y := \sum_{i=1}^{N} Z_i.$$

Da

$$\mathbf{E}[Y|N] = N\mu, \qquad \mathbf{Var}[Y|N] = N\sigma^2,$$

gilt

$$\begin{aligned}\mathbf{E}[Y] &= \mathbf{E}\big[\mathbf{E}[Y|N]\big] = \mathbf{E}[N]\cdot\mu,\\ \mathbf{Var}[Y] &= \mathbf{E}\big[\mathbf{Var}[Y|N]\big] + \mathbf{Var}\big[\mathbf{E}[Y|N]\big] = \mathbf{E}[N]\cdot\sigma^2 + \mathbf{Var}[N]\cdot\mu^2.\end{aligned}$$

7.3. Mehrstufige Experimente

Bisher hatten wir nur höchstens zweistufige Experimente behandelt, etwa hatten wir zunächst eine Erfolgswahrscheinlichkeit U bestimmt, mit der dann eine Münze geworfen wird (Beispiele 7.8 und 7.13). In diesem Abschnitt behandeln wir nun exemplarisch mehrstufige Experimente, vor allem das Beispiel der Pólya-Urne.

Beispiel 7.17 (Pólya-Urne mit zwei Farben)**.** Wir starten mit einer Urne, die zwei Kugeln enthält, nämlich eine grüne (abgekürzt durch 0) und eine schwarze (abgekürzt durch 1). In jedem Zug wählen wir eine Kugel aus der Urne zufällig und legen sie mit einer neuen derselben Farbe zurück. Nun bezeichnen wir mit

- $Z_1 \in \{0,1\}$ das Ergebnis des ersten Zuges,
- $Z_2 \in \{0,1\}$ das Ergebnis des zweiten Zuges (dessen Verteilung von Z_1 abhängt),
- $Z_3 \in \{0,1\}$ das Ergebnis des dritten Zuges (dessen Verteilung von Z_1, Z_2 abhängt),
- $\cdots$

Wir zeigen nun, dass für $a_1, ..., a_n \in \{0,1\}$

$$\mathbf{P}(Z_{n+1} = a_{n+1} | Z_1 = a_1, ..., Z_n = a_n) = \frac{k+1}{n+2}$$

mit

$$k := \#\{j \le n : a_j = a_{n+1}\}.$$

Denn: Nach Hinzufügen der n-ten Kugel liegen insgesamt $n+2$ Kugeln in der Urne. Davon haben $1+k$ Kugeln die Farbe a_{n+1}, nämlich eine Kugel, die bereits am Anfang in der Urne liegt, und k hinzugelegte Kugeln.
Weiter zeigen wir nun noch für $X = \sum_{i=1}^{n} Z_i$, dass

$$\mathbf{P}(X = k) = \frac{1}{n+1},$$

also ist die Gesamtzahl schwarzer Kugeln in der Urne nach $n-1$-maligem Hinzufügen von Kugeln gerade uniform verteilt.
Dies berechnet man wie folgt: zunächst gilt etwa für $(a_1, ..., a_8) = (1,1,0,1,0,0,1,1)$, dass

$$\mathbf{P}(Z_1 = a_1, ..., Z_8 = a_8) = \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{4}\cdot\frac{3}{5}\cdot\frac{2}{6}\cdot\frac{3}{7}\cdot\frac{4}{8}\cdot\frac{5}{9} = \frac{5!3!}{9!}$$

Wir berechnen allgemein für $(a_1, ..., a_n)$ mit $\sum_{i=1}^n a_i = k$

$$\mathbf{P}(Z_1 = a_1, ..., Z_n = a_n) = \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}\frac{1}{\binom{n}{k}}.$$

Da es $\binom{n}{k}$ Folgen $(a_1, ..., a_n) \in \{0,1\}^n$ mit $\sum_{i=1}^n a_i = k$ gibt, folgt dass

$$\mathbf{P}(X = k) = \frac{1}{n+1}, \qquad k = 0, ..., n.$$

Beispiel 7.18 (Pólya Urne mit r Farben)**.** Wir behandeln nun noch den Fall, in dem die Pólya-Urne mit r Kugeln mit r verschiedenen Farben, die mit $1, ..., r$ nummeriert sind, startet. Ansonsten ist alles wir im letzten Beispiel: in jedem Zug ziehen wir eine Kugel aus der Urne und legen eine mit gleicher Farbe mit hinein. Nun ist $Z_i = j$, falls die i-te hinzugefügte Kugel Farbe j hat, $j \in \{1, ..., r\}$. Die Anzahl der bis zum n-ten Zug hinzugefügten Kugeln der Farbe j ist

$$X_j := \sum_{i=1}^n 1_{Z_i = j}.$$

Sei $(a_1, ..., a_n) \in \{1, ..., r\}^n$ mit $\sum_{i=1}^n 1_{a_i=j} = k_j, j = 1, ..., r$ und $k_1 + \cdots + k_r = n$. Es gibt $\binom{n}{k_1 \cdots k_r}$ solche Folgen. Wie für $r = 2$ gilt für jede dieser Folgen

$$\mathbf{P}(Z_1 = a_1, ..., Z_n = a_n) = \frac{k_1! \cdots k_r! \cdot (r-1)!}{(n+r-1)!} = \frac{1}{\binom{n+r-1}{r-1}}\frac{1}{\binom{n}{k_1 \cdots k_r}}$$

und

$$\mathbf{P}(X_1 = k_1, ..., X_r = k_r) = \frac{1}{\binom{n+r-1}{r-1}}.$$

Die Pólya-Urne liefert also uniform verteilte Besetzungen der r Farben.

8. Markov-Ketten

Die Vorlesung hat sich bisher mit Zufallsvariablen und deren Verteilung beschäftigt. Speziell haben wir bereits die Situation von mehreren abhängigen Zufallsvariablen kennen gelernt, die etwa bei zwei- oder mehrstufigen Experimenten auftauchen. In diesem Abschnitt werden wir stochastische Prozesse kennen lernen, also (fast) beliebige Familien von Zufallszahlen. Wie der Name schon sagt, stellt man sich dabei vor, dass die Zufallszahlen nacheinander, dynamisch als Prozess realisiert werden, genau wie bei der Pólya-Urne aus dem letzten Abschnitt. Dabei beschränken wir uns auf den Fall von diskreten Zeitschritten, betrachten also Familien von Zufallsvariablen $\mathcal{X} = (X_t)_{t=0,1,2,\dots}$.

Eine besondere Form der Abhängigkeit von Zufallsvariablen tritt in Markov-Ketten zu Tage. Hier werden der Reihe nach Zufallsvariablen $X_0, X_1, \dots$ realisiert, und zwar so, dass X_{t+1} nur von X_t abhängt. Man sagt auch, X_{t+1} ist unabhängig von $X_1, \dots, X_{t-1}$, wenn X_t bekannt ist. Diese Form der Abhängigkeit wird häufig zur stochastischen Modellierung verwendet. Beispielsweise könnte $X_0, X_1, \dots$ der Preis einer Aktie an Tagen $0, 1, \dots$ sein. Die Markov-Eigenschaft besagt in diesem Fall, dass die Verteilung der Kursänderungen am Tag $t+1$ nur davon abhängt, wie der Kurs X_t am Tag t war.

8.1. Grundlegendes

Wir beginnen mit der Definition von Markov-Ketten. Wir werden vor allem den Fall von zeitlich homogenen Markov-Ketten behandeln. Bei solchen gibt es eine sich zeitlich nicht ändernde stochastische Übergangsvorschrift, wie die Verteilung des Zustandes X_{t+1} ist, wenn der Zustand X_t der Kette zur Zeit t bekannt ist. Diese Vorschrift wird mit Hilfe einer Matrix zusammengefasst, der Übergangsmatrix.

Definition 8.1 (Stochastischer Prozess und Markov-Ketten)**.** *1. Seien I und S Mengen. Ein (S-wertiger) stochastischer Prozess (mit Indexmenge I) ist eine Familie von Zufallsvariablen $\mathcal{X} = (X_t)_{t\in I}$ mit Wertebereich S. Die Menge S heißt auch* Zustandsraum *von $\mathcal{X}$.*

2. Sei $I = \{0, 1, 2, \dots\}$, S abzählbar und $\mathcal{X} = (X_t)_{t\in I}$ ein S-wertiger stochastischer Prozess. Falls

$$\mathbf{P}(X_{t+1} = i \mid X_1, \dots, X_t) = \mathbf{P}(X_{t+1} = i \mid X_t) \tag{8.1}$$

für alle $i \in S$, so heißt $\mathcal{X}$ eine Markov-Kette*. Sie heißt* endlich*, falls S endlich ist.*

3. Sei $\mathcal{X} = (X_t)_{t\in I}$ eine S-wertige Markov-Kette. Existiert eine Matrix $P = (P_{ij})_{i,j\in S}$ mit

$$P_{ij} := \mathbf{P}(X_{t+1} = j \mid X_t = i)$$

für alle t, so heißt $\mathcal{X}$ zeitlich homogen *und P heißt* Übergangsmatrix *von $\mathcal{X}$.*

Bemerkung 8.2. 1. Die Eigenschaft (8.1) bedeutet in Worten: die zukünftige Entwicklung von $\mathcal{X}$ nach t hängt von $X_1, \dots, X_t$ nur durch den aktuellen Zustand X_t ab.

2. Sei P die Übergangsmatrix einer homogenen Markov-Kette $\mathcal{X}$ mit Zustandsraum S. Dann gilt

$$\begin{aligned} 0 \le P_{ij} \le 1, &\qquad i,j \in S, \\ \sum_{j\in S} P_{ij} = 1, &\qquad i \in S. \end{aligned} \tag{8.2}$$

Die erste Eigenschaft ist klar, da die Einträge in P Wahrscheinlichkeiten sind. Außerdem ist

$$1 = \mathbf{P}(X_{t+1} \in S \mid X_t = i) = \sum_{j\in S} \mathbf{P}(X_{t+1} = j \mid X_t = i) = \sum_{j\in S} P_{ij}.$$

Matrizen P mit den Eigenschaften (8.2) heißen *stochastische Matrizen*.

3. Sei $\mathcal{X}$ eine homogene Markov-Kette mit Übergangsmatrix P. Definiere einen gewichteten, gerichteten Graphen (E, K, W) wie folgt: die Menge der Knoten ist E, die Menge der (gerichteten) Kanten ist $K := \{(i,j) : P_{ij} > 0\}$. Das Gewicht der Kante (ij) ist $w_{(ij)} := P_{ij}$ und $W = (w_{(ij)})_{(ij)\in K}$. Der Graph (E, K, W) heißt *Übergangsgraph* von $\mathcal{X}$.

4. Für stochastische Prozesse mit überabzählbarem I bedarf es einiger maßtheoretischer Arbeit, um die Verteilung der Familie $\mathcal{X}$ zu definieren und zu verstehen, durch was diese eindeutig gegeben ist. Diese fortgeschrittene Theorie stochastischer Prozesse wird in der Vorlesung *Stochastischer Prozesse* eingeführt.

Beispiel 8.3 (Irrfahrt im Dreieck)**.** Betrachte eine homogene Markov-Kette $\mathcal{X}$ mit Zustandsraum $\{1,2,3\}$ und Übergangsmatrix

$$P = \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix} \tag{8.3}$$

für $p \in (0,1)$ und $q := 1-p$. Die Kette $\mathcal{X}$ veranschaulicht man sich am besten anhand des Übergangsgraphen; siehe Abbildung 8.1. In jedem Zustand $1,2,3$ ist die Wahrscheinlichkeit, im Uhrzeigersinn zu wandern p, und die Wahrscheinlichkeit gegen den Uhrzeigersinn zu gehen ist q.

Beispiel 8.4 (Ruinproblem)**.** Betrachte folgendes Ruinproblem: zwei Spieler spielen gegeneinander. Spieler 1 startet mit n Euro, Spieler 2 mit $N-n$ Euro. Bei jedem Spiel gewinnt Spieler 1 mit Wahrscheinlichkeit p einen Euro, mit Wahrscheinlichkeit $q = 1-p$ verliert er einen Euro. Das Spiel endet, wenn einer der beiden pleite ist.

Sei X_t das Vermögen von Spieler 1 nach dem t-ten Spiel. In dieser Situation ist $\mathcal{X} = (X_t)_{t=0,1,2,\dots}$ eine endliche, homogene Markov-Kette mit Zustandsraum $\{0, \dots, N\}$ und Über-

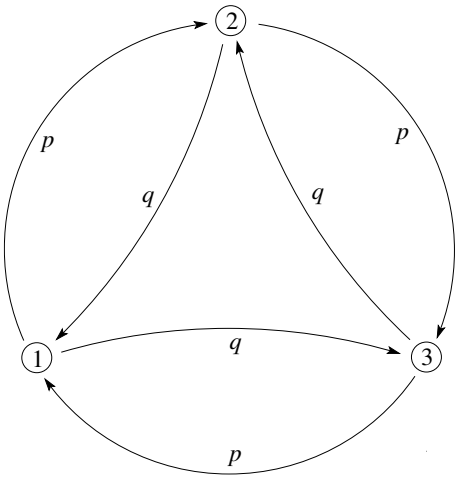


Abbildung 8.1.: Übergangsgraph der Markov-Kette aus Beispiel 8.3.

gangsmatrix

$$P = \begin{pmatrix} 1 & 0 & & & & \\ q & 0 & p & & & \\ & q & 0 & p & & \\ & & \ddots & \ddots & \ddots & \\ & & & q & 0 & p \\ & & & & 0 & 1 \end{pmatrix}.$$

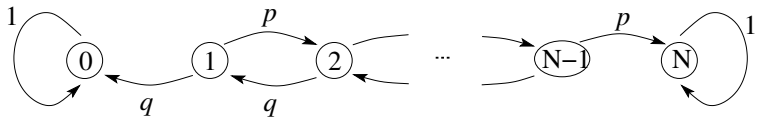


Abbildung 8.2.: Übergangsgraph der Markov-Kette aus Beispiel 8.4.

Der Übergangsgraph ist in Abbildung 8.2 dargestellt. Klar ist, dass nach langer Zeit entweder Spieler 1 oder Spieler 2 gewonnen hat, dass also $X_t \xrightarrow{t\to\infty} 0$ oder $X_t \xrightarrow{t\to\infty} N$. Die Frage ist nur:

Mit welcher Wahrscheinlichkeit gewinnt Spieler 1?

Wir bezeichnen mit

$$p_n := \mathbf{P}(X_t \xrightarrow{t\to\infty} N \mid X_0 = n)$$

die Wahrscheinlichkeit, dass Spieler 1 gewinnt. Wir werden zeigen, dass

$$p_n = \begin{cases} \dfrac{n}{N}, & p = q = \frac{1}{2}, \\ \dfrac{1-(q/p)^n}{1-(q/p)^N}, & \text{sonst.} \end{cases}$$

Zunächst gilt für $n = 1, \dots, N-1$

$$\begin{aligned} p_n &= q \cdot \mathbf{P}(X_t \xrightarrow{t\to\infty} N \mid X_0 = n, X_1 = n-1) + p \cdot \mathbf{P}(X_t \xrightarrow{t\to\infty} N \mid X_0 = n, X_1 = n+1) \\ &= q \cdot \mathbf{P}(X_t \xrightarrow{t\to\infty} N \mid X_0 = n-1) + p \cdot \mathbf{P}(X_t \xrightarrow{t\to\infty} N \mid X_0 = n+1) \\ &= q \cdot p_{n-1} + p \cdot p_{n+1}, \end{aligned}$$

also mit $\Delta p_n := p_n - p_{n-1}$

$$q\Delta p_n = p\Delta p_{n+1}.$$

Im Fall $p = q = \frac{1}{2}$ folgt mit $\sum_{m=1}^{N} \Delta p_m = p_N - p_0 = 1$ daraus bereits

$$p_n = \frac{\sum_{m=1}^{n} \Delta p_m}{\sum_{m=1}^{N} \Delta p_m} = \frac{n\Delta p_1}{N\Delta p_1} = \frac{n}{N}.$$

Im Fall $p \neq q$ setzen wir $u := \frac{q}{p}$ und berechnen iterativ

$$\Delta p_n = u\Delta p_{n-1} = u^2 \Delta p_{n-2} \cdots = u^{n-1}\Delta p_1 = u^{n-1} p_1.$$

Weiter ist

$$1 = \sum_{m=1}^{N} \Delta p_m = p_1 \sum_{m=0}^{N-1} u^m = p_1 \frac{1-u^N}{1-u}.$$

Also

$$p_n = \sum_{m=1}^{n} \Delta p_m = p_1 \sum_{m=1}^{n} u^{m-1} = \frac{1-u}{1-u^N} \frac{1-u^n}{1-u} = \frac{1-u^n}{1-u^N}$$

und die Behauptung ist gezeigt.

Beispiel 8.5 (Ehrenfest'sche Urne)**.** Betrachte folgendes Urnenmodell: In einer Urne gibt es zwei durch eine Trennwand getrennte Kammern. Insgesamt liegen n Kugeln in den beiden Kammern. Wir ziehen eine Kugel rein zufällig aus der Urne und legen sie anschließend in die andere Kammer zurück. In Abbildung 8.3 findet sich eine Illustration.

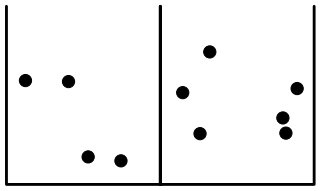


Abbildung 8.3.: Die Ehrenfest'sche Urne mit $n = 10$ Kugeln aus Beispiel 8.5. Im nächsten Schritt wird mit Wahrscheinlichkeit $\frac{4}{10}$ eine Kugel von der linken Kammer in die rechte und mit Wahrscheinlichkeit $\frac{6}{10}$ von der rechten in die linke Kammer gelegt.

Wir betrachten den $\{0,\dots,n\}$-wertigen stochastischen Prozess $\mathcal{X} = (X_t)_{t=0,1,2,\dots}$, wobei X_t die Anzahl der Kugeln in der linken Kammer nach dem t-Schritt darstellt. Dann ist $\mathcal{X}$ eine homogene Markov-Kette, denn der Ausgang des Schrittes $t+1$ hängt nur von X_t ab. Die Übergansmatrix von $\mathcal{X}$ ist gegeben durch

$$P_{ij} = \begin{cases} \frac{i}{n}, & j = i-1, \\ \frac{n-i}{n}, & j = i+1, \\ 0, & \text{sonst.} \end{cases} \tag{8.4}$$

□

Wir kommen nun zu einem wichtigen Zusammenhang zwischen der Verteilung einer Markov-Ketten zum Zeitpunkt t und Potenzen der Übergangsmatrix.

Theorem 8.6. *Sei $\mathcal{X} = (X_t)_{t=0,1,2,\dots}$ eine homogene Markov-Kette mit Übergangsmatrix P und $\mu^{(t)} = (\mu^{(t)}(i))_{i\in S}$,*

$$\mu^{(t)}(i) = \mathbf{P}(X_t = i)$$

für $t = 0,1,2,\dots$ Dann gilt

$$\mu^{(t)} = \mu^{(0)} P^t, \tag{8.5}$$

für $t = 0,1,2,\dots$, wobei die rechte Seite als Multiplikation des Zeilenvektors $\mu^{(0)}$ und der Matrix $P^t = \underbrace{P\cdots P}_{t \text{ mal}}$ zu verstehen ist.

Beweis. Der Beweis geht mittels Induktion über t. Für $t = 0$ ist die Aussage klar. Ist (8.5) für t gezeigt, so gilt

$$\begin{aligned} \mu^{(t+1)}(j) = \mathbf{P}(X_{t+1} = j) &= \sum_{i\in S} \mathbf{P}(X_{t+1} = j \mid X_t = i) \cdot \mathbf{P}(X_t = i) \\ &= \sum_{i\in S} \mu^{(t)}(i) \cdot P_{ij} = \sum_{i\in S} (\mu^{(0)} P^t)_i \cdot P_{ij} = (\mu^{(0)} P^{t+1})_j. \end{aligned}$$

□

Beispiel 8.7 (Irrfahrt im Dreieck)**.** Betrachte die Markov-Kette $\mathcal{X}$ aus Beispiel 8.3 und Übergangsmatrix P aus (8.3). Sei $X_0 = 1$, die Markov-Kette startet also in 1, d.h. $\mu^{(0)} = (1,0,0)$. Weiter berechnen wir

$$P = \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix}, \qquad P^2 = \begin{pmatrix} 2pq & q^2 & p^2 \\ p^2 & 2pq & q^2 \\ q^2 & p^2 & 2pq \end{pmatrix} \tag{8.6}$$

und damit

$$\mu^{(2)} = (2pq, q^2, p^2).$$

Also ist etwa $\mathbf{P}(X_2 = 1) = 2pq$. Dies ist klar, weil $X_2 = 1$ genau dann, wenn die ersten beiden Schritte im Übergangsgraphen aus Abbildung 8.1 einmal mit und einmal entgegen dem Uhrzeigersinn gingen. Es ist $X_2 = 3$ genau dann, wenn zwei Schritte im Uhrzeigersinn realisiert wurden, was Wahrscheinlichkeit p^2 hat.

8.2. Stationäre Verteilungen

Wir studieren nun Markov-Ketten, die schon lange gelaufen sind, also für die Verteilung von X_t für große t. Solche Markov-Ketten können (mehr dazu im nächsten Abschnitt) so beschaffen sein, dass sich die Verteilung nicht mehr viel ändert. Verteilungen auf dem Zustandsraum S, die invariant sind gegenüber Schritten der Markov-Kette, heißen stationär.

Definition 8.8 (Stationäre Verteilung)**.** *Sei $\mathcal{X}$ eine homogene Markov-Kette mit Übergangsmatrix P und π eine Verteilung auf S, gegeben durch einen Vektor $\pi = (\pi_i)_{i\in S}$.*

1. *Gilt*
$$\pi P = \pi,$$
so heißt π stationäre Verteilung *von $\mathcal{X}$.*

2. *Gilt*
$$\pi_i P_{ij} = \pi_j P_{ji}$$
für alle i, j, so heißt π reversible Verteilung.

Bemerkung 8.9 (Eigenschaften von stationären und reversiblen Verteilungen)**.**

1. Sei π stationär und $\mathbf{P}(X_t = i) = \pi_i$. Dann gilt wegen Theorem 8.6, dass
$$\begin{aligned}\mathbf{P}(X_{t+1} = j) &= \sum_{i\in S} \mathbf{P}(X_t = i) \cdot \mathbf{P}(X_{t+1} = j \mid X_t = i) = \sum_{i\in S} \pi_i P_{ij} = (\pi P)_j = \pi_j \\ &= \mathbf{P}(X_t = j).\end{aligned}$$
Das bedeutet, dass X_t und X_{t+1} (und damit auch $X_{t+2}, X_{t+3}, \dots$) identisch verteilt sind.

2. Jede reversible Verteilung ist stationär. Ist nämlich π reversibel, so gilt
$$(\pi P)_j = \sum_{i\in S} \pi_i P_{ij} = \sum_{i\in S} \pi_j P_{ji} = \pi_j,$$
da P Zeilensumme 1 hat.

3. Sei π reversibel, $\mathbf{P}(X_t = i) = \pi_i$ und $i_t, \dots, i_{t+s} \in S$. Dann gilt
$$\begin{aligned}\mathbf{P}(X_t = i_t, \dots, X_{t+s} = i_{t+s}) &= \pi_{i_t} P_{i_t, i_{t+1}} \cdots P_{i_{t+s-1}, i_{t+s}} \\ &= \pi_{i_{t+1}} P_{i_{t+1}, i_{t+2}} \cdots P_{i_{t+s-1}, i_{t+s}} \cdot P_{i_{t+1}, i_t} = \cdots \\ &= \pi_{i_{t+s}} P_{i_{t+s}, i_{t+s-1}} \cdots P_{i_{t+1}, i_t} \\ &= \mathbf{P}(X_t = i_{t+s}, \dots, X_{t+s} = i_t).\end{aligned}$$
Das bedeutet, dass die Wahrscheinlichkeit, die Zustände $i_t, \dots, i_{t+s}$ zu durchlaufen, dieselbe ist wie die, die Zustände in umgekehrter Reihenfolge (reversibel) zu durchlaufen.

Beispiel 8.10 (Irrfahrt im Dreieck)**.** Betrachte die Irrfahrt auf dem Dreieck $\mathcal{X}$ aus Beispiel 8.3 mit Übergangsmatrix P aus (8.3). Für $\pi = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ ist
$$\pi P = \left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right) \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix} = \pi.$$

Damit ist die Gleichverteilung auf $\{1,2,3\}$ stationäre Verteilung von $\mathcal{X}$. Allerdings ist π nicht reversibel (außer für $p=\frac{1}{2}$), denn es gilt etwa

$$\pi_1 P_{12} = \tfrac{1}{3}p \neq \tfrac{1}{3}q = \pi_2 P_{21}.$$

Letztere Beobachtung ist nicht erstaunlich: Sei etwa $p > \frac{1}{2}$. Dann erwarten wir, dass $\mathcal{X}$ das Dreieck öfter im Uhrzeigersinn durchläuft als umgekehrt. Wäre nun π reversibel, würde daraus folgen (siehe Bemerkung 8.9.3) , dass die Wahrscheinlichkeit für einen Durchlauf des Dreiecks gegen und im Uhrzeigersinn dieselbe ist.

Beispiel 8.11 (Ruinproblem)**.** Im Ruinproblem aus Beispiel 8.4 gibt es zwar stationäre Verteilungen, diese sind jedoch nicht sonderlich interessant. Jede Verteilung $\pi = (p', 0, \ldots, 0, q')$ mit $p' \in (0,1)$ und $q' = 1-p'$ ist stationäre (und nicht reversible) Verteilung, wie man leicht nachrechnet. Insbesondere sind $(1,0,\ldots,0)$ und $(0,\ldots,0,1)$ stationär. Das bedeutet, dass die Zustände $X_t = 0$ und $X_t = n$ von der Markov-Kette nicht mehr verlassen werden. Man sagt auch, 0 und n sind *Fallen* für $\mathcal{X}$.

Beispiel 8.12 (Ehrenfest'sche Urne)**.** Sei $\mathcal{X} = (X_t)_{t=0,1,2,\ldots}$ die Anzahl der Kugeln in der linken Kammer einer Ehrenfest'schen Urne, wie in Beispiel 8.5. Hier ist die Binomialverteilung $\pi = B(n, \frac{1}{2})$ reversible Verteilung für $\mathcal{X}$. Es gilt nämlich mit (8.4) für $i = 1, \ldots, n$

$$\pi_i P_{i,i-1} = \binom{n}{i}\frac{1}{2^n}\frac{i}{n} = \binom{n}{i-1}\frac{1}{2^n}\frac{n-i}{n} = \pi_{i-1}P_{i-1,i}.$$

Die Tatsache, dass $B(n,p)$ stationär ist, interpretiert man am besten so: nach langer Zeit ist jede der Kugeln oft von der linken in die rechte Kammer und umgelegt worden, so dass jede Kugel mit Wahrscheinlichkeit $\frac{1}{2}$ jeweils in der linken und rechten Kammer liegt. Diese Verteilung ändert sich nicht mehr, weil in jedem Schritt nur eine der n Kugeln nochmals in die andere Kammer gelegt wird. □

Stationäre Verteilungen existieren für die meisten Markov-Ketten, wie das nächste Resultat zeigt. Im nächsten Abschnitt werden wir dann betrachten, wann diese eindeutig sind.

Theorem 8.13 (Existenz von stationären Verteilungen)**.** *Sei $\mathcal{X} = (X_t)_{t=0,1,2,\ldots}$ eine endliche, homogene Markov-Kette mit Zustandsraum S. Es gibt Folgen $t_n \to \infty$ so, dass*

$$\pi_i := \lim_{n\to\infty} \frac{1}{t_n}\sum_{s=1}^{t_n} \mathbf{P}(X_s = i) \tag{8.7}$$

für alle $i \in S$. Jeder solche Vektor $\pi = (\pi_i)_{i\in S}$ definiert eine Verteilung auf S. Diese ist stationäre Verteilung von $\mathcal{X}$.

Beweis. Sei $S = \{i_1, \ldots, i_m\}$ und

$$\pi_{t,i} := \frac{1}{t}\sum_{s=1}^{t} \mathbf{P}(X_s = i).$$

Klar ist, dass für alle $i \in S$ die Folgen $(\pi_{t,i})_{t=1,2,\ldots}$ Werte im kompakten Intervall $[0,1]$ annehmen. Damit gibt es eine konvergente Teilfolge von $(\pi_{t,i_1})_{t=1,2,\ldots}$. Von dieser Teilfolge gibt es

eine weitere Teilfolge, so dass $(\pi_{t,i_1})_{t=1,2,\ldots}$ und $(\pi_{t,i_2})_{t=1,2,\ldots}$ entlang der Teilfolge konvergiert. Iteriert man dieses Verfahren, findet man eine Teilfolge, so dass (8.7) für alle $i \in S$ gilt.

Klar ist, dass $0 \leq \pi_i \leq 1$ für alle $i \in S$ gilt. Außerdem ist

$$\sum_{i \in S} \pi_i = \lim_{n \to \infty} \frac{1}{t_n} \sum_{s=1}^{t_n} \sum_{i \in S} \mathbf{P}(X_s = i) = 1,$$

und damit definiert π eine Verteilung auf S. Um zu sehen, dass π eine stationäre Verteilung ist, berechnen wir

$$\begin{aligned}
(\pi P)_j &= \sum_{i \in S} \pi_i P_{ij} = \lim_{n \to \infty} \frac{1}{t_n} \sum_{s=1}^{t_n} \sum_{i \in S} \mathbf{P}(X_s = i) \cdot \mathbf{P}(X_{s+1} = j \mid X_s = i) \\
&= \lim_{n \to \infty} \frac{1}{t_n} \sum_{s=1}^{t_n} \mathbf{P}(X_{s+1} = j) \\
&= \frac{1}{t_n}(\mathbf{P}(X_{t_n+1} = i) - \mathbf{P}(X_1 = i)) + \lim_{n \to \infty} \frac{1}{t_n} \sum_{s=1}^{t_n} \mathbf{P}(X_s = j) \\
&= \pi_j
\end{aligned}$$

und die Behauptung ist gezeigt. □

8.3. Markov-Ketten-Konvergenzsatz

In diesem Abschnitt zeigen wir, dass jede endliche, homogene, *aperiodische* und *irreduzible* Markov-Kette konvergiert. Die Konvergenz ist dabei so zu verstehen, dass sich die Verteilung von X_t im Grenzwert großer t nicht mehr ändert, und auch nicht vom Anfangszustand X_0 abhängt. Wir beginnen mit Erklärungen der benötigten Begriffe.

Definition 8.14. *Sei $\mathcal{X}$ eine S-wertige homogene Markov-Kette.*

1. *Der Zustand $i \in S$* kommuniziert *mit $j \in S$, falls es ein t gibt mit*

$$\mathbf{P}(X_t = j \mid X_0 = i) > 0.$$

   In diesem Fall schreiben wir $i \to j$. Falls $i \to j$ und $j \to i$, schreiben wir $i \leftrightarrow j$.

2. *Falls $i \leftrightarrow j$ für alle $i, j \in S$, so heißt $\mathcal{X}$* irreduzibel. *Andernfalls heißt $\mathcal{X}$* reduzibel.

3. *Ein Zustand $i \in S$ heißt* aperiodisch, *falls*

$$d(i) := \mathrm{ggT}\{t : \mathbf{P}(X_t = i \mid X_0 = i) > 0\} = 1.$$

   Andernfalls heißt i periodisch *mit Periode $d(i)$.*

4. *Falls alle $i \in S$ aperiodisch sind, so heißt $\mathcal{X}$* aperiodisch.

Bemerkung 8.15 (Irreduzibilität, Aperiodizität und die Übergangsmatrix)**.**

1. Die Begriffe der Irreduzibilität und Aperiodizität lassen sich auch mittels der Übergangsmatrix P der Markov-Kette $\mathcal{X}$ erklären. Es gilt etwa $i \to j$ genau dann, wenn es ein t gibt mit $(P^t)_{ij} > 0$. Außerdem ist $d(i) = \mathrm{ggT}\{t : P^t_{ii} > 0\}$.

2. Es gibt einen einfachen Zusammenhang zwischen dem Begriff der kommunizierenden Zustände und dem Übergangsgraphen: ein Zustand i kommuniziert mit einem Zustand j, wenn man einen Pfad $i \to j$ im Übergangsgraphen der Markov-Kette findet. Das bedeutet, dass es eine endliche Folge $(i, i_1), (i_1, i_2), \ldots, (i_{n-1}, i_n), (i_n, j)$ im Übergangsgraphen gibt, also der Zustand j von i aus durch eine endliche Folge von Kanten erreicht werden kann.

3. Sei $\mathcal{X}$ eine reduzible Markov-Kette. Der Begriff der Reduzibilität erklärt sich so, dass sich die Markov-Kette $\mathcal{X}$ auf einen Zustandsraum $S' \subseteq S$ *reduzieren* lässt. Das bedeutet, dass aus $X_0 \in S'$ folgt, dass $X_t \in S'$ für alle $t = 1, 2, \ldots$

Beispiel 8.16 (Irrfahrt auf dem Dreieck)**.** Sei $\mathcal{X}$ die Irrfahrt auf dem Dreieck aus Beispiel 8.3. In Beispiel 8.7 haben wir ausgerechnet, dass $P^2 > 0$ ist. Weiter ist auch $P^3 > 0$, was bereits zeigt, dass $\mathcal{X}$ sowohl irreduzibel als auch aperiodisch ist.

Beispiel 8.17 (Ruinproblem)**.** Für die Markov-Kette $\mathcal{X}$ aus dem Ruinproblem, Beispiel 8.4, gilt sicher, dass $0 \not\to i$ für $i = 1, \ldots, n$, und $n \not\to i$ für $i = 0, \ldots, n-1$. Mit anderen Worten: Ist Spieler 1 pleite (d.h. $X_t = 0$ für ein t), so wird er nach den Spielregeln nie wieder Geld bekommen, d.h. alle Zustände $1, \ldots, n$ sind für ihn unerreichbar. Das bedeutet also, $\mathcal{X}$ ist reduzibel.

Beispiel 8.18 (Ehrenfest'sche Urne)**.** Die Markov-Kette $\mathcal{X}$, die die Anzahl der Kugeln in der linken Kammer einer Ehrenfest'schen Urne beschreibt, ist irreduzibel. Betrachtet man nämlich Zustände i, j, etwa mit $i > j$, so genügt es ja, genau $i - j$ Kugeln nacheinander von der linken in die rechte Kammer zu legen. Allerdings ist die Markov-Kette periodisch. Sei etwa $X_0 = 2i$ gerade, dann folgt, dass X_1 ungerade (entweder $2i - 1$ oder $2i + 1$) ist. Damit kann nur zu geraden Zeiten t der Zustand $X_t = 2i$ eintreten und es ist $\mathrm{ggT}\{t : \mathbf{P}(X_t = 2i | X_0 = 2i)\} = 2$.

Nun kommen wir also zum Markov-Ketten-Konvergenzsatz.

Theorem 8.19 (Markov-Ketten-Konvergenzsatz)**.** *Sei $\mathcal{X} = (X_t)_{t=0,1,2,\ldots}$ eine endliche, homogene, irreduzible und aperiodische Markov-Kette. Dann hat $\mathcal{X}$ genau eine stationäre Verteilung π und es gilt für alle $i \in S$*

$$\mathbf{P}(X_t = i) \xrightarrow{t \to \infty} \pi(i). \tag{8.8}$$

Bemerkung 8.20. Vor allem ist an dem Resultat bemerkenswert, dass (8.8) nicht von der Verteilung von X_0 abhängt. Das bedeutet, dass die Markov-Kette nach langer Zeit *vergisst*, in welchem Zustand X_0 sie zur Zeit 0 gestartet ist.

Beispiel 8.21 (Irrfahrt auf dem Dreieck)**.** Von unseren drei Beispielen erfüllt (nur) die Irrfahrt auf dem Dreieck $\mathcal{X}$ aus Beispiel 8.3 die Voraussetzungen des Satzes. Nach Beispiel 8.16 ist $\mathcal{X}$ irreduzibel und aperiodisch. Nach Beispiel 8.10 und obigem Satz ist die Gleichverteilung auf $\{1, 2, 3\}$ die einzige stationäre Verteilung für $\mathcal{X}$. Außerdem gilt

$$\mathbf{P}(X_t = i) \xrightarrow{t \to \infty} \tfrac{1}{3}.$$

Bemerkung 8.22 (Beweistechnik Kopplung)**.** Die Technik, mit der wir das Theorem beweisen werden heißt *Kopplung*. Sind $\mathcal{V}, \mathcal{W}$ stochastische Prozesse mit Zustandsräumen S und

S', so heißt ein stochastischer Prozess $(\mathcal{Y}, \mathcal{Z})$ mit Zustandsraum $S \times S'$ Kopplung von $\mathcal{V}, \mathcal{W}$, wenn $\mathcal{V}$ und $\mathcal{Y}$, sowie $\mathcal{W}$ und $\mathcal{Z}$ identisch verteilt sind. Wichtig ist hier, dass das Paar $(\mathcal{Y}, \mathcal{Z})$ eine gemeinsame Entwicklung definiert, die Prozesse $\mathcal{V}$ und $\mathcal{W}$ jedoch nicht.

Seien etwa $\mathcal{V} = (V_t)_{t=0,1,2,\dots}$ und $\mathcal{W} = (W_t)_{t=0,1,2,\dots}$ homogene Markov-Ketten mit Zustandsraum $\{0, 1\}$ mit Übergangsmatrix

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

mit $q := 1 - p$ und $V_0 = 0, W_0 = 1$. Sowohl $\mathcal{V}$ als auch $\mathcal{W}$ bleiben also jeweils mit Wahrscheinlichkeit p in ihrem Zustand und wechseln ihn mit Wahrscheinlichkeit q. Wir definieren die gekoppelte Markov-Kette $(\mathcal{Y}, \mathcal{Z})$ durch die Übergangsmatrix

$$P' = \begin{pmatrix} p & 0 & 0 & q \\ qp & p^2 & q^2 & pq \\ pq & q^2 & p^2 & qp \\ q & 0 & 0 & p \end{pmatrix},$$

wobei die Reihenfolge der Zustände in den Zeilen und Spalten durch $(0,0), (0,1), (1,0), (1,1)$ gegeben ist. Beispielsweise ist die Wahrscheinlichkeit eines Überganges $(0, 1)$ nach $(1, 0)$ und nach $(1, 1)$ gleich q^2 und pq. Daraus folgt auch, dass die Wahrscheinlichkeit, dass $\mathcal{Y}$ von 0 nach 1 springt, gleich q ist. Durch analoge Überlegungen lässt sich zeigen, dass $\mathcal{V}$ und $\mathcal{Y}$, sowie $\mathcal{W}$ und $\mathcal{Z}$, identisch verteilt sind. Auffällig ist weiterhin, dass die Menge $\{(0,0), (1,1)\}$ nicht verlassen wird. Die Markov-Kette $(\mathcal{Y}, \mathcal{Z})$ lässt sich so interpretieren, dass $\mathcal{Y} = (Y_t)_{t=0,1,2,\dots}$ und $\mathcal{Z} = (Z_t)_{t=0,1,2,\dots}$ sich so lange unabhängig entwickeln, bis $Y_t = Z_t$ gilt. Dann gilt weiterhin $Y_r = Z_r$ für $r \geq t$. Wie man leicht zeigt, gibt es mit Wahrscheinlichkeit 1 ein t mit $Y_t = Z_t$. Man spricht dann auch von einer *erfolgreichen Kopplung.*

Bevor wir den Markov-Ketten-Konvergenzsatz beweisen, benötigen wir noch ein paar Hilfsaussagen.

Lemma 8.23. *Sei $A = \{s_1, s_2, \dots\} \subseteq \mathbb{Z}$ mit $\mathrm{ggT}(A) = 1$.*

1. *Ist A abgeschlossen unter Addition und Subtraktion (d.h. aus $s, s' \in A$ folgt $s+s', s-s' \in A$), dann ist $1 \in A$.*
2. *Ist $A \subseteq \mathbb{N}$ und abgeschlossen unter Addition, dann gibt es ein $t < \infty$ mit $\{t, t+1, t+2, \dots\} \subseteq A$.*

Beweis. 1. Sei a das kleinste positive Element von A. Wir zeigen zunächst, dass $A = \{ka : k \in \mathbb{Z}\}$ gilt. Ist nämlich $c \in A$, so lässt es sich eindeutig als $c = ka + r$ mit $0 \leq r < a$ schreiben. Es muss $r = 0$ gelten, da ansonsten $r = c - ka \in A$ ist und kleiner als a. Damit gilt, dass $c = ka$ für ein $k \in \mathbb{Z}$. Da a das kleinste positive Element von A ist, folgt

$$\mathrm{ggT}(A) = \mathrm{ggT}\{ka : k \in \mathbb{Z}\} = a,$$

woraus $a = 1$ folgt.
2. Wegen 1. gibt es $s_1, \dots, s_k \in A$ und $n_1, \dots, n_k \in \mathbb{Z}$ mit

$$1 = s_1 n_1 + \cdots + s_k n_k.$$

Wenn wir die positiven von den negativen Termen trennen, können wir schreiben, dass $1 = m - p$ mit $m, p \in A$. Wir zeigen nun, dass die Aussage mit $t = p(p-1)$ gilt. Sei nämlich $c \geq t$ mit $c = ap + r$ mit $a \geq p - 1$ und $r < p$. Dann gilt $c = ap + r(m-p) = (a-r)p + rm$. Da $m, p \in A$ und $a - r \geq p - 1 - r \geq 0$, folgt dass $c \in A$. □

Proposition 8.24. *Sei $\mathcal{X}$ eine endliche, homogene, aperiodische Markov-Kette mit Zustandsraum S und Übergangsmatrix P. Dann gibt es $t < \infty$, so dass*

$$\mathbf{P}(X_r = j \mid X_0 = i) > 0$$

für alle $i \in S$ und $r = t, t+1, t+2, \ldots$ gilt.

Beweis. Wegen der Endlichkeit von S genügt es, die Behauptung für ein beliebiges Paar i, j zu zeigen. Sei zunächst $j = i$. Wir setzen

$$A := \{s : \mathbf{P}(X_s = i \mid X_0 = i) > 0\}.$$

Wegen der Aperiodizität von $\mathcal{X}$ ist $\mathrm{ggT}(A) = 1$. Sei weiter $s, s' \in A$. Dann gilt

$$\begin{aligned}\mathbf{P}(X_{s+s'} = i \mid X_0 = i) &= \sum_{j \in S} \mathbf{P}(X_{s+s'} = i \mid X_s = j) \cdot \mathbf{P}(X_s = j \mid X_0 = i) \\ &\geq \mathbf{P}(X_{s+s'} = i \mid X_s = i) \cdot \mathbf{P}(X_s = i \mid X_0 = i) \\ &= \mathbf{P}(X_{s'} = i \mid X_0 = i) \cdot \mathbf{P}(X_s = i \mid X_0 = i) > 0.\end{aligned}$$

Damit ist $s + s' \in A$ und aus Lemma 8.23 folgt nun die Behauptung im Falle $j = i$.

Falls $j \neq i$, können wir wegen der Irreduzibilität ein s so wählen, dass $\mathbf{P}(X_s = j \mid X_0 = i) > 0$ und s' so, dass $\mathbf{P}(X_{r'} = j \mid X_0 = j) > 0$ für $r' \geq s'$. Dann gilt mit $t = s + s'$ und $r \geq t$, dass

$$\begin{aligned}\mathbf{P}(X_r = j \mid X_0 = i) &\geq \mathbf{P}(X_r = j, X_s = j \mid X_0 = i) \\ &= \mathbf{P}(X_s = j \mid X_0 = i) \cdot \mathbf{P}(X_r = j \mid X_s = j) \\ &= \mathbf{P}(X_s = j \mid X_0 = i) \cdot \mathbf{P}(X_{r-s} = j \mid X_0 = j) > 0\end{aligned}$$

und die Behauptung ist gezeigt. □

Beweis von Theorem 8.19. Sei P die Übergangsmatrix der Markov-Kette. Nach Theorem 8.13 existiert eine stationäre Verteilung π. Es genügt nun, (8.8) zu zeigen. Würde nämlich eine weitere stationäre Verteilung $\pi' \neq \pi$ existieren, und würden wir $\mathcal{X}$ so wählen, dass $\mathbf{P}(X_0 = i) = \pi'_i$. Dann würde aus (8.8) folgen, dass

$$0 = \lim_{t \to \infty} \mathbf{P}(X_t = i) - \pi(i) = \pi'(i) - \pi(i)$$

und damit $\pi' = \pi$.

Es bleibt also (8.8) zu zeigen. Der Beweis verwendet eine Kopplung, also einen stochastischen Prozess $(\mathcal{Y}, \mathcal{Z}) = (Y_t, Z_t)_{t=0,1,2,\ldots}$ mit Wertebereich $S \times S$. Wir wählen $(\mathcal{Y}, \mathcal{Z})$ als homogene Markov-Kette mit $\mathbf{P}(Y_0 = i, Z_0 = j) = \mathbf{P}(X_0 = i) \cdot \pi_j$ und Übergangsmatrix $(Q_{ij,kl})_{ij,kl \in S^2}$,

$$Q_{ij,kl} = \begin{cases} P_{ik} P_{jl}, & i \neq j, \\ P_{ik} \delta_{kl}, & i = j. \end{cases}$$

Wir zeigen zunächst, dass sowohl $\mathcal{Y}$ als auch $\mathcal{Z}$ Markov-Ketten mit Übergangsmatrix P sind. Dazu berechnen wir

$$\begin{aligned}\mathbf{P}(Y_{t+1} = k \mid Y_t = i) &= \sum_{j,l \in S} \mathbf{P}(Y_{t+1} = k, Z_{t+1} = l \mid Y_t = i, Z_t = j) \cdot \mathbf{P}(Z_t = j) \\ &= \sum_{j \in S} \mathbf{P}(Z_t = j) \sum_{l \in S} Q_{ij,kl} = \sum_{j \in S} \mathbf{P}(Z_t = j) P_{ik} = P_{ik}.\end{aligned}$$

Genauso ergibt sich, dass $\mathcal{Z}$ eine Markov-Kette mit Übergangsmatrix P ist.

Wie man aus der Übergangsmatrix Q ablesen kann, laufen in der Markov-Kette $(\mathcal{Y}, \mathcal{Z})$ die beiden Ketten $\mathcal{Y}$ und $\mathcal{Z}$ unabhängig voneinander, bis zu dem Zeitpunkt T, an dem $Y_T = Z_T$. Dann gilt, da $Q_{ij,kl} = P_{ik}\delta_{kl}$ für $i = j$, dass $Y_t = Z_t$ für $t \geq T$. Ein spezielles Beispiel für eine solche Kopplung haben wir bereits in Bemerkung 8.22 gesehen.

Da $\mathcal{Y}$ dieselbe Startverteilung hat wie $\mathcal{X}$ und $\mathcal{Z}$ in der stationären Verteilung π startet, folgt, dass

$$\begin{aligned}|\mathbf{P}(X_s = i) - \pi(i)| &= |\mathbf{P}(Y_s = i) - \mathbf{P}(Z_s = i)| \\ &\leq \mathbf{E}[|1_{Y_s = i} - 1_{Z_s = i}|] \leq \mathbf{P}(Y_s \neq Z_s).\end{aligned} \tag{8.9}$$

Da $\mathbf{P}(Y_s \neq Z_s)$ fallend in s ist, genügt es, eine Teilfolge $s_n \to \infty$ zu finden mit $\mathbf{P}(Y_{s_n} \neq Z_{s_n}) \xrightarrow{n \to \infty} 0$. Nach Proposition 11.25 gibt es ein t mit

$$\alpha := \min\{(P^t)_{ik} : i, k \in S\} > 0.$$

Es gilt für jedes $k \in S$

$$\mathbf{P}(Y_t \neq Z_t) = 1 - \sum_{k \in S} \mathbf{P}(Y_t = Z_t = k) \leq 1 - \mathbf{P}(Y_t = k) \cdot \mathbf{P}(Z_t = k) \leq 1 - \alpha^2$$

Damit ist für jedes $n = 1, 2, \ldots$

$$\begin{aligned}\mathbf{P}(Y_{nt} \neq Z_{nt}) &= \mathbf{P}(Y_t \neq Z_t) \cdot \mathbf{P}(Y_{2t} \neq Z_{2t} \mid Y_t \neq Z_t) \cdots \mathbf{P}(Y_{nt} \neq Z_{nt} \mid Y_{(n-1)t} \neq Z_{(n-1)t}) \\ &\leq (1 - \alpha^2)^n \xrightarrow{n \to \infty} 0.\end{aligned}$$

Mit (8.9) ist die Aussage gezeigt. □

Teil II.

Statistik

9. Grundlegendes

Es ist nicht übertrieben zu behaupten, dass in der heutigen Welt immer mehr *Daten* jeglicher Art erhoben werden. Diese zu ordnen und aus Daten Schlussfolgerungen zu ziehen ist Aufgabe der Statistik.

Man teilt dabei diese Aufgaben in zwei Gebiete auf. Die *deskriptive Statistik* dient rein der Beschreibung der Daten, etwa durch geeignete Wahl von Statistiken, die die Daten zusammenfassen. Anders ist dies bei der hier behandelten *schließenden* oder *induktiven Statistik*. Die Aufgabe ist hier, mit Hilfe von stochastischen Modellen Aussagen darüber zu treffen, welchen Annahmen den Daten zugrunde liegen könnten. Manchmal sagt man auch, dass man anhand der Daten etwas *lernt*, weshalb man oft auch von statistischem Lernen spricht.

> Angenommen, $X_1, ...X_n \sim \mathbb{P}$. Was können wir anhand von den Realisierungen von $X_1, ..., X_n$ über $\mathbb{P}$ aussagen (oder lernen)?

9.1. Ein Beispiel

Wir beginnen mit einem Beispiel, das sich um die Erfolgswahrscheinlichkeit beim Münzwurf dreht: eine Münze wird 53 mal geworfen. Dabei ist die Wahrscheinlichkeit für *Kopf* (was wir im Folgenden als Erfolg werten wollen) noch unbekannt. Von den 53 Würfen sind 23 ein Erfolg.

Unsere statistischen Überlegungen gehen nun von der Vorstellung aus, dass die 53 Münzwürfe die Realisierung einer Zufallsvariable $X = (X_1, \dots, X_{53})$ sind, wobei $X_1, \dots, X_{53}$ unabhängig und identisch verteilt sind mit

$$X_i = \begin{cases} 1, & \text{falls der } i\text{-te Wurf } \textit{Kopf} \text{ zeigt,} \\ 0, & \text{sonst.} \end{cases}$$

und es gilt

$$\mathbb{P}(X_i = 1) = p.$$

Jetzt ist $X_1 + \dots + X_n$ die Gesamtzahl der Erfolge. Als Summe von n unabhängigen Bernoulliverteilten Zufallsvariablen ist diese Summe $B(n = 53, p)$-verteilt. Wichtig ist, dass zwar $n = 53$ bereits fest steht (schließlich wissen wir ja, dass wir 53 mal die Münze geworfen haben), nicht jedoch p. In dieser Situation gibt es zwei *statistische Probleme*.

- *Schätzproblem:* Wir können versuchen, den Erfolgsparameter p zu schätzen. Entweder werden wir dazu einen aus den Daten (23 Erfolge aus 53 Versuchen) abgeleiteten Wert $\widehat{p}$ angeben (*Punktschätzer*), oder ein aus den Daten abgeleitetes Intervall $[a, b]$, in dem der wahre Parameter p mit hoher Wahrscheinlichkeit liegt (*Intervallschätzer*).

- *Testproblem:* Stellen wir uns vor, der Werfer der Münze behauptet, dass die Münze fair ist, also $p = \frac{1}{2}$ gilt. Dieser Meinung können wir skeptisch gegenüber stehen, da ja nur 23 aus 53 Würfen (etwa 43 %) ein Erfolg waren. Wir können versuchen, die Hypothese $p = \frac{1}{2}$ zu testen. Das bedeutet, dass wir untersuchen, wie gut die Hypothese mit den Daten in Einklang steht.

Die Erfolgswahrscheinlichkeit schätzen: Wir versuchen nun, die Erfolgswahrscheinlichkeit des Münzwurfes zu schätzen. Dies muss auf Grundlage der Daten erfolgen, also basierend auf dem Wissen, dass 23 aus 53 Erfolge zu verzeichnen waren. Ein einfacher Ansatz ist es, zu vermuten, dass die 23 Erfolge aus der 53 Würfen in etwa der Erfolgsquote entspricht. Also ist

$$\widehat{p} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{23}{53} \approx 0.43$$

ein Schätzer für den unbekannten Parameter p. (Im Folgenden werden wir einen Schätzer für einen Parameter $\bullet$ meistens mit $\widehat{\bullet}$ bezeichnen.) Ein anderer Ansatz (die wir als der Maximum-Likelihood-Methode bezeichnen werden) wäre der, $\hat{p}$ so zu setzen, dass die Wahrscheinlichkeit, 23 Erfolge zu erzielen, maximal wird. Es ist hier also $p \mapsto \binom{53}{23} p^{23}(1-p)^{30}$ zu maximieren. Wir berechnen

$$\frac{\partial}{\partial p} p^{23}(1-p)^{30} = p^{22}(1-p)^{29}(23(1-p) - 30p),$$

und es ergibt sich $53\hat{p} = 23$ oder wieder $\hat{p} = \frac{23}{53}$.

In beiden Fällen hängt $\widehat{p}$ von den Daten ab, die wir uns als Realisierung von einer Zufallsvariable gedacht haben. Also ist $\widehat{p}$ auch eine Zufallsvariable. Warum ist der Schätzer $\widehat{p}$ gut? Nehmen wir an, wir wüssten den wahren Parameter p. Dann leistet $\widehat{p}$ zumindest im Mittel das gewünschte: (Wir schreiben hier und im Folgenden $\mathbf{P}_p(\cdot)$ und $\mathbf{E}_p[\cdot]$, wenn wir Wahrscheinlichkeiten und Erwartungswerte unter der Hypothese ausrechnen wollen, dass p der wahre Parameter ist.)

$$\mathbf{E}_p[\widehat{p}] = \frac{1}{n}\mathbf{E}_p[X_1 + \cdots + X_n] = p.$$

Wir sagen auch, der Schätzer $\widehat{p}$ ist erwartungstreu (oder unverzerrt oder unbiased).

Eine weitere wünschenwerte Eigenschaft eines Schätzers ist, dass er immer besser wird, je größer die zu Grunde liegende Datenmenge ist. Eine große Datengrundlage bedeutet in unserem Fall, dass die Münze oft geworfen wurde, also n groß ist. Aus dem schwachen Gesetz großer Zahlen wissen wir, dass

$$\mathbf{P}_p(|\widehat{p} - p| \geq \varepsilon) = \mathbf{P}_p\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mathbf{E}_p[X_1]\Big| \geq \varepsilon\Big) \xrightarrow{n\to\infty} 0$$

für alle $\varepsilon > 0$. Die Eigenschaft, dass $\widehat{p}$ mit hoher Wahrscheinlichkeit immer näher am wahren Wert p liegt, wenn mehr Daten zur Verfügung stehen, nennen wir *Konsistenz*.

Stellen wir uns vor, in zwei aufeinander folgenden Experimenten bekommen wir zunächst 23 von 53 Erfolge, und dann 23000 von 53000 Erfolge. In beiden Fällen ist $\widehat{p} = \frac{23}{53}$. Es ist jedoch klar, dass wir dem Wert von $\widehat{p}$ im zweiten Experiment eine viel höhere Bedeutung zumessen und wir uns sicherer sind, dass der wahre Wert in der Nähe von $\widehat{p}$ liegt. Diese Sicherheit können wir mit einem *Intervallschätzer*, also einem Intervall, in dem der wahre

Wert mit hoher Wahrscheinlichkeit liegt, zu Ausdruck bringen. Dazu wählen wir ein (kleines) $\alpha \in (0,1)$, etwa $\alpha = 5\,\%$. Aus dem zentralen Grenzwertsatz folgt, dass es eine nach $N(0,1)$ verteilte Zufallsvariable Z gibt, so dass[1]

$$\mathbf{P}_p\Big(-1.96 \leq \frac{n\widehat{p} - np}{\sqrt{np(1-p)}} \leq 1.96\Big) \approx \mathbf{P}\big(-1.96 \leq Z \leq 1.96\big) \approx 0.95.$$

Weiter ist

$$\begin{aligned} 0.95 &\approx \mathbf{P}_p\Big(-1.96 \leq \frac{\widehat{p} - p}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \leq 1.96\Big) \\ &= \mathbf{P}_p\Big(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n} \leq p \leq \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\Big). \end{aligned}$$

Wir haben gerade gezeigt, dass

$$\big[\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}; \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\big]$$

ein *Konfidenzintervall* für $\widehat{p}$ zum *Niveau* $95\,\%$ ist. Das bedeutet, dass der wahre Wert mit Wahrscheinlichkeit etwa $95\,\%$ in diesem (zufälligen) Intervall liegt. Haben wir 23 Erfolge in $n = 53$ Versuchen, ist unser Konfidenzintervall also $[0.30, 0.57]$. Haben wir hingegen 23000 Erfolge bei 53000 Versuchen, ist das Konfidenzintervall etwa $[0.430, 0.438]$, also wesentlich kleiner.

Die Erfolgswahrscheinlichkeit testen: Nehmen wir an, der Werfer der Münze behauptet, sie sei fair, also $p = \frac{1}{2}$. Können wir diese Hypothese aufgrund der Daten verwerfen? Zunächst stellen wir fest, dass wir prinzipiell zwei Arten von Fehlern mit unserer Entscheidung machen können. Wenn wir die Hypothese verwerfen, könnte sie doch wahr sein, und wenn wir die Hypothese nicht verwerfen, könnte sie doch falsch sein.

Da wir nicht grundlos dem Werfer der Münze widersprechen wollen, wollen wir die Wahrscheinlichkeit, dass wir die Hypothese ablehnen (wir dem Werfer der Münze widersprechen), obwohl sie wahr ist (die Hypothese des Werfers richtig ist), kontrollieren. Das bedeutet, dass

$$\mathbf{P}_{p=1/2}(\text{Hypothese verwerfen}) \leq \alpha$$

für ein anfangs gewähltes $\alpha \in (0,1)$ sein soll. Klar ist, dass damit die Hypothese umso seltener abgelehnt werden kann, je kleiner α ist. Nun kommen wir zu der Regel, mit der wir die Hypothese ablehnen wollen. In unserem Beispiel haben wir für die Hypothese $p = \frac{1}{2}$ zu wenig (23 von 53) Erfolge. Wir würden die Hypothese ablehnen wollen, wenn

$$\mathbf{P}_{p=1/2}(\text{Daten extremer als tatsächliche Daten}) \leq \alpha. \tag{9.1}$$

Wir wählen (wie oben beim Konfidenzintervall) $\alpha = 5\,\%$. Für Y_n nach $B(n = 53, p = \frac{1}{2})$ verteilt, erwarten wir 26.5 Erfolge. Um die Wahrscheinlichkeit einer Abweichung, die größer

[1] Wir schreiben in dieser einführenden Bemerkung öfter approximative Formeln mittels '$\approx$'. Es sei bemerkt, dass dieses Symbol keine mathematisch exakten, beweisbaren Aussagen trifft. Für Anwendungen sinnvolle Resultate liefert es allerdings allemal.

ist als die der Daten zu berechnen, betrachten wir eine nach $N(0,1)$ verteilte Zufallsvariable Z und berechnen

$$\begin{aligned}\mathbf{P}_{p=1/2}\big(|X_1+\cdots+X_n-26.5|\geq 3.5\big)\\ &= 1-\mathbf{P}_{p=1/2}\bigg(-\frac{3.5}{\sqrt{np(1-p)}}<\frac{Y_n-np}{\sqrt{np(1-p)}}<\frac{3.5}{\sqrt{np(1-p)}}\bigg)\\ &\approx 1-\mathbf{P}_{p=1/2}\big(-0.962\leq Z\leq 0.962\big)\approx 33.6\,\%\end{aligned}$$

Da dieser Wert größer als $\alpha = 5\,\%$ ist, kann die Hypothese nicht verworfen werden, siehe (9.1).

9.2. Statistisches Modell und Entscheidungstheorie

Wir beginnen nun mit der Formalisierung der obigen Situation.

Definition 9.1 (Statistisches Modell)**.** *Seien S, Θ, Θ' Mengen. Ein* statistisches Modell *ist ein Paar $(X, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, wobei X eine Zufallsvariable mit Zielbereich S ist, bei deren Verteilung noch ein Parameter $\vartheta\in\Theta$ frei, also unbestimmt, ist. Das bedeutet, dass es eine Funktion $\vartheta\mapsto\rho_\vartheta$ gibt mit*[2]

$$\mathbf{P}_\vartheta(X\in da)=\rho_\vartheta(a)da.$$

Die Menge Θ heißt Parameterraum*, die Menge S* Beobachtungsraum*. Jede Zufallsvariable $h(X)$ mit $h: S\to\Theta'$ heißt* Statistik*.*

Beispiel 9.2 (Parametrische und nicht-parametrische Modelle)**.**

1. Sei $\Theta=\{(\mu,\sigma^2):\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+\}$, $S=\mathbb{R}^n$ und $X=(X_1,...,X_n)$ unabhängig normalverteilut, d.h. $(X_i)_*\mathbb{P}_{(\mu,\sigma^2)}=N(\mu,\sigma^2)$. Dann heißt $(X,(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ das *Normalverteilungsmodell.*

2. Sei $\Theta=[0,1]$, $S=\{0,1\}^n$ und $X=(X_1,...,X_n)$ unabhängig mit $(X_i)_*\mathbb{P}_p=B(1,p)$. Dann heißt $(X,(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ das *Binomialmodell.*

3. Ist in Definition 9.1 die Menge Θ endlich-dimensional, also $\Theta\subseteq\mathbb{R}^d$ für ein $d=1,2,...$, so spricht man von einem parametrischen Modell. (Siehe etwa 1. und 2.) Ist Θ hingegen unendlich-dimensional, so spricht man von einem nicht-parametrischen Modell. Ein Beispiel wäre $\Theta=\{F:\mathbb{R}\to[0,1]\text{ Verteilungsfunktion}\}$, $S=\mathbb{R}^n$ und $X=(X_1,...,X_n)$ unabhängig mit
$$\mathbb{P}_F(X_i\leq x)=F(x).$$

[2] Wir wollen im Folgenden die dauernde Unterscheidung zwischen diskreten Zufallsvariablen und Zufallsvariablen mit Dichten durch die Notation $\mathbf{P}(X\in da)$ vermeiden. Ist der Wertebereich S von X diskret und $a\in S$, ist damit
$$\mathbf{P}(X\in da):=\mathbf{P}(X=a)$$
gemeint. Hat X die Dichte $f(a)da$, ist
$$\mathbf{P}(X\in da):=f(a)da.$$

Definition 9.3 (Statistische Fragestellungen)**.** *Zu einer statischen Fragestellung gehört zunächst einmal ein statistisches Modell $(X, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$. Anhand der Daten X muss man nun Schlussfolgerungen über ϑ ziehen. Deshalb gibt es eine (deterministische oder stochastische) Entscheidungsfunktion $h : S \to \aleph$, wobei $\aleph$ auch Entscheidungsraum heißt. Fällt man eine Entscheidung für $\alpha \in \aleph$, was jedoch falsch ist, erhält man einen Verlust $\ell : \Theta \times \aleph \to \mathbb{R}_+$. Die mittlere Verlustfunktion, nämlich die Abbildung $\vartheta \mapsto \mathbb{E}_\vartheta[\ell(\vartheta, h(X))]$, heißt auch Risikofunktion.*

Bemerkung 9.4 (Schätz- und Testprobleme)**.**

1. Ein Punktschätzer ermittelt anhand der Daten X eine Vorstellung über das zugrunde liegende ϑ. Mit anderen Worten will man eine Schätzfunktion $h : S \to \Theta$ finden, die möglichst gut an den wahren Wert ϑ herankommt.

   Hier ist $\aleph = \Theta$, und meist $\ell(\vartheta, \vartheta') = (\vartheta - \vartheta')^2$. Ein durch h gegebene Schätzer hat dann einen Bias (oder eine Verzerrung) von $\mathbb{E}_\vartheta[h(X)] - \vartheta$ und eine Varianz von $\mathbb{V}_\vartheta[h(X)]$. Man berechnet die Risikofunktion leicht durch

   $$\begin{aligned}\mathbb{E}_\vartheta[(h(X) - \theta)^2] &= \mathbb{E}_\vartheta[h(X)^2] - \mathbb{E}_\vartheta[h(X)]^2 + \mathbb{E}_\vartheta[h(X)]^2 - 2\theta\mathbb{E}_\vartheta[h(X)] + \theta^2] \\ &= \mathbb{V}_\vartheta[h(X)] + (\mathbb{E}_\vartheta[h(X)] - \vartheta)^2,\end{aligned}$$

   also die Summe aus Varianz und dem quadratischen Bias.

2. Ein Testproblem liegt dann vor, wenn man sich anhand der Datenlage für oder gegen eine Behauptung (Hypothese) über ϑ entscheidet.

   Hier ist[3] $\Theta = \Theta_0 \uplus \Theta_1$, $\aleph = \{\Theta_0, \Theta_1\}$ und

   $$\ell(\vartheta, \Theta_i) = \begin{cases} 0, & \vartheta \in \Theta_i, \\ 1, & \vartheta \notin \Theta_i. \end{cases}$$

   Die Risikofunktion ist dann

   $$\mathbb{E}_\vartheta[1_{\vartheta\notin h(X)}] = \mathbb{P}_\vartheta[\vartheta \notin h(X)],$$

   also gerade der Wahrscheinlichkeit, sich (bei Gültigkeit von ϑ) falsch zu entscheiden.

Bemerkung 9.5 (Regression und Klassifikation)**.** Wir betrachten nun den Fall, dass $X = ((X_1, Y_1), ..., (X_n, Y_n))$ aus Beobachtungspaaren mit Zustandsraum $S = (S_X \times S_Y)^n$ besteht. Dabei nennen wir X auch *Prädiktor* und Y *Ausgang*. (Man denke etwa an X =Blutdruck und Y =restliche Lebenszeit.) Wir nehmen an (d.h. unser statistisches Modell modelliert), dass es ein r gibt mit $Y = r(X) + \varepsilon$ mit einer Zufallsvariable ε mit $\mathbb{E}[\varepsilon] = 0$. Das statistische Modell besteht also aus $(X, (\mathbb{P}_r)_{r\in\Theta})$, wobei Θ die Menge aller möglichen Zusammenhänge zwischen X_i und Y_i ist.

1. Ein Vorhersageproblem liegt dann vor, wenn wir für einen neuen Datenpunkt X_{n+1} das zugehörige Y_{n+1} *vorhersagen* wollen. Hier ist also $h(X, X_{n+1})$ die Vorhersage von Y_{n+1}. Die Verlustfunktion ist etwa $(h(X, X_{n+1}) - Y_{n+1})^2$, also gerade der quadratischen Abweichung der Vorhersage vom wahren Wert.

2. Ein Klassifikationsproblem ist ein spezielles Vorhersageproblem im Fall von endlichem S_Y. Hier ist $1_{h(X,X_{n+1})\neq Y_{n+1}}$ der Misklassifikationsfehler.

[3] Wir schreiben $A \uplus B$ für die Vereinigung von A und B, falls $A \cap B = \emptyset$.

Bemerkung 9.6 (Frequentistische und Bayesianische Statistik)**.** Oftmals wird streng zwischen der frequentistischen und der Bayesianischen Statistik unterschieden. Der Hauptunterschied ist der Folgende: Modellparameter (z.B. die Erfolgswahrscheinlichkeit im Münzwurf) sind in der frequentistischen Sichtweise immer deterministisch. In der Bayesianischen Statistik werden sie als Zufallsvariablen modelliert.

Nehmen wir das Beispiel aus Abschnitt 9.1. Vor Beginn der Experimentes wissen wir nichts über den Erfolgsparameter des Münzwurfes. Deshalb nehmen wir an, dass $P \sim U([0,1])$. Führen wir nun 53-mal das Experiment durch, und erlangen hierbei 23 Erfolge, so verändert dies unsere Einschätzung über die Verteilung von P. Es ist $X_1 + \cdots + X_{53} \sim B(53, P)$, wobei $P \sim U([0,1])$ die sogenannte apriori-Verteilung ist. Nach Durchführung des Experimentes ist

$$\mathbb{P}(P \in dp | X_1 + \cdots + X_{53} = 23) = \frac{p^{23}(1-p)^{30}}{\int_0^1 x^{23}(1-x)^{30} dx} dp,$$

was man auch *a posteriori* Verteilung nennt.

10. Einführende Konzepte

10.1. Die multivariate Normalverteilung

Die Normalverteilung hat eine wichtige Stellung in der Statistik. Grund hierfür ist der Zentrale Grenzwertsatz: kommt eine zufällige Größe durch viele unabhängige Einflüsse zu Stande, so ist sie annähernd normalverteilt. Wir stellen nun die mehrdimensionale Normalverteilung vor. Wir erinnern daran, dass ein Wahrscheinlichkeitsmaß $\mathbb{P}$ eine mehrdimensionale Dichte $f : \mathbb{R}^d \to \mathbb{R}$ besitzt, falls

$$\mathbb{P}(B_1 \times \cdots \times B_d) = \int_{B_1} \cdots \int_{B_d} f(x_1, ..., x_d) dx_d \cdots dx_1$$

für alle Intervalle $B_1, ..., B_d$.

Definition 10.1 (Multivariate Normalverteilung)**.** *Sei $\mu \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d\times d}$ eine symmetrische, positiv definite Matrix. Eine $\mathbb{R}^d$-wertige Zufallsvariable heißt normalverteilt mit Erwartungswert(vektor) μ und Covarianz(matrix) Σ, falls sie die Dichte*[1]

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |det(\Sigma)|}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big)$$

besitzt. Wir schreiben dann $X \sim N(\mu, \Sigma)$. Im Falle[2] *$\mu = 0$ und $\Sigma = I_d$ spricht man von einer multivariaten Standard-Normalverteilung.*

Bemerkung 10.2 (Singuläre multivariate Normalverteilungen)**.** Bei der obigen Definition der multivariaten Normalverteilung liegt die Wahrscheinlichkeitsmass auf ganz $\mathbb{R}^d$. Man kann die Definition auch erweitern und erlauben, dass die multivariate Normalverteilung nur auf einem linearen Teilraum von $\mathbb{R}^d$ Wahrscheinlichkeitsmasse legt. In einem solchen Fall wäre Σ nur nicht-negativ definit, d.h. hätte auch verschwindende Eigenwerte. Dann braucht man jedoch maßtheoretische Konstruktionen, um die Verteilung überhaupt hinzuschreiben. Da wir in dieser Vorlesung keine solchen Methoden zur Verfügung haben, begnügen wir uns mit obiger Definition.

Lemma 10.3 (Transformationen von multivariaten Normalverteilungen)**.** *Sei $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ positiv definit, $n \leq d$ mit $b \in \mathbb{R}^n$ und $A \in \mathbb{R}^{n\times d}$ mit vollem Rang. Sei $X \sim N(\mu, \Sigma)$ und $Z = AX + b$. Dann ist $Z \sim N(A\mu + b, A\Sigma A^\top)$.*

Insbesondere gilt[3] *für $A = \Sigma^{-1/2}$ und $b = -A\mu$, dass Z multivariat standard-normalverteilt ist.*

[1] Wir bezeichnen mit $x^\top$ einen Zeilen- und mit x einen Spaltenvektor.

[2] Hier bezeichnet I_d die d-dimensionale Einheitsmatrix.

[3] Ein Ergebnis aus der linearen Algebra besagt, dass man aus einer positiv definiten Matrix Σ eine Wurzel ziehen kann, d.h. es gibt ein A mit $A^2 = \Sigma$.

Beweis. Wir zeigen die Behauptung nur im Fall $n = d$, da der allgemeine Fall etwas mehr Maßtheorie erfordert. Es genügt, den Fall $\mu = 0$ zu betrachten. Der allgemeine Fall folgt dann durch Addition von μ. Wir schreiben mit dem Transformationssatz[4]

$$\begin{aligned}\mathbb{P}(AX \in B_1 \times \cdots \times B_d) &= \int_{A^{-1}(B_1\times\cdots\times B_d)} \frac{1}{\sqrt{(2\pi)^d |\det(\Sigma)|}} \exp\Big(-\frac{1}{2} x^\top A^\top (A\Sigma A^\top)^{-1} Ax\Big) dx \\ &= \int_{A^{-1}(B_1\times\cdots\times B_d)} \frac{\det(A)}{\sqrt{(2\pi)^d |\det(A\Sigma A^\top)|}} \exp\Big(-\frac{1}{2} x^\top A^\top (A\Sigma A^\top)^{-1} Ax\Big) dx \\ &= \int_{B_1\times\cdots\times B_d} \frac{1}{\sqrt{(2\pi)^d |\det(A\Sigma A^\top)|}} \exp\Big(-\frac{1}{2} y^\top (A\Sigma A^\top)^{-1} y\Big) dy,\end{aligned}$$

d.h. $AX \sim N(0, A\Sigma A^\top)$. □

Korollar 10.4. *Sei $X = (X_1, ..., X_d) \sim N(0, I_d)$ und $O \in \mathbb{R}^{d\times d}$ orthogonal. Dann ist $OX \sim N(0, I_d)$.*

Beweis. Nach Lemma 10.3 ist $OX \sim N(0, OI_dO^\top) = N(0, I_d)$. □

10.2. Kopplung

In der Statistik müssen wir häufig zwei Verteilungen vergleichen und etwa abschätzen, ob diese ähnlich sind oder nicht. Eine Möglichkeit, dies zu tun, geben wir nun an.

Definition 10.5 (Metrik der Totalvariation)**.** *Seien $\mathbb{P}$ und $\mathbb{Q}$ Wahrscheinlichkeitsmaße (auf derselben σ-Algebra $\mathcal{A}$). Dann ist*

$$||\mathbb{P} - \mathbb{Q}||_{TV} := \sup_{A\in\mathcal{A}} |\mathbb{P}(A) - \mathbb{Q}(A)|$$

der Totalvariationsabstand *von $\mathbb{P}$ und $\mathbb{Q}$.*

Lemma 10.6 (Eine Darstellung des Totalvariationsabstandes)**.** *Seien $\mathbb{P}$ und $\mathbb{Q}$ Wahrscheinlichkeitsverteilungen. Dann gibt es ein A mit $||\mathbb{P} - \mathbb{Q}||_{TV} = (\mathbb{P} - \mathbb{Q})(A)$. Außerdem gilt, falls $\mathbb{P}$ und $\mathbb{Q}$ auf einem diskreten Raum definiert sind,*

$$||\mathbb{P} - \mathbb{Q}||_{TV} = \tfrac{1}{2} \sum_{x\in S} |\mathbb{P}(x) - \mathbb{Q}(x)|.$$

Beweis. Wir setzen $B := \{x : \mathbb{P}(x) \geq \mathbb{Q}(x)\}$. Es gilt

$$\begin{aligned}\sup_{A\in\mathcal{A}} |\mathbb{P}(A) - \mathbb{Q}(A)| &= \sup_{A\in\mathcal{A}} (\mathbb{P}(A) - \mathbb{Q}(A)) = \sup_{A\in\mathcal{A}} \sum_{x\in A} (\mathbb{P}(x) - \mathbb{Q}(x)) \leq \sum_{x\in B} (\mathbb{P}(x) - \mathbb{Q}(x)) \\ &= \mathbb{P}(B) - \mathbb{Q}(B),\end{aligned}$$

woraus die erste Behauptung folgt. Die zweite folgt dann mit

$$\begin{aligned}\sum_{x\in S} |\mathbb{P}(x) - \mathbb{Q}(x)| &= \sum_{x\in B} (\mathbb{P}(x) - \mathbb{Q}(x)) + \sum_{x\in B^c} (\mathbb{Q}(x) - \mathbb{P}(x)) \\ &= (\mathbb{P}(B) - \mathbb{Q}(B)) + (1 - \mathbb{Q}(B) - 1 + \mathbb{P}(B)) = 2(\mathbb{P}(B) - \mathbb{Q}(B)).\end{aligned}$$

□

[4] Dieser besagt, für eine glatte, bijektive Abbildung Φ, dass $\int_{\Phi(\Omega)} f(y)\,\mathrm{d}y = \int_\Omega f(\Phi(x))\,|\det(D\Phi(x))|\,\mathrm{d}x$

Bemerkung 10.7 (Statistische Interpretation)**.** Angenommen, wir haben uns zu entscheiden, ob Daten X nach $\mathbb{P}$ oder nach $\mathbb{Q}$ verteilt sind. Wir wollen hierfür eine Menge A ausmachen, so dass wir uns bei Vorliegen von $X \in A$ für $\mathbb{P}$ und bei $X \notin A$ für $\mathbb{Q}$ entscheiden. Doch wie wählen wir A? Sinnvollerweise wählen wir dies so, dass der Fehler, den wir bei obiger Entscheidungsregel machen, minimal ist. Ein Fehler passiert dabei immer dann wenn X nach $\mathbb{P}$ verteilt ist, aber $X \notin A$ ist, oder wenn X nach $\mathbb{Q}$ verteilt ist, aber $X \in A$ ist. Es gilt also, $\mathbb{P}(A^c) + \mathbb{Q}(A)$ zu minimieren. Es gilt

$$\inf_{A\in\mathcal{A}} (\mathbb{P}(A^c) + \mathbb{Q}(A)) = 1 - \sup_{A\in\mathcal{A}} \mathbb{P}(A) - \mathbb{Q}(A) = 1 - ||\mathbb{P} - \mathbb{Q}||_{TV}.$$

Also legt der Totalvariationsabstand die mögliche Trennschärfe eines Tests $\mathbb{P}$ gegen $\mathbb{Q}$ fest. Je größer der Abstand, desto besser sind die beiden Maße zu trennen.

Beispiel 10.8 (Ziehen mit und ohne Zurücklegen)**.** Zählt man die Anzahl der Erfolge beim Ziehen mit und ohne Zurücklegen, so wird man vermuten, dass bei großer Gesamtkugelzahl das Zurücklegen keine große Rolle spielt. Um diesen Effekt etwas abzuschätzen, sei $\mathbb{P} = Hyp(N, K, n)$ und $\mathbb{Q} = B(n, K/N)$. Wir berechnen für große N und K und $p = K/N$

$$\begin{aligned}
2||\mathbb{P}-\mathbb{Q}|_{TV} &= \sum_{k=0}^{n} \Big| \frac{\binom{N}{K}\binom{N-K}{n-k}}{\binom{N}{n}} - \binom{n}{k} p^k (1-p)^{n-k} \Big| \\
&= \sum_{k=0}^{n} \binom{n}{k} \Big| \frac{K \cdots (K-k+1)(N-K)\cdots(N-K-n+k+1)}{N\cdots(N-n+1)} - \frac{K^k (N-K)^{n-k}}{N^n} \Big| \\
&\approx \sum_{k=0}^{n} \binom{n}{k} \Big| \frac{K^k \Big(1 - K^{-1}\binom{k}{2}\Big)(N-K)^{n-k}\Big(1-(N-K)^{-1}\binom{n-k}{2}\Big)}{N^n\Big(1-N^{-1}\binom{n}{2}\Big)} - \frac{K^k(N-K)^{n-k}}{N^n} \Big| \\
&\approx \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \Big(\frac{1}{K}\binom{k}{2} + \frac{1}{N-K}\binom{n-k}{2} + \frac{1}{N}\binom{n}{2} \Big) \\
&= \frac{1}{2}\Big(\frac{n(n-1)p^2}{K} + \frac{n(n-1)(1-p)^2}{N-K} + \frac{n(n-1)}{N} \Big) = \frac{n(n-1)}{N}.
\end{aligned}$$

Die $\approx$ gelten dabei jeweils ungefähr im Grenzwert großer N. Man sieht also, dass für große N (und moderate n) das Ziehen mit und ohne Zurücklegen im Sinne des Totalvariationsabstandes ähnlich ist.

Beispiel 10.9 (Abstand von Normalverteilungen)**.** Wir berechnen noch den Totalvariations-Abstand von $\mathbb{P} = N(\mu, 1)$ und $\mathbb{Q} = N(\nu, 1)$, d.h. von zwei Normalverteilungen mit gleicher Varianz. Wir behaupten

$$||\mathbb{P} - \mathbb{Q}||_{TV} = 2\Phi\Big(\frac{|\mu - \nu|}{2}\Big) - 1,$$

wobei $x \mapsto \Phi(x)$ die Verteilungsfunktion von $N(0, 1)$ ist.

Zunächst ist für $\mathbb{P}$ mit Dichte f_μ und $\mathbb{Q}$ mit Dichte f_ν immer – wie im Beweis von Lemma 10.6

$$||\mathbb{P} - \mathbb{Q}||_{TV} = \mathbb{P}(f_\mu > f_\nu) - \mathbb{Q}(f_\mu > f_\nu).$$

Nun ist für $\nu > \mu$

$$f_\mu(x) > f_\nu(x) \iff (x-\mu)^2 < (x-\nu)^2 \iff x(\nu-\mu) < \tfrac{1}{2}(\nu^2-\mu^2) \iff 2x < \nu+\mu.$$

Insgesamt folgt also mit $X_*\mathbf{P} = \mathbb{P}$, $Y_*\mathbf{P} = \mathbb{Q}$ und $Z_*\mathbf{P} = N(0,\sigma^2)$

$$\begin{aligned}||\mathbb{P}-\mathbb{Q}||_{TV} &= \mathbf{P}(2X < \nu+\mu) - \mathbf{P}(2Y < \nu+\mu) = \mathbf{P}(2Z < \nu-\mu) - \mathbf{P}(2Z < \mu-\nu)\\ &= 2\Phi\Big(\frac{|\mu-\nu|}{2}\Big) - 1.\end{aligned}$$

Bereits bei Markov-Ketten haben wir Kopplungen kennengelernt. Diese hängen mit dem Totalvariationsabstand wiefolgt zusammen.

Theorem 10.10. *Seien $\mathbb{P}$ und $\mathbb{Q}$ Wahrscheinlichkeitsmaße (auf einer σ-Algebra $\mathcal{A}$). Weiter seien X, Y auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbf{P})$ definiert mit $X_*\mathbf{P} = \mathbb{P}$ und $Y_*\mathbf{P} = \mathbb{Q}$. Dann gilt*

$$||\mathbb{P}-\mathbb{Q}|| \le \mathbf{P}(X \neq Y).$$

Bemerkung 10.11. Man kann auch noch zeigen, dass es einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbf{P})$ mit $X_*\mathbf{P} = \mathbb{P}$ und $Y_*\mathbf{P} = \mathbb{Q}$ gibt, auf dem Gleichheit gilt.

Beweis. Wir schreiben

$$\begin{aligned}||\mathbb{P}-\mathbb{Q}|| &= \sup_{A\in\mathcal{A}} |\mathbf{E}(1_{X\in A} - 1_{Y\in A})| \le \sup_{A\in\mathcal{A}} \mathbf{E}(|1_{X\in A} - 1_{Y\in A}|) = \sup_{A\in\mathcal{A}} \mathbf{E}(|1_{X\in A} - 1_{Y\in A}|, X \neq Y)\\ &\le \mathbf{P}(X \neq Y).\end{aligned}$$

□

Beispiel 10.12. Sei $\mathbb{P} = B(n,p)$ und $\mathbb{Q} = B(n,q)$ mit $q > p$. Seien $U_1, ..., U_n \sim U([0,1])$ unabhängig und $X = \sum_{i=1}^n 1_{U_i \le p}$ sowie $Y = \sum_{i=1}^n 1_{U_i \le q}$. Nun gilt

$$||\mathbb{P}-\mathbb{Q}||_{TV} \le \mathbf{P}(X \neq Y) = 1 - (1-(q-p))^n \le n(q-p).$$

10.3. Stochastische Ordnung

Wie im letzten Beispiel sei $X \sim \mathbb{P} = B(n,p)$ und $Y \sim \mathbb{Q} = B(n,q)$ mit $q \ge p$. Tendenziell wird also Y größer als X sein, da die Erfolgswahrscheinlichkeit ja höher ist. Basierende auf Daten bedeutet das, dass wir uns bei Vorliegen von vielen Erfolgen eher für das Vorliegen von $\mathbb{Q}$ entscheiden werden, und für $\mathbb{P}$ bei wenigen Erfolgen. Das entsprechende Konzept stellen wir nun vor.

Definition 10.13 (Stochastische Ordnung)**.** *1. Seien $\mathbb{P}$ und $\mathbb{Q}$ Wahrscheinlichkeitsmaße auf $\mathbb{R}$. Dann sagen wir, $\mathbb{Q}$ sei* stochastisch größer *als $\mathbb{P}$, falls $\mathbb{Q}([t;\infty)) \ge \mathbb{P}([t;\infty))$ für alle $t \in \mathbb{R}$.*

2. Eine Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ mit $\Theta \subseteq \mathbb{R}$ von Wahrscheinlichkeitsmaßen heißt stochastisch wachsend in ϑ*, falls $\mathbb{P}_\vartheta$ für $\vartheta \ge \vartheta'$ stochastisch größer ist als $\mathbb{P}_{\vartheta'}$.*

3. Seien X, Y reellwertige Zufallsvariable. Dann heißt Y stochastisch größer als X, falls $Y_*\mathbf{P}$ stochastisch größer ist als $X_*\mathbf{P}$, d.h. $\mathbf{P}(Y \ge t) \ge \mathbf{P}(X \ge t)$ für alle $t \in \mathbb{R}$.

Bemerkung 10.14 (Stochastische Ordnung und Verteilungsfunktionen)**.** Seien $F_{\mathbb{P}}$ und $F_{\mathbb{Q}}$ die Verteilungsfunktionen von $\mathbb{P}$ und $\mathbb{Q}$. Es ist genau dann $\mathbb{Q}$ stochastisch größer als $\mathbb{P}$, wenn $F_{\mathbb{P}} \leq F_{\mathbb{Q}}$.

Lemma 10.15 (Stochastische Ordnung und Kopplung)**.** *Seien $\mathbb{P}$ und $\mathbb{Q}$ zwei Wahrscheinlichkeitsmaße auf $\mathbb{R}$. Dann sind folgende Aussagen äquivalent:*

1. *$\mathbb{Q}$ ist stochastisch größer als $\mathbb{P}$.*
2. *Es gibt es einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ und Zufallsvariable X, Y mit $X_*\mathbf{P} = \mathbb{P}, Y_*\mathbf{P} = \mathbb{Q}$ und $X \leq Y$.*

Beweis. 2.⇒1.: Es gilt für alle $t \in \mathbb{R}$

$$\mathbb{Q}([t, \infty)) = \mathbf{P}(Y \geq t) \geq \mathbf{P}(X \geq t) = \mathbb{P}([t, \infty)).$$

1.⇒2.: Seien $F_{\mathbb{P}}$ und $F_{\mathbb{Q}}$ die Verteilungsfunktionen von $\mathbb{P}$ und $\mathbb{Q}$. Nach Voraussetzung ist $F_{\mathbb{Q}} \leq F_{\mathbb{P}}$. Sei $U_*\mathbf{P} = U([0,1])$, d.h. U ist $U([0,1])$-verteilt. Wir definieren $X := \inf\{t : F_{\mathbb{P}}(t) \geq U\}$ und $Y := \inf\{t : F_{\mathbb{Q}}(t) \geq U\}$. Dann gilt

$$\begin{aligned}\mathbf{P}(X \leq t) &= \mathbf{P}(F_{\mathbb{P}}(t) \geq U) = F_{\mathbb{P}}(t),\\ \mathbf{P}(Y \leq t) &= \mathbf{P}(F_{\mathbb{Q}}(t) \geq U) = F_{\mathbb{Q}}(t),\end{aligned}$$

und wegen $\{t : F_{\mathbb{P}}(t) \geq U\} \supseteq \{t : F_{\mathbb{Q}}(t) \geq U\}$ ist auch $X \leq Y$. □

Wir behandeln nun noch eine spezielle Situation, die wir später genauer betrachten werden.

Lemma 10.16. *Sei X eine reellwertige Zufallsvariable mit Dichte f. Falls $f(x) = f(-x)$ für alle x und 0 das einzige Maximum von f ist, dann ist $(X + \mu)^2$ stochastisch wachsend in μ.*

Beweis. Sei F die Verteilungsfunktion von X. Für $t \geq 0$ gilt

$$\begin{aligned}\frac{\partial}{\partial\mu}\mathbf{P}((X+\mu)^2 \geq t^2) = \frac{\partial}{\partial\mu}\mathbf{P}(|X+\mu| \geq t) &= \frac{\partial}{\partial\mu}F(-t+\mu) + F(-t-\mu)\\ &= f(t-\mu) - f(t+\mu) \geq 0.\end{aligned}$$

Daraus folgt, dass $\mu \mapsto \mathbf{P}((X+\mu)^2 \geq t^2)$ wachsend in μ ist. □

Definition 10.17 (χ^2-Verteilung)**.** *Seien $X_1, ..., X_d \sim N(0,1)$ unabhängig und $\mu \in \mathbb{R}^d$. Dann heißt die Verteilung von $Y = \sum_{i=1}^d (\mu_i + X_i)^2$ unzentrierte χ^2-Verteilung mit d Freiheitsgraden und Unzentriertheit $|\mu|^2$. Wir schreiben auch $Y \sim \chi^2_d(|\mu|^2)$.*

Bemerkung 10.18. In der Tat ist zunächst unklar, ob die Verteilung von Y in obiger Definition nur von $|\mu|^2$, und nicht vom gesamten Vektor μ abhängt. Wir beweisen nun

$$\sum_{i=1}^d (\mu_i + X_i)^2 \sim (|\mu| + X_1)^2 + \sum_{i=2}^d X_i^2.$$

Es sei bemerkt, dass hier die rechte Seite nur von $|\mu|^2$ abhängt.
Es gibt eine orthogonale Matrix O mit $O\mu = |\mu|e_1$. Nun ist $OX \sim N(0, I_d)$, also gilt

$$\sum_{i=1}^d (\mu_i + X_i)^2 = |\mu + X|^2 = |O\mu + OX|^2 \sim ||\mu|e_1 + X| = (|\mu| + X_1)^2 + \sum_{i=2}^d X_i^2.$$

Theorem 10.19. *Die Familie $(\chi^2_d(|\mu|^2))_{|\mu|^2}$ ist stochastisch wachsend.*

Beweis. Zunächst behaupten wir folgendes:

> Seien X, Y, Z Zufallsvariable, so dass X, Y und X, Z unabhängig sind. Ist dann Y stochastisch größer als Z, so ist auch $X + Y$ stochastisch größer als $X + Z$. (10.1)

(Für den Beweis hierfür siehe Übung.) Nun ist nach Bemerkung 10.18 $(|\mu|+X_1)^2+\sum_{i=2}^d X_i^2 \sim \chi^2_d(|\mu|^2)$ als Summe zweier unabhängiger Zufallsvariablen. Nach Lemma 10.16 sind die Verteilungen von $(|\mu|+X_1)^2$ stochastisch wachsend in $|\mu|^2$. Nun folgt die Behauptung aus (10.1). □

10.4. Suffizienz

In unserem einführenden Beispiel wollten wir den Erfolgsparameter p eines p-Münzwurfes $X = (X_1, ..., X_{53})$ schätzen, wobei $X_1 + \cdots + X_{53} = 23$ war. Da wir nur Kenntnis dieser Summe hatten, jedoch nicht von den einzelnen Münzwürfen, können wir uns fragen, ob wir den Schätzer von $\hat{p} = 23/53$ verbessern können, wenn wir genauere Kenntnis über die einzelnen Würfe haben. Dies ist nicht der Fall, wie wir sehen werden. Der Grund dafür ist, dass $X_1 + \cdots + X_{53}$ eine für p suffiziente Statistik ist.

Definition 10.20 (Suffiziente Statistik)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\mathbf{P}(X \in da) = \rho_\vartheta(da)$. Eine Zufallsvariable $t(X)$ mit Zielbereich $\widetilde{S}$ heißt suffiziente Statistik für ϑ, falls für alle $b \in \widetilde{S}$*

$$\mathbf{P}_\vartheta(X \in \cdot \mid t(X) \in dt) \text{ nicht von } \vartheta \text{ abhängt.}$$

Beispiel 10.21 (Münzwurf)**.** Sei $X = (X_1, \ldots, X_n) \in \{0, 1\}^n$ ein p-Münzwurf mit noch unbestimmtem p. Der statistische Raum ist also $(X, (\mathbf{P}_p)_{p\in[0,1]})$, wobei

$$\mathbf{P}_p(X = (x_1, \ldots, x_n)) = p^k(1-p)^{n-k}, \text{ falls } k = \sum_{i=1}^n x_i.$$

Die Statistik

$$t(X_1, \ldots, X_n) = X_1 + \cdots + X_n$$

ist suffizient für p. Denn es gilt für $k = x_1 + \cdots + x_n$

$$\begin{aligned}\mathbf{P}_p(X = (x_1, \ldots, x_n) \mid X_1 + \cdots + X_n = k) &= \frac{\mathbf{P}_p(X = (x_1, \ldots, x_n))}{\mathbf{P}_p(X_1 + \cdots + X_n = k)} = \frac{p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} \\ &= \frac{1}{\binom{n}{k}},\end{aligned}$$

unabhängig von p. Für $k \neq \sum_{i=1}^n x_i$ gilt

$$\mathbf{P}_p(X = (x_1, \ldots, x_n) \mid X_1 + \cdots + X_n = k) = 0,$$

ebenfalls unabhängig von p.

Beispiel 10.22 (Uniformes Modell)**.** Wir betrachten das statistische Modell $(X, (\mathbf{P}_\vartheta)_{\vartheta=[0,\infty)})$, wobei $X = (X_1, ..., X_n)$ ein unabhängiger Vektor ist mit $(X_i)_{\mathbf{P}_\vartheta = U([0,\vartheta)}$, d.h. $X_1, ..., X_n$ sind unabhängig und uniform auf $[0, \vartheta]$ verteilt. Wir behaupten nun, dass

$$t(X) := \sup_{i=1,...,n} X_i$$

für ϑ suffizient ist.
Denn: Es gilt etwa für $x < y$

$$\mathbf{P}_\vartheta(X_1 \leq x | t(X) = y) = \frac{x/\vartheta \cdot (n-1)y^{n-2}/\vartheta^{n-1}}{ny^{n-1}/\vartheta^n} = \frac{n-1}{n}\frac{x}{y}$$

unabhängig von ϑ.

Theorem 10.23 (Fisher-Neyman'scher Faktorisierungssatz)**.** *Sei $(X, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $T = t(X)$ mit $t : S \to S'$. Dann sind äquivalent:*

1. *T ist suffizient,*
2. *Es gibt $g_\theta : S' \to \mathbb{R}$ und $h : S \to \mathbb{R}$, so dass*

$$\mathbf{P}_\vartheta(X \in dx) = g_\vartheta(t(x))h(x)dx.$$

Beweis. Wir beweisen die Aussage nur im diskreten Fall. '2. $\Rightarrow$ 1.': Zunächst gilt $\mathbf{P}_\vartheta(X = x|t(X) = t) = 0$ für $t \neq t(x)$, was unabhängig von ϑ ist. Für $t = t(x)$ hingegen ist unter 2.

$$\mathbf{P}_\vartheta(X = x|t(X) = t) = \frac{g_\vartheta(t(x))h(x)}{\sum_{y:t(y)=t} g_\vartheta(t(y))h(y)} = \frac{g_\vartheta(t(x))h(x)}{g_\vartheta(t(x))\sum_{y:t(y)=t} h(y)} = \frac{h(x)}{\sum_{y:t(y)=t} h(y)},$$

was ebenfalls unabhängig von ϑ ist. Für '1. $\Rightarrow$ 2.' setzen wir

$$g_\vartheta(t) := \mathbf{P}_\vartheta(t(X) = t), \qquad h(x) = \mathbf{P}_\vartheta(X = x|t(X) = t(x)).$$

Dann ist $h(x)$ nach Voraussetzung unabhängig von ϑ und es gilt

$$\mathbf{P}_\vartheta(X = x) = \mathbf{P}_\vartheta(X = x, t(X) = t(x)) = h(x)g_\vartheta(t(x))$$

und die Behauptung ist gezeigt. □

Beispiel 10.24 (Münzwurf)**.** Im Beispiel des Münzwurfs aus Beispiel 10.21 ist $t(X) = X_1 + \cdots + X_n$. Hier ist

$$\mathbf{P}_\vartheta(X_1 = x_1, ..., X_n = x_n) = \vartheta^{t(X)}(1-\vartheta)^{n-t(X)},$$

woraus sich die Darstellung aus Theorem 10.23.2 mit $h = 1$ ergibt.

Beispiel 10.25 (Uniformes Modell)**.** Im uniformen Modell aus Beispiel 10.22 ist $t(X) = \sup_{i=1,...,n} X_i$ und

$$\mathbf{P}_\vartheta(X_1 \in dx_1, ..., X_n \in dx_n) = \frac{1}{\vartheta^n} 1_{X_1,...,X_n\in[0,\vartheta)} = \frac{1}{\vartheta^n} 1_{\sup_{i=1,...,n} X_i<\vartheta} \cdot 1_{\inf_{i=1,...,n} X_i\geq 0}.$$

Nun folgt die Suffizienz von $t(X)$ mit $h(x) = 1_{\inf_{i=1,...,n} x_i\geq 0}$ aus Theorem 10.23.

10.5. Exponentialfamilien

Viele Verteilungen, etwa die Normal- Poisson-, Binomial- und Exponentialverteilung, haben eine gemeinsame Struktur, die oftmals direkte Rechnungen ermöglicht. Diese Struktur wird in der folgenden Definition formalisiert.

Definition 10.26 (Exponentialfamilie)**.** *Sei $\Theta \subseteq \mathbb{R}^k$. Ein parametrisches statistisches Modell $(X, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt* k-parametrige Exponentialfamilie *(mit c, t, d, h) für*

$$c_1, ..., c_k, d : \Theta \to \mathbb{R}, \qquad t_1, ..., t_k, h : \mathbb{R}^n \to \mathbb{R},$$

falls

$$\mathbf{P}_\vartheta(X \in dx) = h(x) \cdot \exp\Big(\sum_{j=1}^{k} c_j(\vartheta)t_j(x) - d(\vartheta)\Big)dx = h(x) \cdot \exp\big(c(\vartheta)^\top t(x) - d(\vartheta)\big)dx.$$

Gilt insbesondere $c_j(\vartheta) = \vartheta_j$, also

$$p_\vartheta(x) = h(x) \cdot \exp\big(\vartheta^\top t(x) - d(\vartheta)\Big), \qquad x \in \mathbb{R}^n,$$

so sagt man, die Exponentialfamilie sei in kanonischer Form*.*

Bemerkung 10.27 (1-parametrige Exponentialfamilie)**.** Für den Spezialfall einer 1-parametrigen Exponentialfamilie gibt es Funktionen c, d, t, h mit

$$p_\vartheta(x) = h(x) \cdot \exp\Big(c(\vartheta)t(x) - d(\vartheta)\Big), \qquad x \in \mathbb{R}^n.$$

So ziemlich alle statistischen Modelle, die auf uns bekannten Verteilungen basieren, sind Exponentialfamilien:

Beispiel 10.28. Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell.

1. Ist $\Theta = [0, 1]$ und $X = (X_1, ..., X_n)$ mit $X_1, ..., X_n$ unabhängig und $(X_i)_*\mathbf{P}_\vartheta = B(1, \vartheta)$, (d.h. wir betrachten das Binomialmodell aus Bemerkung 9.2), so gilt

$$\begin{aligned}\mathbf{P}_\vartheta(X = (x_1, ..., x_n)) &= \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i} \\ &= \exp\Big(\sum_{i=1}^n x_i \log\vartheta + \Big(n - \sum_{i=1}^n x_i\Big)\log(1-\vartheta)\Big) \\ &= \exp\Big(\log\frac{\vartheta}{1-\vartheta}\sum_{i=1}^n x_i + n\log(1-\vartheta)\Big)\end{aligned}$$

   und damit haben wir es mit einer 1-parametrigen Exponentialfamilie mit

$$c(\vartheta) = \log\frac{\vartheta}{1-\vartheta}, \qquad t(x) = \sum_{i=1}^n x_i, \qquad d(\vartheta = -n\log(1-\vartheta)$$

   zu tun.

2. Ist $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$ und $X_* \mathbf{P}_{(\mu,\sigma^2)} = N(\nu, \sigma^2)$ (d.h. wir betrachten das Normalverteilungsmodell aus Bemerkung 9.2), so ist

$$\begin{aligned}\mathbf{P}_{(\mu,\sigma^2)}(X \in dx) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)dx \\ &= \exp\Big(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\Big(\frac{\mu}{\sigma^2} + \log(2\pi\sigma^2)\Big)\Big)dx.\end{aligned}$$

Also ist die Familie der (ein-dimensionalen) Normalverteilungen eine 2-parametrige Exponentialfamilie mit

$$\begin{aligned} c_1(\mu,\sigma^2) &= \frac{\mu}{\sigma^2}, & t_1(x) &= x, \\ c_2(\mu,\sigma^2) &= -\frac{1}{2\sigma^2}, & t_2(x) &= x^2, \\ h(x) &= 1, & d(\mu,\sigma^2) &= -\frac{1}{2}\Big(\frac{\mu}{\sigma^2} + \log(2\pi\sigma^2)\Big). \end{aligned}$$

Diese ist nun allerdings nicht in kanonischer Form.

3. Ist $\Theta = [0, \infty)$ und $X = (X_1, ..., X_n)$ mit $X_1, ..., X_n$ unabhängig und $(X_i)_* \mathbf{P}_\vartheta = U([0, \vartheta))$ (d.h. wir betrachten das uniforme Modell), so handelt es sich nicht um eine Exponentialfamilie.

Bemerkung 10.29 (Stichprobe aus einer Exponentialfamilie)**.** Ist $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ eine k-parametrige Exponentialfamilie mit Funktionen $c_1, ..., c_k, d, t_1, ..., t_k, h$, und sind $X_1, ..., X_n$ unabhängig und identisch nach $\mathbb{P}_\vartheta$ verteilt. Dann ist $((X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$, (d.h. die gemeinsame Verteilung von $X_1, ..., X_n$) ebenfalls eine Exponentialfamilie mit

$$c_1, ..., c_k, Nd, \sum_{i=1}^N t_1 \circ \pi_i, ..., \sum_{i=1}^N t_k \circ \pi_i, \prod_{i=1}^N h \circ \pi_i$$

mit der Projektion $\pi_i(x) = x_i$.
Denn: Wir schreiben

$$\mathbf{P}(X_1 \in dx_1, ..., X_n \in dx_n) = h(x_1) \cdots h(x_n) \cdot \exp\Big(\sum_{j=1}^k c_j(\vartheta) \sum_{i=1}^n t_j(x_i) - nd(\vartheta)\Big).$$

Proposition 10.30 (Suffiziente Statistik bei Exponentialfamilien)**.**
Sei $(X, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ eine Exponentialfamilie (mit c, t, d, h). Dann ist die Statistik $T := t(X) = (t_1(X), ..., t_k(X))$ suffizient.

Beweis. Um die Suffizienz von T zu sehen, schreiben wir zunächst

$$\mathbf{P}_\vartheta(X \in dx) = h(x) g_\vartheta(t(x))$$

für

$$g_\vartheta(t(x)) = \exp(c(\vartheta)^\top t(x) - d(\vartheta)).$$

Damit folgt die Aussage aus dem Fisher-Neyman'schen Faktorisierungssatz, Theorem 10.23. □

10.6. Bayes'sche Modelle

Die Formel von Bayes ist wohlbekannt. Auf ihr beruht der große Zweig der Bayesianischen Statistik. Grundlegend ist hier, dass in einem statistischen Modell $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein Vorwissen über die Möglichkeiten besteht, welcher Parameter $\vartheta \in \Theta$ zutrifft. Dies wird in der a-priori-Verteilung zusammengefasst, einer Verteilung auf Θ.

Definition 10.31 (A-priori-Verteilung, a-posteriori-Verteilung)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$. Eine* a-priori-Verteilung *ist die Verteilung π einer Zufallsvariable Ξ auf Θ. In diesem Fall wird durch*

$$\mathbf{P}(\Xi \in A, X \in B) := \int_A \mathbf{P}(\Xi \in d\vartheta)\mathbf{P}_\vartheta(X \in B)$$

die gemeinsame Verteilung von Ξ und X auf $\Theta \times S$ definiert. Die a-posteriori-*Verteilung ist dann die Verteilung*

$$\pi_x(d\vartheta) := \mathbf{P}(\Xi \in d\vartheta | X = x).$$

Bemerkung 10.32 (A-posteriori-Verteilung bei diskreten und stetigen Modellen)**.**

1. Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\Theta \subseteq \mathbb{Z}^k$ (also diskret), und Ξ eine Θ-wertige Zufallsvariable. Für die a-posteriori Verteilung gilt dann

   $$\mathbf{P}(\Xi = \vartheta | X = x) = \frac{\mathbf{P}(\Xi = \vartheta) \cdot \mathbf{P}_\vartheta(X \in dx)}{\sum_{\eta\in\Theta} \mathbf{P}(\Xi = \eta)\mathbf{P}_\eta(X \in dx)}.$$

   Wir bemerken, dass der Nenner nicht von ϑ abhängt und damit nur eine Normierungskonstante darstellt. Deshalb ist es äquivalent,[5]

   $$\mathbf{P}(\Xi = \vartheta | X = x) \sim_x \mathbf{P}(\Xi = \vartheta) \cdot \mathbf{P}_\vartheta(X \in dx)$$

   zu schreiben.

2. Ist $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\Theta \subseteq \mathbb{R}^k$, und ist Ξ eine Θ-wertige Zufallsvariable und $\Xi \sim \pi$ mit Dichte g, so hat die gemeinsame Verteilung von Ξ und X die Dichte $g(d\vartheta) \cdot \mathbf{P}_\vartheta(X \in dx)$. Für die a-posteriori Verteilung gilt dann

   $$\mathbf{P}(\Xi \in d\vartheta | X = x) = \frac{g(\vartheta) \cdot \mathbf{P}_\vartheta(X \in dx)}{\int g(\eta)\mathbf{P}_\eta(X \in dx)d\eta} d\vartheta,$$

   also

   $$\mathbf{P}(\Xi \in d\vartheta | X = x) \sim g(\vartheta) \cdot \mathbf{P}_\vartheta(X \in dx).$$

Es stellt sich heraus, dass die a-priori-Verteilung und die a-posteriori-Verteilung gerade bei Exponentialfamilien einen besonderen Zusammenhang haben. Dies unterstreicht nochmal die gute Handhabbarkeit dieser Verteilungen.

[5] Wir schreiben $a \sim_x b$, falls a/b nur von x abhängt (d.h. a und b sind proportional).

Proposition 10.33 (Konjugierte Familie bei Exponentialfamilien)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ eine 1-parametrige Exponentialfamilie (mit c, t, d, h) mit Dichte. Weiter sei π die a-priori-Verteilung einer $\mathbb{R}$-wertigen Zufallsvariable Ξ mit Dichte*

$$\mathbf{P}_{(r,s)}(\Xi \in d\vartheta) = \frac{\exp\Big(c(\vartheta)r - sd(\vartheta)\Big)}{\int \exp\Big(c(\eta)r - sd(\eta)\Big)d\eta} \sim \exp\Big(c(\vartheta)r - sd(\vartheta)\Big).$$

Dann ist die a-posteriori-Verteilung

$$\mathbf{P}(\Xi \in d\vartheta | X = x) = \mathbf{P}_{(r+t(x),s+1)}(\Xi \in d\vartheta).$$

Beweis. Es gilt

$$\begin{aligned}\mathbf{P}(\Xi \in d\vartheta | X = x) \sim_x \mathbf{P}_{(r,s)}(\Xi \in d\vartheta)\mathbf{P}_\vartheta(X \in dx) &\sim_x \exp\Big(c(\vartheta)(t(x)+r) - (s+1)d(\vartheta)\Big)\\ &\sim \mathbf{P}_{(r+t(x),s+1)}(\Xi \in d\vartheta).\end{aligned}$$

□

Beispiel 10.34 (A-posteriori-Verteilung bei der Normalverteilung)**.** Wir betrachten das Normalverteilungsmodell bei bekanntem σ^2, d.h. das statistische Modell $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ mit $X_1, ..., X_n$ unabhängig und $(X_i)_*\mathbf{P}_\vartheta = N(\vartheta, \sigma^2)$. Also ist

$$\mathbf{P}_\vartheta(X_1 \in dx) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\vartheta)^2}{2\sigma^2}\Big) \sim_x \exp\Big(\frac{\vartheta x}{\sigma^2} - \frac{\vartheta^2}{2\sigma^2}\Big)$$

Nach Bemerkung 10.29 ist damit $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ eine Exponentialfamilie mit

$$t(x) = \sum_{i=1}^n x_i, \qquad c(\vartheta) = \vartheta/\sigma^2 \qquad \text{und} \qquad d(\vartheta) = n\vartheta^2/(2\sigma^2).$$

Angenommen, Ξ ist die a-priori-verteilte Zufallsvariable mit Verteilung $\mathcal{N}(a, b^2)$ für $a, b \in \mathbb{R}$. Wie sieht dann die a-posteriori-Verteilung aus? Und ist diese um den wahren Wert ϑ konzentriert?
Um dies zu beantworten, verwenden wir Proposition 10.33. Wir haben für die a-priori-Verteilung

$$\mathbf{P}(\Xi \in d\vartheta) \sim \exp(-(\vartheta - a)^2/2b^2) \sim \exp(-\vartheta^2/2b^2 + \vartheta a/b^2).$$

Setzen wir

$$r = \sigma^2 a/b^2, \qquad s = \sigma^2/(nb^2),$$

so ist die a-priori-Verteilung also

$$\mathbf{P}_{(r,s)}(\Xi \in d\vartheta) \sim \exp\Big(\frac{\vartheta}{\sigma^2}r - s\frac{n\vartheta^2}{2\sigma^2}\Big).$$

Die a-posteriori-Verteilung ergibt sich damit mit dem Mittelwert $\overline{x}$ zu

$$\begin{aligned}\mathbf{P}(\Xi \in d\vartheta | X = x) = \mathbf{P}_{(r+t(x),s+1)}(\Xi \in d\vartheta) &\sim \exp\Big(\frac{\vartheta a}{b^2} + \frac{\vartheta}{\sigma^2}(x_1 + \cdots + x_n) - \frac{n\vartheta^2}{2\sigma^2} - \frac{\vartheta^2}{2b^2}\Big)\\ &\sim \exp\Big(-\frac{(\vartheta - \overline{x})^2}{2\sigma^2/n} - \frac{(\vartheta - a)^2}{2b^2}\Big) = \exp\Big(-\frac{(\vartheta - \alpha)^2}{2\beta^2}\Big)\end{aligned}$$

für

$$\alpha = \frac{\frac{\overline{x}}{\sigma^2/n} + \frac{a}{b^2}}{\frac{1}{\sigma^2/n} + \frac{1}{b^2}} = \frac{1}{\sigma^2/(nb^2) + 1}\overline{x} + \frac{\sigma^2/(nb^2)}{\sigma^2/(nb^2) + 1}a,$$
$$\beta = \frac{1}{n/\sigma^2 + 1/b^2}.$$

Damit ist gezeigt, dass die a-posteriori-Verteilung für große n um $\overline{x}$ konzentriert ist.

11. Schätzprobleme

Der in einem statistischen Modell freie Parameter ϑ (oder ein $h(\vartheta)$) kann aus Daten, d.h. der Realisierung der Zufallsvariable X geschätzt werden. Führt so eine Schätzung auf einen einzigen Wert, sprechen wir von Punktschätzern.

Definition 11.1 (Punktschätzer, unverzerrte und konsistente Schätzer)**.**

1. *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $m : \Theta \to \Theta'$. Jede Statistik $\widehat{m}(X)$ mit $\widehat{m} : S \to \Theta'$ heißt (Punkt-)Schätzer für $m(\vartheta)$. (Wir schreiben im Folgenden auch $m(\mathbf{P}_\vartheta) \equiv m(\vartheta)$.)*

   Der Bias *oder die* Verzerrung *von $\widehat{m}(X)$ ist gegeben als*

   $$b(\vartheta, m, \widehat{m}) := \mathbf{E}_\vartheta[\widehat{m}(X)] - m(\vartheta).$$

   Ist $b(\vartheta, m, \widehat{m}) = 0$ für alle ϑ, so sagt man, $\widehat{m}$ ist ein unverzerrter *(*erwartungstreuer, unbiased*) Schätzer für m.*

2. *Sei $(X^n, (\mathbf{P}^n_\vartheta)_{\vartheta\in\Theta})_{n=1,2,\dots}$ eine Folge statistischer Modelle mit derselben Parametermenge Θ und $\widehat{m}_1(X^1), \widehat{m}_2(X^2), \dots$ eine Folge von Schätzern für $m(\vartheta)$. Die Folge $\widehat{m}_1(X^1), \widehat{m}_2(X^2), \dots$ heißt konsistent, falls*

   $$\mathbf{P}^n_\vartheta[|\widehat{m}_n(X^n) - m(\vartheta)| \geq \varepsilon] \xrightarrow{n\to\infty} 0$$

   für alle $\varepsilon > 0, \vartheta \in \Theta$.

Bemerkung 11.2 (Varianz eines Schätzers)**.** Die Varianz eines Schätzers ist gegeben als

$$v(\vartheta, \widehat{m}) := \mathbf{V}_\vartheta[\widehat{m}(X)].$$

Mit Bemerkung 9.4 sehen wir, dass

$$\mathbf{E}_\vartheta[(\widehat{m}_n(X^n) - m(\vartheta))^2] = v(\vartheta, \widehat{m}_n) + b(\vartheta, m, \widehat{m})^2.$$

Mit der Chebycheff-Ungleichung (Proposition 6.5) ist also eine Folge $\widehat{m}_1(X^1), \widehat{m}_2(X^2), \dots$ von Schätzern für $m(\vartheta)$ insbesondere dann konsistent, wenn sie approximativ (d.h. im Grenzwert großer n) unverzerrt ist und

$$v(\vartheta, \widehat{m}_n) \xrightarrow{n\to\infty} 0$$

gilt.

Beispiel 11.3 (Erfolgswahrscheinlichkeit beim Münzwurf)**.** Betrachten wir noch einmal das einführende Beispiel der Schätzung des Erfolgsparameters in einem Münzwurf $X = (X_1, ..., X_n)$ in n Versuchen, d.h. $X_1, ..., X_n$ sind unabhängig und $B(1, p)$-verteilt. Wir betrachten zwei Schätzer, nämlich

$$\widehat{p}(X) := \overline{X}, \qquad \widehat{p}'(X) = X_1.$$

Man beachte, dass beide unverzerrt sind, denn

$$\mathbf{E}_p[\widehat{p}] = \mathbf{E}_p[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_p[X_i] = \mathbf{E}_p[X_1] = \mathbf{E}_p[\widehat{p}'] = p.$$

Allerdings ist

$$\mathbf{V}_p[\widehat{p}] = \mathbf{V}_p\Big[\frac{X_1+\cdots+X_n}{n}\Big] = \tfrac{1}{n^2}\sum_{i=1}^{n}\mathbf{V}_p[X_i] = \tfrac{1}{n}\mathbf{V}[X_1] = \tfrac{1}{n}p(1-p),$$
$$\mathbf{V}_p[X_1] = p(1-p),$$

und damit hat $\widehat{p}$ eine kleinere Varianz als $\widehat{p}'$ (was nicht erstaunen sollte).

11.1. Plugin- und momentenbasierte Schätzer

Wir behandeln nun grundlegende Prinzipien des Schätzens. Diese sind Plugin-Schätzer (Definition 11.4), den Spezialfall von momentenbasierten Schätzern (Definition 11.8). Im nächsten Abschnitt behandeln wir dann Maximum-Likelihood-Schätzer (Definition 11.10).

Definition 11.4 (Empirische Verteilung, Plugin-Schätzer)**.** *Sei $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Dann heißt die (zufällige, diskrete) Wahrscheinlichkeitsverteilung auf $\{X_1, ..., X_n\}$, gegeben durch*

$$\widehat{\mathbf{P}}_X(A) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i\in A}$$

die empirische Verteilung *von X. (Es hat also $\widehat{\mathbf{P}}_X$ ein Wahrscheinlichkeitsgewicht von $1/n$ auf $X_1, ..., X_n$, d.h. für $Z \sim \widehat{\mathbf{P}}_X$ ist $\widehat{\mathbf{P}}_X(Z = x_i) = 1/n, i = 1, ..., n$)*

Für $m : \Theta \to \Theta'$ heißt

$$\widehat{m}(X) := m(\widehat{\mathbf{P}}_X)$$

Plugin-Schätzer *für m.*

Beispiel 11.5 (Plugin-Schätzer für Erwartungswert und Varianz)**.** Seien $X_1, ..., X_n$ unter $\mathbf{P}_\vartheta$ identisch verteilt.

1. Wir wollen den Plugin-Schätzer für $m(\vartheta) := \mathbf{E}_\vartheta[X_1]$ angeben. Wir berechnen (und bezeichnen mit $\widehat{\mathbf{E}}_X$ die Erwartung bezüglich $\widehat{\mathbf{P}}_X$ und $Z \sim \widehat{\mathbf{P}}_X$)

$$\widehat{m}(X) = m(\widehat{\mathbf{P}}_X) = \widehat{\mathbf{E}}_X[Z] = \sum_{i=1}^{n} X_i\widehat{\mathbf{P}}_X(Z = X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i =: \overline{X},$$

   d.h. der Plugin-Schätzer ist gerade der Mittelwert der Daten.

2. Nun zum Plugin-Schätzer für $m(\vartheta) := \mathbf{V}_\vartheta[X_1]$ angeben. Wir berechnen (und bezeichnen mit $\widehat{\mathbf{V}}_X$ die Varianz bezüglich $\widehat{\mathbf{P}}_X$)

$$\widehat{m}(X) = \widehat{\mathbf{V}}_X[Z] = \widehat{\mathbf{E}}_X[Z^2] - \widehat{\mathbf{E}}_X[Z]^2 = \Big(\frac{1}{n}\sum_{i=1}^{n} X_i^2\Big) - \overline{X}^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 =: \widetilde{s}(X).$$

Die gerade ermittelten Schätzer für Erwartungswert und Varianz einer Verteilung wollen wir nun etwas näher beleuchten.

Proposition 11.6 (Mittelwert und empirische Varianz als Schätzer für Erwartungswert und Varianz)**.** *Sei $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, wobei $X_1, ..., X_n$ unter $\mathbf{P}_\vartheta$ reellwertig, unabhängig und identisch verteilt sind. Weiter sei $\mathbf{E}_\vartheta[X_1^4] < \infty$ für alle $\vartheta \in \Theta$.*

1. *Der Mittelwert*
$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$
ist ein unverzerrter, konsistenter Schätzer für $\mathbf{E}_\vartheta[X_1]$.

2. *Die empirische Varianz*
$$s^2(X) := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$
ist ein unverzerrter, konsistenter Schätzer für $\mathbf{V}_\vartheta[X_1]$.

Bemerkung 11.7. Der Plugin-Schätzer $\widetilde{s}^2(X)$ für $\mathbf{V}_\vartheta[X_1]$ unterscheidet sich vom unverzerrten Schätzer $s^2(X)$ durch den Faktor $(n-1)/n$ Insbesondere gilt also, da $s^2(X)$ unverzerrt ist,
$$b(\vartheta, \mathbf{V}_\cdot[X_1], \widetilde{s}^2) = \mathbf{E}_\vartheta[\widetilde{s}^2(X) - s^2(X)] = -\frac{1}{n}\mathbf{E}_\vartheta[s^2(X)] = -\frac{1}{n}\mathbf{V}_\vartheta[X_1].$$
Da dies für $n \to \infty$ gegen 0 konvergiert, ist immerhin $\widetilde{s}^2(X)$ nach Bemerkung 11.2 konsistent.

Beweis. 1. Zunächst ist
$$\mathbf{E}_\vartheta[\overline{X}] = \tfrac{1}{n}(\mathbf{E}_\vartheta[X_1] + \cdots + \mathbf{E}_\vartheta[X_n]) = \mathbf{E}_\vartheta[X_1],$$
was bereits die Unverzerrtheit von $\overline{X}$ als Schätzer von μ_ϑ zeigt. Für die Konsistenz schreiben wir für $\varepsilon > 0$
$$\mathbf{P}_\vartheta[|\overline{X} - \mathbf{E}_\vartheta[X_1]| \geq \varepsilon] = \mathbf{P}_\vartheta\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mathbf{E}_\vartheta[X_1]\right| \geq \varepsilon\right) \xrightarrow{n\to\infty} 0$$
nach dem schwachen Gesetz der großen Zahlen.
2. Wir schreiben zunächst
$$\mathbf{E}_\vartheta[s^2(X)] = \tfrac{1}{n-1}\sum_{i=1}^{n}\mathbf{E}_\vartheta[X_i^2 - 2X_i\overline{X} + \overline{X}^2] = \tfrac{n}{n-1}\mathbf{E}_\vartheta[X_1^2 - 2X_1\overline{X} + \overline{X}^2].$$
Nun ist
$$\begin{aligned}\mathbf{E}_\vartheta[X_1^2] &= \mathbf{E}_\vartheta[X_1]^2 + \mathbf{V}_\vartheta[X_1],\\ \mathbf{E}_\vartheta[X_1\overline{X}] &= \mathbf{E}_\vartheta[X_1]^2 + \tfrac{1}{n}\mathbf{V}_\vartheta[X_1],\\ \mathbf{E}_\vartheta[\overline{X}^2] &= \mathbf{E}_\vartheta[X_1\overline{X}],\end{aligned}$$
also
$$\mathbf{E}_\vartheta[s^2(X)] = \tfrac{n}{n-1}\mathbf{E}_\vartheta[X_1^2 - X_1\overline{X}] = \mathbf{V}_\vartheta[X_1],$$
was die Unverzerrtheit bereits zeigt. Die Konsistenz wird in einer Übungsaufgabe nachgeprüft. □

Definition 11.8 (Momentenbasierte Schätzer). *Sei $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, so dass $X_1, ..., X_n$ unabhängig und identisch verteilt sind. Weiter sei*

$$m(\mathbf{P}_\vartheta) = h(m_1(\mathbf{P}_\vartheta), ..., m_\ell(\mathbf{P}_\vartheta))$$

mit

$$m_k(\mathbf{P}_\vartheta) := \mathbf{E}_\vartheta[X_1^k]$$

das k-te Moment von $\mathbf{P}_\vartheta$. (Das heißt, die zu schätzende Funktion $m(\mathbf{P}_\vartheta)$ lässt sich als Funktion der ersten ℓ Momente schreiben.) Dann heißt

$$\widehat{m}_k(X) := \frac{1}{n}\sum_{i=1}^n X_i^k$$

auch k-tes empirisches Moment und der Schätzer

$$\widehat{m}(X) = h(\widehat{m}_1(X), ..., \widehat{m}_\ell(X))$$

heißt momentenbasierter Schätzer für m.

Beispiel 11.9 (Schätzung des Parameters einer Poisson-Verteilung)**.** Momentenbasierte Schätzer müssen nicht eindeutig sein, wie folgendes Beispiel zeigt. Seien $X_1, ..., X_n$ unabhängig und identisch verteilt mit $(X_1)_*\mathbf{P} = \mathrm{Poi}(\vartheta)$. Es gilt

$$m(\mathbf{P}_\vartheta) := \vartheta = m_1(\mathbf{P}_\vartheta) = m_2(\mathbf{P}_\vartheta) - m_1(\mathbf{P}_\vartheta)^2.$$

Deshalb ist sowohl

$$\widehat{m}(X) = \widehat{m}_1(X) := \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$$

als auch

$$\widehat{\widehat{m}}(X) := \widehat{m}_2(X) - \widehat{m}_1(X)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 = \widetilde{s}^2(X)$$

ein momentenbasierter Schätzer für ϑ.

11.2. Maximum-Likelihood-Schätzer

Das Konzept von Maximum-Likelihood-Schätzern geht davon aus, dass bereits Daten $X = x$ erhoben wurden. Der Maximum-Likelihood-Schätzer ist dann dasjenige ϑ, für das die Wahrscheinlichkeit, die Daten zu erhalten, maximiert wird.

Definition 11.10 (Maximum Likelihood)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Die Abbildung*

$$L : \begin{cases} S \times \Theta & \to [0,1] \\ (x,\vartheta) & \mapsto \mathbf{P}_\vartheta(X \in dx) \end{cases}$$

heißt Likelihood-Funktion*. Für eine Abbildung $h : S \mapsto \Theta$ mit*

$$L(x, h(x)) = \max_{\vartheta\in\Theta} L(x,\vartheta)$$

heißt $\widehat{\vartheta}_{ML} = h(X)$ Maximum-Likelihood-Schätzer *von ϑ.*

Bemerkung 11.11 (Interpretation von Maximum-Likelihood-Schätzern)**.** Sei X diskret und $X = x$. Da die Vorstellung die ist, dass die erhobenen Daten Ergebnis eines Zufallsexperiments (d.h. die Realisierung einer Zufallsvariable X) sind, sagt man auch, dass die Daten $X = x$ sind. Ein Maximum-Likelihood-Schätzer ist also ein Parameter ϑ, unter dem die Wahrscheinlichkeit, die Daten $X = x$ zu beobachten – das ist $\mathbf{P}_\vartheta(X = x)$ – maximal ist.

Beispiel 11.12 (Maximum-Likelihood-Schätzer für μ und σ^2 von Normalverteilungen)**.** Wir betrachten den Fall einer unabhängigen, normalverteilten Stichprobe. Sei also $(X, (\mathbf{P}_{(\mu,\sigma^2)})_{(\mu,\sigma^2)\in\Theta})$ mit $\Theta = \mathbb{R} \times \mathbb{R}_+$ so, dass $X_1, \dots, X_n$ unabhängig und identisch nach verteilt sind mit $(X_1)_*\mathbf{P}_{(\mu,\sigma^2)} = N(\mu, \sigma^2)$.

Wir berechnen nun die Maximum-Likelihood-Schätzer für μ und σ^2. Anstatt die Likelihood-Funktion zu maximieren, werden wir dasselbe für deren Logarithmus tun. Wir schreiben

$$\begin{aligned}\log L((X_1, \dots, X_n), (\mu, \sigma^2)) &= \log\Big(\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2}\Big)\Big) \\ &= -n \log \sigma - \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2} + C,\end{aligned}$$

wobei C weder von μ noch von σ abhängt. Ableiten nach μ und σ ergibt

$$\begin{aligned}\frac{\partial \log L((X_1, \dots, X_n), (\mu, \sigma^2))}{\partial \mu} &= \sum_{i=1}^n \frac{X_i - \mu}{\sigma^2}, \\ \frac{\partial \log L((X_1, \dots, X_n), (\mu, \sigma^2))}{\partial \sigma} &= -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^3}.\end{aligned}$$

Für die Maximum-Likelihood-Schätzer $\widehat{\mu}_{ML}$ und $\widehat{\sigma}^2_{ML}$ gilt notwendigerweise

$$\begin{aligned}\sum_{i=1}^n (X_i - \widehat{\mu}_{ML}) &= 0, \\ \frac{n}{\widehat{\sigma}_{ML}} - \sum_{i=1}^n \frac{(X_i - \widehat{\mu}_{ML})^2}{\widehat{\sigma}^3_{ML}} &= 0.\end{aligned}$$

Die Maximum-Likelihood-Schätzer sind also gegeben durch

$$\begin{aligned}\widehat{\mu}_{ML} &= \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}, \\ \widehat{\sigma}^2_{ML} &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \widetilde{s}^2(X).\end{aligned}$$

Insbesondere sehen wir, dass $\overline{X}$ nicht nur erwartungstreu und konsistent (siehe Theorem 8.6) ist, sondern auch ein Maximum-Likelihood-Schätzer für μ. Allerdings ist der Maximum-Likelihood-Schätzer für σ^2 nicht erwartungstreu, wie man aus Proposition 11.6 abliest. Immerhin ist $\widehat{\sigma}^2_{ML}$ für große n annähernd erwartungstreu, da $\widehat{\sigma}^2_{ML} - s^2(X) \xrightarrow{n\to\infty} 0$.

Beispiel 11.13 (Maximum-Likelihood-Schätzer des Parameters einer Poisson-Verteilung)**.** Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\geq 0})$ so, dass $X = (X_1, \dots, X_n)$ und $X_1, \dots, X_n$ unabhängig und identisch nach

$\mathrm{Poi}(\vartheta)$ verteilt ist. Wir wissen bereits, dass $\mathbf{E}_\vartheta[X_1] = \mathbf{V}_\vartheta[X_1] = \vartheta$. Damit folgt, dass sowohl $\overline{X}$ als auch $s^2(X)$ unverzerrte Schätzer für ϑ sind, siehe Proposition 11.6. Wir berechnen nun den Maximum-Likelihood-Schätzer für ϑ (welcher sich als $\overline{X}$ herausstellen wird). Später, in Beispiel 11.26, werden wir sehen, dass dieser dem Schätzer $s^2(X)$ vorzuziehen ist, da er eine kleinere Risikofunktion besitzt.

Für den Maximum-Likelihood-Schätzer für ϑ berechnen wir zunächst wieder die log-Likelihood-Funktion

$$\log L((X_1,\dots,X_n),\vartheta) = \log \prod_{i=1}^n e^{-\vartheta}\frac{\vartheta^{X_i}}{X_i!} = -n\vartheta + \log(\vartheta)\sum_{i=1}^n X_i + C,$$

wobei C nicht von ϑ abhängt. Also ist

$$\frac{\partial \log L((X_1,\dots,X_n),\vartheta)}{\partial\vartheta} = -n + \frac{1}{\vartheta}\sum_{i=1}^n X_i. \tag{11.1}$$

Die log-Likelihood-Funktion ist also maximal für

$$-n + \frac{1}{\widehat{\vartheta}_{ML}}\sum_{i=1}^n X_i = 0, \qquad \widehat{\vartheta}_{ML} = \frac{1}{n}\sum_{i=1}^n X_i.$$

Damit ist $\widehat{\vartheta}_{ML} = \overline{X}$ der Maximum-Likelihood-Schätzer für ϑ. □

Maximum-Likelihood-Schätzer sind in vielen Fällen konsistent, was sicher eine wünschenswerte Eigenschaft ist. Der nächste Satz diese Konsistent von Maximum-Likelihood-Schätzern in einem einfachen Fall.

Theorem 11.14 (Konsistenz von Maximum-Likelihood-Schätzern)**.** *Sei $(X,(\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta\subseteq\mathbb{R}$ endlich, so dass $X=(X_1,\dots,X_n)$ unabhängig und identisch verteilt sind und X_1 eine Zufallsvariable mit Zielbereich $\mathbb{R}$ und einer beschränkten Dichte f_ϑ ist. Dann ist die Folge von Maximum-Likelihood-Schätzern von ϑ für $n=1,2,\dots$ konsistent.*

Bemerkung 11.15. Der Satz gilt auch unter schwächeren Voraussetzungen. (Etwa sollte Θ kompakt sein und $\vartheta\mapsto L(a,\vartheta)$ stetig.)

Beweis von Theorem 11.14. Zunächst ist für alle ϑ,ϑ' wegen der Jensen'schen Ungleichung, und, da $x\mapsto\log(x)$ konkav ist,

$$\mathbf{E}_{\vartheta'}\Big[\log\frac{f_\vartheta(X)}{f_{\vartheta'}(X)}\Big] \le \log\mathbf{E}_{\vartheta'}\Big[\frac{f_\vartheta(X)}{f_{\vartheta'}(X)}\Big] = \log\int f_{\vartheta'}(x)\frac{f_\vartheta(x)}{f_{\vartheta'}(x)}dx = \log\int f_\vartheta(x)dx = \log 1 = 0.$$

Der Maximum-Likelihood-Schätzer maximiert die Funktion, die ϑ auf

$$\frac{1}{n}\log L((X_1,\dots,X_n),\vartheta) - \frac{1}{n}\log L((X_1,\dots,X_n),\vartheta_0) = \frac{1}{n}\sum_{i=1}^n \log\frac{f_\vartheta(X_i)}{f_{\vartheta_0}(X_i)}$$

abbildet. Wegen dem starken Gesetz der großen Zahlen ist

$$\frac{1}{n}\log L((X_1,\dots,X_n),\vartheta) - \frac{1}{n}\log L((X_1,\dots,X_n),\vartheta_0) \xrightarrow{n\to\infty} \mathbf{E}_\vartheta\Big[\log\frac{f_\vartheta(X)}{f_{\vartheta_0}(X)}\Big] \le 0$$

mit $=0$ genau dann, wenn $f_\vartheta = f_{\vartheta_0}$. Da Θ diskret ist, konvergiert die Folge der Maximum-Likelihood-Schätzer für ϑ gegen ϑ_0. □

11.3. Optimalitätskriterien von Schätzern

Natürlich wollen wir möglichst gute Schätzer finden. Vorher ist zu klären, in welchen Sinn diese Qualität der Schätzer denn gemeint ist. Wir beschränken uns hier auf den mittleren quadratischen Fehler, also eine quadratische Verlustfunktion.

Definition 11.16 (Mittlerer quadratischer Fehler)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $m : \vartheta \mapsto m(\vartheta)$.*

1. *Für einen Schätzer $h(X)$ von $m(\vartheta)$ ist der* mittlere quadratische Fehler *(oder die* Risikofunktion*) definiert als*

$$R_{h(X)} : \begin{cases} \Theta & \to [0,\infty], \\ \vartheta & \mapsto \mathbf{E}_\vartheta[(h(X) - m(\vartheta))^2]. \end{cases}$$

2. *Sei $m : \Theta \to \Theta'$ und $\mathcal{S} \subseteq \{h : S \to \Theta'\}$ eine Menge von Schätzern. Falls für ein $h \in \mathcal{S}$ gilt, dass für alle $\vartheta \in \Theta$*
$$R_{h(X)}(\vartheta) = \inf_{h\in\mathcal{S}} R_{h(X)}(\vartheta),$$
so heißt $h(X)$ bester Schätzer in $\mathcal{S}$.

3. *Ist insbesondere $\mathcal{S} = \{h(X) : \mathbf{E}_\vartheta[h(X)] = m(\vartheta), \vartheta \in \Theta\}$ die Menge der unverzerrten Schätzer, so heißt ein bester Schätzer in $\mathcal{S}$ auch* UMVUE *(Uniformly Minimal Variance Unbiased Estimator).*

Mit Hilfe suffizienter Statistiken kann man vorhandene Schätzer besser machen. Dies ist Inhalt folgenden Satzes.

Theorem 11.17 (Satz von Rao-Blackwell)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $m : \Theta \to \Theta'$, $h : S \to \Theta'$ mit $h(X)$ einem Schätzer für $m(\vartheta)$. Ist $t(X)$ suffizient für ϑ, so ist*

$$\widetilde{h}(X) = \mathbf{E}_\vartheta[h(X)|t(X)] \tag{11.2}$$

unabhängig von ϑ und

$$\mathbf{E}_\vartheta[(\widetilde{h}(X) - m(\vartheta))^2] \le \mathbf{E}_\vartheta[(h(X) - m(\vartheta))^2].$$

Bemerkung 11.18 (Interpretation)**.** Das Theorem definiert eine Funktion $\widetilde{h} : S \to \Theta$. Wichtig ist, dass $\widetilde{h}(X)$ nur von $t(X)$ abhängt. Es gilt nämlich

$$\widetilde{h}(x) = \mathbf{E}_\vartheta[h(X)|t(X) = t(x)]$$

nach der Definition der bedingten Erwartung, und die rechte Seite hängt wegen der Definition der bedingten Erwartung nur von $t(x)$ ab. Die Aussage des Satzes ist, dass der Schätzer $\widetilde{h}(X)$ eine kleinere Risikofunktion hat als $h(X)$.

Beispiel 11.19 (Die Schätzer $\widehat{p}$ und $\widehat{p}'$ beim Münzwurf)**.** Sei wieder $X = (X_1, \dots, X_n)$ ein p-Münzwurf mit noch unbestimmten p. In Beispiel 11.3 haben wir zwei erwartungstreue Schätzer für p kennen gelernt, nämlich

$$\widehat{p}(X) = \tfrac{1}{n}(X_1 + \cdots + X_n), \qquad \widehat{p}'(X) = X_1$$

und hatten auch festgestellt, dass $\widehat{p}$ eine kleinere Varianz (also Risikofunktion) besitzt als $\widehat{p}'$. Weiter wissen wir aus Beispiel 10.21, dass $t(X) = X_1 + \cdots + X_n$ suffizient für p ist. Wir können also nun den Schätzer $\widehat{p}'$ mit Hilfe des Satzes von Rao-Blackwell verbessern, indem wir $h(X) = \widehat{p}'(X)$ setzen und $\widetilde{h}(X)$ aus (11.2) berechnen. Dies ergibt aus Symmetriegründen

$$\begin{aligned}\widetilde{h}(X) &= \mathbf{E}_p[\widehat{p}'(X) \mid t(X)] = \mathbf{E}_p[X_1 \mid X_1 + \cdots + X_n] = \tfrac{1}{n}\sum_{i=1}^{n} \mathbf{E}_p[X_i \mid X_1 + \cdots + X_n] \\ &= \tfrac{1}{n}\mathbf{E}_p[X_1 + \cdots + X_n \mid X_1 + \cdots + X_n] = \tfrac{1}{n}\big(X_1 + \cdots + X_n\big) \\ &= \widehat{p}(X).\end{aligned}$$

Insbesondere sehen wir hier ein konkretes Beispiel dafür, dass $\widetilde{h}(X)$ eine echt kleinere Risikofunktion besitzt als $h(X)$.

Beweis von Theorem 11.17. Wir beschränken den Beweis auf den Fall diskreter Zufallsvariablen X. Der Fall von Zufallsvariablen mit Dichten geht analog. Zunächst ist

$$\mathbf{E}_\vartheta[h(X) \mid t(X)] = \sum_{a \in S} h(a)\mathbf{P}_\vartheta(X = a \mid t(X)),$$

was wegen der Suffizienz von $t(X)$ nicht von ϑ abhängt. Wegen (siehe Proposition 7.12)

$$\mathbf{E}_\vartheta[\widetilde{h}(X)] = \mathbf{E}_\vartheta[\mathbf{E}_\vartheta[h(X) \mid t(X)]] = \mathbf{E}_\vartheta[h(X)]$$

gilt mit der Varianzformel (Proposition 7.15)

$$\begin{aligned}\mathbf{E}_\vartheta[(\widetilde{h}(X) - m(\vartheta))^2] &= \big(\mathbf{E}_\vartheta[\widetilde{h}(X)] - m(\vartheta)\big)^2 + \mathbf{Var}_\vartheta[\mathbf{E}_\vartheta[h(X) \mid t(X)]] \\ &\le \big(\mathbf{E}_\vartheta[h(X)] - m(\vartheta)\big)^2 + \mathbf{Var}_\vartheta[\mathbf{E}_\vartheta[h(X)|t(X)]] + \mathbf{E}_\vartheta[\mathbf{Var}_\vartheta[h(X)|t(X)]] \\ &= \mathbf{E}_\vartheta[(h(X) - m(\vartheta))^2]\end{aligned}$$

und die Behauptung ist gezeigt. □

Bemerkung 11.20 (Satz von Lehmann-Scheffe)**.** Betrachten wir die Klasse erwartungstreuer Schätzer. Der Satz von Rao-Blackwell erlaubt die Verbesserung von Schätzern mit Hilfe suffizienter Statistiken. Wir wissen jedoch nicht, ob die Verbesserung optimal ist, d.h. ob es sich bei den verbesserten Schätzern um UMVUE-Schätzern handelt. Eine Antwort auf diese Frage gibt der Satz von Lehmann-Scheffe:
Ist die suffiziente Statistik im Satz von Rao-Blackwell vollständig, d.h. für alle g gilt

$$\mathbf{E}_\vartheta[g(t(X))] = 0, \vartheta \in \Theta \Rightarrow \mathbf{P}_\vartheta(g(t(X)) = 0) = 1, \vartheta \in \Theta,$$

so ist $\widetilde{h}(X)$ im Satz von Rao-Blackwell ein UMVUE. Weiter sind die suffizienten Statistiken von Exponentialfamilien vollständig.
Dies bedeutet etwa, dass im Münzwurf-Beispiel der Schätzer $\overline{X}$ ein UMVUE ist.

Nachdem wir nun gesehen haben, wie man Schätzer besser machen kann, wollen wir nun wissen, wie groß die minimale Varianz eines Schätzers denn sein kann. Hierzu benötigen wir ein neues, von der Likelihood-Funktion abgeleitetes, Konzept. Wir beschränken uns dabei auf folgende Situation:

Definition 11.21 (Fisher-Information)**.** *Sei* $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ *ein statistisches Modell und* $(x,\vartheta) \mapsto L(x,\vartheta) = \mathbf{P}_\vartheta(X \in dx)$ *die Likelihood-Funktion. Folgende Voraussetzungen seien erfüllt:*

1. $\Theta \subseteq \mathbb{R}$ *ist offen;*
2. $\{x : L(x,\vartheta) > 0\}$ *hängt nicht von* ϑ *ab.*
3. *Die Ableitung* $\frac{\partial}{\partial\vartheta}\log L(x,\vartheta)$ *existiert und ist endlich;*
4. *Falls* X *unter* $\mathbf{P}_\vartheta$ *eine Dichte* p_ϑ *hat: Ist* $t : S \to \mathbb{R}$*, so dass* $\mathbf{E}_\vartheta[|t(X)|] < \infty$ *für alle* $\vartheta \in \Theta$*, so gilt*
$$\frac{\partial}{\partial\vartheta}\int t(x)p_\vartheta(x)dx = \int t(x)\frac{\partial}{\partial\vartheta}p_\vartheta(x)dx.$$

Dann heißt $\frac{\partial}{\partial\vartheta}\log L(X,\vartheta)$ Score-Funktion *und*

$$\vartheta \mapsto \mathcal{I}(\vartheta) := \mathbf{E}_\vartheta\Big[\Big(\frac{\partial}{\partial\vartheta}\log L(X,\vartheta)\Big)^2\Big]$$

Fisher-Information.

Bemerkung 11.22 (Fisher-Information einer unabhängigen Stichprobe)**.** Ist $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit Likelihood-Funktion $(x,\vartheta) \mapsto L(x,\vartheta)$ und $\mathcal{I}_1(\vartheta)$ die Fisher-Information. Ist dann $X_1, ..., X_n$ unabhängig mit $X_1, ..., X_n \sim X$ unter $\mathbf{P}_\vartheta$, d.h. $X_1, ..., X_n$ sind eine unabhängige Stichprobe der Verteilung $\mathbf{P}_\vartheta$, $\vartheta \in \Theta$. Dann ist

$$\vartheta \mapsto \mathcal{I}(\vartheta) = \mathbf{E}_\vartheta\Big[\Big(\frac{\partial}{\partial\vartheta}\sum_{i=1}^n \log L(X_i,\vartheta)\Big)^2\Big] = \sum_{i=1}^n \mathbf{E}_\vartheta\Big[\Big(\frac{\partial}{\partial\vartheta}\log L(X_i,\vartheta)\Big)^2\Big] = n\cdot\mathcal{I}_1(\vartheta)$$

die Fisher-Information des Modells $((X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$.

Bemerkung 11.23 (Einparametrige Exponentialfamilien erfüllen Voraussetzungen)**.** Ist die Dichte durch eine ein-parametrige Exponentialfamilie mit

$$p_\vartheta(x) = 1_A(x)\exp(c(\vartheta)t(x) - d(\vartheta))$$

und $\frac{\partial}{\partial\vartheta}c(\vartheta) \neq 0$ für alle $\vartheta \in \Theta$ und $\Theta \subseteq \mathbb{R}$ offen, dann sind die Voraussetzungen von Definition 11.21 erfüllt.

Bemerkung 11.24 (Erwartung der Score-Funktion verschwindet)**.** Im Fall einer Dichte schreiben wir

$$\mathcal{I}(\vartheta) = \int\Big(\frac{\partial}{\partial\vartheta}\log p_\vartheta(x)\Big)^2 p_\vartheta(x)dx = \int\frac{\big(\frac{\partial p_\vartheta(x)}{\partial\vartheta}\big)^2}{p_\vartheta(x)}dx$$

und für die Erwartung der Score-Funktion gilt

$$\mathbf{E}_\vartheta\Big[\frac{\partial}{\partial\vartheta}\log p_\vartheta(X)\Big] = \int\frac{\frac{\partial p_\vartheta(x)}{\partial\vartheta}}{p_\vartheta(x)}p_\vartheta(x)dx = \frac{\partial}{\partial\vartheta}\int p_\vartheta(x)dx = 0.$$

Also gilt

$$\mathcal{I}(\vartheta) = \mathbf{V}_\vartheta\Big[\frac{\partial}{\partial\vartheta}\log L(X,\vartheta)\Big]$$

Theorem 11.25 (Cramér-Rao-Schranke)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\vartheta \mapsto \mathcal{I}(\vartheta)$ die Fisher-Information. Ist $t : S \to \mathbb{R}$, so dass $\mathbf{V}_\vartheta[t(X)] < \infty$ für alle $\vartheta \in \Theta$ und $\Psi(\vartheta) := \mathbf{E}_\vartheta[t(X)]$. Sind die Bedingungen aus Definition 11.21 erfüllt, so ist Ψ differenzierbar und es gilt*

$$\mathbf{V}_\vartheta[t(X)] \geq \frac{(\Psi'(\vartheta))^2}{\mathcal{I}(\vartheta)}, \qquad \vartheta \in \Theta.$$

Ist also insbesondere t ein unverzerrter Schätzer für ϑ, so gilt

$$\mathbf{V}_\vartheta[t(X)] \geq \frac{1}{\mathcal{I}(\vartheta)}, \qquad \vartheta \in \Theta.$$

Gilt Gleichheit, so ist $t(X)$ also ein UMVUE und heißt auch effizient*.*

Beweis. Wir führen den Beweis im Fall kontinuierlicher $\mathbf{P}_\vartheta, \vartheta$, d.h. $X_*\mathbf{P}_\vartheta$ hat die Dichte p_ϑ, $\vartheta \in \Theta$. Zunächst ist

$$\Psi'(\vartheta) = \frac{\partial}{\partial\vartheta}\int t(x)p_\vartheta(x)dx = \int t(x)\frac{\partial}{\partial\vartheta}p_\vartheta(x)dx = \mathbf{E}_\vartheta\Big[\frac{\partial}{\partial\vartheta}t(X)\log p_\vartheta(X)\Big].$$

Daraus folgt, mit Bemerkung 11.24 und der Cauchy-Schwartz-Ungleichung, Proposition 5.10

$$\begin{aligned}(\Psi'(\vartheta))^2 &= \Big(\mathbf{E}_\vartheta\Big[t(X)\frac{\partial}{\partial\vartheta}\log p_\vartheta(X)\Big]\Big)^2 = \mathbf{COV}_\vartheta\Big[t(X), \frac{\partial}{\partial\vartheta}\log p_\vartheta(X)\Big]^2 \\ &\leq \mathbf{V}_\vartheta[t(X)]\cdot\mathbf{V}_\vartheta\Big[\frac{\partial}{\partial\vartheta}\log p_\vartheta(X)\Big] = \mathbf{V}_\vartheta[t(X)]\cdot\mathcal{I}(\vartheta).\end{aligned}$$

Daraus folgen alle Behauptungen. □

Beispiel 11.26 (Effizienz des Schätzers für den Parameter einer Poisson-Verteilung)**.**
Sei $(X, (\mathbf{P}_\lambda)_{\lambda>0})$ so, dass $X = (X_1, \ldots, X_n)$ und $X_1, \ldots, X_n$ unabhängig nach Poi(λ) verteilt ist. Wir haben bereits in Beispiel 11.13 berechnet, dass

$$\widehat{\lambda}_{ML} = \tfrac{1}{n}\sum_{i=1}^n X_i$$

der Maximum-Likelihood-Schätzer für λ ist. Wir wollen nun sehen, ob $\widehat{\lambda}_{ML}$ auch effizient ist. Dafür berechnen wir die Fisher-Information (siehe (11.1) für die Ableitung der log-Likelihood-Funktion)

$$\mathcal{I}(\lambda) = \mathbf{E}_\lambda\Big[\Big(\frac{\partial}{\partial\lambda}\log L(X,\lambda)\Big)^2\Big] = \mathbf{E}_\lambda\Big[\Big(-n + \frac{1}{\lambda}\sum_{i=1}^n X_i\Big)^2\Big] = \mathbf{V}_\lambda\Big[\frac{1}{\lambda}\sum_{i=1}^n X_i\Big] = \frac{n}{\lambda}.$$

Der Schätzer $\widehat{\lambda}_{ML}$ ist also effizient, da

$$\mathbf{V}_\lambda[\widehat{\lambda}_{ML}] = \frac{\lambda}{n} = \frac{1}{\mathcal{I}(\lambda)}.$$

12. Testprobleme

12.1. Grundbegriffe

Neben Schätzproblemen sind Testprobleme das wichtigste Thema der induktiven Statistik. In der empirischen Forschung ist es oftmals so, dass aufgrund eines ersten Verständnisses eine Hypothese über die erworbenen Daten aufgestellt werden kann. Das Überprüfen solcher Hypothesen erfolgt dann mittels statistischer Tests. Wie üblich geht man davon aus, dass die Daten die Realisierung einer Zufallsvariable sind.

Wir beginnen mit der Einführung wichtiger Begriffe wie Teststatistik, Nullhypothese, Alternative, Ablehnungsbereich, Signifikanzniveau und p-Wert. Diese sind teilweise aus der Schule bekannt.

Definition 12.1 (Statistischer Test)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, S der Zielbereich von X, und $\Theta_0, \Theta_A \subseteq \Theta$ disjunkt mit $\Theta_0 \cup \Theta_A = \Theta$. H_0 heißt* Nullhypothese *und H_A heißt* Alternativhypothese*.*

1. *Die Hypothese $H : \vartheta \in \Theta_H$ (die entweder H_0 oder H_A sein kann) heißt* einfach *wenn $\Theta_H = \{\vartheta^*\}$ für ein $\vartheta^* \in \Theta$. Andernfalls heißt H* zusammengesetzt*.*

2. *Eine Abbildung $\varphi : S \to [0, 1]$ heißt* (randomisierter) statistischer Test *von*

$$H_0 : \vartheta \in \Theta_0 \text{ gegen } H_A : \vartheta \in \Theta_A.$$

Hier ist $\varphi(x)$ die Wahrscheinlichkeit, sich für die Alterantive H_A zu entscheiden (also H_0 abzulehnen), falls die Daten $X = x$ sind. (Man stelle sich also vor, dass die Daten $X = x$ vorliegen, und man anschließend einen $\varphi(x)$-Münzwurf macht und daraus entscheidet, ob man die Null- oder Alternativhypothese annimmt. Insbesondere kann es bei gleichen Daten zu unterschiedlichen Entscheidungen kommen.) Man sagt, der Test hat (Signifikanz-)Niveau $\alpha \in [0, 1]$*, falls*

$$\sup_{\vartheta\in\Theta_0} \mathbf{E}_\vartheta(\varphi(X)) \leq \alpha. \tag{12.1}$$

3. *Der Spezialfall eines* nicht-randomisierten, statistischen Tests *ist es, wenn $\varphi(S) = \{0, 1\}$, man sich also immer gleich für Null- oder Alternativhypothese bei Vorliegen von $X = x$ entscheidet. In diesem Fall ist $Y = t(X)$ eine* Teststatistik *und $C := t(\varphi^{-1}(1)) \subseteq S$ der* kritische Bereich *oder* Ablehnungsbereich *des Test. Insbesondere entscheidet man sich genau dann für die Alternative, falls $\varphi(x) = 1$, d.h. $Y = t(x) \in t(\varphi^{-1}(1)) = C$. Nun sagt man auch, dass das Paar (Y, C) der* statistische Test *von*

$$H_0 : \vartheta \in \Theta_0 \text{ gegen } H_A : \vartheta \in \Theta_A$$

ist. Nun hat der Test (T, C) das (Signifikanz-)Niveau $\alpha \in [0, 1]$*, falls*

$$\sup_{\vartheta\in\Theta_0} \mathbf{P}_\vartheta(Y \in C) = \sup_{\vartheta\in\Theta_0} \mathbf{E}_\vartheta(\varphi(X)) \leq \alpha. \tag{12.2}$$

Falls $Y \in C$, sagt man, dass H_0 abgelehnt (und damit H_A angenommen) ist. Falls $Y \notin C$, sagt man, dass H_0 nicht abgelehnt ist (und H_A abgelehnt ist).

4. *Ist $\Theta = (\underline{\vartheta}, \overline{\vartheta})$ ein Intervall (wobei $\underline{\vartheta} = -\infty$ und $\overline{\vartheta} = \infty$ zugelassen sind und die Intervalle auch abgeschlossen sein können), so heißt der Test* einseitig, *falls $\Theta_0 = (\underline{\vartheta}, \vartheta^*)$ oder $\Theta_0 = (\vartheta^*, \overline{\vartheta})$. Falls $\Theta_0 = (\vartheta^+, \vartheta^*)$ mit $\underline{\vartheta} < \vartheta^+ \leq \vartheta^* < \overline{\vartheta}$, so heißt der Test* zweiseitig.

5. *Gilt*

$$\sup_{\vartheta \in \Theta_0} \mathbf{E}_\vartheta[\varphi(X)] < \alpha,$$

so heißt der Test φ konservativ *zum Niveau α.*

6. *Der Test φ heißt* unverfälscht, *falls*

$$\mathbf{E}_{\vartheta_0}[\varphi(X)] \leq \mathbf{E}_{\vartheta_A}[\varphi(X)]$$

für alle $\vartheta_0 \in \Theta_0, \vartheta_A \in \Theta_A$ gilt.

Bemerkung 12.2 (Interpretation und Fehler eines Tests)**.**

1. Einen (nicht-randomisierten) statistischen Test hat man sich am besten so vorzustellen (siehe auch das nächste Beispiel): die Daten sind gegeben durch die Zufallsvariable X. Diese Daten fasst man durch die meist reellwertige Funktion t zusammen zur Teststatistik $Y = t(X)$. Die Daten können entweder nach $\mathbf{P}_\vartheta$ mit $\vartheta \in \Theta_0$ (d.h. die Nullhypothese ist richtig) oder mit $\vartheta \in \Theta_A$ (d.h. die Alternativhypothese ist richtig) verteilt sein. Ziel ist es, die Nullhypothese genau dann (anhand der Daten X) abzulehnen, wenn H_A richtig ist. Der Ablehnungsbereich C ist so gewählt, dass H_0 genau dann abgelehnt wird, wenn $Y \in C$.

2. Bei randomisierten (und nicht-randomisierten) Tests kommt es immer zu einer Entscheidung für H_0 oder H_A. Dabei können zwei verschiedene Arten von Fehler auftreten:

	H_0 abgelehnt	H_0 nicht abgelehnt
H_0 richtig	Fehler erster Art	richtige Entscheidung
H_0 falsch	richtige Entscheidung	Fehler zweiter Art

Gehen wir zunächst davon aus, dass $\vartheta \in \Theta_0$. Hat der Test ein Niveau α, so wissen wir, dass $\mathbf{E}_\vartheta[\varphi(X)] \leq \alpha$ (bzw. $\mathbf{P}_\vartheta[t(X) \in C] \leq \alpha$). Da H_0 genau mit Wahrscheinlichkeit $\varphi(X)$ verworfen wird (verworfen wird, falls $t(X) \in C$), wissen wir also, dass die Nullhypothese im Mittel höchstens mit Wahrscheinlichkeit α abgelehnt wird, wenn sie zutrifft. Damit hat man also die Wahrscheinlichkeit, die Nullhypothese abzulehnen, falls sie zutrifft, durch α beschränkt. Für $\vartheta \in \Theta_0$ und $X = x$ ist also $\varphi(x)$ die Wahrscheinlichkeit für einen *Fehler erster Art*. (Für $\vartheta \in \Theta_0$ und $t(X) \in C$, die Nullyhpothese also irrtümlicherweise verworfen wird, sprechen wir von einem *Fehler erster Art*).

Geht man davon aus, dass $\vartheta \in \Theta_A$, liegt eine Fehlentscheidung mit Wahrscheinlichkeit $1 - \varphi(X)$ vor (dann vor, wenn $Y \notin C$), die Nullhypothese also nicht abgelehnt wird. In diesem Fall sprechen wir von einem *Fehler zweiter Art*. Das Niveau des Tests liefert keinen Anhaltspunkt dafür, mit welcher Wahrscheinlichkeit ein solcher Fehler auftritt.

3. Auf den ersten Blick besteht eine scheinbare Symmetrie zwischen H_0 und H_A. Schließlich lehnen wir mit Wahrscheinlichkeit $\varphi(X)$ (falls $Y \in C$) H_0 genau dann ab (und nehmen H_A an), und mit Wahrscheinlichkeit $1 - \varphi(X)$ (falls $Y \notin C$) lehnen wir H_0 nicht ab (und lehnen damit H_A ab). Allerdings wird diese Symmetrie durch das Niveau des Tests gebrochen. Weiß man, dass φ $((Y, C))$ ein Test zum Niveau α ist, bedeutet das, dass die Wahrscheinlichkeit, die Nullhypothese H_0 abzulehnen, obwohl sie wahr ist, im Mittel höchstens α ist. Mit anderen Worten ist die Wahrscheinlichkeit für einen Fehler erster Art höchstens α. Allerdings hat man keine Kontrolle über den Fehler zweiter Art.

   Wegen dieser Asymmetrie ist in der Praxis die Nullhypothese genau so zu wählen, dass eine Ablehnung der Nullhypothese möglichst sicher auf die Richtigkeit der Alternativhypothese zurückzuführen ist. Wir betrachten das Beispiel 12.5 des Binomialtests, bei dem wir bei $X = 23$ Treffern in $n = 53$ Versuchen eines Münzwurfes testen wollen, ob der Wurf fair gewesen sein kann, d.h. auf die Erfolgswahrscheinlichkeit $p = 1/2$. Zunächst sind wir skeptisch, dass $p = 1/2$ richtig sein kann, da eigentlich zu wenige Erfolge zu verzeichnen waren. Wir legen das Signifikanzniveau α fest (was in der Praxis oft $\alpha = 5\,\%$ ist). Um unsere Vorstellung über die Erfolgswahrscheinlichkeit zu überprüfen, testen wir

$$H_0 : p = 1/2 \qquad \text{gegen} \qquad H_A : p \neq 1/2.$$

   Kommt es nämlich jetzt zu einer Ablehnung von H_0, so wissen wir, dass dies mit Wahrscheinlichkeit höchstens α dann passiert, wenn H_0 wahr ist, die Münze also fair war. Damit können wir uns relativ sicher sein, dass die Ablehnung der Nullhypothese darauf zurückzuführen ist, dass H_A zutrifft. Damit ist unsere Vorstellung, dass die Münze bei einer Ablehnung der Nullhypothese unfair war, höchstwahrscheinlich bestätigt.

4. Ein möglichst großer Ablehnungsbereich bei möglichst kleinem Niveau α ist für jeden Test wünschenswert. Schließlich soll die Nullhypothese in möglichst vielen Fällen abgelehnt werden, ohne dass die Wahrscheinlichkeit, sie irrtümlicherweise abzulehnen, größer als α wird.

5. Die Forderung von unverfälschten Tests ist klar zu verstehen: Da wir H_0 mit Wahrscheinlichkeit $\varphi(X)$ ablehnen, soll zumindest die Wahrscheinlichkeit, dass H_0 abgelehnt wird, unter $\mathbf{P}_{\vartheta_A}$, $\vartheta_A \in \Theta_A$ größer sein als für $\mathbf{P}_{\vartheta_0}$, $\vartheta_0 \in \Theta_0$.

Bemerkung 12.3 (***p*-Werte und alternative Definition eines Tests**)**.** Betrachten wir einen nicht-randomisierten Test (Y, C). Sei $Y = y$, d.h. dass die Teststatistik Y, angewendet auf die echten Daten, ergibt y. Dann heißt der Wert

$$p_y := \sup_{\vartheta \in \Theta_0} \mathbf{P}_\vartheta(Y \text{ extremer als } y)$$

p-Wert des Tests für $Y = y$. Dabei hängt die Bedeutung davon, was 'extremer' heißt davon ab, was genau die Alternative ist. (Dies ist oftmals in konkreten Beispielen einfach zu verstehen, siehe etwa den Binomialtest und den Gauss-Test.) Immer gilt jedoch $p_y \leq p_{y'}$, falls y extremer

als y' ist. Es ist wichtig zu beachten, dass es dadurch einen engen Zusammenhang zwischen dem Niveau α des Tests und dem p-Wert gibt. Ist nämlich (Y, C) ein Test zum Niveau α und

$$C = \{y : y \text{ extremer als } y_0\}$$

für ein y_0, so wird H_0 genau dann abgelehnt, wenn

$$\alpha \geq \sup_{\vartheta \in \Theta_0} \mathbf{P}_\vartheta(Y \text{ extremer als } y_0) = p_{y_0}.$$

Ist $Y = y$ und gilt $p_y \leq p_{y_0}$, so wird H_0 also abgelehnt. Es genügt also, für einen Test zum Niveau α und $Y = y$ den Wert p_y zu bestimmen. Ist $p_y \leq \alpha$, so wird H_0 abgelehnt. Dieses Vorgehen wird bei vielen Statistik-Programmen angewendet, bei denen ausschließlich p-Werte ausgegeben werden. Dabei muss man meist angeben, was genau die Alternative ist (einseitig oder zweiseitig), damit das Programm weiß, in welche Richtungen Abweichungen als extrem zu betrachten sind.

Wir kommen nun zunächst zu zwei konkreten Beispielen für Tests. Den ersten haben wir auch schon in unserem Eingangsbeispiel in Abschnitt 9.1 kennen gelernt.

Proposition 12.4 (Nicht-randomisierter Binomialtest)**.** *Sei $\alpha \in [0,1], n \in \mathbb{N}$ und $(X, (\mathbf{P}_p)_{p\in[0,1]})$ ein statistisches Modell, so dass X unter $\mathbf{P}_p$ nach $B(n,p)$ verteilt ist.*

(a) Ist $\Theta_0 = p^*$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(X, \{0,\dots,k\} \cup \{l,\dots,n\})$ ein unverfälschter Test zum Niveau α, falls

$$\mathbf{P}_{p^*}(X \leq k) \leq \alpha/2, \qquad \mathbf{P}_{p^*}(X \geq l) \leq \alpha/2.$$

(b) Ist $\Theta_0 = [0, p^*]$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(X, \{k,\dots,n\})$ ein unverfälschter Test zum Niveau α, falls

$$\mathbf{P}_{p^*}(X \geq k) \leq \alpha.$$

(c) Ist $\Theta_0 = [p^*, 1]$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(X, \{0,\dots,k\})$ ein unverfälschter Test zum Niveau α, falls

$$\mathbf{P}_{p^*}(X \leq k) \leq \alpha.$$

Beweis. Wir beweisen nur (c), die anderen beiden Aussagen folgen analog. Klar ist, dass der Test unverfälscht ist. Es ist außerdem

$$\sup_{p\in\Theta_0} \mathbf{P}_p(X \in \{0,\dots,k\}) = \mathbf{P}_{p^*}(X \leq k) \leq \alpha$$

nach Voraussetzung. Also folgt bereits die Aussage. □

Beispiel 12.5 (Binomialtest)**.** Sei $\alpha = 5\,\%$, $n = 53$ und $(X, (\mathbf{P}_p)_{p\in[0,1]})$ wie in der Proposition. Wir wollen nun

$$H_0 : p = 1/2 \text{ gegen } H_A : p \neq 1/2$$

testen, wenn wir in 53 Versuchen 23 Erfolge erzielt haben. Nach Proposition 12.4 ist der kritische Bereich von der Form $\{0,\dots,k\} \cup \{l,\dots,53\}$. Es ist

$$\mathbf{P}_{p=1/2}(X \leq 18) + \mathbf{P}_{p=1/2}(X \geq 35) \approx 2.70\,\%.$$

Da $18 < 23 < 35$, liegt 23 nicht im Ablehnungsbereich von H_0. Damit kann die Nullhypothese aufgrund der Daten ($X = 23$) nicht abgelehnt werden. Auf dasselbe Ergebnis kommt man mit Hilfe des p-Wertes. Es ist

$$\mathbf{P}_{p=1/2}(X \text{ extremer als } 23) = \mathbf{P}_{p=1/2}(X \leq 23) + \mathbf{P}_{p=1/2}(X \geq 30) \approx 41.01\,\%.$$

Da dieser Wert größer als $\alpha = 5\,\%$ ist, kann man die Nullhypothese nicht ablehnen.

Beispiel 12.6 (Randomisierter Binomialtest)**.** Betrachten wir den Binomialtest von $H_0 : p \in [0, p^*]$ gegen $H_A : p \in (p^*, 1]$. Im nicht-randomisierten Binomial-Test ist zwar das Signifikanz-Niveau α, jedoch nimmt $\mathbf{P}_{p^*}(X \geq l)$ für $l = 0, ..., n$ nur endlich viele Werte an. Mit anderen Worten wird in vielen Fällen das Signifikanzniveau nicht voll ausgeschöpft, und der Test ist konservativ. Dies wird im *randomisierten Binomial-Test* verbessert:
Für gegebenes $\alpha \in (0, 1)$ sei l minimal mit $\mathbf{P}_{p^*}(X \geq l) \leq \alpha$. Dann definieren wir den randomisierten Test

$$\varphi(x) = \begin{cases} 1, & x = l, ..., n, \\ \frac{\alpha - \mathbf{P}_{p^*}(X \geq l)}{\mathbf{P}_{p^*}(X = l-1)}, & x = l-1, \\ 0, & x = 0, ..., l-2. \end{cases}$$

Dann gilt

$$\mathbf{E}_{p^*}[\varphi(X)] = \mathbf{P}_{p^*}[X \geq l] + \frac{\alpha - \mathbf{P}_{p^*}(X \geq l)}{\mathbf{P}_{p^*}(X = l-1)} \mathbf{P}_{p^*}(X = l-1) = \alpha.$$

Insbesondere ist φ ein (nicht-konservativer) Test zum Niveau α mit einem größeren (wenn auch randomisierten) Ablehnungsbereich.

Proposition 12.7 (Gauss-Test)**.** *Sei $\alpha \in [0, 1], \sigma^2 \in \mathbb{R}_+, \mu^* \in \mathbb{R}$ und $(X = (X_1, \ldots, X_n), (\mathbf{P}_\mu)_{\mu \in \mathbb{R}})$ ein statistisches Modell, so dass $X_1, \ldots, X_n$ unter $\mathbf{P}_\mu$ unabhängig und nach $N(\mu, \sigma^2)$ verteilt sind. Weiter sei*

$$Z := \frac{\overline{X} - \mu^*}{\sqrt{\sigma^2/n}}.$$

und z_p für $p \in [0, 1]$ das p-Quantil von $N(0, 1)$.

(a) Ist $\Theta_0 = \{\mu^*\}$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(Z, (-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = (-\infty, \mu^*]$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(Z, [z_{1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = [\mu^*, \infty)$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(Z, (-\infty, z_\alpha])$ ein unverfälschter Test zum Niveau α.

Beweis. Wieder beweisen wir nur (c). Es ist klar, dass der Test unverfälscht ist. Wir wissen, dass unter $\mathbf{P}_{\mu^*}$ die Zufallsvariable Z nach $N(0, 1)$ verteilt ist. Damit gilt

$$\sup_{\mu \geq \mu^*} \mathbf{P}_\mu(Z \leq z_\alpha) = \mathbf{P}_{\mu^*}(Z \leq z_\alpha) = \alpha,$$

woraus die Behauptung sofort folgt. □

12.2. Intervallschätzer und Tests

Wir kommen nochmal zurück zu Schätzproblemen. Bisher hatten wir nur *Punktschätzer* für (Funktionen des) Parameter(s) betrachtet, uns jedoch – bis auf Abschätzungen über die Varianz des Schätzers – weniger über die mögliche Streuung Gedanken gemacht. Intervallschätzer unterscheiden sich von Punktschätzern vor allem dadurch, dass das Ergebnis nicht ein einziger Wert ist, sondern ein Intervall, in dem der wahre Wert des Parameters (der Funktion des Parameters) mit einer vorgegebenen Wahrscheinlichkeit liegt.

Definition 12.8. *Sei $\alpha \in [0,1]$ und $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell sowie $m : \Theta \to \Theta' \subseteq \mathbb{R}$. Jedes (von X abhängige) Intervall $(\underline{t}(X), \overline{t}(X))$ mit*

$$\mathbf{P}_\vartheta(m(\vartheta) \in (\underline{t}(X), \overline{t}(X))) \geq 1 - \alpha$$

heißt Konfidenzintervall *für $m(\vartheta)$ zum Konfidenzniveau $1-\alpha$.*

Beispiel 12.9 (Konfidenzintervall im Normalverteilungsmodell)**.** Im Normalverteilungsmodell $(X = (X_1, ..., X_n), (\mathbf{P}_\mu = N(\mu, \sigma^2))_{\mu\in\mathbb{R}})$ bei gegebener Varianz suchen wir ein Konfidenzintervall für μ zum Signifikanzniveau $1-\alpha$. Dies ist gegeben als

$$(\overline{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n})$$

(wobei z_x das x-Quantil von $N(0,1)$ ist), denn (mit $Z \sim N(0,1)$)

$$\begin{aligned}
&\mathbf{P}_\mu\Big(\mu \in (\overline{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n})\Big) \\
&\quad = \mathbf{P}_\mu\Big(\frac{|\overline{X} - \mu|}{\sqrt{\sigma^2/n}} \leq z_{1-\alpha/2}\Big) = \mathbf{P}(|Z| \leq z_{1-\alpha/2}) = \mathbf{P}(Z < z_{1-\alpha/2}) - \mathbf{P}(Z < -z_{1-\alpha/2}) \\
&\quad = 1 - \alpha/2 - \alpha/2 = 1 - \alpha.
\end{aligned}$$

Es gibt eine Dualität zwischen Konfidenzintervallen und (nicht-randomisierten) Tests. Das bedeutet, dass man oftmals aus Konfidenzintervallen zum Niveau $1-\alpha$ einen Test zum Niveau α herstellen kann und umgekehrt.

Proposition 12.10 (Konfidenzintervalle und Tests)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $m : \Theta \to \Theta' \subseteq \mathbb{R}$ und $I(X) := (\underline{t}(X), \overline{t}(X))$ ein (zufälliges) Intervall. Dann sind äquivalant:*

1. *Das Intervall $I(X)$ ist ein Konfidenzintervall für $m(\vartheta)$ zum Niveau $1-\alpha$.*
2. *Für jedes $\vartheta^* \in \Theta$ ist*
$$\varphi_{\vartheta^*}(X) := 1_{m(\vartheta^*)\notin I(X)}$$
ein (nicht-randomisierter) Test für $H_0 : \vartheta \in m^{-1}(m(\vartheta^*))$ gegen $H_A : \vartheta \notin m^{-1}(m(\vartheta^*))$ zum Niveau α.

Beweis. 1.⇒2.: Gilt H_0, so ist $\vartheta \in m^{-1}(m(\vartheta^*))$ (also $m(\vartheta) = m(\vartheta^*)$). Daraus berechnen wir

$$\mathbf{E}_\vartheta[\varphi_{\vartheta^*}(X)] = \mathbf{P}_\vartheta(m(\vartheta^*) \notin (\underline{t}(X), \overline{t}(X)) = \mathbf{P}_\vartheta(m(\vartheta) \notin (\underline{t}(X), \overline{t}(X)) \leq \alpha.$$

2.⇒1.: Hier ist für jedes $\vartheta^* \in \Theta$

$$\mathbf{P}_{\vartheta^*}(m(\vartheta^*) \in (\underline{t}(X), \overline{t}(X)) = 1 - \mathbf{E}_{\vartheta^*}[\varphi_\vartheta(X)] \geq 1 - \alpha.$$

□

Beispiel 12.11 (Konfidenzintervall und Test im Normalverteilungsmodell)**.** Für das Normalverteilungsmodell bei bekannter Varianz aus Beispiel 12.9 hatten wir das Konfidenzintervall

$$I(X) = \left(\overline{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n}\right)$$

bestimmt. Wir bemerken, dass

$$\mu \in I(X) \iff Z := \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \in (z_{\alpha/2}, z_{1-\alpha/2}).$$

In der Tat ist nach Proposition 12.7(a) $(Z, (-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, \infty))$ ein (unverfälschter) Test von $H_0 : \Theta_0 = \{\mu\}$ gegen $H_A : \Theta_A = \mathbb{R} \setminus \{\mu\}$ zum Niveau α.

12.3. Optimale Tests

Wie bei Schätzern will man auch bei Tests Optimalität erreichen. Zunächst muss also wieder geklärt werden, was man darunter versteht.

Definition 12.12 (Gütefunktion, Macht eines Tests, Optimaler Test)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und φ ein statistischer Test von $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \Theta_A$ zum Niveau $\alpha \in [0, 1]$.*

1. *Die Funktion*
$$g_\varphi : \begin{cases} \Theta & \to [0,1] \\ \vartheta & \mapsto \mathbf{E}_\vartheta[\varphi(X)] \end{cases}$$
heißt Gütefunktion *von φ. Für $\vartheta \in \Theta_A$ heißt $g_\varphi(\vartheta)$ auch die* Macht *des Tests φ in ϑ.*

2. *Ist φ' ein weiterer Test von H_0 gegen H_A zum Niveau α, so heißt φ* gleichmäßig besser *als φ', falls*
$$g_\varphi(\vartheta) \geq g_{\varphi'}(\vartheta)$$
für alle $\vartheta \in \Theta_A$ gilt.

3. *Ein Test φ zum Signifikanzniveau α heißt UMP (uniformly most powerful) für $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \Theta_A$, falls φ gleichmäßig besser ist als jeder Test φ' von H_0 gegen H_A zum Niveau α, d.h.*
$$g_\varphi(\vartheta) \geq g_{\varphi'}(\vartheta), \qquad \vartheta \in \Theta_A.$$

Bemerkung 12.13 (Idealer Test, Fehler zweiter Art)**.** 1. Ein idealer Test (Y, T) (den es in realen Situationen nie gibt) hätte die Eigenschaft, dass $Y \in C$ nur für $\vartheta \in \Theta_A$ möglich ist und weiter, dass $\mathbf{P}_\vartheta(Y \in C) = 1$ für $\vartheta \in \Theta_A$. Das bedeutet, dass $g_Y(\vartheta) = 1_{\vartheta\in\Theta_A}$ die Gütefunktion eines idealen Tests ist. Für einen solchen Test wäre die Macht für alle $\vartheta \in \Theta_A$ gleich 1.

2. Für $\vartheta \in \Theta_A$ ist $1 - g_\varphi(\vartheta)$ (also 1-*Macht des Tests bei* ϑ) die Wahrscheinlichkeit für einen Fehler zweiter Art.

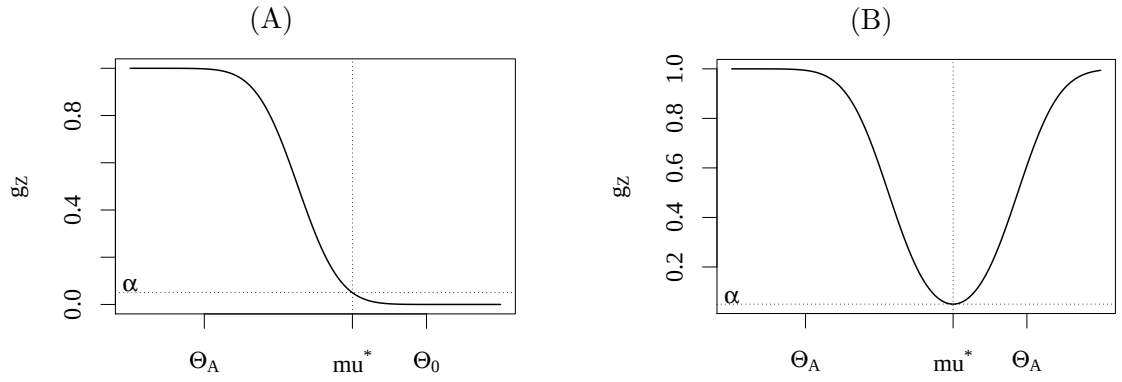


Abbildung 12.1.: Die Gütefunktionen für den einseitigen (A) und den zweiseitigen (B) Gauss-Test aus Beispiel 12.14.

Beispiel 12.14 (Gauss-Test)**.** Betrachten wir die Situation des Gauss-Tests φ aus Proposition 12.7. Sei $\widetilde{Z}$ eine (unter allen $\mathbf{P}_\mu$) nach $N(0,1)$ verteilte Zufallsvariable. Hier gilt im Fall des einseitigen Tests (c)

$$\begin{aligned} g_\varphi(\mu) &= \mathbf{P}_\mu(Z \le z_\alpha) = \mathbf{P}_\mu\Big(\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} + \frac{\mu-\mu^*}{\sqrt{\sigma^2/n}} \le z_\alpha\Big) = \mathbf{P}\Big(\widetilde{Z} \le z_\alpha + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big) \\ &= \Phi\Big(z_\alpha + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big), \end{aligned}$$

wobei $\Phi(.)$ die Verteilungsfunktion von $N(0,1)$ ist. Analog ist für den zweiseitigen Test (a)

$$\begin{aligned} g_\varphi(\mu) &= \mathbf{P}_\mu(Z \le z_{\alpha/2}) + \mathbf{P}_\mu(Z \ge z_{1-\alpha/2}) \\ &= \mathbf{P}\Big(\widetilde{Z} \le z_{\alpha/2} + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big) + \mathbf{P}\Big(Z \ge z_{1-\alpha/2} + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big) \\ &= 1 - \Phi\Big(z_{1-\alpha/2} + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big) + \Phi\Big(z_{\alpha/2} + \frac{\mu^*-\mu}{\sqrt{\sigma^2/n}}\Big). \end{aligned}$$

Diese zwei Gütefunktionen sind in Abbildung 12.1 dargestellt.

Die Optimalität eines Tests kann man zunächst am besten zeigen, wenn sowohl H_0 als auch H_A einfache Hypothesen sind. Hier hilft das Neyman-Pearson-Lemma 12.17, wo gezeigt wird, dass *Likelihood-Quotienten-Tests* optimal sind.

Definition 12.15 (Likelihood-Quotienten-Test)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta = \Theta_0 \cup \Theta_A$, $\Theta_0 = \{\vartheta_0\}, \Theta_A = \{\vartheta_A\}$ und $\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx$. Dann heißt*

$$L(x, \vartheta_0, \vartheta_A) := \frac{p_{\vartheta_A}(x)}{p_{\vartheta_0}(x)}$$

Likelihood-Quotient *und $\varphi : S \to [0,1]$* Likelihood-Quotienten-Test *von H_0 gegen H_A, falls für ein $k \in [0,\infty]$ und $\gamma : S \to [0,1]$*

$$\varphi(x) = \begin{cases} 1, & L(x,\vartheta_0,\vartheta_A) > k, \\ \gamma(x), & L(x,\vartheta_0,\vartheta_A) = k, \\ 0, & L(x,\vartheta_0,\vartheta_A) < k. \end{cases}$$

Beispiel 12.16 (Likelihood-Quotienten-Test im Binomialmodell)**.** Wir wollen im Binomialmodell $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ (d.h. $X_*\mathbf{P}_\vartheta = B(n,\vartheta)$) mit $\Theta = \{p, q\}$ einen Likelihood-Quotiententest für $H_0 : \vartheta = p$ gegen $H_A : \vartheta = q$ (mit $q > p$) aufstellen. Wir berechnen

$$L(x, p, q) = \frac{q^x(1-q)^{n-x}}{p^x(1-p)^{n-x}}.$$

Da $q/p > 1$, sind die Abbildungen $x \mapsto (q/p)^x$ und $x \mapsto \frac{(1-q)^{n-x}}{(1-p)^{n-x}}$ und damit $x \mapsto L(x, p, q)$ monoton wachsend. Damit ist jeder Test φ mit

$$\varphi(x) = \begin{cases} 1, & x = \ell+1, ..., n, \\ \gamma, & x = \ell, \\ 0, & x = 0, ..., \ell-1 \end{cases}$$

ein Likelihood-Quotiententest für $k = L(\ell, p, q)$ zum Niveau

$$\mathbf{E}_p[\varphi(X)] = \gamma\mathbf{P}_p[X = \ell] + \mathbf{P}(X \in \{\ell+1, ..., n\}].$$

Insbesondere ist der randomisierte Test aus Beispiel 12.6 ein Likelihood-Quotienten-Test.

Theorem 12.17 (Neyman-Pearson-Lemma)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta = \Theta_0 \cup \Theta_A$, $\Theta_0 = \{\vartheta_0\}, \Theta_A = \{\vartheta_A\}$ und $\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx$. Sei φ ein Likelihood-Quotienten-Test (mit k und γ) und $\varphi' : S \to [0, 1]$ ein weiterer Test mit $g_{\varphi'}(\vartheta_0) \le g_\varphi(\vartheta_0)$ (d.h. φ' ist Test zum selben Niveau wie φ). Dann gilt*

$$g_\varphi(\vartheta_A) \ge g_{\varphi'}(\vartheta_A),$$

d.h. φ ist UMP-Test zum Niveau $g_\varphi(\vartheta_A)$.

Beweis. Sei zunächst $k \in [0, \infty)$. Es gilt für alle $x \in S$

$$\varphi'(x)(p_{\vartheta_A}(x) - kp_{\vartheta_0}(x)) \le \varphi(x)(p_{\vartheta_A}(x) - kp_{\vartheta_0}(x)).$$

In der Tat: Gilt $p_{\vartheta_A}(x) - kp_{\vartheta_0}(x) > 0$, so ist $\varphi(x) = 1$ und die Behauptung folgt aus $\varphi'(x) \le 1$. Gilt $p_{\vartheta_A}(x) - kp_{\vartheta_0}(x) = 0$, so ist die Aussage trivial, und ist $p_{\vartheta_A}(x) - kp_{\vartheta_0}(x) < 0$, so ist $\varphi(x) = 0$ und die Aussage folgt aus $\varphi'(x) \ge 0$. Nun folgt

$$\begin{aligned}\mathbf{E}_{\vartheta_A}[\varphi'(X)] - k\mathbf{E}_{\vartheta_0}[\varphi'(X)] &= \int \varphi'(x)(p_{\vartheta_A}(x) - kp_{\vartheta_0}(x))dx \\ &\le \int \varphi(x)(p_{\vartheta_A}(x) - kp_{\vartheta_0}(x))dx = \mathbf{E}_{\vartheta_A}[\varphi(X)] - k\mathbf{E}_{\vartheta_0}[\varphi(X)],\end{aligned}$$

also

$$g_{\varphi'}(\vartheta_A) - g_\varphi(\vartheta_A) \le k(g_{\varphi'}(\vartheta_0) - g_\varphi(\vartheta_0)) \le 0.$$

Daraus folgt die Behauptung. Der Fall $k = \infty$ wird in einer Übungsaufgabe behandelt. □

Beispiel 12.18 (Likelihood-Quotienten-Test im Normalverteilungsmodell)**.** Sei $\sigma^2 > 0$ bekannt. Wir betrachten das Normalverteilungsmodell $(X = (X_1, ..., X_n), (\mathbf{P}_\mu)_{\mu\in\mathbb{R}})$. Zunächst schränkten wir unseren Parameterbereich ein und setzen $\Theta_0 = \{\nu_0\}$ und $\Theta_A = \{\nu_A\}$ mit $\nu_A > \nu_0$. Es ist

$$2\sigma^2 \log L(x, \nu_0, \nu_A) = \sum_{i=1}^{n}(x_i - \nu_0)^2 - \sum_{i=1}^{n}(x_i - \nu_A)^2 = n\nu_0^2 - n\nu_A^2 + 2(\nu_A - \nu_0)\sum_{i=1}^{n} x_i.$$

Da monotone wachsende Funktionen von L ebenfalls zu optimalen Tests führen, ist

$$Y = \frac{\sum_{i=1}^{n} X_i - \nu_0}{\sqrt{\sigma^2 n}}$$

mit $Y_*\mathbf{P}_{\nu_0} = N(0, 1)$ eine Teststatistik und für $\alpha \in (0, 1)$ ist mit dem Ablehnungsbereich $C = (Y, [z_{1-\alpha}, \infty))$ der Test (Y, C) UMV zum Niveau α für die Hypothese $H_0 : \mu = \nu_0$ gegen $\mu = \nu_A$.

Bislang können wir optimale Tests nur für einfache Null- und Alternativhypothesen mit Hilfe des Neyman-Pearson-Lemmas angeben. Die Übertragung auf zusammengesetzte Hypothesen gelingt uns zumindest im Falle monotoner Dichtequotienten.

Definition 12.19 (Monotoner Dichtequotient)**.** *Ein statistisches Modell $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta \subseteq \mathbb{R}$ hat einen* monotonen Dichtequotienten *bezüglich $t : S \to \mathbb{R}$, falls für alle $\vartheta < \vartheta'$ der Likelihood-Quotient*

$$x \mapsto \frac{p_{\vartheta'}(x)}{p_\vartheta(x)}$$

eine streng monotone wachsende Funktion in $t(x)$ ist.

Beispiel 12.20 (Ein-parametrige Exponentialfamilie)**.** Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ eine ein-parametrige Exponentialfamilie mit

$$\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx = h(x)\exp\Big(c(\vartheta)t(x) + d(\vartheta)\Big)dx,$$

so hat diese genau einen monotonen Dichtequotienten bezüglich t, falls c streng monoton wachsend ist.
Denn: Für den Likelihood-Quotient ist

$$\log\frac{p_{\vartheta'}(x)}{p_\vartheta(x)} = (c(\vartheta') - c(\vartheta))t(x) + d(\vartheta') - d(\vartheta).$$

Offenbar ist dies für $\vartheta < \vartheta'$ genau dann eine streng monoton wachsende Funktion in $t(x)$, wenn c streng monoton wächst.

Theorem 12.21 (Neyman-Pearson-Lemma unter monotonen Dichtequotienten)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$ und*

$$\Theta_0 = \Theta \cap (-\infty, \theta_0], \qquad \Theta_A = \Theta \cap (\theta_0, \infty),$$

$\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx$ und einem monotonen Dichtequotienten bezüglich $t : S \to \mathbb{R}$. Gilt für ein k und ein γ

$$\varphi(x) = \begin{cases} 1, & t(x) > k, \\ \gamma, & t(x) = k, \\ 0, & t(x) < k, \end{cases}$$

so ist φ ein Likelihood-Quotienten-Test (und damit ein UMP-Test) für jedes Paar $H_0 : \Theta_0 = \{\vartheta\}$ mit $\vartheta \in \Theta_0$ gegen $H_A : \Theta_A = \{\vartheta'\}$ mit $\vartheta' \in \Theta_A$. Weiter ist φ ein UMP-Test für $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \Theta_A$ zum Niveau $\mathbf{P}_{\vartheta_0}(t(X) > k) + \gamma \mathbf{P}_{\vartheta_0}(t(X) = k)$.

Beweis. Zunächst ist zu zeigen, dass $t(x) > k$ genau dann gilt, wenn $L(x, \vartheta, \vartheta') > \widetilde{k}$ für ein passendes $\widetilde{k}$. Da $(\mathbf{P}_\vartheta)_{\vartheta\in\Theta}$ einen monotonen Dichtequotienten bezüglich t besitzt, gibt es ein streng monotones $f_{\vartheta,\vartheta'}$ mit $\frac{p_{\vartheta'}(x)}{p_\vartheta(x)} = f_{\vartheta,\vartheta'}(t(x))$. Insbesondere ist $t(x) > k$ genau dann, wenn $\widetilde{k} := f_{\vartheta,\vartheta'}(k) < \frac{p_{\vartheta'}(x)}{p_\vartheta(x)} = L(x, \vartheta, \vartheta')$.

Weiter betrachten wir nun die Gütefunktion von φ und stellen fest, dass

$$\vartheta \mapsto g_\varphi(\vartheta) = \mathbf{P}_\vartheta(t(X) > k) + \gamma \mathbf{P}_\vartheta(t(X) = k)$$

monoton wachsend ist. Nun ist nach Definition φ ein UMP-Test für $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \{\vartheta'\}$ für jedes $\vartheta' \in \Theta_A$ mit Signifikanzniveau

$$\sup_{\vartheta\in\Theta_0} \mathbf{E}_\vartheta[\varphi(X)] = \sup_{\vartheta\in\Theta_0} g_\varphi(\vartheta) = g_\varphi(\vartheta_0).$$

Daraus folgt dann aber auch $g_\varphi(\vartheta) \geq g_{\varphi'}(\vartheta)$ für alle $\vartheta' \in \Theta_A$. Mit anderen Worten ist φ ein UMP-Test für $H_0 : \vartheta \in \Theta_0$ gegen $H_A : \vartheta \in \Theta_A$. □

Definition 12.22 (Verallgemeinerte Likelihood-Quotienten-Tests)**.** *Sei $(X, (\mathbf{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\mathbf{P}_\vartheta(X \in dx) = p_\vartheta(x)dx$ und $\Theta = \Theta_0 + \Theta_A$. Der* verallgemeinerte Likelihood-Quotient *ist gegeben durch*

$$L(x, \Theta_A, \Theta_0) := \frac{\sup_{\theta\in\Theta_A} p_\vartheta(x)}{\sup_{\theta\in\Theta_0} p_\vartheta(x)}.$$

Gilt für ein k und ein $\gamma : S \to [0, 1]$ für einen Test φ, dass

$$\varphi(x) = \begin{cases} 1, & L(x, \Theta_A, \Theta_0) > k, \\ \gamma(x), & L(x, \Theta_A, \Theta_0) = k, \\ 0, & L(x, \Theta_A, \Theta_0) < k, \end{cases}$$

so heißt φ verallgemeinerter Likelihood-Quotienten-Test (mit k und γ).

Bemerkung 12.23. In der Praxis berechnet man nicht L, sondern

$$\lambda(x) = \max\{L(x, \Theta_0, \Theta_A), 1\} = \frac{\sup_{\theta\in\Theta} p_\vartheta(x)}{\sup_{\theta\in\Theta_0} p_\vartheta(x)},$$

also eine monotone Funktion von L. Die Suprema erhält man aus den Maximum-Likelihood-Schätzern in Θ_0 und in $\Theta \supseteq \Theta_0$.

13. Einige statistische Tests

Es gibt sehr viele statistische Tests. Wir geben hier nun eine Auswahl wichtiger Tests, inklusive der dazugehörigen Teststatistiken und deren Verteilungen an. Wir konzentrieren uns dabei zunächst auf Tests in Normalverteilungsmodellen (etwa auf Gleichheit von Erwartungswerten oder Varianzen). Anschließend gehen wir auf verteilungsfreie Verfahren ein, d.h. dass das zugrunde liegende statistische Modell nicht-parametrisch (also unendlich-dimensional) ist. Schließlich beschäftigen wir uns mit linearen Modellen.

13.1. Aus der Normalverteilung abgeleitete Verteilungen

Definition 13.1 (Die χ^2- t- und F-Verteilung)**.** *Seien $X, X_1, X_2, ...$ unabhängige, nach $N(0,1)$ verteilte Zufallsvariablen.*

1. *Die Verteilung der Zufallsvariable*

$$X_1^2 + \cdots + X_n^2$$

heißt (zentrierte) χ^2-Verteilung mit n Freiheitsgraden *und wird mit $\chi^2(n)$ bezeichnet.*

2. *Die Verteilung von*

$$\frac{X}{\sqrt{(X_1^2 + \cdots + X_n^2)/n}}$$

heißt t-Verteilung mit n Freiheitsgraden *und wird mit $t(n)$ bezeichnet.*

3. *Sei $Y \sim \chi^2(m)$ und $Z \sim \chi^2(n)$. Dann heißt die Verteilung von*

$$\frac{Y/m}{Z/n}$$

F-Verteilung mit m und n Freiheitsgraden *und wird mit $F(m,n)$ bezeichnet.*

Bemerkung 13.2. Von χ^2-, t- und F-Verteilung können jeweils Dichten angegeben werden. Dies ist für uns im Folgenden allerdings nicht interessant, so dass wir nur die Ergebnisse angeben:
Die $\chi^2(n)$-Verteilung hat die Dichte[1]

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}.$$

[1] Wir erinnern an die Definition der Γ-Funktion

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

Sie hat Erwartung n und Varianz $2n$. Die $t(n)$-Verteilung hat die Dichte

$$g_n(x)\frac{1}{\sqrt{n\pi}} \cdot \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \cdot \frac{1}{(1+x^2/n)^{(n+1)/2}}.$$

Der Erwartungswert existiert für $n \geq 2$ und ist dann 0. Die Varianz existiert für $n \geq 3$ und ist dann $n/(n-2)$. Die $F(m,n)$-Verteilung hat die Dichte

$$h_{m,n}(x) = m^{m/2}n^{n/2}\frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}\frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}1_{x>0}.$$

Der Erwartungswert existiert für $n > 2$ und ist dann $n/(n-2)$, die Varianz existiert für $n > 4$ und ist dann $2n^2(m+n-2)/(m(n-2)^2(n-4))$.

Wir kommen nun zu einem Satz, den wir später noch oft brauchen werden. Wir erinnern an die Begriffe des Mittelwertes und der empirischen Varianz aus Proposition 11.6.

Theorem 13.3 (Satz von Fisher)**.** *Sei $n > 1$, $X = (X_1, .., X_n)$ unabhängig nach $N(\mu, \sigma^2)$ verteilt, $\overline{X}$ der Mittelwert und $s^2(X)$ die empirische Varianz. Dann ist*

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

und

$$(n-1)s^2(X)/\sigma^2 \sim \chi^2(n-1).$$

Außerdem sind $\overline{X}$ und $s^2(X)$ stochastisch unabhängig und

$$T := \frac{\overline{X}-\mu}{\sqrt{s^2(X)/n}} \sim t(n-1). \tag{13.1}$$

Bemerkung 13.4 (Interpretation)**.** Stellen wir uns vor, es liegt uns eine Stichprobe $X = (X_1, \ldots, X_n)$ vor, wobei wir davon ausgehen dürfen, dass $X_1, \ldots, X_n$ unabhängig sind und X_i normalverteilt sind. Wollen wir etwa die Verteilung des Mittelwertes berechnen, wissen wir, dass

$$\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$$

nach $N(0,1)$ verteilt ist. Eine gängige Situation ist nun die, dass wir eine Vermutung haben, wie groß μ ist, jedoch nicht wissen, wie groß σ^2 ist. Es liegt nun nahe, σ^2 in der letzten Formel durch $s^2(X)$ zu ersetzen, wie es in T aus (13.1) geschehen ist. Durch das Ersetzen der festen Größe σ^2 durch die Zufallsvariable $s^2(X)$ verändert sich natürlich die Verteilung. Der Satz von Fisher besagt nun, dass die Verteilung der Standardisierung von $\overline{X}$, wenn man σ^2 durch $s^2(X)$ ersetzt, nach $t(n-1)$ verteilt ist. Bemerkenswert ist dabei, dass die Verteilung von T nicht mehr von σ^2 abhängt.

Beweis von Theorem 13.3. Wir setzen $Z = (Z_1, \ldots, Z_n)$ mit $Z_i := (X_i - \mu)/\sigma$. Wir wissen bereits, dass die Z_i unabhängig und nach $N(0,1)$ verteilt sind. Weiter ist

$$\overline{Z} = \tfrac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} = (\overline{X}-\mu)/\sigma$$

und

$$(n-1)s^2(Z) = \sum_{i=1}^{n} \Big(\frac{X_i - \mu - \overline{X} + \mu}{\sigma}\Big)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = (n-1)s^2(X)/\sigma^2.$$

Damit ist

$$T = \frac{\overline{X} - \mu}{\sqrt{s^2(X)/n}} = \frac{\sigma \cdot \overline{Z}}{\sqrt{\sigma^2 s^2(Z)/n}} = \frac{\overline{Z}}{\sqrt{s^2(Z)/n}}$$

und damit genügt es, die Behauptungen für den Vektor Z zu zeigen.

Wir wählen nun eine orthogonale Matrix $O = (o_{ij})_{1\le i,j\le n}$ mit

$$a_{11} = \cdots = a_{1n} = \frac{1}{\sqrt{n}}$$

und setzen $W = OZ$. Wir verwenden im Beweis die aus Korollar 10.4 bekannte Tatsache, dass W ein Vektor unabhängiger, nach $N(0,1)$ verteilter Zufallsvariablen ist. Weiter ist

$$\begin{aligned} W_1 &= a_{11}Z_1 + \cdots + a_{1n}Z_n = \sqrt{n}\cdot\overline{Z}, \\ (n-1)s^2(Z) &= \sum_{i=1}^{n}(Z_i - \overline{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2 \\ &= \sum_{i=1}^{n} Z_i^2 - W_1^2 = \sum_{i=2}^{n} W_i^2 \end{aligned}$$

da O eine orthogonale Matrix ist und damit

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}(OZ)_i^2 = \sum_{i=1}^{n} W_i^2.$$

Insbesondere haben wir eben gezeigt, dass $(n-1)s^2(X)$ die Summe von $n-1$ unabhängigen, nach $N(0,1)$ verteilten Zufallsvariablen ist und damit $\chi^2(n-1)$ verteilt. Da weiter W_1 von $(W_2, \ldots, W_n)$ unabhängig ist, folgt auch, dass $\overline{X}$ von $s^2(X)$ unabhängig ist.

Es bleibt nun zu zeigen, dass T nach $t(n-1)$ verteilt ist. Wir schreiben

$$T = \frac{\overline{Z}}{\sqrt{s^2(Z)/n}} = \frac{\sqrt{n}\overline{Z}}{\sqrt{s^2(Z)}} = \frac{W_1}{\sqrt{(W_2^2 + \cdots + W_n^2)/(n-1)}}.$$

Da $W_1, \ldots, W_n$ unabhängig und nach $N(0,1)$ verteilt sind, folgt, dass T nach $t(n-1)$ verteilt ist. □

13.2. Parametertests bei normalverteilten Daten

Der in Proposition 12.7 besprochene Gauss-Test fällt bereits in die Klasse der Parametertests. Unter diesem Stichwort versteht man statistische Tests, die eine Stichprobe daraufhin testen, ob Parameter der zugrunde liegenden Verteilung gewisse Werte annehmen. Wir werden in diesem Kapitel solche Parametertests für normalverteilten Stichproben kennenlernen, nämlich: den einfachen t-Test, der testet, ob der Erwartungswert einer normalverteilten Stichprobe einen bestimmten Wert annimmt (Proposition 13.5); den doppelten t-Test, der testet, ob die Erwartungswerte von zwei unverbundenen Stichproben identisch sind (Proposition 13.9); den F-Test, der testet, ob die Varianzen zweier normalverteilter Stichproben gleich sind.

Proposition 13.5 (Einfacher t-Test)**.**
Sei $\alpha \in [0,1]$, $\mu^* \in \mathbb{R}$ und $(X = (X_1, \dots, X_n), (\mathbf{P}_{\mu,\sigma^2})_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+})$ ein statistisches Modell, so dass $X_1, \dots, X_n$ unter $\mathbf{P}_{\mu,\sigma^2}$ unabhängig und nach $N(\mu, \sigma^2)$ verteilt sind. Weiter sei

$$T := \frac{\overline{X} - \mu^*}{\sqrt{s^2(X)/n}}.$$

und $t_{n-1,p}$ für $p \in [0,1]$ das p-Quantil von $t(n-1)$.

(a) Ist $\Theta_0 = \{\mu^*\} \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = (-\infty, \mu^*] \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, [t_{n-1,1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = [\mu^*, \infty) \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha}])$ ein unverfälschter Test zum Niveau α.

Beweis. Wir beweisen wir nur (c), da die anderen beiden Aussagen analog folgen. Genau wie im Gauss-Test ist klar, dass der Test unverfälscht ist. Aus dem Satz von Fisher, Theorem 13.3, folgt dass T unter $\mathbf{P}_{\mu^*,\sigma^2}$ nach $t(n-1)$ verteilt ist. Damit gilt

$$\sup_{\mu\geq\mu^*} \mathbf{P}_{\mu,\sigma^2}(T \leq t_{n-1,\alpha}) = \mathbf{P}_{\mu^*,\sigma^2}(T \leq t_{n-1,\alpha}) = \alpha,$$

woraus die Behauptung sofort folgt. □

Bemerkung 13.6 (Vergleich von Gauss-Test und einfachem t-Test)**.** Sowohl der Gauss-Test aus Proposition 12.7, als auch der einfache t-Test basieren auf unabhängigen, normalverteilten Stichproben. Der entscheidende Unterschied der beiden Tests besteht darin, dass beim Gauss-Test die Varianz der zugrunde liegenden Normalverteilung bekannt sein muss, und beim t-Test nicht. Dies sieht man etwa daran, dass beim Gauss-Test die Nullhypothese nur aus einem Bereich für μ, nicht jedoch für σ^2 besteht. Außerdem kann man ja nur bei Kenntnis von σ^2 die Teststatistik Z aus Proposition 12.7 berechnen. Beim t-Test ersetzt man σ^2 durch die empirische Varianz $s^2(X)$, und erhält die Teststatistik T.

Bemerkung 13.7 (Gütefunktion des einfachen t-Tests und Stichprobengröße)**.** Wir betrachten die Situation aus Proposition 13.5(b) mit $\mu^* = 0$ und $\alpha = 5\,\%$. Durch das Signifikanzniveau α wird die Wahrscheinlichkeit für einen Fehler erster Art kontrolliert. Die Wahrscheinlichkeit für einen Fehler zweiter Art wird durch die Gütefunktion angegeben. Ist nämlich $\mu > 0$, so ist

$$\mathbf{P}_{\mu,\sigma^2}(\text{Fehler zweiter Art}) = \mathbf{P}_{\mu,\sigma^2}(T \notin C) = 1 - g_T(\mu).$$

Wir fragen nun, wie gut wir diesen Fehler zweiter Art kontrollieren können, wenn $\mu > 0$ gegeben ist. Die Gütefunktion ist gegeben als

$$\begin{aligned} g_T(\mu) &= \mathbf{P}_{\mu,\sigma^2}(T \in C) = \mathbf{P}_{\mu,\sigma^2}\big(T \geq t_{n-1,1-\alpha}\big) \\ &= \mathbf{P}_{\mu,\sigma^2}\Big(\widetilde{T} + \frac{\mu}{\sqrt{s^2(X)/n}} \geq t_{n-1,1-\alpha}\Big) \end{aligned}$$

für die unter $\mathbf{P}_{\mu,\sigma^2}$ nach $t(n-1)$ verteilte Zufallsgröße

$$\widetilde{T} = \frac{\overline{X} - \mu}{\sqrt{s^2(X)/n}}.$$

Ist n groß, so ist $\widetilde{T}$ etwa nach $N(0,1)$ verteilt und $s^2(X) \approx \sigma^2$. Damit ist für eine nach $N(0,1)$ verteilte Zufallsvariable Z

$$\mathbf{P}_{\mu,\sigma^2}(\text{Fehler zweiter Art}) \approx \mathbf{P}\Big(Z + \frac{\mu}{\sqrt{\sigma^2/n}} \le z_{1-\alpha}\Big).$$

Wollen wir etwa erreichen, dass die Wahrscheinlichkeit für einen Fehler zweiter Art nicht größer als 5 % ist, bedeutet das, dass

$$\frac{\mu}{\sqrt{\sigma^2/n}} \ge 3.29 \tag{13.2}$$

gelten muss, denn

$$\mathbf{P}\Big(Z + 3.29 \le 1.645\Big) = \mathbf{P}\Big(Z \le -1.645\Big) = 5\,\%.$$

Man beachte, dass (13.2) eine Bedingung für die Stichprobengröße liefert. Will man etwa $\mu = 0.1$ (bei $\sigma^2 = 1$) noch mit einem Fehler zweiter Art von 5 % ablehnen können, so muss

$$\sqrt{n} \ge 3.29/0.1 = 32.9, \qquad n \ge 1083$$

gewählt werden.

Korollar 13.8 (Gepaarter t-Test)**.**
Sei $\alpha \in [0,1]$ und $((X,Y) = (X_1,\dots,X_n,Y_1,\dots,Y_n)), (\mathbf{P}_{\mu,\sigma^2})_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+})$ ein statistisches Modell, so dass $Y - X := (Y_1 - X_1,\dots,Y_n - X_n)$ unter $\mathbf{P}_{\mu,\sigma^2}$ unabhängig und nach $N(\mu,\sigma^2)$ verteilt sind. Weiter sei

$$T := \frac{\overline{Y} - \overline{X}}{\sqrt{s^2(Y-X)/n}}.$$

und $t_{n,p}$ für $p \in [0,1]$ das p-Quantil von $t(n)$.

(a) Ist $\Theta_0 = \{0\} \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = (-\infty, 0] \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, [t_{n-1,1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = [\mu^*, \infty) \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha}])$ ein unverfälschter Test zum Niveau α.

Beweis. Man wendet einfach Proposition 13.5 auf den Vektor $Y - X$ an. □

Proposition 13.9 (Doppelter t-Test)**.**
Sei $\alpha \in [0,1]$ und $((X,Y) = (X_1,\dots,X_m,Y_1,\dots,Y_n)), (\mathbf{P}_{\mu_X,\mu_Y,,\sigma^2})_{\mu_X,\mu_Y\in\mathbb{R},\sigma^2\in\mathbb{R}_+})$ ein statistisches Modell, so dass $X_1,\dots,X_m,Y_1,\dots,Y_n$ unter $\mathbf{P}_{\mu_X,\mu_Y,\sigma^2}$ unabhängig sind, sowie $X_1,\dots,X_m$ nach $N(\mu_X,\sigma^2)$ und $Y_1,\dots,Y_n$ nach $N(\mu_Y,\sigma^2)$ verteilt sind. Weiter sei

$$T := \frac{\overline{Y} - \overline{X}}{\sqrt{s^2(X,Y)(m+n)/(mn)}}$$

mit

$$s^2(X,Y) := \frac{1}{m+n-2}\Big(\sum_{i=1}^{m}(X_i-\overline{X})^2+\sum_{i=1}^{n}(Y_i-\overline{Y})^2\Big), \tag{13.3}$$

und $t_{n,p}$ *für* $p \in [0,1]$ *das* p*-Quantil von* $t(n)$*.*

(a) Ist $\Theta_0 = \{(\mu_X,\mu_Y) : \mu_X = \mu_Y\} \times \mathbb{R}_+$*,* $\Theta_A = \Theta \setminus \Theta_0$*, so ist* $(T, (-\infty, t_{n-1,\alpha/2}) \cup (t_{m+n-2,1-\alpha/2}, \infty))$ *ein unverfälschter Test zum Niveau* α*.*

(b) Ist $\Theta_0 = \{(\mu_X,\mu_Y) : \mu_Y \le \mu_X\} \times \mathbb{R}_+$*,* $\Theta_A = \Theta \setminus \Theta_0$*, so ist* $(T, [t_{m+n-2,1-\alpha}, \infty))$ *ein unverfälschter Test zum Niveau* α*.*

(c) Ist $\Theta_0 = \{(\mu_X,\mu_Y) : \mu_Y \ge \mu_X\} \times \mathbb{R}_+$*,* $\Theta_A = \Theta \setminus \Theta_0$*, so ist* $(T, (-\infty, t_{m+n-2,\alpha}])$ *ein unverfälschter Test zum Niveau* α*.*

Beweis. Wir müssen zunächst zeigen, dass T nach $t(m+n-2)$ verteilt ist. Ohne Einschränkung der Allgemeinheit sei $\sigma^2 = 1$. Da $X_1,\dots,X_m,Y_1,\dots,Y_n$ unabhängig sind, sind $s^2(X)$ und $s^2(Y)$ unabhängige Zufallsvariablen. Aus dem Satz von Fisher folgt, dass $\overline{X},\overline{Y},s^2(X),s^2(Y)$ unabhängig sind, $\overline{Y}-\overline{X}$ unter $\mathbf{P}_{0,\sigma^2}$ nach $N(0,\frac{1}{m}+\frac{1}{n})$ und $(m-1)s^2(X)$ nach $\chi^2(m-1)$ und $(n-1)s^2(Y)$ nach $\chi^2(n-1)$ verteilt ist. Damit ist

$$T = \frac{\frac{1}{\sqrt{1/m+1/n}}(\overline{Y}-\overline{X})}{\sqrt{((m-1)s^2(X)+(n-1)s^2(Y))/(m+n-2)}}$$

nach $t(m+n-2)$ verteilt. Der Rest des Beweises folgt wie beim einfachen t-Test, Proposition 13.5. □

Beispiel 13.10 (Geburtsgewichte)**.** In einer Kölner Klinik wurden im Jahr 1985 $m = 269$ Mädchen und $n = 288$ Jungen geboren. (Die einzelnen Geburtsgewichte sind also $X_1,\dots,X_{269}$ und $Y_1,\dots,Y_{288}$.) Das Durchschnittsgewicht und die empirische Varianz der Mädchen in Gramm war $\bar{X} = 3050$ und $s^2(X) = 211600$, das der Jungen $\bar{Y} = 3300$ und $s^2(Y) = 220900$. Es soll zum Signifikanzniveau $\alpha = 0.01$ getestet werden, ob Jungen und Mädchen das gleiche erwartete Geburtsgewicht haben (Fall (a) in Proposition 13.9). Wir berechnen

$$s^2(X,Y) = \frac{1}{269+288-2}(268\cdot s^2(X)+287\cdot s^2(Y)) = 216409$$

und

$$T = \frac{3300-3050}{\sqrt{216409\cdot(269+288)/(269\cdot 288)}} = 6.338 > 2.585 = t_{555,0.995}.$$

Also können wir die Nullhypothese $\mu_X = \mu_Y$ auf dem Signifikanzniveau $\alpha = 0.01$ ablehnen. Unter der (sehr plausiblen) Annahme der Normalverteilung und der Annahme der Varianzhomogenität kann man zum Niveau $\alpha = 0.01$ schließen, dass Jungen im Mittel schwerer sind als Mädchen.

Im doppelten t-Test ist die Annahme der Gleichheit der Varianzen wichtig. Sind die Varianzen nicht gleich, ist nämlich die Teststatistik T unter H_0 nicht $t(m+n-2)$-verteilt. Um die Gleichheit der Varianzen zu überprüfen, gibt es wiederum einen statistischen Test, den F-Test.

Proposition 13.11 (F-Test auf identische Varianzen)**.**
Sei $\alpha \in [0,1]$ und $((X,Y) = (X_1,\dots,X_m,Y_1,\dots,Y_n)), (\mathbf{P}_{\mu_X,\mu_Y,,\sigma_X^2,\sigma_Y^2})_{\mu_X,\mu_Y\in\mathbb{R},\sigma_X^2,\sigma_Y^2\in\mathbb{R}_+})$ ein statistisches Modell, so dass $X_1,\dots,X_m,Y_1,\dots,Y_n$ unter $\mathbf{P}_{\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2}$ unabhängig sind, sowie $X_1,\dots,X_m \sim N(\mu_X,\sigma_X^2)$ und $Y_1,\dots,Y_n \sim N(\mu_Y,\sigma_Y^2)$. Weiter sei

$$F := \frac{s^2(Y)}{s^2(X)},$$

wobei $s^2(X)$ und $s^2(Y)$ die empirischen Varianzen von X und Y sind. Weiter sei $F_{m,n,p}$ für $p \in [0,1]$ das p-Quantil von $F(m,n)$.

(a) Ist $\Theta_0 = \mathbb{R}^2 \times \{(\sigma_X^2,\sigma_Y^2) : \sigma_X^2 = \sigma_Y^2\} \times \mathbb{R}_+$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(F, (-\infty, F_{m-1,n-1,\alpha/2}) \cup (F_{m-1,n-1,1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = \mathbb{R}^2 \times \{(\sigma_X^2,\sigma_Y^2) : \sigma_Y^2 \le \sigma_X^2\}$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(F, [F_{m-1,n-1,1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = \mathbb{R}^2 \times \{(\sigma_X^2,\sigma_Y^2) : \sigma_Y^2 \ge \sigma_X^2\}$, $\Theta_A = \Theta \setminus \Theta_0$, so ist $(F, [0, F_{m-1,n-1,\alpha})$ ein unverfälschter Test zum Niveau α.

Beweis. Zunächst sind $(m-1)s^2(X)/\sigma_X^2 \sim \chi^2(m-1)$ und $(n-1)s^2(Y)/\sigma_Y^2 \sim \chi^2(n-1)$ unabhängig. für $\sigma_X^2 = \sigma_Y^2$ ist also $F \sim F(m-1,n-1)$. Weiter ist $F_*\mathbf{P}_{\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2}$ stochastisch wachsend in σ_Y^2 und fallend in σ_X^2. Der Rest des Beweises folgt wie beim einfachen t-Test, Proposition 13.5. □

Beispiel 13.12 (Geburtsgewichte)**.** Wir betrachten nochmal das Beispiel 13.10 und testen auf Gleichheit der Varianzen zum Signifikanzniveau 0.99. Mit $m = 269, n = 288$ und $s^2(X) = 211600, s^2(Y) = 22099$ ist $F = 1.043951$. Nach der Hypothese $\sigma_X^2 = \sigma_Y^2$ ist diese Teststatistik nach $F(268,287)$ verteilt. Es ist $F_{268,287,0.005} = 0.7326783$ und $F_{268,287,0.995} = 1.362907$. Damit lässt sich die Nullhypothese gleicher Varianzen nicht ablehnen.

13.3. Anpassungstests

Die bisher behandelten Tests basierten alle auf normalverteilten Stichproben. Etwa können wir mit den zuletzt behandelten t-Tests überprüfen, ob der Mittelwert einer Stichprobe vermutlich einen bestimmten Wert annimmt. Auffällig ist, dass diese t-Tests immer nur auf den Erwartungswert der zu Grunde liegenden Verteilung testen. Dies ist bei den χ^2-Tests, die wir in diesem Abschnitt kennen lernen, anders. Beim χ^2-Anpassungstest (Theorem 13.14) wird überprüft, ob die gesamte Verteilung einer Stichprobe (und nicht nur der Erwartungswert) eine bestimmte Form hat.

Beispiel 13.13 (Mendel's Experimente)**.** Der Naturforscher Gregor Mendel kreuzte rot- und weißblühende Erbsenpflanzen. In der ersten Nachkommensgeneration erhielt er ausschließlich rosa-blühende Pflanzen. Kreuzte er diese weiter, erhielt er in der nächsten Generation rot- rosa- und weißblühende Pflanzen. Seine Theorie sagte voraus, dass diese Farben in der zweiten Nachkommensgeneration im Verhältnis 1:2:1 auftreten sollten.

Ziel des χ^2-Anpassungstests in diesem Beispiel ist es, zu überprüfen, ob das Farbverhältnis von 1:2:1 von einer Stichprobe vermutlich eingehalten wird. Betrachten wir dazu $n = 400$ Pflanzen der zweiten Generation, von denen wir annehmen, dass deren Farbe Realisierung

von unabhängigen Experimenten ist. Sei $\mathcal{I} = \{\text{rot, rosa, weiß}\}$ und $X = (X_1, \dots, X_n)$ ein Vektor von unabhängigen, $\mathcal{I}$-wertigen Zufallsvariablen. Falls die Theorie von Mendel stimmt, gilt für alle i

$$\mathbf{P}(X_i = \text{rot}) = \tfrac{1}{4}, \qquad \mathbf{P}(X_i = \text{rosa}) = \tfrac{1}{2}, \qquad \mathbf{P}(X_i = \text{weiß}) = \tfrac{1}{4}.$$

Gleichbedeutend ist es, dass die Verteilungsgewichte der Verteilung von X_i durch den Vektor $\underline{\pi} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$ gegeben sind. Die Stichprobe liefert $S_{\text{weiß}} = 115$ weiße Blüten, $S_{\text{rosa}} = 171$ rosa Blüten und $S_{\text{rot}} = 114$ rote Blüten. Kann aufgrund dieser Beobachtung die Nullhypothese über das 1:2:1-Farbverhältnis abgelehnt werden?

Theorem 13.14 (χ^2-Anpassungstest)**.** *Sei $\alpha \in [0,1]$, $\mathcal{I}$ eine endliche Menge mit $|\mathcal{I}| = r$ und*

$$\Theta = \{\underline{p} = (p_i)_{i\in\mathcal{I}} \text{ Verteilungsgewichte einer Verteilung auf } \mathcal{I}\}.$$

Weiter sei $(X = (X_1, \dots, X_n), (\mathbf{P}_{\underline{p}})_{\underline{p}\in\Theta})$ ein statistisches Modell, so dass $X_1, \dots, X_n$ unter $\mathbf{P}_{\underline{p}}$ unabhängig und nach $\underline{p}$ verteilt sind (d.h. $\mathbf{P}(X_k = i) = p_i$). Setze für $i \in \mathcal{I}$

$$S_i := |\{k : X_k = i\}|.$$

Weiter sei $\chi^2_{m,p}$ für $p \in [0,1]$ das p-Quantil von $\chi^2(m)$ und $\Theta_0 = \{\underline{\pi}\}$, $\Theta_A = \Theta \setminus \Theta_0$ für ein $\underline{\pi} \in \Theta$. Für

$$\chi^2_n := \sum_{i\in\mathcal{I}} \frac{(S_i - n\pi_i)^2}{n\pi_i}.$$

ist $(\chi^2_n, (\chi^2_{r-1,1-\alpha}, \infty))$ im Grenzwert $n \to \infty$ ein Test zum Niveau α.

Bemerkung 13.15 (Approximativer Test)**.** Der χ^2-Anpassungstest ist ein approximativer Test für große Stichproben. Unter H_0 gilt nämlich für jedes x (also insbesondere für $x = \chi^2_{r-1,\alpha}$)

$$\lim_{n\to\infty} \mathbf{P}_{\underline{\pi}}(\chi^2_n > x) = \mathbf{P}(X > x)$$

für eine nach $\chi^2(r-1)$ verteilte Zufallsgröße X. In der Praxis ist die Stichprobengröße natürlich nie beliebig groß. Deswegen hat man Faustregeln für die Anwendbarkeit des χ^2-Anpassungstests aufgestellt. Man verwendet den χ^2-Test dann, wenn entweder

$$\begin{aligned} &n\pi_i \geq 3 \text{ für alle } i \in \mathcal{I} \qquad \text{oder} \\ &r \geq 10 \quad \text{und} \quad n\pi_i \geq 1 \text{ für alle } i \in \mathcal{I} \end{aligned} \tag{13.4}$$

gilt.

Beweisskizze von Theorem 13.14. Wir werden nur einige heuristische Bemerkungen anstelle eines kompletten Beweises des Theorems machen. Zunächst bemerken wir, dass die Teststatistik genau dann klein ist, wenn alle S_i nahe an $n\pi_i$ sind. Unter H_0 ist schließlich auch $\mathbf{E}[S_i] = n\pi_i$. Also misst die Teststatistik quadratische Abweichungen von $(S_i)_{i\in\mathcal{I}}$ von den erwarteten Anzahlen $(n\pi_i)_{i\in\mathcal{I}}$. Sind diese quadratischen Abweichungen zu groß, wird H_0 abgelehnt, da ja $(\chi^2_{r-1,1-\alpha}, \infty)$ der Ablehnungsbereich ist.

Wir zeigen nun noch, dass im speziellen Fall $r = 2$ die Teststatistik χ^2_n approximativ $\chi^2(1)$ verteilt ist. Sei also $\mathcal{I} = \{1, 2\}$. Dann ist S_1 nach $B(n, \pi_1)$ verteilt, $S_2 - n\pi_2 = (n - S_1 - n(1 -$

$\pi_1) = -(S_1 - n\pi_1)$, $\frac{1}{\pi_1} + \frac{1}{\pi_2} = \frac{1}{\pi_1\pi_2} = \frac{1}{\pi_1(1-\pi_1)}$ und, wegen dem zentralen Grenzwertsatz, für jedes $x > 0$, für eine nach $N(0,1)$ verteilte Zufallsvariable Z

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}_{\underline{\pi}}\left(\frac{(S_1 - n\pi_1)^2}{n\pi_1} + \frac{(S_2 - n\pi_2)^2}{n\pi_2} \le x\right) &= \lim_{n\to\infty} \mathbf{P}_{\underline{\pi}}\left(\frac{(S_1 - n\pi_1)^2}{n\pi_1(1-\pi_1)} \le x\right) \\ &= \lim_{n\to\infty} \mathbf{P}_{\underline{\pi}}\left(-\sqrt{x} \le \frac{S_1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} \le \sqrt{x}\right) \\ &= \mathbf{P}(-\sqrt{x} \le Z \le \sqrt{x}) = \mathbf{P}(Z^2 \le x).\end{aligned}$$

Da Z^2 nach $\chi^2(1)$ verteilt ist, haben wir gezeigt, dass approximativ χ^2_n für große n nach $\chi^2(1)$ verteilt ist. □

Beispiel 13.16 (Mendel's Experiment)**.** Wir führen nun den χ^2-Anpassungstest für das Mendel'sche Experiment aus Beispiel 13.13 mit $\alpha = 0.05$ durch. Hier ist $n = 400$,

$$H_0 : \underline{p} = \underline{\pi} := \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}\right).$$

Außerdem liefert die Stichprobe $S_{\text{weiß}} = 115$, $S_{\text{rosa}} = 171$ und $S_{\text{rot}} = 114$. Die erste Bedingung aus (13.4) ist sicher erfüllt, so dass wir den χ^2-Test anwenden können. Wir berechnen

$$\chi^2_{400} = \frac{(115-100)^2}{100} + \frac{(171-200)^2}{200} + \frac{(114-100)^2}{100} \approx 8.46 > 5.99 = \chi^2_{2,0.95},$$

H_0 wird also abgelehnt. (Nun darf man nach den biologischen Ursachen für die Abweichung von den Mendel'schen Verhältnissen forschen: z.B. könnte Selektion gegen Heterozygote im Spiel sein, oder Wechselwirkung des Gens für die Blütenfarbe mit anderen Teilen des Genoms.)

Bemerkung 13.17 (Erweiterung des $\boldsymbol{\chi^2}$-Anpassungstests für unbekannte Parameter)**.** Nicht immer ist die Verteilung, nach der die Daten verteilt sein sollen, so klar wie im Beispiel 13.13. Oftmals will man wissen, ob die Daten einer Verteilungsklasse (etwa die der Poisson-Verteilungen mit Parameter $\lambda > 0$) angehören, die noch einen oder mehr Parameter besitzt. In diesem Fall muss man zunächst die Parameter aus den Daten schätzen und kann erst anschließend den χ^2-Anpassungstest durchführen. In einer solchen Situation ist klar, dass schon durch das Schätzen der Parameter eine Anpassung der Verteilung an die Daten vollzogen wird. Den χ^2-Anpassungstest kann man jedoch nachwievor durchführen, indem man die Anzahl der Freiheitsgrade der χ^2-Verteilung reduziert, falls die verwendeten Schätzer Maximum-Likelihood-Schätzer sind. Man geht also folgendermaßen vor:

1. Schätzung der l fehlenden Parameter der Verteilung mittels Maximum-Likelihood, basierend auf $X_1, \ldots, X_n$.
2. Die Teststatistik χ^2 aus Theorem 13.14 ist dann approximativ (unter den Annahmen (13.4)) nach
   $\chi^2(r - l - 1)$ verteilt.

Also ist dann $(\chi^2_n, (\chi^2_{r-l-1,1-\alpha}, \infty))$ für große n ein Test zum Niveau α.

Beispiel 13.18 (Hufschlagtote)**.** Wir betrachten das klassische Beispiel von Hufschlagtoten der preussischen Armee. Hierbei wurden 14 Regimenter über 20 Jahre beobachtet, und für jedes Regiment und jedes Jahr die Zahl der Hufschlagtoten aufgezeichnet. Folgende Häufigkeiten wurden dabei beobachtet:

Anzahl der Todesfälle	0	1	2	3	4
Häufigkeit	144	91	32	11	2

Es liegt nahe zu vermuten, dass die Anzahl der Hufschlagtoten in jedem Jahr eine Poisson-verteilte Zufallsvariable ist. Um dies zu überprüfen, testen wir mit $\alpha = 0.01$ mittels eines χ^2-Anpassungstests.

Zunächst berechnen wir den Maximum-Likelihood-Schätzer für λ. Dieser ist nach Beispiel 11.13 gegeben durch

$$\widehat{\lambda} = \frac{1}{280}\big(0 \cdot 144 + 1 \cdot 91 + 2 \cdot 32 + 3 \cdot 11 + 4 \cdot 2\big) = 0.7.$$

Damit ergeben sich folgende erwartete Größen:

Anzahl der Todesfälle	0	1	2	3	4
Häufigkeit	139.04	97.33	34.07	7.95	1.39

Da wir fordern, dass alle erwarteten Anzahlen mindestens 5 sind, müssen wir die Fälle mit vielen Hufschlagtoten gruppieren:

Anzahl der Todesfälle	0	1	2	3 oder mehr
Häufigkeit	139.04	97.33	34.07	9.56

Wir setzen also $\mathcal{I} = \{0, 1, 2, 3 \text{ oder mehr}\}$,

$$H_0 : \underline{p} \in \Big\{e^{-\lambda}\Big(\frac{\lambda^0}{0!}, \frac{\lambda^1}{1!}, \frac{\lambda^2}{2!}, \dots\Big) \text{ für ein } \lambda > 0\Big\}.$$

Damit ergibt sich die Teststatistik

$$\begin{aligned}\chi^2_{280} &= \frac{(144 - 139.04)^2}{139.04} + \frac{(91 - 97.33)^2}{97.33} + \frac{(32 - 34.07)^2}{34.07} + \frac{(11 - 7.95)^2}{7.95} + \frac{(2 - 1.61)^2}{1.61} \\ &= 1.95 \leq 9.21 = \chi^2_{2,0.99}.\end{aligned}$$

Damit können wir die Nullhypothese Poisson-verteilter Daten auf dem Signifikanzniveau $\alpha = 0.01$ nicht ablehnen.

13.4. Der Kolmogorov-Smirnov-Test

Sowohl beim Gauß-Test, als auch beim t- und, F-Test haben wir die Annahme gemacht, dass die Daten normalverteilt sind. Diese Annahme lässt sich auch testen. Ein Verfahren hierzu werden wir nun besprechen.

Eine einfache grafische Möglichkeit, sich einen Eindruck zu verschaffen, ob ein Datensatz von reellwertigen Beobachtungen einer bestimmten Verteilung folgt, sind Plots der Quantile

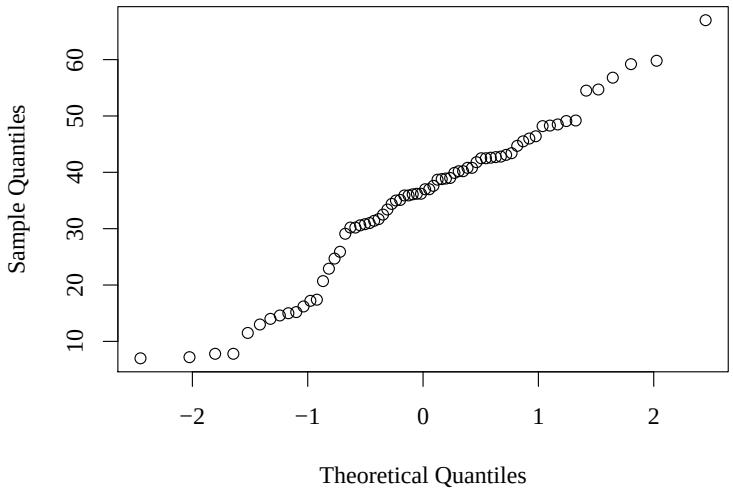


Abbildung 13.1.: QQ-Plot der `precip`-Daten.

oder QQ-Plots. Hier werden die Quantile der empirischen Verteilung gegen Quantile der zu überprüfenden Verteilung geplottet. Etwa ist das 5%-Quantil der empirischen Verteilung der (oder ein) $y \in \mathbb{R}$, so dass unterhalb von y genau 5% aller Datenpunkte zu finden sind. Für den (in R verfügbaren) Datensatz `precip`, der die Niederschlagsmenge (in Zoll) für 70 Städte der USA (und Puerto Rico) angibt, sieht man einen solchen Plot in Figur 13.1.

Natürlich ist es gut, nicht nur einen grafischen Eindruck der möglichen Abweichung der Normalverteilungsannahme zu haben, sondern auch einen statistischen Test. Mit dem hier vorgestellten Kolmogorov-Smirnov-Test kann man testen, ob Daten einer beliebigen, vorgegebenen, stetigen Verteilung folgen. Er basiert auf der empirischen Verteilung der Stichprobe, die wir in Definition 11.4 kennengelernt haben. Der Kolmogorov-Smirnov-Test basiert auf der Tatsache, dass die Verteilungsfunktion der empirischen Verteilung uniform gegen die Verteilungsfunktion der Daten konvergiert. Dieses Resultat formulieren wir zunächst.

Theorem 13.19 (Satz von Glivenko-Cantelli)**.** *Sind $X, X_1, X_2, ...$ unabhängige und identisch verteilte, reellwertige Zufallsgrößen mit Verteilungsfunktion F_X. Dann gilt für die empirische Verteilungsfunktion*

$$S_n(t) := \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq t}$$

und für alle $\varepsilon > 0$

$$\mathbf{P}\big(\sup_{t \in \mathbb{R}} |S_n(t) - F_X(t)| > \varepsilon \big) \xrightarrow{n \to \infty} 0.$$

Beweis. Wir geben nur eine Beweisskizze, die die punktweise, aber nicht die gleichmäßige Konvergenz zeigt. Es sei bemerkt, dass $1_{X_1 \leq t}, 1_{X_2 \leq t}, ...$ unabhängig und identisch verteilt sind mit $\mathbf{E}[1_{X_1 \leq t}] = \mathbf{P}(X_1 \leq t) = F_X(t)$. Damit gilt nach dem Gesetz der großen Zahlen

$$\mathbf{P}\big(|S_n(t) - F_X(t)| > \varepsilon \big) \xrightarrow{n \to \infty} 0$$

für jedes feste $t \in \mathbb{R}$. □

Bemerkung 13.20 (Verteilung von $F_X(X_{(i)})$)**.** Sei $X, X_1, ..., X_n$ unabhängige und identisch verteilte, reellwertige Zufallsvariable mit Dichte und Verteilungsfunktion F_X.

1. Es ist $F_X(X) \sim U[0,1]$.
   Denn: Fast sicher ist X so, dass $F_X^{-1}(X)$ existiert. Daraus folgt
   $$\mathbf{P}(F_X(X) \leq t) = \mathbf{P}(X \leq F_X^{-1}(t)) = F_X(F_X^{-1}(t) = t.$$

2. Seien $U_1, ..., U_n \sim U([0,1])$ unabhängig. Weiter sei $(X_{(1)}, ..., X_{(n)})$ der der Größe nach geordnete Vektor X (sowie $(U_{(1)}, ..., U_{(n)})$ der geordnete Vektor der $U_1, ..., U_n$). Dann gilt $F_X(X_{(i)}) \sim U_{(i)}$.
   Denn: Genau wie oben ist $X_{(i)}$ fast sicher so, dass $F_X^{-1}(X_{(i)})$ existiert. Nun ist
   $$\begin{aligned}\mathbf{P}(F_X(X_{(i)}) \leq t) &= \mathbf{P}(X_{(i)} \leq F_X^{-1}(t)) = \mathbf{P}(X_j \leq F_X^{-1}(t) \text{ für } i \text{ verschiedene } j) \\ &= \mathbf{P}(U_j \leq t \text{ für } i \text{ verschiedene } j) = \mathbf{P}(U_{(i)} \leq t).\end{aligned}$$

Proposition 13.21 (Verteilungsfreiheit von D_n)**.** *Sei $(X = (X_1, ..., X_n), (\mathbf{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell, wobei $X_1, ..., X_n$ unabhängig und identisch verteilt ist und unter jedem $\mathbf{P}_\vartheta$ eine Dichte sowie eine Verteilungsfunktion F_ϑ besitzt. Dann ist für jedes $t \in \mathbb{R}$ und $\vartheta \in \Theta$ die Statistik*
$$D_n(t) := \sup_{t \in \mathbb{R}} |S_n(t) - F_\vartheta(t)|$$
verteilungsfrei, d.h. ihre Verteilung hängt nicht von ϑ ab.

Beweis. Seien $X_{(1)}, ..., X_{(n)}$ die Ordnungsstatistiken von $X_1, ..., X_n$ sowie $X_{(0)} := -\infty$ und $X_{(n+1)} := \infty$. Dann ist
$$S_n(t) = \frac{i}{n} \text{ für } X_{(i)} \leq t < X_{(i+1)}.$$
Wir schreiben nun
$$\begin{aligned}D_n &= \sup_{t \in \mathbb{R}} |S_n(t) - F_\vartheta(t)| = \max_{1 \leq i \leq n} \sup_{X_{(i)} \leq t < X_{(i+1)}} |S_n(t) - F_\vartheta(t)| \\ &= \max_{1 \leq i \leq n} \sup_{X_{(i)} \leq t < X_{(i+1)}} \Big|\frac{i}{n} - F_\vartheta(t)\Big| \\ &= \max_{1 \leq i \leq n} \max\Big(\Big|\frac{i}{n} - F_\vartheta(X_{(i)})\Big|, \Big|\frac{i}{n} - F_\vartheta(X_{(i+1)})\Big|\Big).\end{aligned}$$
Damit ist gezeigt, dass D_n nur von $F_\vartheta(X_{(0)}), ..., F_\vartheta(X_{(n+1)})$ abhängt. Diese Größen haben nach Bemerkung 13.20 dieselbe Verteilung wie die Ordnungsstatistiken eines $U(0,1)$-verteilten Vektors von Zufallsvariablen, und zwar unabhängig von F_ϑ. Daraus folgt die Behauptung. □

Theorem 13.22 (Verteilung von D_n)**.** *Sei $X, X_1, .., X_n$ unabhängig und identisch verteilt mit Dichte sowie F_X die Verteilungsfunktion von X. Dann gilt für $0 < s < (2n-1)/(2n)$*
$$\mathbf{P}\Big(D_n < \frac{1}{2n} + s\Big) = n! \int_{1/(2n)-s}^{1/(2n)+s} \int_{3/(2n)-s}^{3/(2n)+s} \cdots \int_{(2n-1)/(2n)-s}^{(2n-1)/(2n)+s} 1_{0<u_1 \cdots < u_n < 1} du_n \cdots du_1.$$

Beweis. Zunächst bemerken wir, dass immer $D_n \geq 1/2n$ gilt, da F_X stetig ist, S_n aber Sprünge der Größe $1/n$ macht. OBdA nehmen wir wegen der Verteilungsfreiheit von D_n an, dass $F_X(x) = x$, d.h. $X \sim U([0,1])$. Wir schreiben mit $s' := \frac{1}{2n} + s$

$$\begin{aligned}
\mathbf{P}(D_n < s') &= \mathbf{P}\Big(\sup_{t\in[0,1]} |S_n(t) - t| < s'\Big) \\
&= \mathbf{P}\Big(\Big|\frac{i}{n} - t\Big| < s' \text{ für alle } X_{(i)} \leq t < X_{(i+1)}, \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbf{P}\Big(\frac{i}{n} - s' < t < \frac{i}{n} + s' \text{ für alle } X_{(i)} \leq t < X_{(i+1)}, \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbf{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i}{n} + s', \frac{i}{n} - s' < X_{(i+1)} < \frac{i}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbf{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i}{n} + s', \frac{i-1}{n} - s' < X_{(i)} < \frac{i-1}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbf{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i-1}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbf{P}\Big(\frac{2i-1}{2n} - s < X_{(i)} < \frac{2i-1}{2n} + s \text{ für alle } i = 1, ..., n\Big).
\end{aligned}$$

Daraus folgt die Behauptung, da die gemeinsame Verteilung von $X_{(1)}, ..., X_{(n)}$ die Dichte $n! 1_{0\leq u_1<\cdots<u_n}$ hat. □

Beispiel 13.23 (Der Kolmogorov-Smirnov-Test für t-verteilte Daten)**.** Wir wollen die Daten aus dem `precip`-Datensatz auf Normalität testen. Da wir weder Erwartungswert noch Varianz der Normalverteilung, auf die wir testen wollen, wissen, standardisieren wir die Daten und wenden den Kolmogorov-Smirnov-Test an. Das Ergebnis ist wiefolgt:

```
> ks.test((precip - mean(precip))/sd(precip), "pnorm")

One-sample Kolmogorov-Smirnov test

data:  (precip - mean(precip))/sd(precip)
D = 0.10909, p-value = 0.3755
alternative hypothesis: two-sided

Warnmeldung:
In ks.test((precip - mean(precip))/sd(precip), "pnorm") :
  für den Komogorov-Smirnov-Test sollten keine Bindungen vorhanden sein
```

Die von R erwähnte *Bindung* bedeutet, dass es zwei (oder mehr) identische Werte im Datensatz gibt, was für eine stetige Verteilung nicht möglich ist. Mit einem p-Wert von 37.55% kann jedenfalls mit diesem Test die Normalität nicht ausgeschlossen werden. Allerdings muss auch gesagt werden, dass wir den Test nicht korrekt angewendet haben, weil wir zwei Parameter der Verteilung schätzen mussten.