Skript zur Vorlesung

Wahrscheinlichkeitsrechnung und Statistik

ETH Zürich, D-Math

Hans Föllmer
Humboldt Universität Berlin

Hansruedi Künsch
ETH Zürich

mit Ergänzungen von Josef Teichmann
ETH Zürich

Version Februar 2013

Inhaltsverzeichnis

Kapitel 1

Der Wahrscheinlichkeitsbegriff

Es ist erstaunlich, dass sich auch der Zufall, der ja per Definition nicht vorhersagbar ist, mathematisch beschreiben lässt. Sobald der Zufall ins Spiel kommt, ist der genaue Ausgang eines Experiments nicht mehr im Voraus bekannt. Wir nehmen aber an, dass wir alle möglichen Ausgänge beschreiben können:

Der **Grundraum Ω** ist die Menge aller möglichen Ergebnisse oder Fälle ω ("Elementarereignisse"). Das Experiment besteht dann darin, dass eines der möglichen Elementarereignisse "realisiert", d.h. vom Zufall ausgewählt wird.
Beispiel 1.1. *Wurf eines Würfels:* $\Omega = \{1, 2, 3, 4, 5, 6\}$.
Beispiel 1.2. *Unendlich viele Würfe einer Münze:* $\Omega = \{0, 1\}^{\mathbb{N}}$ = *Menge aller binären Folgen.*
Beispiel 1.3. *Bewegung eines mikroskopischen Teilchens in einer Flüssigkeit:* $\Omega = C(\mathbb{R}_+, \mathbb{R}^3)$ = *Menge aller stetigen Funktionen von* $\mathbb{R}_+$ *nach* $\mathbb{R}^3$.

Als nächstes führen wir die Klasse $\mathcal{A}$ der **beobachtbaren Ereignisse** ein. Dies ist eine Menge von Teilmengen von Ω. Für ein $A \in \mathcal{A}$ sagen wir, dass A eintritt, falls das realisierte Ergebnis ω Element von A ist.
Beispiel 1.4. *"Das Resultat ist eine gerade Zahl"*$\leftrightarrow A = \{2, 4, 6\}$.
Beispiel 1.5. *"Der dritte Wurf ist Kopf"*$\leftrightarrow A = \{(\omega_i) \in \Omega; \quad \omega_3 = 1\}$.
Beispiel 1.6. *"Das Teilchen bleibt stets in einer Kugel vom Radius* r *um* $\mathbf{0}$*"*$\leftrightarrow A = \{\mathbf{x}(\cdot) \in \Omega; \quad \sup_t ||\mathbf{x}(t)|| \leq r\}$.

Mit Hilfe der Operationen der Mengenlehre kann man aus gegebenen Ereignissen neue bilden.

Falls Ω endlich oder abzählbar unendlich ist (diskreter Fall), wählen wir meistens als $\mathcal{A}$ die Potenzmenge von Ω. Im allgemeinen kann $\mathcal{A}$ aber nicht aus allen Teilmengen von Ω, bestehen, vgl. Kap. 3, und manchmal will man auch nur gewisse Teilmengen als beobachtbar bezeichnen, vgl. Abschnitt 2.3.4 über Spielsysteme.

Das entscheidende Konzept für die mathematische Behandlung von Zufallsexperimenten ist schliesslich das **Wahrscheinlichkeitsmass** $\mathbb{P}$, eine Abbildung von $\mathcal{A}$ nach $[0, 1]$. $\mathbb{P}(A)$ ist die Wahrscheinlichkeit, dass A eintreten wird. Je grösser diese Wahrscheinlichkeit, desto eher rechnen wir damit, dass A eintritt.

Das Tripel $(\Omega, \mathcal{A}, \mathbb{P})$ heisst ein **Wahrscheinlichkeitsraum**.

Der für die Anwendungen heikelste Punkt ist die Interpretation, bzw. die Festlegung von Wahrscheinlichkeiten. Heuristisch ist die Wahrscheinlichkeit ein Mass für die Ungewissheit

über das Eintreten eines Ereignisses, ausgedrückt als Teil der Gewissheit. Es gibt 3 Ansätze, dies präziser zu definieren, die aber nur zum Teil kompatibel sind:

i) **subjektiv**: Mass des persönlichen Glaubens an das Eintreten von A. Dieses wird bestimmt durch das Wettverhältnis "Einsatz : Gewinn" $= \mathbb{P}(A) : \mathbb{P}(A^c)$, das der Person fair erscheint.

ii) **frequentistisch**: Grenzwert der relativen Häufigkeit des Eintretens von A bei unabhängigen Wiederholungen.

iii) **Gleichverteilung**: $\frac{\text{Anzahl günstige Fälle}}{\text{Anzahl mögliche Fälle}}$ (vgl. den Abschnitt 2.2 über Laplace-Modelle).

Den Zusammenhang zwischen i) und ii) sieht man wie folgt ein: Wenn man die Wette n-mal durchführt mit Einsatz 1 Franken und A dabei n_A-mal auftritt, dann hat man offensichtlich $n - n_A$ Franken verloren und $n_A \mathbb{P}(A^c) / \mathbb{P}(A)$ Franken gewonnen. Wenn also $n_A = n\mathbb{P}(A)$, dann sind Gewinn und Verlust gleich.

Alle diese Ansätze führen zu Schwierigkeiten. Als Ausweg hat Kolmogorov 1933 den axiomatischen Zugang vorgeschlagen: Man kümmert sich nicht darum, was $\mathbb{P}(A)$ bedeutet und wie man $\mathbb{P}$ für ein konkretes Zufallsexperiment erhält; man fordert einfach gewisse Regeln, die für $\mathbb{P}$ gelten sollen, und untersucht deren Konsequenzen. Bevor wir jedoch auf diese Axiome im allgemeinen Fall eingehen, behandeln wir den diskreten Fall ausführlicher.

In der Statistik werden wir uns der Frage zuwenden, wie man aufgrund von Beobachtungen Rückschlüsse auf die zugrunde liegende Wahrscheinlichkeit ziehen kann.

Die frequentistische Interpretation werden wir in dieser Vorlesung insbesondere bei Simulationen benutzen: Wir werden an verschiedenen Stellen viele unabhängige Wiederholungen eines Zufallsexperiments auf dem Computer durchführen und die relative Häufikgeit eines Ereignisses als Approximation der Wahrscheinlichkeit verwenden. Wir sehen dabei darüber hinweg, dass die Experimente auf dem Computer nicht wirklich zufällig sind, sondern dass ein deterministischer Algorithmus dahinter steht.

Die Simulationen und Grafiken in diesem Skript wurden mit der statistischen Software *R* erzeugt. Dies ist eine GNU-Software, die unentgeltlich heruntergeladen werden kann (www.stat.math.ethz.ch/CRAN) und auch auf den ETH-Computern installiert ist.

Kapitel 2

Diskrete Wahrscheinlichkeitsräume

2.1 Grundbegriffe

2.1.1 Die Wahrscheinlichkeit eines Ereignisses

Der Grundraum Ω sei endlich oder abzählbar unendlich. Wir stellen uns auf den axiomatischen Standpunkt und nehmen an, dass für jedes Ergebnis $\omega \in \Omega$ die Wahrscheinlichkeit gegeben ist, dass gerade dieses Ergebnis eintritt. Diese Wahrscheinlichkeit bezeichnen wir mit

$$p(\omega) := \mathbb{P}(\{\omega\}) \in [0,1]. \tag{2.1}$$

Wir setzen ferner voraus, dass diese Wahrscheinlichkeiten $p(\omega)$, die wir auch als Gewichte auffassen können, normiert sind:

$$\sum_{\omega\in\Omega} p(\omega) = 1. \tag{2.2}$$

Das Wahrscheinlichkeitsmass auf der Potenzmenge $\mathcal{A}$ von Ω wird dann festgelegt durch

$$\mathbb{P}(A) = \sum_{\omega\in A} p(\omega)\ ,\ A \in \mathcal{A}. \tag{2.3}$$

Dann gilt offensichtlich:

$$\mathbb{P}(\Omega) = 1, \tag{2.4}$$

und für paarweise disjunkte Ereignisse $A_1, A_2, \ldots$ (d.h. $A_i \cap A_j = \emptyset$ für $i \neq j$) ist

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i), \tag{2.5}$$

weil

$$\sum_{\omega\in\cup_i A_i} p(\omega) = \sum_i \sum_{\omega\in A_i} p(\omega).$$

Bemerkungen

- In der allgemeinen Theorie der Wahrscheinlichkeitsräume werden die Eigenschaften (2.4) und (2.5) als **Axiome** benutzt.

- Auf einer abzählbaren Menge Ω ist jede Abbildung $\mathbb{P} : \mathcal{A} \to [0,1]$, die (2.4) und (2.5) erfüllt, von der Form (2.3) , und zwar mit

$$p(\omega) \equiv \mathbb{P}(\{\omega\}). \tag{2.6}$$

Weitere Rechenregeln: Man sieht sofort, dass die folgenden Regeln gelten

$$\begin{aligned}
\mathbb{P}(A^c) &= 1 - \mathbb{P}(A) && (2.7)\\
\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), && (2.8)\\
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) && (2.9)\\
A \subseteq B &\Rightarrow \mathbb{P}(A) \le \mathbb{P}(B). && (2.10)
\end{aligned}$$

Beispiel 2.1. *Wir betrachten das "Experiment" Anzahl Anrufe während einer festen Zeit bei einer Telefonzentrale. Das bedeutet, dass* $\Omega = \{0, 1, 2, \dots\}$. *Im Abschnitt 2.6.3 werden wir sehen, dass die folgenden Gewichte*

$$p(\omega) = e^{-\lambda} \frac{\lambda^\omega}{\omega!} \qquad (\omega \in \Omega = \{0, 1, 2, \dots\}), \tag{2.11}$$

wobei $\lambda > 0$ *ein Parameter ist, ein vernünftiges Modell ergeben. Dieses Modell heisst die Poisson-Verteilung. Wenn wir das Ereignis* $A =$ *"mindestens 1 Anruf"* $= \{1, 2, \dots\}$, *betrachten, dann ist*

$$\mathbb{P}(A) \overset{(2.7)}{=} 1 - \mathbb{P}(A^c) = 1 - \mathbb{P}(\{0\}) = 1 - e^{-\lambda}.$$

2.1.2 Der Erwartungswert einer Zufallsvariablen

Eine reellwertige Funktion auf Ω heisst auch **Zufallsvariable**. Je nachdem, welches Elementarereignis ω realisiert wird, ändert sich auch der realisierte Wert $X(\omega)$. Für eine Zufallsvariable

$$X : \Omega \to \mathbb{R}^1 \tag{2.12}$$

ist der Wertebereich $X(\Omega)$ auch wieder abzählbar, und durch die Gewichtung

$$x \in X(\Omega) \to \mathbb{P}(X = x) \equiv \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\}) \tag{2.13}$$

ist ein Wahrscheinlichkeitsmass auf $X(\Omega)$ gegeben, die sogenannte **Verteilung der Zufallsvariablen X**. Manchmal ist es bequem, auch die Werte $\pm\infty$ für X zuzulassen.

Jeder Zufallsvariablen X ordnen wir den **Erwartungswert**

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) p(\omega) \tag{2.14}$$

zu, wobei wir sicherstellen müssen, dass die rechte Seite sinnvoll ist. Dies ist z.B. der Fall, wenn $X \ge 0$ (den Wert $+\infty$ lassen wir durchaus zu). Wenn X positive und negative Werte annimmt, benutzen wir die Zerlegung $X = X^+ - X^-$ von X in den Positivteil $X^+ = \max(X, 0)$ und den Negativteil $X^- = (-X)^+$ und setzen

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-) = \sum_{X(\omega) > 0} X(\omega) p(\omega) - \sum_{X(\omega) < 0} (-X(\omega)) p(\omega), \tag{2.15}$$

sofern nicht beide Summen rechts unendlich sind, d.h. $\min(\mathbb{E}(X^+), \mathbb{E}(X^-)) < \infty$. Der Erwartungswert (2.14) lässt sich auch *mit Hilfe der Verteilung* von X ausdrücken:

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x \in X(\Omega)} \sum_{\omega : X(\omega) = x} X(\omega) p(\omega) \\ &= \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(X = x). \end{aligned} \tag{2.16}$$

Beispiel 2.2. *Im Beispiel (2.1) sei X die "Anzahl der Anrufe", also $X(\omega) = \omega$. Dann ist*

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k \mathbb{P}(X = k) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda, \tag{2.17}$$

d.h. der Parameter λ ist gerade die erwartete *Anzahl Anrufe.*

Beispiel 2.3. *In einem einfachen Versicherungskontrakt ist die Leistung der Versicherung*

$$X = \begin{cases} c & \text{falls das Ereignis } A \text{ eintritt,} \\ 0 & \text{sonst,} \end{cases}$$

also zufallsabhängig. Die (deterministische!) Gegenleistung des Versicherungsnehmers ist die Prämie, und ein erster Ansatz für eine faire Prämie ist gerade der Erwartungswert

$$\mathbb{E}(X) = c \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = c \cdot \mathbb{P}(A).$$

Aus der Definition (2.14) ergibt sich sofort die **Linearität des Erwartungswertes:**

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) \quad (a, b \in \mathbb{R}^1) \tag{2.18}$$

(sofern beide Seiten sinnvoll sind).

Zur Berechnung des Erwartungswerts ist folgendes Lemma oft nützlich

Lemma 2.1. *Wenn X nur die Werte $0, 1, 2, \ldots$ annimmt, so ist*

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} \mathbb{P}(X > n). \tag{2.19}$$

Beweis

$$\begin{aligned} \sum_{n=0}^{\infty} \mathbb{P}(X > n) &\overset{(2.5)}{=} \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} \mathbb{P}(X = k) = \sum_{k=1}^{\infty} \sum_{n=0}^{k-1} \mathbb{P}(X = k) \\ &= \sum_{k=1}^{\infty} k \mathbb{P}(X = k) \overset{(2.16)}{=} \mathbb{E}(X). \end{aligned}$$

□

2.2 Laplace-Modelle

Sei Ω *endlich.* In vielen Situationen ist es sinnvoll, den **Laplace-Ansatz** $p(\omega) = \text{const}$. zu machen (Indifferenzprinzip, Prinzip vom unzureichenden Grunde, Symmetrie eines Würfels etc.). Wegen (2.2) folgt

$$p(\omega) = \frac{1}{|\Omega|}, \tag{2.20}$$

und aus (2.3) ergibt sich

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl der möglichen Fälle}} \qquad (A \subseteq \Omega) \tag{2.21}$$

für jedes Ereignis $A \subseteq \Omega$. Man nennt $\mathbb{P}$ auch die **Gleichverteilung** auf Ω.

Beispiel 2.4. Garderobenproblem (Montmort, 1708):

n Mäntel werden zufällig an n Personen verteilt. Was ist die Wahrscheinlichkeit, dass keine Person ihren eigenen Mantel bekommt? Als Modell wählen wir hier die Menge aller Permutationen Ω von $\{1,\dots,n\}$ und die Gleichverteilung $\mathbb{P}$ auf Ω. Für $i = 1,\dots,n$ sei $A_i = \{\omega \in \Omega \mid \omega(i) = i\}$ das Ereignis, dass die i-te Person ihren eigenen Mantel bekommt. Dann interessieren wir uns also für das Ereignis $A = (\cup_{i=1}^n A_i)^c$. Zunächst berechnen wir

$$\begin{aligned}\mathbb{P}(A^c) &= \mathbb{P}(\cup_{i=1}^n A_i) \overset{(2.9)}{=} \sum_{k=1}^n (-1)^{k+1} \sum_{1\le i_1<\dots<i_k\le n} \underbrace{\mathbb{P}(A_{i_1}\cap\dots\cap A_{i_k})}_{=\frac{(n-k)!}{n!}} \\ &= \sum_{k=1}^n (-1)^{k+1}\binom{n}{k}\frac{(n-k)!}{n!} = -\sum_{k=1}^n \frac{(-1)^k}{k!}. \end{aligned} \tag{2.22}$$

Daraus folgt

$$\mathbb{P}(A) = 1 + \sum_{k=1}^n \frac{(-1)^k}{k!} \longrightarrow e^{-1} \quad (n\to\infty). \tag{2.23}$$

Beispiel 2.5. Zufallsauswahl *(→ Meinungsumfragen etc.)*

In einer Urne befinden sich N durchnummerierte Kugeln, K rote und $N-K$ weisse. Es werde eine Stichprobe von n Kugeln mit, bzw. ohne Zurücklegen gezogen. Wenn ω_i die Nummer der beim i-ten Mal gezogenen Kugel bezeichnet, dann ist der Grundraum

$$\begin{aligned} &\textit{Mit Zurücklegen:} && \Omega_1 = \{(\omega_1,\dots,\omega_n) \mid 1 \le \omega_i \le N\}, \\ &\textit{Ohne Zurücklegen:} && \Omega_2 = \{(\omega_1,\dots,\omega_n) \mid 1 \le \omega_i \le N, \omega_i \ne \omega_j\}. \end{aligned}$$

Wir nehmen als $\mathbb{P}_i$ die Gleichverteilung auf Ω_i $(i = 1,2)$ und berechnen die Verteilung der Zufallsvariablen $X =$ Anzahl roter Kugeln in der Stichprobe. Sei $A_i = \{\omega \in \Omega_i \mid 1 \le \omega_j \le K$ für genau k Indizes $j\}$. Dann gilt $\mathbb{P}_i(X = k) = |A_i|/|\Omega_i|$.

Mit Zurücklegen ist $|\Omega_1| = N^n$ und $|A_1| = K^k(N-K)^{n-k}\binom{n}{k}$, also folgt

$$\mathbb{P}_1(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Dies ist die sogenannte **Binomialverteilung** *mit Parameter $p = K/N =$ Anteil roter Kugeln.*

Ohne Zurücklegen ist

$$\begin{aligned} |\Omega_2| &= N(N-1)\dots(N-n+1) = \binom{N}{n} n!, \\ |A_2| &= K(K-1)\dots(K-k+1)(N-K)(N-K-1)\dots(N-K-(n-k)+1)\binom{n}{k} \\ &= \binom{K}{k}\binom{N-K}{n-k} n!. \end{aligned}$$

Daraus folgt

$$\mathbb{P}_2(X=k) \;=\; \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad (k=0,1,\dots \min(n,K)).$$

Dies ist die sogenannte **hypergeometrische Verteilung**.

Man sieht leicht, dass für $N\to\infty,\ K\to\infty,\ K/N\to p$ die hypergeometrische Verteilung gegen die Binomialverteilung konvergiert, dh. die von N abhängigen Wahrscheinlichkeiten $\mathbb{P}_2(X=k)$ konvergieren gegen die Werte $\mathbb{P}_1(X=k)$.

Als Grundraum haben wir hier die Menge aller Ziehungen mit Berücksichtigung der Reihenfolge genommen. Wenn man an der Reihenfolge, in der die Kugeln gezogen werden, nicht interessiert ist, kann man auch die Grundräume

$$\Omega_3 = \{(\omega_1,\dots,\omega_n) \mid 1\le\omega_1\le\omega_2\le\ldots\le\omega_n\le N\},$$

bzw. beim Ziehen ohne Zurücklegen

$$\Omega_4 = \{(\omega_1,\dots,\omega_n) \mid 1\le\omega_1<\omega_2<\ldots<\omega_n\le N\},$$

betrachten. Ob man die Gleichverteilung auf Ω_2 oder Ω_4 wählt, spielt keine Rolle. Man erhält die gleichen Wahrscheinlichkeiten für Ereignisse, die sich nicht auf die Reihenfolge beziehen, denn bei der Reduktion von Ω_2 auf Ω_4 werden jeweils $n!$ Elementarereignisse zu einem neuen Elementarereignis zusammengefasst. Beim Ziehen mit Zurücklegen ist das nicht der Fall: Dem Element $(1,1,\dots,1)\in\Omega_3$ entspricht nur ein Element in Ω_1, dem Element $(1,2,\dots,n)\in\Omega_3$ entsprechen aber $n!$ verschiedene Elemente in Ω_1.

2.3 Irrfahrt

2.3.1 Definition

Die Irrfahrt (random walk, marche aléatoire) ist ein Modell für die zufällige Bewegung eines Teilchens auf dem eindimensionalen Gitter $\mathbb{Z}=\{\dots,-1,0,1,\dots\}$, das in 0 startet und in jeder Periode einen Schritt $+1$ oder -1 macht, jeweils mit Wahrscheinlichkeit $\frac{1}{2}$. Man kann das auch anders interpretieren, z.B. als die Bilanzentwicklung eines Spielers bei einem fairen Glücksspiel, als die Kursentwicklung einer Aktie, usw.

Modell für N Perioden:
Sei Ω die Menge aller binären Folgen der Länge N, also $\Omega=\{\omega=(x_1,\dots,x_N)\mid x_i\in\{+1,-1\}\}$. Ferner betrachten wir die Zufallsvariablen $X_k(\omega)=k$-te Komponente von $\omega=(x_1,\dots,x_N)\in\Omega$ und $S_n(\omega)=\sum_{k=1}^n X_k(\omega)$. X_k ist also der Schritt bzw. Ertrag in der k-ten Periode und S_n die Position bzw. Bilanz nach n Perioden. Wir starten stets im Ursprung, d.h. $S_0(\omega)=0$

Für jedes $\omega\in\Omega$ erhält man eine **Trajektorie** (Pfad, Bilanzentwicklung) $(n,S_n(\omega))$ $(n=0,\dots,N)$. Sei nun $\mathbb{P}$ die **Gleichverteilung auf** Ω, also

$$\mathbb{P}(A)=\frac{|A|}{|\Omega|}=2^{-N}\cdot|A| \qquad (A\subseteq\Omega). \tag{2.24}$$

Definition 2.1. *Die Folge der Zufallsvariablen $S_n(n=0,\dots,N)$ auf dem Wahrscheinlichkeitsraum $(\Omega,\mathbb{P})$ heisst* **Irrfahrt** *(mit Start in 0).*

Wir nehmen also an, dass die binären Folgen $\omega \in \Omega$ bzw. die entsprechenden Trajektorien alle gleichwahrscheinlich sind. Zwei Trajektorien einer Irrfahrt können wir leicht mit folgendem R-code erzeugen, siehe Abbildung 2.1.

```
> set.seed(456)
> X1 <- sample(c(-1,1),size=10000,replace=TRUE)
> X2 <- sample(c(-1,1),size=10000,replace=TRUE)
> Y1 <- cumsum(X1)
> Y2 <- cumsum(X2)
> ymin <- min(c(Y1,Y2))
> ymax <- max(c(Y1,Y2))
> plot(Y1,type="l",ylim=c(ymin,ymax),xlab="Periode",ylab="Position")
> lines(Y2,col="blue")
```

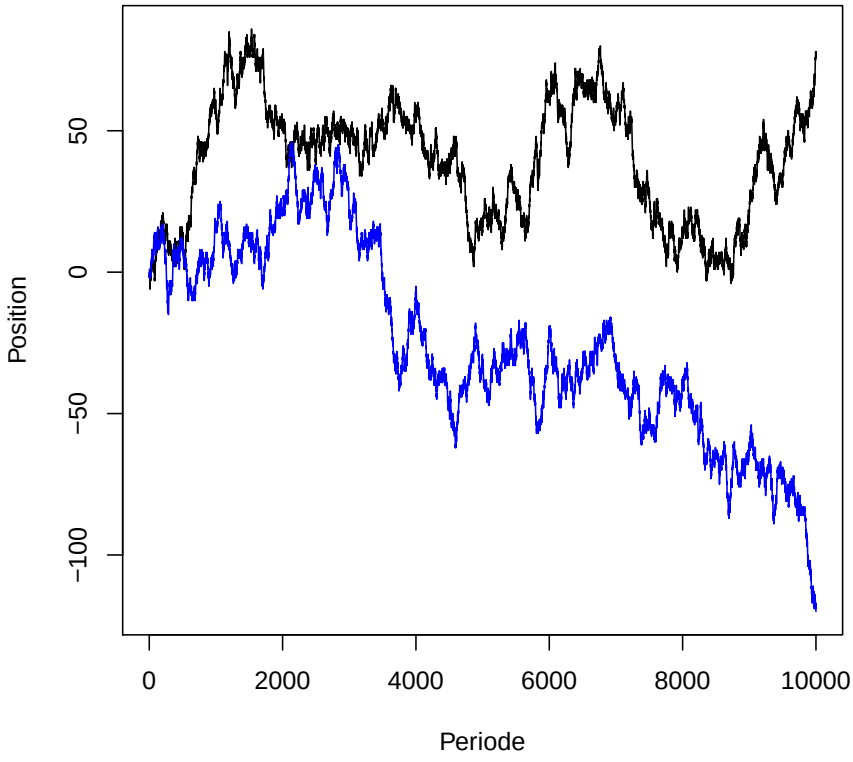

Abbildung 2.1: Zwei Trajektorien mit $N = 10000$ einer Irrfahrt

Aus (2.24) folgt

$$\mathbb{P}(X_k = +1) = \frac{2^{N-1}}{2^N} = \frac{1}{2} \quad (k = 1, \ldots, N). \tag{2.25}$$

Ebenso gilt für beliebige Indizes $1 \leq k_1 < \cdots < k_\ell \leq N$ und für jede Wahl von $x_k = \pm 1$

$$\mathbb{P}(X_{k_1} = x_{k_1}, \ldots, X_{k_\ell} = x_{k_\ell}) = \frac{2^{N-\ell}}{2^N} = 2^{-\ell}. \tag{2.26}$$

Insbesondere bilden die ersten $n < N$ Schritte einer Irrfahrt eine Irrfahrt mit n Perioden.

Aus (2.25) folgt

$$\mathbb{E}(X_k) = (+1) \cdot \mathbb{P}(X_k = +1) + (-1) \cdot \mathbb{P}(X_k = -1) = 0 \tag{2.27}$$

und daraus folgt, wegen der Linearität des Erwartungswerts

$$\mathbb{E}(S_n) = \sum_{k=1}^{n} \mathbb{E}(X_k) = 0 \qquad (n = 0, \ldots, N). \tag{2.28}$$

Satz 2.1. *Für festes n nimmt die Zufallsvariable S_n Werte $x \in \{-n, -n+2, \ldots, n-2, n\}$ an, und zwar mit den folgenden Wahrscheinlichkeiten:*

$$\mathbb{P}(S_n = 2k - n) = \binom{n}{k} 2^{-n} \quad (k = 0, 1, \ldots, n). \tag{2.29}$$

Für alle anderen x ist $\mathbb{P}(S_n = x) = 0$.

Beweis Sei U_n die Anzahl der "Schritte nach oben" bis zum Zeitpunkt n, d.h.

$$U_n = \sum_{k=1}^{n} 1_{\{X_k = +1\}}.$$

Dann ist $S_n = U_n - (n - U_n) = 2U_n - n$.

$$|\{U_n = k\}| = \binom{n}{k} 2^{N-n} \Rightarrow \mathbb{P}(U_n = k) = \binom{n}{k} 2^{N-n} 2^{-N} = \binom{n}{k} 2^{-n}.$$

□

Die Verteilung von S_n ist also eine "linear transformierte Binomialverteilung" mit $p = \frac{1}{2}$.

Einfache Folgerungen:

Die Formel von Stirling besagt

$$n! \sim (\frac{n}{e})^n \sqrt{2\pi n}$$

($a_n \sim b_n$ heisst $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$). Daraus folgt, dass

$$\mathbb{P}(S_{2n} = 0) = \mathbb{P}(S_{2n-1} = 1) = \binom{2n}{n} 2^{-2n} \sim \frac{1}{\sqrt{\pi n}}. \tag{2.30}$$

Ferner gilt

$$\binom{n}{k} = \binom{n}{k-1} \frac{n-k+1}{k} \geq \binom{n}{k-1} \iff k \leq \frac{n+1}{2} \tag{2.31}$$

Also ist für n gerade $\mathbb{P}(S_n = x)$ maximal für $x = 0$, und für n ungerade ist $\mathbb{P}(S_n = x)$ maximal für $x = \pm 1$.

2.3.2 Reflektionsprinzip

Wir untersuchen hier die Verteilung der Zufallsvariable

$$T_a(\omega) = \min\{n > 0 \mid S_n(\omega) = a\} \tag{2.32}$$

(Erstes Erreichen des Niveaus $a \neq 0$, bzw. für $a = 0$ erste Rückkehr nach 0). Dabei setzen wir hier $\min \emptyset = N + 1$.

Die entscheidende Idee ist das folgende Lemma:

Lemma 2.2 (Reflektionsprinzip)**.** *Für $a > 0$ und $b \geq -a$ ist*

$$\mathbb{P}(T_{-a} \leq n, S_n = b) = \mathbb{P}(S_n = -2a - b). \tag{2.33}$$

Beweis Durch Spiegelung des Pfades für Zeiten $\geq T_{-a}$ am Niveau $-a$ erhält man eine Bijektion zwischen Pfaden mit $T_{-a} \leq n, S_n = b$ und solchen mit $S_n = -2a - b$, vergleiche Abbildung 2.2. □

```
> z <- 0
> Yr <- rep(0,100)
> while (z > -7)
+ {
+   X <- X1 <- sample(c(-1,1),size=100,replace=TRUE)
+   Y <- cumsum(X)
+   z <- min(Y)
+ }
> Ta <- min(which(Y==-7))
> if (Y[100] < -7) {
+   Yr <- Y
+   Y <- c(Yr[1:Ta],-14-Yr[(Ta+1):100])
+ } else Yr <- c(Y[1:Ta],-14-Y[(Ta+1):100])
> plot(Y,type="o",ylim=range(c(Y,Yr)),xlab="Periode",ylab="Position")
> abline(h=-7,col="red")
> lines(Yr,col="blue")
```

Aus dem Reflektionsprinzip ergibt sich insbesondere die Verteilung von T_{-a}:
Satz 2.2. *Für $a \neq 0$ gilt*

$$\mathbb{P}(T_{-a} \leq n) = 2\mathbb{P}(S_n < -a) + \mathbb{P}(S_n = -a) = \mathbb{P}(S_n \notin (-a, a]). \tag{2.34}$$

Beweis Mit dem Additionssatz folgt

$$\begin{aligned} \mathbb{P}(T_{-a} \leq n) &= \sum_{b=-\infty}^{\infty} \mathbb{P}(T_{-a} \leq n, S_n = b) \\ &\overset{(2.33)}{=} \sum_{b=-\infty}^{-a} \mathbb{P}(S_n = b) + \sum_{b=-a+1}^{\infty} \mathbb{P}(S_n = -2a - b) \\ &= \mathbb{P}(S_n \leq -a) + \mathbb{P}(S_n \leq -a - 1). \end{aligned}$$

Die letzte Gleichung in (2.34) folgt aus Symmetrie. □
Korollar 2.1. *Für jedes $a \neq 0$ gilt*

$$1.\ \mathbb{P}(T_a > N) \overset{N\to\infty}{\longrightarrow} 0, \quad 2.\ \mathbb{E}(T_a) = \sum_{k=1}^{N+1} k\mathbb{P}(T_a = k) \overset{N\to\infty}{\longrightarrow} \infty.$$

Anschaulich bedeutet die erste Aussage, dass die Irrfahrt mit Wahrscheinlichkeit 1 jedes Niveau erreicht, und die zweite Aussage, dass man *sehr lange* darauf warten muss. Exakt lässt sich das jedoch erst in einem Modell mit unendlich vielen Perioden formulieren. Dort gilt dann

$$\mathbb{P}(T_a < \infty) = \lim_{N\to\infty} \mathbb{P}(T_a \leq N) = 1, \tag{2.35}$$

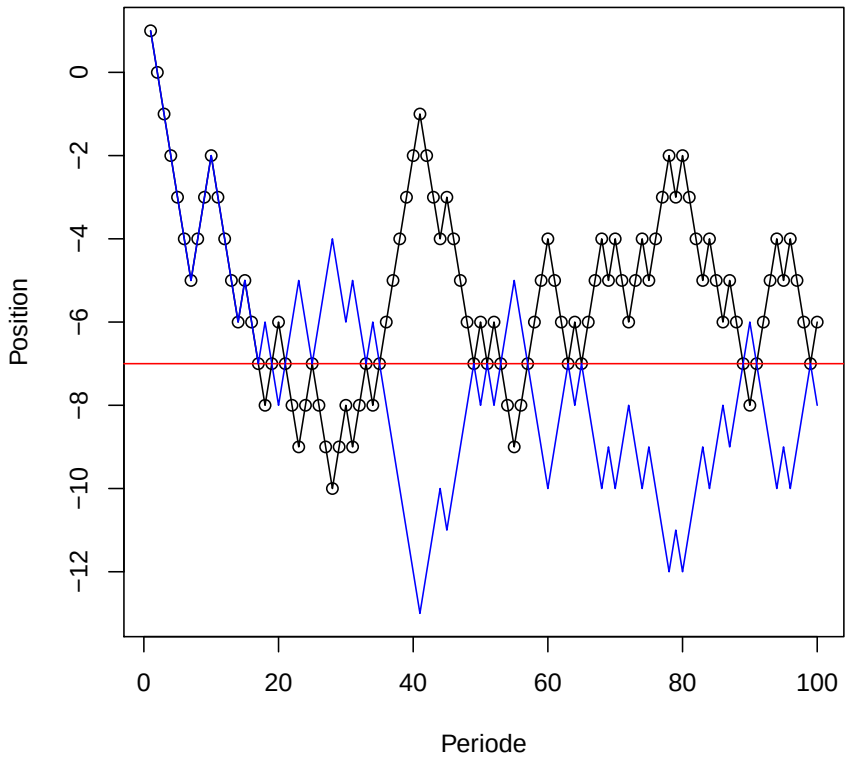

Abbildung 2.2: Ein Pfad der Irrfahrt und der an $-a = -7$ gespiegelte Pfad

und

$$\mathbb{E}(T_a) = \sum_{k=1}^{\infty} k\mathbb{P}(T_a = k) = +\infty. \tag{2.36}$$

In einem Modell mit unendlich vielen Perioden ist Ω jedoch überabzählbar, und die Konstruktion von $\mathbb{P}$ erfordert Masstheorie.

Beweis Sei $a > 0$. Dann gilt

$$\mathbb{P}(T_{-a} > N) \overset{(2.34)}{=} \mathbb{P}(S_N \in (-a, a]) \overset{(2.30),(2.31)}{\leq} \frac{\text{const.}}{\sqrt{\pi N}} \longrightarrow 0.$$

Ferner ist

$$\begin{aligned} \sum_{k=1}^{N+1} k\mathbb{P}(T_{-a} = k) &\overset{(2.19)}{=} \sum_{k=0}^{N} \mathbb{P}(T_{-a} > k) \overset{(2.34)}{=} \sum_{k=0}^{N} \mathbb{P}(S_k \in (-a, a]) \\ &\geq \sum_{k=1}^{N} \mathbb{P}(S_k \in \{0, 1\}) \overset{(2.30)}{\longrightarrow} +\infty. \end{aligned}$$

Für T_a mit $a > 0$ folgt die Behauptung aus Symmetrie. □

Das Korollar 2.1 gilt auch für T_0, die erste Rückkehrzeit nach Null, denn

Satz 2.3.

$$\mathbb{P}(T_0 > 2n) = \mathbb{P}(S_{2n} = 0). \tag{2.37}$$

Beweis Es gibt gleichviele Pfade der Länge $2n$, die stets oberhalb der x-Achse verlaufen, wie es Pfade der Länge $2n-1$ gibt, die nie -1 erreichen. Damit und mit einer Symmetrieüberlegung erhält man

$$\begin{aligned}\mathbb{P}(T_0 > 2n) &= \frac{1}{2}\mathbb{P}(T_{-1} > 2n-1) + \frac{1}{2}\mathbb{P}(T_1 > 2n-1) \;=\; \mathbb{P}(T_{-1} > 2n-1) \\ &\overset{(2.34)}{=} \mathbb{P}(S_{2n-1} \in (-1,1]) \;=\; \mathbb{P}(S_{2n-1} = 1) \overset{(2.30)}{=} \mathbb{P}(S_{2n} = 0)\,.\end{aligned}$$

□

Die Aussage $\mathbb{P}(T_0 > 2n) \to 0$ heisst auch "Die Irrfahrt ist rekurrent".

2.3.3 Das Arkussinus-Gesetz für den letzten Besuch in Null

Sei

$$L(\omega) = \max\{0 \le n \le 2N \mid S_n(\omega) = 0\}$$

der Zeitpunkt des *letzten Besuches in Null* vor $2N$. Vom Zeitpunkt $L(\omega)$ an gibt also einer der beiden Spieler die Führung nicht mehr ab.

Man könnte meinen, dass $L(\omega)$ – zumindest für grosses N – meistens nahe bei $2N$ sein wird, insbesondere also $\mathbb{P}(L \le N) \to 0$ für $N \to \infty$. Stattdessen gilt aber:
Satz 2.4 (Arkussinus-Gesetz)**.** *Die Verteilung von L ist die sogenannte diskrete Arkussinus-Verteilung:*

$$\mathbb{P}(L = 2n) = \mathbb{P}(S_{2n} = 0) \cdot \mathbb{P}(S_{2N-2n} = 0) = 2^{-2N}\binom{2n}{n}\binom{2N-2n}{N-n} \tag{2.38}$$

Die Verteilung ist symmetrisch um N (insbesondere ist $\mathbb{P}(L \le N) \approx \frac{1}{2}$), und U-förmig, siehe Abbildung 2.3. Man muss also damit rechnen, dass der Gewinner die endgültige Führung entweder recht früh oder recht spät übernimmt, und zwar passiert beides mit derselben Wahrscheinlichkeit!

```
> L <- rep(0,10000)
> for (i in (1:10000))
+ {
+   X <- sample(c(-1,1),size=200,replace=TRUE)
+   Y <- c(0,cumsum(X))
+   L[i] <- max(which(Y==0))
+ }
> hist(L,breaks=seq(-1,201,2),main="")
```

Beweis Die Anzahl Pfade bestehend aus $2N$ Schritten mit $L = 2n$ ist das Produkt der Anzahl Pfade bestehend aus $2n$ Schritten mit $S_{2n} = 0$ und der Anzahl Pfade bestehend aus $2N-2n$ Schritten mit $T_0 > 2N-2n$. Also ist wegen der Gleichverteilung aller Pfade $\mathbb{P}(L = 2n) = \mathbb{P}(S_{2n} = 0) \cdot \mathbb{P}(T_0 > 2N-2n)$. Aufgrund von (2.37) ist der zweite Faktor gleich $\mathbb{P}(S_{2N-2n} = 0)$. □

Warum "Arkussinus"?

$$\mathbb{P}(S_{2k} = 0) \approx \frac{1}{\sqrt{\pi k}} \Rightarrow \mathbb{P}(L = 2k) \approx \frac{1}{\pi\sqrt{k(N-k)}} = \frac{1}{N} f(\frac{k}{N}) \text{ mit } f(x) = \frac{1}{\pi\sqrt{x(1-x)}}.$$

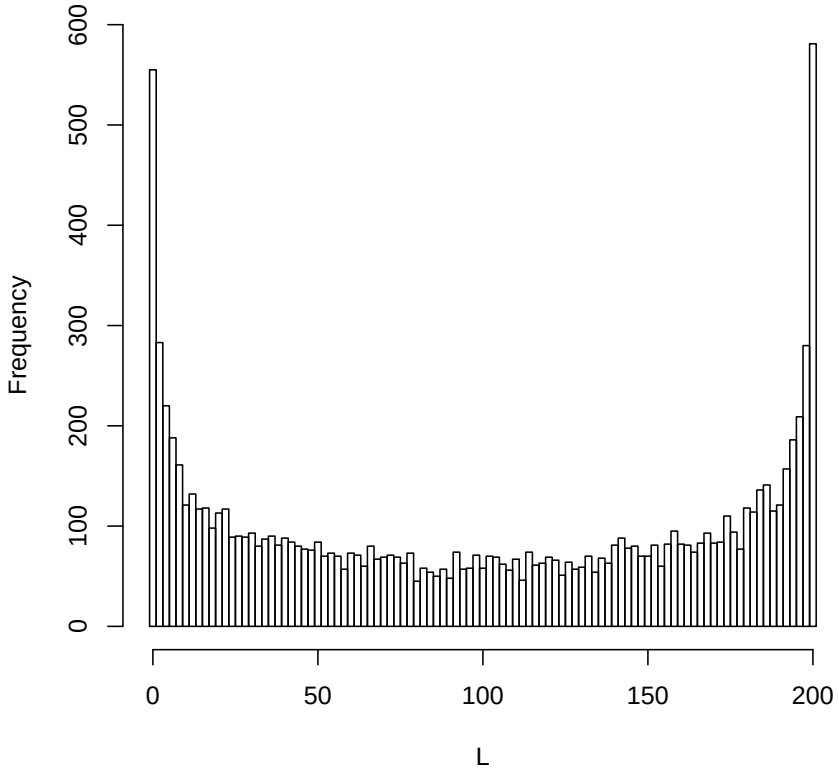

Abbildung 2.3: Histogramm von L für $N = 100$

Daraus folgt

$$\mathbb{P}\left(\frac{L}{2N} \leq z\right) \approx \sum_{k:\frac{k}{N}\leq z} \frac{1}{N} f(\frac{k}{N}) \approx \int_0^z f(x)dx = \frac{2}{\pi} \text{ arcsin } \sqrt{z}. \tag{2.39}$$

2.3.4 Spielsysteme

Nach (2.28) ist für festes n

$$\mathbb{E}\,(S_n) = 0,$$

d.h. der "Ertrag" S_n nach n Perioden ist im Schnitt gleich 0.

Kann man mehr erreichen, wenn man die Bilanzentwicklung $S_n(\omega) \quad (n = 0, \ldots, N)$ in einem zufälligen, aber günstigen Moment $T(\omega)$ stoppt? Kann man also durch geschickte Wahl einer Strategie zum vorzeitigen Abbruch des Spiels im Schnitt einen positiven Gewinn erzielen?

Die Entscheidung, im Zeitpunkt n zu stoppen oder nicht, darf sich natürlich nur auf die Entwicklung der Trajektorie bis zu diesem Zeitpunkt stützen. Es ist also keine "Insider"–Information über die zukünftige Entwicklung zugelassen! Um das mathematisch zu präzisieren, definieren wir zunächst die Klasse derjenigen Ereignisse, die nur vom Verhalten der Trajektorie bis zum Zeitpunkt n abhängen:

Definition 2.2. *Ein Ereignis $A \subseteq \Omega$ heisst* **beobachtbar bis zum Zeitpunkt** *n, wenn es von der Form $\{\omega \mid (X_1(\omega), \ldots, X_n(\omega)) \in C\}$ ist für ein $C \subseteq \{-1, 1\}^n$. Die Menge aller bis zum Zeitpunkt n beobachtbaren Ereignisse bezeichnen wir mit $\mathcal{A}_n$. Für $n = 0$ definieren wir $\mathcal{A}_0 = \{\emptyset, \Omega\}$.*

Bemerkung Jedes $A \in \mathcal{A}_n$ ist offensichtlich eine Vereinigung von Ereignissen der Form $\{\omega \mid X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n\}$. Diese Ereignisse bilden eine Partition von Ω. Ferner ist $A \in \mathcal{A}_n$ genau dann, wenn es von der Form $\{\omega \mid (S_1(\omega), \ldots, S_n(\omega)) \in D\}$ ist, wobei D jetzt eine Teilmenge aller möglichen Verläufe einer Irrfahrt mit n Perioden ist. Jedes $\mathcal{A}_n$ ist abgeschlossen gegen Komplementbildung und gegen (endliche) Vereinigungen und Durchschnitte. Es gilt

$$\mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_N = \text{ Potenzmenge von } \Omega. \tag{2.40}$$

Später wird für aufsteigende Mengensysteme, welche wie $(\mathcal{A}_n)$ abgeschlossen gegen Komplementbildung und abzählbare Vereinigungen sind, der Begriff *Filtration* eingeführt werden.

Definition 2.3. *Eine Abbildung $T : \Omega \to \{0, \ldots, N\}$ heisst* **Stoppzeit** *wenn gilt:*

$$\{\omega \mid T(\omega) = n\} \in \mathcal{A}_n \quad (n = 0, \ldots, N). \tag{2.41}$$

Bemerkung (2.41) ist äquivalent zu $\{T \leq n\} \in \mathcal{A}_n$ für $n = 0, \ldots, N$, weil $\{T \leq n\} = \cup_{k=0}^{n} \{T = k\}$. Es ist auch äquivalent zu $\{T \geq n\} \in \mathcal{A}_{n-1}$ für $n = 1, \ldots, N$, weil $\{T \geq n\} = \{T \leq n-1\}^c$.

Beispiel 2.6. *Sei T_c der erste Zeitpunkt > 0, in dem die Irrfahrt den Level $c \in \mathbb{Z}$ erreicht, also*

$$T_c(\omega) = \min\{n > 0 \;\mid\; S_n(\omega) = c\} \qquad (\min \emptyset := +\infty).$$

Dann ist $\min(T_c, N)$ eine Stoppzeit, denn für $c \geq 0$ und $n \leq N$ gilt

$$\{T_c = n\} = \{S_1 < c, S_2 < c, \ldots, S_n = c\} \in \mathcal{A}_n,$$

Für $c < 0$ argumentiert man analog.

Beispiel 2.7. *Betrachte den (ersten) Zeitpunkt, bei dem die Bilanz S_n maximal ist:*

$$T(\omega) \equiv \min\{n \mid \; S_n(\omega) = \max\{S_k(\omega) \mid k = 0, 1, \ldots N\}\}.$$

Es ist z.B. $\{\omega \mid T(\omega) = 0\} = \{\omega \mid S_1(\omega) \leq 0, S_2(\omega) \leq 0, \ldots, S_n(\omega) \leq 0\} \notin \mathcal{A}_0$, also ist dieses T keine Stoppzeit.

Der folgende Satz gibt die (ernüchternde und vom Experiment bestätigte) Antwort auf die eingangs gestellte Frage.

Satz 2.5. *Für jede Stoppzeit T ist*

$$\mathbb{E}\,(S_T) = 0$$

wobei $S_T(\omega) \equiv S_{T(\omega)}(\omega)$ den bei Benutzung der Stoppzeit T erzielten Ertrag bezeichnet.

Satz 2.5 ist ein Spezialfall eines allgemeinen Satzes über die **Unmöglichkeit von (lohnenden!) Spielsystemen**, den wir jetzt formulieren und beweisen werden.

Ein Spielsystem legt für jede Periode k fest, welcher Betrag $V_k(\omega)$ auf "+1" gesetzt wird, und zwar in Abhängigkeit von der bisherigen Entwicklung. Dieser Betrag kann auch gleich

Null oder negativ sein (dann setzt man $-V_k$ auf "-1"). Der resultierende Ertrag in Periode k ist dann

$$V_k(\omega) \cdot X_k(\omega),$$

und der resultierende Gesamtertrag

$$(V \cdot S)_N \equiv \sum_{k=1}^{N} V_k(\omega) \cdot X_k(\omega).$$

Die Bedingung, dass der Einsatz für Periode k sich nicht schon auf Informationen über die weitere Entwicklung stützen darf, wird präzisiert durch die folgende

Definition 2.4. *Ein* **Spielsystem** *ist eine Folge $V = (V_k)_{k=1,\dots,N}$ von Zufallsvariablen $V_k : \Omega \to \mathbb{R}^1$ derart, dass $V_1 =$ const. und für $k = 2, 3, \dots N$ existieren Funktionen $\phi_k : \{-1, +1\}^{k-1} \to \mathbb{R}$ mit*

$$V_k(\omega) = \phi_k(X_1(\omega), \dots X_{k-1}(\omega)). \tag{2.42}$$

Offensichtlich gilt für jedes Spielsystem

$$\{V_k = c\} \in \mathcal{A}_{k-1} \qquad (c \in \mathbb{R},\ k = 1, \dots, N). \tag{2.43}$$

Umgekehrt folgt aus (2.43), dass V ein Spielsystem ist: Weil V_k nur endlich viele Werte c_j annehmen kann und weil $\{V_k = c_j\} \in \mathcal{A}_{k-1}$, erhalten wir

$$V_k = \sum_j c_j 1_{[V_k = c_j]} = \sum_j c_j 1_{C_j}((X_1, \dots, X_{k-1}))$$

für Mengen $C_j \subseteq \{-1, +1\}^{k-1}$.

Beispiel 2.8. *Jede* **Stoppzeit** *T lässt sich als Spielsystem auffassen, wenn wir setzen*

$$V_k = 1_{\{T \geq k\}} = 1 - 1_{\{T \leq k-1\}}.$$

Die Bedingung (2.43) ist erfüllt, denn $\{T \leq k-1\} \in \mathcal{A}_{k-1}$ (T ist eine Stoppzeit!). Der resultierende Ertrag für dieses Spielsystem ist (wegen $T \leq N$)

$$(V \cdot S)_N = \sum_{k=1}^{N} 1_{\{T \geq k\}} X_k = \sum_{k=1}^{T} X_k = S_T.$$

Beispiel 2.9. "Sukzessives Verdoppeln des Einsatzes bis zur ersten $+1$" *ist offensichtlich ein Spielsystem im Sinne der obigen Definition.*

Beispiel 2.10. *Sei T eine Stoppzeit, und sei*

$$V_k \equiv S_{k-1} I_{\{T \geq k\}} \qquad (k = 1, \dots, N). \tag{2.44}$$

Dieses Spielsystem setzt liefert in Periode k den Ertrag

$$V_k X_k = \frac{1}{2}\left(S_k^2 - S_{k-1}^2 - 1\right) I_{\{T \geq k\}}$$

(denn: $S_k^2 = (S_{k-1} + X_k)^2 = S_{k-1}^2 + 1 + 2S_{k-1} \cdot X_k$) und den Gesamtertrag

$$(V \cdot S)_N = \frac{1}{2}\left(S_T^2 - T\right) \tag{2.45}$$

Satz 2.6. *Für jedes Spielsystem $V = (V_k)_{k=1,\ldots,N}$ ist der erwartete Ertrag*

$$\mathbb{E}\left((V \cdot S)_N\right) = 0$$

Beweis Wegen

$$\mathbb{E}\left((V \cdot S)_N\right) = \sum_{k=1}^{N} \mathbb{E}\left(X_k V_k\right)$$

(Linearität des Erwartungswertes) genügt es, $\mathbb{E}\left(X_k V_k\right) = 0$ zu zeigen. Gemäss der Definition eines Spielsystems ist aber

$$\begin{aligned}\mathbb{E}\left(X_k V_k\right) &= \sum_{x_1,\ldots x_k} x_k \phi_k(x_1,\ldots,x_{k-1}) \mathbb{P}\left(X_1 = x_1,\ldots,X_k = x_k\right) \\ &= \sum_{x_1,\ldots x_{k-1}} \phi_k(x_1,\ldots,x_{k-1})(1 \cdot \mathbb{P}\left(X_1 = x_1,\ldots,X_{k-1} = x_{k-1}, X_k = +1\right) + \\ &\quad (-1) \cdot \mathbb{P}\left(X_1 = x_1,\ldots,X_{k-1} = x_{k-1}, X_k = -1\right)),\end{aligned}$$

und wegen (2.26) ist jeder Term in Klammern auf der rechten Seiten gleich $1 \cdot 2^{-k} - 1 \cdot 2^{-k} = 0$. □

Als Korollar erhalten wir den Stoppsatz 2.5 und darüber hinaus die sogenannte
Korollar 2.2. Waldsche Identität: *Für jede Stoppzeit T ist*

$$\mathbb{E}\left(S_T\right) = 0 \tag{2.46}$$

und

$$\mathbb{E}\left(S_T^2\right) = \mathbb{E}\left(T\right) \tag{2.47}$$

Beweis Wende Satz 2.6 auf die Spielsysteme in den Beispielen 2.8 und 2.10 an. □

Die Waldsche Identität kann man sehr schön an Hand der Stoppzeit, wann ein bestimmtes Level $-a$ das erste Mal erreicht wird, darstellen, siehe Abbildung 2.4.

```
> N <- 100 # Anzahl der Zeitschritte der Irrfahrt
> K <- 1000 # Anzahl der Wiederholungen des Experimentes
> T <- rep(0,K)
> ST <- rep(0,K)
> a <- 10 # sollte eine gerade Zahl sein
> for (z in (1:K))
+ {
+   X <- sample(c(-1,1),size=N,replace=TRUE)
+   Y <- cumsum(X)
+   T[z] <- min((which(Y==a)),N)
+   ST[z] <- Y[T[z]]
+ }
> hist(ST,breaks=max(ST)-min(ST)/2,xlab="S_T",main="")
> mean(ST) # empirischer Mittelwert von $ S_T $
> mean(ST^2) # empirisches zweites Moment von $ S_T $
> mean(T) # empirischer Mittelwert von $ T $
```

Bemerkung In der stochastischen Analysis erkennt man die Irrfahrt als stochastischen Prozess mit Martingaleigenschaft. Der Begriff des Spielsystemes geht im allgemeineren Begriff der vorhersehbaren Strategie (des vorhersehbaren Prozesses) auf.

```
[1] 0.388
[1] 86.104
[1] 84.804
```

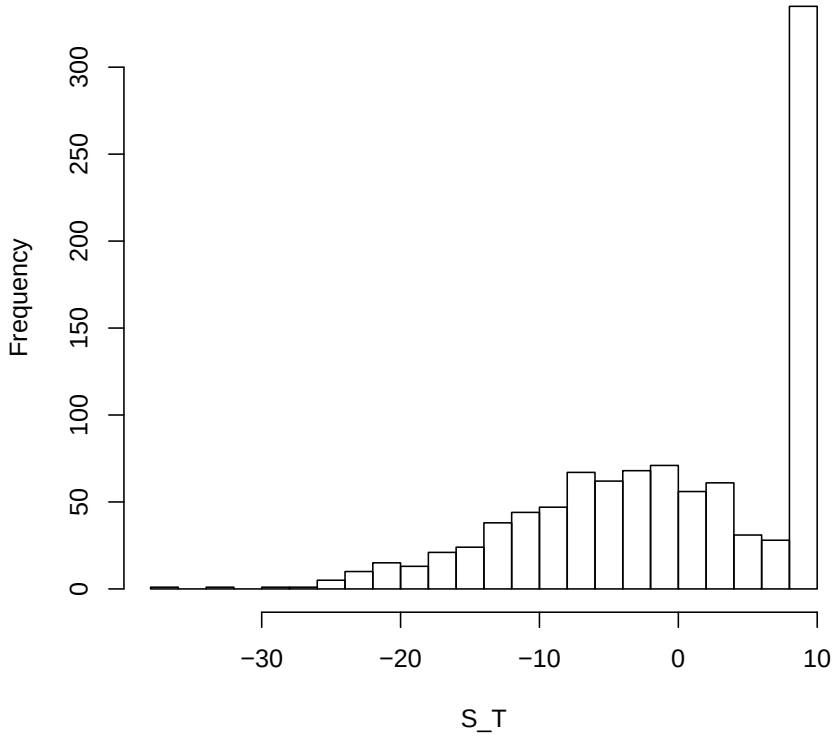

Abbildung 2.4: Verteilung von S_T bei Stoppen an $a = 10$

2.4 Bedingte Wahrscheinlichkeiten

2.4.1 Definition

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein (diskreter) Wahrscheinlichkeitsraum, und sei $B \in \mathcal{A}$ ein Ereignis mit $\mathbb{P}(B) > 0$. Es gibt oft Fälle, wo wir zwar nicht das genaue Ergebnis des Versuchs erfahren, aber wenigstens, dass B eingetreten ist. Dann modifizieren wir die Wahrscheinlichkeiten gemäss

Definition 2.5. *Die* **bedingte Wahrscheinlichkeit von** A **gegeben** B *ist*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2.48}$$

In der frequentistische Interpretation ist ja $\mathbb{P}(C) \approx n_C/n$ = relative Häufigkeit von C. Die bedingte Wahrscheinlichkeit von A gegeben B ist dann analog ungefähr gleich der relativen Häufigkeit von A unter den Versuchen, wo B eingetreten ist. In Formeln

$$\mathbb{P}(A|B) \approx \frac{n_{A\cap B}}{n_B} = \frac{n_{A\cap B}/n}{n_B/n}.$$

Damit sollte obige Definition einleuchten.

$\mathbb{P}(\cdot|B)$ ist eine neue Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$ mit Gewichten

$$p_B(\omega) \propto p(\omega) \quad \forall \omega \in B, \quad p_B(\omega) = 0 \quad \forall \omega \notin B.$$

Insbesondere, wenn $\mathbb{P}$ die Gleichverteilung auf Ω ist, dann ist $\mathbb{P}(\cdot|B)$ die Gleichverteilung auf B.

Beispiel 2.11. *Wurf zweier Würfel. Man wählt* $\Omega = \{(i,j) \mid 1 \leq i,j \leq 6\}$ *und* $\mathbb{P}$ *die Gleichverteilung. Sei* B_k *das Ereignis "Augensumme =k" und* A_i *= "Erster Würfel zeigt i". Dann* $\mathbb{P}(A_i|B_7) = \frac{1}{36} : \frac{1}{6} = \frac{1}{6}$ *für* $i = 1, 2, \ldots, 6$*; d.h. die Information, dass* B_7 *eingetreten ist, nützt nichts für die Prognose von* A_i*. Aber* $\mathbb{P}(A_i|B_{11}) = \frac{1}{2}$ *für* $i = 5, 6$ *und* $\mathbb{P}(A_i|B_{11}) = 0$ *für* $i \leq 4$*.*

Beispiel 2.12. *Für die Irrfahrt gilt* $\mathbb{P}(T_0 > 2n|X_1 = 1) = \mathbb{P}(T_{-1} > 2n-1)$*.*

2.4.2 Berechnung von absoluten Wahrscheinlichkeiten aus bedingten

Oft benützt man (2.48) nicht, um die bedingten Wahrscheinlichkeiten zu berechnen, sondern man postuliert Werte für die bedingten Wahrscheinlichkeiten und berechnet daraus die Wahrscheinlichkeiten des Durchschnitts. Ein typisches Beispiel ist die *Kombination zweier Zufallsexperimente:*

Im i-ten Versuch seien die möglichen Resultate $\omega_i \in \Omega_i$ $(i = 1, 2)$. Im Gesamtversuch haben wir dann den Grundraum

$$\Omega = \{(\omega_1, \omega_2) \mid \omega_i \in \Omega_i\} = \Omega_1 \times \Omega_2. \tag{2.49}$$

Wir betrachten die Ereignisse

$$A_\alpha = \{\omega \in \Omega : \ \omega_2 = \alpha\}, \quad B_\beta = \{\omega \in \Omega : \ \omega_1 = \beta\}, \tag{2.50}$$

welche die Ergebnisse in den beiden Teilversuchen angeben.

Eine Wahrscheinlichkeit $\mathbb{P}$ auf Ω kann man entweder durch die Angaben von $\mathbb{P}(A_\alpha \cap B_\beta)$ oder durch die Angaben von $\mathbb{P}(B_\beta)$ und $\mathbb{P}(A_\alpha|B_\beta)$ festlegen, denn es gilt

$$\mathbb{P}(\{\omega\}) = \mathbb{P}(A_{\omega_2} \cap B_{\omega_1}) = \mathbb{P}(A_{\omega_2}|B_{\omega_1}) \cdot \mathbb{P}(B_{\omega_1}). \tag{2.51}$$

Im zweiten Fall kann man $\mathbb{P}(B_\beta)$ und $\mathbb{P}(A_\alpha|B_\beta)$ beliebig wählen, ausser dass gelten muss:

$$\sum_\beta \mathbb{P}(B_\beta) = 1, \quad \sum_\alpha \mathbb{P}(A_\alpha|B_\beta) = 1 \quad \forall \beta. \tag{2.52}$$

Fasst man Ω als die Menge aller Pfade in einem zweistufigen Baum auf, dann geben die $\mathbb{P}(B_\beta)$ die Wahrscheinlichkeiten auf der obersten Stufe an und $\mathbb{P}(A_\alpha|B_\beta)$ die Wahrscheinlichkeiten auf der nächsten Stufe.

Die absoluten Wahrscheinlichkeiten $\mathbb{P}(A_\alpha)$ erhält man dann wie folgt:

$$\mathbb{P}(A_\alpha) = \sum_\beta \mathbb{P}(A_\alpha \cap B_\beta) = \sum_\beta \mathbb{P}(A_\alpha|B_\beta) \cdot \mathbb{P}(B_\beta). \tag{2.53}$$

Die Mengen (B_β) bilden eine disjunkte Zerlegung (Partition) von Ω, und analog zu (2.53) gilt allgemein

Satz 2.7 (Satz der totalen Wahrscheinlichkeit)**.** *Sei $(B_i)_{i\in I}$ eine disjunkte Zerlegung von Ω (d.h. $\Omega = \bigcup_{i\in I} B_i$, $B_i \cap B_j = \emptyset$ für $i \neq j$). Dann gilt für beliebiges A:*

$$\mathbb{P}(A) = \sum_{i:\mathbb{P}(B_i)>0} \mathbb{P}(A|B_i)\,\mathbb{P}(B_i)\,. \tag{2.54}$$

Beweis

$$A = \bigcup_{i\in I} (A \cap B_i) \overset{(2.5)}{\Rightarrow} \mathbb{P}(A) = \sum_i \underbrace{\mathbb{P}(A \cap B_i)}_{=\mathbb{P}(A|B_i)\mathbb{P}(B_i)}\,.$$

□

Beispiel 2.13. *Irrfahrt, vgl. den Beweis von Satz 2.3*

$$\begin{aligned}\mathbb{P}(T_0 > 2n) &= \mathbb{P}(T_0 > 2n|X_1 = 1)\cdot\mathbb{P}(X_1 = 1) + \mathbb{P}(T_0 > 2n|X_1 = -1)\cdot\mathbb{P}(X_1 = -1)\\ &= \mathbb{P}(T_{-1} > 2n-1)\cdot\frac{1}{2} + \mathbb{P}(T_1 > 2n-1)\cdot\frac{1}{2}.\end{aligned}$$

Beispiel 2.14. *Man wählt zuerst eine von zwei Urnen zufällig und zieht dann aus der gewählten Urne zufällig eine Kugel. Urne 1 enthält k weisse und ℓ rote Kugeln, Urne 2 $n-k$ weisse und $n-\ell$ rote. Dann*

$$\mathbb{P}(\text{ Kugel weiss }) = \frac{k}{k+\ell}\cdot\frac{1}{2} + \frac{n-k}{2n-(k+\ell)}\cdot\frac{1}{2}$$

Für welche Werte von k und ℓ wird dies – bei festem n – maximal? Nach einiger Rechnung (zuerst $k+\ell = m$ festhalten) erhält man $k = 1, \ell = 0$, d.h. man verteilt also die Risiken am besten sehr ungleich. In diesem Fall ist

$$\mathbb{P}(\text{ Kugel weiss }) = \frac{1}{2}\,\frac{3n-2}{2n-1} \overset{n\uparrow\infty}{\longrightarrow} \frac{3}{4}.$$

Ein weiteres nützliches Resultat ist
Satz 2.8. *Für beliebige Ereignisse $A_1, \ldots, A_n$ gilt*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\cdot\mathbb{P}(A_2|A_1)\cdot\mathbb{P}(A_3|A_1\cap A_2)\cdot\cdots\cdot\mathbb{P}(A_n|A_1\cap\cdots\cap A_{n-1})\,, \tag{2.55}$$

sofern die linke Seite > 0 ist.

Beweis Sukzessives Einsetzen der Definition 2.5. □

Beispiel 2.15. *Wie gross ist die Wahrscheinlichkeit, dass n Personen alle an verschiedenen Tagen Geburtstag haben? Sei G_i = Geburtstag der i-ten Person und $A_i = \{G_i \neq G_j$ für alle $j < i\}$. Dann*

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = 1\cdot\frac{364}{365}\cdot\frac{363}{365}\cdot\cdots\cdot\frac{365-n+1}{365}.$$

Man erhält die numerischen Werte $= 0.49$ für $n = 23$, 0.11 für $n = 40$ und 0.03 für $n = 50$. Die Wahrscheinlichkeiten sind also viel kleiner als man naiverweise vermutet.

2.4.3 Bayessche Regel

Aus der Definition der bedingten Wahrscheinlichkeit folgt sofort die Bayessche Formel, welche den Zusammenhang zwischen $\mathbb{P}(A|B)$ und $\mathbb{P}(B|A)$ beschreibt:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\,\mathbb{P}(B)}{\mathbb{P}(A)}.$$

Die bedingte Wahrscheinlichkeit ist also nicht symmetrisch. Die Grössen auf der rechten Seite hängen aber im Allgemeinen voneinander ab. Mit dem Satz von der totalen Wahrscheinlichkeit folgt die Version

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\,\mathbb{P}(B)}{\mathbb{P}(A|B)\,\mathbb{P}(B) + \mathbb{P}(A|B^c)\,(1-\mathbb{P}(B))}. \tag{2.56}$$

Die Wahrscheinlichkeiten $\mathbb{P}(B)$, $\mathbb{P}(A|B)$ und $\mathbb{P}(A|B^c)$ auf der rechten Seite können beliebige Werte annehmen. Betrachten wir zunächst ein typisches Beispiel:

Beispiel 2.16. *Von* 145 *Ihres Alters hat einer die Krankheit K. Für das Ereignis B = "Sie haben K" gilt also a priori* $\mathbb{P}(B) = \frac{1}{145}$. *Sie machen nun einen Test, und es tritt das Ereignis A= "Ergebnis positiv" ein. Nun ist aber kein Test völlig fehlerfrei. Nehmen wir zum Beispiel an:*

$$\mathbb{P}(A|B) = 0.96, \quad \mathbb{P}(A^c|B^c) = 0.94.$$

Dann folgt mit der Formel (2.56)

$$\mathbb{P}(B|A) = \frac{\frac{96}{100}\cdot\frac{1}{145}}{\frac{96}{100}\cdot\frac{1}{145} + \frac{6}{100}\cdot\frac{144}{145}} = \frac{1}{10}$$

(noch kein Grund zur Panik!).

Das erstaunliche Ergebnis in diesem Beispiel wird klarer, wenn wir das Wettverhältnis betrachten:

$$\frac{\mathbb{P}(B|A)}{\mathbb{P}(B^c|A)} = \frac{\mathbb{P}(A|B)}{\mathbb{P}(A|B^c)} \cdot \frac{\mathbb{P}(B)}{\mathbb{P}(B^c)}.$$

Das a posteriori Wettverhältnis (nach dem Eintreten von A) hängt also nicht nur davon ab, wie wahrscheinlich A unter B, bzw. B^c ist, sondern auch vom a priori Wettverhältnis.

Betrachten wir an Stelle von (B, B^c) eine beliebige disjunkte Zerlegung von Ω, so erhalten wir den allgemeinen Satz von Bayes:

Satz 2.9. *Ist* $(B_i)_{i\in I}$ *eine Zerlegung von* Ω *in disjunkte Ereignisse und* $\mathbb{P}(A) \neq 0$, *so ist*

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\cdot\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\cdot\mathbb{P}(B_j)}. \tag{2.57}$$

Beweis

$$\mathbb{P}(B_i|A) \overset{(2.48)}{=} \frac{\mathbb{P}(A|B_i)\cdot\mathbb{P}(B_i)}{\mathbb{P}(A)}.$$

Die Behauptung folgt, wenn man für den Nenner (2.54) einsetzt. □

Dieses Resultat lässt sich wie folgt interpretieren: Wir haben verschiedene Hypothesen B_i mit a priori Wahrscheinlichkeiten $\mathbb{P}(B_i)$; unter der Hypothese B_i hat A die Wahrscheinlichkeit $\mathbb{P}(A|B_i)$. Wenn nun A eintritt, kann man aus den Grössen $\mathbb{P}(B_i)$ und $\mathbb{P}(A|B_i)$ die a posteriori Wahrscheinlichkeiten $\mathbb{P}(B_i|A)$ der verschiedenen Hypothesen berechnen.

Beispiel 2.17. *Wir betrachten einen* **Nachrichtenkanal** *mit Eingangsalphabet I und Ausgangsalphabet J. Sei* A_j *das Ereignis "Signal j wird empfangen"* $(j \in J)$ *und* B_i *das Ereignis "Signal i wird gesendet"* $(i \in I)$.

Die Übertragung ist jedoch nicht fehlerfrei (Rauschen !). Die Übertragungseigenschaften des Kanals *sind durch die bedingten Wahrscheinlichkeiten* $\mathbb{P}(A_j|B_i)$ $(i \in I, j \in J)$ *beschrieben, die* Struktur der Nachrichtenquelle *durch die Wahrscheinlichkeiten* $\mathbb{P}(B_i)$ $(i \in$

I). Daraus kann der Empfänger die bedingten Wahrscheinlichkeiten $\mathbb{P}(B_i|A_j)$ gemäss (2.57) berechnen. Gesucht ist nun eine **Dekodierung** *$\varphi : J \to I$, für die das Ereignis*

$$C_\varphi = \text{“ richtig dekodiert” } = \bigcup_j (A_j \cap B_{\varphi(j)})$$

maximale Wahrscheinlichkeit hat. Es gilt

$$\mathbb{P}(C_\varphi) = \sum_j \mathbb{P}(A_j) \cdot \mathbb{P}\left(B_{\varphi(j)} \middle| A_j\right), \tag{2.58}$$

und, offensichtlich können wir in (2.58) jeden Summanden einzeln maximieren. Wir erhalten also die folgende Lösung: Wähle $\varphi(j)$ für jedes $j \in J$ so, dass

$$\mathbb{P}\left(B_{\varphi(j)} \middle| A_j\right) = \max_i \mathbb{P}(B_i|A_j).$$

Illustration im binären Fall *$I = J = \{0, 1\}$:* *Sei*

$$p_1 \equiv \mathbb{P}(A_1|B_1), \quad p_0 \equiv \mathbb{P}(A_0|B_0), \quad \alpha \equiv \mathbb{P}(B_1)$$

Es gibt vier mögliche Dekodierungen gemäss folgender Tabelle

Dekodierung φ	C_φ	$\mathbb{P}(C_\varphi)$
$\varphi_1 \equiv 1$	B_1	α
$\varphi_2 \equiv 0$	B_0	$1-\alpha$
$\varphi_3(1) = 1, \quad \varphi_3(0) = 0$	$(A_1 \cap B_1) \cup (A_0 \cap B_0)$	$\alpha p_1 + (1-\alpha)p_0$
$\varphi_4(1) = 0, \quad \varphi_4(0) = 1$	$(A_1 \cap B_0) \cup (A_0 \cap B_1)$	$\alpha(1-p_1) + (1-\alpha)(1-p_0)$

Wenn sowohl p_0 als auch p_1 grösser als 0.5 sind, dann ist die optimale Dekodierung gegeben durch

$$\varphi = \begin{cases} \varphi_2 & \text{falls } 0 \le \alpha \le \frac{1-p_0}{1-p_0+p_1} \\ \varphi_1 & \text{falls } \frac{p_0}{p_0+1-p_1} \le \alpha \le 1 \\ \varphi_3 & \text{sonst} \end{cases}$$

2.5 Ausblick: Der bedingte Erwartungswert für diskrete Wahrscheinlichkeitsräume

Sei $(\Omega, \mathcal{A}, P)$ ein diskreter Wahrscheinlichkeitsraum. Für ein Ereignis $B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$, gibt die bedingte Wahrscheinlichkeit $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ an, wie wahrscheinlich das Ereignis A ist, wenn B eingetreten ist. Entsprechend gibt der bedingte Erwartungswert

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(1_B X)}{\mathbb{P}(B)} = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x|B) = \sum_\omega X(\omega)\mathbb{P}(\{\omega\}|B)$$

an, welchen Wert man für die Zufallsvariable X im Mittel erwartet, wenn man Information über das Eintreten von B erhalten hat.

Dieser elementare Begriff von bedingten Erwartungswerten ist jedoch oft nicht ausreichend. Für ein allgemeineres Konzept betrachten wir eine Partition $\mathcal{B} = (B_i)_{i \in I}$ von Ω, d.h. eine Zerlegung von Ω in disjunkte, nicht leere Teilmengen, wobei I eine abzählbare Indexmenge bezeichnet. Dann definieren wir die **Zufallsvariable**

$$\mathbb{E}(X|\mathcal{B})(\omega) = \sum_{i \in I, \mathbb{P}(B_i) > 0} \mathbb{E}(X|B_i)\ 1_{B_i}(\omega).$$

Ihr Wert ist festgelegt, sobald man weiss, welches B_i realisiert wurde (man braucht also den genauen Wert von ω nicht), und sie gibt an, welchen Wert man dann für X erwartet. Diese Zufallsvariable wird daher als bedingte Erwartung von X gegeben $\mathcal{B}$ bezeichnet.

Insbesondere kann die Partition von einer anderen Zufallsvariable Y mit möglichen Werten $y_1, y_2, \ldots$ erzeugt sein, d.h. $B_i = \{\omega \,|\, Y(\omega) = y_i\}$. Dann schreiben wir

$$\mathbb{E}\,(X|Y) = \sum_{y_i} \mathbb{E}\,(X|Y = y_i)\ 1_{y_i}(Y).$$

Satz 2.10. *Sei X eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}\,(X^2) < \infty$ und sei $\mathcal{B} = (B_i)_{i\in I}$ eine Partition von Ω. Dann wird*

$$\mathbb{E}\left(\left(X - \sum_{i\in I, \mathbb{P}(B_i)>0} c_i 1_{B_i}\right)^2\right)$$

minimal für $c_i = \mathbb{E}\,(X1_{B_i})\,/\mathbb{P}\,(B_i)$, d.h. die bedingte Erwartung ergibt die beste Prognose von X auf Grund der Partition $\mathcal{B}$ ("beste" im Sinne des mittleren quadratischen Fehlers).

Beweis Wir schreiben kurz $\sum_i$ für $\sum_{i\in I, \mathbb{P}(B_i)>0}$. Dann gilt für beliebige Wahl der c_i:

$$\begin{aligned}\mathbb{E}\left(X\sum_i c_i 1_{B_i}\right) &= \sum_i c_i \mathbb{E}\,(X1_{B_i}) = \mathbb{E}\left(\sum_i c_i \frac{\mathbb{E}\,(X1_{B_i})}{\mathbb{P}\,(B_i)} 1_{B_i}\right) = \mathbb{E}\left(\sum_{i,j} c_i \frac{\mathbb{E}\left(X1_{B_j}\right)}{\mathbb{P}\,(B_j)} 1_{B_i}1_{B_j}\right)\\ &= \mathbb{E}\left(\mathbb{E}\,(X|\mathcal{B})\sum_i c_i 1_{B_i}\right),\end{aligned}$$

weil $1_{B_i}1_{B_j} = 0$ $(i \neq j)$ und $1_{B_i}1_{B_i} = 1_{B_i}$. Also folgt

$$\mathbb{E}\left((X - \mathbb{E}\,(X|\mathcal{B}))\sum_i c_i 1_{B_i}\right) = 0. \tag{2.59}$$

Insbesondere gilt auch $\mathbb{E}\,((X - \mathbb{E}\,(X|\mathcal{B}))(\mathbb{E}\,(X|\mathcal{B}) - \sum_i c_i 1_{B_i})) = 0$, und damit

$$\mathbb{E}\left((X - \sum_i c_i 1_{B_i})^2\right) = \mathbb{E}\left((X - \mathbb{E}\,(X|\mathcal{B}))^2\right) + \mathbb{E}\left((\mathbb{E}\,(X|\mathcal{B}) - \sum_i c_i 1_{B_i})^2\right).$$

Der zweite Term rechts ist immer positiv ausser wenn $c_i = \mathbb{E}\,(X1_{B_i})\,/\mathbb{P}\,(B_i)$. □

Die Formel (2.59) zeigt, dass $\mathbb{E}\,(X|\mathcal{B})$ die orthogonale Projektion von X auf den Unterraum aller Funktionen der Form $\sum_i c_i 1_{B_i}$ ist bezüglich des Skalarprodukts

$$\langle X, Y\rangle := \sum_\omega X(\omega)Y(\omega)p(\omega).$$

Im Fall, wo Ω abzählbar ist, gehört zu jeder Teil-σ-Algebra von $\mathcal{A}$ eine Partition, und umgekehrt. Im Anhang C ist die Definition der bedingten Erwartung $\mathbb{E}\,(X|\mathcal{B})$ für allgemeine Wahrscheinlichkeitsräume und Teil-σ-Algebren $\mathcal{B}$ ausgeführt. Dort wird die zu (2.59) äquivalente Eigenschaft

$$\mathbb{E}\,(XZ) = \mathbb{E}\,(\mathbb{E}\,(X|\mathcal{B})\,Z)$$

benutzt, um die bedingte Erwartung zu definieren.

Abbildung 2.5 illustriert die Trajektorie einer exponentiellen Irrfahrt und ihrer bedingten Erwartung unter wachsender Information. Die exponentielle Irrfahrt $(Y_n)_{0\leq n\leq N}$ (schwarze Trajektorie) ist durch $Y_n = \exp(S_n/\sqrt{N})$ gegeben, wobei S_n eine gewöhnliche Irrfahrt mit N Perioden ist. Die rote Linie ist eine Trajektorie der bedingten Erwartung $n \mapsto \mathbb{E}(Y_N|\mathcal{A}_n)$. Man kann zeigen, dass

$$\mathbb{E}(Y_N|\mathcal{A}_n) = Y_n \cosh(1/\sqrt{N})^{N-n}.$$

```
> N <- 100 # Anzahl der Zeitschritte der exponentiellen Irrfahrt
> Y <- sample(c(-1,1),size=N,replace=TRUE)
> Y <- c(1,exp(cumsum(Y)/sqrt(N)))
> Ycond <- Y*((exp(1/sqrt(N))+exp(-1/sqrt(N)))*0.5)^(N:0)
> plot(Y,type="l",ylim=range(Y,Ycond),xlab="Periode")
> lines(Ycond,col="red")
```

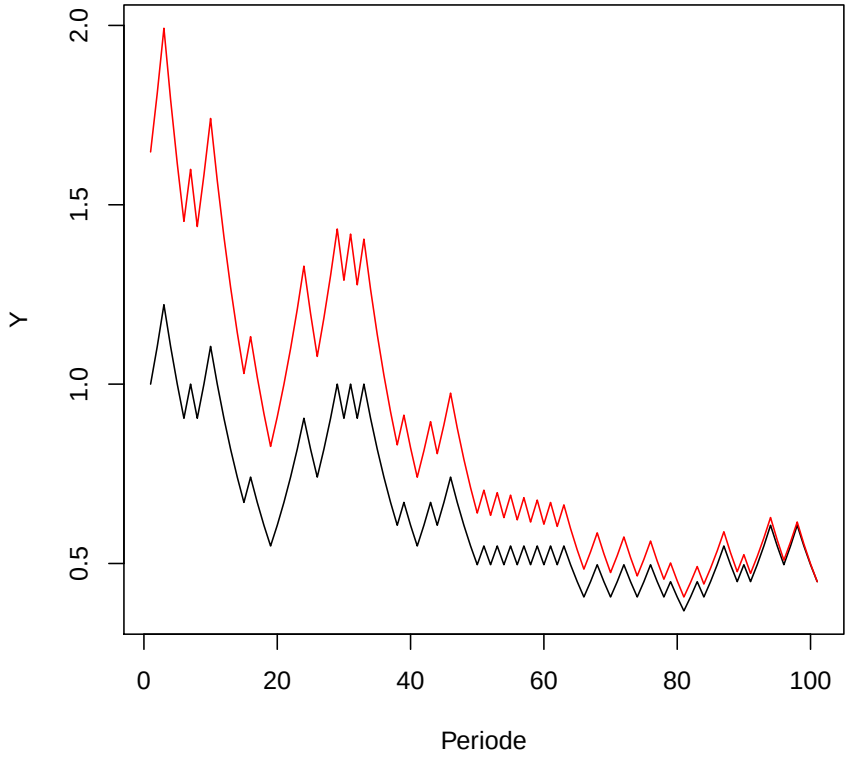

Abbildung 2.5: Exponentielle Irrfahrt mit bedingtem Erwartungswert des Endpunktes unter wachsender Information

2.6 Unabhängigkeit

2.6.1 Definition von Unabhängigkeit

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein (diskreter) Wahrscheinlichkeitsraum.

Definition 2.6. *Eine Kollektion von Ereignissen* $(A_i; i \in I)$ *heisst* **(stochastisch) unabhängig** *wenn gilt:*

$$J \subseteq I \text{ endlich } \Rightarrow \mathbb{P}\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} \mathbb{P}(A_i)\,. \tag{2.60}$$

Bemerkungen

- Unabhängigkeit ist keine Eigenschaft der Ereignisse per se, sondern eine Eigenschaft in Bezug auf die Wahrscheinlichkeitsverteilung $\mathbb{P}$.
- Für zwei Ereignisse A, B mit positiver Wahrscheinlichkeit gilt:

$$A, B \text{ sind unabhängig } \iff \mathbb{P}(A|B) = \mathbb{P}(A) \iff \mathbb{P}(B|A) = \mathbb{P}(B)\,. \tag{2.61}$$

- Die paarweise Unabhängigkeit impliziert noch nicht (2.60). Zum Beispiel sind beim Wurf zweier Münzen die Ereignisse

  A = "Erster Wurf Kopf"
  B = "Zweiter Wurf Kopf"
  C = "Die Ergebnisse der zwei Würfe sind verschieden"

  paarweise unabhängig, aber $\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C)$.
- Bei endlich vielen Ereignissen $A_1, \ldots, A_n$ genügt es nicht, die Produktformel in (2.60) für $J = \{1, \ldots, n\}$ zu fordern. Dies sieht man sofort ein, wenn man $A_1 = \emptyset$ und $A_2 = A_3$ wählt mit $0 < \mathbb{P}(A_2) < 1$.
- Wie bei der bedingten Wahrscheinlichkeiten, *postuliert* man oft die Unabhängigkeit gewisser Ereignisse, um Wahrscheinlichkeiten festzulegen.

Lemma 2.3. *Die Ereignisse* $A_i \quad (i \in I)$ *seien unabhängig. Sei* $B_i = A_i$ *oder* $= A_i^c$. *Dann sind auch die Ereignisse* $B_i \quad (i \in I)$ *unabhängig.*

Beweis Für disjunkte endliche Mengen $J, K \subseteq I$ ist

$$\mathbb{P}\left(\bigcap_{i\in J} A_i \cap \bigcap_{i\in K} A_i^c\right) = \prod_{i\in J} \mathbb{P}(A_i) \prod_{i\in K} \mathbb{P}(A_i^c) \tag{2.62}$$

zu zeigen. Wir benützen Induktion nach $k = |K|$. Für $k = 0$ ist (2.62) richtig, und aus der Gültigkeit für $|K| = k$ folgt

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i\in J} A_i \cap \bigcap_{i\in K} A_i^c \cap A_j^c\right) &= \mathbb{P}\left(\bigcap_{i\in J} A_i \cap \bigcap_{i\in K} A_i^c\right) - \mathbb{P}\left(\bigcap_{i\in J} A_i \cap A_j \cap \bigcap_{i\in K} A_i^c\right) \\ &= \prod_{i\in J} \mathbb{P}(A_i) \prod_{i\in K} \mathbb{P}(A_i^c) \cdot (1 - \mathbb{P}(A_j)),\end{aligned}$$

d.h. (2.62) gilt auch für $\tilde{K} = K \cup \{j\}$. □

Bemerkung: Für endliches I ist die Unabhängigkeit äquivalent zur Gültigkeit von (2.62) für alle $J \subseteq I$, $K = J^c$.

Schliesslich definieren wir noch die Unabhängigkeit von Zufallsvariablen.

Definition 2.7. *Eine Kollektion von diskreten Zufallsvariablen* $(X_i; i \in I)$ *heisst* **unabhängig**, *falls die Ereignisse* $(\{X_i = x_i\}; i \in I)$ *unabhängig sind für jede Wahl von* x_i *aus dem Wertebereich von* X_i.

Lemma 2.4. *Wenn die diskreten Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind, dann gilt*

$$\mathbb{E}\left(\prod_{i=1}^{n} g_i(X_i)\right) = \prod_{i=1}^{n} \mathbb{E}\left(g_i(X_i)\right)$$

für beliebige Funktionen g_i (sofern die Erwartungswerte existieren).

Beweis Es gilt wegen (2.16), bzw. gemäss der Definition der Unabhängigkeit

$$\begin{aligned}\mathbb{E}\left(\prod_{i=1}^{n} g_i(X_i)\right) &= \sum_{x_1,\ldots,x_n} \prod_{i=1}^{n} g_i(x_i)\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) = \sum_{x_1,\ldots,x_n} \prod_{i=1}^{n} g_i(x_i)\mathbb{P}\left(X_i = x_i\right) \\ &= \prod_{i=1}^{n} \sum_{x_i} g_i(x_i)\mathbb{P}\left(X_i = x_i\right) = \prod_{i=1}^{n} \mathbb{E}\left(g_i(X_i)\right).\end{aligned}$$

□

2.6.2 Unabhängige 0-1-Experimente mit Erfolgsparameter p

Wir betrachten das folgende Modell für n 0-1-Experimente: Der Grundraum ist die Menge aller 0 - 1 Folgen der Länge n, d.h

$$\Omega = \{\omega = (x_1, \ldots, x_n) \mid x_i \in \{0, 1\}\}.$$

Die Zufallsvariablen X_i geben das Ergebnis im i-ten Experiment an, d.h.

$$X_i = \text{i-te Komponente von } \omega.$$

Im **Laplace-Modell** ist $\mathbb{P}$ die Gleichverteilung auf Ω, und das impliziert nach (2.25) und (2.26):

$$\mathbb{P}\left(X_i = +1\right) = \frac{1}{2} \quad (i = 1, \ldots, n) \tag{2.63}$$

$$\text{Die Ereignisse } \{X_i = +1\} \quad (i = 1, \ldots, n) \text{ sind unabhängig.} \tag{2.64}$$

Wir geben uns jetzt einen beliebigen **Erfolgsparameter**

$$0 < p < 1$$

vor und suchen ein Wahrscheinlichkeitsmass $\mathbb{P}$ auf Ω, für das (2.64) gilt und ausserdem

$$\mathbb{P}\left(X_i = 1\right) = p \quad (i = 1, \ldots, n). \tag{2.65}$$

Durch diese Bedingungen ist $\mathbb{P}$ festgelegt, denn für ein beliebiges $\omega = (x_1, \ldots, x_n)$ gilt wegen Lemma 2.3

$$\mathbb{P}\left(\{\omega\}\right) = \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i = x_i\}\right) \overset{(2.64)}{=} \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right) \overset{(2.65)}{=} p^k (1-p)^{n-k},$$

also

$$\mathbb{P}\left(\{\omega\}\right) = p^k (1-p)^{n-k} \quad , \text{ falls } \sum_{i=1}^{n} x_i = k \tag{2.66}$$

Umgekehrt ergeben sich aus (2.66), aufgefasst als **Definition** von $\mathbb{P}$, die gewünschten Eigenschaften (2.65), (2.64).

Wenn wir 1 als "Erfolg" interpretieren, dann ist

$$S_n(\omega) = X_1(\omega) + \cdots + X_n(\omega) \tag{2.67}$$

die **Anzahl der Erfolge**. Wegen $\mathbb{E}(X_i) = p$ folgt aus (2.67)

$$\mathbb{E}(S_n) = n \cdot p. \tag{2.68}$$

Die Verteilung von S_n ist gegeben durch

$$\begin{aligned} \mathbb{P}(S_n = k) &= \sum_{\omega=(x_1,\ldots,x_n):\ x_1+\cdots+x_n=k} \mathbb{P}(\{\omega\}) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, \ldots, n). \end{aligned} \tag{2.69}$$

Dies ist die **Binomialverteilung mit Parametern p und n.**

Man kann dies auch mit folgender analytischer Method sehen: für jede reelle Zahl λ gilt

$$\begin{aligned} \mathbb{E}(\exp(\mathrm{i}\lambda S_n)) &= \mathbb{E}(\exp(\mathrm{i}\lambda(X_1 + \cdots + X_n))) && (2.70)\\ &= \mathbb{E}(\exp(\mathrm{i}\lambda X_1)) \ldots \mathbb{E}(\exp(\mathrm{i}\lambda X_n)) && (2.71)\\ &= \left((1-p) + \exp(\mathrm{i}\lambda)p\right)^n && (2.72)\\ &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \exp(\mathrm{i}\lambda k) && (2.73) \end{aligned}$$

mit Unabhängigkeit und der Eigenschaft, dass die Verteilungen der X_i identisch sind. Vergleicht man die letzte Summe mit dem Ausdruck

$$\sum_{k=0}^{n} \mathbb{P}(S_n = k) \exp(\mathrm{i}\lambda k)$$

für $\mathbb{E}(\exp(\mathrm{i}\lambda S_n))$, dann erhält man durch Koeffizientvergleich (das gilt ja für alle λ!) die obige Formel.

Die Binomialwahrscheinlichkeiten lassen sich leicht rekursiv berechnen. Sei $p_n(k) = \binom{n}{k} p^k q^{n-k}$ mit $q = 1 - p$. Dann gilt

$$p_n(k+1) = \frac{n-k}{k+1} \frac{p}{q} p_n(k),$$

insbesondere ist $p_n(k)$ ist maximal für $k \approx np$. In der Nähe von $k = np$ können wir $p_n(k)$ mit Hilfe der Stirling Formel approximieren (cf. (2.30)). Mit einer Taylorentwicklung folgt daraus eine Approximation der Binomialverteilung für grosses n.

Satz 2.11 (de Moivre-Laplace)**.** *Es gilt*

$$p_n(k) = \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k-np)^2}{2npq}\right)(1 + r_n(k)) \tag{2.74}$$

mit

$$\sup\{|r_n(k)| : |k - np| \le A\sqrt{n}\} \to 0 \quad \textit{für alle } A > 0 \quad (n \to \infty)$$

Beweis Aus der Stirling-Formel $k! \sim \sqrt{2\pi k}(\frac{k}{e})^k$ folgt

$$p_n(k) \sim \frac{\sqrt{2\pi n}\quad n^n p^k q^{n-k}}{\sqrt{(2\pi)^2 k(n-k)}\ k^k (n-k)^{n-k}} = \frac{1}{\sqrt{2\pi n \frac{k}{n}(1-\frac{k}{n})}} \exp(n g_p(k/n))$$

wobei

$$g_p(x) = x(\log(p) - \log(x)) + (1-x)(\log(1-p) - \log(1-x)).$$

Wir entwickeln als nächstes g_p in eine Taylorreihe an der Stelle $x = p$. Man erhält $g_p(p) = g_p'(p) = 0$ und $g_p''(p) = -1/(p(1-p))$. Daraus folgt, dass für $|k/n - p| \le A/\sqrt{n}$

$$\exp(n g_p(k/n)) = \exp\left(-\frac{1}{2npq}(k-np)^2 + n\,\mathcal{O}(n^{-3/2})\right).$$

Weiters kann natürlich der erste Ausdruck mit Taylor entwickelt werden, wir erhalten

$$\frac{1}{\sqrt{2\pi n \frac{k}{n}(1-\frac{k}{n})}} = \frac{1}{\sqrt{2\pi np(1-p)}} + \mathcal{O}(\frac{k}{n} - p)$$

Daraus folgt schliesslich die Behauptung.

□

In den Abbildungen 2.6 und 2.7 sind die absoluten und relativen Fehler der Approximation im Satz von de Moivre-Laplace dargestellt. Der absolute Fehler ist im ganzen Bereich klein, der relative Fehler nur in der Nähe des Erwartungswertes.

Durch Approximation des Integrals durch eine Riemannsumme erhält man ferner

$$\begin{aligned}&\mathbb{P}\left(np - A\sqrt{npq} \le S_n \le np + B\sqrt{npq}\right)\\ &\quad\sim \int_{-A\sqrt{pq}}^{B\sqrt{pq}} \frac{1}{\sqrt{2\pi pq}} e^{-x^2/(2pq)} dx = \int_{-A}^{B} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \Phi(B) - \Phi(-A), \quad (2.75)\end{aligned}$$

wobei

$$\Phi(B) = \int_{-\infty}^{B} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

tabelliert ist. Dies ist ein Spezialfall des **zentralen Grenzwertsatzes**, den wir im Kapitel 4.3 ausführlicher behandeln.

Beispiel 2.18. Mendels Versuche in der Genetik. *Die Keimblätter von Gartenerbsen sind gelb oder grün. Gemäss Mendels Theorie enthält jede Pflanze ein Genpaar für die Farbe, entweder ge/ge oder ge/gr oder gr/gr oder gr/ge. Das Gen ge ist dominant, so dass nur Pflanzen mit Gen gr/gr grüne Keimblätter aufweisen. Kreuzt man zwei reine Stämme, so erhält in der zweiten Generation mit Wahrscheinlichkeit* $\frac{1}{4}$ *grüne Keimblätter (alle Genpaare sind gleich wahrscheinlich). Gregot Mendel erhielt unter* 8023 *Pflanzen der zweiten Generation* 2001 *Pflanzen mit grünen Keimblättern. Da* $np = \frac{1}{4}8023 = 2005.75, \sqrt{npq} = 38.8$, *ist*

$$\frac{2001 - np}{\sqrt{npq}} = -0.12.$$

Also ist die Wahrscheinlichkeit für eine mindestens so grosse Abweichung $|S_n - np|$ *wie Mendel sie beobachtete,* $\approx 1 - (\Phi(0.12) - \Phi(-0.12)) \approx 90\%$, *d.h. die Übereinstimmung zwischen Theorie und Experiment ist sehr gut. Beachte dass man dies auch anders interpretieren kann: die Wahrscheinlichkeit für eine so kleine Abweichung vom Wert np ist cirka*

```
> demoivre <- function(m,n,p,A)
+   {
+     w1r <- rbinom(m,n,p)
+     w2r <- rnorm(m,n*p,sqrt(n*p*(1-p)))
+     w1 <- dbinom(0:n,n,p)
+     w2 <- dnorm(0:n,n*p,sqrt(n*p*(1-p)))
+     k1 <- ceiling(n*p-A*sqrt(n*p*(1-p)))
+     k2 <- floor(n*p+A*sqrt(n*p*(1-p)))
+     par(mfrow=c(2,3))
+     hist(w1r)
+     hist(w2r)
+     plot(0:n,w1,type="l",xlab="k",ylab="Binomialverteilung")
+     plot(0:n,w2,type="l",xlab="k",ylab="Normalverteilung")
+     plot(0:n,abs(w1-w2),type="h",xlab="k",ylab="Absoluter Fehler")
+     plot((k1:k2),w1[k1:k2+1]/w2[k1:k2+1]-1,type="h",xlab="k",
+          ylab="Relativer Fehler")
+   }
> demoivre(10^5,50,0.2,3)
```

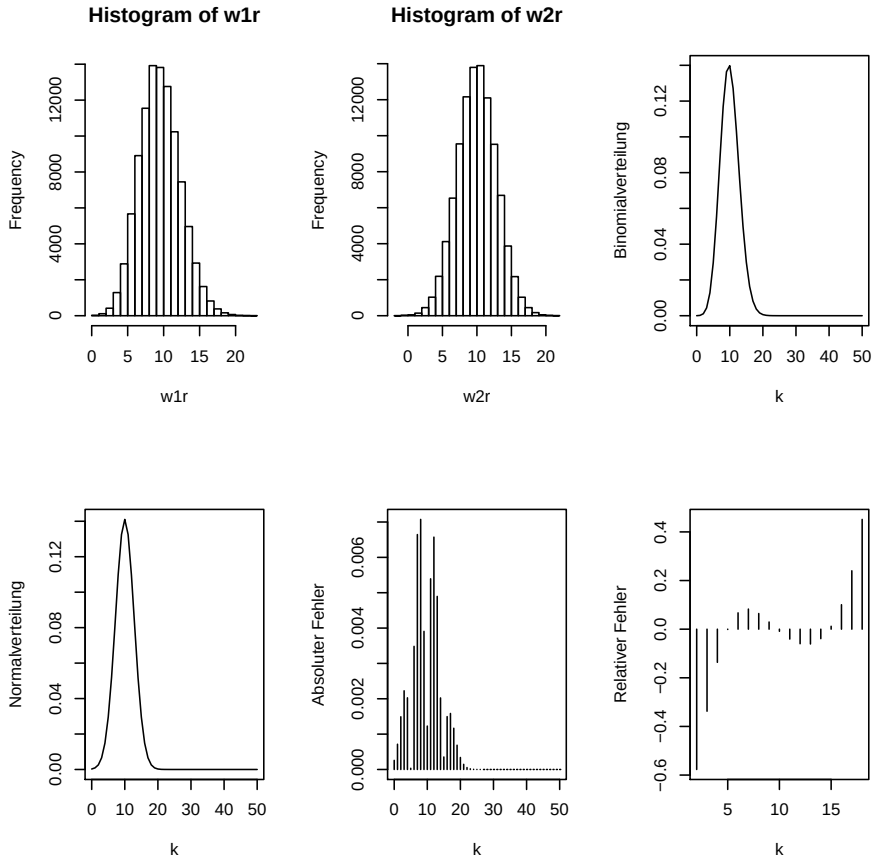

Abbildung 2.6: Illustration des Satzes von de Moivre-Laplace für $p = 0.2$ und $n = 50$.

10%. *In der Tag ergab eine statistische Analyse aller Datensätze Gregor Mendels, dass mit hoher Wahrscheinlichkeit die schon sehr schönen Ergebnisse noch etwas "geschönt" wurden. Man geht allerdings davon aus, dass damit keine unlautere Absicht verbunden war, sondern nur eine unpräzise Durchführung des Experimentes sichtbar werden. Es wird aber*

```
> demoivre(10^5,100,0.5,3)
```

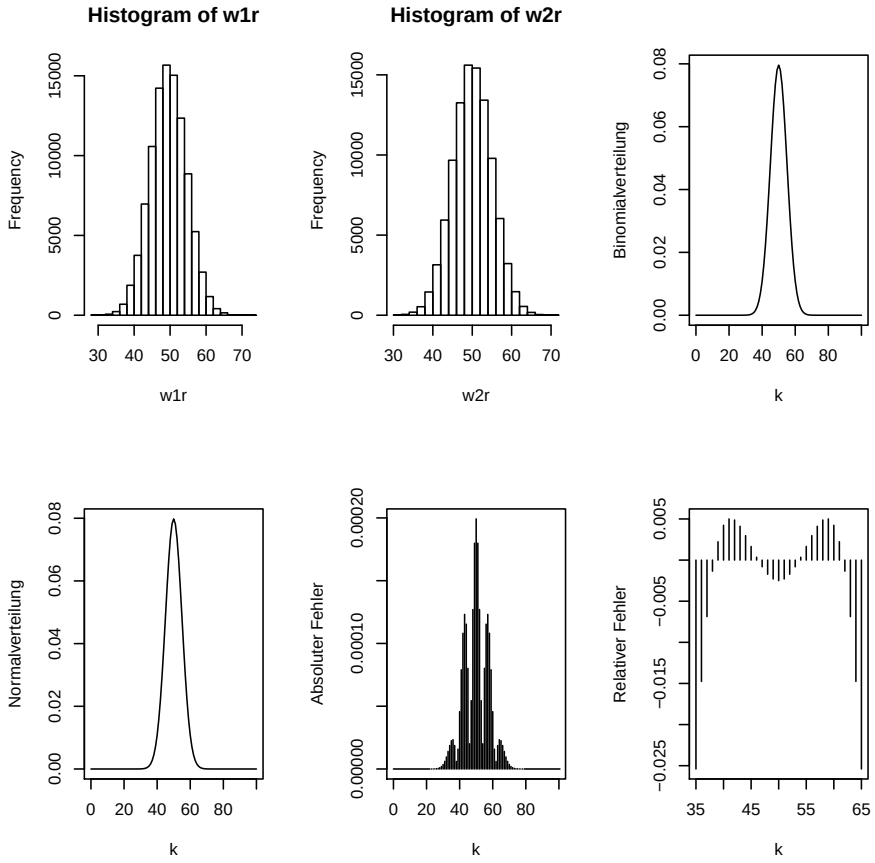

Abbildung 2.7: Illustration des Satzes von de Moivre-Laplace für $p = 0.5$ und $n = 100$.

auch ein weiteres Mal klar, dass jede willentliche Veränderung der Handschrift des Zufalls starke Spuren hinterlässt.

2.6.3 Zusammenhang Binomial- und Poissonverteilung

Die Approximation der Binomialverteilung im Satz von de Moivre-Laplace gilt für p fest und $n \to \infty$. Je näher p bei 0 oder 1 ist, desto schlechter ist sie. Hier besprechen wir die Approximation durch eine Poisson(λ)-Verteilung (siehe 2.11), welche gut ist für n gross und p klein, wobei $np = \lambda$. Das heisst, man hat viele 0-1-Experimente mit kleiner Erfolgswahrscheinlichkeit (Bsp. Radioaktiver Zerfall: viele Teilchen, kleine Zerfallswahrscheinlichkeit für jedes Teilchen; oder Schadenfälle bei einer Versicherung: viele Policen, kleine Schadenwahrscheinlichkeit für jede Police).

Satz 2.12. *Für k fest, $n \to \infty, p \to 0$ mit $np \to \lambda$ gilt:*

$$\binom{n}{k} p^k (1-p)^{n-k} \longrightarrow \frac{\lambda^k}{k!} e^{-\lambda}. \tag{2.76}$$

Beweis Der Beweis ergibt sich aus folgendem Limes:

$$\begin{aligned}\binom{n}{k}p^k(1-p)^{n-k} &= \frac{1}{k!}(np)^k\left(1-\frac{np}{n}\right)^n\frac{n(n-1)\dots(n-k+1)}{n^k}(1-p)^{-k}\\ &\longrightarrow \frac{1}{k!}\lambda^k e^{-\lambda}\cdot 1\cdot 1.\end{aligned}$$

In (2.17) hatten wir gesehen, dass der Erwartungswert der Poissonverteilung gleich λ ist, d.h. $np\to\lambda$ bedeutet, dass der Erwartungswert der Binomialverteilung gegen den Erwartungswert der Poissonverteilung konvergiert.

Beispiel 2.19. *Rutherford (1910) beobachtete die Zerfälle eines radioaktiven Präparats in $n = 2608$ Zeitintervallen von 7.5 Sekunden. n_k ist die Anzahl Intervalle mit genau k Zerfällen. Wenn die Anzahl Zerfälle in 7.5 Sekunden Poisson-verteilt ist, dann sollten nach der frequentistischen Interpretation der Wahrscheinlichkeit $n_k \approx np_k = ne^{-\lambda}\lambda^k/k!$ sein. Um dies zu überprüfen, müssen wir λ wählen. Wir ersetzen dazu den Erwartungswert durch das arithmetische Mittel*

$$\frac{\textit{Gesamtzahl Zerfälle}}{\textit{Anzahl Intervalle}} = \frac{\sum_{k=0}^{\infty} kn_k}{n} = \frac{10097}{2608} = 3.87$$

Die p_k's in der dritten Zeile der folgenden Tabelle sind mit $\lambda = 3.87$ berechnet.

k	=	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
n_k	=	57	203	383	525	532	408	273	139	45	27	10	4	2
np_k	=	54	210	407	525	508	394	254	141	68	29	11	4	1

Die Übereinstimmung scheint sehr gut zu sein.

Eine andere Begründung der Poissonverteilung beruht auf folgendem Satz.

Satz 2.13. *Seien X_1 und X_2 zwei unabhängige Zufallsvariable mit*

$$\mathbb{P}\left(X_1 = k | X_1 + X_2 = \ n\right) = \binom{n}{k}2^{-n}$$

für alle k, n mit $0 \leq k \leq n$. Dann sind X_1 und X_2 Poissonverteilt mit dem gleichen λ.

Um die Bedeutung dieses Satzes zu verstehen, seien X_1 die Anzahl Zerfälle in $[0, T]$ und X_2 die Anzahl Zerfälle in $[T, 2T]$. Die Voraussetzung in obigem Satz bedeutet dann, dass jeder der n Zerfälle in $[0, 2T]$ mit Wahrscheinlichkeit $\frac{1}{2}$ in $[0, T]$ geschah und mit Wahrscheinlichkeit $\frac{1}{2}$ in $[T, 2T]$, unabhängig von den andern Zerfällen.

Beweis

$$\begin{aligned}\frac{1}{n} &= \frac{\mathbb{P}\left(X_1 = n | X_1 + X_2 = n\right)}{\mathbb{P}\left(X_1 = n-1 | X_1 + X_2 = n\right)} = \frac{\mathbb{P}\left(X_1 = n, X_2 = 0\right)}{\mathbb{P}\left(X_1 = n-1, X_2 = 1\right)}\\ &= \frac{\mathbb{P}\left(X_1 = n\right)\cdot\mathbb{P}\left(X_2 = 0\right)}{\mathbb{P}\left(X_1 = n-1\right)\cdot\mathbb{P}\left(X_2 = 1\right)}\end{aligned}$$

Also ist mit $\lambda = \mathbb{P}\left(X_2 = 1\right)/\mathbb{P}\left(X_2 = 0\right)$:

$$\mathbb{P}\left(X_1 = n\right) = \frac{\lambda}{n}\mathbb{P}\left(X_1 = n-1\right) = \ldots = \frac{\lambda^n}{n!}\mathbb{P}\left(X_1 = 0\right).$$

Da sich die Wahrscheinlichkeiten zu eins addieren müssen, folgt $\mathbb{P}\left(X_1 = 0\right) = e^{-\lambda}$. Für X_2 geht es analog. □

```
> rutherford <- function(m,n,lambda)
+   {
+     w1r <- rbinom(m,n,lambda/n)
+     w2r <- rpois(m,lambda)
+     k=2*max(w1r)
+     w1 <- dbinom(0:k,n,lambda/n)
+     w2 <- dpois(0:k,lambda)
+     par(mfrow=c(2,2))
+     hist(w1r)
+     hist(w2r)
+     plot(0:k,w1,type="l",xlab="k",ylab="Binomialverteilung")
+     plot(0:k,w2,type="l",xlab="k",ylab="Poissonverteilung")
+   }
> rutherford(10^5,5000,3)
```

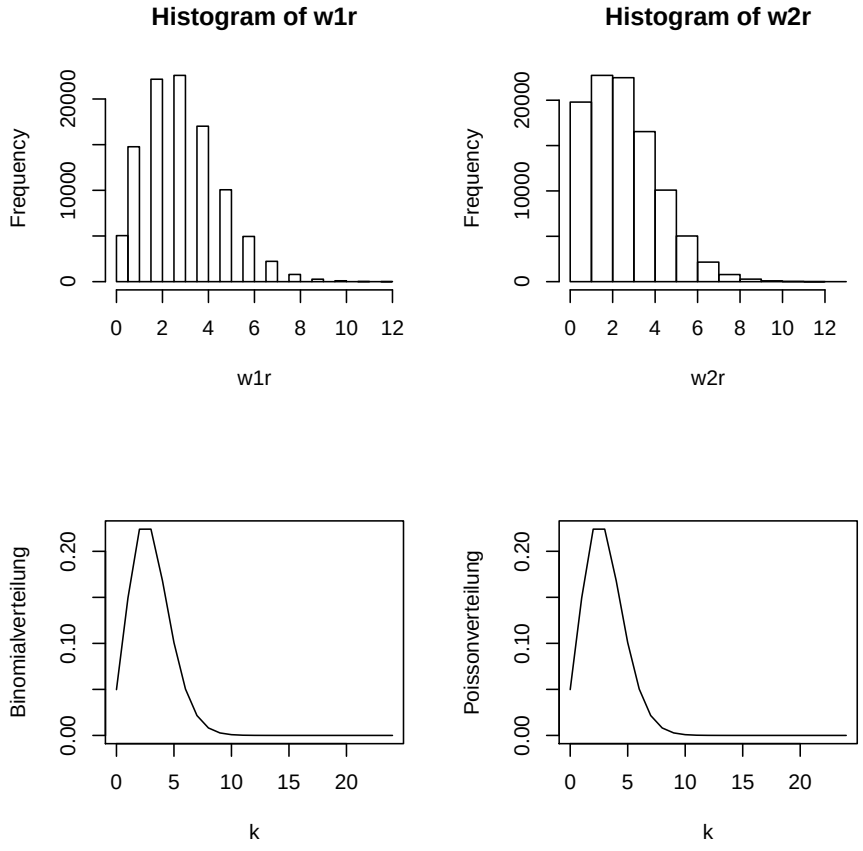

Abbildung 2.8: Illustration der Poissonapproximation für $\lambda = 3$ und $n = 500$.

Schliesslich beweisen wir noch, dass die Summe zweier unabhängiger Poissonverteilter Zufallsvariable wieder Poissonverteilt ist. In den obigen Beispielen ist die Wahl der Intervallänge also irrelevant.

Satz 2.14. *Wenn X_1 und X_2 unabhängig und Poisson(λ_i)-verteilt sind, dann ist $X = X_1 + X_2$ Poisson($\lambda_1 + \lambda_2$)-verteilt.*

Beweis Entweder man approximiert X_i durch 2 unabhängige binomialverteilte Zufallsvariable mit $p = \frac{1}{n}$ und $n_i = [\lambda_i n]$, für die die Additionseigenschaft offensichtlich ist, oder

man rechnet nach:

$$\begin{aligned}
\mathbb{P}(X=k) &= \sum_{j=0}^{k} \mathbb{P}(X_1=j, X_2=k-j) = \sum_{j=0}^{k} \mathbb{P}(X_1=j)\cdot\mathbb{P}(X_2=k-j) \\
&= \sum_{j=0}^{k} e^{-\lambda_1}\frac{\lambda_1^j}{j!}e^{-\lambda_2}\frac{\lambda_2^{k-j}}{(k-j)!} \\
&= e^{-(\lambda_1+\lambda_2)}\frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}\lambda_1^j\lambda_2^{k-j} \\
&= e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^k}{k!}\,.
\end{aligned}$$

□

Kapitel 3

Stetige Modelle

3.1 Allgemeine Wahrscheinlichkeitsräume

Der bisher benutzte Rahmen eines diskreten Wahrscheinlichkeitsraumes erweist sich für viele Situationen und Fragestellungen als zu eng. Viele Phänomene treten erst im Grenzübergang zu einem stetigen Modell deutlich hervor (z.B. beim Arkussinus-Gesetz oder beim zentralen Grenzwertsatz), und andere Fragen lassen sich überhaupt erst in einem überabzählbaren Modell exakt formulieren. Ein Beispiel dafür ist die Frage nach

$$\mathbb{P}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i\leq\frac{1}{2}\right)$$

bei sukzessiven unabhängigen 0-1-Experimenten $X_1, X_2, \ldots$.

Deshalb führen wir jetzt den allgemeinen Begriff des Wahrscheinlichkeitsraumes ein. Dazu braucht man Masstheorie. Wir werden diese nicht systematisch entwickeln, sondern begnügen uns mit Hinweisen, wo diese gebraucht wird. Im Anhang A sind die Hauptresultate der Masstheorie zusammengefasst.

3.1.1 Die Axiome von Kolmogorov

Sei $\Omega \neq \emptyset$ irgendeine nichtleere Menge, $\mathcal{A}$ eine Kollektion von Teilmengen $A \subseteq \Omega$ und $\mathbb{P} : \mathcal{A} \to [0,1]$ eine Abbildung von $\mathcal{A}$ in das Einheitsintervall.

Die Elemente $\omega \in \Omega$ interpretieren wir als die (im Modell in Betracht gezogenen) **möglichen Fälle**, die Teilmengen $A \in \mathcal{A}$ als die (im Modell zugelassenen) **Ereignisse**, und für $A \in \mathcal{A}$ interpretieren wir die Zahl $\mathbb{P}(A)$ als die (im Modell angenommene) **Wahrscheinlichkeit des Ereignisses** A.

Die folgenden **Axiome von Kolmogorov** (1933) verlangen nun, dass die Kollektion $\mathcal{A}$ der Ereignisse abgeschlossen ist unter abzählbaren Mengenoperationen, und dass die Zuordnung $A \to \mathbb{P}(A)$ "konsistent" ist im Sinne der Rechenregeln für Wahrscheinlichkeiten, die uns von den diskreten Modellen her schon vertraut sind.

Definition 3.1. *Das Tripel* $(\Omega, \mathcal{A}, \mathbb{P})$ *heisst ein* **Wahrscheinlichkeitsraum**, *wenn gilt:*

1) $\mathcal{A}$ ist eine **σ-Algebra**, *d.h.*

$$\Omega \in \mathcal{A} \tag{3.1}$$
$$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A} \tag{3.2}$$
$$A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \bigcup_i A_i \in \mathcal{A} \tag{3.3}$$

2) $\mathbb{P}$ ist eine **Wahrscheinlichkeitsverteilung**, *d.h.*

$$\mathbb{P}(\Omega) = 1 \tag{3.4}$$

$$A_1, A_2, \cdots \in \mathcal{A},\ A_i \cap A_j = \emptyset\ (i \neq j) \Rightarrow \mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i) \tag{3.5}$$

In der Sprache der Masstheorie besagt 2), dass $\mathbb{P}$ ein **normiertes Mass** ist, d.h. eine Mengenfunktion $\mathbb{P} : \mathcal{A} \to \mathbb{R}^+$, die im Sinne von (3.5) **σ-additiv** ist und im Sinne von (3.4) **normiert**.

Beispiel 3.1. *Natürlich ist jeder* diskrete *Wahrscheinlichkeitsraum im Sinne von Kap. I ein Wahrscheinlichkeitsraum im Sinne der Definition 3.1; vgl. (2.1). Umgekehrt gilt: Ist Ω abzählbar und $\mathcal{A} = \{A | A \subseteq \Omega\}$, so ist jede Wahrscheinlichkeitsverteilung auf $\mathcal{A}$ von der Form (2.3). Es gilt nämlich*

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{\omega \in A} \{\omega\}\right) \overset{(3.5)}{=} \sum_{\omega \in A} \mathbb{P}(\{\omega\}),$$

und wegen (3.4) ist $p(\omega) \equiv \mathbb{P}(\{\omega\})$, $(\omega \in \Omega)$, eine Gewichtung im Sinne von (2.1).

Beispiel 3.2. Zufällige Wahl einer natürlichen Zahl. *Auf die Frage, mit welcher Wahrscheinlichkeit man bei zufälliger Wahl einer natürlichen Zahl eine gerade Zahl erhält, ist man versucht, die Antwort $\frac{1}{2}$ zu geben. Allgemeiner würde man dann für eine Menge $A \subseteq \Omega = \{1, 2, \dots\}$ den Ansatz*

$$\mathbb{P}(A) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} I_A(n)$$

(relative Häufigkeit von A in Ω) machen, sofern A zur Klasse $\mathcal{A}_0$ derjenigen Mengen gehört, für die dieser Limes existiert. Es gibt aber keine Wahrscheinlichkeitsverteilung im Sinne von Definition 3.1, die mit diesem Ansatz verträglich wäre! Für jedes n ist nämlich $\{n\} \in \mathcal{A}_0$ mit $\mathbb{P}(\{n\}) = 0$, aus (3.5) würde also für jede Menge $A \subseteq \Omega$

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{n \in A} \{n\}\right) \overset{(3.5)}{=} \sum_{n \in A} \mathbb{P}(\{n\}) = 0$$

folgen. Es geht nur dann, wenn man bereit ist, die σ-Additivität (3.5) durch die einfache Additivität (vgl. (3.8)) zu ersetzen, aber eine solche "finite" Wahrscheinlichkeitstheorie ist in ihren technischen Möglichkeiten sehr viel begrenzter.

Beispiel 3.3. Gleichverteilung auf dem Einheitsintervall *Sei*

$$\Omega = [0, 1],$$

und sei $\mathcal{A}$ die kleinste σ-Algebra, welche alle Intervalle $[a, b] \subseteq [0, 1]$ enthält (vgl. 3.7). Dann gibt es genau eine Wahrscheinlichkeitverteilung $\mathbb{P} : \mathcal{A} \to [0, 1]$, bei der die Wahrscheinlichkeit eines Intervalls dessen Länge ist:

$$\mathbb{P}([a, b]) = b - a, \tag{3.6}$$

nämlich das auf $[0,1]$ *eingeschränkte* **Lebesguemass** *(→ Masstheorie). Man kann* $\mathcal{A}$ *zwar noch etwas erweitern zur Klasse der "Lebesgue-messbaren" Mengen, aber es ist nicht klar, ob man bis zur vollen Potenzmenge gehen kann. Mit dem Auswahlaxiom kann man relativ leicht zeigen, dass es keine Wahrscheinlichkeitsverteilung auf der* σ*-Algebra* aller *Teilmengen* $A \subseteq [0,1]$ *gibt, die translationsinvariant ist (d.h.* $\mathbb{P}(\tau_x(A)) = \mathbb{P}(A)$ *für alle* A *und alle* x*, wobei* $\tau_x(y) = y + x \pmod 1$*).*

Für jede feste Zahl $\omega \in [0,1]$ *folgt aus (3.6)* $\mathbb{P}(\{\omega\}) = 0$*. Anders als im vorigen Beispiel 3.2 kann man aber* nicht *für beliebige Teilmengen* $A \in \mathcal{A}$

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{\omega\in A}\{\omega\}\right) = \sum_{\omega\in A}\mathbb{P}(\{\omega\}) = 0$$

schliessen: nach (3.5) ist das nur für abzählbare *Mengen* A *zulässig. Zum Beispiel ist* $\mathbb{P}(A) = 0$ *für* $A =$ *"alle rationalen Zahlen". Bei der Simulation der Gleichverteilung auf einem Taschenrechner ("Zufallszahlen") sieht das natürlich anders aus!*

3.1.2 Einfache Folgerungen

Aus (3.2) und (3.3) folgt zunächst, dass eine σ-Algebra $\mathcal{A}$ abgeschlossen ist gegen *abzählbare* Mengenoperationen. Zum Beispiel:

$$A_1, A_2, \dots \in \mathcal{A} \Rightarrow \bigcap_i A_i = \left(\bigcup_i A_i^c\right)^c \in \mathcal{A},$$

oder das Ereignis $A_\infty =$ "Unendlich viele der Ereignisse A_i treten ein" gehört zu $\mathcal{A}$, denn

$$A_\infty = \bigcap_n \bigcup_{k\geq n} A_k.$$

Bemerkung zur Wahl der σ-Algebra $\mathcal{A}$: Wir haben in Beispiel 3.3 gesehen, dass man als σ-Algebra $\mathcal{A}$ der zugelassenen Ereignisse nicht immer die Klasse *aller* Teilmengen von Ω wählen *kann*. Oft *will* man es auch gar nicht, weil man sich nur für eine gewisse Teilklasse von Ereignissen interessiert (vgl. die Diskussion der Spielsysteme im Abschnitt 1.3.4).

In der Regel ist es so – wie z. B. in Beispiel 3.2 –, dass man eine gewisse Klasse $\mathcal{A}_0$ von Teilmengen im Auge hat, die jedenfalls als Ereignisse zugelassen sein sollten, und dass man dann übergeht zu der **"von $\mathcal{A}_0$ erzeugten σ-Algebra"**:

$$\mathcal{A} = \sigma(\mathcal{A}_0) = \bigcap_{\substack{\mathcal{B}\supseteq\mathcal{A}_0,\\ \mathcal{B}\text{ ist }\sigma\text{−Algebra}}} \mathcal{B}. \tag{3.7}$$

Dies *ist* eine σ-Algebra (Übung), und zwar offensichtlich die **kleinste σ-Algebra, welche $\mathcal{A}_0$ enthält**.

Eine Mengenfunktion $\mathbb{P} : \mathcal{A} \to [0,1]$ mit

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i) \tag{3.8}$$

für je *endlich* viele paarweise disjunkte Ereignisse $A_1, \dots, A_n \in \mathcal{A}$ nennt man auch **additiv**. Aus der Additivität und der Normierung (3.4) ergeben sich die üblichen elementaren

Rechenregeln für Wahrscheinlichkeiten wie im diskreten Fall

$$\begin{aligned} \mathbb{P}(A^c) &= 1-\mathbb{P}(A)\,, \quad \mathbb{P}(A\cup B)=\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cap B)\,, \\ \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n}(-1)^{k+1}\sum_{1\le i_1<\cdots<i_k\le n}\mathbb{P}(A_{i_1}\cap\cdots\cap A_{i_k})\,, \text{ usw.} \end{aligned}$$

Durch die **σ-Additivität** (3.5), bei der auch **abzählbar** viele paarweise disjunkte Ereignisse $A_1, A_2, \dots$ zugelassen sind, kommen zusätzliche **Stetigkeitseigenschaften** ins Spiel. Genauer gilt:

Satz 3.1. *Ist $\mathbb{P}:\mathcal{A}\to[0,1]$ additiv, so sind die folgenden Aussagen äquivalent:*

$$\mathbb{P} \textit{ ist } \sigma\textit{-additiv}, \tag{3.9}$$

$$A_1\subseteq A_2\subseteq\dots\,, \quad A_n\in\mathcal{A}\Rightarrow \mathbb{P}\left(\bigcup_n A_n\right)=\lim_n \mathbb{P}(A_n)\,, \tag{3.10}$$

$$A_1\supseteq A_2\supseteq\dots\,, \quad A_n\in\mathcal{A}\Rightarrow \mathbb{P}\left(\bigcap_n A_n\right)=\lim_n \mathbb{P}(A_n)\,. \tag{3.11}$$

Beweis Für $A_1\subseteq A_2\subseteq\dots$ setzen wir $B_1\equiv A_1, B_n\equiv A_n-A_{n-1}$ $(n\ge 2)$. Diese B_n $(n=1,2,\dots)$ sind paarweise disjunkt, es gilt $\bigcup_n B_n=\bigcup_n A_n$, und aus der Additivität folgt

$$\lim_n \mathbb{P}(A_n)=\lim_n \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right)=\lim_n\sum_{i=1}^{n}\mathbb{P}(B_i)=\sum_{i=1}^{\infty}\mathbb{P}(B_i)\,.$$

Daraus ergibt sich die Äquivalenz von (3.9) und (3.10). Die Äquivalenz von (3.10) und (3.11) folgt durch Komplementbildung. □

Korollar 3.1. *Für beliebige $A_1, A_2, \dots\in\mathcal{A}$ gilt*

$$\mathbb{P}\left(\bigcup_k A_k\right)\le\sum_k\mathbb{P}(A_k)\,.$$

Beweis

$$\mathbb{P}\left(\bigcup_k A_k\right)\overset{(3.10)}{=}\lim_n\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right)\le\lim_n\sum_{k=1}^{n}\mathbb{P}(A_k)=\sum_k\mathbb{P}(A_k)\,.$$

□

Die **Unabhängigkeit von Ereignissen** ist definiert wie in (2.60).

Lemma 3.1 (Borel-Cantelli)**.** *Sei $A_1, A_2, \dots\in\mathcal{A}$ eine Folge von Ereignissen, und sei*

$$A_\infty=\bigcap_n\left(\bigcup_{k\ge n} A_k\right) = \textit{“unendlich viele der } A_k \textit{ treten ein”}.$$

1) *Aus $\sum_k\mathbb{P}(A_k)<\infty$ folgt stets $\mathbb{P}(A_\infty)=0$.*
2) *Sind die Ereignisse $A_1, A_2, \dots$* unabhängig, *so folgt aus $\sum_k\mathbb{P}(A_k)=\infty$ umgekehrt auch $\mathbb{P}(A_\infty)=1$.*

Beweis Die Folge $(\bigcup_{k\ge n} A_k)_n$ ist absteigend. Also folgt aus (3.11) und Korollar 3.1:

$$\mathbb{P}(A_\infty)=\lim_n\mathbb{P}\left(\bigcup_{k\ge n} A_k\right)\le\lim_n\sum_{k\ge n}\mathbb{P}(A_k)=0.$$

Für die zweite Aussage gehen wir aus von

$$\mathbb{P}\left(\bigcap_{k\geq n} A_k^c\right) \overset{(3.11)}{=} \lim_m \mathbb{P}\left(\bigcap_{k=n}^{m} A_k^c\right) \overset{(2.60)}{=} \prod_{k\geq n} \mathbb{P}\left(A_k^c\right) = \prod_{k\geq n} (1-\mathbb{P}\left(A_k\right)).$$

Weil $1-x \leq e^{-x}$ gilt, folgt also für jedes feste n

$$\mathbb{P}\left(\bigcap_{k\geq n} A_k^c\right) \leq \exp(-\sum_{k\geq n} \mathbb{P}\left(A_k\right)) = 0.$$

Da die Folge $(\bigcap_{k\geq n} A_k^c)_n$ aufsteigend ist, folgt mit (3.10)

$$\mathbb{P}\left(A_\infty^c\right) = \mathbb{P}\left(\bigcup_n (\bigcap_{k\geq n} A_k^c)\right) = \lim_n \mathbb{P}\left(\bigcap_{k\geq n} A_k^c\right) = 0.$$

□

3.1.3 Sukzessive unabhängige 0-1-Experimente

Sei

$$\Omega = \{\omega = (x_1, x_2, \dots) | x_i \in \{0,1\}\}$$

die (überabzählbare!) Menge aller 0-1-Folgen, X_i die durch $X_i(\omega) = x_i$ definierte Zufallsvariable und $\mathcal{A}$ die von den Ereignissen $\{\omega | X_i(\omega) = 1\}$ $(i = 1, 2, \dots)$ erzeugte σ-Algebra (vgl. (3.7)). Ferner sei $0 \leq p \leq 1$ ein **"Erfolgsparameter".**

Satz 3.2. *Es gibt genau eine Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf $(\Omega, \mathcal{A})$ derart, dass gilt:*

$$\mathbb{P}\left(X_i = 1\right) = p \quad (i = 1, 2, \dots) \tag{3.12}$$

Die Ereignisse $\{X_i = 1\}$ $(i = 1, 2, \dots)$ sind unabhängig bezüglich $\mathbb{P}$. (3.13)

Zum Beweis: Zunächst folgt für jedes $n \geq 1$ und für jede Wahl von $x_i \in \{0,1\}$ aus (3.12) und (3.13), dass mit $k = \sum_{i=1}^n x_i$ gelten muss

$$\begin{aligned} \mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) &\overset{(3.13)}{=} \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right) \\ &\overset{(3.12)}{=} p^k (1-p)^{n-k}. \end{aligned} \tag{3.14}$$

(Bei der ersten Gleichung wurde noch Lemma 2.3 verwendet).

Damit ist $\mathbb{P}$ festgelegt auf der Kollektion aller Ereignisse, die sich als endliche Vereinigung von Ereignissen der Form $\{X_1 = x_1, \dots, X_n = x_n\}$ darstellen lassen $(= \bigcup_{n\geq 0} \mathcal{A}_n$ in der Notation von Definition 2.2). Für die Erweiterung von $\mathbb{P}$ auf die volle σ-Algebra $\mathcal{A}$ benötigt man nun einen **Fortsetzungssatz der Masstheorie**. □

Für $p = \frac{1}{2}$ folgt die Existenz von $\mathbb{P}$ auch aus der Existenz des Lebesguemasses auf $[0,1]$; vgl. Beispiel 3.4.

Für $p = 1$ (bzw. $= 0$) ist die Konstruktion von $\mathbb{P}$ natürlich sehr einfach:

$$\mathbb{P}\left(A\right) \equiv \begin{cases} 1 & \text{falls } (1,1,1,\dots) \in A \\ 0 & \text{sonst.} \end{cases}$$

Sei nun $0 < p < 1$. Für jedes $\omega = (x_1, x_2, \dots) \in \Omega$ ist dann

$$\mathbb{P}(\{\omega\}) = 0 \tag{3.15}$$

denn mit $k(n) = \sum_{i=1}^{n} x_i$ gilt

$$\mathbb{P}(\{\omega\}) = \mathbb{P}\left(\bigcap_{n=1}^{\infty}\{X_n = x_n\}\right) \overset{(3.11)}{=} \lim_{n\uparrow\infty} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$
$$\overset{(3.14)}{=} \lim_{n\uparrow\infty} p^{k(n)}(1-p)^{n-k(n)} = 0.$$

Satz 3.3. *Sei $[x_1, \dots, x_N]$ ein "binärer Text" mit $x_i \in \{0, 1\}$. Dann ist die Wahrscheinlichkeit, dass irgendwann dieser Text erscheint (und dies nicht nur einmal, sondern sogar unendlich oft !) gleich eins.*

Beweis Wir betrachten die Ereignisse

$$A_k \equiv \{X_{(k-1)N+1} = x_1, \dots, X_{kN} = x_N\} \quad (k = 1, 2, \dots).$$

Diese sind unabhängig und haben alle dieselbe Wahrscheinlichkeit $\mathbb{P}(A_k) > 0$. Mit dem zweiten Teil des Lemmas von Borel-Cantelli folgt also die Behauptung. □

3.1.4 Transformation von Wahrscheinlichkeitsräumen

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $\widetilde{\Omega} \neq \emptyset$ und $\widetilde{\mathcal{A}}$ eine σ-Algebra von Teilmengen $A \subseteq \widetilde{\Omega}$.

Definition 3.2. *Eine Abbildung $\phi : \Omega \to \widetilde{\Omega}$ heisst* **messbar** *(bezüglich $\mathcal{A}$ und $\widetilde{\mathcal{A}}$), wenn gilt:*

$$A \in \widetilde{\mathcal{A}} \Rightarrow \phi^{-1}(A) = \{\omega | \phi(\omega) \in A\} \in \mathcal{A}. \tag{3.16}$$

Bemerkung Wird $\widetilde{\mathcal{A}}$ von irgendeinem Mengensystem $\widetilde{\mathcal{A}}_0$ erzeugt, gilt also $\widetilde{\mathcal{A}} = \sigma(\widetilde{\mathcal{A}}_0)$ im Sinne von (3.7), so genügt es, statt (3.16) zunächst nur

$$A \in \widetilde{\mathcal{A}}_0 \Rightarrow \phi^{-1}(A) \in \mathcal{A} \tag{3.17}$$

zu fordern. Denn

$$\{A \subseteq \widetilde{\Omega} | \phi^{-1}(A) \in \mathcal{A}\}$$

ist eine σ-Algebra, nach (3.17) umfasst sie $\widetilde{\mathcal{A}}_0$, damit aber auch $\widetilde{\mathcal{A}} = \sigma(\widetilde{\mathcal{A}}_0)$, und das ist gleichbedeutend mit (3.16).

Satz 3.4. *Ist $\phi : \Omega \to \widetilde{\Omega}$ messbar, so ist durch*

$$\widetilde{\mathbb{P}}(A) \equiv \mathbb{P}\left(\phi^{-1}(A)\right) \quad (A \in \widetilde{\mathcal{A}}) \tag{3.18}$$

eine Wahrscheinlichkeitsverteilung $\widetilde{\mathbb{P}}$ auf $(\widetilde{\Omega}, \widetilde{\mathcal{A}})$ definiert. $\widetilde{\mathbb{P}}$ heisst das **Bild von** $\mathbb{P}$ *unter ϕ bzw. die* **Verteilung von** ϕ *unter* $\mathbb{P}$.

Beweis Zunächst gilt offensichtlich

$$\widetilde{\mathbb{P}}(\widetilde{\Omega}) = \mathbb{P}\left(\phi^{-1}(\widetilde{\Omega})\right) = \mathbb{P}(\Omega) = 1.$$

Wir müssen also noch die σ-Additivität beweisen. Seien also $A_1, A_2, \dots$ paarweise disjunkte Mengen in $\widetilde{\mathcal{A}}$.. Die Ereignisse $B_n \equiv \phi^{-1}(A_n)$ $(n = 1, 2, \dots)$ gehören dann nach Voraussetzung zu $\mathcal{A}$ und sind offensichtlich paarweise disjunkt. Also

$$\widetilde{\mathbb{P}}\left(\bigcup_i A_i\right) = \mathbb{P}\left(\phi^{-1}\left(\bigcup_i A_i\right)\right) = \mathbb{P}\left(\bigcup_i \phi^{-1}(A_i)\right) \overset{(3.5)}{=} \sum_i \mathbb{P}\left(\phi^{-1}(A_i)\right) = \sum_i \widetilde{P}(A_i).$$

□

Beispiel 3.4. *Wie in Beispiel (3.3) sei* $\Omega = [0,1]$, $\mathcal{A} = \sigma(\{[a,b] \mid 0 \le a \le b \le 1\})$ *und* $\mathbb{P}$ *die* Gleichverteilung auf $[0,1]$, *also das auf* $[0,1]$ *eingeschränkte Lebesguemass. Ferner sei wie im vorigen Abschnitt* $\widetilde{\Omega}$ *die Menge aller 0-1 Folgen und* $\widetilde{\mathcal{A}}$ *die von den Ereignissen* $\{X_i = 0\}$ $(i = 1, 2, \dots)$ *erzeugte* σ*-Algebra.*

Wir definieren nun die Abbildung $\phi : \Omega \longrightarrow \widetilde{\Omega}$ *durch die binäre Darstellung der Zahlen im Einheitsintervall:*

$$\phi(\omega) = (\phi_1(\omega),\ \phi_2(\omega), \dots) \quad (\omega \in [0,1]).$$

Das heisst, wir setzen für $n = 0, 1, \dots$ $\phi_{n+1}(\omega) = 0$ *genau dann, wenn* ω *in einem Intervall* $[k2^{-n}, k2^{-n} + 2^{-(n+1)})$ *liegt für ein* $k \in \{0, \dots, 2^n - 1\}$. *Wenn die binäre Darstellung nicht eindeutig ist, nehmen wir also die Version, die mit lauter Nullen endet.*

Diese Abbildung ist messbar, denn für $n = 0, 1, \dots$ *ist*

$$\phi^{-1}(\{X_{n+1} = 0\}) = \{\phi_{n+1} = 0\} = \bigcup_{k=0}^{2^n-1} [k2^{-n}, k2^{-n} + 2^{-(n+1)}) \in \mathcal{A}$$

(wir benutzen das Resultat (3.17)!).

Wir setzen $\widetilde{\mathbb{P}}$ *gleich dem Bild der Gleichverteilung unter* ϕ. *Dann ist*

$$\widetilde{\mathbb{P}}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(\underbrace{\phi_1 = x_1, \dots, \phi_n = x_n}_{\text{Intervall der Länge } 2^{-n}}) = 2^{-n},$$

und daraus folgt, dass $\widetilde{\mathbb{P}}$ *die Eigenschaften (3.12) und (3.13) mit* $p = \frac{1}{2}$ *hat.*

Aus der Existenz des Lebesguemasses (→ Masstheorie) folgt also die Existenz eines exakten Modells für unendlich viele Würfe einer fairen Münze *(und umgekehrt: vgl. Beispiel 3.9 unten).*

3.2 Zufallsvariable und ihre Verteilung

Sei $\mathcal{B}$ die Borelsche σ-Algebra auf $\mathbb{R}$, das heisst die von allen Intervallen der Form $(-\infty, b]$ mit $b \in \mathbb{R}$ erzeugte σ-Algebra. Man kann zeigen, dass $\mathcal{B}$ alle Intervalle, alle offenen Mengen und alle abgeschlossenen Mengen enthält.

Definition 3.3. *Sei* $(\Omega, \mathcal{A}, \mathbb{P})$ *ein Wahrscheinlichkeitsraum. Eine* **Zufallsvariable** *ist eine messbare Abbildung*

$$X : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}, \mathcal{B}).$$

Die **Verteilung** μ *von* X *ist das Bild von* $\mathbb{P}$ *unter* X, *d.h. für jedes* $A \in \mathcal{B}$

$$\mu(A) = \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}\left(\{\omega | X(\omega) \in A\}\right) = \mathbb{P}\left(X \in A\right).$$

Wenn wir nur an Ereignissen interessiert sind, welche die Zufallsvariable X betreffen, dann können wir den Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ vergessen und mit $(\mathbb{R}, \mathcal{B}, \mu)$ weiterarbeiten.

Beispiel 3.5. *Sei* $(\Omega, \mathcal{A}, \mathbb{P})$ = *das Modell für unendlich viele* $0-1$ *Experimente wie in Abschnitt 3.1.3. Sei* X *die Zufallsvariable "Wartezeit auf die erste Eins" d.h.*

$$X(\omega) = \min\{k \ge 1 | x_k = 1\}$$

wobei wir $\min\emptyset := \infty$ *setzen. Dann ist für* $k = 1, 2, \ldots$

$$\mu(\{k\}) = \mathbb{P}(X(\omega) = k) = \mathbb{P}(\{x_1 = x_2 = \cdots = x_{k-1} = 0, x_k = 1\}) = (1-p)^{k-1}p,$$

und daher für beliebiges $A \in \mathcal{B}$

$$\mu(A) = p\sum_{k\in A}(1-p)^{k-1}.$$

Die Verteilung von $Y = X - 1$ *heisst die* **geometrische Verteilung**.

In diesem Beispiel tritt ∞ als möglicher Wert auf, und es ist von Vorteil, die Theorie so zu verallgemeinern, dass das auch zugelassen ist, vgl. Anhang A.6.

3.2.1 Verteilungsfunktion

Definition 3.4. *Die durch*

$$F(b) = \mathbb{P}(X \le b) = \mu((-\infty, b]) \quad (b \in \mathbb{R})$$

definierte Funktion heisst **Verteilungsfunktion** *von* X *bzw. von* μ.

Aus den bekannten Rechenregeln folgt

$$\begin{aligned}
\mu((a,b]) &= F(b) - F(a) \quad (a < b)\\
\mu(\{a\}) &= \mu\left(\bigcap_{n=1}^{\infty}(a - \frac{1}{n}, a]\right) \overset{(3.11)}{=} \lim_n \mu\left((a - \frac{1}{n}, a]\right)\\
&= F(a) - F(a-) \qquad (= \text{ Sprunghöhe in } a, \text{ cf. (ii) unten}).
\end{aligned}$$

Es gilt sogar ($\to$ Masstheorie), dass man aus F die Verteilung μ, d.h. $\mu(A)$ für alle $A \in \mathcal{B}$, erhalten kann.

Satz 3.5. *Jede Verteilungsfunktion hat die folgenden Eigenschaften*

i) Monotonie: $a \le b \Rightarrow F(a) \le F(b)$

ii) Rechtsstetigkeit: $F(a) = \lim_{h\downarrow 0} F(a+h)$

iii) Normierung: $\lim_{a\to-\infty} F(a) = 0, \lim_{a\to+\infty} F(a) = 1.$

Umgekehrt ist jede Funktion mit diesen 3 Eigenschaften Verteilungsfunktion einer Zufallsvariablen.

Beweis Für $a \le b$ ist $(-\infty, a] \subseteq (-\infty, b]$, also $\mu((-\infty, a]) \le \mu((-\infty, b])$ und damit $F(a) \le F(b)$, d.h. i).

Für $h_n \downarrow 0$ ist $\bigcap_n (a, a+h_n] = \emptyset$, also

$$0 = \mu\left(\bigcap_n (a, a+h_n]\right) \overset{(3.11)}{=} \lim_{n\uparrow\infty} \mu((a, a+h_n]) = \lim_{n\uparrow\infty}(F(a+h_n) - F(a)),$$

d.h. ii). Der Beweis von iii) geht analog wie bei ii).

Für die Umkehrung definieren wir für $0 < t < 1$

$$F^{-1}(t) = \inf\{x | F(x) \ge t\}. \tag{3.19}$$

Gemäss Lemma 3.2 unten gilt

$$F^{-1}(t) \leq x \iff t \leq F(x). \tag{3.20}$$

Wir wählen nun als $(\Omega, \mathcal{A}, \mathbb{P})$ die Gleichverteilung auf $[0,1]$ wie im Beispiel 3.3 und setzen

$$X(\omega) = F^{-1}(\omega).$$

Dann ist

$$\mathbb{P}(X \leq b) = \mathbb{P}(\{\omega | X(\omega) \leq b\}) = \mathbb{P}\left(\{\omega | F^{-1}(\omega) \leq b\}\right) \overset{(3.20)}{=} \mathbb{P}(\{\omega | \omega \leq F(b)\}) = F(b),$$

d.h. F ist die Verteilungsfunktion von X. □

Lemma 3.2. *Wenn F i) - iii) von Satz 3.5 erfüllt und F^{-1} wie in (3.19) definiert ist, dann ist F^{-1} monoton wachsend, linksstetig und es gilt*

i) $F^{-1}(F(x)) \leq x \quad (-\infty < x < \infty)$

ii) $t \leq F(F^{-1}(t)) \quad (0 < t < 1)$.

Beweis Wegen der Voraussetzungen i) und ii) ist

$$\{x | F(x) \geq t\} = [F^{-1}(t), \infty).$$

Damit ist die Monotonie von F^{-1} und $F(F^{-1}(t)) \geq t$ klar. Für $h_n \downarrow 0$ ist

$$\bigcap_n \{x | F(x) \geq t - h_n\} = \{x | F(x) \geq t\},$$

also ist F^{-1} linksstetig. $F^{-1}(F(x)) \leq x$ folgt schliesslich aus $x \in \{x' | F(x') \geq F(x)\}$. □

Die hier durchgeführte Konstruktion einer Zufallsvariable mit vorgegebener Verteilungsfunktion ist auch von praktischer Bedeutung, nämlich bei der Erzeugung von Zufallszahlen mit beliebiger Verteilung aus der Gleichverteilung.

Gemäss der Definition, bzw. dem Lemma gilt

$$\mathbb{P}\left(X < F^{-1}(t)\right) \leq t \leq \mathbb{P}\left(X \leq F^{-1}(t)\right).$$

Man nennt daher die Grösse $F^{-1}(t)$ auch das **t-Quantil** von X. Wichtig ist insbesondere das 50%-Quantil, der sogenannte **Median**.

3.2.2 Typen von Verteilungen

Die beiden wichtigsten Typen sind die **diskreten** und die **absolut stetigen** Verteilungen

Eine Zufallsvariable X heisst **diskret**, wenn es eine abzählbare Menge $A \subset \mathbb{R}$ gibt sodass $\mathbb{P}X \in A\,(=)\,1$. Dann ist die Verteilungsfunktion

$$F(b) = \sum_{x \in A; x \leq b} \mathbb{P}(X = x) = \sum_{x \in A; x \leq b} \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\})$$

eine Treppenfunktion mit Sprungstellen in A und Sprunghöhen $\mathbb{P}(X = x)$.

Eine Zufallsvariable heisst **absolut stetig**, falls eine messbare Funktion $f : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ existiert mit $f(x) \geq 0$ und $\int_{-\infty}^{\infty} f(x)dx = 1$, so dass

$$F(b) = \int_{-\infty}^{b} f(x)dx. \tag{3.21}$$

Die Funktion f heisst die **Dichte** von X. Jede Funktion der Form (3.21) erfüllt offensichtlich i) und iii) von Satz 3.5 und ist stetig. Letzteres ist klar für stückweise stetiges f, allgemein folgt es aus dem Konvergenzsatz von Lebesgue ($\to$ Masstheorie). Falls f stetig an der Stelle x ist, dann $f(x) = F'(x)$. Der Satz von Lebesgue ($\to$ Masstheorie) besagt, dass F ohne zusätzliche Voraussetzungen sogar fast überall differenzierbar ist mit Ableitung f.

Beispiel 3.6. Uniform auf $[a, b]$**:** *Bezeichnung* $\mathcal{U}(a, b)$.

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \textit{sonst.} \end{cases}$$

$$F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x \ge b. \end{cases}$$

Die uniforme Verteilung wird z.B. für Rundungsfehler verwendet.

Beispiel 3.7. Exponential mit Parameter $\alpha > 0$**:** *Bezeichnung:* $\mathrm{Exp}\,(\alpha)$.

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & x \ge 0 \\ 0 & x < 0. \end{cases}$$

$$F(x) = \begin{cases} 1 - e^{-\alpha x} & x \ge 0 \\ 0 & x < 0. \end{cases}$$

Die Exponentialverteilung wird für Warte- oder Überlebenszeiten verwendet. Verteilungsfunktion und Dichte sind in Abbildung 3.1 dargestellt.

```
>  z<-seq(0,10,by=0.01)
> plot(z,pexp(z),type="l")
> lines(z,dexp(z),col="red")
```

Beispiel 3.8. Normal mit Parametern μ, σ^2**:** *Bezeichnung:* $\mathcal{N}(\mu, \sigma^2)$.

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

(Dass $\frac{1}{\sqrt{2\pi}}$ *die richtige Normierung ist, folgt aus der Analysis.) Hier ist* $F(x)$ *nicht in geschlossener Form darstellbar. Es gilt aber*

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = F_{0,1}(\frac{x-\mu}{\sigma}).$$

Die Dichte $f_{0,1}$ *heisst die Standardnormalverteilungsdichte und wird meist mit* φ *bezeichnet. Die Verteilungsfunktion* $F_{0,1}$ *wird meist mit* Φ *bezeichnet und ist tabelliert.* φ *und* Φ *sind in Abbildung 3.2 dargestellt.*

```
> z<-seq(-5,5,by=0.1)
> plot(z,pnorm(z,0,1),type="l")
> lines(z,dnorm(z,0,1),col="red")
```

Die Normalverteilungsdichte hatten wir schon im Satz von de Moivre-Laplace ((Formel 2.74)) als Approximation für die Binomialverteilung kennengelernt. Die Normalverteilung wird verwendet für Messfehler und andere Grössen, die man als Überlagerung vieler kleiner Effekte betrachten kann. Als Rechtfertigung gilt der Zentrale Grenzwertsatz (siehe Kap. 4.3).

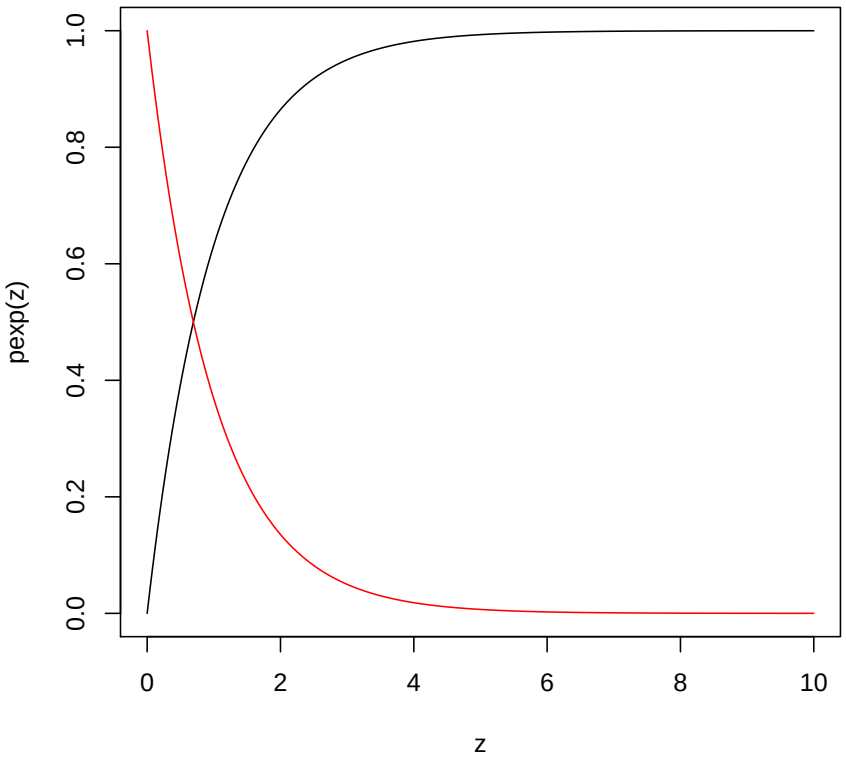

Abbildung 3.1: Dichte und Verteilungsfunktion der Exponentialverteilung mit $\alpha = 1$

Diese beiden Typen (und Mischungen davon) umfassen aber noch nicht alle möglichen Verteilungsfunktionen. Es gibt auch noch solche, die zwar stetig sind, aber für die keine Dichte existiert.

Beispiel 3.9. *Sei $(\Omega, \mathcal{A}, \mathbb{P})$ der Wahrscheinlichkeitsraum für unendlich viele unabhängige $0-1$ Experimente) wie im Abschnitt 3.1.3. Für $\omega = (x_1, x_2, \dots,)$ definieren wir die Zufallsvariable*

$$X(\omega) = \sum_{k=1}^{\infty} x_k 2^{-k},$$

die Umkehrung der Abbildung ϕ in Beispiel 3.4. Um die Verteilungsfunktion von X zu bestimmen, schreiben wir $b \in (0,1)$ in der Binärdarstellung $\sum_{k=1}^{\infty} b_k 2^{-k}$ mit $b_k \in \{0,1\}$ (bei Mehrdeutigkeit nehmen wir die Version mit mit unendlich vielen $b_k = 0$). Dann ist $X(\omega) \leq b$ genau dann, wenn ein n existiert mit $x_k = b_k$ für $k < n$ und $x_n = 0,\ b_n = 1$ oder wenn $x_k = b_k$ für alle k.

Also gilt mit $S_n = \sum_{k=1}^{n} b_k$

$$F(b) = \mathbb{P}\left(\bigcup_{n \geq 1; b_n = 1} \{x_1 = b_1, \dots, x_{n-1} = b_{n-1}, x_n = 0\} \right) = \sum_{n=1}^{\infty} b_n p^{S_{n-1}} q^{n - S_{n-1}}.$$

Für $p = \frac{1}{2}$ erhalten wir insbesondere

$$\mathbb{P}(X \leq b) = \sum_{n=1}^{\infty} b_n 2^{-n} = b,$$

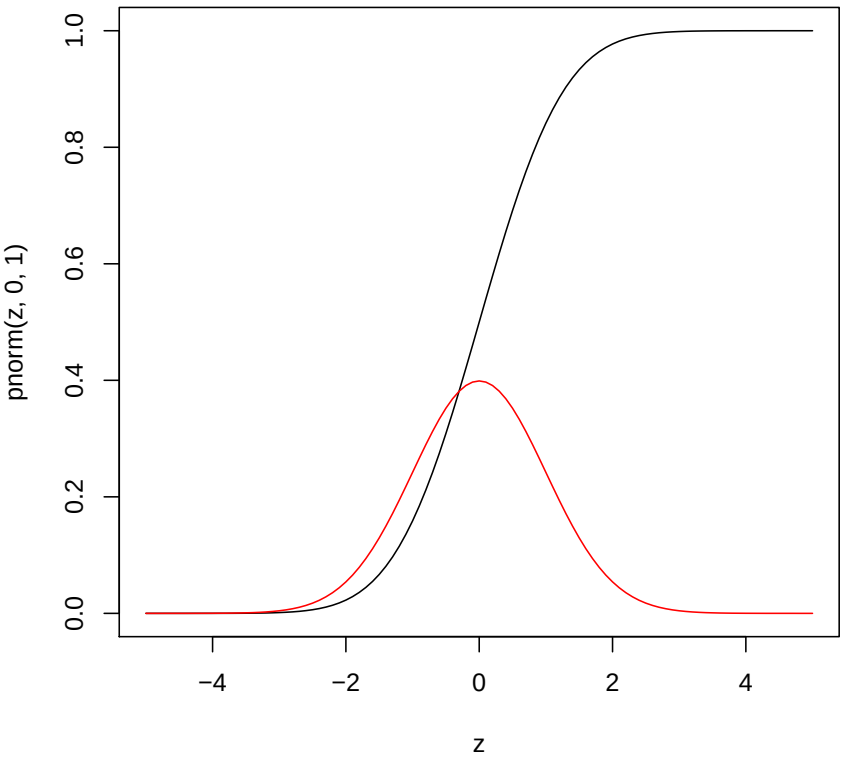

Abbildung 3.2: Dichte und Verteilungsfunktion der Standardnormalverteilung

d.h. die Verteilung von X ist das Lebesguemass (insbesondere ist sie also absolut stetig).

Für $p \neq \frac{1}{2}$ zeigt die folgende Überlegung, dass die Verteilungsfunktion stetig, aber nicht absolut stetig ist: Wenn eine Dichte f_p existieren würde, hätte man für jedes $A \in \mathcal{B}$

$$\mathbb{P}_p(X \in A) = \int_A f_p(x)dx.$$

Insbesondere ist also $\mathbb{P}_p(X \in A) = 0$ für alle A mit $\mathbb{P}_{\frac{1}{2}}(X \in A) = 0$. Aus dem starken Gesetz der grossen Zahlen (siehe Kap. 4.2) folgt aber, dass es ein A gibt mit

$$\mathbb{P}_p(X \in A) = 1, \quad \mathbb{P}_{\frac{1}{2}}(X \in A) = 0.$$

Die Stetigkeit von F (für jedes $p \in (0,1)$) ist leicht einzusehen, denn es gibt höchstens zwei ω mit $X(\omega) = b$, also $\mathbb{P}(X = b) = 0$ für alle b. In der Masstheorie wird gezeigt, dass F sogar fast überall (bezüglich des Lebesguemasses) differenzierbar ist, aber diese Ableitung ist fast überall gleich null, und damit ist F nicht gleich dem Integral der Ableitung.

Abbildung 3.3 zeigt eine empirische Approximation von F für $p = 0.7$: Es werden $K = 10^4$ Werte der Zufallsvariablen erzeugt, wobei wir die unendliche Summe abschneiden bei $m = 50$, und statt $\mathbb{P}(X \leq b)$ wird die relative Häufigkeit berechnet. Man beachte den selbstähnlichen Graphen: Man kann leicht nachrechnen, dass für jedes $b \in [0,1]$ $F(\frac{b}{2}) = qF(b)$ und $F(\frac{b+1}{2}) = q + pF(b)$.

```
> No_density(0.7,50,10000)
```

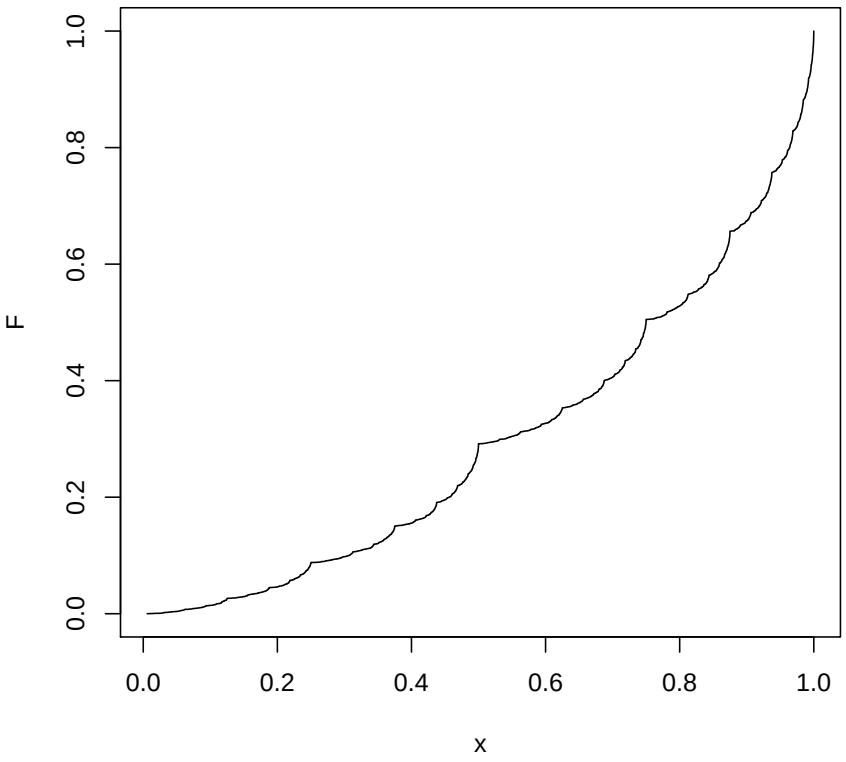

Abbildung 3.3: Eine stetige, aber nicht absolut stetige Verteilungsfunktion

```
> No_density<- function(p,m,K)
+ {
+ x=rbinom(K*m,1,p)
+ dim(x)=c(K,m)
+ z=1:m
+ z=2^(-z)
+ x=x%*%z
+ x=sort(x)
+ F=(1:K)/K
+ plot(x,F,type="l")
+ }
```

3.2.3 Transformation von Zufallsvariablen

Sei X eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ und $g : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ messbar. Dann ist leicht einzusehen, dass

$$Y(\omega) = g(X(\omega))$$

wieder eine Zufallsvariable ist. Sie hat die Verteilungsfunktion

$$F_Y(b) = \mathbb{P}\left(g(X) \leq b\right) = \mathbb{P}\left(X \in g^{-1}((-\infty, b])\right).$$

Beispiel 3.10. *Sei $g(x) = x^2$. Dann ist*

$$F_Y(b) = \mathbb{P}\left(-\sqrt{b} \leq X \leq \sqrt{b}\right) = F_X(\sqrt{b}) - F_X(-\sqrt{b})$$

falls F_X stetig. Falls F_X absolut stetig ist, dann ist auch Y absolut stetig, und zwar ist die Dichte

$$f_Y(b) = \frac{1}{2\sqrt{b}}(f_X(\sqrt{b}) + f_X(-\sqrt{b})).$$

Beispiel 3.11. *Wenn $g(x) = ax + b$ mit $a > 0$ dann ist*

$$F_Y(x) = F_X\left(\frac{x-b}{a}\right).$$

Falls eine Dichte f_X existiert, dann existiert auch f_Y und

$$f_Y(x) = \frac{1}{a} f_X(\frac{x-b}{a}).$$

Durch lineare Transformationen lassen sich also insbesondere die Verteilungen $\mathrm{Exp}\,(\alpha)$ *auf* $\mathrm{Exp}\,(1)$, $\mathcal{U}(a,b)$ *auf* $\mathcal{U}(0,1)$ *und* $\mathcal{N}(\mu,\sigma^2)$ *auf* $\mathcal{N}(0,1)$ *zurückführen.*

Beispiel 3.12. *Wenn g monoton wachsend und differenzierbar ist mit $g'(x) > 0$ für alle x, dann ist $F_Y(b) = F_X(g^{-1}(b))$. Falls die Dichte f_X existiert, dann existiert auch f_Y, und zwar ist*

$$f_Y(x) = \frac{1}{g'(g^{-1}(x))} f_X(g^{-1}(x)).$$

3.3 Erwartungswert

Sei X eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit Verteilung μ. Der Erwartungswert ist eine Kennzahl für die Lage der Verteilung von X; er gibt an, welchen Wert man im Mittel bei vielen unabhängigen Realisierungen erhält (vgl. das Gesetz der grossen Zahlen, Kapitel 4).

Definition 3.5. *Für $X \geq 0$ ist der* **Erwartungswert**

$$\mathbb{E}\,(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega) = \int_{\mathbb{R}} x\mu(dx) \in [0,\infty].$$

Diese Integrale sind im Sinne der Masstheorie zu verstehen. Die Definition geht schrittweise: Zunächst betrachtet man einfache Zufallsvariable $X = \sum_{i=1}^n c_i 1_{A_i}(\omega)$ mit $A_i \in \mathcal{A}$ und setzt $\mathbb{E}\,(X) = \sum c_i \mathbb{P}\,(A_i)$. Dann schreibt man ein beliebiges nichtnegatives X als aufsteigenden Limes von einfachen Zufallsvariablen X_n und setzt $\mathbb{E}\,(X) = \lim_n \mathbb{E}\,(X_n)$. (Man muss zeigen, dass dies nicht davon abhängt, welche Darstellung man wählt).

Für eine Zufallsvariable X, die positive und negative Werte annimmt, setzen wir

$$X^+(\omega) = \max(X(\omega), 0) \quad , \quad X^-(\omega) = \max(-X(\omega), 0)$$

und definieren

$$\mathbb{E}\,(X) = \mathbb{E}\,(X^+) - \mathbb{E}\,(X^-)\,,$$

sofern nicht beide Terme rechts $= +\infty$ sind.

In allen praktischen Fällen können wir den Erwartungswert aber ohne Masstheorie berechnen. Wenn X diskret ist, dann ist $\mathbb{E}\,(X) = \sum_{x_i \in X(\Omega)} x_i \mathbb{P}\,(X = x_i)$.

Wenn die Verteilung von X absolut stetig ist mit stückweise stetiger Dichte, dann ist

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

wobei das Integral im Riemann-Sinn genommen werden kann.

Eigenschaften des Erwartungswertes:

Linearität:

$$\mathbb{E}(\alpha_1 X_1 + \alpha_2 X_2) = \alpha_1 \mathbb{E}(X_1) + \alpha_2 \mathbb{E}(X_2). \tag{3.22}$$

Monotonie:

$$X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y). \tag{3.23}$$

Monotone Stetigkeit:

$$0 \leq X_1 \leq X_2 \leq \cdots \Rightarrow \mathbb{E}\left(\lim_n X_n\right) = \lim_n \mathbb{E}(X_n) \tag{3.24}$$

Konvergenzsatz von Lebesgue: Sei $X_1, X_2, \ldots$ eine f.s. konvergente Folge von Zufallsvariablen. Wenn $|X_n(\omega)| \leq X(\omega)$ für alle n und $\mathbb{E}(X) < \infty$, dann

$$\mathbb{E}\left(\lim_n X_n\right) = \lim_n \mathbb{E}(X_n) \tag{3.25}$$

Beweise: siehe Masstheorie.

Transformation von Zufallsvariablen: Sei $g : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ messbar und $Y(\omega) = g(X(\omega))$. Zur Berechnung von $\mathbb{E}(Y)$ könnte man zuerst die Verteilung von Y gemäss Abschnitt 3.2.3 und dann $\mathbb{E}(Y) = \int x \mu_Y(dx)$ berechnen. Es geht aber einfacher:

$$\mathbb{E}(Y) = \int g(x) \mu_X(dx) = \begin{cases} \sum_{x_i \in X(\Omega)} g(x_i) \mathbb{P}(X = x_i) & (X \text{ diskret}) \\ \int g(x) f(x) dx & (X \text{ absolut stetig}). \end{cases} \tag{3.26}$$

(Beweis siehe Masstheorie).

Insbesondere können wir also

$$\mathbb{E}(X^p) \qquad p = 1, 2, 3 \ldots \ (p - \text{tes Moment}) \tag{3.27}$$

$$\mathbb{E}(|X|^p) \qquad p > 0 \quad (p - \text{tes absolutes Moment}) \tag{3.28}$$

$$\mathbb{E}((X - \mathbb{E}(X))^p) \qquad p = 1, 2, 3 \ldots \ (p - \text{tes zentriertes Moment}) \tag{3.29}$$

direkt aus der Verteilung von X berechnen.

Das zweite zentrierte Moment heisst die **Varianz** von X:

$$\mathbb{V}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) \overset{(3.22)}{=} \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2. \tag{3.30}$$

Die Wurzel der Varianz heisst die **Standardabweichung**:

$$\sigma(X) = \sqrt{\mathbb{V}(X)}. \tag{3.31}$$

Die Standardabweichung misst die Streuung von X um $\mathbb{E}(X)$ und ist damit eine wichtige Kennzahl der Verteilung von X. Wegen (3.22) gilt

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X), \quad \sigma(aX + b) = |a| \sigma(X). \tag{3.32}$$

$\mathcal{L}^p$-Räume: Sei $\mathcal{L}^p = \mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ die Menge der Zufallsvariablen X auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}(|X|^p) < \infty\}$. Durch

$$||X||_p = \mathbb{E}(|X|^p)^{1/p}$$

ist für $p \geq 1$ eine Halbnorm auf $\mathcal{L}^p$ definiert. Wenn wir Zufallsvariable X und Y mit $X = Y$ $\mathbb{P}-$ f.s. identifizieren, wird $\mathcal{L}^p$ zu einem **Banachraum** und $\mathcal{L}^2$ zu einem **Hilbertraum** (siehe Anhang B).

Erwartungswert und Varianz der wichtigsten Verteilungen:

Verteilung	$\mathbb{E}(X)$	$\mathbb{V}(X)$
Binomial (n, p)	np	$np(1-p)$
Hypergeometrisch (n, N, K)	$n\frac{K}{N}$	$n\frac{K}{N}(1-\frac{K}{N})\frac{N-n}{N-1}$
Poisson (λ)	λ	λ
Geometrisch (p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Uniform (a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(α)	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$
Normal (μ, σ^2)	μ	σ^2

Durchführung einiger Rechnungen für obige Tabelle:

Für $X \sim$ Poisson (λ) ist

$$\begin{aligned}\mathbb{E}\left(X^2\right) &= \sum_{k=0}^{\infty} k^2 \mathbb{P}(X=k) = \sum_{k=0}^{\infty}(k(k-1)+k)\mathbb{P}(X=k) \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{k(k-1)}{k!}\lambda^k + \lambda = \lambda^2 + \lambda\end{aligned}$$

Also folgt $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \lambda$.

Für $X \sim \mathcal{U}(0,1)$ ist $\mathbb{E}(X) = \frac{1}{2}$ aus Symmetrie. Ferner

$$\mathbb{E}\left(X^2\right) = \int_0^1 x^2 dx = \frac{1}{3},$$

also $\mathbb{V}(X) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$. Die Resultate für eine beliebige uniforme Verteilung folgen aus (3.22) und (3.32).

Für $X \sim \text{Exp}(1)$ ist mit partieller Integration

$$\mathbb{E}\left(X^k\right) = \int_0^\infty x^k e^{-x} dx = k!,$$

also $\mathbb{E}(X) = 1$ und $\mathbb{V}(X) = 2 - 1^2 = 1$.

3.3.1 Ungleichungen

Für eine nichtlineare Funktion g ist im Allgemeinen

$$\mathbb{E}(g(X)) \neq g(\mathbb{E}(X)).$$

Für ein konvexes oder konkaves g hat man wenigstens eine Ungleichung. Eine Funktion $g : \mathbb{R} \to \mathbb{R}$ heisst konvex, falls es zu jedem x_0 eine Stützgerade $\ell(x) = ax + b$ gibt mit

$$\ell(x) \leq g(x) \quad \forall x, \quad \ell(x_0) = g(x_0).$$

Satz 3.6 (Ungleichung von Jensen)**.** *Für eine Zufallsvariable X mit endlichem Erwartungswert und $g : \mathbb{R} \to \mathbb{R}$ konvex gilt*

$$\mathbb{E}\left(g(X)\right) \geq g(\mathbb{E}\left(X\right)).$$

Beweis Sei ℓ die Stützgerade für $x_0 = \mathbb{E}\left(X\right)$. Dann gilt

$$g(\mathbb{E}\left(X\right)) = \ell(\mathbb{E}\left(X\right)) \overset{(3.22)}{=} \mathbb{E}\left(\ell(X)\right) \overset{(3.23)}{\leq} \mathbb{E}\left(g(X)\right).$$

□

Die folgende Ungleichung ist äusserst nützlich für Grenzwertsätze (→ Kap. 4).

Satz 3.7 (verallgemeinerte Chebyshev-Ungleichung)**.** *Sei g eine nichtnegative, monoton wachsende Funktion auf $\mathbb{R}$. Dann gilt für jedes c mit $g(c) > 0$*

$$\mathbb{P}\left(X \geq c\right) \leq \frac{\mathbb{E}\left(g(X)\right)}{g(c)}.$$

Beweis Offensichtlich ist

$$1_{[X \geq c]} \leq \frac{g(X)}{g(c)}.$$

Also folgt die Behauptung aus (3.22) und (3.23). □

Beispiel 3.13. *Wenn wir Satz (3.7) anwenden auf $Y = |X|$ und $g(x) = \max(x, 0)$, dann folgt*

$$\mathbb{P}\left(|X| > c\right) \leq \frac{\mathbb{E}\left(|X|\right)}{c}.$$

Insbesondere impliziert $\mathbb{E}\left(|X|\right) = 0$, dass $\mathbb{P}\left(X = 0\right) = 1$, da

$$\mathbb{P}\left(X = 0\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \{|X| \leq \frac{1}{n}\}\right) \overset{(3.11)}{=} \lim_n \mathbb{P}\left(|X| \leq \frac{1}{n}\right).$$

Beispiel 3.14. *Wenn wir Satz (3.7) anwenden auf $Y = |X - \mathbb{E}\left(X\right)|$ und $g(x) = (\max(x, 0))^2$, dann folgt die Chebyshev-Ungleichung*

$$\mathbb{P}\left(|X - \mathbb{E}\left(X\right)| > c\right) \leq \frac{\mathbb{V}\left(X\right)}{c^2}.$$

Insbesondere impliziert $\mathbb{V}\left(X\right) = 0$, dass X f.s. konstant ist.

Diese Ungleichungen sind zwar sehr einfach und universell gültig, dafür aber in spezifischen Fällen oft recht grob.

3.4 Mehrere Zufallsvariablen

3.4.1 Begriffe

Seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Wir betrachten den Zufallsvektor

$$\mathbf{X} = (X_1, X_2, \ldots, X_n). \tag{3.33}$$

Auf $\mathbb{R}^n$ sei $\mathcal{B}^n = \sigma(\{A_1 \times A_2 \times \cdots \times A_n | A_i \in \mathcal{B}\})$ die Borel-σ-Algebra. Mengen der Form $A_1 \times \cdots \times A_n$ nennen wir auch (verallgemeinerte) *Rechtecke*. Gemäss (3.17) ist dann die Abbildung

$$\mathbf{X} : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}^n, \mathcal{B}^n)$$

automatisch messbar. Das Bild μ von $\mathbb{P}$ unter $\mathbf{X}$ heisst die **gemeinsame Verteilung** von $X_1, \ldots, X_n$:

$$\mu(A) = \mathbb{P}\left(\mathbf{X}^{-1}(A)\right) = \mathbb{P}\left(\{\omega | \mathbf{X}(\omega) \in A\}\right) = \mathbb{P}[\mathbf{X} \in A] \quad (A \in \mathcal{B}^n). \tag{3.34}$$

Für Anwendungen sind die beiden folgenden Fälle am wichtigsten:

Wenn jedes X_i **diskret** ist, dann ist $\mathbf{X}(\Omega)$ abzählbar und

$$\mu(A) = \sum_{\mathbf{x} \in \mathbf{X}(\Omega) \cap A} \mathbb{P}(\mathbf{X} = \mathbf{x}) = \sum_{\substack{(x_1,\ldots,x_n) \in A \\ x_1 \in X_1(\Omega), \ldots, x_n \in X_n(\Omega)}} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n). \tag{3.35}$$

Die gemeinsame Verteilung ist **absolut stetig**, d.h. es gibt eine messbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ mit $f \geq 0$ und $\int_{\mathbb{R}^n} f(\mathbf{x}) d\mathbf{x} = 1$ derart, dass

$$\mu(A) = \int_A f(\mathbf{x}) d\mathbf{x}. \tag{3.36}$$

Die Funktion f heisst wie im eindimensionalen Fall die Dichte.

Beispiel 3.15. *Sei $A \in \mathcal{B}^n$ mit $0 < \lambda(A) < \infty$, wobei λ das Lebesguemass bezeichnet. Dann ist die* **Gleichverteilung** *auf A:*

$$\mu_A(B) = \frac{\lambda(B \cap A)}{\lambda(A)} \tag{3.37}$$

eine absolut stetige Verteilung. Die Dichte ist konstant gleich $1/\lambda(A)$ auf A und null ausserhalb.

Neben diesen beiden Typen gibt es aber noch andere Verteilungen, z.B. die Gleichverteilung auf $\{\mathbf{x} \in \mathbb{R}^n \mid ||\mathbf{x}|| = 1\}$.

Analog zu $n = 1$ kann man die gemeinsame Verteilungsfunktion definieren:

$$F(b_1, \ldots, b_n) = \mathbb{P}(X_1 \leq b_1, \ldots, X_n \leq b_n) = \mu((-\infty, b_1] \times \cdots \times (-\infty, b_n]).$$

Sie spielt jedoch nur eine untergeordnete Rolle.

Aus der gemeinsamen Verteilung erhält man insbesondere auch die Verteilung von jedem X_i allein, die sogenannte (eindimensionale) **Rand-** oder **Marginalverteilung:**

$$\mu_i(B) = \mathbb{P}(X_i \in B) = \mu(\mathbb{R} \times \mathbb{R} \times \cdots \times \underbrace{B}_{i-te\ Stelle} \times \ldots \mathbb{R}) \quad (B \in \mathcal{B}). \tag{3.38}$$

Speziell für μ diskret

$$\mathbb{P}(X_i = x_i) = \sum_{\substack{x_j \in X_j(\Omega) \\ j \neq i}} \mathbb{P}(X_1 = x_1, \ldots, X_i = x_i, \ldots, X_n = x_n), \tag{3.39}$$

und für μ absolut stetig

$$\mu_i(B) = \int_{\mathbb{R}} \ldots \int_B \ldots \int_{\mathbb{R}} f(\mathbf{x}) d\mathbf{x}.$$

Das heisst, μ_i ist ebenfalls absolut stetig mit Dichte

$$f_i(x_i) = \int_{\mathbb{R}^{n-1}} f(\mathbf{x}) \;\; dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n. \tag{3.40}$$

Umgekehrt kann man aber aus den Randverteilungen nicht die gemeinsame Verteilung bestimmen.

Beispiel 3.16. *Sei $n = 2$, $X_i(\Omega) = \{0, 1\}$ und α ein Parameter mit $-\frac{1}{4} \leq \alpha \leq \frac{1}{4}$. Wir setzen*

$$\begin{array}{rcl} \mathbb{P}(X_1 = 0, X_2 = 0) & = & \mathbb{P}(X_1 = 1, X_2 = 1) \;\; = \;\; \frac{1}{4} + \alpha \\ \mathbb{P}(X_1 = 1, X_2 = 0) & = & \mathbb{P}(X_1 = 0, X_2 = 1) \;\; = \;\; \frac{1}{4} - \alpha \end{array}$$

Dann ist $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \frac{1}{2}$ $(i = 1, 2)$ für jedes α.

Die gemeinsame Verteilung enthält eben noch zusätzliche Information, nämlich solche über die Abhängigkeiten zwischen den Variablen.

Definition 3.6. *$X_1, \dots, X_n$ heissen (stochastisch)* **unabhängig** *falls für alle $A_1, \dots, A_n \in \mathcal{B}$ gilt:*

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdot \dots \cdot \mathbb{P}(X_n \in A_n)$$

bzw.

$$\mu(A_1 \times \dots \times A_n) = \mu_1(A_1) \cdot \dots \cdot \mu_n(A_n).$$

(Man sagt auch, μ ist das Produkt von $\mu_1, \dots, \mu_n$).

Im Fall der Unabhängigkeit ist die gemeinsame Verteilung durch die Randverteilungen festgelegt (denn μ ist festgelegt, durch die Werte auf den verallgemeinerten Rechtecken, $\rightarrow$ Masstheorie).

Satz 3.8. *Seien $X_1, \dots, X_n$ unabhängig. Dann ist μ absolut stetig genau dann, wenn jedes μ_i absolut stetig ist. Ferner gilt $f(\mathbf{x}) = \prod_{i=1}^n f_i(x_i)$.*

Beweis Masstheorie. □

Beispiel 3.17. Standard-Normalverteilung im $\mathbb{R}^n$. *Seien $X_1, \dots, X_n$ unabhängig und $\mathcal{N}(0, 1)$-verteilt. Dann hat die gemeinsame Verteilung die Dichte*

$$f(x_1, \dots, x_n) = (2\pi)^{-n/2} \exp(-\frac{1}{2} \sum_{i=1}^n x_i^2).$$

Die Dichte ist also sphärisch symmetrisch, und man kann zeigen, dass es die einzige sphärisch symmetrische Verteilung ist, bei der die Komponenten unabhängig sind. Daraus hat Maxwell geschlossen, dass der Geschwindigkeitsvektor eines Gasmoleküls diese Verteilung haben muss.

3.4.2 Transformationen

Sei $\mathbf{X}$ ein $n-$dimensionaler Zufallsvektor und $g : (\mathbb{R}^n, \mathcal{B}^n) \rightarrow (\mathbb{R}^m, \mathcal{B}^m)$ eine messbare Abbildung. Dann ist

$$\mathbf{Y}(\omega) = g(\mathbf{X}(\omega)) \tag{3.41}$$

ein $m-$dimensionaler Zufallsvektor. Ferner gilt

$$\mu_{\mathbf{Y}}(A) = \mu_{\mathbf{X}}(g^{-1}(A)).$$

Je nach der Art der Funktion g, kann man $\mu_{\mathbf{Y}}$ mehr oder weniger explizit angeben.

Satz 3.9. *Sei $g : \mathbf{R}^n \to \mathbf{R}^n$ linear und umkehrbar, d.h. $g(\mathbf{x}) = \mathbf{m} + B\mathbf{x}$ mit $\det(B) \neq 0$. Wenn $\mu_{\mathbf{X}}$ absolut stetig ist, dann ist auch $\mu_{\mathbf{Y}}$ absolut stetig und es gilt:*

$$f_{\mathbf{Y}}(\mathbf{x}) = \frac{1}{|\det(B)|} f_{\mathbf{X}}(B^{-1}(\mathbf{x} - \mathbf{m})). \tag{3.42}$$

Beweis Mithilfe einer Substitution erhält man

$$\mu_{\mathbf{Y}}(A) \overset{(3.41)}{=} \mu_{\mathbf{X}}[g^{-1}(A)] = \int_{g^{-1}(A)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int_A f_{\mathbf{X}}(g^{-1}(\mathbf{x})) \frac{1}{|\det(B)|} d\mathbf{x}.$$

□

Beispiel 3.18. *Wenn $\mathbf{X}$ standard-normalverteilt ist, dann*

$$f_{\mathbf{Y}}(\mathbf{x}) = (2\pi)^{-n/2} \frac{1}{\sqrt{|\det \Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^\top \Sigma^{-1}(\mathbf{x} - \mathbf{m})\right) \tag{3.43}$$

*wobei $\Sigma = BB^\top$ und wir $\mathbf{x}$ als Spaltenvektor auffassen. Dies ist die allgemeine nicht degenerierte (da $\det \Sigma \neq 0$) n-**dimensionale Normalverteilung** $\mathcal{N}_n(\mathbf{m}, \Sigma)$.*

Beispiel 3.19. Verteilung von Summen: *Sei $\mathbf{X} = (X_1, X_2)$ eine zweidimensionale absolut stetige Zufallsvariable. Wir setzen $\mathbf{Z} = (X_1, Y)$ mit $Y = X_1 + X_2$. Wegen (3.42) ist $f_{\mathbf{Z}}(x_1, y) = f_{\mathbf{X}}(x_1, y - x_1)$. Also erhalten wir wegen (3.40), dass die Verteilung der Summe $Y = X_1 + X_2$ absolut stetig ist mit Dichte*

$$f_Y(y) = \int_{\mathbb{R}} f_{\mathbf{Z}}(x_1, y) dx_1 = \int_{\mathbb{R}} f_{\mathbf{X}}(x_1, y - x_1) dx_1.$$

Wenn X_1 und X_2 unabhängig sind, hat man

$$f_Y(y) = \int_{-\infty}^{\infty} f_1(x_1) f_2(y - x_1) dx_1 \tag{3.44}$$

d.h. f_Y ist die **Faltung** *von f_1 und f_2. Insbesondere für X_1, X_2 unabhängig, $\mathcal{U}(0,1)$-verteilt, erhalten wir die Dreiecksverteilung*

$$f_Y(y) = \int_0^1 I_{[y-x_1]} dx_1 = \begin{cases} y & (0 \leq y \leq 1) \\ 2 - y & (1 \leq y \leq 2) \\ 0 & \textit{sonst.} \end{cases}$$

Analog berechnet man die Dichte im Fall von 3 oder mehr uniformen Summanden. Abbildung 3.4 zeigt ein Histogramm der Summe zweier unabhängiger uniform $\mathcal{U}(0,1)$-verteilter Zufallsvariablen mit der typischen Zeltform.

3.4.3 Kovarianz und Korrelation

Sei $g : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}, \mathcal{B})$ messbar. Zur Berechnung von $\mathbb{E}(g(\mathbf{X}))$ muss man die Verteilung von $Y = g(\mathbf{X})$ nicht bestimmen, sondern es gilt wie im eindimensionalen Fall

$$\mathbb{E}(g(\mathbf{X})) = \int_{\mathbb{R}^n} g(\mathbf{x}) \mu(d\mathbf{x}). \tag{3.45}$$

Im diskreten, bzw. absolut stetigen Fall heisst das

$$= \sum_{x_i \in X_i(\Omega)} g(x_1, \ldots, x_n) \, \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n), \text{ bzw.} \tag{3.46}$$

$$= \int_{\mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}. \tag{3.47}$$

```
> y<-runif(10000,0,1)
> z<-runif(10000,0,1)
> hist(y+z)
```

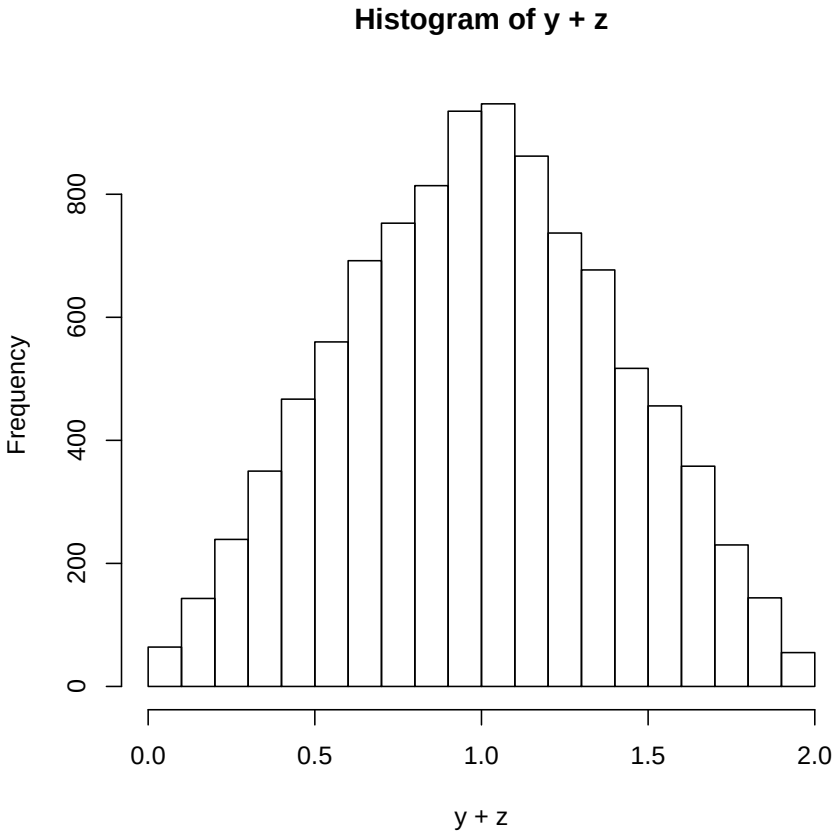

Abbildung 3.4: Histogramm der Verteilung der Summe von 2 unabhängigen $\mathcal{U}(0,1)$-verteilten Zufallsvariablen

Insbesondere können wir das benutzen, um die Kovarianz von zwei Zufallsvariablen zu berechnen:

Definition 3.7. *Die* **Kovarianz** *von X_1 und X_2 ist definiert als*

$$\operatorname{Cov}(X_1, X_2) = \mathbb{E}\left(\Big(X_1 - \mathbb{E}(X_1)\Big)\Big(X_2 - \mathbb{E}(X_2)\Big)\right)$$

Satz 3.10. *Die Kovarianz hat die folgenden Eigenschaften*

i) $\operatorname{Cov}(X, X) = \mathbb{V}(X)$.

ii) $\operatorname{Cov}(X_1, X_2) = \operatorname{Cov}(X_2, X_1)$.

iii) $\operatorname{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\,\mathbb{E}(X_2)$.

iv) $\operatorname{Cov}(X_1, aX_2 + b) = a \operatorname{Cov}(X_1, X_2)$.

v) $\operatorname{Cov}(X_1, X_2 + X_3) = \operatorname{Cov}(X_1, X_2) + \operatorname{Cov}(X_1, X_3)$.

vi) $\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + 2\operatorname{Cov}(X_1, X_2)$.

vii) $|\operatorname{Cov}(X_1, X_2)| \leq \sigma(X_1)\sigma(X_2)$.

viii) *Wenn X_1, X_2 unabhängig sind, dann ist* $\operatorname{Cov}(X_1, X_2) = 0$ *und daher* $\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2)$.

Beweis Die ersten zwei Behauptungen sind offensichtlich aufgrund der Definition. Die nächsten vier Behauptungen folgen durch Ausrechnen und Anwenden der Regeln für den Erwartungswert. Die siebte Behauptung ist nichts anderes als die Cauchy-Schwarz-Ungleichung. Im diskreten Fall folgt die letzte Behauptung aus Lemma 2.4. Analog erhalten wir im absolut stetigen Fall

$$\mathbb{E}(X_1 X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_1(x_1) f_2(x_2) dx_1 dx_2 = \mathbb{E}(X_1)\,\mathbb{E}(X_2)\,.$$

Für den allgemeinen Fall braucht man Masstheorie. □

Beispiel 3.20. *Eine binomial(n,p)-verteilte Zufallsvariable X lässt sich schreiben als $\sum_{i=1}^n X_i$ mit $X_1, \ldots, X_n$ unabhängig und binär. Also folgt $\mathbb{V}(X) = \sum_{i=1}^n \mathbb{V}(X_i) = np(1-p)$, denn $\mathbb{V}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = p - p^2 = p(1-p)$.*

Bemerkung Die Umkehrung von Satz 3.10, viii) ist falsch. Wenn z.B. $X_1 \sim \mathcal{N}(0,1)$ und $X_2 = X_1^2$, dann ist aus Symmetrie $\operatorname{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) = \mathbb{E}(X_1^3) = 0$. Hier sind aber X_1 und X_2 sehr stark abhängig.

Beispiel 3.21. *Bei der n-dimensionalen Normalverteilung (siehe Beispiel 3.18) können wir die Kovarianzen $\operatorname{Cov}(Y_i, Y_j)$ mit Hilfe der Regeln iv) und v) von Satz 3.10 sofort berechnen, weil $\operatorname{Cov}(X_i, X_j) = 0$ für $i \neq j$ und $\operatorname{Cov}(X_i, X_i) = \mathbb{V}(X_i) = 1$. Man erhält $\operatorname{Cov}(Y_i, Y_j) = (B^\top B)_{ij} = \Sigma_{ij}$.*

Wenn $\mathbf{Y}$ eine n-dimensionale Normalverteilung hat und $\operatorname{Cov}(Y_i, Y_j) = 0$ für $i \neq j$, dann zerfällt die gemeinsame Dichte in ein Produkt und damit sind die $Y_1, Y_2, \ldots, Y_n$ unabhängig. Die Umkehrung von Satz 3.10, viii) gilt daher bei gemeinsamer Normalverteilung.

Die Kovarianz verschiedener Paare von Zufallsvariablen lässt sich nicht direkt vergleichen, da sie auch von der Streuung der Variablen abhängt. Eine anschaulichere Kennzahl ist die **Korrelation**.

Definition 3.8. *Die* **Korrelation von** *X_1 und X_2 ist*

$$\rho(X_1, X_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\sigma(X_1)\sigma(X_2)}.$$

Wenn $\rho(X_1, X_2) = 0$ ($\Leftrightarrow \mathrm{Cov}\,(X_1, X_2) = 0$) dann heissen X_1 und X_2 **unkorreliert**. Die Korrelation misst Stärke und Richtung des linearen Zusammenhangs zwischen den beiden Variablen. Aus Satz 3.10 folgen sofort die Eigenschaften

$$\rho(aX_1 + b, cX_2 + d) = \rho(X_1, X_2) \text{ für } a > 0, \quad c > 0. \tag{3.48}$$

$$-1 \leq \rho(X_1, X_2) \leq +1. \tag{3.49}$$

Kapitel 4

Grenzwertsätze

Sei $X_1, X_2, \ldots$ eine Folge von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Wir betrachten die Summen

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

und interessieren uns für das asymptotische Verhalten von S_n für $n \to \infty$. Die **Gesetze der grossen Zahlen** beschreiben die Konvergenz der Mittelwerte $\frac{1}{n}\, S_n$, während der **zentrale Grenzwertsatz** die Form der Verteilung von S_n angibt. Der wichtigste Fall ist der, wo die X_i i.i.d. (unabhängig und identisch verteilt = **i**ndependent and **i**dentically **d**istributed) sind. Wir werden aber auch kurz diskutieren, inwieweit man auf die identische Verteilung verzichten kann. Was passiert, wenn die X_i's abhängig sind, ist ebenfalls untersucht worden; wir gehen aber darauf nicht ein.

4.1 Schwaches Gesetz der grossen Zahlen

Wir nehmen an, dass alle X_i einen gemeinsamen Erwartungswert $\mathbb{E}(X_i) = m$ haben. Wir sagen, dass das **schwache Gesetz der grossen Zahlen** gilt, falls für alle $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > \varepsilon\right) \longrightarrow 0 \quad \text{für } n \to \infty. \tag{4.1}$$

Mit der Chebyshev-Ungleichung folgt:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > \varepsilon\right) \leq \frac{\mathbb{V}(S_n/n)}{\varepsilon^2} = \frac{\mathbb{V}(S_n)}{n^2\varepsilon^2} \tag{4.2}$$

Wenn die X_i i.i.d. sind, dann $\mathbb{V}(S_n) = n\mathbb{V}(X_1)$ also gilt das schwache Gesetz der grossen Zahlen für X_i i.i.d. , $\mathbb{E}(X_i^2) < \infty$. Wenn die X_i unkorreliert sind, gilt $\mathbb{V}(S_n) = \sum_{i=1}^n \mathbb{V}(X_i)$ so dass $\sum_{i=1}^n \mathbb{V}(X_i) = o(n^2)$ hinreichend ist für das schwache Gesetz der grossen Zahlen.
Beispiel 4.1. Gegenbeispiel zum Gesetz der grossen Zahlen: *Sei (X_i) i.i.d. mit Dichte $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ (sog.* **Cauchy-Verteilung***). Dann ist $\mathbb{E}(|X_i|) = \infty$, und mit Hilfe der Faltungsformel (3.44) kann man zeigen, dass $\frac{S_n}{n}$ für alle $n \in \mathbb{N}$ wieder Cauchy-verteilt ist. Das heisst, dass $\frac{S_n}{n}$ immer gleich stark streut, die ausgleichende Wirkung des Zufalls spielt hier also nicht.*

Beispiel 4.2. *Wir verwenden das Gesetz der grossen Zahlen, um den* **Satz von Weierstrass** *zu beweisen, dass die Menge der Polynome dicht ist in $C[0,1]$, versehen mit der Supremums-Norm. Die* **Bernsteinpolynome** *vom Grad n auf $[0,1]$ sind definiert als*

$$B_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k} \quad (k = 0, 1, \ldots, n). \tag{4.3}$$

Eine stetige Funktion f auf $[0,1]$ kann durch die folgende Linearkombination der Bernsteinpolynome approximiert werden

$$B_n^f(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) B_{n,k}(x).$$

Weshalb ist $B_n^f(x) \approx f(x)$? Eine probabilistische Begründung beruht darauf, dass $B_{n,k} = \mathbb{P}(S_n = k)$, wobei S_n die Anzahl Erfolge bei n Würfen mit Erfolgsparameter x bezeichnet. Also ist $B_n^f(x) = \mathbb{E}\left(f(\frac{S_n}{n})\right)$. Nach dem Gesetz der grossen Zahlen ist aber $\frac{S_n}{n} \approx x$ und f ist stetig.

Genauer gilt:

$$\begin{aligned} |B_n^f(x) - f(x)| &= \left|\mathbb{E}\left(f\left(\frac{S_n}{n}\right) - f(x)\right)\right| \le \mathbb{E}\left(\left|f\left(\frac{S_n}{n}\right) - f(x)\right|\right) \\ &\le 2\sup_u |f(u)|\ \mathbb{P}\left(\left|\frac{S_n}{n} - x\right| > \delta\right) + \sup_{|u-v|\le\delta} |f(u) - f(v)|\ \mathbb{P}\left(\left|\frac{S_n}{n} - x\right| \le \delta\right). \end{aligned}$$

Wegen der gleichmässigen Stetigkeit von f ist der zweite Term rechts $\le \varepsilon$ wenn δ klein genug ist. Der erste Term rechts ist wegen der Chebyshev-Ungleichung beschränkt durch

$$2\sup_u |f(u)| \frac{x(1-x)}{n\delta^2} \le \frac{1}{2n\delta^2} \sup_u |f(u)| \le \varepsilon,$$

wenn n gross genug ist. Damit haben wir gezeigt, dass

$$\sup_x |B_n^f(x) - f(x)| \longrightarrow 0 \text{ für } n \to \infty.$$

Abbildung 4.1 zeigt die Bernsteinpolynome vom Grad 10 und die Approximation von der Gauss'schen Glockenkurve.

4.2 Starkes Gesetz der grossen Zahlen

Statt (4.1) versuchen wir jetzt, die stärkere Aussage

$$\lim_n \mathbb{P}\left(\bigcap_{k\ge n} \left\{\left|\frac{S_k}{k} - m\right| \le \varepsilon\right\}\right) = 1 \tag{4.4}$$

für alle $\varepsilon > 0$ zu beweisen, d.h. "das arithmetische Mittel *bleibt* von einem Zeitpunkt n an immer in der Nähe vom Erwartungswert m". Wegen der Stetigkeit von $\mathbb{P}$ (Satz 3.1) ist dies äquivalent zu:

$$\forall \varepsilon > 0 : \ \mathbb{P}\left(\bigcup_n \bigcap_{k\ge n} \left\{\left|\frac{S_k}{k} - m\right| \le \varepsilon\right\}\right) = 1 \quad \text{bzw.} \tag{4.5}$$

```
> x <- seq(0.001,0.999,by=0.001)
> q <- x/(1-x)
> f <- exp(-10*(x-0.5)^2)
> n <- 10
> x.n <- (0:n)/n
> f.n <- exp(-10*(x.n-0.5)^2)
> B <-  matrix(0,nrow=999,ncol=n+1)
> B[,1] <- (1-x)^n #B_{n,0}
> plot(x,B[,1],type="l",col=2,ylim=c(0,max(f)))
> for (k in (1:n)) {
+   B[,k+1] <- ((n-k+1)/k)*q*B[,k] #Rekursive Berechnung von B_{n,k}
+   lines(x,B[,k+1],col=2)
+ }
> lines(x,f)
> lines(x,B %*% f.n,col=4)
```

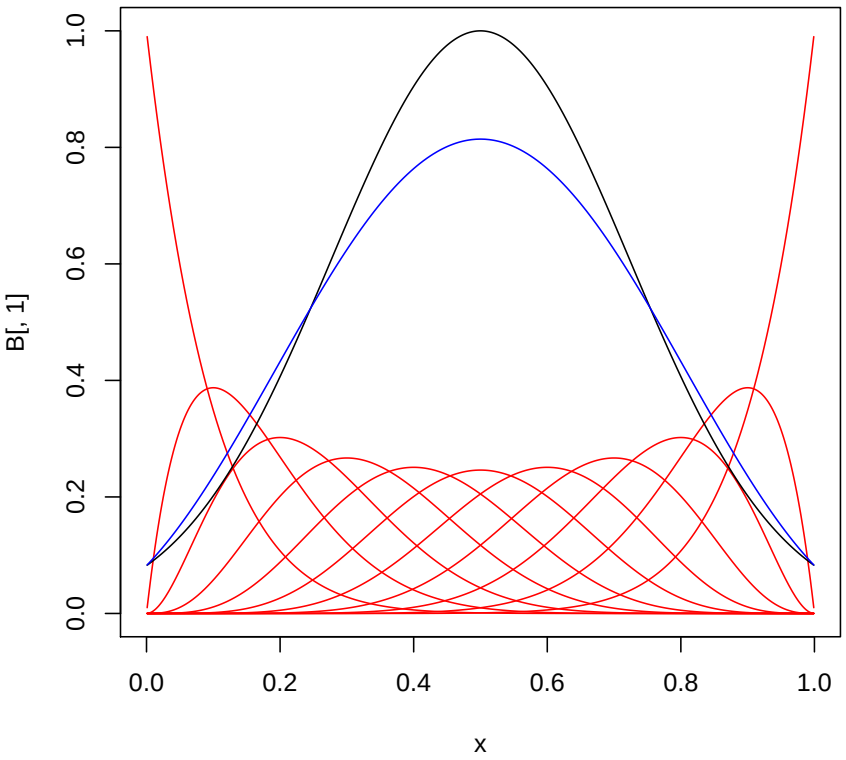

Abbildung 4.1: Bernsteinpolynome vom Grad 10 und Approximation (blau) der Funktion $f(x) = \exp(-10(x-0.5)^2)$ (schwarz).

$$\mathbb{P}\left(\bigcap_{\varepsilon>0}\bigcup_{n}\bigcap_{k\geq n}\left\{\left|\frac{S_k}{k}-m\right|\leq\varepsilon\right\}\right)=\mathbb{P}\left(\lim_{n\to\infty}\frac{S_n}{n}=m\right)=1. \tag{4.6}$$

Wenn $\mathbb{P}\left(\frac{S_n}{n}\to m\right)=1$, dann sagen wir, dass das **starke Gesetz der grossen Zahlen** gilt.

Allgemein definieren wir für Zufallsvariablen $Z, Z_1, Z_2, \ldots$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$

a) **Stochastische Konvergenz** oder Konvergenz in Wahrscheinlichkeit von Z_n gegen Z:

$$\forall\varepsilon>0\quad\lim_{n\to\infty}\mathbb{P}\left(|Z_n-Z|>\varepsilon\right)=0. \tag{4.7}$$

b) **Fast-sichere Konvergenz** von Z_n gegen Z:

$$\mathbb{P}\left(\{\omega|\lim Z_n(\omega)=Z(\omega)\}\right)=1. \tag{4.8}$$

Dann gilt

Satz 4.1. *i) Fast-sichere Konvergenz impliziert stochastische Konvergenz.*

ii) Wenn $\sum_n \mathbb{P}\left(|Z_n-Z|>\varepsilon\right)<\infty\ \ \forall\varepsilon>0$*, dann konvergiert* Z_n *f.s. gegen* Z*.*

Korollar 4.1. *Wenn* (Z_n) *stochastisch gegen* Z *konvergiert, dann existiert eine Teilfolge* (Z_{n_j})*, welche f.s. gegen* Z *konvergiert.*

Beweis des Korollars: Wähle n_j so, dass $\mathbb{P}\left(|Z_{n_j}-Z|>\frac{1}{j}\right)\leq\frac{1}{j^2}$. Dann gilt $\forall\varepsilon>0$ $\sum_j\mathbb{P}\left(|Z_{n_j}-Z|>\varepsilon\right)<\infty$. □

Beweis des Satzes:

i) Mit den gleichen Überlegungen wie oben bei der Äquivalenz von (4.4) und (4.6) ergibt sich:

$$\mathbb{P}\left(\lim_n Z_n=Z\right)=1\Leftrightarrow\lim_n\mathbb{P}\left(\bigcap_{k\geq n}\{|Z_k-Z|\leq\varepsilon\}\right)=1\ \ \forall\varepsilon>0,$$

und daraus folgt natürlich 4.7.

ii) Mit dem Lemma von Borel-Cantelli folgt für jedes $\varepsilon>0$

$$\sum_{n=1}^{\infty}\mathbb{P}\left(|Z_n-Z|>\varepsilon\right)<\infty\Rightarrow\mathbb{P}\left(\bigcap_n\bigcup_{k\geq n}\{|Z_k-Z|>\varepsilon\}\right)=0.$$

Indem wir zum Komplement übergehen, erhalten wir

$$\mathbb{P}\left(\bigcup_n\bigcap_{k\geq n}\{|Z_k-Z|\leq\varepsilon\}\right)=1$$

woraus die Behauptung folgt. □

Gegenbeispiel: Sei (Y_i) i.i.d. mit Werten in $\mathbb{N}$, $\mathbb{P}\left(Y_i\geq k\right)=\frac{1}{k}$, und $Z_n=n^{-1}Y_n$. Dann konvergiert Z_n stochastisch gegen 0, aber $\mathbb{P}\left(Z_n\to 0\right)=0$, denn $\sum_n\mathbb{P}\left(Z_n\geq 1\right)=\sum_n\mathbb{P}\left(Y_n\geq n\right)=\infty$, also sind wegen Borel-Cantelli f.s. unendlich viele Z_n grösser oder gleich 1.

Satz 4.2. *Sei (X_i) i.i.d. mit $\mathbb{E}(X_i^2) < \infty$. Dann konvergiert $\frac{S_n}{n}$ fast sicher gegen $m = \mathbb{E}(X_i)$.*

Beweis : Wir dürfen annehmen, dass $X_i \geq 0$ (zerlege X_i in $X_i^+ = \max(X_i, 0)$ und $X_i^- = \max(-X_i, 0)$). Wegen (4.2) und Satz 4.1 ii) konvergiert die Teilfolge S_{n^2}/n^2 f.s. gegen m. Wir müssen also nur noch S_k/k für $n^2 \leq k \leq (n+1)^2$ untersuchen. Da $S_{k+1} \geq S_k$ ist, folgt:

$$\frac{n^2}{(n+1)^2}\,\frac{S_{n^2}}{n^2} = \frac{S_{n^2}}{(n+1)^2} \leq \frac{S_k}{k} \leq \frac{S_{(n+1)^2}}{n^2} = \frac{(n+1)^2}{n^2}\,\frac{S_{(n+1)^2}}{(n+1)^2}.$$

Die beiden Schranken links und rechts konvergieren f.s. gegen m, also auch $\frac{S_k}{k} \to m$ f.s.

□

Notwendige und hinreichende Bedingungen für das starke Gesetz der grossen Zahlen wurden von Kolmogorov gefunden (ohne Beweis): Für (X_i) i.i.d. gilt $\frac{S_n}{n} \to m \in \mathbb{R}$ fast sicher genau dann, wenn $\mathbb{E}(|X_i|) < \infty$ und $m = \mathbb{E}(X_i)$.

Das **Gesetz vom iterierten Logarithmus** (Hartman-Wintner, 1941) gibt die präzise Antwort auf die Frage, wie stark S_n/n von m abweicht. Sei (X_i) eine i.i.d. Folge von Zufallsvariablen, $\mathbb{E}(X_i) = m$, $\mathbb{V}(X_i) = \sigma^2 < \infty$. Dann gilt mit Wahrscheinlichkeit 1

$$\limsup \frac{S_n - nm}{\sqrt{2\sigma^2 n \log(\log(n))}} = 1, \quad \liminf \cdots = -1$$

(ohne Beweis). Also ist für jedes $\varepsilon > 0$ mit Wahrscheinlichkeit 1

$$|\tfrac{S_n}{n} - m| > (1+\varepsilon)\sigma\sqrt{\tfrac{2\log(\log(n))}{n}} \quad \text{nur für endlich viele } n'\text{s},$$
$$|\tfrac{S_n}{n} - m| > (1-\varepsilon)\sigma\sqrt{\tfrac{2\log(\log(n))}{n}} \quad \text{für unendlich viele } n'\text{s}.$$

```
> monte.carlo.trajectory<-function(n)
+ {
+ integrand <- function(x) {sqrt(x^4+17*x^2)}
+ ## integrate the function from -1 to 1
+ truevalue=integrate(integrand, lower = -1, upper = 1)
+ x=runif(n,-1,1)
+ length=rep(1:n)
+ trajectory=2/length*cumsum(integrand(x))
+ truevalue=rep(truevalue$value,times=n)
+ plot(trajectory,type="l")
+ lines(truevalue,col="red")
+ }
```

4.3 Zentraler Grenzwertsatz

Seien X_i unabhängige Zufallsvariablen mit $\mathbb{E}(X_i) = m_i$ und $\mathbb{V}(X_i) = \sigma_i^2 < \infty$. Wir hatten bereits im Kapitel 1 gesehen, dass für binäre X_i's die Form der Verteilung von S_n durch die Normalverteilung approximiert wird. Dieses Resultat gilt sehr viel allgemeiner. Da die Form der Verteilung unabhängig von Lage und Streuung ist, standardisieren wir S_n, so dass der Erwartungswert $= 0$ und die Varianz $= 1$ ist:

$$S_n^* = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}} = \frac{S_n - \sum_{i=1}^n m_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}. \tag{4.9}$$

```
> monte.carlo.trajectory(10000)
```

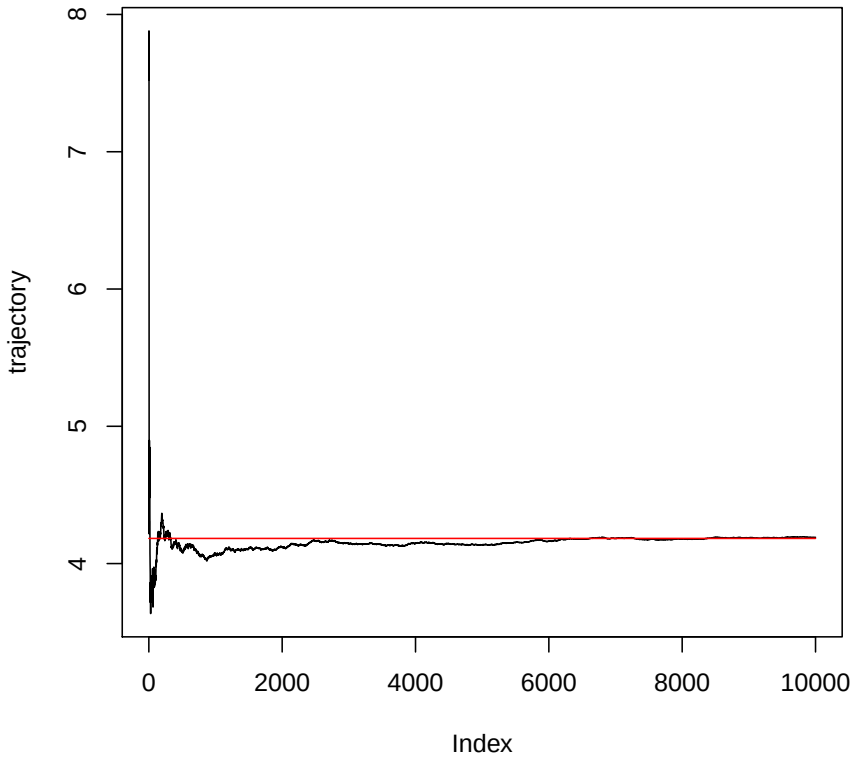

Abbildung 4.2: Monte-Carlo Integration als Funktion von n

Wir wollen nun zeigen, dass die Verteilung von S_n^* unter gewissen Bedingungen gegen die **Standard-Normalverteilung** $\mathcal{N}(0,1)$ konvergiert. Wir verwenden dazu den folgenden Konvergenzbegriff:

Definition 4.1. *Seien μ und μ_n Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B})$. Wir sagen, dass μ_n* **schwach** *gegen μ* **konvergiert**, *falls*

$$\int f d\mu_n \longrightarrow \int f d\mu \tag{4.10}$$

für alle f, welche stetig und beschränkt sind.

Bemerkung: Sei Z_n eine Zufallsvariable mit der Verteilung μ_n, d.h. $\mathbb{P}(Z_n \in B) = \mu_n(B)$ für $B \in \mathcal{B}$. Dann ist $\int f d\mu_n = \mathbb{E}(f(Z_n))$. Also bedeutet schwache Konvergenz von μ_n gegen μ, dass $\mathbb{E}(f(Z_n)) \to \mathbb{E}(f(Z))$ für alle stetigen und beschränkten Funktionen f, wobei $Z_n \sim \mu_n$ und $Z \sim \mu$. Auf welchem Raum Ω diese Zufallsvariablen definiert sind, spielt dabei keine Rolle.

Beispiel 4.3. *$\mu_n = \mathcal{N}(c, \frac{1}{n})$ konvergiert schwach gegen die Verteilung μ, die in c konzentriert ist (Dirac-Mass). Mit einer Substitution erhält man*

$$\int f d\mu_n - \int f d\mu = \frac{1}{\sqrt{2\pi}} \int (f(c + n^{-1/2}x) - f(c)) \exp(-1/(2x^2)) dx,$$

und die Behauptung folgt daher mit dem Konvergenzsatz von Lebesgue.

Lemma 4.1. *Seien μ und μ_n Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B})$ mit Verteilungsfunktionen F und F_n. Dann sind die folgenden Aussagen äquivalent:*

i) $\mu_n \longrightarrow \mu$ schwach.

ii) $F_n(x) \longrightarrow F(x)$ für jede Stetigkeitsstelle x von F.

iii) $\int f d\mu_n \longrightarrow \int f d\mu$ für alle $f \in C_b^3(\mathbb{R})$, wobei $C_b^3(\mathbb{R})$ die Menge aller dreimal stetig differenzierbaren Funktionen auf $\mathbb{R}$ bezeichnet, für die f, f', f'', f''' alle beschränkt sind.

Beispiel 4.4. *Fortsetzung von Beispiel 4.3: Die Verteilungsfunktion F_n von $\mathcal{N}(c, \frac{1}{n})$ ist $\Phi(\sqrt{n}(x-c))$, konvergiert also gegen 0 ($x < c$), bzw. gegen $\frac{1}{2}$ ($x = c$), bzw. gegen 1 ($x > c$). Die Verteilungsfunktion F des Dirac-Masses ist hingegen gleich 0 ($x < c$), bzw. gleich 1 ($x \geq c$), d.h. $F_n(c)$ konvergiert nicht gegen $F(c)$.*

Beweis Die Implikation "i) ⇒ iii)" ist klar.

Für die Implikation "iii) ⇒ ii)", nehmen wir an, es gelte $\int f d\mu_n \to \int f d\mu$ für alle $f \in C_b^3(\mathbb{R})$. Seien $x \in \mathbb{R}$ und $\delta > 0$ fest. Wir wählen ein $f \in C_b^3(\mathbb{R})$ mit

$$I_{(-\infty,x]} \leq f \leq I_{(-\infty,x+\delta]}$$

Dann

$$F_n(x) = \int I_{(-\infty,x]} d\mu_n \leq \int f d\mu_n$$

und

$$\int f d\mu \leq \int I_{(-\infty,x+\delta]} d\mu = F(x+\delta).$$

Daraus folgt

$$\limsup F_n(x) \leq \lim \int f d\mu_n = \int f d\mu \leq F(x+\delta).$$

Analog folgt

$$\liminf F_n(x) \geq F(x-\delta).$$

Wenn jetzt F stetig ist an der Stelle x, dann folgt mit $\delta \to 0$

$$F(x) \le \liminf F_n(x) \le \limsup F_n(x) \le F(x),$$

also gilt überall "=" statt "≤".

Es bleibt noch die Implikation "ii) ⇒ i)" zu zeigen. Wir fixieren dazu ein stetiges, beschränktes f und ein $\varepsilon > 0$. Zunächst bemerken wir, dass die Menge der Stellen, wo F unstetig ist, höchstens abzählbar ist (für jedes k gibt es nur endlich viele Stellen, wo F einen Sprung mit einer Höhe in $(2^{-k}, 2^{-k-1}]$ hat). Wegen ii) gibt es also Stetigkeitsstellen a und b von F, so dass

$$\inf_n \mu_n([a,b]) > 1-\varepsilon, \quad \mu([a,b]) > 1-\varepsilon.$$

Ferner ist f gleichmässig stetig auf $[a,b]$, d.h. es gibt ein δ, so dass $|f(x)-f(y)| \le \varepsilon$ falls $a \le x, y \le b$ und $|x-y| \le \delta$. Wir wählen als nächstes ein m und Stetigkeitsstellen x_i mit $a = x_0 < x_1 \ldots < x_m = b$ und $x_{i+1} - x_i \le \delta$ und setzen

$$f_m = \sum_{i=1}^{m} f(x_{i-1}) I_{(x_{i-1}, x_i]}.$$

Dann gilt

$$\begin{aligned} |f(x) - f_m(x)| &\le \varepsilon \text{ falls } a < x \le b \\ |f(x) - f_m(x)| &\le \sup_x |f(x)| \text{ sonst,} \end{aligned}$$

also ist

$$\left| \int (f - f_m) d\mu_n(x) \right| \le \varepsilon (1 + \sup_x |f(x)|)$$

(und analog mit μ anstelle von μ_n). Ferner ist

$$\int f_m d\mu_n = \sum_{i=1}^{m} f(x_{i-1})(F_n(x_i) - F_n(x_{i-1})),$$

also konvergiert wegen ii) $\int f_m d\mu_n$ gegen $\int f_m d\mu$. Damit ist für n gross genug

$$\begin{aligned} \left| \int f d\mu_n - \int f d\mu \right| &\le \left| \int (f - f_m) d\mu_n \right| + \left| \int f_m d\mu_n - \int f_m d\mu \right| + \left| \int (f - f_m) d\mu \right| \\ &\le \varepsilon (3 + 2 \sup_x |f(x)|). \end{aligned}$$

□

Die Summe zweier unabhängiger normalverteilter Zufallsvariablen ist wieder normalverteilt:

Lemma 4.2. *Wenn X_1, X_2 unabhängig sind und $X_i \sim \mathcal{N}(m_i, \sigma_i^2)$ $(i = 1, 2)$, dann ist $X_1 + X_2 \sim \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.*

Beweis Nachrechnen mit Hilfe von (3.44) und quadratischem Ergänzen. □

Die Normalverteilung ist also ein Fixpunkt bei der Summation von i.i.d. Zufallsvariablen. Der **Zentrale Grenzwertsatz** besagt nun, dass man bei Summation von i.i.d. Zufallsvariablen mit endlichem zweiten Moment stets gegen diesen Fixpunkt konvergiert:

Satz 4.3. *Sei (X_i) i.i.d. mit $\mathbb{E}(X_i) = m$ und $\mathbb{V}(X_i) = \sigma^2 < \infty$. Dann konvergiert die Verteilung von S_n^* schwach gegen $\mathcal{N}(0,1)$, d.h. (gemäss Lemma 4.1, ii))*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - nm}{\sigma\sqrt{n}} \le x\right) = \Phi(x) \quad \forall x \in \mathbb{R}.$$

Wir leiten diesen Satz aus einem viel allgemeineren Resultat ab, bei dem die Summanden nicht i.i.d. sein müssen. Die Verteilung der Summanden darf sogar noch von n abhängen (daher die zwei Indizes n und i).

Satz 4.4 (Lindeberg)**.** *Seien $X_{n,i}$ $(1 \le i \le n, n \in \mathbb{N})$ Zufallsvariablen mit*

$$\begin{aligned}
&a) \quad X_{n,1}, \ldots, X_{n,n} \text{ sind unabhängig } \forall n;\\
&b) \quad \mathbb{E}(X_{n,i}) = 0, \quad \mathbb{E}\left(X_{n,i}^2\right) = \sigma_{n,i}^2 < \infty, \quad \sum_{i=1}^n \sigma_{n,i}^2 = 1;\\
&c) \quad \lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right) = 0 \quad \forall \varepsilon > 0.
\end{aligned}$$

Dann konvergiert die Verteilung von $S_n = X_{n,1} + \cdots + X_{n,n}$ schwach gegen $\mathcal{N}(0,1)$.

Beweis von Satz 4.3 aus Satz 4.4. Setze $X_{n,i} = (X_i - m)/(\sigma\sqrt{n})$. Dann sind a) und b) offensichtlich erfüllt. Ferner folgt mit dem Konvergenzsatz von Lebesgue, dass für alle $\varepsilon > 0$

$$\sum_{i=1}^n \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right) = \frac{1}{\sigma^2}\mathbb{E}\left((X_1 - m)^2 1_{[|X_1 - m| > \varepsilon\sigma\sqrt{n}]}\right) \longrightarrow 0 \quad (n \to \infty).$$

□

Bemerkung: Die Bedingung c) besagt, dass die $X_{n,i}$ klein sein müssen. Es gilt nämlich

$$\max_i \sigma_{n,i}^2 \le \max_i \left(\varepsilon^2 + \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right)\right) \le \varepsilon^2 + \sum_{i=1}^n \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right),$$

d.h. die Bedingung c) impliziert

$$\max_{1\le i\le n} \sigma_{n,i}^2 \longrightarrow 0 \text{ für } n \to \infty. \tag{4.11}$$

Wenn wir also $\varepsilon_n = \sqrt{\max_i \sigma_{n,i}}$ setzen, dann geht ε_n gegen Null und mit der Chebyshev-Ungleichung folgt $\mathbb{P}(|X_{n,i}| \le \varepsilon_n) \to 1$.

Beweis von Satz 4.4. Wir prüfen die Bedingung iii) von Lemma 4.1 nach. Sei also f dreimal stetig differenzierbar mit f, f', f'', f''' beschränkt. Wir wählen Zufallsvariablen $Y_{n,i} \sim \mathcal{N}(0, \sigma_{n,i}^2)$ derart, dass für jedes n

$$Y_{n,1}, \ldots, Y_{n,n}, X_{n,1}, \ldots, X_{n,n} \text{ unabhängig sind.}$$

Dann ist wegen Lemma 4.2 $\sum_{i=1}^n Y_{n,i} \sim \mathcal{N}(0,1)$, also muss man $\mathbb{E}\left(f(\sum_{i=1}^n X_{n,i}) - f(\sum_{i=1}^n Y_{n,i})\right)$ abschätzen Wir tauschen dazu sukzessive ein $X_{n,i}$ gegen ein $Y_{n,i}$ aus.

Sei

$$Z_{n,i} = X_{n,1} + \cdots + X_{n,i-1} + Y_{n,i+1} + \cdots + Y_{n,n}.$$

Mit einer Taylor-Entwicklung folgt dann

$$\begin{aligned}
f(Z_{n,i} + X_{n,i}) - f(Z_{n,i} + Y_{n,i}) \;=\; & f'(Z_{n,i})(X_{n,i} - Y_{n,i}) + \frac{1}{2} f''(Z_{n,i})(X_{n,i}^2 - Y_{n,i}^2)\\
& + R_3(Z_{n,i}, X_{n,i}) + R_1(Z_{n,i}, Y_{n,i}),
\end{aligned}$$

wobei

$$\begin{aligned} |R_1(z,x)| &= |x^3 f'''(z+\theta x)\frac{1}{6}| \le \text{const}\,.\,|x|^3, \quad \text{bzw} \\ |R_2(z,x)| &= |x^2\frac{1}{2}(f''(z+\theta x) - f''(z))| \le \text{const}\,.\,x^2 \end{aligned}$$

und

$$|R_3(z,x)| = |R_1(z,x) + R_2(z,x)| \le \text{const}\,.\,(\varepsilon x^2 + x^2 1_{[|x|>\varepsilon]}).$$

gelten, und die Gleichungen für alle x, z mit von f abhängigen Konstanten gültig sind.

Weil nach Konstruktion $Z_{n,i}$ sowohl von $X_{n,i}$ als auch von $Y_{n,i}$ unabhängig ist, folgt daraus

$$\begin{aligned} &\left|\mathbb{E}\left(f(Z_{n,i}+X_{n,i}) - f(Z_{n,i}+Y_{n,i})\right)\right| \le \left|\mathbb{E}\left(f'(Z_{n,i})\right)\left(\mathbb{E}\left(X_{n,i}\right) - \mathbb{E}\left(Y_{n,i}\right)\right)\right| \\ + \;& \left|\frac{1}{2}\mathbb{E}\left(f''(Z_{n,i})\right)\left(\mathbb{E}\left(X_{n,i}^2\right) - \mathbb{E}\left(Y_{n,i}^2\right)\right)\right| + \mathbb{E}\left(|R(Z_{n,i},X_{n,i})|\right) + \mathbb{E}\left(|R(Z_{n,i},Y_{n,i})|\right) \\ \le \;& \text{const}\,.\Big(\varepsilon\sigma_{n,i}^2 + \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right) + \mathbb{E}\left(|Y_{n,i}|^3\right)\Big). \end{aligned}$$

Damit ist für alle $\varepsilon > 0$

$$\begin{aligned} \left|\mathbb{E}\left(f(\sum_i X_{n,i}) - f(\sum_i Y_{n,i})\right)\right| &= \left|\sum_{i=1}^n \mathbb{E}\left(f(Z_{n,i}+X_{n,i}) - f(Z_{n,i}+Y_{n,i})\right)\right| \\ &\le \text{const}\,.\left(\varepsilon + \sum_{i=1}^n \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right) + \sum_{i=1}^n \mathbb{E}\left(|Y_{n,i}|^3\right)\right). \end{aligned}$$

Die rechte Seite ist $\le 3\,\text{const}\,.\varepsilon$ für n gross genug wegen Voraussetzung c), bzw. weil gilt

$$\sum_{i=1}^n \mathbb{E}\left(|Y_{n,i}|^3\right) = \sum_{i=1}^n \sigma_{n,i}^3\sqrt{\frac{8}{\pi}} \le \max(\sigma_{n,i})\sqrt{\frac{8}{\pi}},$$

was wegen (4.11) gegen null konvergiert. □

Korollar 4.2. *Sei (X_i) eine i.i.d. Folge von d-dimensionalen Zufallsvektoren mit Verteilung μ und sei $f \in L^2(\mathbb{R}^d, \mu)$. Dann gilt*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n f(X_i) = \int_{\mathbb{R}^d} f(x)\mu(dx) = \mathbb{E}\left(f(X_i)\right) =: m, \quad \mathbb{P}\text{-f.s.}$$

und

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\frac{1}{n}\sum_{i=1}^n f(X_i) - m}{\sigma_f/\sqrt{n}} \le x\right) = \Phi(x) \quad \forall x \in \mathbb{R},$$

wobei $\sigma_f^2 := Var[f(X_i)] = \int_{\mathbb{R}^d}(f(x)-m)^2\mu(dx)$. Das heisst der Fehler

$$\frac{1}{n}\sum_{i=1}^n f(X_i) - m$$

ist approximativ normalverteilt mit Mittel 0 und Standardabweichung $\sigma_f/\sqrt{n}$.

Beweis Die erste Aussage folgt aus dem starken Gesetz der grossen Zahlen und die zweite aus dem Zentralen Grenzwertsatz. □

Man kann also Integrale approximieren durch die Erzeugung von Zufallszahlen (Monte-Carlo Verfahren). Beachtenswert ist, dass die Genauigkeit nur von der Anzahl Replikate

```
> monte.carlo(1000,1000)
```

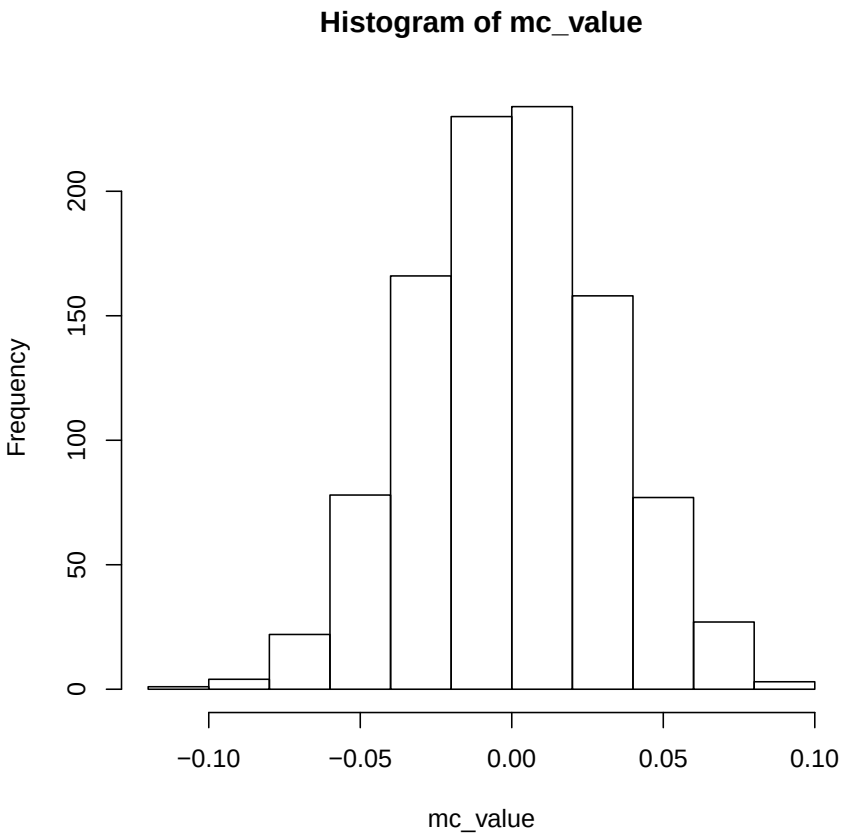

Abbildung 4.3: Fehler eines Monte-Carlo Verfahrens rund um den wahren Wert des Integrals

n und nicht von der Dimension d abhängt. In hohen Dimensionen ist Monte Carlo deterministischen Verfahren zur Approximation von Integralen überlegen ! Wir illustrieren das Verfahren hier jedoch mit einem eindimensionalen Beispiel: In der folgender Abbildung 4.3 sehen wir eine Monte Carlo Berechnung des Integrals $\int_{-1}^{1} \arctan(x)dx = 0$ mit $n = 1000$ Zufallszahlen und $K = 1000$ verschiedenen Versuchen sowie dem Histogramm der Fehler. Die Standardabweichung des Fehlers ist proportional zu $1/\sqrt{n}$.

```
> monte.carlo<-function(n,K)
+ {
+ mc_value<-rep(0, times=K);
+ for (i in 1:K)
+ {
+ x=runif(n,-1,1)
+ mc_value[i]=2/n*sum(atan(x))
+ }
+ hist(mc_value)
+ }
```

Beispiel 4.5. Rundungsfehler. *Seien* $X_1, X_2, \ldots, X_n$ *i.i.d.* $\mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$. *Dann ist* $\mathbb{E}(X_i) =$

0 *und* $\mathbb{V}(X_i) = \frac{1}{12}$, *also mit Satz 4.3:*

$$\mathbb{P}(a \le S_n \le b) = \mathbb{P}\left(a\sqrt{\frac{12}{n}} \le S_n^* \le b\sqrt{\frac{12}{n}}\right) \approx \Phi\left(b\sqrt{\frac{12}{n}}\right) - \Phi\left(a\sqrt{\frac{12}{n}}\right).$$

Wenn Rundungsfehler als unabhängig und gleichverteilt angenommen werden können, dann ist die Wahrscheinlichkeit, dass bei der Addition von $n = 100$ *Zahlen höchstens eine Stelle verloren geht, gleich*

$$\mathbb{P}(-5 < S_{100} < 5) \approx \Phi(\sqrt{3}) - \Phi(-\sqrt{3}) = 0.917$$

(im schlimmsten Fall sind es zwei Stellen).

Beispiel 4.6. Asymptotik des Medians. *Dieses Beispiel zeigt die Flexibilität des Satzes von Lindeberg. Der* **Median** m *einer Verteilung* F *wurde definiert durch* $m = F^{-1}(\frac{1}{2})$. *Seien* $X_1, X_2, \ldots$ *i.i.d. mit Verteilungsfunktion* F *und Median* $m = 0$. *Ferner soll* $F'(0)$ *existieren und* > 0 *sein. Das heisst, dass Beobachtungen in der Nähe des Medians auftreten werden. Ferner sei* Z_n *der sogenannte* **Stichprobenmedian** *von* $X_1, \ldots, X_n$, *d.h.* Z_n *ist die mittlere Beobachtung, oder formelmässig* $Z_n = X_{(k)}$ *mit* $k = [\frac{n}{2} + 1]$, *wobei* $X_{(1)} \le \cdots \le X_{(n)}$ *die der Grösse nach geordneten Zufallsvariablen* $X_1, \ldots, X_n$ *bezeichnen und* $[x]$ *den ganzzahligen Teil von* x. *Wir behaupten, dass die Verteilung von* $\sqrt{n}Z_n$ *schwach gegen* $\mathcal{N}(0, \frac{1}{4F'(0)^2})$ *konvergiert, d.h.*

$$\mathbb{P}(\sqrt{n}\ Z_n \le x) \to \Phi(2F'(0)x).$$

Beweis *Wir setzen* $Y_{n,i} = 1_{[X_i > x/\sqrt{n}]}$. *Dann ist*

$$\sqrt{n}\ Z_n \le x \Leftrightarrow X_{(k)} \le \frac{x}{\sqrt{n}} \Leftrightarrow \sum_{i=1}^{n} Y_{n,i} \le n - k.$$

Es gilt $\mathbb{E}(Y_{n,i}) = p_n$, $\mathbb{V}(Y_{n,i}) = p_n(1-p_n)$ *mit* $p_n = 1 - F(\frac{x}{\sqrt{n}})$. *Wir standardisieren die* $Y_{n,i}$, *um nachher den Satz von Lindeberg anwenden zu können:*

$$X_{n,i} = \frac{Y_{n,i} - p_n}{\sqrt{np_n(1-p_n)}}.$$

Da

$$\sqrt{n}\ Z_n \le x \Leftrightarrow S_n = \sum_{i=1}^{n} X_{n,i} \le a_n = \frac{n - k - np_n}{\sqrt{np_n(1-p_n)}} \longrightarrow 2F'(0)x,$$

gilt für $\delta > 0$ *fest und* n *gross genug*

$$\mathbb{P}(S_n \le 2F'(0)x - \delta) \le \mathbb{P}(S_n \le a_n) = \mathbb{P}(\sqrt{n}Z_n \le x) \le \mathbb{P}(S_n \le 2F'(0)x + \delta).$$

Andrerseits $\mathbb{P}(S_n \le 2F'(0)x \pm \delta) \longrightarrow \Phi(2F'(0)x \pm \delta)$ *wegen Satz 4.4, und daraus folgt die Behauptung.* □

Man sieht aus dem Beweis, dass das Resultat gültig bleibt, solange $k = \frac{n}{2} + o(\sqrt{n})$, *insbesondere also für leicht andere Definitionen des Stichprobenmedians.*

Kapitel 5

Charakteristische Funktionen

Im Folgenden bezeichnen wir mit $\langle x, y\rangle$ das Skalarprodukt für $x, y \in \mathbb{R}^n$, d.h., für $x = (x_1, \ldots, x_n)$ und $y = (y_1, \ldots, y_n)$ ist

$$\langle x, y\rangle = \sum_{j=1}^{n} x_j y_j.$$

Definition 5.1. *Die charakteristische Funktion $\varphi_X : \mathbb{R}^n \to \mathbb{C}$ eines Zufallsvektors $X : \Omega \to \mathbb{R}^n$ ist durch*

$$\varphi_X(u) := \mathbb{E}\left(e^{\mathrm{i}\langle u, X\rangle}\right) = \int_{\mathbb{R}^n} e^{\mathrm{i}\langle u, x\rangle}\mu(dx), \quad u \in \mathbb{R}^n \tag{5.1}$$

definiert, wobei μ die Verteilung von X bezeichnet.

In Definition 5.1 bezeichnet i die imaginäre Einheit der komplexen Zahlen. Für jedes $u \in \mathbb{R}^n$ integrieren wir in (5.1) die komplexwertige Funktion $x \mapsto e^{\mathrm{i}\langle u, x\rangle}$. Diese Funktion hat folgende Zerlegung in Real- und Imaginärteil:

$$e^{\mathrm{i}\langle u, x\rangle} = \cos(\langle u, x\rangle) + \mathrm{i}\sin(\langle u, x\rangle), \quad x \in \mathbb{R}^n.$$

Ausführlich geschrieben bedeutet die Definition (5.1) also

$$\int_{\mathbb{R}^n} \cos(\langle u, x\rangle)\mu(dx) + \mathrm{i}\int_{\mathbb{R}^n} \sin(\langle u, x\rangle)\mu(dx).$$

Beachte, dass die reellwertigen Funktionen $x \mapsto \cos(\langle u, x\rangle)$ und $x \mapsto \sin(\langle u, x\rangle)$ Borelmessbar und beschränkt sind, damit also μ-integrierbar sind. Folglich existiert die charakteristische Funktion $\varphi_X(u)$ für jede Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}^n$.

Beachte den Zusammenhang zwischen der Fouriertransformation der harmonischen Analysis und der charakteristischen Funktion: wenn die Zufallsvariable X eine Dichte p bezüglich dem Lebesguemass hat, dann entspricht die charakteristische Funktion – bis auf eine allfällige Konstante – der Fouriertransformierten von p,

$$\varphi_X(u) := \mathbb{E}\left(e^{\mathrm{i}\langle u, X\rangle}\right) = \int_{\mathbb{R}^n} e^{\mathrm{i}\langle u, x\rangle}p(x)dx.$$

Bevor wir mit der Theorie über charakteristische Funktionen fortfahren, betrachten wir einige konkrete Beispiele. Wir werden nur die ersten vier Beispiele durchrechnen, und uns bei den verbleibenden damit begnügen, das Ergebnis zu präsentieren.

1. Für eine Bernoulli-verteilte Zufallsvariable X mit Parameter p erhalten wir
$$\varphi_X(u) = \mathbb{E}\left(e^{\mathrm{i}\langle u,X\rangle}\right) = e^{\mathrm{i}u0}(1-p) + e^{\mathrm{i}u}p = pe^{iu} + 1 - p.$$

2. Allgemeiner erhalten wir für eine binomialverteilte Zufallsvariable X mit Parametern n und p die charakteristische Funktion
$$\varphi_X(u) = \mathbb{E}\left(e^{\mathrm{i}\langle u,X\rangle}\right) = \sum_{k=0}^{n}\binom{n}{k}e^{\mathrm{i}uk}p^k(1-p)^{n-k} = (pe^{\mathrm{i}u} + 1 - p)^n.$$

3. Für eine Poisson-verteilte Zufallsvariable mit Parameter $\lambda > 0$ berechnen wir die charakteristische Funktion als
$$\begin{aligned}\varphi_X(u) = \mathbb{E}\left(e^{\mathrm{i}\langle u,X\rangle}\right) &= \sum_{k=0}^{\infty} e^{\mathrm{i}uk}\mathbb{P}(X=k) = \sum_{k=0}^{\infty} e^{\mathrm{i}uk}\frac{\lambda^k}{k!}e^{-\lambda}\\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{\mathrm{i}u})^k}{k!} = e^{-\lambda}e^{\lambda e^{\mathrm{i}u}} = e^{\lambda(e^{\mathrm{i}u}-1)}.\end{aligned}$$

4. Für ein $a > 0$ sei X uniform auf dem Intervall $(-a, a)$ verteilt. Dann bekommen wir die charakteristische Funktion
$$\begin{aligned}\varphi_X(u) = \mathbb{E}\left(e^{\mathrm{i}\langle u,X\rangle}\right) &= \frac{1}{2a}\int_{-a}^{a} e^{\mathrm{i}ux}dx = \frac{e^{\mathrm{i}ua} - e^{-\mathrm{i}ua}}{2aiu}\\ &= \frac{2\mathrm{i}\sin au}{2aiu} = \frac{\sin au}{au}.\end{aligned}$$

5. Für eine normalverteilte Zufallsvariable $X \sim \mathcal{N}(0,1)$ erhalten wir
$$\varphi_X(u) = e^{-u^2/2}.$$

6. Hieraus können wir, unter Benutzung der zweiten Aussage von Lemma 5.1 unten, die charakteristische Funktion einer allgemeinen normalverteilen Zufallsvariablen $X \in \mathcal{N}(\mu, \sigma^2)$ berechnen. Sie lautet
$$\varphi_X(u) = e^{\mathrm{i}u\mu - u^2\sigma^2/2}.$$

7. Für eine exponentialverteilte Zufallsvariable X mit Parameter $\lambda > 0$ erhalten wir
$$\varphi_X(u) = \frac{\lambda}{\lambda - \mathrm{i}u}.$$

8. Allgemeiner erhalten wir für eine Gamma-verteilte Zufallsvariable X mit Parametern $\alpha, \lambda > 0$ die charakteristische Funktion
$$\varphi_X(u) = \left(\frac{\lambda}{\lambda - \mathrm{i}u}\right)^{\alpha}.$$

9. Für eine Cauchy-verteilte Zufallsvariable X bekommen wir
$$\varphi_X(u) = e^{-|u|}.$$

Wir zählen nun einige elementare Eigenschaften charakteristischer Funktionen auf.

Lemma 5.1. *Es sei $X : \Omega \to \mathbb{R}^n$ eine Zufallsvariable.*

i) Die charakteristische Funktion $\varphi_X : \mathbb{R}^n \to \mathbb{C}$ ist stetig mit $\varphi_X(0) = 1$ und $|\varphi_X(u)| \leq 1$ für alle $u \in \mathbb{R}^n$.

ii) Es seien $A \in \mathbb{R}^{m\times n}$ eine Matrix und $b \in \mathbb{R}^m$ ein Vektor. Dann gilt für die charakteristische Funktion der neuen $\mathbb{R}^m$-wertigen Zufallsvariablen $AX + b$ die Rechenregel

$$\varphi_{AX+b}(u) = \varphi_X(A^\top u)e^{\mathrm{i}\langle u,b\rangle}, \quad u \in \mathbb{R}^m$$

wobei $A^\top \in \mathbb{R}^{n\times m}$ die zu A transponierte Matrix bezeichnet.

iii) Für zwei unabhängige Zufallsvariablen X und Y gilt

$$\varphi_{X+Y}(u) = \varphi_X(u)\varphi_Y(u), \quad u \in \mathbb{R}^n.$$

iv) Die charakteristische Funktion φ_X ist genau dann reellwertig, wenn die Verteilung von X symmetrisch ist.

Die charakteristische Funktion ist für die Berechnung von Momenten nützlich. Die folgenden beiden Aussagen sind Konsequenzen von Satz 5.1 unterhalb.

- Für eine reellwertige Zufallsvariable X mit $\mathbb{E}(|X|) < \infty$ gilt

$$\mathbb{E}(X) = -\mathrm{i}\varphi'_X(0).$$

- Für eine reellwertige Zufallsvariable X mit $\mathbb{E}(X^2) < \infty$ gilt

$$\mathbb{E}\left(X^2\right) = -\varphi''_X(0).$$

Satz 5.1. *Es sei $X : \Omega \to \mathbb{R}^n$ eine Zufallsvariable mit $\mathbb{E}(\|X\|^m) < \infty$ für ein $m \in \mathbb{N}$. Dann ist $\varphi_X : \mathbb{R}^n \to \mathbb{C}$ eine C^m-Funktion, und für alle $m_1, \ldots, m_n \in \mathbb{N}_0$ mit $m_1 + \ldots + m_n = m$ gilt*

$$\frac{\partial^m}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}\varphi_X(u) = \mathrm{i}^m\mathbb{E}\left(X_1^{m_1} \cdots X_n^{m_n} e^{\mathrm{i}\langle u,X\rangle}\right), \quad u \in \mathbb{R}^n.$$

Der folgende Eindeutigkeitssatz zeigt, dass die charakteristische Funktion φ_X die Wahrscheinlichkeitsverteilung charakterisiert.

Satz 5.2. *Für zwei Zufallsvariablen X_1 und X_2 mit Verteilungen μ_1 und μ_2 auf $\mathbb{R}^n$ und $\varphi_{X_1} = \varphi_{X_2}$ gilt $\mu_1 = \mu_2$.*

Hier ist eine nützliche Konsequenz dieses Theorems.

Korollar 5.1. *Es sei $X = (X_1, \ldots, X_n)$ eine $\mathbb{R}^n$-wertige Zufallsvariable. Dann sind die reellwertigen Zufallsvariablen $X_1, \ldots, X_n$ genau dann unabhängig, wenn für alle $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ gilt*

$$\varphi_X(u_1, \ldots, u_n) = \prod_{j=1}^{n} \varphi_{X_j}(u_j).$$

Nicht jede komplexwertige Funktion $\varphi : \mathbb{R}^n \to \mathbb{C}$ ist zwangsläufig die charakteristische Funktion einer Zufallsvariablen X. Das folgende Resultat, bekannt als Theorem von Bochner, gibt ein Kriterium dafür an, wann eine gegebene komplexwertige Funktion tatsächlich eine charakteristische Funktion ist.

Satz 5.3. *Eine Funktion $\varphi : \mathbb{R}^n \to \mathbb{C}$ ist die charakteristische Funktion einer Zufallsvariablen X genau dann, wenn $\varphi(0) = 1$ gilt, die Funktion in $u = 0$ stetig ist und nichtnegativ definit ist, d.h. für alle $m \in \mathbb{N}$ und alle $u_1, \ldots, u_m \in \mathbb{R}^n$ und $z_1, \ldots, z_m \in \mathbb{C}$ gilt*

$$\sum_{j=1}^{m} \sum_{k=1}^{m} \varphi(u_j - u_k) z_j \overline{z}_k \geq 0.$$

Mit Hilfe des Eindeutigkeitssatzes (Theorem 5.2) und der dritten Aussage von Lemma 5.1 lassen sich leicht folgende Aussagen über die Verteilung von Summen unabhängiger Zufallsvariablen beweisen. Die meisten davon haben wir früher bereits mit der Faltungsformel hergeleitet, was eher umständlich ist.

1. Es seien $X_1, \ldots, X_n$ unabhängig und Bernoulli-verteilt mit Parameter p. Dann ist die Summe $X_1 + \ldots + X_n$ binomialverteilt mit Parametern (n, p).
2. Es seien X, Y unabhängig und Poisson-verteilt mit Parametern $\lambda, \mu > 0$. Dann ist die Summe $X + Y$ Poisson-verteilt mit Parameter $\lambda + \mu$.
3. Es seien X, Y unabhängig und binomialverteilt mit Parametern (n, p) und (m, p). Dann ist die Summe $X + Y$ binomialverteilt mit Parametern $(m + n, p)$.
4. Es seien $X \sim \mathcal{N}(\mu, \sigma^2)$ und $Y \sim \mathcal{N}(\nu, \tau^2)$ unabhängig und normalverteilt. Dann ist auch die Summe $X + Y$ normalverteilt, und zwar gilt $X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \tau^2)$.
5. Es seien X, Y unabhängig und Gamma-verteilt mit Parametern (α, λ) und (β, λ). Dann ist die Summe $X + Y$ Gamma-verteilt mit Parametern $(\alpha + \beta, \lambda)$.
6. Es seien X, Y unabhängig und Cauchy-verteilt. Dann ist das arithmetische Mittel $\frac{1}{2}(X + Y)$ wieder Cauchy-verteilt.

Satz 5.4 (Lévys Stetigkeitssatz)**.** *Sei (μ_n) eine Folge von Verteilungen reellwertiger Zufallsvariablen (X_n) und (φ_{X_n}) bezeichne deren charakteristische Funktionen. Dann gilt:*

i) Falls μ_n schwach gegen eine Verteilung μ einer Zufallsvariablen X konvergiert, d.h. für alle beschränkten und stetigen Funktionen $f : \mathbb{R} \to \mathbb{R}$ gilt

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} f(x)\mu(dx),$$

dann folgt punktweise Konvergenz der charakteristischen Funktionen, d.h. $\varphi_{X_n}(u) \to \varphi_X(u)$ punktweise für alle $u \in \mathbb{R}$.

(ii) Falls $\varphi_{X_n}(u)$ punktweise gegen eine Funktion $g(u)$ konvergiert und diese zusätzlich stetig an 0 ist, dann ist g die charakteristische Funktion einer Zufallsvariablen X: $g(u) = \varphi_X(u)$, und μ_n konvergiert schwach gegen die Verteilung μ von X.

Zum Schluss geben wir einen alternativen Beweis für den zentralen Grenzwertsatz 4.3, der obige Resultate über charakteristische Funktionen benutzt. Dieser Beweis stellt einen direkten Zusammenhang zwischen der Exponentialfunktion einerseits und der Normalverteilung andererseits her, der allen Überlegungen zum zentralen Grenzwertsatz zugrunde liegt.

Aus

$$S^*_{n_1+n_2} = \sqrt{\frac{n_1}{n_1+n_2}} S^*_{n_1} + \sqrt{\frac{n_2}{n_1+n_2}} \frac{\sum_{i=1}^{n_2} X_{n_1+i} - n_2 m}{\sqrt{n_2}\sigma}$$

folgt

$$\varphi_{S^*_{n_1+n_2}}(u) = \varphi_{S^*_{n_1}}\left(u\sqrt{\frac{n_1}{n_1+n_2}}\right) \varphi_{S^*_{n_2}}\left(u\sqrt{\frac{n_2}{n_1+n_2}}\right).$$

Wenn also die Verteilung von S_n^* schwach gegen eine Verteilung mit charakteristischer Funktion φ konvergiert, dann muss wegen i) von Satz 5.4 gelten

$$\varphi(u) = \varphi(u\sqrt{\alpha})\varphi(u\sqrt{1-\alpha})$$

(man lässt n_1 und n_2 gegen unendlich gehen mit $n_1/n_2 \to \alpha$). Mit der Transformation $g(v) := \varphi(\sqrt{v})$ für $v > 0$, erhalten wir also die Cauchy'sche Funktionalgleichung

$$g(v)g(w) = g(v+w) \quad (v, w \geq 0).$$

Die Exponentialfunktion der reellen Analysis ist die einzige, stetige Lösung dieser Gleichung mit $g(0) = 1$ Durch Rücktransformation $\varphi(t) = g(t^2)$, und weil die charakteristische Funktion die Verteilung charakterisiert, kommt man somit auf die Normalverteilung als einzig möglicher Grenzverteilung.

Sei jetzt φ_i die charakteristische Funktion von $X_i - m$. Da die Zufallsvariablen (X_i) i.i.d. sind, hängt φ_i nicht von i ab und wir schreiben der Einfachheit halber φ. Wir zeigen nun dass die charakteristische Funktion von S_n^* punktweise gegen $\exp(-u^2/2)$ konvergiert. Aus der Unabhängigkeit der (X_i) folgt

$$\begin{aligned}
\varphi_{S_n^*}(u) &= \mathbb{E}\left(\exp\left(\mathrm{i}u\frac{\sum_{i=1}^n (X_i - m)}{\sigma\sqrt{n}}\right)\right) \\
&= \mathbb{E}\left(\prod_{i=1}^n \exp\left(\frac{\mathrm{i}u}{\sigma\sqrt{n}}(X_i - m)\right)\right) \\
&= \varphi\left(\frac{u}{\sigma\sqrt{n}}\right)^n
\end{aligned}$$

Nach Voraussetzung existieren erstes und zweites Moment der Zufallsvariablen $X_i - m$, weshalb φ zweimal stetig differenzierbar ist (siehe Satz 5.1). Da $\varphi'(0) = \mathrm{i}\mathbb{E}\,(X_i - m) = 0$ und $\varphi''(0) = -\sigma^2$, ist die Taylorentwicklung von φ um 0 durch

$$\varphi(v) = 1 + 0 - \frac{\sigma^2 v^2}{2} + v^2 h(v)$$

gegeben, wobei $h(v) \to 0$ für $v \to 0$. Mit einem Standardargument der Analysis folgt daher

$$\begin{aligned}
\lim_{n\to\infty} \varphi_{S_n^*}(u) &= \lim_{n\to\infty} \varphi\left(\frac{u}{\sigma\sqrt{n}}\right)^n \\
&= \lim_{n\to\infty} \left(1 - \frac{u^2}{2n} + o(1/n)\right)^n \\
&= \exp\left(-\frac{1}{2}u^2\right).
\end{aligned}$$

Da $u \mapsto e^{-\frac{1}{2}u^2}$ stetig an 0 ist, folgt mit dem Teil ii) von Lévys Stetigkeitsatz (siehe Theorem 5.4), dass die Verteilung von S_n^* schwach gegen eine Zufallsvariable Z konvergiert, deren charakteristische Funktion durch $\varphi_Z(u) = e^{-\frac{1}{2}u^2}$ gegeben ist. Da die charakteristische Funktion die Verteilung charakterisiert, ist $Z \sim \mathcal{N}(0,1)$ verteilt und die Aussage bewiesen. □

Kapitel 6

Einführung in die Statistik

6.1 Was ist Statistik?

Deskriptive Statistik fasst Datensätze zusammen und macht deren Besonderheiten sichtbar, mit Hilfe von Kennzahlen und Grafiken. Wir behandeln hier diesen Teil nicht.

Schliessende Statistik betrachtet die vorliegenden Beobachtungen als Realisierungen von Zufallsvariablen und zieht Rückschlüsse auf die zugrunde liegende Verteilung (und damit auch auf zukünftige Beobachtungen). Das Ziel ist es, Zufallsfehler und echte Effekte unter Angabe der verbleibenden Unsicherheit zu trennen.

Wahrscheinlichkeitstheorie ist deduktiver Natur, Statistik induktiver Natur. Statistik versucht zu erfassen, wie wir aus beschränkter Erfahrung lernen und verallgemeinern können. Wir geben zunächst drei Beispiele, welche die Fragestellungen der schliessenden Statistik illustrieren.

Beispiel 6.1. Aussersinnliche Wahrnehmung.
An der University of California in Davis wurde 1973 folgendes Experiment durchgeführt: Ein Computer wählte zufällig eines von 4 *Symbolen, und ein Medium versuchte, durch aussersinnliche Wahrnehmung das gewählte Symbol herauszufinden.* 15 *Medien machten je* 500 *Versuche und waren insgesamt* 2006 *mal erfolgreich.*

Die Grundfrage: "War hier aussersinnliche Wahrnehmung im Spiel?" lässt sich leider auch nicht mit Statistik beantworten, wohl aber die einfachere Frage: "Kann das Resultat Zufall sein?". Falls nur der Zufall spielt, dann ist die Anzahl Erfolge X binomial$(7500, \frac{1}{4})$-verteilt, und die Wahrscheinlichkeit, dass wir ein Ergebnis erhalten, welches mindestens so gut ist wie das tatsächlich vorliegende, lässt sich mit dem Zentralen Grenzwertsatz sehr genau approximieren:

$$\mathbb{P}\left(X \geq 2006\right) \approx 1 - \Phi\left(\frac{2006 - 1875}{\sqrt{7500 \cdot 0.25 \cdot 0.75}}\right) = 1 - \Phi(3.49) = 0.0002.$$

Entweder ist also etwas extrem Unwahrscheinliches eingetreten, oder die Hypothese "Reiner Zufall" ist falsch. Der Zufall ist also keine zufriedenstellende Erklärung. Andere mögliche Erklärungen sind "Schlechte Durchführung des Experiments" (z. B. der Zufallsgenerator war nicht in Ordnung), oder "Aussersinnliche Wahrnehmung hat stattgefunden"..

Es ist wesentlich, dass man die Wahrscheinlichkeit eines mindestens *so guten Resultates nimmt. Die Wahrscheinlichkeit, genau das vorliegende Resultat zu erhalten, ist nämlich in*

jedem Fall klein; aus dem Satz von de Moivre-Laplace folgt z.B.

$$\mathbb{P}(X = 1875) \approx \frac{1}{\sqrt{2\pi \cdot 7500 \cdot 0.25 \cdot 0.75}} = 0.0106.$$

Wenn wir einmal annehmen, dass aussersinnliche Wahrnehmung mit Wahrscheinlichkeit γ zustande kommt und dass in den andern Fällen einfach geraten wurde, erhalten wir folgendes Modell: X = Anzahl Erfolge $\sim$ Binomial$(7500, p)$ mit $p = \gamma + (1-\gamma)\frac{1}{4} = \frac{3}{4}\gamma + \frac{1}{4}$. Man möchte natürlich wissen, wie gross γ ist. Der naheliegende Schätzwert ist

$$\hat{p} = 2006/7500 = 0.267 \Rightarrow \hat{\gamma} = 0.023.$$

Vermutlich ist dies aber nicht das wahre γ. Wir sind eher an einem Intervall von $\gamma-$Werten interessiert, welches das wahre γ mit grosser Wahrscheinlichkeit $1-\alpha$ einfängt. Aus der Theorie von Abschnitt 4.4 wird folgen, dass $[0.006, 0.040]$ ein solches Intervall ist für $1-\alpha = 0.99$. Auch wenn also aussersinnliche Wahrnehmung im Spiel war, ist sie höchst unzuverlässig.

Für eine detailliertere Analyse verschiedener parapsychologischer Experimente, siehe J. Utts, Replication and Meta-Analysis in Parapsychology, Statistical Science **6** *(1991), 363-403.*

Beispiel 6.2. Vergleich zweier Lehrmethoden.
Die beiden Methoden wurden an je 10 Testpersonen mit ähnlicher Vorbildung und Intelligenz ausprobiert. Bei den Abschlussprüfungen ergaben sich die folgenden Resultate (bereits der Grösse nach geordnet):

Methode	*1:*	*3*	*6*	*18*	*25*	*37*	*48*	*49*	*51*	*81*	*89,*
Methode	*2:*	*9*	*34*	*40*	*57*	*61*	*64*	*75*	*91*	*93*	*98.*

Die wesentlichen Fragestellungen sind: Besteht ein echter Unterschied zwischen beiden Methoden, der nicht nur auf diesen einmaligen Test beschränkt bleibt ? Falls ja, wie viele Punkte macht dieser echte Unterschied aus ?

Modellierung: Wir nehmen an, dass die beobachteten Werte Realisierungen von unabhängigen Zufallsvariablen sind, und zwar
Methode 1: $X_1, \ldots, X_n$ i.i.d $\sim F$,
Methode 2: $Y_1, \ldots, Y_m$ i.i.d.$\sim G$.
Falls beide Methoden gleich gut sind, ist $F = G$ (sogenannte Nullhypothese), während der einfachste Fall eines echten Unterschieds lautet $G(x) = F(x - \Delta)$.

Statistische Methoden für die Frage, ob ein Unterschied besteht:

a) Berechne W =Anzahl Paare (i, j) mit $X_i < Y_j$. Wenn die Nullhypothese stimmt, ist $\mathbb{P}(X_i < Y_j) = \frac{1}{2}$, also $W \approx \frac{n \cdot m}{2}$. Somit verwerfen wir die Nullhypothese, wenn W zu stark von $\frac{n \cdot m}{2}$ abweicht (sogenannter Wilcoxon-Test). Im Beispiel nimmt W den Wert 73 an.

b) Berechne

$$T = \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{1}{n+m-2}\left(\sum (X_i - \overline{X})^2 + \sum (Y_i - \overline{Y})^2\right)}}$$

Wenn die Nullhypothese stimmt, ist $\mathbb{E}(X_i) = \mathbb{E}(Y_i)$, also $T \approx 0$. Somit verwerfen wir die Nullhypothese, wenn $|T|$ bzw. T zu gross ist. Die Definition von T wird verständlicher, wenn man beachtet, dass $\overline{X} - \overline{Y}$ die Varianz $\sigma^2(1/n + 1/m)$ hat,

wenn alle Zufallsvariablen unabhängig sind und die Varianz σ^2 haben. Die zweite Wurzel im Nenner ist einfach eine Schätzung von σ. Dieser Test heisst $t-$Test. Im Beispiel nimmt T den Wert 1.58 *an.*

Um zu einer Entscheidung zu gelangen, müssen wir wissen, wie gross unter der Nullhypothese die Wahrscheinlichkeit einer Abweichung ≥ 23, bzw. ≥ 1.58 ist. Wir schauen also wieder alle Abweichungen an, die mindestens so extrem sind wie die tatsächlich vorliegende. Rechnungen, auf die wir im Abschnitt 6.3.2 genauer eingehen, ergeben

$$\begin{array}{llll} \mathbb{P}(|W-50| \geq 23) & \approx 0.05, & \mathbb{P}(W \geq 73) & \approx 0.025. \\ \mathbb{P}(|T| > 1.58) & \approx 0.12, & \mathbb{P}(T > 1.58) & \approx 0.06. \end{array}$$

falls $F = G$. Man ist also an der Grenze, von dem, was noch als zufällige Abweichung plausibel erscheint. (Konventionell setzt man die Grenze bei 5% *an). Wir sehen auch, dass es zwei mögliche Interpretationen von "mindestens so extrem" gibt (mit, bzw. ohne Absolutbetrag), und dass der Wilcoxon-Test hier empfindlicher ist.*

Für die zweite Fragestellung nach der Grösse des Unterschieds verschieben wir die Y_i um soviel, dass W bzw. T den Wert $\frac{nm}{2}$ bzw. 0 *annimmt. Diese Verschiebung nehmen wir als Schätzung von Δ. Man erhält: $\widehat{\Delta} = 23$ (im Fall von W) bzw. $\widehat{\Delta} = 21.5$ (im Fall von T). Wie zuvor wäre ein Intervall von Δ-Werten, welches das wahre Δ mit vorgegebener Wahrscheinlichkeit einfängt, informativer, siehe Abschnitt 6.4.*

Beispiel 6.3. Blutdruck in Abhängigkeit von Alter und anderen Faktoren.
Ziele solcher Studien sind die Identifikation derjenigen Faktoren, die den Blutdruck beeinflussen, die Erkennung atypischer Fälle, die Prüfung blutdrucksenkender Medikamente etc.. Wir haben die folgenden Grössen

$$\begin{array}{rcl} Y_i & = & \text{Blutdruck der } i-\text{ten Versuchsperson} \\ x_{ij} & = & \text{Wert des } j-\text{ten Faktors der } i-\text{ten Person,} \end{array}$$

wobei zum Beispiel x_{i1} = Alter, x_{i2} = Gewicht, x_{i3} = Geschlecht kodiert als 0/1 *etc.. Als Modell postulieren wir*

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \tag{6.1}$$

d.h. eine lineare Beziehung mit zufälligen Fehlern. Wir nehmen an, dass wir alle wesentlichen Faktoren erfasst haben, so dass wir annehmen können $\mathbb{E}(\varepsilon_i) = 0$. Die Annahme der Linearität der Beziehung ist nicht ganz so einschränkend, wie man zunächst denken könnte, weil wir die x_{ij} transformieren oder durch Kombination neue Faktoren bilden können, z.B. $x_{i,p+1} = \text{Alter}^2$ oder $x_{i,p+1} = $ Alter mal Geschlecht etc.. Zur Überprüfung der Linearität sollte man aber mindestens die sogenannten **Streudiagramme** *$((Y_i, x_{ij});\ i = 1, \ldots, n)$ anschauen für jedes j.*

Die folgenden statistischen Fragen möchte man beantworten:

- *Wie kann man die unbekannten Parameter $\beta_0, \ldots, \beta_p$ schätzen? Welche von mehreren möglichen Schätzverfahren sind gut?*
- *Welche Parameter β_j können gleich null sein? (d.h. der entsprechende Faktor hat keinen Einfluss).*
- *In welchem Bereich wird der Blutdruck einer neuen Versuchsperson mit bestimmten Werten von $x_1, \ldots, x_p$ liegen?*

Das bekannteste Schätzverfahren besteht darin, die Summe der Fehlerquadrate

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j x_{ij} - \beta_0)^2 \tag{6.2}$$

zu minimieren (Kleinste Quadrate, Gauss). Die Lösung $\widehat{\beta} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^T$ *lässt sich formelmässig angeben als*

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

wobei Y *der* $n \times 1$ *Vektor* $(Y_1, \ldots, Y_n)^T$ *ist und* X *die* $n \times (p+1)$*-Matrix mit erster Spalte* $(1, \ldots, 1)'$ *und* $(j+1)$*-ter Spalte* $(x_{1j}, \ldots, x_{nj})'$ *(sofern* X *maximalen Rang hat). Die numerische Berechnung von* $\widehat{\beta}$ *ist ebenfalls bestens untersucht. Andere Schätzverfahren werden wir in Abschnitt 6.2.2 kennenlernen.*

Zusammenfassend halten wir fest, dass die mathematische Statistik mit Klassen von möglichen Verteilungen arbeitet. Diese enthalten einen **strukturellen Parameter**, der üblicherweise p reelle Komponenten hat und direkt mit der ursprünglichen Fragestellung zusammenhängt, sowie meist noch **Störparameter**, welche oft unendlichdimensional sind und nur mittelbar von Interesse. Im ersten Beispiel ist der struktureller Parameter p bzw. γ, und es gibt keinen Störparameter. Im zweiten Beispiel ist der strukturelle Parameter Δ und der Störparameter F. Im dritten Beispiel ist der strukturelle Parameter $\beta_0, \ldots, \beta_p$ und der Störparameter die Verteilung der ε_i's, bzw. deren Varianzen und Kovarianzen, falls wir Normalverteilung annehmen können.

Man unterscheidet ferner die folgenden 3 wichtigsten Problemstellungen:

- **Punktschätzung** eines unbekannten Parameters.
- Prüfverfahren, ob vorgegebene Parameterwerte mit den Daten verträglich sind (**statistischer Test**).
- Angabe von Schranken, die einen unbekannten Parameter mit vorgegebener Wahrscheinlichkeit eingrenzen (**Vertrauens- oder Konfidenzintervall**).

Daneben gibt es z.B. noch Tests für eine bestimmte Verteilung oder einen bestimmten Verteilungstyp (z.B. "Sind die Daten Poisson-verteilt mit beliebigem λ"?).

6.2 Punktschätzungen

Wir verwenden den folgenden Formalismus: Die Beobachtungen seien eine Realisierung eines $\mathbb{R}^n-$wertigen Zufallsvektors $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Als Modellverteilungen für $\mathbf{X}$ haben wir eine Klasse von Verteilungen $(\mu_\theta)_{\theta\in\Theta}$ auf $(\mathbb{R}^n, \mathcal{B}^n)$ zur Verfügung. Eine Punktschätzung von θ ist dann eine im Prinzip beliebige Abbildung

$$T : \mathbb{R}^n \to \Theta.$$

Wir legen also fest, wie unsere Schätzung für beliebige mögliche Daten, nicht nur für die konkret vorliegenden Werte, aussehen würde.

Weil $\mathbf{X}$ ein Zufallsvektor ist, ist $T(\mathbf{X})$ ein zufälliges Element von Θ. Um von der Verteilung von $T(\mathbf{X})$ sprechen zu können, brauchen wir eine σ-Algebra auf Θ, und wir müssen T als messbar voraussetzen. In allen praktisch relevanten Fällen ist das nie ein Problem, und wir nehmen an, dass die Verteilung $\nu_\theta[B] = \mathbb{P}_\theta(T \in B) = \mu_\theta(T^{-1}(B))$ $(B \subset \Theta)$ definiert ist.

Oft ist man aber gar nicht am ganzen Parameter θ interessiert, sondern nur an gewissen Komponenten, dem sogenannten strukturellen Parameter, während man zusätzliche Störparameter gar nicht zu schätzen braucht (vergleiche die vorangegangene Diskussion). Dies berücksichtigen wir, indem wir den Parameter von Interesse $\eta = g(\theta)$ einführen, wobei g meist eine Projektion auf einen niedrig-dimensionalen Raum darstellt. Wenn wir einen Schätzer T von θ haben, dann schätzen wir $\eta = g(\theta)$ meist durch $U = g(T) = g(T(\mathbf{X}))$. Es gibt aber wie gesagt Situationen, wo es einfacher ist, direkt einen guten Schätzer U von η zu konstruieren.

6.2.1 Beurteilung von Schätzern

Wenn wir an $\eta = g(\theta)$ interessiert sind, dann sollte für einen guten Schätzer U von η die Verteilung von U möglichst um $\eta = g(\theta)$ herum konzentriert sein, und zwar nicht nur für ein festes θ, sondern für alle $\theta \in \Theta$.

Wir nehmen hier an, dass $g : \theta \to \mathbb{R}$, d.h. wir sind an einer Komponente von θ interessiert. Das übliche Kriterium ist dann der **Mittlere Quadratische Fehler** (MSE=**m**ean **s**quare **e**rror)

$$\mathbb{E}_\theta \left(|U - g(\theta)|^2 \right),$$

doch gäbe es auch andere Kriterien wie z. B. $\mathbb{P}_\theta \left(|U - g(\theta)| > a \right)$ für ein festes a oder $\mathbb{E}_\theta \left(|U - g(\theta)| \right)$.

Wir führen den **Systematischen Schätzfehler/Bias**

$$b_U(\theta) = \mathbb{E}_\theta \left(U \right) - g(\theta)$$

ein und den sogenannten **Standardfehler**

$$\sigma_U(\theta) = \sqrt{\mathbb{V}_\theta \left(U \right)},$$

der die Grösse der zufälligen Schwankungen von U angibt. Dann können wir den mittleren quadratischen Fehler wie folgt zerlegen:

$$\mathbb{E}_\theta \left(|U - g(\theta)|^2 \right) = \sigma_U^2(\theta) + b_U^2(\theta).$$

Man möchte, dass $\sigma_U(\theta)$ und $|b_U(\theta)|$ beide klein sind, doch typischerweise kann man den einen Fehler nur auf Kosten des andern verkleinern. Am einfachsten wird es, wenn wir verlangen, dass der systematischen Fehler null sein soll. Dann gilt es, den Standardfehler, oder äquivalent dazu, die Varianz zu minimieren.

Definition 6.1. *U heisst* **erwartungstreu** *für $g(\theta)$ falls*

$$\mathbb{E}_\theta \left(U \right) = g(\theta) \quad \forall \theta \in \Theta, \quad \textit{d.h. } b_U(\theta) \equiv 0.$$

Beispiel 6.4. *Normalverteilung: Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Dann ist der unbekannte Parameter $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$, und der übliche Schätzer ist $T = (\overline{X}, S_n^2)$, wobei*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$$

das arithmetische Mittel und die sogenannte Stichprobenvarianz bezeichnen. Man sieht sofort, dass $\overline{X}$ *erwartungstreu ist für* $g_1(\theta) = \mu$*. Um zu sehen, dass* S_n^2 *erwartungstreu ist für* $g_2(\theta) = \sigma^2$*, braucht man eine kleine Rechnung:*

$$\begin{aligned}\mathbb{E}_\theta\left(\sum_{i=1}^n (X_i - \overline{X})^2\right) &= \mathbb{E}_\theta\left(\sum_{i=1}^n (X_i - \mu)^2\right) - n\mathbb{E}_\theta\left((\overline{X} - \mu)^2\right) \\ &= n\sigma^2 - n\sigma^2/n = (n-1)\sigma^2.\end{aligned}$$

(dies erklärt den am Anfang geheimnisvollen Nenner bei S_n^2*!). Man beachte jedoch, dass* S_n *nicht erwartungstreu ist für* $g_3(\theta) = \sigma$*, denn wegen der Jensenschen Ungleichung (Satz 3.6) ist* $\mathbb{E}_\theta\left(\sqrt{S_n^2}\right) < \sqrt{\mathbb{E}_\theta\left(S_n^2\right)} = \sigma$*.*

In der mathematischen Statistik wird bewiesen, dass $\overline{X}$ *und* S_n^2 *minimalen Standardfehler haben unter allen erwartungstreuen Schätzern für* μ *bzw.* σ^2*, und zwar simultan für alle* θ*'s.*

Beispiel 6.5. *Regression: Wir betrachten das Regressionsmodell (6.1) mit fest gegebenen erklärenden Variablen* (x_{ij}) *und zufälligen Fehlern* ε_i *mit* $\mathbb{E}(\varepsilon_i) = 0$*. Der unbekannte Parameter* θ *besteht dann aus dem strukturellen Parameter* β *und Störparametern, die sich auf die Verteilung der Fehler beziehen. Wir betrachten den Kleinste-Quadrate-Schätzer* $\widehat{\beta}$*. Wenn wir den Erwartungswert eines Zufallsvektors komponentenweise definieren, dann folgt aus der Formel (6.2) und der Linearität des Erwartungswertes, dass*

$$\mathbb{E}_\theta\left(\widehat{\beta}\right) = (X'X)^{-1}X'\mathbb{E}_\theta(Y) = (X'X)^{-1}X'X\beta = \beta$$

d.h. $\widehat{\beta}_j$ *ist erwartungstreu für* β_j *(*$j = 0, 1, \ldots p$*).*

Oft wird die Theorie einfacher im Grenzwert von unendlich vielen Beobachtungen, d.h. wir geben die Anzahl Beobachtungen durch einen zusätzlichen Index an: $(X_1, \ldots, X_n) \sim \mu_{n,\theta}$ auf $(\mathbb{R}^n, \mathcal{B}^n)$ und wir haben eine Folge von Schätzern (U_n) von $g(\theta)$, wobei $U_n : \mathbb{R}^n \to \mathbb{R}$ Dann untersuchen wir, ob und wie rasch sich die Folge der Verteilungen von U_n auf $g(\theta)$ konzentriert:

Definition 6.2. (U_n) *heisst* **konsistent** *für* $g(\theta)$ *falls*

$$\mathbb{P}_\theta\left(|U_n - g(\theta)| > \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 0 \quad \forall\ \varepsilon\ >\ 0, \quad \forall\theta.$$

(U_n) *heisst* **asymptotisch normalverteilt** *mit asymptotischer Varianz* $\tau^2(\theta)$*, und wir schreiben* $U_n \approx\sim \mathcal{N}(g(\theta), \frac{1}{n}\tau^2(\theta))$*, falls für alle* θ

$$\mathbb{P}_\theta\left(\sqrt{n}(U_n - g(\theta)) \le x\right) \longrightarrow \Phi(\frac{x}{\tau(\theta)}) \quad \forall x.$$

Bei einem asymptotisch normalverteilten Schätzer bestimmt die asymptotische Varianz die Genauigkeit: je kleiner $\tau^2(\theta)$, desto genauer der Schätzer.

Beispiel 6.6. *Normalverteilung. Es seien wie im vorigen Beispiel* $X_1, \ldots, X_n, \ldots$ *i.i.d.* $\sim \mathcal{N}(\mu, \sigma^2)$*. Von Kapitel 3 wissen wir, dass das arithmetische Mittel* $\overline{X}$ *konsistent und exakt* $\mathcal{N}(\mu, \frac{1}{n}\sigma^2)$*- verteilt ist. Dies zeigt insbesondere, dass die asymptotische Varianz im Allgemeinen vom ganzen Parameter* θ *abhängt, und nicht nur von* $g(\theta)$*. Die Stichproben-Varianz* S_n^2 *zerlegen wir wie oben:*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \frac{n}{n-1} - (\overline{X} - \mu)^2 \frac{n}{n-1} = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 + \ \textit{Rest.}$$

Der erste Term ist wieder ein arithmetisches Mittel, auf das wir das Gesetz der grossen Zahlen und den Zentralen Grenzwertsatz anwenden können, und man kann zeigen, dass der Rest asymptotisch vernachlässigbar ist bei beiden Grenzwertsätzen. Also ist S_n^2 konsistent und asymptotisch $\mathcal{N}(\sigma^2, \frac{1}{n}2\sigma^4)$-verteilt, denn $\mathbb{E}\left((X_i-\mu)^4\right) = 3\sigma^4$.

Das arithmetische Mittel ist nicht der einzig mögliche Schätzer von μ. Als Alternative können wir zum Beispiel den Stichprobenmedian U_n betrachten. Mit einem Symmetrieargument folgt, dass die Verteilung von U_n symmetrisch bezüglich μ ist. Daher ist U_n auch erwartungstreu. Von Beispiel 4.6 her wissen wir ferner, dass der Stichprobenmedian U_n asymptotisch normalverteilt ist mit asymptotischer Varianz $\sigma^2\frac{\pi}{2}$). Der Stichprobenmedian ist bei Normalverteilung also weniger genau als das arithmetische Mittel: er streut etwa $\sqrt{\pi/2} \approx 1.25$ mal so viel.

Bei längerschwänzigen Verteilungen sieht die Sache aber anders aus: Bei der Cauchy-Verteilung mit unbekanntem Lageparameter μ, d.h. X_i i.i.d. mit

$$X_i \sim \frac{1}{\pi}\frac{1}{1+(x-\mu)^2}dx$$

hat z.B. das arithmetische Mittel die gleiche Verteilung für jedes n, ist also nicht konsistent, der Stichprobenmedian hingegen schon.

Die Genauigkeit eines Schätzers ist aber nicht das einzige Kriterium. Eine Rolle spielt auch die Empfindlichkeit eines Schätzers auf vereinzelte Ausreisser (grobe Fehler, Beobachtungen mit anderer Verteilung). Die einfachste Formalisierung dieser Empfindlichkeit ist der sogenannte **Bruchpunkt**, der definiert ist als

$$\varepsilon^*(x_1,\ldots,x_n) = \frac{1}{n}\max\{k\in\mathbb{N}_0\ ;\ \sup\{|U(y_1,\ldots,y_n)|\ ;\ \#\{y_i\neq x_i\}=k\}<\infty\}.$$

Das heisst also, dass U $\varepsilon^* n$ Ausreisser verkraften kann, aber nicht $\varepsilon^* n + 1$.

Beispiel 6.7. *Das arithmetische Mittel hat offensichtlich Bruchpunkt null. Das α-gestutzte Mittel ($0 < \alpha \leq \frac{1}{2}$), das definiert ist durch Weglassen der $k = [\alpha n]$ kleinsten und k grössten Beobachtungen, hat Bruchpunkt $\varepsilon^* = k/n \approx \alpha$. Für $\alpha = \frac{1}{2}$ ist das gestutzte Mittel der Stichprobenmedian, welcher maximalen Bruchpunkt hat.*

6.2.2 Konstruktion von Schätzern

Die wichtigste allgemein anwendbare Methode zur Konstruktion von Schätzern ist die **Maximum-Likelihood-Methode**. Diese geht wie folgt. Zunächst sei die Verteilung der Beobachtungen μ_θ diskret. Dann definieren wir die Likelihoodfunktion $L : \Theta \longrightarrow \mathbb{R}$

$$L(\theta) = \mu_\theta((x_1,\ldots,x_n))$$

für feste Beobachtungen $(x_1,\ldots,x_n)$. $L(\theta)$ gibt an, wie wahrscheinlich die gemachten Beobachtungen sind, wenn die zugrunde liegende Verteilung μ_θ ist. Wenn man θ nicht kennt, ist es plausibel anzunehmen, dass man einen typischen Wert beobachtet hat, d.h. man wird θ schätzen als

$$T = \arg\max_\theta L(\theta).$$

Die Bezeichnung arg max bedeutet, dass wir dasjenige Argument suchen, bei dem die Funktion ihr Maximum annimmt. Falls das Maximum an mehreren Stellen angenommen wird, wählt man willkürlich eine davon. Wenn das Maximum nicht angenommen wird, ist arg max nicht definiert.

Im absolut stetigen Fall definieren wir analog

$$L(\theta) = f_\theta(x_1, \ldots, x_n),$$

wobei f_θ die Dichte bezeichnet.

Statt $L(\theta)$ ist es oft einfacher $\log L(\theta)$ zu maximieren. Oft, aber nicht immer, findet man den Maximum-Likelihood-Schätzer durch Ableiten und null setzen von $\log L(\theta)$.

Beispiel 6.8. *Das sogenannte* **Lokationsmodell**. *Seien $X_1, \ldots, X_n$ i.i.d. mit Dichte $f(x-\theta)$. Dann ist*

$$L(\theta) = f(x_1 - \theta) \ldots f(x_n - \theta).$$

Je nach Wahl von f erhält man dann andere Schätzer, z.B.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \Rightarrow \log L(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2 + \text{const}\,. \Rightarrow T = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Für

$$f(x) = \frac{1}{2}\exp(-|x|) \Rightarrow \log L(\theta) = -\sum_{i=1}^{n}|x_i - \theta| + \text{const}\,.$$

erhalten wir den Stichprobenmedian

$$\begin{array}{lcll} T & \in & [x_{(k)}, x_{(k+1)}] & \textit{falls} \quad n = 2k \\ T & = & x_{(k+1)} & \textit{falls} \quad n = 2k+1, \end{array}$$

wobei $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ die geordnete Stichprobe bezeichnet. Dies kann man wie folgt einsehen. Die Funktion $\log L$ ist stetig und linear auf jedem Intervall, das keine Beobachtungen enthält. Für $\theta < x_{(1)}$ ist die Steigung gleich n, und an einer Stelle $x_{(i)}$, die m mal vorkommt in $(x_1, x_2, \ldots x_n)$, nimmt die Steigung um $2m$ ab. Für $n = 2k+1$ hat $\log L$ also ein Maximum bei $x_{(k+1)}$. Für $n = 2k$ und $x_{(k)} < x_{(k+1)}$ ist das arg max von $\log L$ das ganze Intervall $[x_{(k)}, x_{(k+1)}]$ (die Funktion ist auf diesem Intervall konstant).

Genau gleich argumentiert man im Regressionsmodell (6.1): Bei unabhängigen normalverteilten Fehlern mit konstanter Varianz ist der Maximum-Likelihood-Schätzer für β gleich dem Kleinste-Quadrate-Schätzer. Wenn die Normalverteilung durch die Verteilung mit Dichte $f(x) = \frac{1}{2}\exp(-|x|)$ ersetzt wird, erhält man stattdessen den sogenannten L_1-Schätzer

$$\arg\min_\beta \sum_{i=1}^{n} |Y_i - \sum_{j=1}^{p} \beta_j x_{ij} - \beta_0|.$$

Er wurde ursprünglich von Laplace vorgeschlagen, ist aber weniger verbreitet, unter anderem weil es keine explizite Formel gibt. Heute kann er aber mit Methoden der linearen Programmierung ebenfalls sehr schnell berechnet werden. Wenn die Fehler eine etwas längerschwänzige Verteilung haben als die Normalverteilung, ist der L_1-Schätzer meist besser.

Beispiel 6.9. *Schätzung der Varianz bei Normalverteilung.*
Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ mit $\theta = (\mu, \sigma^2)$. Dann ist

$$\begin{aligned} \log L(\mu, \sigma^2) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n}{2}\log\sigma^2 + \text{const}\,. \\ \Rightarrow \quad T &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i, \quad \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j \right)^2 \right). \end{aligned}$$

Der Maximum-likelihood-Schätzer für die Varianz ist also nicht erwartungstreu.

Der Maximum-likelihood-Schätzer ist nahezu universell anwendbar, und in der mathematischen Statistik wird gezeigt, dass er in der Regel sehr gute asymptotische Eigenschaften hat. Andere allgemeine Methoden sind der **Momentenschätzer** und die **Bayesschätzer**, auf die wir hier aber nicht eingehen.

6.3 Statistische Tests

6.3.1 Problemstellung

Die Beobachtungen seien wieder $\mathbf{X} = (X_1, \ldots, X_n)$ und die möglichen Verteilungen $(\mu_\theta;\ \theta \in \Theta)$. Verschiedene Werte von θ entsprechen verschiedenen Hypothesen, und wir nehmen an, dass wir eine sogenannte Nullhypothese überprüfen wollen, die beschrieben ist durch eine Teilmenge $\theta \in \Theta_0 \subset \Theta$. Häufig bedeutet die Nullhypothese "kein Effekt" oder "reiner Zufall". Das Komplement Θ_0^c bezeichnet man üblicherweise als Alternative.

Aufgrund des beobachteten $\mathbf{X}$ soll man entscheiden, ob die Nullhypothese zutrifft, d.h. ob die Verteilung von $\mathbf{X}$ gleich einem μ_θ mit $\theta \in \Theta_0$, sein kann. Es gibt nur zwei mögliche Entscheidungen: Entweder man behält die Nullhypothese bei, oder man lehnt sie ab. Offensichtlich gibt es dann zwei mögliche Fehlentscheidungen:

- Die Nullhypothese wird abgelehnt (verworfen), obwohl sie richtig ist (Fehler 1. Art),
- Die Nullhypothese wird akzeptiert (beibehalten), obwohl sie falsch ist (Fehler 2. Art).

Ein statistischer Test ist eine Entscheidungsregel basierend auf der Beobachtung, d.h.

$$\varphi : (\mathbb{R}^n, \mathcal{B}^n) \longrightarrow \{0, 1\}$$

ist eine messbare Funktion, wobei $\varphi(\mathbf{x}) = 0$ heisst "Die Nullhypothese wird akzeptiert" und $\varphi(\mathbf{x}) = 1$ "Die Nullhypothese wird verworfen". Eine Entscheidungsregel φ definiert eine messbare Teilmenge $K \subset \mathbb{R}^n$ mit $\varphi = 1_K$. Diese Teilmenge heisst *Verwerfungsbereich oder kritischer Bereich des Tests*. Die Bestimmung eines Tests ist gleichbedeutend mit der Bestimmung des Verwerfungsbereiches.

Offensichtlich ist $\mathbb{E}_\theta(\varphi) = \int \varphi(\mathbf{x})\mu_\theta(d\mathbf{x}) = \mathbb{P}_\theta(\varphi = 1)$ die Wahrscheinlichkeit, die Nullhypothese zu verwerfen. Ein guter Test sollte trennscharf sein im Sinne, dass

$$\mathbb{E}_\theta(\varphi) \quad \text{möglichst klein auf} \quad \Theta_0$$

und

$$\mathbb{E}_\theta(\varphi) \quad \text{möglichst gross auf} \quad \Theta_0^c.$$

Definition 6.3. *Wenn*

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\varphi) \leq \alpha$$

(d.h. die Wahrscheinlichkeit eines Fehlers 1. Art ist $\leq \alpha$), dann heisst φ ein Test zum **Niveau** α. *Für $\theta \notin \Theta_0$ heisst $\mathbb{E}_\theta(\varphi)$ auch die* **Macht des Tests** *an der Stelle $\theta \notin \Theta_0$. Die Macht ist also Eins minus die Wahrscheinlichkeit eines Fehlers 2. Art.*

Üblicherweise wählt man ein Niveau α, z.B. 5% oder 1%, und sucht unter allen Tests mit diesem Niveau denjenigen, der $\mathbb{E}_\theta(\varphi)$ maximiert für ein festes $\theta \notin \Theta_0$, bzw. – sofern möglich – für alle $\theta \notin \Theta_0$ (sog. gleichmässig mächtigster Test). Die beiden Fehlerarten werden also nicht symmetrisch behandelt. Weil ein Test durchgeführt wird, um Kritiker und Skeptiker zu überzeugen, ist das Niveau wichtiger als die Macht.

Es ist zu beachten, dass das Niveau α den Fehler erster Art begrenzt, also eine Ablehnung der Nullhypothese, obwohl sie richtig ist. Bei kleinem α ist die fälschliche Ablehnung der Nullhypothese daher ein seltenes Ereignis. Der Fehler zweiter Art hingegen ist deutlich unbestimmter: er hängt stark davon ab, welche Alternative man betrachtet, und seien Wahrscheinlichkeit konvergiert typischerweise gegen $1-\alpha$, wenn $\theta \notin \Theta_0$ gegen Θ_0 konvergiert. Das heisst, gewisse Alternativen bleiben auch dann plausibel, wenn man die Nullhypothese akzeptiert. Deshalb ist das Beibehalten der Nullhypothese nicht ein Beweis, dass diese richtig ist.

Beispiel 6.10. *Wir betrachten* 20 *unabhängige* $0-1$ *Experimente mit unbekanntem Erfolgsparameter* $p \in \Theta = [0,1]$. *Die Beobachtung ist* $X =$ *Anzahl Erfolge* $\sim$ *Binomial* $(20,p)$, *und die Nullhypothese sei* $\Theta_0 = [0,\frac{1}{2}]$. *Als Beispiel kann man an den Vergleich zweier Behandlungen denken, wo wir die Anzahl Patienten zählen, bei denen die neue Behandlung eine bessere Wirkung hat als die alte. Die Nullhypothese wäre dann, dass die neue Behandlung höchstens gleich gut ist wie die alte.*

Ein grosses X *spricht gegen* $p \leq \frac{1}{2}$, *daher setzen wir*

$$\varphi(x) = \begin{cases} 1 & \textit{falls} \quad x \geq c \\ 0 & \textit{falls} \quad x < c \end{cases}$$

Dann ist $\mathbb{E}_p(\varphi) = \sum_{k=c}^{20} \binom{20}{k} p^k (1-p)^{20-k}$ *monoton wachsend in* p. *Wir können* c *also als Funktion des Niveaus bestimmen, indem wir die Gleichung*

$$2^{-20} \sum_{k=c}^{20} \binom{20}{k} \leq \alpha < 2^{-20} \sum_{k=c-1}^{20} \binom{20}{k}$$

lösen. Insbesondere ergibt ein $\alpha \in [0.021, 0.058]$ *den Wert* $c = 15$.

Wie steht es mit der Macht? Man berechnet

p	0.6	0.7	0.8	0.9
$\mathbb{E}_p(\varphi)$	0.126	0.416	0.804	0.989

Wenn also $p = 0.7$ *eine relevante Verbesserung ist, beträgt die Wahrscheinlichkeit, diese auch zu entdecken, nur etwas mehr als* 40%, *was sicher zu wenig ist. Die einzige Möglichkeit, die Macht zu vergrössern, ohne das Niveau zu vergrössern, ist hier eine grössere Stichprobe.*

Die Fehlerwahrscheinlichkeiten beim Testen beziehen sich auf die Unsicherheit vor der Durchführung. Nachdem der Test durchgeführt wurde und zur Verwerfung der Nullhypothese führte, dann kann man schliessen "Entweder ist die Nullhypothese falsch, oder ein seltenes Ereignis, dessen Wahrscheinlichkeit höchstens $= \alpha$ ist, ist eingetreten". Hingegen ist es nicht korrekt im Falle der Verwerfung der Nullhypothese zu sagen "Die Nullhypothese ist höchstens mit Wahrscheinlichkeit α richtig". Erstens ist die Korrektheit der Nullhypothese im Allgemeinen kein zufälliges Ereignis, hat also auch keine Wahrscheinlichkeit. Und selbst wenn man bereit ist, die Korrektheit der Nullhypothese als zufällig anzusehen, dann ist zweitens $\mathbb{P}(\varphi = 1 | \text{Nullhypothese richtig})$ nicht gleich $\mathbb{P}(\text{Nullhypothese richtig} | \varphi = 1)$: Die erste Wahrscheinlichkeit ist nach Konstruktion kleiner oder gleich α, die zweite Wahrscheinlichkeit muss mit der Bayes'schen Regel berechnet werden und hängt davon ab, wie gross die sogenannte a priori Wahrscheinlichkeit der Nullhypothese ist (man vergleiche das Beispiel 2.16).

In der Praxis wird oft der sogenannte **P-Wert** berechnet. Dazu muss man voraussetzen, dass man für jedes α einen Test φ_α festgelegt hat, und dass diese Tests kompatibel sind

im Sinne, dass

$$\alpha' < \alpha \Rightarrow \varphi_{\alpha'} \leq \varphi_\alpha$$

(wenn eine Beobachtung auf einem bestimmten Niveau als verträglich mit der Nullhypothese angesehen wird, dann ist sie das erst recht, wenn man das Niveau verkleinert). Dann ist der P-Wert von $\mathbf{x}$ definiert als $\pi(\mathbf{x}) = \inf\{\alpha; \varphi_\alpha(\mathbf{x}) = 1\}$, d.h. der P-Wert ist "das kleinste Niveau, bei dem der Test die Nullhypothese gerade noch verwirft". Der P-Wert ist informativer als die Angabe der Testentscheidung auf einem festen Niveau, gewissermassen ein verfeinertes "Signifikanzmass". Wenn der P-Wert kleiner oder gleich α ist, dann verwirft der Test auf dem Niveau α. Der P-Wert darf aber nicht als Wahrscheinlichkeit, dass die Nullhypothese richtig ist, interpretiert werden, vergleiche die Diskussion im vorangehenden Abschnitt.

Der P-Wert hängt ab von $\mathbf{X}$ und ist daher eine Zufallsvariable. Wie sieht seine Verteilung aus ?

Lemma 6.1. *Sei $\pi(\mathbf{X}) = \inf\{\alpha; \varphi_\alpha(\mathbf{X}) = 1\}$ der P-Wert für die Tests (φ_α) mit Niveau α $(0 \leq \alpha \leq 1)$. Wenn $\theta \in \Theta_0$, dann gilt $\mathbb{P}_\theta(\pi(\mathbf{X}) \leq u) \leq u$. Wenn $\mathbb{P}_\theta(\varphi_\alpha(\mathbf{X}) = 1) = \alpha$ für alle α, dann gilt $\mathbb{P}_\theta(\pi(\mathbf{X}) \leq u) = u$, d.h. der P-Wert ist uniform verteilt.*

Beweis Weil $\alpha \mapsto \varphi_\alpha(\mathbf{x})$ monoton wachsend ist, impliziert $\pi(\mathbf{x}) < u$, dass $\varphi_u(\mathbf{x}) = 1$. Daraus folgt

$$\mathbb{P}_\theta(\pi(\mathbf{X}) < u) \leq \mathbb{P}_\theta(\varphi_u(\mathbf{X}) = 1) \leq u.$$

Weil $\mathbb{P}_\theta(\pi(\mathbf{X}) \leq u)$ rechtsstetig und monoton wachsend ist, folgt die erste Behauptung. Ferner folgt aus $\varphi_u(\mathbf{x}) = 1$, dass $\pi(\mathbf{x}) \leq u$, also auf Grund der zusätzlichen Bedingung an $\mathbb{P}_\theta$ dass $u = \mathbb{P}_\theta(\varphi_u(\mathbf{X}) = 1) \leq \mathbb{P}_\theta(\pi(\mathbf{X}) \leq u)$. □

6.3.2 Einige wichtige Tests

Wir besprechen hier einige der üblichen Tests für einfache, häufig auftretende Situationen. In allen praktischen Anwendungen ist der Verwerungsbereich gegeben als Urbild einer sogenannten Teststatistik T: Die Nullhypothese wird verworfen, falls $T > c$, wobei die "kritische Grenze" c vom gewählten Niveau α abhängig ist. Wir beschränken uns darauf, die üblichen Teststatistiken anzugeben, ohne genauer zu untersuchen, ob und in welchem Sinne diese Teststatistiken optimal sind. Es sollte aber jeweils sofort einleuchten, dass grosse Werte von T gegen die Nullhypothese sprechen.

1-Stichproben-t-Test:

Sei

$$X_1, \ldots, X_n \;\; i.i.d. \sim \mathcal{N}(\mu, \sigma^2),\; \Theta = \mathbb{R} \times \mathbb{R}_+ \,,\; \Theta_0 = \{\mu_0\} \times \mathbb{R}_+.$$

Diese Situation trifft man in der Qualitätskontrolle beim Test auf einen Sollwert μ_0 an, sowie bei verbundenen oder gepaarten Stichproben, wo jeweils 2 Behandlungen bei der gleichen Versuchseinheit durchgeführt werden. Im zweiten Fall ist X_i der Unterschied der Resultate für die beiden Behandlungen bei der i-ten Versuchseinheit, und meist hat man $\mu_0 = 0$, d.h. die Nullhypothese besagt, dass es keinen systematischen Unterschied zwischen den beiden Behandlungen gibt.

Als Teststatistik wird in diesem Fall gewählt

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S_n}, \quad S_n^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2, \quad \overline{X} = \frac{1}{n}\sum X_i,$$

und wir verwerfen die Nullhypothese, falls $|T|$ zu gross (sog. zweiseitiger t-Test). Die kritische Grenze wird bestimmt mit

Satz 6.1. *Die Verteilung von T hat für $\mu = \mu_0$, $n \geq 2$ und alle $\sigma > 0$ die Dichte*

$$f_{n-1}(t) = \frac{\Gamma(n/2)}{\sqrt{(n-1)\pi}\Gamma((n-1)/2)} \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$

(sog. $t-$Verteilung mit $n-1$ Freiheitsgraden).

Der Beweis wird z.B. in der Mathematischen Statistik geliefert.

Damit lautet der Test

$$\varphi(x) = 1 \Longleftrightarrow |T| > c(n-1, \alpha)$$

wobei die kritische Grenze $c(n-1,\alpha)$ das $(1-\alpha/2)$-Quantil der t-Verteilung ist. Dieses ist tabelliert für die üblichen α's und nicht allzu grosse $n-1$'s. Im Grenzfall $n \to \infty$ erhält man die Quantile der Standard-Normalverteilung.

Wenn man als Nullhypothese $\mu \leq \mu_0$ hat und als Alternative $\mu > \mu_0$ (d.h. man ist nur an Überschreitungen des Sollwertes, bzw. an Behandlungen, die zu einer Verbesserung führen, interessiert), nimmt man $\varphi(x) = 1 \Leftrightarrow T > c'(n-1,\alpha)$ (einseitiger t-Test). Analog geht es bei der Nullhypothese $\mu \geq \mu_0$.

Wenn man die Standardabweichung σ der Beobachtungen X_i kennt, verwenden wir σ statt S_n und entnehmen die kritischen Grenzen einer Tabelle der Normal- statt der t-Verteilung. Der Test heisst dann auch z-Test.

Vorzeichentest

Sei

$$X_1, \ldots, X_n \text{ } i.i.d. \sim F(x-\mu),\ \Theta = \mathbb{R} \times \{F | F(0) = \frac{1}{2}\},\ \Theta_0 : \mu = \mu_0.$$

Man testet also, ob der Median $= \mu_0$ sein kann. Im Gegensatz zum t-Test nimmt man keine Normalverteilung mehr an. Aus technischen Gründen setzen wir voraus, dass F stetig ist in 0, so dass $\mathbb{P}(X_i = \mu) = 0$.

Die Teststatistik $T = \sum_{i=1}^n 1_{[X_i > \mu_0]}$ ist unter der Nullhypothese Binomial $(n, \frac{1}{2})$-verteilt, und wenn T zu stark abweicht von $\frac{n}{2}$, werden wir die Nullhypothese verwerfen. Wir verwenden also den Test

$$\varphi(x) = 1 \Longleftrightarrow |T - n/2| > c,$$

wobei $c = c(n, \alpha)$ mit Hilfe der Binomialverteilung bestimmt wird.

Vergleich von t-Test und Vorzeichentest: Wenn man tatsächlich Normalverteilung hat, stimmt beim $t-$Test das Niveau exakt und in der Mathematischen Statistik wird gezeigt, dass er maximale Macht hat. Falls F symmetrisch ist, stimmt das Niveau beim $t-$Test genähert (wenn F nicht extrem langschwänzig ist) und beim Vorzeichentest exakt. Bei der Macht kommt es sehr auf F an: Je längerschwänzig F ist, desto besser ist der Vorzeichentest.

2-Stichproben-t-Test

Seien

$$X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2),\quad Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2),\ \text{alle unabhängig.}$$

Dann ist

$$\theta = (\mu_1, \mu_2, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$$

Als Nullhypothese wählen wir $\Theta_0 = \{\mu_1 = \mu_2\}$, d.h. unter der Nullhypothese haben alle Variablen die gleiche Verteilung. Die Teststatistik ist (vgl. Beispiel 6.2)

$$T = \frac{(\overline{X} - \overline{Y})/\sqrt{1/n + 1/m}}{\sqrt{\left(\sum (X_i - \overline{X})^2 + \sum (Y_i - \overline{Y})^2\right)/(n+m-2)}}.$$

Satz 6.2. *Für alle $\theta \in \Theta_0$ hat T die $t-$Verteilung mit $n+m-2$ Freiheitsgraden.*

Ohne Beweis.

Der Test lautet also

$$\varphi(x) = 1 \Longleftrightarrow |T| > c,$$

wobei man $c = c(n+m-2, \alpha)$ aus Tabellen erhält. Analog geht es bei Nullhypothesen $\mu_1 \leq \mu_2$, bzw. $\mu_1 \geq \mu_2$.

2-Stichproben-Wilcoxon-Test (oder Mann-Whitney U-Test)

Seien

$$X_1, \ldots, X_n \sim F, \quad Y_1, \quad \ldots, Y_m \sim G, \text{ alle unabhängig.}$$

Wir wollen jetzt die Verteilungen F und G nicht näher spezifizieren, aber aus technischen Gründen nehmen wir an, dass beide stetig sind. Es ist also $\Theta = \mathcal{F} \times \mathcal{F}$, wobei $\mathcal{F}$ die Menge der stetigen Verteilungsfunktionen bezeichnet. Die Nullhypothese ist $\Theta_0 = \{F = G\}$, und als Teststatistik nehmen wir (vgl. Beispiel 6.2)

$$W = \sum_{i=1}^{n} \sum_{j=1}^{m} 1_{[X_i < Y_j]}$$

Zur Durchführung des Test brauchen wir die Verteilung von W unter der Nullhypothese. Sei $(Z_i)_{1 \leq i \leq n+m}$ die kombinierte Stichprobe, $Z_i = X_i \ \ (i \leq n)$, $Z_i = Y_{i-n} \ \ (i > n)$. Unter der Nullhypothese sind die Z_i i.i.d., also gilt aus Symmetriegründen für alle Permutationen π

$$\mathbb{P}\left(Z_{\pi(1)} < Z_{\pi(2)} < \cdots < Z_{\pi(n+m)}\right) = \frac{1}{(n+m)!}.$$

Da W konstant ist auf $\{Z_{\pi(1)} < \cdots < Z_{\pi(n+m)}\}$, hängt die Verteilung von W nicht von F ab und kann durch Abzählen gefunden werden ($\rightarrow$ Tabellen für kleine n, m). Für grössere n, m verwendet man:

Lemma 6.2. *Unter der Nullhypothese gilt*

$$\mathbb{E}(W) = \frac{nm}{2}, \quad \mathbb{V}(W) = \frac{nm(n+m+1)}{12},$$
$$\mathbb{P}\left(\frac{W - \mathbb{E}(W)}{\sqrt{\mathbb{V}(W)}} \leq x\right) \longrightarrow \Phi(x) \quad (x \in \mathbb{R}; \quad n, m \to \infty).$$

Beweis $\mathbb{E}(W) = nm \cdot \mathbb{P}(X_i < Y_j) = \frac{nm}{2}$ folgt aus Symmetriegründen. Ferner

$$\mathbb{V}(W) = \sum_i \sum_j \sum_k \sum_\ell \operatorname{Cov}\left(1_{[X_i < Y_j]}, 1_{[X_k < Y_\ell]}\right)$$

und

$$\begin{aligned}\mathrm{Cov}\left(1_{[X_i<Y_j]}, 1_{[X_k<Y_\ell]}\right) &= \mathbb{E}\left(1_{[X_i<Y_j]}1_{[X_k<Y_\ell]}\right) - (\mathbb{E}\left(1_{[X_i<Y_j]}\right))^2 \\ &= 0 \text{ falls alle Indizes verschieden} \\ &= \frac{1}{4} \text{ falls } i=k, j=\ell \\ &= \mathbb{P}\left(X_i < \min(Y_j, Y_\ell)\right) - \frac{1}{4} \text{ falls } i=k, j\neq\ell \\ &= \mathbb{P}\left(\max(X_i, X_k) < Y_j\right) - \frac{1}{4} \text{ falls } i\neq k, j=\ell.\end{aligned}$$

Aus Symmetrie $\mathbb{P}\left(X_i < \min(Y_i, Y_\ell)\right) = \mathbb{P}\left(\max(X_i, X_k) < Y_j\right) = \frac{1}{3}$. Also

$$\mathbb{V}(W) = nm\frac{1}{4} + nm(m-1)\frac{1}{12} + nm(n-1)\frac{1}{12} = \frac{nm}{12}(n+m+3-1-1).$$

Auf den Beweis der asymptotischen Normalität verzichten wir. □

Vergleich von $t-$Test und Wilcoxon-Test: In der Mathematischen Statistik zeigt man, dass das Niveau beim Wilcoxon-Test immer exakt stimmt, während es beim $t-$Test meist nur genähert richtig ist. Entscheidend ist jedoch, dass der Wilcoxon-Test oft eine wesentlich grössere Macht hat als der $t-$Test, und dass auch im ungünstigsten Fall seine Macht nur wenig kleiner ist als beim t-Test. Deshalb sollte man eher den Wilcoxon-Test verwenden.

Chi-Quadrat-Anpassungstest

Wir betrachten n unabhängige Wiederholungen eines Experiments mit k möglichen Ausgängen. Die Wahrscheinlichkeit für den i-ten Ausgang sei θ_i. Dann ist

$$\Theta = \{(\theta_1, \ldots, \theta_k) |\ \theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1\}.$$

Wir betrachten zuerst eine einfache Nullhypothese, wo $\Theta_0 = \{\theta_0\} = \{(\theta_{10}, \ldots, \theta_{k0})\}$.
Beispiel 6.11. *Uniforme Zufallszahlengeneratoren. Wir zerlegen* $(0,1]$ *in die Intervalle* $I_i = (\frac{i-1}{k}, \frac{i}{k}]$ *und schauen, in welchem Intervall die Zufallszahl liegt. Wir haben also k Ausgänge, und wenn der Generator korrekt ist, dann sind alle Ausgänge gleich wahrscheinlich, d.h.* $\theta_0 = (\frac{1}{k}, \ldots, \frac{1}{k})$.

Wir arbeiten nicht direkt mit den Resultaten der einzelnen Wiederholungen, sondern schauen nur darauf, wie häufig die verschiedenen Ausgänge vorkommen. Wir betrachten also die Zufallsvariablen N_i = Anzahl Wiederholungen mit Ausgang i, $i = 1, \ldots, k$. Die Verteilung von $(N_1, \ldots, N_k)$ ist die sogenannte **Multinomial-Verteilung**:

$$\mathbb{P}_\theta\left(N_1 = n_1, \ldots, N_k = n_k\right) = \frac{n!}{n_1! \ldots n_k!}\ \theta_1^{n_1} \ldots \theta_k^{n_k}$$

mit $\mathbb{E}(N_i) = n\theta_i$, $\mathbb{V}(N_i) = n\theta_i(1-\theta_i)$ und $\mathrm{Cov}(N_i, N_j) = -n\theta_i\theta_j$ $(i \neq j)$.

Als Teststatistik wählen wir eine gewichtete Summe der quadrierten Abweichungen der N_i's von ihrem Erwartungswert:

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n\theta_{i0})^2}{n\theta_{i0}},$$

und wir verwerfen, falls $\chi^2 > c$ (hier hat man praktisch immer einseitige Tests). Um c als Funktion von k und α zu bestimmen, müssen wir die Verteilung von χ^2 unter der Nullhypothese kennen. Im Prinzip kann man diese aus der Multinomial-Verteilung berechnen. Meist benützt man aber das Resultat (ohne Beweis), dass χ^2 asymptotisch die gleiche Verteilung hat wie $\sum_{i=1}^{k-1} Y_i^2$, wobei die $Y_1, \ldots, Y_{k-1}$ i.i.d. $\mathcal{N}(0,1)$-verteilt sind. Diese Verteilung heisst die **Chiquadrat-Verteilung** mit $k-1$ **Freiheitsgraden**. Sie hat die Dichte

$$f_{k-1}(x) = 2^{-\frac{k-1}{2}} \Gamma\left(\frac{k-1}{2}\right)^{-1} e^{-x/2} x^{(k-3)/2}.$$

Die Grenzen $c = c(k, \alpha)$ kann man daher genähert aus Tabellen der Chiquadrat-Verteilung bestimmen.

Für $k = 2$ folgt obiges Resultat aus dem Zentralen Grenzwertsatz. Es gilt nämlich

$$\chi^2 = \frac{(N_1 - n\theta_{10})^2}{n\theta_{10}} + \frac{(n - N_1 - n(1-\theta_{10}))^2}{n(1-\theta_{10})} = \frac{(N_1 - n\theta_{10})^2}{n\theta_{10}(1-\theta_{10})}.$$

Also hat χ^2 asymptotisch die gleiche Verteilung wie Y^2, wobei $Y \sim \mathcal{N}(0,1)$. Als Faustregel gilt, dass die Approximation brauchbar ist, sofern wenn etwa 80% der $n\theta_{i0} \geq 4$ und der Rest ≥ 1. Sonst muss man Klassen zusammenfassen.

Mit einer kleinen Modifikation können wir auch **zusammengesetzte Nullhypothesen** testen. Sei

$$\Theta_0 = \{(\theta_1(\eta), \ldots, \theta_k(\eta)) | \eta \in E\}, \quad E \subset \mathbb{R}^r \text{ offen }.$$

Ferner sei $\widehat{\eta}$ eine Schätzung von η. Dann verwenden wir die Teststatistik

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n\ \theta_i(\widehat{\eta}))^2}{n\theta_i(\widehat{\eta})}$$

und verwerfen, wenn $\chi^2 > c$.

Die Verteilung von χ^2 ist schwierig zu bestimmen und hängt von der verwendeten Schätzmethode ab. Falls $\widehat{\eta}$ der Maximum-Likelihood-Schätzer ist, d.h.

$$\widehat{\eta} = \arg\min_{\eta} \sum_{i=1}^{k} N_i \log \theta_i(\eta),$$

dann hat χ^2 asymptotisch eine Chiquadrat-Verteilung mit $k - r - 1$ Freiheitsgraden.

Beispiel 6.12. Kontingenztafeln.
Wir haben zwei Merkmale mit p, bzw. q Ausprägungen, und die Nullhypothese lautet, dass die Merkmale unabhängig sind. Ein Beispiel sind die beiden Merkmale Geschlecht und Linkshändigkeit ($p = q = 2$).

Nimmt man jede Kombination von Merkmalen als einen Ausgang, so haben wir genau die Situation für den Chiquadrat-Test:

$$\begin{aligned}
k &= pq \\
\Theta &= \{(\theta_{11}, \theta_{12}, \ldots, \theta_{pq})\}, \\
\theta_{ij} &= \mathbb{P}\,(1.\ \textit{Merkmal}\ = i,\ \ 2.\ \textit{Merkmal}\ = j) \\
E &= \left\{(\eta_1, \ldots, \eta_p), \sum \eta_i = 1\right\} \times \left\{(\xi_1, \ldots, \xi_q), \sum \xi_j = 1\right\}, \\
\theta_{ij}(\eta, \xi) &= \eta_i \xi_j, \quad r = p + q - 2.
\end{aligned}$$

Man findet:

$$\begin{aligned}\widehat{\eta}_i &= N_{i\cdot}/n = \sum_{j=1}^{q} N_{ij}/n,\\ \widehat{\xi}_j &= N_{\cdot j}/n = \sum_{i=1}^{p} N_{ij}/n.\end{aligned}$$

Also ist

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - N_{i\cdot}N_{\cdot j}/n)^2}{N_{i\cdot}N_{\cdot j}/n}$$

und wir haben $k - r - 1 = (p-1)(q-1)$ Freiheitsgrade.

Als Zahlenbeispiel entnehmen wir aus einer amerikanischen Untersuchung

	Männer	*Frauen*	*insgesamt*	
rechtshändig	2′780	3′281	6′061 =	N_1.
andere	311	300	611 =	N_2.
insgesamt	3′091	3′581	6′672 =	n
	$= N_{\cdot 1}$	$= N_{\cdot 2}$		

Das ergibt die unter der Nullhypothese erwarteten Häufigkeiten

	Männer	*Frauen*	*insgesamt*	
rechtshändig	2′808	3′253	6′061 =	N_1.
andere	283	328	611 =	N_2.
insgesamt	3′091	3′581	6′672 =	n
	$= N_{\cdot 1}$	$= N_{\cdot 2}$		

und damit

$$\chi^2 = \frac{(2'780 - 2'808)^2}{2808} + \frac{(3'281 - 3'253)^2}{3253} + \frac{(311 - 283)^2}{283} + \frac{(300 - 328)^2}{328} = 5.68,$$

mit 1 Freiheitsgrad. Auf dem Niveau $\alpha = 5\%$ wird die Nullhypothese verworfen, auf dem Niveau $\alpha = 1\%$ wird sie hingegen beibehalten. Wir haben also eine gewisse, aber nicht sehr starke statistische Evidenz gegen die Unabhängigkeit von Linkshändigkeit und Geschlecht.

In unserer Diskussion wurde angenommen, dass die N_{ij} multinomialverteilt sind, d.h. im Beispiel oben muss man die Personen zufällig aus einer grossen Population auswählen und danach deren Geschlecht und Händigkeit feststellen. Oft führt man aber die Untersuchung so durch, dass die Spaltentotalen im Voraus fixiert werden. In unserem Beispiel würde man also die Anzahl Männer und Frauen im Voraus festlegen und diese separat zufällig auswählen. Dann hat man q Multinomialverteilungen mit p Ausprägungen, und die Nullhypothese lautet, dass die Wahrscheinlichkeiten für jede Ausprägung gleich sind. Man kann zeigen, dass auch in dieser Situation die gleiche Teststatistik angezeigt ist und dass diese unter der Nullhypothese ebenfalls eine genäherte Chiquadrat-Verteilung mit $(p-1)(q-1)$ Freiheitsgraden hat.

6.3.3 Das Neyman-Pearson-Lemma

Im einfachen Fall, wo die Nullhypothese und die Alternative nur aus je einer Verteilung bestehen, kann man unter allen Tests zum Niveau α den mächtigsten bestimmen – zumnidest,

wenn man bereit ist, auch sogenannte randomisierte Tests zu betrachten. Für Anwendungen ist die Annahme, dass nur zwei Verteilungen möglich sind, natürlich zu einfach, aber es ist die Grundsituation, die man zuerst verstehen will, und die die Grundlage bildet für Optimalitätsaussagen in komplizierteren Situationen.

Wir bezeichnen die Nullhypothese mit μ_0 statt μ_{θ_0} und die Alternative mit μ_1 statt μ_{θ_1}. Wir beginnen mit einer heuristischen Überlegung und nehmen an, dass μ_0 und μ_1 diskret sind. Sei K der Verwerfungsbereich $\{\mathbf{x}_i|\varphi(\mathbf{x}_i) = 1\}$. Dann müssen wir $\sum_{\mathbf{x}_i\in K}\mu_1(\mathbf{x}_i)$ maximieren unter der Nebenbedingung $\sum_{\mathbf{x}_i\in K}\mu_0(\mathbf{x}_i) \leq \alpha$. Wenn wir $\mu_0(\mathbf{x}_i)$ als das Gewicht und $\mu_1(\mathbf{x}_i)$ als den Wert des i-ten Gegenstandes interpretieren, dann ist dies das bekannte "Rucksackproblem" aus der diskreten Optimierung (K ist die Menge der Gegenstände, die in den Rucksack gepackt werden). Offensichtlich muss K aus denjenigen $\mathbf{x}_i$ bestehen, für die μ_1 gross und μ_0 klein ist. Die sogenannte "greedy" Strategie wählt diejenigen Gegenstände, bei denen das Verhältnis Wert:Gewicht, d.h. $\mu_1(\mathbf{x}_i)/\mu_0(\mathbf{x}_i)$, am grössten ist. Wegen der Diskretheit führt das jedoch nicht unbedingt zur optimalen Lösung. Der einfachste Ausweg lässt zu, dass wir vom i-ten Gegenstand einen Anteil $\varphi(\mathbf{x}_i) \in [0,1]$ einpacken können. Dann ist es offensichtlich optimal, die Gegenstände mit dem grössten Verhältnis $\mu_1(\mathbf{x}_i)/\mu_0(\mathbf{x}_i)$ auszuwählen.

Im Rahmen von Tests heisst dies, dass wir **randomisierte Tests** zulassen, d.h. wir betrachten messbare Funktionen

$$\varphi : (\mathbb{R}^n, \mathcal{B}^n) \longrightarrow [0,1],$$

mit der Interpretation: "Wenn $\varphi(\mathbf{x}) = \gamma$, dann verwerfen wir die Nullhypothese gemäss einem unabhängigen Zufallsmechanismus mit Wahrscheinlichkeit γ" (also z.B. durch Werfen einer Münze mit Erfolgsparameter γ). Auch bei randomisierten Tests geben $\mathbb{E}_\theta(\varphi)$ $(\theta \in \Theta_0)$ und $1 - \mathbb{E}_\theta(\varphi)$ $(\theta \notin \Theta_0)$ die Fehlerwahrscheinlichkeiten 1. und 2. Art an. Wir müssen also $\mathbb{E}_1(\varphi)$ maximieren unter der Nebenbedingung $\mathbb{E}_0(\varphi) \leq \alpha$.

Randomisierte Tests kann man als gewöhnliche Tests basierend auf der erweiterten Beobachtung $(\mathbf{X}, U)$ auffassen, wobei U unabhängig von $\mathbf{X}$ und uniform auf $[0,1]$ ist: Man setzt einfach $\Phi(\mathbf{x}, u) = 1_{\{u\leq\varphi(\mathbf{x})\}}$. Den letzten Datenpunkt U besorgt man sich dann selber mit einem Zufallsmechanismus. Offensichtlich gilt (mit dem Satz von Fubini)

$$\mathbb{P}_\theta(\Phi(X,U) = 1) = \int\int_0^1 1_{\{u\leq\varphi(\mathbf{x})\}} du\mu_\theta(d\mathbf{x}) = \int \varphi(\mathbf{x})\mu_\theta(d\mathbf{x}),$$

d.h. Niveau und Macht des neuen Tests werden so berechnet wie oben angegeben.

Die Menge der randomisierten Tests ist eine *konvexe* Menge von Funktionen, weshalb die folgende Optimierungsaufgabe mit einem Dualitätsargument gelöst werden kann.

Um unnötige Annahmen über μ_i zu vermeiden, benutzen wir, dass μ_i $(i = 0,1)$ bezüglich $\mu_0 + \mu_1$ immer eine Dichte p_i hat ($\rightarrow$ Masstheorie, Radon-Nikodym-Ableitung), d.h.

$$\mu_i(A) = \int_A p_i(\mathbf{x})(\mu_0(d\mathbf{x}) + \mu_1(d\mathbf{x})) \quad \forall A \in \mathcal{B}^n.$$

Wenn beide μ_i diskret sind, ist

$$p_i(\mathbf{x}) = \frac{\mu_i(\mathbf{x})}{\mu_0(\mathbf{x}) + \mu_1(\mathbf{x})}.$$

Falls beide μ_i absolut stetig sind mit Dichten f_i, dann ist

$$p_i(\mathbf{x}) = \frac{f_i(\mathbf{x})}{f_0(\mathbf{x}) + f_1(\mathbf{x})}.$$

Damit können wir das Hauptresultat formulieren und beweisen:
Satz 6.3 (Neyman-Pearson-Lemma)**.** *Seien μ_0 und μ_1 zwei Wahrscheinlichkeiten mit Dichten p_0 und p_1 bezüglich $\mu_0 + \mu_1$ und sei $\alpha \in [0,1]$ gegeben. Dann*

1. *Es existiert ein randomisierter Test φ und ein $c \in [0,\infty]$ derart dass*

$$\mathbb{E}_0(\varphi) = \alpha, \tag{6.3}$$

$$\varphi(\mathbf{x}) = \begin{cases} 1 & \text{falls} \quad p_1(\mathbf{x}) > c\, p_0(\mathbf{x}) \\ 0 & \text{falls} \quad p_1(\mathbf{x}) < c\, p_0(\mathbf{x}). \end{cases} \tag{6.4}$$

   (wobei $\infty \cdot 0$ als 0 definiert ist).

2. *Jeder Test, der (6.3) und (6.4) erfüllt, ist ein mächtigster Test zum Niveau α.*

3. *Jeder mächtigste Test zum Niveau α erfüllt (6.4) $(\mu_0+\mu_1)$-fast überall. Er erfüllt auch (6.3), ausser wenn es einen Test φ' gibt mit $\mathbb{E}_1(\varphi') = 1$ und $\mathbb{E}_0(\varphi') < \alpha$.*

Bemerkung: $p_1(x)/p_0(x)$ heisst der **Likelihoodquotient** und der Test von 1. der Likelihoodquotiententest. Kurz gesagt ist also der Likelihoodquotiententest optimal.

Beweis

Die Lösung der Optimierungsaufgabe, einen Test zum Niveau α und maximaler Macht zu finden, ist die Suche nach einem optimalen $0 \le \varphi \le 1$, sodass

$$I(\varphi p_0) \le \alpha, \quad I(\varphi p_1) \to \max$$

gilt. Hier bezeichne I das Integral bezüglich dem Mass $\mu_0+\mu_1$. Diese Optimierungsaufgabe kann im Sinne Lagrangescher Multiplikatoren in folgende Aufgabe umgeformt werden: Suche einen Multiplikator $c \ge 0$ und ein $0 \le \varphi \le 1$ mit $I(\varphi p_0) = \alpha$, so dass φ die Zielfunktion

$$\varphi' \mapsto I(\varphi' p_1) - cI(\varphi' p_0) = I(\varphi'(p_1 - cp_0)) \quad (0 \le \varphi' \le 1)$$

maximiert. Diese Zielfunktion wird offensichtlich genau dann maximal, wenn (6.4) fast überall bez. $\mu_0+\mu_1$ erfüllt ist.

Wie kommt diese Umformung des Optimierungsproblemes zustande? Sei C eine kompakte, konvexe Menge in einem euklidischen Vektorraum, dann gilt offensichtlich: $\varphi \in C$ ist eine Lösung des Maximierungsproblemes genau dann wenn

$$I((\varphi' - \varphi)p_1) \le 0$$

fúr alle $\varphi' \in C$ gilt. In anderen Worten: keine Bewegung weg von φ inerhalb von C kann zu einer Vergroesserung des Funktionals $\varphi' \mapsto I(\varphi' p_1)$ führen. Wenn es noch eine Nebenbedingung $I(\varphi' p_0) \le \alpha$ gibt, dann unterscheidet man zwei Fälle:

- Der Maximierer erfüllt die Nebenbedingung $I(\varphi p_0) = \alpha$. Dann ist $\varphi \in C$ mit $I(\varphi p_0) = \alpha$ eine Lösung des Maximierungsproblemes genau dann wenn es ein $c \ge 0$ gibt sodass
$$I((\varphi' - \varphi)(p_1 - cp_0)) \le 0$$
fúr alle $\varphi' \in C$ gilt.

- Der Maximierer nimmt die Nebenbedingung nicht an, das heisst es gibt nur Maximierer mit $I(\varphi p_0) < \alpha$. Dann ist $\varphi \in C$ mit $I(\varphi p_0) \le \alpha$ Lösung des Maximierungsproblems falls
$$I((\varphi' - \varphi)p_1) \le 0$$
fúr alle $\varphi' \in C$ gilt. Das entspricht einem Lagrangeschen Multiplikator $c = 0$.

Der Fall, dass die Menge der $\varphi \in C$ mit $I(\varphi p_0) \leq \alpha$ einpunktig ist, wird durch $c = +\infty$ codiert, weil dann das Maximierungsproblem sofort lösbar ist. Im Falle von Neyman-Pearson entspricht das dem Fall $\alpha = 0$.

Daraus folgt das Vorgehen: suche Lagrangeschen Multiplikator und $\varphi \in C$, sodass die Nebenbedingung angenommen wird, dh $I(\varphi p_0) = \alpha$. Dann hat man jedenfalls einen Maximierer gefunden, denn

$$I((\varphi' - \varphi)(p_1 - cp_0)) \leq 0$$

bedeutet $I(\varphi' p_1) \leq I(\varphi p_1)$ falls $I(\varphi' p_1) \leq I(\varphi p_1) = \alpha$. Es könnte noch andere geben, allerdings muss dann notwendigerweise $c = 0$ gelten.

Man muss also den Vektor $p_1 - cp_0$ so wählen, dass er mit jedem Tangentialvektor $\varphi' - \varphi$ (an einem Randpunkt φ des Argumentebereiches) einen Winkel von mindestens $\pi/2$ bezüglich des durch I induzierten Skalarprodukts $\langle \varphi_1, \varphi_2 \rangle = I(\varphi_1 \varphi_2)$ einschliesst. In diesem Sinne hat das im folgenden konstruierte c die Bedeutung eines Lagrangeschen Multiplikators. In Abbildung 6.1 sieht man das (blaue) Rechteck aller möglichen randomisierten Tests, einen (roten) Vektor, der für $p_1 - c\,p_0$ steht, und einen grünen Vektor, der für $\varphi' - \varphi$ am optimalen φ steht. Offensichtlich ist der eingeschlossene Winkel zwischen dem roten und grünen Vektor grösser als $\pi/2$.

```
> lagrange.graphics <- function(p01,p02,p11,p12,c)
+ {
+ x<-c(0,0,sqrt(p02+p12),sqrt(p02+p12),p11-c*p01+sqrt(p02+p01))
+ y<-c(0,sqrt(p01+p11),sqrt(p01+p02),0,p12-c*p02)
+ plot(x,y)
+ s<-seq(length(x)-2)
+ segments(x[s],y[s],x[s+1],y[s+1],col="blue")
+ segments(x[4],y[4],x[1],y[1],col="blue")
+ arrows(x[4],y[4],x[5],y[5],col="red")
+ arrows(x[4],y[4],x[1],y[1],col="green")
+ }
```

Wir beweisen daher zunächst die erste Aussage, dh die Existenz eines Lagrangeschen Multiplikators $c \in [0, \infty]$. Dazu betrachten wir die kumulative Verteilungsfunktion F der Zufallsvariable $Q = p_1(\mathbf{X})/p_0(\mathbf{X})$ für $\mathbf{X} \sim \mu_0$ (Q ist mit Wahrscheinlichkeit 1 definiert und endlich). Für $0 < \alpha < 1$, wählen wir $c = F^{-1}(1 - \alpha)$ (siehe (3.19) für die Definition der Quantilfunktion F^{-1}). Wenn F an der Stelle c einen Sprung hat, dann ist die Sprunghöhe gleich $P_0(Q = c) = P_0(p_1(\mathbf{X}) = cp_0(\mathbf{X})) > 0$, und wenn wir auf dieser Menge $\varphi = \gamma \in]0, 1[$ setzen mit geeignetem Wert von γ, dann gilt (6.3). Für $\alpha = 0$ setzen wir $c = +\infty$, d.h. $\varphi(\mathbf{x}) = 0$ für alle $\mathbf{x}$ mit $p_0(\mathbf{x}) > 0$ und $\varphi(\mathbf{x}) = 1$ für alle $\mathbf{x}$ mit $p_0(\mathbf{x}) = 0$. Für $\alpha = 1$ setzen wir $c = 0$ und $\varphi(\mathbf{x}) \equiv 1$.

Die Aussagen 2 und 3 folgen nun leicht. Sei φ der Test gemäss Aussage 1, der (6.3) und (6.4) erfüllt. Es gilt dann für jeden anderen randomisierten Test φ'

$$I(\varphi' p_1) \leq I(\varphi p_1) + c(I(\varphi' p_0) - I(\varphi p_0))$$

aufgrund der Konstruktion von ϕ: auf der Menge $p_1 - cp_0 > 0$ is ϕ maximal and auf $p_1 - cp_0 < 0$ verschwindet ϕ, also gilt die Ungleichung für jeden randomisierten Test. Wenn also $I(\varphi' p_0) \leq \alpha = I(\varphi p_0)$ gilt, dann ist $I(\varphi' p_1) \leq I(\varphi p_1)$, dh die zweite Aussage ist bewiesen. Damit ist bewiesen, dass an φ ein maximaler Wert der Zielfunktion angenommen wird. Es könnte aber auch sein, dass es weitere Maximierer gibt, allerdings mit geringerem Niveau.

```
> lagrange.graphics(0.2,0.8,0.7,0.3,1)
```

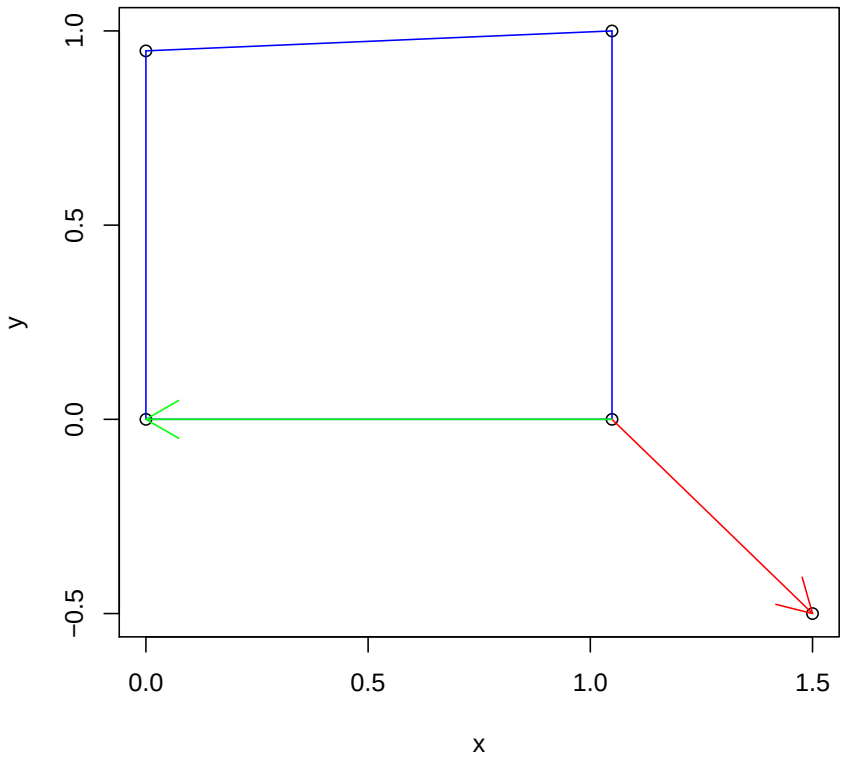

Abbildung 6.1: Visualisierung der Methode der Lagrangeschen Multiplikatoren

Für die dritte Aussage benutzt man, dass die obige Ungleichung strikt ist, ausser wenn (6.4) $(\mu_0+\mu_1)$-fast überall auch für φ' gilt. Aus $I(\varphi' p_1) = I(\varphi p_1)$ folgt ausserdem weiters, dass entweder $I(\varphi' p_0) = I(\varphi p_0)$ oder $c = 0$. Wenn $c = 0$, dann ist $\varphi = 1$ für alle $\mathbf{x}$ mit $p_1(\mathbf{x}) > 0$, dh die Macht von φ und damit auch von φ' ist 1. Der Wert von $I(\varphi' p_0)$ hängt noch davon ab, wie man φ' definiert auf $N = \{\mathbf{x}; p_0(\mathbf{x}) > 0, p_1(\mathbf{x}) = 0\}$. Den minimalen Wert $1 - \mu_0(N)$ erreicht man, wenn man $\varphi' = 0$ setzt auf N, und dieser kann kleiner als α sein. □

Beispiel 6.13. *Binomialverteilung mit Parameter θ. Sei $\theta_0 < \theta_1$. Dann ist*

$$\frac{p_1(x)}{p_0(x)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n r(\theta_0,\theta_1)^x$$

wobei

$$r(\theta_0,\theta_1) = \frac{\theta_1}{1-\theta_1}\frac{1-\theta_0}{\theta_0} > 1,$$

denn $\theta/(1-\theta)$ ist strikt monoton wachsend. Also gilt

$$\frac{p_1(x)}{p_0(x)} \geq c \Leftrightarrow x \geq c'.$$

Der Wert der kritischen Grenze c' ist bestimmt durch (6.3), d.h. er hängt nur von θ_0 und α ab. Also erhalten wir den gleichen optimalen Test für alle $\theta_1 > \theta_0$:

$$\varphi(x) = \begin{cases} 1 & \textit{falls} \quad x > c' \\ \gamma & \textit{falls} \quad x = c' \\ 0 & \textit{falls} \quad x < c' \end{cases}$$

(sogenannter gleichmässig mächtigster Test für $\theta = \theta_0$ gegen $\theta \in (\theta_0, 1]$.)

6.4 Vertrauensintervalle

Die Beobachtungen seien $\mathbf{X} = (X_1, \ldots, X_n)$, die Modellverteilungen für $\mathbf{X}$ seien $(\mu_\theta)_{\theta\in\Theta}$ und der interessierende Parameter sei $g(\theta), g : \Theta \to \mathbb{R}$.

Definition 6.4. *Seien $\underline{T}, \overline{T} \quad \mathbb{R}^n \to \mathbb{R}$ zwei messbare Funktionen mit $\underline{T} < \overline{T}$. Dann heisst $(\underline{T}(\mathbf{X}), \overline{T}(\mathbf{X}))$ ein* **Vertrauensintervall** *für $g(\theta)$ zum Niveau $1-\alpha$, falls*

$$\forall\theta \in \Theta \,:\, \mathbb{P}_\theta\left(\underline{T}(\mathbf{X}) < g(\theta) < \overline{T}(\mathbf{X})\right) > 1-\alpha\,.$$

Ein Vertrauensintervall fängt also den interessierenden Parameter mit Wahrscheinlichkeit $1-\alpha$ ein. Zufällig sind die Intervallgrenzen, nicht der Parameter. Wenn man N unabhängige Wiederholungen des Experiments machen würde, erhielte man N verschiedene Vertrauensintervalle, von denen ungefähr $N(1-\alpha)$ das wahre $g(\theta)$ enthalten.

Beispiel 6.14. *Normalverteilung:*
Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ mit $\theta = (\mu, \sigma^2)$ und $g(\theta) = \mu$.

Ferner sei $t(n-1, 1-\alpha/2)$ das $(1-\frac{\alpha}{2})$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden (siehe Satz 6.1) und

$$S_n^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$$

die Schätzung der Varianz σ^2. Dann ergibt

$$\overline{X} \pm t(n-1, 1-\alpha/2)\frac{S_n}{\sqrt{n}}$$

ein Vertrauensintervall zum Niveau $1-\alpha$*, denn*

$$\overline{X}-t(n-1,1-\alpha/2)\frac{S_n}{\sqrt{n}} \le \mu \le \overline{X}+t(n-1,1-\alpha/2)\frac{S_n}{\sqrt{n}} \Leftrightarrow \frac{\sqrt{n}|\overline{X}-\mu|}{S_n} \le t(n-1,1-\alpha/2).$$

Also folgt das Resultat aus Satz 6.1.

In diesem Beispiel besteht das Vertrauensintervall aus denjenigen Werten μ, für die die Nullhypothese $\mathbb{E}(X_i)=\mu$ akzeptiert wird. Ein solcher Zusammenhang besteht allgemein.

Satz 6.4 (Dualitätssatz)**.** *Sei C eine messbare Teilmenge von $\mathbb{R}^n \times \mathbb{R}$ mit den messbaren Schnitten $A(\gamma)=\{\mathbf{x}\in\mathbb{R}^n|(\mathbf{x},\gamma)\in C\}$ und $B(\mathbf{x})=\{\gamma\in\mathbb{R}|(\mathbf{x},\gamma)\in C\}$. Dann sind die beiden folgenden Aussagen äquivalent:*

i) Für jedes γ ist $\varphi(\mathbf{x})=1_{A(\gamma)^c}(\mathbf{x})$ ein Test der Nullhypothese $g(\theta)=\gamma$ zum Niveau α mit Verwerfungsbereich $A(\gamma)^c$.

ii) $B(\mathbf{X})$ bildet einen Vertrauensbereich für $g(\theta)$ zum Niveau $1-\alpha$.

Beweis i) besagt $\mathbb{P}_\theta(\mathbf{X}\notin A(\gamma))\le\alpha$ für alle θ mit $g(\theta)=\gamma$. ii) besagt $\mathbb{P}_\theta(g(\theta)\in B(\mathbf{X}))>1-\alpha$ für alle θ, für alle γ. Nach Definition von $A(\gamma)$ und $B(\mathbf{x})$ gilt aber

$$g(\theta)\in B(\mathbf{x}) \Leftrightarrow (\mathbf{x},g(\theta))\in C \Leftrightarrow \mathbf{x}\in A(g(\theta)).$$

Daraus folgt die Behauptung. □

Familien von Tests für die Nullhypothesen $g(\theta)=\gamma$ sind also äquivalent zu Vertrauensbereichen. Diese sind häufig, aber nicht immer Intervalle.

Beispiel 6.15. *Binomialverteilung:*
Sei $X\sim$ Binomial(n,p)-verteilt mit $\theta=p\in(0,1)$. Zunächst bestimmen wir ein exaktes Vertrauensintervall. Wir beginnen mit einem Test für $p=p_0$. Wir fixieren α und bestimmen $\underline{k}(p_0)$ und $\overline{k}(p_0)$ derart dass,

$$\begin{aligned}\sum_{j=0}^{\underline{k}-1}\binom{n}{j}p_0^j(1-p_0)^{n-j} &\le \frac{\alpha}{2} < \sum_{j=0}^{\underline{k}}\binom{n}{j}p_0^j(1-p_0)^{n-j}\\ \sum_{j=\overline{k}}^{n}\binom{n}{j}p_0^j(1-p_0)^{n-j} &> \frac{\alpha}{2} \ge \sum_{j=\overline{k}+1}^{n}\binom{n}{j}p_0^j(1-p_0)^{n-j}.\end{aligned}$$

Der Test

$$\varphi=0 \Leftrightarrow X\in\{\underline{k}(p_0),\dots,\overline{k}(p_0)\}$$

hat dann Niveau α. Der entsprechende Vertrauensbereich nach Satz 6.4 ist

$$B(x)=\{p;\ \underline{k}(p)\le x\le\overline{k}(p)\}.$$

Da $\sum_{j=0}^{k}\binom{n}{j}p^i(1-p)^{n-j}$ für festes k monoton fallend in p ist, ergibt sich

$$B(x)=\left[\underline{p}(x),\overline{p}(x)\right],$$

wobei

$$\begin{aligned}\sum_{j=0}^{x}\binom{n}{j}\overline{p}(x)^j(1-\overline{p}(x))^{n-j} &= \frac{\alpha}{2}\quad(x\ne n),\\ \sum_{j=x}^{n}\binom{n}{j}\underline{p}(x)^j(1-\underline{p}(x))^{n-j} &= \frac{\alpha}{2}\quad(x\ne 0).\end{aligned}$$

Für $x=n$ ist $\overline{p}=1$, für $x=0$ ist $\underline{p}=0$.

b) Ein einfacheres, genähertes Vertrauensintervall erhalten wir aus dem Zentralen Grenzwertsatz. Asymptotisch ist

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1).$$

Mit $z_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$ *hat daher der Test*

$$\varphi = 0 \Leftrightarrow |X - np_0| \leq z_\alpha \sqrt{np_0(1-p_0)}$$

das Niveau $\approx \alpha$ *für die Nullhypothese* $p = p_0$*. Der entsprechende Vertrauensbereich ist*

$$B(x) = \{p;\ |x - np| \leq z_\alpha \sqrt{np(1-p)}\}.$$

Durch Umformen erhält man

$$(x - np)^2 \leq z_\alpha^2 np(1-p) \Longleftrightarrow p^2(n + z_\alpha^2) - 2p(x + \frac{z_\alpha^2}{2}) + \frac{x^2}{n} \leq 0,$$

was äquivalent ist zu

$$B(x) = \left[\underline{p}(x), \overline{p}(x)\right]$$

wobei $\underline{p}, \overline{p}$ *die Stellen sind, wo in obiger Ungleichung das Gleichheitszeichen gilt. Man erhält*

$$\begin{aligned} \underline{p}(x) &= \frac{x + z_\alpha^2/2 - z_\alpha \sqrt{x(1-x/n) + z_\alpha^2/4}}{n + z_\alpha^2}, \\ \overline{p}(x) &= \frac{x + z_\alpha^2/2 + z_\alpha \sqrt{x(1-x/n) + z_\alpha^2/4}}{n + z_\alpha^2}. \end{aligned}$$

Anhang A

Grundlagen der Masstheorie

A.1 Mengensysteme

Es sei $\Omega \neq \emptyset$ eine Menge. Im Folgenden bezeichnen wir mit $\mathcal{P}(\Omega)$ die Potenzmenge, also das System aller Teilmengen von Ω.

Definition A.1. *Ein Mengensystem $\mathcal{A} \subset \mathcal{P}(\Omega)$ heisst* σ-Algebra *(über Ω), wenn $\mathcal{A}$ folgende Eigenschaften hat:*

1. $\Omega \in \mathcal{A}$.
2. *Für jedes $A \in \mathcal{A}$ ist $A^c \in \mathcal{A}$.*
3. *Für jede Folge $(A_n)_{n\in\mathbb{N}}$ von Teilmengen aus $\mathcal{A}$ gilt $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.*

Ist $\mathcal{A}$ eine σ-Algebra über Ω und $A \subset \Omega$ eine Teilmenge, so ist auch

$$\mathcal{A}|A := \{A \cap B : B \in \mathcal{A}\}$$

eine σ-Algebra, die sog. *Spur-σ-Algebra* auf A.

σ-Algebren sind die wichtigsten Mengensysteme in der Masstheorie, denn sie werden uns später als Definitionsbereiche von Massen begegnen. Nun lassen sich σ-Algebren wie z.B. die σ-Algebra der Borelschen Teilmengen des $\mathbb{R}^n$ (siehe Definition A.2 unten) i.a. nicht durch unmittelbares Hinschreiben der Elemente angeben. Oft werden σ-Algebren durch Angabe eines sog. Erzeugers definiert. Für ein beliebiges Mengensystem $\mathcal{E} \subset \mathcal{P}(\Omega)$ definieren wir die σ-Algebra

$$\sigma(\mathcal{E}) := \bigcap_{\substack{\mathcal{A}\supset\mathcal{E},\\ \mathcal{A} \text{ ist } \sigma\text{-Algebra}}} \mathcal{A}.$$

Sie heisst die *von $\mathcal{E}$ erzeugte σ-Algebra* über Ω, und $\mathcal{E}$ heisst ein *Erzeuger* von $\sigma(\mathcal{E})$ über Ω. Beachte, dass der Erzeuger einer σ-Algebra keineswegs eindeutig bestimmt ist.

Die für die Zwecke der Masstheorie wichtigste σ-Algebra ist die σ-Algebra der Borelschen Teilmengen des $\mathbb{R}^n$.

Definition A.2. *Es sei $\mathcal{O}^n$ das System der offenen Teilmengen des $\mathbb{R}^n$. Dann heisst die von $\mathcal{O}^n$ erzeugte σ-Algebra $\mathcal{B}(\mathbb{R}^n) := \sigma(\mathcal{O}^n)$ die σ-Algebra der* Borelschen Teilmengen von $\mathbb{R}^n$. *Wir benutzen die Abkürzungen $\mathcal{B}^n := \mathcal{B}(\mathbb{R}^n)$ und $\mathcal{B} := \mathcal{B}^1$.*

Satz A.1. *Jedes der folgenden Mengensysteme ist ein Erzeuger der σ-Algebra $\mathcal{B}^n$ der Borelschen Teilmengen des $\mathbb{R}^n$:*

$$\begin{aligned}\mathcal{O}^n &:= \{U \subset \mathbb{R}^n : U \text{ ist offen}\},\\ \mathcal{C}^n &:= \{A \subset \mathbb{R}^n : A \text{ ist abgeschlossen}\},\\ \mathcal{K}^n &:= \{K \subset \mathbb{R}^n : K \text{ ist kompakt}\},\\ \mathcal{I}^n &:= \{(a,b] : a,b \in \mathbb{R}^n \text{ mit } a \leq b\}.\end{aligned}$$

Für Anwendungen – vor allem im Fall $\Omega = \mathbb{R}^n$ – besonders bequem ist eine Klasse von Erzeugern mit einigen einfachen strukturellen Eigenschaften, die sog. Halbringe.
Definition A.3. *Ein Mengensystem $\mathcal{H} \subset \mathcal{P}(\Omega)$ heisst* Halbring *(über Ω), wenn $\mathcal{H}$ folgende Eigenschaften hat:*

1. $\emptyset \in \mathcal{H}$.
2. *Für alle $A, B \in \mathcal{H}$ ist $A \cap B \in \mathcal{H}$.*
3. *Für alle $A, B \in \mathcal{H}$ gibt es eine natürliche Zahl $m \in \mathbb{N}$ und disjunkte Mengen $C_1, \ldots, C_m \in \mathcal{H}$, so dass $A \setminus B = \bigcup_{k=1}^m C_k$.*

Wie sich herausstellt, besitzen die Borelschen Teilmengen einen Erzeuger, der ein Halbring ist:
Lemma A.1. *$\mathcal{I}^n$ ist ein Halbring über $\mathbb{R}^n$.*

A.2 Masse und Prämasse

Die wichtigste Eigenschaft des elementargeometrischen Volumenbegriffs ist die Additivität: Das Volumen einer disjunkten Vereinigung von Mengen ist gleich der Summe der Volumina der Teilmengen. Bei der folgenden Definition eines Masses wird diese Eigenschaft als Axiom an die Spitze gestellt..
Definition A.4. *Es sei $\mathcal{A}$ ein σ-Algebra über der Menge Ω. Eine Abbildung $\mu : \mathcal{A} \to \overline{\mathbb{R}}$ heisst ein* Mass*, falls μ folgende Eigenschaften hat:*

1. $\mu(\emptyset) = 0$.
2. $\mu \geq 0$.
3. *Für jede Folge $(A_n)_{n\in\mathbb{N}}$ disjunkter Mengen aus $\mathcal{A}$ gilt*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Ist μ ein Mass auf einer σ-Algebra $\mathcal{A}$ über Ω, so heisst $(\Omega, \mathcal{A}, \mu)$ ein *Massraum*. Massräume sind fundamentale Objekte in der Mass- und Integrationstheorie. Da σ-Algebren i.a. nur durch einen Erzeuger gegeben sind, ist zur konkreten Definition von Massen die folgende Definition eines Prämasses naheliegend.
Definition A.5. *Es sei $\mathcal{H}$ ein Halbring über der Menge Ω. Eine Abbildung $\mu : \mathcal{H} \to \overline{\mathbb{R}}$ heisst ein* Prämass*, falls μ folgende Eigenschaften hat:*

1. $\mu(\emptyset) = 0$.
2. $\mu \geq 0$.

3. Für jede Folge $(A_n)_{n\in\mathbb{N}}$ disjunkter Mengen aus $\mathcal{H}$ gilt

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Nach Satz A.1 ist $\mathcal{I}^n$ ein Erzeuger der Borelschen Teilmengen des $\mathbb{R}^n$, der gemäss Lemma A.1 zugleich ein Halbring ist. Das wichtigste Prämass auf $\mathcal{I}^n$ ist das elementargeometrische Volumen

$$\lambda^n((a,b]) := \prod_{j=1}^{n}(b_j - a_j)$$

für $a, b \in \mathbb{R}^n$ mit $a \leq b$.
Lemma A.2. *$\lambda^n : \mathcal{I}^n \to \mathbb{R}$ ist ein Prämass auf $\mathcal{I}^n$.*

Wir nennen λ^n das *Lebesguesche Prämass auf $\mathcal{I}^n$*.

A.3 Fortsetzung eines Prämasses zu einem Mass

Wir sind nun daran interessiert, ein Prämass auf einem Halbring $\mathcal{H}$ zu einem Mass auf die von $\mathcal{H}$ erzeugte σ-Algebra $\sigma(\mathcal{H})$ fortzusetzen. Von besonderem Interesse ist die Fortsetzung des Lebesgueschen Prämasses auf die Borelsche σ-Algebra $\mathcal{B}^n$. Der folgende Fortsetzungssatz geht auf C. Carathéodory (1914) zurück.
Satz A.2. *Jedes Prämass $\mu : \mathcal{H} \to \overline{\mathbb{R}}$ auf einem Halbring $\mathcal{H}$ über Ω kann auf mindestens eine Weise zu einem Mass auf $\sigma(\mathcal{H})$ fortgesetzt werden.*

Nach dem Fortsetzungssatz A.2 lässt sich jedes auf einem Halbring definierte Prämass zu einem Mass auf der σ-Algebra $\sigma(\mathcal{H})$ fortsetzen. Eine solche Fortsetzung ist i.a. allerdings nicht eindeutig bestimmt. Die genauere Untersuchung des Eindeutigkeitsproblems wird ergeben, dass Eindeutigkeit vorliegt, wenn man Ω durch abzählbar viele Mengen endlichen Inhalts überdecken kann.
Definition A.6. *Ein Prämass $\mu : \mathcal{H} \to \overline{\mathbb{R}}$ auf einem Halbring $\mathcal{H}$ über Ω heisst* σ-endlich, *wenn eine Folge $(E_n)_{n\in\mathbb{N}}$ von Mengen aus $\mathcal{H}$ mit $\mu(E_n) < \infty$ für alle $n \in \mathbb{N}$ und $\Omega = \bigcup_{n=1}^{\infty} E_n$ existiert.*
Lemma A.3. *Das Lebesguesche Prämass λ^n auf $\mathcal{I}^n$ ist σ-endlich.*

Für σ-endliche Prämasse gilt der folgende Eindeutigkeitssatz.
Satz A.3. *Jedes σ-endliche Prämass $\mu : \mathcal{H} \to \overline{\mathbb{R}}$ auf einem Halbring $\mathcal{H}$ über Ω kann auf genau eine Weise zu einem Mass auf $\sigma(\mathcal{H})$ fortgesetzt werden.*

Nach Lemma A.3 ist das Lebesguesche Prämass σ-endlich. Gemäss Satz A.3 kann λ^n auf genau eine Weise zu einem Mass auf die Borelsche σ-Algebra $\mathcal{B}^n$ fortgesetzt werden. Wir nennen diese Fortsetzung $\lambda^n : \mathcal{B}^n \to \overline{\mathbb{R}}$ das *Lebesgue-Borel-Mass*. Speziell setzen wir $\lambda := \lambda^1$.

A.4 Vollständige Massräume

Definition A.7. *Ist $\mu : \mathcal{A} \to \overline{\mathbb{R}}$ ein Mass auf einer σ-Algebra $\mathcal{A}$ über Ω, so heisst $A \subset \Omega$ eine* μ-Nullmenge, *wenn $A \in \mathcal{A}$ und $\mu(A) = 0$.*

Definition A.8. *Ein Massraum $(\Omega, \mathcal{A}, \mu)$ heisst* vollständig*, wenn jede Teilmenge einer μ-Nullmenge $A \in \mathcal{A}$ zu $\mathcal{A}$ gehört (und damit selbst eine μ-Nullmenge ist). Ist $(\Omega, \mathcal{A}, \mu)$ vollständig, so nennt man auch μ* vollständig*..*

Ist ein Massraum $(\Omega, \mathcal{A}, \mu)$ unvollständig, so kann man stets das Mass $\mu : \mathcal{A} \to \overline{\mathbb{R}}$ zu einem vollständigen Mass auf eine grössere σ-Algebra fortsetzen. Folgendes Verfahren, die sog. *Vervollständigung*, liefert zu jedem Mass eine vollständige Fortsetzung mit *minimalem* Definitionsbereich.

Satz A.4. *Es seien $(\Omega, \mathcal{A}, \mu)$ ein Massraum, $\mathcal{N}$ das System aller Teilmengen von μ-Nullmengen und*

$$\tilde{\mathcal{A}} := \{A \cup N : A \in \mathcal{A} \text{ und } N \in \mathcal{N}\},$$
$$\tilde{\mu} : \tilde{\mathcal{A}} \to \overline{\mathbb{R}}, \quad \tilde{\mu}(A \cup N) := \mu(A) \text{ für } A \in \mathcal{A} \text{ und } N \in \mathcal{N}.$$

Dann gilt:

1. *$\tilde{\mathcal{A}}$ ist eine σ-Algebra, $\tilde{\mu}$ ist wohldefiniert, und $(\Omega, \tilde{\mathcal{A}}, \tilde{\mu})$ ist ein vollständiger Massraum. $\tilde{\mu}$ ist die einzige Fortsetzung von μ zu einem Mass auf $\tilde{\mathcal{A}}$.*
2. *Jede vollständige Fortsetzung ρ von μ ist eine Fortsetzung von $\tilde{\mu}$.*

Die Vervollständigung $\tilde{\lambda}^n : \mathcal{L}^n \to \overline{\mathbb{R}}$ des Lebesgue-Borel-Masses, die wir wieder mit λ^n bezeichnen, heisst das *Lebesgue-Mass*. Dabei bezeichnet $\mathcal{L}^n := \tilde{\mathcal{B}}^n$ das System der *Lebesgue-messbaren* Mengen. Speziell setzen wir $\mathcal{L} := \mathcal{L}^1$ und $\lambda := \lambda^1$.

Mit Hilfe des Auswahlaxioms lässt sich zeigen, dass die Inklusionen zwischen den Borelschen Teilmengen, den Lebesgue-messbaren Mengen und allen Teilmengen des $\mathbb{R}^n$ strikt sind, d.h.

$$\mathcal{B}^n \subsetneqq \mathcal{L}^n \subsetneqq \mathcal{P}(\mathbb{R}^n).$$

A.5 Messbare Abbildungen

Ist $\mathcal{A}$ eine σ-Algebra über Ω, so nennen wir das Paar $(\Omega, \mathcal{A})$ einen *Messraum* oder einen *messbaren Raum*; die Mengen aus $\mathcal{A}$ heissen *messbare Mengen*. (Dabei wird *nicht* vorausgesetzt, dass auf $\mathcal{A}$ ein Mass definiert ist. Ist zusätzlich $\mu : \mathcal{A} \to \overline{\mathbb{R}}$ ein Mass auf $\mathcal{A}$, so heisst $(\Omega, \mathcal{A}, \mu)$ ein *Massraum*.)

Definition A.9. *Es seien $(\Omega_1, \mathcal{A}_1)$, $(\Omega_2, \mathcal{A}_2)$ Messräume. Eine Funktion $f : \Omega_1 \to \Omega_2$ heisst $\mathcal{A}_1$-$\mathcal{A}_2$-*messbar *oder kurz* messbar*, wenn gilt $f^{-1}(\mathcal{A}_2) \subset \mathcal{A}_1$.*

Hierbei benutzen wir die Schreibweise

$$f^{-1}(\mathcal{A}_2) = \{f^{-1}(B) : B \in \mathcal{A}_2\}.$$

Sollen die zugrundeliegenden σ-Algebren ausdrücklich hervorgehoben werden, so schreiben wir kurz $f : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$.

Satz A.5. *Sind $f : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ eine Abbildung und $\mathcal{E} \subset \mathcal{A}_2$ ein Erzeuger von $\mathcal{A}_2$, so ist f genau dann $\mathcal{A}_1$-$\mathcal{A}_2$-messbar, wenn $f^{-1}(\mathcal{E}) \subset \mathcal{A}_1$.*

Definition A.10. *Sind $f_1, \ldots, f_n$ Funktionen von Ω in Messräume $(\Omega_1, \mathcal{A}_1), \ldots, (\Omega_n, \mathcal{A}_n)$, so ist*

$$\sigma(f_1, \ldots, f_n) = \sigma(\{f_i^{-1}(A) | 1 \leq i \leq n,\ A \in \mathcal{A}_i\})$$

die kleinste σ-Algebra, bezüglich derer die f_i messbar sind. Sie wird als die von $f_1, \ldots, f_n$ erzeugte σ-Algebra bezeichnet.

Wir nennen eine Abbildung $f : \mathbb{R}^m \to \mathbb{R}^n$ kurz *Borel-messbar*, wenn sie $\mathcal{B}^m$-$\mathcal{B}^n$-messbar ist.

Korollar A.1. *Jede auf einer Teilmenge $A \subset \mathbb{R}^m$ definierte stetige Funktion $f : A \to \mathbb{R}^n$ ist Borel-messbar (d.h. $(\mathcal{B}^m|A)$-$\mathcal{B}^n$-messbar).*

Satz A.6. *Sind $(\Omega_1, \mathcal{A}_1)$, $(\Omega_2, \mathcal{A}_2)$, $(\Omega_3, \mathcal{A}_3)$ Messräume und die Abbildungen $f : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$, $g : (\Omega_2, \mathcal{A}_2) \to (\Omega_3, \mathcal{A}_3)$ messbar, so ist auch $g \circ f : (\Omega_1, \mathcal{A}_1) \to (\Omega_3, \mathcal{A}_3)$ messbar.*

A.6 Messbare numerische Funktionen

Für die Zwecke der Integrationstheorie ist es bequem, nicht nur messbare Funktionen $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ zu betrachten, sondern auch Funktionen mit Werten in $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$.

Eine Menge $A \subset \overline{\mathbb{R}}$ heisst *offen*, wenn $A \cap \mathbb{R}$ offen in $\mathbb{R}$ ist und wenn im Falle $+\infty \in A$ (bzw. $-\infty \in A$) ein $a \in \mathbb{R}$ existiert mit $(a, \infty] \subset A$ (bzw. $[-\infty, a) \subset A$).

Die σ-Algebra $\overline{\mathcal{B}} := \mathcal{B}(\overline{\mathbb{R}})$ der Borelschen Teilmengen von $\overline{\mathbb{R}}$ ist per Definition die von den offenen Teilmengen von $\overline{\mathbb{R}}$ erzeugte σ-Algebra. Man erkennt:

$$\overline{\mathcal{B}} = \{B \cup E : B \in \mathcal{B} \text{ und } E \subset \{-\infty, +\infty\}\}.$$

Insbesondere ist $\overline{\mathcal{B}}|\mathbb{R} = \mathcal{B}$.

Im folgenden sei $(\Omega, \mathcal{A})$ stets ein Messraum. Zur Unterscheidung von den reellwertigen Funktionen auf Ω nennen wir die Funktionen $f : \Omega \to \overline{\mathbb{R}}$ *numerische Funktionen*. Eine numerische Funktion heisst *messbar*, wenn sie $\mathcal{A}$-$\overline{\mathcal{B}}$-messbar ist. Für reellwertiges f ist die $\mathcal{A}$-$\overline{\mathcal{B}}$-Messbarkeit gleichbedeutend mit der $\mathcal{A}$-$\mathcal{B}$-Messbarkeit.

Für die zu entwickelnde Integrationstheorie ist die Möglichkeit der Approximation messbarer Funktionen durch Treppenfunktionen von entscheidender Bedeutung.

Definition A.11. *Eine messbare Funktion $f : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$, die nur endlich viele verschiedene (reelle) Werte annimmt, heisst eine* ($\mathcal{A}$-)Treppenfunktion. *Es seien $\mathcal{T}$ die Menge der ($\mathcal{A}$-)Treppenfunktion auf Ω und $\mathcal{T}^+$ die Menge der nicht-negativen Funktionen aus $\mathcal{T}$.*

Wir bezeichnen mit $\mathcal{M}$ die Menge der messbaren *numerischen* Funktionen $f : (\Omega, \mathcal{A}) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ und mit $\mathcal{M}^+$ die Menge der nicht-negativen Funktionen aus $\mathcal{M}$. Folgender Satz ist für die spätere Integraldefinition von entscheidender Bedeutung:

Satz A.7. *Für eine nicht-negative numerische Funktion f auf Ω gilt $f \in \mathcal{M}^+$ genau dann, wenn es eine Folge $(u_n)_{n\in\mathbb{N}}$ von Funktionen aus $\mathcal{T}^+$ mit $u_n \uparrow f$ gibt.*

A.7 Integration von Treppenfunktionen

Für den Rest dieses Kapitels sei $(\Omega, \mathcal{A}, \mu)$ ein Massraum. Messbarkeit von Funktionen $f : \Omega \to \overline{\mathbb{R}}$ ist stets in Bezug auf die σ-Algebra $\mathcal{A}$ zu verstehen.

Bei der Einführung des Integralbegriffs gehen wir in drei Schritten vor: Zunächst definieren wir das Integral für nicht-negative Treppenfunktionen, dehnen die Definition dann mit Hilfe monotoner Folgen auf beliebige Funktionen aus $\mathcal{M}^+$ aus und führen anschliessend den Integralbegriff für integrierbare Funktionen auf den Integralbegriff für Funktionen aus $\mathcal{M}^+$ zurück.

Lemma A.4. *Die Funktion $f \in \mathcal{T}^+$ habe die Darstellungen*

$$f = \sum_{j=1}^{m} \alpha_j 1_{A_j} = \sum_{k=1}^{n} \beta_k 1_{B_k}$$

mit $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n \geq 0$ und $A_1, \ldots, A_m, B_1, \ldots, B_n \in \mathcal{A}$. Dann gilt

$$\sum_{j=1}^{m} \alpha_j \mu(A_j) = \sum_{k=1}^{n} \beta_k \mu(B_k).$$

Nun ist die folgende Definition sinnvoll, denn sie hängt nicht von der Auswahl der Darstellung von f ab.

Definition A.12. *Für $f \in \mathcal{T}^+$, $f = \sum_{j=1}^{m} \alpha_j 1_{A_j}$ mit $\alpha_1, \ldots, \alpha_m \geq$ und $A_1, \ldots, A_m \in \mathcal{A}$ heisst*

$$\int_\Omega f d\mu := \sum_{j=1}^{m} \alpha_j \mu(A_j) \quad (\in [0, \infty])$$

das (μ-)Integral von f *(über Ω).*

A.8 Integration nicht-negativer messbarer Funktionen

Wir erweitern den Integralbegriff durch Bildung monotoner Limiten von Funktionen aus $\mathcal{T}^+$: Zu jedem $f \in \mathcal{M}^+$ gibt es nach Satz A.7 eine Folge $(u_n)_{n\in\mathbb{N}}$ in $\mathcal{T}^+$ mit $u_n \uparrow f$ für $n \to \infty$, und es bietet sich die Definition

$$\int_\Omega f d\mu := \lim_{n\to\infty} \int_\Omega u_n d\mu$$

an. Diese Definition erweist sich als sinnvoll, denn sie hängt nicht von der speziellen Auswahl der Folge $(u_n)_{n\in\mathbb{N}}$ ab. Der Nachweis der Unabhängigkeit von der speziellen Auswahl beruht auf folgendem Satz:

Satz A.8. *Sind $(u_n)_{n\in\mathbb{N}}$, $(v_n)_{n\in\mathbb{N}}$ zwei wachsende Folgen von Funktionen aus $\mathcal{T}^+$ mit $\lim_{n\to\infty} u_n = \lim_{n\to\infty} v_n$, so gilt*

$$\lim_{n\to\infty} \int_\Omega u_n d\mu = \lim_{n\to\infty} \int_\Omega v_n d\mu.$$

Definition A.13. *Es seien $f \in \mathcal{M}^+$ und $(u_n)_{n\in\mathbb{N}}$ eine Folge von Funktionen aus $\mathcal{T}^+$ mit $u_n \uparrow f$. Dann heisst das von der Auswahl der Folge $(u_n)_{n\in\mathbb{N}}$ unabhängige Element*

$$\int_\Omega f d\mu := \lim_{n\to\infty} \int_\Omega u_n d\mu$$

das (μ-) Integral von f *(über Ω).*

Schreibt man $u \in \mathcal{T}^+$ als Limes der konstanten Folge $u_n := u$ ($n \in \mathbb{N}$), so erklärt dies, dass Definition A.13 für Treppenfunktionen denselben Integralwert wie Definition A.12 liefert.

A.9 Integrierbare Funktionen

In einem dritten und letzten Konstruktionsschritt dehnen wir den Integralbegriff auf geeignete messbare Funktionen aus.

Für jede numerische Funktion $f : \Omega \to \overline{\mathbb{R}}$ sind der *Positivteil*

$$f^+ := \max(f, 0)$$

und der *Negativteil*

$$f^- := \max(-f, 0)$$

erklärt, und wir haben die Zerlegungen

$$f = f^+ - f^- \quad \text{und} \quad |f| = f^+ + f^-.$$

Eine Funktion $f : \Omega \to \overline{\mathbb{R}}$ ist genau dann messbar, wenn der Positivteil f^+ und der Negativteil f^- messbar sind, und mit f ist auch $|f|$ messbar.

Definition A.14. *Eine Funktion $f : \Omega \to \overline{\mathbb{R}}$ heisst* (μ-)integrierbar *(über Ω), wenn f* messbar *ist und wenn die zwei Integrale*

$$\int_\Omega f^+ d\mu \quad \textit{und} \quad \int_\Omega f^- d\mu$$

beide endlich *sind, und dann heisst die reelle Zahl*

$$\int_\Omega f d\mu := \int_\Omega f^+ d\mu - \int_\Omega f^- d\mu \tag{A.1}$$

das (μ-) Integral von f *(über Ω) oder das* Lebesgue-Integral von f *(über Ω bez. μ).*

Wenn die Deutlichkeit eine klare Bezeichnung der Integrationsvariablen erfordert, schreiben wir ausführlicher

$$\int_\Omega f d\mu = \int_\Omega f(\omega) d\mu(\omega) \quad \text{oder} \quad \int_\Omega f d\mu = \int_\Omega f(\omega) \mu(d\omega).$$

Eine Funktion $f \in \mathcal{M}^+$ ist genau dann integrierbar, wenn ihr μ-Integral über Ω endlich ist, und dann stimmt das Integral (A.1) mit der früheren Begriffsbildung überein.

Als nächstes sammeln wir ein paar Kriterien für die Integrierbarkeit einer gegebenen Funktion $f : \Omega \to \overline{\mathbb{R}}$.

Satz A.9. *Für jede Funktion $f : \Omega \to \overline{\mathbb{R}}$ sind folgende Aussagen äquivalent:*

1. *f ist integrierbar.*
2. *f^+ und f^- sind integrierbar.*
3. *Es gibt integrierbare Funktionen $u, v \in \mathcal{M}^+$ mit $f = u - v$.*
4. *f ist messbar, und es gibt ein integrierbares $g \in \mathcal{M}^+$ mit $|f| \leq g$.*
5. *f ist messbar und $|f|$ integrierbar.*

Die nächsten beiden Resultate liefern die *Linearität* und die *Monotonie* des Lebesgue-Integrals.

Satz A.10. *Sind $f, g : \Omega \to \overline{\mathbb{R}}$ integrierbar und $\alpha, \beta \in \mathbb{R}$, so ist auch die Funktion $\alpha f + \beta g$ integrierbar, und es gilt*

$$\int_\Omega (\alpha f + \beta g) d\mu = \alpha \int_\Omega f d\mu + \beta \int_\Omega g d\mu.$$

Satz A.11. *Sind $f, g : \Omega \to \overline{\mathbb{R}}$ integrierbar und $f \leq g$, so gilt*

$$\int_\Omega f d\mu \leq \int_\Omega g d\mu.$$

Weiterhin gilt für integrierbare Funktionen die sog. *Dreiecksungleichung.*
Satz A.12. *Ist $f : \Omega \to \overline{\mathbb{R}}$ integrierbar, so gilt*

$$\left| \int_\Omega f d\mu \right| \leq \int_\Omega |f| d\mu.$$

A.10 Integration über messbare Teilmengen

Ist $f : \Omega \to \overline{\mathbb{R}}$ integrierbar, so bieten sich zwei Möglichkeiten zur Definition des Integrals von f über messbare Teilmengen $A \subset \Omega$ an:

1. Wir integrieren die Funktion $f 1_A$ über Ω.
2. Wir bilden den neuen Massraum $(A, \mathcal{A}|A, \mu|(\mathcal{A}|A))$ und integrieren $f|A$ bez. $\mu|(\mathcal{A}|A)$.

Beide Ansätze führen zum gleichen Resultat:
Lemma A.5. *Sind $A \in \mathcal{A}$ und $f : \Omega \to \overline{\mathbb{R}}$, so sind folgende Aussagen äquivalent:*

1. *$f 1_A$ ist μ-integrierbar.*
2. *$f|A$ ist $\mu|(\mathcal{A}|A)$-integrierbar.*

In diesem Fall gilt

$$\int_\Omega f 1_A d\mu = \int_A f|A \, d\mu|(\mathcal{A}|A).$$

Definition A.15. *Ist in der Situation des Lemmas A.5 die Funktion $f 1_A$ integrierbar, so heisst*

$$\int_A f d\mu := \int_\Omega f \mathbb{1}_A d\mu = \int_A f|A \, d\mu|(\mathcal{A}|A).$$

das (μ-)Integral von f über A.

Wir betrachten nun den Spezialfall $\Omega = \mathbb{R}^n$. Ist $A \in \mathcal{L}^n$, so heisst eine $\mathcal{L}^n|A$-messbare Funktion $f : A \to \overline{\mathbb{R}}$ *Lebesgue-messbar*, und für *Lebesgue-integrierbares* $f : A \to \overline{\mathbb{R}}$ schreiben wir

$$\int_A f(x) dx := \int_A f d\lambda^n. \tag{A.2}$$

Im Fall $\Omega = \mathbb{R}$ ist es wegen der Eigenschaft $\lambda(\{x\}) = 0$, $x \in \mathbb{R}$ legitim, für $a, b \in \mathbb{R}$ mit $a < b$ zu definieren

$$\int_a^b f(x) dx := \int_{[a,b]} f(x) dx.$$

Die Definition (A.2) erinnert an das Riemann-Integral einer Funktion f und der folgende Satz zeigt, dass diese Bezeichung in der Tat gerechtfertigt ist.

Satz A.13. *Eine beschränkte Funktion $f : [a, b] \to \mathbb{R}$ ($a, b \in \mathbb{R}^n$ mit $a < b$) ist genau dann Riemann-integrierbar, wenn die Menge ihrer Unstetigkeitsstellen eine λ^n-Nullmenge ist, und dann stimmt das Riemann-Integral von f mit dem Lebesgue-Integral überein.*

A.11 Radon-Nikodym-Ableitung

Als Abschluss führen den Begriff der Radon-Nikodym-Ableitung ein, der beim Beweis des Neyman-Pearson-Lemmas eine Rolle spielt: Seien P, Q zwei Wahrscheinlichkeitsmasse auf dem Messraum $(\Omega, \mathcal{F})$, dann heisst Q absolut stetig bezüglich P falls für alle P-Nullmengen $A \in \mathcal{F}$ gilt $Q(A) = 0$. In Zeichen schreiben wir $Q \ll P$.

Wir können folgenden Satz von Radon-Nikodym formulieren:
Satz A.14. *Sei $Q \ll P$, dann existiert eine P-fast sicher definierte Funktion f auf Ω, die positive, reelle Werte annimmt, sodass*

$$Q(A) = \mathbb{E}_P\left(1_A f\right)$$

für alle $A \in \mathcal{F}$ gilt.

Man bezeichent die ($\mathbb{P}$-fast sicher eindeutig definierte) Funktion f als Radon-Nikodym-Ableitung von Q nach P und schreibt

$$\frac{dQ}{dP} = f.$$

Beispiel A.1. *Seien μ_0 und μ_1 zwei diskrete Verteilungen auf dem Messraum $(\Omega, 2^\Omega)$, dann gilt $\mu_0 \ll (\mu_0 + \mu_1)/2$ and $\mu_1 \ll (\mu_0 + \mu_1)/2$. Die Ableitungen sind, ausserhalb von $\mu_0 + \mu_1$-Nullmengen durch die Quotienten $\frac{2\mu_i}{\mu_0+\mu_1}$ gegeben. Wenn man das ganze bezüglich der dominierenden Masse $\mu_0+\mu_1$, die keine Wahrscheinlichkeitsmasse mehr sind, rechnet, ergeben sich die Ausdrücke beim Neyman-Pearson-Lemma.*

Anhang B

$\mathcal{L}^p$- und L^p- Räume

Im Folgenden sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Räume von Zufallsvariablen, deren p-te Potenz integrierbar ist, heissen $\mathcal{L}^p$-Räume und sind folgendermassen definiert.

Definition B.1. *Für Zufallsvariablen $X : \Omega \to \overline{\mathbb{R}}$ definiert man*

$$\|X\|_p := \left(\int_\Omega |X(\omega)|^p d\mathbb{P}(\omega) \right)^{\frac{1}{p}} = \mathbb{E}\left(|X|^p\right)^{\frac{1}{p}}, \quad \text{falls } p \in [1, \infty)$$

und für $p = \infty$

$$\|X\|_\infty := \inf\{K \geq 0 : \mathbb{P}(\{|X| > K\}) = 0\}.$$

Für jedes $p \in [1, \infty]$ ist $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ der Vektorraum der Zufallsvariablen, für die die obigen Ausdrücke endlich sind, d.h.

$$\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P}) := \{X : \Omega \to \overline{\mathbb{R}} \text{ ist messbar und } \|X\|_p < \infty\}.$$

Wie wir später sehen werden (Satz B.4), impliziert diese Definition, dass $\|\cdot\|$ eine Halbnorm ist, das heisst alle Eigenschaften der Norm sind erfüllt ausser $\|X\|_p = 0 \Rightarrow X = 0$. Zunächst aber widmen wir uns der Vollständigkeit der $\mathcal{L}^p$-Räume.

Definition B.2. *Seien $p \in [1, \infty]$ und $X, X_1, X_2, \ldots \in \mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$. Falls $\|X_n - X\|_p \overset{n\to\infty}{\longrightarrow} 0$, so sagen wir $(X_n)_{n\in\mathbb{N}}$ konvergiert im p-ten Mittel gegen X und schreiben $X_n \overset{L^p}{\to} X$.*

Satz B.1. *Seien $p \in [1, \infty]$ und $X_1, X_2, \ldots \in \mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$. Dann ist äquivalent:*

(i) Es gibt ein $X \in \mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ mit $X_n \overset{L^p}{\to} X$.

(ii) $(X_n)_{n\in\mathbb{N}}$ ist eine Cauchy-Folge in $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$.

B.1 Ungleichungen

Aus der Jensen'schen Ungleichung (3.6) folgt

$$\|X\|_p \leq \|X\|_q \quad (p \leq q).$$

Zum Beweis beachte man, dass $g(x) = |x|^r$ konvex ist für $r \geq 1$, also $\mathbb{E}\left(|X|^q\right) = \mathbb{E}\left(|X|^{p\frac{q}{p}}\right) \geq \left(\mathbb{E}\left(|X|^p\right)\right)^{q/p}$.

Wir kommen nun zu den beiden weiteren wichtigen Ungleichungen, der Hölderschen und der Minkowskischen Ungleichung. Für deren Beweise benötigen wir folgendes Lemma.

Lemma B.1 (Youngsche Ungleichung). *Für $(p,q) \in (1,\infty)$ mit $\frac{1}{p}+\frac{1}{q}=1$ und für $a,b \in [0,\infty)$ gilt*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Beweis Für $a=0$ oder $b=0$, stimmt die Ungleichung klarerweise. Deshalb nehmen wir an, dass $a>0$ und $b>0$. Wir können daher $a = e^{\frac{x}{p}}$ und $b = e^{\frac{y}{q}}$ setzen und die Youngsche Ungleichung lässt sich als

$$e^{\frac{x}{p}} e^{\frac{y}{q}} = ab \leq \frac{a^p}{p} + \frac{b^q}{q} = \frac{e^x}{p} + \frac{e^y}{q}$$

schreiben. Setzt man $\lambda = \frac{1}{p}$ und $1-\lambda = \frac{1}{q}$, so bekommt man

$$e^{\lambda x + (1-\lambda)y} \leq \lambda e^x + (1-\lambda)e^y,$$

was auf Grund der Konvexität der Exponentialfunktion richtig ist. □

Satz B.2 (Höldersche Ungleichung). *Seien $p,q \in [1,\infty]$ mit $\frac{1}{p}+\frac{1}{q}=1$ und $X \in \mathcal{L}^p(\Omega,\mathcal{A},\mathbb{P})$, $Y \in \mathcal{L}^q(\Omega,\mathcal{A},\mathbb{P})$. Dann gilt $XY \in \mathcal{L}^1(\Omega,\mathcal{A},\mathbb{P})$ und*

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

Beweis Die Fälle $p=1$ und $p=\infty$ sowie $X=0$ $\mathbb{P}$-f.s. oder $Y=0$ $\mathbb{P}$-f.s. sind klar. Sei nun also $p \in (1,\infty)$ und X und Y nicht fast sicher 0. Indem wir zu $X/\|X\|_p$ und $Y/\|Y\|_q$ übergehen, können wir $\|X\|_p = \|Y\|_q = 1$ annehmen. Nach Lemma B.1 gilt

$$\|XY\|_1 = \mathbb{E}\left(|X||Y|\right) \leq \frac{1}{p}\mathbb{E}\left(|X|^p\right) + \frac{1}{q}\mathbb{E}\left(|Y|^q\right) = \frac{1}{p}+\frac{1}{q} = 1 = \|X\|_p\|Y\|_q.$$

□

Satz B.3 (Minkowskische Ungleichung). *Für $p,q \in [1,\infty]$ und $X,Y \in \mathcal{L}^p(\Omega,\mathcal{A},\mathbb{P})$, gilt*

$$\|X+Y\|_p \leq \|X\|_p + \|Y\|_p. \tag{B.1}$$

Beweis Der Fall $p=\infty$ ist klar. Sei also $p \in [1,\infty)$. Die linke Seite von (B.1) wird nicht kleiner, wenn wir X und Y durch $|X|$ und $|Y|$ ersetzen. Wir können also o.B.d.A annehmen, dass $X \geq 0$, $Y \geq 0$ und $(X+Y)^p \geq 0$ gilt. Mit Hilfe der Hölderschen Ungleichung angewandt auf $X(X+Y)^{p-1}$ und auf $Y(X+Y)^{p-1}$ mit $q = p/(p-1)$ erhalten wir

$$\begin{aligned}
\|X+Y\|_p^p &= \mathbb{E}\left((X+Y)^p\right) = \mathbb{E}\left(X(X+Y)^{p-1}\right) + \mathbb{E}\left(Y(X+Y)^{p-1}\right) \\
&\leq \|X\|_p \|(X+Y)^{p-1}\|_q + \|Y\|_p \|(X+Y)^{p-1}\|_q \\
&= (\|X\|_p + \|Y\|_p)\|X+Y\|_p^{p-1}.
\end{aligned}$$

Division durch $\|X+Y\|_p^{p-1}$ ergibt dann die gewünschte Ungleichung. □

Satz B.4. *Die Abbildung $\|\cdot\|$ ist eine Halbnorm auf $\mathcal{L}^p(\Omega,\mathcal{A},\mathbb{P})$, das heisst es gilt für alle $X,Y \in \mathcal{L}^p(\Omega,\mathcal{A},\mathbb{P})$ und $\alpha \in \mathbb{R}$*

$$\begin{aligned}
\|X\|_p &\geq 0 \text{ für alle } X \text{ und } \|X\|_p = 0, \text{ falls } X = 0 \ \mathbb{P}\text{-f.s.}, \\
\|\alpha X\|_p &= \alpha \|X\|_p, \\
\|X+Y\|_p &\leq \|X\|_p + \|Y\|_p.
\end{aligned}$$

Beweis Eigenschaften 1 und 2 folgen aus den Eigenschaften des Integrals und die dritte Eigenschaft ist die Minkowskische Ungleichung. □

B.2 L^p-Räume

Wir haben gesehen, dass $\|\cdot\|$ auf $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ nur eine Halbnorm ist. Es gilt nämlich

$$\|X - Y\|_p = 0 \Leftrightarrow X = Y \ \mathbb{P}\text{-f.s.}$$

Um aus $\|\cdot\|_p$ eine echte Norm zu machen, definiert man

$$\mathcal{N} = \{X \text{ ist messbar und } X = 0 \ \mathbb{P}\text{-f.s.}\}.$$

Für jedes $p \in [1, \infty]$ ist $\mathcal{N}$ ein Untervektorraum von $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$. Wir können deshalb den Quotientenraum bilden und X und Y als äquivalent ansehen, falls $X = Y$ $\mathbb{P}$-f.s. gilt.
Definition B.3. *Für jedes $p \in [1, \infty]$ definiert man $L^p(\Omega, \mathcal{A}, \mathbb{P})$ als*

$$L^p(\Omega, \mathcal{A}, \mathbb{P}) = \mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})/\mathcal{N} = \{\overline{X} := X + \mathcal{N} \mid X \in \mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})\}.$$

Für $\overline{X} \in L^p(\Omega, \mathcal{A}, \mathbb{P})$ setzen wir $\|\overline{X}\|_p = \|X\|_p$ und $\int \overline{X} d\mathbb{P} = \int X d\mathbb{P}$ mit $X \in \overline{X}$.

Auf $L^p(\Omega, \mathcal{A}, \mathbb{P})$ ist $\|\cdot\|$ nun eine echte Norm. Ausserdem folgt aus Satz B.1, dass die Norm vollständig ist, d.h. jede Cauchyfolge konvergiert. Ein vollständig normierter Raum heisst Banachraum. Daher haben wir folgenden Satz gezeigt.
Satz B.5 (Fischer-Riesz)**.** *Für $p \in [1, \infty]$ ist $(L^p(\Omega, \mathcal{A}, \mathbb{P}), \|\cdot\|_p)$ ein Banachraum.*

B.3 Hilbertraum $L^2(\Omega, \mathcal{A}, \mathbb{P})$

In diesem Abschnitt betrachten wir nur den Fall $p = 2$.
Definition B.4. *Sei V ein reeller Vektorraum und ist $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ ein Skalarprodukt, so heisst $(V, \langle\cdot,\cdot\rangle)$ ein (reeller) Hilbertraum, falls die durch $\|x\| := \sqrt{\langle x, x\rangle}$ definierte Norm vollständig ist, falls also $(V, \|\cdot\|)$ ein Banachraum ist.*
Definition B.5. *Für $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ definieren wir*

$$\langle X, Y\rangle := \int_\Omega X(\omega) Y(\omega) d\mathbb{P}(\omega) = \mathbb{E}(XY).$$

und für $\overline{X}, \overline{Y} \in L^2(\Omega, \mathcal{A}, \mathbb{P})$

$$\langle \overline{X}, \overline{Y}\rangle := \langle X, Y\rangle,$$

wobei $X \in \overline{X}$ und $Y \in \overline{Y}$.

Man beachte, dass diese Definition unabhängig von der Wahl der Repräsentaten X und Y ist.

Als Korollar zu Satz B.5 erhalten wir:
Korollar B.1. *$(L^2(\Omega, \mathcal{A}, \mathbb{P}), \langle\cdot,\cdot\rangle)$ ist ein (reeller) Hilbertraum.*

Wir geben nun einige Resultate für Hilberträume an, die dann selbstverständlich auch für $L^2(\Omega, \mathcal{A}, \mathbb{P})$ gelten. Von nun an bezeichne V einen Hilbertraum mit Norm $\|\cdot\|$ und Skalarprodukt $\langle\cdot,\cdot\rangle$.
Definition B.6. *Zwei Elemente $v, z \in V$ heissen orthogonal, wenn $\langle v, z\rangle = 0$ gilt. Ein Element x ist orthogonal zu einer Menge $W \subset V$, falls $\langle v, z\rangle = 0$ für alle $z \in W$ gilt.*

Satz B.6. *Sei $W \subset V$ eine Teilmenge von V. Dann ist die Menge $W^\perp$*

$$W^\perp := \{v \in V \mid \langle v, z\rangle = 0 \text{ für alle } z \in W\}$$

ein abgeschlossener linearer Unterraum von V und wird als orthogonales Komplement von W bezeichnet.

Wir betrachten nun das wichtige Konzept von Orthogonalprojektionen.
Satz B.7. *Sei $(V, \langle\cdot,\cdot\rangle)$ ein Hilbertraum und $W \subset V$ ein abgeschlossener linearer Unterraum. Dann existiert für jedes $x \in V$ eine eindeutige Darstellung*

$$x = y + v, \tag{B.2}$$

wobei $y \in W$ und $v \in W^\perp$ ist.

Das Element $y \in W$ heisst Orthogonalprojektion von x auf W und ist das eindeutige Element, das den Abstand minimiert, d.h.

$$\|x - y\| = \min\{\|x - z\| \mid z \in W\}.$$

Wir schreiben auch $y = \Pi_W x$, wobei Π_W den Projektionsoperator bezeichnet. Da $x - y = x - \Pi_W x \in W^\perp$ gilt, folgt insbesondere für alle $z \in W$,

$$\langle x - y, z\rangle = 0. \tag{B.3}$$

Ausserdem impliziert die eindeutige Darstellung (B.2) die Linearität des Projektionsoperators

$$\Pi_W(\alpha x_1 + \beta x_2) = \alpha\Pi_W(x_1) + \beta\Pi_W(x_2). \tag{B.4}$$

Anhang C

Bedingte Erwartung

Im Folgenden betrachten wir den Hilbertraum $L^2(\Omega, \mathcal{A}, \mathbb{P})$ aller quadratintegrierbaren Zufallsvariablen mit Skalarprodukt $\langle X, Y \rangle = \mathbb{E}(XY)$. Sei nun $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra von $\mathcal{A}$. Dann ist $L^2(\Omega, \mathcal{G}, \mathbb{P})$ ein abgeschlossener linearer Unterraum von $L^2(\Omega, \mathcal{A}, \mathbb{P})$. Die bedingte Erwartung ist nun folgendermassen definiert.

Definition C.1. *Sei $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ und $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra. Dann nennt man die Orthogonalprojektion von X auf $L^2(\Omega, \mathcal{G}, \mathbb{P})$ die bedingte Erwartung von X gegeben $\mathcal{G}$. Diese wird mit $\mathbb{E}(X|\mathcal{G})$ bezeichnet. Anders ausgedrückt, $\mathbb{E}(X|\mathcal{G})$ ist das eindeutige Element in $L^2(\Omega, \mathcal{G}, \mathbb{P})$, sodass*

$$\mathbb{E}(XZ) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})\, Z) \tag{C.1}$$

für alle $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ gilt (siehe (B.3)).

Bemerkung: Die bedingte Erwartung ist ein Element in $L^2(\Omega, \mathcal{G}, \mathbb{P})$, das heisst eine Äquivalenzklasse von Zufallsvariablen. Deshalb muss jede Aussage, wie zum Beispiel $\mathbb{E}(X|\mathcal{G}) \geq 0$, mit einem impliziten f.s. verstanden werden.

Satz C.1. *Sei $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ und $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra. Dann gilt:*

i) Die Abbildung $X \to \mathbb{E}(X|\mathcal{G})$ ist linear.

ii) Falls $X \geq 0$, dann gilt $\mathbb{E}(X|\mathcal{G}) \geq 0$.

iii) $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.

Beweis

i) Die Linearität von $X \to \mathbb{E}(X|\mathcal{G})$ folgt aus der Definition der bedingten Erwartung als Orthogonalprojektion, die linear ist (siehe (B.4)).

ii) Wir verwenden die definierende Eigenschaft (C.1) der bedingten Erwartung und setzen $Z = 1_{\{\mathbb{E}(X|\mathcal{G})<0\}}$. Da X f.s. nicht negativ ist, gilt $\mathbb{E}(XZ) \geq 0$, aber

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})\, Z) = \mathbb{E}\left(\mathbb{E}(X|\mathcal{G})\, 1_{\{\mathbb{E}(X|\mathcal{G})<0\}}\right) < 0, \tag{C.2}$$

falls $\mathbb{P}(\{\mathbb{E}(X|\mathcal{G}) < 0\}) > 0$. Das ist ein Widerspruch zu (C.1) und daher folgt $\mathbb{P}(\{\mathbb{E}(X|\mathcal{G}) < 0\}) = 0$.

iii) Es genügt, $Z = 1$ in (C.1) zu setzen.

□

Bemerkung: Die Schlüsseleigenschaft der bedingten Erwartung ist (C.1). Wir führen die Orthogonalprojektion im Hilbertraum $L^2(\Omega, \mathcal{A}, \mathbb{P})$ nur deshalb ein, um den Zusammenhang mit dem Problem des Auffindens einer $\mathcal{G}$-messbaren Zufallsvariable mit minimalem L^2-Abstand von X zu verdeutlichen.

Nun möchten wir die bedingte Erwartung wie oben definiert auf Zufallsvariablen in $L^1(\Omega, \mathcal{A}, \mathbb{P})$ ausdehnen. In diesem Fall kann man die Technik der Hilbertraum Projektion nicht mehr anwenden. Wir weiten die bedingte Erwartung zuerst auf nicht negative Zufallsvariablen aus und bezeichnen den Raum aller Äquivalenzklassen nicht negativer Zufallsvariablen mit $L^+(\Omega, \mathcal{A}, \mathbb{P})$. Die Zufallsvariablen dürfen auch den Wert ∞ annehmen.

Lemma C.1. *Sei $X \in L^+(\Omega, \mathcal{A}, \mathbb{P})$ und $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra. Dann existiert ein eindeutiges Element $\mathbb{E}(X|\mathcal{G}) \in L^+(\Omega, \mathcal{G}, \mathbb{P})$, sodass*

$$\mathbb{E}(XZ) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})\, Z) \tag{C.3}$$

für alle $Z \in L^+(\Omega, \mathcal{G}, \mathbb{P})$. Diese bedingte Erwartung stimmt für $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ mit der bedingten Erwartung von Definition C.1 überein und erfüllt die Eigenschaften von Satz C.1.

Beweis Falls $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ und nicht negativ ist, definieren wir $\mathbb{E}(X|\mathcal{G})$ wie in Definition C.1. Für $Z \in L^+(\Omega, \mathcal{A}, \mathbb{P})$, ist $Z_n = \min(Z, n)$ beschränkt und deshalb insbesondere quadrat-integrierbar. Monotone Konvergenz (zwei mal angewandt) und (C.1) liefert daher

$$\mathbb{E}(XZ) = \lim_{n\to\infty} \mathbb{E}(XZ_n) = \lim_{n\to\infty} \mathbb{E}(\mathbb{E}(X|\mathcal{G})\, Z_n) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})\, Z),$$

womit (C.3) für nicht negative $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ und alle $Z \in L^+(\Omega, \mathcal{G}, \mathbb{P})$ gilt. Sei nun $X \in L^+(\Omega, \mathcal{A}, \mathbb{P})$, dann gilt wieder $X_m = \min(X, m) \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. Da $\mathbb{E}(X_m|\mathcal{G})$ auf Grund von Satz C.1 (ii) eine nicht fallende Folge ist, können wir

$$\mathbb{E}(X|\mathcal{G}) := \lim_{m\to\infty} \mathbb{E}(X_m|\mathcal{G}) \tag{C.4}$$

setzen (wobei erlaubt ist, dass der Limes unendlich ist). Mit monotoner Konvergenz (zweimal angewandt) und Satz C.1 (ii) folgt wieder

$$\begin{aligned}\mathbb{E}(XZ) &= \lim_{m\to\infty} \mathbb{E}(X_m Z) = \lim_{m\to\infty} \mathbb{E}(\mathbb{E}(X_m|\mathcal{G})\, Z) \\ &= \mathbb{E}\left(\lim_{m\to\infty} \mathbb{E}(X_m|\mathcal{G})\, Z\right) = \mathbb{E}(\mathbb{E}(X|\mathcal{G}))\end{aligned}$$

und somit ist (C.3) erfüllt. Mit der Definition der bedingten Erwartung in (C.4), ist klar, dass die Eigenschaften von Satz C.1 erfüllt sind.

Man muss also nur noch die Eindeutigkeit zeigen. Wir nehmen an, es gäbe zwei Elemente $U, V \in L^+(\Omega, \mathcal{A}, \mathbb{P})$, die beide Gleichung (C.3) erfüllen. Sei nun $\Lambda_n = \{U < V \leq n\}$. Auf Grund der $\mathcal{G}$-Messbarkeit von U, V ist $\Lambda_n \in \mathcal{G}$ und $1_{\Lambda_n} \in L^+(\Omega, \mathcal{G}, \mathbb{P})$. Wegen unserer Annahme und (C.3) gilt für $Z = 1_{\Lambda_n}$

$$\mathbb{E}(X1_{\Lambda_n}) = \mathbb{E}(U1_{\Lambda_n}) = \mathbb{E}(V1_{\Lambda_n}).$$

Falls $\mathbb{P}(\Lambda_n) > 0$, muss aber $\mathbb{E}(U1_{\Lambda_n}) < \mathbb{E}(V1_{\Lambda_n})$ gelten, ein Widerspruch. Deshalb folgt $\mathbb{P}(\Lambda_n) = 0$ für alle n und da $\{U < V\} = \cup_{n\geq 1}\Lambda_n$, erhalten wir $\mathbb{P}(U < V) = 0$ und analog $\mathbb{P}(V < U) = 0$ und somit Eindeutigkeit. □

Satz C.2. *Sei $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ und $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra. Dann existiert ein eindeutiges Element $\mathbb{E}(X|\mathcal{G}) \in L^1(\Omega, \mathcal{G}, \mathbb{P})$, sodass*

$$\mathbb{E}(XZ) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})\, Z) \tag{C.5}$$

für alle beschränkten $\mathcal{G}$-messbaren Zufallsvariablen Z. Diese bedingte Erwartung stimmt für $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ mit der bedingten Erwartung von Definition C.1 überein und erfüllt die Eigenschaften von Satz C.1.

Beweis Wir zerlegen X in Positiv- und Negativteil, d.h.

$$X = X^+ - X^-,$$

mit $X^+ = \max(X, 0)$ und $X^- = -\min(X, 0)$, die beide wieder in $L^1(\Omega, \mathcal{A}, \mathbb{P})$ liegen. Definiert man nun die bedingte Erwartung durch

$$\mathbb{E}(X|\mathcal{G}) := \mathbb{E}(X^+|\mathcal{G}) - \mathbb{E}(X^-|\mathcal{G}), \tag{C.6}$$

dann erfüllt $\mathbb{E}(X|\mathcal{G})$ wegen Lemma C.1 Eigenschaft (C.5). Insbesondere liegt $\mathbb{E}(X^+|\mathcal{G})$ und $\mathbb{E}(X^-|\mathcal{G})$ in $L^1(\Omega, \mathcal{G}, \mathbb{P})$, denn auf Grund von Lemma C.1 gilt

$$\mathbb{E}(X^+) = \mathbb{E}(\mathbb{E}(X^+|\mathcal{G})) \quad \text{und} \quad \mathbb{E}(X^-) = \mathbb{E}(\mathbb{E}(X^-|\mathcal{G})).$$

Da die linken Seiten endlich sind (wegen $X^+, X^- \in L^1(\Omega, \mathcal{A}, \mathbb{P})$), folgt $\mathbb{E}(X^+|\mathcal{G}), \mathbb{E}(X^-|\mathcal{G}) \in L^1(\Omega, \mathcal{A}, \mathbb{P})$.

Man muss also nur noch die Eindeutigkeit zeigen. Wir nehmen wieder an, es gäbe zwei Elemente $U, V \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, die beide Gleichung (C.5) erfüllen. Sei nun $\Lambda = \{U < V\}$. Auf Grund der $\mathcal{G}$-Messbarkeit von U, V gilt $\Lambda \in \mathcal{G}$ und 1_Λ ist daher eine beschränkte $\mathcal{G}$-messbare Funktion. Wegen unserer Annahme und (C.5) gilt für $Z = 1_\Lambda$

$$\mathbb{E}(X1_\Lambda) = \mathbb{E}(U1_\Lambda) = \mathbb{E}(V1_\Lambda).$$

Falls $\mathbb{P}(\Lambda) > 0$, dann muss aber $\mathbb{E}(U1_\Lambda) < \mathbb{E}(V1_\Lambda)$ gelten, ein Widerspruch. Deshalb folgt $\mathbb{P}(\Lambda) = 0$ und analog $\mathbb{P}(V < U) = 0$. Die restlichen Aussagen folgen aus (C.6), Lemma C.1 und Satz C.1. □

C.1 Eigenschaften

Wir zählen nun einige Eigenschaften der bedingten Erwartung auf.

Satz C.3. *Sei $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ und $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra.*

i) Sei $\mathcal{H} \subset \mathcal{G}$ eine Teil-σ-Algebra von $\mathcal{G}$. Dann gilt $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})$.

ii) Falls X unabhängig von $\mathcal{G}$ ist, gilt $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$.

Beweis

i) Wegen (C.5) gilt für alle beschränkten $\mathcal{H}$-messbaren Zufallsvariablen Z

$$\mathbb{E}(\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})\, Z) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})\, Z)$$

und für alle beschränkten $\mathcal{G}$-messbaren Zufallsvariablen Z

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})\, Z) = \mathbb{E}(XZ).$$

Da $\mathcal{H} \subset \mathcal{G}$, folgt $\mathbb{E}(\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})\, Z) = \mathbb{E}(XZ)$ für alle beschränkten $\mathcal{H}$- messbaren Zufallsvariablen Z und daher $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})$.

ii) Wir zeigen, dass für alle beschränkten $\mathcal{G}$-messbaren Zufallsvariablen Z

$$\mathbb{E}(XZ) = \mathbb{E}(\mathbb{E}(X)Z)$$

gilt. Nun ist aber $\mathbb{E}(\mathbb{E}(X)Z) = \mathbb{E}(X)\mathbb{E}(Z)$ und $\mathbb{E}(X)\mathbb{E}(Z) = \mathbb{E}(XZ)$, weil X von Z unabhängig ist. Daraus folgt die Aussage.

□

Satz C.4. *Sei $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra, X eine beschränkte ($\mathcal{A}$-messbare) Zufallsvariable und Y eine beschränkte $\mathcal{G}$-messbare Zufallsvariable. Dann gilt*

$$\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G}).$$

Beweis Für alle beschränkten $\mathcal{G}$-messbaren Zufallsvariablen Z muss man zeigen, dass

$$\mathbb{E}(Y\mathbb{E}(X|\mathcal{G})Z) = \mathbb{E}(XYZ).$$

Dies ist äquivalent zu $\mathbb{E}((X - \mathbb{E}(X|\mathcal{G}))YZ) = 0$. Da YZ eine beschränkte $\mathcal{G}$-messbare Zufallsvariable ist, folgt die Behauptung. □

C.2 Jensensche und Höldersche Ungleichung für die bedingte Erwartung

Die folgenden Aussagen gelten alle $\mathbb{P}$-f.s.

Satz C.5. *Sei $g : \mathbb{R} \to \mathbb{R}$ konvex, X und $g(X) \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ und sei $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra. Dann gilt*

$$g(\mathbb{E}(X|\mathcal{G})) \leq \mathbb{E}(g(X)|\mathcal{G}).$$

Beweis Für eine konvexe Funktion von $g : \mathbb{R} \to \mathbb{R}$ gibt es für jedes x eine Stützgerade $\ell(y) = ay + b$ sodass

$$\ell(y) = ay + b \leq g(y),\ l(x) = g(x)$$

gilt. Auf Grund von Satz C.1 und Satz C.2 haben wir für a, b mit $ay + b \leq g(y)$, für alle y, dass

$$\begin{aligned}\mathbb{E}(aX + b|\mathcal{G}) &= a\mathbb{E}(X|\mathcal{G}) + b,\\ \mathbb{E}(aX + b|\mathcal{G}) &\leq \mathbb{E}(g(X)|\mathcal{G})\end{aligned}$$

gilt. Deshalb folgt für jede Stützgerade von g

$$\ell((\mathbb{E}(X|\mathcal{G}))) \leq \mathbb{E}(g(X)|\mathcal{G}).$$

Durch Wahl von ω-abhängigen Stützgeraden folgt die Behauptung. □

Satz C.6. *Seien $1 < p, q < \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$ und $\mathcal{G} \subset \mathcal{A}$ eine Teil-σ-Algebra. Dann gilt für alle Zufallsvariablen $X, Y : \Omega \to \overline{\mathbb{R}}$*

$$\mathbb{E}(|XY||\mathcal{G}) \leq \mathbb{E}(|X|^p|\mathcal{G})^{\frac{1}{p}}\,\mathbb{E}(|Y|^q|\mathcal{G})^{\frac{1}{q}}.$$

Beweis Wir definieren die Mengen $A := \{\mathbb{E}(|X|^p|\mathcal{G}) = 0\}$ und $B := \{\mathbb{E}(|Y|^q|\mathcal{G}) = 0\}$. Da $A \in \mathcal{G}$ ist, gilt

$$\mathbb{E}(|X|^p 1_A) = \mathbb{E}(1_A \mathbb{E}(|X|^p|\mathcal{G})) = 0.$$

Es folgt daher, dass $|X| = 0$ $\mathbb{P}$-f.s. auf A. Dasselbe gilt für Y, d.h. $|Y| = 0$ $\mathbb{P}$-f.s. auf B. Daher haben wir auf $A \cup B$

$$\mathbb{E}(|XY||\mathcal{G})(\omega) = 0 \quad \mathbb{P}\text{-f.s.}$$

und die bedingte Höldersche Ungleichung gilt auf $A \cup B$. Auf der Menge

$$\{\mathbb{E}(|X|^p|\mathcal{G}) = \infty \text{ und } \mathbb{E}(|Y|^q|\mathcal{G}) = 0\} \cup \{\mathbb{E}(|X|^p|\mathcal{G}) > 0 \text{ und } \mathbb{E}(|Y|^q|\mathcal{G}) = \infty\}$$

ist die rechte Seite unendlich und die bedingte Höldersche Ungleichung gilt ebenfalls. Man muss die Ungleichung also nur noch auf der Menge

$$C := \{0 < \mathbb{E}(|X|^p|\mathcal{G}) < \infty \text{ und } 0 < \mathbb{E}(|Y|^q|\mathcal{G}) < \infty\}$$

zeigen. Nun kann man durch $\mathbb{E}(|X|^p|\mathcal{G})^{\frac{1}{p}}$ und $\mathbb{E}(|Y|^q|\mathcal{G})^{\frac{1}{q}}$ dividieren. Daher muss man zeigen, dass

$$\frac{\mathbb{E}(|XY||\mathcal{G})}{\mathbb{E}(|X|^p|\mathcal{G})^{\frac{1}{p}}\,\mathbb{E}(|Y|^q|\mathcal{G})^{\frac{1}{q}}} \leq 1 \tag{C.7}$$

auf C gilt. Sei nun $G \in \mathcal{G}$, $G \subset C$. Dann,

$$\begin{aligned}
&\mathbb{E}\left(\frac{\mathbb{E}(|XY||\mathcal{G})}{\mathbb{E}(|X|^p|\mathcal{G})^{\frac{1}{p}}\,\mathbb{E}(|Y|^q|\mathcal{G})^{\frac{1}{q}}} 1_G\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\frac{|XY|}{\mathbb{E}(|X|^p|\mathcal{G})^{\frac{1}{p}}\,\mathbb{E}(|Y|^q|\mathcal{G})^{\frac{1}{q}}}\middle|\mathcal{G}\right) 1_G\right) \\
&= \mathbb{E}\left(\frac{|X|}{\mathbb{E}(|X|^p|\mathcal{G})^{\frac{1}{p}}} 1_G \frac{|Y|}{\mathbb{E}(|Y|^q|\mathcal{G})^{\frac{1}{q}}} 1_G\right) \\
&\overset{\text{Hölder, Satz B.2}}{\leq} \left(\mathbb{E}\left(\frac{|X|^p}{\mathbb{E}(|X|^p|\mathcal{G})} 1_G\right)\right)^{1/p} \left(\mathbb{E}\left(\frac{|Y|^q}{\mathbb{E}(|Y|^q|\mathcal{G})} 1_G\right)\right)^{1/q} \\
&= \left(\mathbb{E}\Big(\underbrace{\frac{\mathbb{E}(|X|^p|\mathcal{G})}{\mathbb{E}(|X|^p|\mathcal{G})}}_{=1 \text{ f.s. auf } G} 1_G\Big)\right)^{1/p} \left(\mathbb{E}\Big(\underbrace{\frac{\mathbb{E}(|Y|^q|\mathcal{G})}{\mathbb{E}(|Y|^q|\mathcal{G})}}_{=1 \text{ f.s. auf } G} 1_G\Big)\right)^{1/q} \\
&\overset{\frac{1}{p}+\frac{1}{q}=1}{=} \mathbb{E}(1_G).
\end{aligned}$$

□