# Stochastik II

Skript zur Vorlesung im SoSe 2025

Dr. Johannes Brutsche

2025

# 1 Grundbegriffe der Statistik

## 1.1 Einführung

Die Vorlesung *Stochastik II* ist eine Einführung in die *mathematische Statistik*. Ziel der Statistik im Allgemeinen ist es, Daten zu ordnen und daraus Schlüsse zu ziehen, was zu folgender Unterteilung führt:

- *Deskriptive Statistik*: Diese dient dazu, die Daten zu beschreiben, aufzubereiten und zusammenzufassen. Dazu gehören insbesondere graphische Darstellungen und Kennzahlen aller Art.
- *Induktive Statistik* (auch *schließende Statistik*): Hier werden mithilfe stochastischer Modelle aus beobachteten Daten einer Stichprobe Aussagen über die Grundgesamtheit getroffen bzw. über die Annahmen, die dem stochastischen Modell zugrunde liegen.

Diese Vorlesung behandelt ausschließlich Themen der induktiven Statistik. In der vorangegangenen Veranstaltung Stochastik I war das Setting abstrakt gesprochen das Folgende: Gegeben war ein stochastisches Modell in Form eines Wahrscheinlichkeitsraums $(\Omega, \mathcal{A}, \mathbb{P})$ bzw. Zufallsvariablen auf einem solchen samt ihrer Verteilung. Darauf aufbauend wurden Aussagen über die Wahrscheinlichkeit bestimmter Ausgänge getroffen, z.B. der Wahrscheinlichkeit beim dreifachen Würfelwurf eines fairen Würfels mindestens eine Sechs zu würfeln. Die Schlussrichtung der induktiven Statistik ist in gewisser Hinsicht umgekehrt: Die grundlegende Annahme ist hier, dass beobachtete Daten die Realisierungen von Zufallsvariablen sind, über deren Verteilung eine Aussage getroffen werden soll. Eine typische Fragestellung wäre dann etwa, ob davon auszugehen ist, dass ein Würfel fair ist, wenn bei zehnfachem Würfelwurf keine einzige Sechs auftritt (vgl. Abbildung 1). Man beachte, dass dies eine Frage über das Wahrscheinlichkeitsmaß ist, welches der Datengenerierung zugrunde liegt.

Im folgenden Teil der Einleitung werden wir anhand des Beispiels eines $n$-fachen Münzwurfs einige Fragestellungen und Konzepte der Statistik veranschaulichen und sehen, inwiefern die Resultate aus Stochastik I für die Analyse hilfreich und grundlegend sind. Im Anschluss betten wir die grundlegenden Ideen in einen abstrakten Rahmen samt mathematisch präziser Begriffsbildung ein.

### Ein einführendes Beispiel: Erfolgswahrscheinlichkeit beim Münzwurf

Wir betrachten das Modell des $n$-fachen unabhängigen Münzwurfs, wobei die Wahrscheinlichkeit $p$ für das Ereignis 'Kopf' noch unbekannt sei. Wir gehen davon aus, dass wir eine Realisierung $X = (X_1, \dots, X_n)$ von $n$ Münzwürfen beobachten können, wobei $X_1, \dots, X_n$ unabhängig identisch verteilt sind mit

$$X_i = \begin{cases} 1, & \text{falls der } i\text{-te Wurf 'Kopf' zeigt,} \\ 0, & \text{sonst,} \end{cases}$$

und

$$\mathbb{P}(X_i = 1) = p$$

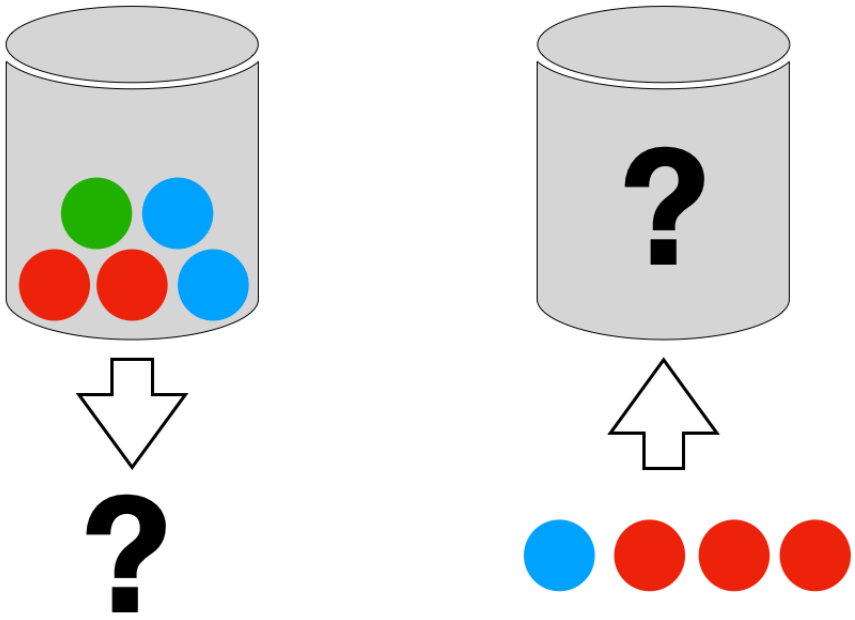


Abbildung 1: Darstellung der Konzepte der Wahrscheinlichkeitsrechnung (Stochastik I) und Statistik (Stochastik II). Das Modell wird hier durch eine Urne symbolisiert, der Pfeil veranschaulicht die Analyserichtung.

für $i = 1, \ldots, n$. Der Wert von $S_n = X_1 + \cdots + X_n$ gibt die Anzahl der Kopf-Würfe in den $n$ Versuchen an und es gilt $S_n \sim \mathcal{B}(n, p)$, d.h. $S_n$ ist binomialverteilt mit den Parametern $n$ und $p$. Man beachte, dass hier $n$ bereits feststeht (da wir wissen, wie häufig wir die Münze geworfen haben), nicht jedoch $p$. In dieser Situation gibt es nun zwei verschiedene Fragestellungen:

- *Schätzung.* Hierbei geht es darum, die Erfolgswahrscheinlichkeit $p$ zu schätzen. Dabei können wir entweder aus den Daten einen möglichen Wert $p(X_1, \ldots, X_n)$ ableiten (*Punktschätzer*) oder ein Intervall $[a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$ bestimmen, in welchem der wahre Parameter $p$ mit hoher Wahrscheinlichkeit liegt (*Konfidenzbereich, Intervallschätzer*).
- *Testproblem.* Anhand der Beobachtungen soll entschieden werden, ob z.B. die Hypothese 'Für den wahren Parameter $p$ gilt $p = 1/2$' plausibel ist. Anders ausgedrückt fragen wir uns, ob die beobachteten Daten diese Aussage stützen oder zu unwahrscheinlich dafür sind.

Für diese verschiedenen Fragestellungen wollen wir nun Lösungen skizzieren:

**Erfolgswahrscheinlichkeit $p$ schätzen.** Ein naheliegender Schätzwert für die Wahrscheinlichkeit $p$, dass die Münze 'Kopf' zeigt, ist der Anteil der 'Kopf'-Würfe an allen $n$ Würfen. Dies führt auf den Schätzer

$$\hat{p}_n := \hat{p}_n(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Wir folgen hier der Konvention, einen Schätzer für einen unbekannten Parameter $\theta$, welcher auf $n$ Beobachtungen basiert, mit $\hat{\theta}_n$ zu bezeichnen. Man beachte, dass $\hat{p}_n$ eine Funktion der Daten $X_1, \dots, X_n$ ist - diese wiederum sind der Modellannahme nach zufällig, sodass auch $\hat{p}_n$ selbst eine Zufallsvariable ist. Um zu sehen, inwiefern der Schätzer $\hat{p}_n$ 'gut' ist, betrachten wir zwei Kriterien. Dabei bezeichnen wir mit $\mathbb{P}_p$ und $\mathbb{E}_p$ die Wahrscheinlichkeit bzw. den Erwartungswert unter der Annahme, dass $p$ der wahre Parameter ist, d.h. dass jedes $X_i$ Bernoulli-verteilt zum Parameter $p$ ist, sowie $X_1, \dots, X_n$ unabhängig sind. Wir erhalten dann:

(i) Unter Ausnutzung der Linearität des Erwartungswertes gilt

$$\mathbb{E}_p\left[\hat{p}_n\right] = \frac{1}{n}\mathbb{E}_p\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_p[X_i] = p. \tag{1}$$

Das bedeutet, dass der Schätzer $\hat{p}_n$ im Mittel das richtige liefert, mit anderen Worten: Wiederholen wir das Prozedere der $n$-fachen Beobachtung des Münzwurfs viele Male, so dürfen wir davon ausgehen, dass unser Verfahren im Schnitt dem wahren Parameter nahekommt. Angenommen, wir haben $N$ unabhängige Stichproben $(X_1^j, \dots, X_n^j)$, $j = 1, \dots, N$, und basierend auf jeder davon errechnen wir einen Schätzer $\hat{p}_n^j$, dann gilt nach dem schwachen Gesetz der großen Zahlen (man prüft leicht, dass $\hat{p}_n^1$ eine von $N$ unabhängige endliche Varianz hat), dass

$$\frac{1}{N}\sum_{j=1}^{N}\hat{p}_n^j \overset{N\to\infty}{\longrightarrow}_{\mathbb{P}_p} \mathbb{E}_p\left[\hat{p}_n^1\right] = p.$$

Die Eigenschaft 1 nennen wir auch *Erwartungstreue*.

(ii) Eine zweite wünschenswerte Eigenschaft ist, dass der Schätzer mit steigender Beobachtungszahl (d.h. besserer Datenlage) auch bessere Ergebnisse liefert. Nach dem schwachen Gesetz der großen Zahlen gilt

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow_{\mathbb{P}_p} \mathbb{E}_p\left[X_1\right] = p.$$

Anders ausgedrückt gilt für alle $\epsilon > 0$, dass

$$\mathbb{P}_p\left(|\hat{p}_n - p| > \epsilon\right) \longrightarrow 0,$$

d.h. $\hat{p}_n$ liegt mit immer größerer Wahrscheinlichkeit nahe am wahren Parameter $p$. Wir nenne die Eigenschaft $\hat{p}_n \longrightarrow_{\mathbb{P}_p} p$ *Konsistenz* des Schätzers.

**Konfidenzintervall für $p$.** Die bisherigen Überlegungen zum Punktschätzer $\hat{p}_n$ sagen noch nichts über seine Verteilung (als Zufallsvariable) aus. Nehmen wir an, dass wir drei Experimente durchführen, wobei wir in einem innerhalb von zehn Würfen sechsmal Kopf werfen, dann innerhalb von 100 Würfen 60-mal und schließlich innerhalb von 1000 Würfen ganze 600-mal. In jedem dieser Fälle würden wir mit unserem Punktschätzer $\hat{p}_n$ die Erfolgswahrscheinlichkeit auf $p = 6/10 = 60/100 = 600/1000 = 0.6$ schätzen. Es ist jedoch klar, dass wir dem Schätzwert im Experiment mit den meisten Würfen die höchste Bedeutung zumessen und davon ausgehen dürfen, dass die Schätzung dort am nächsten am wahren Wert liegt. Diese 'Sicherheit', dass die Schätzung nahe am wahren Parameter

liegt, wollen wir mit einem sogenannten *Intervallschätzer* oder *Konfidenzbereich* präzise beschreiben. Da es sich um eine Verteilungseigenschaft des Schätzers handelt, beruht die Überlegung diesmal auf dem zentralen Grenzwertsatz: Für eine standardnormalverteilte Zufallsvariable $Z \sim \mathcal{N}(0,1)$ gilt

$$\frac{n\hat{p}_n - np}{\sqrt{np(1-p)}} = \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\mathrm{Var}(X_1)}} \xrightarrow{\mathcal{D}} Z.$$

Da $\hat{p}_n \to_{\mathbb{P}_p} p$, folgt (Übungsblatt 1, Aufgabe 2)

$$\frac{1}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \longrightarrow_{\mathbb{P}_p} \frac{1}{\sqrt{p(1-p)}},$$

sodass nach dem Lemma von Slutzky (Proposition $I.5.4$)

$$\frac{n\hat{p}_n - np}{\sqrt{n\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{\mathcal{D}} Z.$$

Damit erhalten wir (die Notation $\approx$ bedeute 'ungefähr', ohne Anspruch auf mathematische Exaktheit), dass für großes $n$

$$\begin{aligned}
0.95 &\approx \mathbb{P}\left(-1.96 \leq Z \leq 1.96\right) \\
&\approx \mathbb{P}_p\left(-1.96 \leq \frac{n\hat{p}_n - np}{\sqrt{n\hat{p}_n(1-\hat{p}_n)}} \leq 1.96\right) \\
&= \mathbb{P}_p\left(\hat{p}_n - 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq p \leq \hat{p}_n + 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right).
\end{aligned}$$

Wir haben also gezeigt, dass der wahre Wert $p$ ungefähr mit Wahrscheinlichkeit 0.95 im Intervall

$$I_n = \left[\hat{p}_n - 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

liegt und nennen dieses Intervall daher ein (approximatives) 95%-Konfidenzintervall. Insbesondere sehen wir, dass dieses Intervall mit steigendem $n$ immer kleiner wird. Für das eingangs erwähnte Beispiel erhalten wir (auf zwei Nachkommastellen gerundet)

$$I_{10} = [0.30, 0.90] \qquad I_{100} = [0.50, 0.70], \qquad I_{1000} = [0.57, 0.63].$$

**Erfolgswahrscheinlichkeit testen.** Nehmen wir nun an, jemand stellt die Behauptung auf, die geworfene Münze sei fair, d.h. $p = 1/2$. Wir wollen nun anhand der Daten entscheiden, ob wir dieser Hypothese zustimmen können oder nicht. Dabei können wir grundsätzlich zwei Arten von Fehlern machen:

- Die Hypothese verwerfen, obwohl sie richtig ist.
- Die Hypothese nicht verwerfen, obwohl sie falsch ist.

Wir gehen später noch detaillierter auf diese beiden Fehler ein. Zunächst wollen wir hier annehmen, dass wir der Person, welche die Fairness der Münze behauptet, nicht ohne guten Grund widersprechen wollen. Das bedeutet, dass wir die Hypothese nicht verwerfen wollen,

falls sie in Wirklichkeit stimmt, d.h. unser Entscheidungsverfahren soll so konzipiert sein, dass die Wahrscheinlichkeit

$$\mathbb{P}_{1/2}\,(\text{Hypothese verwerfen})$$

klein ist, sagen wir z.B. kleiner als 5%. Man beachte, dass wir also etwas an das Verfahren unter der Annahme $p = 1/2$ fordern. Die Idee ist nun, die Hypothese $p = 1/2$ zu verwerfen, falls die Zahl der beobachteten 'Kopf'-Würfe deutlich kleiner oder größer ist, als sie unter $\mathbb{P}_{1/2}$ sein sollte. Eine Möglichkeit wäre es, dies durch die Forderung

$$\mathbb{P}_{1/2}\,(\text{Zahl der beobachteten 'Kopf'-Würfe}) \leq 0.05.$$

mathematisch zu präzisieren. Wir betrachten hier wie schon für das Konfidenzintervall einen approximativen Ansatz: Sei $Z \sim \mathcal{N}(0,1)$, dann gilt nach dem zentralen Grenzwertsatz für $\kappa > 0$,

$$\begin{aligned}\mathbb{P}_{1/2}\,(|S_n - n/2| \geq \kappa) &= 1 - \mathbb{P}_{1/2}\left(-\frac{\kappa}{\sqrt{n/4}} < \frac{S_n - n/2}{\sqrt{n/4}} < \frac{\kappa}{\sqrt{n/4}}\right)\\ &\approx 1 - \mathbb{P}_{1/2}\left(-\frac{\kappa}{\sqrt{n/4}} < Z < \frac{\kappa}{\sqrt{n/4}}\right).\end{aligned}$$

Betrachten wir nun das Beispiel des 100-fachen Münzwurfs mit 55 Beobachtungen 'Kopf'. Unter $p = 1/2$ erwarten wir hier nur 50 Würfe mit 'Kopf', d.h. wir haben $\kappa = 5$ zu viele beobachtet. Die Wahrscheinlichkeit dieser Beobachtung oder einer extremeren ist unter $\mathbb{P}_{1/2}$ approximativ gegeben durch

$$\mathbb{P}_{1/2}\,(|S_n - 50| > 5) \approx 1 - \mathbb{P}_{1/2}\,(-1 < Z < 1) \approx 1 - 0.68 = 0.32.$$

Wir können die Hypothese also nicht verwerfen. Haben wir hingegen 540 Beobachtungen 'Kopf' unter 1000 Würfen, so gilt

$$\mathbb{P}_{1/2}\,(|S_n - 500| > 40) \approx 1 - \mathbb{P}_{1/2}\,(-2.53 < Z < 2.53) \approx 1 - 0.99 = 0.01$$

und die Hypothese $p = 1/2$ kann verworfen werden, da unsere Daten (oder extremere) unter $\mathbb{P}_{1/2}$ lediglich mit einer Wahrscheinlichkeit von 1% auftreten.

## 1.2 Grundbegriffe und Annahmen

Wir formalisieren nun das Setting unseres einführenden Beispiels. Wie dort eingangs erwähnt, gehen wir davon aus, dass beobachtete Daten $x_1, \ldots, x_n$ als Realisierungen von Zufallsvariablen $X_1, \ldots, X_n$ angesehen werden können. Für die mathematische Beschreibung sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein (allgemeiner) Wahrscheinlichkeitsraum und $X_1, \ldots, X_n$ Zufallsvariablen auf $\Omega$, wobei $X_i$ Werte in $E_i$ annehme für $i = 1, \ldots, n$. Wir nennen

$$\mathcal{X} = \bigtimes_{i=1}^{n} E_i$$

den *Beobachtungsraum* (oder *Stichprobenraum*) und definieren ein Mengensystem $\mathcal{F} \subset \mathcal{P}(\mathcal{X})$, sodass $(\mathcal{X}, \mathcal{F})$ ein messbarer Raum ist sowie $X = (X_1, \ldots, X_n)$ eine Zufallsvariable mit Werten in $\mathcal{X}$. Falls der Beobachtungsraum $\mathcal{X}$ abzählbar ist, wählen wir $\mathcal{F} = \mathcal{P}(\mathcal{X})$, ansonsten ist $\mathcal{F}$ eine geeignete $\sigma$-Algebra auf $\mathcal{X}$. Wir beschränken uns in letzterem Falle

auf den in der Stochastik I behandelten Fall $\mathcal{X} = \mathbb{R}^n$ und $\mathcal{F} = \mathscr{B}(\mathbb{R}^n)$ für die Borel'sche $\sigma$-Algebra $\mathscr{B}(\mathbb{R}^n)$. Es sei an dieser Stelle daran erinnert, dass für eine Zufallsvariable $X$ dann $X^{-1}(F) \in \mathcal{A}$ für alle $F \in \mathcal{F}$ gilt. Außerdem sei an das Bildmaß

$$\mathbb{P}^X = \mathbb{P}^{X_1,\dots,X_n}$$

erinnert, welches der gemeinsamen Verteilung von $X_1, \dots, X_n$ entspricht. Dass die Beobachtungen $x_1, \dots, x_n$ Realisierungen von Zufallsvariablen sein sollen bedeutet also, dass $x_i = X_i(\omega)$ für alle $i = 1, \dots, n$ für ein geeignetes $\omega \in \Omega$ bzw. $X(\omega) = (x_1, \dots, x_n)$.

**Die Grundidee**: In der Statistik gehen wir nun davon aus, dass wir weder den Grundraum $(\Omega, \mathcal{A}, \mathbb{P})$ noch die Zufallsvariable $X$ kennen. Damit ist insbesondere das Bildmaß $\mathbb{P}^X$, also die wahre Verteilung der beobachteten Daten, nicht bekannt. Unser Ziel wird es sein, aufgrund der Beobachtungen Rückschlüsse auf Form oder Eigenschaften von $\mathbb{P}^X$ zu ziehen. Dazu nimmt man vereinfachend an, dass die wahre unbekannte Verteilung $\mathbb{P}^X$ innerhalb einer bestimmten Familie $\{\mathbb{P}_\theta : \theta \in \Theta\}$ von Wahrscheinlichkeitsmaßen auf $(\mathcal{X}, \mathcal{F})$ liegt, d.h. dass $\mathbb{P}^X = \mathbb{P}_{\theta_0}$ für ein $\theta_0 \in \Theta$.

**Konvention**: Wir werden durchgehend den zu $\mathbb{P}_\theta$ gehörenden Erwartungswert mit $\mathbb{E}_\theta$ bezeichnen. Analog bezeichnet $\mathrm{Var}_\theta$ die zugehörige Varianz.

**Definition 1.1** (**Statistisches Modell**). *Das Tripel $\mathcal{E} = (\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ aus Beobachtungsraum $\mathcal{X}$, der $\sigma$-Algebra $\mathcal{F}$ und der Familie $\{\mathbb{P}_\theta : \theta \in \Theta\}$ von Wahrscheinlichkeitsmaßen auf $(\mathcal{X}, \mathcal{F})$ heißt* statistisches Modell. *Die Menge $\Theta$ heißt* Parameterraum *oder auch* Parametermenge. *Ist $\Theta$ eine Teilmenge eines endlich-dimensionalen Vektorraums (z.B. $\Theta \subset \mathbb{R}^k$), so spricht man auch von einem* parametrischen Modell, *andernfalls von einen* nicht-parametrischen Modell.

**Bemerkung.** Bei der Definition des parametrischen Modells haben wir bewusst darauf verzichtet, $\Theta$ mit einer (affinen) Vektorraumstruktur zu versehen und lediglich gefordert, dass $\Theta$ Teilmenge eines Vektorraums ist. Schätzen wir etwas wie im Eingangsbeispiel in Abschnitt 1.1 die Erfolgswahrscheinlichkeit $p$ einer Bernoulli-Verteilung, so gilt $\Theta = (0, 1)$ und dieses Intervall ist kein (affiner) Vektorraum. Anders sähe es bei der Erwartungswertschätzung einer Normalverteilung aus, hier ist $\Theta = \mathbb{R}$ eine natürliche Wahl.

**Beispiel 1.2** (**Produktmodell**). Oft nimmt man an, dass die Daten $x_1, \dots, x_n$ Realisierungen unabhängiger, identisch verteilter Zufallsvariablen $X_1, \dots, X_n$ sind. Nehmen diese Werte in $E$ an, so ist $\mathcal{X} = E^n$ und

$$\mathbb{P}_\theta = \bigotimes_{i=1}^{n} \overline{\mathbb{P}}_\theta =: \overline{\mathbb{P}}_\theta^{\otimes n},$$

d.h. $\mathbb{P}_\theta$ ist das $n$-fache Produktmaß eines Wahrscheinlichkeitsmaßes $\overline{\mathbb{P}}_\theta$ auf $E$ (wobei letzterer Raum noch um eine geeignete $\sigma$-Algebra ergänzt werden muss). Für die unbekannte Verteilung $\mathbb{P}^X = \mathbb{P}^{X_1,\dots,X_n}$ gilt in diesem Fall

$$\mathbb{P}^X = \bigotimes_{i=1}^{n} \overline{\mathbb{P}}_{\theta_0},$$

d.h. die unbekannte Verteilung $\mathbb{P}^X$ ist durch die eindimensionale Verteilung $\overline{\mathbb{P}}_{\theta_0} = \mathbb{P}^{X_1}$ bereits eindeutig festgelegt. Bei uns wird entweder $E$ endlich/abzählbar sein und $E^n$ mit der Potenzmenge versehen zu einem messbaren Raum (z.B. $E = \{0,1\}, \mathbb{N}, \mathbb{Z}$), oder $E = \mathbb{R}$ und $\mathbb{R}^n$ mit der entsprechenden Borel-$\sigma$-Algebra $\mathscr{B}(\mathbb{R}^n)$ zu einem solchen. Wir schreiben in letzterem Falle für das Produktmodell auch

$$\mathcal{E} = \left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\}\right) =: \overline{\mathcal{E}}^n$$

für $\overline{\mathcal{E}} = (\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathbb{P}_\theta : \theta \in \Theta\})$ und nennen $\mathcal{E}$ das zu $\overline{\mathcal{E}}$ gehörende $n$-fache Produktmodell (analog endliches/abzählbares $E$).

**Erinnerung**: Ist $\mathbb{P}_\theta$ ein diskretes W-Maß mit Zähldichte $f_\theta$ oder ein stetiges W-Maß mit Riemann-Dichte $f_\theta$, so ist die Zähldichte bzw. Riemann-Dichte $f_{\mathbb{P}_\theta^{\otimes n}}$ des Produktmaßes $\mathbb{P}_\theta^{\otimes n}$ gegeben durch

$$f_{\mathbb{P}_\theta^{\otimes n}}(x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i).$$

**Beispiel 1.3** (**$n$-faches Produktmodell mit Normalverteilungen**)**.** Für die Parametermenge $\Theta = \mathbb{R} \times (0, \infty)$ sei $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)$ die Normalverteilung mit Mittelwert $\mu$ und Varianz $\sigma^2$. Das zugehörige Produktmodell ist

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathcal{N}(\mu, \sigma^2)^{\otimes n} : (\mu, \sigma^2) \in \Theta\}\right)$$

und heißt das $n$-fache Produktmodell mit Normalverteilungen. Man kann auch einen der Parameter als gegeben voraussetzen, dann erhält man z.B. für $\Theta = \mathbb{R}$ das statistische Modell

$$\left(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathcal{N}(\mu, 1)^{\otimes n} : \mu \in \Theta\}\right).$$

Analog können $n$-fache Produktmodell auch für beliebe andere Verteilungen definiert werden.

**Beispiel 1.4** (**Unabhängige, aber nicht identisch verteilte Beobachtungen**)**.** Sind die Zufallsvariablen $X_i$, welche die Daten generieren, nur unabhängig, aber nicht identisch verteilt, so ist die Verteilung $\mathbb{P}_\theta = \bigotimes_{i=1}^n \mathbb{P}_{\theta_i}$ eine Produktverteilung mit nicht-identischen Faktoren und Parameter $\theta = (\theta_1, \ldots, \theta_n)$. Sind alle $\mathbb{P}_{\theta_i}$ parametrisch, so gilt dies auch für $\mathbb{P}_\theta$. Ein Beispiel wäre $\mathbb{P}_\theta = \bigotimes_{i=1}^n \mathrm{Pois}(\lambda_i)$ und $\theta = (\lambda_1, \ldots, \lambda_n) \in \Theta := (0, \infty)^n$.

**Bemerkung.** Jede Zufallsvariable $X : \Omega \longrightarrow E$ mit $E \subset \mathbb{R}$ abzählbar kann auch als Zufallsvariable $X : \Omega \longrightarrow \mathbb{R}$ aufgefasst werden (im Sinne von Definition $I.4.13$), da nach Lemm $I.4.8$ $X(\Omega) \in \mathscr{B}(\mathbb{R})$ für das abzählbare Bild $X(\Omega) \subset E$. Daher beschränken wir uns in vielen Betrachtungen auf das Modell $\mathcal{E} = (\mathbb{R}, \mathscr{B}(\mathbb{R}), \{\mathbb{P}_\theta : \theta \in \Theta\})$ bzw. das zugehörige Produktmodell $\mathcal{E}^n$.

**Beispiel 1.5** (**Nichtparametrisches Modell**)**.** Ein klassisches nicht-parametrisches Beispiel ist z.B. ein Produktmodell aus stetig verteilten Zufallsvariablen, wobei jede eindimensionale Verteilung aus der Menge

$$\{\mathbb{P}_\theta : \ \mathbb{P}_\theta \text{ besitzt eine Riemann-Dichte } \theta\},$$

stammt und die Parametermenge gegeben ist durch

$$\Theta = \left\{p : \mathbb{R} \longrightarrow \mathbb{R}_{\geq 0} : p \text{ ist Riemann-integrierbar mit } \int_{\mathbb{R}} p(x)dx = 1\right\}.$$

**Definition 1.6 (Statistik).** *Es sei $\mathcal{E} = (\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ ein statistisches Modell. Jede (messbare) Funktion $T : \mathcal{X} \longrightarrow \Theta'$ in eine Menge $\Theta'$ heißt* Statistik.

**Bemerkung.** Da $\mathcal{X}$ der Bildraum der Zufallsvariablen $X = (X_1, \ldots, X_n)$ ist, ist auch für jede Statistik $T : \mathcal{X} \longrightarrow \Theta'$ die Abbildung

$$T \circ X : \Omega \longrightarrow \Theta'$$

eine Zufallsvariable. Für einen beobachteten Datenvektor $x = (x_1, \ldots, x_n)$ ist dann $T(x) = T(X(\omega))$ eine mögliche Realisierung dieser Zufallsvariablen.

**Beispiel 1.7.** Wir erinnern nochmals an das Beispiel aus Abschnitt 1.1, das formal dem statistischen Modell

$$\mathcal{E}_{\text{Ber}} = (\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\text{Ber}(p) : p \in (0,1)\})$$

entspricht, wobei $\text{Ber}(p)$ die Bernoulli-Verteilung zum Parameter $p$ bezeichne.

(1) Definieren wir $T_1 : \{0,1\}^n \longrightarrow (0,1)$ durch $T_1(x_1, \ldots, x_n) = \frac{1}{n}\sum_{k=1}^n x_k = \overline{x}_n$, erhalten wir den Punktschätzer $\hat{p}_n$ als $\hat{p}_n = T_1 \circ X$.

(3) Setzen wir $T_2 : \{0,1\}^n \longrightarrow \{I \subset \mathbb{R} : \ I \text{ ist ein Intervall}\}$ mit

$$T_2(x_1, \ldots, x_n) = \left[\overline{x}_n - 1.96\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}, \overline{x}_n + 1.96\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}\right],$$

erhalten wir mit $I_n = T_2 \circ X$ unseren Intervallschätzer (Messbarkeitsfragen klammern wir an dieser Stelle aus).

(3) Ähnlich wie in (2) lässt sich auch ein Test als Statistik $T_3$ schreiben, wobei der Wertebereich von $T_3$ dann durch $\{0,1\}$ gegeben ist (womit wir die Entscheidung des Verwerfens der Hypothese als Zahlen codieren). Näheres dazu später im Kapitel über Tests.

# 2 Parametrische Schätztheorie

Wir beginnen dieses Kapitel mit der Einführung einiger Begriffe und lernen typische Schätzer für die zentralen Größen Erwartungswert und Varianz kennen. Danach werden wir uns etwas systematischer mit der Konstruktion von Schätzern beschäftigen und dabei die Momentenmethode kennenlernen, sowie sogenannte Maximum-Likelihood-Schätzer untersuchen.

## 2.1 Grundbegriffe der Schätztheorie

**Definition 2.1 ((Punkt-)Schätzer).** *Seien $\mathcal{E} = (\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ ein statistisches Modell und $\Theta'$ eine Menge, sowie $g : \Theta \longrightarrow \Theta'$ eine Funktion. Dann heißt jede Statistik $T : \mathcal{X} \longrightarrow \Theta'$* (Punkt-)Schätzer *für $g(\theta)$.*

**Bemerkung.**

(1) Möchte man den Parameter $\theta$ selbst schätzen, ist $\Theta = \Theta'$ und $g(x) = \mathrm{id}_\Theta$ die Identität.

(2) Häufig besteht Interesse an der Schätzung von Erwartungswert und Varianz von $\mathbb{P}_\theta$. Im $n$-fachen Produktmodell mit Exponentialverteilungen $\mathbb{P}_\theta = \mathrm{Exp}(\theta)$ wählt man $\Theta = (0, \infty) = \Theta'$ und

$$g_1(\theta) := \mathbb{E}_\theta[X_1] = \frac{1}{\theta} \quad \text{bzw.} \quad g_2(\theta) = \mathrm{Var}_\theta(X_1) = \frac{1}{\theta^2}.$$

(3) Zur Schätzung des zweiten Moments im $n$-fachen Normalverteilungsmodell mit $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ wählt man $g(\theta) = \mathbb{E}_{(\mu,\sigma^2)}[X_1] = \sigma^2 + \mu^2$. Hier gilt $\Theta' = \mathbb{R} \neq \Theta$.

(4) Oft wird auch die Verkettung $T \circ X$ als Schätzer für $g(\theta)$ bezeichnet.

Nach unserer Definition ist auch jede konstante Funktion $T(x) = c \in \Theta'$ ein möglicher Punktschätzer für $g(\theta)$, auch wenn sie natürlich nur in den seltensten Fällen gute Näherungswerte für $g(\theta)$ ergibt. Um 'gute Näherungswerte' zu beschreiben, führen wir nun zwei Gütekriterien für Schätzer ein.

**Definition 2.2** (**Erwartunstreue und konsistente Schätzer**).

*(a) Es sei $\mathcal{E} = (\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ ein statistisches Modell und $g : \Theta \to \Theta' \subset \mathbb{R}$. Dann heißt ein Schätzer $T$ für $g(\theta)$* erwartungstreu, *falls*

$$\mathbb{E}_\theta[T(X)] = g(\theta) \quad \text{für alle } \theta \in \Theta.$$

*(b) Es sei $(\mathcal{E}^n)_{n\in\mathbb{N}}$ eine Folge statistischer Modelle mit identischer Parametermenge $\Theta$, d.h. $\mathcal{E}^n = (\mathcal{X}^n, \mathcal{F}^n, \{\mathbb{P}^n_\theta : \theta \in \Theta\})$. Weiter sei $g : \Theta \longrightarrow \Theta' \subset \mathbb{R}$ und $(T_n)_{n\in\mathbb{N}}$ eine Folge von Punktschätzern für $g(\theta)$. Dann heißt die Schätzfolge $(T_n(X^n))_{n\in\mathbb{N}}$* konsistent, *falls für alle $\epsilon > 0$*

$$\mathbb{P}^n_\theta \left(|T_n(X^n) - g(\theta)| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0 \qquad \text{für alle } \theta \in \Theta.$$

**Bemerkung.** Für die Konsistenz haben wir angenommen, dass $\Theta' \subset \mathbb{R}$ um den Abstand $|T_n(X^n) - g(\theta)|$ definieren zu können. Die Definition ließe sich daher auf metrische Räume verallgemeinern. Bei der Erwartungstreue hingegen wird benötigt, dass die Zufallsvariable $T(X)$ einen definierten Erwartungswert besitzt. Aus der Stochastik I kennen wir den Erwartungswertbegriff für $\mathbb{R}$-wertige Zufallsvariablen - auch hiervon existieren Verallgemeinerungen, jedoch nicht in derart großer Allgemeinheit. Insbesondere kann die Konsistenz für allgemeinere $\Theta$ bzw. $\Theta'$ formuliert werden als die Erwartungstreue.
**Bemerkung.** Ist $\mathcal{E} = \overline{\mathcal{E}}^n$ ein Produktmodell, so nennen wir einen Schätzer $T_n(X_1, \ldots, X_n)$ konsistent, wenn die Schätzfolge $(T_n(X_1, \ldots, X_n))_{n\in\mathbb{N}}$ konsistent im Sinne obiger Definition ist.

**Beispiel 2.3.** Wir betrachten ein weiteres Mal das $n$-fache Produktmodell mit Bernoulli-Verteilung, d.h. das statistische Modell $\mathcal{E}^n = (\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathrm{Ber}(p) : p \in (0,1)\})$ aus Abschnitt 1.1.

(1) Wie bereits in Abschnitt 1.1 bewiesen ist der Schätzer

$$\hat{p}_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

erwartungstreu und konsistent. Dort haben wir ebenfalls schon die Bedeutung dieser beiden Begriffe diskutiert.

(2) Jeder konstanten Schätzer ungleich dem wahren Parameter ist offensichtlich weder erwartungstreu noch konsistent.

(3) Für $\Theta' = [0, 1]$ ist auch $\hat{p}'_n(X_1, \dots, X_n) = X_1$ ein Schätzer für $p$. Dieser kann nur die Werte 0 oder 1 annehmen, ist jedoch erwartungstreu, denn

$$\mathbb{E}_p[\hat{p}'_n(X_1, \dots, X_n)] = \mathbb{E}_p[X_1] = p.$$

Intuitiv ist jedoch bereits klar, dass $\hat{p}'_n$ mit mehr erhobenen Daten nicht besser wird und in der Tat ist $\hat{p}'_n$ nicht konsistent, da

$$\mathbb{P}_p\left(|\hat{p}'_n(X_1, \dots, X_n) - p| > \epsilon\right) = 1,$$

falls $\epsilon < \min\{p, 1-p\}$, unabhängig von $n$.

Bevor wir uns in den nächsten zwei Unterabschnitten jeweils mit einem allgemeinen Ansatz zur Konstruktion von Schätzern beschäftigen, stellen wir hier noch zwei klassische Schätzer für Erwartungswert und Varianz vor.

**Definition 2.4 (Empirischer Mittelwert und empirische Varianz).** *Sei $X = (X_1, \dots, X_n)$ ein Vektor von Zufallsvariablen. Dann heißt*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(empirischer) Mittelwert *und die Größe*

$$s_n^2(X) = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

empirische Varianz.

Das nächste Resultat gibt an, unter welchen Voraussetzungen diese beiden Schätzer erwartungstreu und konsistent sind.

**Satz 2.5 (Schätzung von Erwartungswert und Varianz).** *Es sei $(\mathcal{E}^n)_{n\in\mathbb{N}} = \left((\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})\right)_{n\in\mathbb{N}}$ eine Folge von Produktmodellen. Dann gilt:*

*(a) Ist $\mathbb{E}_\theta[X_1^2] < \infty$ für alle $\theta \in \Theta$, so ist $(\overline{X}_n)_{n\in\mathbb{N}}$ eine erwartungstreue und konsistente Schätzfolge für den Erwartungswert $g_1(\theta) = \mathbb{E}_\theta[X_1]$.*

*(b) Ist $\mathbb{E}_\theta[X_1^4] < \infty$ für alle $\theta \in \Theta$, so ist $(s_n^2(X))_{n\in\mathbb{N}}$ eine erwartungstreue und konsistente Schätzfolge für die Varianz $g_2(\theta) = \mathrm{Var}_\theta(X_1)$.*

*Beweis.* Aufgrund der Linearität des Erwartungswerts und der identischen Verteilung der $X_1, \dots, X_n$ gilt

$$\mathbb{E}_\theta\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_1] = \mathbb{E}_\theta[X_1] = g_1(\theta)$$

und die Erwartungstreue in (a) folgt. Die Konsistenz ist eine direkte Folge aus dem schwachen Gesetz der großen Zahlen (Satz *I*.3.21), welches anwendbar ist, da mit $\overline{\mathbb{E}}_\theta[X_1^2] < \infty$ auch $\mathrm{Var}_\theta(X_1) < \infty$.
Für Teil (b) berechnen wir zunächst direkt, dass

$$\begin{aligned} s_n^2(X) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2\overline{X}_n\sum_{i=1}^{n}X_i + n\overline{X}_n^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2n\overline{X}_n^2 + n\overline{X}_n^2\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{n}{n-1}\overline{X}_n^2. \end{aligned} \tag{2}$$

Aufgrund der Voraussetzung $\overline{\mathbb{E}}_\theta[X_1^4] < \infty$ folgt $\mathrm{Var}_\theta(X_1^2) < \infty$ für alle $\theta \in \Theta$. Damit konvergiert nach dem schwachen Gesetz der großen Zahlen (Satz *I*.3.21)

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 \longrightarrow_{\mathbb{P}_\theta} \mathbb{E}_\theta[X_1^2].$$

Da wir bereits für Teil (a) gesehen haben, dass $\overline{X}_n \longrightarrow_{\mathbb{P}_\theta} \mathbb{E}_\theta[X_1]$, folgt mit der Identität (2), Lemma *I*.5.13 und der Tatsache, dass für eine reelle Folge $(a_n)_{n\in\mathbb{N}}$ mit $a_n \to a$ und eine Folge von Zufallsvariablen mit $X_n \to_{\mathbb{P}} X$ auch $a_nX_n \to_{\mathbb{P}} aX$ (nachrechnen!)

$$s_n^2(X) \longrightarrow_{\mathbb{P}_\theta} \mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2 = \mathrm{Var}_\theta(X_1).$$

Für die Erwartungstreue nutzen wir, dass

$$\overline{X}_n^2 = \frac{1}{n^2}\sum_{i=1}^{n}X_i^2 + \frac{1}{n^2}\sum_{\substack{i,j=1,\dots,n\\ i\neq i}}X_iX_j.$$

Dann erhalten wir mit (2), der Linearität des Erwartungswerts und der Tatsache, dass $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$ im zweiten Schritt,

$$\begin{aligned} \mathbb{E}_\theta\left[s_n^2(X)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_i^2] + \frac{1}{n(n-1)}\sum_{\substack{i,j=1,\dots,n\\ i\neq j}}\mathbb{E}[X_iX_j] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_1^2] + \frac{1}{n(n-1)}\sum_{\substack{i,j=1,\dots,n\\ i\neq j}}\mathbb{E}_\theta[X_1]\mathbb{E}_\theta[X_1] \\ &= \mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2 = \mathrm{Var}_\theta(X_1). \end{aligned}$$

□

**Bemerkung.**

(1) Damit wir das schwache Gesetz der großen Zahlen anwenden können, brauchen wir in Teil (a) von Proposition 2.9 die Voraussetzung zweiter und in Teil (b) vierter Momente. Das starke Gesetz der Großen Zahlen (→ Wahrscheinlichkeitstheorie) schwächt die Voraussetzung von Satz *I*.3.21 dahingehend ab, dass keine endliche Varianz, sondern nur ein endliches erstes Moment benötigt werden. Damit bleibt Teil (a) in Proposition 2.9 gültig, falls man nur $\mathbb{E}_\theta[|X_1|] < \infty$ fordert und in Teil (b) $\mathbb{E}_\theta[X_1^2] < \infty$. Diese Voraussetzungen sind insofern natürlicher, da sie genügen, damit der zu schätzende Parameter existiert.

(2) Die Normierung mit $n-1$ in der Definition von $s_n^2(X)$ mag auf den ersten Blick überraschen, stellt jedoch sicher, dass $s_n^2(X)$ erwartungstreu ist. Die Konsistenz würde hingegen auch mit einer $1/n$-Normierung gültig bleiben

## 2.2 Momentenschätzer

Wir betrachten in diesem Kapitel ein Produktmodell $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ und bezeichnen mit

$$m_k(\theta) := \mathbb{E}_\theta[X_1^k]$$

das $k$-te Moment von $X_1$. Wir nehmen weiter an, dass $g : \Theta \longrightarrow \Theta'$ gegeben ist durch

$$g(\theta) = h\left(m_1(\theta), \ldots, m_l(\theta)\right)$$

für eine Funktion $h : \mathbb{R}^l \longrightarrow \Theta'$. Das bedeutet, dass $g(\theta)$ eine Funktion der ersten $l$ Momente von $X_1$ ist. Hierzu kennen wir bereits das folgende Beispiel: Ist $\mathbb{P}_\theta = \mathrm{Exp}(\theta)$ mit $\theta \in (0, \infty)$, so gilt $\mathbb{E}_\theta[X_1] = \frac{1}{\theta}$, d.h. $\theta = h(m_1(\theta))$ für $h(x) = 1/x$.

Für jedes Moment $m_1(\theta)$ gibt es analog zum empirischen Mittelwert den naheliegenden Schätzer des sogenannten $k$-ten empirischen Moments

$$\hat{m}_{k,n}(X) = \hat{m}_{k,n}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i^k.$$

Die Idee der Momentenmethode besteht nun darin, $g(\theta)$ durch den Wert $h(\hat{m}_{1,n}(X), \ldots, \hat{m}_{l,n}(X))$ zu schätzen.

**Definition 2.6 (Momentenschätzer).** *Sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein $n$-faches Produktmodell, $m_k(\theta)$ das $k$-te Moment und $\hat{m}_{k,n}(X)$ das $k$-te empirische Moment, sowie $g : \Theta \longrightarrow \Theta'$ gegeben durch*

$$g(\theta) := h\left(m_1(\theta), \ldots, m_l(\theta)\right)$$

*für eine (messbare) Funktion $h : \mathbb{R}^l \longrightarrow \Theta'$. Dann heißt der Schätzer*

$$\hat{g}_n(X) := h\left(\hat{m}_{1,n}(X), \ldots, \hat{m}_{l,n}(X)\right)$$

Momentenschätzer *für $g(\theta)$.*