Stochastik II

Skript zur Vorlesung im SoSe 2025

Dr. Johannes Brutsche

2025

Inhaltsverzeichnis

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

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

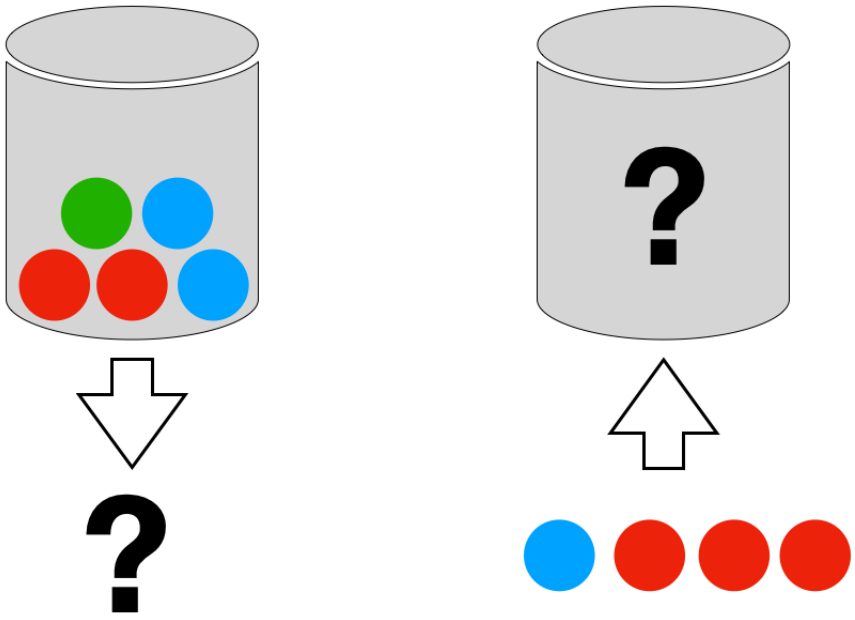

Abbildung 1: Darstellung der Konzepte der Wahrscheinlichkeitsrechnung (Stochastik I) und Statistik (Stochastik II). Das Modell wird hier durch eine Urne symbolisiert, der Pfeil veranschaulicht die Analyserichtung.

für $i = 1, \dots, n$. 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 $p(X_1, \dots, X_n)$ ableiten (*Punktschätzer*) oder ein Intervall $[a(X_1, \dots, X_n), b(X_1, \dots, X_n)]$ 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

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

Wir folgen hier der Konvention, einen Schätzer für einen unbekannten Parameter θ, welcher auf n Beobachtungen basiert, mit $\hat{\theta}_n$ zu bezeichnen. Man beachte, dass $\hat{p}_n$ eine Funktion der Daten $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ 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 $(X_1^j, \ldots, X_n^j)$, $j = 1, \ldots, N$, 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 Ω, 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 σ-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 σ-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 Var_θ 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 σ-Algebra $\mathcal{F}$ und der Familie $\{\mathbb{P}_\theta : \theta \in \Theta\}$ von Wahrscheinlichkeitsmaßen auf $(\mathcal{X}, \mathcal{F})$ heißt* statistisches Modell. *Die Menge Θ heißt* Parameterraum *oder auch* Parametermenge. *Ist Θ 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, Θ mit einer (affinen) Vektorraumstruktur zu versehen und lediglich gefordert, dass Θ 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 σ-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-σ-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_θ oder ein stetiges W-Maß mit Riemann-Dichte f_θ, 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 μ und Varianz σ^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 Θ' 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

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

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 Θ' 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 θ 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 Θ, d.h. $\mathcal{E}^n = (\mathcal{X}^n, \mathcal{F}^n, \{\mathbb{P}_\theta^n : \theta \in \Theta\})$. 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 $\epsilon > 0$*

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

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 Θ bzw. Θ' 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 $\mathcal{E}^n = (\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathrm{Ber}(p)^{\otimes n} : p \in (0,1)\})$ 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 $a_n X_n \to_{\mathbb{P}} aX$ (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 ($\rightarrow$ 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 $g(\theta)$.*

Beispiel 2.7.

(1) Sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein n-faches Produktmodell mit $\mathbb{E}_\theta[X_1^2] < \infty$ für alle $\theta \in \Theta$. Da $g(\theta) = \mathrm{Var}_\theta(X_1) = m_2(\theta) - m_1(\theta)^2$ gilt $g(\theta) = h(m_1(\theta), m_2(\theta))$ für $h : \mathbb{R}^2 \longrightarrow \mathbb{R}$ mit $h(x, y) = y - x^2$. Damit ist der Momentenschätzer für die Varianz gegeben durch

$$\hat{T}_n(X) = \hat{m}_{2,n}(X) - \hat{m}_{1,n}(X)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \frac{n-1}{n} s_n^2(X).$$

Insbesondere ist dieser Momentenschätzer nicht erwartungstreu (vgl. Satz 2.9).

(2) Es sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ das n-fache Produktmodell mit stetigen Gleichverteilungen d.h. $\mathbb{P}_\theta = \mathbb{P}_{(a,b)} = \mathcal{U}([a, b])$ und $\Theta = \{(a, b) \in \mathbb{R}^2 : a < b\}$. Dann gilt $m_1(\theta) = \mathbb{E}_\theta[X_1] = (a+b)/2$ sowie $\mathrm{Var}_\theta(X_1) = (b-a)^2/12$ (Blatt 1, Aufgabe 1). Da entsprechend $\sqrt{3\mathrm{Var}_\theta(X_1)} = (b-a)/2$, gilt

$$a = m_1(\theta) - \sqrt{3\mathrm{Var}_\theta(X_1)} = m_1(\theta) - \sqrt{3(m_2(\theta) - m_1(\theta)^2)}$$
$$b = m_1(\theta) + \sqrt{3\mathrm{Var}_\theta(X_1)} = m_1(\theta) + \sqrt{3(m_2(\theta) - m_1(\theta)^2)}.$$

Daraus ergeben sich die Momentenschätzer

$$\hat{a}_n = \hat{m}_{1,n}(X) - \sqrt{3(\hat{m}_{2,n}(X) - \hat{m}_{1,n}(X)^2)}$$
$$\hat{b}_n = \hat{m}_{1,n}(X) + \sqrt{3(\hat{m}_{2,n}(X) - \hat{m}_{1,n}(X)^2)}.$$

(3) Momentenschätzer sind oft nicht eindeutig. Betrachten wir das n-fache Produktmodell mit Poisson-Verteilungen, so gilt für $\mathbb{P}_\theta = \mathrm{Poi}(\theta)$, dass $\mathbb{E}_\theta[X_1] = \mathrm{Var}_\theta(X_1) = \theta$. Daher kann der unbekannte Parameter $\theta \in (0, \infty)$ entweder durch den Momentenschätzer $\hat{m}_{1,n}(X)$ oder den Momentenschätzer $\hat{m}_{2,n}(X) - \hat{m}_{1,n}(X)^2$ geschätzt werden.

Als nächstes wollen wir uns der Konsistenz von Momentenschätzern widmen, wobei wir uns auf den Fall $g(\theta) \in \mathbb{R}$ beschränken (ebenso wie bei der darauffolgenden Verteilungskonvergenz). Unter geeigneten Voraussetzungen an die Momente erhalten wir aus dem schwachen Gesetz der großen Zahlen direkt die stochastische Konvergenz $\hat{m}_{k,n}(X) \longrightarrow_{\mathbb{P}_\theta} m_k(\theta)$ der empirischen Momente. Dies motiviert die Frage, ob und unter welchen Voraussetzungen, sich die stochastische Konvergenz von einer Folge $(X_n)_{n\in\mathbb{N}}$ auf die zugehörige Folge $(f(X_n))_{n\in\mathbb{N}}$ übertragen lässt. Für die Konvergenz des Momentenschätzers genügt der folgende Spezialfall eines solchen Resultats:

Lemma 2.8 (**Stochastische Konvergenz und Stetigkeit**). *Für $j = 1, \ldots, d$ sei $(X_n^{(j)})_{n\in\mathbb{N}}$ eine Folge reellwertiger Zufallsvariablen mit $X_n^{(j)} \longrightarrow_{\mathbb{P}} c_j \in \mathbb{R}$. Weiter sei $h : \mathbb{R}^d \longrightarrow \mathbb{R}$ eine Funktion, welche in $c = (c_1, \ldots, c_d)$ stetig ist. Dann gilt*

$$h\left(X_n^{(1)}, \ldots, X_n^{(d)}\right) \longrightarrow_{\mathbb{P}} h\left(c_1, \ldots, c_d\right).$$

Beweis. Es sei $\epsilon > 0$ und $\delta > 0$ passend gewählt, sodass $|h(x) - h(c)| < \epsilon$ für alle $x = (x_1, \dots, x_d)$ mit $\|x - c\|_\infty < \delta$. Hierbei bezeichnet $\|\cdot\|_\infty$ die Maximumsnorm im $\mathbb{R}^d$, d.h. $\|(x_1, \dots, x_n)\|_\infty = \max_{i=1,\dots,d} |x_i|$. Die Wahl eines solchen δ ist möglich, da h in c bezüglich jeder Norm stetig ist (da auf dem $\mathbb{R}^d$ alle Normen äquivalent sind). Es folgt dann

$$\begin{aligned}\mathbb{P}\left(\left|h\left(X_n^{(1)}, \dots, X_n^{(d)}\right) - h\left(c_1, \dots, c_d\right)\right| > \epsilon\right) &\leq \mathbb{P}\left(\left\|\left(X_n^{(1)} - c_1, \dots, X_n^{(d)} - c_d\right)\right\|_\infty > \delta\right) \\ &\leq \sum_{j=1}^{d} \mathbb{P}\left(\left|X_n^{(j)} - c_j\right| > \delta\right) \longrightarrow 0,\end{aligned}$$

wobei die zweite Ungleichung die endliche Additivität des Maßes $\mathbb{P}$ nutzt. □

Satz 2.9 (Konsistenz von Momentenschätzern). *Es sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein n-faches Produktmodell und $\mathbb{E}_\theta[X_1^{2l}] < \infty$ für alle $\theta \in \Theta$. Weiter sei $h : \mathbb{R}^l \longrightarrow \mathbb{R}$ stetig und $h\left(\hat{m}_{1,n}(X), \dots, \hat{m}_{l,n}(X)\right)$ der Momentenschätzer für $g(\theta) = h(m_1(\theta), \dots, m_l(\theta))$. Dann ist dieser konsistent, d.h.*

$$h\left(\hat{m}_{1,n}(X), \dots, \hat{m}_{l,n}(X)\right) \longrightarrow_{\mathbb{P}_\theta} h(m_1(\theta), \dots, m_l(\theta)).$$

Beweis. Aufgrund der Momentenannahme $\mathbb{E}_\theta[X_1^{2l}] < \infty$ existiert die Varianz $\mathrm{Var}_\theta(X_1^j)$ für alle $j = 1, \dots, l$ und $\theta \in \Theta$. Damit folgt aus dem schwachen Gesetz der großen Zahlen $\hat{m}_{j,n}(X) \longrightarrow_{\mathbb{P}_\theta} m_j(\theta)$ für $j = 1, \dots, l$ und die Aussage folgt aus Lemma 2.8. □

Als nächstes wenden wir uns der Verteilungskonvergenz der Momentenschätzer zu. Aus dem zentralen Grenzwertsatz erhält man direkt, dass

$$\sqrt{n}\frac{\hat{m}_{k,n}(X) - m_k(\theta)}{\sqrt{m_{2k}(\theta) - m_k(\theta)^2}} \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, 1).$$

Aus Proposition I.5.4 erhält man hieraus die alternative Formulierung

$$\sqrt{n}\left(\hat{m}_{k,n}(X) - m_k(\theta)\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\big(0, m_{2k}(\theta) - m_k(\theta)^2\big).$$

Analog zur Konsistenz fragen wir uns nun, ob sich diese Verteilungskonvergenz auf eine Funktion von $\hat{m}_{k,n}(X)$ übertragen lässt. Um das Resultat in größerer Allgemeinheit zu fassen, stellt sich zunächst sogar eine andere Frage, nämlich diejenige nach der Verteilungskonvergenz des Vektors

$$\left(\hat{m}_1(X), \dots, \hat{m}_l(X)\right).$$

Anders als bei der stochastischen Konvergenz kann diese nicht auf die Verteilungskonvergenz der Komponenten zurückgeführt werden, da die die Abhängigkeitsstruktur der Komponenten untereinander nicht akkurat berücksichtigen würde. Die Antwort besteht jedoch in einer mehrdimensionalen Version des zentralen Grenzwertsatzes.

Definition 2.10 (**Mehrdimensionale Normalverteilung**). *Es sei $d \in \mathbb{N}$, $\mu \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d\times d}$ eine symmetrische positiv definite Matrix. Wir sagen, dass ein Vektor $(X_1, \dots, X_d) \in \mathbb{R}^d$* mehrdimensional normalverteilt ist mit Erwartungswert μ und Kovarianzmatrix Σ*, wenn die gemeinsame Dichte dieses Vektors gegeben ist durch*

$$f_{X_1,\dots,X_d}(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

wobei $x \in \mathbb{R}^d$. Wir schreiben dann auch $(X_1, \dots, X_d) \sim \mathcal{N}_d(\mu, \Sigma)$.

Für die angekündigte Verallgemeinerung des zentralen Grenzwertsatzes noch eine Bemerkung: Für einen Zufallsvektor $X = (X_1, \dots, X_d)$ definieren wir den Erwartungswert komponentenweise, d.h. $\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d]) \in \mathbb{R}^d$.

Theorem 2.11 (**Mehrdimensionaler zentraler Grenzwertsatz**). *Es seien $X_1, \dots, X_n$ unabhängig identisch verteilte d-dimensionale Zufallsvektoren mit $\mathbb{E}[X_1] = \mu \in \mathbb{R}^d$ und positiv definiter Kovarianzmatrix $\Sigma = (\Sigma_{ij})_{1\le i,j\le d}$, wobei $\Sigma_{ij} = \mathrm{Cov}(X_{1,i}, X_{1,j})$ ist für den i-ten Eintrag $X_{1,i}$ des Vektors X_1. Dann gilt für das Stichprobenmittel $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$, dass*

$$\sqrt{n}\,(Z_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}_d(0, \Sigma).$$

Beweis. Der Beweis kann in Analogie zum eindimensionalen Fall geführt werden. Wir verzichten an dieser Stelle darauf und verweisen auf Theorem 21.3 in *Probability essentials von J.Jacod und P.Protter (2000, Universitext, Springer).* □

Beispiel 2.12. Es seien $X_1, \dots, X_n \sim \mathbb{P}_\theta$ unabhängig identisch verteilt mit $\mathbb{E}[X_1^{2l}] < \infty$ für alle θ. Dann sind auch $Y_i = (X_i, X_i^2, \dots, X_i^l)^T$, $i = 1, \dots, n$, unabhängig und identisch verteilt. Weiter gilt $\mathbb{E}_\theta[Y_1] = (m_1(\theta), \dots, m_l(\theta))^T$ und die Kovarianzmatrix $\Sigma(\theta) = (\Sigma_{ij}(\theta))_{1\le i,j\le d}$ gegeben durch $\Sigma_{ij} = \mathrm{Cov}_\theta(X_1^i, X_1^j)$ sei positiv definit (im Allgemeinen ist nur klar, dass sie positiv semidefinit ist, vgl. Bemerkung *I*.3.27). Dann folgt nach Theorem 2.11

$$\sqrt{n}\Big(\hat{m}_{1,n}(X) - m_1(\theta), \dots, \hat{m}_{l,n}(X) - m_l(\theta)\Big) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}_d(0, \Sigma).$$

Satz 2.13 (**δ-Methode**). *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge d-dimensionaler Zufallsvektoren und $\mu \in \mathbb{R}^d$ eine Konstante, sodass*

$$\sqrt{n}\,(X_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}_d(0, \Sigma)$$

für eine symmetrische, positiv definite Matrix $\Sigma \in \mathbb{R}^{d\times d}$. Es sei weiter $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ stetig differenzierbar, sodass für den Gradienten ∇g gelte, dass $\nabla g(\mu) \neq 0$. Dann gilt

$$\sqrt{n}\,(g(X_n) - g(\mu)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \nabla g(\mu)\Sigma(\nabla g(\mu))^T\right).$$

Beweis. Wir führen den Beweis nur für $d = 1$, d.h. $\Sigma = \sigma^2 > 0$. Nach dem Mittelwertsatz gilt

$$g(X_n) = g(\mu) + g'(\xi_n)(X_n - \mu)$$

für eine zufällige Zwischenstelle ξ_n zwischen X_n und μ. Da $(X_n - \mu) = \frac{1}{\sqrt{n}}(\sqrt{n}(X_n - \mu))$ folgt aus den Propositionen $I.5.4$ und $I.5.7$, dass $X_n \longrightarrow_{\mathbb{P}} \mu$. Wegen $|X_n - \mu| \geq |\xi_n - \mu|$, folgt aus $X_n \longrightarrow_{\mathbb{P}} \mu$ entsprechend $\xi_n \longrightarrow_{\mathbb{P}} \mu$. Da g' nach Voraussetzung stetig ist, folgt aus Lemma 2.8

$$g'(\xi_n) \longrightarrow_{\mathbb{P}} g'(\mu).$$

Nutzen wir nun zusätzlich die Identität

$$\sqrt{n}\,(g(X_n) - g(\mu)) = g'(\xi_n)\sqrt{n}(X_n - \mu),$$

die vorausgesetzte Verteilungskonvergenz von $\sqrt{n}(X_n - \mu)$, sowie Proposition $I.5.4$(iii) erhalten wir für $Z \sim \mathcal{N}(0, \sigma^2)$

$$\sqrt{n}\,(g(X_n) - g(\mu)) \xrightarrow{\mathcal{D}} g'(\mu)Z \sim \mathcal{N}\left(0, \sigma^2(g'(\mu))^2\right)$$

und die Behauptung folgt. □

Bemerkung. Ähnlich wie in Lemma 2.8 gilt nach dem sogenannten *Continuous-Mapping-Theorem* für die schwache Konvergenz, dass $X_n \xrightarrow{\mathcal{D}} X$ bereits $g(X_n) \xrightarrow{\mathcal{D}} g(X)$ für stetiges g impliziert. Der Vorteil der δ-Methode besteht an dieser Stelle darin, dass die Verteilung des Grenzwerts explizit angegeben werden kann.

Satz 2.14 (**Asymptotische Normalität von Momentenschätzern**). *Sei* $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ *ein n-faches Produktmodell mit* $\mathbb{E}_\theta[X_1^{2l}] < \infty$ *sowie* $\Sigma(\theta) = (\Sigma_{ij}(\theta))_{1\leq i,j\leq l}$ *positiv definit für* $\Sigma_{ij}(\theta) = \mathrm{Cov}_\theta(X_1^i, X_1^j)$ *für alle* $\theta \in \Theta$. *Außerdem sei* $h : \mathbb{R}^l \longrightarrow \mathbb{R}$ *stetig differenzierbar, sodass* $v_\theta := \nabla h(m_1(\theta), \ldots, m_l(\theta))^T \neq 0$ *und* $h\left(\hat{m}_{1,n}(X), \ldots, \hat{m}_{l,n}(X)\right)$ *der Momentenschätzer für* $g(\theta) = h(m_1(\theta), \ldots, m_l(\theta))$. *Dann gilt*

$$\sqrt{n}\Big(h\left(\hat{m}_{1,n}(X), \ldots, \hat{m}_{l,n}(X)\right) - h(m_1(\theta), \ldots, m_l(\theta))\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, v_\theta^T \Sigma v_\theta\right).$$

Beweis. Die Aussage folgt direkt aus Beispiel 2.12 und Satz 2.13. □

Beispiel 2.15. Wir betrachten das n-fache Produktmodell mit $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)$ und $\theta \in \Theta = \mathbb{R} \times (0, \infty)$. Dann ist der Momentenschätzer für die Varianz gegeben durch $\hat{\sigma}_n^2 = \hat{m}_{2,n}(X) - \hat{m}_{1,n}(X)^2 = h(\hat{m}_{1,n}(X), \hat{m}_{2,n}(X))$ mit $h : \mathbb{R}^2 \longrightarrow \mathbb{R}$ gegeben durch $h(x, y) = y - x^2$. Dann erhalten wir unmittelbar $\nabla h(\mu, \sigma^2) = (-2\mu, 1)$ und durch Nachschlagen der Momente der Normalverteilung

$$\Sigma = \begin{pmatrix} \mathrm{Var}_\theta(X_1) & \mathbb{E}_\theta[X_1]\mathbb{E}_\theta[X_1^2] \\ \mathbb{E}_\theta[X_1]\mathbb{E}_\theta[X_1^2] & \mathrm{Var}_\theta(X_1^2) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 4\mu^2\sigma^2 + 2\sigma^4 \end{pmatrix}.$$

Dann liefert Satz 2.14 direkt

$$\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma^2\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\sigma^4),$$

vergleiche auch Beispiel $I.5.12$, in welchem die gleiche asymptotische Verteilung für den empirischen Varianzschätzer $s_n^2(X) = \frac{n}{n-1}\hat{\sigma}_n^2$ hergeleitet wurde.

2.3 Maximum-Likelihood-Schätzer

Wir betrachten erneut das Beispiel aus Abschnitt 1.1 und unterstellen diesmal, dass wir die Münze mit 'Kopfwahrscheinlichkeit' p genau 100 mal geworfen haben, wobei 44 Mal 'Kopf' geworfen wurde. Die Zufallsvariable, deren Realisierung wir beobachten, ist in diesem Fall binomialverteilt mit Parameter $n = 100$ und unbekanntem $p \in (0,1)$. Die Eintrittswahrscheinlichkeit unseres beobachteten Ereignisses ist daher gegeben durch

$$\binom{100}{44} p^{100-44}(1-p)^{44} = \binom{100}{44} p^{56}(1-p)^{44}.$$

Da wir wissen, dass 44-mal 'Kopf' geworfen wurde (und wir diesem Ereignis daher eine hohe Wahrscheinlichkeit zugestehen wollen), besteht ein neuer Ansatz zur Schätzung von p darin, diese Eintrittswahrscheinlichkeit in p zu maximieren. Dieses Verfahren ist die sogenannte *Maximum-Likelihood-Methode.*

Definition 2.16 (Likelihoodfunktion). *Es sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein n-faches Produktmodell. Die Likelihoodfunktion $L : \mathbb{R}^n \times \Theta \longrightarrow \mathbb{R}$ ist im Falle diskreter Maße $\mathbb{P}_\theta$ gegeben durch*

$$L(x;\theta) = L(x_1,\dots,x_n;\theta) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i)$$

und im Falle von stetigen Verteilungen mit Riemann-Dichte f_θ durch

$$L(x;\theta) = L(x_1,\dots,x_n;\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

Definition 2.17 (Maximum-Likelihood-Schätzer). *Es sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein n-faches Produktmodell und L die zugehörige Likelihoodfunktion. Ein Schätzer $\hat{T}_{\mathrm{ML}} : \mathbb{R}^n \longrightarrow \Theta$, für den gilt*

$$L\left(x; \hat{T}_{\mathrm{ML}}(x)\right) = \max_{\theta\in\Theta} L(x;\theta)$$

heißt Maximum-Likelihood-Schätzer *(oder kürzer* ML-Schätzer*) für θ und wird oft mit $\hat{T}_{\mathrm{ML}} = \hat{\theta}_{\mathrm{ML}}$ bezeichnet.*

Bemerkung. Da die Logarithmusfunktion $\log : (0,\infty) \longrightarrow \mathbb{R}$ streng monoton wächst, haben die Likelihoodfunktion $L(x;\theta)$ und die sogenannte log-Likelihoodfunktion

$$l(x;\theta) = \log(L(x;\theta))$$

den gleichen Maximierer. Da $L(x;\theta)$ als ein Produkt definiert ist, welches beim Logarithmieren in eine Summe zerfällt, ist das Maximieren der log-Likelihoodfunktion rechentechnisch meist deutlich einfacher.

Beispiel 2.18. Wir betrachten das Binomialmodell vom Anfang des Kapitels: Hier war noch die Funktion

$$p \mapsto \binom{100}{44} p^{56}(1-p)^{44}$$

zu maximieren, was nach der vorangehenden Bemerkung das gleiche ist, wie

$$p \mapsto \log \binom{100}{44} + 56 \log(p) + 44 \log(1-p)$$

zu maximieren. Von dieser Funktion kann leicht mittels Differentialrechnung das Maximum $\hat{p}_{\mathrm{ML}} = 56/100$ bestimmt werden.

Beispiel 2.19 (**ML-Schätzer im Normalverteilungsmodell**). Wir betrachten das n-fache Produktmodell $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ mit Normalverteilungen, d.h. $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)$ und $\Theta = \mathbb{R} \times (0, \infty)$. Die Likelihoodfunktion ist gegeben durch

$$L(x_1, \ldots, x_n; (\mu, \sigma^2)) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

und die zugehörige log-Likelihoodfunktion durch

$$l(x_1, \ldots, x_n; (\mu, \sigma^2)) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Da diese Funktion für Parameter aus ganz Θ zweimal stetig differenzierbar ist, können wir das Maximum mittels Methoden der Analysis II bestimmen: Die notwendige Bedingung ist in diesem Fall (wir schreiben $x = (x_1, \ldots, x_n)$)

$$\nabla l(x; (\mu, \sigma^2)) = \left(\tfrac{1}{\sigma^2} \textstyle\sum_{i=1}^{n} (x_i - \mu), \quad -\tfrac{n}{2\sigma^2} + \tfrac{1}{2\sigma^4} \textstyle\sum_{i=1}^{n} (x_i - \mu)^2\right) = 0,$$

was auf die einzige Nullstelle

$$\left(\mu_0, \sigma_0^2\right) = \left(\overline{x}_n, \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2\right)$$

führt. Die Hessematrix ist in diesem Fall gegeben durch

$$H^2\big(l(x; (\mu, \sigma^2))\big) = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) \\ -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix},$$

also für unseren Kandiaten der Extremalstelle

$$H^2\big(l(x; (\mu_0, \sigma_0^2))\big) = \begin{pmatrix} -\frac{n}{\sigma_0^2} & 0 \\ 0 & -\frac{n}{2\sigma_0^4} \end{pmatrix}.$$

Diese Matrix ist offensichtlich negativ definit, sodass es sich bei (μ_0, σ_0^2) um ein lokales Maximum handelt. Betrachten der Grenzfälle $\mu \to \pm\infty$ und $\sigma^2 \to \infty$ für welche die log-Likelihoodfunktion gegen $-\infty$ konvergiert, zeigt, dass es sich dabei auch um ein globales Maximum handelt. Daher gilt

$$\hat{\mu}_{\mathrm{ML}}(X) = \overline{X}_n \qquad \text{und} \qquad \hat{\sigma}^2_{\mathrm{ML}}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Insbesondere sehen wir, dass $\hat{\sigma}^2_{\mathrm{ML}}(X) = \frac{n-1}{n} s_n^2(X)$, d.h. der ML-Schätzer ist in diesem Fall nicht erwartungstreu.

Beispiel 2.20 (**ML-Schätzer für die Uniformverteilung**). Wir betachten das n-fache Produktmodell $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ mit stetiger Uniformverteilung auf $[0,\theta]$, d.h. $\mathbb{P}_\theta = \mathcal{U}([0,\theta])$ und $\Theta = (0,\infty)$. In diesem Fall ist die Likelihoodfunktion gegeben durch

$$L(x_1,\dots,x_n;\theta) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(x_i).$$

Die Likelihoodfunktion ist in θ weder stetig noch differenzierbar (beachte den Indikator). Nichtsdestotrotz kann man den Maximierer $\hat\theta_{\mathrm{ML}}$ explizit bestimmen: Zunächst ist klar, dass $\hat\theta_{\mathrm{ML}} \geq \max_{i=1,\dots,n} x_i$, denn andernfalls wäre einer der Faktoren $\mathbb{1}_{[0,\theta]}(x_i)$ gleich null und entsprechend die ganze Likelihoodfunktion. Nun gilt für jedes $\theta \geq \max_{i=1,\dots,n}$, dass $L(x_1,\dots,x_n;\theta) = \theta^{-n}$, was eine monoton fallende Funktion in θ ist. Entsprechend erhalten wir

$$\hat\theta_{\mathrm{ML}}(X) = \max_{i=1,\dots,n} X_i.$$

Bemerkung. Die bisherigen Beispiel von ML-Schätzern sind insofern eher untypisch, als dass sich der ML-Schätzer bzw. das Maximum der Likelihoodfunktion explizit berechnen lässt. In vielen Fällen ist das *nicht* der Fall, hier muss der ML-Schätzer numerisch berechnet werden. Dennoch ist es eine große Stärke der ML-Methode, dass sie auch für komplizierte Likelihoodfunktion bzw. Parametermengen Θ anwendbar ist, da man bis auf Kenntnis der (Zähl-)Dichten a priori keine weiteren Annahmen an das Modell benötigt.

Satz 2.21 (**Jensen'sche Ungleichung**). *Es sei $(\Omega,\mathcal{A},\mathbb{P})$ ein Wahrscheinlichkeitsraum, $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ eine konvexe Funktion. Ist X eine reellwertige Zufallsvariable mit $\mathbb{E}[|\varphi(X)|] < \infty$, so gilt*

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

Beweis. Blatt 2, Aufgabe 3. □

Satz 2.22 (**Konsistenz von ML-Schätzern für endliche Parametermenge**). *Es sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein n-faches Produktmodell mit endlicher Parametermenge $\Theta \subset \mathbb{R}$. Weiter gelte $\mathbb{E}_\theta[(\log(f_{\theta'}(X_1)))^2] < \infty$ für alle $\theta,\theta' \in \Theta$, wobei f_θ die Zähl- bzw. Riemanndichte von X_1 unter $\mathbb{P}_\theta$ bezeichne. Dann ist die Folge $(\hat\theta_{\mathrm{ML}})_{n\in\mathbb{N}}$ von Maximum-Likelihood-Schätzern für θ konsistent.*

Beweis. Es sei θ_0 der wahre Parameter. Dann maximiert der ML-Schätzer $\hat\theta_{\mathrm{ML}}$ neben der Likelihoodfunktion auch die Funktion

$$\ell_n(\theta) := \frac{1}{n}\log\left(L(X_1,\dots,X_n;\theta)\right) - \frac{1}{n}\log\left(L(X_1,\dots,X_n;\theta_0)\right) = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right).$$

Aufgrund der Momentenannahme für $\log(f_\theta(X_1))$ gilt nach dem schwachen Gesetz der großen Zahlen, dass

$$\frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right) \longrightarrow_{\mathbb{P}_{\theta_0}} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right)\right].$$

Da $x \mapsto \log(x)$ konkav ist (d.h. $x \mapsto -\log(x)$ ist konvex), folgt aus der Jensen'schen Ungleichung im Falle stetiger Zufallsvariablen

$$\mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right)\right] \leq \log \mathbb{E}_{\theta_0}\left[\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right] = \log \int_{\mathbb{R}} \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x)dx = \log \int_{\mathbb{R}} f_\theta(x)dx = 0,$$

und analog für diskrete Zufallsvariablen. Da $x \mapsto \log(x)$ sogar strikt konkav ist, gilt Gleichheit lediglich für $\theta = \theta_0$, d.h. $\delta := \min_{\theta \neq \theta_0} \mathbb{E}_{\theta_0}[-\ell_n(\theta)] > 0$. Wählen wir nun $0 < \epsilon < \min\{|\theta - \theta_0| : \theta \in \Theta \setminus \{\theta_0\}\}$ folgt damit aus $\ell_n(\theta_0) = 0$ dass

$$\begin{aligned}
\mathbb{P}_{\theta_0}\left(\left|\hat{\theta}_{\mathrm{ML}} - \theta_0\right| > \epsilon\right) &\leq \mathbb{P}_{\theta_0}\left(\bigcup_{\theta \neq \theta_0} \{\ell_n(\theta) \geq 0\}\right) \\
&\leq \sum_{\theta \neq \theta_0} \mathbb{P}_{\theta_0}\left(\ell_n(\theta) \geq 0\right) \\
&\leq \sum_{\theta \neq \theta_0} \mathbb{P}_{\theta_0}\left(\ell_n(\theta) - \mathbb{E}_{\theta_0}[\ell_n(\theta)] \geq -\mathbb{E}_{\theta_0}[\ell_n(\theta)]\right) \\
&\leq \sum_{\theta \neq \theta_0} \mathbb{P}_{\theta_0}\left(|\ell_n(\theta) - \mathbb{E}_{\theta_0}[\ell_n(\theta)]| \geq \delta\right) \longrightarrow 0.
\end{aligned}$$

□

Bemerkung. Der Beweis von Satz 2.22 hat entscheidend die Endlichkeit von Θ genutzt, um sicherzustellen, dass $\delta > 0$. Außerdem ging die Endlichkeit in die letzte Rechnung in Form der endlichen Additivität ein. Dieser letzte Schritt ließe sich leicht auch für abzählbar unendliche Mengen durchführen, jedoch nicht für überabzählbare (und häufig ist die Parametermenge eine solche). Letzteren Fall kann man z.B. unter Stetigkeitsannahmen der Likelihoodfunktion im Parameter θ jedoch auf den abzählbaren Fall zurückführen.
Grundsätzlich gilt, dass sich beide oben genannten Probleme durch zusätzliche Regularitätsannahmen umlaufen lassen. Die Theorie der Konsistenz für Maximum-Likelihood-Schätzer ist sogar in einem sehr allgemeinen Rahmen entwickelt (→ weiterführende Vorlesung Statistik).

Bemerkung. Neben der Konsistenz lässt sich auch die Verteilungskonvergenz von ML-Schätzern für sehr viele Modelle bestimmen. Insbesondere existiert eine Menge an 'typischen' Regularitätsvoraussetzungen, unter welchen man zeigen kann, dass der ML-Schätzer asymptotisch normalverteilt ist (vgl. Satz 2.14 für Momentenschätzer).

2.4 Mittlerer quadratischer Fehler

Wie schon anhand des Beispiels in Abschnitt 1.1 gesehen, sagt weder die Konsistenz noch Erwartungstreue eines Schätzers etwas über seine Verteilung aus. Wenn wir das datengenerierende Experiment wiederholen würde, würden wir (leicht) andere Daten und somit auch leicht andere Werte für den Parameterschätzer erhalten. Das bedeutet insbesondere, dass jeder konkrete Schätzwert eine limitierte Aussagekraft hat und ein natürliches Interesse besteht, die Genauigkeit des Schätzers zu quantifizieren.

Definition 2.23 (Mittlerer quadratischer Fehler). *Es sei $\mathcal{E} = (\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein statistisches Modell mit $\Theta = \Theta' \subset \mathbb{R}^k$ und $g : \Theta \longrightarrow \Theta'$, sowie T ein Schätzer für $g(\theta)$. Dann definieren wir den* mittleren quadratischen Fehler *als*

$$\mathrm{MSE}_\theta(T) := \mathbb{E}_\theta \left[\|T(X) - g(\theta)\|_2^2 \right],$$

wobei $\|\cdot\|_2$ die Euklidische Norm im $\mathbb{R}^k$ bezeichne.

Bemerkung.

(1) Für $l(x, y) = \|x - y\|_2^2$ können wir $\mathrm{MSE}_\theta(T)$ als Abbildung $R_T : \Theta \longrightarrow \mathbb{R}$ auffassen, die gegeben ist durch

$$R(\theta) := \mathbb{E}_\theta \left[l(T(X), g(\theta)) \right].$$

Man nennt diese Funktion R_T auch Risikofunktion.

(2) Die Funktion l in Teil (1) wird auch *Verlustfunktion* genannt. Wir haben hier die spezielle Wahl des quadrierten Euklidischen Abstands getroffen. Aber auch andere Verlustfunktionen werden in der Literatur benutzt, z.B. der L^1-Verlust $l(x, y) = \|x - y\|_1$.

(3) Gilt $k = 1$, so ist

$$\mathrm{MSE}_\theta(T) = \mathbb{E}_\theta \left[(T(X) - g(\theta))^2 \right].$$

Beispiel 2.24. Wir betrachten das n-fache Produktmodell mit Bernoulliverteilungen und die beiden Schätzer $\hat{p}_1(X) = X_1$ und $\hat{p}_2(X) = \overline{X}_n$. Beide Schätzer sind erwartungstreu, sodass

$$\mathrm{MSE}_p(\hat{p}_1) = \mathbb{E}_p \left[(\hat{p}_1 - p)^2 \right] = \mathrm{Var}_p(X_1) = p(1-p)$$

und

$$\mathrm{MSE}_p(\hat{p}_2) = \mathbb{E}_p \left[(\hat{p}_2 - p)^2 \right] = \mathrm{Var}_p(\overline{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}_p(X_i) = \frac{1}{n} \mathrm{Var}_p(X_1) = \frac{1}{n} p(1-p).$$

Wie erwartet hat der Schätzer $\hat{p}_2$ einen deutlich geringeren quadratischen Fehler (falls $n \neq 1$). Insbesondere verringert sich dieser mit wachsendem n. Man beachte ebenfalls, dass der MSE für verschiedenen Schätzer miteinander vergleichen werden kann.

Definition 2.25 (Bias). *Es sei $\mathcal{E} = (\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein statistisches Modell mit $\Theta = \Theta' \subset \mathbb{R}$ und $g : \Theta \longrightarrow \Theta'$, sowie T ein Schätzer für $g(\theta)$. Dann definieren wir den* Bias *als*

$$\mathrm{Bias}_\theta(T) := \mathbb{E}_\theta \left[T(X) \right] - g(\theta).$$

Bemerkung. Für einen erwartungstreuen Schätzer T gilt per Definition $\mathrm{Bias}_\theta(T) = 0$ für alle $\theta \in \Theta$.

Lemma 2.26 (Bias-Varianz-Zerlegung). *Sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein statistisches Modell mit $\Theta = \Theta' \subset \mathbb{R}$ und $g : \Theta \longrightarrow \Theta'$, sowie T ein Schätzer für $g(\theta)$. Dann gilt für alle $\theta \in \Theta$, dass*

$$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + \mathrm{Bias}_\theta(T)^2.$$

Beweis. Mit der Linearität des Erwartungswertes erhalten wir

$$\begin{aligned}
\mathrm{MSE}_\theta(T) &= \mathbb{E}_\theta\left[(T(X) - g(\theta))^2\right] \\
&= \mathbb{E}_\theta\left[(T(X) - \mathbb{E}_\theta[T(X)] + \mathbb{E}_\theta[T(X)] - g(\theta))^2\right] \\
&= \mathbb{E}_\theta\left[(T(X) - \mathbb{E}_\theta[T(X)])^2\right] + 2\mathbb{E}_\theta\left[(T(X) - \mathbb{E}_\theta[T(X)])(\mathbb{E}_\theta[T(X)] - g(\theta))\right] \\
&\qquad + \mathbb{E}_\theta\left[(\mathbb{E}_\theta[T(X)] - g(\theta))^2\right] \\
&= \mathbb{E}_\theta\left[(T(X) - \mathbb{E}_\theta[T(X)])^2\right] + (\mathbb{E}_\theta[T(X)] - g(\theta))^2 \\
&= \mathrm{Var}_\theta(T) + \mathrm{Bias}_\theta(T)^2.
\end{aligned}$$

□

Lemma 2.27 (MSE und Konsistenz). *Es sei $\mathcal{E} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^{\otimes n} : \theta \in \Theta\})$ ein statistisches Modell mit $\Theta = \Theta' \subset \mathbb{R}$ und $g : \Theta \longrightarrow \Theta'$, sowie $(T_n)_{n\in\mathbb{N}}$ ein Folge von Schätzer für $g(\theta)$. Gilt $\mathrm{MSE}_\theta(T_n) \longrightarrow 0$, so ist die Schätzfolge $(T_n)_{n\in\mathbb{N}}$ konsistent.*

Beweis. Dies folgt direkt aus der Chebychev-Ungleichung, denn für $\epsilon > 0$ gilt

$$\mathbb{P}_\theta\left(|T_n(X) - g(\theta)| > \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}\left[(T_n(X) - g(\theta))^2\right] = \frac{1}{\epsilon^2}\mathrm{MSE}_\theta(T_n) \longrightarrow 0.$$

□

Beispiel 2.28. Wir betrachten das n-fache Produktmodell mit Normalverteilungen, d.h. $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)$, $\theta \in \mathbb{R} \times (0, \infty)$ und den empirischen Mittelwert $\overline{X}_n$ als Schätzer für μ. Nach Satz 2.9 ist $\overline{X}_n$ erwartungstreu und es gilt somit

$$\begin{aligned}
\mathrm{MSE}_{(\mu,\sigma^2)}\left(\overline{X}_n\right) &= \mathrm{Var}_{(\mu,\sigma^2)}\left(\overline{X}_n\right) \\
&= \mathbb{E}_{(\mu,\sigma^2)}\left[\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu)\right)^2\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n \mathbb{E}_{(\mu,\sigma^2)}\left[(X_i - \mu)^2\right] + \sum_{i\neq j} \mathbb{E}_{(\mu,\sigma^2)}\left[(X_i - \mu)(X_j - \mu)\right]\right) = \frac{\sigma^2}{n}.
\end{aligned}$$

So erhalten wir zum einen die Varianz des Schätzer $\overline{X}_n$ und zum anderen folgt aus Lemma 2.27 die Konsistenz.

2.5 Die Cramér-Rao-Ungleichung

Im Falle eines erwartungstreuen Schätzer bedeutet nach der Bias-Varianz-Zerlegung, dass der mittlere quadratische Fehler eines Schätzers T dann klein ist, wenn die Varianz von T klein ist. Insbesondere können wir ein weiteres Gütekriterium für einen (erwartungstreuen) Schätzer formulieren, nämlich dass er eine möglichst kleine Varianz besitzen soll. In Aufgabe 2 von Blatt 2 haben wir bereits verschiedene Schätzer für den gleichen Parameter kennengelernt, deren Varianzen sich unterscheiden. Hat man eine möglichst kleine Varianz als Ziel vor Augen, stellt sich natürlich unmittelbar die Frage, was die kleinste zu erreichende Varianz für einen Schätzer in einem gegeben Modell ist. Das nächste Resultat liefert hierauf eine Antwort.

Theorem 2.29 (Satz von Cramér-Rao). *Es sei $\mathcal{E} = (\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d), \{\mathbb{P}_\theta : \theta \in \Theta\})$ ein statistisches Modell mit $\Theta \subset \mathbb{R}$ offen. Es bezeichne f_θ die Zähl- oder Riemanndichte von $\mathbb{P}_\theta$ und es gelte:*

(i) Die Menge $M_f = \{x \in \mathbb{R}^d : f_\theta(x) > 0\}$ ist unabhängig von $\theta \in \Theta$ und für jedes $x \in M_f$ existiere $\frac{\partial}{\partial\theta} \log(f_\theta(x))$.

(ii) $\mathbb{E}_\theta \left[\frac{\partial}{\partial\theta} \log(f_\theta(X))\right] = 0$.

(iii) $0 < I_d(\theta) := \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial\theta} \log(f_\theta(X))\right)^2\right] < \infty$.

Es sei weiter $g : \Theta \longrightarrow \Theta' \subset \mathbb{R}$ und T ein Schätzer für $g(\theta)$ mit $\mathbb{E}_\theta[|T(X)|] < \infty$ für alle $\theta \in \Theta$ und der Eigenschaft

(iv) $\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)] = \mathbb{E}_\theta \left[T(X)\frac{\partial}{\partial\theta} \log(f_\theta(X))\right]$.

Dann gilt

$$\mathrm{Var}_\theta(T(X)) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)]\right)^2}{I_d(\theta)} \qquad \text{für alle } \theta \in \Theta.$$

Insbesondere gilt für einen erwartungstreuen Schätzer T

$$\mathrm{Var}_\theta(T(X)) \geq \frac{(g'(\theta))^2}{I_d(\theta)} \qquad \text{für alle } \theta \in \Theta,$$

wobei der Zähler Eins ist, falls $\Theta' = \Theta$ und $g = \mathrm{id}_\Theta$, d.h. falls der Parameter selbst geschätzt wird.

Beweis. Mit den Voraussetzungen (ii),(iv) und der Cauchy-Schwarz-Ungleichung für den Erwartungswert (vgl. Satz I.3.26 oder Stochastik I, Blatt 6, A2(b)) erhalten wir

$$\begin{aligned}
\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)]\right)^2 &= \mathbb{E}_\theta \left[T(X)\frac{\partial}{\partial\theta} \log\left(f_\theta(X)\right)\right]^2 \\
&= \left[\mathrm{Cov}_\theta \left(T(X), \frac{\partial}{\partial\theta} \log\left(f_\theta(X)\right)\right)\right]^2 \\
&\leq \mathrm{Var}_\theta\left(T(X)\right) \mathrm{Var}_\theta \left(\frac{\partial}{\partial\theta} \log\left(f_\theta(X)\right)\right).
\end{aligned}$$

Nach Voraussetzung (ii) erhalten wir

$$\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log\left(f_\theta(X)\right)\right) = I_d(\theta),$$

sodass die erste Aussage des Satzes nach Umstellen der obigen Ungleichung folgt. Die zweite Aussage folgt unmittelbar, da wegen der Erwartungstreue $\mathbb{E}_\theta[T(X)]) = g(\theta)$. □

Bemerkung.

(1) Die Größe $I_d(\theta)$ in Theorem 2.29 heißt Fisher-Information.

(2) Die Regularitätsvoraussetzungen von Theorem 2.29 bedeuten eigentlich lediglich, dass sich Differentiation und Erwartungswertbildung vertauschen lassen. Im stetigen Fall folgt (ii) aus der Annahme der Vertauschbarkeit von Differentiation und Integration, denn dann gilt

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log\left(f_\theta(X)\right)\right] &= \int_{\mathbb{R}^d}\frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(x)}f_\theta(x)dx\\ &= \int_{\mathbb{R}^d}\frac{\partial}{\partial\theta}f_\theta(x)dx = \frac{\partial}{\partial\theta}\int_{\mathbb{R}^d}f_\theta(x)dx = \frac{\partial}{\partial\theta}1 = 0.\end{aligned}$$

(3) Häufig besteht der Vektor $X = (X_1,\dots,X_d)$ aus unabhängig identisch verteilten Zufallsvariablen $X_1,\dots,X_d$ mit (Zähl-)Dichte $f_\theta^{(1)}$. In diesem Fall gilt

$$f_\theta(x_1,\dots,x_d) = \prod_{i=1}^d f_\theta^{(1)}(x_i)$$

und somit

$$I_d(\theta) = \mathbb{E}_\theta\left[\left(\sum_{i=1}^d\frac{\partial}{\partial\theta}\log\left(f_\theta^{(1)}(X_i)\right)\right)^2\right] = \sum_{i=1}^d\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\left(f_\theta^{(1)}(X_i)\right)\right)^2\right] = dI_1(\theta).$$

Beispiel 2.30. Wir betrachten das n-fache Produktmodell mit Normalverteilungen, d.h. $\mathbb{P}_\theta = \mathcal{N}(\mu,\sigma^2)$, $\theta\in\mathbb{R}\times(0,\infty)$. Für den Parameter μ erhalten wir

$$\frac{\partial}{\partial\mu}\log\left(f_{(\mu,\sigma^2)}(x)\right) = \frac{\partial}{\partial\mu}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) = \frac{x-\mu}{\sigma^2}.$$

Mit (3) aus der vorangegangen Bemerkung erhalten wir

$$I_n(\mu) = n\mathbb{E}_{(\mu,\sigma^2)}\left[\left(\frac{\partial}{\partial\mu}\log\left(f_{(\mu,\sigma^2)}(X_1)\right)\right)^2\right] = \frac{n}{\sigma^4}\mathbb{E}_{(\mu,\sigma^2)}\left[(X_1-\mu)^2\right] = \frac{n}{\sigma^2}.$$

Damit hat jeder erwartungstreue Schätzer T_μ für den Parameter μ nach Theorem 2.29 bestenfalls die Varianz

$$\mathrm{Var}_{(\mu,\sigma^2)}\left(T_\mu(X)\right) \geq \frac{\sigma^2}{n}.$$

Nach Beispiel 2.28 wird die untere Schranke für den empirischen Mittelwert $T_\mu(X) = \overline{X}_n$ angenommen.

Bemerkung (Exponentialfamilien). Wie im vorangegangen Beispiel gesehen, gibt es Fälle, in welchen Gleichheit in der Cramér-Rao-Ungleichung gilt. Es stellt sich entsprechend die Frage, ob man charakterisieren kann, wann dies der Fall ist. Grundlegend für den Beweis der Ungleichung war die Cauchy-Schwarz-Ungleichung

$$\left[\mathrm{Cov}_\theta\left(T(X), \frac{\partial}{\partial\theta}\log\left(f_\theta(X)\right)\right)\right]^2 \leq \mathrm{Var}_\theta\left(T(X)\right)\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log\left(f_\theta(X)\right)\right).$$

Man überzeugt sich leicht davon, dass hier Gleichheit gilt, falls für jedes $\theta \in \Theta$ Konstanten $a_1(\theta)$ und $a_2(\theta)$ existieren sodass

$$\mathbb{P}_\theta\left(\frac{\partial}{\partial\theta}\log\left(f_\theta(X)\right) = a_1(\theta)T(X) + a_2(\theta)\right) = 1.$$

Integration nach θ liefert dann

$$\mathbb{P}_\theta\Big(f_\theta(X) = \exp\big(A_1(\theta)T(X) + A_2(\theta) + S(X)\big)\Big) = 1,$$

wobei A_1 und A_2 die Stammfunktionen von a_1 bzw. a_2 bezeichnen. Verteilungsfamilien, deren Zähl- oder Riemanndichten von ebendieser Form

$$f_\theta(x) = \exp\big(A_1(\theta)T(x) + A_2(\theta) + S(x)\big)$$

sind, bezeichnet man als Exponentialfamilien. Für diese lassen sich einerseits eine ganze Reihe interessanter Resultate herleiten, zum anderen sind viele uns bereits bekannte Verteilungen (z.B. die Binomial, Exponential, Poisson- und Normalverteilung) von dieser Gestalt. → Mathematische Statistik.

Bemerkung (ML-Schätzer). Es seien $X_1, \ldots, X_n$ unabhängig identisch nach $\mathbb{P}_\theta$ verteilt, wobei $\theta \in \Theta \subset \mathbb{R}$ offen und X_1 eine Zähl- bzw. Riemanndichte $f_\theta^{(1)}$ besitzt, welche die Voraussetzungen von Satz 2.29 erfüllt. Weiter sei für jedes $x \in M_f$ die Funktion

$$\theta \mapsto \log\left(f_\theta^{(1)}(x)\right)$$

zweimal stetig differenzierbar und es gelte die stochastische Konvergenz

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log\left(f_{\theta_n^*}(X_i)\right) \longrightarrow_{\mathbb{P}_\theta} -\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log\left(f_\theta^{(1)}(X_1)\right)\right)^2\right]$$

für jeden konsistenten Schätzer $\theta_n^* \longrightarrow_{\mathbb{P}_\theta} \theta$. Dann gilt mit $\mathbb{P}_\theta$-Wahrscheinlichkeit, die gegen 1 konvergiert:

- Es existiert eine Lösung $\hat{\theta}_n = \hat{\theta}_n(X)$ der Maximum-Likelihood-Gleichung

$$\frac{\partial}{\partial\theta}\log\left(f_\theta(X)\right) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log\left(f_\theta^{(1)}(X_i)\right) = 0.$$

- Im Punkt $\hat{\theta}_n$ nimmt die Funktion $\theta \mapsto \log\left(f_\theta(X)\right)$ ein lokales Maximum an und $\hat{\theta}_n$ ist ein konistenter Schätzer für den Parameter θ.

Gelten die obigen Voraussetzungen kann man weiter zeigen, dass

$$\sqrt{n}\left(\hat{\theta}_n(X) - \theta\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, I_1(\theta)^{-1}\right),$$

wobei $I(\theta)$ die Fisher-Information fon $f_\theta^{(1)}$ bezeichne (vgl. Aufgabenblatt 4). Somit nimmt der ML-Schätzer asymptotisch die nach der Cramér-Rao-Ungleichung minimale Varianz an. Man spricht auch davon, dass der ML-Schätzer (unter obigen Voraussetzungen) ASYMPTOTISCH EFFIZIENT ist.

3 Konfidenzbereiche

In diesem Kapitel wenden wir uns den Bereichsschätzern oder auch Konfidenzintervallen zu, welche uns bereits im einführenden Beispiel in Abschnitt 1.1 begegnet sind.

3.1 Konstruktionsprinzip und Quantile

Definition 3.1 (Konfidenzbereich). *Es sei $\mathcal{E} = (\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ ein statistisches Modell, Θ' eine Menge und $g : \Theta \longrightarrow \Theta'$ eine Funktion. Weiter sei $\mathcal{S} \subset \mathcal{P}(\Theta')$ eine Familie von Teilmengen von (oder eine σ-Algebra auf) Θ', sowie $\alpha \in (0,1)$. Dann heißt jede Abbildung $C : \mathcal{X} \longrightarrow \mathcal{S}$* KONFIDENZBEREICH ZUM (KONFIDENZ-)NIVEAU $1-\alpha$ FÜR $g(\theta)$, *falls für alle $\theta \in \Theta$ gilt*

$$\mathbb{P}_\theta\left(g(\theta) \in C(X)\right) = \mathbb{P}_\theta\left(\{x \in \mathcal{X} :\ g(\theta) \in C(x)\}\right) \geq 1 - \alpha.$$

Bemerkung.

(1) Ist $x = (x_1, \ldots, x_n) \in \mathcal{X}$ eine Realisierung des Zufallsexperiments, so wird häufig auch $C(x)$ als Konfidenzbereich bezeichnet.

(2) Man beachte, dass die Menge $\{x \in \mathcal{X} :\ g(\theta) \in C(x)\}$ messbar sein muss, d.h. dieses Ereignis muss in der σ-Algebra liegen, auf welcher $\mathbb{P}_\theta$ definiert ist.

(3) Die Wahl $c(x) = \Theta'$ liefert einen Konfidenzbereich für jedes Niveau $1-\alpha$, beinhaltet allerdings auch keine Infomration über die Lage von $g(\theta)$. Generell ist es wünschenswert, ein möglichst großes Konfidenzniveau $1-\alpha$ (d.h. kleines α) und gleichzeitig einen möglichst kleinen Konfidenzbereicht $C(X)$ zu finden. Eine solche simultane Optimierung ist allerdings nicht möglich, wie das Eingangsbeispiel bereits andeutet. Vielmehr verfährt man in der Praxis so, dass man sich $\alpha \in (0,1)$ vorgibt und für dieses feste α einen möglichst kleinen Konfidenzbereich konstruiert.

Bemerkung (Konfidenzintervalle). Ist $\Theta' = \mathbb{R}$, so wählt man als Mengensystem meist

$$\mathcal{S} = \{[a,b], [a,\infty), (-\infty, b] :\ a, b \in \mathbb{R}\},$$

das heißt die Menge aller abgeschlossenen bzw. halboffenen Intervalle. In diesem Fall gilt $C(X) = [a(X), b(X)]$ bzw. $C(X) = [a(X), \infty)$ oder $C(X) = (-\infty, b(X)]$ für Funktionen $a, b : \mathcal{X} \longrightarrow \mathbb{R}$. Man spricht in diesem Fall von KONFIDENZINTERVALLEN ZUM NIVEAU $1-\alpha$, wenn für alle $\theta \in \Theta$ gilt

$$\mathbb{P}_\theta\left(a(X) \leq g(\theta) \leq b(X)\right) \geq 1 - \alpha$$

(und analog für die halboffenen Intervalle).

Bemerkung (Interpretation). Man beachte die folgende Interpretation des Begriffs Konfidenzbereich: Die zufällige Menge $C(X)$ enthält den unbekannten Parameter θ mit Wahrscheinlichkeit $\geq 1-\alpha$. Das bedeutet nicht, dass für die Daten $x \in \mathcal{X}$ die Menge $C(x)$ den wahren Parameter θ mit Wahrscheinlichkeit $1-\alpha$ enthält, denn diese Aussage ist entweder richtig oder falsch.
Um uns einer Antwort auf die Frage zu nähern, wie man Konfidenzbereiche oder -intervalle bestimmen kann, starten wir mit einem Beispiel.

Beispiel 3.2 (**Konfidenzbereich für μ im Normalverteilungsmodell bei bekannter Varianz**). Wir betrachten das n-fache Produktmodell mit Normalverteilungen mit bekannter Varianz, d.h. $\mathbb{P}_\mu = \mathcal{N}(\mu, \sigma^2)$ für vorgegebenes $\sigma^2 > 0$, und interessieren uns für ein Konfidenzintervall zum Niveau $1-\alpha$ für den Parameter μ. Man überlegt sich leicht, dass in diesem Modell

$$\overline{X}_n - \mu \sim \frac{\sigma}{\sqrt{n}} Z$$

für $Z \sim \mathcal{N}(0,1)$. Wir wählen nun $a, b \in \mathbb{R}$ so, dass $\mathbb{P}(a \leq Z \leq b) = 1-\alpha$. Dann gilt

$$1-\alpha = \mathbb{P}_\mu\left(a \leq \frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) \leq b\right) = \mathbb{P}\mu\left(\overline{X}_n - b\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n - a\frac{\sigma}{\sqrt{n}}\right)$$

und ein mögliches Konfidenzintervall zum Niveau $1-\alpha$ ist gegeben durch

$$C(X) = \left[\overline{X}_n - b\frac{\sigma}{\sqrt{n}}, \overline{X}_n - a\frac{\sigma}{\sqrt{n}}\right].$$

Man beachte, dass die Werte $a, b \in \mathbb{R}$ durch die Bedingung $\mathbb{P}(a \leq Z \leq b) = 1-\alpha$ nicht eindeutig bestimmt sind. Häufig wählt man sie jedoch symmetrisch in dem Sinne, dass $\mathbb{P}(Z < a) = \mathbb{P}(Z > b) = \alpha/2$.

Wir fassen die drei wesentlichen Konstruktionsschritte aus Beispiel 3.2 zusammen:

(i) Bilden einer Statistik $T(X) = T(X, \theta)$ aus den Daten $X = (X_1, \ldots, X_n)$ und dem einzugrenzenden Parameter θ. Ziel ist es hierbei, dass die Verteilung von $T(X, \theta)$ selbst nicht mehr von θ abhängt.

(ii) Bestimmung eines Intervalls $I = [a, b] \subset \mathbb{R}$ mit $\mathbb{P}_\theta(T(X, \theta) \in I) = 1-\alpha$. Da die Verteilung von $T(X, \theta)$ nicht von θ abhängt, gilt dies auch für a und b.

(iii) Auflösen der Ungleichung $a \leq T(X, \theta) \leq b$ nach dem in $T(X, \theta)$ enthaltenen Parameter θ liefert das gesuchte Konfidenzintervall.

Der schwierige Schritt besteht hierbei meistens in Punkt (i). Bevor wir diesen im Normalverteilungsmodell weiter untersuchen, widmen wir uns der Konstruktion des Intervalls in (ii). Bereits in Beispiel 3.2 hatten wir diskutiert, dass a und b meist so gewählt werden, dass $\mathbb{P}_\theta(T(X, \theta) < a) = \mathbb{P}_\theta(T(X, \theta) > b) = \alpha/2$. Dies führt auf den folgenden Begriff des Quantils.

Definition 3.3 (**Quantilfunktion und Quantile**). *Es sei F_X die Verteilungsfunktion einer (stetigen oder diskreten) Verteilung auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Dann ist die zugehörige* Quantilfunktion *$F_X^{-1} : (0,1) \longrightarrow \mathbb{R}$ definiert durch die verallgemeinerte Inverse der Verteilungsfunktion, d.h.*

$$F_X^{-1}(p) := \inf\{z \in \mathbb{R} : \ F_X(z) \geq p\}.$$

*Der Wert $F_X^{-1}(p)$ heißt p-*Quantil von F_X *(bzw. der zugehörigen Verteilung $\mathbb{P}^X$).*

Bemerkung (Eindeutigkeit der Quantile). Man beachte, dass nach der obigen Definition die p-Quantile aufgrund der Rechtsstetigkeit der Verteilungsfunktion F_X für alle $p \in (0,1)$ eindeutig bestimmt sind. Man findet in der Literatur auch den alternativen Ansatz, dass jede Zahl q_p ein p-Quantil genannt wird, wenn sie

$$\mathbb{P}(X < q_p) \leq p \leq \mathbb{P}(X \leq q_p)$$

erfüllt. Dies legt den Wert von q_p nicht eindeutig fest, falls die Verteilungsfunktion stückweise konstant ist: Ist etwas X Bernoulli-verteilt, so gilt für jede Zahl $a \in (0,1)$, dass $\mathbb{P}(X < a) = 1/2 = \mathbb{P}(X \leq a)$, sodass in diesem Fall jede solche Zahl a ein 0.5-Quantil ist. Unsere Definition über die verallgemeinerte Inverse wählt immer den kleinsten der auf diese Weise in Frage kommenden Werte aus.

Lemma 3.4 (**Eigenschaften der Quantilfunktion**). *Es sei F_X die Verteilungsfunktion einer Zufallsvariablen auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ und F_X^{-1} die zugehörige Quantilfunktion. Dann gilt:*

(a) F_X^{-1} ist monoton wachsend.

(b) Ist F_X stetig, so gilt $F_X(F_X^{-1}(p)) = p$ für alle $p \in (0,1)$.

(c) Ist F_X bijektiv, so ist F_X^{-1} die Umkehrfunktion.

(d) Besitzt die Verteilung von X eine Riemann-Dichte f, so ist $F_X^{-1}(p)$ eine Lösung der Gleichung $F_X(y) = p$ und es gilt

$$\mathbb{P}\left(X < F_X^{-1}(p)\right) = \mathbb{P}\left(X \leq F_X^{-1}(p)\right) = p.$$

Gilt zusätzlich $f(x) > 0$ für alle $x \in \mathbb{R}$, so ist F_X bijektiv und $F_X^{-1}(p)$ die eindeutige Lösung von $F_X(y) = p$.

Beweis. (a) Übungsaufgabe, Blatt 4.

(b) Sei $p \in (0,1)$. Ist $F_X : \mathbb{R} \longrightarrow (0,1)$ stetig, so existiert aufgrund des Zwischenwertsatzes ein $z_0 \in \mathbb{R}$ mit $F_X(z_0) = p$. Da F_X monoton wachsend ist muss auch für jedes z mit $z < z_0$ und $F_X(z) \geq p$ gelten, dass $F_X(z) = p$ und die Behauptung folgt aufgrund der Stetigkeit von F_X auch für das Infimum aller dieser z.

(c) Da F_X monoton wachsend ist folgt aus der Bijektivität, dass F_X auch stetig ist. Da eine eindeutige Umkehrfunktion existieren muss und nach (b) bereits $F_X(F_X^{-1}(p)) = p$ für alle $p \in (0,1)$ gilt, muss F_X^{-1} bereits diese Umkehrfunktion sein.

(d) Durch die Existenz der Riemann-Dichte ist F_X stetig und daher $F_X^{-1}(p)$ eine Lösung von $F_X(y) = p$ nach (b). Da Zufallsvariablen mit Riemann-Dichten keine Punktmassen haben gilt außerdem

$$\mathbb{P}\left(X < F_X^{-1}(p)\right) = \mathbb{P}\left(X \leq F_X^{-1}(p)\right) = F_X\left(F_X^{-1}(p)\right) = p.$$

Der letzte Teil der Aussage folgt aus (c), da $F_X' = f > 0$ und F_X somit streng monoton wächst.

□

Beispiel 3.5. Es sei $Z \sim \mathcal{N}(0,1)$ und Φ_α^{-1} das α-Quantil der Standardnormalverteilung. Dann gilt nach Lemma 3.4 für $\alpha \in (0,1)$, dass

$$\mathbb{P}\left(Z < \Phi_{\alpha/2}^{-1}\right) = \frac{\alpha}{2},$$

sowie

$$\mathbb{P}\left(Z > \Phi_{1-\alpha/2}^{-1}\right) = 1 - \mathbb{P}\left(Z \leq \Phi_{1-\alpha/2}^{-1}\right) = \frac{\alpha}{2},$$

sodass insgesamt

$$\mathbb{P}\left(\Phi_{\alpha/2}^{-1} \leq Z \leq \Phi_{1-\alpha/2}^{-1}\right) = \alpha.$$

Aufgrund der Symmetrie der Standard-Normalverteilung gilt überdies $\Phi_{\alpha/2}^{-1} = -\Phi_{1-\alpha/2}^{-1}$. Damit erhalten wir in Beispiel 3.2 das Konfidenzintervall

$$C(X) = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi_{1-\alpha/2}^{-1}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}\Phi_{1-\alpha/2}^{-1}\right].$$

Man beachte, dass dieses Intervall symmetrisch um $\overline{X}_n$ ist.

Beispiel 3.6.

(1) Für die Exponentialverteilung mit Parameter $\lambda > 0$ erhalten wir die Verteilungsfunktion $F(t) = (1 - e^{-\lambda t})\mathbb{1}_{[0,\infty)}(t)$. Diese ist stetig, sodass wir das α-Quantil t_α nach Lemma 3.4(b) als Lösung der Gleichung $1 - e^{-\lambda t_\alpha} = \alpha$ erhalten, also

$$t_\alpha = -\frac{\log(1-\alpha)}{\lambda}.$$

(2) Im wichtigen Fall der Normalverteilung existiert keine geschlossene Form, da sich die Gleichung

$$\int_{-\infty}^{t_\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dy = \alpha$$

nicht explizit nach t_α auflösen lässt. Die Werte der Quantile der Normalverteilung müssen entsprechend entweder in sogenannten *Quantiltabellen* nachgeschlagen werden oder können in jeder Statistiksoftware über entsprechende Befehle abgerufen werden.

3.2 Konfidenzintervalle für normalverteilte Daten

In diesem Abschnitt wollen wir Konfidenzintervalle für μ und σ^2 im n-fachen Produktmodell mit Normalverteilungen $\mathcal{N}(\mu, \sigma^2)$ bestimmen. Dies ist schwieriger als in Beispiel 3.2, da wir σ^2 nicht mehr als bekannt voraussetzen wollen. Dort hatten wir ausgenutzt, dass

$$\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1).$$

Falls die Varianz σ^2 unbekannt ist, liegt es nahe, diese in der Statistik auf der linken Seite durch ihren Schätzer $s_n^2(X)$ zu ersetzen, d.h. die Größe

$$T_n = \frac{\overline{X}_n - \mu}{\sqrt{s_n^2(X)/n}}$$

zu betrachten. Es ist zu erwarten, dass sich dadurch die Verteilung ändert. Wir werden diese zunächst bestimmen und dabei die bemerkenswerte Feststellung treffen können, dass die Verteilung von T_n weder von μ noch von σ^2 abhängt. Darauf aufbauend können wir dann wie in Beispiel 3.2 Konfidenzintervalle bestimmen. Für die Verteilungsbestimmung von T_n benötigen wir zunächst zwei von der Normalverteilung abgeleitete Verteilungen.

Definition 3.7 (Die χ^2- und die t-Verteilung). *Es seien $X, X_1, \ldots, X_n$ unabhängige, standardnormalverteilte Zufallsvariablen.*

(a) Die Verteilung der Zufallsvariablen

$$X_1^2 + \cdots + X_n^2$$

heißt χ^2-Verteilung mit n Freiheitsgraden und wird mit χ_n^2 bezeichnet.

(b) Die Verteilung von

$$\frac{X}{\sqrt{(X_1^2 + \cdots + X_n^2)/n}}$$

heißt t-Verteilung mit n Freiheitsgraden und wird mit t_n bezeichnet.

Bemerkung (Dichten der t- und χ^2-Verteilung).

(1) Die χ_n^2-Verteilung besitzt eine stetige Riemann-Dichte mit Träger $(0, \infty)$.

(2) Die t_n-Verteilung besitzt ebenfalls eine stetige Riemann-Dichte, die auf ganz $\mathbb{R}$ strikt positiv ist. Weiter ist sie symmetrisch um 0. Bezeichnet $t_{n,\alpha}$, das α-Quantil, so folgt entsprechend wie bei der Normalverteilung, dass $t_{n,\alpha/2} = -t_{n,1-\alpha/2}$.

Wir erinnern an dieser Stelle an die mehrdimensionale Normalverteilung aus Definition 2.10. Sind $X_1, \ldots, X_d$ stochastisch unabhängige $\mathcal{N}(0,1)$-verteilte Zufallsvariablen, so erhalten wir die Dichte des daraus aufgebauten d-dimensionalen Zufallsvektors $X = (X_1, \ldots, X_n)$ als Produkt der Marginaldichten (vgl. Definition $I.4.22$). Damit erhält man sofort, dass $X \sim \mathcal{N}_d(0, I_d)$ d-dimensional normalverteilt ist mit Erwartungswert 0 und d-dimensionaler Einheitsmatrix I_d als Kovarianzmatrix.

Lemma 3.8 (**Orthogonale Transformation von normalverteilten Vektoren**). *Es seien $X_1, \dots, X_d$ stochastisch unabhängige $\mathcal{N}(0,1)$-verteilte Zufallsvariablen und $X = (X_1, \dots, X_d)$ der entsprechende d-dimensionale Zufallsvektor, sowie $O \in \mathbb{R}^{d\times d}$ eine orthogonale Matrix (d.h. $O^T O = I_d$). Dann gilt $Z = (Z_1, \dots, Z_d)^T = OX \sim \mathcal{N}_d(0, I_d)$, insbesondere sind $Z_1, \dots, Z_d$ unabhängige $\mathcal{N}(0,1)$-verteilte Zufallsvariablen.*

Beweis. Es sei $U \subset \mathbb{R}^d$ eine offene Menge, sowie $O^T(U)$ das Bild dieser Menge unter der Abbildung O^T, d.h. $O^T(U) := \{x \in \mathbb{R}^d : \exists z \in U \text{ mit } x = O^T z\}$. Weiter bezeichne f_{0,I_d} die Dichte von $\mathcal{N}_d(0, I_d)$. Dann gilt nach der mehrdimensionalen Transformationsformel

$$\mathbb{P}(OX \in U) = \mathbb{P}\left(X \in O^T(U)\right) = \int_{O^T(U)} f_{0,I_d}(x)dx = \int_U f_{0,I_d}(Ox)|\det(O)|dx.$$

Da O orthogonal, gilt $|\det(O)| = 1$. Damit erhalten wir insbesondere

$$f_{0,I_d}(Ox) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(Ox)^T I_d(Ox)\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}x^T O^T Ox\right) = f_{0,I_d}(x)$$

für alle $x \in \mathbb{R}^d$ und somit die Behauptung, da

$$\mathbb{P}(Z \in U) = \mathbb{P}(OX \in U) = \int_U f_{0,I_d}(x)dx.$$

Der Zusatz über die Unabhängigkeit folgt, da die Dichte $f_{0,I_d}(x_1, \dots, x_n) = \prod_{i=1}^n f_{0,I_1}(x_i)$ faktorisiert (nachrechnen!), vgl. Definition I.4.22 und Bemerkung I.4.23. □

Theorem 3.9 (**Satz von Fisher**). *Sei $n > 1$, $X = (X_1, \dots, X_n)$ ein Vektor von unabhängig identisch $\mathcal{N}(\mu, \sigma^2)$-verteilten Zufallsvariablen, sowie $\overline{X}_n$ der empirische Mittelwert und $s_n^2(X)$ die empirische Varianz. Dann gilt*

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$$

und

$$\frac{n-1}{\sigma^2} s_n^2(X) \sim \chi^2_{n-1}.$$

Außerdem sind $\overline{X}_n$ und $s_n^2(X)$ stochastisch unabhängig und

$$T_n := \frac{\overline{X}_n - \mu}{\sqrt{s_n^2(X)/n}} \sim t_{n-1}.$$

Bemerkung. Betrachtet man die Definition der t-Verteilung, ist es naheliegend, dass wir zeigen können, dass T_n t_{n-1}-verteilt ist, vorausgesetzt, dass wir zeigen können, dass $\overline{X}_n$ und $s_n^2(X)$ stochastisch unabhängig sind. Dies ist der überraschende Teil der Aussage von Theorem 3.9, denn a priori ist dies überhaupt nicht klar, da beide Größen aus den Zufallsvariablen $X_1, \dots, X_n$ zusammengesetzt sind!

Beweis von Theorem 3.9. Zunächst halten wir fest, dass $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ klar ist (vgl. Bemerkung *I*.4.14). Für die anderen Aussagen betrachten wir den Vektor $Z = (Z_1, \dots, Z_n)$ mit $Z_i := (X_i - \mu)/\sigma \sim \mathcal{N}(0,1)$. Mit den X_i sind auch $Z_1, \dots, Z_n$ unabhängig. Weiter ist

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} = \frac{\overline{X}_n - \mu}{\sigma}$$

und damit

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

Damit ist insbesondere

$$T_n = \frac{\overline{X}_n - \mu}{\sqrt{s_n^2(X)/n}} = \frac{\sigma \overline{Z}_n}{\sqrt{\sigma^2 s_n^2(Z)/n}} = \frac{\overline{Z}_n}{\sqrt{s_n^2(Z)/n}}$$

und es genügt die Behauptungen für den Vektor Z zu zeigen. Zu diesem Zweck wählen wir nun eine orthogonale Matrix $O = (o_{ij})_{1 \le i,j \le n}$ mit der Eigenschaft, dass

$$o_{11} = o_{12} = \cdots = o_{1n} = \frac{1}{\sqrt{n}}.$$

Eine solche Wahl ist möglich, da der erste Zeilenvektor Euklidische Länge 1 hat und man die anderen Zeilen durch Ergänzen dieses Vektors zu einer ONB des $\mathbb{R}^n$ erhalten kann (die Matrix O ist dadurch natürlich nicht eindeutig festgelegt). Wir setzen nun $W = OZ$ und wissen nach Lemma 3.8, dass $W = (W_1, \dots, W_n)$ ein Vektor unabhängiger identisch $\mathcal{N}(0,1)$-verteilter Zufallsvariablen ist. Weiter ist durch Wahl von O

$$W_1 = o_{11} Z_1 + \cdots + o_{1n} Z_n = \sqrt{n}\ \overline{Z}_n$$

und

$$(n-1)s_n^2(Z) = \sum_{i=1}^{n} \left(Z_i - \overline{Z}_n \right)^2 = \sum_{i=1}^{n} Z_i^2 - n\overline{Z}_n^2 = \sum_{i=1}^{n} Z_i^2 - W_1^2 = \sum_{i=2}^{n} W_i^2,$$

wobei wir im letzten Schritt verwendet haben, dass die Euklidische Länge unter orthogonaler Transformation erhalten bleibt, d.h.

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (OZ)_i^2 = \sum_{i=1}^{n} W_i^2.$$

Insbesondere haben wir gezeigt, dass $(n-1)s_n^2(Z)$ die Summe von $n-1$ quadrierten unabhängig identisch $\mathcal{N}(0,1)$-verteilten Zufallsvariablen ist und somit nach Definition χ^2_{n-1}-verteilt. Da W_1 unabhängig von $(W_2, \dots, W_n)$ ist folgt weiter, dass $\overline{Z}_n$ und $s_n^2(Z)$ unabhängig sind.
Für die letzte Aussage schreiben wir

$$T_n = \frac{\overline{Z}_n}{\sqrt{s_n^2(Z)/n}} = \frac{\sqrt{n}\overline{Z}_n}{\sqrt{s_n^2(Z)}} = \frac{W_1}{\sqrt{(W_2^2 + \cdots + W_n^2)/(n-1)}}.$$

Da $W_1, W_2, \dots, W_n$ unabhängig identisch $\mathcal{N}(0,1)$-verteilt sind, folgt direkt aus der Definition der t-Verteilung, dass $T_n \sim t_{n-1}$. □

Wir kommen nun zum Hauptresultat dieses Abschnitts.

Theorem 3.10 (**Konfidenzintervall im Normalverteilungsmodell**). *Es sei $n \geq 2$ und $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \{\mathbb{P}_{\mu,\sigma^2}^{\otimes n} : (\mu, \sigma^2) \in \mathbb{R} \times (0,\infty)\})$ das n-fache Produktmodell mit $\mathcal{N}(\mu, \sigma^2)$-Verteilungen. Weiter sei $\alpha \in (0,1)$ und $t_{n,\alpha}$ und $\chi^2_{n,\alpha}$ das α-Quantil der t_n- bzw. χ^2_n-Verteilung.*

(a) Das Intervall

$$\left[\overline{X}_n - \sqrt{s_n^2(X)/n} \cdot t_{n-1,1-\alpha/2}, \overline{X}_n + \sqrt{s_n^2(X)/n} \cdot t_{n-1,1-\alpha/2}\right]$$

ist ein Konfidenzintervall für μ zum Niveau $1-\alpha$.

(b) Das Intervall

$$\left[\frac{n-1}{\chi_{n-1,1-\alpha/2}} s_n^2(X), \frac{n-1}{\chi_{n-1,\alpha/2}} s_n^2(X)\right]$$

ist ein Konfidenzintervall für σ^2 zum Niveau $1-\alpha$.

Beweis. (a) Nach dem Satz von Fisher (Satz 3.9) gilt im betrachteten Modell

$$T_n = \frac{\overline{X}_n - \mu}{\sqrt{s_n^2(X)/n}} \sim t_{n-1}.$$

Dann gilt mit Lemma 3.4(d) und der Tatsache, dass die t-Verteilung stetig ist mit um Null symmetrischer Dichte (vgl. die Bemerkung nach Definition 3.7), dass

$$\begin{aligned}
&\mathbb{P}_{\mu,\sigma^2}\left(\overline{X}_n - \sqrt{s_n^2(X)/n} \cdot t_{n-1,1-\alpha/2} \leq \mu \leq \overline{X}_n + \sqrt{s_n^2(X)/n} \cdot t_{n-1,1-\alpha/2}\right) \\
&\quad = \mathbb{P}_{\mu,\sigma^2}\left(T_n - t_{n-1,1-\alpha/2} \leq 0 \leq T_n + t_{n-1,1-\alpha/2}\right) \\
&\quad = \mathbb{P}_{\mu,\sigma^2}\left(T_n \leq t_{n-1,1-\alpha/2}\right) - \mathbb{P}_{\mu,\sigma^2}\left(T_n \leq -t_{n-1,1-\alpha/2}\right) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha.
\end{aligned}$$

(b) Wieder nach dem Satz von Fisher ist

$$\chi^2 := (n-1)s_n^2(X)/\sigma^2 \sim \chi^2_{n-1}$$

und wir erhalten

$$\begin{aligned}
&\mathbb{P}_{\mu,\sigma^2}\left(\frac{n-1}{\chi_{n-1,1-\alpha/2}} s_n^2(X) \leq \sigma^2 \leq \frac{n-1}{\chi_{n-1,\alpha/2}} s_n^2(X)\right) \\
&\quad = \mathbb{P}_{\mu,\sigma^2}\left(\frac{\chi^2}{\chi_{n-1,1-\alpha/2}} \leq 1 \leq \frac{\chi^2}{\chi_{n-1,\alpha/2}}\right) \\
&\quad = \mathbb{P}_{\mu,\sigma^2}\left(\chi^2 \leq \chi_{n-1,1-\alpha/2}\right) - \mathbb{P}_{\mu,\sigma^2}\left(\chi^2 \leq \chi_{n-1,\alpha/2}\right) = 1 - \alpha.
\end{aligned}$$

Man beachte, dass der zweite Schritt nutzt, dass die Quantilfunktion jeder Verteilung nach Lemma 3.4(a) monoton wachsend ist.

□

Beispiel 3.11. Ein Medikament werde daraufhin untersucht, ob es den Schlaf von Probanden verlängert. Dazu wird die Schlafdauerdifferenz vor und nach der Einnahme von zehn Probanden notiert und es wurden beobachtet:

$$1.0,\ 0.8,\ 1.1,\ 0.1,\ -0.1,\ 4.4,\ 5.5,\ 1.6,\ 4.6 \text{ und } 3.4.$$

Wir nehmen an, dass diese Beobachtungen Realisierungen unabhängiger Normalverteilungen $\mathcal{N}(\mu, \sigma^2)$ mit unbekanntem Mittelwert μ und unbekannter Varianz σ^2 sind. Wir erhalten

$$\overline{X}_{10} = 2.33 \quad \text{und} \quad s^2_{10}(X) = 4.01$$

und schlagen nach, dass $\chi^2_{9,0.975} = 19.02$, $t_{9,0.975} = 2.262$, womit das 95%-Konfidenzintervall

$$[0.897, 2.762] \quad \text{für } \mu$$

und

$$[1.897, 13.361] \quad \text{für } \sigma^2$$

gegeben ist. Hätten wir hingegen unterstellt, dass die Daten Realisierungen unabhängiger $\mathcal{N}(\mu, 4)$-Verteilungen sind, so ist $\Phi^{-1}_{0.975} = 1.96$ und Beispiel 3.5 liefert uns das 95%-Konfidenzintervall

$$[1.090, 3.570] \quad \text{für } \mu.$$

Wir sehen, dass dieses Intervall kleiner ist als das zuvor erhaltene. Das ist darauf zurückzuführen, dass für kleine Stichprobengrößen n die Verteilung von t_n größere *Tail-Wahrscheinlichkeiten* besitzt als die Standardnormalverteilung, d.h. dass $\mathbb{P}(Z > t) \leq \mathbb{P}(X > t)$ für $Z \sim \mathcal{N}(0,1)$, $X \sim t_n$. In Quantile übersetzt bedeutet das, dass $\Phi^{-1}_{1-\alpha/2} \leq t_{n,1-\alpha/2}$ für kleine α und kleine n.

3.3 Asymptotische Konfidenzbereiche

In allen bisherigen Beispielen ist es uns gelungen, eine Statistik $T(X) = T(X, \theta)$ zu konstruieren, deren Verteilung unabhängig von $\theta \in \Theta$ ist. Gleichzeitig ist dieser Schritt im Allgemeinen schwierig bis unmöglich. Basierend auf der asymptotischen Verteilung von Statistiken ist es jedoch oft möglich, Konfidenzbereiche zu bestimmen, die ein gefordertes Niveau $1 - \alpha$ approximativ einhalten.

Definition 3.12 (Asymptotischer Konfidenzbereich). *Es sei $(\mathcal{E}^n)_{n\in\mathbb{N}}$ eine Folge statistischer Modelle $\mathcal{E}^n = (\mathcal{X}^n, \mathcal{F}^n, \{\mathbb{P}^n_\theta : \theta \in \Theta\})$, Θ' eine Menge und $g : \Theta \longrightarrow \Theta'$ eine Funktion sowie $\mathcal{S} \subset \mathcal{P}(\Theta')$ eine Familie von Teilmengen von Θ'. Weiter sei für jedes $n \in \mathbb{N}$ eine Abbildung $C_n : \mathcal{X}^n \longrightarrow \mathcal{S}$ gegeben, sowie $\alpha \in (0,1)$. Dann heißt die Folge $(C_n)_{n\in\mathbb{N}}$* asymptotischer Konfidenzbereich zum Niveau $1 - \alpha$, *falls für alle $\theta \in \Theta$ und $X^n \sim \mathbb{P}^n_\theta$*

$$\liminf_{n\to\infty} \mathbb{P}^n_\theta \left(g(\theta) \in C_n(X^n)\right) \geq 1 - \alpha.$$

Bemerkung. In vielen Fällen ist der 'lim inf' ein 'lim' und das '$\geq$' ein '$=$'.

Beispiel 3.13. Wir betrachten das n-fache Produktmodell mit Bernoulli-Verteilungen aus Abschnitt 1.1, d.h. $X_1, \dots, X_n$ sind unabhängig identisch Bernoulli-verteilt zum Parameter $p \in (0,1)$. Wir hatten dort bereits gesehen, dass

$$Z_n := \frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{\mathcal{D}_p} \mathcal{N}(0,1).$$

Setzen wir $X = (X_1, \dots, X_n)$ und bezeichnen mit $\Phi^{-1}_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung sowie

$$p_n^-(X) = \overline{X}_n - \Phi^{-1}_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}},$$
$$p_n^+(X) = \overline{X}_n + \Phi^{-1}_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}},$$

so erhalten wir (unter Beachtung der Tatsache, dass Normalverteilungen keine Punktmassen haben)

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}_p\left(p_n^-(X) \le p \le p_n^+(X)\right) &= \lim_{n\to\infty}\left(-\Phi^{-1}_{1-\alpha/2} \le -Z_n \le \Phi^{-1}_{1-\alpha/2}\right)\\ &= \Phi\left(\Phi^{-1}_{1-\alpha/2}\right) - \Phi\left(-\Phi^{-1}_{1-\alpha/2}\right)\\ &= \Phi\left(\Phi^{-1}_{1-\alpha/2}\right) - \Phi\left(\Phi^{-1}_{\alpha/2}\right)\\ &= 1-\alpha.\end{aligned}$$

Damit ist $[p_n^-(X), p_n^+(X)]$ ein asymptotisches $(1-\alpha)$-Konfidenzintervall für den Parameter p (vgl. Abschnitt 1.1).

Korollar 3.14 (Konfidenzintervall für Momentenschätzer). *Das statistische Modell $\mathcal{E}$ und $h : \mathbb{R}^l \to \mathbb{R}$ erfüllen die Voraussetzungen von Satz 2.14 sowie $\mathbb{E}_\theta[X_1^{4l}] < \infty$ für alle $\theta \in \Theta$. Weiter sei $\widehat{v_\theta} := \nabla h(\hat{m}_{1,n}(X), \dots, \hat{m}_{l,n}(X))^T$, $\hat{\Sigma} = (\hat{\Sigma}_{ij})_{1\le i,j\le l}$ mit*

$$\hat{\Sigma}_{ij} = \hat{m}_{i+j,n}(X) - \hat{m}_{i,n}(X)\hat{m}_{j,n}(X)$$

sowie $\widehat{g(\theta)}_{\mathrm{MS}} = h(\hat{m}_{1,n}(X), \dots, \hat{m}_{l,n}(X))$ der Momentenschätzer für $g(\theta) = h(m_1(\theta), \dots, m_l(\theta)) \in \mathbb{R}$. Ist $\alpha \in (0,1)$, so ist

$$\left[\widehat{g(\theta)}_{\mathrm{MS}} - \sqrt{\frac{\widehat{v_\theta}^T\hat{\Sigma}\widehat{v_\theta}}{n}}\Phi^{-1}_{1-\alpha/2}, \widehat{g(\theta)}_{\mathrm{MS}} + \sqrt{\frac{\widehat{v_\theta}^T\hat{\Sigma}\widehat{v_\theta}}{n}}\Phi^{-1}_{1-\alpha/2}\right]$$

ein asymptotisches $(1-\alpha)$-Konfidenzintervall für $g(\theta)$.

Beweis. Sei $v_\theta = \nabla h(m_1(\theta), \dots, m_l(\theta))$. Da ∇h nach Voraussetzung stetig ist, erhalten wir mit Lemma 2.8 und dem schwachen Gesetz großer Zahlen dass $\widehat{v_\theta} \to_{\mathbb{P}_\theta} v_\theta$. Aus der

Darstellung $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ erhalten wir wiederum mit dem schwachen Gesetz großer Zahlen und $\mathbb{E}_\theta[X_1^{4l}] < \infty$, dass

$$\hat{\Sigma}_{ij} \longrightarrow_{\mathbb{P}_\theta} \Sigma_{ij} = \mathrm{Cov}_\theta(X_1^i, X_1^j)$$

für alle $1 \leq i, j \leq l$. Fassen wir für einen Vektor $x \in \mathbb{R}^d$ und eine Matrix $M \in \mathbb{R}^{d\times d}$ die Abbildung $f(x, M) = x^T M x$ als Abbildung $f : \mathbb{R}^{d(d+1)} \to \mathbb{R}$ auf, so liefert uns Lemma 2.8

$$\widehat{v_\theta}^T \hat{\Sigma} \widehat{v_\theta} \longrightarrow_{\mathbb{P}_\theta} v_\theta^T \Sigma v_\theta.$$

Damit impliziert Satz 2.14 zusammen mit Proposition $I.5.4$(iii), dass

$$\sqrt{\frac{n}{\widehat{v_\theta}^T \hat{\Sigma} \widehat{v_\theta}}} \left(\widehat{g(\theta)}_{\mathrm{MS}} - g(\theta) \right) \xrightarrow{\mathcal{D}_\theta} Z,$$

mit $Z \sim \mathcal{N}(0,1)$ und die Behauptung folgt. □

Bemerkung. Für Maximum-Likelihood-Schätzer, welche asymptotisch normalverteilt sind (vgl. Bemerkung am Ende von Abschnitt 2.5) gilt ein analoges Resultat mit identischem Beweis.

Lemma 3.15 (Konservative Konfidenzbereiche für Parametervektoren). *Es sei $\mathcal{E} = (\mathcal{X}, \mathcal{F}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ ein statistisches Modell mit $\Theta = \Theta_1 \times \cdots \times \Theta_d$. Weiter seien $\alpha_1, \ldots, \alpha_d \in (0,1)$ mit $\sum_{i=1}^d \alpha_i < 1$ und für jedes $i = 1, \ldots, d$ sei $C_i : \mathcal{X} \longrightarrow \mathcal{P}(\Theta_i)$ ein $(1-\alpha_i)$-Konfidenzbereich für θ_i, d.h.*

$$\mathbb{P}_\theta\left(\theta_i \in C_i(X)\right) \geq 1 - \alpha_i \qquad \text{für alle } \theta \in \Theta.$$

Dann ist die Abbildung $C : \mathcal{X} \longrightarrow \Theta$ gegeben durch

$$C(x) = C_1(x) \times \ldots C_d(x)$$

ein Konfidenzbereich für den Vektor $\theta = (\theta_1, \ldots, \theta_d)$ zum Niveau $1 - \sum_{i=1}^d \alpha_i$. Eine analoge Aussage gilt für asymptotische Konfidenzbereiche.

Beweis. Die Aussage folgt aus der endlichen Additivität von $\mathbb{P}_\theta$, da

$$\mathbb{P}_\theta\left(\theta \notin C(X)\right) = \mathbb{P}_\theta\left(\bigcup_{i=1}^d \{\theta_i \notin C_i(X)\}\right) \leq \sum_{i=1}^d \mathbb{P}_\theta\left(\theta_i \notin C_i(X)\right) \leq \sum_{i=1}^d \alpha_i.$$

□

4 Statistische Tests

Neben der Schätztheorie ist die Theorie statistischer Tests ein weiteres wichtiges Gebiet der induktiven Statistik. Basierend auf den erhobenen Daten lassen sich oftmals Hypothesen über deren Verteilung formulieren. Das Überprüfen dieser Hypothesen erfolgt dann mithilfe statistischer Tests.

4.1 Grundbegriffe der Testtheorie

Ziel der Schätztheorie war es, den wahren und unbekannten Parameter θ, welcher die den Daten zugrunde liegenden Verteilung bestimmt, möglichst genau zu bestimmen ($\to$ Punktschätzer) oder seine Lage auf einen bestimmten Bereich einzuschränken ($\to$ Konfidenzbereiche). Die Aufgabe statistischer Tests ist es, zu entscheiden ob der Parameter θ innerhalb einer vorgegebenen Teilmengen $\Theta_0 \subset \Theta$ der Parametermenge enthalten ist oder in deren Komplement $\Theta_1 = \Theta \subset \Theta_0$ liegt.

Wir unterstellen in diesem Kapitel durchgängig ein Produktmodell

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

d.h. wir nehmen an, dass wir n Realisierungen unabhängig identisch nach $\mathbb{P}_\theta$ verteilter reellwertiger Zufallsvariablen beobachten. Weiter nennen wir

$$\begin{aligned} H_0 : \theta \in \Theta_0 \qquad & \text{die Nullhypothese und} \\ H_1 : \theta \in \Theta_1 \qquad & \text{die Alternative.} \end{aligned}$$

Nachfolgend schreiben bezeichnen wir ein Testproblem häufig durch (Θ_0, Θ_1), d.h. durch Angabe seiner Null- und Alternativhypothese.

Beispiel 4.1.

(1) In Abschnitt 1.1 haben wir das n-fache Produktmodell mit Bernoulliverteilungen betrachtet und dort die Hypothese aufgestellt, dass die Münze fair sei. Dies bedeutet formal, dass $\Theta_0 = \{\frac{1}{2}\}$ und $\Theta_1 = (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$. Man schreibt für das Testproblem auch

$$H_0 : p = \frac{1}{2} \qquad \text{gegen} \qquad H_1 : p \neq \frac{1}{2}.$$

(2) Ist $\Theta = \mathbb{R}$, so sind die Hypothesen meist von einer der Formen

$$\begin{aligned} &H_0 : \theta \leq \theta_0 \quad \text{gegen} \quad H_1 : \theta > \theta_0, && \text{d.h. } \Theta_0 = (-\infty, \theta_0], \Theta_1 = (\theta_0, \infty), \\ &H_0 : \theta \geq \theta_0 \quad \text{gegen} \quad H_1 : \theta < \theta_0, && \text{d.h. } \Theta_0 = [\theta_0, \infty), \Theta_1 = (-\infty, \theta_0), \\ &H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_1 : \theta \neq \theta_0, && \text{d.h. } \Theta_0 = \{\theta_0\}, \Theta_1 = (-\infty, \theta_0) \cup (\theta_0, \infty). \end{aligned}$$

In den ersten beiden Fällen spricht man auch von einseiten Testproblem, im letzten Fall von einem zweiseitigen Testproblem.

Wir kommen nun zur zentralen Definition eines statistischen Tests.

Definition 4.2 (Statistischer Test). *Ein* statistischer Test für das Testproblem (Θ_0, Θ_1) *ist eine Entscheidungsregel der Form einer (messbaren) Funktion $\phi : \mathbb{R}^n \longrightarrow \{0, 1\}$, wobei $\phi(x) = k$ bedeutet, sich für Θ_k zu entscheiden.*

Bemerkung (Ablehnungsbereich). Sezten wir für einen statistischen Test ϕ die Menge $C := \{x \in \mathbb{R}^n : \phi(x) = 1\}$, so gilt $\phi(x) = \mathbb{1}_C(x)$. Die Menge C gibt diejenigen Beobachtungen an, bei denen wir uns für die Alternative entscheiden und wird Ablehnungsbreich (von H_0) genannt. Faktisch werden Tests fast immer darüber konstruiert, dass dieser Ablehnungsbereich spezifiziert wird. In vielen Fällen erfolgt die Konstruktion dergestalt, dass für eine Statistik $T(x)$ ein Ablehnungsbereich C' konstruiert wird, d.h. der resultierende Test hat die Form $\phi(x) = \mathbb{1}_{C'}(T(x))$.

Bei der Testentscheidung für oder gegen die Nullhypothese sind zwei verschiedene Fehler/Irrtümer möglich:

- **Fehler 1. Art**: Die Nullhypothese H_0 wird verworfen, obwohl sie zutrifft.
- **Fehler 2. Art**: Die Nullhypothese H_0 wird nicht verworfen, obwohl sie falsch ist.

Dies führt auf das nachfolgenden Schema:

	H_0 abgelehnt	H_0 nicht abgelehnt
H_0 richtig	Fehler 1. Art	richtige Entscheidung
H_0 falsch	richtige Entscheidung	Fehler 2. Art

Man überlegt sich leicht, dass es unmöglich ist, das Testverfahren so zu konstruieren, dass sowohl der Fehler 1. Art als auch der Fehler 2. Art beliebig klein werden (so macht man z.B. nie einen Fehler 1. Art, wenn man sich immer für H_0 entscheidet). Daher benötigen wir irgendeine zusätzliche Restriktion: Unser Ansatz besteht darin, dass wir uns ein Level vorgeben, wie groß bzw. wahrscheinlich der Fehler 1. Art noch sein darf. Unter dieser Restriktion können wir dann versuchen einen Test zu finden, dessen Fehler 2. Art ebenfalls klein ist. Es sei an dieser Stelle daran erinnert, dass unser Vorgehen bei der Konstruktion von Konfidenzintervallen ähnlich war: Hier haben wir uns eine Irrtumswahrscheinlichkeit α vorgegeben, sodass wir mit $\mathbb{P}(g(\theta) \in C(X)) \geq 1 - \alpha$ zufrieden waren und anschließend versucht, möglichst genaue (d.h. kleine) Bereiche $C(X)$ zu konstruieren, die dieser Bedingung genügen. Formalisiert sieht der Ansatz für Tests dann folgendermaßen aus:

Definition 4.3 (**Niveau-α-Test**). *Sei $\alpha \in [0,1]$. Ein Test ϕ für das Testproblem (Θ_0, Θ_1) heißt* Niveau-α-Test*, falls $\mathbb{E}_\theta[\phi(X)] \leq \alpha$ für alle $\theta \in \Theta_0$.*

Bemerkung (Asymmetrie von H_0 und H_1/Wahl der Hypothesen).

(1) Um zu überprüfen, ob ein Test ϕ ein bestimmtes Niveau hat, ist es notwendig die Verteilung von ϕ für jedes $\theta \in \Theta_0$ zu kennen. Konstruieren wir den Test als $\phi = \mathbb{1}_C(T(X))$, so bedeutet dies, dass wir Interesse an der Verteilung einer Statistik $T(X)$ unter allen $\mathbb{P}_\theta$ für $\theta \in \Theta_0$ haben.

(2) Durch das Niveau α stellen wir sicher, dass ein Fehler 1. Art nur mit Wahrscheinlichkeit α auftritt, denn

$$\alpha \geq \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\phi(X) = 1) = \mathbb{P}_\theta(\text{Entscheidung für } H_1) \quad \text{für alle } \theta \in \Theta_0.$$

Dies bricht die Symmetrie zwischen H_0 und H_1, denn während wir eine Kontrolle über den Fehler 1. Art haben, fehlt diese für den Fehler 2. Art.

(3) In der Praxis wählt man die Nullhypothese nun so, dass die Ablehnung dieser möglichst sicher auf die Richtigkeit der Alternative zurückzuführen ist. Wollen wir etwa ein Medikament auf seine Wirksamkeit testen, so tun wir dass normalerweise nur, wenn wir davon ausgehen, dass diese vorhanden ist. Daher wählen wir

$$H_0: \text{'das Medikament wirkt nicht'} \quad \text{gegen} \quad H_1: \text{'das Medikament wirkt'}.$$

Kommt es nun zu einer Ablehung von H_0, so wissen wir, dass dies mit Wahrscheinlichkeit höchstens α dann passiert, falls H_0 wahr sein sollte (d.h. falls das Medikament unwirksam ist). Damit ist relativ sicher, dass eine Ablehung von H_0 darauf zurückzuführen ist, dass H_1 zutrifft (in diesem Fall, dass das Medikament wirkt).

(4) Lehnt ein Test die Nullhypothese umgekehrt *nicht* ab, so kann und sollte daraus nicht geschlossen werden, dass H_0 richtig ist!

Satz 4.4 (Binomialtest). *Es sei $\mathcal{E}$ das n-fache Produktmodell mit Bernoulli-Verteilungen, d.h. $\mathbb{P}_p = \mathrm{Ber}(p)$, $p \in \Theta = (0,1)$ und $\alpha \in [0,1]$.*

(a) Ist $\Theta_0 = \{p_0\}$, $\Theta_1 = (0,p_0) \cup (p_0,1)$, so ist

$$X \mapsto \mathbb{1}_{\{0,1,\dots,k\}\cup\{l,\dots,n\}}(n\overline{X}_n)$$

ein Test zum Niveau α (für das Testproblem (Θ_0,Θ_1)), falls

$$\mathbb{P}_{p_0}(n\overline{X}_n \le k) \le \alpha/2 \qquad \text{und} \qquad P_{p_0}(n\overline{X}_n \ge l) \le \alpha/2.$$

(b) Ist $\Theta_0 = (0,p_0]$, $\Theta_1 = (p_0,1)$, so ist

$$X \mapsto \mathbb{1}_{\{l,\dots,n\}}(n\overline{X}_n)$$

ein Test zum Niveau α, falls $\mathbb{P}_{p_0}(n\overline{X}_n \ge l) \le \alpha$.

(c) Ist $\Theta_0 = [p_0,1)$, $\Theta_1 = (0,p_0)$, so ist

$$X \mapsto \mathbb{1}_{\{0,1,\dots,k\}}(n\overline{X}_n)$$

ein Test zum Niveau α, falls $\mathbb{P}_{p_0}(n\overline{X}_n \le k) \le \alpha$.

Beweis. Teil (a) folgt direkt, da

$$\mathbb{E}_{p_0}\left[\mathbb{1}_{\{0,1,\dots,k\}\cup\{l,\dots,n\}}(n\overline{X}_n)\right] \le \mathbb{P}_{p_0}\left(n\overline{X}_n \in \{0,1,\dots,k\}\right) + \mathbb{P}_{p_0}\left(n\overline{X}_n \in \{l,\dots,n\}\right) \le \alpha.$$

Für Teil (b) stellen wir fest, dass $n\overline{X}_n = \sum_{i=1}^n X_i \sim \mathrm{Bin}(n,p)$ binomialverteilt ist mit Parametern n und p. Damit gilt

$$\sup_{p\in\Theta_0} \mathbb{E}_p\left[\mathbb{1}_{\{l,\dots,n\}}(n\overline{X}_n)\right] = \sup_{p\in(0,p_0]} \mathbb{P}_p\left(n\overline{X}_n \in \{l,\dots,n\}\right).$$

Um nun zu sehen, wann die Wahrscheinlichkeit auf der rechten Seite am größten ist, sei $(\Omega,\mathcal{F},\mathbb{P})$ ein Wahrscheinlichkeitsraum auf welchem auf $[0,1]$ uniform verteilte Zufallsvariablen $U_1,\dots,U_n$ definiert sind. Dann ist $N_p := \#\{k : U_k \le p\}$ binomialverteilt mit den Parametern n und p und es gilt offensichtlich $N_p \le N_{p'}$ für $p \le p'$. Damit erhalten wir

$$\sup_{p\in(0,p_0]} \mathbb{P}_p\left(n\overline{X}_n \in \{l,\dots,n\}\right) = \sup_{p\in(0,p_0]} \mathbb{P}\left(N_p \ge l\right) = \mathbb{P}(N_{p_0} \ge l) = \mathbb{P}_{p_0}\left(n\overline{X}_n \ge l\right)$$

und entsprechend

$$\sup_{p\in\Theta_0} \mathbb{E}_p\left[\mathbb{1}_{\{l,\dots,n\}}(n\overline{X}_n)\right] = \mathbb{P}_{p_0}\left(n\overline{X}_n \ge l\right) \le \alpha,$$

sodass es sich beim Test in (b) um einen Level-α-Test handelt. (c) folgt analog zu (b). □

Bemerkung (Coupling). Innerhalb des Beweises von Satz 4.4 mussten wir die Wahrscheinlichkeiten $\mathbb{P}(X \geq m)$ und $\mathbb{P}(Y \geq m)$ miteinander vergleichen, wobei $X \sim \text{Bin}(n, p_1)$ und $Y \sim \text{Bin}(n, p_2)$. Dies ist uns gelungen, indem wir (auf einem potentiell anderen Wahrscheinlichkeitsraum) eine Familie $(N_p)_{p\in(0,1)}$ von Zufallsvariablen definiert haben mit $N_p \sim \text{Bin}(n, p)$, für welche $N_{p_1} \leq N_{p_2}$ für $p_1 \leq p_2$ offensichtlich ist. Eine solche Konstruktion ist als Kopplung (engl. Coupling) bekannt.

Beispiel 4.5. Wir betrachten abermals die Situation aus Abschnitt 1.1 mit $n = 100$ und 55 Beobachtungen 'Kopf' (codiert als Eins). Weiter wählen wir $\alpha = 5\%$ und wollen

$$H_0 : p = 1/2 \quad \text{gegen} \quad H_1 : p \neq 1/2$$

testen. Für $X \sim \text{Bin}(100, 1/2)$ gilt

$$\mathbb{P}(X \leq 39) = \mathbb{P}(X \geq 61) \approx 1.76\% \quad \text{und} \quad \mathbb{P}(X \leq 40) = \mathbb{P}(X \geq 60) \approx 2.84\%.$$

Nach Satz 4.4 lehnen wir H_0 dann ab, wenn $n\overline{X}_n \in \{0, 1, \ldots, 39\} \cup \{61, \ldots, 100\}$. Da in unserem Fall $n\overline{X}_n = 55$, können wir H_0 nicht ablehnen.

Bemerkung (p-Wert). Es sei $\phi = \mathbb{1}_C(T(X))$ ein Test für eine Statistik $T(X)$. Seien x die Beobachtungen und $t = T(x)$ der Wert der Teststatistik für die Beobachtungen. Dann heißt der Wert

$$p_t = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta \left(T(X) \text{ ist extremer als } t\right)$$

der p-Wert des Tests für $T(X) = t$. Was genau 'extremer' bedeutet hängt davon ab, was genau die Alternative des betrachteten Testproblems ist. In jedem Fall gilt $p_y \leq p_{y'}$, wenn y extremer als y' ist. In den meisten Beispiel ist jedoch klar, was darunter zu verstehen ist: So ist der p-Wert für den einseitigen Binomialtest aus Satz 4.4(b) gerade gegeben durch die Wahrscheinlichkeit $\mathbb{P}_{p_0}(n\overline{X}_n \geq n\overline{x}_n)$.
Man beachte, dass ein enger Zusammenhang zwischen dem p-Wert und dem Niveau α eines Tests besteht: Ist ϕ ein Niveau-α-Test für die Wahl

$$C = \{T(x) : T(x) \text{ ist extremer als } t_0\}$$

für ein t_0, so gilt

$$p_{t_0} = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(T(X) \text{ extremer als } t_0) \leq \alpha.$$

Gilt nun $X = x$ und $p_{T(x)} \leq p_{t_0}$, so ist $T(x)$ extremer als t_0 und damit im Ablehnungsbereich. Damit kann eine Entscheidung über H_0 oder H_1 auch anhand des Werts $p_{T(x)}$ bestimmt werden - gilt nämlich $p_{T(x)} \leq \alpha$, so ist H_0 zu verwerfen. Dieses Vorgehen wird bei vielen Statistik-Programmen angewendet, bei denen lediglich p-Werte ausgegeben werden.
Beispiel: Eine solche Wahl von C liegt z.B. im Binomialtest von Satz 4.4(b) vor, hier ist $T(X) = n\overline{X}_n$ und $t_0 = l$, für das dort spezifizierte l und 'extremer' bedeutet dort '$\geq$'.

4.2 Zusammenhang mit Konfidenzintervallen

Ziel dieses Abschnitts ist es, einen Zusammenhang zwischen Tests und Konfidenzbereichen herzustellen. Der nachfolgende Satz ermöglicht es insbesondere, basierend auf bereits konstruierten Konfidenzbereichen, statistische Tests zu konstruieren.

Satz 4.6 (**Dualität zwischen Konfidenzbereichen und Tests**).

(a) Seien $\alpha \in (0,1)$, $C : \mathcal{X} \longrightarrow \mathcal{P}(\Theta)$ ein Konfidenzbereich zum Niveau $1-\alpha$ für θ und $\theta_0 \in \Theta$ vorgegeben. Dann ist der Test

$$\phi(x) := \mathbb{1}_{\{\theta_0 \notin C(x)\}}$$

ein Niveau-α-Test für das zweiseitige Testproblem $(\{\theta_0\}, \Theta \setminus \{\theta_0\})$.

(b) Seien $\alpha \in (0,1)$ und $(\phi_\theta)_{\theta\in\Theta}$ eine Familie von Tests, sodass für jedes $\theta_0 \in \Theta$ gilt, dass ϕ_{θ_0} ein Niveau-α-Test für das Testproblem $(\{\theta_0\}, \Theta\setminus\{\theta_0\})$ ist. Dann ist

$$C(x) := \{\theta \in \Theta : \phi_\theta(x) = 0\}$$

ein Konfidenzbereich zum Niveau $1-\alpha$ für θ.

Beweis. (a) Offensichtlich handelt es sich bei ϕ um einen Test und es genügt zu zeigen, dass er das Niveau α hat. Dies folgt jedoch direkt, da

$$\mathbb{E}_{\theta_0}[\phi] = \mathbb{P}_{\theta_0}(\theta_0 \notin C(X)) = 1 - \mathbb{P}_{\theta_0}(\theta_0 \in C(X)) \leq 1 - (1-\alpha) = \alpha.$$

(b) Es sei $\theta_0 \in \Theta$. Nach Voraussetzung ist ϕ_{θ_0} ein Niveau-α-Test für das Testproblem $(\{\theta_0\}, \Theta \setminus \{\theta_0\})$, d.h. es gilt $\mathbb{E}_{\theta_0}[\phi_{\theta_0}] \leq \alpha$. Nach Definition von C gilt dann

$$\mathbb{P}_{\theta_0}(\theta_0 \in C(X)) = \mathbb{P}_{\theta_0}(\phi_{\theta_0}(X) = 0) = 1 - \mathbb{P}_{\theta_0}(\phi_{\theta_0} = 1) = 1 - \mathbb{E}_{\theta_0}[\phi_{\theta_0}] \geq 1 - \alpha.$$

Da θ_0 beliebig war, folgt die Behauptung.

□

Beispiel 4.7 (**Zweiseitiger Gauß-Test**). Wir betrachten das n-fache Produktmodell mit Normalverteilungen mit bekannter Varianz, d.h. $\mathbb{P}_\mu = \mathcal{N}(\mu, \sigma^2)$ für vorgegebenes $\sigma^2 > 0$. Nach Beispiel 3.2 und 3.5 sowie Satz 4.6 ist dann ein Niveau-α-Test für das Testproblem

$$H_0 : \mu = \mu_0 \quad \text{gegen} \quad H_1 : \mu \neq \mu_0$$

gegeben durch

$$\mathbb{1}_{\left\{\mu \notin \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}_{1-\alpha/2}\right]\right\}}.$$

Definiert man die Teststatistik

$$Z_{\mu_0} := \frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}},$$

so kann dieser Test auch geschrieben werden (Übung!) als

$$\mathbb{1}_{\left(-\infty, \Phi^{-1}_{\alpha/2}\right) \cup \left(\Phi^{-1}_{1-\alpha/2}, \infty\right)}(Z_{\mu_0}).$$

Mittels einer darauf aufbauenden Überlegung erhält man auch Tests für die einseitigen Testprobleme wie im nachfolgenden Abschnitt mehrfach dargestellt.

4.3 Tests im Normalverteilungsmodell

Basierend auf dem vorangegangen Abschnitt und den Konfidenzintervallen aus Abschnitt 3.2, stellen wir in diesem Abschnitt Tests im Normalverteilungsmodell vor. Dabei diskutieren wir sowohl Einstichproben- als auch Zweistichprobenprobleme.

Satz 4.8 (**Gauß-Test**). *Es sei $\mathcal{E}$ das n-fache Produktmodell mit Normalverteilungen mit bekannter Varianz, d.h. $\mathbb{P}_\mu = \mathcal{N}(\mu, \sigma^2)$ für vorgegebenes $\sigma^2 > 0$ und $\mu \in \Theta = \mathbb{R}$. Weiter sei $\alpha \in (0,1)$, $\mu_0 \in \mathbb{R}$ und*

$$Z_{\mu_0} = \frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}}.$$

(a) Ist $\Theta_0 = \{\mu_0\}$, $\Theta_1 = (-\infty, \mu_0) \cup (\mu_0, \infty)$, so ist ein Niveau-$\alpha$-Test gegeben durch

$$X \mapsto \mathbb{1}_{\left(-\infty, \Phi^{-1}_{\alpha/2}\right) \cup \left(\Phi^{-1}_{1-\alpha/2}, \infty\right)}(Z_{\mu_0}).$$

(b) Ist $\Theta_0 = (-\infty, \mu_0]$, $\Theta_1 = (\mu_0, \infty)$, so ist ein Niveau-$\alpha$-Test gegeben durch

$$X \mapsto \mathbb{1}_{\left[\Phi^{-1}_{1-\alpha}, \infty\right)}(Z_{\mu_0}).$$

(c) Ist $\Theta_0 = [\mu_0, \infty)$, $\Theta_1 = (-\infty, \mu_0)$, so ist ein Niveau-$\alpha$-Test gegeben durch

$$X \mapsto \mathbb{1}_{\left(-\infty, \Phi^{-1}_{\alpha}\right]}(Z_{\mu_0}).$$

Beweis. Teil (a) ist eine direkte Konsequenz aus Satz 4.6 zusammen mit Beispiel 3.5, vgl. Beispiel 4.7. Für Teil (b) überlegen wir uns, dass für $\mu \in (-\infty, \mu_0]$

$$Z_{\mu_0} = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} + \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}},$$

wobei der erste Summand unter $\mathbb{P}_\mu$ standardnormalverteilt ist und der zweite deterministisch und negativ. Damit gilt

$$\sup_{\mu \le \mu_0} \mathbb{P}_\mu \left(Z_{\mu_0} \ge \Phi^{-1}_{1-\alpha}\right) = \mathbb{P}_{\mu_0} \left(Z_{\mu_0} \ge \Phi^{-1}_{1-\alpha}\right) = \alpha,$$

wobei der letzte Schritt nutzt, dass Z_{μ_0} unter $\mathbb{P}_{\mu_0}$ standardnormalverteilt ist. Teil (c) folgt analog. □

Bemerkung. Man beachte, dass die hier genutzt Darstellung des Gauß-Tests gerade von der Bauart ist, dass eine Statistik spezifiziert wird (hier Z_{μ_0}) und ein dazugehöriger Ablehnungsbereich, vgl. die Bemerkung im Anschluss an Definition 4.2.

Falls die Varianz im Normalverteilungsmodell unbekannt ist, können wir den Gauß-Test nicht länger verwenden. Wir haben mit Theorem 3.10 und dem Satz von Fisher, welcher diesem Resultat zugrunde liegt, jedoch bereits die notwendige Vorarbeit geleistet, um den in diesem Fall zu verwendenden t-Test beschreiben zu können.

Satz 4.9 (*t*-**Test**). *Es sei $\mathcal{E}$ das n-fache Produktmodell mit Normalverteilungen, d.h. $\mathbb{P}_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2)$ und $(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Weiter sei $\alpha \in (0, 1)$, $\mu_0 \in \mathbb{R}$ und*

$$T_{\mu_0} = \frac{\overline{X}_n - \mu_0}{\sqrt{s_n^2(X)/n}}$$

und es bezeichne $t_{n,\alpha}$ das α-Quantil der t_n-Verteilung.

(a) Ist $\Theta_0 = \{\mu_0\}$, $\Theta_1 = (-\infty, \mu_0) \cup (\mu_0, \infty)$, so ist ein Niveau-$\alpha$-Test gegeben durch

$$X \mapsto \mathbb{1}_{(-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty)}(T_{\mu_0}).$$

(b) Ist $\Theta_0 = (-\infty, \mu_0]$, $\Theta_1 = (\mu_0, \infty)$, so ist ein Niveau-$\alpha$-Test gegeben durch

$$X \mapsto \mathbb{1}_{[t_{n-1,1-\alpha}, \infty)}(T_{\mu_0}).$$

(c) Ist $\Theta_0 = [\mu_0, \infty)$, $\Theta_1 = (-\infty, \mu_0)$, so ist ein Niveau-$\alpha$-Test gegeben durch

$$X \mapsto \mathbb{1}_{(-\infty, t_{n-1,\alpha}]}(T_{\mu_0}).$$

Beweis. Teil (a) folgt direkt aus Satz 4.6 und Theorem 3.10. Teil (b) und (c) folgen analog zum Vorgehen im Beweis von Satz 4.8. □

Sowohl der Gauß- als auch der t-Test gehören zu sogenannten Einstichprobenproblemen. Damit ist gemeint, dass die beobachteten Daten einer Stichprobe eines n-fachen Produktmodells entsprechen. Wir wollen nun unterstellen, dass unsere Daten aus zwei Gruppen bestehen: einer Stichprobe $X_1, X_2, \ldots, X_n$ von unabhängig identisch nach $\mathbb{P}_\theta$ verteilten Daten, sowie einer weiteren Stichprobe $Y_1, Y_2, \ldots, Y_m$ von unabhängig identisch nach $\tilde{\mathbb{P}}_\vartheta$ verteilten Daten. Ziel von Testproblemen in einem solchen Modell ist es meist, Hypothesen über Unterschiede in beiden Stichproben zu testen. Für die beiden Stichproben unterscheidet man die folgenden Fälle:

- *Gepaarte Stichproben.* Hier gilt $n = m$ und X_i hängt auf eine bestimmte Weise mit Y_i zusammen. Zum Beispiel beschreibt X_i den Blutdruck eines Patienten i vor Gabe eines Medikaments und Y_i den danach. Insbesondere werden X_i und Y_i nicht als unabhängig angenommen. Stattdessen kann man z.B. eine Verteilungsannahme an $X_i - Y_i$ stellen.

- *Ungepaarte Stichproben.* Hier besteht nicht notwendigerweise ein 'physikalischer' Zusammenhang zwischen den Stichproben $X = (X_1, \ldots, X_n)$ und $Y = (Y_1, \ldots, Y_m)$. Man unterstellt typischerweise, dass sie unabhängig voneinander sind. So könnte etwa X ein Vektor von Geburtsgewichten von Kindern in einer Klinik A sein, wohingegen Y Geburtsgewichte von Kindern in einer anderen Klinik B beschreibt.

Im zweiten Fall spricht man auch von Zweistichprobenproblemen. Im folgenden betrachten wir exemplarisch wieder den Normalverteilungsfall. Dabei stellen wir für beide Varianten eine Abwandlung des t-Tests vor, den gepaarten t-Test und den doppelten t-Test.

Satz 4.10 (**Gepaarter t-Test**). *Wir betrachten das statistische Modell $\mathcal{E} = (\mathbb{R}^{2n}, \mathscr{B}(\mathbb{R}^{2n}), \{\mathbb{P}_{\mu,\sigma^2} : (\mu,\sigma^2) \in \mathbb{R}\times(0,\infty)\})$, wobei unter $\mathbb{P}_{\mu,\sigma^2}$ die Zufallsvariablen $Y - X = (Y_1 - X_1, \dots, Y_n - X_n)$ unabhängig identisch $\mathcal{N}(\mu,\sigma^2)$-verteilt sind. Wir definieren*

$$T = \frac{\overline{Y}_n - \overline{X}_n}{\sqrt{s_n^2(Y-X)/n}}$$

und es bezeichne $t_{n,\alpha}$ das α-Quantil der t_n-Verteilung. Dann ist für $\alpha \in [0,1]$

$$(X,Y) \mapsto \mathbb{1}_{(-\infty, t_{n-1,\alpha/2})\cup(t_{n-1,1-\alpha/2},\infty)}(T)$$

ein Test zum Niveau α für das Testproblem

$$H_0 : \mu = 0 \quad \textit{gegen} \quad H_1 : \mu \neq 0.$$

Beweis. Dies folgt direkt aus Satz 4.9 und der Feststellung, dass $\overline{(X-Y)}_n = \overline{Y}_n - \overline{X}_n$. □

Beispiel 4.11. In Beispiel 3.11 wurden Schlafdauern vor und nach Einnahme eines Medikaments notiert. Sind X und Y die Schlafdauern vor und nach Einnahme, so ist

$$Y - X = (1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4).$$

Wir haben dort bereits berechnet, dass $\overline{Y}_{10} - \overline{X}_{10} = 2.33$ und $s^2(Y-X) = 4.01$, womit $T = 2.33/\sqrt{4.01/10} \approx 3.68$. Zum Niveau von 5% ist der Ablehungsbereich des gepaarten t-Test in Satz 4.10 durch

$$C = (-\infty, -2.262) \cup (2.262, \infty)$$

gegeben, sodass die Hypothese H_0 (Schlafdauer ändert sich durch Medikamenteneinnahme nicht) hier verworfen werden kann.

Satz 4.12 (**Doppelter t-Test**). *Wir betrachten das statistische Modell $\mathcal{E} = (\mathbb{R}^{n+m}, \mathscr{B}(\mathbb{R}^{n+m}), \{\mathbb{P}_{\mu_X,\mu_Y,\sigma^2} : (\mu_X,\mu_Y,\sigma^2) \in \mathbb{R}^2\times(0,\infty)\})$, wobei unter $\mathbb{P}_{\mu_X,\mu_Y,\sigma^2}$ die Zufallsvariablen $X_1, \dots, X_n, Y_1, \dots, Y_m$ unabhängig seien, $X_1, \dots, X_n$ identisch $\mathcal{N}(\mu_X,\sigma^2)$-verteilt sind und $Y_1, \dots, Y_m$ identisch $\mathcal{N}(\mu_Y,\sigma^2)$-verteilt. Wir definieren*

$$T = \frac{\overline{Y}_m - \overline{X}_n}{\sqrt{s_{n,m}^2(X,Y)(m+n)/(mn)}}$$

mit

$$s_{n,m}^2(X,Y) = \frac{1}{m+n-1}\left(\sum_{i=1}^n (X_i - \overline{X}_n)^2 + \sum_{i=1}^m (Y_i - \overline{Y}_m)^2\right)$$

und es bezeichne $t_{n,\alpha}$ das α-Quantil der t_n-Verteilung. Dann ist für $\alpha \in [0,1]$

$$(X,Y) \mapsto \mathbb{1}_{(-\infty, t_{m+n-2,\alpha/2})\cup(t_{m+n-2,1-\alpha/2},\infty)}(T)$$

ein Test zum Niveau α für das Testproblem

$$H_0 : \mu_X = \mu_Y \quad \textit{gegen} \quad H_1 : \mu_X \neq \mu_Y.$$

Beweis. Nach Aufgabe 4 von Blatt 05 ist

$$I_{n,m} = \left[\overline{X}_n - \overline{Y}_m - \frac{s^2_{n,m}(X,Y)}{\sqrt{\frac{1}{n}+\frac{1}{m}}} t_{m+n-2,1-\alpha/2}, \overline{X}_n - \overline{Y}_m + \frac{s^2_{n,m}(X,Y)}{\sqrt{\frac{1}{n}+\frac{1}{m}}} t_{m+n-2,1-\alpha/2} \right]$$

ein Konfidenzintervall für $\mu_X - \mu_Y$ zum Niveau $1-\alpha$. Damit ist nach Satz 4.6 die Abbildung

$$(X,Y) \mapsto \mathbb{1}_{\{0 \notin I_{n,m}\}}$$

ein Niveau-α-Test für das angegebene Testproblem (beachte, dass H_0 als $\mu_X - \mu_Y = 0$ geschrieben werden kann). Umstellen liefert den im Satz angegebenen Ablehnungsbereich für T. □

Beispiel 4.13. In einer Kölner Klinik wurden im Jahr 1985 $m = 269$ Mädchen und $n = 288$ Jungen geboren. $X_1, \ldots, X_{269}$ seien die Geburtsgewichte der Mädchen und $Y_1, \ldots, Y_{288}$ diejenigen der Jungen. Es gilt $\overline{X}_{269} = 3050$, $s^2_{269}(X) = 211600$ und $\overline{Y}_{288} = 3300$, $s^2_{288}(Y) = 220900$. Wir wollen zum Niveau $\alpha = 0.01$ die Nullhypothese testen, dass Mädchen und Jungen das gleiche erwartete Geburtsgewicht haben. Dazu berechnen wir

$$s^2_{269,288}(X,Y) = \frac{1}{269+288-2}\left(268 s^2_{269}(X) + 287 s^2_{288}(Y)\right) = 216409$$

und

$$T = \frac{3300-3050}{\sqrt{216409 \cdot (269+288)/(269 \cdot 288)}} = 6.338 > 2.585 = t_{555,0.995}.$$

Damit können wir die Nullhypothese gleichen Geburtsgewichts zum Niveau 0.01 ablehnen. Unter der Annahme einer Normalverteilung und der Annahme gleicher Varianzen in beiden Stichproben kann also geschlossen werden, dass bei Geburt Jungen im Mittel schwerer als Mädchen sind.

Bemerkung (Suffizienz). Man beachte im vorangegangen Beispiel, dass wir die Teststatistik T nur auf Basis von $\overline{X}_n$, $\overline{Y}_m$, $s^2_n(X)$ und $s^2_m(Y)$ berechnen konnten und damit eine Entscheidung für das Testproblem treffen konnten. Insbesondere war eine genaue Kenntnis der einzelnen Werte $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ dazu nicht notwendig. Die ganze relevante Information für das gegebene Testproblem kann also in vier Variablen anstelle von $n+m$ Datenpunkten codiert werden. Die Frage nach der *relevanten* Information für eine statistische Fragestellung führt zum Begriff der *Suffizienz* → Mathematische Statistik.

Bemerkung (Einseitige Testprobleme). Analog zum einfachen t-Test aus Satz 4.9 existieren auch für den gepaarten und den doppelten t-Test Varianten für die einseitigen Hypothesen. Die Teststatistik bleibt in beiden Fällen unverändert, nur der Ablehnungsbereich muss entsprechend modifiziert werden.

Bemerkung (Annahme gleicher Varianzen). Der doppelte t-Test in Satz 4.12 setzt voraus, dass jede Komponente von $X = (X_1, \ldots, X_n)$ und $Y = (Y_1, \ldots, Y_m)$ die gleiche Varianz σ^2 hat. Sind die Varianzen nicht gleich, so ist die Teststatistik in diesem Satz nicht länger t_{n+m-2}-verteilt. Um die Gleichheit der Varianzen zu überprüfen, gibt es ebenfalls statistische Test, etwa den F-Test. Außerdem sei bemerkt, dass es auch eine Variante des doppelten t-Test im Falle ungleicher Varianzen gibt → Behrens-Fisher-Problem und Lösungsansätze.

4.4 Gütefunktion

Mit der Einführung des Niveaus α eines Test haben wir eine Asymmetrie zwischen dem Fehler 1. und dem Fehler 2. Art etabliert. Dabei beschränkt das Niveau den Fehler 1. Art und sagt nichts über den Fehler 2. Art aus. Wir wollen nun einen Begriff für dessen Größe einführen und anhand dessen Tests miteinander vergleichen - denn es ist natürlich wünschenswert, dass ein Niveau-α-Test möglichst selten einen Fehler 2. Art liefert.

Definition 4.14 (Gütefunktion, Trennschärfe). *Es sei ϕ ein statistischer Test für das Testproblem (Θ_0, Θ_1). Dann heißt die Funktion*

$$G_\phi : \begin{cases} \Theta \longrightarrow [0,1] \\ \theta \quad \mapsto \quad \mathbb{E}_\theta[\phi(X)] \end{cases}$$

Gütefunktion. Für $\theta \in \Theta_1$ nennt man den Wert $G_\phi(\theta)$ Trennschärfe (englisch: Power) des Tests.

Bemerkung.

(1) Ist ϕ ein Test zum Niveau α für das Testproblem (Θ_0, Θ_1), so gilt $\sup_{\theta\in\Theta_0} G_\phi(\theta) \leq \alpha$.

(2) Bei Vorliegen von $\theta \in \Theta_1$ ist die Wahrscheinlichkeit für einen Fehler 2. Art gegeben durch
$$\beta_\phi(\theta) := \mathbb{P}_\theta(\text{Fehler 2. Art}) = \mathbb{P}_\theta(\phi(X) = 0) = 1 - G_\phi(\theta).$$

Beispiel 4.15 (Gütefunktion und Stichprobengröße). Wir betrachten den einseitigen Gaußtest ϕ_Z aus Satz 4.8 für $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$. Es sei $\tilde{Z}$ eine unter allen $\mathbb{P}_\mu$ standardnormalverteilte Zufallsvariable. Dann gilt

$$G_{\phi_Z}(\mu) = \mathbb{P}_\mu\left(Z_{\mu_0} \leq \Phi_\alpha^{-1}\right) = \mathbb{P}_\mu\left(\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} + \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}} \leq \Phi_\alpha^{-1}\right) = \mathbb{P}_\mu\left(\tilde{Z} \leq \Phi_\alpha^{-1} + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}}\right).$$

Da die Verteilungsfunktion monoton wachsend ist, übersetzt sich eine gewünschte Mindesttrennschärfe $G_{\phi_Z}(\mu)$ für $\mu < \mu_0$ in eine Ungleichung der Form

$$\frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \geq c.$$

Dies Ungleichung ist erfüllt, falls $n \geq c\sigma(\mu_0 - \mu)^{-1} \geq 0$, d.h. eine Mindestanforderung an die Trennschärfe ist nur dann erfüllt, wenn die Stichprobengröße größer als eine gewisse (hier explizit bestimmbare) untere Schranke ist.

Basierend auf der Gütefunktion formulieren wir nun ein Optimalitätskriterium für Niveau-α-Tests.

Definition 4.16 (Gleichmäßig bester Test). *Ein Niveau-α-Test ϕ für das Testproblem (Θ_0, Θ_1) heißt gleichmässig bester Test zum Niveau α, falls seine Trennschärfe größer ist als diejenige aller anderen Tests zum Niveau α, d.h. falls*

$$G_\phi(\theta) \geq G_\psi(\theta) \quad \text{für alle } \theta \in \Theta_1$$

für jeden Niveau-α-Test ψ für das Testproblem (Θ_0, Θ_1).

4.5 Likelihood-Quotienten-Tests und Neyman-Pearson-Lemma

Nicht für jedes Testproblem muss ein gleichmäßig bester Test existieren. Wir werden jedoch sehen, dass ein solcher existiert, falls wir zwei einfache Hypothesen testen, d.h. falls Θ_0 und Θ_1 einelementige Mengen sind. Darüber hinaus, kann man einen gleichmäßig besten Test in diesem Fall direkt angeben. Davon ausgehend kann man gleichmäßig beste Tests für statistische Modelle mit sogenannten monotonen Dichtequotienten konstruieren. Wir erinnern an dieser Stellen an die Likelihoodfunktion $L(x;\theta)$ aus Definition 2.16.

Definition 4.17 (**Likelihood-Quotienten-Test**). *Es gelte $L(x;\theta_0) > 0$ für alle $x \in \mathbb{R}^n$. Ein Test ϕ für das einfache Testproblem $(\{\theta_0\},\{\theta_1\})$ heißt* Likelihood-Quotienten-Test *(kurz: LQ-Test), falls er von der Form*

$$\phi(x) = \begin{cases} 0, & \text{falls } \frac{L(x;\theta_1)}{L(x;\theta_0)} < c, \\ 1, & \text{falls } \frac{L(x;\theta_1)}{L(x;\theta_0)} \geq c \end{cases}$$

für den sogenannten kritischen Wert $c > 0$ *ist.*

Bemerkung. Für diskrete Modelle gibt die Likelihoodfunktion $L(x;\theta_i)$ die Wahrscheinlichkeit an, unter $\mathbb{P}_{\theta_i}$ die beobachteten Daten x zu erhalten. Somit ist der Quotient $L(x;\theta_1)/L(x;\theta_0)$ umso größer, je weniger wahrscheinlich die Daten unter dem Maß $\mathbb{P}_{\theta_0}$ sind. Daher sprechen große Werte des Likelihood-Quotienten gegen das Vorliegen von $\mathbb{P}_{\theta_0}$. Der Likelihood-Quotienten-Test implementiert diese Heuristik, indem er für zu große Werte des Quotienten (also zu kleiner Wahrscheinlichkeit für die Daten unter $\mathbb{P}_{\theta_0}$ in Relation zur Wahrscheinlichkeit unter $\mathbb{P}_{\theta_1}$) die Nullhypothese ablehnt.

Das nachfolgende Resultat geht auf Neyman und Pearson zurück und ist seit Anfang der 1930er Jahre bekannt.

Theorem 4.18 (**Neyman-Pearson-Lemma**). *Es sei ϕ ein Likelihood-Quotienten Test für das einfache Testproblem $(\{\theta_0\},\{\theta_1\})$ und es gelte*

$$\mathbb{P}_{\theta_0}\left(\frac{L(X;\theta_1)}{L(X;\theta_0)} \geq c\right) = \alpha.$$

Dann ist ϕ ein gleichmäßig bester Test für $(\{\theta_0\},\{\theta_1\})$ zum Niveau α.

Beweis. Nach Voraussetzung gilt

$$\mathbb{E}_{\theta_0}[\phi(X)] = \mathbb{P}_{\theta_0}(\phi(X) = 1) = \mathbb{P}_{\theta_0}\left(\frac{L(X;\theta_1)}{L(X;\theta_0)} \geq c\right) = \alpha,$$

insbesondere ist ϕ ein Niveau-α-Test. Es sei ψ ein weiterer Test für das Testproblem $(\{\theta_0\},\{\theta_1\})$ mit $\mathbb{E}_{\theta_0}[\psi(X)] \leq \alpha$. Um zu zeigen, dass ϕ ein gleichmäßig bester Test ist, muss $\mathbb{E}_{\theta_1}[\phi(X)] - \mathbb{E}_{\theta_1}[\psi(X)] \geq 0$ für diesen beliebig gewählten Niveau-α-Test ψ gelten. Wir zeigen dies für den Fall, dass $\mathbb{P}_{\theta_j}$, $j = 0, 1$, eine Riemann-Dichte besitzen, d.h. insbesondere, dass $L(x;\theta_j) = f_j(x) = \prod_{i=1}^n f_j(x_i)$ (vgl. Definition 2.16). Der Fall mit diskreten

Maßen funktioniert analog, die nachfolgend auftretenden Integrale sind in diesem Fall Summen bzw. Reihen. Es gilt

$$\begin{aligned}\mathbb{E}_{\theta_1}[\phi(X)] - \mathbb{E}_{\theta_0}[\psi(X)] &= \int_{\mathbb{R}^n} (\phi(x) - \psi(x))\, f_1(x) dx \\ &= \int_{\mathbb{R}^n} (\phi(x) - \psi(x))\, (f_1(x) - c f_0(x))\, dx + c \int_{\mathbb{R}^n} (\phi(x) - \psi(x))\, f_0(x) dx.\end{aligned}$$

Wir zeigen nun, dass beide Summanden ≥ 0 sind. Für den zweiten folgt dies einfach, da

$$\int_{\mathbb{R}^n} (\phi(x) - \psi(x))\, f_0(x) dx = \mathbb{E}_{\theta_0}[\phi(X)] - \mathbb{E}_{\theta_0}[\psi(X)] = \alpha - \mathbb{E}_{\theta_0}[\psi] \geq 0.$$

Für den ersten Summanden überlegen wir uns, dass für den Integranden

$$J(x) := (\phi(x) - \psi(x))\, (f_1(x) - c f_0(x)) \geq 0$$

gilt. Dazu unterscheiden wir drei Fälle:

- Ist $\phi(x) = \psi(x)$, so gilt $J(x) = 0$.
- Ist $\phi(x) > \psi(x)$, so muss $\phi(x) = 1$ sein und da ϕ ein LQ-Test mit kritischem Wert c ist, gilt in diesem Fall

  $$c \leq \frac{L(x;\theta_1)}{L(x;\theta_0)} = \frac{f_1(x)}{f_0(x)} \quad \Leftrightarrow \quad c f_0(x) \leq f_1(x).$$

  Damit sind beide Faktoren von J nicht-negativ und $J(x) \geq 0$.
- Ist umgekehrt $\phi(x) < \psi(x)$, so muss $\phi(x) = 0$ gelten, was analog zum vorherigen Fall $c f_0(x) > f_1(x)$ impliziert. Damit sind beide Faktoren von J negativ, entsprechend erhalten wir $J(x) \geq 0$.

□

Beispiel 4.19 (LQ-Test im Binomialmodell). Es sei $0 < p_0 < p_1 < 1$ und $\mathbb{P}_{p_j} = \mathrm{Bin}(n, p_j)$, $i = 1, 2$, und die Beobachtungen x seien die Anzahl der Erfolge. Dann gilt

$$\frac{L(x;p_1)}{L(x;p_0)} = \frac{p_1^x (1-p_1)^{n-x}}{p_0^x (1-p_0)^{n-x}} = \left(\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right)^x \left(\frac{1-p_1}{1-p_0}\right)^n.$$

Da $x \mapsto x/(1-x)$ auf $(0,1)$ monoton wachsend ist, ist der Term in der vorderen Klammer strikt größer Null und der Likelihoodquotient somit monoton wachsend in x. Damit gilt

$$\frac{L(x;p_1)}{L(x;p_0)} \geq c \quad \Leftrightarrow \quad x \geq x^*(c).$$

Um den kritischen Wert c zu bestimmen, müssen wir also die Gleichung

$$\alpha = \mathbb{P}(X \geq x^*(c)) = \sum_{k=x^*(c)}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

nach $x^*(c)$ (und anschließend $x^*(c)$ nach c) auflösen, wobei unter $\mathbb{P}$ gilt $X \sim \mathrm{Bin}(n, p_0)$. Man beachte, dass dies nicht für alle $\alpha \in [0,1]$ möglich ist, da X nur Werte zwischen $0, 1, \ldots, n$ annimmt und $\mathbb{P}(X \geq x^*(c))$ somit auch nur endlich viele verschiedene Werte annehmen kann.

Bemerkung (Randomisierte Tests). Man beachte, dass im Neyman-Pearson-Lemma das Niveau α nicht frei wählbar ist. Die Aussage gilt nur für solche α, für welche die Bestimmungsgleichung $\mathbb{P}_{\theta_0}(L(X;\theta_1)/L(X;\theta_0) \geq c) = \alpha$ eine Lösung c hat. Die Nicht-Lösbarkeit tritt tpyischerweise in diskreten Modellen für manche $\alpha \in [0,1]$ auf, vgl. Beispiel 4.19. Für allgemeine α gibt es nur einen optimalen randomisierten Test: Hier setzt man

$$\phi(x) = \begin{cases} 0, & \text{falls } \frac{L(x;\theta_1)}{L(x;\theta_0)} < c, \\ \gamma(x), & \text{falls } \frac{L(x;\theta_1)}{L(x;\theta_0)} = c, , \\ 1, & \text{falls } \frac{L(x;\theta_1)}{L(x;\theta_0)} > c \end{cases}$$

wobei $\gamma : \mathbb{R}^n \to [0,1]$ eine Funktion ist. Der Wert $\phi(x)$ des Tests ist dann zu lesen als die Wahrscheinlichkeit, sich bei vorliegen der Daten x für die Alternative zu entscheiden. Falls also $L(x;\theta_1)/L(x;\theta_0) = c$, treffen wir die Entscheidung über Ablehnung der Nullhypothese zufällig ($\to$ Mathematische Statistik).

Definition 4.20 (Monotoner Dichtequotient). *Es sei $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ und $T : \mathbb{R}^n \longrightarrow \mathbb{R}$ eine Statistik. Wir sagen, dass $\mathcal{P}$ einen* (streng) monotonen Dichtequotienten *in T hat, falls für alle $\theta \leq \theta'$ eine (streng) monoton wachsende Abbildung $f_{\theta,\theta'} : \mathbb{R} \longrightarrow \mathbb{R}$ existiert mit*

$$\frac{L(x;\theta')}{L(x;\theta)} = f_{\theta,\theta'}(T(x)).$$

Beispiel 4.21. Es seien $X_1, \ldots, X_n$ unabhängig identisch nach $\mathbb{P}_\mu$ verteilt, wobei $\mathbb{P}_\mu = \mathcal{N}(\mu, \sigma^2)$ mit $\sigma^2 > 0$ als gegeben vorausgesetzt ist. Dann gilt

$$L(x_1, \ldots, x_n; \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right)$$

und entsprechend

$$\frac{L(x,\mu')}{L(x;\mu)} = \exp\left(-\frac{1}{2\sigma^2}(\mu - \mu')\sum_{i=1}^n x_i + \frac{\mu^2 - (\mu')^2}{2\sigma^2}\right).$$

Für $\mu \leq \mu'$ ist dies eine streng monoton wachsende Funktion in $T(x) = \sum_{i=1}^n x_i$ (oder auch in $\tilde{t}(x) = \overline{x}_n$).

Lemma 4.22 (Erhalt der Monotonie). *Es sei $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$ eine Familie von Wahrscheinlichkeitsmaßen mit monotonem Dichtequotienten in T und $g : \mathbb{R} \longrightarrow \mathbb{R}$ eine monoton wachsende Funktion mit $\mathbb{E}_\theta[g(T(X))] < \infty$ für alle $\theta \in \mathbb{R}$. Dann ist die Abbildung*

$$\theta \mapsto \mathbb{E}_\theta[g(T(X))]$$

monoton wachsend.

Beweis. Der Beweis wurde für den Spezialfall eines diskreten Maßes und $T(x) = x$ in Aufgabe 3 auf Blatt 06 geführt. Der allgemeine Fall folgt analog. □

Korollar 4.23 (**Gleichmäßig beste Tests für einseitige Testprobleme**). *Es sei $\{\mathbb{P}_\theta^{\otimes n} : \theta \in \mathbb{R}\}$ eine Familie mit streng monotonem Dichtequotienten in T und $\theta_0 \in \mathbb{R}$. Wir setzen*

$$\phi(x) = \begin{cases} 0, & \text{falls } T(x) < k, \\ 1, & \text{falls } T(x) \geq k \end{cases}$$

und nehmen an, dass k so gewählt ist, dass $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$. Dann ist ϕ ein gleichmäßig bester Test für das Testproblem $(\{\theta \leq \theta_0\}, \{\theta > \theta_0\})$ zum Niveau α.

Beweis. Wir wählen zunächst $\theta_1 > \theta_0$. Dann existiert eine streng monoton wachsende Funktion f_{θ_0,θ_1} mit $L(x) := L(x;\theta_1)/L(x;\theta_0) = f_{\theta_0,\theta_1}(T(x))$. Setzen wir $c = f_{\theta_0,\theta_1}(k)$, so erhalten wir

$$T(x) < k \iff L(x) < c \quad \text{und} \quad T(x) \geq k \iff L(x) \geq c.$$

Damit ist ϕ ein LQ-Test mit kritischem Wert c für das Testproblem $(\{\theta_0\}, \{\theta_1\})$ und da nach Voraussetzung $\mathbb{P}_{\theta_0}(L(x) \geq c) = \mathbb{P}_{\theta_0}(T(x) \geq k) = \alpha$, handelt es sich nach dem Neyman-Pearson-Lemma 4.18 um einen gleichmäßig besten Test zum Niveau α, d.h. $\mathbb{E}_{\theta_1}[\phi(X)] \geq \mathbb{E}_{\theta_1}[\psi(X)]$ für jeden weiteren Niveau-α-Test ψ. Da diese Konstruktion für alle $\theta_1 \in \{\theta > \theta_0\}$ gilt, folgt, dass ϕ ein gleichmäßig bester Test für das Testproblem $(\{\theta_0\}, \{\theta > \theta_0\})$ zum Niveau α ist.
Da die Funktion $x \mapsto \mathbb{1}_{[k,\infty)}(x)$ monoton wachsend ist, folgt mit Lemma 4.22

$$\sup_{\theta \leq \theta_0} \mathbb{E}_\theta[\phi(X)] = \sup_{\theta \leq \theta_0} \mathbb{E}_\theta[\mathbb{1}_{[k,\infty)}(T(X))] = \mathbb{E}_{\theta_0}[\mathbb{1}_{[k,\infty)}(T(X))] = \alpha$$

und ϕ ist von Niveau α für die Nullhypothese $\{\theta \leq \theta_0\}$. Damit folgt die Behauptung. □

Beispiel 4.24 (**Einseitiger Gaußtest ist gleichmäßig bester Test**). Der einseitige Gauß-Test aus Satz 4.8(b) hat die Form

$$\phi(X) = \begin{cases} 0, & \text{falls } \overline{X}_n < \mu_0 + \Phi_{1-\alpha}^{-1}\sqrt{\sigma^2/n}, \\ 1, & \text{falls } \overline{X}_n \geq \mu_0 + \Phi_{1-\alpha}^{-1}\sqrt{\sigma^2/n}. \end{cases}$$

Da nach Beispiel 4.21 im n-fachen Normalverteilungsmodell mit bekannter Varianz einen streng monotonen Dichtequotienten in $T(X) = \overline{X}_n$ hat und

$$\mathbb{P}_{\mu_0}\left(\overline{X}_n \geq \mu_0 + \Phi_{1-\alpha}^{-1}\sqrt{\sigma^2/n}\right) = \mathbb{P}_{\mu_0}\left(\frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}} \geq \Phi_{1-\alpha}^{-1}\right) = \alpha$$

folgt nach Korollar 4.23, dass es sich beim einseitigen Gauß-Test um einen gleichmäßig besten Test handelt.

Bemerkung (Weitere beste Tests). Man kann auch zeigen, dass der zweiseitige Gauß-Test ein gleichmäßig bester Test ist, sowie der einseitige Test aus Satz 4.8(c). Außerdem handelt es sich beim t-Test aus Satz 4.9 um einen gleichmäßig besten Test. Da hier σ^2 ebenfalls unbekannt ist und die Parametermenge somit $\mathbb{R} \times (0, \infty)$ ist (dies ist keine total geordnete Menge), muss unsere bisherige Theorie für einen Beweis jedoch erst noch erweitert werden → bedingte Tests (Mathematische Statistik).

4.6 Beispiel eines nicht-parametrischen Test: Test auf Verteilungsgleichheit

Ziel dieses Abschnitts ist es, eine nicht-parametrische Alternative zum doppelten t-Test vorzustellen, den sogenannten Wald-Wolfowitz-Runs-Test. Hierbei handelt es sich um ein Zwei-Stichprobenproblem: Wir beobachten $X_1, \dots, X_m$, welche unabhängig und identisch nach $\mathbb{P}_{\theta_X}$ verteilt seien und $Y_1, \dots, Y_n$, welche ebenfalls unabhängig und identisch verteilt mit Verteilung $\mathbb{P}_{\theta_Y}$ sind. Ziel ist es dann

$$H_0 : \theta_X = \theta_Y \quad \text{gegen} \quad H_1 : \theta_X \neq \theta_Y$$

zu testen. Hier können θ_X, θ_Y jedoch aus einer unendlich-dimensionalen Menge stammen, so könnte X_1 etwa eine Riemann-Dichte θ_X und Y_1 eine Riemann-Dichte θ_Y besitzen.

Bevor wir uns diesem Problem widmen, wollen wir zunächst eine 0-1-Folge auf *Zufälligkeit* testen. Formal sei dazu $E = \{x \in \{0,1\}^n : x_1 + \dots + x_n = n_1\}$ und $n_0 := n - n_1$ die Anzahl der Nullen. Für $x \in E$ definieren wir die Anzahl der Runs als

$$r(x) := 1 + \sum_{i=2}^{n} \mathbb{1}_{\{x_i \neq x_{i-1}\}}.$$

So gilt etwa $r((0,1,1,0,0,0,1,0,0,1,1)) = 6$. In einer zufälligen Folge würden wir erwarten, dass die Zahl der Runs nicht zu groß und nicht zu klein ist, denn anschaulich ist keine der Folgen $(0,0,0,0,0,1,1,1,1,1)$ oder $(0,1,0,1,0,1,0,1,0,1)$ zufällig auch wenn beide aus gleich vielen Nullen und Einsen bestehen. In der Tat gilt das folgende Resultat:

Satz 4.25 (Verteilung der Runs unter Zufälligkeit). *Es sei $\mathbb{P}$ die Gleichverteilung auf $(E, \mathcal{P}(E))$. Dann gilt für die Anzahl der Runs*

$$\mathbb{P}(r(X) = r) = \begin{cases} 2\frac{\binom{n_0-1}{r/2-1}\binom{n_1-1}{r/2-1}}{\binom{n_0+n_1}{n_0}}, & \text{für } r \text{ gerade,} \\ \frac{\binom{n_0-1}{(r-1)/2}\binom{n_1-1}{(r-3)/2}+\binom{n_0-1}{(r-3)/2}\binom{n_1-1}{(r-1)/2}}{\binom{n_0+n_1}{n_0}}, & \text{für } r \text{ ungerade.} \end{cases}$$

Beweis. Sei zunächst r gerade. Dann gibt es genau $r/2$ Runs bestehend aus Nullen und $r/2$ Runs bestehend aus Einsen. Wir betrachten zunächst die Anzahl der Runs mit Nullen und bestimmen die Zahl der Möglichkeiten, die n_0 möglichen Nullen auf $r/2$ verschiedene Runs der Länge ≥ 1 aufzuteilen. Weiter nehmen wir an, dass X mit einer Null starte und überlegen, dass sich jedes solche X als Folge von $r/2$ vielen $|$ und n_0 vielen Nullen schreiben lässt (etwa $0|000|0|00|$). Hierbei steht jeder $|$ für einen Run aus Einsen und es ist klar, dass die Folge mit $|$ endet, wenn sie mit 0 begonnen hat. Da zwischen zwei $|$ mindestens eine Null stehen muss (da zwischen zwei Einser-Runs immer ein Run mit Nullen liegt), erhalten wir durch Streichen einer Null zwischen zwei $|$ und vor dem ersten eine bijektive Abbildung

$$\Big\{x \in \{0,|\}^{n_0+r/2} : x_1 = 0, x_{n_0+r/2} = |, \#\{j : x_j = 0\} = n_0, (x_j, x_{j+1}) \neq (|,|)\ \forall j\Big\}$$
$$\longrightarrow \{x \in \{0,|\}^{n_0} : x_{n_0} = |, \#\{j : x_j = 0\} = n_0 - r/2, \}$$

(so wird 0|000|0|00| zu |00||0|). Die Anzahl der Elemente in der letzten Menge entspricht nun der Anzahl, $r/2-1$ mal | auf n_0-1 Stellen zu verteilen und diese Anzahl ist bekanntermaßen als $\binom{n_0-1}{r/2-1}$ gegeben. Analog gibt es $\binom{n_1-1}{r/2-1}$ Möglichkeiten, die $r/2$ Runs mit Einsen aufzuteilen, womit die Anzahl aller Folgen, welche mit Null beginnen und aus $r/2$ Runs Nullen und Einsen bestehen gerade $\binom{n_0-1}{r/2-1}\binom{n_1-1}{r/2-1}$ ist. Ebenso viele Folgen mit entsprechender Anzahl Runs gibt es, welche mit Eins starten (daher die 2 in der Behauptung des Satzes). Dividiert man diese Anzahl durch die Anzahl aller möglichen 0-1-Folgen der Länge n_0+n_1 mit n_0 vielen Nullen, erhält man die Behauptung.
Ist r ungerade, gibt es entweder $(r+1)/2$ Runs mit Nullen und $(r-1)/2$ Runs mit Einsen oder umgekehrt, wobei die Folge immer mit der Ziffer begonnen werden muss, von der mehr Runs vorhanden sind. Dann führt eine analoge kombinatorische Überlegung wie oben zum Ergebnis. □

Korollar 4.26 (Runs-Test). *Es seien $X_1,\dots,X_n \in \{0,1\}$ mit $X_1+\cdots+X_n = n_1$ und $\alpha \in [0,1]$. Weiter sei R eine Zuvallsvariable mit der in Satz 4.25 spezifizierten Verteilung und $k,l \in \mathbb{N}$ so gewählt, dass $\mathbb{P}(B \le k) \le \alpha/2$ und $\mathbb{P}(B \ge l) \le \alpha/2$. Dann ist*

$$(X_1,\dots,X_n) \mapsto \mathbb{1}_{\{1,\dots,k\}\cup\{l,\dots,n\}}(r(X))$$

ein Niveau-α-Test für das Testproblem

$$H_0 : X \text{ rein zufällig} \quad \text{gegen } H_1 : X \text{ nicht rein zufällig.}$$

Beweis. Unter der Nullhypothese ist $r(X)$ wie in Satz 4.25 verteilt und es gilt

$$\mathbb{E}_{H_0}\left[\mathbb{1}_{\{1,\dots,k\}\cup\{l,\dots,n\}}(r(X))\right] \le \mathbb{P}_{H_0}\left(r(X)\le k\right) + \mathbb{P}_{H_0}\left(r(X)\ge l\right) \le \alpha.$$

□

Bemerkung (Asymptotische Verteilung von $r(X)$). Es gelte $n_0 = n_0(n) \longrightarrow \infty$, $n_1 = n_1(n) \longrightarrow \infty$ mit $n_0/n \longrightarrow p$ und $n_1/n \longrightarrow q = 1-p$. Dann ist die Statistik

$$\frac{r(X) - 2npq}{2pq\sqrt{n}}$$

approximativ $\mathcal{N}(0,1)$-verteilt. Man beachte, dass dies keine direkte Konsequenz aus dem zentralen Grenzwertsatz ist, da die Zufallsvariablen $\mathbb{1}_{\{X_i \neq X_{i-1}\}}$ für $i = 2,\dots,n$ nicht unabhängig sind. Für große n kann diese asymptotische Verteilung genutzt werden, um die Größen k,l in Korollar 4.26 approximativ über Quantile der Normalverteilung zu bestimmen.

Nun kehren wir zu unserer ursprünglichen Frage zurück: Es seien $X_1,\dots,X_m$ unabhängig und identisch nach $\mathbb{P}_{\theta_X}$ und $Y_1,\dots,Y_n$ unabhängig identisch nach $\mathbb{P}_{\theta_Y}$ verteilt. Dann definieren wir $Z = (Z_1,\dots,Z_{n+m}) = (X,Y) \in \mathbb{R}^{m+n}$ und bezeichnen mit $Z_{(1)},\dots,Z_{(m+n)}$ die Ordnungsstatistiken. Dabei ist $Z_{(i)}$ das Element an i-ter Stellen, wenn man die Einträge von Z von klein nach groß ordnet. Insbesondere gilt also

$$Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(m+n)}.$$

Schließlich führen wir einen weiteren Vektor $W \in \{0,1\}^{m+n}$ ein durch

$$W = \left(\mathbb{1}_{\{Z_{(1)} \in \{X_1,\dots,X_m\}\}}, \dots, \mathbb{1}_{\{Z_{(m+n)} \in \{X_1,\dots,X_m\}\}} \right).$$

Unter der Nullhypothese $H_0 : \theta_X = \theta_Y$ ist W ein rein zufälliger Vektor und damit ist die Anzahl der Runs von W unter H_0 so verteilt wir in Satz 4.25 spezifiziert. Damit erhalten wir den folgenden Test.

Korollar 4.27 (**Wald-Wolfowitz-Runs-Test**). *Es seien X, Y, Z und W wie oben spezifiziert. Weiter sei R eine Zuvallsvariable mit der in Satz 4.25 spezifizierten Verteilung und $k \in \mathbb{N}$ so gewählt, dass $\mathbb{P}(B \leq k) \leq \alpha$. Dann ist*

$$(X, Y) \mapsto \mathbb{1}_{\{0,1,\dots,k\}}(r(W))$$

ein Niveau-α-Test für das Testproblem

$$H_0 : \theta_X = \theta_Y \quad \text{gegen } H_1 : \theta_X \neq \theta_Y.$$

Beweis. Dies folgt direkt aus Satz 4.25 und unser Vorüberlegung, dass W unter H_0 eine rein zufällige 0-1-Folge ist. □

Bemerkung. Man beachte, dass wir die Nullhypothese nur verwerfen, falls zu wenige Runs auftreten, was dafür spricht, dass sich die Punkte aus X und Y im Vektor Z zu Clustern zusammentun. Sind etwa alle Werte von X kleiner als diejenigen von Y, so gilt $W = (1, \dots, 1, 0, \dots, 0)$ und die Anzahl der Runs ist 2.

5 Lineare Modelle

In diesem Abschnitt betrachten wir eine neue Klasse statistischer Modelle, die sogenannten linearen Modelle. Dabei sind unsere Beobachtungen gegeben durch

$$(Y_1, X_1^1, \dots, X_1^d), \dots, (Y_n, X_n^1, \dots, X_n^d) \in \mathbb{R} \times \mathbb{R}^d$$

und wir unterstellen, dass

$$Y_i = \beta_0 + \beta_1 X_i^1 + \cdots + \beta_d X_i^d + \epsilon_i \quad \text{für alle } i = 1, \dots, n. \tag{3}$$

Anschaulich gesprochen nehmen wir also an, dass bis auf Fehler ϵ_i ein linearer Zusammenhang zwischen Y_i und $(X_i^1, \dots, X_i^d)$ besteht. Die Variablen Y_i heißen ZIELVARIABLEN und die $(X_i^1, \dots, X_i^d)$ COVARIATEN oder auch UNABHÄNGIGE VARIABLEN.

Für eine kompaktere Notation setzen wir $X_i = (1, X_i^1, \dots, X_i^d)$ und $\beta = (\beta_0, \beta_1, \dots, \beta_d)$, sodass

$$Y_i = \langle X_i, \beta \rangle + \epsilon_i \quad \text{für alle } i = 1, \dots, n.$$

Definieren wir darüber hinaus $Y = (Y_1, \dots, Y_n)$, $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ und die Matrix $X = (x_{ij})_{i=1,\dots,n, j=0,1,\dots d}$ mit x_{ij} als j-tem Eintrag von X_i, so erhalten wir (3) als

$$Y = X\beta + \epsilon.$$

Abbildung 2 zeigt einen Beispieldatensatz für ein EINFACHES REGRESSIONSMODELL, d.h. einem Regressionsmodell mit nur einer Covariaten ($d = 1$).

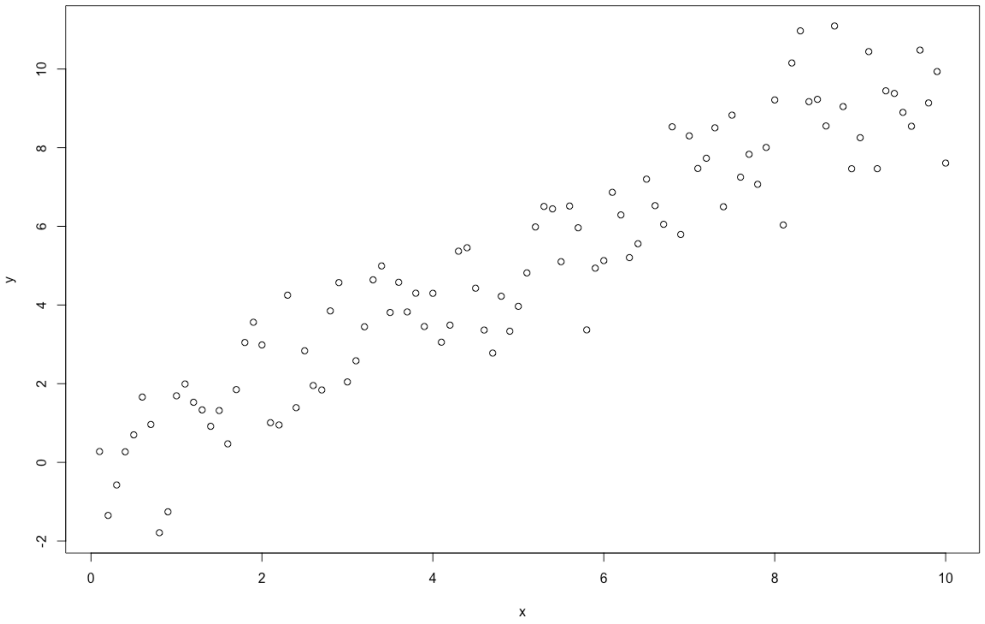

Abbildung 2: Darstellung von Paaren (X, Y), wobei $Y = X + \epsilon, \epsilon \sim \mathcal{N}(0, 1)$.

Der lineare Zusammenhang zwischen Zielvariable und unabhängiger Variable ist in Abbildung 2 bereits gut zu erahnen. Ziel der sogenannten LINEAREN REGRESSION ist es, einen möglichst guten linearen Fit an die Daten zu bestimmen, d.h. eine Gerade, welche die Daten 'möglichst gut' beschreibt. Dies entspricht einer Schätzung von $\beta_0, \beta_1, \ldots, \beta_d$ in (3). In Abbildung 3 ist eine solche Gerade dargestellt.

5.1 Schätzung der Modellparameter

Um einen Schätzer $\hat{\beta}$ für β zu konstruieren, überlegen wir uns, dass wir mithilfe eines solchen Schätzers Vorhersagen $\hat{Y} = X\hat{\beta}$ treffen können (in Abbildung 3 entspricht das den Punkten auf der Regressionsgeraden für entsprechende Werte von X). Ein lineares Modell beschreibt unsere Daten entsprechend gut, wenn die RESIDUEN $Y - \hat{Y} \in \mathbb{R}^n$ klein sind. Nun hat man eine Wahl bei der Metrik - der klassische, auf Gauß zurückgehende, Ansatz besteht darin, die Euklidische Norm $\|\cdot\|_2$ zu verwenden, d.h. die Summe der Residuenquadrate als Referenzmaß zu nutzen. Entsprechend versucht man $\hat{\beta}$ so zu wählen, dass dieser Wert minimal wird.

Definition 5.1 (RSS, Kleinster-Quadrate-Schätzer). *Für Beobachtungen $(Y_i, X_i^1, \ldots, X_i^d)$ und $X_i = (1, X_i^1, \ldots, X_i^d)$ definieren wir die* SUMME DER RESIDUENQUADRATE *(englisch: residual sum of squares) als*

$$\mathrm{RSS}(\beta) := \sum_{k=1}^{n} (Y_i - \langle X_i, \beta \rangle)^2 .$$

Wir definieren den KLEINSTER-QUADRATE-SCHÄTZER $\hat{\beta}$ *als*

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^{d+1}} \mathrm{RSS}(\beta).$$

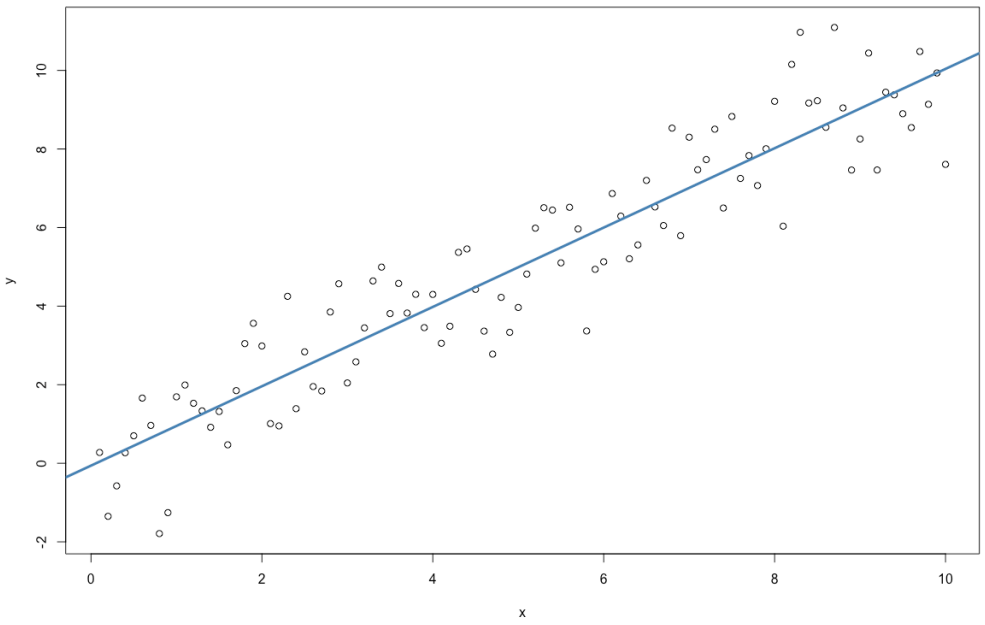

Abbildung 3: Darstellung von Paaren (X, Y), wobei $Y = X + \epsilon, \epsilon \sim \mathcal{N}(0, 1)$ mitsamt der Regressionsgeraden.

Bemerkung (Darstellung RSS). Wir werden im Folgenden immer wieder benutzen, dass (nachrechnen!)

$$\mathrm{RSS}(\beta) = (Y - X\beta)^T(Y - X\beta) = Y^TY - 2Y^TX\beta + \beta^TX^TX\beta.$$

Beispiel 5.2 (Einfache Regression). Bevor wir den allgemeinen Fall besprechen, betrachten wir das Modell (3) mit $d = 1$, d.h.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{für alle } i = 1, \dots, n.$$

In diesem Fall gilt

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

und wir suchen die Minimierer dieser Funktion, d.h.

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname*{argmin}_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Die Bedingungen erster und zweiter Ordnung ergeben, dass in diesem Modell

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_iY_i - n\overline{X}_n\overline{Y}_n}{\sum_{i=1}^n X_i^2 - n\overline{X}_n^2} \quad \text{und} \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}_n$$

gilt (Übungsblatt 07).

Bemerkung (Notation). Man beachte, dass wir in der Notation unterschlagen, dass Y, X, ϵ und auch der Schätzer $\hat{\beta}$ von der Stichprobengröße n abhängen.

Theorem 5.3 (Multiple Regression). *Falls $X^TX \in \mathbb{R}^{(d+1)\times(d+1)}$ invertierbar ist, so ist das Minimum von $\beta \mapsto \mathrm{RSS}(\beta)$ eindeutig und liegt bei*

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

Für die Vorhersage $\hat{Y} = X\hat{\beta}$ gilt

$$Y - \hat{Y} = \left(I_n - X(X^TX)^{-1}X^T\right)\epsilon$$

und die Residuen $Y - \hat{Y}$ stehen sowohl auf der Vorhersage $\hat{Y}$ als auch auf den Spalten von X senkrecht.

Beweis. Eine notwendige Bedingung für das Vorliegen eines Minimums ist, dass (nachrechnen!)

$$0 = \frac{1}{2}\nabla\mathrm{RSS}(\beta) = -Y^TX + \beta^TX^TX = \left(X^TX\beta - X^TY\right)^T,$$

was für $\beta = (X^TX)^{-1}X^TY$ erfüllt ist. In der Tat handelt es sich hierbei um ein Minimum, denn die Hessematrix von $\mathrm{RSS}(\beta)$ ist gegeben durch

$$H(\mathrm{RSS}(\beta)) = X^TX$$

und diese Matrix ist positiv definit, denn für alle $y \in \mathbb{R}^{d+1}$ mit $y \neq 0$ gilt

$$y^T(X^TX)y = \langle Xy, Xy\rangle = \|Xy\|_2^2 > 0.$$

Dabei ist ≥ 0 klar und $\|Xy\|_2^2 = 0$ impliziert $Xy = 0$, was wiederum bedeutet, dass $(X^TX)y = 0$ und dies ist aufgrund der Invertierbarkeit von (X^TX) nur für $y = 0$ möglich. Die zweite Behauptung folgt durch direktes Nachrechnen, denn

$$\begin{aligned}
Y - \hat{Y} &= Y - X\hat{\beta} \\
&= \left(I_n - X(X^TX)^{-1}X^T\right)Y \\
&= \left(I_n - X(X^TX)^{-1}X^T\right)(X\beta + \epsilon) \\
&= X\beta + \epsilon - X(X^TX)^{-1}(X^TX)\beta - X(X^TX)^{-1}X^T\epsilon \\
&= \left(I_n - X(X^TX)^{-1}X^T\right)\epsilon.
\end{aligned}$$

Die behaupteten Orthogonalitäten folgen wiederum durch Einsetzen und Ausmultiplizieren, denn es gilt $(X^TX)^T = X^TX$ und somit

$$\begin{aligned}
(Y - \hat{Y})^T\hat{Y} &= Y^TX(X^TX)^{-1}X^TY - Y^TX\left((X^TX)^{-1}\right)^TX^TX(X^TX)^{-1}X^TY \\
&= Y^TX(X^TX)^{-1}X^TY - Y^TX\left((X^TX)^T\right)^{-1}X^TY = 0,
\end{aligned}$$

sowie

$$\begin{aligned}
(Y - \hat{Y})^TX &= Y^TX - Y^TX\left((X^TX)^{-1}\right)^TX^TX \\
&= Y^TX - Y^TX\left((X^TX)^T\right)^{-1}X^TX = 0.
\end{aligned}$$

□

Als nächstes Stellen wir uns die Frage, ob der Schätzer $\hat{\beta}$ erwartungstreu und konsistent ist. Wir werden sehen, dass er beide Eigenschaften hat, sofern die Fehler ϵ_i in (3) gewisse Eigenschaften erfüllen.

Definition 5.4 (Gauß-Markov-Bedingungen). *Wir sagen, dass die Zufallsvariablen $\epsilon_1, \dots, \epsilon_n$ den* Gauss-Markov-Bedingungen *genügen, wenn für ein $\sigma^2 > 0$ gilt, dass*

$$\mathbb{E}_\beta[\epsilon_i] = 0 \quad \textit{und} \quad \mathrm{Cov}_\beta(\epsilon_i, \epsilon_j) = \sigma^2 \delta_{ij}$$

für alle $1 \leq i, j \leq n$.

Bemerkung.

(1) Für die Gleichungen der Gauß-Markov-Bedingungen schreiben wir auch

$$\mathbb{E}_\beta[\epsilon] = 0 \quad \text{und} \quad \mathrm{Cov}_\beta(\epsilon, \epsilon) = \mathbb{E}_\beta[\epsilon\epsilon^T] = \sigma^2 I_n,$$

wobei alle Gleichungen in Vektorschreibweise gelesen werden.

(2) Eine stärkere, aber ebenfalls typische Annahme ist, dass $\epsilon_1, \dots, \epsilon_n$ unabhängig und identisch $\mathcal{N}(0, \sigma^2)$-verteilt sind.

Bemerkung (Kovarianzmatrix). Wir schreiben im Folgenden für Zufallsvektoren $X \in \mathbb{R}^m$ und $Y \in \mathbb{R}^n$

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY^T] - \mathbb{E}[X]\mathbb{E}[Y]^T \in \mathbb{R}^{m \times n},$$

wobei die Erwartungswerte komponentenweise zu verstehen sind. Ist $A \in \mathbb{R}^{k \times n}$ eine Matrix und $b \in \mathbb{R}^k$, so gilt $\mathbb{E}[AY] = A\mathbb{E}[Y]$ und (nachrechnen!)

$$\mathrm{Cov}(AY + b, AY + b) = A\mathrm{Cov}(Y, Y)A^T.$$

Bemerkung (Spur einer Matrix). Sei $A = (a_{ij})_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix. Dann ist die Spur von A definiert als $\mathrm{Tr}(A) = \sum_{i=1}^n a_{ii}$.

Theorem 5.5 (Erwartungstreue und Konsistenz von $\hat{\beta}$). *Es gelten die Gauß-Markov-Bedingungen. Dann gilt:*

(a) $\mathbb{E}_\beta[Y] = X\beta$ und $\hat{\beta}$ ist ein erwartungstreuer Schätzer für β.

(b) $\mathrm{Cov}_\beta(\hat{\beta}, \hat{\beta}) = \sigma^2(X^T X)^{-1}$.

(c) Ist zusätzlich $\mathrm{Tr}((X^T X)^{-1}) \longrightarrow 0$ für $n \to \infty$, so ist die Folge von Parameterschätzer $(\hat{\beta}_n)_{n \in \mathbb{N}}$ basierend auf n Beobachtungen eine konsistente Schätzfolge für β.

Beweis. (a) Es ist klar, dass $E_\beta[Y] = \mathbb{E}_\beta[X\beta + \epsilon] = X\beta + \mathbb{E}_\beta[\epsilon] = X\beta$ unter den Gauß-Markov-Bedingungen. Mit der Darstellung von $\hat{\beta}$ aus Theorem 5.3 erhalten wir dann

$$\mathbb{E}_\beta[\hat{\beta}] = (X^T X)^{-1} X^T \mathbb{E}_\beta[Y] = (X^T X)^{-1} X^T X\beta = \beta.$$

(b) Wieder mit der Darstellung $\hat{\beta} = (X^TX)^{-1}X^TY$ aus Theorem 5.3 und $((X^TX)^{-1})^T = ((X^TX)^T)^{-1} = (X^TX)^{-1}$ erhalten wir

$$\begin{aligned}\mathrm{Cov}_\beta(\hat{\beta},\hat{\beta}) &= (X^TX)^{-1}X^T\mathrm{Cov}_\beta[Y,Y]X(X^TX)^{-1}\\ &= (X^TX)^{-1}X^T\mathrm{Cov}_\beta[\epsilon,\epsilon]X(X^TX)^{-1}\\ &= (X^TX)^{-1}X^T\sigma^2I_nX(X^TX)^{-1}\\ &= (X^TX)^{-1}X^TX\sigma^2I_n(X^TX)^{-1}\\ &= \sigma^2(X^TX)^{-1}.\end{aligned}$$

(c) Aus Teil (b) folgt zunächst, dass $\mathrm{Var}_\beta(\hat{\beta}_i) = \mathrm{Cov}_\beta(\hat{\beta},\hat{\beta})_{ii} = \sigma^2((X^TX)^{-1})_{ii}$. Wir hatten bereits im Beweis von Theorem 5.3 gesehen, dass X^TX positiv definit ist und daher ist auch $(X^TX)^{-1}$ positiv definit (nachrechnen!). Insbesondere gilt für den i-ten Basisvektor e_i in $\mathbb{R}^{d+1}$, dass

$$0 < e_i^T(X^TX)^{-1}e_i = ((X^TX)^{-1})_{ii}.$$

Damit impliziert $\mathrm{Tr}((X^TX)^{-1}) \longrightarrow 0$, dass $\mathrm{Var}_\beta(\hat{\beta}_i) \longrightarrow 0$ für $i = 0, 1, \ldots, d$. Mit $\mathbb{E}_\beta[\hat{\beta}_i] = \beta_i$ und der Chebychev-Ungleichung erhalten wir dann

$$\mathbb{P}_\beta\left(\left|\hat{\beta}_i - \beta_i\right| > \delta\right) \leq \frac{\mathrm{Var}_\beta(\hat{\beta}_i)}{\delta^2} \longrightarrow 0,$$

d.h. $\hat{\beta}_i \longrightarrow_{\mathbb{P}_\beta} \beta_i$. Die Konsistenz des Vektors folgt dann aus Lemma 2.8 für die Identitätsfunktion.

□

Neben den Modellparametern $\beta = (\beta_0, \beta_1, \ldots, \beta_d)$ kann man auch den Parameter σ^2 der Gauß-Markov-Bedingungen schätzen.

Satz 5.6 (Ein Schätzer für σ^2). *Es gelten die Gauß-Markov-Bedingungen und wir definieren*

$$\hat{\sigma}^2 := \frac{1}{n-d-1}\mathrm{RSS}(\hat{\beta}) = \frac{1}{n-d-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Dann gilt:

(a) $\hat{\sigma}^2$ ist ein erwartungstreuer Schätzer für σ^2.

(b) Sind $\epsilon_1, \ldots, \epsilon_n$ unabhängig und identisch verteilt mit $\mathbb{E}_\beta[\epsilon_1^4] < \infty$, so ist die Folge von Schätzern $(\hat{\sigma}_n^2)_{n\in\mathbb{N}}$ basierend auf n Beobachtungen konsistent.

Beweis. (a) Aus Theorem 5.3 erhalten wir zunächst, dass

$$\begin{aligned}\mathrm{RSS}(\hat{\beta}) &= (Y-\hat{Y})^T(Y-\hat{Y})\\ &= \epsilon^T(I_n - X(X^TX)^{-1}X^T)^T(I_n - X(X^TX)^{-1}X^T)\epsilon\\ &= \epsilon^T(I_n - X(X^TX)^{-1}X^T)\epsilon,\end{aligned}$$

wobei der letzte Schritt benutzt, dass $I_n - X(X^TX)^{-1}X^T$ symmetrisch und idempotent (d.h. für eine quadratische Matrix A, dass $A^2 = A$) ist (nachrechnen!). Damit berechnen wir nun

$$\begin{aligned}\mathbb{E}_\beta[\mathrm{RSS}(\hat\beta)] =& \mathbb{E}_\beta\left[\epsilon^T(I_n - X(X^TX)^{-1}X^T)\epsilon\right]\\ &= \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}_\beta\left[\epsilon_i(I_n - X(X^TX)^{-1}X^T)_{ij}\epsilon_j\right]\\ &= \sigma^2\sum_{i=1}^n (I_n - X(X^TX)^{-1}X^T)_{ii}\\ &= \sigma^2\mathrm{Tr}\left(I_n - X(X^TX)^{-1}X^T\right)\\ &= \sigma^2\left(\mathrm{Tr}(I_n) - \mathrm{Tr}\left(X^TX(X^TX)^{-1}\right)\right)\\ &= \sigma^2(n-d-1),\end{aligned}$$

wobei der vorletzte Schritt nutzt, dass $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$ und $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ für Matrizen A, B. Der letzte Schritt wiederum nutzt, dass $X^TX \in \mathbb{R}^{(d+1)\times(d+1)}$.

(b) Nach dem schwachen Gesetz der großen Zahlen gilt $\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \longrightarrow_{\mathbb{P}_\beta} \sigma^2$ und da $\frac{n-d-1}{n} \to 1$ erhalten wir hieraus

$$\frac{1}{n-d-1}\epsilon^T\epsilon \longrightarrow_{\mathbb{P}_\beta} \sigma^2.$$

Da $(X^TX)^{-1}$ positiv definit ist (vgl. Beweis von Theorem 5.5(c)), erhalten wir $\epsilon^TX(X^TX)^{-1}X^T\epsilon \geq 0$ und mit einer analogen Rechnung wie in Teil (a) dann

$$\begin{aligned}\mathbb{E}_\beta\left[\left|\epsilon^TX(X^TX)^{-1}X^T\epsilon\right|\right] &= \mathbb{E}_\beta\left[\epsilon^TX(X^TX)^{-1}X^T\epsilon\right]\\ &= \sigma^2\mathrm{Tr}\left(X(X^TX)^{-1}X^T\right)\\ &= \sigma^2\mathrm{Tr}\left(XX^T(X^TX)^{-1}\right)\\ &= \sigma^2(d+1).\end{aligned}$$

Insbesondere folgt mit der Markov-Ungleichung

$$\mathbb{P}_\beta\left(\frac{1}{n-d-1}\left|\epsilon^TX(X^TX)^{-1}X^T\epsilon\right| > \delta\right) \leq \frac{\mathbb{E}_\beta\left[\left|\epsilon^TX(X^TX)^{-1}X^T\epsilon\right|\right]}{(n-d-1)\delta} \longrightarrow 0.$$

Aus der ersten Rechnung in Teil (a) erhalten wir die Darstellung

$$\hat\sigma^2 = \frac{1}{n-d-1}\left(\epsilon^T\epsilon - \epsilon^TX(X^TX)^{-1}X^T\epsilon\right)$$

und die Konsistenz folgt nun mithilfe von Lemma I.5.13.

□

5.2 Das Bestimmtheitsmaß R^2

Zwar können wir mit dem Wert $\mathrm{RSS}(\hat\beta)$ messen, wie nah unsere berechnete Regressionsgerade an den tatsächlichen Daten liegt, jedoch ist dieser Wert schwierig zu interpretieren: So wächst der der $\mathrm{RSS}(\hat\beta)$ zum Beispiel mit der Anzahl der Datenpunkte an und es ist daher nicht klar, welche Werte wünschenswert sind. Um den Fit der Regressionsgeraden an die Daten zu beschreiben, wird häufig das sogenannte Bestimmtheitsmass R^2 genutzt.

Definition 5.7 (Bestimmtheitsmaß). *Im Regressionsmodell ist das* Bestimmtheitsmass *definiert als*

$$R^2 := \frac{\left(\sum_{i=1}^n (Y_i - \overline{Y}_n)(\hat{Y}_i - \overline{Y}_n)\right)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2 \sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2}.$$

Bemerkung. Das Bestimmtheitsmaß R^2 entspricht dem Quadrate des empirischen Korrelationskoeffizienten der Vektoren Y und $\hat{Y}$.

Satz 5.8 (Darstellung von R^2). *Es gilt*

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2} = 1 - \frac{\mathrm{RSS}(\hat{\beta})}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}$$

Beweis. Zunächst gilt mit der Notation $E_n = (1, 1, \ldots, 1) \in \mathbb{R}^n$, dass

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \overline{Y}_n)(\hat{Y}_i - \overline{Y}_n) &= \left(Y - \overline{Y}_n E_n\right)^T \left(\hat{Y} - \overline{Y}_n E_n\right) \\
&= \left(Y - \hat{Y} + \hat{Y} - \overline{Y}_n E_n\right)^T \left(\hat{Y} - \overline{Y}_n E_n\right) \\
&= \left(\hat{Y} - \overline{Y}_n E_n\right)^T \left(\hat{Y} - \overline{Y}_n E_n\right) \\
&= \sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2,
\end{aligned}$$

wobei der vorletzte Schritt benutzt, dass $(Y - \hat{Y})^T \hat{Y} = (Y - \hat{Y})^T E_n = 0$. Beides ist eine Konsequenz aus Theorem 5.3, nach welchem die Residuen $Y - \hat{Y}$ sowohl auf der Vorhersage $\hat{Y}$ als auch der ersten Spalte von X (welche E_n entspricht) senkrecht stehen. Mit dieser Darstellung folgt nun direkt

$$R^2 = \frac{\left(\sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2\right)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2 \sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2} = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2},$$

also die erste behauptete Darstellung des Satzes. Da $Y - \hat{Y}$ auf E_n senkrecht steht, erhalten wir $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$ und da $Y - \hat{Y}$ auf $\hat{Y}$ senkrecht steht auch $\hat{Y}^T \hat{Y} = (\hat{Y} - Y + Y)^T \hat{Y} = Y^T \hat{Y}$. Daraus folgt

$$\begin{aligned}
\sum_{i=1}^n \left((Y_i - \overline{Y}_n)^2 - (\hat{Y}_i - \overline{Y}_n)^2\right) &= \sum_{i=1}^n \left(Y_i^2 - \hat{Y}_i^2\right) + 2\overline{Y}_n \sum_{i=1}^n (\hat{Y}_i - Y_i) \\
&= Y^T Y - \hat{Y}^T \hat{Y} \\
&= Y^T \left(Y - \hat{Y}\right) \\
&= \left(Y - \hat{Y}\right)^T \left(Y - \hat{Y}\right) \\
&= \mathrm{RSS}(\hat{\beta}).
\end{aligned}$$

Mithilfe der bereits gezeigten ersten Identität folgt nun die zweite, denn

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2} = \frac{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2 - \mathrm{RSS}(\hat{\beta})}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2} = 1 - \frac{\mathrm{RSS}(\hat{\beta})}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}.$$

□

Bemerkung (Interpretation von R^2). Liegt ein Bestimmtheitsmaß von R^2 vor, so sagt, dass die Regressionsgerade einen Anteil von R^2 an der Varianz der Daten erklärt. Dies folgt aus der ersten Darstellung in Satz 5.8, denn die ERKLÄRTE VARIANZ ist gerade $\sum_{i=1}^n (\hat{Y}_i - \overline{Y}_n)^2$, wohingegen der Nenner $\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$ die GESAMTVARIANZ beschreibt.

5.3 Das Gauß-Markov-Theorem

In diesem Abschnitt werden wir zeigen, dass der Kleinste-Quadrate-Schätzer unter allen erwartungstreuen linearen Schätzern derjenige mit der kleinsten Varianz ist. Diese Aussage ist als GAUSS-MARKOV-THEOREM bekannt.
Für $\ell \in \mathbb{R}^{d+1}$ ist $\ell^T \beta = \sum_{i=1}^{d+1} \ell_i \beta_{i-1}$ eine Linearkombination der Koeffizienten β. Für jeden Schätzer $\hat{\beta}$ ist dann $\ell^T \hat{\beta}$ ein naheliegender Schätzer für diese Linearkombination $\ell^T \beta$. Da $\ell^T \beta \in \mathbb{R}$ können wir die Varianz für verschiedene Schätzer dieses Wertes vergleichen.

Definition 5.9 (BLUE). *Für jedes $\ell \in \mathbb{R}^{d+1}$ sei ein $c_\ell \in \mathbb{R}^n$ gegeben. Im Regressionsmodell heißt jeder Schätzer $Y \mapsto c_\ell^T Y$* LINEAR. *Er heißt* ERWARTUNGSTREU *(für β), falls*

$$\mathbb{E}_\beta[c_\ell^T Y] = \ell^T \beta$$

für jedes $\ell \in \mathbb{R}^{d+1}$. Er heißt BESTER LINEARER ERWARTUNGSTREUER SCHÄTZER *(engl. best linear unbiased estimator, kurz BLUE), falls $\mathrm{Var}_\beta(c_\ell^T Y) \leq \mathrm{Var}_\beta(d_\ell^T Y)$ für jeden anderen linearen erwartungstreuen Schätzer $Y \mapsto d_\ell^T Y$.*

Beispiel 5.10. *Setzen wir $c_\ell = X(X^T X)^{-1} \ell$ und nehmen an, dass die Gauß-Markov-Bedingungen gelten, so gilt nach Theorem 5.5*

$$\mathbb{E}_\beta[c_\ell^T Y] = \ell^T \mathbb{E}_\beta[\hat{\beta}] = \ell^T \beta.$$

und der Schätzer $Y \mapsto c_\ell^T Y = \ell^T \hat{\beta}$ ist ein linearer erwartungstreuer Schätzer.

Theorem 5.11 (Gauß-Markov-Theorem). *Es gelten die Gauß-Markov-Bedingungen. Dann ist der Schätzer $Y \mapsto \ell^T (X^T X)^{-1} X^T Y = \ell^T \hat{\beta}$ ein BLUE.*

Beweis. Es sei $Y \mapsto d_\ell^T Y$ ein weiterer erwartungstreuer linearer Schätzer. Aus der Erwartungstreue folgt

$$\ell^T \beta = \mathbb{E}_\beta[d_\ell^T Y] = d_\ell^T X \beta,$$

also $X^T d_\ell = \ell$ für alle $\ell \in \mathbb{R}^{d+1}$. Mithilfe von Theorem 5.5(b) erhalten wir

$$\begin{aligned} \mathrm{Var}_\beta(d_\ell^T Y) - \mathrm{Var}_\beta(\ell^T \hat{\beta}) &= d_\ell^T \mathrm{Cov}_\beta(Y, Y) d_\ell - \ell^T \mathrm{Cov}_\beta(\hat{\beta}, \hat{\beta}) \ell \\ &= \sigma^2 d_\ell^T d_\ell - \sigma^2 d_\ell^T X (X^T X)^{-1} X^T d_\ell \\ &= \sigma^2 d_\ell^T \left(I_n - X(X^T X)^{-1} X^T \right) d_\ell. \end{aligned}$$

Im Beweis von Satz 5.6(a) haben wir gesehen, dass $\mathrm{RSS}(\hat{\beta}) = \epsilon^T(I_n - X(X^TX)^{-1}X^T)\epsilon$. Da die linke Seite dieser Gleichung nicht-negativ ist und sie für jeden Fehlervektor $\epsilon \in \mathbb{R}^n$ gilt, folgt, dass $I_n - X(X^TX)^{-1}X^T$ eine positiv semi-definite Matrix ist. Entsprechend folgt

$$\mathrm{Var}_\beta(d_\ell^T Y) - \mathrm{Var}_\beta(\ell^T\hat{\beta}) \geq 0$$

und damit die Behauptung. □

Bemerkung. Die Betrachtung von erwartungstreuen Schätzern für $\ell^T\beta$ haben wir genutzt, um Varianzen eindimensionaler Größen vergleichen zu können. Man kann alternativ auch zeigen, dass für den Kleinste-Quadrate-Schätzer $\hat{\beta} = (X^TX)^{-1}X^TY$ und einen beliebigen anderen erwartungstreuen Schätzer $\tilde{\beta}$ gilt, dass

$$\mathrm{Cov}_\beta(\tilde{\beta}) - \mathrm{Cov}_\beta(\hat{\beta})$$

eine positiv semi-definite Matrix ist.

5.4 Tests im Regressionsmodell

Zum Abschluss wollen wir noch einen Test entwickeln, der die Frage beantwortet, ob die unabhängigen Variablen X_i^j für ein $j \in \{1, \ldots, d\}$ überhaupt einen Einfluss auf die abhängigen Variablen Y_i haben. Wir wollen also einen Test entwickeln für die Hypothesen

$$H_0 : \beta_j = 0 \quad \text{gegen} \quad H_1 : \beta_j \neq 0.$$

Unter den Gauß-Markov-Bedingungen ist eine naheliegende Statistik die standardisierte Größe

$$\frac{\hat{\beta}_j}{\sqrt{\mathrm{Var}_\beta(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\sqrt{\sigma^2((X^TX)^{-1})_{jj}}},$$

wobei wir die Varianz von $\hat{\beta}_j$ aus Theorem 5.5 erhalten. Da wir σ^2 nicht kennen, lässt sich diese Statistik jedoch nicht auf Basis der Daten berechnen, sodass wir noch σ^2 durch den Schätzer $\hat{\sigma}^2$ aus Satz 5.6 ersetzen. Um die Verteilung dieser Statistik zu berechnen, benötigen wir folgende Lemmata.

Lemma 5.12. *Es sei $Y \sim \mathcal{N}_n(0, \Sigma)$ n-dimensional normalverteilt, wobei $\Sigma^2 = \Sigma$ und $\mathrm{rang}(\Sigma) = r$. Dann gilt $Y^T\Sigma Y \sim \chi_r^2$.*

Beweis. Da $\Sigma^2 = \Sigma$, sind alle Eigenwerte von Σ entweder 0 oder 1 (denn gilt $\Sigma v = \lambda v$, so erhalten wir $\lambda v = \Sigma v = \Sigma^2 v = \lambda^2 v$, also $\lambda^2 = \lambda$, was nur für $\lambda \in \{0, 1\}$ gelten kann). Weiter ist Σ als Kovarianzmatrix symmetrisch, weswegen Σ diagonalisierbar ist, d.h. es existiert eine Orthogonalmatrix O, sodass für $D = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$ gerade

$$\Sigma = ODO^T.$$

Nach der Bemerkung vor Theorem 5.5 gilt $\mathrm{Cov}_\beta(O^TY, O^TY) = O^T\mathrm{Cov}_\beta(Y, Y)O = O^T\Sigma O$, sodass $O^TY \sim \mathcal{N}_n(0, O^T\Sigma O) = \mathcal{N}_n(0, D)$. Da aufgrund der Rangvoraussetzung an Σ genau r Eigenwerte von Σ eins sind, d.h. D genau r-mal die Eins auf der Diagonalen enthält ist $Y^T\Sigma Y = Y^TODO^TY$ die Summe von r quadrierten, unabhängigen standardnormalverteilten Zufallsvariablen. Damit gilt $Y^T\Sigma Y \sim \chi_r^2$. □

Lemma 5.13. *Es seien Y_i jeweils d_i-dimensionale Zufallsvektoren für $i = 1, 2$ mit $(Y_1, Y_2)^T \sim \mathcal{N}_{d_1+d_2}(\mu, \Sigma)$, wobei $\Sigma = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ eine Blockdiagonalmatrix mit $\mathrm{Cov}(Y_1, Y_1) = A$ und $\mathrm{Cov}(Y_2, Y_2) = B$. Dann sind Y_1 und Y_2 unabhängig.*

Beweis. Es genügt es nach Definition $I.4.22$/Bemerkung $I.4.23$ zu zeigen, dass für die gemeinsame Dichte $f_{(Y_1,Y_2)}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2)$ gilt (wobei $y_1 \in \mathbb{R}^{d_1}, y_2 \in \mathbb{R}^{d_2}$). Mit der Determinantenformel für Blockmatrizen erhalten wir $\det(\Sigma) = \det(A)\det(B)$ und es gilt

$$\Sigma^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix}.$$

Aus letzerer Darstellung erhalten wir für $\mu = (\mu_1, \mu_2)^T \in \mathbb{R}^{d_1+d_2}$ (nachrechnen!)

$$\begin{aligned} &\left((y_1, y_2)^T - (\mu_1, \mu_2)^T\right)^T \Sigma^{-1} \left((y_1, y_2)^T - (\mu_1, \mu_2)^T\right) \\ &\qquad = (y_1 - \mu_1)^T A^{-1} (y_1 - \mu_1) + (y_2 - \mu_2)^T B^{-1} (y_2 - \mu_2). \end{aligned}$$

Folglich erhalten wir für $y = (y_1, y_2)^T$

$$\begin{aligned} f_{(Y_1,Y_2)}(y_1, y_2) &= \frac{1}{(2\pi)^{(d_1+d_2)/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right) \\ &= \frac{1}{(2\pi)^{d_1/2}\sqrt{\det(A)}} \exp\left(-\frac{1}{2}(y_1-\mu_1)^T A^{-1} (y_1-\mu_1)\right) \\ &\qquad \cdot \frac{1}{(2\pi)^{d_2/2}\sqrt{\det(B)}} \exp\left(-\frac{1}{2}(y_2-\mu_2)^T B^{-1} (y_2-\mu_2)\right). \end{aligned}$$

□

Wir kommen nun zum Hauptresultat unseres Abschnitts. Hierfür bezeichne $\mathbb{P}_{\beta_j=0}$ das Maß $\mathbb{P}_\beta$ bei welchem $\beta_j = 0$ für den wahren Parameter β_j gilt.

Theorem 5.14 (Test auf $\beta_j = 0$). *Wir betrachten das Regressionsmodell (3) mit $\epsilon_1, \ldots, \epsilon_n$ unabhängig identisch $\mathcal{N}(0, \sigma^2)$-verteilt. Dann gilt unter $\mathbb{P}_{\beta_j=0}$*

$$T_j(Y) = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2((X^TX)^{-1})_{jj}}} \sim t_{n-d-1},$$

wobei $\hat{\sigma}^2$ in Satz 5.6 gegeben ist. Bezeichnet $t_{n,\alpha}$ das α-Quantil der t_n-Verteilung, so ist

$$Y \mapsto \mathbb{1}_{(-\infty, t_{n-d-1,\alpha/2}) \cup (t_{n-d-1,1-\alpha/2}, \infty)}\left(T_j(Y)\right)$$

ein Niveau-α-Test für das Testproblem

$$H_0 : \beta_j = 0 \quad \text{gegen} \quad H_1 : \beta_j \neq 0.$$

Beweis. Man beachte zunächst, dass unsere Annahme an die Fehler ϵ_i impliziert, dass die Gauß-Markov-Bedingungen gelten. Außerdem müssen wir nur $T_j(Y) \sim t_{n-d-1}$ unter

$\mathbb{P}_{\beta_j=0}$ zeigen, der Rest folgt dann direkt aus der Definition der Quantile.
Mit der expliziten Form von $\hat{\beta}$ aus Theorem 5.3 und der Kovarianz aus Theorem 5.5(b) erhalten wir $\hat{\beta} \sim \mathcal{N}_{d+1}(\beta, \sigma^2(X^TX)^{-1})$. Insbesondere gilt also $\hat{\beta}_j \sim \mathcal{N}(0, \sigma^2((X^TX)^{-1})_{jj})$ unter $\mathbb{P}_{\beta_j=0}$, bzw. äquivalent dazu

$$\frac{\hat{\beta}_j}{\sqrt{\sigma^2((X^TX)^{-1})_{jj}}} \sim \mathcal{N}(0,1).$$

Damit folgt die erste Behauptung des Theorems aus der Definition der t-Verteilung, wenn wir zeigen können, dass $(n-d-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-d-1}$ unter $\mathbb{P}_{\beta_j=0}$ und $\hat{\sigma}^2$ unter diesem Maß unabhängig von $\hat{\beta}_j$ ist.

- **Verteilung:** Da

  $$\beta^TX^T\left(I_n - X(X^TX)^{-1}X^T\right)X\beta = \beta^TX^TX\beta - \beta^TX^TX(X^TX)^{-1}X^TX\beta = 0,$$

  folgt mit der Darstellung von $\mathrm{RSS}(\hat{\beta})$ aus dem Beweis von Satz 5.6

  $$Y^T\left(I_n - X(X^TX)^{-1}X^T\right)Y = \mathrm{RSS}(\hat{\beta}) = (n-d-1)\hat{\sigma}^2.$$

  Wir setzen dazu $A = I_n - X(X^TX)^{-1}X^T$ und bemerken, dass $A^2 = A$. Damit folgt $(n-d-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-d-1}$ aus Lemma 5.12, sobald wir gezeigt haben, dass $AY/\sigma \sim \mathcal{N}_n(0, A)$ und $\mathrm{rang}(I_n - X(X^TX)^{-1}X^T) = n-d-1$. Die erste Aussage folgt direkt, da $\mathrm{Cov}_{\beta_j=0}(Y,Y) = \sigma^2 I_n$ und $A^2 = A$. Für die zweite Eigenschaft bemerken wir, dass A nur die Eigenwerte 0 oder 1 haben kann (vgl. Beweis von Lemma 5.12). Da A symmetrisch, mithin diagonalisierbar und da der Rang invariant unter Basiswechsel ist, entspricht er damit der Anzahl der Eigenwerte $\neq 0$. Es genügt daher zu zeigen, dass die Summe der Eigenwerte $n-d-1$ ergibt und da die Spur einer Matrix ebenfalls invariant unter Basistransformation ist, genügt es entsprechend $\mathrm{Tr}(A) = n-d-1$ zu zeigen (denn damit entspricht die Spur gerade der Summe aller Eigenwerte). Diese hatten wir bereits im Beweis von Satz 5.6(a) berechnet als

  $$\mathrm{Tr}(A) = \mathrm{Tr}\left(I_n - X(X^T)^{-1}X^T\right) = \mathrm{Tr}\left(I_n\right) - \mathrm{Tr}\left(XX^T(X^T)^{-1}\right) = n - (d+1).$$

- **Unabhängigkeit:** Es gilt (nachrechnen!)

  $$\begin{pmatrix}\hat{\beta}\\ (I_n - X(X^TX)^{-1}X^T)Y\end{pmatrix} = \begin{pmatrix}\beta\\ 0\end{pmatrix} + \begin{pmatrix}(X^TX)^{-1}X^T\\ I_n - X(X^TX)^{-1}X^T\end{pmatrix}\epsilon,$$

  sodass der Vektor auf der linken Seite mehrdimensional normalverteilt ist. Da

  $$\begin{aligned}&\mathrm{Cov}_{\beta_j=0}(\hat{\beta}, I_n - X(X^TX)^{-1}X^TY)\\ &\quad= \mathbb{E}_{\beta_j=0}\left[(X^TX)^{-1}X^T\epsilon\epsilon^T(I_n - X(X^TX)^{-1}X^T)\right]\\ &\quad= \sigma^2\left((X^TX)^{-1}X^T - (X^TX)^{-1}X^TX(X^TX)^{-1}X^T\right) = 0,\end{aligned}$$

  folgt aus Lemma 5.13, dass $\hat{\beta}$ und $(I_n - X(X^TX)^{-1}X^T)Y$ unabhängig sind. Da $\hat{\beta}_j$ eine Funktion von $\hat{\beta}$ und wegen

  $$\begin{aligned}&Y^T(I_n - X(X^TX)^{-1}X^T)^T(I_n - X(X^TX)^{-1}X^T)Y\\ &\qquad= Y^T(I_n - X(X^TX)^{-1}X^T)Y = \mathrm{RSS}(\hat{\beta})\end{aligned}$$

  $\hat{\sigma}^2$ eine Funktion von $(I_n - X(X^TX)^{-1}X^T)Y$ ist, folgt die behauptete Unabhängigkeit auch für diese beiden Größen.

□