Mathematische Statistik

von Peter Pfaffelhuber

Version: 10. Februar 2016

WARNUNG: Dieses Skript enthält vermutlich noch viele Fehler. Es wurde teilweise in Eile geschrieben. Für alle Fehler bin ich selbst verantwortlich. Ich hoffe, in Zukunft eine Version mit weniger Fehlern bereit stellen zu können.

Inhaltsverzeichnis

1 Einleitung

Die Mathematische Statistik ist ein eher theoretisches Teilgebiet der Stochastik. Im Gegensatz zur angewandten Statistik, die sich der Methodenentwicklung für Datenanalyse verschrieben hat, geht es in der theoretischen Statistik darum, Eigenschaften solcher Methoden festzustellen, etwa Optimalitätskriterien für Schätzer und Tests. In diesem Kurzskript soll überblicksartig ein kurzer Einblick in dieses Gebiet gegeben werden. Komplettiert wird es (neben den Übungen) durch die Vorstellung statistischer Methoden, die an anderer Stelle im Rahmen dieser Vorlesung zusammengefasst vorgestellt werden.

1.1 Wiederholungen aus der Wahrscheinlichkeitstheorie

Wir setzen Kenntnisse aus den Vorlesungen *Stochastik I*, *Stochastik II* und *Wahrscheinlichkeitstheorie* voraus. Zur Sicherheit jedoch wiederholen wir einige Begriffe, die im Folgenden unerlässlich sein werden. Alle Räume in diesem Skript (z.B. $E, E', E'', \aleph, ...$) seien vollständige und separable metrische Räume, wenn nicht anders angegeben. Im Folgenden sei $(E, \mathcal{A} = \mathcal{B}(E), \mathbb{P})$ ein Wahrscheinlichkeitsraum.

Bemerkung 1.1 (Maß mit Dichte). Das Bildmaß einer Zufallsvariable X, bezeichnet mit $X_*\mathbb{P}$, ist gegeben als $X_*\mathbb{P}(A) = \mathbb{P}(X \in A), A \in \mathcal{B}(E)$. Es hat Dichte p bezüglich λ^n (dem n-dimensionalen Lebesgue-Maß), falls für alle $A \in \mathcal{B}(\mathbb{R})$

$$\mathbb{P}(X \in A) = \int 1_A(x)p(x)\lambda^n(dx).$$

In diesem Fall gilt dann für $f : E \to \mathbb{R}$

$$\mathbb{E}[f(X)] = \int f(x)p(x)\lambda^n(dx),$$

falls eine der beiden Seiten existiert.

Bemerkung 1.2 (Unabhängigkeit, Messbarkeit). Zufallsvariablen X, Y sind (bezüglich $\mathbb{P}$) unabhängig, wenn $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$ für alle $A, B \in \mathcal{B}(E)$. Wir schreiben dann auch $X \perp_{\mathbb{P}} Y$.
Seien X, T Zufallsvariablen. Wir erinnern daran, dass $\sigma(T) \subseteq \mathcal{A}$ die von T erzeugte σ-Algebra ist. Nach einem Satz aus der Wahrscheinlichkeitstheorie ist genau dann X messbar bezüglich $\sigma(T)$ oder T-messbar, falls es eine Funktion g gibt mit $X = g(T)$.

Bemerkung 1.3 (Bedingte Erwartung, bedingte Verteilung). Sei $(E, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, X eine integrierbare Zufallsvariable und $\mathcal{G} \subseteq \mathcal{F}$ eine (Teil-)σ-Algebra. Die bedingte Erwartung von X gegeben $\mathcal{G}$ (bezeichnet mit $\mathbb{E}[X|\mathcal{G}]$) ist die einzige $\mathcal{G}$-messbare Zufallsvariable, für die

$$\mathbb{E}[X, G] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}], G]$$

für alle $G \in \mathcal{G}$ gilt.
Wir erinnern an zwei Eigenschaften der bedingten Erwartung: Für eine weitere σ-Algebra $\mathcal{H} \subseteq \mathcal{G}$ gilt

$$\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}].$$

Für eine weitere reellwertige Zufallsvariable Y gilt

$$\mathbb{E}\big[X\mathbb{E}[Y|\mathcal{G}]\big] = \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]Y\big] = \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\mathbb{E}[Y|\mathcal{G}]\big], \tag{1.1}$$

falls alle Erwartungen existieren.

Wir definieren weiterhin $\mathbb{P}(A|\mathcal{H}) := \mathbb{E}[1_A|\mathcal{H}]$ für $A \in \mathcal{F}$. Angemerkt sei, dass es sich bei $A \mapsto \mathbb{P}(A|\mathcal{H})$ zunächst nicht notwendigerweise um ein Wahrscheinlichkeitsmaß handelt, da bedingte Erwartungen zunächst nur $\mathbb{P}$-fast sicher definiert sind, es also immer Ausnahme-Nullmengen geben kann. Ist jedoch E ein vollständiger und separabler metrischer Raum (was wir hier annehmen werden), so existiert (nach einem Satz aus der Wahrscheinlichkeitstheorie) die *reguläre Version der bedingten Verteilung*, d.h. ein stochastischer Kern κ von E nach E, so dass für $\mathbb{P}$-f.a. $\omega \in E$

$$\kappa(\omega, B) = \mathbb{P}(X \in B|\mathcal{G})(\omega).$$

Bemerkung 1.4 (Varianzzerlegung). Neben der bedingten Erwartung kann man auch die bedingte Varianz

$$\mathbb{V}[X|\mathcal{G}] := \mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])^2|\mathcal{G}] = \mathbb{E}[X^2|\mathcal{G}] - \mathbb{E}[X|\mathcal{G}]^2$$

definieren. Es gilt dann die Varianzzerlegung

$$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}[X|\mathcal{G}]] + \mathbb{V}[\mathbb{E}[X|\mathcal{G}]]. \tag{1.2}$$

Bemerkung 1.5 (Notation). Für $x, y \in \mathbb{R}^n$ ist $x^\top y$ das euklidische Skalarprodukt. Allgemeiner bezeichnen wir für $A \in \mathbb{R}^{m\times n}$ und $x \in \mathbb{R}^n$ die Matrixmultiplikation mit $A\,\underline{x}$ und mit $A^\top$ die Transponierte von A.
Wir bezeichnen das n-dimensionale Lebesgue-Maß (auf $\mathbb{R}^n$) mit λ^n, und zur Vereinfachung der Notation n-dimensionale Zählmaß ebenfalls mit λ^n (also ist in diesem Fall $\lambda^n = \sum_{\underline{x}\in\mathbb{Z}^n} \delta_x$). Das n-dimensionale Produktmaß eines Maßes μ bezeichnen wir im Allgemeinen mit μ^n. Etwa ist für die Standardnormalverteilung $\mathcal{N}(0,1)$ die Verteilung einer unabhängigen Stichprobe gerade $\mathcal{N}(0,1)^n$.
Für die vollständigen, separablen metrische Räume (D, r_D), (E, r_E), ... seien $\mathcal{B}(D), \mathcal{B}(E), ...$ die Borel'schen σ-Algebren.
Hat μ die Dichte f bezüglich ν, so schreiben wir $\mu = f\cdot\nu$. Das Dirac-Maß auf $x \in E$ bezeichnen wir mit δ_x. Das Bildmaß von X unter μ bezeichnen wir mit $X_*\mu$.

1.2 Beispiele

Als theoretische Wissenschaft mit Anwendungsbezügen lebt die Statistik von guten Beispielen, anhand denen man die zu entwickelnde Theorie ausprobieren kann. Auch wenn im Verlauf des Skriptes noch weitere Beispiele auftreten werden, sammeln wir hier drei besonders wichtige, die wir zunächst ohne großen Formalismus vorstellen.

Beispiel 1.6 (Beispiel *Bern*). Ein Bernoulli-Experiment besteht aus einer (endlich oder unendlich oft) unabhängig wiederholten Durchführung eines Zufallsexperiments, in dem jede Durchführung entweder einen *Erfolg* oder einen *Misserfolg* liefert. Es wird beschrieben durch einen Zufallsvektor $X = (X_1, X_2, ...)$ und eine Wahrscheinlichkeitsverteilung $\mathbb{P}_\theta$, so dass für $x_i \in \{0,1\}, i = 1, 2, ...$

$$\mathbb{P}_\theta((X_1, ..., X_n) = (x_1, ..., x_n)) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \qquad (Bern\ 0)$$

für ein $\theta \in [0,1]$ gilt. Hierbei ist θ die Wahrscheinlichkeit eines Erfolges (in jeder der Durchführungen). Bei einer gegebenen Folge von Erfolgen und Misserfolgen $x_1, ..., x_n$ wären naheliegende Fragen etwa:

- Wie groß ist θ?
- Ist $\theta = \frac{1}{2}$?

Beispiel 1.7 (Beispiel *Norm*). Der zentrale Grenzwertsatz betont die Bedeutung der Normalverteilung. Für statistische Fragestellungen bedeutet dies, dass man – zumindest approximativ – oft eine Stichprobe von Daten $X_1, ..., X_n$ erhebt, die unabhängig und normalverteilt sind. Da man typischerweise die Stichprobe aus der gleichen Grundgesamtheit zieht, sollten hierbei die Erwartungswerte und Varianzen gleich sein. Hier ist also $X = (X_1, ..., X_n)$ und $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ hat für $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ die Dichte

$$p_{(\mu,\sigma^2)}(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x_i - \mu)^2}{\sigma} \Big) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(- \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma} \Big) \quad (Norm\ 0)$$

bezüglich des Lebesgue-Maßes in $\mathbb{R}^n$. Bei einer Stichprobe $X_1, ..., X_n$ könnte man etwa fragen:

- Wie groß ist μ, wie groß ist σ^2?
- Ist $\mu = \mu_0$ für einen vorgegebenen Wert μ_0, wenn man weiß, wie groß σ^2 ist?
- Ist $\mu = \mu_0$ für einen vorgegebenen Wert μ_0, wenn man nicht weiß, wie groß σ^2 ist?

Beispiel 1.8 (Beispiel *Unif*). Etwas pathologischer klingt folgendes Beispiel: Wir nehmen an, dass Daten $X = (X_1, ..., X_n)$ unabhängig uniform auf $[0, \theta]$ verteilt sind (für ein $\theta \in \mathbb{R}_+$). Das heißt, dass $\mathbb{P}_\theta$ so ist, dass $X_* \mathbb{P}_\theta = p_\theta \cdot \lambda^n$ mit

$$p_\theta(x) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{0 \le x_i \le \theta} = \frac{1}{\theta^n} \mathbb{1}_{\max_{i=1,...,n} x_i \le \theta}$$

für $x \in \mathbb{R}_+^n$. Hier könnte man etwa fragen:

- Wie groß ist θ?

Für alle Beispiele kann man also sagen, dass *Daten* X erhoben wurden, deren Verteilung von einem Parameter θ abhängen. Dies führt zur ersten Definition.

Definition 1.9 (Statistisches Modell). *Sei $(\Omega, \mathcal{A})$ ein Messraum.*

1. *Ein* statistisches Modell (auf E) *ist ein Paar $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$, wobei $X : \Omega \to E$ messbar und $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$ eine Familie von Wahrscheinlichkeitsmaßen auf $\mathcal{A}$ ist. Hierbei heißt $\mathcal{P}$ auch der* Parameterraum*.*
2. *Das statistische Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ heißt* identifizierbar*, wenn $\theta = \theta'$ genau dann, wenn $X_* \mathbb{P}_\theta = X_* \mathbb{P}_{\theta'}$ gilt. (Im weiteren Verlauf werden wir immer annehmen dass statistische Modelle diese Eigenschaft haben.)*

3. *Weiter heißt das statistische Modell* $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ regulär*, falls* $E = \mathbb{R}^n$ *für ein* n *und für alle* $\theta \in \mathcal{P}$

$$X_* \mathbb{P}_\theta = p_\theta \cdot \lambda^n$$

für eine geeignete Dichtefunktion p_θ *bzw. eine geeignete Zähldichte* p_θ *gilt. (Wir erinnern daran, dass wir* λ^n *sowohl für das Lebesgue-Maß in* $\mathbb{R}^n$ *oder das Zählmaß auf* $\mathbb{Z}^n$ *verwenden.) Im ersten Fall sprechen wir von einem regulären stetigen Modell, im zweiten Fall von einem regulären, diskreten Modell.*

Bemerkung 1.10 (Parametrische und nicht-paramatrische statistische Modelle). Ist $\mathcal{P} \subseteq \mathbb{R}^k$ für ein k, so spricht man oft von parametrischen Modellen. Die Idee ist, dass $\mathbb{P}_\theta$ von einem Vektor θ von verschiedenen Parametern abhängt, etwa Mittelwert und Varianz einer Normalverteilung. In allen anderen Fällen spricht man von nicht-parametrischen Modellen. Im allgemeinsten Fall ist $\mathcal{P}$ die Menge aller Wahrscheinlichkeitsmaße auf $\mathcal{B}(\mathbb{R}^n)$.

Beispiel 1.11 (Beispiel *Norm*). Im Fall von Beispiel 1.7 setzen wir $\theta = (\mu, \sigma^2)$, also

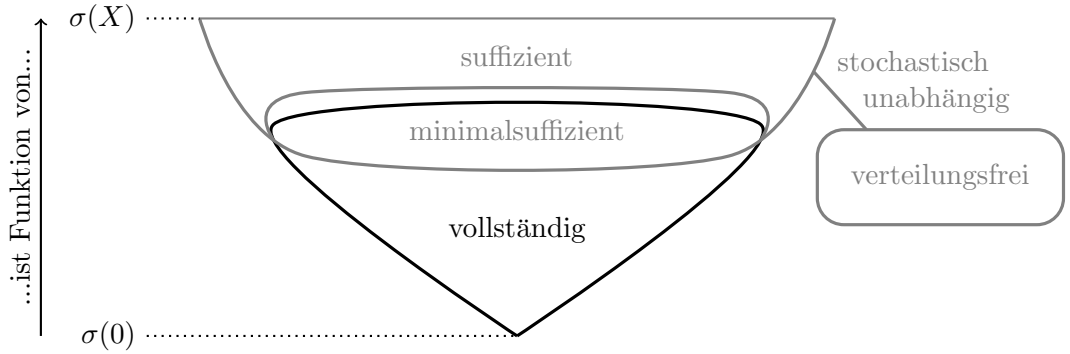
$$(X, \{\mathbb{P}_{\theta=(\mu,\sigma^2)} = \mathcal{N}(\mu, \sigma^2)^n : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}). \tag{Norm 1a}$$

Zu dieser Situation sagen wir auch, dass sowohl μ als auch σ^2 unbekannt sind. In einigen Situation werden wir annehmen, dass wir etwa σ^2 bereits kennen (aber μ nicht). Dann setzen wir für dieses σ^2 das statistische Modell

$$(X, \{\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^n : \theta \in \mathbb{R}\}). \tag{Norm 1b}$$

2 Grundlegende Konzepte

In diesem und dem nächsten Abschnitt führen wir die Konzepte Suffizienz, Minimalsuffizienz, Vollständigkeit und Verteilungsfreiheit von Statistiken ein. (Eine Statistik ist dabei einfach eine $\sigma(X)$-messbare Zufallsvariable, d.h. für eine Abbildung t gilt $T = t(X)$.) Folgende Grafik illustriert kurz die Zusammenhänge.



Die wichtigsten Sätze des Abschnittes sind der Satz von Bahadur, Theorem 2.16, der besagt, dass suffiziente, vollständige Statistiken minimalsuffizient sind. Weiter besagt der Satz von Basu, Theorem 2.19, dass verteilungsfreie und suffiziente Statistiken unabhängig sind.

2.1 Suffizienz

Suffiziente Statistiken sind solche, die alle (zu statistischen Zwecken) nötigen Informationen über die Daten enthält.

Definition 2.1 (Suffiziente Statistik). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell auf E und $T = t(X)$ für $t : E \to E'$ für einen vollständigen metrischen Raum (E', r') messbar. Dann heißt T* suffizient*, falls unter $\mathbb{P}_\theta$ eine reguläre Version der bedingten Verteilung von X gegeben T existiert, die nicht von θ abhängt. Insbesondere hängt also für $A \in \mathcal{B}(E)$ die bedingte Erwartung $\mathbb{P}_\theta(X \in A|T)$ nicht von θ ab.*

In jedem statistischen Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ist X suffizient. Weitere Beispiele für Suffizienz lassen sich mit Hilfe des Fisher-Neyman'schen Faktorisierungssatzes ausmachen, siehe Theorem 2.5.

Bemerkung 2.2 (Umformulierung).

1. Im Falle eines regulären, diskreten statistischen Modells bedeutet die bedingte Unabhängigkeit von X gegeben $T = t(X)$ (unter $\mathbb{P}_\theta$) gerade, dass (für $t = t(x)$)

$$\mathbb{P}_\theta(X = x|T = t) = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(t(X) = t)} = \frac{\mathbb{P}_\theta(X = x)}{\sum_{y:t(y)=t} \mathbb{P}_\theta(X = y)} = \frac{p_\theta(x)}{\sum_{y:t(y)=t} p_\theta(y)}$$

   nicht von θ abhängt. (Im Fall $t \neq t(x)$ ist natürlich $\mathbb{P}_\theta(X = x|t(X) = t) = 0$.)

2. Die reguläre Version der bedingten Verteilung von X gegeben T ist (unter $\mathbb{P}_\theta$) der $\mathbb{P}_\theta$-fast sicher eindeutige stochastische Kern $\kappa_{X,T}$ (von E nach E'), so dass

$$\kappa_{X,T,\theta}(\omega, A) = \mathbb{P}_\theta(X \in A|T)(\omega)$$

   $\mathbb{P}_\theta$-f.s. gilt; siehe auch Bemerkung 1.3. Diese existiert für vollständige und separable metrische Räume (E, r_E) nach einem Satz aus der Wahrscheinlichkeitstheorie. Diese reguläre Version der bedingten Verteilung ist dabei durch die Gleichung

$$\mathbb{E}_\theta[\mathbb{P}_\theta(X \in A|T), X \in B] = \mathbb{P}_\theta(X \in A \cap B)$$

   für $A \in \mathcal{B}(E)$ und $B \in \sigma(T) = t^{-1}(\mathcal{B}(E'))$ eindeutig definiert. Genauer ist $\mathbb{P}_\theta(X \in A|T)$ die $\mathbb{P}_\theta$-f.s. eindeutige, T-messbare Zufallsvariable, die diese Gleichung erfüllt.

Beispiel 2.3 (Beispiel *Bern*). Da das Beispiel 1.6 diskret ist, können wir hier nachrechnen, dass $T := t(X) := \sum_{i=1}^n X_i$ suffizient ist.
Beweis siehe Übung.

Zunächst erscheint es schwierig, suffiziente Statistiken auszumachen. Allerdings geben wir mit Theorem 2.5 eine einfache Charakterisierung von suffizienten Statistiken in Termen der Dichte von X. Erst einmal benötigen wir jedoch ein Lemma.

Lemma 2.4 (Vorbereitung des Fisher-Neyman'schen Faktorisierungssatzes). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein reguläres, stetiges statistisches Modell mit $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$, sowie $t : E \to E'$ messbar und $T = t(X)$. Dann gilt:*

1. *Es gibt $c_1, c_2, ... \geq 0$ mit $\sum_{i=1}^\infty c_i = 1$ und $\theta_1, \theta_2, ... \in \mathcal{P}$, so dass $\mathbb{P}_\theta \ll \nu^* := \sum_{i=1}^\infty c_i \mathbb{P}_{\theta_i}$ für alle $\theta \in \mathcal{P}$.*

2. *Ist T suffizient und $A \in \mathcal{B}(E)$, so gilt $\nu^*(X \in A|T) \overset{\nu^*-f.s.}{=} \mathbb{P}_\theta(X \in A|T)$ für alle $\theta \in \mathcal{P}$.*

3. *Gilt $p_\theta(x) = h(x)g_\theta(t(x))$ für ($\mathbb{P}_\theta$-f.a.) $x \in E, \theta \in \mathcal{P}$, dann gibt es eine T-messbare Version von $\frac{d\mathbb{P}_\theta}{d\nu^*}$, nämlich*
$$\frac{d\mathbb{P}_\theta}{d\nu^*} = \frac{g_\theta(t(x))}{\sum_{i=1}^\infty c_i g_{\theta_i}(t(x))}.$$

4. *Falls $\frac{d\mathbb{P}_\theta}{d\nu^*}$ für alle $\theta \in \mathcal{P}$ nur von $t(x)$ abhängt, so ist $T = t(X)$ suffizient.*

Beweis. 1. Sei $P = p \cdot \lambda^n$ ein Wahrscheinlichkeitsmaß auf $\mathbb{R}^n$ mit $p > 0$ (etwa eine n-dimensionale, nicht-entartete Normalverteilung). Sei weiter

$$\mathcal{Q} = \Big\{ \sum_{i=1}^\infty c_i \mathbb{P}_{\theta_i} \text{ für } \theta_1, \theta_2, ... \in \mathcal{P} \text{ und } c_1, c_2, ... \geq 0 \text{ mit } \sum_{i=1}^\infty c_i = 1 \Big\}$$

die Menge der Konvexkombinationen aus $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$. Wir setzen

$$\mathcal{C} := \Big\{ C \in \mathcal{B}(E) : \exists Q \in \mathcal{Q} : Q(C) > 0 \text{ und } \frac{dQ}{dP}|_C > 0 \Big\}.$$

Dann ist $\mathcal{C}$ nicht leer und $0 < \sup_{C \in \mathcal{C}} P(C) =: c \leq 1$. Wähle nun $C_1, C_2, ... \in \mathcal{C}$ mit $\sup_i P(C_i) = c$ und $D := \bigcup_i C_i$ sowie $Q_1, Q_2, ... \in \mathcal{Q}$ so, dass $Q_n(C_n) > 0$ und $\frac{dQ_n}{dP}|_{C_n} > 0$. Wir wählen
$$\nu^* = \sum_{i=1}^\infty 2^{-i} Q_i \in \mathcal{Q}.$$

Dann gilt sowohl $\nu^*(D) > 0$ als auch $d\nu^*/dP|_D = \sum_{i=1}^\infty 2^{-i} dQ_n/dP|_D > 0$ und damit $D \in \mathcal{C}$. Wir müssen zeigen, dass für $\theta \in \mathcal{P}$ aus $\nu^*(A) = 0$ folgt, dass $\mathbb{P}_\theta(A) = 0$. Wir setzen $C := \{d\mathbb{P}_\theta/dP > 0\}$ und schreiben

$$\begin{aligned}\mathbb{P}_\theta(A) &= \mathbb{P}_\theta(A \cap D) + \mathbb{P}_\theta(A \cap D^c \cap C^c) + \mathbb{P}_\theta(A \cap D^c \cap C) \\ &\leq \int_{A \cap D} \frac{d\mathbb{P}_\theta}{dP} \cdot \Big(\frac{d\nu^*}{dP}\Big)^{-1} d\nu^* + \mathbb{P}_\theta(C^c) + \mathbb{P}_\theta(C \cap D^c).\end{aligned}$$

Die ersten beiden Summanden verschwinden. Angenommen, $\mathbb{P}_\theta(C \cap D^c) > 0$, dann wäre $C \cup D \in \mathcal{C}$ und $P(C \cup D) = P(D) + P(C \cap D^c) > P(D)$ im Widerspruch zur Maximalität von $P(D)$. Damit folgt $\mathbb{P}_\theta(A) = 0$.
2. Sei T suffizient. Für $A \in \mathcal{B}(E)$ hängt $\mathbb{P}_\theta(X \in A|T)$ ($\mathbb{P}_\theta$-f.s.) nicht von θ ab, es gibt also eine Funktion ψ mit $\mathbb{P}_\theta(X \in A|T) \overset{\mathbb{P}_\theta-\text{fs}}{=} \psi(A, T)$ für alle θ. Deshalb gilt auch

$$\nu^*(X \in A|T) = \sum_{i=1}^\infty c_i \mathbb{P}_{\theta_i}(X \in A|T) \overset{\nu^*-\text{fs}}{=} \psi(A, T) \overset{\mathbb{P}_\theta-\text{fs}}{=} \mathbb{P}_\theta(X \in A|T).$$

Die Behauptung folgt nun auch ν^*-f.s., wenn man ψ auf $\mathbb{P}_\theta$-Nullmengen geeignet definiert.
3. Es gilt
$$\frac{d\nu^*}{d\lambda^n} = \sum_{i=1}^\infty c_i \frac{d\mathbb{P}_{\theta_i}}{d\lambda^n} = \sum_{i=1}^\infty c_i h(x) g_{\theta_i}(t(x))$$

und weiter, da $\mathbb{P}_\theta \ll \nu^*$ für alle $\theta \in \mathcal{P}$ (d.h. falls $\nu^*(A) = 0$, so ist zwar $d\nu^*/d\lambda^n$ auf A nicht invertierbar, aber es gilt $d\mathbb{P}_\theta/d\lambda^n = 0$) und

$$\frac{d\mathbb{P}_\theta}{d\nu^*} = \frac{d\mathbb{P}_\theta}{d\lambda^n}\Big(\frac{d\nu^*}{d\lambda^n}\Big)^{-1} = \frac{g_\theta(t(x))}{\sum_{i=1}^\infty c_i g_{\theta_i}(t(x))}.$$

Damit ist $\frac{d\mathbb{P}_\theta}{d\nu^*}$ eine Funktion von $t(x)$ und die Behauptung folgt.
4. Für $A \in \mathcal{B}$, $B \in \sigma(T)$ zeigen wir

$$\mathbb{E}_\theta[\mathbb{P}_\theta(X \in A|T), X \in B] = \mathbb{E}_\theta[\mathbb{P}_{\nu^*}(X \in A|T), X \in B].$$

Dann ist nämlich $\mathbb{P}_\theta(X \in A|T)$ ($\mathbb{P}_\theta$-fs) unabhängig von θ und T ist suffizient. Wir schreiben, da $\frac{d\mathbb{P}_\theta}{d\nu^*}$ nach Voraussetzung messbar bezüglich $\sigma(T)$ ist,

$$\begin{aligned}\mathbb{E}_\theta[\mathbb{P}_\theta(X \in A|T), X \in B] &= \mathbb{E}_\theta[1_{X\in A}1_{X\in B}] = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}, X \in A \cap B\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\mathbb{P}_{\nu^*}(X \in A|T), X \in B\Big] \\ &= \mathbb{E}_\theta[\mathbb{P}_{\nu^*}(X \in A|T), X \in B]\end{aligned}$$

und die Behauptung ist gezeigt. □

Theorem 2.5 (Fisher-Neyman'scher Faktorisierungssatz). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell und $T = t(X)$ mit $t : \mathbb{R}^n \to \mathbb{R}^m$ messbar. Dann sind äquivalent:*

1. *T ist suffizient,*
2. *Es gibt $g_\theta : \mathbb{R}^m \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}$, so dass*

$$p_\theta(x) = g_\theta(t(x))h(x).$$

Bemerkung 2.6 (Diskreter Fall). Im diskreten Fall ist der Beweis einfach. '2. ⇒ 1.': Nach Bemerkung 2.2 gilt $\mathbb{P}_\theta(X = x|t(X) = t) = 0$ für $t \neq t(x)$, was unabhängig von θ ist. Für $t = t(x)$ hingegen ist unter 2.

$$\mathbb{P}_\theta(X = x|t(X) = t) = \frac{g_\theta(t(x))h(x)}{\sum_{y:t(y)=t} g_\theta(t(y))h(y)} = \frac{g_\theta(t(x))h(x)}{g_\theta(t(x))\sum_{y:t(y)=t} h(y)} = \frac{h(x)}{\sum_{y:t(y)=t} h(y)},$$

was ebenfalls unabhängig von θ ist. Für '1. ⇒ 2.' setzen wir

$$g_\theta(t) := \mathbb{P}_\theta(t(X) = t), \qquad h(x) = \mathbb{P}_\theta(X = x|t(X) = t(x)).$$

Dann ist $h(x)$ nach Voraussetzung unabhängig von θ und es gilt

$$p_\theta(x) = \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x, t(X) = t(x)) = h(x)g_\theta(t(x))$$

und die Behauptung ist gezeigt.

Beweis im stetigen Fall. '1. $\Rightarrow$ 2.': Sei ν^* wie in Lemma 2.4.1. Nach Lemma 2.4.2 gilt $\mathbb{P}_\theta(X \in A|T) = \nu^*(X \in A|T)$ (ν^*-f.s.) für alle $\theta \in \mathcal{P}$ und für $A \in \mathcal{B}(E)$. Wir schreiben nun mittels (1.1)

$$\begin{aligned}\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}, X \in A\Big] &= \mathbb{P}_\theta(X \in A) = \mathbb{E}_\theta[\mathbb{P}_\theta(X \in A|T)] \\ &= \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\nu^*(X \in A|T)\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\nu^*(X \in A|T)\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\Big|T\Big]\Big] \\ &= \mathbb{E}_{\nu^*}\Big[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\Big|T\Big], X \in A\Big]\end{aligned}$$

wegen (1.1). Da A beliebig war und $T = t(X)$, gilt

$$\frac{d\mathbb{P}_\theta}{d\nu^*} = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\theta}{d\nu^*}\Big|t(X)\Big] =: g_\theta(t(X))$$

und mit $h := \frac{d\nu^*}{d\lambda^n}$ gilt

$$p_\theta(x) = \frac{d\mathbb{P}_\theta}{d\nu^*}\frac{d\nu^*}{d\lambda^n} = g_\theta(t(x))h(x).$$

'2. $\Rightarrow$ 1.' Dies ist eine Folgerung aus Lemma 2.4.3 und Lemma 2.4.4. □

Beispiel 2.7 (Beispiel *Norm*). Bei normalverteilten Daten wie in Beispiel 1.7 und 1.11 schreiben wir für die Dichte

$$\begin{aligned}p_\theta(x) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big(-\sum_{i=1}^n\frac{(x_i-\mu)^2}{2\sigma^2}\Big) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big(-\frac{n\mu^2}{2\sigma^2}\Big)\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\Big).\end{aligned}$$

Im Fall von unbekanntem μ und σ^2, d.h. im statistischen Modell (Norm 1a) folgt mit dem Fisher-Neyman Kriterium, dass $T = \Big(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\Big)$ suffizient ist. Weiter können wir hier ablesen, dass im statistischen Modell (Norm 1b) (bei bekanntem σ^2) bereits $\sum_{i=1}^n X_i$ suffizient ist.

Beispiel 2.8 (Beispiel *Unif*). Für die Situation aus Beispiel 1.8 sehen wir mit dem Fisher-Neyman'schen Kriterium, dass $T := t(X) := \max_{i=1,\dots,n} X_i$ suffizient ist. Beweis siehe Übung.

Definition 2.9 (Minimalsuffizienz). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell und $T = t(X)$ für $t : E \to E'$ eine suffiziente Statistik. Dann heißt T* minimalsuffizient*, falls es für jede weitere suffiziente Statistik $U = u(X)$ für $u : E \to E''$ eine messbare Abbildung $g : E'' \to E'$ gibt mit $T \overset{\mathbb{P}_\theta - fs}{=} g(U)$ für alle $\theta \in \mathcal{P}$.*

Für obige Beispiele ist es nicht einfach nachzuprüfen, ob die angegebenen suffizienten Statistiken auch minimalsuffizient sind. Folgender Satz erleichtert aber das Auffinden minimalsuffizienter Statistiken.

Theorem 2.10 (Kriterium für Minimalsuffizienz). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell mit $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ und $t : E \to E'$ eine messbare Abbildung, so dass*

$$t(y) = t(x) \iff \text{Es gibt ein } \ell(x,y) > 0, \text{ so dass für alle } \theta \in \mathcal{P}\text{: } p_\theta(y) = p_\theta(x)\ell(x,y).$$

Dann ist $T = t(X)$ minimalsuffizient.

Beweis. Zunächst ist klar, dass die für $t(y) = t(x)$ angegebene Bedingung symmetrisch in x und y, also wohldefiniert ist. (Mit $\ell(x,y) > 0$ ist nämlich auch $\ell(y,x) := 1/\ell(x,y) > 0$.)
Wir zeigen zunächst, dass T suffizient ist. Sei ν^* wie in Lemma 2.4. Es gilt für $t(x) = t(y)$

$$\begin{aligned}\frac{d\mathbb{P}_\theta}{d\nu^*}(x) &= \frac{d\mathbb{P}_\theta}{d\lambda^n}(x) \Big/ \frac{d\nu^*}{d\lambda^n}(x) = \frac{p_\theta(x)}{\sum_{i=1}^\infty c_i p_{\theta_i}(x)} \\ &= \frac{p_\theta(x)\ell(x,y)}{\sum_{i=1}^\infty c_i p_{\theta_i}(x)\ell(x,y)} = \frac{p_\theta(y)}{\sum_{i=1}^\infty c_i p_{\theta_i}(y)} = \frac{d\mathbb{P}_\theta}{d\nu^*}(y).\end{aligned}$$

Damit hängt $\frac{d\mathbb{P}_\theta}{d\nu^*}$ nur von $t(x)$ ab und die Behauptung folgt nach Lemma 2.4.4.
Es folgt nun der Beweis, dass T minimalsuffizient ist. Sei hierzu $U = u(X)$ eine weitere suffiziente Statistik. Nach dem Fisher-Neyman'schen Faktorisierungssatz gibt es g_θ und h, so dass $p_\theta(x) = g_\theta(u(x))h(x)$. (Man kann oBdA annehmen, dass $h > 0$ ist.) Wir müssen zeigen, dass aus $u(x) = u(y)$ folgt, dass $t(x) = t(y)$. Dann nämlich gibt es eine Funktion g mit $t(x) = g(u(x))$. Sei also $u(x) = u(y)$ und damit

$$\frac{p_\theta(y)}{p_\theta(x)} = \frac{g_\theta(u(y))h(y)}{g_\theta(u(x))h(x)} = \frac{h(y)}{h(x)}.$$

Daraus folgt $t(x) = t(y)$ mit der Wahl $\ell(x,y) = h(y)/h(x)$. □

Beispiel 2.11 (Beispiel *Bern*). Wir haben bereits gesehen, dass $T = \sum_{i=1}^n X_i$ suffizient ist. Außerdem gilt $p_\theta(x) = p_\theta(y)$ genau dann, wenn $t(x) = t(y)$. Wählt man also $\ell(x,y) = 1$ in Theorem 2.10, so erhält man die Minimalsuffizienz von T.

Beispiel 2.12 (Beispiel *Unif*). In Beispiel *Unif* ist $t(x) := \max_{i=1,\dots,n} x_i$ suffizient; siehe Beispiel 2.8. Für festes θ ist

$$\frac{p_\theta(y)}{p_\theta(x)} = \frac{\theta^n 1(t(y) \le \theta)}{\theta^n 1(t(x) \le \theta)} = \frac{1(t(y) \le \theta)}{1(t(x) \le \theta)}.$$

Nun $t(x) = t(y)$ genau dann, wenn $\frac{p_\theta(y)}{p_\theta(x)} = 1$ für alle θ. Damit ist $T = t(X)$ minimalsuffizient.

2.2 Vollständigkeit und Verteilungsfreiheit

Manchmal benötigt man suffiziente Statistiken, die weitere Eigenschaften erfüllen. Eine solche Eigenschaft wird nun beschrieben. Sie hilft insbesondere, minimalsuffiziente Statistiken zu finden; siehe Theorem 2.16.

Definition 2.13 (Vollständige Statistik). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell. Eine Statistik $T = t(X)$ für ein $t : E \to E'$ heißt* (beschränkt) vollständig, *falls für alle (beschränkten) messbaren Funktionen g gilt, dass*

$$\mathbb{E}_\theta[g(T)] = 0 \text{ für alle } \theta \in \mathcal{P} \implies g(T) \overset{\mathbb{P}_\theta - fs}{=} 0 \text{ für alle } \theta \in \mathcal{P}.$$

Beispiel 2.14 (X nicht beschränkt vollständig). Wir zeigen nun, dass im Allgemeinen für ein statistisches Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ die Zufallsvariable X nicht vollständig ist. Sei hierzu etwa für ein $n \geq 2$ die Verteilung $\mathbb{P}_\theta = \text{Poi}(\theta)^n$ für $\theta \in \mathcal{P} := \mathbb{R}_+$ die n-dimensionale Poisson-Verteilung mit Parameter θ, d.h. unter $\mathbb{P}_\theta$ hat $X = (X_1, ..., X_n)$ Werte in $\mathbb{R}^n$ und $X_1, ..., X_n$ sind unabhängig und identisch Poisson verteilt zum Parameter $\theta \geq 0$. Für ein beschränktes (aber auf $\mathbb{Z}_+$ nicht konstantes) $f : \mathbb{Z}_+ \to \mathbb{R}$ sei $g : \mathbb{Z}_+^n \to \mathbb{R}$ gegeben durch $g(x_1, ..., x_n) := f(x_1) - f(x_2)$. Dann ist offensichtlich (da $X_1 \sim X_2$ für alle $\theta \in \mathcal{P}$)

$$\mathbb{E}_\theta[g(X)] = \mathbb{E}_\theta[f(X_1)] - \mathbb{E}_\theta[f(X_2)] = 0,$$

aber $\mathbb{P}_\theta(f(X_1) \neq f(X_2)) > 0$, d.h. $g(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$ gilt nicht. Deshalb ist X nicht beschränkt vollständig (und damit nicht vollständig).

Beispiel 2.15 (Poisson-verteilte Statistik T). Für ein statistisches Modell $\{\mathbb{P}_\theta : \theta \in \mathbb{R}_+\}$ sei $T_*\mathbb{P}_\theta = \text{Poi}(\theta)$. Dann ist T beschränkt vollständig.
Denn: Sei g messbar und beschränkt, so dass

$$\mathbb{E}_\theta[g(T)] = e^{-\theta} \sum_{t=0}^{\infty} g(t) \frac{\theta^t}{t!} = 0$$

für alle $\theta \geq 0$. Dies ist die Potenzreihenentwicklung einer Funktion $\theta \mapsto 0$. Da diese Funktion analytisch ist, ist die Darstellung eindeutig und damit gilt $g(t) = 0$ für $t = 0, 1, 2, ...$, d.h. $g(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$.

Theorem 2.16 (Satz von Bahadur). *Sei $\{\mathbb{P}_\theta : \theta \in \mathbb{R}_+\}$ ein statistisches Modell und $T = t(X)$ für $t : E \to \mathbb{R}^k$ für ein $k \geq 0$ beschränkt vollständig und suffizient. Dann ist T minimalsuffizient.*

Beweis. Wir setzen $t(x) = (t_1(x), ..., t_k(x))$. Sei $U = u(X)$ für $u : E \to E''$ eine weitere suffiziente Statistik. Wir müssen zeigen, dass T eine Funktion von U ist. Wir setzen $S = s(T)$ mit $s = (s_1, ..., s_k)$ und $s_i(t) = (1+e^{t_i})^{-1}$. Dann ist $s : \mathbb{R}^k \to \mathbb{R}^k$ injektiv und S ist beschränkt. Wir setzen für $i = 1, ..., k$

$$H_i(U) = \mathbb{E}_\theta[S_i(T)|U],$$
$$L_i(T) = \mathbb{E}_\theta[H_i(U)|T].$$

Da T und U suffizient sind, hängen diese Funktionen nicht von θ ab und mit S sind auch H_i und L_i beschränkt, $i = 1, ..., k$. Weiter gilt für $\theta \in \mathcal{P}$

$$\mathbb{E}_\theta[S_i(T) - L_i(T)] = \mathbb{E}_\theta[H_i(U) - \mathbb{E}_\theta[H_i(U)|T]] = 0$$

und da T beschränkt vollständig ist folgt $S_i(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} L_i(T)$ für alle $\theta \in \mathcal{P}$. Wir zerlegen nun die Varianz von $L_i(T)$ und von $H_i(U)$ mittels

$$\mathbb{V}_\theta[L_i(T)] = \mathbb{E}_\theta[\mathbb{V}_\theta[L_i(T)|U]] + \mathbb{V}_\theta[\mathbb{E}_\theta[L_i(T)|U]] = \mathbb{E}_\theta[\mathbb{V}_\theta[L_i(T)|U]] + \mathbb{V}_\theta[H_i(U)],$$
$$\mathbb{V}_\theta[H_i(U)] = \mathbb{E}_\theta[\mathbb{V}_\theta[H_i(U)|T]] + \mathbb{V}_\theta[L_i(T)]$$

und damit $\mathbb{V}_\theta[L_i(T)|U] = \mathbb{V}_\theta[H_i(U)|T] \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$, also auch $\mathbb{V}_\theta[S_i(T)|U] \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$. Demnach gilt

$$S_i(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} \mathbb{E}_\theta[S_i(T)|U] \overset{\mathbb{P}_\theta-\text{fs}}{=} H_i(U).$$

Also gilt $T_i = s_i^{-1}(H_i(U))$ für $i = 1, ..., k$ und T ist eine Funktion von U. Damit folgt die Behauptung. □

In Theorem 2.32 werden wir sehen, dass die suffizienten Statistiken aus Beispiel *Bern* und *Norm* vollständig sind. Beispiel *Unif* behandeln wir zunächst getrennt.

Beispiel 2.17 (Beispiel *Unif*). Wir haben bereits (siehe Beispiele 1.8, 2.8 und 2.12) gesehen, dass $T := t(X) := \max_{i=1,\dots,n} X_i$ minimalsuffizient ist. Wir zeigen nun, dass T auch vollständig ist. Also würde hier zusammen mit der Suffizienz aus Beispiel 2.8 und dem Satz von Bahadur nochmals folgen, dass T minimalsuffizient ist.
Unter $\mathbb{P}_\theta$ gilt

$$\mathbb{P}_\theta(T \leq t) = 1_{t\leq\theta}\Big(\frac{t}{\theta}\Big)^n,$$

also ist $t \mapsto 1_{t\leq\theta} n t^{n-1}/\theta^n$ die Dichte von T. Gilt nun $\mathbb{E}_\theta[g(T)] = 0$ für alle $\theta > 0$, so folgt

$$\int_0^\theta t^{n-1} g(t)dt = 0.$$

Da dies für alle $\theta \geq 0$ gilt, ist

$$\int_a^b t^{n-1} g(t)dt = 0.$$

Dies ist nur möglich, wenn $g \overset{\lambda-\text{fü}}{=} 0$.

Wir kommen nun zum Begriff der verteilungsfreien Statistik. Eine solche beinhaltet keine (statistisch verwendbaren) Informationen über die Daten. Das Hauptresultat über verteilungsfreie Statistiken ist Theorem 2.19, der den Zusammenhang zu suffizienten Statistiken herstellt; siehe auch die Abbildung am Anfang des Kapitels.

Definition 2.18 (Verteilungsfreie Statistik). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell. Eine Statistik $T = t(X)$ für ein $t : E \to E'$ heißt* verteilungsfrei*, falls $T_*\mathbb{P}_\theta$ unabhängig von θ ist. Eine verteilungsfreie Statistik T heißt* maximal*, wenn es für jede andere verteilungsfreie Statistik U eine Funktion g gibt mit $U = g(T)$.*

Theorem 2.19 (Satz von Basu). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell. Weiter sei $T = t(X)$ für $t : E \to E'$ und $U = u(X)$ für $u : E \to E''$.*[1]

1. *Ist T beschränkt vollständig und suffizient sowie U verteilungsfrei, dann gilt $T \perp_{\mathbb{P}_\theta} U$ für alle $\theta \in \mathcal{P}$.*
2. *Angenommen, für alle $\theta, \theta' \in \mathcal{P}$ gibt es eine Menge $A \in \mathcal{B}(E)$ mit $\mathbb{P}_\theta(X \in A), \mathbb{P}_{\theta'}(X \in A) > 0$. Ist $T \perp_{\mathbb{P}_\theta} U$ für alle $\theta \in \mathcal{P}$ und T suffizient, dann ist U verteilungsfrei.*
3. *Sei $T \perp_{\mathbb{P}_\theta} U$ für alle $\theta \in \mathcal{P}$ und U verteilungsfrei. Wenn $\sigma(T, U) = \sigma(X)$, dann ist T suffizient.*

Beweis. 1. Sei $A \in \mathcal{B}(E'')$. Offenbar gilt

$$\mathbb{P}_\theta(U \in A) = \mathbb{E}_\theta[\mathbb{P}_\theta(U \in A|T)].$$

Weiter hängt (wegen der Verteilungsfreiheit von U) weder $\mathbb{P}_\theta(U \in A)$ noch (wegen der Suffizienz von T) die Größe $\mathbb{P}_\theta(U \in A|T)$ von θ ab. Damit ist $g : t \mapsto \mathbb{P}_\theta(U \in A) - \mathbb{P}_\theta(U \in A|T)$ eine beschränkte messbare Funktion (unabhängig von θ) mit $\mathbb{E}_\theta[g(T)] = 0$ für alle $\theta \in \mathcal{P}$.

[1] Wir erinnern an die Schreibweise $X \perp_{\mathbb{P}_\theta} Y$, falls X und Y unter $\mathbb{P}_\theta$ stochastisch unabhängig sind.

Wegen der beschränkten Vollständigkeit von T ist damit $\mathbb{P}_\theta(U \in A) \overset{\mathbb{P}_\theta-\text{fs}}{=} \mathbb{P}_\theta(U \in A|T)$. Dies aber bedeutet die Unabhängigkeit von U und T.
2. Da T suffizient ist, hängt $\mathbb{P}_\theta(U \in A|T)$ für alle $A \in \mathcal{B}(E'')$ nicht von θ ab. Da $\mathbb{P}_\theta(U \in A|T) = \mathbb{P}_\theta(U \in A)$ wegen der Unabhängigkeit von U und T, hängt also $\mathbb{P}_\theta(U \in A)$ nicht von θ ab. Dies ist aber gerade die Verteilungsfreiheit von U, da A beliebig war.
3. Es ist zu zeigen, dass für alle $A \in \mathcal{B}(E)$ gilt, dass $\mathbb{P}_\theta(X^{-1}(A)|T)$ unabhängig von θ ist. In der Tat genügt es mittels einer monotonen Klasse, für einen schnittstabilen Erzeuger $\mathcal{E}$ von $\sigma(X)$ zu zeigen, dass für alle $E \in \mathcal{E}$ $\mathbb{P}_\theta(E|T)$ unabhängig von θ ist.

Sei nun $B \in \mathcal{B}(E')$ und $C \in \mathcal{B}(E'')$. Es gilt

$$\mathbb{P}_\theta(T \in B, U \in C|T) = 1_{T\in B}\mathbb{P}_\theta(U \in C)$$

und dies ist (wegen der Verteilungsfreiheit von U) unabhängig von θ. Da $\{T^{-1}(B) \cap U^{-1}(C) : B \in \mathcal{B}(E'), C \in \mathcal{B}(E'')\}$ schnittstabil ist und nach Voraussetzung $\sigma(X)$ erzeugt, folgt die Aussage. □

Beispiel 2.20 (Beispiel *Norm*). Für festes σ^2 betrachten wir das statistische Modell aus (Norm 1*b*), also das Normalverteilungsmodell mit bekanntem σ^2. Weiter setzen wir

$$T := \overline{X} := \frac{1}{n}\sum_{i=1}^n X_i, \qquad U = \frac{1}{n-1}\sum_{i=1}^n (X_i - T)^2.$$

Aus der Rechnung in Beispiel 2.7 sieht man zusammen mit dem Fisher-Neyman'schen Faktorisierungssatzes (Theorem 2.5), dass T suffizient ist. (Weiter werden wir in Theorem 2.32 sehen, dass T auch vollständig, also sogar minimalsuffizient ist.) Außerdem ist U verteilungsfrei, denn: Der Vektor $(X_1 - T, ..., X_n - T)$ ist unabhängig von θ normalverteilt mit

$$\begin{aligned}\mathbb{E}_\theta[X_i - T] &= 0,\\ \mathbb{COV}_\theta[X_i - T, X_j - T] &= \sigma^2(\delta_{ij} - \tfrac{2}{n} + \tfrac{1}{n}) = \sigma^2(\delta_{ij} - \tfrac{1}{n}).\end{aligned}$$

Nun ist U eine Funktion dieses Vektors, also verteilungsfrei.) Nach dem Theorem von Basu, Theorem 2.19.2, sind diese beiden Vektoren also unabhängig. (Dieses Ergebnis ist auch Teil des bekannten Satzes von Fisher.)

2.3 Exponentialfamilien

Viele Verteilungen, etwa die Normal- Poisson-, Binomial- und Exponentialverteilung, haben eine gemeinsame Struktur, die oftmals direkte Rechnungen ermöglicht. Diese Struktur wird in der folgenden Definition formalisiert.

Definition 2.21 (Exponentialfamilie). *Sei $\mathcal{P} \subseteq \mathbb{R}^k$. Eine Familie $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$ von Wahrscheinlichkeitsmaßen auf $\mathbb{R}^n$ (auf $\mathbb{Z}^n$) heißt* k-parametrige Exponentialfamilie *(mit c, t, d, h) für*

$$c_1, ..., c_k, d : \mathcal{P} \to \mathbb{R}, \qquad t_1, ..., t_k, h : \mathbb{R}^n \to \mathbb{R},$$

falls $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ und

$$p_\theta(x) = h(x) \cdot \exp\Big(\sum_{j=1}^k c_j(\theta)t_j(x) - d(\theta)\Big) = h(x) \cdot \exp\big(c(\theta)^\top t(x) - d(\theta)\big), \qquad x \in \mathbb{R}^n.$$

Gilt insbesondere $c_j(\theta) = \theta_j$, also

$$p_\theta(x) = h(x) \cdot \exp\Big(\theta^\top t(x) - d(\theta)\Big), \qquad x \in \mathbb{R}^n,$$

so sagt man, die Exponentialfamilie sei in kanonischer Form *und wir definieren den* kanonischen Parameterraum

$$\Gamma := \Big\{\theta \in \mathbb{R}^k : \int h(x) e^{\theta^\top t(x)} d\lambda^n(x) < \infty\Big\}.$$

Bemerkung 2.22 (1-parametrige Exponentialfamilie). Für den Spezialfall einer 1-parametrigen Exponentialfamilie gibt es Funktionen c, d, t, h mit

$$p_\theta(x) = h(x) \cdot \exp\Big(c(\theta)t(x) - d(\theta)\Big), \qquad x \in \mathbb{R}^n.$$

Beispiel 2.23 (Beispiel *Norm*). Wir betrachten das statistische Modell (Norm $1a$) für $n = 1$, also $\theta := (\mu, \sigma^2)$ mit $\mathcal{P} = \mathbb{R} \times \mathbb{R}_+$ und $\mathbb{P}_\theta := \mathcal{N}(\mu, \sigma^2)$, und damit

$$p_{(\mu,\sigma^2)}(x) = \exp\Big(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\Big(\frac{\mu}{\sigma^2} + \log(2\pi\sigma^2)\Big)\Big).$$

Also ist die Familie der (ein-dimensionalen) Normalverteilungen eine 2-parametrige Exponentialfamilie mit

$$\begin{aligned} c_1(\mu, \sigma^2) &= \frac{\mu}{\sigma^2}, & t_1(x) &= x, \\ c_2(\mu, \sigma^2) &= -\frac{1}{2\sigma^2}, & t_2(x) &= x^2, \\ h(x) &= 1, & d(\mu, \sigma^2) &= -\frac{1}{2}\Big(\frac{\mu}{\sigma^2} + \log(2\pi\sigma^2)\Big). \end{aligned}$$

Diese ist nun allerdings nicht in kanonischer Form.

Beispiel 2.24 (Beispiel *Bern*). Sei $(X, \{\mathbb{P}_\theta : \theta \in (0,1)\})$ wie in Beispiel 1.6. Dann ist also mit $(Bern\ 0)$

$$\begin{aligned} p_\theta(x) &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= \exp\Big(\sum_{i=1}^n x_i \log\theta + \Big(n - \sum_{i=1}^n x_i\Big)\log(1-\theta)\Big) \\ &= \exp\Big(\log\frac{\theta}{1-\theta}\sum_{i=1}^n x_i + n\log(1-\theta)\Big) \end{aligned}$$

und damit ist $\{\mathbb{P}_\theta : \theta \in (0,1)\}$ eine 1-parametrige Exponentialfamilie.

Beispiel 2.25 (Beispiel *Unif*). Die Familie $\{\mathbb{P}_\theta : \theta \in \mathbb{R}_+\}$ aus Beispiel 1.8 ist keine Exponentialfamilie.

Beispiel 2.26 (Beispiel *Exp*). Sei $\mathcal{P} = \mathbb{R}_+$ und $\mathbb{P}_\theta = p_\theta \cdot \lambda$ die Exponentialverteilung mit Parameter θ. Dann ist

$$p_\theta(x) = 1_{x \geq 0} e^{-\theta x + \log\theta}$$

und damit ist die Familie der Exponentialverteilungen eine 1-parametrige Exponentialfamilie und obige Darstellung ist die kanonische Form.

Beispiel 2.27 (Beispiel *Pois*). Sei $\Lambda = \mathbb{R}_+$. Für die Poisson-Verteilung mit Parameter λ schreiben wir $\mathbb{P}_\lambda = p_\lambda \cdot \mu$ mit

$$p_\lambda(x) = \frac{1_{x\geq 0}}{x!} \exp\big((\log\lambda)x - \lambda\big)$$

Diese ist damit eine 1-parametrige Exponentialfamilie. Um obige Darstellung in kanonische Form zu bringen, setzen wir $\theta := \log\lambda$ und schreiben für $\theta \in \mathbb{R}$

$$p_\theta(x) = \frac{1_{x\geq 0}}{x!} \exp\big(\theta x - e^\theta\big).$$

Bemerkung 2.28 (Sample aus einer Exponentialfamilie). Ist $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$ eine k-parametrige Exponentialfamilie auf $\mathbb{R}^n$ mit Funktionen $c_1, ..., c_k, d, t_1, ..., t_k, h$, und sind $X_1, ..., X_N$ unabhängig und identisch nach $\mathbb{P}_\theta$ verteilt. Dann ist die gemeinsame Verteilung von $X_1, ..., X_N$ eine Exponentialfamilie mit

$$c_1, ..., c_k, Nd, \sum_{i=1}^N t_1 \circ \pi_i, ..., \sum_{i=1}^N t_k \circ \pi_i, \prod_{i=1}^N h \circ \pi_i$$

mit der Projektion $\pi_i(x) = x_i$.
Denn: Als gemeinsame Dichte schreiben wir

$$p_\theta(x_1, ..., x_N) = h(x_1)\cdots h(x_N) \cdot \exp\Big(\sum_{j=1}^k c_j(\theta) \sum_{i=1}^N t_j(x_i) - Nd(\theta)\Big).$$

Proposition 2.29 (Suffiziente Statistik bei Exponentialfamilien).
Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine Exponentialfamilie (mit c, t, d, h). Dann ist die Statistik $T := t(X) = (t_1(X), ..., t_k(X))$ suffizient.

Beweis. Um die Suffizienz von T zu sehen, schreiben wir zunächst

$$p_\theta(x) = h(x) g_\theta(t(x))$$

für

$$g_\theta(t(x)) = \exp(c(\theta)^\top t(x) - d(\theta)).$$

Damit folgt die Aussage aus dem Fisher-Neyman'schen Faktorisierungssatz, Theorem 2.5. □

Beispiel 2.30 (Lineare Regression). Bei der linearen Regression beobachten wir $(x_1, Y_1), ..., (x_n, Y_n)$ (mit $x_i = (x_{i0} = 1, x_{i1}, ..., x_{im}) \in \mathbb{R}^{m+1}$) und gehen davon aus, dass

$$Y_i = x_i\beta + \varepsilon_i \qquad (\text{also } Y = x^\top\beta + \varepsilon)$$

für $\beta_0, ..., \beta_m \in \mathbb{R}$ mit $\varepsilon_1, ..., \varepsilon_n$ unabhängig und identisch nach $\mathcal{N}(0, \sigma^2)$ verteilt. Hierbei sind $x_1, ..., x_n$ bekannt (und fest, also keine Parameter des Modells) und $Y_1, ..., Y_n$ werden als

Zufallsvariablen betrachtet. Wir haben also ein statistisches Modell $(Y, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta = (\beta, \sigma^2) \in \mathbb{R}^{m+2}\})$ und wir schreiben für die gemeinsame Dichte von $Y = (Y_1, ..., Y_n)$

$$\begin{aligned} p_\theta(y) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_i - x_i\beta)^2}{2\sigma^2}\Big) \\ &= \exp\Big(- \frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \frac{1}{\sigma^2}\beta^\top \sum_{i=1}^n y_i x_i - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i\beta)^2 - \tfrac{n}{2}\log(2\pi\sigma^2)\Big) \end{aligned}$$

Wählen wir

$$\begin{aligned} c_j(\beta) &= \frac{\beta_j}{\sigma^2}, & t_j(y) &= \sum_{i=1}^n y_i x_{ij}, & j &= 0, ..., m, \\ c_{m+1}(\beta) &= -\frac{1}{2\sigma^2}, & t_{m+1}(y) &= \sum_{i=1}^n y_i^2, \end{aligned}$$

$$d(\beta) = \frac{1}{2\sigma^2}\sum_{i=1}^n (\beta^\top x_i)^2 + \tfrac{n}{2}\log(2\pi\sigma^2),$$

so sehen wir, dass $\{\mathbb{P}_\theta : \theta \in \mathbb{R}^{m+2}\}$ eine $m+2$-parametrige Exponentialfamilie ist. Weiter ist

$$\Big(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i x_{i1}, ..., \sum_{i=1}^n Y_i x_{im}, \sum_{i=1}^n Y_i^2\Big)$$

suffizient.

Lemma 2.31 (Kanonischer Parameterraum konvex). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine Exponentialfamilie in kanonischer Form (mit c, t, d, h). Dann ist der kanonische Parameterraum konvex und $\theta \mapsto e^{d(\theta)}$ ist konvex.*

Beweis. Sei $\theta, \eta \in \Gamma$ und $0 < \alpha < 1$. Dann gilt, da die Exponentialfunktion konvex ist,

$$\begin{aligned} e^{d(\alpha\theta + (1-\alpha)\eta)} &:= \int h(x) \exp\big((\alpha\theta + (1-\alpha)\eta)^\top t(x)\big) d\lambda^n(x) \\ &\le \int h(x)\big(\alpha e^{\theta^\top t(x)} + (1-\alpha) e^{\eta^\top t(x)}\big) d\lambda^n(x) \\ &= \alpha e^{d(\theta)} + (1-\alpha) e^{d(\eta)}. \end{aligned}$$

Dies zeigt alle Behauptungen. □

Theorem 2.32 (Vollständigkeit der suffizienten Statistik). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine k-parametrige Exponentialfamilie in kanonischer Form (mit (θ, t, d, h)). Ist[2] $\mathbb{R}^k \supseteq \mathcal{P}^\circ \neq \emptyset$, so ist $T = t(X)$ vollständig (und wegen der Suffizienz aus Proposition 2.29 mit Theorem 2.16 auch minimalsuffizient).*

[2] Mit A° bezeichnen wir das Innere der Menge A.

Beweis. Wir beweisen die Behauptung nur im Fall $k = 1$, da die anderen Fälle mit Induktion folgen. Angenommen, T sei nicht vollständig. Dann gibt es $g : \mathbb{R} \to \mathbb{R}$ messbar und $\mathbb{E}_\theta[g(T)] = 0$ für alle $\theta \in \mathcal{P}$, aber $\mathbb{P}_{\theta_0}(g(T) > 0) > 0$ für ein $\theta_0 \in \mathcal{P}^\circ$. (Wir können oBdA θ_0 im Inneren von $\mathcal{P}$ wegen der Stetigkeit der Dichte von T und monotoner Konvergenz wählen.) Anders ausgedrückt heißt das, dass für alle $\theta \in \mathcal{P}$

$$\mathbb{E}_\theta[g^+(T)] = \mathbb{E}_\theta[g^-(T)] \tag{$*$}$$

gilt, aber

$$\mathbb{E}_{\theta_0}[g^+(T)] = \mathbb{E}_{\theta_0}[g^-(T)] =: w \in (0, \infty)$$

für ein $\theta_0 \in \mathcal{P}^\circ$. Wir definieren die beiden Wahrscheinlichkeitsmaße

$$P := \tfrac{1}{w} g^+ \circ T \cdot \mathbb{P}_{\theta_0}, \qquad Q := \tfrac{1}{w} g^- \circ T \cdot \mathbb{P}_{\theta_0}.$$

Nun schreiben wir $(*)$ um in

$$\begin{aligned}
\mathbb{E}_P[e^{(\theta-\theta_0)T}] &= \tfrac{1}{w}\mathbb{E}_{\theta_0}[e^{(\theta-\theta_0)T} g^+(T)] \\
&= \frac{1}{w} \int e^{(\theta-\theta_0)t(x)} g^+(t(x)) h(x) e^{\theta_0 t(x) - d(\theta_0)} d\lambda(x) \\
&= \frac{e^{d(\theta)-d(\theta_0)}}{w} \int g^+(t(x)) h(x) e^{\theta t(x) - d(\theta)} d\lambda(x) \\
&= \frac{e^{d(\theta)-d(\theta_0)}}{w} \mathbb{E}_\theta[g^+(T)] = \frac{e^{d(\theta)-d(\theta_0)}}{w} \mathbb{E}_\theta[g^-(T)] \\
&= \tfrac{1}{w}\mathbb{E}_{\theta_0}[e^{(\theta-\theta_0)T} g^-(T)] \\
&= \mathbb{E}_Q[e^{(\theta-\theta_0)T}].
\end{aligned}$$

Wegen der Eindeutigkeit der Laplace-Transformierten bedeutet dies $P = Q$ und damit auch $g^+ \overset{\mathbb{P}_{\theta_0}-\text{fs}}{=} g^-$, also $g \overset{\mathbb{P}_{\theta_0}-\text{fs}}{=} 0$ und damit (wegen der Form der Exponentialfamilie) auch $g \overset{\mathbb{P}_\theta-\text{fs}}{=} 0$ für alle $\theta \in \mathcal{P}$. □

Bemerkung 2.33 (Exponentialfamilien in nicht-kanonischer Form). Das Ergebnis aus Theorem 2.32 lässt sich durch Umparametrisierung auf Exponentialfamilien übertragen, die nicht in kanonischer Form vorliegen. Man sieht etwa, dass die suffizienten Statistiken aus Beispiel 2.23, 2.24, 2.26 und 2.27 minimalsuffizient sind.

Exponentialfamilien erlauben oft explizite Berechnungen. Dies illustrieren wir an zwei Ergebnissen, von denen wir das erste ohne Beweis angeben.

Lemma 2.34 (Differenzierbarkeit nach θ). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine Exponentialfamilie in kanonischer Form (mit θ, t, d, h) und der kanonische Parameterraum Γ habe nicht-leeres Inneres, $\Gamma^\circ \neq \emptyset$. Weiter sei $\phi : \mathbb{R}^k \to \mathbb{R}$ messbar und $\theta \in \Gamma^\circ$ so, dass $\eta \mapsto \mathbb{E}_\eta[|\phi(t(X)|)e^{\eta^\top t(X)}] < \infty$ für η in einer Umgebung von θ. Dann ist die Abbildung*

$$f : \eta \mapsto \mathbb{E}_\eta[\phi(t(X))e^{\eta^\top t(X)}]$$

in einer Umgebung von θ analytisch mit

$$\frac{\partial f(\eta)}{\partial \eta_i} = \mathbb{E}_\eta[t_i(X)\phi(t(X))e^{\eta^\top t(X)}].$$

Proposition 2.35 (Laplace-Transformierte von Exponentialfamilien). *Sei* $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ *eine* k*-parametrige Exponentialfamilie in kanonischer Form. Ist* $\theta \in \mathcal{P}^\circ$*, dann existiert die Laplace-Transformierte von* $t(X)$ *mit*

$$\mathbb{E}_\theta[e^{\eta^\top t(X)}] = \exp(d(\eta+\theta) - d(\theta)),$$

falls $|\eta|$ *klein genug ist und für* $l_1, ..., l_k \geq 0$ *mit* $l_1 + \cdots + l_k = l$

$$\mathbb{E}_\theta\Big[\prod_{i=1}^k t_i(X)^{\ell_i}\Big] = e^{-d(\theta)} \frac{\partial^\ell}{\partial \eta_1^{\ell_1} \cdots \partial \eta_k^{\ell_k}} e^{d(\eta)}\Big|_{\eta=\theta}.$$

Insbesondere gilt also

$$\mathbb{E}_\theta[t_i(X)] = \frac{\partial d(\eta)}{\partial \eta_i}\Big|_{\eta=\theta},$$
$$\mathbb{COV}_\theta[t_i(X), t_j(X)] = \frac{\partial^2 d(\eta)}{\partial \eta_i \partial \eta_j}\Big|_{\eta=\theta}.$$

Beweis. Wir berechnen

$$\begin{aligned}\mathbb{E}[e^{\eta^\top t(X)}] &= \int \exp(\eta^\top t(x)) \cdot h(x) \cdot \exp(\theta^\top t(x) - d(\theta))\lambda^n(dx) \\ &= e^{d(\eta+\theta)-d(\theta)} \int h(x) \cdot \exp((\theta+\eta)^\top t(x) - d(\theta+\eta))\lambda^n(dx) \\ &= e^{d(\eta+\theta)-d(\theta)},\end{aligned}$$

da das Integral über eine Dichte eins ist. Die zweite Behauptung folgt aus Lemma 2.34. □

2.4 Bayes'sche Modelle

Die Formel von Bayes ist wohlbekannt. Auf ihr beruht der große Zweig der Bayesianischen Statistik. Grundlegend ist hier, dass in einem statistischen Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein Vorwissen über die Möglichkeiten besteht, welcher Parameter $\theta \in \mathcal{P}$ zutrifft. Dies wird in der a-priori-Verteilung zusammengefasst, einer Verteilung auf $\mathcal{P}$.

Definition 2.36 (A-priori-Verteilung, a-posteriori-Verteilung). *Sei* $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ *ein statistisches Modell und* $\mathcal{P}$ *ein vollständiger, separabler metrischen Raum. Eine* a-priori-Verteilung *ist eine Wahrscheinlichkeitsverteilung* π *auf* $\mathcal{P}$*. Ist für alle* $A \in \mathcal{B}(E)$ *die Abbildung* $\theta \mapsto \mathbb{P}_\theta(X \in A)$ *messbar, so wird durch* $(\theta, A) \mapsto \mathbb{P}_\theta(X \in A)$ *ein Markov-Kern von* $\mathcal{P}$ *nach* $\mathcal{B}(E)$ *definiert und für* $\Theta \sim \pi$ *wird durch*

$$P(\Theta \in A, X \in B) := \int_A \pi(d\theta)\mathbb{P}_\theta(X \in B)$$

die gemeinsame Verteilung von Θ *und* X *auf* $\mathcal{B}(\mathcal{P})\otimes\mathcal{B}(E)$ *definiert. Die* a-posteriori-*Verteilung ist dann die reguläre Version der bedingten Verteilung* $P(\Theta \in .|X)$*, also*

$$\pi_x = P(\Theta \in .|X = x).$$

Bemerkung 2.37 (A-posteriori-Verteilung bei regulären Modellen). Ist $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell und $\mathcal{P} \subseteq \mathbb{R}^k$, und ist $\pi = g \cdot \lambda^k$, so hat die gemeinsame Verteilung von Θ und X die Dichte $g(d\theta) \cdot p_\theta(dx)$. Die a-posteriori Verteilung π_x hat dann die Dichte

$$p_x(\theta) = \frac{p_\theta(x)g(\theta)}{\int p_\eta(x)g(\eta)d\eta}.$$

Beispiel 2.38 (Beispiel *Bern*). Wieder ist $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ durch (*Bern* 0) gegeben. Weiter sei die a-priori-Verteilung die Beta-Verteilung $\pi = \beta(k, l)$, d.h. $\pi = p_{kl} \cdot \lambda$ mit[3]

$$p_{kl}(x) = \frac{\Gamma(k)\Gamma(l)}{\Gamma(k+l)} x^{k-1}(1-x)^{l-1} \sim x^{k-1}(1-x)^{l-1},$$

wobei wir mit $\sim$ ausdrücken, dass der restliche Faktor unabhängig von x ist. Mit Bemerkung 2.37 folgt für die Dichte der a-posteriori-Verteilung π_x

$$p_x(\theta) \sim \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\theta^{k-1}(1-\theta)^{l-1} = \theta^{k-1+\sum x_i}(1-\theta)^{l+n-1-\sum x_i}.$$

Dies ist also eine $\beta(k + \sum x_i, l + n - \sum x_i)$-Verteilung.

Das letzte Beispiel lässt sich auf allgemeine Exponentialfamilien verallgemeinern. (Aus Beispiel 2.24 wissen wir, dass das letzte Beispiel eine solche Verteilungsklasse behandelt.)

Proposition 2.39 (Konjugierte Familie bei Exponentialfamilie). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ eine k-parametrige Exponentialfamilie (mit c, t, d, h). Weiter sei $\pi_s = p_s \cdot \lambda^k$ eine a-priori-Verteilung mit*

$$p_s(\theta) = \frac{\exp\Big(\sum_{j=1}^k c_j(\theta)s_j - s_{k+1}d(\theta)\Big)}{\int \exp\Big(\sum_{j=1}^k c_j(\eta)s_j - s_{k+1}d(\eta)\Big)\lambda^k(d\eta)}$$

gegeben. Dann ist die a-posteriori-Verteilung gerade $P(\Theta \in .|X = x) = p_x \cdot \lambda^k$ mit

$$p_x(\theta) = p_{(s_1+t_1,\ldots,s_k+t_k,s_{k+1}+1)}(\theta).$$

Beweis. Wir schreiben $a \sim_x b$, falls a/b nur von x abhängt (d.h. a und b sind proportional). Es gilt

$$\begin{aligned} p_x(\theta) \sim_x p_\theta(x)\pi_s(x) &\sim_x \exp\Big(\sum_{j=1}^k c_j(\theta)(t_j(x) + s_j) - (s_{k+1}+1)d(\theta)\Big) \\ &\sim_x p_{(s_1+t_1,\ldots,s_k+t_k,s_{k+1}+1)}(\theta). \end{aligned}$$

□

Beispiel 2.40 (A-posteriori-Verteilung bei der Normalverteilung). Wir betrachten das Normalverteilungsmodell bei bekanntem σ^2, gegeben durch (Norm 1b). Angenommen, die a-priori-Verteilung π für θ ist selbst eine Normalverteilung, nämlich $\mathcal{N}(a, b^2)$ für $a, b \in \mathbb{R}$. Wie

[3] Für die Gamma-Funktion gilt etwa $\Gamma(k) = (k-1)!$ für $k = 1, 2, \ldots$ sowie $\Gamma(x+1) = x\Gamma(x)$ für $x \in \mathbb{R}$.

sieht dann die a-posteriori-Verteilung aus? Und ist diese um den wahren Wert θ konzentriert? In der Übung wird gezeigt werden, dass dies wieder eine Normalverteilung $\mathcal{N}(\alpha, \beta)$ ist mit

$$\alpha = \frac{1}{\sigma^2/(nb^2)+1}\overline{x} + \frac{\sigma^2/(nb^2)}{\sigma^2/(nb^2)+1}a,$$
$$\beta = \frac{1}{n/\sigma^2 + 1/b^2}.$$

Insbesondere ist für große n die a-posteriori-Verteilung um $\overline{x}$ konzentriert.

3 Entscheidungstheorie

Statistische Fragestellungen formuliert man oft als Entscheidungsprobleme. Diese zeichnen sich dadurch aus, dass aufgrund von Daten in einem (statistischen) Modell immer eine Entscheidung über die verwendeten Parameter getroffen werden muss. Bei einem statistischen Test ist diese Entscheidung etwa, ob der Parameter größer oder kleiner als ein vorgegebener Wert ist, bei einem Schätzproblem fällt eine Entscheidung über die (vermutete) Größe eines Parameters.

3.1 Einführung

Zentral sind bei Entscheidungen die Begriffe des Entscheidungsraumes und der Entscheidungsfunktion. Um über die Qualität der Entscheidung(sfunktion) zu urteilen, gibt es außerdem eine Verlustfunktion.

Definition 3.1 (Entscheidungsraum, -funktion, Verlustfunktion). *Sei* $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ *ein statistisches Modell.*

1. *Für einen* Entscheidungsraum $\aleph$ *heißt jede messbare Abbildung* $d : E \to \aleph$ nicht-randomisierte Entscheidungsfunktion. *Eine* (randomisierte) Entscheidungsfunktion *ist ein Markov-Kern* $\delta(.,.)$ *von* E *nach* $\mathcal{B}(\aleph)$. *Die Menge der nicht-randomisierten Entscheidungsfunktionen bezeichnen wir mit* $\mathcal{D}_{nr}$, *die Menge der randomisierten Entscheidungsfunktionen mit* $\mathcal{D}$. *Für* $d \in \mathcal{D}_{nr}$ *sei* $\delta_d(x, A) := 1_{d(x)\in A}$ *die zugehörige randomisierte Entscheidungsfunktion.*

2. *Eine* Verlustfunktion *ist eine messbare Abbildung* $\ell : \mathcal{P} \times \aleph \to \mathbb{R}_+$.

3. *Ein* statistisches Entscheidungsproblem *ist ein Tripel* $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ *aus einem statistischen Modell, einem Entscheidungsraum und einer Verlustfunktion.*

4. *Für ein statistisches Entscheidungsproblem* $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ *heißt die Abbildung*

$$R_\delta(\theta) := R(\theta, \delta) := \mathbb{E}_\theta\Big[\int \ell(\theta, y)\delta(X, dy)\Big] \tag{3.1}$$

Risikofunktion. *Für eine nicht-randomisierte Entscheidung ist dies gleich*

$$R_d(\theta) := R(\theta, d) := \mathbb{E}_\theta\big[\ell(\theta, d(X))\big].$$

Die Menge $\mathcal{R} := \{R_\delta : \delta \in \mathcal{D}\}$ *heißt* Risikomenge.

Bemerkung 3.2 (Interpretation). In einem statistischen Modell steht X für die (zufällig entstandenen) Daten. Bei Vorliegen des (unbekannten Parameters) $\theta \in \mathcal{P}$ sind diese nach $\mathbb{P}_\theta$ verteilt. Für eine Entscheidungsfunktion δ ist nun $\delta(x, A)$ die Wahrscheinlichkeit, sich für ein $a \in A$ zu entscheiden, wenn $X = x$ vorliegt. (Bei einer nicht-randomisierten Entscheidungsfunktion d ist $d(x) = a$ die Wahl der Entscheidung a im Entscheidungsraum $\aleph$.) Bei einer Entscheidung für a hat man, falls θ vorliegt, einen Verlust zu verzeichnen. Dieser wird mit $\ell(\theta, a)$ bezeichnet. Da die Daten als zufällig angesehen werden, kann man sich fragen, welchen Verlust man denn (bei Vorliegen von θ und für die Entscheidung δ) erwartet. Dies ist gerade die Risikofunktion $R_\delta(\theta)$.

Bemerkung 3.3 (Punkt-Schätzproblem). 1. Aus der Vorlesung *Stochastik* bekannt ist folgendes Problem:
Ein Versuch, der entweder einen Erfolg oder einen Misserfolg bringt, wird n-mal wiederholt, wobei $S = k$-mal ein Erfolg eintritt. Nun soll die Wahrscheinlichkeit für einen Erfolg (in einem der n Versuchen) geschätzt werden.
Dies erinnert an das statistische Modell aus Beispiel *Bern*. Wir setzen hier $\mathcal{P} = [0, 1]$, $\mathbb{P}_\theta$ wie in (*Bern* 0), $S = X_1 + \cdots + X_n$ und $\aleph = \mathcal{P}$. Eine offensichtlich Wahl für eine nicht-randomisierte Entscheidungsfunktion ist $d(x_1, ..., x_n) = (x_1 + \cdots + x_n)/n =: \overline{x}$. Als Verlustfunktion könnte etwa $\ell(\theta, a) = (\theta - a)^2$ dienen, also $\ell(\theta, d(x_1, ..., x_n)) = (\theta - \overline{x})^2$. Die Risikofunktion berechnet sich dann zu

$$R_d(\theta) = \mathbb{E}_\theta[(\theta - \overline{X})^2] = \mathbb{V}_\theta[\overline{X}] = \frac{\theta(1-\theta)}{n}.$$

2. Allgemeiner ist ein Punkt-Schätzproblem ein statistisches Entscheidungsproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ mit dem statistischen Raum $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$. Im Allgemeinen wollen wir nicht θ schätzen, sondern $g(\theta)$ für eine Funktion $g : \mathcal{P} \to \aleph$, wobei $\aleph$ auch der Entscheidungsraum ist. (Wie hierbei die Entscheidungsfunktion δ (oder d im nicht-randomisierten Fall) aussieht, hängt vom jeweiligen Beispiel ab.) Verlustfunktionen, die der Bemessung einer falschen Entscheidung dienen sollen, sind (im Falle eines normierten Raumes $\aleph$) etwa der

$$\begin{aligned} \text{Laplace-Verlust } \ell(\theta, a) &:= |a - g(\theta)|, \\ \text{Gauß-Verlust } \ell(\theta, a) &:= |a - g(\theta)|^2, \\ \text{0-1-Verlust } \ell(\theta, a) &:= 1_{|a-g(\theta)|>\varepsilon}. \end{aligned}$$

Die Risikofunktion $R_\delta(\theta)$ ist dann der (unter $\mathbb{P}_\theta$) erwartete Verlust unter der (etwa nicht-randomisierten) Entscheidungsfunktion d, also beim Laplace-Verlust etwa

$$R_\delta(\theta) = \mathbb{E}_\theta[|d(X) - g(\theta)|].$$

Man nennt Entscheidungsfunktionen bei solchen Schätzproblemen *(Punkt-)Schätzer* und $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem.

Bemerkung 3.4 (Bereichs-Schätzproblem). 1. Aus der Vorlesung *Stochastik* bekannt ist folgendes Problem:
Für eine Stichprobe $x_1, ..., x_n$ normalverteilter Daten (d.h. $X_1, ..., X_n$ sind unabhängig und identisch nach $\mathcal{N}(\theta, \sigma^2)$ verteilt) bei bekannter Varianz σ^2. Es wird ein von den Daten abhängiger Bereich gesucht, so dass θ mit Wahrscheinlichkeit $1 - \alpha$ (etwa für

$\alpha = 5\%$) in diesem liegt. Es ist naheliegend, dass dieses Intervall symmetrisch um $\overline{x} := (x_1 + \cdots + x_n)/n$ liegt.
Dies erinnert an das statistische Modell *Norm* mit bekannter Varianz σ^2, d.h. an das statistische Modell $(X, \{\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\})$, so dass (*Norm* 0) gilt. Um das Intervall aufzustellen, verwenden wir, dass $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, also $(\overline{X} - \mu)/\sqrt{\sigma^2/n} \sim \mathcal{N}(0,1)$. Sei q_α das α-Quantil der Standardnormalverteilung[4]. Dann gilt

$$\begin{aligned} 1 - \alpha &= \mathbb{P}_\theta(q_{\alpha/2} \leq (\overline{X} - \theta)/\sqrt{\sigma^2/n} \leq q_{1-\alpha/2}) \\ &= \mathbb{P}_\theta(|\overline{X} - \theta| \leq q_{1-\alpha/2}\sqrt{\sigma^2/n}) \\ &= \mathbb{P}_\theta(\overline{X} - q_{1-\alpha/2}\sqrt{\sigma^2/n} \leq \theta \leq \overline{X} + q_{1-\alpha/2}\sqrt{\sigma^2/n}) \end{aligned}$$

Der gesuchte Datenbereich ist also $\{y : \overline{x} - q_{1-\alpha/2}\sqrt{\sigma^2/n} \leq \overline{y} \leq \overline{x} + q_{1-\alpha/2}\sqrt{\sigma^2/n}\}$.

2. Allgemeiner ist ein Bereichs-Schätzproblem ein statistisches Entscheidungsproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ mit dem statistischen Raum $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$. Wieder wollen wir nicht θ schätzen, sondern $g(\theta)$ für eine Funktion $g : \mathcal{P} \to \Upsilon$, wobei hier nun $\aleph = \mathcal{B}(\Upsilon)$ ist. (Übrigens ist damit $\aleph$ kein metrischer Raum, und man muss die Messbarkeit der folgenden Abbildungen überprüfen.) Eine nicht-randomisierte Entscheidungsfunktion ist hier wieder eine Funktion $d : E \to \aleph$. (Nun muss etwa für alle $g \in \Upsilon$ die Bedingung $\{x : g \in d(x)\} \in \mathcal{B}(E)$ gelten.) Als Verlustfunktion bietet sich die Wahl

$$\ell(\theta, B) := \begin{cases} 1, & g(\theta) \notin B, \\ 0, & g(\theta) \in B \end{cases}$$

an. Das Risiko ist somit

$$R_d(\theta) = \mathbb{E}_\theta[\ell(\theta, d(X))] = \mathbb{P}_\theta[g(\theta) \notin d(X)].$$

Entscheidungsfunktionen in einer solchen Situation heißen auch *Bereichs-Schätzer*.

Bemerkung 3.5 (Test-Problem). 1. Aus der Vorlesung *Stochastik* bekannt ist folgendes Problem (siehe *einfacher t-Test*):
Für eine Stichprobe $x_1, ..., x_n$ normalverteilter Daten (d.h. $X_1, ..., X_n$ sind unabhängig und identisch nach $\mathcal{N}(\theta = (\mu, \sigma^2))$ verteilt) bei unbekannter Varianz σ^2. Es soll getestet werden, ob die (Null-)Hypothese $\mu = \mu_0$ aufgrund der Daten verworfen werden kann oder nicht. Dabei soll die Wahrscheinlichkeit, die Nullhypothese irrtümlicherweise zu verwerfen, höchstens ein vorgegebenes α sein.
Dies erinnert an das statistische Modell *Norm*, d.h. an das statistische Modell $(X, \{\mathbb{P}_\theta = \mathcal{N}(\theta) : \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\})$, so dass (*Norm* 0) gilt. Um die Hypothese $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$ zu testen, verwendet man, dass für $s^2(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ die Statistik

$$T := t(X) = \frac{\overline{X} - \mu}{\sqrt{s^2(X)/n}}$$

unter $\mathbb{P}_{(\mu,\sigma^2)}$ (für jedes $\sigma^2 \in \mathbb{R}_+$) nach t_{n-1}-verteilt ist.[5] Man stellt hier den kritischen (oder Ablehnungs-)Bereich $C \subseteq \mathbb{R}$ (der nur von α abhängt) so auf, dass H_0 gerade dann

[4] Es ist also für $Z \sim \mathcal{N}(0,1)$ gerade $\mathbb{P}(Z \leq q_\alpha) = \alpha$. Insbesondere ist auch $q_\alpha = -q_{1-\alpha}$.
[5] Dies ist eine studentische t-Verteilung mit $n-1$ Freiheitsgraden. Ihr α-Quantil bezeichnen wir mit $t_{n-1,\alpha}$.

abgelehnt wird, wenn $t(X) \in C$ ist. Es soll außerdem $\mathbb{P}_{(\mu_0,\sigma^2)}(t(X) \in C) \leq \alpha$ gelten. In diesem Beispiel ergibt sich mit $C = (-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty)$ gerade, dass

$$\mathbb{P}_{(\mu_0,\sigma^2)}(t(X) \in C) = \mathbb{P}_{(\mu_0,\sigma^2)}(t(X) \leq t_{n-1,\alpha/2}) + \mathbb{P}_{(\mu_0,\sigma^2)}(t(X) \geq t_{n-1,1-\alpha/2}) = \alpha,$$

was ja gefordert war.

2. Allgemeiner ist ein Test-Problem ein statistisches Entscheidungsproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ mit dem statistischen Raum $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ mit $\aleph = \{H_0, H_1\}$. Die (randomisierte) Entscheidungsfunktion δ ist eindeutig durch eine Funktion $\varphi : E \to [0,1]$ mittels $\delta(x, \{H_1\}) := \varphi(x)$ und $\delta(x, \{H_0\}) := 1 - \varphi(x)$ bestimmt. (Hier ist $\varphi(x)$ die Wahrscheinlichkeit, sich bei Daten x für die Hypothese H_1 zu entscheiden.) Die Entscheidungsfunktion δ ist genau dann nicht-randomisiert, falls $\varphi = 1_C$ für ein geeignetes $C \in \mathcal{B}(E)$. Als Verlustfunktion bietet sich die Neyman-Pearson-Verlustfunktion an. Hierbei ist für $\mathcal{P}_0, \mathcal{P}_1$ mit $\mathcal{P}_0 \uplus \mathcal{P}_1 = \mathcal{P}$ gerade

$$\ell(\theta, H_1) = \begin{cases} 0, & \theta \in \mathcal{P}_1 \quad \text{richtige Entscheidung,} \\ 1, & \theta \in \mathcal{P}_0 \quad \text{Fehler 1. Art: Entscheidung für } H_1, \text{ aber } H_0 \text{ trifft zu.} \end{cases}$$

$$\ell(\theta, H_0) = \begin{cases} 0, & \theta \in \mathcal{P}_0 \quad \text{richtige Entscheidung,} \\ 1, & \theta \in \mathcal{P}_1 \quad \text{Fehler 2. Art: Entscheidung für } H_0, \text{ aber } H_1 \text{ trifft zu.} \end{cases}$$

Die Hypothese H_i besteht dabei gerade daraus, dass eine der Verteilungen aus $\mathcal{P}_i$ (auf die Daten) zutrifft, $i = 0, 1$. Die Risikofunktion ist damit

$$\begin{aligned} R_\delta(\theta) &= \mathbb{E}_\theta[\ell(\theta, H_0)\delta(X, \{H_0\}) + \ell(\theta, H_1)\delta(X, \{H_1\})] \\ &= 1_{\theta \in \mathcal{P}_1}\mathbb{E}_\theta[1 - \varphi(X)] + 1_{\theta \in \mathcal{P}_0}\mathbb{E}_\theta[\varphi(X)] = \begin{cases} 1 - \mathbb{E}_\theta[\varphi(X)], & \theta \in \mathcal{P}_1, \\ \mathbb{E}_\theta[\varphi(X)], & \theta \in \mathcal{P}_0. \end{cases} \end{aligned}$$

Wie wir sehen, wird das Risiko eindeutig durch die *Gütefunktion*

$$\beta_\varphi : \theta \mapsto \mathbb{E}_\theta[\varphi(X)]$$

bestimmt. Im Fall einer nicht-randomisierten Entscheidungsfunktion $\varphi = 1_C$ ist also

$$R_\delta(\theta) = \begin{cases} \mathbb{P}_\theta(X \notin C), & \theta \in \mathcal{P}_1, \\ \mathbb{P}_\theta(X \in C), & \theta \in \mathcal{P}_0. \end{cases}$$

Bei diesem allgemeinen Vorgehen fällt auf, dass nirgends das Signifikanzniveau α aus obigem Beispiel eingeht. Das liegt daran, dass wir zunächst alle Entscheidungsfunktionen zugelassen haben. Sinnvoller ist es jedoch, sich auf eine Klasse von Entscheidungsfunktionen einzuschränken, etwa auf solche δ, für die $R_\delta(\theta) \leq \alpha$ für $\theta \in \mathcal{P}_0$ gilt.
Man nennt Entscheidungsfunktionen δ (oder φ) bei solchen Testproblemen auch *Tests*.

Bemerkung 3.6 (Randomisierter Test). Wir geben noch das Beispiel eines randomisierten Tests an. Sei hierzu $(X, \{\mathbb{P}_\theta : \theta \in [0,1]\})$ das statistische Modell aus Beispiel 1.6 sowie $H_0 = \{\theta \leq 0.5\}$ und $H_1 = \{\theta > 0.5\}$. Für ein $\alpha \in (0,1)$ wollen wir einen Test φ mit $\mathbb{E}_\theta[\varphi(X)] \leq \alpha$ für $\theta \in H_0$ angeben, d.h. die Wahrscheinlichkeit, sich für die Alternative zu entscheiden, soll bei Vorliegen von $\theta \in H_0$ höchstens α betragen. Man sagt auch, der Fehler 1.

Art soll höchstens α sein. Sei hierzu $q_{\theta,1-\alpha}$ ein $1-\alpha$-Quantil einer $B(n,\theta)$-Verteilung[6]. Dann könnten wir etwa

$$\varphi(x) = \begin{cases} 0, & \sum_{i=1}^n x_i \le q_{0.5,1-\alpha}, \\ 1, & \sum_{i=1}^n x_i > q_{0.5,1-\alpha} \end{cases}$$

setzen. Für $\theta \in H_0$ gilt dann

$$\mathbb{E}_\theta[\varphi(X)] = \mathbb{P}_\theta\Big[\sum_{i=1}^n X_i > q_{0.5,1-\alpha}\Big] \le \mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i > q_{0.5,1-\alpha}\Big] \le \alpha.$$

Schön wäre es, wenn wir sogar ein $\psi \ge \varphi$ angeben können, für das ebenfalls $\mathbb{E}_\theta[\psi(X)] \le \alpha$ für $\theta \in H_0$ gilt. Dies können wir erreichen, indem wir

$$\psi(x) = \begin{cases} 0, & \sum_{i=1}^n x_i < q_{0.5,1-\alpha}, \\ p := \frac{\mathbb{P}_{0.5}\left[\sum_{i=1}^n x_i \le q_{0.5,1-\alpha}\right] - (1-\alpha)}{\mathbb{P}_{0.5}\left[\sum_{i=1}^n x_i = q_{0.5,1-\alpha}\right]}, & \sum_{i=1}^n x_i = q_{0.5,1-\alpha}, \\ 1, & \sum_{i=1}^n x_i > q_{0.5,1-\alpha}. \end{cases}$$

Das bedeutet: falls $\sum_{i=1}^n x_i = q_{0.5,1-\alpha}$, so entscheiden wir uns mit Wahrscheinlichkeit p für die Alternative, und mit $1-p$ für die Nullhypothese. Die Entscheidung ist also zufällig oder randomisiert. Dann ist für $\theta \in H_0$

$$\begin{aligned} \mathbb{E}_\theta[\psi(X)] \le \mathbb{E}_{0.5}[\psi(X)] &= 1 - \mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i \le q_{0.5,1-\alpha}\Big] + p\mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i = q_{0.5,1-\alpha}\Big] \\ &= 1 - \mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i \le q_{0.5,1-\alpha}\Big] + \mathbb{P}_{0.5}\Big[\sum_{i=1}^n X_i \le q_{0.5,1-\alpha}\Big] - (1-\alpha) = \alpha, \end{aligned}$$

also ist der Fehler 1. Art ebenfalls höchstens α.

Es sollte klar sein, dass *gute* Entscheidungskriterien gerade solche sind, die eine kleine Risikofunktion aufweisen. Allerdings gibt es verschiedene Möglichkeiten dafür, dass eine Entscheidungsfunktion *klein* ist. Mit diesen werden wir uns in Abschnitt 3.3 beschäftigen. Wir geben nur noch ein Beispiel, in dem klar wird, dass naiv angegebene Entscheidungsfunktionen auch *schlecht* sein können.

Beispiel 3.7 (Schätzung des Lageparameters einer Uniformverteilung).
Sei $U(\theta-1/2,\theta+1/2)$ die Uniformverteilung auf $(\theta-1/2,\theta+1/2)$ und $\mathbb{P}_\theta = U(\theta-1/2,\theta+1/2)^n$ das n-fache Produkt, $\theta \in \mathbb{R}$. Wir wollen für das statistische Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathbb{R}\})$ den Lageparameter θ schätzen. Hierzu geben wir zwei verschiedene, nicht-randomisierte Entscheidungsfunktionen an. Wir wählen (mit $g(\theta) := \theta$)[7]

$$d_1(X) = \overline{X}, \qquad d_2(X) = \tfrac{1}{2}(X_{(1)} + X_{(n)})$$

[6] Wir erinnern daran, dass für eine Verteilung $\mathbb{P}$ mit Verteilungsfunktion F das α-Quantil durch jede Zahl x mit $F(x-) \le \alpha \le F(x)$ gegeben ist.

[7] Für einen Vektor $x = (x_1,...,x_n) \in \mathbb{R}^n$ ist $x_{(i)}$ das it-kleinste Element. Insbesondere ist also $x_{(1)}$ das kleinste und $x_{(n)}$ das größte Element. Die Zahlen $x_{(1)},...,x_{(n)}$ heißen auch *Ordnungsstatistiken*.

und verwenden den Gauß-Verlust. Das Risiko berechnet sich für d_1 zu

$$R_{d_1}(\theta) = \mathbb{E}_\theta[(\overline{X} - \theta)^2] = \mathbb{V}_\theta[\overline{X}] = \frac{1}{n}\mathbb{V}_\theta[X_1] = \frac{1}{12n}.$$

Für d_2 benötigen wir ein paar Vorüberlegungen. Um die gemeinsame Verteilung von $X_{(1)}$ und $X_{(n)}$ zu berechnen, schreiben wir (für $\theta = 1/2$)

$$\mathbb{P}_{1/2}(X_{(1)} > x, X_{(n)} < y) = (y - x)^n.$$

Also hat $(X_{(1)}, X_{(n)})$ die gemeinsame Dichte $(x, y) \mapsto n(n-1)(y-x)^{n-2}$. Daraus berechnen wir

$$\begin{aligned}
\mathbb{E}_{1/2}[X_{(1)}] &= 1 - \mathbb{E}_{1/2}[1 - X_{(1)}] = 1 - \int_0^1 n(1-x)^n dx = \frac{1}{n+1},\\
\mathbb{E}_{1/2}[X_{(n)}] &= \int_0^1 nyy^{n-1} dy = \frac{n}{n+1},\\
\mathbb{V}_{1/2}[X_{(1)}] &= \mathbb{V}_{1/2}[1 - X_{(1)}] = \int_0^1 n(1-x)^{n+1} dx - \frac{n^2}{(n+1)^2} = \frac{n}{n+2} - \frac{n^2}{(n+1)^2}\\
&= \frac{n}{(n+2)(n+1)^2},\\
\mathbb{V}_{1/2}[X_{(n)}] &= \frac{n}{(n+2)(n+1)^2},\\
\mathbb{COV}_{1/2}[X_{(1)}, X_{(n)}] &= \int_0^1 \int_0^y n(n-1)xy(y-x)^{n-2} dxdy - \frac{n}{(n+1)^2}\\
&= \int_0^1 \int_0^y ny(y-x)^{n-1} dxdy - \frac{n}{(n+1)^2}\\
&= \int_0^1 y^{n+1} dy - \frac{n}{(n+1)^2} = \frac{1}{n+2} - \frac{1}{(n+1)^2} = \frac{1}{(n+1)^2(n+2)},
\end{aligned}$$

wobei die Varianzen und Covarianz unabhängig von θ sind. Wir erhalten

$$\begin{aligned}
R_{d_1}(\theta) &= \mathbb{E}_\theta[(\tfrac{1}{2}(X_{(1)} + X_{(n)}) - \theta)^2] = \mathbb{V}_\theta[\tfrac{1}{2}(X_{(1)} + X_{(n)})]\\
&= \tfrac{1}{4}(\mathbb{V}[X_{(1)}] + \mathbb{V}[X_{(n)}] + 2\mathbb{COV}[X_{(1)}, X_{(n)}])\\
&= \frac{1}{2}\Big(\frac{n}{(n+2)(n+1)^2} + \frac{1}{(n+1)^2(n+2)}\Big) = \frac{1}{2(n+1)(n+2)}.
\end{aligned}$$

Wir sehen also, dass das Risiko von d_2 gerade für große n deutlich kleiner ist als das von d_1.

3.2 Die Rolle suffizienter Statistiken

Suffiziente Statistiken enthalten alle wichtigen Informationen über die Daten. Deswegen ist es sinnvoll, dass Entscheidungsfunktionen nur von solchen Statistiken abhängen.

Proposition 3.8. *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, δ eine (randomisierte) Entscheidungsfunktion (also ein Markov-Kern von E nach $\mathcal{B}(\aleph)$) und R_δ ihre Risikofunktion. Ist $T = t(X)$ (für $t : E \to E'$) eine suffiziente Statistik, so gibt es einen Markov-Kern ε von E' nach $\mathcal{B}(\aleph)$, so dass $\varepsilon \circ t$ (mit $\varepsilon \circ t(x, A) := \varepsilon(t(x), A)$) eine Entscheidungsfunktion ist mit $R_{\varepsilon \circ t} = R_\delta$.*

Beweis. Für $A \in \mathcal{B}(\aleph)$ setzen wir

$$\varepsilon(t, A) := \mathbb{E}_\theta[\delta(X, A)|T = t].$$

(Da T suffizient ist, hängt die rechte Seite nicht von θ ab. Wir unterdrücken den Subskript θ im Folgenden.) Deshalb gilt für integrierbare Funktionen $h : \aleph \to \mathbb{R}$, dass

$$\mathbb{E}\Big[\int h(a)\delta(X, da)\Big|T = t\Big] = \int h(a)\varepsilon(t, da).$$

(Zunächst überprüft man die Aussage mit Indikatorfunktionen, dann mit einfachen Funktionen und anschließend mit allgemeinen messbaren Funktionen h.) Nach Definition gilt dann

$$\begin{aligned} R_{\varepsilon\circ t}(\theta) &= \mathbb{E}_\theta\Big[\int \ell(\theta, a)\varepsilon(t(X), da)\Big] = \mathbb{E}_\theta\Big[\mathbb{E}\Big[\int \ell(\theta, a)\delta(X, da)\Big|T\Big]\Big] \\ &= \mathbb{E}_\theta\Big[\int \ell(\theta, a)\delta(X, da)\Big] = R_\delta(\theta). \end{aligned}$$

□

Man beachte, dass auch bei nicht-randomisierten δ der Markov-Kern ε (und damit $\varepsilon \circ t$) randomisiert sein kann.

Proposition 3.9. *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem. Sei $\aleph \subseteq \mathbb{R}^m$ konvex und $\ell(\theta, .)$ für alle $\theta \in \mathcal{P}$ eine konvexe Verlustfunktion. Für eine (randomisierte) Entscheidungsfunktion δ definieren wir die nicht-randomisierte Entscheidungsfunktion*

$$d(x) := \int a\delta(x, da),$$

zumindest für $x \in E$, für die dieses Integral existiert. Für solche $x \in E$ und alle $\theta \in \mathcal{P}$ ist dann

$$\ell(\theta, d(x)) \leq \int \ell(\theta, a)\delta(x, da).$$

Beweis. Zunächst ist $d(x) \in \aleph$, da $\aleph$ konvex ist. Weiter gilt mit der Jensen'schen Ungleichung

$$\ell(\theta, d(x)) = \ell\Big(\theta, \int a\delta(x, da)\Big) \leq \int \ell(\theta, a)\delta(x, da).$$

□

Man kann mit Hilfe suffizienter Statistiken Entscheidungsfunktionen besser machen. Vor allem für Schätzprobleme wird folgendes Resultat schöne Anwendungen haben.

Theorem 3.10 (Rao-Blackwell). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem. Sei $\aleph \subseteq \mathbb{R}^m$ konvex und $\ell(\theta, .)$ für alle $\theta \in \mathcal{P}$ eine konvexe Verlustfunktion, $T = t(X)$ eine suffiziente Statistik und d eine nicht-randomisierte Entscheidungsfunktion mit $\mathbb{E}_\theta[|d(X)|] < \infty$ für alle $\theta \in \mathcal{P}$ sowie R_d ihre Risikofunktion. Für*

$$e(t) := \mathbb{E}_\theta[d(X)|T = t]$$

(wobei die rechte Seite unabhängig von θ ist) gilt dann für alle $\theta \in \mathcal{P}$

$$R_{e\circ t}(\theta) \leq R_d(\theta).$$

Beweis. Wir setzen

$$\varepsilon(t, A) := \mathbb{P}(d(X) \in A | T = t),$$

also

$$\int h(a)\varepsilon(t, da) = \mathbb{E}[h(d(X))|T = t],$$

falls das Integral existiert. (Wieder zeigt man das, indem man zuerst einfache Funktionen einsetzt und danach durch ein Approximationsargument messbare Funktionen.) Daraus folgt insbesondere

$$\int a\varepsilon(t, da) = \mathbb{E}[d(X)|T = t] = e(t).$$

Dann gilt wegen Proposition 3.9

$$\ell(\theta, e \circ t(x)) \leq \int \ell(\theta, a)\varepsilon(t(x), da),$$

also

$$R_{eot}(\theta) \leq R_\varepsilon(\theta) = \mathbb{E}_\theta\Big[\int \ell(\theta, a)\varepsilon(T, da)\Big] = \mathbb{E}_\theta[\mathbb{E}_\theta[\ell(\theta, d(X))|T]] = R_d(\theta).$$

□

Beispiel 3.11 (Beispiel *Unif*). Sei $(X, \{\mathcal{P}_\theta : \theta \in (0, 1)\})$ wie in Beispiel 1.8, $\aleph = (0, 1)$, $g(\theta) = \theta$ und ℓ der Gauß-Verlust $\ell(\theta, a) = (\theta - a)^2$. Wir betrachten den Schätzer $d(x) = 2\overline{x}$. Für diesen ist immerhin $\mathbb{E}_\theta[d(X)] = \theta$. Wir wissen aus Beispiel 2.8, dass $T = t(X) = \max_{i=1,\ldots,n} X_i$ suffizient ist. Wir verwenden Theorem 3.10 und schreiben aus Symmetriegründen

$$\mathbb{E}_\theta[2\overline{X}|t(X)] = 2\mathbb{E}_\theta[X_1|t(X)] = 2\Big(\frac{1}{n}t(X) + \frac{n-1}{n}\frac{t(X)}{2}\Big) = \frac{n+1}{n}t(X).$$

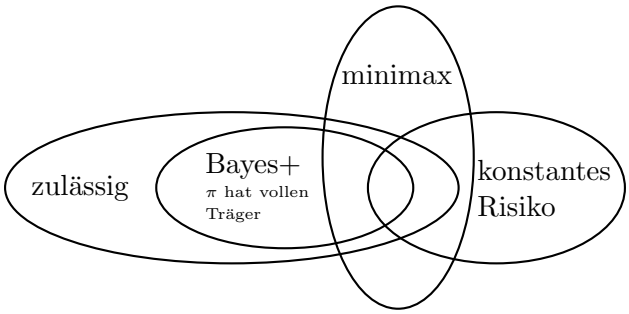
Nun lesen wir ab, dass der Schätzer

$$d'(x) := \frac{n+1}{n}\max_{i=1,\ldots,n} x_i$$

eine kleinere Risikofunktion hat.

3.3 Zulässige, Bayes, Minimax-Entscheidungsfunktionen

In diesem Abschnitt geht es um den Vergleich von Entscheidungsfunktionen mittels deren Risikofunktionen, sowie um bestimmte Optimalitätskriterien. Wie werden hier speziell zulässige, minimax und Bayes-Entscheidungsfunktionen betrachten. Oftmals macht es dabei keinen Sinn, optimale Entscheidungskriterien unter allen möglichen Entscheidungsfunktionen zu suchen, sondern nur aus einer Teilmenge. (Wir erinnern daran, dass wir die Menge aller Entscheidungsfunktionen mit $\mathcal{D}$ bezeichnet hatten.) Wir illustrieren einige Ergebnisse dieses Abschnitts in einer Grafik.



In Lemma 3.14 zeigen wir etwa, dass zulässige Entscheidungsfunktionen mit konstantem Risiko minimax sind. Lemma 3.19 besagt, dass Bayes-Entscheidungsfunktionen (unter einer Voraussetzung an die a-priori-Verteilung) zulässig sind. Die minimax-Eigenschaft von (allgemeinen) Bayes-Entscheidungsfunktionen wird dann mit Theorem 3.20 geklärt.

Beispiel 3.12. 1. Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}), \aleph, \ell)$ mit $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$, $\aleph = \mathbb{R}$ und ℓ der Gauß-Verlust. Wir wollen also den Mittelwert einer Normalverteilung (bei bekannter Varianz) schätzen, wenn wir nur einmal aus ihr ziehen. Die offensichtliche Wahl $\delta_1(x, a) = x$ (d.h. wir schätzen den Parameter θ durch die Beobachtung x) führt zur Risikofunktion

$$\theta \mapsto R_{\delta_1}(\theta) = \mathbb{E}_\theta[(X - \theta)^2] = 1.$$

Hingegen führt die Wahl $\delta_2(x, a) = b$ für ein festes $b \in \mathbb{R}$ (d.h. wir schätzen den Parameter θ immer durch dieselbe Zahl b, unabhängig von unserer Beobachtung) zur Risikofunktion

$$\theta \mapsto R_{\delta_2}(\theta) = \mathbb{E}_\theta[(b - \theta)^2] = (b - \theta)^2.$$

Insbesondere sehen wir, dass weder $R_{\delta_1} \leq R_{\delta_2}$ noch $R_{\delta_2} \leq R_{\delta_1}$ gilt. Man kann also nicht sagen, dass δ_1 besser wäre als δ_2. Ein Ausweg hier ist es, sich nur auf erwartungstreue Schätzer einzuschränken, d.h. auf Entscheidungsfunktionen δ mit $\mathbb{E}_\theta[\int \delta(X, da)] = \theta$, und unter diesen die kleinste Risikofunktion zu suchen.

2. Ganz ähnlich verhält es sich bei einem Test. Verwenden wir hierzu die Situation aus Bemerkung 3.5.2. Es ist verlockend, einfach $C = E$ zu wählen (d.h. die Hypothese wird immer verworfen), weil dadurch zumindest für $\theta \in \mathcal{P}_0$ die Risikofunktion verschwindet. Allerdings wird durch diese Wahl die Risikofunktion für $\theta \in \mathcal{P}_1$ maximal. Wählt man andererseits $C = \emptyset$, so ist die Situation genau andersherum. Deshalb fragt man hier eher nach dem Minimum aller Risikofunktionen, die $R_\delta(\theta) \leq \alpha$ für alle $\theta \in \mathcal{P}_0$ erfüllen. Das bedeutet, dass man das Signifikanzniveau α festlegt und damit den maximalen Fehler erster Art.

Definition 3.13 (Zulässige, Minimax-Entscheidungsfunktionen). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, $\delta, \delta' \in \mathcal{D}$ Entscheidungsfunktionen und $R_\delta, R_{\delta'}$ ihre Risikofunktionen. Weiter sei $\mathcal{D}_0 \subseteq \mathcal{D}$.*

1. *Gilt für δ, δ', dass $R_\delta \leq R_{\delta'}$ und $R_\delta(\theta) < R_{\delta'}(\theta)$ für mindestens ein $\theta \in \mathcal{P}$, so sagen wir, δ* dominiert *δ'.*

2. *Wir sagen, die Entscheidungsfunktion δ hat* konstantes Risiko*, falls $\theta \mapsto R_\delta(\theta)$ konstant ist.*

3. *Gibt es für eine Entscheidungsfunktion $\delta \in \mathcal{D}_0$ ein $\delta' \in \mathcal{D}_0$, das δ dominiert, so heißt δ auch* unzulässig für $\mathcal{D}_0$, *andersfalls* zulässig für $\mathcal{D}_0$.

4. *Die Entscheidungsfunktion $\delta \in \mathcal{D}_0$ heißt* minimax für $\mathcal{D}_0$, *falls*

$$\sup_{\theta\in\mathcal{P}} R_\delta(\theta) = \inf_{\delta'\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta).$$

Für $\mathcal{D}$ zulässige Entscheidungsfunktionen heißen auch zulässig, für $\mathcal{D}$ unzulässige auch unzulässig, und für $\mathcal{D}$ minimax-Entscheidungsfunktionen heißen auch minimax-Entscheidungsfunktionen.

Lemma 3.14 (Zusammenhänge zulässige, minimax-Entscheidungsfunktionen). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, $\delta \in \mathcal{D}_0 \subseteq \mathcal{D}$ eine Entscheidungsfunktion und R_δ ihre Risikofunktion.*

1. *Ist δ die einzige minimax-Entscheidungsfunktion für $\mathcal{D}_0$, so ist δ zulässig für $\mathcal{D}_0$.*

2. *Ist δ zulässig für $\mathcal{D}_0$ mit konstantem Risiko, so ist δ minimax für $\mathcal{D}_0$.*

Beweis. 1. Angenommen, δ ist nicht zulässig für $\mathcal{D}_0$. Dann gibt es ein $\delta' \in \mathcal{D}_0$ mit ($\delta' \neq \delta$ und) $R_{\delta'} \leq R_\delta$ und ein $\theta \in \mathcal{P}$ mit $R_{\delta'}(\theta) < R_\delta(\theta)$. Deshalb ist

$$\inf_{\varepsilon\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_\varepsilon(\theta) = \sup_{\theta\in\mathcal{P}} R_\delta(\theta) \geq \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta) \geq \inf_{\varepsilon\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_\varepsilon(\theta).$$

Damit ist δ' ebenfalls minimax für $\mathcal{D}_0$ im Widerspruch zur Voraussetzung.
2. Angenommen, es ist $R_\delta(\theta) =: c$ unabhängig von θ und δ ist nicht minimax für $\mathcal{D}_0$. Dann gibt es $\delta' \in \mathcal{D}_0$ mit

$$\sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta) < \sup_{\theta\in\mathcal{P}} R_\delta(\theta) = c.$$

Daraus folgt $R_{\delta'} < R_\delta$ im Widerspruch zur Zulässigkeit für $\mathcal{D}_0$ von δ. □

Definition 3.15 (Bayes'sches Risiko). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, δ eine Entscheidungsfunktion und R_δ ihre Risikofunktion. Sei weiter π ein Wahrscheinlichkeitsmaß auf $\mathcal{P}$ und $\Theta \sim \pi$. Dann ist das* Bayes-Risiko *(bezüglich π) gegeben als*

$$r_\delta(\pi) := \mathbb{E}_\pi[R_\delta(\Theta)] := \int R_\delta(\theta)\pi(d\theta).$$

Für $\mathcal{D}_0 \subseteq \mathcal{D}$ heißt $\delta \in \mathcal{D}_0$ eine $\mathcal{D}_0$-Bayes-Entscheidungsfunktion *bezüglich π, falls*

$$r_\delta(\pi) \leq r_{\delta'}(\pi)$$

für alle $\delta' \in \mathcal{D}_0$. Eine $\mathcal{D}$-Bayes-Entscheidungsfunktion *bezüglich π heißt auch* Bayes-Entscheidungsfunktion *bezüglich π.*

Oftmals sind Bayes-Entscheidungsfunktionen gar nicht so schwer zu finden. Hierbei besonders hilfreich ist das nächste Resultat.

Theorem 3.16 (Konstruktion von Bayes-Entscheidungsfunktionen). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem. Sei π eine a-priori-Verteilung, π_x die zugehörige a-posteriori-Verteilung und $\Theta_x \sim \pi_x$.*

1. *Sei δ eine Entscheidungsfunktion, so dass für alle x*

$$\mathbb{E}\Big[\int \ell(\Theta_x, a)\delta(x, da)\Big] = \inf_{m\in\mathcal{M}_1(\aleph)} \mathbb{E}\Big[\int \ell(\Theta, a)m(da)\Big].$$

Dann ist δ eine Bayes-Entscheidungsfunktion bezüglich π.

2. *Sei $d \in \mathcal{D}_{nr}$ eine nicht-randomisierte Entscheidungsfunktion, so dass für alle x*

$$\mathbb{E}[\ell(\Theta_x, d(x)] = \inf_{a\in\aleph} \mathbb{E}[\ell(\Theta, a)].$$

Dann ist δ eine Bayes-Entscheidungsfunktion bezüglich π für $\mathcal{D}_{nr}$.

Beweis. Wir zeigen nur 1. da der Beweis für 2. analog funktioniert. Wir schreiben das Bayes-Risiko für $\Theta \sim \pi$ als

$$r_\delta(\pi) = \mathbb{E}[R_\delta(\Theta)] = \mathbb{E}\Big[\mathbb{E}\Big[\int \ell(\Theta, a)\delta(X, da)\Big|X\Big]\Big].$$

Da die Verteilung von Θ gegeben X gerade π_X ist, die Erwartung sicher dann minimiert, wenn

$$\mathbb{E}\Big[\int \ell(\Theta, a)\delta(X, da)\Big|X = x\Big] = \mathbb{E}\Big[\int \ell(\Theta_x, a)\delta(x, da)\Big]$$

minimal ist. Daraus folgt die Behauptung. □

Korollar 3.17 (Bayes-Schätzer). *Wir betrachten das Schätzproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph = \mathbb{R}, g, \ell)$ für den Gauß-Verlust ℓ, ein reguläres statistisches Modell $\{\mathbb{P}_\theta : \theta \in \mathcal{P}\}$ mit $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ und $\mathcal{P} \subseteq \mathbb{R}^k$. Weiter sei $\pi = p \cdot \lambda^k \in \mathcal{M}_1(\mathcal{P})$ die a-priori-Verteilung und π_x die a-posteriori-Verteilung. (Nach Bemerkung 2.37 hat diese die Dichte*

$$p_x(\theta) = \frac{p_\theta(x)p(\theta)}{\int p_\eta(x)p(\eta)d\eta}.\Big)$$

Dann ist der nicht-randomisierte Schätzer (falls das Integral existiert)

$$d(x) := \int g(\theta)\pi_x(d\theta).$$

ein Bayes-Schätzer.

Beweis. Nach Proposition 3.9 müssen wir nur zeigen, dass d ein Bayes-Schätzer bezüglich π für $\mathcal{D}_{nr}$ ist. Nach Theorem 3.16 ist ein Bayes-Schätzer gegeben, wenn

$$\mathbb{E}[\ell(\Theta_x, d(x))] = \min_a \mathbb{E}[\ell(\Theta_x, a)].$$

Nun ist

$$\mathbb{E}[\ell(\Theta_x, a)] = \mathbb{E}[(g(\Theta_x) - a)^2] \geq \mathbb{E}[(g(\Theta_x) - \mathbb{E}[g(\Theta_x)])^2]$$

mit '=' genau dann, wenn

$$a = \mathbb{E}[g(\Theta_x)] = \int g(\theta)\pi_x(d\theta) = d(x).$$

Daraus folgt die Behauptung. □

Beispiel 3.18 (Beispiel *Norm* mit σ^2 bekannt). Wir betrachten für (bekanntes) $\sigma^2 > 0$ wie in (Norm 1*b*) das Schätzproblem $((X, \{\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^n : \theta \in \mathbb{R}\}), \aleph = \mathbb{R}, g = \mathrm{id}, \ell)$ für den Gauß-Verlust ℓ. Wir zeigen, dass für die a-priori-Verteilung $\pi_b = \mathcal{N}(0, b^2)$ der Bayes-Schätzer durch

$$d_b(x) := \frac{nb^2}{nb^2 + \sigma^2}\overline{x}$$

gegeben ist. (Diesen liest man am besten als Konvexkombintation aus dem a-priori-Schätzer 0 – gegeben durch die a-priori-Verteilung, die Erwartungswert 0 hat – und dem Mittelwert.) Aus Beispiel 2.40 können wir die a-posteriori-Verteilung ablesen. Diese ist gerade

$$\pi_x := \mathcal{N}\Big(\frac{nb^2}{nb^2 + \sigma^2}\overline{x}, \frac{\sigma^2 b^2}{nb^2 + \sigma^2}\Big).$$

Nach Proposition 3.17 ist nun der Bayes-Schätzer bezüglich π gegeben durch

$$d(x) = \int \theta \pi_x(d\theta) = \frac{nb^2}{nb^2 + \sigma^2}\overline{x}.$$

Lemma 3.19 (Zusammenhänge zulässige, Bayes-Entscheidungsfunktionen). *Betrachte dieselbe Situation wie in Lemma 3.14. Weiter sei $\pi \in \mathcal{M}_1(\mathcal{P})$.*

1. *Sei $\theta \mapsto R_{\delta'}(\theta)$ für alle $\delta \in \mathcal{D}_0$ stetig und π habe vollen Träger. Ist δ Bayes-Entscheidungsfunktion bezüglich π für $\mathcal{D}_0$, so ist δ zulässig bezüglich $\mathcal{D}_0$.*
2. *Ist δ die einzige Bayes-Entscheidungsfunktion bezüglich π und $\mathcal{D}_0$, so ist δ zulässig bezüglich $\mathcal{D}_0$.*

Beweis. 1. Angenommen, δ ist nicht zulässig $\mathcal{D}_0$. Dann gibt es ein $\delta' \in \mathcal{D}_0$ mit ($\delta' \neq \delta$ und) $R_{\delta'} \leq R_\delta$ und ein $\theta \in \mathcal{P}$ mit $R_{\delta'}(\theta) < R_\delta(\theta)$. Wegen der Stetigkeit von $R_{\delta'}$ gibt es ein $r > 0$, so dass für $\theta' \in B_r(\theta)$ gerade $R_{\delta'}(\theta') < R_\delta(\theta) - r$. Damit gilt

$$\begin{aligned} r_{\delta'}(\pi) &= \mathbb{E}_\pi[R_{\delta'}(\Theta), \Theta \in B_r(\theta)] + \mathbb{E}_\pi[R_{\delta'}(\Theta), \Theta \notin B_r(\theta)] \\ &\leq \mathbb{E}_\pi[R_\delta(\Theta)] - r\mathbb{P}(\Theta \in B_r(\theta)) < \mathbb{E}_\pi[R_\delta(\Theta)] = r_\delta(\pi) \end{aligned}$$

im Widerspruch dazu, dass δ Bayes-Entscheidungsfunktion bezüglich π für $\mathcal{D}_0$ ist.
2. Sei δ' so, dass $R_{\delta'} \leq R_\delta$. Dann gilt

$$r_{\delta'}(\pi) = \mathbb{E}[R_{\delta'}(\Theta)] \leq \mathbb{E}[R_\delta(\Theta)] = r_\delta(\pi) = \inf_{\varepsilon \in \mathcal{D}_0} r_\varepsilon(\pi).$$

Damit ist auch δ' eine Bayes-Entscheidungsfunktion für $\mathcal{D}_0$. Diese ist aber nach Voraussetzung eindeutig und damit $\delta' = \delta$. Damit ist δ zulässig für $\mathcal{D}_0$. □

Theorem 3.20 (Hodges-Lehmann). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem, $\delta, \delta_1, \delta_2, ... \in \mathcal{D}_0 \subseteq \mathcal{D}$ Entscheidungsfunktionen, $R_\delta, R_{\delta_1}, R_{\delta_2}, ...$ ihre Risikofunktionen und $\pi, \pi_1, \pi_2, ...$ Wahrscheinlichkeitsmaße auf $\mathcal{P}$ sowie $\Theta \sim \pi, \Theta_1 \sim \pi_1, ...$*

1. *Ist δ_n eine Bayes-Entscheidungsfunktion bezüglich π_n für $\mathcal{D}_0$, $n = 1, 2, ...$ und*

$$\sup_{\theta \in \mathcal{P}} R_\delta(\theta) \leq \limsup_{n \to \infty} r_{\delta_n}(\pi_n),$$

dann ist δ auch eine minimax-Entscheidungsfunktion für $\mathcal{D}_0$.

2. *Ist δ eine Bayes-Entscheidungsfunktion bezüglich π für $\mathcal{D}_0$ mit konstantem Risiko, so ist δ auch minimax für $\mathcal{D}_0$.*

Beweis. 1. Für $\delta \in \mathcal{D}_0$ und $k = 1, 2, ...$ gilt

$$\sup_{\theta\in\mathcal{P}} R_\delta(\theta) \geq \mathbb{E}_{\pi_k}[R_\delta(\Theta_k)] = r_\delta(\pi_k) \geq r_{\delta_k}(\pi_k).$$

Daraus folgt

$$\inf_{\delta'\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta) \limsup_{k\to\infty} r_{\delta_k}(\pi_k) \geq \sup_{\theta\in\mathcal{P}} R_\delta(\theta) \geq \inf_{\delta'\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta),$$

woraus insbesondere $\sup_{\theta\in\mathcal{P}} R_\delta(\theta) = \inf_{\delta'\in\mathcal{D}_0} \sup_{\theta\in\mathcal{P}} R_{\delta'}(\theta)$ folgt. Deshalb ist δ eine minimax-Entscheidungsfunktion für $\mathcal{D}_0$.
2. ist ein Spezialfall von 1. Für δ mit konstantem Risiko ist nämlich $\sup_{\theta\in\mathcal{P}} R_\delta(\theta) = \mathbb{E}_\pi[R_\delta(\Theta)]$. Damit ist δ auch eine Bayes-Entscheidungsfunktion für $\mathcal{D}_0$. □

Als Beispiel zeigen wir nun die Optimalität (im Sinne der Zulässigkeit) des arithmetischen Mittels zur Schätzung des Erwartungswertes bei normalverteilten Daten.

Proposition 3.21 (Zulässigkeit des arithmetischen Mittels). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem mit $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^n$ für $\sigma^2 > 0$, $n \in \mathbb{N}$, sowie $\aleph = \mathbb{R}$ und ℓ der Gauß-Verlust für die Funktion $g(\theta) = \theta$. Dann ist die Entscheidungsfunktion $d(x) := \overline{x}$ zulässig und minimax.*

Beweis. Die Risikofunktion von d ist

$$R_d(\theta) = \mathbb{E}_\theta[(\overline{X} - \theta)^2] = \frac{\sigma^2}{n}.$$

Angenommen, d ist nicht zulässig. Dann gibt es einen – nach Proposition 3.9 nichtrandomisierte – Schätzer d', der d dominiert. Es gilt also $R_{d'} \leq \sigma^2/n$ und es gibt ein $\theta \in \mathbb{R}$ mit $R_{d'}(\theta) < \sigma^2/n$. Wegen der Stetigkeit der Risikofunktion (in θ für alle Entscheidungsregeln) gibt es also ein $r > 0$, so dass $R_{d'}(\eta) < \sigma^2/n - r$ für $|\eta - \theta| < r$.

Sei nun $\pi_b = \mathcal{N}(0, b^2)$ für $b > 0$. Einerseits ist damit $r_{d'}(\pi_b) < \sigma^2/n$. Andererseits ist nach Beispiel 3.18 der Bayes-Schätzer bezüglich π_b gerade

$$d_b(x) := \frac{nb^2}{nb^2 + \sigma^2}\overline{x}$$

mit Bayes-Risiko

$$\begin{aligned} r_{d_b}(\pi_b) &= \int \mathbb{E}_\theta[(d_b(X) - \theta)^2]\pi_b(d\theta) = \int \mathbb{V}_\theta[d_b(X)] + (\mathbb{E}_\theta[d_b(X)] - \theta)^2\pi_b(d\theta) \\ &= \frac{n^2b^4}{(nb^2+\sigma^2)^2}\frac{\sigma^2}{n} + \frac{\sigma^4b^2}{(nb^2+\sigma^2)^2} = \frac{\sigma^2}{n}\frac{n^2b^4 + n\sigma^2b^2}{(nb^2+\sigma^2)^2} \\ &= \frac{\sigma^2}{n}\Big(\frac{n^2b^4 + 2nb^2\sigma^2 + \sigma^4 - nb^2\sigma^2 - \sigma^4}{(nb^2+\sigma^2)^2}\Big) = \frac{\sigma^2}{n}\Big(1 - \frac{nb^2\sigma^2 + \sigma^4}{(nb^2+\sigma^2)^2}\Big). \end{aligned}$$

Da d_b ein Bayes-Schätzer bezüglich π_b ist, gilt

$$\frac{2nb^2\sigma^2 + \sigma^4}{(nb^2+\sigma^2)^2} \geq \tfrac{\sigma^2}{n} - r_{d_b}(\pi_b) \geq \tfrac{\sigma^2}{n} - r_{d'}(\pi_b) \geq \frac{1}{\sqrt{2\pi b^2}} \int_{\theta-r}^{\theta+r} re^{-\eta^2/(2b^2)}d\eta.$$

Die linke Seite ist offenbar $O(1/b^2)$ für $b \to \infty$, die rechte jedoch $O(1/b)$. Dies ist offenbar ein Widerspruch und damit ist die Zulässigkeit von d gezeigt. Nun ist d auch minimax nach Lemma 3.14. □

Korollar 3.22 (Unbekannte Varianz). *Betrachte dieselbe Situation wie in Proposition 3.21, jedoch mit unbekannter Varianz σ^2, d.h. das statistische Modell $(X, \{\mathbb{P}_\theta : \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\})$ und $g(\mu, \sigma^2) = \mu$. Dann ist $d(x) := \overline{x}$ zulässig und minimax-Schätzer.*

Beweis. Wie im Beweis von Proposition 3.21 ist nur die Zulässigkeit von d zu zeigen. Angenommen, d wäre unzulässig. Dann gibt es einen (oBdA nicht-randomisierten) Schätzer d' mit $R_{d'} \leq R_d$ und ein $\theta = (\mu, \sigma^2)$ mit $R_{d'}(\theta) < R_d(\theta)$. Da $R_{d'}$ und R_d nicht davon abhängen, ob σ^2 bekannt oder unbekannt ist, würde d' den Schätzer nun bei bekannter Varianz σ^2 dominieren. Dies wiederspricht aber der Aussage von Proposition 3.21 und die Behauptung ist gezeigt. □

Bemerkung 3.23 (Höher-dimensionale Normalverteilungen). Interessanterweise lässt sich Proposition 3.21 (und damit Korollar 3.22) nur bedingt auf höhere Dimensionen verallgemeinern. Ist nämlich in der gleichen Situation $X = (X_1, ..., X_n)$ mit $X_i = (X_{i1}, ..., X_{ik}) \in \mathbb{R}^k$ nach $\mathbb{E}_\theta[X_{ij}] = \theta_j$ und $\mathbb{COV}[X_{ij}, X_{il}] = \delta_{jl}$, so kann man zeigen, dass für die Verlustfunktion $\ell(\theta, a) = \sum_{i=1}^k (\theta_i - a_i)^2 = (\theta - a)^\top (\theta - a)$ der Schätzer $d(x) = \overline{x}$ für $k \geq 3$ nicht mehr zulässig ist. Ein dominierender Schätzer ist durch den *James-Stein-Schätzer*

$$d'(x) := \Big(1 - \frac{k-2}{x^\top x}\Big)\overline{x}$$

gegeben. (Übrigens ist d' ebenfalls nicht zulässig und ist von

$$d''(x) := \Big(1 - \frac{k-2}{x^\top x}\Big)^+ \overline{x}$$

dominiert. Aber auch d'' ist nicht zulässig.)
Nun zum Beweis der Unzulässigkeit von d im Fall $k > 2$. OBdA sei $n = 1$, also $d(x) = x$, andersfalls ändert sich nur die Varianz von $d(X)$. Wir berechnen die Risikofunktion von d

$$R_d(\theta)\mathbb{E}_\theta[(\theta - d(X))^\top(\theta - d(X))] = k,$$

da $(\theta - X)^\top(\theta - X) \sim \chi^2_k$. Wir werden nun $R_{d'}(\theta) < k$ zeigen. Es ist nämlich

$$\begin{aligned} R_{d'}(\theta) &= \mathbb{E}_\theta\Big[\Big(X - \theta - \frac{(k-2)X}{X^\top X}\Big)^\top \Big(X - \theta - \frac{(k-2)X}{X^\top X}\Big)\Big] \\ &= \mathbb{E}_\theta[(X-\theta)^\top(X-\theta)] - 2(k-2)\mathbb{E}_\theta\Big[\frac{(X-\theta)^\top X}{X^\top X}\Big] + (k-2)^2\mathbb{E}_\theta\Big[\frac{1}{X^\top X}\Big] \\ &= k - 2(k-2)\sum_{i=1}^k \mathbb{E}_\theta\Big[\frac{(X_i - \theta_i)X_i}{X^\top X}\Big] + (k-2)^2\mathbb{E}_\theta\Big[\frac{1}{X^\top X}\Big]. \end{aligned}$$

Nun ist mit partieller Integration

$$\begin{aligned}\sum_{i=1}^k \mathbb{E}_\theta\Big[\frac{(X_i-\theta_i)X_i}{X^\top X}\Big] &= \sum_{i=1}^k \frac{1}{\sqrt{2\pi k}}\int \frac{x_1}{x^\top x}(x_i-\theta_i)\exp\big(-(x-\theta)^\top(x-\theta)/2\big)dx\\ &= \sum_{i=1}^k \frac{1}{\sqrt{2\pi k}}\int \frac{x^\top x-2x_1^2}{(x^\top x)^2}\exp\big(-(x-\theta)^\top(x-\theta)/2\big)dx\\ &= \frac{k-2}{\sqrt{2\pi k}}\int \frac{1}{x^\top x}\exp\big(-(x-\theta)^\top(x-\theta)/2\big)dx\\ &= (k-2)\mathbb{E}_\theta\Big[\frac{1}{X^\top X}\Big]\end{aligned}$$

und damit

$$R_\theta(d') = k-(k-2)\mathbb{E}_\theta\Big[\frac{1}{X^\top X}\Big] < k = R_\theta(d).$$

Wir wollen nun die entwickelte Theorie auf die Beispiele *Bern* und *Unif* anwenden.

Beispiel 3.24 (Beispiel *Bern*). Wir betrachten das Schätzproblem $((X,\{\mathbb{P}_\theta : \theta\in(0,1)\},\aleph=(0,1),g=\mathrm{id},\ell)$ für den Gauß-Verlust ℓ. Speziell werden wir zeigen, dass $d(x)=(\sum x_i)/n$ eine zulässige, aber keine minimax-Entscheidungsfunktion ist.
Wir wollen zunächst die Zulässigkeit von d zeigen. Offenbar minimiert d das Risiko genau dann für $\ell(\theta,a)=(\theta-a)^2$, wenn d das Risiko für $\ell'(\theta,a)=(\theta-a)^2/(\theta(1-\theta))$ minimiert. Sei nun $\pi=U(0,1)$ eine a-priori-Verteilung. Da dies genau die $\beta(1,1)$-Verteilung ist, sehen wir aus Beispiel 2.38, dass $\pi_x=\beta(1+\sum x_i, n+1-\sum x_i)$ die a-posteriori-Verteilung ist. Um den Bayes-Schätzer zu bestimmen, ist hierfür nach Theorem 3.16.2

$$\mathbb{E}[\ell(\Theta_x,a)] = \frac{\Gamma(n+2)}{\Gamma(1+\sum x_i)\Gamma(n+1-\sum x_i)}\int_0^1 (\theta-a)^2\theta^{\sum x_i-1}(1-\theta)^{n-1-\sum x_i}d\theta$$

zu minimieren. Offenbar liegt dieses Minimum genau beim Erwartungswert der $\beta(\sum x_i, n-\sum x_i)$-Verteilung, also bei[8] $a=(\sum x_i)/n$. Damit ist d Bayes-Schätzer und nach Lemma 3.19 auch zulässig.
Um zu zeigen, dass d nicht minimax ist sei $d_{a,b}(x) := \frac{a+\sum x_i}{a+b+n}$. Für die a-priori-Verteilung $\pi=\beta(a,b)$ ist die a-posteriori-Verteilung $\beta(a+\sum x_i, b+n-\sum x_i)$ und der Bayes-Schätzer ist gerade $d_{a,b}(x)$. Das Risiko dieses Schätzers ist

$$\begin{aligned}R_{d_{a,b}}(\theta) &= \mathbb{E}_\theta[(d_{a,b}(X)-\theta)^2] = \mathbb{V}_\theta[d_{a,b}(X)] + (\mathbb{E}_\theta[d_{a,b}(X)]-\theta)^2\\ &= \frac{1}{(a+b+n)^2}(n\theta(1-\theta)+(a+n\theta-(a+b+n)\theta)^2)\\ &= \frac{1}{(a+b+n)^2}(\theta^2((a+b)^2-n)+\theta(n-2a(a+b))+a^2).\end{aligned}$$

Für $a=b=\sqrt{n}/2$ ist dieser konstant $\frac{n/4}{(n+\sqrt{n})^2}$, und damit ist $d_{\sqrt{n}/2,\sqrt{n}/2}$ nach Theorem 3.20 minimax. Allerdings ist

$$\sup_{\theta\in(0,1)} R_d(\theta) = \sup_{\theta\in(0,1)} R_{d_{0,0}}(\theta) = \sup_{\theta\in(0,1)}\frac{n\theta(1-\theta)}{n^2} = \frac{1}{4n} > \frac{n/4}{(n+\sqrt{n})^2}$$

und damit ist d nicht minimax.

[8] Wir verwenden, dass für $\beta(p,q)$ der Erwartungswert durch $p/(p+q)$ und die Varianz durch $pq/((p+q+1)(p+q)^2)$ gegeben ist.

4 Testtheorie

In diesem Kapitel sei immer $((X, \{\mathbb{P}_\theta : \theta \in \Theta\}), \aleph, \ell)$ ein statistisches Entscheidungsproblem mit $\aleph = \{H_0, H_1\}$ (d.h. die Entscheidung besteht immer zwischen der Null-Hypothese H_0 und der Alternativ-Hypothese H_1). Wir erinnern an die Neyman-Pearson'schen Verlustfunktion

$$\ell(\theta, H_1) = \begin{cases} 0, & \theta \in \mathcal{P}_1, \\ \ell_0, & \theta \in \mathcal{P}_0 \end{cases},$$
$$\ell(\theta, H_0) = \begin{cases} 0, & \theta \in \mathcal{P}_0, \\ \ell_1, & \theta \in \mathcal{P}_1 \end{cases}.$$

für $\mathcal{P}_0, \mathcal{P}_1$ mit $\mathcal{P} = \mathcal{P}_0 \uplus \mathcal{P}_1$. Wir werden immer $\ell_0 = \ell_1 = 1$ betrachten. (Der allgemeinere Fall ergibt sich dann meist ebenfalls.) Die Entscheidung $a = H_0$ bedeutet, dass man sich für die Nullhypothese entscheidet (bzw. dass man sie nicht verwirft) und $a = H_1$ bedeutet, dass man die Nullhypothese verwirft und sich für die Alternative entscheidet. Eine Entscheidungsfunktion (bzw. der Test) δ wir eindeutig durch die Funktion

$$\varphi(x) := \delta(x, \{H_1\})$$

definiert, d.h. durch die Wahrscheinlichkeit, sich für die Alternativ-Hypothese zu entscheiden bei Vorliegen der Daten x. (Damit ist automatisch $\delta(x, \{H_0\}) = 1 - \varphi(x)$.) Das Entscheidungsproblem (oder auch Test-Problem) heißt *einfach*, wenn sowohl $\mathcal{P}_0$ als auch $\mathcal{P}_1$ ein-elementig sind. Andernfalls heißt es *zusammengesetzt*. Wir erinnern an die Gütefunktion $\beta_\varphi : \theta \mapsto \mathbb{E}_\theta[\varphi(X)]$. Das Risiko von δ (oder φ) ist gegeben durch

$$R_\varphi(\theta) := R_\delta(\theta) = \mathbb{E}_\theta\Big[\int \ell(\theta, a)\delta(X, da)\Big] = \begin{cases} \mathbb{E}_\theta[\varphi(X)] = \beta_\varphi(\theta), & \theta \in \mathcal{P}_0, \\ 1 - \mathbb{E}_\theta[\varphi(X)] = 1 - \beta_\varphi(\theta), & \theta \in \mathcal{P}_1. \end{cases}$$

Weiter sei Φ die Menge aller möglichen Tests (gegeben entweder durch δ oder φ) und $\{\theta \mapsto R_\delta(\theta) : \delta \in \Phi\}$ die Risikomenge, d.h. die Menge der Risikofunktionen aller Tests. Für einfache Tests mit $\mathcal{P}_0 = \{\mathbb{P}_0\}, \mathcal{P}_1 = \{\mathbb{P}_1\}$ identifizieren wir

$$\mathcal{R} = \{(\mathbb{E}_0[\varphi(X)], 1 - \mathbb{E}_1[\varphi(X)]) : \varphi \in \Phi\}. \tag{R}$$

4.1 Bayes-Tests

Wie wir in Theorem 3.16 gesehen haben, ist es oftmals gar nicht schwierig, Bayes-Entscheidungsfunktionen aufzustellen. Dies wollen wir nun für Tests durchführen, die entsprechenden Entscheidungsfunktionen heißen dann Bayes-Tests.

Proposition 4.1 (Bayes-Tests). *Sei π ein Wahrscheinlichkeitsmaß auf $\mathcal{P}$. Dann ist $\varphi \in \Phi$ genau dann ein Bayes-Test (d.h. eine Bayes-Entscheidungsfunktion für das Test-Problem) bezüglich π, wenn (für die a-posteriori-Verteilung π_x)*

$$\varphi(x) = \begin{cases} 1, & \pi_x(\mathcal{P}_0) < \pi_x(\mathcal{P}_1), \\ 0, & \pi_x(\mathcal{P}_0) > \pi_x(\mathcal{P}_1). \end{cases}$$

(Auf dem Bereich $\{x : \pi_x(\mathcal{P}_0) = \pi_x(\mathcal{P}_1)\}$ ist φ nicht eindeutig bestimmt.)

Beweis. Wir berechnen das Bayes-Risiko für $\Theta \sim \pi$ für $\Theta_x \sim \pi_x$ und den Test $\psi \in \Phi$ als

$$\begin{aligned} r_\delta(\pi) &= \mathbb{E}\Big[\mathbb{E}\Big[\int \ell(\Theta, a)\delta(X, da)\Big|X\Big]\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\ell(\Theta, H_1)\psi(X) + \ell(\Theta, H_0)(1-\psi(X))\Big|X\Big]\Big] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{\Theta_X \in \mathcal{P}_0}\psi(X) + \mathbb{1}_{\Theta_X \in \mathcal{P}_1}(1-\psi(X))|X]] \\ &= \mathbb{E}[\pi_X(\mathcal{P}_1) + (\pi_X(\mathcal{P}_0) - \pi_X(\mathcal{P}_1))\psi(X)]. \end{aligned}$$

Die Funktion ψ, die die rechte Seite minimiert, muss genau die angegebene Form haben. Daraus folgt die Behauptung. □

Korollar 4.2 (Bayes-Tests für einfache Tests). *Sei $\mathcal{P}_0 = \{\mathbb{P}_0 = p_0 \cdot \lambda\}, \mathcal{P}_1 = \{\mathbb{P}_1 = p_1 \cdot \lambda^n\}$ (d.h. wir betrachten einen einfachen Test für ein reguläres statistisches Modell) und $\pi_k(\mathcal{P}_0) = \frac{k}{k+1}$ (und damit $(\pi(\mathcal{P}_1) = \frac{1}{k+1})$ für ein $k \in [0, \infty]$[9]. Dann ist φ_k genau dann ein Bayes-Test bezüglich π_k, wenn φ_k gegeben ist als*

$$\begin{aligned} \varphi_k(x) &:= \begin{cases} 1, & \frac{p_1(x)}{p_0(x)} > k, \\ 0, & \frac{p_1(x)}{p_0(x)} < k \end{cases}, \\ \varphi_0(x) &:= 1_{p_1(x)>0}, \\ \varphi_\infty(x) &:= 1_{p_0(x)>0}. \end{aligned}$$

Beweis. Wir zeigen nur den Fall $0 < k < \infty$, die anderen beiden Fälle zeigt man direkt. Es gilt

$$\pi_x(\{H_0\}) = \frac{p_0(x)k}{p_0(x)k + p_1(x)}, \qquad \pi_x(\{H_1\}) = \frac{p_1(x)}{p_0(x)k + p_1(x)}.$$

Aus Proposition 4.1 folgt, dass φ_k Bayes-Tests sind mit

$$\pi_x(\{\mathbb{P}_0\}) < \pi_x(\{\mathbb{P}_1\}) \text{ genau dann, wenn } \frac{p_1(x)}{p_0(x)} > k.$$

□

4.2 Likelihood-Quotienten-Tests

Die soeben konstruierten Bayes-Tests werden gerade durch die Quotienten der Dichten der Alternative und der Nullhypothese bestimmt. Diese Dichten sind – bei gegebenen Daten – außerdem als Likelihood (des Parameters) bekannt. Deshalb nennt man diese Tests auch Likelihood-Quotienten-Tests.

Definition 4.3 (Likelihood-Quotienten-Test). *Sei $\mathcal{P}_0 = \{\mathbb{P}_0 = p_0 \cdot \lambda^n\}, \mathcal{P}_1 = \{\mathbb{P}_1 = p_1 \cdot \lambda^n\}$. Für $k \in [0, \infty]$ heißt φ_k aus Korollar 4.2* Likelihood-Quotienten-Test *(oder LQ-Test) mit kritischem Wert k. Genauer setzen wir*

$$\varphi_k(x) := \varphi_{k,\gamma}(x) := \begin{cases} 1, & \frac{p_1(x)}{p_0(x)} > k, \\ \gamma(x), & \frac{p_1(x)}{p_0(x)} = k, \\ 0, & \frac{p_1(x)}{p_0(x)} < k. \end{cases}$$

Die Menge $\left\{x : \frac{p_1(x)}{p_0(x)} = k\right\}$ heißt Randomisierungsbereich *von φ_k.*

[9] Wir setzen $\infty/\infty = 1$.

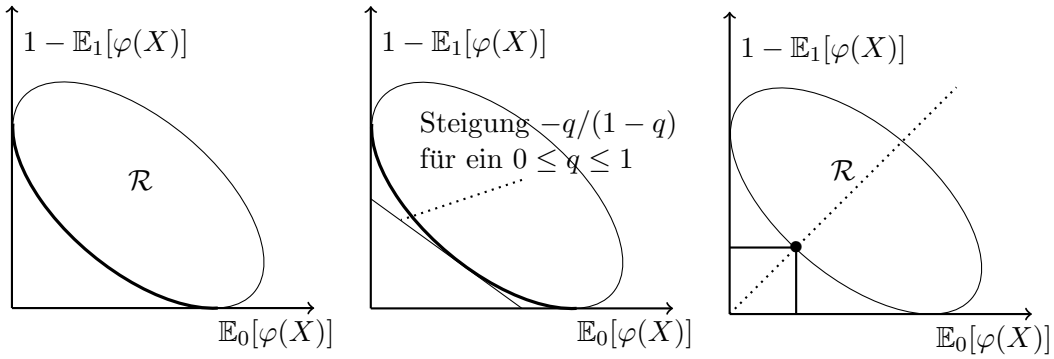
Beispiel 4.4 (Beispiel *Bern*). Im Beispiel *Bern* seien $\theta_0, \theta_1 \in (0,1)$ verschieden. Ein Likelihood-Quotiententest ist von der Form

$$\varphi(x) = \begin{cases} 1, & \frac{\theta_1^{\sum x_i}(1-\theta_1)^{n-\sum x_i}}{\theta_0^{\sum x_i}(1-\theta_0)^{n-\sum x_i}} > k, \\ 0, & \frac{\theta_1^{\sum x_i}(1-\theta_1)^{n-\sum x_i}}{\theta_0^{\sum x_i}(1-\theta_0)^{n-\sum x_i}} < k, \end{cases}$$

also für ein geeignetes k'

$$\varphi(x) = \begin{cases} 1, & \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_i} > k', \\ 0, & \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_i} < k', \end{cases}$$

Bemerkung 4.5 (Grafische Darstellung der Risikomenge für einfache Tests). Die Menge aller Tests Φ ist konvex, da die Konvexkombination zweier $[0,1]$-wertiger Funktionen wieder eine solche Funktion ergibt. Demnach ist auch die in (R) definierte Risikomenge für einfache Tests eine konvexe Teilmenge des $\mathbb{R}^2_+$. Weiter ist sowohl $\varphi = 0$ als auch $\varphi = 1$ erlaubt, so dass $\mathcal{R}$ sowohl die x- als auch die y-Achse berührt. Hieraus lassen sich die zulässigen, Bayes-Tests (also LQ-Tests) und minimax-Tests ablesen.



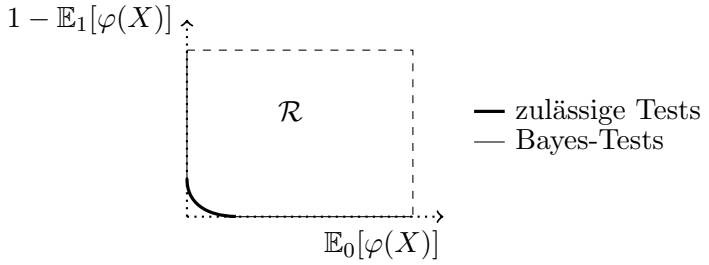
Zulässige Tests Bayes-Tests Minimax-Test

Zulässige Tests ergeben sich für solche φ, für die zwar $R_\varphi \in \mathcal{R}$, aber kein $(x,y) \le R_\varphi$ (in der üblichen Halbordnung in $\mathbb{R}^2_+$) in $\mathcal{R}$ ist.
Ein *Bayes-Test* φ bezüglich $\pi = (q, 1-q)$ minimiert gerade das Skalarprodukt

$$r_\varphi(\pi) = (q,1-q)(\mathbb{E}_0[\varphi(X)], 1 - \mathbb{E}_1[\varphi(X)]) = \min_{(x,y)\in\mathcal{R}} (q,1-q)(x,y).$$

Schreibt man $(x,y) = a(0,1) + b(1-q,-q)$ (d.h. eine parallelverschobene Gerade mit Steigung $-q/(1-q)$, so ist $(q,1-q)(x,y) = a(1-q)$. Dieses Minimum ergibt sich also gerade als die am wenigsten verschobene Gerade mit Steigung $-q/(1-q)$, die $\mathcal{R}$ berührt.
Wir wissen aus Theorem 3.20, dass ein zulässiger Test mit konstantem Risiko ein *minimax-Test* ist. Dies ist gerade für $\mathbb{E}_0[\varphi(X)] = 1 - \mathbb{E}_1[\varphi(X)]$ gegeben, also für Schnittpunkte von $\mathcal{R}$ mit $\{(x,x) : x \ge 0\}$.

Es gibt in diesem Bild die Möglichkeit, dass die Menge der Bayes-Tests größer ist als die der zulässigen Tests, was wir nun veranschaulichen wollen.



Hier bestehen die minimalen Geraden, die $\mathcal{R}$ berühren, ebenfalls aus den horizontalen und vertikalen Begrenzungen von $\mathcal{R}$. Diese sind jedoch keine zulässigen Tests, weil es Punkte $(x, y) \in \mathcal{R}$ gibt, die diese Tests dominieren.

Proposition 4.6 (Zulässige und minimax-LQ-Tests). *Sei $\mathcal{P}_0 = \{\mathbb{P}_0 = p_0 \cdot \lambda^n\}, \mathcal{P}_1 = \{\mathbb{P}_1 = p_1 \cdot \lambda^n\}$ und φ_k für $k \in [0, \infty]$ ein Likelihood-Quotienten-Test mit kritischem Parameter k. Dann gilt:*

1. *Ist φ ein zulässiger Test, so gibt es $k \in [0, \infty]$ mit $\varphi = \varphi_k$.*
2. *Gilt $\mathbb{E}_0[\varphi_k(X)] > 0, \mathbb{E}_1[\varphi_k(X)] < 1$, so ist φ_k zulässig.*
3. *Genau dann ist φ ein minimax-Test, wenn es $k \in [0, \infty]$ gibt mit $\varphi = \varphi_k$ und $\mathbb{E}_0[\varphi_k(X)] = \mathbb{E}_1[1 - \varphi_k(X)]$.*

Beweis. 1. Wir verwenden die grafische Darstellung aus Bemerkung 4.5. Ist φ mit $(\mathbb{E}_0[\varphi(X), 1 - \mathbb{E}_1[\varphi(X)]) \in \mathcal{R}$ zulässig (also Element des unteren Randes der Risikomenge), so lesen wir ab, dass es eine minimale berührende Gerade gibt, die $\mathcal{R}$ gerade in $(\mathbb{E}_0[\varphi(X), 1 - \mathbb{E}_1[\varphi(X)])$ berührt. Damit ist φ ein Bayes-Test, also nach Korollar 4.2 auch ein LQ-Test.
2. Wir müssen ausschließen, dass $k = 0$ oder $k = \infty$. Für $k \in (0, \infty)$ ist nämlich φ_k Bayes-Test zu einer a-priori-Verteilung mit vollem Träger und damit nach Lemma 3.19 zulässig. Offenbar ist $\mathbb{P}_i[p_i(X) > 0] = 1$, $i = 0, 1$. Wäre $k = 0$, so gilt aber $\mathbb{E}_1[\varphi_0(X)] = \mathbb{P}_1(p_1(X) > 0) = 1$ im Widerspruch zur Voraussetzung und wäre $k = \infty$, so gilt $\mathbb{E}_0[\varphi_\infty(X)] = \mathbb{P}_0[p_0(X) > 0] = 1$.
3. '$\Leftarrow$': Offenbar ist φ_k Bayes-Test mit konstantem Risiko. Deshalb folgt die Behauptung aus Theorem 3.20.2.
'$\Rightarrow$': Wir zeigen zunächst, dass $\mathbb{E}_0[\varphi(X)] = 1 - \mathbb{E}_1[\varphi(X)]$ gelten muss. Angenommen, es wäre $\lambda := 1 - \mathbb{E}_1[\varphi(X)] - \mathbb{E}_0[\varphi(X)] > 0$. Wir definieren dann die neue Entscheidungsfunktion

$$\psi(X) := \frac{1}{1+\lambda}\varphi(X) + \frac{\lambda}{1+\lambda} \in \Phi.$$

Nun gilt

$$\begin{aligned}\mathbb{E}_0[\psi(X)] &= \frac{1}{1+\lambda}\mathbb{E}_0[\varphi(X)] + \frac{1}{1+\lambda}(1 - \mathbb{E}_1[\varphi(X)] - \mathbb{E}_0[\varphi(X)]) \\ &= \frac{1}{1+\lambda}(1 - \mathbb{E}_1[\varphi(X)]) \\ &= 1 - \frac{1}{1+\lambda}\mathbb{E}_1[\varphi(X)] - \frac{\lambda}{1+\lambda} = 1 - \mathbb{E}_1[\psi(X)].\end{aligned}$$

Weiter gilt

$$\begin{aligned}\max\big(\mathbb{E}_0[\psi(X)], 1-\mathbb{E}_1[\psi(X)]\big) &= \frac{1}{1+\lambda}(1-\mathbb{E}_1[\varphi(X)]) < 1-\mathbb{E}_1[\varphi(X)] \\ &= \max\big(\mathbb{E}_0[\varphi(X)], 1-\mathbb{E}_1[\varphi(X)]\big).\end{aligned}$$

Dies steht im Widerspruch dazu, dass φ minimax ist. Im Fall $\lambda < 0$ führt eine ähnliche Konstruktion zum Ziel. Nun ist also $(\mathbb{E}_0[\varphi(X)], 1-\mathbb{E}_0[\varphi(X)])) \in \mathcal{R} \cap \{(x,x) : x \in \mathbb{R}_+\}$. Aus der grafischen Darstellung aus Bemerkung 4.5 und der Konvexität der Risikomenge $\mathcal{R}$ folgt, dass φ ein Bayes-Test sein muss und damit ein LQ-Test. □

Auch für zusammengesetzte Hypothesen kann man Likelihood-Quotiententests definieren. Dies wollen wir nun tun, auch wenn wir in diesem Fall nur Beispiele angeben und die Optimalität der Tests nicht weiter untersuchen werden.

Definition 4.7 (Likelihood-Quotienten-Tests für zusammengesetzte Alternativen). *Sei $(X, \{\mathcal{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}_0\}, \aleph, \ell\})$ ein reguläres statistisches Modell, $\mathcal{P} = \mathcal{P}_0 \uplus \mathcal{P}_1$ eine Partition von $\mathcal{P}$ und $H_0 : \theta \in \mathcal{P}_0$, $H_1 : \theta \in \mathcal{P}_1$, sowie $\aleph = \{H_0, H_1\}$ und ℓ der Neyman-Pearon-Verlust. Dann heißt ein Test φ_k mit*

$$\varphi_k(x) := \varphi_{k,\gamma}(x) := \begin{cases} 1, & \lambda(x) := \frac{\sup_{\theta\in\mathcal{P}_0} p_\theta(x)}{\sup_{\theta\in\mathcal{P}} p_\theta(x)} < k, \\ \gamma(x), & \lambda(x) = k, \\ 0, & \lambda(x) > k \end{cases}$$

Likelihood-Quotienten-Test *mit kritischem Wert k.*

Bemerkung 4.8 (Mögliche Werte für k). Anders als bei Likelihood-Quotienten-Tests von einfachen Hypothesen ist die Definition bei zusammengesetzten Hypothesen so, dass immer $\lambda \leq 1$ ist. Es machen also nur $k \in [0,1]$ Sinn. Diese Werte legen natürlich auch das Signifikanzniveau fest.

Beispiel 4.9 (Test auf den Parameter einer Exponentialverteilung). Sei $\mathcal{P} = [\theta_0, \infty)$ und $\mathbb{P}_\theta = \exp(\theta)^n$. Es ist unter $\mathbb{P}_\theta$ also $X = (X_1, ..., X_n)$ unabhängig und $X_i \sim \exp(\theta)$. Weiter sei $\mathcal{P}_0 = \{\theta_0\}$ und $\mathcal{P}_1 = (\theta_0, \infty)$. Nun ist φ genau dann ein Likelihood-Quotienten-Test, falls

$$\varphi(x) := 1_{\{\sum_{i=1}^n x_i > c\}}$$

für ein $c > 0$.
Denn: Zunächst berechnen wir $\sup_{\theta\in\mathcal{P}} p_\theta(x)$. Wir schreiben hierfür

$$\log p_\theta(x) = \log(\theta^n e^{-\theta(x_1+\cdots+x_n)}) = n\log\theta - \theta(x_1+\cdots+x_n)$$

und damit $\frac{d}{d\theta}\log p_\theta(x) = \frac{n}{\theta} - (x_1+\cdots+x_n)$, also

$$\log\frac{p_{\theta_0}(x)}{\sup_{\theta\in\mathcal{P}} p_\theta(x)} = \begin{cases} \log p_{\theta_0(x)} - \log p_{1/\overline{x}}(x) = n\log\theta_0 - \theta_0 n\overline{x} + n\log\overline{x} - n, & \overline{x} < 1/\theta_0, \\ 0, & \overline{x} \geq 1/\theta_0. \end{cases}$$

Damit ist ein Likelihood-Quotienten-Test von der Form $\varphi(x) = 1_{\log\overline{x}-\theta_0\overline{x}<k} 1_{\overline{x}>1/\theta_0}$. Nun ist $\overline{x} \mapsto \log\overline{x} - \theta_0\overline{x}$ für $\overline{x} \geq 1/\theta_0$ monoton fallend, d.h. φ ist von der Form $\varphi(x) = 1_{\overline{x}>k'}$.

Bereits bekannte Tests, etwa der t-Test, sind ebenfalls Likelihood-Quotienten-Tests.

Proposition 4.10 (Einfacher t-Test ist ein Likelihood-Quotienten-Test). *Sei $(X, \{\mathbb{P}_\theta = \mathcal{N}(\mu,\sigma^2)^n : \theta = (\mu,\sigma)^2 \in \mathcal{P}_0 := \mathbb{R}\times\mathbb{R}_+\})$ das Normalverteilungsmodell, $\mathcal{P}_0 = \{\mathbb{P}_\theta : \theta = (\mu_0,\sigma^2) \in \{\mu_0\}\times\mathbb{R}_+\}$, $\mathcal{P}_1 = \mathcal{P}\setminus\mathcal{P}_0$, $\aleph = \{H_0, H_1\}$ und ℓ der Neyman-Pearson-Verlust. Dann ist φ genau dann ein Likelihood-Quotienten-Test, wenn $\varphi(x) = 1_{|t(x)|\geq c}$ für ein geeignetes c mit*

$$t(x) = \frac{\overline{x}-\mu_0}{\sqrt{s^2(x)/n}}.$$

Beweis. Sei $\mathbb{P}_\theta = p_\theta\cdot\lambda^n$ mit

$$p_\theta(x) = (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{\sum_{i=1}^n(x_i-\theta)^2}{2\sigma^2}\Big).$$

Bekannt ist, dass der Maximum-Likelihood-Schätzer für $\theta = (\mu,\sigma^2)$ immer durch $(\overline{x}, (n-1)s^2(x)/n)$ gegeben ist. Bei gegebenen μ ist der Maximum-Likelihood-Schätzer von σ^2 durch $\hat{s}^2(x) := \frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2$ gegeben. Um also p_θ zu maximieren, berechnen wir erst

$$\begin{aligned}\sup_{\sigma^2\in\mathbb{R}_+} p_\theta(x) &= (2\pi\hat{s}^2(x))^{-n/2}\exp\Big(-\frac{\sum_{i=1}^n(x_i-\mu)^2}{2\hat{s}^2(x)}\Big)\\ &= \Big(\frac{2\pi}{n}\sum_{i=1}^n(x_i-\mu)^2\Big)^{-n/2}e^{-n/2}\end{aligned}$$

und damit

$$\begin{aligned}\lambda(x) &= \Big(\frac{\sum_{i=1}^n(x_i-\mu_0)^2}{\sum_{i=1}^n(x_i-\overline{x})^2}\Big)^{-n/2} = \Big(\frac{\sum_{i=1}^n(x_i-\overline{x})^2+(\overline{x}-\mu_0)^2}{\sum_{i=1}^n(x_i-\overline{x})^2}\Big)^{-n/2}\\ &= \Big(1+|t(x)|^2\frac{n-1}{n}\Big)^{-n/2}.\end{aligned}$$

Die Behauptung folgt nun daraus, dass $x\mapsto(1+(n-1)x/n)^{-n/2}$ monoton ist. □

Bemerkung 4.11 (Approximative Verteilung des Likelihood-Quotienten). Um das Signifikanzniveau eines Tests φ zu bestimmen, benötigt man $\sup_{\theta\in\mathcal{P}_0}\mathbb{E}_\theta[\varphi(X)]$. Da bei Likelihood-Quotienten-Tests $\varphi = \varphi_{k,\gamma}$ von den Parametern k und γ abhängt, heißt das, dass man diese so bestimmen muss, dass

$$\mathbb{E}_\theta[\varphi_{k,\gamma}(X)] = \mathbb{P}_\theta[\lambda(X)\leq k] + \mathbb{E}_\theta[\gamma(X),\lambda(X)=k] \leq \alpha,\quad \theta\in\mathcal{P}_0.$$

Hierzu benötigt man also die Verteilung des Likelihood-Quotienten λ.

Sei $\mathcal{P}\subseteq\mathbb{R}^p$ und $\mathcal{P}_0 = \{\eta\}$. Wir werden (mit Hilfe der noch zu zeigenden Aussagen aus Abschnitt 5.4) nun zeigen, dass (unter gewissen Regularitätsannahmen)

$$-2\log\lambda(X) =: \Lambda(X) \overset{n\to\infty}{\Longrightarrow} Z\sim\chi^2_p.$$

Sei $\hat{\theta}$ der Maximum-Likelihood-Schätzer für θ. Dann schreiben wir

$$\begin{aligned}\sum_{k=1}^n\log p_\eta^n(x_k) = \sum_{k=1}^n\Big(&\log p_{\hat{\theta}}(x_k) + \sum_{i=1}^p\frac{\partial}{\partial\theta_i}\log p_{\hat{\theta}}(x_k)(\eta_i-\hat{\theta}_i)\\ &+\frac{1}{2}\sum_{i,j=1}^k\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\log p_{\hat{\theta}}(x_k)(\eta_i-\hat{\theta}_i)(\eta_j-\hat{\theta}_j)\Big) + o((\eta-\hat{\theta})^2).\end{aligned}$$

Nun verschwindet der zweite Term, da $\hat{\theta}$ die Maximalstelle von $\log p_\theta$ ist, und damit ist für $X \sim \mathbb{P}_\eta^n$

$$\begin{aligned}\Lambda(X) &= 2\sum_{k=1}^{n} \log p_{\hat{\theta}}(X_k) - 2\log p_\eta(X_k) \\ &= -\sum_{k=1}^{n}\sum_{i,j=1}^{p} \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j} \log p_{\hat{\theta}}(X_k)(\eta_i - \hat{\theta}_i)(\eta_j - \hat{\theta}_j) + o((\eta-\hat{\theta})^2) \\ &\overset{n\to\infty}{\approx} -n\sum_{i,j=1}^{p} \mathbb{E}_\eta\Big[\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\log p_\eta(X)\Big](\eta_i - \hat{\theta}_i)(\eta_j - \hat{\theta}_j),\end{aligned}$$

da $\hat{\theta}$ konsistent ist (siehe Theorem 5.30). Nun ist nach Theorem 5.31 und Bemerkung 5.18 für

$$I(\theta) := -\Big(\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\Big]\Big)_{i,j=1,\dots,k}$$

gerade (für die Einheitsmatrix E_k)

$$\sqrt{n}(I(\eta))^{1/2}(\hat{\theta}-\eta) \overset{n\to\infty}{\Longrightarrow} \mathcal{N}(0, E_p),$$

also

$$\Lambda(X) \approx n(\hat{\theta}-\eta)I(\eta)(\hat{\theta}-\eta) \overset{n\to\infty}{\Longrightarrow} Z_1^2 + \cdots + Z_p^2 \sim \chi_p^2$$

4.3 Beste Tests

Optimale Tests minimieren die Risikofunktion. Allerdings ist es nicht sinnvoll, alle Tests zuzulassen, sondern nur solche, deren Fehler erster Art kleiner als ein vorgegebenes α sind. Diese Tests heißen auch beste Tests.

Definition 4.12 (Niveau eines Tests). *1. Für $\alpha \in [0,1]$ ist*

$$\Phi_\alpha := \{\varphi \in \Phi : \mathbb{E}_\theta[\varphi(X)] \le \alpha \text{ für alle } \theta \in \mathcal{P}_0\}$$

die Menge aller Tests zum Niveau α.

2. Ein Test φ heißt (gleichmäßig) bester Tests zum Niveau α *(oder* (Uniformly) Most Powerful Test oder UMP-Test*), falls*

$$\mathbb{E}_\theta[\varphi(X)] = \sup_{\psi\in\Phi_\alpha} \mathbb{E}_\theta[\psi(X)], \qquad \theta \in \mathcal{P}_1.$$

Das folgende Resultat erlaubt die Konstruktion bester Tests im Falle von einfachen Hypothesen.

Theorem 4.13 (Neyman-Pearson-Lemma). *Wir betrachten das reguläre statistische Modell $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ mit $\mathcal{P}_i = \{\mathbb{P}_i\}, i = 0,1$ und $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$. Sei $\alpha \in (0,1)$.*

1. Es gibt einen LQ-Test $\varphi_{k,\gamma}$ mit $\gamma \in [0,1]$ konstant und $\mathbb{E}_0[\varphi(X)] = \alpha$.

2. Ist φ ein LQ-Test mit $\mathbb{E}_0[\varphi(X)] = \alpha$, so ist φ bester Test zum Niveau α.

3. *Ist φ ein bester Test zum Niveau α, so ist φ ein LQ-Test. Es gilt dann entweder $\mathbb{E}_0[\varphi(X)] = \alpha$ oder $\mathbb{E}_1[\varphi(X)] = 1$.*

Beweis. 1. Wir setzen

$$k := \inf\{y : \mathbb{P}_0(p_1(X)/p_0(X) > y) \leq \alpha\}$$

und haben damit k bereits bestimmt. Es gilt $\mathbb{P}_0(p_1(X)/p_0(X) \geq k) \geq \alpha \geq \mathbb{P}_0(p_1(X)/p_0(X) > k)$. Für die Wahl von γ gibt es nun zwei Möglichkeiten. Ist $\mathbb{P}_0(p_1(X)/p_0(X) = k) = 0$, so setzen wir $\gamma = 0$. Andernfalls setzen wir

$$\gamma := \frac{\alpha - \mathbb{P}_0(p_1(X)/p_0(X) > k)}{\mathbb{P}_0(p_1(X)/p_0(X) = k)} \in (0, \infty).$$

Nun gilt nach Definition des LQ-Tests $\varphi^*_{k,\gamma}$

$$\mathbb{E}_0[\varphi^*_{k,\gamma}(X)] = \mathbb{P}_0(p_1(X)/p_0(X) > k) + \gamma \mathbb{P}_0(p_1(X)/p_0(X) = k) = \alpha.$$

2. Sei $\varphi = \varphi^*_k$ für ein $k \in [0, \infty]$ sowie $\psi \in \Phi_\alpha$. Wir müssen zeigen, dass

$$\mathbb{E}_1[\varphi^*_k(X)] \geq \mathbb{E}_1[\psi(X)]$$

und schreiben hierfür

$$\begin{aligned}
&\mathbb{E}_1[\varphi^*(X)] - \mathbb{E}_1[\psi(X)] \\
&\qquad = \int (\varphi^*(x) - \psi(x))(p_1(x) - k p_0(x))\lambda^n(dx) + k \int (\varphi^*_k(x) - \psi(x)) p_0(x) \lambda^n(dx) \\
&\qquad = \int (1_{p_1(x) > k p_0(x)} - \psi(x))(p_1(x) - k p_0(x))\lambda^n(dx) + k \mathbb{E}_0[\varphi^*_k(X) - \psi(X)] \\
&\qquad \geq \int 1_{p_1(x) > k p_0(x)} (1 - \psi(x))(p_1(x) - k p_0(x))\lambda^n(dx) \geq 0.
\end{aligned}$$

3. Sei φ bester Test zum Niveau α und φ^*_k der LQ-Test zum Niveau α aus 1. Nach 2. ist φ^*_k ebenfalls bester Test zum Niveau α und wir müssen zeigen, dass $\varphi = \varphi^*_k$. In der Rechnung aus dem Beweis von 2. muss Gleichheit in beiden Abschätzungen gelten, es gilt also $\mathbb{E}_0[\varphi(X)] = \mathbb{E}_0[\varphi^*_k(X)] = \alpha$ sowie $(\varphi^*_k - \varphi)(p_1 - k p_0) \overset{\lambda^n\text{-f.ü.}}{=} 0$. Daraus folgt $\{x : \varphi^*_k(x) \neq \varphi(x)\} \subseteq \{x : p_1(x)/p_0(x) = k\}$, d.h. φ muss ein LQ-Test sein, der höchstens auf dem Randomisierungsbereich nicht mit φ^*_k übereinstimmt. □

Ziel des restlichen Kapitels ist es, die Optimalität von Tests bei zusammengesetzten Hypothesen zu zeigen. Dies ist vor allem dann möglich, wenn das statistische Modell stochastisch geordnet ist.

Definition 4.14 (Monotone Dichtequotienten). *Sei $(\mathcal{P}, \leq)$ total geordnet und $(X, \{\mathcal{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell und $T = t(X)$ für $t : \mathbb{R}^n \to \mathbb{R}$. Dann hat $\mathcal{P}$ einen monotonen Dichtequotienten in t, wenn für $\theta, \theta' \in \mathcal{P}$ mit $\theta \leq \theta'$ eine monotone Abbildung $p_{\theta,\theta'}$ existiert mit*

$$\frac{p_{\theta'}}{p_\theta} = p_{\theta,\theta'} \circ t$$

($(p_\theta + p_{\theta'}) \cdot \lambda^n$-fast sicher).

Beispiel 4.15 (Exponentialfamilie). Sei $p_\theta(x) = h(x)\exp\big(c(\theta)^\top t(x) - d(\theta)\big)$, d.h. $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n\})$ ist eine ein-parametrige Exponentialfamilie. Dann ist

$$\frac{p_{\theta'}(x)}{p_\theta(x)} = \exp\big((c(\theta') - c(\theta))^\top t(x) - (d(\theta') - d(\theta))\big),$$

also hat $\mathcal{P}$ genau dann einen monotonen Dichtequotienten in t, wenn $\mathcal{P}$ total geordnet ist und c monoton ist.

Beispiel 4.16 (Beispiel *Norm*). Für das Modell aus Beispiel 2.23 mit bekanntem $\sigma^2 > 0$ ist

$$p_\theta(x) = \exp\Big(-\frac{x^2}{2\sigma^2}\Big) \cdot \exp\Big(\frac{\theta x}{\sigma^2} - \frac{1}{2}\Big(\frac{\theta^2}{\sigma^2} + \log(2\pi\sigma^2)\Big)\Big),$$

es handelt sich also um eine 1-parametrige Exponentialfamilie mit

$$c(\theta) = \frac{\theta}{\sigma^2}, \qquad t(x) = x,$$
$$h(x) = \exp\Big(-\frac{x^2}{2\sigma^2}\Big), \qquad d(\theta) = -\frac{1}{2}\Big(\frac{\theta^2}{\sigma^2} + \log(2\pi\sigma^2)\Big).$$

insbesondere ist c monoton, und damit hat $\mathcal{P}$ einen monotonen Dichtequotienten.

Proposition 4.17 (Stochastische Monotonie). *Sei $(\mathcal{P}, \leq)$ total geordnet und $(X, \{\mathcal{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell und $T = t(X)$ für $t : \mathbb{R}^n \to \mathbb{R}$. Sei f isoton (und so, dass $\mathbb{E}_\theta[f(t(X))]$ für alle $\theta \in \mathcal{P}$ existiert), und hat $\mathcal{P}$ einen monotonen Dichtequotienten in t, dann ist $\theta \mapsto \mathbb{E}_\theta[f(t(X))]$ ebenfalls isoton.*

Beweis. Wir schreiben für $\theta \leq \theta'$

$$\begin{aligned}
\mathbb{E}_{\theta'}[f(t(X))] - \mathbb{E}_\theta[f(t(X))] &= \int (f(t(y)) - f(t(x)))p_{\theta'}(y)p_\theta(x)\lambda^n(dx)\lambda^n(dy) \\
&= \int 1_{t(y)>t(x)}((f(t(y)) - f(t(x)))p_{\theta'}(y)p_\theta(x) \\
&\qquad + (f(x) - f(y))p_{\theta'}(x)p_\theta(y))\lambda^n(dx)\lambda^n(dy) \\
&= \int 1_{t(y)>t(x)}(f(t(y)) - f(t(x)))(p_{\theta'}(y)p_\theta(x) - p_{\theta'}(x)p_\theta(y))\lambda^n(dx)\lambda^n(dy) \\
&\geq 0,
\end{aligned}$$

da $f \circ t$ isoton ist und $\frac{p_{\theta'}(y)}{p_\theta(y)} = p_{\theta,\theta'}(t(y)) \geq p_{\theta,\theta'}(t(x)) = \frac{p_{\theta'}(x)}{p_\theta(x)}$ auf $\{t(y) > t(x)\}$. □

Theorem 4.18 (Beste Tests bei einseitigen, zusammengesetzten Hypothesen). *Sei $(\mathcal{P}, \leq)$ total geordnet und $(X, \{\mathcal{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell, $T = t(X)$ für $t : \mathbb{R}^n \to \mathbb{R}$ und habe $\mathcal{P}$ einen streng monotonen Dichtequotienten in t. Weiter sei $\alpha \in [0, 1]$ und $\theta_0 \in \mathcal{P}$ so, dass $\mathcal{P}_0 = \{\theta \leq \theta_0\}, \mathcal{P}_1 = \{\theta > \theta_0\}$. Dann gilt:*

1. *Sei*

$$\psi(x) := \psi_{m,\gamma}(x) := \begin{cases} 1, & t(x) > m, \\ \gamma, & t(x) = m, \\ 0, & t(x) < m \end{cases}$$

für ein $m \in [0, \infty]$ und $\gamma \in [0, 1]$, so dass $\mathbb{E}_{\theta_0}[\psi(X)] = \alpha$. Dann ist ψ bester Test zum Niveau α.

2. *Die Gütefunktion* $\beta_\psi : \theta \mapsto \mathbb{E}_\theta[\psi(X)]$ *ist isoton und für* $\theta < \theta_0$ *gilt*

$$\mathbb{E}_\theta[\psi(X)] = \inf\{\mathbb{E}_\theta[\varphi(X)] : \varphi \in \Phi, \mathbb{E}_{\theta_0}[\varphi(X)] \geq \alpha\},$$

d.h. ψ *minimiert den Fehler erster Art.*

Beweis. Wir beginnen mit dem Beweis von 1. Sei zunächst $\theta_1 > \theta_0$. Wir betrachten den einfachen Test mit $\mathcal{P}_i = \{\theta_i\}, i = 0, 1$ und zeigen, dass ψ auch in diesem Fall ein LQ-Test ist. Wegen des streng monotonen Dichtequotienten ist für $k := p_{\theta_0,\theta_1}(m)$

$$\begin{aligned}
\{t(x) > m\} &= \{p_{\theta_0,\theta_1}(t(x)) > p_{\theta_0,\theta_1}(m)\} = \{p_{\theta_1}(x)/p_{\theta_0}(x) > k\},\\
\{t(x) = m\} &= \{p_{\theta_0,\theta_1}(t(x)) = p_{\theta_0,\theta_1}(m)\} = \{p_{\theta_1}(x)/p_{\theta_0}(x) = k\},\\
\{t(x) < m\} &= \{p_{\theta_0,\theta_1}(t(x)) < p_{\theta_0,\theta_1}(m)\} = \{p_{\theta_1}(x)/p_{\theta_0}(x) < k\}.
\end{aligned}$$

Damit ist ψ also LQ-Test und $\mathbb{E}_{\theta_0}[\psi(X)] = \alpha$ nach Voraussetzung. Nach dem Neyman-Pearson-Lemma, Theorem 4.13, ist also ψ bester Test zum Niveau α, d.h. es gilt $\mathbb{E}_{\theta_1}[\psi(X)] = \sup_{\varphi\in\Phi_\alpha} \mathbb{E}_{\theta_1}[\varphi(X)]$. Da dies aber für alle $\theta_1 \in \mathcal{P}_1$ gilt, folgt, dass ψ bester Test zum Niveau α für $\mathcal{P}_0 = \{\theta_0\}, \mathcal{P}_1 = \{\theta > \theta_0\}$ ist. Wegen $\psi = 1_{\{.\geq m\}} \circ t$, der Monotonie von $1_{\{.\geq m\}}$ und da $\mathcal{P}$ einen monotonen Dichtequotienten in t hat, folgt aus Proposition 4.17, dass $\theta \mapsto \mathbb{E}_\theta[\psi(X)]$ monoton ist (d.h. die in 2. behauptete Isotonie der Gütefunktion ist gezeigt). Insbesondere gilt für $\theta \leq \theta_0$, dass $\mathbb{E}_\theta[\psi(X)] \leq \mathbb{E}_{\theta_0}[\psi(X)] = \alpha$, also $\psi \in \Phi_\alpha$. Nach Definition ist damit ψ bester Test zum Niveau für $\mathcal{P}_0 = \{\theta \leq \theta_0\}$, $\mathcal{P}_1 = \{\theta > \theta_0\}$. Es ist noch zu zeigen, dass ψ den Fehler erster Art minimiert. Sei hierzu $\theta < \theta_0$. Genau wie in 1. zeigt man, dass $1-\psi$ ein LQ-Test zum Niveau $1-\alpha$ ist für $\mathcal{P}_0 = \{\theta_0\}, \mathcal{P}_1 = \{\theta < \theta_0\}$. Insbesondere ist $1-\psi$ bester Test zum Niveau $1-\alpha$, d.h. für $1-\varphi \in \Phi_{1-\alpha}$ (oder $\mathbb{E}_{\theta_0}[1-\varphi(X)] \leq 1-\alpha$) ist $\mathbb{E}_\theta[1-\varphi] \leq \mathbb{E}_\theta[1-\psi], \theta < \theta_0$, also auch $\mathbb{E}_\theta[\psi] \leq \mathbb{E}_\theta[\varphi]$ für $\mathbb{E}_\theta[\varphi(X)] \geq \alpha$ und damit die Behauptung. □

Korollar 4.19 (Gauß-Test). *Sei* $\mathcal{P} = \mathbb{R}$, $(X, \{\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^N)$ *für ein* $\sigma^2 > 0$ *wie in* (Norm 1*b*) *und* $\theta_0 \in \mathbb{R}$ *mit* $\mathcal{P}_0 = (-\infty, \theta_0], \mathcal{P}_1 = (\theta_0, \infty)$. *Weiter sei* q_α *das* α*-Quantil von* $\mathcal{N}(0, 1)$. *Dann ist*

$$\psi(x) := 1\Big(\frac{\overline{x} - \theta_0}{\sqrt{\sigma^2/n}} \geq q_\alpha\Big)$$

bester Test zum Niveau α.

Beweis. Nach Beispiel 2.23 und Bemerkung 2.28 ist $\mathcal{N}(\theta, \sigma^2)^n$ eine ein-parametrige Exponentialfamilie mit $t(x) = \overline{x}$. Deshalb hat diese Familie nach Proposition 4.17 einen monotonen Dichtequotienten. Weiter ist $\mathbb{E}_{\theta_0}[\psi(X)] = \mathbb{P}(Z \geq q_\alpha) = \alpha$ für $Z \sim \mathcal{N}(0, 1)$. Damit ist ψ nach Theorem bester Test zum Niveau α. □

Bemerkung 4.20 (Weitere beste Tests). Wir haben nun den einfachsten Fall eines besten Tests behandelt. Ebenfalls beste Tests sind zweiseitige Gauß-Tests (hier genügen relativ einfache Abschätzungen), aber auch einfache t-Tests sowie χ^2-Tests. Bei den t-Tests ist σ^2 unbekannt, und deshalb ist $\mathcal{P}$ nicht mehr total geordnet. Deshalb benötigt man dort den Begriff des bedingten Tests, um die Optimalität des Tests zu zeigen.

5 Schätztheorie

Im Folgenden sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ mit einem statistischen Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ und $g : \mathcal{P} \to \aleph$ eine Abbildung auf den Entscheidungsraum $\aleph$ (typischerweise mit $\aleph \subseteq \mathbb{R}^k$ für ein k). In diesem Kapitel betrachten wir nur nicht-randomisierte Entscheidungsfunktionen oder Schätzer $d : E \to \aleph$. Wir sagen hier, dass d ein Schätzer für g ist.

5.1 Grundlagen

Es gibt verschiedene Prinzipien, mit denen man zu Schätzern kommt. Wir stellen in diesem Abschnitt das Substitutionsprinzip (und darauf aufbauend die Momentenmethode) und das Maximum-Likelihood-Prinzip vor. Aber erst gibt es ein paar Grundbegriffe.

Definition 5.1 (Bias, Mean Squared Error). *Sei d ein Schätzer für g und $\aleph \subseteq \mathbb{R}$. Dann heißt*

$$b_\theta(d) := \mathbb{E}_\theta[d(X)] - g(\theta)$$

der Bias *oder die* Verzerrung *von d. Im Fall $b_\theta(d) = 0$ für alle $\theta \in \mathcal{P}$ heißt d* unverzerrt *oder* erwartungstreu *oder* unbiased. *Weiter heißt*

$$\mathbb{E}_\theta[(d(X) - g(\theta))^2]$$

die mittlere quadratische Abweichung *oder der* Mean Squared Error.

Lemma 5.2 (Zerlegung des Risikos). *Sei $\aleph \subseteq \mathbb{R}$, d ein Schätzer für g und ℓ der Gauß-Verlust, d.h. $\ell(\theta, a) = |a - g(\theta)|^2$. Dann gilt für die Risikofunktion*

$$R_d(\theta) = \mathbb{V}_\theta[d(X)] + b_\theta(d)^2.$$

Beweis. Wir schreiben

$$\begin{aligned} R_d(\theta) &= \mathbb{E}_\theta[(d(X) - g(\theta))^2] = \mathbb{E}_\theta[(d(X)^2] - \mathbb{E}_\theta[(d(X)]^2 + \mathbb{E}_\theta[(d(X)]^2 - 2g(\theta)\mathbb{E}_\theta[d(X)] + g(\theta)^2 \\ &= \mathbb{V}_\theta[d(X)] + b_\theta(d)^2. \end{aligned}$$

□

Definition 5.3 (Plug-in-Schätzer, Maximum-Likelihood-Schätzer). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem.*

1. *Sei $X = (X_1, ..., X_n)$ ein Vektor unabhängiger und identisch verteilter Zufallsvariablen. Dann heißt die (zufällige) Wahrscheinlichkeitsverteilung*

$$P_X := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

   empirische Verteilung von X.

2. *Sei $X = (X_1, ..., X_n)$ ein Vektor identisch verteilter Zufallsvariablen und $g(\theta) = \mathbb{E}_\theta[f(X_1)]$ für eine Funktion f, falls der Erwartungswert existiert. Dann heißt*

$$\widehat{g}(X) := E_X[f] = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

   Plug-in-Schätzer für g. Plug-In-Schätzer für $g(\theta) = \mathbb{E}_\theta[X_1]$ heißen auch momentenbasierte Schätzer.

3. *Sei* $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ *ein reguläres statistisches Problem und* $\aleph = \mathcal{P}$. *Existiert eine messbare Abbildung* $\widehat{\theta} : E \to \aleph$ *mit*

$$p_{\widehat{\theta}(x)}(x) = \sup_{\theta \in \mathcal{P}} p_\theta(x),$$

so heißt $\widehat{\theta}$ Maximum-Likelihood-Schätzer *von* $g(\theta) = \theta$. *Weiter heißt für festes* x *die Funktion* $\theta \mapsto p_\theta(x)$ Likelihood *und* $\theta \mapsto \log p_\theta(x)$ Log-Likelihood.

Bemerkung 5.4 (Einfache Eigenschaften). 1. Sei $X = (X_1, ..., X_n)$ ein Vektor identisch verteilter Zufallsvariablen. Dann ist jeder Plug-In-Schätzer $\widehat{g}$ für g (mit $g(\theta) := \mathbb{E}_\theta[f(X_1)]$) unverzerrt.
Denn: Es gilt

$$\mathbb{E}_\theta[\widehat{g}(X)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[f(X_i)] = \mathbb{E}_\theta[f(X_1)] = g(\theta).$$

2. Sei $d : E \to \aleph$ ein Schätzer für $g : \mathcal{P} \to \aleph$ und $f : \aleph \to \aleph$. Ist d ein Maximum-Likelihood-Schätzer für g, so ist $d \circ f$ ein Maximum-Likelihood-Schätzer für $g \circ f$. Ist d unverzerrt, so ist jedoch der Schätzer $d \circ f$ für $g \circ f$ nicht unbedingt unverzerrt.
Denn: Der Wert, an dem eine Funktion ihr Maxmimum annimmt, verändert sich nicht, wenn man die Funktion mit einer weiteren Funktion verknüpft, was die erste Behauptung zeigt. Für die zweite Behauptung sei etwa d ein Plug-In-Schätzer für $g(\theta) = \mathbb{E}_\theta[X_1]$, sowie $f : x \mapsto x^2$. Wäre $d \circ f$ unverzerrt, so müsste

$$\mathbb{E}_\theta[d(X)^2] = \mathbb{E}_\theta[f(d(X))] = f(g(\theta)) = (g(\theta))^2 = (\mathbb{E}_\theta[d(X)])^2,$$

also $\mathbb{V}_\theta[d(X)] = 0$, was im Allgemeinen sicher falsch ist.

Proposition 5.5 (Maximum-Likelihood-Schätzer in Exponentialfamilien). *Sei* $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ *eine* k*-parametrige Exponentialfamilie mit* c, t, d, h *für* $c_1, ..., c_k$ *injektiv, also*

$$p_\theta(x) = h(x) \cdot \exp\big(c(\theta)^\top t(x) - d(\theta)\big).$$

Sei $C = \{c(\theta) : \theta \in \mathcal{P}\}^\circ$. *Falls die Gleichung (in* θ*)*

$$\mathbb{E}_\theta[t_i(X)] = t_i(x), \qquad i = 1, ..., k$$

(in θ*) eine Lösung* $x \mapsto \widehat{\theta}(x)$ *besitzt mit* $c(\widehat{\theta}(x)) \in C$, *dann ist* $\widehat{\theta}$ *der eindeutige Maximum-Likelihood-Schätzer für* θ.

Beweis. Wir zeigen die Behauptung nur für $k = 1$. Sei oBdA $c = \mathrm{id}$. Den allgemeinen Fall zeigt man mit Bemerkung 5.4.2. Wir berechnen

$$\frac{\partial}{\partial \theta_i} \log p_\theta(x) = t(x) - d'(\theta), \qquad \frac{\partial^2}{\partial \theta_i^2} \log p_\theta(x) = -d''(\theta).$$

Daraus folgt mit Proposition 2.35, dass $\mathbb{E}_\theta[t(X)] = d'(\theta)$ und $\mathbb{V}_\theta[t(X)] = -d''(\theta) y 0$. Damit ist gezeigt, dass die Log-Likelihood-Funktion wegen der negativen Ableitung strikt konkav ist und $\widehat{\theta}$ genau dann ein Maximum-Likelihood-Schätzer ist, wenn $t(x) = \mathbb{E}_{\widehat{\theta}(x)}[t(X)]$ gilt. (Die Eindeutigkeit folgt mit der Konkavität.) □

Beispiel 5.6 (Beispiel *Norm*). Für das Beispiel aus Beispiel 2.23 erinnern wir an Bemerkung 2.28 und sehen, dass es sich um eine 2-parametrige Exponentialfamilie handelt mit $t_1(x) = \sum_{i=1}^n x_i, t_2(x) = \sum_{i=1}^n x_i^2$. Wir betrachten also etwa die Gleichung

$$\mathbb{E}_{(\mu(x),\sigma^2(x))}[t_1(X)] = t_1(x),$$

die genau dann erfüllt ist, wenn

$$n\mu(x) = t_1(x), \qquad \text{also für } \mu(x) = \overline{x}.$$

Damit ist bereits $\overline{X}$ der maximum-Likelihood-Schätzer für μ. Weiter betrachten wir die Gleichung

$$\mathbb{E}_{(\mu(x),\sigma^2(x))}[t_2(X)] = t_2(x)$$

die genau dann erfüllt ist, wenn

$$n(\sigma^2(x) + \mu^2(x)) = t_2(x), \qquad \text{also für } \sigma^2(x) = \frac{1}{n}t_2(x) - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2.$$

Damit haben wir den maximum-Likelihood-Schätzer für σ^2 hergeleitet.

Beispiel 5.7 (Lineare Regression). Wir kehren zurück zur linearen Regression aus Beispiel 2.30. Wieder leiten wir Maximum-Likelihood-Schätzer für die Parameter $\beta_0, ..., \beta_m, \sigma^2$ her. Für $t(y)^\top = (t_0(y), ..., t_m(y))$ schreiben wir

$$\mathbb{E}_{(\beta(y),\sigma^2(y))}[t(Y)]^\top = xx^\top\beta(y) = t(y)^\top = xy,$$
$$\mathbb{E}_{(\beta(y),\sigma^2(y))}[t_{m+1}(Y)] = (x^\top\beta)^2 + \sigma^2(y) = t_{m+1}(y) = \sum_{i=1}^n y_i^2 = yy^\top.$$

Aus der ersten Gleichung lesen wir ab, dass

$$\hat{\beta}(y) = (xx^\top)^{-1}xy$$

und aus der zweiten, dass

$$\widehat{\sigma^2}(y) = yy^\top - (x^\top(xx^\top)^{-1}xy)^2$$

die Maximum-Likelihood-Schätzer für die Parameter β und σ^2 sind.

5.2 UMVUE-Schätzer

Um es vorwegzunehmen: UMVUE steht für *Uniformly Minimum Variance Unbiased Estimator*. Solche Schätzer minimieren also unter allen unverzerrten Schätzern die Varianz.

Definition 5.8 (UMVUE-Schätzer). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem. Ein Schätzer d für g heißt* UMVUE *oder* Uniformly Minimum Variance Unbiased Estimator *für g, falls er unverzerrt ist und*

$$\mathbb{V}_\theta[d(X)] = \inf_{e \text{ unverzerrt}} \mathbb{V}_\theta[e(X)], \qquad \theta \in \mathcal{P}.$$

Bemerkung 5.9 (Rao-Blackwell). Wir erinnern kurz an den Satz von Rao-Blackwell, Theorem 3.10, für ein Schätzproblem $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ im Falle des Gauß-Verlustes. Sei $T = t(X)$ suffizient und d ein nicht-randomisierter Schätzer für g mit $\mathbb{E}_\theta[|d(X)|] < \infty$ für alle $\theta \in \mathcal{P}$. Dann hat der Schätzer $e \circ t$ für g mit

$$e(t) := \mathbb{E}_\theta[d(X)|T = t]$$

geringeres Risiko als d.

Theorem 5.10 (Lehmann-Scheffé). *Sei $((X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem mit Gauß-Verlust ℓ, $T = t(X)$ eine vollständige, suffiziente Statistik und d ein unverzerrter Schätzer für g. Dann ist $e \circ t$ mit*

$$e(t) := \mathbb{E}_\theta[d(X)|T = t]$$

ein UMVUE für g. Ist außerdem $\mathbb{V}_\theta[e(t((X)))] < \infty$ für alle $\theta \in \mathcal{P}$, so ist $e \circ t$ der einzige UMVUE für g.

Beweis. Da d unverzerrt ist, folgt mit der Turmeigenschaft der bedingten Erwartung, dass auch $e \circ t$ unverzerrt ist. Wir zeigen nun, dass e und damit $e \circ t$ unabhängig von der Wahl von d ist. Dann hat insbesondere $e \circ t$ minimale Varianz unter allen unverzerrten Schätzern, ist also UMVUE. Seien also d_1, d_2 unverzerrte Schätzer für g. Weiter seien $e_i \circ t$ für $e_i(t) = \mathbb{E}_\theta[d_i(X)|T = t]$ zwei unverzerrte Schätzer. Insbesondere gilt

$$\mathbb{E}_\theta[e_1(T) - e_2(T)] = \mathbb{E}_\theta[d_1(X) - d_2(X)] = g(\theta) - g(\theta) = 0$$

für alle $\theta \in \mathcal{P}$. Da T vollständig ist folgt $e_1(T) \overset{\mathbb{P}_\theta-\text{fs}}{=} e_2(T)$ für alle $\theta \in \mathcal{P}$, was zu zeigen war. Für die Eindeutigkeit sei d' ein weiterer UMVUE. Einerseits ist dann nach dem Satz von Rao-Blackwell

$$\mathbb{E}_\theta[(\mathbb{E}_\theta[d'(X)|T] - g(\theta))^2] \leq \mathbb{E}_\theta[(d'(X) - g(\theta))^2],$$

und da d' ein UMVUE ist, gilt hier Gleichheit, also $d'(X) = \mathbb{E}_\theta[d'(X)|T]$. Andererseits haben wir oben gezeigt, dass $e(T) = \mathbb{E}_\theta[d'(X)|T]$, da d' unverzerrt ist. Insgesamt gilt also $d' = e \circ t$. □

Beispiel 5.11 (Beispiel *Unif*). Wir betrachten das Schätzproblem mit $g(\theta) = \theta$ und Gauß-Verlust ℓ. Hierzu verwendet wir die Statistik $T = t(X) = \max_{1 \leq i \leq n} X_i$. Wir haben bereits in Beispiel 2.8 und 2.17 gesehen, dass T vollständig und suffizient ist. Aus Beispiel 3.11 wissen wir außerdem, dass

$$d(X) := \frac{n+1}{n} \max_{1 \leq i \leq n} X_i$$

unverzerrt ist und dieser Schätzer eine kleinere Risikofunktion hat als $2\overline{x}$. Außerdem gilt $\mathbb{E}_\theta[d(X)|T] = d(X)$, da $d(X)$ eine Funktion von T ist. Nach dem Satz von Lehmann und Scheffé ist also d ein UMVUE. Wegen $\mathbb{V}_\theta[d(X)] < \infty$ für alle $\theta \in \mathcal{P}$ ist dieser UMVUE auch eineutig.

Die Argumentation des letzten Beispiels lässt sich verallgemeinern, was wir nun für Exponentialfamilien tun wollen.

Korollar 5.12 (UMVUE bei Exponentialfamilien). *Sei $(X, \{\mathbb{P}_\theta = p_\theta \cdot \lambda^n : \theta \in \mathcal{P}\})$ eine k-parametrige Exponentialfamilie mit c, t, d, h, also*

$$p_\theta(x) = h(x) \cdot \exp\big(c(\theta)^\top t(x) - d(\theta)\big).$$

Weiter sei $g(\theta) = \mathbb{E}_\theta[f(t(X))]$. Dann ist $d(X) := f(t(X))$ ein UMVUE für g. Ist außerdem $\mathbb{V}_\theta[d(X)] < \infty$ für alle $\theta \in \mathcal{P}$, so gibt es nur diesen einen UMVUE.

Beweis. Zunächst wissen wir aus Proposition 2.29 und Theorem 2.32, dass $T = t(X) = (t_1(X), ..., t_k(X))$ suffizient und vollständig ist. Weiter ist d unverzerrt und $\mathbb{E}_\theta[d(X)|T] = d(X)$, da $d(X)$ messbar bezüglich T ist. Nach dem Satz von Lehmann und Scheffé ist damit d ein UMVUE für g. Die letzte Behauptung wurde schon in Theorem 5.10 bewiesen. □

Beispiel 5.13 (Beispiel *Norm*). Für das Beispiel *Norm* aus Beispiel 2.23 haben wir in Beispiel 5.6 gesehen, dass es sich um eine 2-parametrige Exponentialfamilie handelt mit $t_1(x) = \sum_{i=1}^n x_i, t_2(x) = \sum_{i=1}^n x_i^2$. Wir betrachten nun $g_1(\mu, \sigma^2) = \mu$ und $g_2(\mu, \sigma^2) = \sigma^2$ und schreiben

$$\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2 = \frac{1}{n-1}\Big(\Big(\sum_{i=1}^n x_i^2\Big) - 2n\overline{x}^2 + n\overline{x}^2\Big) = \frac{1}{n-1}t_2(x) - \frac{1}{(n-1)n}t_1(x)^2,$$

also

$$\mathbb{E}_\theta\Big[\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2\Big] = \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 = \sigma^2$$

und damit sind

$$\begin{aligned} g_1(\theta) &= \mathbb{E}_\theta[\tfrac{1}{n}t_1(X)] = \mathbb{E}_\theta[\overline{X}] = \mu, \\ g_2(\theta) &= \mathbb{E}_\theta\Big[\frac{1}{n-1}t_2(X) - \frac{1}{(n-1)n}t_1^2(X)\Big] = \sigma^2 \end{aligned}$$

eindeutige UMVUEs.

Beispiel 5.14 (Lineare Regression). Wir betrachten noch einmal die lineare Regression aus Beispiel 2.30. Für $t(y)^\top = (t_0(y), ..., t_m(y))$ schreiben wir für $f(t(y)) = (xx^\top)^{-1}t(y)$

$$\mathbb{E}_{(\beta,\sigma^2)}[f(t(Y))] = \mathbb{E}_{(\beta,\sigma^2)}[(xx^\top)^{-1}t(Y)]\mathbb{E}_{(\beta,\sigma^2)}[(xx^\top)^{-1}xY] = (xx^\top)^{-1}xx^\top\beta = \beta.$$

Damit ist $(xx^\top)^{-1}xy$ ein UMVUE für β (und ein Maximum-Likelihood-Schätzer nach Beispiel 5.7).

5.3 Information und die Cramér-Rao-Schranke

Zwar haben wir bereits obere Schranken für das Risiko eines Schätzers ermittelt, jedoch gibt es auch untere Schranken, insbesondere die in Theorem 5.22 vorgestellte Cramér-Rao-Schranke. Für diese benötigen wir den Begriff der (Fisher-)Information. Diese beschreibt die Krümmung (im Sinne der zweiten Ableitung) der log-Likelihood in Bezug auf die Parameter. Ist diese Zahl groß, bedeutet das eine hohe Krümmung, und damit ist die log-Likelihood für nahe Parameter deutlich kleiner. Anders ausgedrückt haben wir über die log-Likelihood viel über den wahren Parameter gelernt, es gibt also viel "Information". Um die zweite Ableitung sinnvoll berechnen zu können, benötigen wir zunächst ein paar Regularitätsannahmen.

Annahme 5.15 (Regularitätsannahmen der Fisher-Information). *Für ein reguläres statistisches Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ auf $\mathbb{R}^n$ mit $\mathcal{P} \subseteq \mathbb{R}^m$ und $\mathbb{P}_\theta = p_\theta \cdot \lambda^n$ geben wir folgende Regularitätsannahmen an:*

(A1) Es gibt ein $N \in \mathcal{B}(\mathbb{R}^n)$, so dass für alle $\theta \in \mathcal{P}$ und $i = 1, ..., k$ die Ableitung $\partial p_\theta(x)/\partial\theta_i$ für $x \notin N$ existiert.

(A2) Für jedes messbare ϕ gilt, falls der Erwartungswert existiert,

$$\frac{\partial}{\partial\theta_i}\mathbb{E}_\theta[\phi(X)] = \int \phi(x)\frac{\partial p_\theta(x)}{\partial\theta_i}d\lambda^n(x).$$

(A3) Die Menge

$$C := \{x : p_\theta(x) > 0\}$$

ist unabhängig von θ.

(A4) Es gilt (A2) und außerdem auch

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathbb{E}_\theta[\phi(X)] = \int \phi(x)\frac{\partial^2 p_\theta(x)}{\partial\theta_i\partial\theta_j}d\lambda^n(x)$$

für jedes messbare ϕ, für das der Erwartungswert existiert.

Beispiel 5.16 (Beispiele *Bern*, *Norm*, *Unif*). Für Exponentialfamilien ist Annahme 5.15 nach Lemma 2.34 immer erfüllt, insbesondere also für Beispiele *Bern* und *Norm*. Für Beispiel 1.8 ist jedoch sowohl (A1) als auch (A3) verletzt.

Definition 5.17 (Fisher-Information). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell, das die Annahmen (A1)–(A3) erfüllt. Dann heißt*

$$I(\theta) := I(\mathbb{P}_\theta) := \Big(\mathbb{COV}\Big[\frac{\partial}{\partial\theta_i}\log p_\theta(X), \frac{\partial}{\partial\theta_j}\log p_\theta(X)\Big]\Big)_{i,j=1,...,k}$$

Fisher-Informations-Matrix.

Bemerkung 5.18 (Berechnung der Fisher-Information unter (A4)). Gilt statt (A2) sogar (A4), so kann man schreiben

$$\begin{aligned}\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\Big] &= \mathbb{E}_\theta\Big[\frac{1}{p_\theta(X)^2}\Big(p_\theta(X)\frac{\partial^2 p_\theta(X)}{\partial\theta_i\partial\theta_j} - \frac{\partial p_\theta(X)}{\partial\theta_i}\cdot\frac{\partial p_\theta(X)}{\partial\theta_j}\Big)\Big] \\ &= 0 - \mathbb{E}_\theta\Big[\frac{\partial\log p_\theta(X)}{\partial\theta_i}\frac{\partial\log p_\theta(X)}{\partial\theta_i}\Big] = -(I(\theta))_{i,j}.\end{aligned}$$

Dies liefert also eine alternative Berechnung der Fisher-Information.

Beispiel 5.19 (Exponentialfamilie). Im Falle einer Exponentialfamilie in kanonischer Form ist

$$p_\theta(x) = h(x)\exp(\theta^\top t(x) - d(\theta))$$

und damit mit Proposition 2.35

$$\frac{\partial}{\partial\theta_i}\log p_\theta(x) = t_i(x) - \frac{\partial d(\theta)}{\partial\theta_i} = t_i(X) - \mathbb{E}_\theta[t_i(X)].$$

Weiter ist

$$(I(\theta))_{i,j} = \mathbb{COV}_\theta[t_i(X), t_j(X)].$$

Beispiel 5.20 (Fisher-Information einer unabhängigen Stichprobe). Für ein statistisches Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ haben wir die Fisher-Information definiert. Nun betrachten wir den Fall von Daten $X_1, ..., X_n$, bei denen die X_i unter allen $\mathbb{P}_\theta$ unabhängig sind. Anders gesagt betrachten wir das statistische Modell $(X = (X_1, ..., X_n), \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\})$, wobei $\mathbb{P}_\theta^n$ das n-fache Produktmaß ist. Hier ist die Dichte gegeben durch $x = (x_1, ..., x_n) \mapsto p_\theta(x_1) \cdots p_\theta(x_n)$ und damit gilt

$$\begin{aligned} I(\mathbb{P}_\theta^n) &= \mathbb{COV}_\theta^n\Big[\frac{\partial}{\partial\theta_j}\log p_\theta(X), \frac{\partial}{\partial\theta_k}\log p_\theta(X)\Big]_{jk} \\ &= \mathbb{COV}_\theta^n\Big[\sum_{i=1}^n \frac{\partial}{\partial\theta_j}\log p_\theta(X_i), \sum_{i=1}^n \frac{\partial}{\partial\theta_k}\log p_\theta(X_i)\Big]_{jk} \\ &= \sum_{i=1}^n \mathbb{COV}_\theta\Big[\frac{\partial}{\partial\theta_j}\log p_\theta(X_i), \frac{\partial}{\partial\theta_k}\log p_\theta(X_i)\Big]_{jk} \\ &= nI(\mathbb{P}_\theta). \end{aligned}$$

Proposition 5.21 (Mittlere Ableitung der log-Likelihood verschwindet). *Es gelte (A1)–(A3). Dann ist für $i = 1, ..., m$*

$$\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_i}\log p_\theta(X)\Big] = 0.$$

Beweis. Wir schreiben

$$\begin{aligned} \mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_i}\log p_\theta(X)\Big] &= \int p_\theta(x)\frac{\partial}{\partial\theta_i}\log p_\theta(x)d\lambda^n(dx) = \int p_\theta(x)\frac{1}{p_\theta(x)}\frac{\partial}{\partial\theta_i}p_\theta(x)d\lambda^n(dx) \\ &= \frac{\partial}{\partial\theta_i}\mathbb{E}_\theta[1] = 0, \end{aligned}$$

wobei die Vertauschung von Integral und Ableitung nach (A3) erlaubt ist. □

Theorem 5.22 (Cramér-Rao-Schranke). *Es gelte (A1)–(A3) sowie $\mathcal{P} \subseteq \mathbb{R}$ und $I(\theta) > 0$ für alle $\theta \in \mathcal{P}$. Sei $T = t(X)$ für $t : \mathbb{R}^m \to \mathbb{R}$ eine Statistik und es existiere $\Psi(\theta) := \mathbb{E}_\theta[T]$. Dann gilt*

$$\mathbb{V}_\theta[T] \geq \frac{\Psi'(\theta)^2}{I(\theta)}.$$

Ist insbesondere $\mathbb{E}_\theta[T] = \theta$ (d.h. T ist ein erwartungstreuer Schätzer für θ), so gilt

$$\mathbb{V}_\theta[T] \geq \frac{1}{I(\theta)}.$$

Beweis. Wir schreiben mit Proposition 5.21 und der Cauchy-Schwartz-Ungleichung

$$\begin{aligned} \Psi'(\theta) &= \int t(x)\frac{\partial}{\partial\theta}p_\theta(x)\lambda^n(dx) = \mathbb{E}_\theta\Big[t(x)\frac{\partial}{\partial\theta}\log p_\theta(x)\Big] \\ &= \mathbb{E}_\theta\Big[\big(t(X) - \mathbb{E}_\theta[T]\big)\frac{\partial}{\partial\theta}\log p_\theta(X)\Big] \\ &\leq \big(\mathbb{V}_\theta[T]\cdot I(\theta)\big)^{1/2}. \end{aligned}$$

Daraus folgt die Behauptung. □

Beispiel 5.23 (1-parametrige Exponentialfamilie). Wir betrachten den Fall aus Beispiel 5.19 mit $\theta \in \mathbb{R}$ und der suffizienten Statistik $T = t(X)$. Wir wissen (etwa aus Beispiel 5.19 und Proposition 2.35), dass $\mathbb{E}_\theta[t(X)] = d'(\theta)$ und $I(\theta) = \mathbb{V}_\theta[t(X)] = d''(\theta)$. Damit gilt

$$\mathbb{V}_\theta[t(X)] = d''(\theta) = \frac{d''(\theta)^2}{I(\theta)} = \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[t(X)])^2}{I(\theta)},$$

und damit gilt Gleichheit in der Schranke von Theorem 5.22.

Andresherum ist im Beweis der Cramér-Rao-Schranke klar: Gleichheit gilt genau dann, wenn Gleichheit in der Cauchy-Schwartz-Ungleichung gilt, also wenn $t(X) - \mathbb{E}_\theta[T]$ und $\frac{\partial}{\partial\theta}\log p_\theta(X)$ linear abhängig sind, also wenn für geeignete a und b

$$\frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)t(x) + b(\theta)$$

oder

$$\log p_\theta(x) = t(x)\int_{\theta_0}^{\theta} a(\eta)d\eta + \int_{\theta_0}^{\theta} b(\eta)d\eta.$$

Das bedeutet, dass p_θ die Dichte einer Exponentialfamilie ist. Die Cramér-Rao-Schranke ist damit (falls (A1)–(A3) gelten) genau für 1-parametrige Exponentialfamilien scharf.

Beispiel 5.24 (Beispiel *Unif*). Aus Beispiel 2.25 wissen wir, dass es sich nich um eine Exponentialfamilie handelt. Außerdem sind die Regularitätsannahmen (A1) und (A3) nicht erfüllt; siehe Beispiel 5.16. Deshalb ist nicht klar, ob die Cramér-Rao-Schranke in diesem Modell existiert. Um zu sehen, ob die Schranke doch funktioniert, verwenden wir Bemerkung 5.18, so dass

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2}\mathbb{E}_\theta[\log p_\theta(X)] = 1/\theta^2. \tag{5.1}$$

Wir setzen $t(X) = X$, so dass $\mathbb{V}_\theta[t(X)] = \theta^2/12$. Außerdem ist $\frac{d}{d\theta}\mathbb{E}_\theta[X] = 1/2$ und damit

$$\mathbb{V}_\theta[t(X)] = \frac{\theta^2}{12} < \frac{\theta^2}{4} = \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[X])^2}{I(\theta)}.$$

Also gilt die Cramér-Rao-Schranke hier nicht.

Bemerkung 5.25 (Unabhängige Stichprobe). Für das statistische Modell $(X^n = (X_1, ..., X_n), \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\})$ der n-fachen unabhängigen Versuchswiederholung, $n = 1, 2, ...$ aus Beispiel 5.20 erhalten wir folgendes Resultat (falls (A1)–(A3) gelten):
Ist $\mathcal{P} \subseteq \mathbb{R}$ und ist d^n ein Schätzer für θ im n-ten Modell, so gilt

$$\mathbb{V}_\theta[d^n(X^n)] \geq \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[d^n(X^n)])^2}{nI(\theta)}.$$

Asymptotisch bedeutet das, falls $\sqrt{n}(\mathbb{E}_\theta[d^n(X^n) - g(\theta)] \xrightarrow{n\to\infty} 0$ und $\frac{d}{d\theta}(\mathbb{E}_\theta[d^n(X^n)] - g(\theta)) \xrightarrow{n\to\infty} 0$

$$\begin{aligned}\liminf_{n\to\infty} n\mathbb{E}_\theta[(d^n(X^n) - g(\theta))^2] &= \liminf_{n\to\infty} n\mathbb{V}_\theta[(d^n(X^n)] + n(\mathbb{E}_\theta[d^n(X^n) - g(\theta)])^2 \\ &= \liminf_{n\to\infty} n\mathbb{V}_\theta[(d^n(X^n)] \geq \liminf_{n\to\infty} \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[d^n(X^n)])^2}{I(\theta)} = \frac{(g'(\theta))^2}{I(\theta)}.\end{aligned}$$

5.4 Asymptotik von Maximum-Likelihood-Schätzern

Das letzte Resultat behandelt schon die Asymptotik einer Folge von Schätzern. Dies wollen wir nun noch für Maximum-Likelihood-Schätzer ausbauen. Einfach gesagt konvergieren Maximum-Likelihood-Schätzer (unter einigen Regularitätsannahmen) für unabhängige Stichproben gegen den wahren Parameter (Theorem 5.30). Noch genauer sind sie asymptotisch normalverteilt und die Inverse der Fisher-Information gibt die Covarianzmatrix (Theorem 5.31). Es folgen nun weitere Annahmen, insbesondere die der unabhängigen, identisch verteilten Daten.

Bemerkung 5.26 (Weitere Annahmen). (A5) Es ist $(X^n = (X_1^n, ..., X_n^n), \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\})$ ein reguläres statistisches Modell, so dass X^n ein unabhängiger, identisch verteilter Vektor ist mit

$$\mathbb{P}^n(X_1^n \in .) = p_\theta \cdot \lambda^m.$$

Insbesondere hat $\mathbb{P}^n$ die Dichte

$$\prod_{i=1}^n p_\theta(x_i).$$

(A6) Für alle x ist

$$\eta \mapsto \frac{\partial^2}{\partial\eta_i \partial\eta_j} \log p_\eta(x)$$

stetig und endlich und es gilt für alle $\theta \in \mathcal{P}$

$$\eta \mapsto \frac{\partial^2}{\partial\eta_i \partial\eta_j} \mathbb{E}_\theta\Big[\log p_\eta(X)\Big] = \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_i \partial\eta_j} \log p_\eta(X)\Big].$$

Bemerkung 5.27 (Maximum-Likelihood-Schätzer). Gilt (A1)–(A5), so ist der Maximum-Likelihood-Schätzer $d^n(x^n)$ für g, falls er im Inneren von $\mathcal{P}$ liegt, Lösung der Gleichung

$$\sum_{i=1}^n \frac{\partial}{\partial\theta} \log p_\theta(x_i^n)\Big|_{\theta = d^n(x^n)} = 0. \tag{Lik}$$

Definition 5.28 (Konsistenz). *Sei $((X^n, \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\}), \aleph, g, \ell)$ ein Schätzproblem, und d^n ein Schätzer für g, $n = 1, 2, ...$ Dann heißt die Folge d^n konsistent, falls*

$$\mathbb{P}_\theta(|d^n(X^n) - g(\theta)| \geq \varepsilon) \xrightarrow{n\to\infty} 0$$

für alle $\varepsilon > 0$ und alle $\theta \in \mathcal{P}$. Dies ist insbesondere dann der Fall, wenn $X^1, X^2, ...$ auf demselben Wahrscheinlichkeitsraum definiert sind und $d^n(X^n) \xrightarrow{n\to\infty}_{fs} g(\theta)$.

Bemerkung 5.29 (Konsistenz bei unabhängigen Daten). Unter Annahme (A5) ist unter $\mathbb{P}_\theta^n$ der Vektor X^n unabhängig und identisch nach $\mathbb{P}_\theta$ verteilt. Wir können dann oBdA annehmen, dass alle X^n auf demselben Wahrscheinlichkeitsraum definiert sind. Wir schreiben dann auch X anstelle von X^n (und denken uns, dass d^n ja nur von den ersten n Einträgen von X abhängt). Insbesondere kann dann die fast sichere Konvergenz $d^n(X) \xrightarrow{n\to\infty} g(\theta)$ unter $\mathbb{P}_\theta$ gelten.

Theorem 5.30 (Konsistenz von Maximum-Likelihood-Schätzern). *Sei $\theta \in \mathcal{P}$ und $d^n(X)$ der Maximum-Likelihood-Schätzer für θ (basierend auf den ersten n Beobachtungen). Sei*

$$Z(M,x) := \inf_{\eta\in M} \log\frac{p_\theta(x)}{p_\eta(x)}.$$

Angenommen, für jedes $\eta \neq \theta$ gibt es ein $\varepsilon(\eta) > 0$, so dass $\mathbb{E}_\theta[Z(B_{\varepsilon(\eta)}, X)] > 0$. Weiter existiere ein $\mathcal{C}$ kompakt mit $\theta \in \mathcal{C}$ und $\mathbb{E}_\theta[Z(\mathcal{P}\setminus\mathcal{C}, X)] > 0$. Dann ist

$$\lim_{n\to\infty} d^n(X) \overset{\mathbb{P}_\theta-fs}{=} \theta.$$

Beweis. Für beliebiges $\varepsilon > 0$ müssen wir zeigen, dass

$$\mathbb{P}_\theta(\limsup_{n\to\infty} |d^n(X) - \theta| > \varepsilon) = 0.$$

(Die Aussage folgt dann mit einem Grenzwert $\varepsilon \downarrow 0$.) Da $\mathcal{C}\setminus B_\varepsilon(\theta)$ kompakt ist und $\{B_{\varepsilon(\eta)}(\eta) : \eta \in \mathcal{C}\setminus B_\varepsilon(\theta)\}$ eine offene Überdeckung der kompakten Menge $\mathcal{C}\setminus B_\varepsilon(\theta)$ ist, gibt es eine endliche Teilüberdeckung $A_1 := B_{\varepsilon(\eta_1)}(\eta_1), ..., A_m := B_{\varepsilon(\eta_m)}(\eta_m)$. Mit $A_0 := \mathcal{P}\setminus\mathcal{C}$ ist $\mathcal{P} = B_\varepsilon(\theta)\cup A_0\cup\cdots\cup A_m$ und $\mathbb{E}_\theta[Z(A_j, X_i)] =: c_j > 0$, $j = 0, ..., m$. Nach dem starken Gesetz der großen Zahlen ist

$$\frac{1}{n}\sum_{i=1}^n Z(A_j, X_i) \xrightarrow{n\to\infty,\mathbb{P}_\theta-\text{fs}} c_j.$$

Dann gilt

$$\begin{aligned}
&\mathbb{P}_\theta(\limsup_{n\to\infty}|d^n(X)-\theta| \geq \varepsilon)\\
&\qquad\leq \mathbb{P}_\theta\Big(\text{ Es gibt } \eta_1,\eta_2,... \notin B_\varepsilon(\theta), \text{ so dass}\\
&\qquad\qquad\qquad \frac{1}{n}\sum_{i=1}^n \log p_{\eta_i}(X_i) \geq \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \text{ für unendlich viele } n\Big)\\
&\qquad\leq \sum_{j=0}^m \mathbb{P}\Big(\inf_{\eta\in A_j}\frac{1}{n}\sum_{i=1}^n \log\frac{p_\theta(X_i)}{p_\eta(X_i)} \leq 0 \text{ unendlich oft}\Big)\\
&\qquad\leq \sum_{j=0}^m \mathbb{P}_\theta\Big(\frac{1}{n}\sum_{i=1}^n Z(A_j,X_i) \leq 0 \text{ unendlich oft}\Big) = 0
\end{aligned}$$

und die Behauptung ist gezeigt. □

Theorem 5.31 (Asymptotische Normalität des Maximum-Likelihood-Schätzers). *Es gelte (A1)–(A6) und es sei für jedes $\theta \in \mathcal{P}^\circ$*[10]

$$\sup_{j,k}\mathbb{E}_\theta\Big[\sup_{|\eta-\theta|<r}\Big|\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X_1)\Big|_{\eta=\theta} - \frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X_1)\Big|\Big] \xrightarrow{r\to 0} 0. \qquad (*)$$

Weiter sei $d^n(X)$ konsistent und die Fisher-Information $I(\theta)$ sei für alle $\theta \in \mathcal{P}$ nicht-singulär. Dann gilt für $X \sim \mathbb{P}_\theta^\infty$

$$\sqrt{n}(d^n(X) - \theta) \xRightarrow{n\to\infty} \mathcal{N}(0, I^{-1}(\theta)).$$

[10] Für eine Menge A sei A° das Innere.

Beweis. Sei

$$\ell_x^n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i).$$

Dann ist, wegen der stetigen Differenzierbarkeit,

$$\nabla \ell_x^n(\theta) := \Big(\frac{\partial}{\partial \theta_j}\ell_x^n(\theta)\Big)_j = \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_j}\log p_\theta(x_i)\Big)_j.$$

Es gilt

$$\mathbb{E}_\theta[\nabla \ell_X^n(\theta)] = 0$$

wegen Proposition 5.21 sowie

$$n\mathbb{COV}_\theta[\nabla\ell_X^n(\theta), \nabla\ell_X^n(\theta)] = I(\theta) = \Big(-\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_\theta(X)\Big]\Big)_{jk}.$$

Damit folgt aus dem mehrdimensionalen Zentralen Grenzwertsatz bereits, dass für $X \sim \mathbb{P}_\theta^\infty$

$$\sqrt{n}\nabla\ell_X^n(\theta) \xRightarrow{n\to\infty} \mathcal{N}(0, I(\theta)). \tag{$**$}$$

Außerdem ist

$$\nabla\ell_x^n(d^n(x)) = 0,$$

da die Funktion $\theta \mapsto \ell_x^n(\theta)$ bei $d^n(x)$ ein Maximum besitzt und mit einer Taylor-Entwicklung von $\nabla\ell_x^n(\eta)$ für $\eta = d^n(x)$ um θ gilt

$$0 = \nabla\ell_x^n(d^n(x)) = \nabla\ell_x^n(\theta) + B_n(d^n(x) - \theta)$$

für

$$B_n = \Big(\frac{\partial^2}{\partial\eta_j\partial\eta_k}\ell_X^n(\eta)\Big|_{\eta=\eta_{nj}^*}\Big)_{jk}$$

für ein η_{nj}^* zwischen θ_j und $d^n(x)_j$ für $j = 1, ..., m$. Da $d^n(X) \xrightarrow{n\to\infty}_p \theta$, gilt auch $\eta_{nj}^* \xrightarrow{n\to\infty}_p \theta_j$. Zusammen mit $(**)$ haben wir damit gezeigt, dass

$$\begin{aligned}
&\sqrt{n}B_n(d^n(X) - \theta)\\
&\quad= \sqrt{n}\Big(\Big(\frac{\partial^2}{\partial\eta_j\partial\eta_k}\ell_X^n(\eta)\Big|_{\eta=\eta_{nj}^*} - \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X)\Big|_{\eta=\eta_{nj}^*}\Big]\Big)_{jk}\\
&\qquad+ \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X)\Big|_{\eta=\eta_{nj}^*} - \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_\theta(X)\Big]\Big)_{jk}\\
&\qquad\qquad- I(\theta)\Big)(d^n(X) - \theta) \xRightarrow{n\to\infty} \mathcal{N}(0, I(\theta)).
\end{aligned}$$

Nun gilt für $X \sim \mathbb{P}_\theta^\infty$ wegen dem Gesetz der großen Zahlen und Annahme $(*)$

$$\begin{aligned}
\frac{\partial^2}{\partial\eta_j\partial\eta_k}\ell_X^n(\eta)\Big|_{\eta=\eta_{nj}^*} - \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X)\Big|_{\eta=\eta_{nj}^*}\Big] &\xrightarrow{n\to\infty}_p 0,\\
\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log p_\eta(X)\Big|_{\eta=\eta_{nj}^*} - \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_\theta(X)\Big] &\xrightarrow{n\to\infty}_p 0.
\end{aligned}$$

Damit folgt zunächst $B_n \xrightarrow{n\to\infty}_p -I(\theta)$, woraus $\sqrt{n}(d^n(X-\theta)) = O(1)$ folgt und damit (aus Slutskys Theorem)

$$-\sqrt{n}I(\theta)(d^n(X)-\theta) \xRightarrow{n\to\infty} \mathcal{N}(0, I(\theta)).$$

Da die Matrix-Multiplikation mit $I(\theta)^{-1}$ eine stetige Abbildung ist, folgt damit das Ergebnis. □

Anwendungen der Statistik

von Peter Pfaffelhuber

Version: 9. Oktober 2022

Inhaltsverzeichnis

1 Regression

1.1 Einleitung

Oftmals will man mit Daten Zusammenhänge bestimmen, etwa zwischen der Größe einer Wohnung und dem Mietpreis, oder der Verkehrsdichte und der Durschschnittsgeschwindigkeit von Fahrzeugen. Weiter kann die *Zielvariable* oder *Beobachtung* (hier etwa Mietpreis und Durschschnittsgeschwindigkeit) von weiteren Einflüssen (*Covariate* oder *unabhängige Variable*) abhängen, etwa von der Lage der Wohnung, oder der Breite der Straße. In den meisten Situationen kann man ein Regressionsmodell verwenden, um Korrelationen zwischen Covariaten und Zielvariablen herauszufinden. Dieses ist für n Messungen (also etwa n Wohnungen oder n bestimmte Durchschnitsgeschwindigkeiten) und k Einflussgrößen von der Form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, ..., n \tag{1.1}$$

Hier sind $y_1, ..., y_n$ die Beobachtungen und $x_{i1}, ..., x_{ik}$ sind die Werte der Einflussgrößen auf die i-te Beobachtung. Um dies in ein statistisches Modell umzuwandeln, seien $\epsilon_1, ..., \epsilon_n$ (und damit auch $y_1, ..., y_n$) Zufallsvariable, und mit $x_{i0} := 1$ schreiben wir besser mit Vektoren[1]

$$Y_i = x_{i\cdot}\beta + \epsilon_i, \qquad i = 1, ..., n$$

oder mit $x = (x_{ij})_{i=1,...,n,j=0,...,k}$

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

□

Bemerkung 1.1 (Einfache Regression). Der einfachste Fall tritt ein, wenn es nur eine einzige Covariate gibt; siehe auch Beipsiel 1.2. In diesem Fall verändert sich das Regressionsmodell zu

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, ..., n$$

Im Gegensatz dazu nennt man (1.1) für $k > 1$ *multiple Regression*.

Beispiel 1.2 (Regressionsanalyse mit `R`). Wir verwenden einen Datensatz `faithful` aus den 1980er Jahren, der in [R] verfügbar ist und dessen ersten Zeilen wir mittels

```
> head(faithful)
```

ansehen[2], was

```
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

[1] Für uns ist im Folgenden x ein Spaltenvektor und $x^\top$ ein Zeilenvektor.

[2] Das Kommando `head` liefert nur die ersten Zeilen des Datensatzes. Will man den Datensatz ganz ansehen, gibt man `faithful` ein.

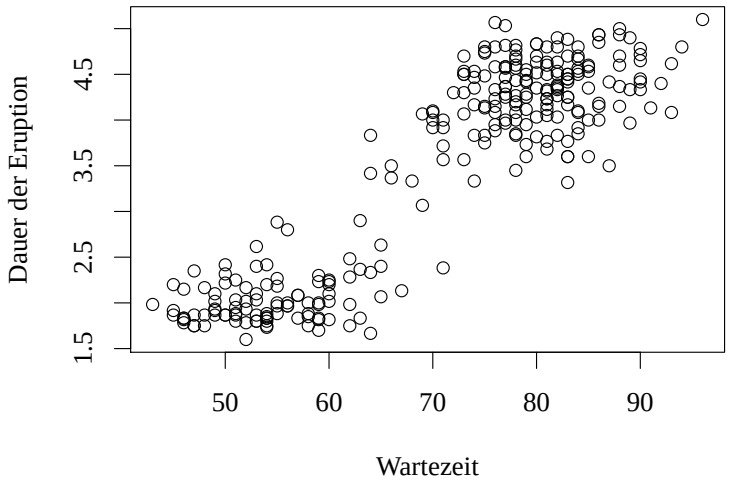


Abbildung 1.1: Das Datenbeispiel aus dem `faithful`-Datensatz, der in R zur Verfügung steht.

liefert. Die Größe `waiting` steht für die Wartezeit bis zur nächsten Eruption des *Old Faithful Gaysier* im Yellowstone National Park der USA und `eruptions` für dessen Dauer. Um uns einen ersten Eindruck zu verschaffen, ob diese beiden Größen korreliert sind, plotten wir einfach mal die Datenpunkte. Mit

```
> duration = faithful$eruptions
> waiting = faithful$waiting
```

weisen wir die beiden Spalten des Datensatzes den Vektoren `duration` und `waiting` zu. Den gewünschten Plot erzeugen wir durch

```
> plot(waiting, duration, xlab="Wartezeit", ylab="Dauer der Eruption")
```

Das Ergebnis ist in Abbildung 1.1 abgebildet. [3]

Offenbar besteht ein Zusammenhang zwischen der Wartezeit und der Dauer der Eruption. Wir werden in den folgenden Kapiteln herleiten, wie man sinnvollerweise eine *Regressionsgerade* durch die Datenwolke legt, die gut passt. Der entsprechende R-Befehl wird

```
> lm(eruptions ~ waiting, data=faithful)
```

lauten. Dies liefert den Output

[3] Um das Bild in ein Skript wie dieses hier einzubetten, ist es natürlich praktisch, wenn es als `pdf` vorliegt. In R habe ich deswegen die Befehle

```
> pdf(file = "fig1.pdf", width=7, height=5, family="Times", onefile=FALSE)
> par(mar=c(5,4,1,1), cex=1.5)
```

vor den `plot`-Befehl gestellt. (Der `par`-Befehl verkleinert die Ränder des Bildes für eine bessere Optik.) Nicht vergessen darf man allerdings, nach dem `plot`-Befehl auch noch

```
> dev.off()
```

einzugeben, erst dann kann die `pdf`-Datei fehlerfrei dargestellt werden.

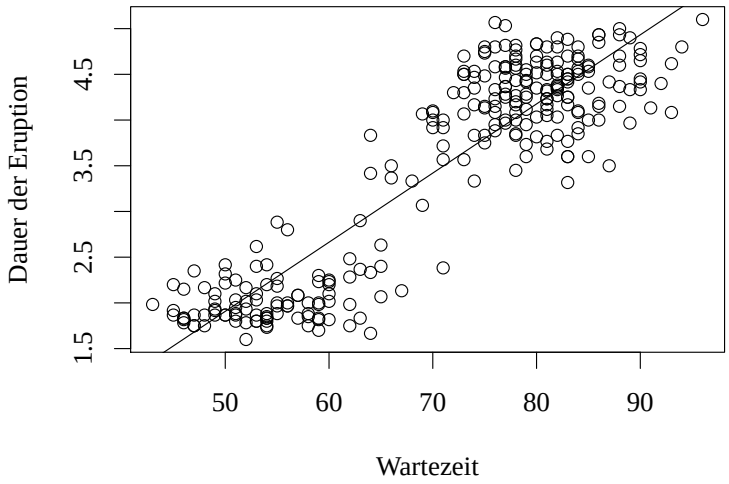


Abbildung 1.2: Die von R berechnete Regressionsgerade im `faithful`-Datensatz.

```
Coefficients:
(Intercept)      waiting
   -1.87402      0.07563
```

Das bedeutet, dass R die Gerade

$$\hat{Y} = -1.87402 + 0.07563x$$

für die Wartezeit Y und die Dauer der Eruption x gefunden hat. Dies können wir auch grafisch in Abbildung 1.2 veranschaulichen.[4] Im Folgenden wollen wir diese Regressiongerade und ihre Eigenschaften diskutieren. Wir gehen dabei gleich zum Fall der multiplen Regression, in dem `waiting` auch mehr als eine Variable beinhalten hätte können. In Kapitel 1.7 kommen wir noch einmal auf das Beispiel zurück.

1.2 Das Modell

Das statistische Modell besteht aus den Daten Y und deren Verteilungen. Letztere hängen nur von den Werten β und den Verteilungen von ϵ ab. Wir bezeichnen die Verteilungen deswegen auch mit $\mathbb{P}_\beta$ (und spezifizieren damit die Abhängigkeit von der Verteilung von ϵ nicht genauer). Oftmals werden wir Annahmen über die Verteilung von ϵ treffen.

Annahme 1.3 (Gauß-Markov-Bedingungen). *Es gilt für ein* $\sigma^2 > 0$

$$\mathbb{E}_\beta[\epsilon_i] = 0, \qquad \mathbb{COV}_\beta[\epsilon_i, \epsilon_j] = \sigma^2 \delta_{ij}.$$

[4] Praktisch ist hier der Befehl `abline`. Um die Gerade zu plotten, habe ich die Befehle

```
> coeffs=coefficients(lm(eruptions ~ waiting, data=faithful))
> coeffs=as.vector(coeffs)
> abline(coeffs)
```

benutzt. Der erste Befehl gibt die beiden Koeffizienten in einer Liste aus, der zweite wandelt diese in einen Vektor um und der dritte zeichnet die Regressionsgerade.

Hierfür schreiben wir auch

$$\mathbb{E}_\beta[\epsilon] = 0, \qquad \mathbb{COV}_\beta[\epsilon, \epsilon] = \mathbb{E}_\beta[\epsilon\epsilon^\top] = \sigma^2 I$$

für die $k \times k$-Einheitsmatrix I, wobei alle Gleichungen in Vektorschreibweise gelesen werden.

Stärker ist die Annahme, dass die Daten sogar unabhängig normalverteilt sind und gleiche Varianz haben.

Annahme 1.4 (Normalverteilungsannahme). *Für ein σ^2 ist $\epsilon_1, ..., \epsilon_n$ unabhängig und nach $\mathcal{N}(0, \sigma^2)$ verteilt. (Insbesondere sind alle Varianzen identisch.)*

Ein erstes Ziel ist es, die Parameter β zu bestimmen bzw. zu schätzen. Als Konsequenz erhält man dann die Vorhersage $\hat{Y} = x\hat{\beta}$. Der Fit des Modells ist umso besser, je kleiner die Residuen $Y - \hat{Y}$ sind. Deshalb versucht man, die Summe der Residuenquadrate zu minimieren, also suchen wir β, so dass[5]

$$RSS(\beta) := \sum_{i=1}^{n} (Y_i - x_{i\cdot}\beta)^2 = (Y - x\beta)^\top (Y - x\beta) = Y^\top Y - 2Y^\top x\beta + \beta^\top x^\top x\beta$$

minimal wird. Wir nehmen im Folgenden immer an, dass $x^\top x$ invertierbar ist (ansonsten müssen wir mit Pseudo-Inversen arbeiten). Eine notwendige Bedingung ist damit

$$0 = \tfrac{1}{2}\nabla RSS(\beta) = -Y^\top x + \beta^\top x^\top x = (x^\top x\beta - x^\top Y)^\top,$$

also ist ein Extremum von $\beta \mapsto RSS(\beta)$ bei

$$\hat{\beta} = (x^\top x)^{-1} x^\top Y.$$

Theorem 1.5 (Multiple Regression). *Falls $x^\top x$ invertierbar ist, so ist das Minimum von $RSS(\beta)$ eindeutig und bei*

$$\hat{\beta} = (x^\top x)^{-1} x^\top Y.$$

Für die Vorhersage

$$\hat{Y} := x\hat{\beta} (= x(x^\top x)^{-1} x^\top Y)$$

gilt

$$Y - \hat{Y} = (I - x(x^\top x)^{-1} x^\top)\epsilon.$$

Außerdem stehen die Residuen $Y - \hat{Y}$ sowohl auf den Vorhersagen $\hat{Y}$, als auch auf den Spalten von x senkrecht.

Bemerkung 1.6 (Minimales RSS). Den minimalen Wert der *Residual Sum of Squares* bezeichnen wir mit

$$RSS := RSS(\hat{\beta}) = \sum_{i=1}^{n} (Y_i - x_{i\cdot}\hat{\beta})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (Y - \hat{Y})^\top (Y - \hat{Y}).$$

[5] RSS steht für *Residual Sum of Squares.*

Beweis von Theorem 1.5. Die Tatsache, dass der Gradient von $RSS(\beta)$ bei $\hat{\beta}$ verschwindet, haben wir oben bereits nachgerechnet. Weiter ist die Hesse-Matrix von $RSS(\beta)$ (für alle β) durch $x^\top x$ gegeben ist, also durch eine positiv-definite Matrix. Für die zweite Behauptung setzen wir $\hat{Y}$ in das Modell ein und erhalten

$$\begin{aligned} Y - \hat{Y} &= (I - x(x^\top x)^{-1}x^\top)Y = (I - x(x^\top x)^{-1}x^\top)(x\beta + \varepsilon) \\ &= x\beta + \epsilon - x\beta - x(x^\top x)^{-1}x^\top\epsilon = (I - x(x^\top x)^{-1}x^\top)\epsilon. \end{aligned} \tag{$\circ$}$$

Weiter schreiben wir

$$\begin{aligned} (Y - \hat{Y})^\top\hat{Y} &= Y^\top x(x^\top x)^{-1}x^\top Y - Y^\top x(x^\top x)^{-1}x^\top x(x^\top x)^{-1}x^\top Y = 0, \\ (Y - \hat{Y})^\top x &= Y^\top x - Y^\top x(x^\top x)^{-1}x^\top x = 0, \end{aligned}$$

woraus die behauptete Orthogonalität folgt. □

1.3 Schätzung der Modellparameter

Zwar haben wir nun Schätzer für $\hat{\beta}$ erhalten, allerdings wissen wir noch nichts über ihre Eigenschaften, etwa die Unverzerrtheit und Konsistenz. In diesem Abschnitt zeigen wir, dass $\hat{\beta}$ beide Eigenschaften besitzt (Theorem 1.7), und geben einen unverzerrten und konsistenten Schätzer für σ^2 an (Theorem 1.8).

Theorem 1.7 (Unverzerrtheit, Konsistenz von $\hat{\beta}$). *Gelten die Gauß-Markov-Bedingungen, so ist $\mathbb{E}_\beta[Y] = x\beta$ und $\hat{\beta}$ ist ein unverzerrter Schätzer für β. Weiter gilt*

$$\mathbb{COV}_{\beta,\sigma^2}[\hat{\beta}, \hat{\beta}] = \sigma^2(x^\top x)^{-1}.$$

Gilt $tr((x^\top x)^{-1}) \xrightarrow{n\to\infty} 0$, so ist $\hat{\beta}$ ein konsistenter Schätzer für β.

Beweis. Es gilt

$$\mathbb{E}_{\beta,\sigma^2}[\hat{\beta}] = (x^\top x)^{-1}x^\top(x\beta) = \beta,$$

woraus die Unverzerrtheit von $\hat{\beta}$ folgt. Weiter ist

$$\begin{aligned} \mathbb{COV}_{\beta,\sigma^2}[\hat{\beta}, \hat{\beta}] &= ((x^\top x)^{-1}x^\top)\mathbb{COV}_{\beta,\sigma^2}[Y, Y]x(x^\top x)^{-1}) \\ &= ((x^\top x)^{-1}x^\top)\mathbb{COV}_{\beta,\sigma^2}[\epsilon, \epsilon]x(x^\top x)^{-1}) \\ &= ((x^\top x)^{-1}x^\top)\sigma^2 I x(x^\top x)^{-1}) = \sigma^2(x^\top x)^{-1}. \end{aligned}$$

Für die Konsistenz ist zunächst klar, dass $\mathbb{V}_{\beta,\sigma^2}[\hat{\beta}_i] = \sigma^2((x^\top x)^{-1})_{ii}$. Da $(x^\top x)^{-1}$ als positiv-definite Matrix positive Diagonaleinträge hat, so folgt aus der Bedingung $\mathrm{tr}((x^\top x)^{-1}) \xrightarrow{n\to\infty} 0$, dass für $i = 1, ..., k$

$$\mathbb{V}_{\beta,\sigma^2}[\hat{\beta}_i] \xrightarrow{n\to\infty} 0$$

und die Behauptung folgt. □

Zwar haben wir nun einen unverzerrten und konsistenten Schätzer für β, jedoch sollten wir auch in der Lage sein, σ^2 zu schätzen.

Theorem 1.8 (Ein Schätzer für σ^2). *Gelten die Gauß-Markov-Bedingungen, so ist*

$$\widehat{\sigma^2} := \frac{1}{n-k-1} RSS = \frac{1}{n-k-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

ein unverzerrter und konsistenter Schätzer für σ^2.

Beweis. Zunächst ist mit $(\circ)$

$$\begin{aligned} RSS &= (Y-\hat{Y})^\top (Y-\hat{Y}) = \epsilon^\top (I - x(x^\top x)^{-1}x^\top)(I - x(x^\top x)^{-1}x^\top)\epsilon \\ &= \epsilon^\top (I - x(x^\top x)^{-1}x^\top)\epsilon. \end{aligned} \tag{$*$}$$

Wir berechnen mit Theorem 1.5[6]

$$\begin{aligned} \mathbb{E}_\beta[RSS] &= \mathbb{E}_\beta[\epsilon^\top (I - x(x^\top x)^{-1}x^\top)\epsilon] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}_\beta\Big[\epsilon_i (I - x(x^\top x)^{-1}x^\top)_{ij}\epsilon_j\Big] \\ &= \sigma^2 \sum_{i=1}^{n} ((I - x(x^\top x)^{-1}x^\top))_{ii} = \sigma^2 \mathrm{tr}(I - x(x^\top x)^{-1}x^\top) \\ &= \sigma^2(\mathrm{tr}(I) - \mathrm{tr}(x^\top x(x^\top x)^{-1})) = \sigma^2(n-k-1), \end{aligned}$$

woraus die Unverzerrtheit folgt. Für die Konsistenz schreiben wir mit $(*)$

$$\widehat{\sigma^2} = \frac{1}{n-k-1}(\epsilon^\top \epsilon - \epsilon^\top x(x^\top x)^{-1}x^\top \epsilon).$$

Nach dem Gesetz der großen Zahlen ist $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 \xrightarrow{n\to\infty}_{fs} \sigma^2$, also auch $\frac{1}{n-k-1}\epsilon^\top\epsilon \xrightarrow{n\to\infty}_{fs} \sigma^2$. Außerdem ist $(x^\top x)^{-1}$ positiv semi-definit und damit

$$\begin{aligned} \mathbb{E}_\beta[|\epsilon^\top x(x^\top x)^{-1}x^\top \epsilon|] &= \mathbb{E}_\beta[\epsilon^\top x(x^\top x)^{-1}x^\top \epsilon] = \mathbb{E}_\beta[\mathrm{tr}(\epsilon^\top x(x^\top x)^{-1}x^\top \epsilon)] \\ &= \mathrm{tr}(x(x^\top x)^{-1}x^\top \mathbb{E}_\beta[\epsilon\epsilon^\top]) = \sigma^2 \mathrm{tr}(x(x^\top x)^{-1}x^\top) = \sigma^2(k+1), \end{aligned}$$

also $\frac{1}{n}\epsilon^\top x(x^\top x)^{-1}x^\top \epsilon \xrightarrow{n\to\infty}_{L^1} 0$. Insgesamt folgt also die Konsistenz

$$\widehat{\sigma^2} \xrightarrow{n\to\infty}_{p} \sigma^2.$$

□

In Theorem 1.5 und im Beweis des letzten Theorems spielte die Matrix $I - x(x^\top x)^{-1}x^\top$ eine zentrale Rolle. Sie hat wichtige Eigenschaften, die wir nun sammeln. Wir wiederholen zunächst den Begriff der Idempotenz.

[6] Wir verwenden hier die wohlbekannten Tatsachen aus der linearen Algebra, dass für Matrizen A, B

$$\begin{aligned} \mathrm{tr}(A+B) &= \mathrm{tr}(A) + \mathrm{tr}(B), \\ \mathrm{tr}(AB) &= \sum_i \sum_j A_{ij}B_{ji} = \sum_i \sum_j A_{ji}B_{ij} = \mathrm{tr}(BA). \end{aligned}$$

Bemerkung 1.9 (Idempotente Matrix). Eine quadratische Matrix A heißt idempotent, wenn $A^2 = A$. Eine solche Matrix hat als Eigenwerte nur 0 und 1.
Denn: Ist $Av = \lambda v$ für ein $v \neq 0$, so gilt auch $Av = A^2v = \lambda Av = \lambda^2 v$ und damit $\lambda = \lambda^2$. Dies ist aber nur für $\lambda \in \{0, 1\}$ möglich.

Lemma 1.10 (Eigenschaften von $I - x(x^\top x)^{-1}x^\top$). *Die Matrix*

$$\Sigma := I - x(x^\top x)^{-1}x^\top$$

ist idempotent, symmetrisch und positiv semi-definit. Weiter ist $(x(x^\top x)^{-1}x^\top)_{ii} \leq 1$ für alle i und $rg(\Sigma) = n - k - 1$.

Beweis. Die Symmetrie und Idempotenz von Σ leitet man direkt her. Weiter ist klar, dass im letzten Beweis $RSS \geq 0$, ganz egal, welche Werte ϵ annimmt. Nun folgt die positive Semi-Definitheit von Σ aus $(*)$. Für die nächste Behauptung bemerken wir, dass die Diagonaleinträge einer positiv semi-definiten Matrix nicht-negativ sind. (Wäre der i-te Diagonaleintrag Σ_{ii}, so wäre $e_i^\top \Sigma e_i = \Sigma_{ii} < 0$, ein Widerspruch.) Es bleibt, die Aussage über den Rang von Σ zu zeigen. Da als Eigenwerte von Σ nur 0 und 1 in Betracht kommen (siehe Bemerkung 1.9), genügt es zu zeigen, dass die Summe der Eigenwerte von Σ gerade $n-k-1$ ist. Hierfür genügt es, $\mathrm{tr}(\Sigma) = n-k-1$ zu zeigen, wobei $\mathrm{tr}(\Sigma)$ die Spur von Σ ist (und bekanntermaßen invariant unter Ähnlichkeitstransformationen ist). Die Behauptung folgt nun aus

$$\mathrm{tr}(\Sigma) = \mathrm{tr}(I) - \mathrm{tr}(x(x^\top x)^{-1}x^\top) = n - \mathrm{tr}(x^\top x(x^\top x)^{-1}) = n - k - 1.$$

□

1.4 Fit der Regressionsgeraden

Wir wollen nun untersuchen, wie gut der Fit der Regressionsgeraden $\hat{Y} = x\beta$ an die Daten Y ist. Am besten geht dies durch die empirische Korrelation von Y und $\hat{Y}$.

Definition 1.11 (Bestimmtheitsmaß). *Das Bestimmtheitsmaß ist definiert als*

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

Wir wollen das Bestimmtheitsmaß nun durch die RSS ausdrücken, um einen klareren Zusammenhang zu sehen.

Proposition 1.12 (Darstellung des Bestimmtheitsmaßes). *Es gilt*

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

Bemerkung 1.13 (Interpretation). Liegt ein Bestimmtheitsmaß von R^2 vor, so sagt man auch, dass die Regressionsgerade einen Anteil von R^2 an der Varianz der Daten erklärt. Grund hierfür ist die erste Darstellung aus der Proposition. Die *erklärte Varianz* ist ja gerade $\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$, und die *Gesamtvarianz* ist $\sum_{i=1}^n (Y_i - \overline{Y})^2$.

Beweis. Zunächst zeigen wir die beiden Identitäten

$$RSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - (\hat{Y}_i - \overline{Y})^2,$$
$$\sum_{i=1}^{n} (Y_i - \overline{Y})(\hat{Y}_i - \overline{Y}) = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2.$$

Sind diese gezeigt, so folgt die Aussage einfach aus

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

Für die erste Identität wissen wir aus Theorem 1.5, dass $\hat{Y} - Y$ auf der ersten Spalte von x, also auf 1, senkrecht steht. Deshalb ist $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} Y_i$. Damit ergibt sich, da $\hat{Y}^\top$ auf $Y - \hat{Y}$ senkrecht steht, $\hat{Y}^\top \hat{Y} = (\hat{Y} - Y + Y)^\top \hat{Y} = Y^\top \hat{Y}$ und

$$\begin{aligned}\sum_{i=1}^{n} (Y_i - \overline{Y})^2 - (\hat{Y}_i - \overline{Y})^2 &= \sum_{i=1}^{n} Y_i^2 - \hat{Y}_i^2 = Y^\top Y - \hat{Y}^\top \hat{Y} \\ &= Y^\top (Y - \hat{Y}) = (Y - \hat{Y})^\top (Y - \hat{Y}) = RSS.\end{aligned}$$

Für die zweite Identität schreiben wir

$$\begin{aligned}\sum_{i=1}^{n} (Y_i - \overline{Y})(\hat{Y}_i - \overline{Y}) &= (Y - \overline{Y}I)^\top (\hat{Y} - \overline{Y}I) = (Y - \hat{Y} + \hat{Y} - \overline{Y}I)^\top (\hat{Y} - \overline{Y}I) \\ &= (\hat{Y} - \overline{Y}I)^\top (\hat{Y} - \overline{Y}I) = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2\end{aligned}$$

und alle Aussagen sind gezeigt. □

1.5 Das Gauß-Marov-Theorem

Die Schätzer $\hat{\beta}$ haben wir mit der Methode der kleinsten Quadrate erhalten. Nun geben wir einen berühmten Satz, dass dieses Vorgehen in der Tat in gewissem Sinn optimal ist.

Definition 1.14 (BLUE). *Im Regressionsmodell heißt jeder Schätzer $y \mapsto c_\ell^\top y$* linear. *Er heißt unverzerrt (für β), falls*

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

für alle ℓ. Weiter heißt er Best Linear Unbiased Estimator (BLUE) *(für β), wenn er unverzerrt ist und $\mathbb{V}_\beta[c_\ell^\top Y] \leq \mathbb{V}_\beta[d_\ell^\top Y]$ für jeden linearen unverzerrten Schätzer $y \mapsto d_\ell^\top y$ ist.*

Bemerkung 1.15 (Ein linearer unverzerrter Schätzer). Aus Theorem 1.7 wissen wir bereits, dass $\hat{\beta} = (x^\top x)^{-1} x^\top Y$ ein unverzerrter Schätzer (für β) ist. Setzen wir $c_\ell = x(x^\top x)^{-1}\ell$, so ist damit $\mathbb{E}_\beta[c_\ell^\top Y] = \ell^\top \mathbb{E}_\beta[\hat{\beta}] = \ell^\top \beta$ und damit ist $y \mapsto c_\ell^\top y$ ein unverzerrter, linearer Schätzer. Das folgende Resultat zeigt, dass es sich auch um einen BLUE handelt.

Theorem 1.16 (Gauß-Markov-Theorem). *Sei $\hat{\beta} = (x^\top x)^{-1} x^\top Y$. Falls die Gauß-Markov-Bedingungen gelten, ist $y \mapsto \ell^\top (x^\top x)^{-1} x^\top y = \ell^\top \hat{\beta}$ ein BLUE.*

Beweis. Sei $y \mapsto d_\ell^\top y$ ein weiterer linearer, unverzerrtter Schätzer für β, also

$$\ell^\top \beta = \mathbb{E}_\beta[d_\ell^\top Y] = d_\ell^\top x\beta.$$

Da dies für alle ℓ gelten muss, ist also $x^\top d_\ell = \ell$. Wir schreiben nun mit Hilfe von Theorem 1.7

$$\begin{aligned}\mathbb{V}_\beta[d_\ell^\top Y] - \mathbb{V}_\beta[\ell^\top \hat{\beta}] &= d_\ell^\top \mathbb{COV}_\beta[Y,Y]d_\ell - \ell^\top \mathbb{COV}_\beta[\hat{\beta},\hat{\beta}]\ell \\ &= \sigma^2 d_\ell^\top d_\ell - \sigma^2 d_\ell^\top x(x^\top x)^{-1}x^\top d_\ell = \sigma^2 d_\ell^\top (I - x(x^\top x)^{-1}x^\top)d_\ell \geq 0\end{aligned}$$

wegen Lemma 1.10. □

1.6 Statistische Tests im Regressionsmodell

Oft will man herausfinden, ob man bei einer Regression auch mit weniger Covariaten auskommt. Könnte man etwa auf die i-te Covariate verzichten, so würde das auf ein Modell mit $\beta_i = 0$ hinauslaufen. Mit anderen Worten wollen wir im Regressionsmodell $H_0 : \beta_i = 0$ gegen $H_1 : \beta_i \neq 0$ testen. Etwas allgemeiner beschreiben wir im Folgenden Tests von $H_0 : A\beta - \gamma = 0$ für $A \in \mathbb{R}^{m\times(k+1)}$ mit Rang $m \leq k+1$ und $\gamma \in \mathbb{R}^m$ gegen $H_1 : A\beta - \gamma \neq 0$. Die Teststatistik wird dann eine F-Verteilung besitzen, die wir zunächst definieren.

Definition 1.17 (F-Verteilung). *Seien $X_1, ..., X_k, Y_1, ..., Y_l$ unabhängig und nach $\mathcal{N}(0,1)$ verteilt. Dann heißt die Verteilung von*

$$\frac{(X_1^2 + \cdots + X_k^2)/k}{(Y_1^2 + \cdots + Y_l^2)/l}$$

F-Verteilung mit Freiheitsgraden k und l oder $F_{k,l}$. Ihr p-Quantil bezeichnen wir mit $F_{k,l,p}$.

Bemerkung 1.18 (Äquivalente Formulierung). Bekanntermaßen hat $X_1^2 + \cdots + X_k^2$ (für $X_1, ..., X_k$ unabhängig nach $\mathcal{N}(0,1)$ verteilt) gerade eine χ_k^2-Verteilung (d.h. eine χ^2-Verteilung mit k Freiheitsgraden. Sind also $Z_1 \sim \chi_k^2$ und $Z_2 \sim \chi_l^2$ zwei unabhängige χ^2-Verteilungen, so ist

$$\frac{Z_1/k}{Z_2/l} \sim F_{k,l}.$$

Wir werden zwei Eigenschaften von mehrdimensionalen Normalverteilugen benötigen, die wir nun wiederholen.

Bemerkung 1.19 (Mehrdimensionale Normalverteilung). Sei $b \in \mathbb{R}^k$ und Σ symmetrisch und positiv semi-definit.

1. Ist $Y \sim \mathcal{N}(b, \Sigma)$, dann ist $AY \sim \mathcal{N}(Ab, A\Sigma A^\top)$.
   Denn: Es gilt $\mathbb{E}[AY] = Ab$ und

   $$\mathbb{COV}[AY, AY] = \mathbb{E}[(AY - Ab)(AY - Ab)^\top] = \mathbb{E}[AYY^\top A^\top] - Abb^\top A^\top = A\Sigma A^\top$$

2. Ist $Y \sim \mathcal{N}(0, \Sigma)$ und Σ eine idempotente Matrix von Rang r. Dann ist $Y^\top \Sigma Y \sim \chi_r^2$.
   Denn: Da Σ symmetrisch ist, und Σ nach Bemerkung 1.9 als Eigenwerte nur 0 und 1 hat, gibt es ein O orthogonal und $D = \mathrm{diag}(1, ..., 1, 0, ..., 0)$ mit $\mathrm{rg}(D) = r$, so dass $ODO^\top = \Sigma$. Damit ist $O^\top Y \sim \mathcal{N}(0, O^\top \Sigma O) = \mathcal{N}(0, D)$ und $Y^\top \Sigma Y = Y^\top ODO^\top Y \sim \chi_r^2$.

Theorem 1.20 (χ^2-Verteilungen im Regressionsmodell). *Es gelte Annahme 1.4. Ist $A\beta - \gamma = 0$, so ist unter $\mathbb{P}_\beta$ mit $\hat{\beta} = (x^\top x)^{-1}x^\top Y$*

$$\frac{1}{\sigma^2}(A\hat{\beta} - \gamma)^\top (A(x^\top x)^{-1}A^\top)^{-1}(A\hat{\beta} - \gamma) \sim \chi^2_m,$$
$$\frac{1}{\sigma^2}Y^\top (I - x(x^\top x)^{-1}x^\top)Y \sim \chi^2_{n-k-1}$$

und die Zufallsvariablen in den beiden Zeilen sind unabhängig.

Teilt man die beiden Zufallsvariablen des letzten Theorems durcheinander, so erhält man sofort eine F-verteilte Zufallsgröße, die später als Teststatistik dient.

Korollar 1.21 (Verteilung der Teststatistik). *Es gelte Annahme 1.4. Ist $A\beta - \gamma = 0$, so ist unter $\mathbb{P}_\beta$*

$$F := \frac{(A\hat{\beta} - \gamma)^\top (A(x^\top x)^{-1}A^\top)^{-1}(A\hat{\beta} - \gamma)}{m\widehat{\sigma^2}} \sim F_{m,n-k-1}$$

mit $\widehat{\sigma^2}$ wie in Theorem 1.8.

Beweis von Theorem 1.20. Nach Theorem 1.7 ist $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(x^\top x)^{-1})$. Damit ist, falls $A\beta - \gamma = 0$ nach Bemerkung 1.19.1

$$A\hat{\beta} - \gamma \sim \mathcal{N}(A\beta - \gamma, \sigma^2 A(x^\top x)^{-1}A^\top) = \mathcal{N}(0, \sigma^2 A(x^\top x)^{-1}A^\top).$$

Da $A(x^\top x)^{-1}A^\top$ positiv definit ist, gibt es $(A(x^\top x)^{-1}A^\top)^{-1}$ und auch die Wurzel $(A(x^\top x)^{-1}A^\top)^{-1/2}$. Es ist

$$\tfrac{1}{\sqrt{\sigma^2}}(A(x^\top x)^{-1}A^\top)^{-1/2}(A\hat{\beta} - \gamma) \sim \mathcal{N}(0, I),$$

also auch

$$\tfrac{1}{\sigma^2}(A\hat{\beta} - \gamma)^\top (A(x^\top x)^{-1}A^\top)^{-1}(A\hat{\beta} - \gamma) \sim \chi^2_m.$$

Für die zweite Zufallsvariable erinnern wir an Lemma 1.10, wo wir gezeigt haben, dass

$$\Sigma := I - x(x^\top x)x^\top$$

symmetrisch, nicht-negativ definit, idempotent und von Rang $n - k - 1$ ist. Da $\frac{1}{\sqrt{\sigma^2}}(I - x(x^\top x)^{-1}x^\top)Y = \frac{1}{\sqrt{\sigma^2}}\Sigma Y \sim N(0, \Sigma)$, ist nach Bemerkung 1.19.2

$$\frac{1}{\sigma^2}Y^\top (I - x(x^\top x)^{-1}x^\top)Y = \frac{1}{\sigma^2}Y^\top \Sigma Y \sim \chi^2_{n-k-1}.$$

Um die Unabhängigkeit einzusehen, schreiben wir

$$\begin{aligned}\mathbb{COV}_\beta[\hat{\beta}, \Sigma Y] &= \mathbb{E}_\beta[(x^\top x)^{-1}x^\top \epsilon\epsilon^\top (I - x(x^\top x)^{-1}x^\top)] \\ &= \sigma^2((x^\top x)^{-1}x^\top - (x^\top x)^{-1}x^\top x(x^\top x)^{-1}x^\top) = 0.\end{aligned}$$

Damit sind die beiden normalverteilten Zufallsvariablen $\hat{\beta}$ und $(I - x(x^\top x)^{-1}x^\top)Y$ unabhängig. Die Zufallsvariable der ersten Zeile des Theorems ist eine Funktion von $\hat{\beta}$ und die in der zweiten Zeile ist wegen

$$Y^\top (I - x(x^\top x)^{-1}x^\top)Y = Y^\top \Sigma Y = (\Sigma Y)^\top \Sigma Y$$

eine Funktion von ΣY. Damit sind beide Zufallsvariablen unabhängig. □

Beispiel 1.22 (Test auf $\beta_i = 0$). Wollen wir die Nullhypothese $H_0 : \beta_i = 0$ testen, So setzen wir $A = e_i^\top$ (dem i-ten kanonischen Basisvektor) und $\gamma = 0$. Im Beweis von Theorem 1.20 haben wir gesehen, dass

$$\frac{1}{\sqrt{\sigma^2}}(e_i^\top (x^\top x)^{-1} e_i)^{-1/2} e_i^\top \hat{\beta} = \frac{1}{\sqrt{\sigma^2((x^\top x)^{-1})_{ii}}}\hat{\beta}_i \sim \mathcal{N}(0,1)$$

unabhängig von $\frac{1}{\sigma^2}\widehat{\sigma^2} \sim \chi^2_{n-k-1}$ ist. Damit ist

$$T_i := \frac{\hat{\beta}_i}{\sqrt{((x^\top x)^{-1})_{ii}\widehat{\sigma^2}}} \sim t_{n-k-1}.$$

Deshalb ist für $\alpha \in (0,1)$ das Tupel (T, C) mit $C = (-\infty, t_{n-k-1,\alpha/2}) \cup (t_{n-k-1,1-\alpha/2}, \infty)$ ein Test von H_0 gegen $H_1 : \beta_i \neq 0$ zum Niveau α.

Beispiel 1.23 (Test auf $\beta = 0$). Will man testen, ob überhaupt ein Zusammenhang zwischen den Covariaten und Zielvariablen besteht, so überprüft man die Nullhypothese $H_0 : \beta_1 = \cdots = \beta_k = 0$. (Man beachte, dass $\beta_0 \neq 0$ zugelassen ist.) Hierzu verwenden wir in Korollar 1.21

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

und $\gamma = 0$. Damit ist (F, C) mit $C = (F_{k,n-k-1,1-\alpha}, \infty)$ ein Test von H_0 zum Signifikanzniveau α.

1.7 Ein R-Beispiel

Wir kommen noch einmal zurück zu den Daten von Geysir-Ausbrüchen aus Beispiel 1.2. Wir nehmen an, dass die Daten normalverteilt sind. Hierzu sehen wir uns nun die Ausgabe von

```
> summary(lm(eruptions ~ waiting, data=faithful))
```

an:

```
Call:
lm(formula = eruptions ~ waiting, data = faithful)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Zunächst wird hier eine Zusammenfassung der Residuen (`residuals`), also $Y - \hat{Y}$ angegeben. Dies geschieht durch Angabe des minimalen und maximalen Wertes, sowie durch Angabe der drei Quartile. Als nächstes werden die Werte $\hat{\beta}_0$ und $\hat{\beta}_1$ im Modell

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

angegeben. Die hier berichteten Standardfehler (`Std. Error`) sind durch

$$\widehat{\text{s.e.}}(\hat{\beta}_i) := \sqrt{\widehat{\sigma^2}(x^\top x)^{-1}_{ii}}$$

mit $\widehat{\sigma^2}$ aus Theorem 1.8 gegeben. Diese Formel begründet sich mit Theorem 1.7, wobei σ^2 durch einen Schätzer ersetzt wurde. Der nachfolgende t-Wert (`t value`) ist wie in Beispiel 1.22 berechnet. Der entsprechende p-Wert (`Pr(>|t|)`) ist sowohl für β_0 als auch für β_1 so klein, dass selbst zu einem sehr kleinen Signifikanzniveau die Hypothese $\beta_0 = 0$ bzw. $\beta_1 = 0$ nicht abgelehnt werden kann. Der residuale Standardfehler (`Residual standard error`) ist gerade $\sqrt{\widehat{\sigma^2}}$, Das Bestimmtheitsmaß (`Multiple R-squared`) haben wir in Proposition 1.12 bestimmt. (Der `Adjusted R-squared` ergibt sich dabei aus $1 - \widehat{\sigma^2}(n-1)/(\sum_{i=1}^n Y_i - \overline{Y})^2$; vergleiche mit Proposition 1.12) Schließlich wird die F-Statistik angegeben, die sich beim Test von $\beta_1 = 0$ zu $\beta_1 \neq 0$ ergibt; siehe Beispiel 1.23.

Wichtige Formeln

$$
\begin{aligned}
Y &= x\beta + \epsilon && \\
\hat{Y} &= x\hat{\beta} && \text{Theorem 1.5} \\
\hat{\beta} &= (x^\top x)^{-1} x^\top Y && \text{Theorem 1.5} \\
Y - \hat{Y} &= (I - x(x^\top x)^{-1}x^\top)Y = (I - x(x^\top x)^{-1}x^\top)\epsilon && \text{Theorem 1.5 und } (\circ) \\
RSS &= (Y - \hat{Y})^\top (Y - \hat{Y}) = \epsilon^\top (I - x(x^\top x)^{-1}x^\top)\epsilon && \text{Bemerkung 1.6 und } (*) \\
\mathbb{COV}_\beta[\hat{\beta}, \hat{\beta}] &= \sigma^2 (x^\top x)^{-1} && \text{Theorem 1.7} \\
\widehat{\sigma^2} &= \frac{1}{n-k-1} RSS && \text{Theorem 1.8} \\
R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{RSS}{\sum_{i=1}^n (Y_i - \overline{Y})^2} && \text{Proposition 1.12}
\end{aligned}
$$

2 Varianzanalyse

2.1 Einleitung

Bei der (ein-faktoriellen) Varianzanalyse will man Unterschiede zwischen metrischen Merkmalen X verschiedener Gruppen herausfinden. (Man stelle sich etwa die Wirkung p verschiedener Behandlungsmethoden auf ein biometrisches Merkmal bei einer Krankheit vor.) Man betrachtet also p Populationen und Stichprobengrößen $n_1, ..., n_p$. Die zu messende metrische Größe X wird auch *Faktor* genannt, die einzelnen Gruppen als *Levels* oder *Faktorstufen*.

Beispiel 2.1 (Insektensprays). Wir verwenden den in R verfügbaren Datensatz `InsectSprays` mittels

```
> attach(InsectSprays)
> a<-data(InsectSprays)
```

Der Datensatz enthält eine Untersuchung von sechs verschiedenen Insektensprays und deren Auswirkungen auf die gefundene Zahl der Insekten auf einem damit behandelten Gebiet. Mit den obigen Befehlen stehen nun die Variablen `spray` und `count` zur Verfügung.

```
> spray
 [1] A A A A A A A A A A A A B B B B B B B B B B B C C C C C C C C C C C C D D
[39] D D D D D D D D D E E E E E E E E E E E E F F F F F F F F F F F F
Levels: A B C D E F
> count
 [1] 10  7 20 14 14 12 10 23 17 20 14 13 11 17 21 11 16 14 17 17 19 21  7 13  0
[26]  1  7  2  3  1  2  1  3  0  1  4  3  5 12  6  4  3  5  5  5  5  2  4  3  5
[51]  3  5  3  6  1  1  3  2  6  4 11  9 15 22 15 16 13 10 26 26 24 13
```

Die sechs verschiedenen Sprays sind mit `A` bis `F` gekennzeichnet. Da wir uns damit befassen wollen, ob die verschiedenen Sprays gleiche oder unterschiedliche Effekte auf die Insektenzahlen haben, verschaffen wir uns zunächst einen Überblick über die Daten.[7]

```
> tapply(count, spray, mean)
        A         B         C         D         E         F
14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

Beispielsweise sehen wir so, dass bei Gruppen `E` und `F` die Mittelwerte stark voneinander abweichen.

Eine weitere sinnvolle Methode, sich einen Überblick über die Daten zu verschaffen, ist es, eine Grafik zu erstellen. In unserem Fall bietet es sich an, einen *Box(-Whisker)-Plot* zu verwenden. Dieser wird von

[7] Der Aufruf von `tapply(count, spray, mean)` wendet die Funktion `mean` auf die Zielvariable `count` an, wobei sie die Faktoren `spray` unterscheidet. Die Vektoren `count` und `spray` müssen gleich lang sein. Etwa liefert

```
> tapply(count, spray, length)
A B C D E F
12 12 12 12 12 12
```

da alle Faktoren 12-mal in `spray` vorkommen.

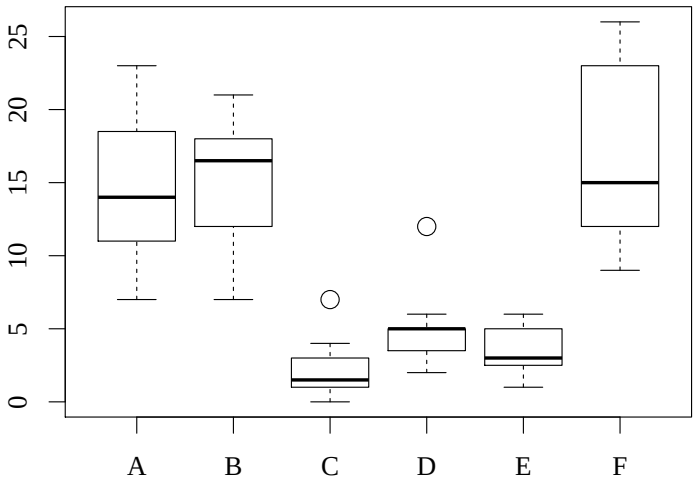


Abbildung 2.1: Der Box-Plot der Auswirkungen sechs verschiedener Insektensprays.

```
> boxplot(count ~ spray)
```

erzeugt; siehe Abbildung 2.1. Hier wird für alle Levels eine Box angelegt. Der horizontale Strich innerhalb der *Box* stellt den Median dar, die Begrenzungen der *Box* das erste und dritte Quartil (und die *Box* damit den Interquartilbereich). Die *Whiskers*[8] reichen maximal bis zum kleinsten bzw. größten Wert der Daten und sind maximal die anderthalbfache Breite des Interquartilbereiches lang. Daten außerhalb dieses Bereichs werden als einzelne Punkte dargestellt.

2.2 Das Modell

Für die Varianzanalyse habe im Modell der Faktor innerhalb der Population k einen Mittelwert von β_k, $k = 1, ..., p$. Weiterhin werden wir wie auch bei der Regression eine gemeinsame Varianz von σ^2 annehmen. Die Modellannahmen lauten also

$$Y_{ki} = \beta_k + \epsilon_{ki}, \qquad k = 1, ..., p, i = 1, ..., n_k,$$

mit $\epsilon \sim N(0, \sigma^2 I)$ und $n = n_1 + \cdots + n_p$. Ziel der Varianzanalyse ist es, die Nullhypothese

$$H_0 := \beta_1 = \cdots = \beta_p$$

gegen $H_1 : \beta_k \neq \beta_\ell$ *für ein Paar* k, ℓ zu testen. Hierfür berechnen wir zunächst die Stichprobenmittel innerhalb der *Faktorstufen* sowie das Gesamtmittel

$$\overline{Y}_{k\bullet} := \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}, \qquad \overline{Y} := \frac{1}{n} \sum_{k=1}^{p} \sum_{i=1}^{n_k} Y_{ki}.$$

Beispiel 2.2 (Insektensprays, Modellannahmen). Da wir annehmen, dass $Y_{ki} \sim N(\beta_k, \sigma^2)$, ist

$$\overline{Y}_{k\bullet} \sim N(\beta_k, \sigma^2/n_k),$$

[8] Dies bezeichnet auch die Schnurrhaare einer Katze.

insbesondere sollten $\sqrt{n_k}\overline{Y}_{k\bullet}$ dieselben Varianzen haben. Diese können wir erwartungstreu schätzen, indem wir die empirischen Varianzen innerhalb des k-ten Levels der Stichprobe betrachten.

```
> tapply(count, spray, sd)
       A        B        C        D        E        F
4.719399 4.271115 1.975225 2.503028 1.732051 6.213378
```

Für Level `C` ergibt sich eine kleine Varianz, so dass sich eine Abweichung der Modellannahmen entstehen könnte. Im Moment gehen wir noch nicht darauf ein, die Hypothese der gleichen Varianzen zu testen.

Eine weitere Möglichkeit, die (Un-)Gleichheit der Varianzen zu sehen, ist die grafische Darstellung der Residuen $Y_{ki} - \overline{Y}_{k\bullet}$. Bereits in Abbildung 2.1 sieht man jedoch, dass kleineres $\overline{Y}_{k\bullet}$ mit einer eher kleineren Streuung einhergeht. □

Zurück zum Test von H_0 gegen H_1. Grundgedanke der Varianzanalyse ist die Varianzzerlegung, also die Zerlegung der Stichprobenvarianz (Sum of sQuares Total)

$$SQT := \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y})^2 \qquad \text{(Sum of sQuares Total).}$$

Diese Zerlegung ist folgendermaßen gegeben.

Proposition 2.3 (Varianzzerlegung). *Es gilt*

$$SQT = SQE + SQR$$

für SQT wie oben und

$$SQE := \sum_{k=1}^{p} n_k(\overline{Y}_{k\bullet} - \overline{Y})^2 \qquad \textit{(Sum of sQuares Explained)},$$

$$SQR := \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y}_{k\bullet})^2. \qquad \textit{(Sum of sQuares Residual)}$$

Beweis. Wir schreiben

$$\begin{aligned}\sum_{k=1}^{p} n_k(\overline{Y}_{k\bullet} - \overline{Y})^2 + \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y}_{k\bullet})^2 &= \sum_{k=1}^{p} n_k(Y_{k\bullet}^2 - \overline{Y}^2) + \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki}^2 - \overline{Y}_{k\bullet}^2)\\ &= \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki}^2 - \overline{Y}^2) = \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y})^2.\end{aligned}$$

□

Mit diesem Resultat geben wir nun eine Teststatistik an, mit der man H_0 testen kann.

Theorem 2.4 (Varianzanalyse). *Ist $\beta_1 = \cdots = \beta_p = 0$, so ist*

$$SQT/\sigma^2 \sim \chi^2_{n-1}, \qquad SQE/\sigma^2 \sim \chi^2_{p-1}, \qquad SQR/\sigma^2 \sim \chi^2_{n-p},$$

und

$$\frac{SQE/(p-1)}{SQR/(n-p)} \sim F_{p-1,n-p}.$$

Beweis. OBdA sei $\mu = 0, \sigma^2 = 1$, da der allgemeine Fall durch lineare Transformation in diesen überführt werden kann. Für die erste Aussage ändern wir die Nummerierung der Y's zu $Y_1, ..., Y_n$. Sei $O \in \mathbb{R}^{n\times n}$ eine orthogonale Matrix mit $O_{11} = \cdots = O_{1n} = 1/\sqrt{n}$. Dann ist $Z := OY \sim N(0, I)$ (wobei I die Einheitsmatrix ist) und $Z_1 = (OY)_1 = \sqrt{n}\overline{Y}$ sowie

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = \sum_{i=1}^{n} Y_i^2 - Z_1^2 = \sum_{i=2}^{n} Z_i^2 \sim \chi^2_{n-1}.$$

Die Aussage über SQR ergibt sich analog, da $\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y}_{k\bullet})^2 \sim \chi^2_{n_k-1}$, $k = 1, ..., p$ und diese Zufallsvariablen für verschiedene k unabhängig sind. Bei der Aussage über SQE setzen wir $W_k := \sqrt{n_k}\overline{Y}_{k\bullet}$. Dann ist $W_1, ..., W_p$ unabhängig mit $W_k \sim N(0, 1)$ und genau wie oben folgt $SQE \sim \chi^2_{p-1}$. Es bleibt noch, die Unabhängigkeit von SQR und SQE zu zeigen. Hierzu bemerken wir, dass SQE eine Funktion von $(\overline{Y}_{k\bullet} - \overline{Y})_{k=1,...,p}$ ist, und SQR eine Funktion von $(Y_{ki} - \overline{Y}_{k\bullet})_{k=1,...,p,i=1,...,n_k}$. Diese beiden Vektoren sind unabhängig, da

$$\mathbb{COV}[n_k(\overline{Y}_{k\bullet} - \overline{Y}), Y_{\ell i} - \overline{Y}_{\ell\bullet}] = \delta_{k\ell} - \delta_{k\ell} - \frac{n_k}{n} + \frac{n_k}{n} = 0.$$

□

Beispiel 2.5 (Insektensprays). Wir führen nun eine Varianzanalyse für das Beispiel 2.1 durch. Dies funktioniert mit der Funktion `aov` (was für *Analysis Of Variance* steht)

```
> aov.out = aov(count ~ spray, data=InsectSprays)
> summary(aov.out)
```

Dies liefert

```
            Df Sum Sq Mean Sq F value Pr(>F)
spray        5   2669   533.8    34.7 <2e-16 ***
Residuals   66   1015    15.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Da es $p = 6$ Faktoren gibt, wird die Anzahl von $p - 1 = 5$ Freiheitsgraden für SQE berichtet. Dabei ist $SQE = 2669$ und $SQR = 1015$ mit $n - p = 66$ Freiheitsgraden. Zähler und Nenner der Teststatistik werden in der nächsten Spalte berichtet, also $SQE/(p-1)$ und $SQR/(n-p)$. Der entsprechende p-Wert ist

```
> 1 - pF(34.7,5,66)
[1] 0
```

was R kleiner als $2 \cdot 10^{-16}$ berichtet, der internen Rechengenauigkeit.

Die Varianzanalyse verläuft nach dem letzten Theorem folgendermaßen ab:

Modell der einfaktoriellen Varianzanalyse

Annahme	$Y_{ki} = \beta_k + \epsilon_{ki} \quad (k = 1, \ldots, p, i = 1, \ldots, n_k)$
Dabei sind	
$Y_{11}, \ldots, Y_{pn_p}$	gegebene Merkmalsausprägungen eines Merkmals gemessen in Levels $1, \ldots, p$
β_k	erwarteter Effekt der k-ten Faktorstufe auf die Ausprägung des Merkmals
$\epsilon_{11}, \ldots, \epsilon_{p,n_p}$	Zufallsvariablen, die die Abweichung der Messdaten des k-ten Levels messen. Diese sind unabhängig, identisch verteilt mit $\epsilon_{ki} \sim N(0, \sigma^2)$.
Hypothesen	$H_0 : \beta_1 = \ldots = \beta_p \quad$ gegen $\quad H_1 : \beta_k \neq \beta_\ell$ für ein Paar k, ℓ
Teststatistik	$F = \dfrac{SQE/(p-1)}{SQR/(n-p)} \sim F(p-1, n-p)$
Ablehnungsbereich	$F >$ $(1-\alpha)$-Quantil von $F(p-1, n-p)$
p-Wert	$1 - P_{F(p-1,n-p)}(F)$

2.3 Verbindung zu Regression

Die Varianzanalyse lässt sich mit der Regression vergleichen. Wir können die Modellannahmen auch schreiben als $Y = x\beta + \epsilon$ mit

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & & & \cdots & 1 \\ \vdots & & & & \vdots \\ 0 & & & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \vdots \\ \epsilon_{p1} \\ \vdots \\ \epsilon_{pn_p} \end{pmatrix}.$$

In diesem Fall ist

$$x^\top x = \begin{pmatrix} n_1 & 0 & \cdots & \cdots & 0 \\ 0 & n_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & \cdots & n_p \end{pmatrix}, \qquad (x^\top x)^{-1} = \begin{pmatrix} n_1^{-1} & 0 & \cdots & \cdots & 0 \\ 0 & n_2^{-1} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & \cdots & n_p^{-1} \end{pmatrix},$$

und damit ist

$$\hat\beta = (x^\top x)^{-1} x^\top Y = \Big(\frac{1}{n_1}(Y_{11} + \cdots + Y_{1n_1}), \cdots, \frac{1}{n_p}(Y_{p1} + \cdots + Y_{pn_p})\Big)^\top =: \big(\overline{Y}_{1\bullet}, \cdots, \overline{Y}_{p\bullet}\big)^\top$$

der kleinste-Quadrate-Schätzer von β. Das ist auch nicht erstaunlich, ist doch $\overline{Y}_{k\bullet}$ der Mittelwert der Beobachtungen in Klasse k. Weiter schreiben wir

$$\hat Y = x\hat\beta = (\underbrace{\overline{Y}_{1\bullet}, \cdots, \overline{Y}_{1\bullet}}_{n_1-\text{mal}}, \cdots, \underbrace{\overline{Y}_{p\bullet}, \cdots, \overline{Y}_{p\bullet}}_{n_p-\text{mal}}, \cdots,)^\top,$$

$$RSS = (Y - \hat Y)^\top (Y - \hat Y) = \sum_{k=1}^{p} \sum_{i=1}^{n_k} (Y_{ki} - \overline{Y}_{k\bullet})^2 = SQR.$$

Aus dem Beweis von Proposition 4.2 aus dem Skript *Regression* folgt damit die Varianzzerlegung

$$\sum_{k=1}^{p} \sum_{i=1}^{n_k} (Y_{ki} - \overline{Y})^2 = \sum_{k=1}^{p} \sum_{i=1}^{n_k} (Y_{ki} - \overline{Y}_{k\bullet})^2 + \sum_{k=1}^{p} n_k (\overline{Y}_{k\bullet} - \overline{Y})^2$$

Den Test $\beta_1 = \cdots = \beta_p$ werden wir nun für den einfacheren Fall $p = 3$ und $n_1 = n_2 = n_3 =: q$ anhand von Korollar 6.5 aus dem Skript *Regression* erklären. Hierzu setzen wir $A\beta - \gamma = 0$ für $\gamma = 0$ und

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p}$$

(und also $m = 2$). Nun ist für den Zähler der Statistik aus Korollar 6.5 des Skripts zur *Regression*

$$A(x^\top x)^{-1} A^\top = \frac{1}{q} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \qquad (A(x^\top x)^{-1} A^\top)^{-1} = \frac{q}{3} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

und damit

$$\begin{aligned} A\hat\beta &= (\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}, \overline{Y}_{2\bullet} - \overline{Y}_{3\bullet})^\top, \\ (A\hat\beta)^\top (A(x^\top x)^{-1} A^\top)^{-1} A\hat\beta &= \frac{q}{3} (A\hat\beta)^\top \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} A\hat\beta \\ &= \frac{2q}{3} ((\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet})^2 + (\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet})(\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet}) + (\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet})^2) \\ &= \frac{2q}{3} (\overline{Y}_{1\bullet}^2 + \overline{Y}_{2\bullet}^2 + \overline{Y}_{3\bullet}^2 - \overline{Y}_{1\bullet}\overline{Y}_{2\bullet} - \overline{Y}_{1\bullet}\overline{Y}_{3\bullet} - \overline{Y}_{2\bullet}\overline{Y}_{3\bullet}) \\ &= \frac{q}{3} (3(\overline{Y}_{1\bullet}^2 + \overline{Y}_{2\bullet}^2 + \overline{Y}_{3\bullet}^2) - (\overline{Y}_{1\bullet} + \overline{Y}_{2\bullet} + \overline{Y}_{3\bullet})^2) \\ &= q(\overline{Y}_{1\bullet}^2 + \overline{Y}_{2\bullet}^2 + \overline{Y}_{3\bullet}^2 - 3\overline{Y}^2) = SQE \end{aligned}$$

und

$$\widehat{\sigma^2} = \frac{RSS}{n-p} = \frac{SQR}{n-p}.$$

Damit folgt, dass die Statistik aus Korollar 6.5 identisch mit der aus Theorem 2.4 ist.

2.4 Erweiterungen

Bemerkung 2.6 (Untersuchung bei signifikantem Ergebnis, Tukey's Test). Kann man nun H_0 ablehnen, stellt man sich sofort die Frage, zwischen welchen Levels der Unterschied der Mittelwerte denn für dieses Ergebnis entscheidend war. Hierfür kann man einen *Post-Hoc*-Test an die Varianzanalyse anschließen. Einer dieser Tests ist *Tukey's Test*. Er basiert auf der t-Range-Statistik. Im Modell der Varianzanalyse sei (falls $n_1 = \cdots = n_p$)

$$Q := \frac{\max_k \hat{\beta}_k - \min_k \hat{\beta}_k}{\sqrt{\widehat{\sigma^2}/n_k}}$$

Um etwas über diese Verteilung zu erfahren, stehen die R-Befehle `ptukey` und `qtukey` zur Verfügung. Ausführung des Tukey-Tests für den `InsectSprays`-Datensatz liefert

```
> TukeyHSD(aov.out)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = count ~ spray)

$spray
           diff        lwr       upr     p adj
B-A   0.8333333  -3.866075  5.532742 0.9951810
C-A -12.4166667 -17.116075 -7.717258 0.0000000
D-A  -9.5833333 -14.282742 -4.883925 0.0000014
E-A -11.0000000 -15.699409 -6.300591 0.0000000
F-A   2.1666667  -2.532742  6.866075 0.7542147
C-B -13.2500000 -17.949409 -8.550591 0.0000000
D-B -10.4166667 -15.116075 -5.717258 0.0000002
E-B -11.8333333 -16.532742 -7.133925 0.0000000
F-B   1.3333333  -3.366075  6.032742 0.9603075
D-C   2.8333333  -1.866075  7.532742 0.4920707
E-C   1.4166667  -3.282742  6.116075 0.9488669
F-C  14.5833333   9.883925 19.282742 0.0000000
E-D  -1.4166667  -6.116075  3.282742 0.9488669
F-D  11.7500000   7.050591 16.449409 0.0000000
F-E  13.1666667   8.467258 17.866075 0.0000000
```

Hier sieht man, welche paarweisen Vergleiche signifikant sind, wenn man die letzte Spalte betrachtet. Zu beachten ist hier, dass *gleichzeitig* insgesamt 15 Tests zum Signifikanzniveau 5% durchgeführt werden würden, so dass mit mindestens einem signifikanten Ergebnis zu rechnen ist, auch wenn H_0 zutrifft. R reagiert darauf, indem das Signifikanzniveau angepasst ist. Auf dieses Thema werden wir später zurückkommen.

Bemerkung 2.7 (Ungleiche Varianzen). Neben der Varianzanalyse von oben gibt es noch die Möglichkeit, in R eine Varianzanalyse ohne die Annahme der gleichen Varianzen durchzuführen.

```
> oneway.test(count~spray)

One-way analysis of means (not assuming equal variances)

data:  count and spray
F = 36.0654, num df = 5.000, denom df = 30.043, p-value = 7.999e-12
```

Hier wird die Anzahl der Freiheitsgrade von SQR an die unterschiedlichen Varianzen angepasst.

Bemerkung 2.8 (Zwei-faktorielle Varianzanalyse). Gibt es nicht nur einen Faktor, sondern zwei, hilft die Zwei-faktorielle Varianzanalyse weiter. Hinter dieser steckt das Modell

$$Y_{k\ell i} = \beta_{k\bullet} + \beta_{\bullet\ell} + \epsilon_{k\ell i},$$

wobei $\beta_{k\bullet}$ den Effekt des Levels k für den ersten Faktor und $\beta_{\bullet\ell}$ den Effekt des Levels ℓ für den zweiten Faktor beschreibt. Ähnliche Tests wie oben können auch für eine zwei-faktorielle Varianzanalyse durchgeführt werden.

3 Überprüfen von Modellannahmen

Sowohl bei der Regression, als auch bei der Varianzanalyse, haben wir angenommen, dass verschiedene Stichproben dieselbe Varianz aufweisen, oder sogar alle normalverteilt mit den gleichen Varianzen sind. Um Fehlinterpretationen der statistischen Verfahren auszuschließen, sollte man diese Annahmen überprüfen. Einige Tests, die hierfür zur Verfügung stehen, wollen wir hier vorstellen.

3.1 Gleichheit von Varianzen...

...bei zwei Stichproben

Seien $X_1, ..., X_m$ unabhängig und nach $N(\mu_X, \sigma_X^2)$ verteilt, sowie $Y_1, ..., Y_n$ unabhängig und nach $N(\mu_Y, \sigma_Y^2)$ verteilt. Wir wollen die Hypothese $H_0 : \sigma_X^2 = \sigma_Y^2$ testen. Glücklicherweise ist dies einfach zu bewerkstelligen, da die empirischen Varianzen unabhängig sind und verteilt sind nach $(m-1)s^2(X)/\sigma_X^2 \sim \chi^2_{m-1}$ und $(n-1)s^2(Y)/\sigma_Y^2 \sim \chi^2_{n-1}$. Daraus ergibt sich bereits der F-Test auf ungleiche Varianzen

F-Test auf gleiche Varianzen

Annahme	$X_1, ..., X_m \sim N(\mu_X, \sigma_X^2), Y_1, ..., Y_n \sim N(\mu_Y, \sigma_Y^2)$
Hypothese	$H_0 : \sigma_X^2 = \sigma_Y^2 \quad$ gegen $\quad H_1 : \sigma_X^2 \neq \sigma_Y^2$
Teststatistik	$F = \dfrac{s^2(X)}{s^2(Y)} \sim F(m-1, n-1)$
Ablehnungsbereich	$F \in (-\infty, F_{m-1,n-1,\alpha/2}) \cup (F_{m-1,n-1,1-\alpha/2}, \infty)$
p-Wert	$2(1 - P_{F(m-1,n-1)}(F)) \wedge 2(1 - P_{F(n-1,m-1)}(1/F))$

Beispiel 3.1 (Verletzung der Modellannahmen). Der F-Test testet auf Gleichheit zweier Varianzen von normalverteilten Stichproben. Damit lautet

$$H_0 : X \text{ und } Y \text{ sind normalverteilt mit } \sigma_X^2 = \sigma_Y^2.$$

Insbesondere steckt bereits in H_0 die Annahme der Normalverteilung der Daten. Wird also die Nullhypothese verworfen werden, so kann dies bedeuten, dass die Normalverteilungsannahme nicht stimmt. Als Veranschaulichung nehmen wir exponentialverteilte Daten und vergleichen deren Varianz mit normalverteilten Daten:

```
> x<-rexp(100)
> y<-rnorm(100)
> var.test(x,y)

F test to compare two variances
```

```
data:  x and y
F = 1.5575, num df = 99, denom df = 99, p-value = 0.02854
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.047926 2.314755
sample estimates:
ratio of variances
          1.557463
```

Obwohl die Varianzen der beiden Stichproben gleich sind, wird also H_0 aufgrund der unterschiedlichen Verteilung von X und Y auf dem Niveau von 5% abgelehnt.

Levene- und Brown-Forsythe-Test

Annahme	$(X_{ki})_{k=1,...,p,i=1,...,n_k}$ unabhängig, $(X_{ki})_{i=1,...,n_k}$ identisch verteilt, $k = 1, ..., p$
Hypothese	$H_0 : \mathbb{V}[X_{k1}] = \mathbb{V}[X_{\ell 1}], k, \ell = 1, ..., p$ gegen $H_1 : \mathbb{V}[X_{k1}] \neq \mathbb{V}[X_{\ell 1}]$ für ein Paar k, ℓ
Teststatistik	$W = \frac{\sum_{k=1}^{p} n_k(\overline{Z}_{k\bullet} - \overline{Z})^2/(p-1)}{\sum_{k=1}^{p}\sum_{i=1}^{n_k}(Z_{ki} - \overline{Z}_{k\bullet})^2/(n-p)} \overset{\text{approx}}{\sim} F(p-1, n-p)$ $Z_{ki} = \lvert X_{ki} - \overline{X}_{k\bullet}\rvert$, wobei $\overline{X}_{k\bullet} = \begin{cases} \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki}, & \text{Levene-Test} \\ \text{Median von } (X_{ki})_{i=1,...,n_k}, & \text{Brown-Forsythe-Test} \end{cases}$ $\overline{Z}_{k\bullet} := \frac{1}{n_k}\sum_{i=1}^{n_k} Z_{ki}, \overline{Z} := \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} Z_{ki}$
Ablehnungsbereich	$F \in (F_{p-1,n-p,1-\alpha}, \infty)$
p-Wert	$1 - P_{F(p-1,n-p)}(F)$

...bei k Stichproben

Liegen nicht zwei, sonder k Stichproben vor (etwa bei einer Varianzanalyse), könnten paarweise F-Tests Aufschluss über die Gleichheit der Varianzen geben, aber es gibt auch Alternativen. Oft verwendet werden hier der Levene-Test und der Brown–Forsythe-Test. Diese beschreiben wir lediglich, ohne auf genaue Eigenschaften einzugehen.

Beispiel 3.2. Wir verwenden dieselben simulierten Daten wie in Beispiel 3.1. Hier wird nun die Hypothese der gleichen Varianzen nicht verworfen. Beim Levene-Test handelt es sich um einen Test, der robuster ist gegen die Verletzung der Modellannahme der Normalverteilung.

```
> library(lawstat)
> x<-rexp(100)
> y<-rnorm(100)
> data = c(x,y)
> group = c(rep(1,100), rep(2, 100))
> levene.test(data, group)

modified robust Brown-Forsythe Levene-type test based on the absolute
deviations from the median

data:  data
Test Statistic = 0.0306, p-value = 0.8613
```

3.2 Testen der Normalverteilungsannahme

Sowohl beim t-Test, χ^2-Test, als auch bei der Regression und der Varianzanalyse haben wir die Annahme gemacht, dass die Daten normalverteilt sind. Diese Annahme lässt sich auch testen. Verfahren hierzu werden wir nun besprechen.

QQ-Plots

Eine einfache grafische Möglichkeit, sich einen Eindruck zu verschaffen, ob ein Datensatz von reellwertigen Beobachtungen einer bestimmten Verteilung folgt, sind Plots der Quantile oder QQ-Plots. Hier werden die Quantile der empirischen Verteilung gegen Quantile der zu überprüfenden Verteilung geplottet. Etwa ist das 5%-Quantil der empirischen Verteilung der (oder ein) $y \in \mathbb{R}$, so dass unterhalb von y genau 5% aller Datenpunkte zu finden sind.

In R sind solche QQ-Plots einfach zu bekommen. Hierzu verwenden wir den Datensatz `precip`, der die Niederschlagsmenge (in Zoll) für 70 Städte der USA (und Puerto Rico) angibt.

```
> head(precip)
    Mobile      Juneau     Phoenix Little Rock Los Angeles  Sacramento
      67.0        54.7         7.0        48.5        14.0        17.2
```

Für den QQ-Plot gibt es den Befehl

```
> qqnorm(precip),
```

der die empirischen Quantile gegen die einer Standardnormalverteilung plottet; siehe Abbildung 3.1.

Der Kolmogorov-Smirnov-Test

Natürlich ist es gut, nicht nur einen grafischen Eindruck der möglichen Abweichung der Normalverteilungsannahme zu haben, sondern auch einen statistischen Test. Mit dem hier vorgestellten Kolmogorov-Smirnov-Test kann man testen, ob Daten einer beliebigen, vorgegebenen, stetigen Verteilung folgen. Er basiert auf der empirischen Verteilung der Stichprobe.

Definition 3.3 (Empirische Verteilung). *Sei $X = (X_1, ..., X_n)$ ein Vektor von Zufallsgrößen. Die empirische Verteilung von X ist gegeben als*

$$\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

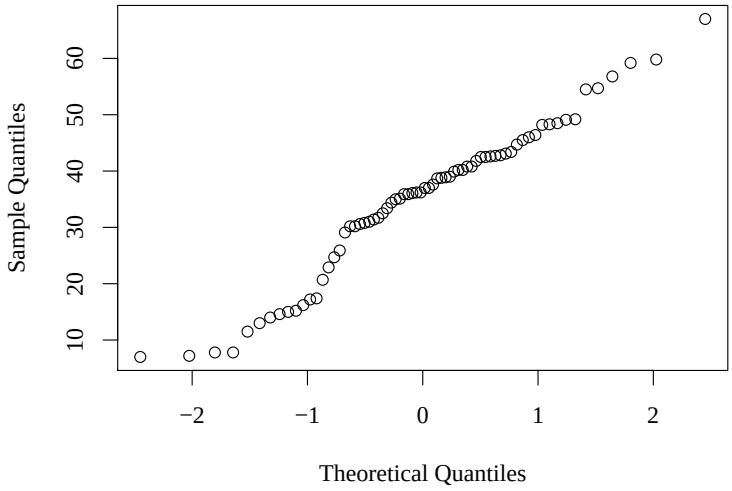


Abbildung 3.1: QQ-Plot der `precip`-Daten.

Sind die Zufallsvariablen reellwertig, dann ist die empirische Verteilnugsfunktion die Verteilungsfunktion der empirischen Verteilung und gegeben als

$$t \mapsto S_n(t) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(-\infty; t] = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq t}.$$

Bemerkung 3.4 (Satz von Glivenko-Cantelli). Aus der Vorlesung Wahrscheinlichkeitstheorie ist bekannt: Sind $X, X_1, X_2, ...$ unabhängige und identisch verteilte Zufallsgrößen mit Verteilungsfunktion F_X. Dann gilt

$$D_n := \sup_{t \in \mathbb{R}} |S_n(t) - F_X(t)| \xrightarrow{n\to\infty}_{fs} 0.$$

Um dies einzusehen, sei bemerkt, dass $1_{X_1 \leq t}, 1_{X_2 \leq t}, ...$ unabhängig und identisch verteilt sind mit $\mathbb{E}[1_{X_1 \leq t}] = \mathbb{P}(X_1 \leq t) = F_X(t)$. Damit ist mit dem Gesetz der großen Zahlen zumindest erklärt, warum $S_n(t) - F_X(t) \xrightarrow{n\to\infty}_{fs} 0$ für jedes feste t gilt.

Bemerkung 3.5 (Verteilung von $F_X(X_{(i)})$). Sei X eine Zufallsvariable mit Dichte und habe Verteilungsfunktion F_X.

1. Es ist $F_X(X) \sim U[0, 1]$.
   Denn: Fast sicher ist X so, dass $F_X^{-1}(X)$ existiert. Daraus folgt

   $$\mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t)) = F_X(F_X^{-1}(t) = t.$$

2. Seien $X, X_1, ..., X_n$ unabhängig und identisch verteilt und $U_1, ..., U_n \sim U([0, 1])$ unabhängig. Dann gilt $F_X(X_{(i)}) \sim U_{(i)}$.
   Denn: Genau wie oben ist $X_{(i)}$ fast sicher so, dass $F_X^{-1}(X_{(i)})$ existiert. Nun ist

   $$\begin{aligned}\mathbb{P}(F_X(X_{(i)}) \leq t) &= \mathbb{P}(X_{(i)} \leq F_X^{-1}(t)) = \mathbb{P}(X_j \leq F_X^{-1}(t) \text{ für } i \text{ verschiedene } j) \\ &= \mathbb{P}(U_j \leq t \text{ für } i \text{ verschiedene } j) = \mathbb{P}(U_{(i)} \leq t).\end{aligned}$$

Proposition 3.6 (Verteilungsfreiheit von D_n). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein reguläres, stetiges statistisches Modell. Dann ist für jedes $t \in \mathbb{R}$ die Statistik $D_n(t)$ verteilungsfrei.*

Beweis. Seien $X_{(1)}, ..., X_{(n)}$ die Ordnungsstatistiken von $X_1, ..., X_n$ sowie $X_{(0)} := -\infty$ und $X_{(n+1)} := \infty$. Dann ist

$$S_n(t) = \frac{i}{n} \text{ für } X_{(i)} \leq t < X_{(i+1)}.$$

Wir schreiben nun

$$\begin{aligned} D_n &= \sup_{t\in\mathbb{R}} |S_n(t) - F_X(t)| = \max_{1\leq i\leq n} \sup_{X_{(i)}\leq t<X_{(i+1)}} |S_n(t) - F_X(t)| \\ &= \max_{1\leq i\leq n} \sup_{X_{(i)}\leq t<X_{(i+1)}} \Big|\frac{i}{n} - F_X(t)\Big| \\ &= \max_{1\leq i\leq n} \max\Big(\Big|\frac{i}{n} - F_X(X_{(i)})\Big|, \Big|\frac{i}{n} - F_X(X_{(i+1)})\Big|\Big). \end{aligned}$$

Damit ist gezeigt, das D_n nur von $F_X(X_{(0)}), ..., F_X(X_{(n+1)})$ abhängt. Diese Größen haben nach Bemerkung 3.5 dieselbe Verteilung wie die Ordnungsstatistiken eines $U(0,1)$-verteilten Vektors von Zufallsvariablen, und zwar unabhängig von F_X. Daraus folgt die Behauptung. □

Kolmogorov-Smirnov-Test

Annahme	$X_1, ..., X_n$ reellwertig, unabhängig und stetig identisch verteilt
Hypothese	$H_0 : X_i$ hat Verteilungsfunktion F_X gegen $H_1 : X_i$ hat eine andere Verteilungsfunktion
Teststatistik	$D_n := \sup_{t\in\mathbb{R}} \lvert S_n(t) - F_X(t)\rvert$ $S_n(t) := \frac{1}{n}\sum_{i=1}^n 1_{X_i\leq t}$ Verteilung $(D_n)_*\mathbb{P}$ von D_n ist in Theorem 3.7 angegeben
Ablehnungsbereich	$D_n >$ $(1-\alpha)$-Quantil von $(D_n)_*\mathbb{P}$
p-Wert	$(D_n)_*\mathbb{P}((D_n, \infty))$

Theorem 3.7 (Verteilung von D_n). *Sei $X, X_1, .., X_n$ unabhängig und identisch verteilt mit Dichte sowie F_X die Verteilungsfunktion von X. Dann gilt für $0 < s < (2n-1)/(2n)$*

$$\mathbb{P}\Big(D_n < \frac{1}{2n} + s\Big) = n! \int_{1/(2n)-s}^{1/(2n)+s} \int_{3/(2n)-s}^{3/(2n)+s} \cdots \int_{(2n-1)/(2n)-s}^{(2n-1)/(2n)+s} 1_{0<u_1\cdots<u_n<1} du_n \cdots du_1.$$

Beweis. Zunächst bemerken wir, dass immer $D_n \geq 1/2n$ gilt, da F_X stetig ist, S_n aber Sprünge der Größe $1/n$ macht. OBdA nehmen wir wegen der Verteilungsfreiheit von D_n an,

dass $F_X(x) = x$, d.h. $X \sim U([0,1])$. Wir schreiben mit $s' := \frac{1}{2n} + s$

$$\begin{aligned}
\mathbb{P}(D_n < s') &= \mathbb{P}(\sup_{t\in[0,1]} |S_n(t) - t| < s') \\
&= \mathbb{P}\Big(\Big|\frac{i}{n} - t\Big| < s' \text{ für alle } X_{(i)} \le t < X_{(i+1)}, \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{i}{n} - s' < t < \frac{i}{n} + s' \text{ für alle } X_{(i)} \le t < X_{(i+1)}, \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i}{n} + s', \frac{i}{n} - s' < X_{(i+1)} < \frac{i}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i}{n} + s', \frac{i-1}{n} - s' < X_{(i)} < \frac{i-1}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i-1}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{2i-1}{2n} - s < X_{(i)} < \frac{2i-1}{2n} + s \text{ für alle } i = 1, ..., n\Big).
\end{aligned}$$

Daraus folgt die Behauptung, da die gemeinsame Verteilung von $X_{(1)}, ..., X_{(n)}$ die Dichte $n! 1_{0\le u_1<\cdots<u_n}$ hat. □

Beispiel 3.8 (Der Kolmogorov-Smirnov-Test für t-verteilte Daten). Es ist bekannt, dass die t-Verteilung mit k Freiheitsgraden für große k gegen $N(0,1)$ konvergiert. Wir wollen nun testen, ob der Unterschied der t-Verteilung mit $k = 10$ Freiheitsgraden und der $N(0,1)$-Verteilung erkennbar ist. Wir verwenden hierzu verschiedene Stichprobengrößen. Es ergibt etwa

```
> data = rt(1000, df=10)
> ks.test(data, "pnorm")
One-sample Kolmogorov-Smirnov test

data:  data
D = 0.0348, p-value = 0.178
alternative hypothesis: two-sided
```

also kann in dieser Stichprobe der Größe 1000 die Normalverteilungsannahme nicht verworfen werden. In einer deutlich größeren Stichprobe allerdings schon, wie wir nun sehen.

```
> data = rt(10000, df=10)
> ks.test(data, "pnorm")
One-sample Kolmogorov-Smirnov test

data:  data
D = 0.0207, p-value = 0.0003925
alternative hypothesis: two-sided
```

Der Lilliefour-Test

Will man prüfen, ob ein Datensatz einer Normalverteilung folgt, so kennt man zunächst die Parameter μ und σ^2 nicht. Deshalb ist es nicht möglich, den Kolmogorov-Smirnov-Test direkt anzuwenden, da man nicht weiß, gegen welche Verteilung genau getestet werden soll. Es liegt

nun nahe, zunächst μ und σ^2 etwa durch $\overline{x}$ und $s^2(x)$ aus den Daten zu testen und anschließend die Normalverteilungsannahme dadurch zu überprüfen, ob die Daten x einer $N(\overline{x}, s^2(x))$-Verteilung folgen. Allerdings verändert sich durch das Schätzen der Modellparameter aus den Daten die Verteilung der Teststatistik D_n. Die neue Verteilung von D_n kann man mittels Simulation ermitteln.

4 Nicht-parametrische Statistik

Die statistischen Verfahren, die wir bisher kennengelernt haben, basieren auf statistischen Modellen, die immer eine bestimmte Klasse von Verteilungen voraussetzen; man erinnere sich beispielsweise an die Normalverteilungsannahme bei der linearen Regression. Ist eine solche Annahme nicht gerechtfertigt oder verletzt, so greift man auf nicht-parametrische Verfahren zurück. Das statistische Modell ist hier viel flexibler, so dass unter sehr wenigen Grundannahmen Aussagen getroffen werden können. Formal ist es so: Die Parametermenge $\mathcal{P}$ eines statistischen Modells $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ist oftmals eine Teilmenge eines $\mathbb{R}^k$, etwa beim Normalverteilungsmodell $(X, \{\mathbb{P}_{\theta=(\mu,\sigma^2)} = N(\mu, \sigma^2)^n : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\})$. Ist diese Annahme zu restriktiv, so müssen wir $\mathcal{P}$ als viel größere Menge annehmen, so dass $\mathcal{P}$ keine Teilmenge eines $\mathbb{R}^k$ mehr ist. In genau diesem Fall spricht man von nicht-parametrischer Statistik. Etwa könnte $\mathcal{P} = \{\theta : \mathbb{R} \to \mathbb{R}_+ \text{ Dichte bzgl } \lambda\}$ die Menge der regulären, stetigen Modelle (mit $E = \mathbb{R}$) bezeichnen oder $\mathcal{P} = \{\theta : \mathbb{R} \to \mathbb{R}_+ \text{ Dichte bzgl } \lambda \text{ mit } \theta(m + x) = \theta(m - x) \text{ für ein m}\}$ die Menge der bezüglich $m \in \mathbb{R}$ symmetrischen regulären, stetigen Modelle. Wir wollen in diesem Abschnitt statistische Verfahren mit solchen *großen* Parametermengen $\mathcal{P}$ angeben.

4.1 Quantil-Tests

Wir beginnen mit dem einfachen Beispiel eines Tests auf ein Quantil. Wir verwenden das statistische Modell $(X, \{\mathbb{P}_\theta^n : \theta \in \mathcal{P}\})$ mit $\mathcal{P} = \{\theta : \mathbb{R} \to \mathbb{R}_+ \text{ Dichte bzgl } \lambda\}$ der stetigen, regulären Modelle. Wir bezeichnen mit $\kappa_{\theta,p}$ das p-Quantil von $\mathbb{P}_\theta$. Laut Definition gilt

$$\mathbb{P}_\theta(X_1 \leq \kappa_{\theta,p}) = p,$$

außerdem ist $\sum_{i=1}^n 1_{X_i \leq \kappa_{\theta,p}} \sim B(n, p)$. Daraus lässt sich bereits ein Test auf ein vorgegebenes Quantil ableiten.

Beispiel 4.1 (Schlafdauern). Wir erinnern an das Datenbeispiel aus dem t-Test. Ein Medikament wir daraufhin untersucht, ob es den Schlaf von Probanden verlängert. Dazu wird jeweils die Schlafdauerdifferenz bei zehn Patienten notiert. Man erhält

$$1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4.$$

Wir wollen nun testen, ob der Median (das 50%-Quantil) 0 ist oder nicht.

```
> a<-c(1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4)
> length(a)
[1] 10
> sum(a>0)
[1] 9
> binom.test(c(9,1), 0.5)

Exact binomial test

data:  c(9, 1)
number of successes = 9, number of trials = 10, p-value = 0.02148
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
```

```
 0.5549839 0.9974714
sample estimates:
probability of success
                   0.9
```

Vorzeichentest auf ein Quantil

Annahme	$X_1, ..., X_n$ unabhängig, verteilt nach einer Verteilung $\mathbb{P}_\theta = \theta \cdot \lambda$
Hypothese	$H_0 : \kappa_{\theta,p} = \kappa^*$ für ein vorgegebenes κ^* gegen $H_1 : \kappa_{\theta,p} \neq \kappa^*$
Teststatistik	$Q := \sum_{i=1}^{n} 1_{X_i \leq \kappa^*} \sim B(n,p)$ unter H_0
Ablehnungsbereich	$\{0, \ldots, k, l, \ldots, n\}$ mit $B(n,p)(1,...,k), B(n,p)(l,...,n) \leq \alpha/2$
p-Wert	$B(n,p)(1,...,Q' \wedge Q, Q \vee Q', ..., n)$ mit $Q' = 2np - Q$

4.2 Tests auf Zufälligkeit

In einer Warteschlange stehen 6 Frauen und 5 Männer, etwa in der Reihenfolge `F`, `M`, `M`, `F`, `M`, `M`, `M`, `F`, `F`, `F`, `F`. Ist diese Folge eine *zufällige* Folge?

Um diese Frage zunächst zu formalisieren, sei $E = \{x \in \{0,1\}^n : x_1 + \cdots + x_n = n_1\}$ und $n_0 := n - n_1$. Weiter bezeichne für $x \in E$

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

die Anzahl der *Runs* in x. Etwa ist $r(0,1,1,0,1,1,1,0,0,0,0) = 5$. Außerdem bezeichne $\mathbb{P}$ die Gleichverteilung auf E.

Theorem 4.2 (Verteilung der Anzahl der Runs unter Zufälligkeit). *Es gilt für $X \sim \mathbb{P}$ und $R = r(X)$*

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

Beweis. Sei zunächst r gerade. Dann gibt es genau $r/2$ Runs mit 0 und $r/2$ runs mit 1. Sehen wir uns zunächst die $r/2$ Runs mit 0 an. Es gibt insgesamt $\binom{n_0-1}{r/2-1}$ Möglichkeiten, die n_0 möglichen 0er auf $r/2$ verschiedene Runs (der Länge ≥ 1) zu verteilen. (Denn: Jede solche Möglichkeit lässt sich als Reihung, etwa 0|000|0|...|0 mit genau $r/2 - 1$ mal | und n_0 mal 0

aufschreiben. Da zwischen zwei | mindestens eine 0 stehen muss, gibt es eine Bijektion dieser Reihungen auf die Darstellungen |00||...|, bei der zwischen zwei | (und vor der ersten und nach der letzten) eine 0 entfernt wurde. Die Anzahl dieser Möglichkeiten ist nun gegeben, wenn man die Möglichkeiten abzählt, $r/2-1$ mal | auf insgesamt $r/2-1+n_0-r/2 = n_0-1$ Stellen zu verteilen. Dies ist bekanntlich $\binom{n_0-1}{r/2-1}$.) Die gesuchte Wahrscheinlichkeit ergibt sich nun aus dem Quotienten der Anzahl der Möglichkeiten, $r/2$ Runs mit 0 und $r/2$ Runs mit 1 zu erhalten, und der Gesamtzahl an Möglichkeiten, n_0 mal 0 auf insgesamt n_0+n_1 Plätze aufzuteilen. Der Vorfaktor 2 entsteht dadurch, dass entweder mit 0 oder mit 1 begonnen werden kann.

Für r ungerade bemerken wir, dass entweder $(r+1)/2$ Runs mit 0 und $(r-1)/2$ Runs mit 1 oder umgekehrt vorliegen, wobei die Folge immer mit der Ziffer begonnen werden muss, von der mehr Runs vorhanden sind. Dieselben kombinatorischen Überlegungen wie oben führen auf das Ergebnis. Man beachte hierbei $(r+1)/2-1 = (r-1)/2$ und $(r-1)/2-1 = (r-3)/2$. □

Proposition 4.3 (Erwartungswert und Varianz von R). *Es gilt, falls $n_0 \to \infty, n_1 \to \infty$ und so, dass $n_0/n \to p, n_1/n \to q := 1-p$*

$$\frac{1}{n}\mathbb{E}[R] \xrightarrow{n\to\infty} 2pq,$$
$$\frac{1}{n}\mathbb{V}[R] \xrightarrow{n\to\infty} 4p^2q^2.$$

Beweis. Wir berechnen zunächst für $i,j = 2,...,n$ mit $j > i$

$$\mathbb{E}[1_{X_i\neq X_{i-1}}] = \frac{n_0}{n}\frac{n_1}{n-1} + \frac{n_1}{n}\frac{n_0}{n-1} = 2\frac{n_0}{n}\frac{n_1}{n-1} = 2pq + O(1/n),$$
$$\mathbb{E}[1_{X_i\neq X_{i-1}}1_{X_j\neq X_{j-1}}] = \begin{cases} \frac{n_0n_1(n_0-1)+n_1n_0(n_1-1)}{n(n-1)(n-2)} = \frac{n_0n_1}{n(n-1)}, & j = i+1, \\ 4\frac{n_0n_1(n_0-1)(n_1-1)}{n(n-1)(n-2)(n-3)}, & j > i+1. \end{cases}$$

Damit sehen wir, dass

$$\mathbb{V}[1_{X_i\neq X_{i-1}}] = \mathbb{E}[1_{X_i\neq X_{i-1}}] - \mathbb{E}[1_{X_i\neq X_{i-1}}]^2 = 2pq(1-2pq) + O(1/n)$$

und für $j = i+1$

$$\begin{aligned}\mathbb{COV}[1_{X_i\neq X_{i-1}}, 1_{X_j\neq X_{j-1}}] &= \frac{n_0n_1}{n(n-1)} - 4\frac{n_0^2n_1^2}{n^2(n-1)^2} \\ &= \frac{n_0n_1}{n(n-1)}\Big(1 - 4\frac{n_0n_1}{n(n-1)}\Big) = pq(1-4pq) + O(1/n)\end{aligned}$$

sowie für $j > i+1$

$$\begin{aligned}
\tfrac{1}{4}\mathbb{COV}[1_{X_i\neq X_{i-1}}, 1_{X_j\neq X_{j-1}}] &= \frac{n_0n_1(n_0-1)(n_1-1)}{n(n-1)(n-2)(n-3)} - \frac{n_0^2n_1^2}{n^2(n-1)^2} \\
&= \frac{n_0n_1}{n(n-1)}\Big(\frac{(n_0-1)(n_1-1)}{(n-2)(n-3)} - \frac{n_0n_1}{n(n-1)}\Big) \\
&= \frac{n_0n_1}{n(n-1)}\frac{n(n-1)(n_0-1)(n_1-1) - n_0n_1(n-2)(n-3)}{n(n-1)(n-2)(n-3)} \\
&= \frac{n_0n_1}{n(n-1)}\frac{-nn_0n_1 - n^2n_0 - n^2n_1 + 5n_0n_1n + O(n^2)}{n(n-1)(n-2)(n-3)} \\
&= \frac{1}{n}pq(4pq-p-q) + O(1/n^2) \\
&= -\frac{1}{n}pq(1-4pq) + O(1/n^2).
\end{aligned}$$

Daraus ergibt sich für die Varianz

$$\begin{aligned}
\mathbb{V}[R] &= n\mathbb{V}[1_{X_2\neq X_1}] + 2n\mathbb{COV}[1_{X_2\neq X_1}, 1_{X_3\neq X_2}] + n^2\mathbb{COV}[1_{X_2\neq X_1}, 1_{X_4\neq X_3}] + O(1) \\
&= n(2pq(1-2pq) + 2pq(1-4pq) - 4pq(1-4pq)) + O(1) = 4np^2q^2 + O(1)
\end{aligned}$$

□

Bemerkung 4.4 (R approximativ normalverteilt). Zwar sind die Zufallsvariablen $1_{X_i\neq X_{i-1}}, i = 2, ..., n$ nicht unabhängig, jedoch kann man für R doch einen zentralen Grenzwertsatz angeben. Genauer ist (für große n) die Statistik

$$\frac{R - 2npq}{2\sqrt{n}pq}$$

approximativ $N(0,1)$-verteilt.

Tests auf die Anzahl von Runs in einer zufälligen Folge

Annahme	$X_1, ..., X_n \in \{0,1\}$ mit $X_1 + \cdots + X_n = n_1$
Hypothese	$H_0 : X$ rein zufällig gegen $H_1 : X$ nicht rein zufällig
Teststatistik	$R = 1 + \sum_{i=2}^{n} 1_{X_i\neq X_{i-1}}$ unter H_0 verteilt wie in Theorem 4.2, approximativ wie in Bemerkung 4.4.
Ablehnungsbereich	ergibt sich aus der Verteilung von R
p-Wert	ergibt sich aus der Verteilung von R

Beispiel 4.5 (Zufälligkeit von Zufallszahlgeneratoren). Ein linearer Kongruenzgenerator für Pseudo-Zufallszahlen ist bekanntermaßen gegeben durch die Rekursionsvorschrift (mit einem Startwert $x_0 \in \{0, ..., m-1\}$)

$$x_i = ax_{i-1} + b \mod m.$$

Typischerweise ist hier $m = 2^e$ für eine implementierte Wortlänge e. Eine R-Implemetierung könnte also etwa sein (siehe auch POSIX.1-2001)

```
> myrand<-function(n, seed=1) {
    res<-rep(seed,n)
    for(i in 2:n) {
      res[i] = (res[i-1] * 1103515245 + 12345) %% 32768;
    }
    res/32768
}
```

Wir wollen nun sehen, ob eine so generierte Folge dem Test auf Zufälligkeit standhält. Wir laden zunächst das entsprechende R-Paket.

```
> install.package("randtests")
> library("randtests")
```

In einer Stichprobe der Größe 10000 kann die Zufälligkeit nicht verworfen werden.

```
> x<-myrand(10000)
> runs.test(x)

Runs Test

data:  x
statistic = 1.3544, runs = 5067, n1 = 5092, n2 = 4908, n = 10000,
p-value = 0.1756
alternative hypothesis: nonrandomness
```

4.3 Der Wald-Wolfowitz-Runs-Test

Wir wenden uns nun – im Gegensatz zur Situation in Abschnitt 4.1 – Tests mit zwei unabhängigen Stichproben zu. Insbesondere geben wir nun eine nicht-parametrische Alternative zum doppelte t-Test an. Hierzu sei $X_1, ..., X_m$ unabhängig und identisch nach $\mathbb{P}_\theta = \theta \cdot \lambda$ und $Y_1, ..., Y_n$ unabhängig und identisch nach $\mathbb{P}_{\theta'} = \theta' \cdot \lambda$ verteilt. Ziel ist es, den Test $H_0 : \theta = \theta'$ zu testen. Seien hierzu $X_{(1)}, ..., X_{(m)}$ und $Y_{(1)}, ..., Y_{(n)}$ die Ordnungsstatistiken von X und Y. Weiter sei $Z = (X, Y)$ und $Z_{(1)}, ..., Z_{(m+n)}$ die Ordnungsstatistiken der gemeinsamen Stichprobe $X_1, ..., X_m, Y_1, ..., Y_n$. Im weiteren verwenden wir den Vektor

$$W := (1_{Z_{(1)} \in \{X_1, ..., X_m\}}, ..., 1_{Z_{(m+n)} \in \{X_1, ..., X_m\}}).$$

Unter H_0 ist W ein rein zufälliger Vektor in $\{0, 1\}^{m+n}$ mit genau n mal 0 und m-mal 1. Die Verteilung der Anzahl von Runs in W haben wir also im letzten Kapitel hergeleitet. Einzig für die Berechnung des Ablehnungsbereiches bemerken wir, dass H_0 nur dann abgelehnt wird, wenn die Anzahl der Runs zu klein ist. (Etwa seien alle X_i kleiner als alle Y_j. Dann ist $W = 1, ..., 1, 0, ..., 0$ und die Anzahl der Runs ist 2.)

Beispiel 4.6 (Der Runs-Test mit t-verteilten Daten). Schon beim Überprüfen von Modellannahmen haben wir untersucht, welche t-Verteilungen von einer Normalverteilung zu unterscheiden sind. Dies wollen wir nochmal vertiefen, indem wir den Runs-Test auf einen Datensatz t- und einen Datensatz normalverteilter Daten anwenden. Wir verwenden hier 10 Freiheitsfrage für die t-Verteilung.

```
> set.seed(1)
> x<-rnorm(100)
> y<-rt(100, df=10)
> perm<-sort(c(x,y), index.return=TRUE)$ix
> w<-as.numeric(perm<=100)
> runs.test(w)

Runs Test

data:  w
statistic = -0.8507, runs = 95, n1 = 100, n2 = 100, n = 200, p-value =
0.395
alternative hypothesis: nonrandomness
```

Der Wald-Wolfowitz-Runs-Test

Annahme	$X_1, ..., X_m$ unabhängig, verteilt nach einer Verteilung $\mathbb{P}_\theta = \theta \cdot \lambda$ $Y_1, ..., Y_n$ unabhängig, verteilt nach einer Verteilung $\mathbb{P}_{\theta'} = \theta' \cdot \lambda$
Hypothese	$H_0 : \theta = \theta'$ gegen $H_1 : \theta \neq \theta'$
Teststatistik	$R := r(W)$, unter H_0 verteilt nach Theorem 4.2, approximativ with in Bemerkung 4.4 mit $Z = (X, Y)$ und $W := (1_{Z_{(1)} \in \{X_1,...,X_m\}}, ..., 1_{Z_{(m+n)} \in \{X_1,...,X_m\}})$.
Ablehnungsbereich	ergibt sich aus der Verteilung von R
p-Wert	ergibt sich aus der Verteilung von R

4.4 Der Kruskal-Wallis-Test

Nachdem wir nun eine nicht-parametrische Version des doppelten t-Tests kennengelernt haben, kommt nun eine nicht-parametrische Version der einfaktoriellen Varianzanalyse. Wir erinnern daran, dass hierfür Y_{ki} die i-te Messung der k-ten Gruppe ist, wobei wir die Gleichheit der Verteilungen von p Gruppen testen wollen. Etwas genauer seien hier $Y_{k\bullet} = Y_{k1}, ..., Y_{kn_k}$ unabhängig und nach $\mathbb{P}_{\theta_k} \sim \theta_k \cdot \lambda$ verteilt, $k = 1, ..., p$. Wie im Wald-Wolfowitz-Test definieren wir $Y_{\bullet\bullet} = (Y_{ki})_{k=1,...,p, i=1,...,n_k}$ und $Z = Y_1, ..., Y_n$ die als Vektor geschriebenen Daten $Y_{\bullet\bullet}$. Für

die Ordnungsstatistiken $Z_{(1)}, ..., Z_{(n)}$ verwenden wir den Vektor $R = (R_1, ..., R_p)$ mit

$$R_k = \sum_{i=1}^{n} i1(Z_{(i)} \in \{Y_{k1}, ..., Y_{kn_k}\}),$$

d.h. R_k ist die Summe der Ränge der Größen $Y_{k\bullet}$ in Z. Nun ist die Summe aller Ränge immer gleich

$$\sum_{k=1}^{p} R_k = \sum_{i=1}^{n} i \sum_{k=1}^{p} 1(Z_{(i)} \in \{Y_{k1}, ..., Y_{kn_k}\}) = \sum_{i=1}^{n} i = \binom{n+1}{2}.$$

Gilt außerdem $H_0 = \theta_1 = \cdots = \theta_p$, so gilt für die erwartete Summe der Ränge von Gruppe k

$$\mathbb{E}[R_k] = \sum_{i=1}^{n} i\mathbb{P}(Z_{(i)} \in \{Y_{k1}, ..., Y_{kn_k}\}) = \frac{n_k}{n}\binom{n+1}{2} = \frac{n_k(n+1)}{2}.$$

Damit können wir nun den Kruskal-Wallis-Test angeben. Allerdings ist die Verteilung der Teststatistik S (siehe unten) nur für kleine Werte von p einfach anzugeben.

Kruskal-Wallis-Test (nicht-parametrische einfaktorielle Varianzanalyse)

Annahme	$Y_{k1}, ..., Y_{kn_k}$ unabhängig, nach $\mathbb{P}_{\theta_k} = \theta_k \cdot \lambda$ verteilt, $k = 1, ..., p$
Dabei sind	
$Y_{11}, \ldots, Y_{pn_p}$	gegebene Merkmalsausprägungen eines Merkmals gemessen in Levels $1, \ldots, p$
Hypothesen	$H_0 : \theta_1 = \cdots = \theta_p$ gegen $H_1 : \theta_k \neq \theta_\ell$ für ein Paar k, ℓ
Teststatistik	$S = \sum_{k=1}^{p} \left(R_k - \frac{n_k(n+1)}{2}\right)^2$
Ablehnungsbereich	durch Verteilung von S gegeben
p-Wert	durch Verteilung von S gegeben

Beispiel 4.7. Wir verwenden dieselben normalverteilten Daten X und t-verteilten Daten Y aus dem letzten Beispiel. Nun ergibt sich

```
> a<-list(x,y)
> kruskal.test(a)
Kruskal-Wallis rank sum test

data:  a
Kruskal-Wallis chi-squared = 0.0539, df = 1, p-value = 0.8164
```

Also wird auch hier die Nullhypothese nicht verworfen. R berechnet hier nicht das S von oben, sondern eine normalisierte Version davon, wodurch die Teststatistik approximativ χ^2-verteilt ist mit $p - 1$ Freiheitsgraden.

5 Bootstrap

5.1 Aus Verteilungsschätzern abgeleitete Schätzer

Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell. Man kann immer einen Schätzer von $\mathbb{P}_\theta$, etwa die empirische Verteilung, angeben. Um $\mathbb{P}_\theta$ zu schätzen ist es auch möglich, θ durch ein $\hat{\theta}$ zu schätzen und anschließend $\mathbb{P}_\theta$ durch $\mathbb{P}_{\hat{\theta}}$. Mit einem Schätzer von $\mathbb{P}_\theta$ kann man nun Schätzer für alle $g(\mathbb{P}_\theta)$, $\theta \in \mathcal{P}$ für beliebiges, messbares g angeben, nämlich $g(\mathbb{P}_{\hat{\theta}})$. Man spricht hier auch von *Plugin-Schätzern*.

Beispiel 5.1 (Parametrische und nicht-parametrische Schätzer). 1. Sei

$$(X, \{\mathbb{P}_\theta^n : \mathbb{P}_\theta \text{ Wahrscheinlichkeitsmaß auf } \mathbb{R}\})$$

das nicht-parametrische Modell für unabhängige Daten. Dann ist

$$\mathbb{P}_{\hat{\theta}} := \frac{1}{n}\sum_{i=1}^n \delta_{X_i},$$

die empirische Verteilung von X ein Schätzer für $\mathbb{P}_\theta$. Ist nun etwa

$$g(\mathbb{P}_\theta) = \int f(x)\mathbb{P}_\theta(dx) = \mathbb{E}_\theta[f(X_1)],$$

für ein messbares f, dann ist

$$g(\mathbb{P}_{\hat{\theta}}) = \mathbb{E}_{\hat{\theta}}[f(Y_1)] = \int f(x)\mathbb{P}_{\hat{\theta}}(dx) = \frac{1}{n}\sum_{i=1}^n f(X_i)$$

ein Schätzer für $g(\mathbb{P}_\theta)$.

2. Sei $n = 2m-1$ ungerade und das statistische Modell von 1. gegeben. Weiter sei $g(\mathbb{P}_\theta)$ der Median von $\mathbb{P}_\theta$. Für $\mathbb{P}_{\hat{\theta}}$ wie oben ist damit

$$g(\mathbb{P}_{\hat{\theta}}) = g\Big(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}\Big) = X_{(m)}$$

der Stichproben-Median (wobei $X_{(1)}, ..., X_{(n)}$ die Ordnungsstatistiken von X sind).

3. Sei

$$(X, \{\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)^n : \theta = (\mu, \sigma^2) \in \mathcal{P} = \mathbb{R} \times \mathbb{R}_+\})$$

das Normalverteilungsmodell mit unbekannter Varianz. Dann ist bekanntlich $\hat{\theta} = (\overline{X}, s^2(X))$ ein (erwartungstreuer, konsistenter) Schätzer von θ. Außerdem ist $\mathbb{P}_{(\overline{X}, s^2(X))} = \mathcal{N}(\overline{X}, s^2(X))$ ein Schätzer für $\mathbb{P}_\theta$. Ist etwa

$$g(\mathbb{P}_\theta) = \mathbb{P}_\theta(\overline{X} \in A)$$

die Wahrscheinlichkeit für $\overline{X} \in A$ unter $\mathbb{P}_\theta$ (und damit eine Funktion von $\mathbb{P}_\theta$), dann ist $\overline{X}_*\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2/n)$, also[9]

$$g(\mathbb{P}_{\hat{\theta}}) = \mathbb{P}_{\overline{X}, s^2(X)/n}(\overline{Y} \in A).$$

[9]Die Notation ist hier etwas schwierig: Wir vereinbaren, dass $X \sim \mathbb{P}_\theta$ (so wie immer) und $Y \sim \mathbb{P}_{\hat{\theta}} = \mathbb{P}_{\hat{\theta}(X)}$, d.h. wir verwenden Y immer als nach der geschätzten Verteilung verteilte Zufallsvariable. Diese Unterscheidung ist deshalb wichtig, weil ja $\hat{\theta}$ von X abhängt.

5.2 Bias- und Varianzschätzung

Für einen Schätzer $g(\mathbb{P}_{\hat{\theta}})$ von $g(\mathbb{P}_\theta)$ gibt es Bias und Varianz, nämlich

$$b_{\theta,g} := \mathbb{E}_\theta[g(\mathbb{P}_{\hat{\theta}})] - g(\mathbb{P}_\theta), \qquad v_{\theta,g} := \mathbb{V}_\theta[g(\mathbb{P}_{\hat{\theta}})].$$

Da wir zunächst nichts außer den Daten zur Verfügung haben, möchten wir aus den Daten den Bias und die Varianz des Schätzers $g(\mathbb{P}_{\hat{\theta}})$ schätzen. Für einfache Funktionen g gibt es hierbei gute Möglichkeiten.

Beispiel 5.2 (Parametrische und nicht-parametrische Schätzer). Wir behandeln nun nochmal 1.-3. aus dem letzten Beispiel.

1. Da bekanntermaßen

$$\mathbb{E}_\theta[g(\mathbb{P}_{\hat{\theta}})] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta[f(X_i)] = \mathbb{E}_\theta[f(X_1)] = g(\mathbb{P}_\theta),$$

   weiß man, dass der Schätzer $g(\mathbb{P}_{\hat{\theta}})$ unverzerrt ist. Damit setzen wir

$$\hat{b}_{\theta,g} = 0$$

   als unverzerrten Schätzer des Bias. Weiter ist

$$\mathbb{V}_\theta[g(\mathbb{P}_{\hat{\theta}})] = \frac{1}{n}\mathbb{V}_\theta[f(X_1)].$$

   Um diese Varianz zu schätzen, nehmen wir wegen der Unabhängigkeit der Daten

$$\hat{v}_{\theta,g} := \frac{1}{n}s^2(f(X))$$

   als unverzerrten Schätzer für die Varianz von $g(\mathbb{P}_{\hat{\theta}})$.

2. Wir wollen nun Bias und Varianz des Stichprobenmedians $X_{(m)}$ als Schätzer für den Median von $\mathbb{P}_\theta$ herausfinden. Ist $\mathbb{P}_\theta = p_\theta \cdot \lambda$, so hat $X_{(m)}$ die Dichte

$$x \mapsto \binom{n}{m} m\mathbb{P}_\theta(X_1 \le x)^m \mathbb{P}_\theta(X_1 > x)^m p_\theta(x).$$

   Hier ist nun allerdings unklar, wie weiter zu verfahren ist. Um den Median momentenbasiert zu schätzen, müsste man $\mathbb{E}_\theta[X_{(m)}]$ berechnen, was nur für ein parametrisches Modell machbar ist. Genau dasselbe Problem besteht beim Maximum-Likelihood-Ansatz.

3. Wir berechnen, da $(\overline{X}, s^2(X)/n)$ unter $\mathbb{P}_\theta$ unabhängig sind mit $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ und $(n-1)s^2(X)/\sigma^2 \sim \chi^2_{n-1}$

$$\mathbb{E}_\theta[g(\mathbb{P}_{\hat{\theta}})] = \mathbb{E}_\theta[\mathbb{P}_{\overline{X},s^2(X)/n}(\bar{Y} \in A)] = \int\int \mathbb{P}_{(\tilde{\mu},\tilde{\sigma}^2)}(Y \in A)\mathbb{P}(s^2(X) \in d\tilde{\sigma}^2)\mathbb{P}(\overline{X} \in d\tilde{\mu})$$

   Wieder ist nun allerdings etwas unklar, wie weiter zu verfahren ist.

Bei 2. und 3. gibt es nun dieselbe Möglichkeit wie bei der Herleitung des Schätzers selbst, $b_{\theta,g}$ und $v_{\theta,g}$ zu schätzen. Man kann nämlich die Schätzung dadurch bewerkstelligen, dass man θ durch $\hat{\theta}$ ersetzt. Diese Plugin-Schätzer für Bias und Varianz eines Schätzers heißen ideale Bootstrap-Schätzer.

Definition 5.3 (Idealer Bootstrap-Schätzer). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell, $\hat{\theta}$ ein Schätzer für θ und $g : \mathbb{P}_\theta \mapsto g(\mathbb{P}_\theta) \in \mathbb{R}$. Dann heißen*

$$\hat{b}_{\theta,g,boot} := \mathbb{E}_{\hat{\theta}}[g(\mathbb{P}_{\hat{\theta}})] - g(\mathbb{P}_{\hat{\theta}}), \qquad \hat{v}_{\theta,g,boot} := \mathbb{V}_{\hat{\theta}}[g(\mathbb{P}_{\hat{\theta}})].$$

ideale Bootstrap-Schätzer für Bias und Varianz. Etwas genauer spezifizieren wir die Abhängigkeiten der nach $\mathbb{P}_\theta$ verteilten Daten X und gemäß $\mathbb{P}_{\hat{\theta}} = \mathbb{P}_{\hat{\theta}(X)}$ gezogenen Zufallsvariablen Y. Es ergeben sich

$$\hat{b}_{\theta,g,boot}(X) := \mathbb{E}_{\hat{\theta}(X)}[g(\mathbb{P}_{\hat{\theta}(Y)})] - g(\mathbb{P}_{\hat{\theta}(X)}), \qquad \hat{v}_{\theta,g,boot}(X) := \mathbb{V}_{\hat{\theta}(X)}[g(\mathbb{P}_{\hat{\theta}(Y)})]. \qquad (*)$$

Für diese sind also Eigenschaften der Verteilung von $g(\mathbb{P}_{\hat{\theta}(Y)})$ unter $P_{\hat{\theta}(X)}$ zu bestimmen.

Bemerkung 5.4 (Nicht-parametrischer und parametrischer Bootstrap). In den Beispielen sieht man zwei verschiedene Situationen: In 1. und 2. schätzen wir $\mathbb{P}_\theta$ durch die empirische Verteilung. In 3. schätzen wir $\mathbb{P}_\theta$, indem wir $\theta = (\mu, \sigma^2)$ schätzen, und diese Schätzer für θ in $\mathbb{P}_\theta$ einsetzen. Hier kommt also die empirische Verteilung nicht vor. Im ersten Fall heißt der Schätzer $g(\mathbb{P}_{\hat{\theta}})$ nicht-parametrisch, im zweiten Fall parametrischer Bootstrap-Schätzer. Die Schätzer für Bias und Varianz heißen entsprechend nicht-parametrischer und parametrischer Bootstrap-Schätzer.

Beispiel 5.5 (Schätzung des Bias und der Varianz). Wir verwenden dieselben statistischen Modelle wie in Beispiel 5.1 und 5.2.

1. Zwar haben wir bereits Schätzer für $b_{\theta,g}$ und $v_{\theta,g}$ kennen gelernt, jedoch wollen wir auch noch berechnen, was sich durch den idealen Bootstrap-Schätzer ergibt. Wir erhalten

$$g(\mathbb{P}_{\hat{\theta}(Y)}) = \frac{1}{n}\sum_{i=1}^{n} f(Y_i) =: \overline{f(Y)}$$

und damit

$$\begin{aligned}
\hat{b}_{\theta,g,\text{boot}}(X) &= \mathbb{E}_{\hat{\theta}(X)}[\overline{f(Y)}] - \overline{f(X)} = \mathbb{E}_{\hat{\theta}(X)}[f(Y_1)] - \overline{f(X)} \\
&= \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \overline{f(X)} = 0, \\
\hat{v}_{\theta,g,\text{boot}}(X) &= \mathbb{V}_{\hat{\theta}(X)}\Big[\frac{1}{n}\sum_{i=1}^{n} f(Y_i)\Big] = \frac{1}{n}\mathbb{V}_{\hat{\theta}(X)}[f(Y_1)] = \frac{1}{n}\Big(\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \overline{f(X)})^2\Big) \\
&= \frac{n-1}{n^2} s^2(f(X)).
\end{aligned}$$

Damit sind die idealen Bootstrap-Schätzer für Bias und Varianz fast identisch mit den in Beispiel 5.2 erhaltenen.

2. Wir benötigen Aussagen über die Verteilung von $g(\mathbb{P}_{\hat{\theta}(Y)}) = Y_{(m)}$, wobei $Y \sim \mathbb{P}_{\hat{\theta}(X)}$, also über den Median einer nach der empirischen Verteilung gezogenen Stichprobe. Wir berechnen

$$\begin{aligned}
p_k &:= \mathbb{P}_{\hat{\theta}(X)}(Y_{(m)} = X_{(k)}) \\
&= \mathbb{P}_{\hat{\theta}(X)}(Y_{(m)} \leq X_{(k)}) - \mathbb{P}_{\hat{\theta}(X)}(Y_{(m)} \leq X_{(k-1)}) \\
&= \mathbb{P}_{\hat{\theta}(X)}\begin{pmatrix}\text{einer aus } X_{(k+1)},...,X_{(n)} \text{ wird} \\ \text{höchstens } m\text{-mal gezogen}\end{pmatrix} - \mathbb{P}_{\hat{\theta}(X)}\begin{pmatrix}\text{einer aus } X_{(k)},...,X_{(n)} \text{ wird} \\ \text{höchstens } m\text{-mal gezogen}\end{pmatrix} \\
&= \sum_{i=0}^{m}\binom{n}{i}\Big[\Big(\frac{n-k}{n}\Big)^i\Big(\frac{k}{n}\Big)^{n-i} - \Big(\frac{n-k+1}{n}\Big)^i\Big(\frac{k-1}{n}\Big)^{n-i}\Big]
\end{aligned}$$

(und bemerken, dass p_k nicht von X abhängt). Damit sind also

$$\hat{b}_{\theta,g,\text{boot}}(X) = \mathbb{E}_{\hat{\theta}(X)}[Y_{(m)}] - X_{(m)} = \sum_{k=1}^{n} p_k X_{(k)} - X_{(m)},$$

$$\hat{v}_{\theta,g,\text{boot}}(X) = \mathbb{V}_{\hat{\theta}(X)}[Y_{(m)}] = \sum_{k=1}^{n} p_k\Big(X_{(k)} - \sum_{j=1}^{n} p_j X_{(j)}\Big)^2$$

Schätzer für Bias und Varianz von $X_{(m)}$ als Median von $\mathbb{P}_\theta$.

3. Hier benötigen wir Eigenschaften der Verteilung von $g(\mathbb{P}_{\hat{\theta}(Y)}) = \mathbb{P}_{\overline{Y},s^2(Y)/n}(\overline{Z} \in A)$ (mit $Z \sim \mathbb{P}_{\overline{Y},s^2(Y)/n}$), wobei $Y \sim \mathbb{P}_{\overline{X},s^2(X)/n}$. Wir erhalten

$$b_{\theta,g,\text{boot}}(X) = \mathbb{E}_{\overline{X},s^2(X)}[\mathbb{P}_{\overline{Y},s^2(Y)/n}[\overline{Z} \in A]] - \mathbb{P}_{\overline{X},s^2(X)/n}(\overline{Y} \in A),$$

$$v_{\theta,g,\text{boot}}(X) = \mathbb{V}_{\overline{X},s^2(X)}[\mathbb{P}_{\overline{Y},s^2(Y)/n}[\overline{Z} \in A]].$$

Bemerkung 5.6 (Berechnung unter $\mathbb{P}_{\hat{\theta}(X)}$). Da es sich bei $\mathbb{P}_{\hat{\theta}(X)}$ bei Beispielen 1. und 2. um eine empirische (und damit diskrete) Verteilung handelt, können Erwartungswerte bezüglich $\mathbb{P}_{\hat{\theta}(X)}$ als Summen geschrieben werden. Die Anzahl der Summanden lässt sich reduzieren auf die möglichen Anordnungen von $Y_1,...,Y_n$ (die nach der empirischen Verteilung gezogen werden) auf die Daten $X_1,...,X_n$. Da jedes X_i öfter vorkommen kann, führt dies auf eine Summe mit $\binom{2n-1}{n}$ Summanden. In Beispielen 1. und 2. lässt sich die Summe glücklicherweise geschickt umschreiben, dass deutlich weniger Summanden zu berechnen sind. Dies muss allerdings nicht immer sein.

Anstatt die Erwartungswerte bezüglich $\mathbb{P}_{\hat{\theta}(X)}$ auszurechnen, kann man diese auch mittels des Gesetzes großen Zahlen approximieren. Dies führt auf einen neuen Schätzer von $b_{\theta,g}$ und $v_{\theta,g}$.

Definition 5.7 (Bootstrap-Schätzer). *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein statistisches Modell und $\mathbb{P}_{\hat{\theta}(X)}$ ein Schätzer für $\mathbb{P}_\theta$, sowie g eine reellwertige Funktion. Sei $B \in \mathbb{N}$ und $Y^1,...,Y^B \in \mathbb{R}^n$ unabhängig und nach $\mathbb{P}_{\hat{\theta}(X)}$ verteilt. Dann heißen*

$$b^B_{\theta,g,boot}(X) := \frac{1}{B}\sum_{b=1}^{B} g(\mathbb{P}_{\hat{\theta}(Y^b)}) - g(\mathbb{P}_{\hat{\theta}(X)}),$$

$$v^B_{\theta,g,boot}(X) := \frac{1}{B-1}\sum_{b=1}^{B}\Big(g(\mathbb{P}_{\hat{\theta}(Y^b)}) - \frac{1}{B}\sum_{c=1}^{B} g(\mathbb{P}_{\hat{\theta}(Y^c)})\Big)^2$$

approximative Bootstrap-Schätzer von Bias und Varianz.

5.3 Awendungen

Beispiel 5.8 (Anwendungen). Das interessante an den approximativen Bootstrap-Schätzern $b^B_{\theta,g,\text{boot}}$ und $v^B_{\theta,g,\text{boot}}$ ist, dass man sie gut simulieren kann und kein theoretisches Wissen mehr über $\mathbb{P}_\theta$ nötig ist. Weiter gilt wegen des Gesetzes großer Zahlen

$$b^B_{\theta,g,\text{boot}} \xrightarrow{B\to\infty} b_{\theta,g,\text{boot}},$$
$$v^B_{\theta,g,\text{boot}} \xrightarrow{B\to\infty} v_{\theta,g,\text{boot}},$$

also handelt es sich immerhin um approximativ ideale Schätzer. Wir illustrieren dies nun an unseren drei Beispielen

1. Sei etwa $\mathbb{P}_\theta$ eine χ^2_θ-Verteilung und $f = \text{id}$. Wir schätzen also $g(\mathbb{P}_\theta) = \mathbb{E}_\theta[X]$ mittels $\overline{X}$. Für $n = 100$ und $\theta = 1$ ergibt sich:

   ```
   > library(stats)
   > x<-rchisq(100, df=theta)
   > mean(x)
   [1] 0.8185859
   ```

   Wir geben nun an, wie man den Bias und die Varianz dieses Schätzers (die exakt in Beispiel 5.2 berechnet wurden und 0 und 2/100 sind) approximieren kann. Zunächst berechnen wir den Schätzer für die Varianz aus Beispiel 5.2.

   ```
   > sd(x)^2/n
   [1] 0.02063456
   ```

   Wie wir aus Beispiel 5.5 wissen, ist der ideale Bootstrap-Schätzer des Bias und der Varianz recht ähnlich.

   ```
   > (n-1)*sd(x)^2/n^2
   [1] 0.02042822
   ```

   Um diesen letzten Schätzer zu approximieren, verwenden wir nun den approximativen Bootstrap-Schätzer. Sei hierzu $B = 1000$.

   ```
   > B=1000; ghat=rep(0,B)
   > for(i in 1:B) ghat[i]<-mean(sample(x, n, replace=TRUE))
   ```

   Nun stehen im Vektor `ghat` die Mittelwerte von 1000 Bootstrap-Stichproben. Wir können nun Bias und Varianz wie in Definition 5.7 schätzen.

   ```
   > mean(ghat) - mean(x)
   [1] 0.006199645
   > sd(ghat)^2
   [1] 0.01943494
   ```

   Der Bias wird also fast auf 0 und die Varianz auch fast richtig geschätzt.

2. Die Boostrap-Schätzung des Medians durch den Stichproben-Median $X_{(m)}$ wird in der Übung behandelt.

3. Sei $\theta = (\mu, \sigma^2) = (0, 1), n = 100$ und $A = (-\infty, 0.1645)$, so dass $\mathbb{P}_\theta(\overline{X} \in A) \approx 0.95$. Wir berechnen nun den approximativen Bootstrap-Schätzer des Bias und der Varianz von $\mathbb{P}_\theta(\overline{X} \in A)$ durch $\mathbb{P}_{\overline{X}, s^2(X)/n}(\overline{Y} \in A)$.

```
pnorm(0.1645, mean=0, sd=0.1)
[1] 0.9500151
n=100; B=1000
x<-rnorm(n)
ghat=rep(0,B)
for(i in 1:B) {
  y<-rnorm(n, mean=mean(x), sd=sd(x))
  ghat[i]<-pnorm(0.1645, mean=mean(y), sd=sd(y))
}
mean(ghat) - pnorm(0.1645, mean=mean(x), sd=sd(x))
[1] -0.0004691647
sd(ghat)^2
[1] 0.001487795
```

Wir zeigen nun noch zwei Eigenschaften des idealen und approximativen Bootstrap-Schätzers. Beide haben gleichen Erwartungswert, jedoch hat der ideale eine kleinere Varianz als der approximative.

Proposition 5.9 (Vergleich approximativer und idealer Bootstrap-Schätzer). *Für ein statistisches Modell $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ sei $\mathbb{P}_{\hat{\theta}(X)}$ ein Schätzer von $\mathbb{P}_\theta$ und g eine reellwertige Funktion. Sei $g(\mathbb{P}_{\hat{\theta}(X)})$ der Schätzer von $g(\mathbb{P}_\theta)$. Dann gilt für die idealen und approximativen Bootstrap-Schätzer des Bias und der Varianz von $g(\mathbb{P}_{\hat{\theta}(X)})$*

$$\begin{aligned} &\mathbb{E}_\theta[b^B_{\theta,g,boot}(X)] = \mathbb{E}_\theta[b_{\theta,g,boot}(X)], && \mathbb{V}_\theta[b^B_{\theta,g,boot}(X)] \geq \mathbb{V}_\theta[b_{\theta,g,boot}(X)], \\ &\mathbb{E}_\theta[v^B_{\theta,g,boot}(X)] = \mathbb{E}_\theta[v_{\theta,g,boot}(X)], && \mathbb{V}_\theta[v^B_{\theta,g,boot}(X)] \geq \mathbb{V}_\theta[v_{\theta,g,boot}(X)] \end{aligned}$$

Beweis. Wir schreiben für den Schätzer des Bias mit Hilfe der Turmeigenschaft für die bedingte Erwartung und der Varianzzerlegung

$$\begin{aligned} \mathbb{E}_\theta[b^B_{\theta,g,\text{boot}}(X)] &= \mathbb{E}_\theta[g(\mathbb{P}_{\hat{\theta}(Y)}) - g(\mathbb{P}_{\hat{\theta}(X)})] = \mathbb{E}_\theta[\mathbb{E}_{\hat{\theta}(X)}[g(\mathbb{P}_{\hat{\theta}(Y)})] - g(\mathbb{P}_{\hat{\theta}(X)})] \\ &= \mathbb{E}_\theta[b_{\theta,g,\text{boot}}(X)], \\ \mathbb{V}_\theta[b^B_{\theta,g,\text{boot}}(X)] &\geq \mathbb{V}_\theta[\mathbb{E}_{\hat{\theta}(X)}[b^B_{\theta,g,\text{boot}}(X)]] = \mathbb{V}_\theta[\mathbb{E}_{\hat{\theta}(X)}[g(\mathbb{P}_{\hat{\theta}(Y)})] - g(\mathbb{P}_{\hat{\theta}(X)})] \\ &= \mathbb{V}_\theta[b_{\theta,g,\text{boot}}(X)]. \end{aligned}$$

Für den Schätzer der Varianz erhalten wir

$$\begin{aligned}
\mathbb{E}_\theta[v^B_{\theta,g,\mathrm{boot}}(X)] &= \frac{B}{B-1}\mathbb{E}_\theta\Big[\Big(g(\mathbb{P}_{\hat\theta(Y^1)}) - \frac{1}{B}\sum_{c=1}^{B} g(\mathbb{P}_{\hat\theta(Y^c)})\Big)^2\Big] \\
&= \frac{B}{B-1}\mathbb{E}_\theta\Big[\mathbb{E}_{\hat\theta(X)}\Big[\Big(g(\mathbb{P}_{\hat\theta(Y^1)})^2 - \frac{2}{B}g(\mathbb{P}_{\hat\theta(Y^1)}\sum_{c=1}^{B} g(\mathbb{P}_{\hat\theta(Y^c)}) + \frac{1}{B^2}\sum_{c,d=1}^{B} g(\mathbb{P}_{\hat\theta(Y^c)}g(\mathbb{P}_{\hat\theta(Y^d)}\Big]\Big] \\
&= \frac{B}{B-1}\mathbb{E}_\theta\Big[\mathbb{E}_{\hat\theta(X)}\Big[\Big(1-\frac{1}{B}\Big)g(\mathbb{P}_{\hat\theta(Y^1)})^2 - \frac{B-1}{B}g(\mathbb{P}_{\hat\theta(Y^1)})g(\mathbb{P}_{\hat\theta(Y^2)})\Big]\Big] \\
&= \mathbb{E}_\theta\big[\mathbb{E}_{\hat\theta(X)}[g(\mathbb{P}_{\hat\theta(Y)})^2]\big] - \mathbb{E}_{\hat\theta(X)}[g(\mathbb{P}_{\hat\theta(Y)})]^2\big] = \mathbb{E}_\theta\big[\mathbb{V}_{\hat\theta(X)}[g(\mathbb{P}_{\hat\theta(Y)})]\big] \\
&= \mathbb{E}_\theta[v_{\theta,g,\mathrm{boot}}(X)],
\end{aligned}$$

und die letzte Ungleichung wird als Übungsaufgabe behandelt. □

6 Der E(xpectation)-M(aximization)-Algorithmus

In bestimmten Fällen ist es schwierig, Maximum-Likelihood-Schätzer direkt anzugeben. Wir werden hier einen besonderen Fall diskutieren, in dem es einen Ausweg über den EM-Algorithmus gibt.

6.1 Maximum-Likelihood-Schätzer in Mischungsmodellen

Wir nehmen ein statistisches Modell

$$(X, \{\mathbb{P}_\theta = ((1-\pi)p_{\theta_0} + \pi p_{\theta_1})^n \cdot \lambda^n : \theta = (\pi, \theta_0, \theta_1), \pi \in [0,1], \theta_0, \theta_1 \in \mathcal{P}\}) \tag{6.1}$$

an. Das bedeutet: da $\mathbb{P}_\theta$ ein Produktmaß ist, sind die Datenpunkte $X_1, ..., X_n$ unabhängig. Die Verteilung von X_1 ist jedoch eine Mischung aus $p_{\theta_0} \cdot \lambda$ und $p_{\theta_1} \cdot \lambda$. Das stellt man sich am besten so vor: Sei $Z_1 = 1$ mit Wahrscheinlichkeit π und $Z_1 = 0$ mit Wahrscheinlichkeit $(1-\pi)$. Im Anschluss ziehen wir eine nach $p_{\theta_{Z_1}}$-verteilte Zufallsvariable X_1. Dann ist nämlich gerade $X_1 \sim ((1-\pi)p_{\theta_0} + \pi p_{\theta_1}) \cdot \lambda$.

Will man in einer solchen Situation einen Maximum-Likelihood-Schätzer für $\theta = (\pi, \theta_0, \theta_1)$ angeben, so hätte man

$$\ell(\theta; X) = \sum_{i=1}^{n} \log\big((1-\pi)p_{\theta_0}(X_i) + \pi p_{\theta_1}(X_i)\big)$$

zu maximieren, was wegen der Summe innerhalb von log nicht einfach ist. Viel einfacher wäre es, wenn wir über X_i wüssten, ob es sich um eine Ziehung aus $p_{\theta_0} \cdot \lambda$ oder aus $p_{\theta_1} \cdot \lambda$ handelt. Wäre nämlich $Z_i = 0$ oder $Z_i = 1$ je nachdem, welcher Fall eintritt, so ist die log-Likelihood

$$\ell'(\theta; X, Z) = \sum_{i=1}^{n} Z_i \log \pi + \sum_{i=1}^{n} (1-Z_i)\log(1-\pi) + \sum_{i=1}^{n} (1-Z_i)\log p_{\theta_0}(X_i) + Z_i p_{\theta_1}(X_i). \tag{6.2}$$

Hierbei meinen wir allerdings die Likelihood im statistischen Modell

$$((X,Z), \{\mathbb{P}_\theta = (p_\theta \cdot \lambda)^n : \theta = (\pi, \theta_0, \theta_1)\}) \text{ mit } p_\theta(x,z) = \pi^z (1-\pi)^{1-z} p_{\theta_z}(x). \tag{6.3}$$

Obwohl sicherlich die Maximierung im statistischen Modell 6.3 einfacher wird, kennen wir normalerweise Z nicht. Wir sprechen hier auch davon, dass Z eine *latente* oder *unbeobachtete* Variable ist. Der Trick, den der EM-Algorithmus verwendet, ist es, diese Variablen Z_i durch ihre Erwartung zu ersetzen, und anschließend die Likelihood zu maximieren.

Wir bemerken noch, dass es sich bei der Dichte im statistischen Modell (6.1) nicht um eine Exponentialfamilie handelt, in (6.3) aber schon. Auch dies spricht zumindest dafür, dass Berechnungen im Modell (6.3) einfacher sein werden.

6.2 Der Algorithmus

Wir beginnen damit, in der obigen Situation den rekursiv definierten EM-Algorithmus anzugeben, wobei x die erhobenen Daten sind.

Der EM-Algorithmus

1. Starte für $t = 0$ mit Anfangswerten $\hat{\theta}^{(t=0)}$.
2. E-Schritt: Berechne für $t = 0, 1, 2, ...$
$$Q(\theta, \hat{\theta}^{(t)}; x) := \mathbb{E}_{\hat{\theta}^{(t)}}[\ell'(\theta, X, Z)|X = x].$$
3. M-Schritt: Berechne
$$\hat{\theta}^{(t+1)} := \arg\max\{Q(\theta, \theta^{(t)}; x) : \theta \in \mathcal{P}\}.$$
4. Iteriere 2. und 3. bis $\hat{\theta}^{(t)}$ konvergiert.

Für den speziellen Fall des obigen Mischungsmodells berechnen wir zunächst für $X = x$

$$\mathbb{E}_{\theta'}[\ell'(\theta; x, Z)] = n(\pi' \log \pi + (1 - \pi') \log(1 - \pi)) + \sum_{i=1}^{n} (1 - \pi') \log p_{\theta_0}(x_i) + \pi' \log p_{\theta_1}(x_i).$$

Nun ergibt sich folgendes:

Der EM-Algorithmus für das Mischungsmodell

1. Starte für $t = 0$ mit Anfangswerten $\hat{\theta}^{(t=0)} = (\hat{\pi}, \hat{\theta}_0, \hat{\theta}_1)$.
2. Erwartungswert-Schritt: Berechne

$$\hat{Z}_i^{(t+1)} = \mathbb{E}_{\hat{\theta}^{(t)}}[Z_i | X_i = x_i] = \frac{\hat{\pi}^{(t)} p_{\hat{\theta}_1^{(t)}}(x_i)}{(1-\hat{\pi}^{(t)}) p_{\hat{\theta}_0^{(t)}}(x_i) + \hat{\pi}^{(t)} p_{\hat{\theta}_1^{(t)}}(x_i)}, \qquad i = 1, ..., n$$

   Damit ergibt sich

$$Q(\theta, \theta^{(t)}; x) = \sum_{i=1}^{n} (\hat{Z}_i^{(t+1)} \log \pi + (1 - \hat{Z}_i^{(t+1)}) \log(1-\pi)) \\ + \sum_{i=1}^{n} (1 - \hat{Z}_i^{(t+1)}) \log p_{\theta_0}(x) + \hat{Z}_i^{(t+1)} \log p_{\theta_1}(x).$$

3. Maximierungs-Schritt: Wir führen zunächst die Maximierung bezüglich π durch. Wir erhalten als notwendige Bedingung

$$\sum_{i=1}^{n} (1-\pi) \hat{Z}_i^{(t+1)} = \sum_{i=1}^{n} (1 - \hat{Z}_i^{(t+1)}) \pi, \text{ also } \hat{\pi}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i^{(t+1)}.$$

   Weiter berechnen wir (mit (6.2))

$$\hat{\theta}^{(t+1)} := \arg\max\{Q(\theta, \hat{\theta}^{(t)}; x) : \theta \in \mathcal{P}\}.$$

4. Iteriere 2. und 3. bis $\hat{\theta}^{(t)}$ konvergiert.

6.3 Beispiele

Beispiel 6.1 (Mischung aus Normalverteilungen). Hier sei für $\theta_i = (\mu_i, \sigma_i^2)$ die Verteilung $\mathbb{P}_{\theta_i} = \mathcal{N}(\mu_i, \sigma_i^2), i = 0, 1$. Hier ist also (wenn μ_0 und μ_1 genügend weit auseinander liegen) die Dichte $(1-\pi)p_{\theta_0} + \pi p_{\theta_1}$ bi-modal (d.h. sie hat zwei Maxima).

In Maximierungs-Schritt des EM-Algorithmus ist also noch

$$\sum_{i=1}^{n} (1 - \hat{Z}_i) \log p_{\theta_0}(x) + \hat{Z}_i \log p_{\theta_1}(x)$$

zu maximieren. Im Fall der Normalverteilungen müssen wir also

$$-\sum_{i=1}^{n}(1-\hat{Z}_i)\Big(\frac{1}{2}\log\sigma_0^2+\frac{(x_i-\mu_0)^2}{2\sigma_0^2}\Big),$$
$$-\sum_{i=1}^{n}\hat{Z}_i\Big(\frac{1}{2}\log\sigma_1^2+\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\Big)$$

über (μ_0,σ_0^2) bzw. (μ_1,σ_1^2) maximieren. Wir berechnen

$$\sigma_0^2\frac{\partial}{\partial\mu_0}\sum_{i=1}^{n}(1-\hat{Z}_i)\Big(\frac{1}{2}\log\sigma_0^2+\frac{(x_i-\mu_0)^2}{2\sigma_0^2}\Big)=\sum_{i=1}^{n}(1-\hat{Z}_i)(x_i-\mu_0),$$
$$2\sigma_0^2\frac{\partial}{\partial\sigma_0^2}\sum_{i=1}^{n}(1-\hat{Z}_i)\Big(\frac{1}{2}\log\sigma_0^2+\frac{(x_i-\mu_0)^2}{2\sigma_0^2}\Big)=\sum_{i=1}^{n}(1-Z_i)\Big(1-\frac{(x_i-\mu_0)^2}{2\sigma_0^2}\Big)$$

und damit sind die Maximierer

$$\hat{\mu}_0=\frac{\sum_{i=1}^{n}(1-\hat{Z}_i)x_i}{\sum_{i=1}^{n}(1-\hat{Z}_i)},$$
$$\hat{\sigma}_0^2=\frac{\sum_{i=1}^{n}(1-\hat{Z}_i)(x_i-\hat{\mu}_0)^2}{\sum_{i=1}^{n}(1-\hat{Z}_i)}.$$

Analog ergeben sich

$$\hat{\mu}_1=\frac{\sum_{i=1}^{n}\hat{Z}_i x_i}{\sum_{i=1}^{n}\hat{Z}_i},$$
$$\hat{\sigma}_1^2=\frac{\sum_{i=1}^{n}\hat{Z}_i(x_i-\hat{\mu}_0)^2}{\sum_{i=1}^{n}\hat{Z}_i}.$$

Beispiel 6.2 (Gauß'sches Mischungsmodell). Für $\pi=0.5$ wählen wir $(\mu_0,\sigma_0^2)=(-2,1)$ und $(\mu_1,\sigma_1^2)=(2,1)$.

```
n=1000
pi=0.5
z<-rbinom(n,1,pi)
x<-0*z

for(i in 1:n) {
  if(z[i]==0) x[i]<-rnorm(1, mean=-2, sd=1)
  else x[i]<-rnorm(1, mean=2, sd=1)
}
```

Nachdem unsere Daten nun in `x` gespeichert sind, wollen wir die Parameter schätzen.

```
# theta = c(pi, mu0, sigma20, mu1, sigma21) = (0.5,-2,1,2,1)
thetaold=0
theta = c(0.2, -1, 1, 1, 1)
```

```
eps=10^(-4)

while(norm(as.matrix(theta-thetaold))>eps) {
  thetaold=theta
  hatZ<-(theta[1]*dnorm(x, mean=theta[4], sd=sqrt(theta[5])))/
      (theta[1]*dnorm(x, mean=theta[4], sd=sqrt(theta[5]))
      + (1-theta[1])*dnorm(x, mean=theta[2], sd=sqrt(theta[3])))
  theta[1] = mean(hatZ)
  theta[2] = sum((1-hatZ)*x)/sum(1-hatZ)
  theta[3] = sum((1-hatZ)*(x-theta[2])^2)/sum(1-hatZ)
  theta[4] = sum(hatZ*x)/sum(hatZ)
  theta[5] = sum(hatZ*(x-theta[4])^2)/sum(hatZ)
  cat(theta, "\n")
}
0.4080493 -1.657099 1.873421 2.146217 1.0034
0.4221161 -1.757708 1.571493 2.157208 0.8661055
0.4324705 -1.818895 1.393268 2.143772 0.8500342
0.4408061 -1.860213 1.292241 2.121252 0.8673303
0.4472676 -1.889556 1.228207 2.099996 0.8913398
0.4521783 -1.910826 1.184815 2.082438 0.9139556
0.4558773 -1.926362 1.154515 2.06858 0.9331244
0.4586528 -1.937761 1.133011 2.057859 0.9486591
0.4607324 -1.946157 1.117571 2.049651 0.9609445
0.4622902 -1.952364 1.106378 2.043406 0.9705129
0.4634575 -1.956969 1.098198 2.038673 0.9778897
0.4643328 -1.960395 1.092182 2.035094 0.983537
0.4649893 -1.96295 1.087734 2.032393 0.987839
0.465482 -1.96486 1.084432 2.030356 0.9911048
0.465852 -1.966289 1.081973 2.028822 0.9935777
0.4661299 -1.967359 1.080138 2.027666 0.9954469
0.4663386 -1.968162 1.078766 2.026797 0.9968578
0.4664955 -1.968764 1.077738 2.026142 0.9979217
0.4666134 -1.969217 1.076968 2.02565 0.9987234
0.466702 -1.969556 1.076391 2.025279 0.9993273
0.4667686 -1.969812 1.075957 2.025001 0.9997818
0.4668187 -1.970003 1.075631 2.024791 1.000124
0.4668563 -1.970147 1.075387 2.024634 1.000381
0.4668847 -1.970256 1.075203 2.024515 1.000575
0.4669059 -1.970337 1.075065 2.024426 1.000721
0.4669219 -1.970398 1.074961 2.024359 1.00083
0.466934 -1.970445 1.074883 2.024309 1.000913
0.466943 -1.970479 1.074824 2.024271 1.000975
0.4669498 -1.970505 1.07478 2.024242 1.001021
0.4669549 -1.970525 1.074747 2.024221 1.001056
0.4669588 -1.970539 1.074722 2.024205 1.001083
```

Wir wollen nun noch begründen, warum der EM-Algorithmus (zumindest lokale) Maxima der Likelihood ansteuert.

Proposition 6.3. *Für*

$$\ell(\theta; x) := \log \int p_\theta(x, z')\lambda(dz')$$

gilt

$$\ell(\hat\theta^{(t+1)}; x) \geq \ell(\hat\theta^{(t)}; x)$$

mit "=" genau dann, wenn $\hat\theta^{(t+1)} = \hat\theta^{(t)}$.

Beweis. Wir schreiben zunächst

$$p_{\theta,x}(z) := \frac{p_\theta(x, z)}{\int p_\theta(x, z')\lambda(dz')}$$

als die bedingte Dichte von p_θ gegeben $X = x$. Dann ist nämlich

$$\ell(\theta; x) = \log p_\theta(x, z) - \log p_{\theta,x}(z).$$

Nehmen wir nun auf der rechten Seite Erwartungswerte bezüglich der Verteilung mit Dichte $p_{\hat\theta^{(t)},x}(z)$, so folgt (mit $R(\theta, \theta'; x) := \mathbb{E}_{\theta'}[\log p_{\theta,X}(Z)|X = x]$)

$$\begin{aligned}\ell(\theta; x) &= \mathbb{E}_{\hat\theta^{(t)}}[\log p_\theta(X, Z)|X = x] - \mathbb{E}_{\hat\theta^{(t)}}[\log p_{\theta,X}(Z)|X = x] \\ &= Q(\theta, \hat\theta^{(t)}; x) - R(\theta, \hat\theta^{(t)}; x).\end{aligned}$$

Nun ist

$$\ell(\hat\theta^{(t+1)}; x) - \ell(\hat\theta^{(t)}; x) = Q(\hat\theta^{(t+1)}, \hat\theta^{(t)}; x) - Q(\hat\theta^{(t)}, \hat\theta^{(t)}; x) - (R(\hat\theta^{(t+1)}, \hat\theta^{(t)}; x) - R(\hat\theta^{(t)}, \hat\theta^{(t)}; x))$$

und wir haben

$$Q(\hat\theta^{(t+1)}, \hat\theta^{(t)}; x) - Q(\hat\theta^{(t)}, \hat\theta^{(t)}; x) \geq 0$$

da $\hat\theta^{(t+1)}$ der Maximierer von $\theta \mapsto Q(\theta, \hat\theta^{(t)}; x)$ ist, und für jedes θ, θ' ist

$$\begin{aligned}R(\theta', \theta) - R(\theta, \theta) &= \mathbb{E}_\theta[\log p_{\theta',X}(Z)|X = x] - \mathbb{E}_\theta[\log p_{\theta,X}(Z)|X = x] \\ &= \mathbb{E}_\theta\Big[\log \frac{p_{\theta',X}(Z)}{p_{\theta,X}(Z)}|X = x\Big] \leq \log \mathbb{E}_\theta\Big[\frac{p_{\theta',X}(Z)}{p_{\theta,X}(Z)}|X = x\Big] \\ &= \log \int \frac{p_{\theta',x}(z)}{p_{\theta,x}(z)} p_{\theta,x}(z)\lambda(dz) = 0.\end{aligned}$$

mit "=" genau dann, wenn $\theta = \theta'$. Daraus folgen nun alle Behauptungen. □

Der EM-Algorithmus konvergiert wegen der Beschränktheit der Likelihood immer. Allerdings ist unklar, ob er nicht in einem lokalen Maximum der Likelihood konvergiert. Eine mögliche Strategie, dies zu umgehen, ist es, in verschiedenen Startpunkten zu beginnen, und dann die maximale Likelihood auszuwählen.

7 Die Hauptkomponentenanalyse

Eigentlich ist die Hauptkomponentenanalyse (englisch: *Principal Component Analysis*) eine Methode der deskriptiven Statistik. Wir behandeln sie hier dennoch, da sie häufig verwendet wird, und auch Verbindungen zu linearen Modellen aufweist.

7.1 Einführung

Wir gehen von einem Datensatz $x \in \mathbb{R}^{n\times p}$ aus, wobei – wie im Regressionsmodell – x_{ij} die j-te Covariate des i-ten beobachteten Items ist. Um x exakt zu beschreiben, benötigen wir natürlich alle $n \times p$ Einträge. Die Hauptkomponentenanalyse versucht nun, mit weniger Daten zumindest die wichtigsten Eigenschaften der Daten abzubilden. Die Grundidee ist, anstatt x die Matrix $B^\top x$ für ein $B \in \mathbb{R}^{p\times q}$ für ein $q < p$ zu betrachten. Damit werden die Daten auf $n\times q$ reduziert. Relevante Information über die Daten entstehen durch Varianzen von Kenngrößen. (D.h. hat eine Größe eine kleine Varianz, so kann man aus ihr keine starken Aussagen über die Daten ablesen.) Deshalb versucht man, B so zu wählen, dass (die q Komponenten von) $B^\top x$ möglichst große Varianz besitzen.

Versuchen wir uns zunächst am Fall $q = 2$, wobei wir nicht die Varianz, sondern die empirische Varianz maximieren wollen. Gesucht ist also zunächst ein $\alpha_1 \in \mathbb{R}^p$, so dass $s^2(x\alpha_1)$ maximal wird. Zunächst ist (mit $x1 = (x, ..., x)$ für ein $x \in \mathbb{R}$ und $\overline{x} = (\frac{1}{n}\sum_{i=1}^n x_{ij})_{j=1,...,p}$

$$\begin{aligned}
\overline{x\alpha_1} &= \frac{1}{n}\sum_{i=1}^n (x\alpha_1)_i = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p x_{ij}\alpha_{1j} = \sum_{j=1}^p \alpha_{1j}\overline{x}_{\cdot j} = \alpha_1^\top \overline{x},\\
s^2(x\alpha_1) &= \frac{1}{n-1}\sum_{i=1}^n ((x\alpha_1)_i - \overline{x\alpha_1})^2 = \frac{1}{n-1}(x\alpha_1 - 1\overline{x\alpha_1})^\top (x\alpha_1 - 1\overline{x\alpha_1})\\
&= \frac{1}{n-1}(\alpha_1^\top x^\top x\alpha_1 - \alpha_1^\top \overline{x}1^\top 1\overline{x}^\top \alpha_1) = \frac{1}{n-1}\alpha_1^\top S\alpha_1
\end{aligned}$$

für $S := (x - 1\overline{x}^\top)^\top (x + 1\overline{x}^\top)$. Um ein Maximum von $s^2(x\alpha_1)$ zu finden, verwenden wir noch die Nebenbedinung $||\alpha_1||_2 = \alpha_1^\top \alpha_1 = 1$. Damit müssen wir mittels Lagrange-Multiplikatoren das Maximierungsproblem

$$\alpha_1^\top S\alpha_1 \to \max \text{ mit } \alpha_1^\top \alpha_1 = 1$$

lösen. Hierzu setzen wir

$$\nabla_{\alpha_1}\alpha_1^\top S\alpha_1 - \lambda_1(\alpha_1^\top \alpha_1 - 1) = 2S\alpha_1 - 2\lambda_1\alpha_1 = 2(S - \lambda_1 I)\alpha_1 = 0$$

an (nachdem wir uns an das Kapitel *Extrema unter Nebenbedingungen* aus der Analysis erinnert haben). Also muss α_1 ein Eigenvektor von S zum Eigenwert λ_1 sein. Um $\alpha_1^\top S\alpha_1$ zu maximieren, muss weiter α_1 der Eigenvektor zum größten Eigenwert λ_1 von S sein. Wir sagen auch, $x\alpha_1$ ist die erste Hauptkomponente von x.

Um die zweite Hauptkomponente $x\alpha_2$ von x zu bestimmen, benötigen wir einen Vektor α_2, so dass $\alpha_1^\top \alpha_2 = 0, \alpha_2^\top \alpha_2 = 1$ (d.h. α_2 ist senkrecht auf α_1) und $s^2(x\alpha_2)$ maximal ist. Hierzu stellen wir das Maximierungsproblem

$$\alpha_2^\top S\alpha_2 \to \max \text{ mit } \alpha_2^\top \alpha_2 = 1, \alpha_2^\top \alpha_1 = 0$$

auf, das wir durch

$$\nabla_{\alpha_2}(\alpha_2^\top S\alpha_2 - \lambda_2(\alpha_2^\top \alpha_2 - 1) - \phi\alpha_2^\top \alpha_1) = 2(S - \lambda_2 I)\alpha_2 - \phi\alpha_1 = 0$$

ansetzen. Multiplikation mit $\alpha_1^\top$ von links ergibt $\phi = 0$, und damit muss α_2 Eigenvektor von S sein. Um das Maximierungsproblem zu lösen, muss also λ_2 der zweitgrößte Eigenwert sein, und α_2 der dazugehörige Eigenvektor.

Iteriert man dieses Vorgehen weiter, führt dies zur Definition der Hauptkomponentenanalyse.

Definition 7.1 (Hauptkomponentenanalyse, PCA). *1. Seien $X_1, ..., X_p$ Zufallsvariable mit $\mathbb{E}[X_i] = 0$ und $\mathbb{COV}[X_i, X_j] = \Sigma_{ij}$ für alle $i, j = 1, ..., p$ und eine Matrix $\Sigma \in \mathbb{R}^{p\times p}$. (Notwendigerweise ist dann $\Sigma \in \mathbb{R}_+^{p\times p}$ symmetrisch und positiv semi-definit[10].) Wir nehmen an, dass alle Eigenwerte von Σ verschieden und verschieden von 0 sind. Seien $\lambda_1 > \cdots > \lambda_p$ die Eigenwerte von Σ mit Eigenvektoren $\alpha_1, ..., \alpha_p$ für $||\alpha_k||_2 = 1$. (Da Σ symmetrisch ist, ist also $\alpha_1, ..., \alpha_p$ ein Orthonormalsystem.) Dann heißt $Z_k := \alpha_k^\top X$ (oder $Z_k^\top = X^\top \alpha_k$) die k-te Hauptkomponente (von Σ) und α_k heißt der Vektor der Gewichte der k-ten Hauptkomponente.*

2. Sei $x \in \mathbb{R}^{n\times p}$ (man denke etwa an n unabhänige Realisierungen der Zufallsvariablen X aus 1.), $\overline{x}_k = \frac{1}{n}\sum_{i=1}^n x_{ik}$ und $S = (s_{k\ell})_{1\le k,\ell\le p} \in \mathbb{R}^{p\times p}$, definiert als

$$s_{k\ell} := \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)(x_{i\ell} - \overline{x}_\ell).$$

(Auch hier ist S symmetrisch und positiv semi-definit.) Wir nehmen an, dass alle Eigenwerte von S verschieden und verschieden von 0 sind.

Seien $\lambda_1 > \cdots > \lambda_p$ die Eigenwerte von S mit Eigenvektoren $\alpha_1, ..., \alpha_p$ für $||\alpha_k||_2 = 1$. Dann heißt $z_k := x\alpha_k$ die k-te Hauptkomponente (von S) und α_k heißt der Vektor der Gewichte der k-ten Hauptkomponente.

7.2 Die Hauptkomponentenanalyse in R

Wir verwenden den Datensatz `iris`, der in `R` implementiert ist. Dieser beschreibt je 50 Pflanzen der Gattungen *Iris setosa*, *versicolor* und *virginica*.

```
data(iris)
iris.pca <- prcomp(iris[,1:4], center = TRUE, scale = TRUE)
```

Dies führt bereits die Hauptkomponentenanalyse für die ersten vier Variablen des Datensatzes durch. Die fünfte Variable kann nicht verwendet werden, weil sie nicht numerisch ist. `center=TRUE` gibt hierbei an, dass in der Tat (genau wie in den Formeln oben) die Matrix S mit den um $\overline{x}$ verschobenen Daten gefüllt wird. `scale=TRUE` bedeutet, dass die Variablen – bevor die Hauptkomponentenanalyse durchgeführt wird – auf eine empirische Varianz von 1 gebracht werden. Das Ergebnis ist:

```
iris.pca
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:
```

[10] Die positive Semi-Definitheit folgt aus $\alpha^\top \Sigma\alpha = \sum_{i,j}\alpha_i \mathbb{COV}[X_i, X_j]\alpha_j = \mathbb{V}\Big[\sum_i \alpha_i X_i\Big] \ge 0$ für alle $\alpha \in \mathbb{R}^p$

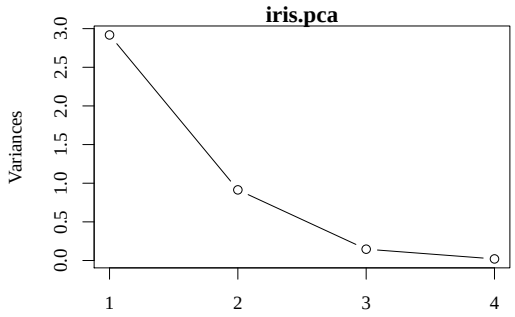


Abbildung 7.1: Plot der erklärten Varianz (d.h. das Quadrat der in `iris.pca` gespeicherten Standardabweichungen) im `iris`-Datensatz.

```
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Die `Standard deviations` geben an, wieviel der Gesamtvarianz durch die vier Hauptkomponenten erklärt wird. Die angefügte Tabelle gibt die Gewichts-Vektoren der vier Hauptkomponenten an. Wir erzeugen nun noch mit diesen Daten zwei Grafiken.

```
plot(iris.pca, type="l")
biplot(iris.pca)
```

Um den letzten Plot etwas übersichtlicher zu gestalten, führen wir ihn nochmal durch, geben nun aber Farben für die drei Arten bei. Hierbei erkennen wir, dass bereits die erste Hauptkomponente gut zwischen den drei verschiedenen Arten unterscheiden kann.

```
raw <- iris.pca$x[,1:2]
plot(raw[,1], raw[,2], col="white", pch=20)
points(raw[1:50,1], raw[1:50,2], col="red", pch=20)
points(raw[51:100,1], raw[1:50,2], col="blue", pch=20)
points(raw[101:150,1], raw[1:50,2], col="green", pch=20)
```

7.3 Optimalität der Hauptkomponenten

Bemerkung 7.2 (Notation).

1. Wie bereits früher schreiben wir $\mathbb{COV}[X, X] = \Sigma$.

2. In obiger Situation schreiben wir $\Lambda \in \mathbb{R}^{p\times p}$ für die Diagonalmatrix mit Einträgen $\lambda_1, ..., \lambda_p$. Weiter ist $A := (\alpha_1, ..., \alpha_p) \in \mathbb{R}^{p\times p}$ die Matrix mit (Spalten-)Vektoren $\alpha_1, ...\alpha_p$. Wir verwenden außerdem noch die Schreibweise $A_q := (\alpha_1, ..., \alpha_q)$ für die $p \times q$-Matrix der ersten q Spalten von A und $A_q^* := (\alpha_{p-q+1}, ..., \alpha_p)$ für die $p \times q$-Matrix der letzten q Spalten von A.

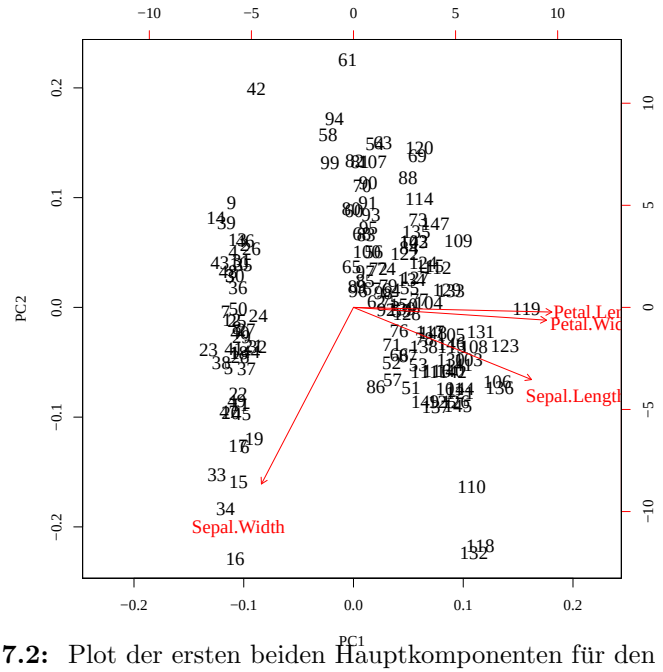


Abbildung 7.2: Plot der ersten beiden Hauptkomponenten für den `iris`-Datensatz.

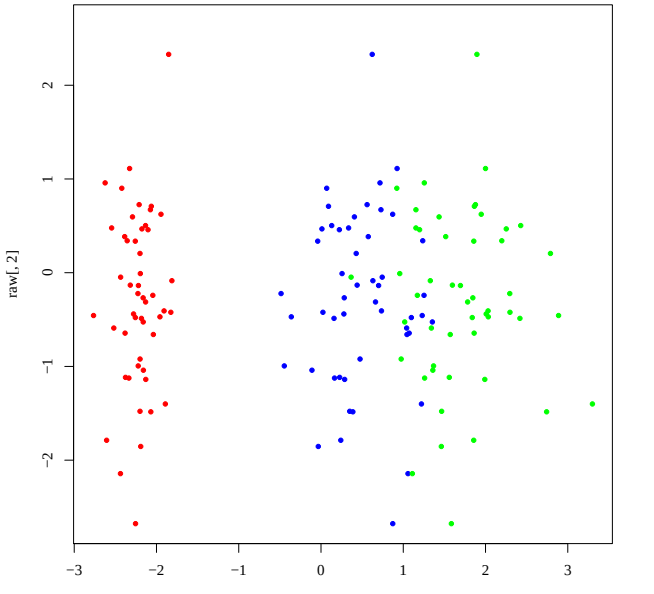


Abbildung 7.3: Gefärbter Plot der ersten beiden Hauptkomponenten für den `iris`-Datensatz.

3. Mit dieser Notation gilt also

$$Z = A^\top X, \qquad z = xA$$

und (da A eine orthogonale Matrix ist)

$$A^\top \Sigma A = A^{-1}\Sigma A = \Lambda \text{ oder } A\Lambda A^\top = \Sigma$$

bzw.

$$A^\top S A = A^{-1} S A = \Lambda \text{ oder } A\Lambda A^\top = S.$$

Bemerkung 7.3 (Covarianzmatrizen und die Spur). Sei $X = (X_1, ..., X_p)$ so verteilt, dass $\mathbb{COV}[X, X] = \Sigma \in \mathbb{R}^{p\times p}$ und $B \in \mathbb{R}^{p\times q}$.

1. Bereits bei Regressionen haben wir folgendes festgestellt:
   Es gilt
   $$\mathbb{COV}[B^\top X] = B^\top \Sigma B.$$
   Insbesondere ist also
   $$\operatorname{tr}(B^\top \Sigma B) = \sum_{j=1}^{p} \mathbb{V}[(B^\top X)_j].$$
   Denn: Sei oBdA $\mathbb{E}[X] = 0$. Wir schreiben
   $$\mathbb{COV}[B^\top X, B^\top X] = \mathbb{E}[B^\top X (B^\top X)^\top] = \mathbb{E}[B^\top X X^\top B] = B^\top \Sigma B.$$

2. Es gilt (mit A wie Bemerkung 7.2)
   $$\mathbb{COV}[A^\top X] = A^\top \Sigma A = \Lambda.$$

Theorem 7.4 (Optimierung der Varianz durch die Hauptkomponenten). *Die Matrix $\Sigma \in \mathbb{R}^{p\times p}$ erfülle die Bedingungen aus Definition 7.1 (insbesondere ist A die Matrix der Eigenvektoren von Σ) und $1 \le q \le p$. Dann gilt*[11]

$$arg\max\{tr(B^\top \Sigma B) : B \in \mathbb{R}^{p\times q} \text{ orthogonal } \} = A_q$$

und

$$arg\min\{tr(B^\top \Sigma B) : B \in \mathbb{R}^{p\times q} \text{ orthogonal } \} = A_q^*.$$

Bemerkung 7.5.

1. Wir bemerken, dass das Theorem sowohl auf eine Covarianzmatrix Σ wie in Definition 7.1.1 anwendbar ist, als auch auf eine Matrix S wie in Definition 7.1.2.

2. Sei X wie in Definition 7.1 und $B \in \mathbb{R}^{p\times q}$. Dann ist für $Y := B^\top X$ nach der letzten Bemerkung gerade $\mathbb{COV}[Y, Y] = B^\top \Sigma B$. Ist nun $q = 1$, so ist also
   $$\mathbb{V}[B^\top X] = \operatorname{tr}(\mathbb{COV}[B^\top X, B^\top X]) = \operatorname{tr}(B^\top \Sigma B).$$
   Nach dem Theorem wird diese gerade dann maximiert, wenn $B = A_1 = \alpha_1$, der Eigenvektor zum größten Eigenwert von Σ. Für $q = 2$ haben wir mittels α_1 bereits $\mathbb{V}[B^\top X]$ maximiert. Das Theorem sagt nun auch, dass der auf α_1 orthogonale und normierte Vektor b, der $\mathbb{V}[b^\top X]$ maximiert, gerade α_2 ist. Genau dies haben wir bereits zu Beginn des Kapitels (zumindest für die empirische Covarianzmatrix S) nachgerechnet.

[11] Wir nennen für $q \le p$ eine Matrix $B \in \mathbb{R}^{p\times q}$ orthogonal, wenn es eine orthogonale Matrix $D \in \mathbb{R}^{p\times p}$, so dass die ersten q Spalten von D und B identisch sind.

Beweis. Sei $B = (\beta_1, ..., \beta_q)$. Da $\alpha_1, ..., \alpha_p$ eine Basis von $\mathbb{R}^p$ bilden, gibt es ein (eindeutiges) $C = (c_1^\top, ..., c_p^\top) \in \mathbb{R}^{p\times q}$ (also $c_j^\top \in \mathbb{R}^q$, $j = 1, ..., p$) mit

$$\beta_k = \sum_{j=1}^{p} c_{jk}\alpha_j, \quad k = 1, ..., q, \quad \text{oder auch} \quad B = AC.$$

Da A, B orthogonal sind, ist auch C orthogonal und

$$\operatorname{tr}(B^\top \Sigma B) = \operatorname{tr}(C^\top A^\top \Sigma A C) = \operatorname{tr}(C^\top \Lambda C) = \sum_{j=1}^{p} \operatorname{tr}(c_j^\top \lambda_j c_j) = \sum_{j=1}^{p} \lambda_j c_j^\top c_j = \sum_{j=1}^{p}\sum_{k=1}^{q} \lambda_j c_{jk}^2. \tag{7.1}$$

Sei nun $D = (d_1^\top, ..., d_p^\top) \in \mathbb{R}^{p\times p}$ orthogonal, so dass die ersten q Spalten von C und D übereinstimmen (also $d_{jk} = c_{jk}, j = 1, ..., p, k = 1, ..., q$). Dann ist

$$\sum_{k=1}^{q} c_{jk}^2 \le \sum_{k=1}^{p} d_{jk}^2 = 1$$

und

$$\sum_{j=1}^{p}\sum_{k=1}^{q} c_{jk}^2 = \sum_{k=1}^{q}\sum_{j=1}^{p} d_{jk}^2 = \sum_{k=1}^{q} 1 = q.$$

Die rechte Seite aus (7.1) ist also sicher dann maximal, wenn man $(c_1, ..., c_p)$ so wählt, dass

$$\sum_{k=1}^{q} c_{jk}^2 = \begin{cases} 1, & j = 1, ..., q, \\ 0, & j = q+1, ..., p. \end{cases} \tag{7.2}$$

Ist aber $B = A_q$, so gilt $C = I_q = (e_1^\top, ..., e_q^\top, 0^\top, ..., 0^\top)$, wobei I_q aus den ersten q Spalten der $p \times p$-Einheitsmatrix besteht und damit gilt hierfür (7.2).

Weiter ist die rechte Seite aus (7.1) sicher dann minimal, wenn man $(c_1, ..., c_p)$ so wählt, dass

$$\sum_{k=1}^{q} c_{jk}^2 = \begin{cases} 0, & j = 1, ..., p-q, \\ 1, & j = p-q+1, ..., p. \end{cases} \tag{7.3}$$

Ist aber $B = A_q^*$, so gilt $C = I_q^* = (0^\top, ..., 0^\top, e_1^\top, ..., e_q^\top)$, wobei I_q^* aus den letzten q Spalten der $p \times p$-Einheitsmatrix besteht und damit gilt hierfür (7.3). □

7.4 Die Hauptkomponentenanalyse in der Regression

Das Regressionsmodell

$$Y = x\beta + \epsilon$$

mit $Y \in \mathbb{R}^n, x \in \mathbb{R}^{n\times p}, \beta \in \mathbb{R}^p$ und $\epsilon \in \mathbb{R}^n$ haben wir bereits kennen gelernt. Wir gehen im Folgenden von den Gauß-Markov-Bedingungen mit $\mathbb{COV}[Y_i, Y_j] = \sigma^2\delta_{ij}$ aus.

Bei der (multiplen) Regression hatten wir uns gefragt, ob die Hypothesen $\beta_k = 0$ verworfen werden können oder nicht, $k = 1, ..., p$. Wünschenswert wäre es natürlich, wenn man die Parameter β möglichst genau schätzen könnte. Und wie das nächste Resultat zeigt, kann man die Parameter am genauesten (im Sinne einer kleinen Varianz) schätzen, wenn man die Covariate x durch die Hauptkomponenten $z = xA$ von x ersetzt.

Theorem 7.6 (Hauptkomponenten in der Regression). *Sei $Y = x\beta + \epsilon$ wie oben, $z = xB$ für $B \in \mathbb{R}^{p\times p}$ orthogonal, also*

$$Y = z\gamma + \epsilon \text{ mit } \gamma = B^\top \beta.$$

Weiter sei $\hat{\gamma} := (z^\top z)^{-1} z^\top Y$ der (übliche) Schätzer für γ. Dann gilt für alle $q \leq p$

$$arg\min \Big\{ \sum_{j=1}^{q} \mathbb{V}_\gamma[\hat{\gamma}_j] : B \in \mathbb{R}^{p\times q} \text{ orthogonal } \Big\} = A_q.$$

Beweis. Die Covarianzmatrix von $\hat{\gamma}$ ist gegeben durch

$$\frac{1}{\sigma^2}\mathbb{COV}_\gamma[\hat{\gamma},\hat{\gamma}] = (z^\top z)^{-1} = (B^\top x^\top x B)^{-1} = B^\top (x^\top x)^{-1} B.$$

Wir müssen nun $\mathrm{tr}(B_q^\top (x^\top x)^{-1} B_q)$ für $q = 1, ..., p$ minimieren. Wenden wir Theorem 7.4 mit $(x^\top x)^{-1}$ anstelle von Σ an, so ist dieses Minimum genau dann gegeben, wenn $B = C_q^*$, wobei C die Eigenvektoren von $(x^\top x)^{-1}$ zu den Eigenwerten in fallender Reihenfolge als Einträge enthält. Nun sind diese Eigenvektoren dieselben[12] wie die von $x^\top x$, allerdings in umgekehrter Reihenfolge. Also können wir $C_q^* = A_q$ wählen, und das Ergebnis ist gezeigt. □

[12] Sei A eine (symmetrische) invertierbare Matrix mit Eigenwerten $\lambda_1, ...\lambda_p$ und zugehörigen Eigenvektoren $\alpha_1, ..., \alpha_p$. Dann hat A^{-1} die Eigenwerte $\lambda_1^{-1}, ..., \lambda_p^{-1}$ mit denselben Eigenvektoren $\alpha_1, ..., \alpha_p$.
Denn: Es gilt $\alpha_i = A^{-1}A\alpha_i = \lambda_i A^{-1}\alpha_i$, also $A^{-1}\alpha_i = \lambda_i^{-1}\alpha_i$.

8 Einführung in die Zeitreihenanalyse

8.1 Einleitung

In diesem Kapitel beschäftigen wir uns mit stochastischen Prozessen, also mit Familien von Zufallsvariablen. Wir werden nur zeit-diskrete Prozesse betrachten, also $(X_t)_{t=1,2,...}$. Wir werden annehmen, dass

$$X_t = m_t + Y_t, \qquad t = 1, 2, ...$$

wobei wir den *Trend* $(m_t)_{t=1,2,...}$ als deterministisch annehmen, und $(Y_t)_{t=1,2,...}$ als stationären, stochastischen Prozess.

Definition 8.1 ((Stationärer, zeitdiskreter) stochastischer Prozess). *1. Ein* stochastischer Prozess *(mit Indexmenge I ist eine Familie von Zufallsvariablen $X = (X_t)_{t\in I}$. Er heißt quadratisch integrierbar, falls $\mathbb{E}[X_t^2] < \infty$ für alle $t \in I$.*

2. Ist I diskret, so heißt der stochastische Prozess zeitdiskret *oder eine* Zeitreihe*. Beispiele sind $I = \mathbb{N}$ und $I = \mathbb{Z}$.*

3. Eine quadratisch integrierbare Zeitreihe X heißt (schwach) stationär, wenn $t \mapsto \mathbb{E}[X_t]$ und $t \mapsto \mathbb{COV}[X_t, X_{t+h}]$ (für alle h) konstante Funktionen sind. In diesem Fall heißt $\mathbb{E}[X_1]$ der Erwartungswert und $h \mapsto \gamma(h) := \mathbb{COV}[X_1, X_{h+1}]$ Autokovarianz-Funktion der Zeitreihe. (Hier darf h auch negativ sein; damit ist dann $\gamma(h) = \gamma(-h)$.)

4. Sei $I = \mathbb{Z}$ und X stationär. Dann ist $B(X_t) := X_{t-1}$ der Shift-Operator.

Im Folgenden sei stets $I = \mathbb{N}$ oder $I = \mathbb{Z}$.

Beispiel 8.2. 1. Sei $X = (X_t)_{t\in I}$ eine Familie unabhängiger und identisch verteilter, quadratisch integrierbarer Zufallsgrößen. Dann ist X stationär und hat die Auto-Kovarianzfunktion $\gamma(h) = \delta_0(h)\mathbb{V}[X_1]$.

2. Sei $Z = (Z_t)_{t\in\mathbb{Z}}$ eine Familie unabhängiger, identisch verteilter und quadratisch integrierbarer Zufallsgrößen mit $\mathbb{E}[Z_0] = 0$ und $X = (X_t)_{t\in\mathbb{Z}}$ mit $X_t = Z_t + aZ_{t-1}$. Dann ist X stationär und hat die Auto-Kovarianzfunktion

$$\gamma(h) = \mathbb{COV}[Z_1 + aZ_0, Z_{h+1} + aZ_h] = \begin{cases} (1+a^2)\mathbb{V}[Z_0], & h = 0, \\ a\mathbb{V}[Z_0], & h = 1, \\ 0, & \text{sonst.} \end{cases}$$

3. Sei $(X_t)_{t=1,2,...}$ eine ergodische Markov-Kette mit stationärer Verteilung ν. Dann ist $(X_t)_{t=1,2,...}$ genau dann eine stationäre Zeitreihe, wenn $X_1 \sim \nu$.

4. Eine stationäre Zeitreihe mit Trend ist ein diskreter stochastischer Prozess $(X_t)_{t=1,2,...}$, der sich als $X_t = m_t + Y_t$ für ein deterministisches $(m_t)_{t=1,2,...}$ und eine stationäre Zeitreihe $(Y_t)_{t=1,2,...}$ schreiben lässt.

8.2 Elimination eines Trends

Für eine Zeitreihe X mit $X_t = m_t + Y_t$ wie in Beispiel 8.2.4 wollen wir nach Beobachtung von $X_1, ..., X_t$ die Größe von X_{t+1} voraussagen. Um dies zu tun, werden wir zunächst $(m_t)_{t=1,2,...}$ schätzen, um anschließend mit $(X_t - m_t)_{t=1,2,...}$ eine stationäre Situation weiter studieren zu können.

Bemerkung 8.3 (Elimination eines Trends). Es gibt verschiedene Arten, einen Trend zu eliminieren. Wir stellen hier drei davon vor:

1. Nimmt man an, dass $m_s = \sum_{i=0}^{k} a_i s^i$, so bleibt nun, die Parameter $a_0, ..., a_k$ so zu schätzen, dass $\sum_{r=0}^{s}(X_r - m_r)^2$ minimiert wird. Dieser Aufgabe haben wir uns jedoch schon beim Thema *Regression* gestellt, denn wir müssen nun

$$\sum_{r=0}^{s}(X_r - w_{r\cdot}a)^2$$

(mit $w_{ri} = r^i$) minimieren. Wir haben gesehen, dass diese Minimierung durch

$$\hat{a} = (w^\top w)^{-1} w^\top X$$

gegeben ist (falls $w^\top w$ invertierbar ist). Damit ist also

$$\hat{m}_s = \sum_{i=0}^{k} \hat{a}_i s^i.$$

2. Für ein $q > 1$ kann man

$$\hat{m}_s := \frac{1}{2q+1} \sum_{r=s-q}^{s+q} X_{r \vee 0 \wedge t}$$

schätzen.

3. Eine Polynom-Funktion erkennt man bekanntlich daran, dass irgendeine Ableitung verschwindet. Deshalb definieren wir eine (Art) Ableitung mittels $BX_s := X_{s-1}$, und dann $\nabla^k X_s := (1 - B)^k X_s$, also etwa

$$\nabla X_s = X_s - X_{s-1}, \qquad \nabla^2 X_s = X_s - 2X_{s-1} + X_{s-2}, ...$$

Ist nun $m_s = \sum_{i=0}^{k} a_i s^i$, dann ist

$$\nabla^k X_s = k! a_k + \nabla^k Y_s,$$

also ein stationärer Prozess mit Erwartungswert $k!a_k$. Zwar können wir durch dieses Vorgehen den Prozess $(m_s)_{s=0,1,2,...}$ nicht schätzen, jedoch haben wir unseren Ausgangsprozess auf einen stationären Prozess zurückgeführt. Den ursprünglichen Prozess erhalten wir dann wieder durch Summation, da etwa

$$X_t = X_1 + \sum_{s=1}^{t} \Delta X_s,$$

$$X_t = X_1 + t\Delta X_1 + \sum_{s=1}^{t} \Delta X_s - \Delta X_1$$

$$= X_1 + t\Delta X_1 + \sum_{s=1}^{t}\sum_{r=1}^{s} \Delta X_r - \Delta X_{r-1} = X_1 + t\Delta X_1 + \sum_{s=1}^{t}\sum_{r=1}^{s} \Delta^2 X_r$$

etc.

8.3 Vorhersage stationärer Prozesse

Nachdem wir nun wissen, wie wir aus einer Zeitreihe einen Trend eliminieren, beschäftigen wir uns mit der Vorhersage in stationären Prozessen. Das bedeutet, dass wir $X_1, ..., X_t$ beobachten und daraus X_{t+1} vorhersagen wollen. Am einfachsten geht dies mit Projektionseigenschaften in Hilbert-Räumen.

Definition 8.4 (Hilbert-Raum). *Sei $(\mathcal{H}, \langle ., . \rangle)$ ein Vektor-Raum, versehen mit einem Skalarprodukt. Ist die Norm $x \mapsto ||x|| := \sqrt{\langle x, x \rangle}$ vollständig, so heißt $(\mathcal{H}, \langle ., . \rangle)$ Hilbert-Raum.*

Für $x, y \in \mathcal{H}$ schreiben wir $x \perp y$ genau dann, wenn $\langle x, y \rangle = 0$ und für $\mathcal{M} \subseteq \mathcal{H}$ schreiben wir

$$\mathcal{M}^\perp := \{y : x \perp y \text{ für alle } x \in \mathcal{M}\}$$

für das orthogonale Komplement von $\mathcal{M}$.

Bemerkung 8.5 (Parallelogramm-Identität etc.).

1. Die Norm $||.||$ auf $\mathcal{H}$ definiert eine Topologie auf $\mathcal{H}$. Eine Folge $x_1, x_2, ... \in \mathcal{H}$ heißt *konvergent* gegen $x \in \mathcal{H}$, falls $||x_n - x|| \xrightarrow{n\to\infty} 0$. (Wir schreiben dann $x_n \xrightarrow{n\to\infty} x$.) Weiter ist $\mathcal{M} \subseteq \mathcal{H}$ abgeschlossen, wenn aus $x_1, x_2, ... \in \mathcal{M}$, $x \in \mathcal{H}$ mit $x_n \xrightarrow{n\to\infty} x$ folgt, dass $x \in \mathcal{M}$.

2. Wie in jedem Vektor-Raum mit Skalarprodukt gilt in einem Hilbert-Raum die Parallelogramm-Identität
$$||x + y||^2 + ||x - y||^2 = 2||x||^2 + 2||y||^2.$$
Denn: Wegen der Bilinearität und Symmetrie des Skalarprodukts erhält man
$$\begin{aligned} ||x + y||^2 + ||x - y||^2 &= \langle x + y, x + y \rangle + \langle x - y, x - y \rangle \\ &= \langle x, x \rangle + 2\langle x, y \rangle + \langle y, y \rangle + \langle x, x \rangle - 2\langle x, y \rangle + \langle y, y \rangle \\ &= 2||x||^2 + 2||y||^2. \end{aligned}$$

3. Ist $x \perp y$, so gilt $||x + y||^2 = ||x||^2 + ||y||^2$.
Denn: Wir berechnen direkt
$$||x + y||^2 = \langle x + y, x + y \rangle = ||x||^2 + ||y||^2 + 2\langle x, y \rangle = ||x||^2 + ||y||^2.$$

4. Das Skalarprodukt ist eine stetige Abbildung.
Denn: Ist $x, x_1, x_2, ... \in \mathcal{H}$ mit $||x_n - x|| \xrightarrow{n\to\infty} 0$, so ist für jedes $y \in \mathcal{H}$
$$|\langle x_n, y \rangle - \langle x, y \rangle|^2 = |\langle x_n - x, y \rangle|^2 \leq ||x_n - x||^2 ||y||^2 \xrightarrow{n\to\infty} 0$$
wegen der Cauchy-Schwartz'schen Ungleichung.

Lemma 8.6 (Orthogonales Komplement ist abgeschlossener Teil-Vektorraum). *Für jedes $\mathcal{M} \subseteq \mathcal{H}$ ist $\mathcal{M}^\perp$ ein abgeschlossener Teil-Vektorraum von $\mathcal{H}$.*

Beweis. Man prüft einfach nach, dass $0 \in \mathcal{M}^\perp$ und dass mit $x_1, x_2 \in \mathcal{M}^\perp$ auch $\lambda x_1 \in \mathcal{M}^\perp$ und $x_1 + x_2 \in \mathcal{M}^\perp$. Damit ist $\mathcal{M}^\perp$ ein Teil-Vektorraum. Ist nun $x_1, x_2, ... \in \mathcal{M}^\perp$ und $||x_n - x|| \xrightarrow{n\to\infty} 0$, so ist für jedes $y \in \mathcal{M}$ auch $\langle x, y \rangle = \lim_{n\to\infty} \langle x_n, y \rangle = 0$ wegen der Stetigkeit des Skalarprodukts. □

Theorem 8.7 (Projektionstheorem). *Sei $(\mathcal{H}, \langle ., . \rangle)$ ein Hilbert-Raum, $\mathcal{M} \subseteq \mathcal{H}$ ein abgeschlossener Teil-Vektorraum und $x \in \mathcal{H}$.*

1. *Es gibt ein eindeutiges $\hat{x} \in \mathcal{M}$, so dass*

$$||x - \hat{x}|| = \inf_{y \in \mathcal{M}} ||x - y||. \qquad (*)$$

2. *Sei $\hat{x} \in \mathcal{M}$. Dann gilt $(*)$ genau dann, wenn $(x - \hat{x}) \in \mathcal{M}^{\perp}$.*

Beweis. 1. Zunächst zur Existenz. Sei $d := \inf_{y \in \mathcal{M}} ||x-y||$. Dann gibt es eine Folge $y_1, y_2, ... \in \mathcal{M}$ mit $||x - y_n|| \xrightarrow{n \to \infty} d$. Nun gilt mit der Parallelogrammidentität

$$\begin{aligned} 0 \leq ||y_m - y_n||^2 &= -||(y_m + y_n) - 2x||^2 + 2||y_n - x||^2 + 2||y_m - x||^2 \\ &\leq -4d^2 + 2(|y_n - x||^2 + ||y_m - x||^2) \xrightarrow{m,n \to \infty} 0. \end{aligned}$$

Damit ist $y_1, y_2, ...$ eine Cauchy-Folge und es gibt $\hat{x} \in \mathcal{H}$ mit $||y_n - \hat{x}|| \xrightarrow{n \to \infty} 0$. Da $\mathcal{M}$ abgeschlossen ist, ist $\hat{x} \in \mathcal{M}$ und

$$||x - \hat{x}|| = \lim_{n \to \infty} ||x - y_n|| = d.$$

Für die Eindeutigkeit nehmen wir an, es gäbe $\hat{y} \in \mathcal{M}$, so dass $(*)$ auch für $\hat{y}$ gilt. Dann ist

$$0 \leq ||\hat{x} - \hat{y}||^2 = -||(\hat{x} + \hat{y}) - 2x||^2 + 2(||\hat{x} - x||^2 + ||\hat{y} - x||^2) \leq -4d^2 + 4d^2 = 0,$$

also $\hat{x} = \hat{y}$.
2. $\Leftarrow$: Für jedes $y \in \mathcal{M}$ gilt

$$||x - y||^2 = \langle x - \hat{x} + \hat{x} - y, x - \hat{x} + \hat{x} - y \rangle = ||x - \hat{x}||^2 + ||\hat{x} - y||^2 \geq ||x - \hat{x}||^2$$

woraus $(*)$ folgt.
$\Rightarrow$: Sei $(x - \hat{x}) \notin \mathcal{M}^{\perp}$. Dann gibt es $y \in \mathcal{M}$ mit $a := \langle x - \hat{x}, y \rangle > 0$. Dann ist für $\tilde{x} := \hat{x} + ay/||y||^2 \in \mathcal{M}$

$$\begin{aligned} ||x - \tilde{x}||^2 &= \langle x - \hat{x} + \hat{x} - \tilde{x}, x - \hat{x} + \hat{x} - \tilde{x} \rangle \\ &= ||x - \hat{x}||^2 + a^2/||y||^2 - 2a\langle y/||y||^2, x - \hat{x} \rangle = ||x - \hat{x}||^2 - a^2/||y||^2 < ||x - \hat{x}||^2. \end{aligned}$$

Damit erfüllt $\hat{x}$ also nicht $(*)$ und die Behauptung ist gezeigt. □

Definition 8.8 (Vorhersage-Gleichungen). *Gegeben $x \in \mathcal{H}$ und einen abgeschlossenen Teil-Vektorraum $\mathcal{M} \subseteq \mathcal{H}$, lauten die Vorhersagegleichungen*

$$\langle x - \hat{x}, y \rangle = 0 \text{ für alle } y \in \mathcal{M},$$

die man nach $\hat{x} \in \mathcal{M}$ auflösen muss. In diesem Sinne ist also $\hat{x} \in \mathcal{M}$ eine Vorhersage für $x \in \mathcal{H}$ (unter der Bedingung, dass $||x - \hat{x}||$ minimal ist). Man schreibt auch $\hat{x} = \mathcal{P}_{\mathcal{M}} x$, wobei man $\mathcal{P}_{\mathcal{M}}$ als Projektionsoperator bezeichnet.

Ist span$(y_1, y_2, ...) = \mathcal{M}$, so sind die Vorhersagegleichungen genau dann erfüllt, wenn

$$\langle x - \hat{x}, y_n \rangle = 0 \text{ für alle } n = 1, 2, ...$$

Beispiel 8.9 (Regression). Bei der Regression hatten wir (für $k < n$) Daten $x_1, ..., x_n \in \mathbb{R}^{k+1}$ und $y_1, ..., y_n \in \mathbb{R}$. Gesucht war $\hat{\beta} \in \mathbb{R}^{k+1}$, so dass $||y - x\beta||$ minimal wird (wobei $x = (x_1, ..., x_n)^\top$).
Sei also $\mathcal{H} = \mathbb{R}^n$ und $\mathcal{M} = \text{span}(x_{\cdot 0}, ..., x_{\cdot k})$ ein (maximal) k+1-dimensionaler, abgeschlossener Unter-Vektorraum. Da wir also $\inf_{z \in \mathcal{M}} ||z - x||$ bestimmen wollen, wird dies nach Theorem genau von dem $\hat{y} \in \mathcal{M}$ gelöst, für das

$$\langle y - \hat{y}, x_{\cdot i} \rangle = 0, i = 0, ..., k$$

gilt. Mit $\hat{y} = x\hat{\beta}$ muss also $x^\top y = x^\top x\hat{\beta}$ gelten. Ist $x^\top x$ invertierbar, bedeutet dies, dass

$$\hat{\beta} = (x^\top x)^{-1} x^\top y.$$

Genau dasselbe Ergebnis haben wir bereits in Theorem 2.3 des Regressions-Skriptes erhalten.

8.4 Vorhersage von stationären Zeitreihen

In diesem Kapitel gehen wir von einem stationären, stochastischen Prozess $(X_s)_{s=1,2,...}$ mit $\mathbb{E}[X_t^2] < \infty$ aus, den wir für Zeiten $s = 1, ..., t$ beobachtet haben. Wir wollen durch die Beobachtungen $X_1, ..., X_t$ den Wert X_{t+1} vorhersagen.

Proposition 8.10 (Vorhersage von stationären Prozessen). *Sei $\mathcal{L}^2(\mathbb{P})$ der Hilbert-Raum der quadratisch integrierbaren Zufallsvariablen, versehen mit dem Skalarprodukt $\langle X, Y \rangle := \mathbb{E}[XY]$. Sei $(X_s)_{s=1,2,...}$ ein stationärer stochastischer Prozess mit Erwartungswert 0 und Auto-Kovarianzfunktion γ. Weiter sei $\mathcal{M} := span(X_1, ..., X_t)$. Sei $\phi_{11}, ..., \phi_{tt} \in \mathbb{R}$ so, dass*

$$\sum_{s=1}^{t} \phi_{ts} \gamma(s - r) = \gamma(r), r = 1, ..., t, \text{ oder } \Gamma_t \Phi_t = \gamma_t$$

mit $\Gamma_t := (\gamma(s - r))_{r,s=1,...,t}, \Phi_t := (\phi_{ts})_{s=1,...,t}, \gamma_t := (\gamma(s))_{s=1,...,t}$. Dann ist

$$\hat{X}_{t+1} := \sum_{s=1}^{t} \phi_{ts} X_{t+1-s} \tag{8.1}$$

die Projektion von X_{t+1} auf $\mathcal{M}$ (und damit die beste lineare Vorhersage von X_{t+1} gegeben $X_1, ..., X_t$).

Beweis. Es gilt zu zeigen, dass das angegebene $\hat{X}_{t+1}$ die Vorhersagegleichungen erfüllt. Diese sind für $r = 1, ..., t$

$$0 = \langle X_{t+1} - \hat{X}_{t+1}, X_{t+1-r} \rangle = \left\langle X_{t+1} - \sum_{s=1}^{t} \phi_{ts} X_{t+1-s}, X_{t+1-r} \right\rangle = \gamma(r) - \sum_{s=1}^{t} \phi_{ts} \gamma(s - r).$$

Daraus folgt bereits die Behauptung. □

Ist Γ_t in der obigen Proposition für alle t invertierbar, so berechnet sich die Vorhersage $\hat{X}_{t+1}$ mit $\Phi_t = \Gamma_t^{-1} \gamma_t$ und (8.1). Wir geben nun einen Algorithmus an, mit dem man diese Vorhersagen rekursiv berechnen kann. Der Vorteil dabei ist, dass man bei der Vorhersage des nächsten Wertes der Zeitreihe auf bereits berechnetes zurückgreifen kann.

Theorem 8.11 (Der Innovations-Algorithmus). *Sei $(X_s)_{s=1,2,...}$ ein stationärer stochastischer Prozess mit Erwartungswert 0 und Auto-Kovarianzfunktion γ. Sei Γ_t aus Proposition 8.10 für alle t invertierbar. Dann ist die Vorhersage $\hat{X}_{t+1}$ und $v_{n+1} := ||X_{n+1} - \hat{X}_{n+1}||^2$ gegeben durch die Rekursion*

$$\hat{X}_{t+1} = \begin{cases} 0, & t = 0, \\ \sum_{s=1}^{t} \theta_{ts}(X_{t+1-s} - \hat{X}_{t+1-s}), & t \geq 1 \end{cases} \tag{8.2}$$

mit

$$\begin{aligned} v_1 &:= \gamma(0), \\ \theta_{t,t-s} &:= v_{s+1}^{-1}\Big(\gamma(t-s) - \sum_{r=0}^{s-1} \theta_{s,s-r}\theta_{t,t-r}v_{r+1}\Big), \qquad s = 0, ..., t-1, \\ v_{t+1} &:= \Gamma(t+1, t+1) - \sum_{r=0}^{t-1} \theta_{t,t-r}^2 v_{r+1}. \end{aligned}$$

Bemerkung 8.12. Man überzeugt sich leicht davon, dass die Parameter θ_{ts} und v_t in der Reihenfolge $v_1; \theta_{11}, v_2; \theta_{22}, \theta_{21}, v_3; \theta_{33}, \theta_{32}, \theta_{31}, v_4; ...$ rekursiv berechnet werden können.

Beweis. Das System $(X_1 - \hat{X}_1, ..., X_t - \hat{X}_t)$ ist orthogonal, da $X_r - \hat{X}_r \in \text{span}(X_1, ..., X_{r-1})^\perp = \text{span}(X_1 - \hat{X}_1, ..., X_{r-1} - \hat{X}_{r-1})^\perp$ für $r = 1, ..., t$ gilt. Nimmt man in (8.2) das Skalarprodukt mit $X_{s+1} - \hat{X}_{s+1}$, so erhält man für $s = 0, ..., t-1$

$$\langle \hat{X}_{t+1}, X_{s+1} - \hat{X}_{s+1} \rangle = \sum_{r=1}^{t} \theta_{tr} \langle X_{t+1-r} - \hat{X}_{t+1-r}, X_{s+1} - \hat{X}_{s+1} \rangle = \theta_{t,t-s} v_{s+1}.$$

Da $(X_{t+1} - \hat{X}_{t+1}) \perp (X_{s+1} - \hat{X}_{s+1})$, sieht man, dass

$$\begin{aligned} \theta_{t,t-s} &= v_{s+1}^{-1} \langle X_{t+1}, X_{s+1} - \hat{X}_{s+1} \rangle \\ &= v_{s+1}^{-1}\Big(\gamma(t-s) - \sum_{r=1}^{s} \theta_{sr} \langle X_{t+1}, X_{s+1-r} - \hat{X}_{s+1-r} \rangle\Big) \\ &= v_{s+1}^{-1}\Big(\gamma(t-s) - \sum_{r=0}^{s-1} \theta_{s,s-r} \langle \hat{X}_{t+1}, X_{r+1} - \hat{X}_{r+1} \rangle\Big) \\ &= v_{s+1}^{-1}\Big(\gamma(t-s) - \sum_{r=0}^{s-1} \theta_{s,s-r}\theta_{t,t-r}v_{r+1}\Big). \end{aligned}$$

Da $\hat{X}_{t+1} \perp (X_{t+1} - \hat{X}_{t+1})$, ist $||X_{t+1}||^2 = ||\hat{X}_{t+1}||^2 + ||X_{t+1} - \hat{X}_{t+1}||^2$ und damit

$$\begin{aligned} v_{t+1} = ||X_{t+1} - \hat{X}_{t+1}||^2 &= ||X_{t+1}||^2 - ||\hat{X}_{t+1}||^2 = \gamma(0) - \sum_{r=1}^{t} \theta_{tr}^2 v_{t+1-r} \\ &= \gamma(0) - \sum_{r=0}^{t-1} \theta_{t,t-r}^2 v_{r+1}. \end{aligned}$$

Damit sind alle Aussagen gezeigt. □

8.5 AR(I)MA-Prozesse

Wir behandeln nun eine äußerst wichtige Klasse von Zeitreihen.

Definition 8.13 (AR(I)MA-Prozess). *Sei $X = (X_t)_{t\in\mathbb{Z}}$ eine Zeitreihe und B der Shift-Operator aus Definition 8.1.*

1. *Die Zeitreihe X heißt ARMA-Prozess (der Ordnung $p, q \in \mathbb{N}$), falls X stationär ist und es eine unabhängige Familie $Z = (Z_t)_{t\in\mathbb{Z}}$ gibt mit $Z_t \sim \mathcal{N}(0, \sigma^2)$, sowie $\phi_1, ..., \phi_p, \theta_1, ..., \theta_q \in \mathbb{R}$ mit*

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \qquad t \in \mathbb{Z}.$$

   Wir schreiben für diese Gleichungen auch

$$\phi(B)X_t = \theta(B)Z_t, \quad t \in \mathbb{Z}$$

   mit

$$\phi(x) := 1 - \phi_1 x - \cdots - \phi_p x^p, \qquad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

   (mit $B^i(X) := X_{t-i}$). Hier steht AR für Auto-Regressive *und MA für* Moving Average.

2. *Ein ARMA-Prozess X heißt* kausal*, wenn es eine Folge $\psi_0, \psi_1, ... \in \mathbb{R}$ gibt mit $\sum_{s=0}^{\infty} |\psi_s| < \infty$ und*

$$X_t = \sum_{s=0}^{\infty} \psi_s Z_{t-s}.$$

3. *X heißt ARIMA-Prozess (der Ordnung $p, d, q \in \mathbb{N}$), falls $\Delta^d X$ ein kausaler ARMA-Prozess der Ordnung p, q ist.*

Beispiel 8.14 (MA-Prozess). Ist $\phi = 1$ und $\theta_1, ..., \theta_q$ wie oben, also

$$X_t = \theta(B)Z_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

so sprechen wir von einem AR-Prozess der Ordnung q. Dieser Prozess ist immer stationär, da (mit $\theta_0 := 0$) für $h = 0, 1, 2, ...$

$$\mathbb{E}[X_t] = \sum_{i=0}^{q} \theta_i \mathbb{E}[Z_{t-i}] = 0,$$

$$\mathbb{COV}[X_t, X_{t+h}] = \sum_{i=0}^{q}\sum_{j=0}^{q} \theta_i\theta_j \mathbb{COV}[Z_{t-i}Z_{t+h-j}] = \begin{cases} \sum_{i=0}^{q-h} \theta_i\theta_{i+h}\sigma^2, & h = 0, ..., q, \\ 0, & \text{sonst.} \end{cases}$$

Außerdem ist X trivialerweise immer kausal.

Beispiel 8.15 (AR-Prozess). Ist $\theta = 1$ und $\phi_1, ..., \phi_p$ wie oben, also

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$$

und ist X stationär, so sprechen wir von einem MA-Prozess der Ordnung p.

Als Beispiel betrachten wir $p = 1$ mit $|\phi_1| < 1$. Es gilt also iterativ für $n = 1, 2, ...$

$$\begin{aligned} X_t &= Z_t + \phi_1 X_{t-1} = Z_t + \phi_1 Z_{t-1} + \phi_1^2 X_{t-2} = \cdots \\ &= Z_t + \phi_1 Z_{t-1} + \cdots + \phi_1^s Z_{t-s} + \phi_1^{s+1} X_{t-s-1}. \end{aligned}$$

Ist nun X stationär, so ist $t \mapsto \mathbb{E}[X_t^2]$ konstant, und falls $\mathbb{E}[X_0^2] < \infty$ folgt

$$\mathbb{E}\Big[\Big(X_t - \sum_{s=0}^{t} \phi_1^s Z_{t-s}\Big)^2\Big] = \phi_1^{t+1} \mathbb{E}[X_0^2] \xrightarrow{t\to\infty} 0.$$

Also folgt (zumindest im L^2-Sinne)

$$X_t = \sum_{s=0}^{\infty} \phi_1^s Z_{t-s}, \qquad (*)$$

also ist X kausal.
Ist also andersherum X_t durch die letzte Gleichung gegeben (mit $|\phi_1| < 1$), so ist für $h = 0, 1, 2, ...$

$$\begin{aligned} \mathbb{E}[X_t] &= \sum_{s=0}^{\infty} \phi_1^s \mathbb{E}[Z_{t-s}] = 0, \\ \mathbb{COV}[X_t, X_{t+h}] &= \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \phi_1^r \phi_1^s \mathbb{COV}[Z_{t-r}, Z_{t+h-s}] =, \sigma^2 \sum_{r=0}^{\infty} \phi_1^{2r+h} = \sigma^2 \frac{\phi_1^h}{1-\phi_1^2}. \end{aligned}$$

Also ist dann X stationär.

Proposition 8.16 (Berechnung von γ). *Sei X ein kausaler ARMA-Prozess mit $\phi(B)X_t = \theta(B)Z_t$ und $X_t = \sum_{s=0}^{\infty} \psi_s Z_{t-s}$, $t \in \mathbb{Z}$. Dann gilt*

$$\gamma(h) - \sum_{s=1}^{p} \phi_s \gamma(h-s) = \begin{cases} \sigma^2 \sum_{j=h}^{q} \theta_j \psi_{j-h}, & h = 0, ..., q, \\ 0, & h = q+1, q+2, ... \end{cases}$$

Beweis. Geht man von den Gleichungen $\phi(B)X_t = \theta(B)Z_t$ aus, und bildet auf beiden Seiten das Skalarprodukt mit X_{t-h}, so erhält man direkt

$$\begin{aligned} \gamma(h) - \phi_1 \gamma(h-1) - ... - \phi_p \gamma(h-p) &= \sum_{j=1}^{q} \sum_{s=0}^{\infty} \theta_j \psi_s \langle Z_{t-h-s}, Z_{t-j} \rangle \\ &= \sigma^2 \sum_{j=1}^{q} \theta_j \psi_{j-h} 1_{j \geq h} = \sigma^2 \sum_{j=h}^{q} \theta_j \psi_{j-h}. \end{aligned}$$

□

Bemerkung 8.17 (Numerische Berechnung von γ). Die Gleichungen der letzten Proposition kann man verwenden, um rekursiv die Funktion γ zu berechnen. Zunächst stellt man die Gleichungen für $h = 0, ..., p$ auf. Die linken Seiten hängen für diese $p+1$ Gleichungen nur von $\gamma(0), ..., \gamma(p)$ ab (wegen der Symmetrie $\gamma(j) = \gamma(-j)$). Löst man dieses lineare Gleichungssystem auf, so kann man anschließend die Werte für $\gamma(p+1), \gamma(p+2), ...$ rekursiv berechnen.

8.6 Zeitreihen mit R

Wir generieren zunächst eine (kausale, stationäre) AR-Zeitreihe der Ordnung $p = 1$ und plotten diese; siehe Figur 8.1. Der Befehl `ts` macht aus dem Vektor `x` eine Zeitreihe, für die R im Folgenden weitere Befehle bereitstellt.

```
end<-200
x<-rep(0,end)
z<-rnorm(end)
phi1<-0.9
for(i in 2:end) {
  x[i]<-z[i] + phi1 * x[i-1]
}
dat<-ts(x)
plot(dat, type='l')
```

Um Vorhersagen in Zeitreihen machen zu können, bietet sich das Paket `forecast` an,[13] das wir mit

```
install.packages("forecast")
library("forecast")
```

laden. Zwar wissen wir alle Parameter unserer Zeitreihe ($p = 1, q = 0, \phi_1 = 0.9$), aber R stellt auch eine Funktion zur Schätzung der Parameter bereit (deren Funktion wir nicht besprechen werden). Mit

```
>auto.arima(dat, d=0)
Series: dat
ARIMA(1,0,0) with zero mean

Coefficients:
         ar1
      0.9022
s.e.  0.0292

sigma^2 estimated as 0.8581:  log likelihood=-269.32
AIC=542.65   AICc=542.71   BIC=549.25
```

sehen wir, dass R einen ARMA-Prozess der Ordnung $(1, 0)$ mit $\phi_1 = 0.9022$ vorschlägt.[14]

Nun veranschaulichen wir noch, wie man eine Vorhersage des i-ten Datenpunktes macht und grafisch mit der echten Zeitreihe vergleichen kann.

[13] Übrigens hätte es hier auch ein Simulationstool für ARMA-Modelle gegeben. Obige Zeitreihe hätten wir auch mit

```
plot(arima.sim(list(ar=(0.9)), n=200))}
```

simulieren können.

[14] Wir hätten auch

```
>arima(x = dat, order = c(1, 0, 0), include.mean = FALSE)
Call:
```

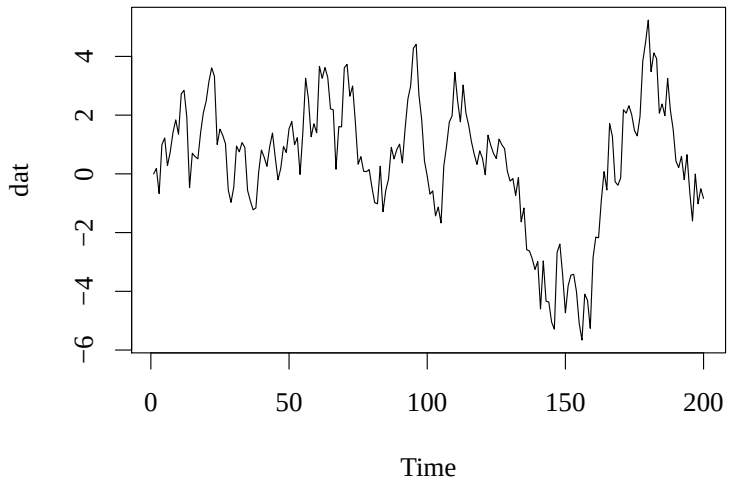


Abbildung 8.1: Eine AR-Zeitreihe.

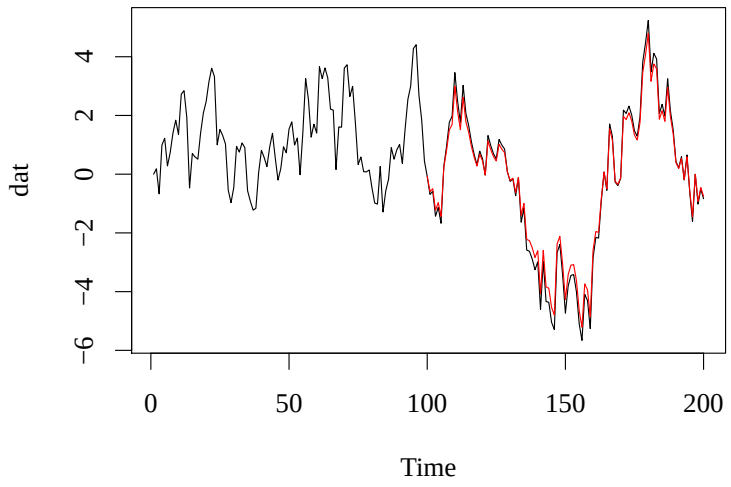


Abbildung 8.2: Eine AR-Zeitreihe und ihre Vorhersage ab $t = 100$ (basierend auf den vergangenen Daten).

```
datloc<-rep(0,200)
for(i in 100:end) {
  fit<-arima(dat[1:i], order=c(1,0,0))
  datloc[i]<-predict(fit, n.ahead=1)$pred[1]
}
plot(arima.sim(list(ar=(0.9)), n=200))}
lines(100:end, datloc(100:end), col="red")
```

```
arima(x = dat, order = c(1, 0, 0), include.mean = FALSE)

Coefficients:
         ar1
      0.9022
s.e.  0.0292

sigma^2 estimated as 0.8581:  log likelihood = -269.32,  aic = 542.65
```

verwenden können und damit die Ordnung vorgeben.

Literatur

[Bro91] P. J. Brockwell, R. A. Davis. Time Series: Theory and Methods. Second Edition. *Springer*, 1991.

[Fah07] L. Fahrmeir, T. Kneib and S. Lang. Regression. Modelle, Methoden und Anwendungen. *Springer*, 2. Auflage, 2009.

[GC03] J.-D. Gibbons, S. Chakraborti. Nonparametric Statistical Inference. Fourth Edition. DEKKER Series, 2003.

[HTF08] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. *Springer*, 2008.

[Jol02] I. T. Jolliffe. Principal Component Analysis. Second Edition *Springer*, 2002.

[KN06] J.-P. Kreiß, G. Neuhaus. Einführung in die Zeitreihenanalyse. *Springer*, 2006.

[R] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`, 2013.

[Sen90] A. Sen and M. Srivastava. Regression Analysis. Theory, Methods, and Applications. *Springer*, 1990.

[ST95] J. Shao, D. Tu. The Jackknife and Bootstrap. *Springer*, 1995.