Mathematische Statistik

Mathias Vetter

10. Februar 2016

Inhaltsverzeichnis

Kapitel 1

Bedingte Erwartungen

Wir interessieren uns in diesem Kapitel für bedingte Erwartungen, die eine wesentliche Grundlage nicht nur für die Analyse statistischer Verfahren darstellen, sondern auch in der Stochastik insgesamt von großer Bedeutung sind.

Bemerkung 1.1. Wir betrachten im Folgenden eine integrierbare Zufallsvariable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$. Diese modelliert ein zufälliges Experiment, bei dem reelle Messwerte beobachtet werden. Entsprechend enthält die Abbildung

$$\omega \mapsto X(\omega)$$

nach Durchführung des Experiments und Beobachtung des Ergebnisses die volle Information über dessen Ausgang.

Vor Beginn des Experiments sind dagegen noch keinerlei Informationen über dessen Ausgang bekannt. Die beste Annäherung an X ist sein Erwartungswert, d.h. die konstante Abbildung

$$\omega \mapsto \mathbb{E}[X].$$

Ausgangspunkt für den Begriff der bedingten Erwartung ist die Frage, wie man den Grad der Information über X misst, wenn nur ein Teilausgang des Experiments bekannt ist.

Definition 1.2. Es seien $(\Omega, \mathcal{A})$ ein Messraum mit Maßen μ und ν. Dann heißt ν *absolut stetig* bzgl. μ (oder μ *dominierend* zu ν), wenn jede μ-Nullmenge auch eine ν-Nullmenge ist, d.h.

$$\mu(A) = 0 \implies \nu(A) = 0 \quad \forall A \in \mathcal{A}.$$

Man schreibt: $\nu \ll \mu$.

Satz 1.3. (Radon-Nikodým) *Es seien $(\Omega, \mathcal{A})$ ein Messraum mit Maßen μ und ν. Ist μ σ-endlich, dann sind äquivalent:*

(i) ν ist absolut stetig bzgl. μ.

(ii) ν besitzt eine Dichte bzgl. μ, d.h. es existiert eine $\mathcal{A}$-messbare, nicht-negative Funktion f mit

$$\nu(A) = \int_A f d\mu \quad \forall A \in \mathcal{A}.$$

Als Notation für die Dichte von ν bzgl. μ verwendet man oft

$$f = \frac{d\nu}{d\mu}.$$

Beweis: vgl. Satz 17.10 in Bauer (1992). □

Definition 1.4. Es sei $\mathcal{F} \subset \mathcal{A}$ eine Unter-σ-Algebra. Eine Zufallsvariable $Y : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$ heißt bedingte Erwartung von X gegeben $\mathcal{F}$, falls gilt:

(i) Y ist $\mathcal{F}$-$\mathcal{B}$-messbar (kurz: $\mathcal{F}$-messbar), d.h. $Y^{-1}(B) \in \mathcal{F}$ für alle $B \in \mathcal{B}$,

(ii) für jedes $A \in \mathcal{F}$ ist
$$\mathbb{E}[X1_A] = \mathbb{E}[Y1_A].$$

Satz 1.5. *Die bedingte Erwartung von X gegeben $\mathcal{F}$ existiert und ist fast sicher eindeutig.*

Beweis:

Eindeutigkeit: Es seien Y und Y' Zufallsvariablen, die (i) und (ii) aus Definition 1.4 erfüllen. Wir setzen dann $A = \{Y > Y'\}$ und wissen aus (i), dass $A \in \mathcal{F}$ gilt. Dann folgt aus (ii)

$$0 = \mathbb{E}[Y1_A] - \mathbb{E}[Y'1_A] = \mathbb{E}[(Y - Y')1_A].$$

Wegen $(Y - Y')1_A \geq 0$ folgt $\mathbb{P}(A) = 0$. Analog erhält man $\mathbb{P}(Y < Y') = 0$.

Existenz: Wir zerlegen gemäß $X = X^+ - X^-$ die Zufallsvariable X in Positiv- und Negativteil. Mittels
$$\mathbb{Q}^{\pm}(A) = \mathbb{E}[X^{\pm}1_A], \quad A \in \mathcal{F},$$
lassen sich zwei Maße auf $(\Omega, \mathcal{F})$ definieren. Offenbar sind beide Maße absolut stetig bzgl. $\mathbb{P}$. Nach Satz 1.3 existieren dann $\mathcal{F}$-messbare Dichten $Y^{\pm}$ mit

$$\mathbb{Q}^{\pm}(A) = \int_A Y^{\pm} d\mathbb{P} = \mathbb{E}[Y^{\pm}1_A].$$

Offenbar erfüllt $Y = Y^+ - Y^-$ die definierenden Eigenschaften der bedingten Erwartung. □

Bemerkung 1.6. Wir verwenden im Folgenden für die bedingte Erwartung von X gegeben $\mathcal{F}$ die Bezeichnung $Y = \mathbb{E}[X|\mathcal{F}]$, die nach Satz 1.5 im Sinne $\mathbb{P}$-fast sicherer Gleichheit zu verstehen ist. Ist $X = 1_C$ für ein $C \in \mathcal{A}$, so heißt

$$\mathbb{P}(C|\mathcal{F}) = \mathbb{E}[1_C|\mathcal{F}]$$

die bedingte Wahrscheinlichkeit von C gegeben $\mathcal{F}$.

Definition 1.7. Ist Y eine weitere (nicht notwendigerweise integrierbare) Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$, so setzen wir

$$\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)],$$

wobei $\sigma(Y)$ die von Y erzeugte σ-Algebra bezeichnet.

Lemma 1.8. *Es seien $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $\mathcal{F} \subset \mathcal{A}$ eine Unter-σ-Algebra und X, Y reellwertige, integrierbare Zufallsvariablen. Dann gilt:*

(i) Für alle $\alpha, \beta \in \mathbb{R}$ ist

$$\mathbb{E}[\alpha X + \beta Y|\mathcal{F}] = \alpha\mathbb{E}[X|\mathcal{F}] + \beta\mathbb{E}[Y|\mathcal{F}]. \tag{1.1}$$

(ii) Gilt $X \leq Y$ $\mathbb{P}$-f.s., so auch $\mathbb{E}[X|\mathcal{F}] \leq \mathbb{E}[Y|\mathcal{F}]$ $\mathbb{P}$-f.s.

Beweis:

(i) Die rechte Seite von (1.1) ist $\mathcal{F}$-messbar, und für $A \in \mathcal{F}$ gilt aufgrund der Linearität der (unbedingten) Erwartung:

$$\begin{aligned}\mathbb{E}[(\alpha X + \beta Y)1_A] &= \mathbb{E}[\alpha 1_A X + \beta 1_A Y] = \alpha\mathbb{E}[1_A X] + \beta\mathbb{E}[1_A Y] \\ &= \alpha\mathbb{E}[1_A\mathbb{E}[X|\mathcal{F}]] + \beta\mathbb{E}[1_A\mathbb{E}[Y|\mathcal{F}]] \\ &= \mathbb{E}\left[1_A\left(\alpha\mathbb{E}[X|\mathcal{F}]] + \beta\mathbb{E}[Y|\mathcal{F}]\right)\right].\end{aligned}$$

Dabei haben wir im vorletzten Schritt die definierende Eigenschaft (ii) aus Definition 1.4 verwendet.

(ii) Dieser Beweis wird als Übungsaufgabe geführt. □

Beispiel 1.9.

(i) Ist $\mathcal{F} = \{\emptyset, \Omega\}$, so ist $\mathbb{E}[X|\mathcal{F}]$ nach Definition 1.4 (i) konstant, und mit der Wahl von $A = \Omega$ in (ii) folgt

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

Dies entspricht dem Fall, dass keine zusätzlichen Informationen über den Ausgang des Experiments verfügbar sind.

(ii) Ist $\sigma(X) \subset \mathcal{F}$, so erfüllt X selbst bereits Definition 1.4 (i). Man erhält $\mathbb{E}[X|\mathcal{F}] = X$, da X Eigenschaft (ii) immer erfüllt.

(iii) Es sei $\mathcal{F} = \sigma(A) = \{\Omega, A, A^c, \emptyset\}$ für ein $A \in \mathcal{A}$ mit $\mathbb{P}(A) \in (0, 1)$. Wegen der $\mathcal{F}$-Messbarkeit muss $\mathbb{E}[X|A] = \mathbb{E}[X|\sigma(A)]$ auf A und A^c jeweils konstant sein. Es folgt aus Definition 1.4 (ii)

$$\mathbb{E}[X|A] = \left(\mathbb{P}(A)^{-1}\int_A X d\mathbb{P}\right)1_A + \left(\mathbb{P}(A^c)^{-1}\int_{A^c} X d\mathbb{P}\right)1_{A^c}.$$

Allgemein gilt für $\Omega = A_1 \cup \ldots A_n$ disjunkt mit $\mathbb{P}(A_i) > 0$ für alle i die Identität

$$\mathbb{E}[X|A_1, \ldots, A_n] = \sum_{i=1}^{n}\left(\mathbb{P}(A_i)^{-1}\int_{A_i} X d\mathbb{P}\right)1_{A_i}.$$

Satz 1.10. (monotone Konvergenz) *Es seien $X_n \geq 0$ integrierbare Zufallsvariablen mit $X_n \nearrow X$. Dann existiert eine $\mathcal{F}$-messbare Zufallsvariable Y, so dass $\mathbb{E}[X_n|\mathcal{F}] \nearrow Y$. Insbesondere erfüllt diese*

$$\mathbb{E}[X1_A] = \mathbb{E}[Y1_A] \quad \forall A \in \mathcal{F}$$

und stimmt im Fall der Integrierbarkeit von X mit der üblichen bedingten Erwartung überein.

Beweis: Wegen Lemma 1.8 (ii) ist $\mathbb{E}[X_n|\mathcal{F}]$ monoton wachsend. Also existiert Y und ist als punktweiser Grenzwert von $\mathcal{F}$-messbaren Funktionen selbst wieder $\mathcal{F}$-messbar. Zudem folgt aus $X_n \nearrow X$ auch $X_n 1_A \nearrow X 1_A$, und wir erhalten aus dem klassischen Satz von der monotonen Konvergenz

$$\mathbb{E}[X1_A] = \lim_{n\to\infty} \mathbb{E}[X_n 1_A] = \lim_{n\to\infty} \mathbb{E}[\mathbb{E}[X_n|\mathcal{F}]1_A] = \mathbb{E}[Y1_A].$$

□

Bemerkung 1.11.

(i) Satz 1.10 zeigt, dass bedingte Erwartungen auch allgemein für nicht-negative Zufallsvariablen X existieren, indem man mittels $X_n = \min(X, n)$ approximiert.

(ii) Lemma 1.8 und Satz 1.10 liefern nur einige Eigenschaften von Erwartungswerten, die sich auf bedingte Erwartungen verallgemeinern lassen. Weitere Beispiele sind etwa die Cauchy-Schwarz- und die Jensen-Ungleichung sowie eine Variante des Satzes von der majorisierten Konvergenz.

Satz 1.12. (Faktorisierung) *Sind X und Y Zufallsvariablen mit $\mathbb{E}[|X|] < \infty$ und $\mathbb{E}[|XY|] < \infty$ und ist Y $\mathcal{F}$-messbar, so gilt*

$$\mathbb{E}[XY|\mathcal{F}] = Y\mathbb{E}[X|\mathcal{F}].$$

Beweis: Die $\mathcal{F}$-Messbarkeit von $Y\mathbb{E}[X|\mathcal{F}]$ folgt direkt. Ansonsten verwenden wir die „maßtheoretische Induktion“:

(i) Für $Y = 1_B$, $B \in \mathcal{F}$, und $A \in \mathcal{F}$ gilt

$$\mathbb{E}[XY1_A] = \mathbb{E}[X1_A 1_B] = \mathbb{E}[X1_{A\cap B}] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}]1_{A\cap B}] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}]Y1_A],$$

da $A \cap B \in \mathcal{F}$.

(ii) Ist Y eine einfache Funktion, so folgt die Aussage aus (i) mittels Linearität (vgl. Lemma 1.8).

(iii) Für $Y \geq 0$ und $X \geq 0$ verwendet man (ii) und den Satz von der monotonen Konvergenz (vgl. Satz 1.10).

(iv) Allgemein zerlegt man X und Y jeweils in Positiv- und Negativteil und verwendet (iii) und Linearität. □

Satz 1.13. (Turmeigenschaft) *Ist X eine Zufallsvariable mit $\mathbb{E}[|X|] < \infty$ und sind $\mathcal{F}_1 \subset \mathcal{F}_2$ Unter-σ-Algebren von $\mathcal{A}$, so gilt:*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{F}_2]|\mathcal{F}_1] = \mathbb{E}[X|\mathcal{F}_1] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}_1]|\mathcal{F}_2].$$

Beweis: Die zweite Identität folgt direkt aus Satz 1.12, da $\mathbb{E}[X|\mathcal{F}_1]$ $\mathcal{F}_2$-messbar ist. Für die erste Identität wähle $A \in \mathcal{F}_1$. Dann gilt

$$\int_A \mathbb{E}[\mathbb{E}[X|\mathcal{F}_2]|\mathcal{F}_1]d\mathbb{P} = \int_A \mathbb{E}[X|\mathcal{F}_2]d\mathbb{P} = \int_A X d\mathbb{P} = \int_A \mathbb{E}[X|\mathcal{F}_1]d\mathbb{P},$$

jeweils aufgrund von Eigenschaft (ii) aus Definition 1.4. □

Satz 1.14. (Unabhängigkeit) *Es seien X eine Zufallsvariable mit $\mathbb{E}[|X|] < \infty$ und $\mathcal{F}$ und $\mathcal{G}$ Unter-σ-Algebren, so dass $\mathcal{G}$ und $\sigma(\sigma(X), \mathcal{F})$ unabhängig sind. Dann gilt:*

$$\mathbb{E}[X|\sigma(\mathcal{F}, \mathcal{G})] = \mathbb{E}[X|\mathcal{F}].$$

Beweis: Die $\sigma(\mathcal{F}, \mathcal{G})$-Messbarkeit von $\mathbb{E}[X|\mathcal{F}]$ ist offensichtlich. Ansonsten gilt für $F \in \mathcal{F}$ und $G \in \mathcal{G}$:

$$\mathbb{E}[X1_{F\cap G}] = \mathbb{E}[1_G(X1_F)] = \mathbb{E}[1_G]\mathbb{E}[X1_F] = \mathbb{E}[1_G]\mathbb{E}[\mathbb{E}[X|\mathcal{F}]1_F] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}]1_{F\cap G}],$$

wobei wir zweimal die Unabhängigkeit von $\mathcal{G}$ und $\sigma(\sigma(X), \mathcal{F})$ verwendet haben. Damit gilt

$$\mathbb{E}[X1_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}]1_A] \tag{1.2}$$

für einen durchschnittstabilen Erzeuger von $\sigma(\mathcal{F}, \mathcal{G})$, denn $\sigma(\mathcal{F}, \mathcal{G})$ entspricht aufgrund der Unabhängigkeit von $\mathcal{F}$ und $\mathcal{G}$ der Produkt-σ-Algebra von $\mathcal{E}$ und $\mathcal{F}$.. Man kann zudem leicht zeigen, dass die Menge aller A, die (1.2) erfüllen, ein Dynkin-System bilden. Also gilt (1.2) für alle $A \in \sigma(\mathcal{F}, \mathcal{G})$. □

Korollar 1.15. *Es seien $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ und $\mathcal{H} \subset \mathcal{A}$ eine Unter-σ-Algebra.*

(i) Es gilt stets der Satz von der iterierten Erwartung:

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

(ii) Ist X unabhängig von $\mathcal{H}$, so gilt

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

Beweis:

(i) folgt zum Beispiel aus Satz 1.13 mit $\mathcal{F}_1 = \{\Omega, \emptyset\}$.

(ii) folgt aus Satz 1.14 mit $\mathcal{F} = \{\Omega, \emptyset\}$ und $\mathcal{G} = \mathcal{H}$. □

Beispiel 1.16.

(i) Angenommen, X und Y sind diskrete Zufallsvariablen, d.h. es existieren $I_X, I_Y \subset \mathbb{R}$ abzählbar mit $\mathbb{P}(X \in I_X) = \mathbb{P}(Y \in I_Y) = 1$. Setze für $x \in I_X$ und $y \in I_Y$ mit $\mathbb{P}(Y = y) > 0$ nun

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} =: \frac{p_{xy}}{p_y}.$$

Dann ergibt sich mit

$$g(y) = \begin{cases} \sum_{x\in I_X} x\frac{p_{xy}}{p_y}, & p_y > 0, \\ 0, & p_y = 0, \end{cases}$$

die Identität $\mathbb{E}[X|Y] = g(Y)$, sofern $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ liegt.

(ii) Angenommen, X und Y besitzen eine gemeinsame Dichte $f_{X,Y}(x,y)$ bzgl. des Lebesgue-Maßes und marginale Dichten

$$f_X(x) = \int f(x,y)dy \quad \text{bzw.} \quad f_Y(y) = \int f(x,y)dx.$$

Setzt man dann

$$g(y) = \begin{cases} \frac{\int x f_{X,Y}(x,y)dx}{f_Y(y)}, & f_Y(y) > 0, \\ 0, & f_Y(y) = 0, \end{cases}$$

ergibt sich wieder die Identität $\mathbb{E}[X|Y] = g(Y)$, sofern $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ liegt.

Beweis:

(i) Die Funktion g verschwindet außerhalb einer abzählbaren Menge, ist also Borel-messbar. Insbesondere ist $g(Y)$ dann $\sigma(Y)$-messbar. Ist nun $B \in \mathcal{B}$ und $F = \{Y \in B\} \in \sigma(Y)$ beliebig, so folgt

$$\begin{aligned} \mathbb{E}[g(Y)1_F] &= \sum_{y \in B \cap I_Y} \sum_{x \in I_X} x \frac{p_{xy}}{p_y} 1_{\{p_y > 0\}} p_y \\ &= \sum_{y \in B \cap I_Y} \sum_{x \in I_X} x p_{xy} = \sum_{x \in I_X} \sum_{y \in B \cap I_Y} x p_{xy} = \mathbb{E}[X 1_F] \end{aligned}$$

aus dem Satz von Fubini.

(ii) Die Abbildung f_Y ist Borel-messbar. Da zudem

$$\mathbb{E}[X] = \int_{\mathbb{R}\times\mathbb{R}} x f_{X,Y}(x,y)dxdy$$

endlich ist, folgt aus dem Satz von Fubini auch die Messbarkeit von

$$y \mapsto \int x f_{X,Y}(x,y)dx$$

und also die von g. Nun gilt wieder mit $F = \{Y \in B\} \in \sigma(Y)$ beliebig

$$\begin{aligned} \mathbb{E}[g(Y)1_F] &= \int_B g(y) f_Y(y) 1_{\{f_Y(y)>0\}} dy = \int_B \frac{\int x f_{X,Y}(x,y)dx}{f_Y(y)} f_Y(y) 1_{\{f_Y(y)>0\}} dy \\ &= \int_B 1_{\{f_Y(y)>0\}} \int x f_{X,Y}(x,y)dxdy \\ &= \int x \int_B 1_{\{f_Y(y)>0\}} f_{X,Y}(x,y)dydx = \mathbb{E}[X 1_F]. \end{aligned}$$

□

Bemerkung 1.17. Wir haben in Beispiel 1.16 gesehen, dass $\mathbb{E}[X|Y]$ in beiden Fällen die Form $g(Y)$ für eine messbare Abbildung g besitzt. Dies ist zwangsläufig so, wie das folgende Resultat zeigt, dessen reellwertige Version im Rahmen dieser Ausarbeitung genügt. Da wir hier Abbildungen zwischen mehr als zwei Räumen betrachten, werden wir kurzfristig wieder die längere Notation für Messbarkeit einer Abbildung verwenden.

Satz 1.18. (Faktorisierungslemma) *Es seien $Y : \Omega \to \Omega'$ eine $\mathcal{A}$-$\mathcal{A}'$-messbare und $Z : \Omega \to \overline{\mathbb{R}}$ eine $\mathcal{A}$-$\overline{\mathcal{B}}$-messbare Abbildung. Dann ist Z genau dann $\sigma(Y)$-$\overline{\mathcal{B}}$-messbar, wenn eine $\mathcal{A}'$-$\overline{\mathcal{B}}$-messbare Abbildung $g : \Omega' \to \overline{\mathbb{R}}$ existiert mit $Z = g \circ Y$.*

Beweis:

$\Longleftarrow$ Y ist $\sigma(Y)$-$\mathcal{A}'$-messbar und g ist $\mathcal{A}'$-$\overline{\mathcal{B}}$-messbar. Daraus folgt die Aussage über $Z = g \circ Y$ sofort.

$\Longrightarrow$ Sei zunächst $Z = \sum_{i=1}^n \alpha_i 1_{A_i}$ mit $A_i \in \sigma(Y)$ für alle i. Dann existieren $A_i' \in \mathcal{A}'$ mit $A_i = Y^{-1}(A_i')$, und die Behauptung gilt mit $g = \sum_{i=1}^n \alpha_i 1_{A_i'}$. Jedes nicht-negative $\sigma(Y)$-$\overline{\mathcal{B}}$-messbare Z lässt sich als Grenzwert von einfachen Funktionen obigen Typs darstellen, so dass sich die allgemeine Aussage mit den üblichen maßtheoretischen Argumenten ergibt. □

Satz 1.19. *Es seien $X : \Omega \to \mathbb{R}$ und $Y : \Omega \to \Omega'$ Zufallsvariablen und X integrierbar. Dann ist jede $\mathcal{A}'$-$\mathcal{B}$-messbare Funktion $g : \Omega' \to \mathbb{R}$ mit $\mathbb{E}[X|Y] = g(Y)$ auch $\mathbb{P}^Y$-integrierbar und erfüllt*

$$\int_{A'} g d\mathbb{P}^Y = \int_{\{Y \in A'\}} X d\mathbb{P} \quad \forall A' \in \mathcal{A}'. \tag{1.3}$$

Sie ist zudem fast sicher eindeutig bestimmt.

Umgekehrt, ist $g : \Omega' \to \overline{\mathbb{R}}$ zugleich $\mathcal{A}'$-$\mathcal{B}$-messbar und $\mathbb{P}^Y$-integrierbar und erfüllt (1.3), so ist $g \circ Y$ eine Version von $\mathbb{E}[X|Y]$.

Beweis:

$\Longrightarrow$ Die Transformationsformel liefert für jedes $A' \in \mathcal{A}'$

$$\int_{\{Y \in A'\}} X d\mathbb{P} = \int_{\{Y \in A'\}} \mathbb{E}[X|Y] d\mathbb{P} = \int_{\{Y \in A'\}} g \circ Y d\mathbb{P} = \int_{A'} g d\mathbb{P}^Y.$$

Da X $\mathbb{P}$-integrierbar ist, ist auch g $\mathbb{P}^Y$-integrierbar, und insbesondere gilt (1.3). Ist h eine weitere $\mathcal{A}'$-$\mathcal{B}$-messbare Funktion mit $\mathbb{E}[X|Y] = h(Y)$, so folgt

$$\int_{A'} g d\mathbb{P}^Y = \int_{A'} h d\mathbb{P}^Y \quad \forall A' \in \mathcal{A}'.$$

Die Zerlegung in Positiv- und Negativteil liefert

$$\int_{A'} (g^+ + h^-) d\mathbb{P}^Y = \int_{A'} (h^+ + g^-) d\mathbb{P}^Y \quad \forall A' \in \mathcal{A}'.$$

Beide Integranden sind nicht-negativ, d.h. es lässt sich wie im Beweis von Satz 1.5 durch geschickte Wahl von A' schließen, dass $g^+ + h^- = h^+ + g^-$ $\mathbb{P}^Y$-f.s. gilt, also nach Subtraktion auch $g = h$ $\mathbb{P}^Y$-f.s.

$\Longleftarrow$ Die Abbildung $g \circ Y$ ist $\sigma(Y)$-$\mathcal{B}$-messbar und erfüllt nach Transformationsformel und Voraussetzung

$$\int_{\{Y \in A'\}} g \circ Y d\mathbb{P} = \int_{A'} g d\mathbb{P}^Y = \int_{\{Y \in A'\}} X d\mathbb{P} \quad \forall A' \in \mathcal{A}'.$$

Damit gilt

$$\int_C g \circ Y d\mathbb{P} = \int_C X d\mathbb{P} \quad \forall C \in \sigma(Y)$$

und also $g \circ Y = \mathbb{E}[X|Y]$. □

Beispiel 1.20. Es seien $(\Omega', \mathcal{A}')$ mit $\{y\} \in \mathcal{A}'$ für ein $y \in \Omega'$ gegeben. In diesem Fall ergibt (1.3) die Identität

$$g(y)\mathbb{P}(Y = y) = \int_{\{Y=y\}} X d\mathbb{P}.$$

Ist $\mathbb{P}(Y = y) > 0$, erhält man

$$g(y) = \frac{1}{\mathbb{P}(Y = y)} \int_{\{Y=y\}} X d\mathbb{P} =: \mathbb{E}[X|Y = y].$$

In vielen Fällen, etwa im Fall stetiger Verteilungen, gilt jedoch $\mathbb{P}(Y = y) = 0$ für viele oder alle $y \in \Omega'$, und die obige Definition ist nicht zugänglich. Jedoch steht im Allgemeinen immer der Wert von g an der Stelle y zur Verfügung.

Definition 1.21. Für eine integrierbare Zufallsvariable $X : \Omega \to \mathbb{R}$ und eine Zufallsvariable $Y : \Omega \to \Omega'$ sei g eine der Bedingung (1.3) in Satz 1.19 genügende, $\mathcal{A}'$-$\mathcal{B}$-messbare und $\mathbb{P}^Y$-integrierbare Funktion. Dann heißt $g(y)$ für jedes $y \in \Omega'$ der bedingte Erwartungswert von X unter der Bedingung $Y = y$. Wir schreiben:

$$\mathbb{E}[X|Y = y] := g(y).$$

Bemerkung 1.22. Im Allgemeinen ist $\mathbb{E}[X|Y]$ eine Zufallsvariable, die je nach Realisation von $\omega \mapsto Y(\omega)$ einen anderen Wert herausgibt. $\mathbb{E}[X|Y = y]$ ist die reelle Zahl, die diesen Wert angibt, falls $Y(\omega) = y$ ist.

Beispiel 1.23. Wir hatten in Beispiel 1.16 schon zwei typische bedingte Erwartungswerte betrachtet. Für diskrete X und Y ergibt sich

$$\mathbb{E}[X|Y = y] = g(y) = \begin{cases} \sum_{x \in I_X} x \frac{p_{xy}}{p_y}, & p_y > 0, \\ 0, & p_y = 0, \end{cases}$$

während wir im Lebesgue-stetigen Fall

$$\mathbb{E}[X|Y = y] = g(y) = \begin{cases} \frac{\int x f_{X,Y}(x,y) dx}{f_Y(y)}, & f_Y(y) > 0, \\ 0, & f_Y(y) = 0, \end{cases}$$

erhalten haben. Beide Identitäten ähneln den bekannten Formeln für Erwartungswerte im diskreten bzw. stetigen Fall, nur dass die unbedingten Wahrscheinlichkeiten p_x bzw. Dichten $f_X(x)$ durch die bedingten Wahrscheinlichkeiten oder Dichten

$$\mathbb{P}(X = x|Y = y) = \frac{p_{xy}}{p_y} \quad \text{bzw.} \quad f_{X|Y}(x, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

ersetzt wurden. Sind X und Y unabhängig, ergeben sich die üblichen unbedingten Erwartungswerte (vgl. Korollar 1.15).

Bemerkung 1.24. Wir hatten zu Beginn des Kapitels die bedingte Erwartung $\mathbb{E}[X|\mathcal{F}]$ intuitiv als beste Annäherung an X eingeführt, wenn nur ein Teilausgang des Experiments in Form von $\mathcal{F}$ bekannt ist. Dies werden wir zuletzt präzisieren.

Satz 1.25. *Es sei $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. Dann gilt für jede Zufallsvariable $Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ die Ungleichung*

$$\mathbb{E}[(X-Y)^2] \geq \mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}])^2] \tag{1.4}$$

mit Gleichheit genau dann, wenn $Y = \mathbb{E}[X|\mathcal{F}]$.

Beweis: Wir nehmen zunächst an, dass $\mathbb{E}[(\mathbb{E}[X|\mathcal{F}])^2] < \infty$ gilt. Dann folgt aus

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|\mathcal{F}]] = \mathbb{E}[Y\mathbb{E}[X|\mathcal{F}]]$$

und

$$\mathbb{E}[X\mathbb{E}[X|\mathcal{F}]] = \mathbb{E}[\mathbb{E}[X\mathbb{E}[X|\mathcal{F}]|\mathcal{F}]] = \mathbb{E}[(\mathbb{E}[X|\mathcal{F}])^2]$$

(vgl. jeweils Korollar 1.15 und Satz 1.12) für jedes $\mathcal{F}$-messbare Y die Ungleichung

$$\begin{aligned}
&\mathbb{E}[(X-Y)^2] - \mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}])^2] \\
=&\mathbb{E}[X^2 - 2XY + Y^2] - \mathbb{E}[X^2 - 2X\mathbb{E}[X|\mathcal{F}] + \mathbb{E}[X|\mathcal{F}]^2] \\
=&\mathbb{E}[Y^2 - 2Y\mathbb{E}[X|\mathcal{F}] + \mathbb{E}[X|\mathcal{F}]^2] = \mathbb{E}[(Y - \mathbb{E}[X|\mathcal{F}])^2] \geq 0.
\end{aligned}$$

Zuletzt beweisen wir also $\mathbb{E}[(\mathbb{E}[X|\mathcal{F}])^2] < \infty$. Dazu betrachten wir zunächst für $N \in \mathbb{N}$ die Zufallsvariable $\min(|X|, N)$, die gemäß Lemma 1.8 die Bedingung

$$\mathbb{E}[(\mathbb{E}[\min(|X|, N)|\mathcal{F}])^2] \leq N^2 < \infty$$

erfüllt. Aus $a^2 \leq 2(a-b)^2 + 2b^2$ und (1.4) mit $Y = 0$ folgt

$$\begin{aligned}
&\mathbb{E}[(\mathbb{E}[\min(|X|, N)|\mathcal{F}])^2] \\
\leq&2\mathbb{E}[\{\min(|X|, N) - \mathbb{E}[\min(|X|, N)|\mathcal{F}]\}^2] + 2\mathbb{E}[(\min(|X|, N))^2] \\
\leq&4\mathbb{E}[(\min(|X|, N))^2] \leq 4\mathbb{E}[X^2].
\end{aligned}$$

Zudem folgt mit Satz 1.10

$$\mathbb{E}[\min(|X|, N)|\mathcal{F}] \nearrow \mathbb{E}[|X||\mathcal{F}].$$

Insgesamt erhält man

$$\mathbb{E}[(\mathbb{E}[X|\mathcal{F}])^2] \leq \mathbb{E}[(\mathbb{E}[|X||\mathcal{F}])^2] = \lim_{N\to\infty} \mathbb{E}[(\mathbb{E}[\min(|X|, N)|\mathcal{F}])^2] \leq 4\mathbb{E}[X^2] < \infty.$$

□

Kapitel 2

Grundlagen der Punktschätzung

In diesem Kapitel beginnen wir mit der Untersuchung von Punktschätzern und einer Analyse hinsichtlich ihrer Qualität. Punktschätzer werden verwendet, um bestimmte Eigenschaften einer unbekannten Verteilung anhand eines gegebenen Datensatzes zu schätzen.

Beispiel 2.1. Bei der Einführung eines Medikaments werden häufig Tests an Versuchstieren durchgeführt, um die Qualität eines Wirkstoffs in Abhängigkeit von der Dosis zu testen. Beobachten wir ein einzelnes Versuchstier, können wir erkennen, ob es bei Zufuhr der Dosis x geheilt wird oder nicht. Wir nehmen daher formal

$$Y \sim B(1, p(x))$$

an, wobei $B(n, p)$ eine Binomialverteilung mit Parametern n und p bezeichnet und $p(x)$ für die Heilungswahrscheinlichkeit bei Dosis x steht.

Typischerweise beobachten wir mehrere Tiere, wobei wir annehmen, dass die Zufallsvariablen $Y_1, \dots, Y_n$ unabhängig mit Verteilung $Y_i \sim B(1, p(x_i))$ sind. Das Ziel wäre dann allgemein die Schätzung der unbekannten Funktion $p : [0, \infty) \to [0, 1]$. Dies ist oftmals schwierig, weshalb man zu einem *parametrischen Modell* übergeht. Ein typisches Beispiel wäre

$$p(x) = 1 - \exp(-\beta x), \quad \beta > 0,$$

so dass die Schätzung von p auf die Schätzung des Parameters $\beta > 0$ zurückgeführt werden kann.

Annahme 2.2. Es seien $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{B})$ ein Messraum und $X : \Omega \to \mathcal{X}$ eine Zufallsvariable. Mit dem Bildmaß

$$P(B) = \mathbb{P}^X(B) = \mathbb{P}(X^{-1}(B)), \quad B \in \mathcal{B},$$

wird $(\mathcal{X}, \mathcal{B}, P)$ ebenfalls zu einem Wahrscheinlichkeitsraum. $\mathcal{X}$ spielt dabei die Rolle des Raumes, in dem unsere Beobachtungen liegen, und wird daher auch als *Stichprobenraum* bezeichnet. Entsprechend beobachten wir eine *Stichprobe* $x = X(\omega)$. Die σ-Algebra $\mathcal{B}$ ist im Allgemeinen trotz derselben Notation nicht die Borel-σ-Algebra.

Definition 2.3. Es seien $\mathcal{X} \neq \emptyset$, $\mathcal{B}$ eine σ-Algebra auf $\mathcal{X}$ und $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathcal{X}, \mathcal{B})$, wobei $|\Theta| \geq 2$ und $P_\vartheta \neq P_{\vartheta'}$ für $\vartheta \neq \vartheta'$ gilt. Dann wird Θ als *Parameterraum* und $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ als *statistisches Experiment* bezeichnet.

Bemerkung 2.4. Man interpretiert Definition 2.3 im Allgemeinen so, dass man an der wahren Verteilung $P \in \mathcal{P}$ einer Zufallsvariable $X : \Omega \to \mathcal{X}$ interessiert ist. Auf Basis der Stichprobe $x = X(\omega)$ soll eine Entscheidung über das unbekannte P getroffen werden. Durch die Identifikation von $\mathcal{P}$ mit der Parametermenge Θ ist die Entscheidung für $P \in \mathcal{P}$ äquivalent zur Entscheidung für einen Parameter $\vartheta \in \Theta$. Dies vereinfacht das Schätzproblem üblicherweise erheblich.

Beispiel 2.5. In Beispiel 2.1 haben wir unabhängige Zufallsvariablen $Y_1, \dots, Y_n$ mit $Y_i \sim B(1, 1 - \exp(-\beta x_i)) = P_i^\beta$ betrachtet. Formal beobachten wir also die Zufallsvariable $X = (Y_1, \dots, Y_n)^T$ aus dem statistischen Experiment

$$\mathcal{X} = \{0,1\}^n, \quad \mathcal{B} = \mathcal{P}(\mathcal{X}), \quad \mathcal{P} = \{\otimes_{i=1}^n P_i^\beta \mid \beta > 0\}, \quad \Theta = [0, \infty).$$

Definition 2.6. Es seien $(\Gamma, \mathcal{A}_\Gamma)$ ein Messraum und $\gamma : \Theta \to \Gamma$ eine Abbildung. Eine messbare Abbildung

$$g : (\mathcal{X}, \mathcal{B}) \to (\Gamma, \mathcal{A}_\Gamma)$$

heißt *(Punkt-)schätzer* für $\gamma(\vartheta)$. Beobachtet man $x = X(\omega)$, so heißt $g(x)$ *Schätzung* für $\gamma(\vartheta)$.

Beispiel 2.7.

(i) Es seien $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2) = P_{\mu,\sigma^2}$, also $X = (X_1, \dots, X_n)^T$ und

$$\mathcal{X} = \mathbb{R}^n, \quad \mathcal{B} = \mathcal{B}^n, \quad \mathcal{P} = \{\otimes_{i=1}^n P_{\mu,\sigma^2} \mid \mu \in \mathbb{R}, \sigma^2 > 0\}, \quad \Theta = \mathbb{R} \times (0, \infty).$$

Typische Punktschätzer für den Parameter $\vartheta = (\mu, \sigma^2)$ sind durch

$$g : \begin{cases} \mathbb{R}^n \to \Theta \\ x \mapsto \begin{pmatrix} \overline{x}_n \\ \hat{s}_n^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i \\ \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2 \end{pmatrix} \end{cases}$$

gegeben, deren Qualität wir im Folgenden u.a. untersuchen werden. In diesem Fall haben wir formal $\gamma(\vartheta) = \vartheta$ verwendet.

(ii) Es seien reellwertige $X_1, \dots, X_n$ i.i.d. $\sim F$ gegeben, wobei $F(x) = \mathbb{P}(X_i \leq x)$ die unbekannte Verteilungsfunktion von X bezeichnet. In diesem Fall ist

$$\Theta = \{F \mid F \text{ reellwertige Verteilungsfunktion}\}$$

ein unendlichdimensionaler Parameterraum. Interessieren wir uns nur für

$$\gamma : \begin{cases} \Theta \to \Gamma = [0,1] \\ F \mapsto F(0) = \mathbb{P}(X_i \leq 0), \end{cases}$$

so ist ein Punktschätzer für $\gamma(\vartheta)$ durch

$$g : \begin{cases} \mathbb{R}^n \to \Gamma \\ x \mapsto \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \leq 0\}} \end{cases}$$

gegeben.

Bemerkung 2.8. Wir sehen anhand von Beispiel 2.7, weshalb wir in Definition 2.6 die Funktion $\gamma : \Theta \to \Gamma$ eingeführt haben. Oft sind wir nicht am Parameter ϑ interessiert, sondern an Funktionalen davon. Dies könnte eine Komponente von ϑ oder eine Funktion davon sein (etwa in Teil (i) der Erwartungswert μ oder die Standardabweichung σ) oder ein allgemeineres Funktional von ϑ wie $F(0)$ in Teil (ii).

Definition 2.9. Eine messbare Funktion $L : \Gamma \times \Gamma \to [0, \infty)$ heißt *Verlustfunktion*. Für einen Punktschätzer $g : \mathcal{X} \to \Gamma$ heißt

$$R(\cdot, g) : \begin{cases} \Theta \to [0, \infty] \\ \vartheta \mapsto R(\vartheta, g) = \mathbb{E}_\vartheta[L(\gamma(\vartheta), g(X))] = \int_{\mathcal{X}} L(\gamma(\vartheta), g(x)) P_\vartheta(dx) \end{cases}$$

das *Risiko* (von g) unter L.

Bemerkung 2.10. Ist ϑ der wahre Parameter und $g(x)$ eine Schätzung für $\gamma(\vartheta)$, so misst $L(\gamma(\vartheta), g(x))$ den dadurch entstehenden Verlust. Typische Verlustfunktionen basieren daher auf dem Abstand von $\gamma(\vartheta)$ und $g(x)$, sofern Γ ein metrischer Raum ist. Ein Beispiel ist etwa der quadratische Verlust $L(x, y) = (x - y)^2$ im reellen Fall. Das Risiko entspricht dann dem erwarteten Verlust, falls $\gamma(\vartheta)$ der wahre Parameter ist, und dient im Folgenden als Grundlage für die Beurteilung der Qualität eines Schätzers.

Definition 2.11. Es seien L eine Verlustfunktion und $\gamma(\vartheta)$ ein zu schätzender Parameter. Bezeichnet $\mathcal{K}$ die Menge aller Punktschätzer für $\gamma(\vartheta)$, so heißt $g^* \in \mathcal{K}$ mit

$$R(\vartheta, g^*) = \inf_{g \in \mathcal{K}} R(\vartheta, g) \quad \forall \vartheta \in \Theta$$

ein *gleichmäßig bester Schätzer*. Analog definiert man gleichmäßig beste Schätzer in Teilklassen $\widetilde{\mathcal{K}} \subset \mathcal{K}$.

Beispiel 2.12. Im Allgemeinen ist weder ein Schätzer gleichmäßig besser als ein anderer, noch existieren gleichmäßig beste Schätzer über die gesamte Klasse. Als einfaches Beispiel betrachten wir

$$\mathcal{X} = \mathbb{R}, \quad \mathcal{P} = \{P_\mu = \mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R}\}, \quad \gamma(\mu) = \mu$$

mit dem quadratischen Verlust. Offenbar gilt für den trivialen Schätzer $g_\nu(x) = \nu$, der sich unabhängig von der Stichprobe x immer für ν entscheidet,

$$R(\mu, g_\nu) = \mathbb{E}_\mu[(\mu - \nu)^2] = (\mu - \nu)^2 \quad \forall \mu \in \mathbb{R}$$

und insbesondere $R(\nu, g_\nu) = 0$. Es folgt, dass kein g_ν gleichmäßig besser als ein anderer Schätzer g_μ ist. Zudem müsste ein gleichmäßig bester Schätzer g^* demnach $R(\mu, g^*) = 0$ für alle $\mu \in \mathbb{R}$ erfüllen, um lokal in μ nicht schlechter als der jeweilige Schätzer g_μ zu sein. Formal müsste also

$$\mathbb{E}_\mu[(g^*(X) - \mu)^2] = 0 \quad \forall \mu \in \mathbb{R} \qquad \Longleftrightarrow \qquad g^*(x) = \mu \ P_\mu\text{-f.s.} \quad \forall \mu \in \mathbb{R}$$

gelten. Dies ist offensichtlich nicht möglich.

Bemerkung 2.13. Als Konsequenz aus Beispiel 2.12 bestehen im Wesentlichen zwei Möglichkeiten, trotzdem „beste“ Schätzer zu erhalten:

(i) Einschränkung der Klasse $\mathcal{K}$ oder

(ii) Abweichung von der Bedingung eines gleichmäßig kleinsten Risikos.

Wir werden uns im weiteren Verlauf dieses Kapitels mit der ersten Variante befassen.

Definition 2.14.

(i) Ein Schätzer $g^* \in \mathcal{K}$ heißt *zulässig*, falls kein $g \in \mathcal{K}$ mit

$$R(\vartheta, g) \leq R(\vartheta, g^*) \quad \forall \vartheta \in \Theta \qquad \text{und} \qquad R(\vartheta^*, g) < R(\vartheta^*, g^*) \quad \text{für ein } \vartheta^* \in \Theta$$

existiert.

(ii) Eine Klasse $\widetilde{\mathcal{K}} \subset \mathcal{K}$ heißt *vollständig*, falls für alle $g \in \mathcal{K} \backslash \widetilde{\mathcal{K}}$ ein $\widetilde{g} \in \widetilde{\mathcal{K}}$ mit

$$R(\vartheta, \widetilde{g}) \leq R(\vartheta, g) \quad \forall \vartheta \in \Theta$$

existiert. Enthält $\widetilde{\mathcal{K}}$ keine echte vollständige Teilklasse, so bezeichnen wir $\widetilde{\mathcal{K}}$ als *minimalvollständig*.

Bemerkung 2.15. In Beispiel 2.12 sind die Schätzer $g_\nu(x) = \nu$ zulässig, da jeder potentiell bessere Schätzer g^* auch $R_\nu(\nu, g^*) = 0$ und also $g^*(x) = \nu P_\nu$-f.s. erfüllen muss. Dies zeigt insbesondere, dass auch a priori schlechte Schätzer zulässig sind, und macht deutlich, dass die Einschränkung auf Klassen zulässiger oder vollständiger Schätzer das Entscheidungsproblem nur selten vereinfacht. Wir werden daher im Folgenden deutlich stärker eingeschränkte Klassen betrachten.

Definition 2.16.

(i) Es sei g ein Schätzer für $\gamma : \Theta \to \mathbb{R}$. Dann heißt

$$B_\vartheta(g) = \mathbb{E}_\vartheta[g(X)] - \gamma(\vartheta)$$

der *Bias* (oder die *Verzerrung*) von g. Der Schätzer heißt *erwartungstreu* (oder *unverzerrt*), falls

$$B_\vartheta(g) = 0 \quad \forall \vartheta \in \Theta.$$

(ii) Ein Schätzer g^* heißt *erwartungstreuer Schätzer mit gleichmäßig kleinster Varianz* (oder *UMVU-Schätzer* für “uniformly minimum variance unbiased”), falls

$$g^* \in \mathcal{E}_\gamma = \{g \mid g \text{ ist erwartungstreu für } \gamma \text{ und liegt in } L^2(P_\vartheta) \text{ für alle } \vartheta \in \Theta\}$$

und

$$\mathrm{Var}_\vartheta(g^*(X)) = \mathbb{E}_\vartheta[(g^*(X) - \gamma(\vartheta))^2] = \inf_{g \in \mathcal{E}_\gamma} \mathrm{Var}_\vartheta(g(X)) \quad \forall \vartheta \in \Theta. \tag{2.1}$$

Bemerkung 2.17.

(i) Schätzer, die Bedingungen vom Typ (2.1) nur für ein $\vartheta \in \Theta$ erfüllen, heißen lokal optimal. Diese sind in der Praxis häufig nutzlos, da der wahre Parameter ϑ ja gerade nicht bekannt ist.

(ii) Wählt man die quadratische Verlustfunktion $L(x,y) = (x-y)^2$, so heißt für einen Schätzer $g \in L^2(P_\vartheta)$ für alle $\vartheta \in \Theta$ der Ausdruck

$$MSE_\vartheta(g) = R(\vartheta, g) = \mathbb{E}_\vartheta[(g(X) - \gamma(\vartheta))^2] = \mathrm{Var}_\vartheta(g(X)) + B_\vartheta^2(g)$$

die *mittlere quadratische Abweichung* von g. Ist g erwartungstreu, so folgt

$$MSE_\vartheta(g) = \mathrm{Var}_\vartheta(g(X)).$$

UMVU-Schätzer sind daher gleichmäßig beste Schätzer bzgl. der quadratischen Verlustfunktion, wenn man die Klasse der Schätzer auf $\mathcal{E}_\gamma$ einschränkt.

(iii) Analog lässt sich der Begriff der Erwartungstreue für Schätzer g für $\gamma : \Theta \to \mathbb{R}^d$ definieren, indem man diese Eigenschaft komponentenweise fordert.

Definition 2.18.

(i) Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(0,1)$. Dann besitzt die Zufallsvariable $Z = \sum_{i=1}^n X_i^2$ eine χ^2*-Verteilung mit* n *Freiheitsgraden.* Notation: $Z \sim \chi_n^2$. Die Dichte f von Z ist durch

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} 1_{(0,\infty)}(x)$$

gegeben, wobei $\Gamma(\cdot)$ die Gamma-Funktion bezeichnet.

(ii) Es seien Y eine standardnormalverteilte und Z eine unabhängige χ_n^2-verteilte Zufallsvariable. Dann heißt die Verteilung von

$$T = \frac{Y}{\sqrt{Z/n}}$$

eine t*-Verteilung mit* n *Freiheitsgraden.* Notation: $T \sim t_n$. Die Dichte g von T ist durch

$$g(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

gegeben.

Bemerkung 2.19. Ist $Z \sim \chi_n^2$, so gilt

$$\mathbb{E}[Z] = \sum_{i=1}^n \mathbb{E}[X_i^2] = n$$

und

$$\mathrm{Var}(Z) = \sum_{i=1}^n \mathrm{Var}(X_i^2) = 2n.$$

Lemma 2.20. *Es seien* $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{N}(\mu, \sigma^2)$. *Dann gelten*

$$\hat{s}_n^2(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2 \sim \frac{\sigma^2}{n} \chi_{n-1}^2$$

und

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

und beide Schätzer sind unabhängig.

Beweis: Die Verteilung von $\overline{X}_n$ ergibt sich aus klassischen Eigenschaften der Normalverteilung. Wir setzen ansonsten

$$Y_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

und $Y = (Y_1, \ldots, Y_n)^T$ und wählen (etwa nach dem Gram-Schmidt-Verfahren) eine Orthogonalmatrix A, deren letzte Zeile

$$(A_{nj})_{j=1,\ldots,n} = \left(\frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}} \right) = v^T$$

erfüllt. Offenbar gilt für $Z = AY$ wegen der Orthogonalität von A die Relation

$$\sum_{i=1}^{n} Z_i^2 = Z^T Z = Y^T A^T A Y = Y^T Y = \sum_{i=1}^{n} Y_i^2,$$

und wir erhalten $Z \sim \mathcal{N}(0, \mathbb{I}_n)$ wegen $\mathrm{Cov}(Z) = AA^T$. Nun gilt

$$\sqrt{n}\overline{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\sigma Y_i + \mu) = \sigma v^T Y + \sqrt{n}\mu = \sigma Z_n + \sqrt{n}\mu$$

und

$$\begin{aligned} n\hat{s}_n^2(X) &= \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \sigma^2 \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2 \\ &= \sigma^2 \left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}_n^2 \right) = \sigma^2 \left(\sum_{i=1}^{n} Y_i^2 - \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \right\}^2 \right) \\ &= \sigma^2 \left(\sum_{i=1}^{n} Z_i^2 - Z_n^2 \right) = \sigma^2 \sum_{i=1}^{n-1} Z_i^2. \end{aligned}$$

Da Z ein Vektor aus i.i.d. standardnormalverteilten Zufallsvariablen ist, ergibt sich die Aussage über die Verteilung von $\hat{s}_n^2(X)$ direkt aus Definition 2.18. Zudem sind $\overline{X}_n$ und $\hat{s}_n^2(X)$ als Funktionen unabhängiger Zufallsvariablen selbst unabhängig. □

Korollar 2.21. *Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Dann besitzt*

$$T_n = \frac{\sqrt{n-1}(\overline{X}_n - \mu)}{\hat{s}_n(X)}$$

eine t-Verteilung mit $n-1$ Freiheitsgraden.

Beweis: Schreibt man

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sqrt{\frac{\sigma^2(n-1)}{n\hat{s}_n^2(X)}},$$

folgt die Aussage direkt aus Lemma 2.20 und Definition 2.18. □

Beispiel 2.22. Mit Hilfe von Lemma 2.20 können wir die beiden Punktschätzer aus Beispiel 2.7 in Hinblick auf ihre Erwartungstreue untersuchen. Es seien $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, und wir verwenden $\overline{X}_n$ zur Schätzung von μ bzw. $\hat{s}_n^2(X)$ zur Schätzung von σ^2. Mit

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{und} \qquad \hat{s}_n^2(X) \sim \frac{\sigma^2}{n}\chi^2_{n-1}$$

folgt zwar, dass $\overline{X}_n$ erwartungstreu für μ ist, $\hat{s}_n^2(X)$ erfüllt wegen Bemerkung 2.19 allerdings bloß

$$\mathbb{E}_\vartheta[\hat{s}_n^2(X)] = \frac{n-1}{n}\sigma^2$$

und ist damit nur "asymptotisch erwartungstreu". Außerdem ergeben sich

$$MSE_\vartheta(\overline{X}_n) = \mathrm{Var}_\vartheta(\overline{X}_n) = \frac{\sigma^2}{n}$$

bzw.

$$MSE_\vartheta(\hat{s}_n^2(X)) = \mathrm{Var}_\vartheta(\hat{s}_n^2(X)) + B_\vartheta^2(\hat{s}_n^2) = \frac{\sigma^4}{n^2}2(n-1) + \frac{\sigma^4}{n^2} = \frac{2n-1}{n^2}\sigma^4.$$

Für beide Folgen von Schätzern gilt also $MSE_\vartheta(g_n) \to 0$ für $n \to \infty$.

Satz 2.23. (Cramér-Rao-Ungleichung) *Es sei μ ein σ-endliches Maß, und es gelte $P_\vartheta \ll \mu$ für alle $\vartheta \in \Theta \subset \mathbb{R}$. Ferner seien Θ offen und $g : \mathcal{X} \to \mathbb{R}$ ein Schätzer. Wir nehmen an, dass die folgenden Regularitätsannahmen gelten, wenn $f(\cdot, \vartheta)$ die Dichte von P_ϑ bzgl. μ bezeichnet:*

(i) *Die Menge $M_f := \{x \in \mathcal{X} \mid f(x, \vartheta) > 0\}$ ist unabhängig von ϑ.*

(ii) *Die partielle Ableitung $\frac{\partial}{\partial\vartheta} f(x, \vartheta)$ existiert für alle $x \in \mathcal{X}$.*

(iii) *(a)* $\mathbb{E}_\vartheta[\frac{\partial}{\partial\vartheta} \log f(X, \vartheta)] = 0$,

(b) $\mathbb{E}_\vartheta[g(X)\frac{\partial}{\partial\vartheta} \log f(X, \vartheta)] = \frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta[g(X)]$.

(iv) $0 < I(f(\cdot, \vartheta)) = \mathbb{E}_\vartheta[(\frac{\partial}{\partial\vartheta} \log f(X, \vartheta))^2] < \infty$.

Dann gilt für alle $\vartheta \in \Theta$ die Ungleichung

$$\mathrm{Var}_\vartheta(g(X)) \geq \frac{(\frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta[g(X)])^2}{I(f(\cdot, \vartheta))}. \tag{2.2}$$

Beweis: Wir verwenden die Notation

$$U_\vartheta(x) = \begin{cases} 0, & \text{falls } x \notin M_f, \\ \frac{\partial}{\partial\vartheta} \log f(x, \vartheta), & \text{sonst.} \end{cases}$$

Für die Zufallsvariablen $U_\vartheta := U_\vartheta(X)$ und $g = g(X)$ gilt dann $\mathbb{E}_\vartheta[U_\vartheta] = 0$ wegen Eigenschaft (iii) (a) und

$$\mathrm{Var}_\vartheta(U_\vartheta) = \mathbb{E}_\vartheta[U_\vartheta^2] = I(f(\cdot, \vartheta)) \in (0, \infty)$$

nach Eigenschaft (iv). Damit erhält man mit Hilfe der Cauchy-Schwarz-Ungleichung

$$\Big(\frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta[g(X)]\Big)^2 = \mathbb{E}_\vartheta[gU_\vartheta]^2 = \mathrm{Cov}_\vartheta(g,U_\vartheta)^2$$
$$\leq \mathrm{Var}_\vartheta(g)\,\mathrm{Var}_\vartheta(U_\vartheta) = I(f(\cdot,\vartheta))\mathrm{Var}_\vartheta(g(X)),$$

wobei wir im ersten Schritt Eigenschaft (iii) (b) verwendet haben. □

Definition 2.24.

(i) Die Größe $I_\vartheta(f) = I(f(\cdot,\vartheta))$ in Satz 2.23 heißt *Fisher-Information* von $\mathcal{P} = \{P_\vartheta \mid \vartheta\in\Theta\}$ bzgl. ϑ.

(ii) Falls g ein Schätzer ist, für den in (2.2) das Gleichheitszeichen gilt, so heißt g *effizient* für $\gamma(\vartheta) = \mathbb{E}_\vartheta[g(X)]$.

Bemerkung 2.25.

(i) Eigenschaft (iii) in den Regularitätsbedingungen zur Cramér-Rao-Ungleichung lässt sich relativ leicht interpretieren, indem man die beiden Bedingungen äquivalent formuliert:

(a) $\int_\mathcal{X} \frac{\partial}{\partial\vartheta}\log f(x,\vartheta) f(x,\vartheta)\mu(dx) = \int_\mathcal{X} \frac{\partial}{\partial\vartheta} f(x,\vartheta)\mu(dx) = \frac{\partial}{\partial\vartheta}\int_\mathcal{X} f(x,\vartheta)\mu(dx)$,

(b) $\int_\mathcal{X} g(x)\frac{\partial}{\partial\vartheta}\log f(x,\vartheta) f(x,\vartheta)\mu(dx) = \int_\mathcal{X} g(x)\frac{\partial}{\partial\vartheta} f(x,\vartheta)\mu(dx) = \frac{\partial}{\partial\vartheta}\int_\mathcal{X} g(x) f(x,\vartheta)\mu(dx)$.

Man erkennt, dass beide Annahmen der Vertauschung von Integral und Ableitung entsprechen, was in vielen Fällen leicht nachzuweisen ist.

(ii) Satz 2.23 gibt insbesondere eine untere Schranke für die Varianz eines Schätzers für $\gamma(\vartheta) = \mathbb{E}_\vartheta[g(X)]$ an und kann daher im Prinzip verwendet werden, um UMVU-Schätzer zu bestimmen: Sind im statistischen Modell die Regularitätsbedingungen zur Anwendung der Cramér-Rao-Ungleichung erfüllt, so ist jeder erwartungstreue und effiziente Schätzer auch UMVU-Schätzer.

(iii) Ist $X = (X_1,\dots,X_n)^T$ und $X_1,\dots,X_n$ i.i.d. $\sim P^1_\vartheta \ll \mu^1$, dann gilt

$$P_\vartheta = \otimes_{j=1}^n P^1_\vartheta \ll \otimes_{j=1}^n \mu^1 \qquad \text{und} \qquad f(x,\vartheta) = \prod_{j=1}^n f^1(x_j,\vartheta),$$

wenn $f^1(\cdot,\vartheta) = \frac{dP^1_\vartheta}{d\mu^1}$ die μ_1-Dichte der Verteilung P^1_ϑ von X_i bezeichnet. Man erhält dann leicht

$$I(f(\cdot,\vartheta)) = n\; I(f^1(\cdot,\vartheta)).$$

Beispiel 2.26. Es seien $X_1,\dots,X_n$ i.i.d. $\sim\mathcal{N}(\mu,1)$ mit Dichte

$$f^1(x,\mu) = \frac{1}{\sqrt{2\pi}}\;\exp\left(-\frac{1}{2}(x-\mu)^2\right).$$

Dann erhält man für die Fisher-Information

$$I(f^1(\cdot,\mu)) = \mathbb{E}_\mu\Big[\Big(\frac{d}{d\mu}\log f^1(X_1,\mu)\Big)^2\Big] = \mathbb{E}_\mu[(X-\mu)^2] = 1$$

Insbesondere ergibt sich aus Bemerkung 2.25 (iii) als Fisher-Information der gemeinsamen Verteilung $I(f(\cdot,\mu)) = n$, und man erhält die Cramér-Rao-Ungleichung

$$\mathrm{Var}_\mu(g(X)) \geq \frac{(\frac{\partial}{\partial\mu}\mathbb{E}_\mu[g(X)])^2}{n} = \frac{1}{n},$$

wobei die letzte Identität für erwartungstreue Schätzer für μ gilt. Insbesondere ist $g(x) = \overline{x}_n$ ein UMVU-Schätzer (vgl. Beispiel 2.22).

Definition 2.27. Es sei

$$SYM(k) = \{A \in \mathbb{R}^{k\times k} \mid A \text{ ist symmetrisch}\}$$

die Menge der symmetrischen Matrizen. Für $A, B \in SYM(k)$ schreiben wir

$$\begin{aligned} A \geq 0 \quad &\Longleftrightarrow \quad A \text{ ist nicht-negativ definit,} \\ A \geq B \quad &\Longleftrightarrow \quad A - B \geq 0. \end{aligned}$$

Die durch $\geq$ definierte Halbordnung auf $SYM(k)$ heißt *Löwner-Ordnung*. Mit

$$\begin{aligned} NND(k) &= \{A \in SYM(k) \mid A \geq 0\}, \\ PD(k) &= \{A \in NND(k) \mid |A| \neq 0\}, \end{aligned}$$

$|A| = \det(A)$, bezeichnen wir die Mengen der symmetrischen, nicht-negativ definiten bzw. der symmetrischen, positiv definiten Matrizen.

Satz 2.28. (mehrdimensionale Cramér-Rao-Ungleichung) *Es sei μ ein σ-endliches Maß, und es gelte $P_\vartheta \ll \mu$ für alle $\vartheta \in \Theta \subset \mathbb{R}^d$. Ferner seien Θ offen und $g = (g_1, \ldots, g_k)^T : \mathcal{X} \to \mathbb{R}^k$ ein Schätzer. Wir setzen zuletzt*

$$G(\vartheta) := \Big(\frac{\partial}{\partial\vartheta_j}\mathbb{E}_\vartheta[g_i(X)]\Big)_{i,j} \in \mathbb{R}^{k\times d}.$$

Dann gilt unter analogen Regularitätsbedingungen wie in Satz 2.23 für die Kovarianzmatrix von g die Ungleichung

$$\mathrm{Cov}_\vartheta(g(X)) \geq G(\vartheta)I^{-1}(f(\cdot,\vartheta))G(\vartheta)^T$$

bzgl. der Löwner-Ordnung. Hier bezeichnet

$$I(f(\cdot,\vartheta)) = \Big(\mathbb{E}\Big[\frac{\partial}{\partial\vartheta_i}\log f(X,\vartheta)\frac{\partial}{\partial\vartheta_j}\log f(X,\vartheta)\Big]\Big)_{i,j=1}^d \in \mathbb{R}^{d\times d}$$

die Fisher-Informationsmatrix von $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ bzgl. ϑ, deren Invertierbarkeit vorausgesetzt wird.

Beweis: Man geht wie im Beweis von Satz 2.23 vor. Einziger wesentlicher Unterschied ist die folgende Cauchy-Schwarz-Ungleichung für $\mathbb{R}^p$- bzw. $\mathbb{R}^q$-wertige Zufallsvektoren Y und Z, wonach

$$\mathbb{E}[YY^T] \geq \mathbb{E}[YZ^T](\mathbb{E}[ZZ^T])^{-1}\mathbb{E}[ZY^T]$$

im Sinne der Löwner-Ordnung gilt (vgl. Tripathi (1999)), sofern $\mathbb{E}[ZZ^T]$ invertierbar ist. Der formale Beweis sowie eine Diskussion der Regularitätsbedingungen wird als Übung diskutiert. □

Beispiel 2.29. In der Situation von Beispiel 2.22 ist

$$f^1(x,\vartheta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Big) .$$

Damit erhalten wir für den *Score-Vektor* die Identität

$$U_\vartheta := \Big(\frac{\partial}{\partial\mu}\log f^1(X_1,\vartheta), \frac{\partial}{\partial\sigma^2}\log f^1(X_1,\vartheta)\Big)^T = \begin{pmatrix} (X_1-\mu)/\sigma^2 \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X_1-\mu)^2 \end{pmatrix},$$

und als Fisher-Information ergibt sich

$$I(f^1(\cdot,\vartheta)) = \mathbb{E}_\vartheta[U_\vartheta U_\vartheta^T] = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & \frac{1}{2}\sigma^{-4} \end{pmatrix} = \frac{1}{n} I(f(\cdot,\vartheta)).$$

Aus Satz 2.28 folgt, dass für jeden Schätzer g für (μ,σ^2) die Ungleichung

$$\mathrm{Cov}_\vartheta(g(X)) \geq G(\vartheta) \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} G(\vartheta) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$$

gilt. Dabei ergibt sich die letzte Identität, falls g erwartungstreu ist. Insbesondere erhält man für

$$\widetilde{g}(x) = \Big(\overline{x}_n, \frac{n}{n-1}\hat{s}_n^2(x)\Big)^T = \begin{pmatrix} \overline{x}_n \\ \frac{1}{n-1}\sum_{i=1}^n (x_j - \overline{x}_n)^2 \end{pmatrix}$$

die Ungleichung

$$\mathrm{Cov}_\vartheta(\widetilde{g}(X)) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix} \geq I(f(\cdot,\vartheta)).$$

Insbesondere ist der Schätzer $\widetilde{g}$ nicht effizient.

Bemerkung 2.30. Wir haben in den Beispielen 2.26 und 2.29 ohne Beweis angenommen, dass die Regularitätsbedingungen zur Anwendung der jeweiligen Cramér-Rao-Ungleichung erfüllt sind. Wir werden im Folgenden eine Klasse von parametrischen Familien kennenlernen, bei der die Vertauschung von Integral und Ableitung immer möglich ist und zu der insbesondere die Normalverteilung gehört. Zudem lässt sich innerhalb dieser Klasse für bestimmte Schätzer nachweisen, dass die untere Schranke in der Cramér-Rao-Ungleichung stets angenommen wird.

Proposition 2.31. *Es seien μ ein σ-endliches Maß und $P_\vartheta \ll \mu$ mit μ-Dichte*

$$f(x,\vartheta) = c(\vartheta)h(x)\exp(\vartheta T(x))$$

für alle $\vartheta \in \Theta$. Dann gilt in der Cramér-Rao-Ungleichung (2.2) für $g(x) = T(x)$ stets Gleichheit.

Beweis: Wir zeigen später in Satz 2.38, dass die Funktion c beliebig oft differenzierbar ist und dass Ableitung und Integral vertauscht werden dürfen. Da $f(\cdot,\vartheta)$ eine Dichte ist, erhält man zunächst

$$c(\vartheta) = \left(\int_{\mathcal{X}} h(x)\exp(\vartheta T(x))\mu(dx) \right)^{-1} . \tag{2.3}$$

Ferner ergibt sich

$$\begin{aligned} 0 &= \frac{\partial}{\partial\vartheta}\int_{\mathcal{X}} c(\vartheta)h(x)\exp(\vartheta T(x))\mu(dx) \\ &= \int_{\mathcal{X}} h(x)\left(c'(\vartheta)+c(\vartheta)T(x)\right)\exp(\vartheta T(x))\mu(dx). \end{aligned} \tag{2.4}$$

Aus (2.3) und (2.4) lässt sich

$$\begin{aligned} \mathbb{E}_\vartheta[T(x)] &= c(\vartheta)\int_{\mathcal{X}} h(x)T(x)\exp(\vartheta T(x))\mu(dx) \\ &= -c'(\vartheta)\int_{\mathcal{X}} h(x)\exp(\vartheta T(x))\mu(dx) \\ &= -\frac{c'(\vartheta)}{c(\vartheta)} = (-\log c(\vartheta))' \end{aligned} \tag{2.5}$$

ableiten. Nach Definition der Dichte $f(\cdot,\vartheta)$ ergibt sich daher als Fisher-Information

$$I(f(\cdot,\vartheta)) = \mathbb{E}_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right)^2\right] = \mathbb{E}_\vartheta[(T(X)+(\log c(\vartheta))')^2] = \mathrm{Var}_\vartheta(T(X)).$$

Außerdem erhält man durch Differentiation in (2.5)

$$\begin{aligned} \frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta[T(X)] &= \int_{\mathcal{X}} c'(\vartheta)h(x)T(x)\exp(\vartheta T(x))\mu(dx) + \int_{\mathcal{X}} c(\vartheta)h(x)T^2(x)\exp(\vartheta T(x))\mu(dx) \\ &= \frac{c'(\vartheta)}{c(\vartheta)}\int_{\mathcal{X}} c(\vartheta)h(x)T(x)\exp(\vartheta T(x))\mu(dx) + \mathbb{E}_\vartheta[T^2(X)] \\ &= \mathbb{E}_\vartheta[T^2(X)] - (\mathbb{E}_\vartheta[T(X)])^2 = \mathrm{Var}_\vartheta(T(X)). \end{aligned}$$

Damit wird die Cramér-Rao-Ungleichung zu

$$\mathrm{Var}_\vartheta(T(X)) \geq \frac{\left(\frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta[T(X)]\right)^2}{I(f(\cdot,\vartheta))} = \mathrm{Var}_\vartheta(T(X)).$$

□

Definition 2.32.

(i) Eine Familie von Verteilungen $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ wird als *exponentielle Familie* oder *Exponentialfamilie* bezeichnet, wenn reellwertige Funktionen $c, Q_1, \ldots, Q_k : \Theta \to \mathbb{R}$ bzw. messbare Abbildungen $h, T_1, \ldots, T_k : \mathcal{X} \to \mathbb{R}$ sowie ein $\mathcal{P}$-dominierendes σ-endliches Maß μ existieren, so dass die μ-Dichten von P_ϑ die Form

$$f(x,\vartheta) = c(\vartheta)h(x)\exp\left(\sum_{j=1}^{k} Q_j(\vartheta)T_j(x)\right)$$

besitzen.

(ii) $\mathcal{P}$ wird als *k-parametrige exponentielle Familie* bezeichnet, falls sowohl die Funktionen $1, Q_1, \ldots, Q_k$ als auch die Funktionen $1, T_1, \ldots, T_k$ (letztere auf dem Komplement jeder μ-Nullmenge) linear unabhängig sind.

Bemerkung 2.33.

(i) Man kann als Verallgemeinerung von Proposition 2.31 zeigen, dass in allgemeinen exponentiellen Familien in der Cramér-Rao-Ungleichung für $(T_1, \dots, T_k)^T$ stets das Gleichheitszeichen gilt. Diese Eigenschaft charakterisiert exponentielle Familien zudem eindeutig (vgl. Theorem 3.4.2 in Bickel and Doksum (2001)).

(ii) Ohne Beschränkung der Allgemeinheit lässt sich $h(x) = 1$ annehmen, da man andernfalls zum Maß μ' mit $\mu'(dx) = h(x)\mu(dx)$ übergeht.

(iii) Wir betrachten ausschließlich k-parametrige exponentielle Familien.

Beispiel 2.34.

(i) Es sei $X \sim B(n, \vartheta)$ binomialverteilt mit Zähldichte

$$f(k, \vartheta) = \binom{n}{k} \vartheta^k (1-\vartheta)^{n-k} = (1-\vartheta)^n \binom{n}{k} \exp\left(k \log\left(\frac{\vartheta}{1-\vartheta}\right)\right).$$

In diesem Fall sind $c(\vartheta) = (1-\vartheta)^n$, $h(k) = \binom{n}{k}$, $Q_1(\vartheta) = \log\left(\frac{\vartheta}{1-\vartheta}\right)$ und $T_1(k) = k$.

(ii) Es sei $X \sim \mathcal{N}(\mu, \sigma^2)$ normalverteilt, also $\vartheta = (\mu, \sigma^2)^T$. Dann erhält man für die Dichte bzgl. des Lebesgue-Maßes

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

Man erhält $c(\vartheta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\mu^2}{2\sigma^2})$, $Q_1(\vartheta) = \frac{\mu}{\sigma^2}$, $T_1(x) = x$, $Q_2(\vartheta) = -\frac{1}{2\sigma^2}$ und $T_2(x) = x^2$.

(iii) Es sei $X \sim Po(\lambda)$ Poisson-verteilt mit Parameter $\lambda > 0$. Dann erhält man für die Zähldichte

$$f(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \frac{1}{x!} \exp(x \log \lambda).$$

In diesem Fall ergeben sich $c(\lambda) = e^{-\lambda}$, $h(x) = \frac{1}{x!}$, $T_1(x) = x$ und $Q_1(\lambda) = \log \lambda$.

Bemerkung 2.35.

(i) Ist $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine exponentielle Familie, so lässt sich leicht nachweisen, dass auch $\mathcal{P}^{(n)} := \left\{\otimes_{j=1}^n P_\vartheta \mid \vartheta \in \Theta\right\}$ eine exponentielle Familie ist.

(ii) Mit der Notation $Q(\vartheta) := (Q_1(\vartheta), \dots, Q_k(\vartheta))^T$ erhält man einen neuen Parameterraum $Q(\Theta)$, der wieder mit Θ bezeichnet wird. Die μ-Dichten haben dann die Form

$$f(x, \vartheta) = c(\vartheta) \exp\left(\sum_{j=1}^k \vartheta_j T_j(x)\right), \quad \vartheta \in \Theta,$$

und die Menge

$$\Theta^* := \left\{\vartheta \in \mathbb{R}^k \mid f(\cdot, \vartheta) \in L^1(\mu)\right\}$$

heißt *natürlicher Parameterraum der exponentiellen Familie.*

Beispiel 2.36. Für die natürlichen Parameterräume der exponentiellen Familien aus Beispiel 2.34 erhält man

(i) $X \sim B(n,\vartheta) : \Theta^* = \{\log(\frac{\vartheta}{1-\vartheta}) \mid \vartheta \in (0,1)\} = \mathbb{R}$.

(ii) $X \sim \mathcal{N}(\vartheta,\sigma^2) : \Theta^* = \{(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}) \mid \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\} = \mathbb{R} \times \mathbb{R}^-$.

(iii) $X \sim Po(\lambda) : \Theta^* = \{\log \lambda \mid \lambda \in \mathbb{R}^+\} = \mathbb{R}$.

Satz 2.37. *Der natürliche Parameterraum Θ^* einer k-parametrigen exponentiellen Familie ist konvex und hat nichtleeres Inneres.*

Beweis: Es seien $\vartheta = (\vartheta_1,\dots,\vartheta_k)^T$ und $\vartheta' = (\vartheta'_1,\dots,\vartheta'_k)^T$ jeweils Elemente von Θ^*. Mit Hilfe der Hölder-Ungleichung ergibt sich für $\alpha \in (0,1)$

$$\begin{aligned}
&\int \exp\Big\{\sum_{j=1}^k (\alpha\vartheta_j + (1-\alpha)\vartheta'_j)T_j(x)\Big\}\mu'(dx)\\
=\ &\int \Big(\exp\Big\{\sum_{j=1}^k \vartheta_j T_j(x)\Big\}\Big)^{\alpha}\Big(\exp\Big\{\sum_{j=1}^k \vartheta'_j T_j(x)\Big\}\Big)^{1-\alpha}\mu'(dx)\\
\le\ &\left(\int \exp\Big\{\sum_{j=1}^k \vartheta_j T_j(x)\Big\}\mu'(dx)\right)^{\alpha}\left(\int \exp\Big\{\sum_{j=1}^k \vartheta'_j T_j(x)\Big\}\mu'(dx)\right)^{1-\alpha} < \infty.
\end{aligned}$$

Damit ist auch $\alpha\vartheta + (1-\alpha)\vartheta' \in \Theta^*$ und Θ^* also konvex. Dass Θ^* nichtleeres Inneres besitzt, folgt unmittelbar aus der Annahme, dass $1, Q_1(\vartheta),\dots,Q_k(\vartheta)$ linear unabhängig sind. □

Satz 2.38. *Es sei $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine k-parametrige exponentielle Familie mit Dichten*

$$f(x,\vartheta) = c(\vartheta)\exp\left(\sum_{j=1}^k \vartheta_j T_j(x)\right).$$

Ferner sei $\Theta^{**} \subset \Theta^*$ offen und $\varphi \in L^1(P_\vartheta)$ für alle $\vartheta \in \Theta^{**}$. Dann ist die Funktion

$$\beta : \begin{cases} \Theta^{**} \to \mathbb{R} \\ \vartheta \mapsto \beta(\vartheta) := \int \varphi(x)\exp\left(\sum_{j=1}^k \vartheta_j T_j(x)\right)\mu(dx) \end{cases}$$

beliebig oft differenzierbar, und es gilt

$$\left(\frac{\partial}{\partial\vartheta_1}\right)^{l_1}\dots\left(\frac{\partial}{\partial\vartheta_k}\right)^{l_k}\beta(\vartheta) = \int \varphi(x)T_1^{l_1}(x)\dots T_k^{l_k}(x)\exp\left(\sum_{j=1}^k \vartheta_j T_j(x)\right)\mu(dx).$$

Beweis: vgl. Theorem 2.7.1 in Lehmann and Romano (2005). □

Bemerkung 2.39. Bevor sich die Frage nach der Qualität eines gegebenen Schätzers stellt, werden Methoden benötigt, um überhaupt geeignete Kandidaten zu erhalten. Ist $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine k-parametrige exponentielle Familie, so folgt aus Proposition 2.31, dass $T(x) = (T_1(x),\dots,T_k(x))^T$ ein effizienter Schätzer für $\mathbb{E}_\vartheta[T(X)]$ ist.

Beispiel 2.40. Sind $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, dann ist die gemeinsame Dichte durch

$$f(x, \vartheta) = c(\mu, \sigma^2) \cdot \exp\left(-\frac{n}{2\sigma^2} \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) + \frac{n\mu}{\sigma^2} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) \right)$$

mit geeigneten Konstanten $c(\mu, \sigma^2)$ gegeben (vgl. Beispiel 2.34). Offensichtlich erhält man dann mit

$$T(x) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^T$$

einen effizienten Schätzer für $(\mu, \sigma^2 + \mu^2)^T$.

Bemerkung 2.41. Besonders zwei Verfahren zur Konstruktion von Schätzern sind klassisch, falls etwa keine exponentielle Familie vorliegt:

(i) Es seien $X_1, \dots, X_n$ i.i.d. $\sim P_\vartheta$ reellwertige Zufallsvariablen, $\vartheta \in \Theta \subset \mathbb{R}^k$ und $\gamma : \Theta \to \Gamma \subset \mathbb{R}^l$. Ferner sei $m_j = \mathbb{E}_\vartheta[X_1^j] = \int x^j P_\vartheta(dx)$, $j = 1, \dots, k$, und für das zu schätzende Funktional gelte

$$\gamma(\vartheta) = f(m_1, \dots, m_k).$$

Dann erhält man einen Schätzer durch

$$\hat{\gamma}(x) = f(\hat{m}_1, \dots, \hat{m}_k),$$

wobei $\hat{m}_j := \frac{1}{n} \sum_{k=1}^{n} x_k^j$ den Mittelwertschätzer für das j-te Moment bezeichnet. Man beachte, dass wegen des starken Gesetzes der großen Zahlen gilt:

$$\hat{m}_j \to \mathbb{E}_\vartheta[X_1^j] \quad P_\vartheta\text{-f.s.}$$

Diese Schätzmethode wird als *Momentenmethode* bezeichnet.

(ii) Es sei $X \sim P_\vartheta \ll \mu$ mit Dichte $f(x, \vartheta)$, $\vartheta \in \mathbb{R}^k$, und $\gamma(\vartheta) = \vartheta$. Ein Schätzer $\hat{\theta} = \hat{\theta}(x)$ heißt *Maximum-Likelihood-Schätzer* oder kurz *ML-Schätzer*, falls

$$f(x, \hat{\theta}) = \sup_{\vartheta \in \Theta} f(x, \vartheta).$$

Beispiel 2.42. Es seien $X_1, \dots X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)^T$ und $\vartheta = (\mu, \sigma^2)^T$. Offenbar gilt

$$\vartheta = (m_1, m_2 - m_1^2)^T,$$

und mit der Momentenmethode erhält man die Schätzer

$$\hat{\gamma} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} {x_i}^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \right)^T = (\overline{x}_n, \hat{s}_n^2(x))^T.$$

Man kann zeigen: Dieselben Schätzer erhält man auch mittels der ML-Methode. Dies ist kein generelles Phänomen.

Kapitel 3

Bayes- und Minimax-Schätzer

Wir haben im vergangenen Kapitel gesehen, dass es selten möglich ist, gleichmäßig beste Schätzer zu erhalten. Eine alternative Variante zum Vergleich von Risikofunktionen sind Integration oder Maximumbildung.

Definition 3.1. Es seien $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ ein statistisches Experiment, $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ und $(\Theta, \mathcal{A}_\Theta)$ ein Messraum. Ein Wahrscheinlichkeitsmaß π auf $\mathcal{A}_\Theta$ heißt *a-priori-Verteilung* für ϑ. Für einen Schätzer $g \in \mathcal{K}$ und das zugehörige Risiko $R(\cdot, g)$ heißt

$$R(\pi, g) = \int_\Theta R(\vartheta, g)\pi(d\vartheta)$$

das *Bayes-Risiko von g bzgl. π*. Ein Schätzer $g^* \in \mathcal{K}$ heißt *Bayes-Verfahren* (oder *Bayes-Schätzer*), sofern er das Bayes-Risiko in der Klasse aller Schätzer minimiert, d.h.

$$R(\pi, g^*) = \inf_{g \in \mathcal{K}} R(\pi, g). \tag{3.1}$$

Die rechte Seite in (3.1) heißt *Bayes-Risiko*.

Bemerkung 3.2.

(i) In der Bayes'schen Interpretation ist der Parameter ϑ ebenfalls zufällig und die Realisation einer Zufallsvariablen $\theta : (\Omega, \mathcal{A}) \to (\Theta, \mathcal{A}_\Theta)$ mit Verteilung π. Man erhält das folgende Diagramm:

$$\begin{array}{ccc} (\Omega, & \mathcal{A}, \mathbb{P}) & \xrightarrow{X} (\mathcal{X}, \mathcal{B}) \\ & \theta \downarrow & \\ (\Theta, & \mathcal{A}_\Theta) & \end{array}$$

(ii) Der Ausdruck $R(\pi, g)$ beschreibt eine gemittelte Risikofunktion, wobei Werte von ϑ gemäß ihrer Wahrscheinlichkeit gewichtet werden. Die Verteilung π beschreibt ein Vorwissen des Statistikers über den unbekannten Parameter.

Annahme 3.3. Im Folgenden seien $(\Theta, \mathcal{A}_\Theta) = (\mathbb{R}^l, \mathcal{B}^l)$ und $(\mathcal{X}, \mathcal{B}) = (\mathbb{R}^n, \mathcal{B}^n)$, und es bezeichne $Q^{X,\theta}$ die Verteilung von (X, θ) auf $(\mathcal{X} \times \Theta, \mathcal{B} \otimes \mathcal{A}_\Theta)$. Man setzt P_ϑ als die bedingte Verteilung von X gegeben $\theta = \vartheta$, d.h.

$$P_\vartheta = Q^{X|\theta=\vartheta}.$$

π ist die Marginalverteilung von θ unter $Q^{X,\theta}$. Für die gemeinsame Verteilung von (X, θ) gilt dann nach dem Satz von der iterierten Erwartung (vgl. Korollar 1.15)

$$Q^{X,\theta}(A) = \int_\Theta \int_{\mathcal{X}} 1_A(x, \vartheta) P_\vartheta(dx)\pi(d\vartheta).$$

Ferner existiert die *a-posteriori-Verteilung* $Q^{\theta|X=x}$ von θ gegeben $X = x$.

Bemerkung 3.4. In der Bayes'schen Statistik werden die Größen π und P_ϑ wie folgt interpretiert: Vor Durchführung des Experiments ist $\pi = Q^\theta$ die vom Statistiker angenommene Verteilung für ϑ (a-priori-Verteilung). Nach Beobachtung von $X(\omega) = x$ ändert sich das Wissen über ϑ von π zu $Q^{\theta|X=x}$ (a-posteriori-Verteilung). Für das Risiko bei der Schätzung von $\gamma(\vartheta)$ erhält man dann die folgenden Darstellungen

$$\begin{aligned} R(\pi, g) &= \int_\Theta R(\vartheta, g)\pi(d\vartheta) = \int_\Theta \int_{\mathcal{X}} L(\gamma(\vartheta), g(x)) P_\vartheta(dx)\pi(d\vartheta) \\ &= \int_{\Theta\times\mathcal{X}} L(\gamma(\vartheta), g(x)) Q^{X,\theta}(dx, d\vartheta) \\ &= \int_{\mathcal{X}} \int_\Theta L(\gamma(\vartheta), g(x)) Q^{\theta|X=x}(d\vartheta) Q^X(dx) = \int_{\mathcal{X}} R_\pi^x(g) Q^X(dx). \end{aligned}$$

Die Größe

$$R_\pi^x(g) := \int_\Theta L(\gamma(\vartheta), g(x)) Q^{\theta|X=x}(d\vartheta)$$

heißt *a-posteriori-Risiko* von g bei gegebenem $X = x$.

Satz 3.5.

(i) g^* ist genau dann ein Bayes-Schätzer für ϑ, wenn g^* das a-posteriori Risiko $R_\pi^X(g)$ Q^X-f.s. minimiert, d.h.

$$R_\pi^x(g^*) = \inf_{g\in\mathcal{K}} R_\pi^x(g) = \inf_{a\in\Theta} \int L(\vartheta, a) Q^{\theta|X=x}(d\vartheta) \quad Q^X\text{-f.s.}$$

gilt.

(ii) Es seien $\Theta \subset \mathbb{R}$, $L(\vartheta, a) = (\vartheta - a)^2$ und $\int \vartheta^2 Q^{\theta|X=x}(d\vartheta) < \infty$ Q^X-f.s. Dann ist der Bayes-Schätzer für ϑ durch

$$g^*(x) = \mathbb{E}[\theta|X = x] = \int_\Theta \vartheta Q^{\theta|X=x} d\vartheta$$

gegeben.

Beweis:

(i) $R(\pi, g)$ wird gemäß Bemerkung 3.4 minimal bzgl. g genau dann, wenn $R_\pi^X(g)$ minimal bzgl. g Q^X-f.s. wird. Beachte dabei, dass wir über die Klasse aller Schätzer minimieren. Im Falle der Existenz muss g^* das Risiko $R_\pi^X(g)$ Q^X-f.s. minimieren, da man sonst einen besseren Schätzer mit kleinerem Risiko konstruieren könnte.

(ii) Gemäß Satz 1.25 wird der Ausdruck $\int_\Theta (\vartheta - a)^2 Q^{\theta|X=x}(d\vartheta)$ minimal für $a = \mathbb{E}[\theta|X = x]$. □

Bemerkung 3.6. Falls $P_\vartheta \ll \mu$ mit μ-Dichte $f(x|\vartheta)$ von $Q^{X|\theta=\vartheta}$ gilt und ferner $\pi \ll \nu$ mit ν-Dichte $h(\vartheta)$ ist, dann gilt für die gemeinsame Verteilung von (X, Q)

$$Q^{X,\theta} \ll \mu \otimes \nu$$

mit zugehöriger Dichte $f(x|\vartheta)\, h(\vartheta)$. Außerdem hat die a-posteriori Verteilung $Q^{\theta|X=x}$ die ν-Dichte

$$f(\vartheta|x) = \begin{cases} \frac{f(x|\vartheta)h(\vartheta)}{\int_\Theta f(x|\vartheta)h(\vartheta)\nu(d\vartheta)}, & \text{falls } \int_\Theta f(x|\vartheta)h(\vartheta)\nu(d\vartheta) > 0, \\ \text{beliebige } \nu\text{-Dichte}, & \text{falls } \int_\Theta f(x|\vartheta)h(\vartheta)\nu(d\vartheta) = 0. \end{cases}$$

Für das a-posteriori- und das Bayes-Risiko erhält man dann

$$R^x_\pi(g) = \frac{\int_\Theta L(\vartheta, g(x)) f(x|\vartheta)h(\vartheta)\nu(d\vartheta)}{\int_\Theta f(x|\vartheta)h(\vartheta)\nu(d\vartheta)},$$
$$R(\pi, g) = \int_{\mathcal{X}} \int_\Theta L(\vartheta, g(x)) f(x|\vartheta)h(\vartheta)\nu(d\vartheta)\mu(dx).$$

Beispiel 3.7. Es seien $\Theta = (0, 1)$, $\mathcal{X} = \{0, \ldots, n\}$ und

$$P_\vartheta(X = x) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x} .$$

Wir betrachten den quadratischen Verlust $L(x, y) = (x - y)^2$ bei der Schätzung von ϑ.

(i) Mit Hilfe von Proposition 2.31 und Beispiel 2.34 zeigt man leicht, dass $g(x) = \frac{x}{n}$ ein erwartungstreuer Schätzer mit gleichmäßig kleinster Varianz ist. Konkret gilt:

$$\mathrm{Var}_\vartheta(g(X)) = \frac{\vartheta(1 - \vartheta)}{n}.$$

(ii) Ist $\pi \sim \mathcal{U}(0, 1)$ gleichverteilt auf $(0, 1)$ und μ das Zählmaß, dann ist

$$f(x|\vartheta) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x}$$

und man erhält für die Dichte der gemeinsamen Verteilung $Q^{X,\theta}$

$$\binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x} 1_{(0,1)}(\vartheta) 1_{\{0,\ldots,n\}}(x).$$

Man bestimmt die a-posteriori Verteilung $Q^{\theta|X=x}$ oft dadurch, dass man $f(x|\vartheta)$ und $h(\vartheta)$ multipliziert und bereits anhand des Produkts die Verteilung erkennt. Die Standardisierung im Nenner ergibt sich dann sofort. In diesem Fall ist das durch die Gleichverteilung π besonders einfach und man erhält die Dichte

$$\frac{\vartheta^x (1 - \vartheta)^{n-x} 1_{(0,1)}(\vartheta)}{\beta(x + 1, n - x + 1)} ,$$

wobei für $a > -1, b > -1$

$$\beta(a,b) = \int_0^1 \vartheta^{a-1}(1-\vartheta)^{b-1} d\vartheta$$

das Beta-Integral bezeichnet. Der Bayes-Schätzer bei quadratischem Verlust lautet gemäß Satz 3.5 dann

$$g^*(x) = \mathbb{E}[\theta|X=x] = \frac{\int_0^1 \vartheta^{x+1}(1-\vartheta)^{n-x} d\vartheta}{\beta(x+1, n-x+1)} = \frac{\beta(x+2, n-x+1)}{\beta(x+1, n-x+1)} = \frac{x+1}{n+2}.$$

Eine einfache Rechnung ergibt für das Bayes-Risiko

$$\begin{aligned} R(\pi, g^*) &= \int_0^1 R(\vartheta, g^*) d\vartheta = \int_0^1 \mathbb{E}_\vartheta \left[\frac{(X+1-\vartheta(n+2))^2}{(n+2)^2} \right] d\vartheta \\ &= \frac{1}{(n+2)^2} \int_0^1 \left(n\vartheta - n\vartheta^2 + 1 - 4\vartheta + 4\vartheta^2 \right) d\vartheta = \frac{1}{6(n+2)}. \end{aligned}$$

Beispiel 3.8. Es seien $X_1, \ldots, X_n$ i.i.d. $\sim P_\mu^1 = \mathcal{N}(\mu, \sigma^2)$ mit bekanntem $\sigma^2 > 0$, so dass sich $\mathcal{X} = \mathbb{R}^n$ und $P_\mu = \otimes_{i=1}^n P_\mu^1$ ergeben. Wir nehmen an, dass die a-priori-Verteilung π durch eine Normalverteilung mit Dichte

$$h(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\mu - \mu_0)^2 \right)$$

gegeben ist. Mittels der Dichte

$$f(x|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2 \right).$$

erhält man für die a-posteriori-Verteilung für μ aus dem Produkt der beiden Dichten

$$Q^{\theta|X=x} \sim \mathcal{N}\left(g_{\mu_0,\tau^2}(x), \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1} \right)$$

mit

$$g_{\mu_0,\tau^2}(x) = \left(1 + \frac{\sigma^2}{n\tau^2} \right)^{-1} \overline{x}_n + \left(\frac{n\tau^2}{\sigma^2} + 1 \right)^{-1} \mu_0.$$

Bei quadratischer Verlustfunktion ist dann g_{μ_0,τ^2} der Bayes-Schätzer für μ.

Man beachte: Für $\tau^2 \to \infty$ gilt $g_{\mu_0,\tau^2}(x) \approx \overline{x}_n$ (geringe Vorinformation über μ), während man für $\tau^2 \to 0$ die Approximation $g_{\mu_0,\tau^2}(x) \approx \mu_0$ (große Vorinformation über μ) erhält.

Definition 3.9. Es sei g ein Schätzer für $\gamma(\vartheta)$. Dann heißt die Größe

$$R^*(g) = \sup_{\vartheta \in \Theta} R(\vartheta, g)$$

maximales Risiko von g, und

$$\inf_{g \in \mathcal{K}} R^*(g) = \inf_{g \in \mathcal{K}} \sup_{\vartheta \in \Theta} R(\vartheta, g)$$

wird als *Minimax-Risiko* bezeichnet. Ein Schätzer $g^* \in \mathcal{K}$ heißt *Minimax-Schätzer* bzw. *minimax-optimal*, falls g^* das maximale Risiko minimiert, d.h.

$$R^*(g^*) = \inf_{g \in \mathcal{K}} R^*(g)$$

gilt.

Bemerkung 3.10.

(i) Bei Verwendung von Minimax-Schätzern schützt man sich gegen große Verluste.

(ii) Es sei $\mathcal{M} = \{\pi \mid \pi \text{ Wahrscheinlichkeitsmaß auf } \mathcal{A}_\Theta\}$. Dann zeigt man leicht:

$$R^*(g) = \sup_{\pi \in \mathcal{M}} R(\pi, g).$$

Definition 3.11. Eine a-priori-Verteilung π^* auf $\mathcal{A}_\Theta$ heißt *ungünstigste a-priori-Verteilung* genau dann, wenn gilt:

$$\inf_{g \in \mathcal{K}} R(\pi^*, g) \geq \inf_{g \in \mathcal{K}} R(\pi, g) \quad \forall \pi \in \mathcal{M}.$$

In diesem Fall ist das Bayes-Risiko als Funktion von π also maximal.

Satz 3.12.

(i) Ist g_π ein Bayes-Schätzer für $\gamma(\vartheta)$ bzgl. der a-priori-Verteilung π und gilt

$$R(\pi, g_\pi) = \sup_{\vartheta \in \Theta} R(\vartheta, g_\pi), \tag{3.2}$$

dann ist g_π auch ein Minimax-Schätzer.

(ii) Ist g_π ein eindeutig bestimmter Bayes-Schätzer mit der Eigenschaft (3.2), dann ist g_π ebenfalls ein eindeutig bestimmter Minimax-Schätzer.

(iii) Gilt (3.2), dann ist die a-priori-Verteilung π eine ungünstigste a-priori-Verteilung.

Beweis: Für alle Schätzer $g \in \mathcal{K}$ gilt

$$\sup_{\vartheta \in \Theta} R(\vartheta, g) \geq \int_\Theta R(\vartheta, g)\pi(d\vartheta) \geq \int_\Theta R(\vartheta, g_\pi)\pi(d\vartheta) = R(\pi, g_\pi) = \sup_{\vartheta \in \Theta} R(\vartheta, g_\pi),$$

wobei die erste Ungleichung aufgrund des Mittelns über alle ϑ und die zweite nach Definition des Bayes-Schätzers gilt. Die letzte Identität folgt aus (3.2). Damit folgt, dass g_π ein Minimax-Schätzer ist. Im Falle der Eindeutigkeit von g_π gilt

$$\int_\Theta R(\vartheta, g)\pi(d\vartheta) > \int_\Theta R(\vartheta, g_\pi)\pi(d\vartheta),$$

also ist der Minimax-Schätzer eindeutig bestimmt.

Zuletzt gilt für jedes Wahrscheinlichkeitsmaß μ die Ungleichung

$$\inf_{g \in \mathcal{K}} \int_\Theta R(\vartheta, g)\mu(d\vartheta) \leq \int_\Theta R(\vartheta, g_\pi)\mu(d\vartheta) \leq \sup_{\vartheta \in \Theta} R(\vartheta, g_\pi).$$

Gilt (3.2), so ist der Ausdruck auf der rechten Seite identisch zu

$$R(\pi, g_\pi) = \inf_{g \in \mathcal{K}} \int_\Theta R(\vartheta, g)\pi(d\vartheta),$$

und damit ist π eine ungünstigste a-priori-Verteilung. □

Bemerkung 3.13. Satz 3.12 ist für die Bestimmung von Minimax-Schätzern hilfreich. Unter Umständen ist für eine Bayes-Schätzung g_π mit a-priori-Verteilung π die Risikofunktion konstant, d.h. es gilt $R(\vartheta, g_\pi) = c$ für alle $\vartheta \in \Theta$. In diesem Fall erhält man

$$\sup_{\vartheta \in \Theta} R(\vartheta, g_\pi) = c = \int_\Theta R(\vartheta, g_\pi)\pi(d\vartheta) = R(\pi, g_\pi).$$

Wegen Satz 3.12 ist g_π dann Minimax-Schätzer und π eine ungünstigste a-priori Verteilung. Diese muss jedoch nicht immer existieren.

Beispiel 3.14. Es seien $\Theta = (0,1)$, $\mathcal{X} = \{0, 1, \dots, n\}$ und

$$P_\vartheta(X = x) = \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x}.$$

Wir betrachten wieder den quadratischen Verlust $L(x,y) = (x-y)^2$ und verwenden als a-priori-Verteilung im Gegensatz zu Beispiel 3.7 eine Beta-Verteilung $\pi_{a,b} \sim B(a,b)$ mit Parametern a, b und Dichte

$$h(\vartheta) = \frac{1}{\beta(a,b)} \vartheta^{a-1}(1-\vartheta)^{b-1} 1_{(0,1)}(\vartheta).$$

Mit ähnlichen Argumenten wie in Beispiel 3.7 zeigt man, dass die a-posteriori Verteilung die Dichte

$$\frac{1}{\beta(a+x, n-x+b)} \vartheta^{a+x-1}(1-\vartheta)^{n-x+b-1} 1_{(0,1)}(\vartheta)$$

besitzt, d.h. es gilt

$$Q^{\vartheta|X=x} \sim B(a+x, n-x+b).$$

Für $Z \sim \beta(p,q)$ lässt sich

$$\mathbb{E}[Z] = \frac{p}{p+q} \quad \text{und} \quad \operatorname{Var}(Z) = \frac{pq}{(p+q)^2(p+q+1)}$$

nachweisen. Damit ist nach Satz 3.5

$$g_{a,b}(x) = \frac{x+a}{n+a+b}$$

der Bayes-Schätzer bezüglich $\pi_{a,b}$. Für dessen Risiko ergibt sich

$$\begin{aligned} R(\vartheta, g_{a,b}) =& \mathbb{E}_\vartheta[(g_{a,b}(X) - \vartheta)^2] = \frac{E_\vartheta[(X - n\vartheta + a - \vartheta(a+b))^2]}{(n+a+b)^2} \\ =& \frac{\vartheta^2[-n + (a+b)^2] + \vartheta[n - 2a(a+b)] + a^2}{(n+a+b)^2}. \end{aligned}$$

Wählt man $a^* = b^* = \sqrt{n}/2$, dann gilt

$$R(\vartheta, g_{a^*,b^*}) = \frac{n}{4(n+\sqrt{n})^2},$$

und die Risikofunktion ist konstant. Nach Satz 3.12 ist damit

$$g_{a^*,b^*}(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}}$$

Minimax-Schätzer für ϑ und $\pi_{\sqrt{n}/2,\sqrt{n}/2}$ eine ungünstigste a-priori-Verteilung.

Definition 3.15. Es sei

$$r_\pi := \inf_{g \in \mathcal{K}} R(\pi, g)$$

für $\pi \in \mathcal{M}$. Eine Folge $(\pi_m)_{m\in\mathbb{N}}$ von Wahrscheinlichkeitsmaßen heißt *ungünstigste Folge von a-priori-Verteilungen*, falls gilt:

(i) $\lim_{m\to\infty} r_{\pi_m} = r$.

(ii) Für alle $\pi \in \mathcal{M}$ gilt $r_\pi \leq r$.

Satz 3.16. *Es sei $(\pi_m)_{m\in\mathbb{N}}$ eine Folge von Wahrscheinlichkeitsmaßen und es existieren ein $r \in \mathbb{R}$ mit $\lim_{m\to\infty} r_{\pi_m} = r$ und ein $g^* \in \mathcal{K}$ mit*

$$\sup_{\vartheta\in\Theta} R(\vartheta, g^*) = r.$$

Dann gilt:

(i) g^* ist Minimax-Schätzer.

(ii) $(\pi_m)_{m\in\mathbb{N}}$ ist eine Folge ungünstigster a-priori Verteilungen.

Beweis:

(i) Für alle $g \in \mathcal{K}$ gilt

$$\sup_{\vartheta\in\Theta} R(\vartheta, g) \geq \int_\Theta R(\vartheta, g)\pi_m(d\vartheta) \geq r_{\pi_m} \longrightarrow r = \sup_{\vartheta\in\Theta} R(\vartheta, g^*).$$

Damit ist g^* Minimax-Schätzer.

(ii) Für alle $\pi \in \mathcal{M}$ gilt

$$r_\pi \leq R(\pi, g^*) = \int_\Theta R(\vartheta, g^*)\pi(d\vartheta) \leq \sup_{\vartheta\in\Theta} R(\vartheta, g^*) = r.$$

Damit ist $(\pi_m)_{m\in\mathbb{N}}$ eine Folge ungünstigster a-priori Verteilungen. □

Beispiel 3.17. Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ mit bekannter Varianz $\sigma^2 > 0$. Wir verwenden als a-priori-Verteilung π_m eine Normalverteilung mit Dichte

$$\frac{1}{\sqrt{2\pi m}} \exp\left\{-\frac{(\mu - \mu_0)^2}{2m}\right\}.$$

Diese Verteilung ergibt sich aus der folgenden heuristischen Interpretation: Eine ungünstigste a-priori-Verteilung bedeutet für den statistischen Anwender die Entscheidung mit dem größtmöglichen Risiko. Es ist naheliegend zu vermuten, dass dies mit der unsichersten Vorinformation einhergeht. In unserem Fall entspräche dies einer Verteilung gemäß dem Lebesgue-Maß λ auf $\mathbb{R}$, das allerdings kein Wahrscheinlichkeitsmaß ist. Wir approximieren λ daher durch π_m für $m \to \infty$. Nach Beispiel 3.8 ist der Bayes-Schätzer bezüglich π_m bei quadratischem Verlust durch

$$g_m(x) = \left(1 + \frac{\sigma^2}{nm}\right)^{-1} \overline{x}_n + \left(1 + \frac{nm}{\sigma^2}\right)^{-1} \mu_0$$

gegeben. Für jedes $\mu \in \mathbb{R}$ ergibt sich

$$\begin{aligned} R(\mu, g_m) &= \mathbb{E}_\mu[(g_m(X) - \mu)^2] \\ &= \mathbb{E}_\mu \left[\left(\left(1 + \frac{\sigma^2}{nm}\right)^{-1} (\overline{x}_n - \mu) + \left(1 + \frac{nm}{\sigma^2}\right)^{-1} (\mu_0 - \mu) \right)^2 \right] \\ &= \frac{\sigma^2}{n} \left(1 + \frac{\sigma^2}{nm}\right)^{-2} + \left(1 + \frac{nm}{\sigma^2}\right)^{-2} (\mu - \mu_0)^2 \longrightarrow \frac{\sigma^2}{n} =: r. \end{aligned}$$

Für das Bayes-Risiko bezüglich π_m erhält man daher nach dem Satz von der majorisierten Konvergenz

$$r_{\pi_m} = R(\pi_m, g_m) = \int_{\mathbb{R}} R(\mu, g_m) \pi_m(d\mu) \longrightarrow r,$$

denn

$$R(\mu, g_m) \leq \frac{\sigma^2}{n} + (\mu - \mu_0)^2.$$

Zuletzt gilt für $g^*(x) = \overline{x}_n$ offenbar

$$R(\mu, g^*) = \mathbb{E}_\mu[(\overline{X}_n - \mu)^2] = \frac{\sigma^2}{n} = r,$$

so dass sich nach Satz 3.16 ergibt, dass g^* der Minimax-Schätzer und $(\pi_m)_m$ wie erwartet eine Folge ungünstigster a-priori-Verteilungen ist.

Kapitel 4

Suffizienz und Vollständigkeit

Oftmals sind wir mit dem statistischen Problem konfrontiert, auf Basis von Beobachtungen $X_1, \dots, X_n$ eine Entscheidung etwa zur Schätzung eines Parameters zu treffen. Ein Ziel in der statistischen Anwendung ist daher die Reduktion dieses Problems auf handlichere Größen als den unter Umständen hochdimensionalen Vektor $X = (X_1, \dots, X_n)^T$. In diesem Kapitel erläutern wir Konzepte, die dies ermöglichen. Wie zuvor seien $(\mathcal{X}, \mathcal{B}, P)$ ein Wahrscheinlichkeitsraum mit $P \in \mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$. Ferner sei $(\mathcal{T}, \mathcal{D})$ ein Messraum.

Definition 4.1. Eine $\mathcal{B}$-$\mathcal{D}$-messbare Abbildung $T : \mathcal{X} \to \mathcal{T}$ heißt *Statistik*.

Beispiel 4.2. Es seien $X_1, \dots, X_n$ i.i.d. $\sim B(1, \vartheta)$ mit gemeinsamer Zähl-Dichte

$$f(x, \vartheta) = \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i} = \vartheta^{T(x)}(1-\vartheta)^{n-T(x)},$$

wobei

$$T(x) = \sum_{i=1}^n x_i, \quad x = (x_1, \dots, x_n)^T,$$

gesetzt wurde. Wählt man $u_1, \dots, u_n \in \{0, 1\}$, so gilt

$$P_\vartheta\left(X_i = u_i \text{ für alle } i = 1, \dots, n | T(X) = k\right) = \begin{cases} 0, & \text{falls } \sum_{i=i}^n u_i \neq k \\ \frac{1}{\binom{n}{k}}, & \text{falls } \sum_{i=1}^n u_i = k, \end{cases} \tag{4.1}$$

denn alle Kombinationen von k Erfolgen bei n Versuchen besitzen dieselbe Wahrscheinlichkeit. Wir erkennen aus (4.1), dass man bei bereits vorhandener Kenntnis von $T(X)$ aus der zusätzlichen Kenntnis des Vektors $X = (X_1, \dots, X_n)^T$ keine weiteren Informationen über ϑ gewinnt.

Definition 4.3.

(i) Eine σ-Algebra $\mathcal{C} \subset \mathcal{B}$ heißt *suffizient* für ϑ, falls für alle $B \in \mathcal{B}$ gilt:

$$k_B = P_\vartheta(B|\mathcal{C}) \quad \forall \vartheta \in \Theta,$$

d.h. die auf $\mathcal{C}$ bedingten P_ϑ-Wahrscheinlichkeiten sind unabhängig von ϑ.

(ii) Eine Statistik $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ heißt *suffizient* für ϑ, falls $\sigma(T)$ suffizient für ϑ ist.

Bemerkung 4.4.

(i) Aus dem Faktorisierungslemma (Satz 1.18) folgt sofort, dass T genau dann suffizient für ϑ ist, wenn für jedes $B \in \mathcal{B}$ eine von $\vartheta \in \Theta$ unabhängige Funktion $h_B : \mathcal{T} \to \mathbb{R}$ mit

$$h_B(t) = P_\vartheta(B|T = t)$$

existiert.

(ii) In Beispiel 4.2 ist

$$T(x) = \sum_{i=1}^{n} x_i$$

suffizient für ϑ. Die dort diskutierte Interpretation der Suffizienz greift auch im allgemeinen Fall aus Definition 4.3: Bei Kenntnis von $\mathcal{C}$ erhält man aus $\mathcal{B}$ keine zusätzlichen Informationen über ϑ.

(iii) Sei $g : \mathcal{X} \to \mathbb{R}$ aus $L^1(\mathcal{P}) = \bigcap_{\vartheta\in\Theta} L^1(P_\vartheta)$. Dann gilt:

(a) Ist $\mathcal{C}$ suffizient für ϑ, so existiert eine von ϑ unabhängige Version k der bedingten Erwartung

$$k = \mathbb{E}_\vartheta[g|\mathcal{C}].$$

Diese Relation ergibt sich aus Definition 4.3 und der üblichen „maßtheoretischen Induktion“.

Beachte: $k = k(\cdot)$ ist eine Zufallsvariable, und $x \in \mathcal{X}$ spielt die Rolle von $\omega \in \Omega$.

(b) Ist $T : \mathcal{X} \to \mathcal{T}$ suffizient für ϑ, so existiert eine von ϑ unabhängige Version h des bedingten Erwartungswerts

$$h(t) = \mathbb{E}_\vartheta[g|T = t].$$

Beispiel 4.5. Es sei Q eine endliche Gruppe von messbaren Transformationen von $(\mathcal{X}, \mathcal{B})$ auf sich. Mit dem Mengensystem

$$\mathcal{C} = \mathcal{C}(Q) = \{B \in \mathcal{B} \mid B = \pi(B)\ \forall \pi \in Q\}$$

bezeichnen wir das System der Q-invarianten Mengen. Da Q eine Gruppe ist, ist jede Transformation bijektiv, und man weist leicht nach, dass $\mathcal{C}$ eine σ-Algebra ist.

Ist dann $\mathcal{P}$ invariant bezüglich Q, d.h.

$$P_\vartheta^\pi(B) = P_\vartheta(\pi^{-1}(B)) = P_\vartheta(B) \quad \forall B \in \mathcal{B}\ \forall \pi \in Q \text{ und } \forall \vartheta \in \Theta,$$

und $g \in L^1(\mathcal{P})$, so ist

$$k(x) = \frac{1}{q} \sum_{\pi\in Q} g(\pi(x))$$

mit $q = |Q|$ eine von ϑ unabhängige Version von $\mathbb{E}_\vartheta[g|\mathcal{C}]$. Also sind $\mathcal{C}$ (und jede Statistik T mit $\sigma(T) = \mathcal{C}$) suffizient für ϑ.

Beweis: Es ist $k(x) = k(\pi(x))$, da Q eine Gruppe ist, und also ist k $\mathcal{C}$-messbar. Aus der Transformationsformel erhält man zudem für alle $C \in \mathcal{C}$ die Relation

$$\begin{aligned}\int_C k(x)P_\vartheta(dx) &= \frac{1}{q}\sum_{\pi\in Q}\int_C g(\pi(x))P_\vartheta(dx) = \frac{1}{q}\sum_{\pi\in Q}\int_{\pi C} g(x)P_\vartheta^\pi(dx)\\ &= \frac{1}{q}\sum_{\pi\in Q}\int_{\pi C} g(x)P_\vartheta(dx) = \int_C g(x)P_\vartheta(dx).\end{aligned}$$

□

Beispiel 4.6. Es seien $X_1, \ldots, X_n$ i.i.d. mit Verteilungsfunktion F und

$$\mathcal{P} = \{P \sim F^n \mid F : \mathbb{R} \to [0,1] \text{ Verteilungsfunktion }\}, \quad F^n(x) = \prod_{i=1}^n F(x_i).$$

Ist $\mathcal{X} = \mathbb{R}^n, \mathcal{B} = \mathcal{B}^n$ und

$$Q = S_n = \{\pi : \mathbb{R}^n \to \mathbb{R}^n \mid \pi \text{ Permutation}\}$$

die Permutationsgruppe, dann gilt $P^\pi = P$ für alle $\pi \in S_n$ und $\mathcal{P}$ ist invariant bzgl. S_n. Insbesondere ist $\mathcal{C} = \mathcal{C}(S_n)$ suffizient.

Definition 4.7. Es seien $\mathbb{R}^n_\leq = \{x \in \mathbb{R}^n \mid x_1 \leq \ldots \leq x_n\}$ und $x_{(j)}$ die j-kleinste Zahl unter $x_1, \ldots, x_n$. Die Statistik

$$T : \begin{cases} \mathbb{R}^n \to \mathbb{R}^n_\leq \\ x \mapsto x_{(\cdot)} = (x_{(1)}, \ldots, x_{(n)})^T \end{cases}$$

heißt *Ordnungsstatistik* von $x_1, \ldots, x_n$.

Bemerkung 4.8.

(i) Man kann zeigen:

$$\sigma(T) = \mathcal{C}(S_n),$$

d.h. die Statistik T ist suffizient für F. Mit anderen Worten: Bei Untersuchungen von F reicht es aus, die der Größe nach sortierten Beobachtungen zu betrachten.

(ii) Die Statistik

$$\widetilde{T} : \begin{cases} \mathbb{R}^n \to \mathbb{R}^n \\ x \mapsto \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2, \ldots, \sum_{i=1}^n x_i^n\right)^T \end{cases}$$

ist ebenfalls suffizient für F, denn es gilt $\sigma(T) = \sigma(\widetilde{T})$ (vgl. Example 2.4.1 in Lehmann and Romano, 2005).

Definition 4.9. Es sei $\mathcal{P} := \{P_\vartheta \mid \vartheta \in \Theta\}$ eine Klasse von Wahrscheinlichkeitsmaßen. Dann heißt ein Maß ν zu $\mathcal{P}$ *äquivalent*, wenn

$$\nu(N) = 0 \iff P_\vartheta(N) = 0 \quad \forall \vartheta \in \Theta$$

gilt.

Der folgende Satz klärt, dass man in einem statistischen Modell mit $P_\vartheta \ll \mu$ für alle $\vartheta \in \Theta$ immer zu einem minimalen Maß $\nu \ll \mu$ übergehen kann, das zu $\mathcal{P}$ äquivalent ist. Außerdem wird ein Kriterium für Suffizienz angegeben, das direkt mit der Dichte bzgl. ν zusammenhängt.

Satz 4.10. (Halmos-Savage) *Es seien μ ein σ-endliches Maß und $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsmaßen mit $\mathcal{P} \ll \mu$, d.h. $P_\vartheta \ll \mu$ für alle $\vartheta \in \Theta$. Dann gilt:*

(i) Es existiert ein zu $\mathcal{P}$ äquivalentes Maß ν der Form

$$\nu = \sum_{i=1}^{\infty} c_i P_{\vartheta_i}, \quad \text{wobei } c_i \geq 0 \text{ für alle } j \text{ und } \sum_{i=1}^{\infty} c_i = 1 \tag{4.2}$$

gelten. Insbesondere wird nur über eine geeignete abzählbare Teilklasse der Wahrscheinlichkeitsmaße P_ϑ summiert.

(ii) Die σ-Algebra $\mathcal{C} \subset \mathcal{B}$ ist suffizient für ϑ genau dann, wenn eine $\mathcal{C}$-messbare Version von $\frac{dP_\vartheta}{d\nu}$ für alle $\vartheta \in \Theta$ existiert.

(iii) $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ ist suffizient für ϑ genau dann, wenn für alle $\vartheta \in \Theta$ eine $\mathcal{D}$-messbare Funktion $q_\vartheta(\cdot)$ mit

$$\frac{dP_\vartheta}{d\nu}(x) = q_\vartheta(T(x))$$

existiert.

Beweis:

(i) Wir zeigen die Aussage nur im Fall eines endlichen Maßes μ und setzen

$$\mathcal{J} := \{\nu \mid \nu \text{ Wahrscheinlichkeitsmaß der Form (4.2)}\}.$$

Außerdem bezeichnen wir mit

$$q = \frac{dQ}{d\mu}, \quad Q \in \mathcal{J},$$

eine Version der Dichte von Q bzgl. μ. Offensichtlich genügt es, die Existenz eines Maßes $Q_0 \in \mathcal{J}$ zu zeigen, so dass

$$Q_0(A) = 0 \Longrightarrow Q(A) = 0 \quad \forall Q \in \mathcal{J} \tag{4.3}$$

gilt, denn wegen $P_\vartheta \in \mathcal{J}$ folgt dann $P_\vartheta \ll Q_0$ für alle $\vartheta \in \Theta$. Die umgekehrte Relation gilt nach Konstruktion.

Zum Beweis von (4.3) betrachten wir das Mengensystem

$$\mathcal{C} := \{B \in \mathcal{B} \mid \exists Q \in \mathcal{J} \text{ mit } Q(B) > 0 \text{ und } q(x) > 0 \text{ für } \mu\text{-fast alle } x \in B\}$$

und setzen $\sup_{B \in \mathcal{C}} \mu(B) = r$. Dann gibt es Mengen $B_i \in \mathcal{C}$ mit $\mu(B_i) \to r$, und für

$$B_0 = \bigcup_{i \in \mathbb{N}} B_i$$

gilt $\mu(B_0) = r$. Bezeichnet Q_i das zu $B_i \in \mathcal{C}$ gewählte Q, setzen wir

$$Q_0 = \sum_{i=1}^{\infty} c_i Q_i \in \mathcal{J}$$

für $c_i > 0$ mit $\sum_i c_i = 1$. Q_0 besitzt die μ-Dichte

$$\frac{dQ_0}{d\mu}(x) = q_0(x) = \sum_{i=1}^{\infty} c_i q_i(x).$$

Offensichtlich ist $q_0(x) > 0$ für μ-fast alle $x \in B_0$, und damit ist $B_0 \in \mathcal{C}$.

Seien nun $Q_0(A) = 0$ und $Q \in \mathcal{J}$ beliebig. Für dieses Q bezeichne q die μ-Dichte, und wir setzen $B = \{x \mid q(x) > 0\}$. Dann erhalten wir

$$0 = Q_0(A \cap B_0) = \int\limits_{A \cap B_0} q_0 d\mu.$$

Wegen $q_0 > 0$ auf B_0 folgt $\mu(A \cap B_0) = 0$ und damit auch $Q(A \cap B_0) = 0$ für jedes $Q \in \mathcal{J}$. Außerdem ergibt sich aus der Definition von B sofort $Q(A \cap B_0^c \cap B^c) = 0$.

Wäre zuletzt $Q(A \cap B_0^c \cap B) > 0$, so wären sowohl B_0 als auch $A \cap B_0^c \cap B$ Elemente von $\mathcal{C}$. Insbesondere folgt, dass auch deren (disjunkte) Vereinigung in $\mathcal{C}$ liegt, indem man eine geeignete Linearkombination der beiden Darstellungen von Q und Q_0 gemäß (4.2) wählt. Da zudem auch $\mu(A \cap B_0^c \cap B) > 0$ gilt, würden sich

$$\mu(B_0 \cup (A \cap B_0^c \cap B)) = \mu(B_0) + \mu(A \cap B_0^c \cap B) > \mu(B_0)$$

und ein Widerspruch zur Maximalität von B_0 in $\mathcal{C}$ ergeben. Damit gilt neben $Q(A \cap B_0) = 0$ auch $Q(A \cap B_0^c) = 0$ und also (4.3).

(ii) Es sei zunächst $\mathcal{C}$ suffizient. Dann gibt es für jedes $B \in \mathcal{B}$ eine $\mathcal{C}$-messbare Funktion k_B, unabhängig von ϑ, mit der Eigenschaft

$$\int_C k_B dP_\vartheta = \int_C 1_B dP_\vartheta \quad \forall C \in \mathcal{C}\ \forall \vartheta \in \Theta.$$

Aus (i) erhält man

$$\int_C k_B d\nu = \int_C 1_B d\nu \quad \forall C \in \mathcal{C},$$

also gilt $k_B = \mathbb{E}_\nu[1_B|\mathcal{C}]$. Es ergibt sich

$$P_\vartheta(B) = \int_{\mathcal{X}} 1_B dP_\vartheta = \int_{\mathcal{X}} k_B dP_\vartheta = \int_{\mathcal{X}} \mathbb{E}_\nu[1_B|\mathcal{C}] dP_\vartheta.$$

Bezeichnet nun $f_\vartheta^{\mathcal{C}} = \frac{dP_\vartheta^{\mathcal{C}}}{d\nu}$ die ν-Dichte von P_ϑ eingeschränkt auf die σ-Algebra $\mathcal{C}$, so folgt aufgrund der $\mathcal{C}$-Messbarkeit (Satz 1.3) von $f_\vartheta^{\mathcal{C}}$ gemäß Satz 1.12

$$P_\vartheta(B) = \int_{\mathcal{X}} \mathbb{E}_\nu[1_B|\mathcal{C}] f_\vartheta^{\mathcal{C}} d\nu = \int_{\mathcal{X}} \mathbb{E}_\nu[1_B f_\vartheta^{\mathcal{C}}|\mathcal{C}] d\nu = \int_B f_\vartheta^{\mathcal{C}} d\nu.$$

Also gilt $\frac{dP_\vartheta}{dv} = f_\vartheta^{\mathcal{C}}$ und $f_\vartheta^{\mathcal{C}}$ ist $\mathcal{C}$-messbar.

Für den Beweis der Rückrichtung seien $B \in \mathcal{B}$ und k_B eine Version von $\mathbb{E}_\nu[1_B|\mathcal{C}]$ mit dem Maß ν aus (i). Dann gilt mit $f_\vartheta = \frac{dP_\vartheta}{dv}$ für alle $C \in \mathcal{C}$

$$\int_C 1_B dP_\vartheta = \int_{B\cap C} f_\vartheta d\nu = \int_C 1_B f_\vartheta d\nu = \int_C \mathbb{E}_\nu[1_B f_\vartheta|\mathcal{C}]d\nu.$$

Nach Voraussetzung ist f_ϑ $\mathcal{C}$-messbar. Also

$$\int_C 1_B dP_\vartheta = \int_C f_\vartheta \mathbb{E}_\nu[1_B|\mathcal{C}]d\nu = \int_C f_\vartheta k_B d\nu = \int_C k_B dP_\vartheta.$$

Also ist $\mathcal{C}$ suffizient gemäß Definition 4.3.

(iii) folgt aus (ii) und dem Faktorisierungslemma (Satz 1.18). □

Satz 4.11. (Neyman-Kriterium) *Es seien μ ein σ-endliches Maß und $\mathcal{P} \ll \mu$. Dann gilt:*

(i) Eine σ-Algebra $\mathcal{C} \subset \mathcal{B}$ ist suffizient für $\vartheta \in \Theta$ genau dann, wenn eine $\mathcal{B}$-messbare Funktion $r : \mathcal{X} \to \mathbb{R}$ und für alle $\vartheta \in \Theta$ eine $\mathcal{C}$-messbare Funktion $f_\vartheta : \mathcal{X} \to \mathbb{R}$ existieren, so dass

$$\frac{dP_\vartheta}{d\mu}(x) = r(x) f_\vartheta(x) \quad \mu\text{-f.s.}$$

(ii) Eine Statistik $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ ist suffizient für $\vartheta \in \Theta$ genau dann, wenn eine $\mathcal{B}$-messbare Funktion $r : \mathcal{X} \to \mathbb{R}$ und für alle $\vartheta \in \Theta$ eine $\mathcal{D}$-messbare Funktion $q_\vartheta : \mathcal{T} \to \mathbb{R}$ existieren, so dass

$$\frac{dP_\vartheta}{d\mu}(x) = r(x) q_\vartheta(T(x)) \quad \mu\text{-f.s.}$$

Beweis: Wir zeigen nur (i), da sich (ii) wieder mit Satz 1.18 ergibt.

Für die Hinrichtung sei ν das Wahrscheinlichkeitsmaß aus Satz 4.10. Dann gilt $P_\vartheta \ll \nu \ll \mu$, und mit einer zweifachen Anwendung von Satz 1.3 erhält man

$$\frac{dP_\vartheta}{d\mu}(x) = \frac{dP_\vartheta}{d\nu}(x)\frac{d\nu}{d\mu}(x).$$

Dabei ist der erste Faktor $\mathcal{C}$-messbar nach Satz 4.10 (ii).

Für den Beweis der Rückrichtung seien f_ϑ $\mathcal{C}$-messbar und

$$\frac{dP_\vartheta}{d\mu}(x) = r(x) f_\vartheta(x).$$

Sind c_i und f_{ϑ_i} die Konstanten aus (4.2) und die zugehörigen μ-Dichten, so setzen wir

$$\widetilde{f}_\vartheta(x) = \begin{cases} f_\vartheta(x) \Big/ \sum_{i=1}^\infty c_i f_{\vartheta_i}(x), & \text{falls mindestens ein } f_{\vartheta_i}(x) > 0, \\ 0, & \text{sonst.} \end{cases}$$

Dann ist die Funktion $\widetilde{f}_\vartheta$ $\mathcal{C}$-messbar und es gilt

$$\int_B \widetilde{f}_\vartheta dv = \int_B \widetilde{f}_\vartheta \sum_{i=1}^\infty c_i dP_{\vartheta_i} = \int_B \widetilde{f}_\vartheta \sum_{i=1}^\infty c_i r f_{\vartheta_i} d\mu = \int_B f_\vartheta r d\mu = P_\vartheta(B).$$

Damit ist $\widetilde{f}_\vartheta$ eine $\mathcal{C}$-messbare Version von $\frac{dP_\vartheta}{dv}$, und nach Satz 4.10 (ii) ist $\mathcal{C}$ suffizient für ϑ. □

Beispiel 4.12.

(i) Es sei $\mathcal{P}^1$ eine k-parametrige exponentielle Familie mit Dichten

$$\frac{dP_\vartheta}{d\mu}(x) = c(\vartheta)h(x)\exp\left(\sum_{j=1}^{k} Q_j(\vartheta)T_j(x)\right).$$

Nach Satz 4.11 ist $T = (T_1, \ldots, T_k)^T$ eine suffiziente Statistik für ϑ. Dies überträgt sich auf den Fall, wenn $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{P} = \{\otimes_{i=1}^n P \mid P \in \mathcal{P}^1\}$ ist. Dann ist $\sum_{i=1}^n T(x_i)$ eine suffiziente Statistik für ϑ.

(ii) Speziell ergibt sich im Fall $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, nach Beispiel 2.40, dass die Statistik

$$T(x) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)^T$$

suffizient für $(\mu, \sigma^2)^T$ ist.

(iii) Sind $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{U}[0, \vartheta]$ mit $\vartheta > 0$, dann erhält man für die gemeinsame Dichte

$$f(x, \vartheta) = \left(\frac{1}{\vartheta}\right)^n \prod_{j=1}^{n} 1_{[0,\vartheta]}(x_j) = \left(\frac{1}{\vartheta}\right)^n 1_{[0,\vartheta]}(\max_{j=1}^{n} x_j).$$

Nach Satz 4.11 ist $T(x) = \max_{j=1}^{n} x_j$ suffiziente Statistik für ϑ.

Bemerkung 4.13. Es seien $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ und $\widetilde{T} : (\mathcal{X}, \mathcal{B}) \to (\widetilde{\mathcal{T}}, \widetilde{\mathcal{D}})$ zwei Statistiken und ohne Einschränkung sei $\mathcal{T} = T(\mathcal{X})$ bzw. $\widetilde{\mathcal{T}} = \widetilde{T}(\mathcal{X})$. Ist nun T suffizient für ϑ und existiert eine messbare, bijektive Abbildung $b : \mathcal{T} \to \widetilde{\mathcal{T}}$ mit $\widetilde{T} = b \circ T$ und $b(\mathcal{D}) = \widetilde{\mathcal{D}}$, so ist wegen

$$\widetilde{T}^{-1}(\widetilde{D}) = T^{-1}(b^{-1}(b(D))) = T^{-1}(D), \quad \widetilde{D} \in \widetilde{\mathcal{D}} \text{ und } D = b^{-1}(\widetilde{D}) \in \mathcal{D},$$

auch $\widetilde{T}$ suffizient für ϑ.

Kurz: Bijektive Transformationen suffizienter Statistiken bleiben suffizient.

Beispiel 4.14. Im Fall $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ ist nach Beispiel 4.12 (ii)

$$T(x) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)^T$$

suffizient für $(\mu, \sigma^2)^T$. Nach Bemerkung 4.13 überträgt sich diese Eigenschaft auf die Statistik $(\overline{x}_n, \hat{s}_n^2)$. In diesem Fall ist

$$b(x, y) = \left(\frac{x}{n}, \frac{1}{n}y - \left(\frac{1}{n}x\right)^2\right)^T, \quad \mathcal{T} = \left\{(x, y) \mid x \in \mathbb{R}, y \geq \frac{1}{n}x^2\right\} \text{ und } \widetilde{\mathcal{T}} = \mathbb{R} \times \mathbb{R}^+,$$

wie man z.B. mit Hilfe der Cauchy-Schwarz-Ungleichung sieht.

Satz 4.15. (Satz von Rao-Blackwell) *Es seien $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine Familie von Verteilungen auf $(\mathcal{X}, \mathcal{B})$ und $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ suffizient für ϑ. Außerdem seien $\gamma : \Theta \to \Gamma \in \mathbb{R}^l$ und $L : \Gamma \times \Gamma \to \mathbb{R}$ eine Verlustfunktion, so dass die Abbildung $y \mapsto L(\gamma(\vartheta), y)$ für alle $\vartheta \in \Theta$ konvex ist. Ist g nun ein erwartungstreuer Schätzer für $\gamma(\vartheta)$ mit $\mathbb{E}_\vartheta[L(\gamma(\vartheta), g)] < \infty$ für alle $\vartheta \in \Theta$, dann gilt:*

(i) Es existiert ein $\sigma(T)$-messbarer erwartungstreuer Schätzer k mit gleichmäßig nicht größerer Risikofunktion, d.h.

$$R(\vartheta, k) \leq R(\vartheta, g) \quad \forall \vartheta \in \Theta, \tag{4.4}$$

nämlich jede Version von $k = \mathbb{E}[g|T]$. (Dieser Ausdruck hängt aufgrund der Suffizienz von T nicht von ϑ ab. Daher ist k ein sinnvoller Schätzer. Insbesondere ist der Verzicht auf den Index ϑ in $\mathbb{E}_\vartheta[g|T]$ hier und im Folgenden angemessen.)

(ii) Ist $y \mapsto L(\gamma(\vartheta), \cdot)$ strikt konvex für alle $\vartheta \in \Theta$, dann gilt für alle $\vartheta \in \Theta$ das Gleichheitszeichen in (4.4) genau dann, wenn $g = h \circ T$ mit $h(t) = \mathbb{E}[g|T = t]$ ist.

Beweis: Man verwendet die bedingte Jensen-Ungleichung: Ist f konvex, dann gilt

$$f(\mathbb{E}[g(X)|V]) \leq \mathbb{E}[f(g(X))|V] \quad \mathbb{P}^V\text{-f.s.} \tag{4.5}$$

Ist f sogar strikt konvex, dann gilt Gleichheit in (4.5) genau dann, wenn $g = \mathbb{E}[g|V]$.

Im Beweis von (i) bedienen wir uns des Satzes von der iterierten Erwartung, um

$$\mathbb{E}_\vartheta[k] = \mathbb{E}_\vartheta[\mathbb{E}_\vartheta[g|T]] = \mathbb{E}_\vartheta[g] = \gamma(\vartheta) \quad \forall \vartheta \in \Theta$$

zu erhalten. Also ist k erwartungstreu für ϑ. Aufgrund der bedingten Jensen-Ungleichung ergibt sich dann

$$L(\gamma(\vartheta), k) = L(\gamma(\vartheta), \mathbb{E}_\vartheta[g|T]) \leq \mathbb{E}_\vartheta[L(\gamma(\vartheta), g)|T] \quad \forall \vartheta \in \Theta,$$

und durch Integration bezüglich P_ϑ erhält man

$$R(\vartheta, k) = \mathbb{E}_\vartheta[L(\gamma(\vartheta), k)] \leq \mathbb{E}_\vartheta[L(\gamma(\vartheta), g)] = R(\vartheta, g).$$

Im Fall der strikten Konvexität gilt Gleichheit genau dann, wenn $g = \mathbb{E}[g|T]$ ist. □

Korollar 4.16. *Als wichtigste Anwendung von Satz 4.15 ergibt sich im eindimensionalen Fall mit $L(x, y) = (x-y)^2$, dass für einen erwartungstreuen Schätzer g und eine suffiziente Statistik T der Schätzer $k = \mathbb{E}[g|T]$ die Ungleichung*

$$\mathrm{Var}_\vartheta(k) \leq \mathrm{Var}_\vartheta(g) \quad \forall \vartheta \in \Theta$$

erfüllt, wobei Gleichheit genau dann gilt, wenn $g = \mathbb{E}[g|T]$ ist.

Beispiel 4.17.

(i) Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{U}[0, \vartheta]$ und $\vartheta > 0$. Nach Beispiel 4.12 ist

$$x_{(n)} = \max_{i=1}^{n} x_i$$

eine suffiziente Statistik. Offensichtlich ist

$$g(x) = \frac{2}{n} \sum_{i=1}^{n} x_i$$

ein erwartungstreuer Schätzer für ϑ, und man erhält mit Korollar 4.16, dass

$$k(x) = \mathbb{E}[g(X)|X_{(n)} = x] = \frac{n+1}{n} x$$

ein erwartungstreuer Schätzer mit nicht größerer Varianz ist.

Man erhält obige Darstellung von k, indem man

$$\mathbb{E}[X_{(i)}|X_{(n)} = x] = \frac{i}{n}x \tag{4.6}$$

nachweist. Aus (4.6) folgt dann

$$\mathbb{E}[g(X)|X_{(n)}] = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[X_{(i)}|X_{(n)}] = \frac{2}{n}\sum_{i=1}^{n}\frac{i}{n}X_{(n)} = \frac{n+1}{n}X_{(n)}.$$

Die Varianz der beiden Schätzer g und k berechnet man, indem man verwendet, dass für die i-te Ordnungsstatistik unabhängiger $\mathcal{U}[0,1]$-verteilter Zufallsvariablen

$$\mathbb{P}(X_{(i)} \leq x) = \sum_{j=i}^{n}\binom{n}{j}x^j(1-x)^{n-j} = \frac{1}{\beta(i,n-i+1)}\int_0^x t^{i-1}(1-t)^{n-i}dt$$

gilt, wobei $\beta(a,b)$ das Beta-Integral bezeichnet. Also gilt $X_{(i)} \sim B(i,n-i+1)$ mit der Beta-Verteilung aus Beispiel 3.14, und man erhält nach Skalierung mit ϑ

$$\mathrm{Var}_\vartheta(g) = \frac{\vartheta^2}{3n} \text{ sowie } \mathrm{Var}_\vartheta(k) = \frac{\vartheta^2}{n(n+2)}.$$

Offensichtlich besitzt k für $n \geq 2$ eine kleinere Varianz.

(ii) Es seien $X_1,\ldots,X_n$ i.i.d. $\sim F$ und $\mathcal{P} = \{P \sim F \mid F \text{ Verteilungsfunktion}\}$, und für $z \in \mathbb{R}$ sei $\gamma : F \mapsto F(z)$ ein zu schätzendes Funktional.

Offensichtlich ist mit $x = (x_1,\ldots,x_n)^T$ ein erwartungstreuer Schätzer für $F(z)$ durch $g(x) = 1_{\{x_1 \leq z\}}$ gegeben. Die Statistik $X_{(\cdot)} = (X_{(1)},\ldots,X_{(n)})^T$ ist nach Bemerkung 4.8 suffizient für F. Wegen

$$\mathbb{E}[1_{\{X_1 \leq z\}}|X_{(\cdot)}] = \frac{1}{n}\sum_{j=1}^{n}1_{\{X_j \leq z\}} \tag{4.7}$$

ist

$$\hat{F}_n(z) = \frac{1}{n}\sum_{j=1}^{n}1_{\{X_j \leq z\}}$$

ein erwartungstreuer Schätzer für $F(z)$. Die Funktion $z \to \hat{F}_n(z)$ heißt *empirische Verteilungsfunktion* von $X_1,\ldots,X_n$.

Zum Nachweis von (4.7) verwendet man, dass gemäß Bemerkung 4.8

$$\sigma(X_{(\cdot)}) = \mathcal{C}(S_n)$$

gilt. Da $\hat{F}_n(z)$ invariant bezüglich Permutationen der X_j ist, ist $\hat{F}_n(z)$ $\mathcal{C}(S_n)$-messbar. Außerdem gilt für alle i,j

$$\int_B 1_{\{X_i \leq z\}}dP = \int_B 1_{\{X_j \leq z\}}dP \quad \forall B \in \mathcal{C}(S_n),$$

so dass sich (4.7) nach Definition der bedingten Erwartung sofort ergibt.

Bemerkung 4.18. Gute Schätzer faktorisieren im Allgemeinen immer über einer suffizienten Statistik. Dadurch ergibt sich eine Reduktion des Datenmaterials, indem man etwa in der Situation aus Beispiel 4.12 nicht den Vektor $x = (x_1, \dots, x_n)^T$, sondern nur noch $(\overline{x}_n, \hat{s}_n^2)^T$ betrachtet. Wünschenswert ist eine maximale Reduktion durch *minimalsuffiziente* Statistiken.

Definition 4.19. Eine suffiziente Statistik $T^* : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ heißt *minimalsuffizient* für $\vartheta \in \Theta$, falls T^* über jeder suffizienten Statistik T für ϑ faktorisiert.

Beispiel 4.20. Es sei $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ mit $\Theta = \{\vartheta_0, \dots, \vartheta_k\}$ eine Familie äquivalenter Verteilungen mit μ-Dichten f_{ϑ_i}, $i = 0, \dots, k$. Dann ist die Statistik

$$T^*(x) = \left(\frac{f_{\vartheta_1}(x)}{f_{\vartheta_0}(x)}, \dots, \frac{f_{\vartheta_k}(x)}{f_{\vartheta_0}(x)} \right)^T$$

minimalsuffizient für ϑ.

Beweis: Setze $P_i = P_{\vartheta_i}$ und $f_i = f_{\vartheta_i}$ für $i = 0, \dots, k$. Wählt man dann P_0 als das dominierende Maß im Neyman-Kriterium (Satz 4.11), folgt

$$\frac{dP_i}{dP_0} = \frac{\frac{dP_i}{d\mu}}{\frac{dP_0}{d\mu}} = \frac{f_i}{f_0} = \pi_i \circ T^* \quad \text{für } i = 1, \dots, k,$$

mit der Projektion π_i auf die i-te Komponente. Mit $dP_0/dP_0 = 1$ ist damit T^* suffizient. Umgekehrt gilt für jede suffiziente Statistik T

$$f_i(x) = h(x) g_i(T(x)), \quad i = 0, \dots, k,$$

für geeignete Funktionen h und g_i. Also ergibt sich

$$\frac{f_i(x)}{f_0(x)} = \frac{g_i(T(x))}{g_0(T(x))}.$$

Damit lässt sich T^* als Funktional von T schreiben. □

Lemma 4.21. *Es sei $\mathcal{P}$ eine Familie äquivalenter Verteilungen und $\mathcal{P}_0 \subset \mathcal{P}$ eine endliche Teilfamilie. Dann ist jede Statistik, die suffizient für $\mathcal{P}$ und minimalsuffizient für $\mathcal{P}_0$ ist, auch minimalsuffizient für $\mathcal{P}$.*

Beweis: Es seien T eine derartige Statistik und S für $\mathcal{P}$ suffizient. Dann ist S auch suffizient für $\mathcal{P}_0$, und es existiert aufgrund der Minimalsuffizienz von T eine messbare Funktion h mit

$$T = h \circ S \quad \mathcal{P}_0\text{-f.s.}$$

Daraus folgt

$$T = h \circ S \quad \mathcal{P}\text{-f.s.},$$

denn $\mathcal{P}$ und $\mathcal{P}_0$ sind nach Voraussetzung äquivalent. □

Satz 4.22. *Es sei $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine k-parametrige exponentielle Familie mit Dichten*

$$\frac{dP_\vartheta}{d\mu}(x) = c(\vartheta) h(x) \exp\left(\sum_{j=1}^{k} Q_j(\vartheta) T_j(x) \right).$$

Hat $Z = \{(Q_1(\vartheta), \dots, Q_k(\vartheta))^T \mid \vartheta \in \Theta\}$ *ein nichtleeres Inneres, so ist* $T(x) = (T_1(x), \dots, T_k(x))^T$ *minimalsuffizient für* ϑ*.*

Beweis: Nach Satz 4.11 ist T suffizient. Sei nun $\mathcal{P}_0 = \{P_{\vartheta_i} \mid i = 0, \dots, k\}$, eine Teilfamilie. Wir erhalten aus Beispiel 4.20 und Bemerkung 4.13, dass

$$\widetilde{T}(x) = \left(\sum_{j=1}^{k} (Q_j(\vartheta_1) - Q_j(\vartheta_0)) T_j(x), \dots, \sum_{j=1}^{k} (Q_j(\vartheta_k) - Q_j(\vartheta_0)) T_j(x) \right)^T$$

minimalsuffizient für $\mathcal{P}_0$ ist. Insbesondere gilt

$$\widetilde{T} = \Delta Q \cdot T = (Q_j(\vartheta_i) - Q_j(\vartheta_0))_{i,j} \cdot T$$

im Sinne der Matrix-Vektor-Multiplikation. Wählt man die endliche Teilklasse so, dass man eine invertierbare Matrix ΔQ erhält, folgt

$$T = (\Delta Q)^{-1} \widetilde{T},$$

und die Aussage folgt aus Lemma 4.21. Eine solche Wahl von ΔQ ist wegen des nicht-leeren Inneren möglich. □

Bemerkung 4.23. Nach Satz 4.15 reicht es aus, über eine suffiziente Statistik T faktorisierende, erwartungstreue Schätzer zu betrachten, um Kandidaten für UMVU-Schätzer zu finden. Wir suchen im Folgenden nach Bedingungen, wann die Klasse dieser Schätzer nicht zu reichhaltig oder bestenfalls einelementig ist, um direkt optimale Schätzer zu erhalten.

Definition 4.24.

(i) Es sei $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine Klasse von Wahrscheinlichkeitsmaßen auf $(\mathcal{X}, \mathcal{B})$. Dann heißt eine σ-Algebra $\mathcal{B}_0 \subset \mathcal{B}$ *vollständig* für $\mathcal{P}$, falls für alle $\mathcal{B}_0$-messbaren Funktionen $g : \mathcal{X} \to \mathbb{R}$ mit der Eigenschaft

$$\mathbb{E}_\vartheta[g] = 0 \quad \forall \vartheta \in \Theta$$

bereits $g = 0$ P_ϑ-f.s. für alle $\vartheta \in \Theta$ gilt.

(ii) Eine Statistik $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ heißt *vollständig* für $\vartheta \in \Theta$, falls $\sigma(T)$ vollständig für $\vartheta \in \Theta$ ist, d.h.

$$\mathbb{E}_\vartheta[g \circ T] = 0 \quad \forall \vartheta \in \Theta$$

impliziert $g \circ T = 0$ P_ϑ-f.s. für alle $\vartheta \in \Theta$

Satz 4.25. *Es sei* $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ *eine* k*-parametrige exponentielle Familie mit Dichten*

$$\frac{dP_\vartheta}{d\mu}(x) = c(\vartheta) h(x) \exp\left(\sum_{j=1}^{k} Q_j(\vartheta) T_j(x) \right).$$

Hat $Z = \{(Q_1(\vartheta), \dots, Q_k(\vartheta))^T \mid \vartheta \in \Theta\}$ *ein nichtleeres Inneres, so ist* $T(x) = (T_1(x), \dots, T_k(x))^T$ *vollständig für* ϑ*.*

Beweis: Wir gehen zum natürlichen Parameterraum über und setzen $\vartheta_j = Q_j(\vartheta)$. Ohne Beschränkung der Allgemeinheit sei zudem $[-a,a]^k \subset Z \subset \Theta^*$ mit $a > 0$. Andernfalls ist eine affine Transformation durchzuführen. Wir erhalten als Dichte

$$\frac{dP_\vartheta}{d\mu}(x) = c(\vartheta) \exp\left(\sum_{j=1}^k \vartheta_j T_j(x)\right).$$

Sei nun g messbar, und für alle $\vartheta \in [-a,a]^k$ gelte

$$0 = \mathbb{E}_\vartheta[g(T)] = \int c(\vartheta) g(t) \exp\left(\sum_{j=1}^k \vartheta_j t_j\right) \mu^T(dt).$$

Mit der Zerlegung $g = g^+ - g^-$ ist diese Aussage äquivalent zu

$$\int g^+(t) \exp\left(\sum_{j=1}^k \vartheta_j t_j\right) \mu^T(dt) = \int g^-(t) \exp\left(\sum_{j=1}^k \vartheta_j t_j\right) \mu^T(dt) \quad \forall \vartheta \in [-a,a]^k. \tag{4.8}$$

Insbesondere erhält man für $\vartheta = 0$

$$A = \int g^+(t)\mu^T(dt) = \int g^-(t)\mu^T(dt).$$

Zu zeigen ist $g^+ = g^-$ μ^T-f.s. Im Fall $A = 0$ ergibt sich direkt die Behauptung, da g^+ und g^- nicht-negativ sind (und also jeweils verschwinden). Andernfalls definiert man Wahrscheinlichkeitsmaße $P^\pm$ via

$$P^\pm(B) = \frac{1}{A} \int_B g^\pm(t)\mu^T(dt).$$

Dann ist (4.8) äquivalent zu

$$\int \exp\left(\sum_{j=1}^k \vartheta_j t_j\right) P^+(dt) = \int \exp\left(\sum_{j=1}^k \vartheta_j t_j\right) P^-(dt).$$

Wir setzen nun $\vartheta_j = \xi_j + i\eta_j$ mit $|\xi_j| \le a$, $j = 1,\dots,k$. Offenbar existieren die Integrale auch für diese Parameter. Dann betrachten wir die Funktionen

$$f_l^\pm : \begin{cases} \{z \in \mathbb{C} \mid |\mathrm{Re}(z)| \le a\} & \to \mathbb{C} \\ z & \mapsto \int \exp\left(\sum_{j\neq l} \vartheta_j t_j + z t_l\right) P^\pm(dt)\ . \end{cases}$$

Diese Funktionen sind analytisch und damit holomorph, wie man mittels einer komplexen Erweiterung von Satz 2.38 zeigt. Da auf $\mathbb{C} \cap [-a,a]$ die Identität $f_l^+(z) = f_l^-(z)$ gilt, folgt $f_l^+ = f_l^-$ auf $\{z \in \mathbb{C} \mid |\mathrm{Re}(z)| \le a\}$ mit Hilfe des Identitätssatzes für holomorphe Funktionen auf Gebieten. Da der Index l beliebig ist, ergibt sich insbesondere als Konsequenz für alle $\eta = (\eta_1,\dots,\eta_k)^T$

$$\int \exp\left(i\sum_{j=1}^k t_j\eta_j\right) P^+(dt) = \int \exp\left(i\sum_{j=1}^k t_j\eta_j\right) P^-(dt).$$

Aus dem Eindeutigkeitssatz für charakteristische Funktionen folgen dann unmittelbar $P^+ = P^-$ und damit $g^+ = g^-$ μ^T-f.s. □

Beispiel 4.26.

(i) Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. Da die gemeinsame Dichte durch

$$\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\frac{\sum_{j=1}^n {x_j}^2}{2\sigma^2}\Big)$$

gegeben ist, liefert das Neyman-Kriterium aus Satz 4.11, dass die Statistiken

$$\begin{aligned}
T_1(x) &= (x_1, \ldots, x_n)^T, \\
T_2(x) &= (x_1^2, \ldots, x_n^2)^T, \\
T_3(x) &= (x_1^2 + \ldots + x_m^2, x_{m+1}^2, \ldots, x_n^2)^T, \quad 1 < m < n, \\
T_4(x) &= \sum_{j=1}^n x_j^2
\end{aligned}$$

jeweils suffizient mit einer wachsenden Datenreduktion sind.

(ii) Man sieht leicht: $T_j = h_{ij}(T_i)$, $i < j$, für geeignete Funktionen h_{ij}. Daher sind T_1, T_2 und T_3 nicht minimalsuffizient. Die Statistik T_4 ist dagegen minimalsuffizient gemäß Satz 4.22.

(iii) T_1, T_2 und T_3 sind nicht vollständig, wie man anhand der Wahl von $g(x) = x_1$ für T_1, $h(x) = x_1 - x_2$ für T_2 und $k_m(x) = x_1 - mx_2$ für T_3 sieht. Hingegen ist T_4 gemäß Satz 4.25 vollständig.

Beispiel 4.27.

(i) In Beispiel 4.17 (i) ist

$$x_{(n)} = \max_{i=1}^n x_i$$

vollständig für $\vartheta \in \mathbb{R}^+$, denn wegen

$$P_\vartheta(X_{(n)} \leq x) = P_\vartheta(X_j \leq x \text{ für alle } j = 1, \ldots, n) = \Big(\frac{x}{\vartheta}\Big)^n$$

ergibt sich

$$f_\vartheta^{(n)}(x) = \frac{n}{\vartheta^n} x^{n-1} 1_{[0,\vartheta]}(x)$$

als Dichte von $X_{(n)}$ bzgl. dem Lebesgue-Maß λ. Aus

$$\mathbb{E}_\vartheta[g(X_{(n)})] = \int_0^\vartheta g(x) f_\vartheta^{(n)}(x) dx = \frac{n}{\vartheta^{n-1}} \int_0^\vartheta g(x) x^{n-1} dx = 0 \quad \forall \vartheta > 0$$

erhält man sofort $g \equiv 0$ λ-f.s., also $g \equiv 0$ P_ϑ-f.s. für alle $\vartheta > 0$.

(ii) Es seien $X_1, \ldots, X_n$ i.i.d. $\sim F \ll \lambda$ und

$$\mathcal{P} = \{P \sim F^n \mid F \text{ Verteilungsfunktion }\}.$$

Dann ist die Ordnungsstatistik $X_{(\cdot)}$ vollständig für F. Der Beweis verwendet die Vollständigkeit von $X_{(\cdot)}$ bzgl. einer geeigneten Unterklasse von exponentiellen Familien (vgl. Example 4.3.4 in Lehmann and Romano, 2005).

Satz 4.28. (Satz von Lehmann-Scheffé) *Es seien $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine Klasse von Wahrscheinlichkeitsmaßen auf $(\mathcal{X}, \mathcal{B})$, $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ eine suffiziente und vollständige Statistik für $\vartheta \in \Theta$, $\gamma : \Theta \to \Gamma \subset \mathbb{R}^l$ ein erwartungstreu schätzbares Funktional und $L : \Gamma \times \Gamma \to \mathbb{R}$ eine Verlustfunktion, so dass für jedes feste $\vartheta \in \Theta$ die Funktion $L(\gamma(\vartheta), \cdot)$ konvex ist.*

Dann gibt es (fast sicher) genau einen erwartungstreuen Schätzer der Form $h \circ T$ für $\gamma(\vartheta)$. Dieser besitzt die gleichmäßig kleinste Risikofunktion unter allen erwartungstreuen Schätzern.

Beweis: Die Existenz des Schätzers ergibt sich direkt aus dem Satz von Rao-Blackwell (Satz 4.15). Zum Nachweis der Eindeutigkeit sei $g \circ T$ ein weiterer Schätzer für $\gamma(\vartheta)$. Dann gilt

$$0 = \mathbb{E}_\vartheta[h(T) - g(T)] \quad \forall \vartheta \in \Theta,$$

und aus der Vollständigkeit erhält man

$$g = h \ P_\vartheta\text{-f.s.} \quad \forall \vartheta \in \Theta.$$

Ist zuletzt k ein erwartungstreuer Schätzer für $\gamma(\vartheta)$, dann ist auch $k \circ T = \mathbb{E}[k|T]$ erwartungstreu für $\gamma(\vartheta)$ und man erhält

$$R(\vartheta, k) \geq R(\vartheta, k \circ T) = R(\vartheta, h \circ T) \quad \forall \vartheta \in \Theta,$$

wieder mit Hilfe des Satzes von Rao-Blackwell. □

Korollar 4.29. *Im Fall $l = 1$ mit der Verlustfunktion $L(x, y) = (x - y)^2$ gilt: Ist T suffizient und vollständig, dann ist ein von T abhängender erwartungstreuer Schätzer fast sicher eindeutig bestimmt und der UMVU-Schätzer.*

Beispiel 4.30.

(i) Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{U}[0, \vartheta]$. Dann ist $\frac{n+1}{n} x_{(n)}$ nach den Beispielen 4.17 und 4.27 der UMVU-Schätzer für ϑ.

(ii) Es seien $X_1, \ldots, X_n$ i.i.d. $\sim F \ll \lambda$. Dann ist die empirische Verteilungsfunktion $\hat{F}_n(z)$ ebenfalls nach den Beispielen 4.17 und 4.27 der UMVU-Schätzer für $F(z)$.

(iii) Es seien $X_1, \ldots, X_n$ i.i.d. $\sim B(1, \vartheta)$, $\vartheta \in [0, 1]$. Dann ist die Statistik

$$T(x) = \sum_{j=1}^{n} x_j$$

nach Beispiel 4.2 suffizient für ϑ. Zum Nachweis der Vollständigkeit verwende man entweder Beispiel 2.34 und Satz 4.25, oder man beachte direkt, dass die Identität

$$0 = \mathbb{E}_\vartheta[h(T)] = \sum_{j=0}^{n} h(j) \binom{n}{j} \vartheta^j (1 - \vartheta)^{n-j} \quad \forall \vartheta \in [0, 1]$$

mit einem Argument über Polynome die Aussage

$$h(j) = 0 \quad \text{für alle } j = 0, \ldots, n$$

impliziert. Damit ist $k(x) = \overline{x}_n$ der UMVU-Schätzer für $\vartheta \in \Theta$, wie sich auch aus der Cramér-Rao-Ungleichung ergibt (vgl. Beispiel 2.26).

Bemerkung 4.31. Unter denselben Annahmen wie in Bemerkung 4.13 gilt für die Statistiken

$$T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D}) \text{ und } \widetilde{T} = b \circ T : (\mathcal{X}, \mathcal{B}) \to (\widetilde{\mathcal{T}}, \widetilde{\mathcal{D}})$$

die Implikation: Ist T vollständig und suffizient, dann ist auch $\widetilde{T}$ vollständig und suffizient.

Dadurch ergibt sich etwa in Beispiel 4.30 (iii), dass $g(x) = \frac{n}{n-1}\overline{x}_n(1-\overline{x}_n)$ der UMVU-Schätzer für $\gamma(\vartheta) = \vartheta(1-\vartheta)$ ist, denn

$$\mathbb{E}_\vartheta[g(X)] = \frac{1}{n(n-1)}\mathbb{E}_\vartheta\left[(n-1)\sum_{i=1}^{n} X_i - \sum_{i\neq j} X_i X_j\right] = \vartheta(1-\vartheta).$$

Hier haben wir $X_i = X_i^2$ sowie die Verkleinerung des Parameterraumes auf $\Theta = [0, 1/2]$ bzw. $\Gamma = [0, 1/4]$ verwendet, um Bijektivität zu erhalten.

Kapitel 5

Asymptotische Eigenschaften von Schätzern

In diesem Kapitel sei $X^{(n)} = (X_1, \dots, X_n)^T$ ein n-dimensionaler Zufallsvektor mit Werten im Stichprobenraum $\mathcal{X}_n = \mathcal{X}^n$ und einer Verteilung P aus der Klasse $\mathcal{P}^n = \{P_\vartheta^n \mid \vartheta \in \Theta\}$. Für jedes n sei

$$T_n : \begin{cases} \mathcal{X}_n \to \Gamma \\ x^{(n)} \mapsto T_n(x^{(n)}) \end{cases}$$

ein Schätzer für $\gamma(\vartheta)$. Eine Minimalforderung an die Statistik T_n lautet, dass bei großem Stichprobenumfang T_n nahe beim wahren Wert $\gamma(\vartheta)$ liegt. In welchem Sinne dies gelten soll, werden wir im Verlauf des Abschnitts präzisieren.

Definition 5.1. Es seien $T_n : \mathcal{X}_n \to \Gamma$ ein Schätzer mit Werten in einem metrischen Raum. Zudem seien alle Experimente auf einem gemeinsamen Wahrscheinlichkeitsraum definiert, und es gelte $P_\vartheta^n \ll Q_\vartheta$ für alle n.

(i) T_n heißt *(schwach) konsistent* für $\gamma(\vartheta)$ genau dann, wenn T_n für alle $\vartheta \in \Theta$ in Q_ϑ-Wahrscheinlichkeit gegen $\gamma(\vartheta)$ konvergiert, d.h.

$$T_n \xrightarrow{Q_\vartheta} \gamma(\vartheta) \quad \forall \vartheta \in \Theta.$$

(ii) T_n heißt *stark konsistent* für $\gamma(\vartheta)$ genau dann, wenn T_n für alle $\vartheta \in \Theta$ Q_ϑ-fast sicher gegen $\gamma(\vartheta)$ konvergiert, d.h.

$$T_n \longrightarrow \gamma(\vartheta) \; Q_\vartheta\text{-f.s.} \quad \forall \vartheta \in \Theta.$$

Beispiel 5.2. Wir betrachten die Momentenmethode aus Beispiel 2.41. Es seien also $X_1, \dots, X_n$ i.i.d. $\sim P_\vartheta$ reellwertige Zufallsvariablen, $\vartheta \in \Theta \subset \mathbb{R}^k$ und $\gamma : \Theta \to \Gamma \subset \mathbb{R}^l$. Ferner sei $m_j = \mathbb{E}_\vartheta[X_1^j] = \int x^j P_\vartheta(dx)$, $j = 1, \dots, k$, und für das zu schätzende Funktional gelte

$$\gamma(\vartheta) = f(m_1, \dots, m_k).$$

Wählt man

$$\hat{\gamma}(x) = f(\hat{m}_1, \dots, \hat{m}_k),$$

wobei $\hat{m}_j := \frac{1}{n} \sum_{k=1}^n x_k^j$ den Mittelwertschätzer für das j-te Moment bezeichnet, und ist f stetig, so gilt wegen des starken Gesetzes der großen Zahlen

$$\hat{\gamma}(X) \longrightarrow \gamma(\vartheta) \quad Q_\vartheta\text{-f.s.}$$

mit $Q_\vartheta = \otimes_{i=1}^{\mathbb{N}} P_\vartheta$.

Satz 5.3. (Satz von Cramér-Wold) *Es sei $(X_n)_n$ eine Folge von d-dimensionalen Zufallsvektoren. Dann gilt die schwache Konvergenz $X_n \xrightarrow{\mathcal{L}} X$ genau dann, wenn*

$$y^T X_n \xrightarrow{\mathcal{L}} y^T Y \quad \forall y \in \mathbb{R}^d \tag{5.1}$$

gilt.

Beweis: Nach dem Stetigkeitssatz von Lévy gilt:

$$X_n \xrightarrow{\mathcal{L}} X \quad \Longleftrightarrow \quad \mathbb{E}[\exp(iu^T X_n)] \longrightarrow \mathbb{E}[\exp(iu^T X)] \ \forall u \in \mathbb{R}^d.$$

Letztere Aussage ist mit $u = ty$, $t \in \mathbb{R}$, äquivalent zu (5.1). □

Satz 5.4. (Zentraler Grenzwertsatz) *Es seien $X_1, \ldots, X_n$ i.i.d. d-dimensionale Zufallsvariablen mit $E[X_j] = \mu \in \mathbb{R}^d$ und $\mathrm{Cov}(X_j) = \Sigma > 0$. Dann gilt für den Vektor der Mittelwerte*

$$Z^{(n)} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

die Konvergenz in Verteilung

$$\sqrt{n}(Z^{(n)} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

Beweis: Dieser Beweis wird als Übungsaufgabe geführt. □

Definition 5.5. Es sei $T_n : \mathcal{X} \to \mathbb{R}^l$ eine Folge von Schätzern.

(i) Setzt man $\mu_n(\vartheta) = \mathbb{E}_\vartheta[T_n]$, so heißt T_n *asymptotisch erwartungstreu* für $\gamma(\vartheta)$, falls

$$\mu_n(\vartheta) - \gamma(\vartheta) \longrightarrow 0$$

für alle $\vartheta \in \Theta$ gilt.

(ii) T_n heißt *asymptotisch normalverteilt*, falls

$$(\mu_n(\vartheta))_{n\in\mathbb{N}} \subset \mathbb{R}^l \text{ und } (\Sigma_n(\vartheta))_{n\in\mathbb{N}} \subset PD(l)$$

mit $||\Sigma_n(\vartheta)|| \longrightarrow 0$ für alle $\vartheta \in \Theta$ existieren, so dass die Konvergenz in Verteilung

$$\Sigma_n^{-\frac{1}{2}}(\vartheta)(T_n - \mu_n(\vartheta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{I}_l)$$

für alle $\vartheta \in \Theta$ gilt.

Lemma 5.6. (Lemma von Slutsky) *Es seien $(Z_n)_n$ und $(Y_n)_n$ Folgen von d-dimensionalen Zufallsvektoren. Gilt*

$$Z_n \xrightarrow{\mathcal{L}} Z \quad \textit{und} \quad Y_n \xrightarrow{\mathbb{P}} y_0$$

für eine d-dimensionale Zufallsvariable Z und einen konstanten Vektor $y_0 \in \mathbb{R}^d$, so folgt:

(i) $Z_n + Y_n \xrightarrow{\mathcal{L}} Z + y_0$.

(ii) $Y_n^T Z_n \xrightarrow{\mathcal{L}} y_0^T Z$.

Beweis: vgl. Theorem 11.2.11 in Lehmann and Romano (2005). □

Beispiel 5.7. Es seien $X_1, \ldots, X_n$ i.i.d. $\sim B(1,p)$ mit $p \in (0,1)$. Dann ist $T_n = \overline{X}_n$ erwartungstreu für p und es gilt der zentrale Grenzwertsatz

$$\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

wegen $\sigma_n^2 = \mathrm{Var}(\overline{X}_n) = \frac{p(1-p)}{n}$. Ersetzt man im Nenner p durch $\overline{X}_n$, bleibt die asymptotische Normalität wegen Lemma 5.6 erhalten. Daher ergibt sich die Approximation

$$P_p(|\overline{X}_n - p| < \varepsilon) \approx 2\Phi\Big(\varepsilon\sqrt{\frac{n}{\overline{X}_n(1-\overline{X}_n)}}\Big) - 1 \quad \forall p \in (0,1),$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet.

Zum Beispiel ergeben sich für $n = 100$ und $\overline{X}_n = 0.85$ die Wahrscheinlichkeiten

$$P_p(|\overline{X}_n - p| < \varepsilon) \approx \begin{cases} 83.84\%, & \varepsilon = 0.05 \\ 99.46\%, & \varepsilon = 0.1. \end{cases}$$

Satz 5.8. (Delta-Methode) *Es sei $(X_n)_{n\in\mathbb{N}}$ sei eine Folge von k-dimensionalen Zufallsvariablen. Ferner seien $\mu \in \mathbb{R}^k$ und $\Sigma \in NND(k)$ sowie eine Folge $(c_n)_{n\in\mathbb{N}}$ mit $c_n \to 0$ gegeben, so dass die schwache Konvergenz*

$$\frac{X_n - \mu}{c_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,\Sigma)$$

gilt. Ist nun $g : \mathbb{R}^k \to \mathbb{R}^m$ in μ stetig differenzierbar mit Jacobi-Matrix $D \in \mathbb{R}^{m\times k}$, so gilt

$$Z_n = \frac{g(X_n) - g(\mu)}{c_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, D\Sigma D^T).$$

Beweis: Zunächst ergibt sich

$$X_n - \mu = c_n \frac{X_n - \mu}{c_n} \xrightarrow{\mathcal{L}} 0$$

aus dem Lemma von Slutsky (Lemma 5.6), und wenn man nutzt, dass Konvergenz in Verteilung und Konvergenz in Wahrscheinlichkeit für einen deterministischen Grenzwert äquivalent sind, ergibt sich insbesondere $X_n \to \mu$ in Wahrscheinlichkeit.

Aus dem Mittelwertsatz der Differentialrechnung folgt nun

$$Z_n = \frac{g(X_n) - g(\mu)}{c_n} = g'(\mu)\frac{X_n - \mu}{c_n} + \{g'(\xi_n) - g'(\mu)\}\frac{X_n - \mu}{c_n},$$

mit einer Zwischenstelle ξ_n, für die

$$||\xi_n - \mu|| \le ||X_n - \mu||$$

gilt. Daher folgt $\xi_n \xrightarrow{\mathbb{P}} \mu$, und da g stetig differenzierbar ist, ergibt sich auch $g'(\xi_n) \xrightarrow{\mathbb{P}} g'(\mu)$. Die Aussage folgt dann wieder aus Lemma 5.6 und mit Eigenschaften der Normalverteilung. □

Bemerkung 5.9. In der Situation von Beispiel 5.2 seien $\mathbb{E}_\vartheta[X_1^{2k}] < \infty$ für alle $\vartheta \in \Theta$ und $\gamma : \mathbb{R}^k \to \mathbb{R}^l$ stetig differenzierbar in $\mu = (m_1, \ldots, m_k)^T$ mit Jacobi-Matrix D. Mit

$$\Sigma = (m_{i+j} - m_i m_j)_{i,j=1}^k$$

folgt dann

$$\sqrt{n}(\hat{\gamma}(X) - \gamma(\vartheta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, D\Sigma D^T)$$

für alle $\vartheta \in \Theta$.

Beispiel 5.10.

(i) Es seien $X_1, \ldots, X_n$ i.i.d. mit $\mathbb{E}_\vartheta[X_i] = \mu$ und $\mathrm{Var}_\vartheta(X_i) = \sigma^2$. Aus dem zentralen Grenzwertsatz folgt

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Als Schätzer für μ^2 ergibt sich nach der Momentenmethode der (nur asymptotisch erwartungstreue) Schätzer $T_n = \overline{X}_n^2$. Man erhält

$$\sqrt{n}(T_n - \mu^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\mu^2\sigma^2).$$

(ii) Es seien

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \varrho\sigma\tau \\ \varrho\sigma\tau & \tau^2 \end{pmatrix} \right), \quad i = 1, \ldots, n,$$

i.i.d. mit Parameter $\vartheta = (\mu_1, \mu_2, \sigma^2, \tau^2, \varrho)^T$. Der Schätzer

$$\begin{aligned} \hat{\varrho}_n &= \frac{SQ_{xy}}{\sqrt{SQ_{xx}}\sqrt{SQ_{yy}}} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y}_n)^2}} \\ &= \frac{\sum_{i=1}^n x_i y_i - n\overline{x}_n\overline{y}_n}{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}_n^2}\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}_n^2}} \end{aligned}$$

heißt *Pearson-Korrelationskoeffizient*. Ohne Einschränkung lässt sich $\mu_1 = \mu_2 = 0$ und $\sigma = \tau = 1$ annehmen, da $\hat{\varrho}_n$ invariant unter affin-linearen Transformationen ist.

Es gilt: Der Vektor $S_n = (SQ_{xx}, SQ_{yy}, SQ_{xy})^T$ erfüllt mit $m = (1, 1, \varrho)^T$ und

$$V = 2\begin{pmatrix} 1 & \varrho^2 & \varrho \\ \varrho^2 & 1 & \varrho \\ \varrho & \varrho & (1+\varrho^2)/2 \end{pmatrix}$$

die Konvergenz in Verteilung

$$\sqrt{n}(S_n - m) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V). \tag{5.2}$$

Um (5.2) nachzuweisen, verwendet man zunächst Lemma 5.6 und den Zentralen Grenzwertsatz, um die Konvergenzen in Wahrscheinlichkeit

$$\sqrt{n}(\overline{X}_n\overline{Y}_n) \xrightarrow{\mathbb{P}} 0, \quad \sqrt{n}(\overline{X}_n)^2 \xrightarrow{\mathbb{P}} 0, \quad \sqrt{n}(\overline{Y}_n)^2 \xrightarrow{\mathbb{P}} 0$$

zu erhalten. Man folgert dann schnell, dass

$$\sqrt{n}(S_n - m) - \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - m\right) \xrightarrow{\mathbb{P}} 0$$

mit $Z_i = (X_i^2, Y_i^2, X_iY_i)^T$ gilt. Beachtet man nun noch, dass eine leichte Rechnung

$$\mathrm{Cov}(Z_i) = \mathbb{E}[Z_i Z_i^T] - \mathbb{E}[Z_i]\mathbb{E}[Z_i]^T = V$$

liefert, so ergibt sich (5.2) aus dem Zentralen Grenzwertsatz für die Vektoren Z_i. Die Funktion

$$g(x_1, x_2, x_3) = \frac{x_3}{\sqrt{x_1 x_2}}$$

liefert zuletzt $g(S_n) = \hat{\varrho}_n$, und für die Jacobi-Matrix an der Stelle m ergibt sich $D = (-\frac{1}{2}\varrho, -\frac{1}{2}\varrho, 1)$. Also erhält man

$$\sqrt{n}(\hat{\varrho}_n - \varrho) \xrightarrow{\mathcal{L}} \mathcal{N}(0, DVD^T) = \mathcal{N}(0, (1-\varrho^2)^2).$$

Definition 5.11. Es sei $T_n : \mathcal{X} \to \mathbb{R}^l$ eine Folge von Schätzern, die asymptotisch erwartungstreu und asymptotisch normalverteilt ist.

Unter denselben Regularitätsbedingungen wie in Satz 2.23 heißt T_n *asymptotisch effizient*, falls

$$\lim_{n\to\infty} \Sigma_n(\vartheta) I(f_n(\cdot, \vartheta)) = \mathbb{I}_l \quad \forall \vartheta \in \Theta$$

gilt, wobei $I(f_n(\cdot, \vartheta))$ die Fisher-Information von $\mathcal{P}^n$ bzgl. ϑ bezeichnet.

Bemerkung 5.12. Die Aussage aus der obigen Definition lässt sich wie folgt interpretieren: Ist T_n erwartungstreu, dann gilt wegen der Cramer-Rao-Ungleichung

$$\mathrm{Cov}_\vartheta(T_n) \geq I^{-1}(f_n(\cdot, \vartheta))$$

im Sinne der Löwner-Ordnung.

Da aber

$$\Sigma_n^{-\frac{1}{2}}(\vartheta)(T_n - \mu_n(\vartheta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{I}_l)$$

gilt, erhält man unter den üblichen Regularitätsannahmen näherungsweise für die Kovarianzmatrix von T_n

$$\mathrm{Cov}_\vartheta(T_n) \approx \Sigma_n(\vartheta) \approx I^{-1}(f_n(\cdot, \vartheta)),$$

d.h. für großen Stichprobenumfang ist T_n näherungsweise erwartungstreu und erreicht im Wesentlichen die untere Schranke der Cramér-Rao-Ungleichung.

Beispiel 5.13. Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Dann gilt gemäß Beispiel 2.29 für den erwartungstreuen Schätzer

$$g_n(x) = \begin{pmatrix} \overline{x}_n \\ \frac{1}{n-1}\sum(x_i - \overline{x}_n)^2 \end{pmatrix}$$

die Identität

$$\mathrm{Cov}_\vartheta(g_n) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix} = \Sigma_n(\mu, \sigma^2).$$

Die Inverse der Fisher-Information ist durch

$$I(f_n(\cdot,\vartheta))^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$$

gegeben, und also ist g_n asymptotisch effizient.

Bemerkung 5.14. Wir interessieren uns im Folgenden für asymptotische Eigenschaften des Maximum-Likelihood-Schätzers. Dazu seien $X_1, \dots, X_n$ i.i.d. $\sim P_\vartheta$, $\vartheta \in \Theta$, mit μ-Dichten $f(\cdot,\vartheta)$. Wir bezeichnen mit

$$\ell(\cdot,\vartheta) = \log f(\cdot,\vartheta)$$

die *Log-Likelihood-Funktion* und setzen analog zu Bemerkung 2.41 (ii)

$$\hat{\theta}_n(x) = \arg\sup_{\vartheta\in\Theta} f(x,\vartheta) = \arg\sup_{\vartheta\in\Theta} \ell(x,\vartheta) = \arg\sup_{\vartheta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(x_i,\vartheta)$$

als Maximum-Likelihood-Schätzer für ϑ bei Beobachtung von $x = (x_1,\dots,x_n)^T$ (im Fall seiner Existenz).

Definition 5.15. Für zwei Wahrscheinlichkeitsmaße $\mathbb{P}$ und $\mathbb{Q}$ auf einen Messraum $(\mathcal{X},\mathcal{B})$ heißt

$$KL(\mathbb{P}|\mathbb{Q}) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)(x)\mathbb{P}(dx), & \text{falls } \mathbb{P} \ll \mathbb{Q}, \\ \infty, & \text{sonst,} \end{cases}$$

der *Kullback-Leibler-Abstand* von $\mathbb{P}$ und $\mathbb{Q}$.

Lemma 5.16. *Für den Kullback-Leibler-Abstand gilt*

$$KL(\mathbb{P}|\mathbb{Q}) \geq 0.$$

Außerdem ist $KL(\mathbb{P}|\mathbb{Q}) = 0$ genau dann, wenn $\mathbb{P} = \mathbb{Q}$.

Beweis: Aus der Jensen-Ungleichung folgt

$$\begin{aligned} \int_{\mathcal{X}} \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)(x)\mathbb{P}(dx) &= \int_{\mathcal{X}} -\log\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)(x)\mathbb{P}(dx) \\ &\geq -\log\int_{\mathcal{X}}\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)(x)\mathbb{P}(dx) = -\log\int_{\mathcal{X}}\mathbb{Q}(dx) = 0. \end{aligned}$$

Insbesondere gilt Gleichheit genau dann, wenn $\frac{d\mathbb{Q}}{d\mathbb{P}}(x) = 1$ $\mathbb{P}$-fast sicher. □

Satz 5.17. *Es seien $X_1,\dots,X_n$ i.i.d. $\sim P_\vartheta$, $\vartheta \in \Theta$, mit Log-Likelihoodfunktion $\ell(\cdot,\vartheta)$. Zudem gelten die folgenden Bedingungen:*

(i) $\Theta \subset \mathbb{R}^k$ *ist kompakt.*

(ii) Die Abbildungen

$$\eta \mapsto L(\eta,\vartheta) = \mathbb{E}_\vartheta[\ell(X_i,\eta)] \quad \textit{und} \quad \eta \mapsto L_n(\eta) = \frac{1}{n}\sum_{i=1}^n \ell(X_i,\eta)$$

sind stetig bzw. $\otimes_{i=1}^n P_\vartheta$-f.s. stetig.

(iii) Mit $Q_\vartheta = \otimes_{i=1}^{\mathbb{N}} P_\vartheta$ *gilt*

$$\sup_{\eta\in\Theta} |L_n(\eta) - L(\eta,\vartheta)| \to 0$$

in Q_ϑ*-Wahrscheinlichkeit.*

Dann ist der Maximum-Likelihood-Schätzer $\hat\theta_n(x)$ *konsistent für* ϑ*.*

Beweis: Für jedes η gilt

$$L_n(\eta) \to L(\eta,\vartheta) = \int \ell(x,\eta) f(x,\vartheta)\mu(dx) = \int \ell(x,\vartheta) f(x,\vartheta)\mu(dx) - KL(\vartheta|\eta)$$

μ-fast sicher, und die rechte Seite wird gemäß Lemma 5.16 maximal für $\eta = \vartheta$. Insbesondere besitzt $\eta \mapsto L(\eta,\vartheta)$ ein eindeutiges Maximum in ϑ.

Wir zeigen nun: Die Funktion

$$\arg\max : C(\Theta,\mathbb{R}) \to \Theta,$$

die jeder stetigen Funktion f den Wert $m_f = \arg\max_{\eta\in\Theta} f(\eta)$ zuordnet, ist stetig an den Stellen f, an denen m_f eindeutig ist. Dann folgt die Aussage aus

$$\vartheta = \arg\max L(\eta,\vartheta) \quad \text{und} \quad \hat\theta_n(x) = \arg\max L_n(\eta)$$

und Bedingung (iii).

Wir beweisen die Hilfsaussage. Es sei f_n eine Folge mit $||f_n - f||_\infty \to 0$. Zu zeigen ist $m_{f_n} \to m_f$. Es folgt zunächst wegen

$$\begin{aligned} f_n(m_{f_n}) - f(m_f) &\le f(m_{f_n}) - f(m_f) + ||f_n - f||_\infty \le ||f_n - f||_\infty \to 0, \\ f_n(m_{f_n}) - f(m_f) &\ge f_n(m_{f_n}) - f_n(m_f) - ||f_n - f||_\infty \ge -||f_n - f||_\infty \to 0, \end{aligned}$$

die Konvergenz $f_n(m_{f_n}) \to f(m_f)$. Da Θ kompakt ist, besitzt (m_{f_n}) einen Häufungspunkt m. Zunächst entlang einer Teilfolge gilt $f(m) = \lim_{n\to\infty} f_n(m_{f_n})$, wobei der letztgenannte Limes nach Vorbereitung allgemein existiert und gleich $f(m_f)$ ist. Insbesondere ergibt sich $m = m_f$ aufgrund der Eindeutigkeit des Maximums. Damit ist jeder Häufungspunkt der Folge (m_{f_n}) gleich m_f. □

Bemerkung 5.18.

(i) Da $\eta \mapsto L_n(\eta)$ als (fast sicher) stetig angenommen wurde, existiert der Maximum-Likelihood-Schätzer auf einem Kompaktum stets (fast sicher). Er kann außerdem messbar gewählt werden (vgl. Hilfssatz 6.7 in Witting and Müller-Funk (1995)).

(ii) Die Schwierigkeit in der Anwendung von Satz 5.17 besteht im Nachweis von Bedingung (iii). Dazu verwendet man typischerweise folgendes Resultat: Es seien $\Gamma \subset \mathbb{R}^k$ kompakt und $X_n(\gamma)$ Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega,\mathcal{A},\mathbb{P})$ mit $X_n(\gamma) \xrightarrow{\mathbb{P}} X(\gamma)$ für jedes $\gamma\in\Gamma$, wobei $\gamma \mapsto X_n(\gamma)$ und $\gamma \mapsto X(\gamma)$ jeweils stetig sind. Dann gilt

$$\sup_{\gamma\in\Gamma} |X_n(\gamma) - X(\gamma)| \xrightarrow{\mathbb{P}} 0$$

genau dann, wenn für alle $\varepsilon > 0$

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{|\gamma_1-\gamma_2|<\delta} |X_n(\gamma_1) - X_n(\gamma_2)| \ge \varepsilon \right) = 0$$

gilt.

Satz 5.19. (Asymptotische Effizienz des Maximum-Likelihood-Schätzers) *Es seien $X_1, \dots, X_n$ i.i.d. $\sim P_\vartheta$, $\vartheta \in \Theta$, mit Log-Likelihoodfunktion $\ell(\cdot, \vartheta)$. Zusätzlich gelten die folgenden Bedingungen:*

(i) $\Theta \subset \mathbb{R}^k$ *ist kompakt, und* ϑ *liegt im Innern von* Θ.

(ii) $\eta \mapsto \ell(x, \eta)$ *ist stetig auf* Θ *und zweimal stetig differenzierbar in einer Umgebung* U *von* ϑ *für fast alle* $x \in \mathcal{X}$.

(iii) Es existieren $H_0, H_2 \in L^1(P_\vartheta)$ *und* $H_1 \in L^2(P_\vartheta)$ *mit*

$$\sup_{\eta \in \Theta} ||\ell(x, \eta)|| \leq H_0(x), \quad \sup_{\eta \in U} ||\dot{\ell}(x, \eta)|| \leq H_1(x), \quad \sup_{\eta \in U} ||\ddot{\ell}(x, \eta)|| \leq H_2(x)$$

für alle $x \in \mathcal{X}$.

(iv) Die Fisher-Informationsmatrix (zu einer Beobachtung)

$$I(f(\cdot, \vartheta)) = \mathbb{E}_\vartheta[\dot{\ell}(X, \vartheta)\dot{\ell}(X, \vartheta)^T]$$

ist positiv definit.

Dann ergibt sich die asymptotische Normalität

$$\sqrt{n}\left(\hat{\theta}_n - \vartheta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(f(\cdot, \vartheta))^{-1}).$$

Ferner gilt die Formel

$$I(f(\cdot, \vartheta)) = -\mathbb{E}_\vartheta[\ddot{\ell}(X, \vartheta)].$$

Beweis: Wir weisen zunächst die Bedingungen aus Satz 5.17 nach, um die Konsistenz des Maximum-Likelihood-Schätzers verwenden zu können. Dabei ist Kompaktheit von Θ klar wegen (i), und

$$\eta \mapsto L_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, \eta)$$

ist aufgrund von (ii) fast sicher stetig. Außerdem gilt

$$|L(\eta_1, \vartheta) - L(\eta_2, \vartheta)| \leq \int_{\mathcal{X}} |\ell(x, \eta_1) - \ell(x, \eta_2)| f(x, \vartheta) \mu(dx) \to 0$$

für $\eta_1 \to \eta_2$ aufgrund von (ii) und (iii) mit Hilfe von majorisierter Konvergenz. Zuletzt gilt

$$\begin{aligned}
&\limsup_{n \to \infty} \sup_{||\eta_1 - \eta_2|| < \delta} |L_n(\eta_1) - L_n(\eta_2)| \\
\leq &\limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \sup_{||\eta_1 - \eta_2|| < \delta} |\ell(X_i, \eta_1) - \ell(X_i, \eta_2)| \\
= &\mathbb{E}_\vartheta \left[\sup_{||\eta_1 - \eta_2|| < \delta} |\ell(X_i, \eta_1) - \ell(X_i, \eta_2)| \right] \quad Q_\vartheta = \otimes_{i=1}^{\mathbb{N}} P_\vartheta\text{-f.s.}
\end{aligned}$$

nach dem starken Gesetz der großen Zahlen. Aufgrund der gleichmäßigen Stetigkeit auf dem Kompaktum Θ und wieder durch majorisierte Konvergenz erhalten wir die

Konvergenz dieses Ausdrucks gegen Null für $\delta \to 0$, also auch die dritte Bedingung von Satz 5.17 gemäß Bemerkung 5.18.

Es sei nun A_n das k-dimensionale Rechteck mit den Endpunkten $\hat{\theta}_n$ und ϑ. Aufgrund der Konsistenz von $\hat{\theta}_n$ und wegen $\vartheta \in \text{int}(\Theta)$ folgt $\lim_{n\to\infty} Q_\vartheta(A_n \subset \text{int}(\Theta)) = 1$. Auf A_n gilt zudem nach Definition

$$\dot{L}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}(X_i, \hat{\theta}_n) = 0,$$

und der Mittelwertsatz liefert

$$-\dot{L}_n(\vartheta) = \dot{L}_n(\hat{\theta}_n) - \dot{L}_n(\vartheta) = \ddot{L}_n(\tilde{\theta}_n)(\hat{\theta}_n - \vartheta) \tag{5.3}$$

für ein geeignetes $\tilde{\theta}_n \in A_n$.

Offenbar gilt:

$$\mathbb{E}[\dot{\ell}(X_i, \vartheta)] = \int_{\mathcal{X}} \dot{\ell}(x, \vartheta) f(x, \vartheta) \mu(dx) = 0$$

mit Hilfe von Bemerkung 2.25 (i), wobei man wieder majorisierte Konvergenz verwendet, um Ableitung und Integral zu vertauschen. Außerdem gilt

$$\text{Cov}(\dot{\ell}(X_i, \vartheta)) = I(f(\cdot, \vartheta))$$

nach Definition. Mit Satz 5.4 folgt

$$\sqrt{n}\dot{L}_n(\vartheta) \xrightarrow{\mathcal{L}} N(0, I(f(\cdot, \vartheta))).$$

Um aus (5.3) das Resultat folgern zu können, befassen wir uns zuletzt mit $\ddot{L}_n(\tilde{\theta}_n)$ und weisen die Konvergenz

$$\ddot{L}_n(\tilde{\theta}_n) \xrightarrow{Q_\vartheta} -I(f(\cdot, \vartheta)) \tag{5.4}$$

nach. Da $I(f(\cdot, \vartheta))$ nach Voraussetzung invertierbar ist, folgt dann

$$\lim_{n\to\infty} Q_\vartheta(\ddot{L}_n(\tilde{\theta}_n) \text{ ist invertierbar}) = 1.$$

Entsprechend ergibt sich mit einem geeigneten $B_n \xrightarrow{Q_\vartheta} 0$

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \vartheta) &= -\ddot{L}_n(\tilde{\theta}_n)^{-1}\dot{L}_n(\vartheta) 1_{\{A_n \subset \text{int}(\Theta)\} \cap \{\ddot{L}_n(\tilde{\theta}_n) \text{ ist invertierbar}\}} + B_n \\ &\xrightarrow{\mathcal{L}} I(f(\cdot, \vartheta))^{-1} N(0, I(f(\cdot, \vartheta))) = N(0, I(f(\cdot, \vartheta))^{-1})\end{aligned}$$

aus dem Lemma von Slutsky (Lemma 5.6).

Zum Beweis von (5.4) verwenden wir zunächst

$$\ddot{\ell}(x, \vartheta) = \frac{\ddot{f}(x, \vartheta)}{f(x, \vartheta)} - \dot{\ell}(x, \vartheta)\dot{\ell}(x, \vartheta)^T,$$

wodurch man (wieder mit majorisierter Konvergenz wie in Bemerkung 2.25 (i))

$$\mathbb{E}_\vartheta[\ddot{\ell}(X_i, \vartheta)] + I(f(\cdot, \vartheta)) = \mathbb{E}_\vartheta[\ddot{f}(X_i, \vartheta)/f(X_i, \vartheta)] = \int \ddot{f}(x, \vartheta)\mu(dx) = \ddot{1} = 0$$

erhält. Offenbar gilt

$$\ddot{L}_n(\vartheta) \to \mathbb{E}_\vartheta[\ddot{\ell}(X_i,\vartheta)] = -I(f(\cdot,\vartheta)) \quad Q_\vartheta\text{-f.s.}$$

nach dem starken Gesetz der großen Zahlen. Setzt man $\Omega_\delta^n = \{||\tilde{\theta}_n - \vartheta|| < \delta\}$, folgt

$$\mathbb{E}_\vartheta[|\ddot{L}_n(\tilde{\theta}_n) - \ddot{L}_n(\vartheta)|1_{\Omega_\delta^n}] \leq \mathbb{E}_\vartheta\left[\sup_{||\theta-\vartheta||<\delta} ||\ddot{\ell}(X_i,\theta) - \ddot{\ell}(X_i,\theta)||\right] \to 0$$

für $\delta \to 0$, nach Stetigkeit von $\ddot{\ell}$ und majorisierter Konvergenz. (5.4) ergibt sich dann aus

$$\lim_{\delta\to 0}\lim_{n\to\infty} Q_\vartheta(\Omega_\delta^n) = 1.$$

□

Bemerkung 5.20. Satz 5.17 und Satz 5.19 sind Spezialfälle zur Konsistenz bzw. zur asymptotischen Normalität von Minimum-Kontrast-Schätzern, die als Minimierer des empirischen Konstrasts

$$\eta \mapsto \frac{1}{n}\sum_{i=1}^n k(\eta, X_i)$$

für eine geeignete Funktion k definiert werden. Unter vergleichbaren Bedingungen lassen sich auch in diesem allgemeinen Fall ähnliche Resultate herleiten, wobei die Schätzer nicht mehr notwendigerweise asymptotisch effizient (wie für $k = -\ell$) sind.

Beispiel 5.21. Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \exp(\lambda)$. Um den Maximum-Likelihood-Schätzer für λ zu bestimmen, benötigen wir die n-dimensionale Dichte bzw. deren Log-Likelihood-Funktion, d.h.

$$f_n(x,\lambda) = \lambda^n \exp\left(-\lambda\sum_{i=1}^n x_i\right) 1_{(0,\infty)}\left(\min_{i=1}^n x_i\right)$$

bzw.

$$\ell_n(x,\lambda) = n\log(\lambda) - \lambda\sum_{i=1}^n x_i + \log\left(1_{(0,\infty)}\left(\min_{i=1}^n x_i\right)\right).$$

Leitet man den letzten Ausdruck nach λ ab und setzt die Ableitung gleich Null, so ergibt sich als Bedingung für den Maximum-Likelihood-Schätzer

$$\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0,$$

also

$$\hat{\lambda}_n = \overline{x}_n^{-1}.$$

Bevor wir dessen Asymptotik herleiten, berechnen wir die Fisher-Information des eindimensionalen Experiments. Es gilt

$$\ell_1(x,\lambda) = \log(\lambda) - \lambda x + \log(1_{(0,\infty)}(x)).$$

Mit Hilfe von

$$\mathbb{E}_\lambda[X] = \frac{1}{\lambda} \quad \text{und} \quad \mathrm{Var}_\lambda(X) = \frac{1}{\lambda^2}$$

ergibt sich für die Fisher-Information also

$$I(f(\cdot,\lambda)) = \mathbb{E}_\lambda\left[\left(X - \frac{1}{\lambda}\right)^2\right] = \mathrm{Var}_\lambda(X) = \frac{1}{\lambda^2}.$$

Aus dem Zentralen Grenzwertsatz (vgl. Satz 5.4) folgt

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right).$$

Wendet man die Delta-Methode (vgl. Satz 5.8) auf $g(x) = x^{-1}$ an, so ergibt sich mit $g'(\lambda^{-1}) = -\lambda^2$ zuletzt

$$\sqrt{n}\left(\hat{\lambda}_n - \lambda\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \lambda^2\right).$$

Dies entspricht exakt der Aussage von Satz 5.19. Wir beachten zudem, dass

$$\mathrm{Var}_\lambda(\hat{\lambda}_n) = n^{-1}\lambda^2 = (nI(f(\cdot,\lambda)))^{-1}$$

bereits eine Konsequenz aus Proposition 2.31 über exponentielle Familien ist.

Kapitel 6

Grundbegriffe der Testtheorie

In diesem Abschnitt sollen Hypothesen über den unbekannten Parameter $\vartheta \in \Theta$ untersucht werden. Dazu sei wieder $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ mit $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ ein statistisches Experiment.

Beispiel 6.1. Wir betrachten die vereinfachte Situation einer klinischen Studie, in der ein bekanntes Medikament A mit einem neuen Medikament B verglichen werden soll. Für das Medikament A ist aus langjähriger Erfahrung bekannt, dass die Heilungswahrscheinlichkeit 65% beträgt. Das neue Medikament B wurde an 100 Patienten getestet, von denen 80 geheilt werden konnten. Ist B also besser als A?

Bezeichnet man mit p die unbekannte Heilungswahrscheinlichkeit von B, dann kann die obige Frage mathematisch wie folgt beschrieben werden: Wir testen

$$H : p \leq 0.65 \quad \text{vs.} \quad K : p > 0.65.$$

Definition 6.2. Es sei $\Theta = \Theta_H \cup \Theta_K$ eine Partition des Parameterbereichs.

(i) Die Menge Θ_H heißt *Nullhypothese* (bzw. *Hypothese*), und die Menge Θ_K heißt *Alternative*.

(ii) Ein *randomisierter Test* ist eine messbare Abbildung

$$\varphi : (\mathcal{X}, \mathcal{B}) \to ([0,1], \mathcal{B}|_{[0,1]}).$$

Dabei gibt $\varphi(x)$ die Wahrscheinlichkeit für die Entscheidung $K : \vartheta \in \Theta_K$ an, falls $x = X(\omega)$ beobachtet wurde. Mit $\Phi := \{\varphi \mid \varphi \text{ ist randomisierter Test}\}$ bezeichnen wir die Menge aller Tests.

(iii) Für einen Test φ heißen $\mathcal{K} := \{x \in \mathcal{X} \mid \varphi(x) = 1\}$ *kritischer Bereich* und $\mathcal{R} := \{x \in \mathcal{X} \mid 0 < \varphi(x) < 1\}$ *Randomisierungsbereich*. Ein Test φ heißt *nichtrandomisiert*, falls $\mathcal{R} = \emptyset$ bzw. $\varphi(\mathcal{X}) \subset \{0,1\}$.

Beispiel 6.3. In der Situation von Beispiel 6.1 ist die Statistik $\overline{X}_n$ der UMVU-Schätzer für p. Eine naheliegende Entscheidungsregel besteht dann darin, für große Werte von $\overline{X}_n$ für $K : p > 0.65$ zu entscheiden. Zum Beispiel könnte man $\varphi : \mathbb{R}^n \to [0,1]$ mit

$$\varphi(x) = \begin{cases} 1, & \text{falls } \overline{X}_n > 0.7, \\ 0, & \text{falls } \overline{X}_n \leq 0.7, \end{cases} \tag{6.1}$$

wählen.

Bemerkung 6.4. Bei einer Entscheidung für H oder K mit Hilfe eines Tests φ können zwei Fehler auftreten:

- Fehler 1. Art: $\vartheta \in \Theta_H$, aber φ entscheidet für $\vartheta \in \Theta_K$.
- Fehler 2. Art: $\vartheta \in \Theta_K$, aber φ entscheidet für $\vartheta \in \Theta_H$.

Graphisch:

		Realität	
		Θ_H	Θ_K
Entscheidung	Θ_H	richtig	Fehler 2. Art
	Θ_K	Fehler 1. Art	richtig

Dabei treten beide Fehler mit bestimmten Wahrscheinlichkeiten auf.

Beispiel 6.5. Für den Test (6.1) aus Beispiel 6.3 erhält man für die Wahrscheinlichkeit einer Entscheidung für K

$$P_p(\varphi(X) = 1) = P_p(\overline{X}_n > 0.7).$$

Die exakte Berechnung dieser Wahrscheinlichkeit ist mühselig. Einfacher ist die Verwendung des Zentralen Grenzwertsatzes. Bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung, so ergibt sich aus $1 - \Phi(t) = \Phi(-t)$ die Approximation

$$P_p\Big(\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} > \frac{\sqrt{n}(0.7 - p)}{\sqrt{p(1-p)}}\Big) \approx \Phi\Big(\frac{\sqrt{n}(p - 0.7)}{\sqrt{p(1-p)}}\Big).$$

Zudem erlaubt das Lemma von Slutsky, im Nenner p durch $\overline{X}_n$ zu ersetzen. In unserer Situation, mit $n = 100$ und $\overline{x}_{100} = 80$, erhält man also näherungsweise

$$P_p(\varphi(X) = 1) \approx \Phi\Big(\frac{10(p - 0.7)}{0.4}\Big).$$

Dieser Ausdruck hängt nur noch von p ab. Beachtet man die Monotonie bzgl. p, so ergibt sich zum Beispiel

$$P_p(\text{Fehler 1. Art}) \approx \begin{cases} 0, & p = 0.5, \\ 0.006, & p = 0.6, \end{cases}$$

und als obere Schranke

$$P_p(\text{Fehler 1. Art}) \leq \sup_{p \leq 0.65} P_p(\varphi(X) = 1) \approx 0.106.$$

Aus Symmetriegründen gilt andererseits

$$P_p(\text{Fehler 2. Art}) \approx \begin{cases} 0, & p = 0.9, \\ 0.006, & p = 0.8, \\ 0.5, & p = 0.7, \end{cases}$$

sowie

$$P_p(\text{Fehler 2. Art}) \leq \sup_{p > 0.65} P_p(\varphi(X) = 0) \approx 0.894.$$

Bemerkung 6.6. Wünschenswert ist natürlich die simultane Minimierung der Wahrscheinlichkeiten für den Fehler 1. und 2. Art durch die Wahl eines geeigneten Tests φ. Ein solcher Test existiert allerdings nicht, wie man an dem folgendem Beispiel sieht.

$$\varphi_1(x) \equiv 1 \ \forall x \in \mathcal{X} \implies \begin{cases} \text{Wahrscheinlichkeit für Fehler 1. Art} & = 1, \\ \text{Wahrscheinlichkeit für Fehler 2. Art} & = 0. \end{cases}$$

$$\varphi_0(x) \equiv 0 \ \forall x \in \mathcal{X} \implies \begin{cases} \text{Wahrscheinlichkeit für Fehler 1. Art} & = 0, \\ \text{Wahrscheinlichkeit für Fehler 2. Art} & = 1. \end{cases}$$

In der Praxis geht man daher wie folgt vor: Man legt eine Schranke α für die Wahrscheinlichkeit eines Fehlers 1. Art fest (typischerweise $0.01 \leq \alpha \leq 0.1$) und wählt dann einen Test φ, der die Wahrscheinlichkeit für einen Fehler 2. Art unter dieser Nebenbedingung minimiert. In der Regel wählt man dabei die Variante mit den schwerwiegenderen Konsequenzen als Alternative, da man die Wahrscheinlichkeit des Fehlers einer fälschlichen Entscheidung für K durch α kontrollieren kann.

Definition 6.7. Es sei φ ein Test für die Hypothesen: $H : \vartheta \in \Theta_H$ vs. $K : \vartheta \in \Theta_K$.

(i) Die Funktion

$$\beta_\varphi : \begin{cases} \Theta \to [0,1] \\ \vartheta \mapsto \mathbb{E}_\vartheta[\varphi(X)] \end{cases}$$

heißt *Gütefunktion von* φ.

(ii) φ heißt *Test zum Niveau* α, falls für alle $\vartheta \in \Theta_H$ gilt: $\beta_\varphi(\vartheta) \leq \alpha$. Wir setzen $\Phi_\alpha := \{\varphi \in \Phi \mid \varphi \text{ ist Test zum Niveau } \alpha\}$.

(iii) φ heißt *unverfälscht zum Niveau* α, falls $\varphi \in \Phi_\alpha$ ist und

$$\beta_\varphi(\vartheta) \geq \alpha \ \forall \vartheta \in \Theta_K$$

gilt. Wir setzen $\Phi_{\alpha\alpha} := \{\varphi \in \Phi_\alpha \mid \varphi \text{ ist unverfälscht}\}$.

Bemerkung 6.8.

(i) Ist φ nichtrandomisiert, dann bezeichnet $\beta_\varphi(\vartheta) = P_\vartheta(\varphi(X) = 1)$ die Wahrscheinlichkeit, für K zu entscheiden, falls $\vartheta \in \Theta$ der wahre Parameter ist. Insbesondere ergibt sich:

- Für $\vartheta \in \Theta_H$ ist $\beta_\varphi(\vartheta)$ die Wahrscheinlichkeit eines Fehler 1. Art.
- Für $\vartheta \in \Theta_K$ ist $1 - \beta_\varphi(\vartheta)$ die Wahrscheinlichkeit eines Fehler 2. Art.

Eine analoge Interpretation ergibt sich für randomisierte Tests.

(ii) Ist φ ein Test zum Niveau α, dann ist die Wahrscheinlichkeit für einen Fehler 1. Art für alle $\vartheta \in \Theta_H$ kleiner oder gleich α.

(iii) Ist φ unverfälscht, dann ist für jedes $\vartheta \in \Theta_K$ die Wahrscheinlichkeit einer Entscheidung für Θ_K nicht kleiner als für alle $\vartheta \in \Theta_H$.

Beispiel 6.9. Für die Gütefunktion des Tests aus Beispiel 6.3 erhält man die Näherung

$$\beta_\varphi : \begin{cases} [0,1] \to [0,1] \\ p \mapsto \beta_\varphi(p) \approx \Phi\Big(\frac{\sqrt{n}(p-0{,}7)}{\sqrt{\overline{x}_n(1-\overline{x}_n)}}\Big). \end{cases}$$

Zumindest approximativ ist φ ein Test zum Niveau $\alpha = 0.106$.

Definition 6.10.

(i) Ein Test $\varphi^* \in \Phi_\alpha$ heißt *gleichmäßig bester Test zum Niveau* α, falls

$$\beta_{\varphi^*}(\vartheta) = \sup_{\varphi \in \Phi_\alpha} \beta_\varphi(\vartheta) \quad \forall \vartheta \in \Theta_K.$$

Solche Tests werden als *UMP-Tests* (für "uniformly most powerful") bezeichnet.

(ii) Ein Test $\varphi^* \in \Phi_{\alpha\alpha}$ heißt *gleichmäßig bester unverfälschter Test*, falls

$$\beta_{\varphi^*}(\vartheta) = \sup_{\varphi \in \Phi_{\alpha\alpha}} \beta_\varphi(\vartheta) \quad \forall \vartheta \in \Theta_K.$$

Solche Tests werden als *UMPU-Tests* (für "uniformly most powerful unbiased") bezeichnet.

Satz 6.11. *Es seien* $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ *eine Klasse von Verteilungen auf* $(\mathcal{X}, \mathcal{B})$, $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{T}, \mathcal{D})$ *eine suffiziente Statistik für* ϑ *und* $\varphi : \mathcal{X} \to [0,1]$ *ein Test. Dann existiert ein Test* $\psi \circ T$*, der dieselbe Gütefunktion wie* φ *besitzt, nämlich*

$$\psi \circ T = \mathbb{E}[\varphi | T].$$

Beweis: Wir setzen $\psi(t) = \mathbb{E}[\varphi | T = t]$. Offenbar gilt $\psi(T(x)) \in [0,1]$ für alle $x \in \mathcal{X}$, und wegen des Satzes über die iterierte Erwartung folgt

$$\beta_{\psi \circ T}(\vartheta) = \mathbb{E}_\vartheta[\psi \circ T] = \mathbb{E}_\vartheta[\mathbb{E}_\vartheta[\varphi | T]] = \mathbb{E}_\vartheta[\varphi] = \beta_\varphi(\vartheta).$$

□

Bemerkung 6.12.

(i) Satz 6.11 zeigt, dass man sich bei der Konstruktion von Tests auf Verfahren beschränken kann, die von einer suffizienten Statistik abhängen.

(ii) Das einfachste Beispiel besteht in der Untersuchung von *einfachen Hypothesen*, d.h. $\Theta = \{\vartheta_0, \vartheta_1\}$ mit $\Theta_H = \{\vartheta_0\}$ und $\vartheta_K = \{\vartheta_1\}$. Wählt man z.B. $\mu = P_{\vartheta_0} + P_{\vartheta_1}$, so dominiert μ beide Maße, und wir bezeichnen mit

$$p_i = \frac{dP_{\vartheta_i}}{d\mu}, \ i = 0, 1,$$

die jeweiligen μ-Dichten von P_{ϑ_i}. Wir betrachten im Folgenden den Quotienten p_1/p_0, wobei wir $p_0 = 0$ und $p_1 > 0$ $p_1/p_0 = \infty$ und für $p_0 = 0$ und $p_1 = 0$ p_1/p_0 beliebig setzen. Dann gilt:

(a) p_1/p_0 ist (bis auf Nullmengen) unabhängig von der Wahl von μ.

(b) p_1/p_0 ist minimalsuffizient für ϑ (vgl. Beispiel 4.20).

(c) Ein UMP-Test zum Niveau α maximiert

$$\beta_\varphi(\vartheta_1) = \mathbb{E}_{\vartheta_1}[\varphi(X)] = \int p_1(x)\varphi(x)\mu(dx)$$

unter der Nebenbedingung

$$\beta_\varphi(\vartheta_0) = \mathbb{E}_{\vartheta_0}[\varphi(X)] = \int p_0(x)\varphi(x)\mu(dx) \leq \alpha.$$

Definition 6.13. In der Situation einfacher Hypothesen heißt ein Test φ *Neyman-Pearson-Test* (kurz: NP-Test), falls ein $c \in [0, \infty]$ existiert, so dass φ die Darstellung

$$\varphi(x) = \begin{cases} 1, & \text{falls } p_1(x) > cp_0(x), \\ 0, & \text{falls } p_1(x) < cp_0(x), \end{cases}$$

besitzt.

Satz 6.14. (Neyman-Pearson-Lemma) *Es seien $\Theta = \{\vartheta_0, \vartheta_1\}$ mit $\Theta_H = \{\vartheta_0\}$ und $\Theta_K = \{\vartheta_1\}$ einfache Hypothesen.*

(i) Ein NP-Test φ^* ist ein UMP-Test zum Niveau $\alpha = \mathbb{E}_{\vartheta_0}[\varphi^*(X)]$ für die Hypothesen $H : \vartheta = \vartheta_0$ vs. $K : \vartheta = \vartheta_1$.

(ii) Für alle $\alpha \in [0, 1]$ existiert ein NP-Test φ mit $\mathbb{E}_{\vartheta_0}[\varphi(X)] = \alpha$ für die Hypothesen $H : \vartheta = \vartheta_0$ vs. $K : \vartheta = \vartheta_1$.

(iii) Ist φ' ein UMP-Test zum Niveau α für $H : \vartheta = \vartheta_0$ vs. $K : \vartheta = \vartheta_1$, dann ist φ' μ-f.ü. ein NP-Test. Ist $\mathbb{E}_{\vartheta_0}[\varphi'(X)] < \alpha$, so folgt $\mathbb{E}_{\vartheta_1}[\varphi'(X)] = 1$.

Beweis:

(i) Es sei φ^* NP-Test mit der zugehörigen Konstante c^*. Ferner sei φ ein weiterer Test zum Niveau α, d.h. es gelte $\beta_\varphi(\vartheta_0) \leq \alpha := \beta_{\varphi^*}(\vartheta_0)$. Zu zeigen ist, dass

$$\beta_{\varphi^*}(\vartheta_1) - \beta_\varphi(\vartheta_1) = \int p_1(\varphi^* - \varphi)d\mu = \int (\varphi^* - \varphi)(p_1 - c^* p_0)d\mu + \int c^* p_0(\varphi^* - \varphi)d\mu$$

nicht-negativ ist. Dabei entspricht der hintere Ausdruck

$$c^*(\beta_{\varphi^*}(\vartheta_0) - \beta_\varphi(\vartheta_0)) \geq 0.$$

Der vordere Term ist wegen

$$\varphi^* - \varphi > 0 \Longrightarrow \varphi^* > 0 \Longrightarrow p_1 \geq c^* p_0$$

bzw.

$$\varphi^* - \varphi < 0 \Longrightarrow \varphi^* < 1 \Longrightarrow p_1 \leq c^* p_0$$

nicht-negativ.

(ii) Für $c \in \mathbb{R}$ sei

$$\alpha(c) := P_{\vartheta_0}\Big(\frac{p_1(X)}{p_0(X)} > c\Big).$$

Offensichtlich ist $1 - \alpha(c)$ eine Verteilungsfunktion.

Ist $\alpha \neq 0$, bestimme man c^* so, dass $\alpha(c^*) \leq \alpha \leq \alpha(c^*-)$ gelte, wobei $\alpha(c-)$ den linksseitigen Grenzwert von α an der Stelle c bezeichnet. Andernfalls sei $c^* = \infty$. Wir setzen nun $\gamma^* = \frac{\alpha - \alpha(c^*)}{\alpha(c^*-0) - \alpha(c^*)}$, falls $\alpha(c^*) > \alpha(c^* - 0)$ gilt, (und $\gamma^* = 0$ sonst) und definieren den NP-Test

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } p_1(x) > c^* p_0(x), \\ 0, & \text{falls } p_1(x) < c^* p_0(x), \\ \gamma^*, & \text{falls } p_1(x) = c^* p_0(x). \end{cases}$$

Falls $c^* < \infty$ gilt, so folgt

$$\begin{aligned} \beta_{\varphi^*}(\vartheta_0) &= \int_{\{p_1 > c^* p_0\}} dP_{\vartheta_0} + \gamma^* \int_{\{p_1 = c^* p_0\}} dP_{\vartheta_0} \\ &= \alpha(c^*) + \gamma^*(\alpha(c^* - 0) - \alpha(c^*)) = \alpha. \end{aligned}$$

Gilt $c^* = \infty$, so folgt

$$\beta_{\varphi^*}(\vartheta_0) = P_{\vartheta_0}(p_1(X) > c^* p_0(X)) = 0,$$

da das Ereignis nur auf der P_{ϑ_0}-Nullmenge eintreten kann, auf der $p_0(x) = 0$ ist.

(iii) Es sei φ' ein UMP-Test zum Niveau α und andererseits φ^* der Test aus (ii) zum selben α. Einerseits gilt $\beta_{\varphi^*}(\vartheta_1) \leq \beta_{\varphi'}(\vartheta_1)$ aus der UMP-Eigenschaft, andererseits nach Teil (i) auch $\beta_{\varphi'}(\vartheta_1) \leq \beta_{\varphi^*}(\vartheta_1)$. Damit erhält man

$$0 = \int (\varphi^* - \varphi') p_1 d\mu = \int (\varphi^* - \varphi')(p_1 - c^* p_0) d\mu + \int c^* p_0 (\varphi^* - \varphi') d\mu = I + II.$$

Wegen $\beta_{\varphi^*}(\vartheta_0) = \alpha \geq \beta_{\varphi'}(\vartheta_0)$ und $c^* \geq 0$ folgt $II \geq 0$, und aufgrund von $p_1 > c^* p_0 \Longrightarrow \varphi^* - \varphi' = 1 - \varphi' \geq 0$ bzw. $p_1 < c^* p_0 \Longrightarrow \varphi^* - \varphi' = -\varphi' \leq 0$ gilt das auch für I. Damit folgt $I = II = 0$.

Sei nun $S = \{x \mid \varphi'(x) \neq \varphi^*(x)\} \cap \{x \mid p_1(x) \neq c^* p_0(x)\}$. Mit der Argumentation von oben gilt auf S die Eigenschaft $(\varphi^* - \varphi')(p_1 - c^* p_0) > 0$, und es folgt wegen $I = 0$ sofort $\mu(S) = 0$. Außerhalb von S ist φ' ein NP-Test.

Aus $II = 0$ folgt $c^* = 0$ oder $\beta_{\varphi'}(\vartheta_0) = \beta_{\varphi^*}(\vartheta_0) = \alpha$. Falls $\beta_{\varphi'}(\vartheta_0) < \alpha$ ist, erhält man also $c^* = 0$ und damit $\varphi^*(x) = 1$ für alle x mit $p_1(x) > 0$. Also gilt $\beta_{\varphi^*}(\vartheta_1) = \int \varphi^* p_1 d\mu = 1$. Da φ' ein UMP-Test ist, erhält man $\beta_{\varphi'}(\vartheta_1) = \beta_{\varphi^*}(\vartheta_1) = 1$. □

Bemerkung 6.15. Der UMP-Test φ^* für $H : \vartheta = \vartheta_0$ vs. $K : \vartheta = \vartheta_1$ ist außerhalb der Menge $S_= = \{x \mid p_1(x) = c^* p_0(x)\}$ eindeutig bestimmt. Auf $S_=$ kann der optimale Test φ^* beliebig definiert werden, so dass $\beta_{\varphi^*}(\vartheta_0) = \alpha$ gilt. Eine derartige Möglichkeit im Beweis von Teil (ii) gegeben worden.

Korollar 6.16. *Jeder NP-Test φ^* mit $\beta_{\varphi^*}(\vartheta_0) \in (0, 1)$ ist unverfälscht. Insbesondere gilt $\alpha := \beta_{\varphi^*}(\vartheta_0) < \beta_{\varphi^*}(\vartheta_1)$.*

Beweis: Der Test $\varphi \equiv \alpha$ ist ein Test zum Niveau α. Da φ^* der UMP-Test zum selben Niveau ist, erhält man $\alpha = \beta_\varphi(\vartheta_1) \le \beta_{\varphi^*}(\vartheta_1) = \beta$. Falls $\alpha = \beta < 1$ gilt, so ist $\varphi \equiv \alpha$ ein UMP-Test und nach Satz 6.14 (iii) ein NP-Test. Insbesondere folgt $p_1 = p_0$ μ-f.s., also $P_{\vartheta_0} = P_{\vartheta_1}$. Dies ist ausgeschlossen. □

Beispiel 6.17. Es seien $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ mit bekannter Varianz σ^2. Wir untersuchen das Testproblem

$$H : \mu = \mu_0 \quad \text{vs.} \quad K : \mu = \mu_1$$

mit $\mu_0 < \mu_1$. Für die Dichte von $X_1, \dots, X_n$ erhält man dann

$$p_j(x) = (2\pi\sigma^2)^{-n/2} \exp\Big\{ -\frac{1}{2\sigma^2}\Big(\sum_{i=1}^n x_i^2 - 2\mu_j \sum_{i=1}^n x_i + n\mu_j^2 \Big) \Big\}, \quad j = 0, 1.$$

Als Ungleichung für den Dichte-Quotienten (oder *Likelihood-Quotienten*), der zur Konstruktion des NP-Tests benötigt wird, ergibt sich

$$\frac{p_1(x)}{p_0(x)} = \exp\Big\{ \frac{1}{\sigma^2} \sum_{i=1}^n x_i(\mu_1 - \mu_0) \Big\} \cdot f(\sigma^2, \mu_1, \mu_0) > \tilde{c},$$

wobei die Konstante $f(\sigma^2, \mu_1, \mu_0) > 0$ nicht von x abhängt. Diese Ungleichung ist wegen $\mu_1 > \mu_0$ äquivalent zu

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^n x_i > c.$$

Für die Berechnung der Fehlerwahrscheinlichkeit beachtet man, dass

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

gilt. Bezeichnet u_β das β-Quantil der Standardnormalverteilung, also den Wert mit $\Phi(u_\beta) = \beta$, erhält man

$$\alpha(c) = P_{\mu_0}\Big(\overline{X}_n > c \Big) = \alpha \quad \Longleftrightarrow \quad c = \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}.$$

Da stetige Verteilungen vorliegen, kann man $\gamma = 0$ wählen und erhält mit

$$\varphi(x) = \begin{cases} 1, & \text{falls } \overline{x}_n > \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \\ 0, & \text{falls } \overline{x}_n \le \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \end{cases}$$

einen UMP-Test für $H : \mu = \mu_0$ vs. $K : \mu = \mu_1$.

Beispiel 6.18. Es seien $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{U}[0, \vartheta]$. Wir untersuchen das Testproblem $H : \vartheta = \vartheta_0$ vs. $K : \vartheta = \vartheta_1$ mit $\vartheta_0 < \vartheta_1$. Für die gemeinsame Dichte von $X_1, \dots, X_n$ erhält man mit $x_{(n)} = \max_{i=1}^n x_i$

$$p_j(x) = \Big(\frac{1}{\vartheta_j} \Big)^n 1_{[0, \vartheta_j]}(x_{(n)}).$$

Damit ist der Likelihood-Quotient durch

$$L(x) = \frac{p_1(x)}{p_0(x)} = \begin{cases} \left(\frac{\vartheta_0}{\vartheta_1}\right)^n, & \text{falls } x_{(n)} \le \vartheta_0, \\ \infty, & \text{falls } x_{(n)} \in (\vartheta_0, \vartheta_1], \\ \text{beliebig}, & \text{sonst}, \end{cases}$$

gegeben, und es gilt

$$\alpha(c) = P_{\vartheta_0}(p_1(X) > cp_0(X)) = \begin{cases} 1, & \text{falls } c < \left(\frac{\vartheta_0}{\vartheta_1}\right)^n \\ 0, & \text{falls } c \ge \left(\frac{\vartheta_0}{\vartheta_1}\right)^n. \end{cases}$$

Für $\alpha \in (0,1)$ gilt

$$\alpha(c) \le \alpha \le \alpha(c-0) \quad \Longleftrightarrow \quad c = c^* = \left(\frac{\vartheta_0}{\vartheta_1}\right)^n,$$

und nach dem Neyman-Pearson-Lemma hat jeder UMP-Test die Form

$$\varphi(x) = 1_{\{L(x)>c^*\}} + \gamma(x)1_{\{L(x)=c^*\}} = 1_{\{x_{(n)}>\vartheta_0\}} + \gamma(x)1_{\{x_{(n)}\le\vartheta_0\}}.$$

Einige mögliche Festlegungen sind

$$\begin{aligned} \varphi_1(x) &= 1\{x_{(n)} > c_1\}, \quad c_1 = \vartheta_0 \sqrt[n]{1-\alpha}, \\ \varphi_2(x) &= 1\{x_{(n)} > \vartheta_0\} + 1\{x_{(n)} < c_2\}, \quad c_2 = \vartheta_0 \sqrt[n]{\alpha}, \\ \varphi_3(x) &= 1\{x_{(n)} > \vartheta_0\} + \alpha 1\{x_{(n)} \le \vartheta_0\}, \quad \text{(vgl. Satz 6.14 (ii))}. \end{aligned}$$

Auf der Menge $\{x \mid p_1(x) \ne c^* p_0(x)\} = \{x \mid x_{(n)} > \vartheta_0\}$ gilt stets $\varphi_j(x) = 1_{\{x_{(n)}>\vartheta_0\}}$, und in allen drei Fällen ist $\beta_{\varphi_j}(\vartheta_0) = \mathbb{E}_{\vartheta_0}[\varphi_j(X)] = \alpha$.

Bemerkung 6.19. In der Regel sind einfache Hypothesen nicht praxisrelevant, aber aus zweierlei Gründen für das Verständnis der Testtheorie wichtig:

(i) Sie geben ein intuitives Gefühl dafür, wie Tests zu konstruieren sind: Zunächst benötigt man sogenannte *Konfidenzbereiche* $c(X) \subset \Theta$, in denen sich der unbekannte Parameter mit Wahrscheinlichkeit $1-\alpha$ befindet. In Beispiel 6.17 haben wir etwa verwendet, dass mit $c(X) = [\overline{X}_n - u_{1-\alpha}\sigma/\sqrt{n}, \infty)$

$$P_{\mu_0}(\mu_0 \in c(X)) = P_{\mu_0}\left(\overline{X}_n \le \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right) = 1-\alpha$$

gilt. Grundsätzlich liefert jede Wahl eines Konfidenzbereichs mit $P_{\mu_0}(\mu_0 \in c(X)) = 1-\alpha$ einen Test zum Niveau α, indem man $\varphi(x) = 1_{\{\mu_0 \notin c(X)\}}$ setzt, zum Beispiel auch

$$c'(X) = [\overline{X}_n - u_{1-\alpha/2}\sigma/\sqrt{n}, \overline{X}_n + u_{1-\alpha/2}\sigma/\sqrt{n}].$$

Zusätzlich weiß man bei einfachen Hypothesen jedoch, in welche Richtung eine Abweichung entdeckt werden soll, weshalb der Test aus Beispiel 6.17 optimal ist.

(ii) Formale Resultate wie das Neyman-Pearson-Lemma sind ein wichtiges Hilfsmittel, um die Optimalität von Tests für kompliziertere Hypothesen zu zeigen.

Definition 6.20. Es seien $\Theta \subset \mathbb{R}$, $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\} \ll \mu$ für ein σ-endliches Maß μ und $T : (\mathcal{X}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ eine Statistik. Die Verteilungsfamilie $\mathcal{P}$ heißt *Klasse mit (streng) isotonem Dichtequotienten in* T, falls für alle $\vartheta_0, \vartheta_1 \in \Theta$ mit $\vartheta_0 < \vartheta_1$ eine (streng) monoton wachsende Funktion $H_{\vartheta_0,\vartheta_1} : \mathbb{R} \to [0, \infty]$ mit

$$\frac{p_{\vartheta_1}(x)}{p_{\vartheta_0}(x)} = H_{\vartheta_0,\vartheta_1}(T(x)) \quad P_{\vartheta_0} + P_{\vartheta_1}\text{-f.s.}$$

existiert.

Beispiel 6.21.

(i) In der Situation von Beispiel 6.17 war

$$\frac{p_{\mu_1}(x)}{p_{\mu_0}(x)} = \exp\left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i(\mu_1 - \mu_0)\right\} \cdot f(\sigma^2, \mu_1, \mu_0).$$

Wegen $f(\sigma^2, \mu_1, \mu_0) \geq 0$ und $\mu_1 > \mu_0$ ist der Likelihoodquotient streng isoton in $T(x) = \overline{x}_n$. Allgemein lässt sich dies für jede 1-parametrige Exponentialfamilie mit Dichte

$$p_\vartheta(x) = c(\vartheta)h(x)\exp\left(Q(\vartheta)T(x)\right)$$

zeigen, sofern Q monoton wachsend ist.

(ii) In der Situation von Beispiel 6.18 war

$$\frac{p_{\vartheta_1}(x)}{p_{\vartheta_0}(x)} = \left(\frac{\vartheta_0}{\vartheta_1}\right)^n 1_{[0,\vartheta_0]}(x_{(n)}) + \infty 1_{(\vartheta_0,\vartheta_1]}(x_{(n)}) \quad P_{\vartheta_0} + P_{\vartheta_1}\text{-f.s.},$$

was offensichtlich isoton in $T(x) = x_{(n)}$ ist.

Satz 6.22. *Es seien $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine Klasse mit isotonem Dichtequotienten in T, $\vartheta_0 \in \Theta$ und $\alpha \in (0, 1)$. Ferner sei ein Test der Form*

$$\varphi^*(x) = 1_{\{T(x)>c\}} + \gamma 1_{\{T(x)=c\}}$$

mit $c := \inf\{t \mid P_{\vartheta_0}(T(X) > t) \leq \alpha\}$ und

$$\gamma = \begin{cases} \frac{\alpha - P_{\vartheta_0}(T(x)>c)}{P_{\vartheta_0}(T(X)=c)}, & \text{falls } P_{\vartheta_0}(T(X) = c) > 0, \\ 0, & \text{falls } P_{\vartheta_0}(T(X) = c) = 0, \end{cases}$$

gegeben. Dann gilt:

(i) $\beta_{\varphi^*}(\vartheta_0) = \alpha$, und φ^* ist ein UMP-Test zum Niveau α für die einseitigen Hypothesen

$$H : \vartheta \leq \vartheta_0 \quad \text{vs.} \quad K : \vartheta > \vartheta_0.$$

(ii) Für alle $\vartheta < \vartheta_0$ gilt: $\beta_{\varphi^*}(\vartheta) = \inf\{\beta_\varphi(\vartheta) \mid \varphi \in \Phi \text{ mit } \beta_\varphi(\vartheta_0) = \alpha\}$.

(iii) Die Gütefunktion: $\beta_{\varphi^*} : \vartheta \mapsto \beta_{\varphi^*}(\vartheta)$ ist strikt isoton für alle ϑ mit $\beta_{\varphi^*}(\vartheta) \in (0, 1)$.

(iv) Für alle $\vartheta' \in \Theta$ ist φ^* ein UMP-Test zum Niveau $\alpha' := \beta_{\varphi^*}(\vartheta')$ für die Hypothesen $H' : \vartheta \leq \vartheta'$ vs. $K' : \vartheta > \vartheta'$.

Beweis:

(i) Gilt $P_{\vartheta_0}(T(X) = c) = 0$, dann folgt

$$\beta_{\varphi^*}(\vartheta_0) = \mathbb{E}_{\vartheta_0}[\varphi^*(X)] = P_{\vartheta_0}(T(X) > c) = \alpha.$$

Andernfalls gilt $P_{\vartheta_0}(T(X) = c) > 0$, also

$$\beta_{\varphi^*}(\vartheta_0) = P_{\vartheta_0}(T(X) > c) + \gamma P_{\vartheta_0}(T(X) = c) = \alpha.$$

Ansonsten sei $\vartheta_0 < \vartheta_1$ und $H_{\vartheta_0,\vartheta_1}(T(x)) = p_{\vartheta_1}(x)/p_{\vartheta_0}(x)$ wie in Definition 6.20. Aufgrund der Isotonie gilt

$$H_{\vartheta_0,\vartheta_1}(T(x)) > H_{\vartheta_0,\vartheta_1}(c) = s \Longrightarrow T(x) > c,$$

und analog mit $<$. Es ergibt sich

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } \frac{p_{\vartheta_1}(x)}{p_{\vartheta_0}(x)} = H_{\vartheta_0,\vartheta_1}(T(x)) > s, \\ 0, & \text{falls } \frac{p_{\vartheta_1}(x)}{p_{\vartheta_0}(x)} = H_{\vartheta_0,\vartheta_1}(T(x)) < s. \end{cases}$$

Wegen $\beta_{\varphi^*}(\vartheta_0) = \alpha$ ist φ^* ein NP-Test zum Niveau α, und nach Satz 6.14 (i) gilt:

$$\beta_{\varphi^*}(\vartheta_1) = \sup\{\beta_\varphi(\vartheta_1) \mid \varphi \in \Phi \text{ mit } \beta_\varphi(\vartheta_0) = \alpha\}. \tag{6.2}$$

Da φ^* unabhängig von ϑ_1 ist, erhält man (6.2) für alle $\vartheta_1 > \vartheta_0$.

Betrachtet man $\varphi'(x) = 1 - \varphi^*(x)$, so kann man analog

$$\beta_{\varphi'}(\vartheta_2) = \sup\{\beta_\varphi(\vartheta_2) \mid \varphi \in \Phi \text{ mit } \beta_\varphi(\vartheta_0) = 1 - \alpha\} \tag{6.3}$$

für alle $\vartheta_2 < \vartheta_0$ zeigen. Für $\overline{\varphi} \equiv \alpha$ gilt $\beta_{\overline{\varphi}}(\vartheta_0) = \alpha$, und man erhält aus (6.3) die Ungleichung

$$1 - \beta_{\varphi^*}(\vartheta_2) = \beta_{\varphi'}(\vartheta_2) \geq \beta_{1-\overline{\varphi}}(\vartheta_2) = 1 - \beta_{\overline{\varphi}}(\vartheta_2) = 1 - \alpha$$

für alle $\vartheta_2 < \vartheta_0$. Insbesondere folgt $\beta_{\varphi^*}(\vartheta_2) \leq \alpha$, also $\varphi^* \in \Phi_\alpha$. Nach (6.2) ist φ^* ein UMP-Test.

(ii) Diese Aussage folgt direkt aus (6.3), da $\beta_{\varphi'} = 1 - \beta_{\varphi^*}$ ist.

(iii) Diese Aussage folgt direkt aus Korollar 6.16, da φ^* für zwei beliebige $\vartheta_1 < \vartheta_2$ ein NP-Test ist.

(iv) Diese Aussage lässt sich genau wie Teil (i) beweisen. □

Beispiel 6.23. Es seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ bei bekannter Varianz σ^2. Aus Beispiel 6.17 folgt, dass die Verteilungsfamilie mit Dichten

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

einen streng isotonen Dichtequotienten in $T(x) = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}_n$ besitzt. Nach Satz 6.22 ist ein UMP-Test zum Niveau α für die Hypothesen $H : \mu \leq \mu_0$ vs. $K : \mu > \mu_0$ dann durch

$$\varphi^*(x) = 1_{\{\overline{x}_n > c\}} + \gamma 1_{\{\overline{x}_n = c\}}$$

gegeben. Die Wahl von c wird anhand von $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ vollzogen. Es gilt $P_{\mu_0}(\overline{X}_n = c) = 0$, also $\gamma = 0$. Entsprechend wird c so gewählt, dass

$$P_{\mu_0}(\overline{X}_n > c) = \alpha \quad \Longleftrightarrow \quad c = \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$$

gilt. Also ist ein UMP-Test für die Hypothesen $H : \mu \leq \mu_0$ vs. $K : \mu > \mu_0$ durch den bekannten Test

$$\varphi^*(x) = 1_{\{\overline{X}_n > \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\}}$$

aus Beispiel 6.17 gegeben. Er heißt *einseitiger Gauß-Test*.

Bemerkung 6.24.

(i) Es gibt auch eine heuristische Herleitung für den einseitigen Gauß-Test. Da $\overline{X}_n$ der UMVU-Schätzer für μ ist, besteht eine naheliegende Entscheidungsregel darin, bei "großen" Werten von $\overline{X}_n$ für K zu entscheiden und ansonsten die Hypothese beizubehalten. Es ergibt sich ein Test der Form

$$\varphi^*(x) = 1_{\{\overline{x}_n > c\}}.$$

Die Festlegung von c erfolgt dann durch Untersuchung des Fehlers 1. Art. Für alle $\mu \leq \mu_0$ gilt

$$\begin{aligned} \beta_{\varphi^*}(\mu) =& P_\mu(\overline{X}_n > c) = P_\mu\Big(\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} > \frac{c - \mu}{\sqrt{\sigma^2/n}}\Big) \\ =& 1 - \Phi\Big(\frac{c - \mu}{\sqrt{\sigma^2/n}}\Big) \leq 1 - \Phi\Big(\frac{c - \mu_0}{\sqrt{\sigma^2/n}}\Big) \end{aligned} \tag{6.4}$$

aufgrund der Monotonie der Verteilungsfunktion Φ. Entsprechend erhält man

$$\begin{aligned} \beta_{\varphi^*}(\mu) \leq \alpha \ \forall \mu \leq \mu_0 \quad &\Longleftrightarrow \quad \Phi\Big(\frac{c - \mu_0}{\sqrt{\sigma^2/n}}\Big) \geq 1 - \alpha \\ &\Longleftrightarrow \quad c \geq \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}. \end{aligned}$$

Will man die Irrtumswahrscheinlichkeit α voll ausschöpfen, so wählt man das minimale

$$c = \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}.$$

(ii) Diese Methode liefert keine Aussage über die Optimalität des Tests. Dafür kann sie aber auch in allgemeineren Situationen verwendet werden, etwa falls σ^2 unbekannt ist. In diesem Fall "studentisiert" man, d.h. man ersetzt σ^2 durch den Schätzer

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2$$

und erhält analog zu (6.4) mit Hilfe von Korollar 2.21

$$\beta_{\varphi^*}(\mu) = P_\mu\Big(\frac{\overline{X}_n - \mu}{\sqrt{\hat{\sigma}_n^2/n}} > \frac{c - \mu}{\sqrt{\hat{\sigma}_n^2/n}}\Big) = 1 - F_{t_{n-1}}\Big(\frac{c - \mu}{\sqrt{\hat{\sigma}_n^2/n}}\Big),$$

wobei $F_{t_{n-1}}$ die Verteilungsfunktion der t-Verteilung mit $n-1$ Freiheitsgraden bezeichnet. Wie zuvor ergibt sich als sinnvolle Wahl

$$c = \mu_0 + \frac{\hat{\sigma}_n}{\sqrt{n}} t_{n-1,1-\alpha},$$

wobei $t_{n-1,\beta}$ das entsprechende β-Quantil bezeichnet. Als Test erhält man

$$\varphi^*(x) = 1_{\{\overline{x}_n > \mu_0 + \frac{\hat{\sigma}_n}{\sqrt{n}} t_{n-1,1-\alpha}\}},$$

den sogenannten *Ein-Stichproben-t-Test*.

Bemerkung 6.25. Für Hypothesen der Form

$$H : \vartheta = \vartheta_0 \quad \text{vs.} \quad K : \vartheta \neq \vartheta_0$$

existieren im Allgemeinen keine UMP-Tests, da diese für alle Hypothesen der Form

$$H' : \vartheta = \vartheta_0 \quad \text{vs.} \quad K' : \vartheta = \vartheta_1$$

mit $\vartheta_0 \neq \vartheta_1$ optimal sein müssen. Für derartige Hypothesen ist gemäß Satz 6.14 der Neyman-Pearson-Test optimal, der im Fall $\vartheta_1 > \vartheta_0$ (und bei Existenz eines isotonen Dichtequotienten) die Form

$$\varphi(x) = 1_{\{T(x)>c\}} + \gamma(x) 1_{\{T(x)=c\}}$$

und im Fall $\vartheta_1 < \vartheta_0$ die Darstellung

$$\varphi'(x) = 1_{\{T(x)<c'\}} + \gamma'(x) 1_{\{T(x)=c'\}}$$

besitzt. Dies ergibt einen Widerspruch.

Satz 6.26. *Es sei $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ eine 1-parametrige Exponentialfamilie mit μ-Dichte*

$$p_\vartheta(x) = c(\vartheta) h(x) \exp\left(Q(\vartheta) T(x)\right),$$

wobei Q monoton wachsend ist. Dann existiert ein UMPU-Test für die Hypothesen

$$H : \vartheta \in [\vartheta_1, \vartheta_2] \quad \textit{vs.} \quad K : \vartheta \notin [\vartheta_1, \vartheta_2],$$

nämlich

$$\varphi^*(x) = \begin{cases} 1, & \textit{falls } T(x) \notin [c_1, c_2], \\ \gamma_i, & \textit{falls } T(x) = c_i, \\ 0, & \textit{falls } T(x) \in [c_1, c_2], \end{cases}$$

wobei die Konstanten c_i, γ_i gemäß

$$\beta_\varphi(\vartheta_1) = \beta_\varphi(\vartheta_2) = \alpha$$

bestimmt werden.

Beweis: Der Satz beruht auf einer verallgemeinerten Version des Neyman-Pearson-Lemmas für

$$H : \vartheta \in \{\vartheta_1, \ldots, \vartheta_k\} \quad \text{vs.} \quad K : \vartheta = \vartheta_{k+1}$$

und findet sich z.B. als Theorem 3.7.1 in Lehmann and Romano (2005). □

Bemerkung 6.27. Auch in mehrparametrigen Exponentialfamilien existieren ähnliche Resultate.

Kapitel 7

Asymptotische Eigenschaften von Tests

In diesem Abschnitt sei $X^{(n)} = (X_1, \ldots, X_n)^T$ für jedes $n \in \mathbb{N}$ ein Vektor von Zufallsvariablen, der einer Verteilung P aus der Klasse $\mathcal{P}^n = \{P_\vartheta^n \mid \vartheta \in \Theta\}$ folgt. Zudem sei

$$\varphi_n : \begin{cases} \mathcal{X}_n \to [0,1] \\ x^{(n)} \mapsto \varphi_n(x^{(n)}) \end{cases}$$

ein Test für

$$H : \vartheta \in \Theta_H \quad \text{vs.} \quad K : \vartheta \in \Theta_K.$$

Definition 7.1.

(i) Die Folge (φ_n) besitzt *asymptotisch das Niveau* α, falls

$$\limsup_{n\to\infty} \sup_{\vartheta\in\Theta_H} \beta_{\varphi_n}(\vartheta) \leq \alpha.$$

(ii) Die Folge (φ_n) heißt *konsistent für* H *gegen* K, falls

$$\lim_{n\to\infty} \beta_{\varphi_n}(\vartheta) = 1 \ \forall \vartheta \in \Theta_K.$$

Bemerkung 7.2. Beide Eigenschaften sind Minimalforderungen an eine Folge von Tests, um sicherzustellen, dass das Niveau (zumindest asymptotisch) eingehalten und die Wahrscheinlichkeit für einen Fehler 2. Art klein wird.

Beispiel 7.3. Es seien $X_1, \ldots, X_m$ i.i.d. $\sim N(\mu_1, \sigma^2)$ und $Y_1, \ldots, Y_n$ i.i.d. $\sim N(\mu_2, \tau^2)$ zwei unabhängige Stichproben. Unser Ziel ist,

$$H : \mu_1 \leq \mu_2 \quad \text{vs.} \quad K : \mu_1 > \mu_2$$

zu testen. Intuitiv lehnt man die Nullhypothese ab, falls der erste Stichprobenmittelwert deutlich größer als der zweite ist.

(i) Die beiden Varianzen σ^2 und τ^2 seien unbekannt, aber identisch. Beachte dazu: Es gilt nach Lemma 2.20

$$\overline{X}_m - \overline{Y}_n \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right)$$

und

$$\hat{\sigma}^2_{m,n} = \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(x_i-\overline{x}_m)^2+\sum_{j=1}^{n}(y_j-\overline{y}_n)^2\right) \sim \frac{\sigma^2}{m+n-2}\chi^2_{m+n-2}.$$

Im Grenzfall $\mu_1 = \mu_2$ folgt wieder mit Hilfe von Korollar 2.21, dass

$$T_{m,n} = \sqrt{\frac{mn}{m+n}}\frac{\overline{X}_m - \overline{Y}_n}{\hat{\sigma}_{m,n}} \sim t_{m+n-2},$$

und man erhält einen (UMPU-)Test zum Niveau α durch

$$\varphi_{m,n}(x) = 1_{\{T_{m,n} > t_{m+n-2,1-\alpha}\}}.$$

Dieser Test heißt *Zwei-Stichproben-t-Test.*

(ii) Im Fall $\sigma^2 \neq \tau^2$ gilt

$$\overline{X}_m - \overline{Y}_n \sim \mathcal{N}\left(\mu_1-\mu_2, \frac{\sigma^2}{m}+\frac{\tau^2}{n}\right).$$

Will man diese unbekannte Varianz schätzen, verwendet man

$$\hat{s}^2_{m,n} = \frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(x_i-\overline{x}_m)^2 + \frac{1}{n}\frac{1}{n-1}\sum_{j=1}^{n}(y_j-\overline{y}_n)^2.$$

Die exakte Verteilung der Statistik

$$T^*_{m,n} = \frac{\overline{X}_m - \overline{Y}_n}{\hat{s}_{m,n}}$$

unter H ist unbekannt ("Behrens-Fisher-Problem"), so dass man die Konvergenz in Verteilung

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_1-\mu_2)}{\hat{s}_{m,n}} \xrightarrow{\mathcal{L}} N(0,1)$$

aus dem Zentralen Grenzwertsatz verwendet und

$$\varphi^*_{m,n}(x) = 1_{\{T^*_{m,n} > u_{1-\alpha}\}}$$

definiert. Formal gilt, wenn $m \to \infty$, $n \to \infty$ und $n/m \to \lambda \in (0,\infty)$ konvergiert:

$$\beta_{\varphi^*_{m,n}}(\mu_1,\mu_2) = P_{\mu_1,\mu_2}\left(\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_1-\mu_2)}{\hat{s}_{m,n}} > -\frac{(\mu_1-\mu_2)}{\hat{s}_{m,n}} + u_{1-\alpha}\right)$$
$$\longrightarrow \begin{cases} \alpha, & \text{falls } \mu_1 = \mu_2, \\ 1, & \text{falls } \mu_1 > \mu_2, \\ 0, & \text{falls } \mu_1 < \mu_2, \end{cases}$$

also besitzt $\varphi^*_{m,n}$ unter diesen Voraussetzungen asymptotisch das Niveau α und ist konsistent.

Bemerkung 7.4. Ein allgemeines Prinzip zur Konstruktion von Tests für

$$H : \vartheta \in \Theta_H \quad \text{vs.} \quad K : \vartheta \in \Theta_K$$

ist die *Likelihood-Quotienten-* bzw. *LQ-Methode*. Bezeichnet $f_n(x^{(n)}, \vartheta)$ die Dichte von P_ϑ^n bzgl. einem gemeinsamen dominierenden Maß μ, so ist der *Likelihood-Quotient* durch

$$\lambda(x^{(n)}) = \frac{\sup_{\vartheta \in \Theta_H} f_n(x^{(n)}, \vartheta)}{\sup_{\vartheta \in \Theta} f_n(x^{(n)}, \vartheta)}$$

gegeben. Als *Likelihood-Quotienten-* bzw. *LQ-Test* wird

$$\varphi_n(x^{(n)}) = 1_{\{\lambda(x^{(n)}) < c\}}$$

bezeichnet, wobei c derart festgelegt wird, dass

$$\sup_{\vartheta \in \Theta_H} P_\vartheta \left(\lambda(X^{(n)}) < c \right) \leq \alpha$$

gilt. Da die Verteilung des Likelihood-Quotienten $\lambda(X^{(n)})$ im Allgemeinen nicht zugänglich ist, interessiert man sich für dessen Asymptotik.

Annahme 7.5. Wir verwenden die folgenden Annahmen:

(i) Es sei $\Theta \subset \mathbb{R}^d$ und es existieren $\Delta \subset \mathbb{R}^c$ offen und $h : \Delta \to \Theta$ mit $\Theta_H = h(\Delta)$, wobei h zweimal stetig differenzierbar sei und dessen Jacobi-Matrix stets vollen Rang besitze.

(ii) Es seien $X_1, \ldots, X_n$ i.i.d. P_ϑ, und sowohl in $\mathcal{P} = \{P_\vartheta \mid \vartheta \in \Theta\}$ als auch in der Teilfamilie $\mathcal{P}_h = \{P_{h(\eta)} \mid \eta \in \Delta\}$ gelten die Annahmen aus Satz 5.19 zur Existenz und asymptotischen Normalität der entsprechenden ML-Schätzer $\hat{\theta}_n$ bzw. $\hat{\eta}_n$.

Beispiel 7.6. Es seien $X_1, \ldots, X_m$ i.i.d. $\sim N(\mu_1, \sigma^2)$ und $Y_1, \ldots, Y_n$ i.i.d. $\sim N(\mu_2, \sigma^2)$ zwei unabhängige Stichproben. Ähnlich wie in Beispiel 7.3 interessieren wir uns für

$$H : \mu_1 = \mu_2 \quad \text{vs.} \quad K : \mu_1 \neq \mu_2.$$

Unter der Nullhypothese sind drei Parameter unbekannt, und wir setzen $\Theta = \mathbb{R}^2 \times \mathbb{R}^+$. Gilt die Alternative, so genügen zwei Parameter, also $\Delta = \mathbb{R} \times \mathbb{R}^+$. Die Abbildung

$$h : \begin{cases} \Delta \to \Theta \\ (\mu, \sigma^2)^T \mapsto (\mu, \mu, \sigma^2)^T, \end{cases}$$

parametrisiert die Nullhypothese und besitzt die Jacobi-Matrix

$$D_h = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix},$$

die vollen Rang hat. Dass die Bedingungen aus Satz 5.19 erfüllt sind, wissen wir bereits aus früheren Beispielen.

Satz 7.7. *Unter den Annahmen 7.5 gilt*

$$T_n = -2\log(\lambda(X^{(n)})) = 2\left(\log f_n(X^{(n)}, \hat{\theta}_n) - \log f_n(X^{(n)}, h(\hat{\eta}_n))\right) \xrightarrow{\mathcal{L}} \chi^2_{d-c},$$

falls $\vartheta \in \mathcal{P}_h$.

Beweis: Wir setzen wie zuvor

$$\ell(x, \vartheta) = \log f(x, \vartheta),$$

wobei f die Dichte eines Experiments bezeichnet, und betrachten zunächst

$$\begin{aligned}
T_n^{(1)} &= 2\left(\log f_n(X^{(n)}, \hat{\theta}_n) - \log f_n(X^{(n)}, \vartheta)\right) \\
&= 2\sum_{i=1}^n \left(\ell(X_i, \hat{\theta}_n) - \ell(X_i, \vartheta)\right) \\
&= 2(\hat{\theta}_n - \vartheta)^T \sum_{i=1}^n \dot{\ell}(X_i, \vartheta) + (\hat{\theta}_n - \vartheta)^T \sum_{i=1}^n \ddot{\ell}(X_i, \tilde{\vartheta}_n)(\hat{\theta}_n - \vartheta) \\
&= 2(\hat{\theta}_n - \vartheta)^T \big(\sum_{i=1}^n \dot{\ell}(X_i, \vartheta) + \sum_{i=1}^n \ddot{\ell}(X_i, \tilde{\vartheta}_n)(\hat{\theta}_n - \vartheta)\big) - (\hat{\theta}_n - \vartheta)^T \sum_{i=1}^n \ddot{\ell}(X_i, \tilde{\vartheta}_n)(\hat{\theta}_n - \vartheta),
\end{aligned}$$

wobei $\dot{\ell}$ und $\ddot{\ell}$ den jeweiligen Gradienten bzw. die jeweilige Hesse-Matrix bezeichnen und $\tilde{\vartheta}_n$ eine geeignete Zwischenstelle ist. Der erste Ausdruck lässt sich mit der Notation aus Satz 5.19 als

$$\begin{aligned}
&2n(\hat{\theta}_n - \vartheta)^T \big(\frac{1}{n}\sum_{i=1}^n \dot{\ell}(X_i, \vartheta) + \frac{1}{n}\sum_{i=1}^n \ddot{\ell}(X_i, \tilde{\vartheta}_n)(\hat{\theta}_n - \vartheta)\big) \\
=&2n(\hat{\theta}_n - \vartheta)^T \big(\dot{L}_n(\vartheta) + \ddot{L}_n(\tilde{\vartheta}_n)(\hat{\theta}_n - \vartheta)\big)
\end{aligned}$$

schreiben, und (5.3) zeigt, dass der Term verschwindet. Beachte: Gilt $X \sim \mathcal{N}_d(0, \Sigma)$ für ein $\Sigma > 0$, so folgt nach Definition der χ^2-Verteilung

$$X^T \Sigma^{-1} X \sim \chi^2_d.$$

Daher ergibt sich nach Satz 5.19 und wegen

$$\ddot{L}_n(\tilde{\vartheta}_n) \xrightarrow{Q_\vartheta} -I(f(\cdot, \vartheta))$$

aus dem Beweis von Satz 5.19 die Konvergenz in Verteilung

$$T_n^{(1)} = -\sqrt{n}(\hat{\theta}_n - \vartheta)^T \ddot{L}_n(\tilde{\vartheta}_n)\sqrt{n}(\hat{\theta}_n - \vartheta) \xrightarrow{\mathcal{L}} A \sim \chi^2_d.$$

Analog beweist man

$$T_n^{(2)} = 2\left(\log f_n(x^{(n)}, h(\hat{\eta}_n)) - \log f_n(x^{(n)}, h(\eta))\right) \xrightarrow{\mathcal{L}} B \sim \chi^2_c.$$

Ferner lässt sich zeigen, dass $A - B$ und B unabhängig sind. Gilt $\vartheta \in \mathcal{P}_h$, folgt dann

$$T_n = 2\left(\log f_n(x^{(n)}, \hat{\theta}_n) - \log f_n(x^{(n)}, h(\hat{\eta}_n))\right) \xrightarrow{\mathcal{L}} A - B.$$

Dass die Verteilung von $A - B$ die gewünschte Form hat, folgt nicht direkt aus der Definition der χ^2-Verteilung, jedoch zum Beispiel mit einem Argument über charakteristische Funktionen. □

Bemerkung 7.8.

(i) Satz 7.7 zeigt, dass der Text

$$\varphi_n(x^{(n)}) = \begin{cases} 1, & -2\log(\lambda(x^{(n)})) > \chi^2_{d-c,1-\alpha}, \\ 0, & -2\log(\lambda(x^{(n)})) \leq \chi^2_{d-c,1-\alpha}, \end{cases}$$

ein asymptotischer Niveau-α-Test für

$$H : \vartheta \in \Theta_H \quad \text{vs.} \quad K : \vartheta \in \Theta_K$$

ist.

(ii) Die Folge φ_n ist konsistent, wie man anhand von

$$-\frac{2}{n}\log(\lambda(X^{(n)})) = \frac{2}{n}\sum_{i=1}^{n}\Big(\ell(X_i,\hat{\theta}_n) - \ell(X_i,h(\hat{\eta}_n))\Big)$$
$$\xrightarrow{Q_\vartheta} \mathbb{E}_\vartheta[\log f(X_1,\vartheta) - \log f(X_1,h(\eta))] = KL(\vartheta|h(\eta)) > 0$$

erkennt, wobei $h(\eta)$ unter den Annahmen 7.5 existiert und unter der Alternative von ϑ verschieden ist. Also folgt

$$-2\log(\lambda(X^{(n)})) \xrightarrow{Q_\vartheta} \infty.$$

Beispiel 7.9. (Bartlett-Test) Es seien $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ unabhängig, $i = 1,\dots,r$ und $j = 1,\dots,n_i$ mit $n_i \to \infty$ für alle i, und wir möchten

$$H : \sigma_1^2 = \ldots = \sigma_r^2 \quad \text{vs.} \quad K : \sigma_i^2 \neq \sigma_j^2 \text{ für ein } i \neq j$$

testen.

In diesem Fall wählt man $\Theta = \mathbb{R}^r \times (\mathbb{R}^+)^r$ und $\Delta = \mathbb{R}^r \times \mathbb{R}^+$ sowie

$$h\big((x_1,\dots,x_r,y)^T\big) = (x_1,\dots,x_r,y,\dots,y)^T,$$

also $d = 2c$ und $c = r+1$. Als ML-Schätzer ergibt sich im allgemeinen Modell wie üblich $\hat{\theta}_n = (\hat{\mu}_1,\dots,\hat{\mu}_r,\hat{s}_1^2,\dots,\hat{s}_r^2)^T$ mit

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} = \overline{x_{i\cdot}} \quad \text{bzw.} \quad \hat{s}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_{i\cdot}})^2.$$

Man erhält

$$f_n(x^{(n)},\hat{\theta}_n) = \prod_{i=1}^{r}(2\pi\hat{s}_i^2)^{-n_i/2}\exp(-n_i/2).$$

Unter der Nullhypothese ergibt sich der ML-Schätzer $\hat{\eta}_n$ als Maximierer von

$$f_n(x^{(n)},\hat{\eta}_n) = \prod_{i=1}^{r}(2\pi\sigma^2)^{-n_i/2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2\Big).$$

Auch in diesem Modell erhält man zunächst $\hat{\mu}_i = \overline{x_{i\cdot}}$, danach mit $n = \sum_{i=1}^r n_i$ aber

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\overline{x_{i\cdot}})^2 = \sum_{i=1}^{r}\frac{n_i}{n}\hat{s}_i^2.$$

Hier betrachtet man also

$$f_n(x^{(n)}, \hat{\eta}_n) = (2\pi\hat{\sigma}^2)^{-n/2} \exp(-n/2)$$

Als Teststatistik ergibt sich demnach insgesamt

$$T_n = -2 \log \lambda(x^{(n)}) = n \log \hat{\sigma}^2 - \sum_{i=1}^{r} n_i \log(\hat{s}_i^2),$$

und der *Bartlett-Test* lautet

$$\varphi_n(x^{(n)}) = \begin{cases} 1, & T_n > \chi^2_{r-1,1-\alpha}, \\ 0, & T_n \leq \chi^2_{r-1,1-\alpha}. \end{cases}$$

Beispiel 7.10. (Unabhängigkeitstest in der Kontingenztafel) Gegeben seien zwei statistische Merkmale A und B (z.B. Geschlecht, Alter, Einkommen, Bildungsstand), wobei Faktor A aus r Kategorien und Faktor B aus s Kategorien besteht. Beobachtet werden n Personen, die bei beiden Merkmalen in jeweils eine der Kategorien fallen. Graphisch ergibt sich eine Tafel der Form

		Faktor B			
		1	...	s	Summe
Faktor A	1	X_{11}	...	X_{1s}	$X_{1\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	X_{r1}	...	X_{rs}	$X_{r\cdot}$
	Summe	$X_{\cdot 1}$	...	$X_{\cdot s}$	n

Dabei bezeichnen

$$X_{i\cdot} = \sum_{j=1}^{s} X_{ij} \quad \text{bzw.} \quad X_{\cdot j} = \sum_{i=1}^{r} X_{ij}$$

die Anzahl der Beobachtungen in Kategorie i bzw. j. Als Modell wählt man die Multinomialverteilung mit Parametern

$$(X_{11}, \ldots, X_{rs})^T \sim \mathcal{M}(n, p_{11}, \ldots, p_{rs}),$$

wobei $\sum_{i,j}^{r,s} p_{ij} = 1$. Damit ergibt sich als Wahrscheinlichkeitsverteilung

$$f(x^{(n)}, p) = P(X_{ij} = x_{ij}\ \forall i,j) = \frac{n!}{\prod_{i,j=1}^{r,s} x_{ij}!} \prod_{i,j=1}^{r,s} p_{ij}^{x_{ij}}, \quad x_{ij} \in \{0, \ldots, n\} \text{ mit } \sum_{i,j=1}^{r,s} x_{ij} = n.$$

Als ML-Schätzer in diesem allgemeinen Modell ergibt sich wie im Spezialfall der Binomialverteilung

$$\hat{p}_{ij} = \frac{x_{ij}}{n}.$$

und wir erhalten

$$f(x^{(n)}, \hat{p}) = \frac{n!}{\prod_{i,j=1}^{r,s} x_{ij}!} \prod_{i,j=1}^{r,s} \left(\frac{x_{ij}}{n}\right)^{x_{ij}}.$$

Wir interessieren uns für die Frage, ob die beiden Merkmale stochastisch unabhängig sind. In unserem Modell entspricht das

$$H : p_{ij} = p_i q_j\ \forall i,j \quad \text{vs.} \quad K : p_{ij} \neq p_i q_j \text{ für ein } i \neq j$$

mit

$$p_i = p_{i\cdot} = \sum_{j=1}^{s} p_{ij} \quad \text{bzw.} \quad q_j = p_{\cdot j} = \sum_{i=1}^{r} p_{ij}.$$

Es ergibt sich $d = rs-1$ und $c = r+s-2$, also $d-c = (r-1)(s-1)$. Im alternativen Modell gilt

$$f(x^{(n)}, p, q) = \frac{n!}{\prod_{i,j=1}^{r,s} x_{ij}!} \prod_{i,j=1}^{r,s} (p_i q_j)^{x_{ij}} = \frac{n!}{\prod_{i,j=1}^{r,s} x_{ij}!} \prod_{i=1}^{r} p_i^{x_{i\cdot}} \prod_{j=1}^{s} q_j^{x_{\cdot j}}.$$

Als Schätzer erhält man hier

$$\hat{p}_i = \frac{x_{i\cdot}}{n} \quad \text{bzw.} \quad \hat{q}_j = \frac{x_{\cdot j}}{n}$$

und also

$$f(x^{(n)}, \hat{p}, \hat{q}) = \frac{n!}{\prod_{i,j=1}^{r,s} x_{ij}!} \prod_{i,j=1}^{r,s} \left(\frac{x_{i\cdot} x_{\cdot j}}{n^2}\right)^{x_{ij}}.$$

Es ergibt sich

$$T_n = -2 \log \lambda(x^{(n)}) = 2 \sum_{i=1}^{r} \sum_{j=1}^{s} x_{ij} \log \left(\frac{x_{ij}}{\frac{x_{i\cdot} x_{\cdot j}}{n}} \right)$$

und der χ^2-*Unabhängigkeitstest* lautet

$$\varphi_n(x^{(n)}) = \begin{cases} 1, & T_n > \chi^2_{(r-1)(s-1),1-\alpha}, \\ 0, & T_n \leq \chi^2_{(r-1)(s-1),1-\alpha}. \end{cases}$$

Oft verwendet man die aufgrund einer Taylor-Entwicklung und des starken Gesetzes der großen Zahlen asymptotisch äquivalente Teststatistik

$$\tilde{T}_n = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(x_{ij} - \frac{x_{i\cdot} x_{\cdot j}}{n}\right)^2}{\frac{x_{i\cdot} x_{\cdot j}}{n}}.$$

Sowohl unter der Hypothese als auch unter der Alternative gilt mit

$$V_n = \sqrt{\frac{\tilde{T}_n}{n(\min(r,s)-1)}}$$

die Konvergenz

$$V_n^2 \xrightarrow{\mathbb{P}} \frac{1}{\min(r,s)-1} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}.$$

Man verwendet V_n daher als ein empirisches Maß für die Abhängigkeit der beiden Merkmale.

Im Zahlenbeispiel

		Jahreseinkommen				
		1	2	3	4	Summe
Anzahl der Kinder	0	2161	3577	2184	1636	9558
	1	2755	5081	2222	1052	11110
	2	936	1753	640	306	3635
	3	225	419	96	38	778
	4+	39	98	31	14	182
	Summe	6116	10928	5173	3046	25263

mit einer Einteilung des Einkommens pro Familie in vier Klassen ergibt sich

$$\tilde{T}_n = 568.566 \quad \text{und} \quad \chi^2_{12,0.95} = 21.026,$$

der Test wird zum Niveau 5% also verworfen. Jedoch ist die Abhängigkeit aufgrund von $V_n = 0.087$ eher schwach.

Kapitel 8

Das lineare Modell

In diesem Abschnitt betrachten wir eine spezielle Klasse von Modellen, in denen die Parameter linear auftauchen.

Beispiel 8.1. (Lineare Regression) Wir nehmen an, dass ein (physikalischer, ökonomischer oder biologischer) Zusammenhang zwischen zwei Größen x und y durch

$$y = b_0 + b_1 x \tag{8.1}$$

beschrieben werden kann. Dabei sind b_0 und b_1 unbekannte Parameter, die anhand von n Beobachtungen geschätzt werden sollen. Offenbar genügen bereits zwei Beobachtungen, wenn der Zusammenhang (8.1) exakt besteht. In den meisten Fällen werden jedoch Messfehler oder Modellungenauigkeiten dazu führen, dass nur

$$Y_i = b_0 + b_1 x_i + \varepsilon_i$$

beobachtet werden kann, wobei die ε_i den zufälligen Messfehler bezeichnen, der $\mathbb{E}[\varepsilon_i] = 0$ und $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ erfüllt. In Vektornotation gilt dann

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

bzw. kurz $Y = Xb + \varepsilon$. Typische statistische Fragestellungen betreffen neben der Schätzung von b_0 und b_1 etwa Tests auf $b_1 = 0$ (kein Einfluss von x auf y) oder auch auf $b_0 = 0$.

Beispiel 8.2. (Einfaktorielle Varianzanalyse) In einem Experiment soll der Einfluss verschiedener Futtersorten auf die Gewichtszunahme von Versuchstieren untersucht werden. Dazu werden n Tiere auf a Gruppen verteilt, wobei in jeder Gruppe n_i Tiere mit Futtersorte i gefüttert werden ($i = 1, \ldots, a$). Als Modell der *einfaktoriellen Varianzanalyse* wird

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, a, \ j = 1, \ldots, n_i$$

bzw. in Vektornotation

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{a1} \\ \vdots \\ Y_{an_a} \end{pmatrix} = \begin{pmatrix} \mathbb{1}_{n_1} & 0 & \dots & 0 \\ 0 & \mathbb{1}_{n_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbb{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_a \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{a1} \\ \vdots \\ \varepsilon_{an_a} \end{pmatrix}$$

verwendet, wobei $\mathbb{1}_m \in \mathbb{R}^m$ den m-dimensionalen Vektor bezeichnet, der nur aus Einsen besteht. Wichtige Fragen sind hier erneut die Schätzung der Größen μ_i oder ein Test auf $H : \mu_1 = \ldots = \mu_a$.

Definition 8.3. Es seien $X \in \mathbb{R}^{n\times k}$ und $b \in \mathbb{R}^k$ mit $n \geq k$ sowie eine n-dimensionale Zufallsvariable $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ gegeben.

(i) Gilt $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$, so heißt $Y = Xb + \varepsilon$ *lineares Modell mit Normalverteilungsannahme* bzw. *LMN*.

(ii) Es sei $Z \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$. Gilt für alle $i_j \in \{1, \ldots, n\}$

$$\mathbb{E}[\varepsilon_{i_1}\varepsilon_{i_2}\varepsilon_{i_3}\varepsilon_{i_4}] = \mathbb{E}[Z_{i_1}Z_{i_2}Z_{i_3}Z_{i_4}],$$

so heißt $Y = Xb + \varepsilon$ *lineares Modell mit Momentenannahme* bzw. *LMM*.

(iii) Die Matrix X heißt in beiden Modellen *Designmatrix*.

Bemerkung 8.4. Bezeichnen r den Rang der Matrix X und

$$R(X) = \{Xb \mid b \in \mathbb{R}^k\} \subset \mathbb{R}^n$$

das r-dimensionale Bild von X, so ist eine naheliegende Schätzung für Xb die orthogonale Projektion PY von Y auf $R(X)$. Jeder Wert $\hat{b}$ mit $PY = X\hat{b}$ ist dann ein sinnvoller Schätzer für b.

Definition 8.5. Für eine Matrix $A \in \mathbb{R}^{m\times n}$ heißt $G \in \mathbb{R}^{n\times m}$ eine *verallgemeinerte Inverse*, falls

$$AGA = A$$

gilt. Wir bezeichnen mit

$$A^- = \{G \in \mathbb{R}^{n\times n} \mid AGA = A\}$$

die Menge aller verallgemeinerten Inversen von A.

Bemerkung 8.6. Wir schreiben in Formeln A^- statt G, falls die Gültigkeit der Formel nicht von der speziellen Wahl der verallgemeinerten Inversen abhängt. Zum Beispiel gilt also

$$AA^-A = A.$$

Beispiel 8.7. Zu

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

sind verallgemeinerte Inversen zum Beispiel durch

$$G_1 = \frac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{und} \quad G_2 = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

gegeben.

Lemma 8.8. (Range inclusion) *Es seien $X \in \mathbb{R}^{n\times k}$ und $V \in \mathbb{R}^{n\times s}$. Dann gilt:*

(i) $R(X) \subset R(V) \quad \Longleftrightarrow \quad VV^-X = X.$

(ii) Gelten eine der beiden Seiten aus (i) und $V \geq 0$ (also $s = n$), so folgt

(a) $X^TV^-X \geq 0$,
(b) $R(X^T) = R(X^TV^-X)$.

Beweis:

(i) Die Rückrichtung ist klar, denn jedes $z = Xy$ lässt sich offenbar als $z = Vw$ darstellen. Daher sei $R(X) \subset R(V)$, also $X = VW$ für eine geeignete Matrix W. Dann gilt für jede verallgemeinerte Inverse G von V

$$VGX = VGVW = VW = X.$$

(ii) (a) Es sei G eine verallgemeinerte Inverse von V. Dann gilt mit $W = GX$ aufgrund von (i) auch $X = VW$, und wegen der Symmetrie von V folgt

$$X^TGX = W^TV^TGVW = W^TVW \geq 0.$$

(b) Erinnerungen:

- Da V symmetrisch ist, besitzt V reelle Eigenwerte. Insbesondere sind diese nicht-negativ, wenn $V \geq 0$ gilt. Daher

$$V = \sum_{j=1}^{n} \lambda_j z_j z_j^T,$$

wobei die z_j orthonormal und die $\lambda_j \geq 0$ sind. Insbesondere existiert

$$V^\alpha = \sum_{j=1}^{n} \lambda_j^\alpha z_j z_j^T,$$

und es gilt $V^\alpha V^\beta = V^{\alpha+\beta}$, $\alpha, \beta \geq 0$.

- Für reelle Matrizen A und B und ihren Rang r gilt: $r(A) = r(A^TA)$ und $r(A \cdot B) \leq \min\{r(A), r(B)\}$.

Insbesondere folgt $r(VW) = r(W^TVW)$ wegen

$$r(VW) = r(V^{1/2}V^{1/2}W) \leq r(V^{1/2}W) = r(W^TVW) \leq \min\{r(W^T), r(VW)\}.$$

Verwendet man (a), ergibt sich

$$r(X^TV^-X) = r(W^TVW) = r(VW) = r(VV^-X) = r(X).$$

Dann gilt auch $r(X^T) = r(X^TV^-X)$, und die Aussage folgt aufgrund von $R(X^TV^-X) \subset R(X^T)$. □

Satz 8.9. *Im linearen Modell $Y = Xb + \varepsilon$ sind die orthogonalen Projektionen auf den Untervektorraum $R(X)$ bzw. auf dessen orthogonales Komplement*

$$R(X)^{\perp} = \{z \in \mathbb{R}^n \mid z^T X = 0\}$$

durch die Matrizen

$$P = X(X^T X)^- X^T \quad \textit{und} \quad R = \mathbb{I}_n - X(X^T X)^- X^T$$

gegeben.

Beweis: Wir beweisen die Aussage nur für P und verwenden, dass P genau dann eine orthogonale Projektion ist, wenn P idempotent (d.h. $P^2 = P$) ist und $P^T = P$ gilt. Dies ist in diesem Fall gegeben, weil

$$(X^T X)(X^T X)^- X^T = X^T \tag{8.2}$$

wegen Lemma 8.8 (i) und (ii) (b) gilt und daher

$$P^2 = X(X^T X)^- X^T X(X^T X)^- X^T = X(X^T X)^- X^T = P$$

folgt. $P^T = P$ folgt direkt aus Lemma 8.8 (ii) (a).

Zu klären ist noch, dass P auf den richtigen Unterraum abbildet. Dazu sei $z = Xb$. Dann gilt wieder wegen (8.2)

$$P_0 z = P_0 Xb = X(X^T X)^- X^T Xb = Xb = z.$$

Zudem bildet P wegen $P = XA$ mit $A = (X^T X)^- X^T$ auf $R(X)$ ab. □

Bemerkung 8.10.

(i) Naheliegende Schätzer für Xb und σ^2 im linearen Modell erhält man gemäß Satz 8.9 durch

$$X\hat{b} = PY = X(X^T X)^- X^T Y \tag{8.3}$$

bzw.

$$\hat{\sigma}^2 = \frac{||Y - X\hat{b}||_2^2}{n-r} = \frac{||RY||_2^2}{n-r} = \frac{Y^T RY}{n-r},$$

wobei $r = r(X)$ ist und die letzte Identität aus der Idempotenz von R folgt. Die Wahl des Nenners erklärt sich durch Korollar 8.13.

(ii) Schätzer $\hat{b}$ für b sind im Allgemeinen ($r < k$) nicht eindeutig bestimmt. Sofern $X^T X$ invertierbar ist, ergibt sich aus (8.3) jedoch

$$X^T X\hat{b} = X^T Y \quad \text{bzw.} \quad \hat{b} = (X^T X)^{-1} X^T Y,$$

und der Schätzer ist wegen $\mathbb{E}[Y] = Xb$ sogar erwartungstreu.

Beispiel 8.11. Es sei $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, also

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

mit $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Wegen $X^T X = n$ ist

$$P = \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & \dots & 1 \end{pmatrix},$$

also

$$PY = \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \sum_{j=1}^{n} Y_j = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \overline{Y}_n \quad \text{ein Schätzer für} \quad X\mu = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu.$$

Insbesondere verwendet man $\overline{Y}_n$ als Schätzer für μ. Zudem gilt

$$\hat{\sigma}^2 = \frac{||Y - X\hat{b}||_2^2}{n-1} = \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \overline{Y}_n)^2.$$

Lemma 8.12. *Es sei Y eine n-dimensionale Zufallsvariable mit $\mathbb{E}[Y] = \mu$ und $\mathrm{Var}(Y) = V \geq 0$. Ferner seien Matrizen $A, B \in \mathbb{R}^{n \times n}$ gegeben.*

(i) Es gilt: $\mathbb{E}[Y^T AY] = \mu^T A\mu + Spur(AV)$.

(ii) Besitzt Y Momente bis zur vierten Ordnung wie eine $\mathcal{N}(\mu, V)$-Verteilung, so gilt zusätzlich:

(a) $\mathrm{Cov}(Y, Y^T AY) = 2VA\mu$.

(b) $\mathrm{Cov}(Y^T AY, Y^T BY) = 2Spur(AVBV)$, sofern $\mu = 0$.

Beweis: Diese Eigenschaften lassen sich direkt nachrechnen. □

Korollar 8.13. *Im linearen Modell mit Momentenannahme gilt: $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.*

Beweis: Nach Definition und wegen Lemma 8.12 erhält man

$$\mathbb{E}[\hat{\sigma}^2] = \frac{\mathbb{E}[Y^T RY]}{n-r} = \frac{1}{n-r}\big(\mu^T R\mu + \mathrm{Spur}(\sigma^2 R\mathbb{I}_n)\big) = \frac{1}{n-r}\big(\mu^T R\mu + \sigma^2 \mathrm{Spur}(R)\big)$$

mit $\mu = Xb$. Offenbar gilt wieder nach Lemma 8.8

$$R\mu = Xb - X(X^T X)^- X^T Xb = 0.$$

Zudem gilt für jede orthogonale Projektion Q aufgrund der Idempotenz, dass alle Eigenwerte entweder 0 oder 1 sind. Insbesondere entspricht die Anzahl der Einsen dem Rang der Matrix. Es folgt

$$\mathrm{Spur}(Q) = r(Q) = n - r$$

für $Q = R$. □

Satz 8.14. (Gauß-Markov) *Im linearen Modell mit Momentenannahme und $r(X) = k$ gilt:*

(i) $\hat{b}$ und $\hat{\sigma}^2$ sind erwartungstreu für b und σ^2 und unkorreliert.

(ii) $\hat{b}$ *ist der beste lineare erwartungstreue Schätzer für* b*, d.h. für alle* $\tilde{b} = LY$ *mit* $\mathbb{E}[\tilde{b}] = b$ *gilt*

$$\mathrm{Var}(\tilde{b}) \geq \mathrm{Var}(\hat{b}) = \sigma^2 (X^T X)^{-1}$$

im Sinne der Löwner-Ordnung.

(iii) $\hat{\sigma}^2$ *ist der beste quadratische erwartungstreue Schätzer für* σ^2*, d.h. für alle* $\tilde{\sigma}^2 = Y^T A Y$ *mit* $\mathbb{E}[\tilde{\sigma}^2] = \sigma^2$ *gilt*

$$\mathrm{Var}(\tilde{\sigma}^2) \geq \mathrm{Var}(\hat{\sigma}^2).$$

Beweis:

(i) Die Erwartungstreue folgt gemäß Bemerkung 8.10 (ii) bzw. Korollar 8.13. Zudem gilt

$$\mathrm{Cov}(\hat{b}, \hat{\sigma}^2) = \frac{1}{n-k} (X^T X)^{-1} X^T \,\mathrm{Cov}(Y, Y^T R Y) = \frac{2\sigma^2}{n-k} (X^T X)^{-1} X^T R X b$$

nach Lemma 8.12 (ii) (a). Die Aussage folgt wieder mit $RXb = 0$.

(ii) Offenbar gilt

$$b = \mathbb{E}[\tilde{b}] = L\mathbb{E}[Y] = LXb$$

für alle b, so dass L ein Linksinverses zu X ist. Dann ergibt sich

$$\begin{aligned} 0 &\leq \big((X^T X)^{-1} X^T - L\big)\big((X^T X)^{-1} X^T - L\big)^T \\ &= (X^T X)^{-1} - (X^T X)^{-1} X^T L^T - LX(X^T X)^{-1} + LL^T \\ &= LL^T - (X^T X)^{-1}. \end{aligned}$$

Zuletzt gilt

$$\mathrm{Var}(\tilde{b}) = \mathrm{Var}(LY) = L\,\mathrm{Var}(Y) L^T = \sigma^2 LL^T \geq \sigma^2 (X^T X)^{-1} = \mathrm{Var}(\hat{b}),$$

wie man direkt nachrechnet.

(iii) Diese Aussage lässt sich ähnlich zeigen wie (ii), wobei man Lemma 8.12 (ii) (b) verwendet. □

Lemma 8.15. *Es seien* $Y \sim \mathcal{N}(0, V)$ *sowie* $A \in \mathbb{R}^{p \times n}$ *und* $B \in \mathbb{R}^{q \times n}$*. Dann gilt:*

(i) AY *und* BY *sind unabhängig, falls* $AVB^T = 0$*.*

(ii) Ist $q = n$ *und* B *eine orthogonale Projektion, so impliziert* $AVB = 0$ *die Unabhängigkeit von* $Y^T A Y$ *und* BY*.*

Beweis: (i) folgt direkt aus Eigenschaften der Normalverteilung, während sich (ii) wie im Beweis von Lemma 2.20 ergibt. □

Satz 8.16. *Im linearen Modell mit Normalverteilungsannahme und* $r(X) = k$ *gilt:* $(\hat{b}, \hat{\sigma}^2)^T$ *ist der gleichmäßig beste erwartungstreue Schätzer für* $(b, \sigma^2)^T$*, und beide Komponenten sind unabhängig.*

Beweis: $Y \sim \mathcal{N}(Xb, \sigma^2 \mathbb{I}_n)$ besitzt eine Dichte der Form

$$f(y) = c(\sigma^2) \exp\left(-\frac{1}{2\sigma^2}||y - Xb||_2^2\right) = \tilde{c}(\sigma^2, b) \exp\left(-\frac{1}{2\sigma^2}(y^T y - 2b^T X^T y)\right).$$

Es handelt sich also um eine $(k+1)$-dimensionale Exponentialfamilie, und gemäß Satz 4.22 und Satz 4.25 sind $y^t y$ und $X^t y$ suffizient und vollständig für $(b, \sigma^2)^T$. Dies überträgt sich nach Bemerkung 4.31 auf $(\hat{b}, \hat{\sigma}^2)^T$. Satz 4.28 liefert dann die UMVU-Eigenschaft, und Unabhängigkeit folgt gemäß Lemma 8.15 (ii) mit $A = R$ und $B = P$. □

Beispiel 8.17. Wir betrachten wieder die lineare Regression aus Beispiel 8.1, d.h.

$$Y_i = b_0 + b_1 x_i + \varepsilon_i$$

mit $\mathbb{E}[\varepsilon_i] = 0$ und $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ bzw.

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Setzt man

$$\overline{x}_n = \frac{1}{n}\sum_{j=1}^n x_j \quad \text{und} \quad \overline{x^2}_n = \frac{1}{n}\sum_{j=1}^n x_j^2,$$

so ergibt sich

$$X^T X = n \begin{pmatrix} 1 & \overline{x}_n \\ \overline{x}_n & \overline{x^2}_n \end{pmatrix} \quad \text{bzw.} \quad (X^T X)^{-1} = \frac{1}{\sum_{j=1}^n (x_j - \overline{x}_n)^2} \begin{pmatrix} \overline{x^2}_n & -\overline{x}_n \\ -\overline{x}_n & 1 \end{pmatrix},$$

sofern nicht alle Einträge der Matrix dieselben sind. (In diesem Fall wären b_0 und b_1 auch ohne Fehlerterme ε nicht zu bestimmen.) Außerdem gilt

$$X^T Y = \begin{pmatrix} \sum_{j=1}^n Y_j \\ \sum_{j=1}^n x_j Y_j \end{pmatrix}.$$

Man erhält

$$\hat{b} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix} \text{ mit } \hat{b}_0 = \overline{Y}_n - \hat{b}_1 \overline{x}_n \text{ und } \hat{b}_1 = \frac{\sum_{j=1}^n (x_j - \overline{x}_n)(Y_j - \overline{Y}_n)}{\sum_{j=1}^n (x_j - \overline{x}_n)^2}$$

als besten linearen erwartungstreuen Schätzer für b. Das allgemeine Prinzip zum Erhalt dieser Schätzer wird auch als Methode der kleinsten Quadrate bezeichnet. Zudem ergibt sich

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{j=1}^n (Y_j - \hat{b}_0 - \hat{b}_1 x_j)^2.$$

Bemerkung 8.18. Oft ist man nicht an b, sondern an $K^T b$ für ein $K \in \mathbb{R}^{k \times s}$ interessiert. Falls $r(X) = k$ ist, ist eine naheliegende Schätzung durch

$$K^T \hat{b} = K^T (X^T X)^{-1} X^T Y$$

gegeben. Gilt $r(X) = k$ nicht, aber $R(K) \subset R(X^T)$, so ist sogar

$$K^T\hat{b} = K^T(X^TX)^-X^TY$$

gemäß Lemma 8.8 (i) und (ii) (b) eindeutig bestimmt. In diesem Fall gilt

$$\mathbb{E}[K^T\hat{b}] = K^T(X^TX)^-X^TXb = K^Tb.$$

Definition 8.19. Für $K \in \mathbb{R}^{k\times s}$ mit $r(K) = s$ heißt K^Tb *schätzbar*, falls $R(K) \subset R(X^T)$ gilt.

Beispiel 8.20. Möchte man

$$H_0 : K^Tb = 0 \quad \text{vs.} \quad H_1 : K^Tb \neq 0$$

testen, so soll nicht zugleich $Xb_1 = Xb_2$ und $K^Tb_1 \neq K^Tb_2$ gelten können, da b nur über Xb in das Modell $Y = Xb + \varepsilon$ eingeht. Man fordert also

$$Xb_1 = Xb_2 \quad \Longrightarrow \quad K^Tb_1 = K^Tb_2.$$

Setzt man

$$N(A) = \{y \mid Ay = 0\},$$

so bedeutet dies $N(X) \subset N(K^T)$ bzw. $R(X^T)^\perp \subset R(K)^\perp$ bzw. $R(K) \subset R(X^T)$.

Satz 8.21. *Es sei K^Tb schätzbar. Dann gilt:*

(i) Im linearen Modell mit Momentenannahme ist $K^T\hat{b}$ der beste lineare erwartungstreue Schätzer für K^Tb, und es gilt

$$\mathrm{Var}(K^T\hat{b}) = \sigma^2K^T(X^TX)^-K \in \mathbb{R}^{s\times s}.$$

(ii) Im linearen Modell mit Normalverteilungsannahme ist $K^T\hat{b}$ ist der gleichmäßig beste erwartungstreue Schätzer für K^Tb.

Beweis: Dieser Satz lässt sich genau wie Satz 8.14 (ii) bzw. wie Satz 8.16 beweisen. □

Beispiel 8.22.

(i) Wir betrachten wieder die lineare Regression aus Beispiel 8.1, d.h.

$$Y_i = b_0 + b_1x_i + \varepsilon_i$$

mit $\mathbb{E}[\varepsilon_i] = 0$ und $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ bzw.

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Möchte man

$$H_0 : b_0 = 0 \quad \text{vs.} \quad H_1 : b_0 \neq 0$$

testen, so wählt man $K = \begin{pmatrix}1 & 0\end{pmatrix}^T$ und zum Beispiel

$$X = \begin{pmatrix}1 & 0\\ \vdots & \vdots\\ 1 & 0\end{pmatrix}.$$

Offensichtlich gilt $R(K) \subset R(X^T)$. Wir erhalten außerdem

$$X^TX = \begin{pmatrix}n & 0\\ 0 & 0\end{pmatrix} \quad \text{mit einer verallgemeinerten Inversen} \quad G = \frac{1}{n}\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}.$$

Es ergibt sich

$$K^T\hat{b} = K^TGX^TY = \frac{1}{n}\sum_{j=1}^{n} Y_j = \overline{Y}_n.$$

(ii) Interessiert man sich für die einfaktorielle Varianzanalyse aus Beispiel 8.2 (mit $a = 3$), also

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, 3,\ j = 1, \dots, n_i$$

bzw. in Vektornotation

$$\begin{pmatrix}Y_{11}\\ \vdots\\ Y_{1n_1}\\ Y_{21}\\ \vdots\\ Y_{2n_2}\\ Y_{31}\\ \vdots\\ Y_{3n_3}\end{pmatrix} = \begin{pmatrix}\mathbb{I}_{n_1} & 0 & 0\\ 0 & \mathbb{I}_{n_2} & 0\\ 0 & 0 & \mathbb{I}_{n_3}\end{pmatrix}\begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\end{pmatrix} + \begin{pmatrix}\varepsilon_{11}\\ \vdots\\ \varepsilon_{1n_1}\\ \varepsilon_{21}\\ \vdots\\ \varepsilon_{2n_2}\\ \varepsilon_{31}\\ \vdots\\ \varepsilon_{3n_3}\end{pmatrix},$$

so wählt man zum Beispiel

$$K^T = \begin{pmatrix}1 & -1 & 0\\ 0 & 1 & -1\end{pmatrix},$$

wenn man

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j \text{ für ein } i \neq j$$

testen möchte.

Bemerkung 8.23. Wir betrachten allgemein

$$H_0 : K^Tb = 0 \quad \text{vs.} \quad H_1 : K^Tb \neq 0$$

mit $R(K) \subset R(X^T)$.

Die wesentliche Größe bei der Konstruktion eines entsprechenden Tests ist der Abstand von Y zum Hypothesenraum

$$L_{H_0} = \{Xb \mid K^Tb = 0, b \in \mathbb{R}^k\} \subset R(X).$$

Es lässt sich nachrechnen, dass die orthogonale Projektion auf L_{H_0} durch $P_{H_0} = P_0 - P_1$ mit

$$P_0 = X(X^TX)^-X^T \quad \text{und} \quad P_1 = X(X^TX)^-K(K^T(X^TX)^-K)^-K^T(X^TX)^-X^T$$

gegeben ist, wobei P_1 ebenfalls eine orthogonale Projektion ist. Eine sinnvolle Testgröße ist dann der Abstand der Projektionen von Y auf $R(X)$ und L_{H_0}. Nach dem Satz des Pythagoras betrachtet man also zum Beispiel

$$||(\mathbb{I}_n - P_{H_0})Y||_2^2 - ||(\mathbb{I}_n - P_0)Y||_2^2 = ||P_1Y||_2^2 = Y^TP_1Y,$$

wobei sich die letzte Identität ergibt, da P_1 idempotent ist.

Satz 8.24. *Es seien $Y \sim \mathcal{N}(\mu, \sigma^2\mathbb{I}_n)$ und $P \in \mathbb{R}^{n\times n}$ mit $P^T = P$. Dann gilt: P ist eine orthogonale Projektion genau dann, wenn*

$$Q = \frac{(Y-\mu)^TP(Y-\mu)}{\sigma^2} \sim \chi^2_{r(P)}.$$

Beweis: Ohne Einschränkung sei $\mu = 0$.

$\Longrightarrow$ Gilt $P^2 = P$, so existiert $A \in \mathbb{R}^{n\times n}$ mit $A^TA = AA^T = \mathbb{I}_n$ und

$$A^TPA = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix} \subset \mathbb{R}^{n\times n},\ r = r(P).$$

Also gilt $Z = A^TY \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$ und

$$Q = \frac{1}{\sigma^2}Z^TA^TPAZ = \frac{1}{\sigma^2}\sum_{j=1}^r Z_j^2 = \sum_{j=1}^r \Big(\frac{Z_j}{\sigma}\Big)^2 \sim \chi^2_r.$$

$\Longleftarrow$ Wegen $P^T = P$ existiert B mit $B^TB = BB^T = \mathbb{I}_n$ und

$$B^TPB = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),$$

wobei die λ_i die reellen Eigenwerte von P sind. Setzt man $X = B^TY \sim \mathcal{N}(0, \sigma^2\mathbb{I}_n)$, so ergibt sich diesmal

$$Q = \frac{1}{\sigma^2}X^TB^TPBX.$$

Wegen $Q \sim \chi^2_r$ gilt dann zugleich

$$\mathbb{E}[\exp(itQ)] = (1-2it)^{-r/2}$$

und

$$\mathbb{E}[\exp(itQ)] = \mathbb{E}[\exp(i(t/\sigma^2)X^TB^TPBX)] = \mathbb{E}\Big[\exp\Big(it\sum_{j=1}^n \lambda_j\Big(\frac{X_j}{\sigma}\Big)^2\Big)\Big]$$

$$= \prod_{j=1}^n \mathbb{E}\Big[\exp\Big(it\lambda_j\Big(\frac{X_j}{\sigma}\Big)^2\Big)\Big] = \prod_{j=1}^n (1-2i\lambda_jt).$$

Da Polynome durch ihre Linearfaktoren eindeutig bestimmt sind, folgt (passend sortiert) $\lambda_1 = \ldots = \lambda_r = 1$ und $\lambda_j = 0$ für $j > r$. Insbesondere ergibt sich

$$P^2 = B\Lambda B^TB\Lambda B^T = B\Lambda^2B^T = B\Lambda B^T = P.$$

□

Bemerkung 8.25. Ist $Y \sim \mathcal{N}(\mu, \sigma^2 \mathbb{I}_n)$ und $P \in \mathbb{R}^{n\times n}$ eine orthogonale Projektion mit $r = r(P) = \text{Spur}(P)$, so folgt mit

$$A^T P A = \begin{pmatrix} \mathbb{I}_r & 0 \\ 0 & 0 \end{pmatrix}$$

wie im Beweis von Satz 8.24, dass $Z = A^T Y \sim \mathcal{N}(A^T\mu, \sigma^2\mathbb{I}_n)$ gilt. Insbesondere ergibt sich

$$Q = \frac{1}{\sigma^2} Y^T P Y = \frac{1}{\sigma^2} Z^T A^T P A Z = \sum_{j=1}^{r} \left(\frac{Z_j}{\sigma}\right)^2.$$

Mit einem Argument über charakteristische Funktionen kann man zeigen: Die Verteilung von Q hängt nur von r und

$$\delta^2 = \frac{1}{\sigma^2} \sum_{j=1}^{r} ((A^T\mu)_j)^2 = \frac{\mu^T P \mu}{\sigma^2}$$

ab, wobei $(A^T\mu)_j$ die j-te Zeile von $A^T\mu$ bezeichnet. Sie heißt χ^2-*Verteilung mit r Freiheitsgraden und Nichtzentralitätsparameter δ^2*. Notation: $Q \sim \chi^2_{r,\delta^2}$.

Definition 8.26. Es seien $X \sim \chi^2_m$ und $Y \sim \chi^2_n$ unabhängig.

(i) Die Verteilung von

$$F = \frac{\frac{1}{m}X}{\frac{1}{n}Y}$$

heißt *F-Verteilung mit m und n Freiheitsgraden*. Als Notation verwenden wir $F \sim F_{m,n}$.

(ii) Ist $X \sim \chi^2_{m,\delta^2}$, so besitzt

$$F = \frac{\frac{1}{m}X}{\frac{1}{n}Y}$$

ein *F-Verteilung mit m und n Freiheitsgraden und Nichtzentralitätsparameter δ^2*. Notation: $F \sim F_{m,n,\delta^2}$.

Satz 8.27. (F-Test im Modell mit Normalverteilungsannahme) *Im Modell mit Normalverteilungsannahme sei $R(K) \subset R(X^T)$ sowie $t = r(K)$ und $r = r(X)$.*

(i) Es gilt:

$$F = \frac{\frac{1}{t}||P_1 Y||_2^2}{\frac{1}{n-r}||RY||_2^2} \sim F_{t,n-r,\delta^2}, \quad \delta^2 = \frac{1}{\sigma^2}(K^T b)^T \left(K^T (X^T X)^- K\right)^- K^T b.$$

Dabei haben wir wieder die Notation $R = \mathbb{I}_n - P_0$ verwendet.

(ii) Der durch

$$\varphi(y) = \begin{cases} 1, & \text{falls } F > F_{t,n-r,1-\alpha}, \\ 0, & \text{falls } F \le F_{t,n-r,1-\alpha}, \end{cases}$$

gegebene F-Test für

$$H_0 : K^T b = 0 \quad \text{vs.} \quad H_1 : K^T b \neq 0$$

besitzt das Niveau α.

Beweis: Es genügt offenbar, Aussage (i) zu beweisen. Nach Bemerkung 8.25 gilt

$$\frac{1}{\sigma^2}||P_1 Y||_2^2 = \frac{1}{\sigma^2} Y^T P_1 Y \sim \chi^2_{t,\delta^2}$$

mit

$$\delta^2 = \frac{1}{\sigma^2}(Xb)^T P_1 Xb = \frac{1}{\sigma^2}(K^T b)^T \left(K^T (X^T X)^- K\right)^- K^T b$$

nach Definition von P_1. Analog lässt sich

$$\frac{1}{\sigma^2}||RY||_2^2 \sim \chi^2_{n-r}$$

zeigen, wobei man wegen $RXb = 0$ eine zentrale χ^2-Verteilung erhält. Wegen $P_1 R = 0$ und Lemma 8.15 sind die beiden Zufallsvariablen unabhängig, und die Aussage folgt nach Definition der F-Verteilung. □ □

Literaturverzeichnis

Bauer, H. (1992). *Maß- und Integrationstheorie*. Walter de Gruyter & Co., Berlin.

Bickel, P. J. and K. A. Doksum (2001). *Mathematical statistics*. Holden-Day, Inc., San Francisco, Calif.-Düsseldorf-Johannesburg.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses*. Springer, New York.

Tripathi, G. (1999). A matrix extension of the Cauchy-Schwarz inequality. *Econom. Lett. 63*(1), 1–3.

Witting, H. and U. Müller-Funk (1995). *Mathematische Statistik. II*. B. G. Teubner, Stuttgart.