Mathematische Statistik

Prof. Dr. Friedrich Götze
mitgeschrieben von Arthur Sinulis

12. April 2014

Inhaltsverzeichnis

0 Grundlagen

Definition 0.1:

(*i*) Ein statistisches Experiment ist eine Familie $\mathcal{P}$ von Wahrscheinlichkeitsmaßen über einem Maßraum $(\Omega, \mathcal{A})$.

(*ii*) Ein Parameter(vektor) ist eine Abbildung $\tau : \mathcal{P} \to \mathbb{R}^k, k \geq 1$.

(*iii*) Eine Familie $\mathcal{P}$ wird parametrisch genannt, falls es einen Parameter $\vartheta : \mathcal{P} \to \mathbb{R}^k$ gibt und ϑ injektiv ist (d.h. $\vartheta : \mathcal{P} \to \vartheta(\mathcal{P})$ ist Bijektion). Im Allgemeinen wird Stetigkeit & Differenzierbarkeit gefordert.
ϑ heißt Parametrisierung von $\mathcal{P}$ und $\Theta \coloneqq \vartheta(\mathcal{P}) \subset \mathbb{R}^k$ heißt Parameterraum. Θ kann auch eine endliche Menge sein (Graphen sind auch Parameter!). In diesem Fall ist

$$\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}.$$

ϑ identifiziert $P \in \mathcal{P}$.

(*iv*) Falls k nicht endlich ist, dann heißt $\mathcal{P}$ nichtparametrische Familie, d.h. hier ist $k = \infty$.

(*v*) Ein semiparametrisches statistisches Modell ist eine nichtparametrische Familie $\mathcal{P}$ von Wahrscheinlichkeitsmaßen zusammen mit einem Parametervektor $\tau : \mathcal{P} \to \mathbb{R}^k$, wobei jedoch $\tau^{-1}(\vartheta)$ unendlichdimensional ist.

Beispiel 0.2:

(*i*) Sei $\mathcal{P} = \{Q \text{ W-Maß auf } (\mathbb{R}, \mathcal{B}^1) : Q \ll \lambda\}$, d.h. zu $Q \in \mathcal{P}$ existiert wegen dem Satz von Radon-Nikodym eine Dichte $f_Q \in L^1(\mathbb{R}, \mathcal{B}^1, \lambda)$ mit

$$Q(A) = \int_A f_Q(x)\lambda(dx) \text{ für alle } A \in \mathcal{B}^1$$

mit

$$\tau : \mathcal{P} \to L^1(\mathbb{R}, \mathcal{B}^1, \lambda), Q \to f_Q$$

ist eine nichtparametrische Familie.

(*ii*) $\mathcal{P}$ wie oben. Setze

$$\mathcal{P}_0 \coloneqq \{Q \in \mathcal{P} : \int |x|\, f_Q(x)\lambda(dx) < \infty\}$$

und

$$\tau : \mathcal{P}_0 \to \mathbb{R} \text{ mit } \tau(Q) \coloneqq \int xQ(dx), Q \in \mathcal{P}_0$$

ist der Mittelwertparameter von Q. Dies definiert ein semiparametrisches Modell.

(*iii*) $\mathcal{P} = \{Q \text{ W-Maß mit } Q(A) = \int_A \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \lambda(dx), \sigma > 0, \mu \in \mathbb{R}\}$ ist ein Beispiel für eine parametrische Familie - die Familie der Normalverteilungen. Es gilt

$$\vartheta(\mathcal{N}(\mu, \sigma^2)) = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+ .$$

(*iv*) Sei $\mathcal{P}_n = \{P_{p,n}, 0 \leq p \leq 1\}, n \in \mathbb{N}$ fest, mit

$$P_{p,n}(j) = \begin{cases} \frac{n!}{(n-j)!j!} p^j (1-p)^{n-j} & 0 \leq j \leq n \\ 0 & \text{sonst} \end{cases}$$

ist eine parametrische Familie. Hier gilt

$$\vartheta(P_{p,n}) = p.$$

(*v*) Setze $\Omega = \{1, \ldots, 6\}^3$. Für $w \in \Omega$ setze

$$P_1(\omega) = 6^{-3}$$

bzw.

$$P_2(\omega) = \frac{1}{6 \cdot 5 \cdot 4} \mathbb{1}_{\Omega^*}$$

mit $\Omega^* \coloneqq \{\omega \in \Omega : w_1 \neq w_2 \neq w_3\}$.

Definition 0.3:
Eine Stichprobe / Beobachtung X bzgl. des Experiments $(\mathcal{P}, (\Omega, \mathcal{A}))$ ist eine Zufallsvariable

$$X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$$

mit Verteilung $P_X \in Q = \{P_X : P \in \mathcal{P}\}$, wobei $\mathcal{X}$ ein vollständiger, metrischer, separabler Raum ist. $\mathcal{X}$ heißt Stichprobenraum.

Beispiel 0.4 (Wiederholte Stichproben)**:**
Sei $X = (\omega_1, \ldots, \omega_n) \in \Omega^n$ mit Verteilung

$$Q_n \coloneqq \underbrace{P \bigotimes \ldots \bigotimes P}_{n \text{ mal}}.$$

Dann beschreibt X eine Stichprobe aus einer n-maligen unabhängigen Wiederholung des Experiments $(\mathcal{P}, \Omega, \mathcal{A}), Q = \{\bigotimes_{j=1}^n P, P \in \mathcal{P}\}$ ist das Produktexperiment auf $(\Omega^n, \bigotimes_{j=1}^n \mathcal{A})$. X heißt Stichprobe der Länge n.

Ziel: Aus einer Stichprobe der Länge n schließe zurück auf die Eigenschaften des W-Maßes P, welches X generiert, d.h. X ist verteilt mit $P \in \mathcal{P}$. Das ist Gegenstand der schließenden Statistik im Gegensatz zur explorativen Datenanalyse. Ω kann komplex sein.

Beispiel 0.5 (Münzexperiment)**:**
$\omega_1, \ldots, \omega_n \in \{0, 1\}$ seien unabhängige Münzwürfe, wobei Kopf = 1, Zahl = 0. Definiere: $P_\alpha(\{1\}) = \alpha = 1 - P_\alpha(\{0\})$ mit $0 \leq \alpha \leq 1$. Setze $\Omega = \{0, 1\}^n, \mathcal{A} = \mathcal{P}(\Omega)$. Dann hat $X \coloneqq (X_1, \ldots, X_n)$ die Verteilung $Q_\alpha = \bigotimes_{j=1}^n P_\alpha$.
Dies ist es ein parametrisches Produktexperiment $Q = \{Q_\alpha, 0 \leq \alpha \leq 1\}$, $X = (X_1, \ldots, X_n), Q_\alpha((\omega_1, \ldots, \omega_n)) = \alpha^{\sum_i \omega_i}(1-\alpha)^{n-\sum_i \omega_i}$ die Wahrscheinlichkeit von $X = (\omega_1, \ldots, \omega_n)$ gegeben die Erzeugung durch P_α^n.
Die sogenannte Likelihood-Funktion $[0, 1] \ni \alpha \mapsto Q_\alpha(\omega_1, \ldots, \omega_n)$ hängt nur von der Summe $\sum_j \omega_j$ ab, und

$$\widehat{\alpha} = \frac{1}{n} \sum_j \omega_j$$

ist die empirische Häufigkeit von Köpfen in n Würfen.
Idee eines Schätzers für den unbekannten Parameter $\alpha \in [0, 1]$: Rahmen für das Schließen von der Stichprobe X auf die erzeugende Verteilung P entsteht durch die Entscheidungs- oder Spieltheorie, sowohl für das Schätzen wie auch für das Testen von Hypothesen über $\mathcal{P}$.
Die Wahl eines $P \in \mathcal{P}$ zur Beobachtung X stellt eine „Entscheidung“ dar. Allgemeiner ist man an speziellen Eigenschaften oder Parametern von P interessiert, z.B. einem Parameter $\tau(P)$, welcher kategorial ist, d.h. $\tau : \mathcal{P} \to \{0, \ldots, k\} = D, k \geq 1$. Dies zerlegt $\mathcal{P} = \tau^{-1}(0) \cup \ldots \cup \tau^{-1}(k)$ und man muss mittels X eine Entscheidung treffen.

Definition 0.6:
Sei $D = \tau(\mathcal{P})$ die Menge der Entscheidungen, wobei D ein vollständiger, metrischer Raum mit Borel-σ-Algebra sei. Dann ist eine Entscheidungsregel eine messbare Abbildung

$$\delta : (\mathcal{X}, \mathcal{B}) \to (D, \mathcal{D}),$$

wobei $\mathcal{X}$ der Stichprobenraum ist. δ liegt im Bildbereich des Parameters τ, d.h. $\delta(x) \in \tau(\mathcal{P})$.

Sei $\tau : \mathcal{P} \to \{0, \ldots, k\} =: D$. Für $k = 1$ ist dies das Problem des Hypothesentestens. Entscheide, ob $\tau(P) = 0$ (Hypothese) oder $\tau(P) = 1$ (Alternative), z.B. beim n-maligen Münzwurf.
Sei z.B. die Hypothese gegeben durch $H_0 \coloneqq \{Q_\alpha : 0 \leq \alpha < 0.5\}$. Dann bedeutet $\tau(Q_\alpha) = 0$, dass $0 \leq \alpha < 0.5$ und $\tau(Q_\alpha) = 1$, dass die Alternative $A \coloneqq \{Q_\alpha : 0.5 \leq \alpha \leq 1\}$ vorliegt.

Beispiel:
Sei P_0 die Gleichverteilung auf $I_0 = [-2, 1]$ und P_1 die Gleichverteilung auf

$I_1 = [-1, 2]$. Beobachte $x \in \mathbb{R}$ erzeugt von P_0 oder P_1, wobei $|x| \leq 2$.
Mögliche Entscheidungsregel:

$$\delta(x) = \begin{cases} 0 & x \leq -1 \\ 1 & x \geq 1 \\ \varphi(x) & |x| \leq 1 \end{cases}$$

Wähle $\varphi(x) \in \{0, 1\}$.
Dann ist der Fehler erster Art: $\alpha_0 = P_0(x : \delta(x) = 1) = P_0(\varphi(x) = 1)$ und der Fehler zweiter Art $\alpha_1 = P_1(x : \delta(x) = 0) = P_1(\varphi(x) = 0)$.

0.1 Verlust für fehlerhafte Entscheidungen

Definition:
Sei $(\mathcal{P}, (\mathcal{X}, \mathcal{B}))$ ein Experiment, $(D, \mathcal{D})$ der Entscheidungsraum, $\tau : \mathcal{P} \to D$ ein Parameter.
Eine für jedes $P \in \mathcal{P}$ in der zweiten Koordinate messbare Funktion

$$L : \mathcal{P} \times D \to [0, \infty)$$

heißt Verlustfunktion, falls

$$L(P, \tau(P)) = 0 \text{ für alle } P \in \mathcal{P},$$

d.h. $\tau(P)$ ist die korrekte Entscheidung und sollte daher Verlust 0 haben. Falls $\tau(\mathcal{P}) \subset \mathbb{R}^d$, sollte L monoton wachsend für größere Abweichungen von $\delta(x)$ und $\tau(P)$ sein, wo $X \sim P$, d.h.

$$\|\tau(P) - \tau\| \to L(P, \tau)$$

sollte monoton wachsend sein.

Beispiel 0.7:
Beim n-maligen Münzwurf mit $0 \leq \alpha \leq 1, 0 \leq \beta \leq 1$ setze z.B.

(a) $L(Q_\alpha, \beta) = (\alpha - \beta)^2$

(b) $L(Q_\alpha, \beta) = \|\alpha - \beta\|$

(c) $L(Q_\alpha, \beta) = \frac{(\alpha-\beta)^2}{\alpha(1-\alpha)}$

Sei $x_1, \ldots, x_N \in \mathcal{X}$ eine N-fache unabhängige Wiederholung eines Experiments mit identischer Verteilung P. Dann ist der mittlere „empirische“ Verlust einer Entscheidungsregel δ in $x = (x_1, \ldots, x_N)$:

$$\widehat{L}_N = \frac{1}{N} \sum_{j=1}^{N} L(P, \delta(x_j)) \to \int L(P, \delta(x)) P(dx) =: R(P, \delta) \; P - \text{f.s.}$$

falls $R(P, \delta) < \infty$ für alle $P \in \mathcal{P}$ (Kolmogorov-Gesetz).

Definition 0.8:
Seien die Voraussetzungen wie in Beispiel 0.7. Dann heißt die Abbildung

$$(P, \delta) \to R(P, \delta) \coloneqq \int L(P, \delta(x)) P(dx), P \in \mathcal{P}$$

Risiko der Entscheidungsregel δ für Stichproben aus $P \in \mathcal{P}$. $P \to R(P, \delta)$ heißt Risiko-Funktion.

Beispiel 0.9:

(*i*) Münzwurf: $X_j \in \{0,1\}, \mathbb{E}_{Q_\alpha} X_j = \alpha$ und L wie in (0.7)a. Dann gilt für $\widehat{\alpha} = \frac{1}{n}(x_1 + \ldots + x_n)$

$$\begin{aligned} R(Q_\alpha, \widehat{\alpha}) &= \mathrm{Var}_{Q_\alpha}(\widehat{\alpha}) = \int (\alpha - \widehat{\alpha})^2 dQ_{\widehat{\alpha}} \\ &= \frac{n}{n^2} Var(X_j) = \frac{\alpha(1-\alpha)}{n} \end{aligned}$$

d.h. die Risiko-Funktion ist

$$\alpha \mapsto \frac{\alpha(1-\alpha)}{n}.$$

(*ii*) Sei $\{\mathcal{N}(\mu, 1), \mu \in \mathbb{R}\}$ eine Normalverteilungsfamilie mit Erwartungswert $\mu \in \mathbb{R}$ und Varianz 1. Setze als Verlustfunktion

$$L(\mu, \tau) = (\mu - \tau)^2$$

und $\tau(P_\mu) = \mu$. $(x_1, \ldots, x_n)$ sei das Produktexperiment auf P_μ^n. Schätze

$$\delta(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \ldots + x_n).$$

Da $\int \delta(x_1, \ldots, x_n) dP_\mu^n = \mu$, gilt

$$\begin{aligned} R(P_\mu, \delta) &= \int (\mu - \delta)^2 dP_\mu^n = Var_\mu(\frac{1}{n}(x_1 + \ldots + x_n)) \\ &= \frac{n}{n^2} Var(X_1) = \frac{1}{n} \ \forall \mu. \end{aligned}$$

Frage: Gibt es eine beste Entscheidungsregel bezüglich der Risiko-Funktion? Beim Münzwurf: $L(\alpha, \beta) = (\alpha - \beta)^2$. Erster Schätzer:

$$\widehat{\alpha} = \frac{1}{n}(x_1 + \ldots + x_n)$$

Zweiter Schätzer:

$$\widehat{\beta} = \beta, 0 \leq \beta \leq 1 \ \ \text{const}$$

Da $\alpha \to R(Q_\alpha, \delta)$ halbgeordnet sind, gibt es keinen „beste“ Schätzer.
Gibt es für eine Verlustfunktion L und ein Experiment $(\mathcal{P}, (\Omega, \mathcal{A}))$ und eine Klasse $\mathcal{E}$ von Entscheidungsregeln wenigstens eine minimax-Lösung, d.h. ein $\delta_0 \in \mathcal{E}, P_0 \in \mathcal{P}$ mit

$$R(P_0, \delta_0) = \min_{\delta \in \mathcal{E}} \left(\max_{P \in \mathcal{P}} R(P, \delta) \right)?$$

Im Allgemeinen nicht, da $(P, \delta) \to R(P, \delta)$ keine konvexe Funktion auf einer konvexen Menge $\mathcal{P}$ und $\mathcal{E}$ ist.
Idee: Einbettung von $\mathcal{P}$ und $\mathcal{E}$ in konvexe Mengen von W-Maßen bzw. konvexen Mengen, die $\mathcal{E}$ umfassen mittels $\mathcal{P} \ni P \to \delta_P$ W-Maß auf $\mathcal{P}$ mit

$$\mathcal{E}_P(A) = \mathbb{1}_{P \in A}$$

(Dirac-Maß) und $A \in \mathcal{F}$ als σ-Algebra über $\mathcal{P}$.
Dann gilt

$$R(P, \delta) = \int R(Q, \delta) \delta_P(dQ).$$

Die Menge der W-Maße auf $\mathcal{P}$ ist eine konvexe Menge. Sei $\mathcal{P}_C$ diese Menge von W-Maßen auf $\mathcal{P}$ mit einer σ-Algebra $\mathcal{F}$.
Statt die Risiko-Funktion $P \to R(P, \delta)$ für eine Entscheidungsregel zu vergleichen, wählt man a-priori Information über die unbekannte $P \in \mathcal{P}$ in der Form einer Wahrscheinlichkeitsverteilung Π auf $\mathcal{P}$. Dann ergibt sich ein mittleres Bayessches Risiko zu Π

$$R(\Pi, \delta) = \int R(P, \delta) \Pi(dP),$$

vorausgesetzt, dass das Integral existiert. Π ist ein „Super“-Experiment auf der Ebene der W-Maße. Ebenso ist die Menge der Entscheidungsregeln $\delta : \mathcal{X} \to D$ im Allgemeinen nicht konvex. Hier kann man genauso randomisieren.
Zu $x \in \mathcal{X}$ wähle ein W-Maß auf D, $\delta(x, \cdot)$ auf $(D, \mathcal{D})$, d.h. man würfelt nach der Beobachtung x mit der Verteilung $\delta(x, \cdot)$, welche Entscheidung man fällt.
Definiere das W-Maß auf D

$$\delta \mapsto \delta_{\delta(x)}$$

mit

$$\delta_{\delta(x)}(A) = \mathbb{1}_{\delta(x) \in A}.$$

Die Menge der randomisierten Entscheidungsregeln $\mathcal{E}_C$ ist konvex.
Also:

1) Natur wählt ein $\Pi \in \mathcal{P}_C$.

2) Zusatzexperiment $P \sim \Pi$, wählt P mit Π.

3) Beobachte x aus einer Zufallsvariablen mit $\mathcal{X} \ni X \sim P$.

4) Wähle das W-Maß $x \to \delta(x, \cdot)$.

5) Treffe Entscheidung mittels einer $(D, \mathcal{D})$-messbaren Zufallsvariablen mit $d \sim \delta(x, \cdot)$.

Dann ist das Risiko

$$R(\Pi, \delta) : \mathcal{P}_C \times \mathcal{E}_c \to \mathbb{R}$$

gegeben durch

$$R(\Pi, \delta) = \iiint L(P, e)\delta(x, de)P(dx)\Pi(dP).$$

Definition 0.10:
Eine randomisierte Entscheidungsregel ist ein Markov-Kern

$$\delta : \mathcal{X} \to \{\text{W-Maße über } (D, \mathcal{D})\}, \delta \in \mathcal{E}_C$$

d.h. $x \to \delta(x, A)$ ist messbar auf $\mathcal{X}$ für alle $A \in \mathcal{D}$ und $A \to \delta(x, A)$ ist ein W-Maß auf $\mathcal{D}$.

Idee: Beobachte $x \in \mathcal{X}$ und treffe Entscheidung mit einem weiteren Zufallsexperiment mit Verteilung $\delta(x, \cdot)$ auf D.
Zum Beispiel beim Münzwurf

$$\tilde{\alpha}(x_1, \ldots, x_n) = \left(\frac{1}{n}(x_1 + \ldots + x_n) + y\right)_{[0,1]}$$

wo $x_j \in \{0, 1\}$ und y ist von x_j unabhängige Zufallsvariable mit Verteilung $\mathcal{N}(0, \varepsilon^2)$.
Die Entscheidungsregeln in $\mathcal{E}$ bzw die W-Maße des Experiments in $\mathcal{P}$ sind die Eckpunkte der konvexen Mengen $\mathcal{E}_C$ und $\mathcal{P}_C$.

Beispiel:
Sei $Q = \{Q_p, Q_q\}$ ein einfacher Münzwurf mit Erfolgswahrscheinlichkeit $0 \leq p \leq 1, 0 \leq q \leq 1, p \neq q$, d.h. $Q_p(X = 1) = p$ oder $Q_q(X = 1) = q$. Sei die Entscheidungsregel

$$\delta : \underbrace{\{0, 1\}}_{\mathcal{X}} \times \underbrace{\{p, q\}}_{D} \to [0, 1]$$

gegeben durch

$$\delta_{\alpha_0,\alpha_1}(0, q) = \alpha_0 = 1 - \delta(0, p)$$

und

$$\delta_{\alpha_0,\alpha_1}(1, q) = \alpha_1 = 1 - \delta(1, p).$$

Dann ist der Verlust

$$L(Q_\beta, d) = \mathbb{1}_{\beta=d}$$

und das Risiko ist

$$R(Q_\beta, \delta_{\alpha_0,\alpha_1}) = \begin{cases} \alpha_1 p + \alpha_0(1 - p) & \beta = p \\ (1 - \alpha_0)(1 - q) + (1 - \alpha_1)q & \beta = q \end{cases},$$

sodass man eine Abbildung

$$[0,1]^2 \to (\alpha_0, \alpha_1) \to (R(Q_p, \delta_{\alpha_0,\alpha_1}), R(Q_q, \delta_{\alpha_0,\alpha_1}))$$

erhält, und somit das Risiko im $\mathbb{R}^2$ liegt, d.h. nicht total geordnet ist.

Definition 0.11:

i) Eine Entscheidungsregel δ_1 heißt besser als die Regel δ_2 ($\delta_1 \ll \delta_2$), falls für alle $P \in \mathcal{P}$

$$R(P, \delta_1) \leq R(P, \delta_2)$$

gilt, wobei die Ungleichung für mindestens ein $P \in \mathcal{P}$ strikt ist. Es gibt bezüglich dieser besser-Relation im Allgemeinen viele minimalen Entscheidungsregeln δ.

ii) Gibt es zur Entscheidungsregel δ eine Entscheidungsregel δ' mit $\delta' \ll \delta$, so heißt δ nicht zulässig, sonst zulässig.

iii) Eine Menge $\mathcal{R}$ von Entscheidungsregeln heißt vollständig, falls es für alle $\delta \in \mathcal{R}$ ein δ' gibt mit $\delta' \ll \delta$.

iv) $\mathcal{R}$ heißt minimal vollständig, falls $\mathcal{R}$ keine echte vollständige Teilmenge enthält.

v) $\mathcal{R}$ heißt wesentlich vollständig, falls es für alle Entscheidungsregeln δ ein $\delta' \in \mathcal{R}$ gibt mit $R(P, \delta') \leq R(P, \delta)$ für alle $P \in \mathcal{P}$.

vi) δ, δ' heißen äquivalent, falls $R(P, \delta') = R(P, \delta)$ für alle $P \in \mathcal{P}$. Falls $\mathcal{R}$ minimal vollständig ist und $\delta \in \mathcal{R}$ und δ' äquivalent sind, dann $\delta' \notin \mathcal{R}$. Im Fall, dass $\mathcal{R}$ wesentlich minimal vollständig ist, so enthält $\mathcal{R}$ zu jedem minimalen Element $\delta \in \mathcal{R}$ nur einen Repräsentanten der Risiko-Äquivalenzklasse.

Definition 0.12:
$\delta_0 \in \mathcal{E}_C$ heißt Bayes-optimale Entscheidungsregel zu $\Pi \in \mathcal{P}_C$, falls

$$R(\Pi, \delta) \geq R(\Pi, \delta_0) \text{ für alle } \delta \in \mathcal{E}_C.$$

Satz 0.13 (Le Cam, 1955)**:**
Sei $\mathcal{E}$ ein polnischer Raum. Die Klasse aller Entscheidungsregeln δ, für die für jedes $\varepsilon > 0$ gilt: es existiert Π_ε Bayes Vorbewertung auf $\mathcal{P}$ mit

$$R(\Pi_\varepsilon, \delta) - \inf_{\tau \in \mathcal{E}} R(\Pi_\varepsilon, \tau) < \varepsilon$$

ist wesentlich vollständig, falls $\delta \mapsto R(P, \delta)$ unterhalbstetig und $\mathcal{E}$ kompakt in geeigneter Topologie (genauer: die schwache Topologie) ist.

Beweis. [8], Kapitel 8, Satz 42.3 und Satz 47.7. ∎

Definition 0.14:
Eine Minimax-Entscheidungsregel ist eine Entscheidungsregel $\delta_m \in \mathcal{E}_C$ mit Minimax-Risiko

$$MR = \sup_{P\in\mathcal{P}} R(P,\delta_m) = \inf_{\delta\in\mathcal{E}_C} \sup_{\Pi\in\mathcal{P}_C} R(\Pi,\delta) = \inf_{\delta\in\mathcal{E}} \sup_{P\in\mathcal{P}} R(P,\delta).$$

Bemerkung 0.15:

(*i*) Die Menge $\mathcal{P}_B = \{\delta \in \mathcal{E}_C : \delta$ Bayes-optimale Entscheidungsregeln zu einem $\Pi_\delta \in \mathcal{P}_C\}$ ist wesentlich vollständig.

(*ii*) Ist δ_m eine Minimax-Entscheidungsregel, so gilt

$$MR = \sup_{P\in\mathcal{P}} R(P,\delta_m) = \inf_{\tau\in\mathcal{E}} \sup_{Q\in\mathcal{P}} R(Q,\tau) \geq \sup_{Q\in\mathcal{P}} \inf_{\tau\in\mathcal{E}} R(Q,\tau).$$

Beweis.

(*i*) Sei $\mathcal{P}_B$ nicht wesentlich vollständig. Dann existiert eine Entscheidungsregel δ aus $\mathcal{E}$, sodass für alle $\delta' \in \mathcal{P}_B$ ein $P = P(\delta) \in \mathcal{P}$ existiert, sodass

$$R(P,\delta') > R(P,\delta).$$

Daraus folgt

$$R(\Pi,\delta) \leq R(\Pi,\delta')\ \forall\delta' \in \mathcal{P}_B\ \forall\Pi \in \mathcal{P}_C\,. \qquad (*)$$

Da δ Bayes-Entscheidungsregel ist, gibt es ein $\Pi_\delta \in \mathcal{P}_C$ mit

$$R(\Pi_\delta,\delta) \leq R(\Pi_\delta,\delta') \qquad (**)$$

woraus folgt

$$R(\Pi_\delta,\delta') = R(\Pi_\delta,\delta) \leq R(\Pi_\delta,\tau) \ \text{ für alle } \tau \in \mathcal{E},$$

d.h. δ' ist auch eine Bayes-optimale Entscheidungsregel zu Π_δ und damit $\delta' \in \mathcal{P}_B$.

(*ii*) $MR \geq \inf R(Q,\tau) \geq \sup_{Q\in\mathcal{P}} \inf_{\tau\in\mathcal{E}} R(Q,\tau)$.

∎

Satz 0.16 (Le Cam, Wald, 1945)**:**
Sind $\mathcal{P}$ und $\mathcal{D}$ kompakt in geeigneter Topologie, $\mathcal{P}_C$ und $\mathcal{E}_C$ mit der schwach$*$-Topologie versehen und

$$(\Pi, \delta) \to R(\Pi, \delta)$$

auf $\mathcal{P}_C \times \mathcal{E}_C$ unterhalbstetig, so gilt der Minimax-Satz:
Es existiert ein $\Pi_0 \in \mathcal{P}_C, \delta_0 \in \mathcal{E}_C$, sodass

$$MR = \inf_{\delta \in \mathcal{E}_C} \sup_{\Pi \in \mathcal{P}_C} R(\Pi, \delta) = \sup_{\Pi \in \mathcal{P}_C} \inf_{\delta \in \mathcal{E}_C} R(\Pi, \delta) = R(\Pi_0, \delta_0)$$

und $R(\cdot, \delta_0)$ ist konstant. Π_0 heißt ungünstigste Vorbewertung und δ_0 ist eine Bayes-optimale Entscheidungsregel zu Π_0.

Beweis. [8], Satz 46.3, S. 241. ∎

1 Suffiziente Statistiken und exponentielle Familien

Satz 1.0:

(*i*) *Seien μ, ν σ-endliche Maße auf $(\mathcal{X}, \mathcal{B})$ (d.h. $\mathcal{X} = \bigcup_{i=1}^{\infty} A_i$ mit $\mu(A_i) < \infty \ \forall i \in \mathbb{N}$) mit $\nu \ll \mu$. Dann gibt es nach dem Satz von Radon-Nikodym eine fast-überall Äquivalenzklasse von Dichten $\frac{d\nu}{d\mu}$, sodass für $f \in \frac{d\nu}{d\mu}$ gilt*

$$\nu(A) = \int_A f(x)\mu(dx). \qquad (*)$$

(*ii*) *Sei g ν-integrierbar und $\nu \ll \mu$. Dann gilt*

$$\int g d\nu = \int g f d\mu$$

mit f wie oben.

(*iii*) *Sei $(\mathcal{Z}, \mathcal{S})$ ein Maßraum und $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Z}, \mathcal{S})$ messbar. Dann gilt der Transformationssatz*

$$\int f d(\mu \circ T) = \int f \circ T d\mu$$

wobei $\mu \circ T(B) = \mu(T^{-1}(B))$ das induzierte Maß von T auf $(\mathcal{Z}, \mathcal{S})$ ist.

(*iv*) *Sei μ ein W-Maß auf $(\Omega, \mathcal{A})$, $f : \Omega \to [0, \infty)$ messbar und $(\mathcal{Z}, \mathcal{S})$ ein Maßraum. Dann heißt eine Abbildung $\varphi : \mathcal{Z} \to [0, \infty)$ eine bedingte faktorisierte Erwartung von f gegeben $T : \Omega \to \mathcal{Z}$ (messbar), falls*

$$\nu_T(B) \coloneqq \int_{T^{-1}(B)} f d\mu = \int_{T^{-1}(B)} \varphi \circ T d\mu = \int_B \varphi \ d\mu \circ T$$

auf $(\mathcal{Z}, \mathcal{S})$ gilt, d.h. $\varphi \circ T$ ist eine Version der Radon-Nikodym-Dichte $\frac{d\nu_T}{d\mu_T}$ und ist $T^{-1}(\mathcal{S}) - \mathcal{B}$-messbar, d.h.

$$\varphi \circ T = \mathbb{E}_\mu \left(f \mid T^{-1}(\mathcal{S}) \right).$$

$\varphi \circ T$ ist auf $\{\omega \in \Omega : T(\omega) = a\}$ konstant und eine Mittelung von f über $\{T = a\}$ mittels μ.
Im diskreten Fall: $T : \Omega \to \{1, 2, \ldots, k\}$ mit $\mu(T^{-1}(j)) > 0$ gilt offensichtlich

$$\varphi(j) = \int_{T^{-1}(j)} f d\mu \frac{1}{\mu(T^{-1}(j))}, j = 1, \ldots, k$$

d.h. $\omega \mapsto \varphi(T(\omega))$ ist konstant auf $\{T = j\}$.

(v) *Seien* $(\Omega, \mathcal{A}), (\mathcal{Z}, \mathcal{S})$ *Maßräume. Eine Funktion*

$$M : \Omega \times \mathcal{S} \to [0, 1]$$

heißt Markov-Kern, falls

(a) $\omega \mapsto M(\omega, C)$ *ist* $\mathcal{A}$*-messbar für alle* $C \in \mathcal{S}$*.*

(b) $C \mapsto M(\omega, C)$ *ist ein W-Maß für alle* $\omega \in \Omega$*.*

Ist Ω *ein vollständiger metrischer Raum, so gibt es zu jedem W-Maß* P *auf* $(\Omega, \mathcal{A})$ *und zu einer messbaren Abbildung* $T : \Omega \to \mathcal{Z}$ *einen Markov-Kern* M *auf* $\mathcal{Z} \times \mathcal{A}$*, sodass* $M(\cdot, A), A \in \mathcal{A}$ *die bedingte faktorisierte Erwartung von* $\mathbb{1}_A$ *gegeben* T *bezüglich* P *ist, d.h.*

$$\int_C M(z, A) P \circ T(dz) = \int_{T^{-1}(C)} M(T(\omega), A) P(dw) = P(A \cap T^{-1}(C)) \tag{1.1}$$

für alle $C \in \mathcal{S}$ *und* $A \in \mathcal{A}$ *gilt.*
Allgemeiner ist also für einen Maßraum $(Y, \mathcal{F})$ *und eine Abbildung* $S : \Omega \to Y$ *ein Markov-Kern* $M(z, B), B \in \mathcal{F}$ *definiert, der die bedingte Erwartung von* $\mathbb{1}_{S^{-1}(B)}$ *gegeben* T *bzgl.* P *ist, falls*

$$\int_C M(z, B) P \circ T(dz) = P(T^{-1}(C) \cap S^{-1}(B)) \tag{1.2}$$

für alle $C \in \mathcal{S}, B \in \mathcal{F}$ *gilt.*
Hier ist $M(z, B)$ *für* $A = S^{-1}(B)$ *die bedingte Wahrscheinlichkeit* $P \circ S(B \mid T^{-1}(z))$*, falls* $P(T^{-1}(z)) > 0$ *bedingt von* S *gegeben* T*.*

(vi) *Unter den gleichen Voraussetzungen wie bei v) sei* M *auf* $\mathcal{Z} \times \mathcal{F}$ *eine bedingte Verteilung von* S *gegeben* T*. Dann gilt für jede messbare Funktion* $f : Y \times \mathcal{Z} \to \mathbb{R}$*, dass*

$$Z \ni z \mapsto \varphi(z) \coloneqq \int f(y, z) M(z, dy)$$

die bedingte Erwartung von $f(S, T)$ *gegeben* T *unter* P *ist, falls* f P*-integrierbar ist. D.h. es gilt*

$$\int_{T^{-1}(A)} \varphi \circ T dP = \int_{T^{-1}(A)} f(S, T) dP$$

für alle $A \in \mathcal{S}$*.*

Beispiel:
Sei $T : \Omega \to Z, S : \Omega \to Y$, wobei Y, Z endliche Mengen sind. Setze

$$M(z, B) = P(S \in B \mid T = z) = \frac{P\left(S(\omega) \in B \cap T(\omega) = z\right)}{P(T(\omega) = z)},$$

falls $P(T(\omega) = z) > 0$ ist.
Für eine Funktion $f : Y \times Z \to \mathbb{R}$ gilt dann

$$\mathbb{E}_P(f(S, T) \mid T = z) = \sum_{y \in Y} f(y, z) M(z, \{y\}).$$

Definition 1.1:
Sei $\mathcal{P}$ eine Familie von W-Maßen über $(\mathcal{X}, \mathcal{B})$ und μ ein σ-endliches Maß über $(\mathcal{X}, \mathcal{B})$. Dann heißt $\mathcal{P}$ dominiert durch μ, falls $P \ll \mu$ für alle $P \in \mathcal{P}$, in Zeichen $\mathcal{P} \ll \mu$. In diesem Fall gibt es nach Satz 1.0 Funktionen f_P derart dass für alle $A \in \mathcal{B}$ gilt

$$P(A) = \int_A f_P(x) \mu(dx).$$

Die Familie

$$\mathcal{Q} \coloneqq \{\sum_{i=1}^{\infty} \alpha_i P_i \mid P_i \in \mathcal{P}, \alpha_i \geq 0, \sum_{i=1}^{\infty} \alpha_i = 1\}$$

heißt die zu $\mathcal{P}$ gehörige konvexe Familie. Es gilt $\mathcal{Q} \subset \mathcal{P}_C$.

Satz 1.2:
Falls $\mathcal{P} \ll \mu, \mu$ σ-endlich, dann gibt es ein W-Maß $Q_* \in \mathcal{Q}$ mit $\mathcal{Q} \ll Q_*$.

Beweis. Sei OE μ endlich: Sei $\mathcal{X} = \cup_{i=1}^{\infty} A_i, \infty > \mu(A_i) > 0$. Ersetze μ durch das W-Maß

$$\tilde{\mu}(A) \coloneqq \sum_{k=1}^{\infty} \frac{\mu(A \cap A_k)}{\mu(A_k)} 2^{-k}$$

für $A \in \mathcal{B}$. Dann $\mathcal{P} \ll \mu \ll \tilde{\mu}$.
Wähle zu $P \in \mathcal{P}$ eine Version der Dichte $h_P \in \frac{dP}{d\mu}$. Sei $S_P \coloneqq \{x \in \mathcal{X} : h_P(x) > 0\}, \mathcal{S} \coloneqq \{S_P : P \in \mathcal{P}\}$ und sei $\mathcal{S}_*$ die Menge der abzählbaren Vereinigungen von Mengen aus $\mathcal{S}$. Dann ist $\mathcal{S}_*$ abgeschlossen bezüglich abzählbaren Vereinigungen. Daher gibt es ein $S_0 \in \mathcal{S}_*$ mit

$$\mu(S_0) = \sup\{\mu(A) : A \in \mathcal{S}_*\}$$

denn $\sup(\cdot) = \lim_n \mu(A_n)$ für eine Folge $(A_n)_n \subset \mathcal{S}_*$ und

$$\tilde{A}_n \coloneqq \cup_{j=1}^{n} A_j \uparrow \cup_{j=1}^{\infty} A_j = S_0 \in \mathcal{S}_*$$

und mit Hilfe der Maßstetigkeit von μ.
Es gilt $\mu(A \cap S_0^c) = 0$ für alle $A \in \mathcal{S}_*$ nach Konstruktion, d.h. speziell

$$\mu(S_P \cap S_0^c) = 0 \ \forall P \in \mathcal{P}\,. \tag{i}$$

Aus $S_0 \in \mathcal{S}_*$ folgt, dass es eine Folge $(P_n)_n \subset \mathcal{P}$ gibt mit

$$S_0 = \cup_{n=1}^{\infty} S_{P_n}.$$

Definiere

$$Q_* := \sum_{n=1}^{\infty} 2^{-n} P_n,$$

so gilt $\mathcal{Q} \ll Q_*$, denn

$$Q_*(A) = 0 \Rightarrow P_n(A) = \int \mathbb{1}_A(x) h_{P_n}(x) \mu(dx) = 0 \text{ für alle } n \in \mathbb{N},$$

d.h. $\mathbb{1}_A(\cdot) h_{P_n}(\cdot) = 0$ μ-f.ü. oder $\mu(A \cap S_{P_n}) = 0$. Daraus folgt

$$\mu(A \cap S_0) = 0. \tag{ii}$$

Für beliebiges $Q \in \mathcal{Q}$ und $S_Q := \{h_Q > 0\}$, wo $h_Q = \sum_{i=1}^{\infty} \alpha_{i,Q} h_{P_{i,Q}}$ mit $\sum_i \alpha_{i,Q} = 1$ gilt

$$\mu(A \cap S_Q) = \mu(A \cap S_0 \cap S_Q) + \mu(A \cap S_Q \cap S_0^c) \leq \mu(A \cap S_0) + \mu(S_Q \cap S_0^c) = 0,$$

da $S_Q \subset \cup_{i=1}^{\infty} S_{P_i,Q}$.
Also folgt $Q(A) = \int \mathbb{1}_A(x) h_Q(x) \mu(dx) = 0$ für beliebiges $Q \in \mathcal{Q}$. ∎

1.1 Suffizienz von Statistiken

Definition 1.3:
Eine Statistik ist eine messbare Abbildung T vom Raum der Stichproben $(\mathcal{X}, \mathcal{B})$ in einen Maßraum $(\mathcal{Z}, \mathcal{S})$. Wir nehmen an, dass $\mathcal{X}$ ein vollständiger, separabler, metrischer Raum und $\mathcal{B}$ die Borel-σ-Algebra darauf ist.

Definition 1.4:
Eine Statistik $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Z}, \mathcal{S})$ heißt suffizient für die Familie $\mathcal{P}$ auf $(\mathcal{X}, \mathcal{B})$, falls es für jedes $A \in \mathcal{B}$ eine faktorisierte bedingte Erwartung $\varphi_A : \mathcal{Z} \to [0, \infty)$ von $\mathbb{1}_A$ gegeben T gibt mit

$$\int_B \varphi(A) dP \circ T = \int_{T^{-1}(B)} \varphi_A \circ T dP = \int_{T^{-1}(B)} \mathbb{1}_A dP =: P_A(T^{-1}(B))$$

für alle $P \in \mathcal{P}, B \in \mathcal{S}$, d.h. φ_A ist gleich für alle W-Maße $P \in \mathcal{P}$.
Ist $\mathcal{P} \ll \mu$, dann gilt für alle $A \in \mathcal{B}$:

$$\varphi_A \in \bigcap_{P \in \mathcal{P}} \frac{dP_A \circ T}{dP \circ T} \Leftrightarrow T \text{ ist suffizient.}$$

Satz 1.5 (Faktorisierungssatz)**:**
Sei $\mathcal{P}$ eine Familie von W-Maßen über $(\mathcal{X}, \mathcal{B})$ mit $\mathcal{P} \ll \mu, \mu$ σ-endlich und Q_* das dominierende Maß aus Satz 1.2. Dann ist eine Statistik $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Z}, \mathcal{S})$ genau dann suffizient für $\mathcal{P}$, wenn es für alle $P \in \mathcal{P}$ eine $\mathcal{B}^1 - \mathcal{S}-$messbare Funktion

$$g_P : \mathcal{Z} \to [0, \infty)$$

mit der Eigenschaft gibt, dass

$$\mathcal{X} \ni x \mapsto g_P(T(x)) \tag{1.6}$$

eine Version von $\frac{dP}{dQ_*}$ gibt, d.h. es gilt für alle $A \in \mathcal{B}$:

$$P(A) = \int_A g_P(T(x)) Q_*(dx).$$

Beweis. (i): Ist T suffizient, so existiert $\varphi_A \in \bigcap_{P \in \mathcal{P}} \frac{dP_A \circ T}{dP \circ T}$, d.h. es gilt wegen Satz 1.0(iii)

$$\int_B \varphi_A(z) P \circ T(dz) = \int_{T^{-1}(B)} \mathbb{1}_A(x) P(dx) \quad \text{für alle } B \in \mathcal{S} \text{ und } P \in \mathcal{P}\,. \quad (i)$$

Wegen Satz 1.2 folgt für $B \in \mathcal{S}$ und die Folge $(P_j)_{j \in \mathbb{N}}$ von Q_* aus (i)

$$\int_B \varphi_A d\left(\sum_{j=1}^n \alpha_j P_j \circ T\right) = \int_{T^{-1}(B)} \mathbb{1}_A d\left(\sum_{j=1}^n \alpha_j P_j\right) \to \int_{T^{-1}(B)} \mathbb{1}_A dQ_*.$$

Aus dieser Konvergenz folgt mit $\mathbb{1}_A$ ersetzt durch $f_k = \sum_{j=1}^k \beta_j \mathbb{1}_{A_j}$ ebenfalls Konvergenz für $n \to \infty$. Für $0 \leq g \leq 1$ und eine Folge von elementaren Funktionen $0 \leq f_i \uparrow g$ (gleichmäßig!) folgt dann auch

$$\lim_{n\to\infty} \int_{T^{-1}(B)} g d \left(\sum_{j=1}^n \alpha_j P_j \circ T \right) = \int_{T^{-1}(B)} g dQ_*.$$

Wähle hier $g = \varphi_A$, um

$$\int_B \varphi_A dQ_* \circ T = \int_{T^{-1}(B)} \mathbb{1}_A dQ_* \qquad (ii)$$

für $B \in \mathcal{S}$ zu erhalten. Daraus folgt für jede $Q_* \circ T$ integrierbare Funktion $g : \mathcal{Z} \to \mathbb{R}$

$$\int \varphi_A g dQ_* \circ T = \int \mathbb{1}_A g \circ T dQ_* \qquad (iii)$$

mittels Approximation von g durch $\sum_j \gamma_j \mathbb{1}_{B_j}$ und dominierter Konvergenz. Angewandt auf $g_i = g_P \in \frac{dP \circ T}{dQ_* \circ T}$ folgt daraus mittels 1.0(ii) und (iii)

$$\int \varphi_A g_P dQ_* \circ T = \int \varphi_A dP \circ T = \int \mathbb{1}_A g_P \circ T dQ_*. \qquad (iv)$$

Dies ergibt zusammen mit (i) und $B = \mathcal{Z}$

$$P(A) = \int \mathbb{1}_A g_P \circ T dQ_*$$

für alle $A \in \mathcal{B}$.

■

Es fehlt noch die Rückrichtung. Zuerst allerdings zwei Beispiele.

Beispiel 1.6:

(a) Sei $\mathcal{X} = \{0,1\}^n$ eine Stichprobe eines n-fachen Münzwurfes. Setze $\mathcal{P} = \{Q_\alpha^n : 0 \leq \alpha \leq 1\}$.
Die Idee: Setze $T_n : \mathcal{X} \to \{0, \ldots, n\} = \mathcal{Z}, T_n(x) = x_1 + \ldots + x_n$. Dann ist

$$Q_\alpha^n(\{x\} \mid T_n^{-1}(j)) = \begin{cases} \binom{n}{j}^{-1} & j = T_n(x) \\ 0 & \text{sonst} \end{cases}$$

unabhängig von α!
Es gilt aber

$$Q_\alpha^n(\{x\}) = \alpha^{T_n(x)}(1-\alpha)^{n-T_n(x)}$$

sowie

$$Q_\alpha^n(T_n^{-1}(j)) = \binom{n}{j} \alpha^j (1-\alpha)^{n-j}.$$

(*b*) Sei $(\mathcal{X}, \mathcal{B}) = (\mathbb{R}^n, \mathcal{B}^n)$ und $\mathcal{P} = \{\mathcal{N}(\mu, 1)^n, \mu \in \mathbb{R}\}$ mit Lebesgue-Dichten

$$\begin{aligned}\varphi_{\mu,n}(x) &= \frac{1}{\sqrt{2\pi}^n} \exp\left(-\sum_{j=1}^{n}(x_j - \mu)^2/2\right) \\ &= \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2}\sum_{j}^{n} x_j^2\right) \exp\left(\mu \sum_{j}^{n} x_j - \frac{n\mu^2}{2}\right) \\ &= h_n(x) g_\mu(T_n(x))\end{aligned}$$

mit $T_n(x) = \sum_j x_j$. $T_n(x)$ ist suffizient für $\mathcal{N}(\mu, 1)^n$. Zeige dazu, dass die relativen Dichten $\varphi_{\mu,n}(x \mid T_n = \tau) = \Phi(x, \tau)$ unabhängig von μ sind.

Beispiel 1.7:
Für den Münzwurf gilt:

1) $T : \mathcal{X} \to \mathcal{X}, x \mapsto x$ ist immer suffizient nach dem Faktorisierungssatz.

2) $T_n : \mathcal{X} \to \{0, \ldots, n\}, x \mapsto x_1 + \ldots + x_n$ ist auch suffizient, mit größeren Niveaumengen $T_n^{-1}(z) \subset \mathcal{X}$. Ist hier $T_n^{-1}(z)$ „maximal“, d.h. T_n minimal suffizient?

Rückrichtung von Satz 1.5.
Es existiert $g_P : \mathcal{Z} \to [0, \infty)$, sodass $x \mapsto g_P(T(x))$ eine Version $\frac{dP}{dQ_*}, P \in \mathcal{P}$ ist. Daraus folgt $g_P \in \frac{dP \circ T}{dQ_* \circ T}$, denn aus $P(A) = \int_A g_P \circ T dQ_*$ folgt mit $A = T^{-1}(C), C \in \mathcal{S}$ nach Satz 1.0(ii)

$$P \circ T(C) = \int_C g_P dQ_* \circ T.$$

Für $A \in \mathcal{B}$ sei $\varphi_A \in \frac{dQ_{*,A} \circ T}{dQ_* \circ T}$, d.h. mittels 1.0(ii) erhält man

$$\int \varphi_A \mathbb{1}_B dQ_* \circ T = \int \mathbb{1}_B \circ T \mathbb{1}_A dQ_* \qquad (v)$$

für alle $B \in \mathcal{B}$ und somit auch für alle $Q_* \circ T$-integrierbaren Funktionen h mittels Approximation.
Daher gilt

$$\int \varphi_A h dQ_* \circ T = \int \mathbb{1}_A h \circ T dQ_*.$$

Angewandt auf $h = \mathbb{1}_B g_P$ folgt für alle $B \in \mathcal{S}$ und $P \in \mathcal{P}$

$$\begin{aligned}
\int_B \varphi_A dP \circ T &= \int \varphi_A \mathbb{1}_B dP \overset{1.0}{=} \int \varphi_A \circ T \mathbb{1}_B \circ T dP \\
&\overset{(1.6)}{=} \int \varphi_A \circ T \mathbb{1}_B \circ T g_P \circ T dQ_* \overset{(1.0)iii}{=} \int \varphi_A \mathbb{1}_B g_P dQ_* \circ T \\
&= \underbrace{\int \mathbb{1}_A \mathbb{1}_B \circ T g_P \circ T dQ_*}_{\int \mathbb{1}_A \mathbb{1}_B \circ T dP} = P(A \cap T^{-1}(B)) \\
&= P_A \circ T(B),
\end{aligned}$$

d.h. $\varphi_A \in \cap_{P \in \mathcal{P}} \frac{dP_A \circ T}{dP \circ T}$. ∎

Im Folgenden wird immer $\mathcal{P} \ll \mu, \mu$ σ-endlich sein.

Definition 1.8:
Eine suffiziente Statistik $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Z}, \mathcal{S})$ heißt minimal suffizient, falls für jede suffiziente Statistik $S : (\mathcal{X}, \mathcal{B}) \to (Y, \mathcal{F})$ eine messbare Funktion $H : (Y, \mathcal{F}) \to (\mathcal{Z}, \mathcal{S})$ existiert, sodass $T = H \circ S$ $\mathcal{P}$-f.ü. (d.h. P-f.ü. für alle $P \in \mathcal{P}$).

Das bedeutet: Zu jeder Niveaumenge $T^{-1}(z) \subset \mathcal{X}$ existiert ein $y \in H^{-1}(z)$ mit $S^{-1}(y) \subset T^{-1}(z)$ auf $\mathcal{P}$-Nullmengen, d.h. $T^{-1}(\mathcal{S})$ ist die gröbste suffiziente σ-Algebra aus $\mathcal{B}$, unter der die bedingte Erwartung von f gegeben $T^{-1}(\mathcal{S})$,

$$\mathbb{E}_P\left(f \mid T^{-1}(\mathcal{S})\right) : \mathcal{X} \to \mathbb{R},$$

nicht von $P \in \mathcal{P}$ abhängt.

Satz 1.9:
Falls $\mathcal{B}$ abzählbar erzeugt ist, existiert eine minimal suffiziente Statistik.

Beweis. [6], S. 13. ∎

Satz 1.10:
Falls $\mathcal{B}$ abzählbar erzeugt, $(\mathcal{Z}, \mathcal{S})$ ein polnischer Raum und $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Z}, \mathcal{S})$ suffizient für $\mathcal{P}$ ist, wobei $h_P(\cdot)$ eine Faktorzerlegung bezüglich $Q_* \in \mathcal{Q}$ ist, und gibt es eine abzählbare Teilmenge $\mathcal{P}_0 \subset \mathcal{P}$ mit

$$\forall z, z' \in T(\mathcal{X}) : h_P(z') = h_P(z)\ \forall P \in \mathcal{P}_0 \Rightarrow z = z',$$

so ist T minimal suffizient.

Beweis. [6], Theorem 1.4.4, S. 14. ∎

Beispiel:
Beim Münzwurf ist die Statistik $T_n(x) = \sum_j x_j$ (Zahl der Köpfe) auch minimal suffizient. Denn mit Hilfe von Satz 1.10 sieht man, dass

$$P_n(x,\alpha) = h_\alpha(T_n(x)), h_\alpha(z) = \binom{n}{z}\alpha^z(1-\alpha)^{n-z} = \binom{n}{z}\left(\frac{\alpha}{1-\alpha}\right)^z(1-\alpha)^n$$

Damit folgt

$$h_\alpha(z) = h_\alpha(z')\ \forall\alpha \in [0,1] \Leftrightarrow \left(\frac{\alpha}{1-\alpha}\right)^{z-z'} = 1$$

für alle $\alpha \in [0,1]$, also folgt daraus $z = z'$.

Definition 1.11:
Eine Familie $\mathcal{P}$ heißt (beschränkt) vollständig, falls für jede messbare (beschränkte) Funktion $f : (\mathcal{X}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B}^1)$ gilt:

$$\int f dP = 0\ \forall P \in \mathcal{P} \Rightarrow f = 0\,\mathcal{P}\text{ −f.ü.}$$

Beispiel 1.12:
Für den Münzwurf ist $T_n : \{0,1\}^n \to \{0,\ldots,n\}, (x_1,\ldots,x_n) \mapsto x_1 + \ldots + x_n$ eine suffiziente Statistik. $\mathcal{P} = \{P_\alpha^n \circ T_n : 0 \le \alpha \le 1\}$ ist beschränkt vollständig, denn es gilt

$$\begin{aligned}\int f \circ T_n dP_\alpha^n &= \int f dP_\alpha^n \circ T_n = \sum_{j=0}^n \binom{n}{j}\alpha^j(1-\alpha)^{n-j}f(j) = 0\ \forall\alpha\in[0,1]\\ &\Leftrightarrow \sum_{j=0}^n g_n(j)\alpha^j = 0 \text{ mit } g_n(j) = \sum_{k+l=j}\binom{n}{l}\binom{n-l}{k}(-1)^k f(l)\\ &\Leftrightarrow g_n(j) = 0\ \forall j.\end{aligned}$$

Nun gilt für $j = 0$:

$$0 = g_n(0) = \binom{n}{0}^2 f(0) \Rightarrow f(0) = 0$$

und induktiv für alle $f(j), j = 1,\ldots,n$.

Proposition 1.13:
Sei $T : (\mathcal{X},\mathcal{B}) \to (\mathcal{Z},\mathcal{S})$, wo $(\mathcal{Z},\mathcal{S})$ ein polnischer Raum, suffizient für $\mathcal{P}$. Falls $\mathcal{P}\circ T$ beschränkt vollständig ist, so ist T minimal suffizient.

Beweis. Sei $S : (\mathcal{X}, \mathcal{B}) \to (Y, \mathcal{F})$ eine andere suffiziente Statistik. Dann gibt es zu $A \in \mathcal{B}$ ein $\varphi_A : Y \to [0,1]$ mit

$$P(A) = \int \varphi_A \circ S dP$$

für alle $P \in \mathcal{P}$. Da $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Z}, \mathcal{S})$ suffizient ist, so existiert eine bedingte Erwartung von $\varphi_A \circ S : \mathcal{X} \to [0,1]$ gegeben T bezüglich P, $\psi_A : \mathcal{Z} \to [0,1]$, welche nicht von P abhängt. Insbesondere gilt

$$P(A) = \int \varphi_A \circ S dP = \int \psi_A \circ T dP$$

für alle $P \in \mathcal{P}$. Für $A \coloneqq T^{-1}(C), C \in \mathcal{S}$ gilt daher

$$\int \psi_{T^{-1}(C)} \circ T dP = P(T^{-1}(C)) \overset{\text{def}}{=} \int \mathbb{1}_C \circ T dP$$

d.h.

$$\int \psi_{T^{-1}(C)} \circ T - \mathbb{1}_C \circ T dP = \int \varphi_{T^{-1}(C)} - \mathbb{1}_C dP \circ T = 0 \; \forall P \in \mathcal{P}\,.$$

Daraus folgt $\psi_{T^{-1}(C)} = \mathbb{1}_C \; Q_* \circ T$-f.s. wegen der beschränkten Vollständigkeit. Also ist $\mathbb{1}_C$ die bedingte faktorisierte Erwartung von $\varphi_{T^{-1}(C)} \circ S$ gegeben T bezüglich Q_*. Es gibt nach Integration also eine Funktion $\chi : \mathcal{X} \to [0,1]$ bezüglich Q_* mit $0 \leq \chi \leq 1$ Q_*-f.ü. und

$$\int \mathbb{1}_{C_0} \mathbb{1}_C dQ_* \circ T = \int_{T^{-1}(C_0)} \chi dQ_*$$

sowie

$$0 = \int_{T^{-1}(C_0)} (\mathbb{1}_C \circ T - \chi) dQ_* \; \forall C_0 \in \mathcal{S},$$

und angewandt auf $C_0 = C$ und $C_0 = C^c$ folgt hieraus $\mathbb{1}_C \circ T = \chi$ Q_*-f.ü., denn

$$\int \underbrace{\mathbb{1}_{T^{-1}(C)}(1-\chi)}_{\geq 0} dQ_* = 0 = \int \underbrace{\mathbb{1}_{T^{-1}(C^c)}(0-\chi)}_{\leq 0} dQ_*.$$

Daher ist $\chi = \varphi_{T^{-1}(C)} \circ S = \mathbb{1}_C \circ T \; = \mathbb{1}_{T^{-1}(C)} Q_*$-f.ü. und $T^{-1}(\mathcal{S}) \subset S^{-1}(\mathcal{F})$, d.h. für $C \in \mathcal{S}$ gibt es ein $A \in \mathcal{B}$ mit $Q_*(T^{-1}(C) \backslash S^{-1}(A)) = 0$.
Es gibt nun eine Folge von $T^{-1}(\mathcal{S}) - \mathcal{S}-$messbaren Funktionen $T_n : \mathcal{X} \to \mathcal{Z}$ mit nur endlich vielen Werten mit $T_n \to T \; \mathcal{P}-$f.ü. ([6], Lemma 1.10.6, S. 56). Daraus lässt sich dann eine messbare Funktion

$$H : (\mathcal{Y}, \mathcal{F}) \to (\mathcal{Z}, \mathcal{S})$$

mit $T = H \circ S$ konstruieren (vergleiche auch MIT, Proposition 5.13). ∎

Bemerkung:
Aus $\mathcal{P}$ vollständig folgt im Allgemeinen nicht, dass $\mathcal{P}_n = \{P^n, P \in \mathcal{P}\}$ vollständig ist.
Ist $X = (X_1, \ldots, X_n) \sim P^n$ eine Stichprobe, wobei $P \in \mathcal{P}$, so gilt natürlich, dass $X_\pi = (X_{\pi(1)}, \ldots, X_{\pi(n)}) \sim P^n$ für jede Permutation π von $\{1, \ldots, n\}$.
Eine Statistik $S_n : \mathcal{X}^n \to \mathcal{Z}$, die maximal invariant unter Permutationen ist, d.h. $S_n(x') = S_n(x_\pi) \Leftrightarrow x_\pi = x'$ für eine Permutation π heißt Ordnungsstatistik.
Für $\mathcal{X} = \mathbb{R}$ ist das klar: $S_n(x_1, \ldots, x_n) = (x_{1:n}, \ldots, x_{n:n}) \in \mathbb{R}^n$ mit $(x_1, \ldots, x_n) = (x_{\cdot:n})_\pi$ mit π Permutation und $x_{1:n} \leq \ldots \leq x_{n:n}$.
Offensichtlich gilt

a) S_n ist suffizient für $\{P^n : P \in \mathcal{P}\}$.

b) $f : \mathcal{X}^n \to \mathcal{Z}$ mit $f = g \circ S_n$ mit $g : \mathcal{X}^n \to \mathcal{Z} \Leftrightarrow f$ ist symmetrisch.

Definition 1.14:
Eine Familie $\{P^n : P \in \mathcal{P}\}$ heißt symmetrisch (beschränkt) vollständig, falls für jede symmetrische (beschränkte) Funktion $f_n : \mathcal{X}^n \to \mathbb{R}$ gilt

$$\int f_n dP^n = 0 \ \forall P \in \mathcal{P} \Rightarrow f_n = 0 \ P^n - \text{f.ü.} \ \forall P \in \mathcal{P}\,.$$

Bemerkung:

i) Ist $\{P^n : P \in \mathcal{P}\}$ symmetrisch vollständig für ein $n \geq 1$, so ist $\mathcal{P}$ vollständig. Dazu wähle $f_n(x_1, \ldots, x_n) = \sum_j f(x_j)$, wobei f wie in Definition 1.11.

ii) Es braucht „große“ Familien $\mathcal{P}$, damit $\{P^n : P \in \mathcal{P}\}$ symmetrisch vollständig ist.

Satz 1.15:
Sei $\mathcal{P}$ vollständig und abgeschlossen unter Konvexkombinationen. Dann ist $\{P^n : P \in \mathcal{P}\}$ symmetrisch vollständig für jedes $n \in \mathbb{N}$.

Beweis. [6], Satz 1.5.10, S. 19-21.
Idee: Sei $\mathcal{P}$ vollständig, dann:

i) $\{P_1 \otimes \ldots \otimes P_n, P_i \in \mathcal{P}\}$ ist vollständig.

ii) $\int f_n d\left(\sum_j \alpha_j P_j\right)^k = 0$ für alle Konvexkombinationen α_j, dann folgt

$$\sum_{(i_1,\ldots,i_n),\text{Permutation}} \int f_n(dP_{i_1} \times \ldots \times P_{i_n})\alpha_1 \cdots \alpha_n = 0$$

d.h. $\int f_n dP_1 \times \ldots \times P_n = 0 \Rightarrow f_n = 0$ wegen Teil *i)*.

■

1.2 Anwendung: Exponentielle Familien

Definition 1.16:
Eine Familie von W-Maßen $\mathcal{P}$ auf $(\mathcal{X}, \mathcal{B})$ heißt exponentiell, falls $\mathcal{P} \ll \mu, \mu$ σ-endlich und

$$\frac{dP}{d\mu}(x) = c(P)g(x)\exp\left(\sum_{j=1}^{n} a_j(P)T_j(x)\right),$$

wo $g, T_j : (\mathcal{X}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B}^1)$ messbar und $g(x) \geq 0, a_j \in \mathbb{R}$.

Bemerkung 1.17:

(*i*) $g(x)$ kann stets als Konstant 1 gewählt werden mittels

$$\tilde{\mu}(A) = \int_A g(x)\mu(dx)$$

als neuem dominierenden Maß für $\mathcal{P}$.

(*ii*) Für $P, P' \in \mathcal{P}$ gilt $P \ll P'$ und $P' \ll P$, d.h. P, P' sind äquivalent.

(*iii*) Wir nehmen an:

1) $\mathcal{P} \ni P \mapsto (a_1(P), \ldots, a_n(P))$ hat maximalen Rang, d.h.

$$\sum_j a_j(P)c_j = c_0 \ \forall P \in \mathcal{P} \Longrightarrow c_0 = c_1 = \ldots = c_n = 0.$$

2) Die Funktionen $T_j(x)$ sind affin unabhängig, d.h.

$$\sum_j c_j T_j(x) = c_0 \ \mu\text{-f.s.} \Longrightarrow c_0 = c_1 = \ldots = c_n = 0.$$

3) Ist $\mathcal{P}$ exponentiell, so auch $\mathcal{P}^n = \{P^n : P \in \mathcal{P}\}$ mit

$$\begin{aligned}\frac{dP^n}{d\mu^n}(x_1, \ldots, x_n) &= \prod_{r=1}^{n} c(P)\exp\left(\sum_{j=1}^{m} a_j(P)T_j(x_r)\right) \\ &= c(P)^n \exp\left(\sum_{j=1}^{m} a_j(P)\left(\sum_{r=1}^{n} T_j(x_r)\right)\right)\end{aligned}$$

Beispiel 1.18:

a) Die Poissonverteilung $P_\lambda(X = k) = \frac{\lambda^k}{k!}\exp(-\lambda)$ für $\lambda > 0$.

b) Die Binomialverteilung mit $B_\alpha(X = k) = \binom{n}{k}\alpha^k(1-\alpha)^{n-k}$ wobei

$$\underbrace{\binom{n}{k}}_{g(k)} \exp\left(k \underbrace{\left(\frac{\alpha}{1-\alpha}\right)}_{a_j(P)}\right) \underbrace{(1-\alpha)^n}_{c(P)}.$$

c) Die Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ mit Lebesgue-Dichte

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right) \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

wobei hier $a_1(\mathcal{N}(\mu, \sigma^2)) = -\frac{1}{2\sigma^2}$ und $a_2(\mathcal{N}(\mu, \sigma^2)) = \frac{\mu}{\sigma^2}$, $T_1(x) = x^2$ und $T_2(x) = x$.

d) Die χ^2-Verteilung mit $f \in \mathbb{N}$ Freiheitsgraden und Lebesgue-Dichte

$$\chi^2_f(x) = \frac{x^{f-1}}{2^f\Gamma(\frac{f}{2})} \exp\left(-\frac{x}{2}\right) \mathbb{1}_{\{x>0\}},$$

wobei Γ die Gamma-Funktion bezeichnet.

e) Die Beta-Verteilung mit Parametern $p, q > 1$ und der Dichte

$$x^{p-1}(1-x)^{q-1}\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\mathbb{1}_{[0,1]}.$$

f) Exponentielle Verteilung mit Dichte

$$\vartheta^{-1} \exp\left(-\frac{x}{\vartheta}\right) \mathbb{1}_{\{x\geq 0\}}$$

mit $\vartheta > 0$.

g) Nicht exponentiell ist z.B. die Cauchy-Verteilung mit Lebesgue-Dichte

$$\frac{1}{\pi}\frac{1}{1+(x-\vartheta)^2}, \vartheta \in \mathbb{R}\,.$$

h) Ebenfalls nicht exponentiell ist die logistische Verteilung mit Dichte

$$\frac{\exp(-(x-\vartheta)}{1+\exp(-(x-\vartheta))^2}, \vartheta \in \mathbb{R}\,.$$

Die Bildvektoren der Abbildung

$$a : \mathcal{P} \to \mathbb{R}^m, P \mapsto (a_1(P), \ldots, a_m(P))$$

für eine exponentielle Familie heißen natürliche Parameter. Es gilt

$$\Theta := \{a(P) : P \in \mathcal{P}\} \subset \Theta_*$$

wobei Θ_* die maximale Teilmenge des $\mathbb{R}^m$, sodass die μ-Dichte existiert.

Proposition 1.19:
Θ_* *ist konvex.*

Beweis. Seien $\vartheta, \vartheta' \in \Theta_*$ mit $\vartheta = (\vartheta_1, \ldots, \vartheta_m), \vartheta' = (\vartheta'_1, \ldots, \vartheta'_m)$ und $\alpha \in [0, 1]$. Dann gilt

$$\begin{aligned}
& \int \exp\left(\sum_{j=1}^m (\alpha\vartheta_j + (1-\alpha)\vartheta'_j)T_j\right) d\mu \\
= & \int \exp\left(\sum_{j=1}^m \vartheta_j T_j\right)^\alpha \exp\left(\sum_{j=1}^m \vartheta'_j T_j\right)^{1-\alpha} d\mu \\
\leq & \left(\int \exp\left(\sum_j \vartheta_j T_j\right) d\mu\right)^\alpha \left(\int \exp\left(\sum_j \vartheta'_j T_j\right) d\mu\right)^{1-\alpha} < \infty.
\end{aligned}$$

■

Satz 1.20:
Sei $f : \mathcal{X} \to \mathbb{R}$ beschränkt und messbar. Dann gilt

i) $(z_1, \ldots, z_m) \mapsto \int f \exp\left(\sum_j z_j T_j\right) d\mu$ *ist eine analytische Funktion (d.h. holomorph in jeder Variablen) auf $G_{\Theta_*} \subset \mathbb{C}^n$, wo*

$$G_{\Theta_*} = \{(\mu_1 + i\vartheta_1, \ldots, \mu_m + i\vartheta_m) : (\mu_1, \ldots, \mu_n) \in \mathrm{Int}(\Theta_*), \vartheta_j \in \mathbb{R}\}$$

und wo Int *das Innere bezeichnet.*

ii) *Die Ableitungen von i) können durch Differentation unter dem Integral gewonnen werden.*

iii) *Ist $\{P_\vartheta, \vartheta \in \Theta\}$ eine exponentielle Familie, Θ natürlich. Falls $P_{\vartheta_n} \Rightarrow P_\vartheta$ mit $\vartheta_n, \vartheta \in \Theta$, dann gilt $\vartheta_n \to \vartheta$.*

Beweis. Es reicht den Fall $m = 1$ zu zeigen:

$$\chi(\alpha) = \int f \exp(\alpha T) d\mu$$

mit $\alpha \in \text{Int}(\Theta_*)$. Es existiert ein $\varepsilon > 0$ sodass $\chi(\alpha)$ für $\alpha = a + ib$ und $|a - a_0| < \varepsilon$ existiert. Sei $|c_n| < \varepsilon, c_n \to 0, c_n \in \mathbb{C}, \alpha_0 = a_0 + ib$. Dann gilt

$$\frac{1}{c_n}\left(\chi(\alpha_0 + c_n) - \chi(\alpha_0)\right) = \int \underbrace{\frac{1}{c_n} \exp\left(c_n T - 1\right)}_{h_n} f d\mu \to \int f T \exp(\alpha_0 T) d\mu$$

mit dem Satz von Lebesgue, wobei gilt

$$\begin{aligned} |h_n| &\leq \left| \sum_{j=1}^{\infty} \frac{c_n^{j-1} T^j}{j!} \right| \leq \sum_{j=1}^{\infty} \frac{\varepsilon^{j-1} |T|^j}{j!} \\ &\leq \frac{1}{\varepsilon} \exp(\varepsilon |T|) \leq \frac{1}{\varepsilon} \left(\exp(\varepsilon T) + \exp(-\varepsilon T)\right). \end{aligned}$$

(iii) : [6], Proposition 1.6.8, S. 24. ∎

Bemerkung:
Da $a \mapsto c(a) := \int g \exp\left(\sum_{j=1}^m a_j T j\right) d\mu$ in $\text{Int}(\Theta_*)$ stetig ist, gilt dies auch für die μ Dichten

$$a \mapsto p_a(x) = \frac{1}{c(a)} g \exp\left(\sum_{j=1}^m a_j T_j(x)\right)$$

Satz 1.21:
Sei

$$\frac{dP}{d\mu}(x) = \exp\left(\sum_{j=1}^m a_j(P) T_j(x)\right) h(x) c(P) \text{ für } P \in \mathcal{P}$$

mit $h \geq 0$ eine exponentielle Familie und $\mathcal{P}^n := \{P^n : P \in \mathcal{P}\}$. Falls $\text{Int}(a(\mathcal{P})) = \text{Int}(\Theta) \neq \emptyset$ ist, dann ist mit $T(x_1, \ldots, x_n) = \left(\sum_{r=1}^n T_j(x_r)\right)_{j=1,\ldots,m}$ die Familie $\mathcal{P}^n \circ T$ vollständig und T ist minimal suffizient.

Beweis. Es reicht die Behauptung für $h \equiv 1$ und $n = 0$ zu zeigen. Setze

$$g(\vartheta_1, \ldots, \vartheta_m) := \int f(t_1, \ldots, t_m) \exp\left(\sum_{j=1}^m \vartheta_j t_j\right) \mu \circ T(dt_1, \ldots, dt_m).$$

Sei $g(\vartheta_1, \ldots, \vartheta_m) = 0$ für alle $(\vartheta_1, \ldots, \vartheta_m) \in I \subset \text{Int}(\Theta)$, wobei $I = (-\varepsilon, \varepsilon)^m$, wo OE $0 \in \text{Int}(\Theta)$.
Da g auf $(-\varepsilon, \varepsilon)^m + i\,\mathbb{R}^m$ nach Satz 1.20 analytisch fortsetzbar ist folgt mittels

des Eindeutigkeitssatzes für analytische Funktionen, dass $g_j(\vartheta_j) \equiv 0$ für $\vartheta_j \in (-\varepsilon, \varepsilon) + i\,\mathbb{R}$. Mit der Zerlegung von $f = f_+ - f_-$ in positiven und negativen Teil setze

$$g_\pm(\vartheta) = \int f_\pm \exp\left(\sum_{j=1}^m \vartheta_j t_j\right) d\mu \circ T.$$

Sei $c \coloneqq g_+(\vartheta) > 0$. Falls $c = 0$ so folgt $f_+ = 0$ $\mu\circ T-$f.ü. und $f_- = 0$ $\mu\circ T-$f.ü. Ansonsten seien $Q_\pm$ die W-Maße mit $\mu \circ T-$Dichten $\frac{1}{c} f_\pm(t)$. Dann

$$\int \exp\left(\sum_{j=1}^m \vartheta_j t_j\right) dQ_+ = \int \exp\left(\sum_{j=1}^m \vartheta_j t_j\right) dQ_-$$

für $\vartheta_j \in (-\varepsilon, \varepsilon) + i\,\mathbb{R}$, d.h. insbesondere folgt für $\vartheta_j = -i\mu_j, \mu_j \in \mathbb{R}$, d.h. Gleichheit der multivariaten charakteristischen Funktion von $Q_\pm$. Nach Wahrscheinlichkeitstheorie I folgt $Q_+ = Q_-$ und damit $f_+ = f_-$ $\mu \circ T-$f.ü., d.h. $f = f_+ - f_- = 0$ $\mu \circ T$-f.ü. ∎

2 Parametrische Schätzfunktion

In diesem Kapitel sei $\mathcal{P}$ eine Familie von W-Maßen mit $\mathcal{P} \ll \mu$, wobei μ σ-endlich, und $\kappa : \mathcal{P} \to (Y, \mathcal{F})$ sei ein Operator.

Definition 2.1:

(*a*) Ein Schätzer für κ ist eine messbare Abbildung $\widehat{\kappa} : (\Omega, \mathcal{A}) \to (Y, \mathcal{F})$. $\widehat{\kappa}(x)$ heißt Schätzung.

(*b*) Ein randomisierter Schätzer ist ein Markov-Kern K auf $\Omega \times \mathcal{F}$. Falls $x \in \Omega$ beobachtet wird, so ist die Schätzung eine Realisation eines Zufallsexperiments mit Verteilung $K(x, \cdot)$.

$\widehat{\kappa}$ kann dabei durch den Übergangskern $(x, A) \mapsto \mathbb{1}_A(\widehat{\kappa}(x))$ reproduziert werden. In Anwendungen ist üblicherweise $(\Omega, \mathcal{A})$ ersetzt durch $(\Omega^n, \mathcal{A}^n)$, d.h. n unabhängigen, identisch verteilten Beobachtungen.

Im Folgenden werden wir nicht randomisierte Schätzer betrachten.
Vergleichskonzepte für Schätzer
Seien Stichproben $(\Omega^n, \mathcal{A}^n)$ und W-Maße $P \in \mathcal{P}$ gegeben.
1. Forderung: Vergleich von Schätzern $\widehat{\kappa}_n$ anhand der Verteilungen $P^n \circ \widehat{\kappa}_n^{-1}, P \in \mathcal{P}$.
2. Forderung: Vergleiche die Konzentration von $P^n \circ \widehat{\kappa}_n^{-1}$ um den wahren Wert $\kappa(P)$ für alle $P \in \mathcal{P}$.
3. Forderung: Suche Schätzer, deren „Qualität“ abhängig von $P \in \mathcal{P}$ keine großen Varianz aufweisen, in dem Sinne, dass sie alle W-Maße „gleich behandeln“.
Zur 1. Forderung: Prinzipien wie „Schneide $\alpha\%$ der größten oder kleinsten Beobachtung ab und bilde den Mittelwertschätzer“ oder „Finde das Maximum der Likelihood-Funktion gegeben die Stichprobe“ sind fragwürdig ohne die Untersuchung der Verteilungen.
Zur 2. Forderung: Bezeichne mit C_* die Menge aller offenen, konvexen und beschränkten Mengen aus dem $\mathbb{R}^k$, die symmetrisch um 0 sind. Für $k = 1$ sind dies um 0 symmetrische Intervalle.

Definition 2.2:

(*i*) $\widehat{\kappa}_1$ heißt besser als $\widehat{\kappa}_2$ um $\kappa(P)$ konzentriert, falls

$$P^n\left(\widehat{\kappa}_1 - \kappa(P) \in C\right) > P^n\left(\widehat{\kappa}_2 - \kappa(P) \in C\right) \ \forall C \in C_*.$$

(*ii*) Sei $L : \mathcal{P} \times \mathbb{R}^k \to [0, \infty)$ und $\tau \mapsto L(P, \tau)$ messbar für alle $P \in \mathcal{P}$. L heißt Verlustfunktion, falls $L(P, \kappa(P)) = 0$. L heißt subkonvex, falls

$$\{u \in \mathbb{R}^k : L(P, u) \leq r\}$$

konvex ist für alle $r \geq 0, P \in \mathcal{P}$ und konvex, falls $\tau \mapsto L(P,\tau)$ konvex ist für jedes $P \in \mathcal{P}$.
Hier heiße $\widehat{\kappa}_1$ von geringerem Risiko als $\widehat{\kappa}_2$, falls mit

$$R(P,\widehat{\kappa}) \coloneqq \int L(P,\widehat{\kappa}(\omega))P^n(d\omega)$$

gilt

$$R(P,\widehat{\kappa}_1) < R(P,\widehat{\kappa}_2).$$

$R(P,\widehat{\kappa})$ heißt Risiko oder mittlerer Verlust von $\widehat{\kappa}$ unter P.

Proposition 2.3:
Es sind äquivalent:

i) $P(\widehat{\kappa}_1 - \kappa(P) \in C) > P(\widehat{\kappa}_2 - \kappa(P) \in C)$ *für alle* $C \in C_*$.

ii) *Für jede subkonvexe, symmetrische Verlustfunktion* $L(P,\cdot)$ *ist die Verteilung des Verlusts* $P \circ L(P,\widehat{\kappa}_1)$ *mehr um* 0 *konzentriert als die Verteilung* $P \circ L(P,\widehat{\kappa}_2)$.

iii) *Für jede subkonvexe, symmetrische Verlustfunktion* $L(P,\cdot)$ *ist das Risiko von* $\widehat{\kappa}_1$ *kleiner als das von* $\widehat{\kappa}_2$.

Beispiel:

a) $L(P,\tau) = c_P \left\|\tau - \kappa(P)\right\|^2$ ist konvex.

b) $L(P,\tau) = 1 - \mathbb{1}_C(\tau - \kappa(P))$ mit $C \in C_*$ ist subkonvex.

c) Für $k = 1$ ist die Funktion $L(P,\tau) = |\tau - \kappa(P)|$ konvex.

Bemerkung:

- Konvexe Verlustfunktionen sind subkonvex.
- *ii)* zeigt, dass Konzentration und Verlust komplementäre Funktionen sind.
- Für randomisierte Schätzer $\widehat{\kappa}$ auf $\mathcal{X}^n \times \mathcal{B}^k$ wird der mittlere Verlust definiert durch

$$R(P,\widehat{\kappa}) = \int\int L(P,t)\widehat{\kappa}(\omega,dt)P^n(d\omega).$$

Beweis.
i) $\Rightarrow$ *ii)* : Die Menge $C_r = \{u \in \mathbb{R}^p : L(P,u) \leq r\}$ ist messbar, konvex und symmetrisch um $\kappa(P)$. Da $P(L(P,\widehat{\kappa}) \leq r) = P(\widehat{\kappa} \in C_r)$ folgt die Behauptung.
ii) $\Rightarrow$ *iii)* : Es gilt

$$R(P,\widehat{\kappa}) = \int L(P,\widehat{\kappa})dP = \int L(P,s)\ P \circ \widehat{\kappa}(ds) = \int_0^\infty P \circ \widehat{\kappa}(L(P,\cdot) > r)\ dr$$

woraus die Behauptung folgt.
iii) $\Rightarrow$ *i)* : $L(P,\cdot) \coloneqq 1 - \mathbb{1}_C$ ist symmetrisch um $\kappa(P)$ und subkonvex, falls C symmetrisch um $\kappa(P)$ und konvex. ∎

Definition 2.4:
Eine Funktion $f : \mathbb{R}^p \to [-\infty, \infty]$ heißt superkonvex / unimodal, falls die Menge $\{x \in \mathbb{R}^p : f(x) \geq r\}$ konvex ist für alle $r \in \mathbb{R}$.

Proposition 2.5:

i) Jede konkave Funktion ist superkonvex.

ii) Ist $f : \mathbb{R}^p \to \mathbb{R}$ superkonvex und $m : \mathbb{R} \to \mathbb{R}$ monoton wachsend, so ist $m \circ f$ superkonvex.

Satz 2.6 (Satz von Andersen, Faltungsformel)**:**
Sei $f : \mathbb{R}^p \to [0, \infty)$ superkonvex.

i) Ist $f(x) = f(-x)$ λ^p−f.ü. und $\int f d\lambda^p < \infty$ und $C \in \mathcal{B}^p$ konvex und symmetrisch um 0, so ist

$$r \mapsto \psi_{x_0}(r) \coloneqq \int_{C + r x_0} f(x)\lambda^p(dx), r > 0$$

monoton fallend für alle $x_0 \in \mathbb{R}^p$.

ii) Sei f zusätzlich eine Dichte und C wie in (i). Dann gilt für das Wahrscheinlichkeitsmaß $Q(A) \coloneqq \int_A f d\lambda^p$ und ein beliebiges W-Maß P auf $(\mathbb{R}^p, \mathcal{B}^p)$

$$Q(C) \geq Q * P(C) = \int Q(C - y)P(dy)$$

Zum Beweis benötigen wir das folgende Lemma.

Lemma 2.7 (Ungleichung von Brunn-Minkowski)**:**
Seien $A, B \neq \emptyset$ beschränkte, messbare und konvexe Mengen im $\mathbb{R}^p$. Dann gilt

$$\lambda^p(\alpha A + (1-\alpha)B)^{1/p} \geq \alpha\lambda^p(A)^{1/p} + (1-\alpha)\lambda^p(B)^{1/p}.$$

Beweis. Siehe [3]. ∎

Beweis zu Satz 2.6. ii): Es gilt

$$Q * P(C) = \int \underbrace{Q(C-y)}_{\leq Q(C)} P(dy) \leq \int Q(C)P(dy) = Q(C)$$

unter Zuhilfenahme von Teil i).
i): Sei nun $u > 0$ und $D_u := \{x \in \mathbb{R}^p : f(x) \geq u\}$ und C eine beliebige, symmetrische, konvexe Menge. Dann ist $(C - x_0) \cap D_u$ die Spiegelung von $(C + x_0) \cap D_u$, d.h. es gilt

$$\lambda^p((C - x_0) \cap D_u) = \lambda^p((C + x_0) \cap D_u). \qquad (*)$$

Setze nun für $r \in [0, 1]$ $\alpha := \frac{1+r}{2}$, so gilt

$$C + rx_0 \supset \alpha C + (1 - \alpha)C + rx_0 = \alpha(C + x_0) + (1 - \alpha)(C - x_0),$$

woraus wegen der Konvexität von C und D_u

$$(C + rx_0) \cap D_u \supset \alpha\left((C + x_0) \cap D_u\right)) + (1 - \alpha)\left((C - x_0) \cap D_u\right) \qquad (**)$$

folgt. Mit der Brunn-Minkowski-Ungleichung folgt mit Hilfe von $(*)$ und $(**)$ nun

$$H(u) := \lambda^p((C + rx_0) \cap D_u) \geq \lambda^p((C + x_0) \cap D_u) =: H^*(u),$$

denn

$$\begin{aligned} H(u) &\geq \lambda^p((\alpha(C + x_0) + (1 - \alpha)(C - x_0)) \cap D_u) \\ &\geq \left[\alpha\lambda^p((C + x_0) \cap D_u))^{1/p} + (1 - \alpha)\lambda^p((C - x_0) \cap D_u)^{1/p}\right]^p \\ &= \left[(\alpha + (1 - \alpha))\lambda^p((C + x_0) \cap D_u)^{1/p}\right]^p = \lambda^p((C + x_0) \cap D_u) \\ &= H^*(u). \end{aligned}$$

Folglich gilt

$$\begin{aligned} \int_{C+rx_0} f d\lambda^p - \int_{C+x_0} f d\lambda^p &= -\int_0^\infty u dH^*(u) + \int_0^\infty u dH(u) \\ &= \int_0^\infty u d(H^*(u) - H(u)) \\ &\overset{\text{p. I.}}{=} \underbrace{\lim_{b\to\infty} b(H^*(b) - H(b))}_{=0} - \underbrace{\lim_{a\to 0} a(H^*(a) - H(a))}_{=0} \\ &+ \int_0^\infty \underbrace{H(u) - H^*(u)}_{\geq 0} \lambda^p(du). \end{aligned}$$

Die ersten zwei Terme verschwinden, woraus die Monotonie folgt. Ebenfalls folgt aus $(*)$ die Monotonie in $|r|$. ∎

Beispiel 2.8:
Sei für eine Konvarianz-Matrix $\Sigma > 0$ $\mathcal{N}(0, \Sigma)$ die k-dimensionale Normalverteilung. Diese ist symmetrisch & superkonvex.

Satz 2.9:
Für $\Sigma_2 > 0$ gilt genau dann $\mathcal{N}(0, \Sigma_1)(C) \geq \mathcal{N}(0, \Sigma_2)(C)$ für alle konvexen, um 0 symmetrischen Mengen, wenn $\Sigma_1 \leq \Sigma_2$, d.h. $\Sigma_2 - \Sigma_1 \geq 0$.

Beweis. Faltungsformel: $\mathcal{N}(0, \Sigma_2) = \mathcal{N}(0, \Sigma_1) * \mathcal{N}(0, \Sigma_2 - \Sigma_1)$. ∎

Zur 3. Forderung: Gleichbehandlung aller W-Maße aus $\mathcal{P}$

Definition 2.10:

i) Klassisches Konzept: Ein Schätzer $\widehat{\kappa}$ für κ heißt erwartungstreu (mean unbiased), falls
$$\int \widehat{\kappa} dP = \kappa(P)$$
für alle $P \in \mathcal{P}$.

ii) Alternatives Konzept: Ein Schätzer $\widehat{\kappa}$ von κ heißt mediantreu (median unbiased), falls für alle $P \in \mathcal{P}$ gilt:
$$P(\widehat{\kappa} \geq \kappa(P)) \geq \frac{1}{2} \leq P(\widehat{\kappa} \leq \kappa(P)).$$

iii) Verlustfunktionsabhängiges Konzept: Sei L eine Verlustfunktion der Form $(P, t) \mapsto L(\kappa(P), t)$. Dann heißt der Schätzer $\widehat{\kappa}$ für κ L-treu, falls der mittlere Verlust $y \mapsto \int L(y, \widehat{\kappa}(x)) P(dx)$ in $\kappa(P)$ sein Minimum für alle $P \in \mathcal{P}$ annimmt.

Proposition 2.11:
Aus $L(\kappa, t) = |\kappa - t|^p$ erhalten wir Erwartungstreue für $p = 2$ und Mediantreue für $p = 1$.

Beweis. $p = 2$: Sei $Q := P \circ \widehat{\kappa}^{-1}, P \in \mathcal{P}$ und $\mu = \mathbb{E}_P \widehat{\kappa}$ definiert. Dann gilt
$$MQE = \int (x - y)^2 Q(dy) = \underbrace{\int (y - \mu)^2 Q(dy)}_{\text{Varianz von } \widehat{\kappa}} + \underbrace{(\mu - \kappa)^2}_{\text{Bias}},$$
woraus die Behauptung folgt.
$p = 1$: Sei der Median von $Q := P \circ \widehat{\kappa}^{-1}$ OE in Null und $\kappa > 0$. Dann gilt
$$|y| \leq \begin{cases} |y - \kappa| + \kappa & y > 0 \\ |y - \kappa| - \kappa & y \leq 0 \end{cases}$$

d.h. es gilt

$$\int |y|\, Q(dy) \leq \int |y-\kappa|\, Q(dy) + \kappa(Q(0,\infty) - Q((-\infty,0])) \leq \int y - \kappa Q(dy)$$

wobei $\kappa(Q((0,\infty)) - Q((-\infty,0])) \leq 0$, da 0 der Median von Q ist. ∎

Diskussion:

i) Das klassische Konzept ist vernünftig bei unabhängigen Wiederholungen $j = 1, \ldots, N$ der Schätzung $\widehat{\kappa}(x_j)$. Hier ist es sinnvoll, Erwartungstreue zu fordern, denn nach dem Gesetz der großen Zahlen gilt P-stochastisch

$$\frac{1}{n}\left(\sum_{j=1}^{n} \widehat{\kappa}(x_j)\right) \to \int \widehat{\kappa} dP,$$

d.h. dies wäre fatal für $\int \widehat{\kappa} d\,\mathbb{P} \neq \kappa(\mathbb{P})$.

ii) In den meisten Fällen ist Mediantreue adäquat, d.h. das Zentrum der Wahrscheinlichkeitsverteilung von $P^n \circ \widehat{\kappa}_n$ bei $\kappa(P)$ zu haben. Z.B. will man statt $\kappa(P) \in \mathbb{R}$ $(h \circ \kappa)(P)$ für eine strikt monotone Funktion $h : \mathbb{R} \to \mathbb{R}$ schätzen. Dann ist $h \circ \widehat{\kappa}$ mediantreu für $h \circ \kappa$, falls $\widehat{\kappa}$ mediantreu für κ ist. Dies gilt im Allgemeinen nicht für Erwartungstreue.

Beispiel:

(*a*) $\mathcal{P} = \{\mathcal{N}(\vartheta, 1)^n, \vartheta \in \mathbb{R}\}$. Der Schätzer $\bar{x}_n = \frac{1}{n}(x_1 + \ldots + x_n)$ ist erwartungstreu und mediantreu, denn

$$\int \bar{x}_n d\mathcal{N}_{\vartheta,1}^n = \frac{1}{n}\sum_j \int x_j d\mathcal{N}_{\vartheta,1}^n = \frac{1}{n}\sum_j \vartheta = \vartheta.$$

Da $\mathcal{N}_{\vartheta,1}^n \circ \bar{x}_n = \mathcal{N}(\vartheta, 1/n)$ gilt $\mathcal{N}_{\vartheta,1/n}(\bar{x}_n \leq \vartheta) = \frac{1}{2}$.

(*b*) Setze $B_n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 \leq x_n\}$ und definiere

$$\tilde{x}_n := \begin{cases} 2\bar{x}_n & x_1, \ldots, x_n \in B_n \\ 0 & \text{sonst} \end{cases}$$

Dann gilt $\mathcal{N}_{\vartheta,1}(\tilde{X}_n = 0) = \frac{1}{2}$, d.h. $\tilde{x}_n$ ist nicht mediantreu, aber

$$\int \tilde{x}_n d\mathcal{N}_{\vartheta,1}^n \;=\; \frac{2}{n}\int (x_1 + \ldots + x_n)\mathcal{N}_{\vartheta,1} d(x_1) \cdots \mathcal{N}_{\vartheta,1}(dx_n) = \vartheta.$$

(*c*) Das quadratische Risiko von $\bar{x}_n$ und $\tilde{x}_n$ ist

$$R(\vartheta, \bar{x}_n) = \int (\bar{x}_n - \vartheta)^2 d\mathcal{N}_{\vartheta,1}^n = \mathrm{Var}_\vartheta(\bar{x}_n) = \frac{1}{n} \; \forall \vartheta$$

und

$$R(\vartheta, \tilde{x}_n) > \frac{1}{2}\vartheta^2 > \frac{1}{n}$$

für $n \to \infty$, falls $\vartheta \neq 0$.

(*d*) Radioaktiver Zerfall - Wahrscheinlichkeit, dass ein Atom bis zur Zeit $t > 0$ nicht zerfallen ist: $\exp(-\vartheta \cdot t), \vartheta > 0$.
Die Halbwertszeit ist $\tau(\vartheta) = \frac{\log 2}{\vartheta}$, d.h. $P_\vartheta(\text{Zerfall bis } \tau(\vartheta)) = \frac{1}{2}$.
Sei $\widehat{\tau}$ ein erwartungstreuer Schätzer für ϑ auf der Basis von $X \sim P_\vartheta$, d.h.

$$\int_0^\infty \widehat{\tau}(x)\vartheta e^{-\vartheta x} dx = \vartheta$$

für alle $\vartheta > 0$. Division durch ϑ auf beiden Seiten sowie Ableitung nach ϑ (dominierte Konverenz) liefert

$$\int_0^\infty \widehat{\tau}(x)(-xe^{-\vartheta x}) dx \equiv 0,$$

woraus $\widehat{\tau}(x) \cdot x = 0$ $P_\vartheta-$f.s. folgt, da $\{e^{-\vartheta x} dx, \vartheta > 0\}$ eine exponentielle Familie und damit vollständig ist. Da $\widehat{\tau}(x) = 0$ $P_\vartheta-$f.s. $\Rightarrow \vartheta = 0$, Widerspruch.
Aber natürlich ist $\widehat{\kappa}(x) = x \log(2)$ erwartungstreu für $\frac{\log 2}{\vartheta}$, d.h. der Halbwertszeit. Für mediantreue Schätzer ist dagegen die monotone Transformation $\vartheta \mapsto \frac{\log 2}{\vartheta}$ kein Problem.

(*e*) Schätzung von Median- und Mittelwertfunktionalen für Lokationsfamilien.

Proposition 2.12 (Erwartungstreue Schätzer für Mittelwerte von Lokalisationsfamilien)**:**
Sei $\mathcal{P}_n$ eine Lokalisationsfamilie, d.h. $\mathcal{P}_n \coloneqq \{P_\mu^n : \frac{dP_\mu}{d\lambda} = f(x-\mu), \mu \in \mathbb{R}\}$, wobei f eine Lebesgue-Dichte mit $\sigma^2 = \int x^2 f(x) dx < \infty$ und $\int x f(x) dx = 0$ sei. Dann ist $\overline{x}_n = \frac{1}{n}(x_1 + \ldots + x_n)$ ein erwartungstreuer Schätzer für $\mu = \tau(P_\mu) = \int x f(x-\mu) dx = \int x dP_\mu(x)$ mit quadratischem Risiko

$$\begin{aligned} R(P_\mu^n, \overline{x}_n) &= \int (\mu - \overline{x}_n)^2 f(x_1 - \mu) \cdots f(x_n - \mu) dx_1, \ldots, dx_n \\ &= \frac{n}{n^2} \int x^2 f(x) dx = \frac{\sigma^2}{n}. \end{aligned}$$

(f): Median: Sei $\mathcal{P}_n$ eine Lokationsfamilie mit positiver Dichte $f(x) > 0$ für alle $x \in \mathbb{R}$ und $\int_0^\infty f(x)dx = \frac{1}{2}$. Dann ist $\tau(P_\mu) = \mu$ der Median von P_μ, d.h. $\frac{1}{2} = P_\mu(x \leq \mu) = P_\mu(x \geq \mu)$.
1): Sei n ungerade. Dann gilt $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)} < x_{(5)}$ P_μ-f.s. Dann ist der Stichprobenmedian $\tilde{x}_n := x_{(\frac{n+1}{2})}$ für die geordnete Stichprobe ein mediantreuer Schätzer.
Die Idee von (e) und (f): Definiere das empirische Maß zu einer Stichprobe $(x_1, \ldots, x_n)$: $\widehat{P}_n := \frac{1}{n}\sum_{j=1}^n \delta_{x_j}$. Es gilt nach dem Gesetz der großen Zahlen $\left|\widehat{P}_n(A) - P_\mu(A)\right| \to 0$ $P_\mu^{\mathbb{N}}-$stochastisch für alle $A \in \mathcal{B}^1$.
Daraus folgt $\tau(\widehat{P}_n) \to \tau(P_\mu)$ P_μ^n-stochastisch für τ als Mittelwert respektive Median.
Wenn n gerade ist, gilt $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)}$. Falls $f(x) = f(-x)$ für alle $x \in \mathbb{R}$ gilt, dann ist der Stichprobenmedian $\tilde{x}_n := \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$ ein mediantreuer Schätzer.

Beweis. Da der Median von $x_1 + \mu, \ldots, x_n + \mu$ gerade $\tilde{x}_n + \mu$, sei o.E. $\mu = 0$. Dann gilt, da $(x_1, \ldots, x_n) \to \tilde{x}_n$ symmetrisch bzgl. Vertauschung in x_j ist,

$$P_0^n(\tilde{x}_n \leq 0) = \int \ldots \int \mathbb{1}_{\{\tilde{x}_n \leq 0\}} f(x_1) \cdots f(x_n) dx_1 \cdots dx_n$$

Sei $F(t) = \int_{-\infty}^t f(x)dx$. Dann ist $t \mapsto F(t)$ strikt monoton wachsend, da $F'(t) = f(t) > 0$ λ-f.ü.
Ferner gilt für alle $0 \leq u \leq 1$

$$P_0(F(X_j) \leq u) = u,$$

d.h. $Y_j = F(X_j)$ sind i.i.d. mit Gleichverteilung in $[0, 1]$. Also gilt

$$\mathbb{1}_{\{\tilde{x}_n \leq 0\}} = \mathbb{1}_{\{\tilde{y}_n \leq F(0) = \frac{1}{2}\}}$$

und

$$P_0^n(\tilde{x}_n \leq 0) = \int_0^1 \cdots \int_0^1 \mathbb{1}_{\{\tilde{y}_n \leq 1/2\}} dy_1 \ldots dy_n.$$

Man zeige nun

$$\begin{aligned}
\int \mathbb{1}_{\{\tilde{y}_n \leq a\}} dy_1 \cdots dy_n &= \frac{n!}{((n-1)/2)!((n-1)/2)!} \int_0^a \int_0^{y_{(n+1)/2}} dy_1 \cdots \int_0^{y_{(n+1)/2}} dy_{(n-1)/2} \\
&\quad \int_{y_{(n+1)/2}}^1 dy_{(n+1)/2+1} \cdots \int_{y_{(n+1)/2}}^1 dy_n dy_{(n+1)/2} \\
&= \frac{n!}{\left(\frac{n-1}{2}\right)!^2} \int_0^a u^{(n-1)/2}(1-u)^{(n-1)/2}\, du
\end{aligned}$$

Das Integral ist genau dann $\frac{1}{2}$, wenn $a = \frac{1}{2}$. ∎

2.1 Erwartungstreue Schätzer und konvexe Verlustfunktion

Sei $\kappa : \mathcal{P} \to \mathbb{R}^p$ ein Funktional, $L : \mathcal{P} \times \mathbb{R}^p \to [0, \infty)$ eine konvexe, messbare Verlustfunktion. Problem: Finde einen Schätzer $\widehat{\kappa} : \mathcal{X} \to \mathbb{R}$ mit

$$\int \widehat{\kappa} dP = \kappa(P) \quad \text{für alle } P \in \mathcal{P}, \tag{2.14}$$

welcher das Risiko

$$R(P, \widehat{\kappa}) = \int L(P, \widehat{\kappa}(x)) P(dx) \tag{2.15}$$

für $\mathcal{P}$ minimiert, wobei $\widehat{\kappa}$ so gewählt ist, dass obige Gleichungen definiert sind. Ein Schätzer $\widehat{\kappa}$ mit (2.14) heißt erwartungstreu.

Proposition 2.16:
Ist L strikt konvex, so gibt es zu jedem $P \in \mathcal{P}$ höchstens einen erwartungstreuen Schätzer mit minimalen Risiko $R(P, \cdot)$.

Beweis. Seien $\widehat{\kappa}_1$ und $\widehat{\kappa}_2$ erwartungstreu. Dann ist $\widehat{\kappa}_0 = \frac{1}{2}(\widehat{\kappa}_1 + \widehat{\kappa}_2)$ erwartungstreu und es gilt

$$L(P, \widehat{\kappa}_0(x)) \leq \frac{1}{2} L(P, \widehat{\kappa}_1(x)) + \frac{1}{2} L(P, \widehat{\kappa}_2(x)). \tag{$*$}$$

Daraus folgt

$$R(P, \widehat{\kappa}_0) \leq \frac{1}{2}(R(P, \widehat{\kappa}_1) + R(P, \widehat{\kappa}_2)) \tag{$**$}$$

Sind $\widehat{\kappa}_1$ und $\widehat{\kappa}_2$ minimal für P, so auch $\widehat{\kappa}_0$ und es gilt die Gleichheit in $(**)$. Daraus folgt die P-f.s. Gleichheit in $(*)$. Also folgt $\widehat{\kappa}_1(x) = \widehat{\kappa}_2(x)$ P-f.ü. ∎

Satz 2.17 (Satz von Rao-Blackwell)**:**
Sei $\kappa : \mathcal{P} \to \mathbb{R}^p$ ein Funktional und $T : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Z}, \mathcal{S})$ eine suffiziente Statistik für $\mathcal{P}$. Sei $\widehat{\kappa} : \mathcal{X} \to \mathbb{R}^p$ ein Schätzer für $\kappa(P)$ und

$$\tilde{\kappa}(\cdot) = \mathbb{E}_P(\widehat{\kappa} | T = \cdot) : \mathcal{Z} \to \mathbb{R}^p$$

eine faktorisierte bedingte Erwartung von $\widehat{\kappa}$ gegeben T (d.h. $\tilde{\kappa} \circ T$ ist bedingte Erwartung bzgl. $T^{-1}(\mathcal{S})$).
Dann ist $\tilde{\kappa}$ unabhängig von $P \in \mathcal{P}$ definiert und es gilt:

(*i*) *Für jede konvexe Verlustfunktion L gilt*

$$R(P, \tilde{\kappa} \circ T) \leq R(P, \widehat{\kappa}) \quad \textit{für alle } P \in \mathcal{P}\,. \tag{2.18}$$

(*ii*) *Falls L strikt konvex ist, so gilt in* (2.18) *die strikte Ungleichung, falls nicht $\tilde{\kappa} \circ T = \widehat{\kappa}$ P-f.s.*

(iii) Ist $\widehat{\kappa}$ erwartungstreu, so auch $\tilde{\kappa} \circ T$.

Beweis. *(i)*: Sei $P \in \mathcal{P}$ beliebig. Da $L(P, \cdot)$ konvex ist, gilt nach der bedingten Jensen-Ungleichung

$$L(P, \underbrace{\tilde{\kappa} \circ T}_{=\mathbb{E}(\widehat{\kappa}|\sigma(T))}) \leq \mathbb{E}\left(L(P, \widehat{\kappa}) \mid \sigma(T)\right) \qquad (*)$$

Integration von $(*)$ über P liefert (2.18).
(ii): Gleichheit klar.
(iii): Folgt aus $\int \tilde{\kappa} \circ T dP = \int \widehat{\kappa} dP = \kappa(P)$, da $\Omega \in \sigma(T)$. ∎

Satz 2.18 (Satz von Lehmann-Scheffé)**:**
Ist T suffizient und $\mathcal{P} \circ T$ vollständig, so minimiert jeder erwartungstreue Schätzer $\bar{\kappa} \circ T$ für $\kappa(P)$ das Risiko zu konvexen Verlustfunktionen in der Klasse der erwartungstreuen Schätzer.

Beweis. Sei $\bar{\kappa} \circ T$ erwartungstreu für $\kappa(P)$ und $\widehat{\kappa}$ ein weiterer erwartungstreuer Schätzer für $\kappa(P)$. Dann gibt gibt es nach Satz 2.17 einen Schätzer $\tilde{\kappa} \circ T$ mit

$$R(P, \tilde{\kappa} \circ T) \leq R(P, \widehat{\kappa}) \text{ für alle } P \in \mathcal{P}\,.$$

Aus Satz 2.17 *(iii)* folgt

$$\int (\tilde{\kappa} \circ T - \bar{\kappa} \circ T)\, dP = 0 \text{ für alle } P \in \mathcal{P}\,.$$

Da $\mathcal{P} \circ T$ vollständig ist gilt also $\tilde{\kappa} = \bar{\kappa}$ $P \circ T-$f.s. Daher ist für alle $P \in \mathcal{P}$

$$R(P, \bar{\kappa} \circ T) = R(P, \tilde{\kappa} \circ T) \leq R(P, \widehat{\kappa}).$$

∎

Aus diesem Satz folgt: es gibt für vollständige Familien höchstens einen erwartungstreuen Schätzer mit minimalem Risiko!

Korollar 2.19:
Sei $\mathcal{P}$ eine exponentielle Familie mit Dichten $x \mapsto c(P)h(x) \exp\left(\sum_j a_j(P) T_j(x)\right)$ und $\{a_1(P), \ldots, a_k(P), P \in \mathcal{P}\}$ habe nichtleeres Inneres. Falls für eine Stichprobengröße m und ein Funktional $\kappa : \mathcal{P} \to \mathbb{R}^p$ ein erwartungstreuer Schätzer $\tilde{\kappa}(x_1, \ldots, x_m)$ für $\kappa(P)$ existiert, so gibt es für alle $n \geq m$ einen erwartungstreuen Schätzer für $\kappa(P)$, nämlich $\widehat{\kappa}(T(x_1, \ldots, x_n))$, mit minimalem konvexen Risiko unter allen erwartungstreuen Schätzern, wobei

$$T(x_1, \ldots, x_n) := \left(\sum_{r=1}^{n} T_j(x_r)\right)_{j=1,\ldots,k}.$$

Beweis. Übung. ∎

Beispiel:

1) Binomial-Verteilung $B(n, \alpha)$. Der Schätzer $\bar{x}_n = \frac{1}{n}(x_1 + \ldots + x_n)$ für α ist erwartungstreu mit minimalem konvexem Risiko in der Klasse der erwartungstreuen Schätzer.

2) Normalverteilungsfamilie $\{\mathcal{N}(\mu, 1), \mu \in \mathbb{R}\}$. Auch $\bar{x}_n$ wie im ersten Teil.

3) Da $\frac{1}{n}(x_1^2 + \ldots + x_n^2)$ ein erwartungstreuer Schätzer für den Parameter σ^2 aus $\{\mathcal{N}(0, \sigma^2), \sigma^2 > 0\}$ ist, ist $\widehat{\sigma}_n^2 = \frac{1}{n}(x_1^2 + \ldots + x_n^2)$ mit minimalem konvexem Risiko in der Klasse der erwartungstreuen Schätzer.

2.2 Untere Schranken für erwartungstreue Schätzer

Sei $\mathcal{P} = \{P_\vartheta, \vartheta \in \mathbb{R}\} \ll \mu$ eine parametrische Familie, $\widehat{\kappa} : \Omega \to \mathbb{R}$ ein erwartungstreuer Schätzer für $\kappa(P_\vartheta) =: \kappa(\vartheta), \kappa : \mathbb{R} \to \mathbb{R}$ und $f(w, \vartheta) \in \frac{dP_\vartheta}{d\mu}$. Wir wollen die Konzentration um den wahren (rellen) Parameter untersuchen. Dazu seien:

i) $\kappa(\vartheta)$ ist stetig differenzierbar.

ii) $l_\vartheta(w, \vartheta) := \frac{\frac{d}{d\vartheta} f(w,\vartheta)}{f(w,\vartheta)}$ ist f.ü. definiert mit $\int l_\vartheta(w, \vartheta)^2 P_\vartheta(dw) < \infty$.

iii) $\int l_\vartheta(w, \vartheta) P_\vartheta(dw) = 0$.

iv) $\int \widehat{\kappa}(w) l_\vartheta(w, \vartheta) P_\vartheta(dw) = \frac{d}{d\vartheta} \int \widehat{\kappa}(w) P_\vartheta(dw)$.

Dann gilt:

Satz 2.20 (Satz von Cramér-Rao)**:**
Falls $\int \widehat{\kappa}^2 dP_\vartheta < \infty$*, so gilt*

$$\mathrm{Var}_\vartheta\left(\widehat{\kappa}\right) := \int \left(\widehat{\kappa}(x) - \kappa(\vartheta)\right)^2 P_\vartheta(dx) \geq \frac{\kappa'(\vartheta)^2}{\int l_\vartheta(x, \vartheta)^2 P_\vartheta(dx)}.$$

Der Nenner heißt auch Fischer-Information.

Beweis. Es sei $g(x, \vartheta) = \widehat{\kappa}(x) - \kappa(\vartheta)$. Dann folgt aus $\int \widehat{\kappa}(x) P_\vartheta(dx) = \kappa(\vartheta)$ durch Differenzieren nach ϑ mit (iv)

$$\kappa'(\vartheta) = \int \widehat{\kappa}(x) l_\vartheta(x, \vartheta) P_\vartheta(dx),$$

d.h. wegen $\int l_\vartheta(x, \vartheta) P_\vartheta(dx) = \frac{d}{d\vartheta} 1 = 0$ gilt

$$\int \left(\widehat{\kappa}(x) - \kappa(\vartheta)\right) l_\vartheta(x, \vartheta) P_\vartheta(dx) = \kappa'(\vartheta)$$

und nach der Cauchy-Schwarzschen Ungleichung folgt

$$\int \left(\widehat{\kappa}(x) - \kappa(\vartheta)\right)^2 P_\vartheta(dx) \int l_\vartheta(x, \vartheta)^2 P_\vartheta(dx) \geq \kappa'(\vartheta)^2.$$

■

Die Abhängigkeit von der Stichprobengröße n:
Die obigen Beispiele mit Schätzerrisiken der Ordnung $\frac{1}{n}$ zeigen, dass das Risiko ab- bzw. die Konzentration für $n \to \infty$ zunimmt.

Definition 2.21:
Seien $\widehat{\kappa}_{1,n}, \widehat{\kappa}_{2,n}, n \in \mathbb{N}$ Folgen von Schätzern zur Stichprobenanzahl $n = 1, 2, \ldots$ für $\kappa(P)$. Seien $R(P, \widehat{\kappa}_{j,n})$ die Risiken bezüglich der Verlustfunktion $L(P, \kappa)$. Definiere für festes $n \in \mathbb{N}$

$$m(n) \coloneqq \inf\{m \in \mathbb{N} : R\,(P, \widehat{\kappa}_{2,m}) \leq R(P, \widehat{\kappa}_{1,n})\}.$$

Dann heißt $\frac{n}{m(n)}$ die relative Effizienz von $\widehat{\kappa}_{2,n}$ zu $\widehat{\kappa}_{1,n}$. Falls $\lim_{n\to\infty} \frac{n}{m(n)}$ existiert, so heißt der Grenzwert asymptotische relative Effizienz.

$m(n)$ kann dahingehend interpretiert werden, wie groß die Stichprobenanzahl sein muss, damit der Schätzer $\widehat{\kappa}_2$ besser ist als $\widehat{\kappa}_1$. Je größer $m(n)$ ist, desto besser ist $\widehat{\kappa}_2$.

Beispiel 2.22:
Sei $\mathcal{P} \coloneqq \{\mathcal{N}(0, \sigma^2), \sigma^2 > 0\}$. Der Schätzer $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_j x_j^2$ hat minimales konvexes Risiko aller erwartungstreuen Schätzer für σ^2 (Beweis später). Es gilt:

(*i*) $\frac{n\widehat{\sigma}_n^2}{\sigma^2}$ hat eine χ_n^2-Verteilung.

(*ii*) Sei $S_n^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$. Dann ist S_n^2 erwartungstreu für $\mathcal{N}(\mu, \sigma^2)$ für $\mu \in \mathbb{R}, \sigma^2 > 0$ und $(n-1)\frac{S_n^2}{\sigma^2}$ ist χ_{n-1}^2-verteilt.

Also ist S_{n+1}^2 genauso verteilt wie $\widehat{\sigma}_n^2$, d.h. die relative Effizienz von S_n^2 zu $\widehat{\sigma}_n^2$ ist $\frac{n}{n+1}$.

3 Asymptotische Schätztheorie

In diesem Kapitel geht es um Folgen von Schätzern $\widehat{\tau}_n(\omega_1, \ldots, \omega_n)$ für ein Funktional $\tau(P)$ und ihre asymptotische Verteilung für $n \to \infty$. Sei dazu $\mathcal{P}$ eine Familie von W-Maßen.

Definition 3.1:

(*i*) Eine Schätzerfolge $(\widehat{\tau}_n : \Omega^n \to (Y, \mathcal{B}))_{n\in\mathbb{N}}$, wobei (Y, d) ein polnischer Raum ist, heißt konsistent für $\tau : \mathcal{P} \to Y$, falls $\widehat{\tau}_n \to \tau(P)$ $P^{\mathbb{N}}$-stochastisch für alle $P \in \mathcal{P}$, d.h. es gilt

$$\lim_n P^n(w \in \Omega^n : d(\widehat{\tau}_n(\omega), \tau(P)) > \varepsilon) = 0$$

für alle $\varepsilon > 0$.

(*ii*) $\widehat{\tau}_n$ heißt gleichmäßig konsistent auf $\mathcal{P}$, falls

$$\lim_n \sup_{P\in\mathcal{P}} P^n(\omega \in \Omega^n : d(\widehat{\tau}_n(\omega), \tau(P)) > \varepsilon) = 0$$

für alle $\varepsilon > 0$.
Ist $\mathcal{P}$ mit einer Topologie versehen, z.B. für eine parametrische Familie $\mathcal{P} = \{P_\vartheta, \vartheta \in \Theta\}$, mit Θ topologischer Raum, so gibt es den Begriff der lokal gleichmäßigen Konsistenz (d.h. gleichmäßig auf kompakten Teilmengen von $\mathcal{P}$).

Frage: Wie erhält man eine konsistente Schätzerfolge für ein Funktional?
Quelle: Konstruiere eine konsistente Schätzerfolge für das Maß $P \in \mathcal{P}$. Hierzu sei $\mathcal{P}$ eine Familie von W-Maßen auf $(\mathbb{R}^p, \mathcal{B}^p)$.
1): Es gilt, dass die Folge der empirischen Maße

$$F_n^\omega(A) := \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_A(\omega_j)$$

eine konsistente Schätzerfolge für $P(A)$ für festes $A \in \mathcal{B}^p$ ist. Denn nach dem schwachen Gesetz der großen Zahlen gilt

$$\lim_n P(|F_n^\omega(A) - P(A)| > \varepsilon) = 0$$

für alle $\varepsilon > 0$ und alle $P \in \mathcal{P}$.
2): Seien $A_k \in \mathcal{B}^p$ abzählbar viele Rechtecke, die die Borel-σ-Algebra erzeugen, d.h.

$$A_k = \{\omega = (\omega_1, \ldots, \omega_p) : \omega_1 \le a_1, \ldots, \omega_p \le a_p\} \ \forall k \in \mathbb{N}$$

wobei $a_j \in \mathbb{Q}$ und sei

$$d(P, P') := \sum_{k=1}^{\infty} |P(A_k) - P'(A_k)| \, 2^{-k}.$$

Dann ist d eine Metrik auf $\mathcal{P}$ und es gilt $d(P, P') = 0 \Leftrightarrow P = P'$ auf $\mathcal{B}^p$.

Lemma 3.2:

(*i*) *Die Schätzerfolge* $(x_1, \ldots, x_n) \to F_n^x$ *ist konsistent für* P *gleichmäßig auf* $\mathcal{P}$ *bzgl.* d*, d.h. es gilt*

$$\lim_n \sup_{P \in \mathcal{P}} P^n(x \in X^n : d(F_n^x, P) > \varepsilon) = 0.$$

(*ii*) *Sei* $\mathcal{P} := \{P_\vartheta : \vartheta \in \Theta\}, \Theta \subset \mathbb{R}^p$ *kompakt,* $\vartheta \mapsto P_\vartheta(A_k)$ *stetig für alle* k *und* $\vartheta \mapsto P_\vartheta$ *injektiv (identifizierend).*
Dann gibt es eine konsistente Schätzerfolge $\widehat{\vartheta}_n : X^n \to \Theta$ *für* ϑ*, die gleichmäßig konsistent auf* Θ *ist, d.h. es gilt*

$$\lim_n \sup_{\vartheta \in \Theta} P_\vartheta^n(x \in X^n : \left\| \widehat{\vartheta}_n(x) - \vartheta \right\| > \varepsilon) = 0$$

für alle $\varepsilon > 0$.

Beweis. [6], S. 189 - 191.

Skizze:

(*i*): $d(F_n^x, P) = \sum_k |F_n^x(A_k) - P(A_k)| \, 2^{-k} \overset{P\text{-stoch.}}{\longrightarrow} 0$, da $F_n^x(A_k) - P(A_k) \overset{P}{\to} 0$ für $n \to \infty, 1 \le k \le k_n$, sodass mit Tschebycheff gilt

$$\sup_{P \in \mathcal{P}} P^n \left(\underbrace{|F_n^x(A_k) - P(A_k)|}_{= F_n^x(A_k) - \mathbb{E}F_n^x(A_k)} > \varepsilon \right) \le \frac{1}{n\varepsilon^2} \sup_{P \in \mathcal{P}} \mathrm{Var}_P(\mathbb{1}_{A_k}) \le \frac{1}{4n\varepsilon^2}.$$

(*ii*): $\vartheta \mapsto d(P_\vartheta, P)$ ist stetig für alle W-Maße P auf $\mathbb{R}^p$.
(1): Minimum-Distanz-Schätzer $\vartheta(P) := \arg\min d(P_\vartheta, P)$ existiert, da Θ kompakt ist.
(2): Setze $P = F_n^x$ und definiere $\widehat{\vartheta}_n(x) = \vartheta(F_n^x)$. Dies ist der gesuchte Schätzer. Dann

$$d(P_{\vartheta(P)}, P) \le 2d(P, P_\vartheta) \; \forall \vartheta \in \Theta.$$

Technische Probleme:
(*i*): Man muss zeigen, dass $x \mapsto \widehat{\vartheta}_n(x)$ messbar ist.
(*ii*): Zeige, dass $\{P_\vartheta, \vartheta \in \Theta\}$ $d-$kompakt ist und $(\mathcal{P}, d) \to (\Theta, \|\cdot\|), P_\vartheta \mapsto \vartheta$ gleichmäßig stetig ist. Dann existiert zu $\varepsilon > 0$ ein $\delta > 0$ derart dass

$$P_\vartheta^n(x \in X^n : \left\| \widehat{\vartheta}_n(x) - \vartheta \right\| > \varepsilon) \le P_\vartheta^n(x \in X^n : d(P_{\vartheta(F_n^x)}, P_\vartheta) > \delta).$$

Mit Teil (*i*) folgt die Behauptung. ∎

Konkrete Schätzerfolgen: Maximum-Likelihood-Schätzer
Sei $\Theta \subset \mathbb{R}^p$ und $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ eine parametrische Familie mit Dichten $p_\vartheta \in \frac{dP_\vartheta}{d\mu}$.
Idee: Schätze den Parameter ϑ durch

$$\widehat{\vartheta}_n(x_1, \ldots, x_n) = \arg\max_{\vartheta \in \Theta} p_\vartheta(x_1) \cdots p_\vartheta(x_n),$$

wo $p_\vartheta(x_1) \cdots p_\vartheta(x_n) \in \frac{dP_\vartheta^n}{d\mu^n}$ die gemeinsame Dichte von $(x_1, \ldots, x_n)$ unter P_ϑ^n ist. $\widehat{\vartheta}_n(x)$ heißt Maximum-Likelihood-Schätzer.
$\vartheta \mapsto L^{(n)}(\vartheta, x) \coloneqq p_\vartheta(x_1) \cdots p_\vartheta(x_n)$ heißt Likelihood-Funktion und

$$l^{(n)}(\vartheta, x) = \frac{1}{n} \log(L^{(n)}(\vartheta, x)) = \frac{1}{n} \sum_{j=1}^{n} \log p_\vartheta(x_j)$$

heißt Log-Likelihood-Funktion.

Beispiel 3.3 (ML-Schätzer für exponentielle Familien)**:**
Sei $\mathcal{P} = \{P_\vartheta, \vartheta \in \Theta\}$ eine dominierte Familie mit μ-Dichte

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

Dann maximiere $\log\left(\frac{dP_\vartheta^n}{d\mu^n}\right)$ in ϑ, d.h. maximiere

$$n \log(c(\vartheta)) + \sum_{j=1}^{p} a_j(\vartheta) \left(\sum_{\nu=1}^{n} T_j(x_\nu) \right)$$

in ϑ.
Dann ist die Lösung von

$$n \left(\frac{d}{d\vartheta_l} c(\vartheta) \right) \frac{1}{c(\vartheta)} + \sum_{j=1}^{k} \frac{d}{d\vartheta_l} a_j(\vartheta) \left(\sum_{\nu=1}^{n} T_j(x_\nu) \right) = 0$$

der Maximum-Likelihood-Schätzer $\widehat{\vartheta}_n$.

(a) Sei $\mathcal{P} = \{\mathcal{N}(\vartheta, 1), \vartheta \in \mathbb{R}\}$ mit $p = 1$. Dann gilt die Extremalbedingung

$$\frac{d}{d\vartheta} \left(-\frac{\vartheta^2}{2} n + \frac{2\vartheta}{2} \sum_{\nu=1}^{n} x_\nu \right) = 0 \Leftrightarrow \widehat{\vartheta}_n = \frac{1}{n} \sum_{\nu} x_\nu = \bar{x}_n$$

(*b*) Sei $\mathcal{P} = \{\mathcal{N}(0, \vartheta^2), \vartheta > 0\}$ mit λ-Dichte

$$\varphi_{0,\vartheta^2}(x) = \frac{1}{\sqrt{2\pi}\vartheta} \exp\left(-\frac{x^2}{2\vartheta^2}\right).$$

Dann ist der ML-Schätzer

$$\log\left(\prod_{\nu=1}^{n} \varphi_{0,\vartheta^2}(x_\nu)\right) = n\log(\frac{1}{\sqrt{2\pi}}) + n\log(\frac{1}{\vartheta}) - \frac{\sum x_\nu^2}{2\vartheta^2}$$

maximal in ϑ, wenn

$$-n\vartheta\frac{1}{\vartheta^2} + \frac{\sum_j x_j^2}{\vartheta^3} = 0 \Leftrightarrow \widehat{\vartheta}_n = \sqrt{\frac{1}{n}\sum_j x_j^2}.$$

Idee: Die Likelihood-Funktion

$$\tau \mapsto l^{(n)}(\tau, x) \coloneqq \frac{1}{n}\sum_{\nu=1}^{n} \log(p_\tau(x_\nu)) \to \int \log(p_\tau) dP_\vartheta,$$

falls der Erwartungswert existiert. Allerdings gilt

$$\begin{aligned}\int \log p_\tau dP_\vartheta - \int \log(p_\vartheta) dP_\vartheta &= \int \log\left(\frac{p_\tau}{p_\vartheta}\right) dP_\vartheta \overset{\text{Jensen}}{\leq} \log\left(\int \frac{p_\tau}{p_\vartheta} p_\vartheta d\mu\right) \\ &= \log\left(\int p_\tau d\mu\right) = 0\end{aligned}$$

d.h.

$$\tau \mapsto H(\tau, \vartheta) \coloneqq \int (\log(p_\tau)) p_\vartheta d\mu \leq \int (\log(p_\vartheta)) p_\vartheta d\mu = H(\vartheta, \vartheta).$$

Dies heißt auch Kullback-Leibler-Distanz (relative Information, relative Distanz, Entropie, ...) von P_ϑ und P_τ.
Ersetze P_ϑ durch $F_n^x, x \sim P_\vartheta^n$. Dann maximiert der ML-Schätzer $\widehat{\vartheta}_n(x)$

$$\tau \mapsto \int \log(p_\tau) dF_n^x = \underbrace{\int \log(p_\tau) dP_\vartheta}_{H(\tau,\vartheta)} + \int \log(p_\tau) d(\underbrace{F_n^x - P_\vartheta}_{\text{empirischer Prozess}})$$

Der erste Summand hat das Maximum für $\tau = \vartheta$, und es gilt

$$\int \log(p_\tau) d(F_n^x - P_\vartheta) = \frac{1}{n}\sum_{\nu=1}^{n} \left(\log p_\tau(x_\nu) - \int (\log(p_\tau) dP_\vartheta\right) \to 0$$

P_ϑ^n-stochastisch.

Beispiel (Trunkierung)**:**
Seien $Z_{\tau_j}, j = 1, \ldots, n$ i.i.d. und gleichmäßig integrierbar in τ mit $\mathbb{E}_{\vartheta_0} Z_{\tau_j} = 0$. Definiere die trunkierte Größe

$$Z_{\tau_j} = Z_{\tau_j} \mathbb{1}_{\{|Z_{\tau_j}| \leq N\}} + Z_{\tau_j} \mathbb{1}_{\{|Z_{\tau_j}| > N\}} =: \widehat{Z}_{\tau_j} + \tilde{Z}_{\tau_j}.$$

Dann gilt $\overline{Z}_{\tau_j} \coloneqq \widehat{Z}_{\tau_j} - \mathbb{E}_{\vartheta_0} \widehat{Z}_{\tau_j} \in L^2(P_\vartheta)$. Mit der Ungleichung von Tschebyscheff folgt

$$P_{\vartheta_0}\left(\left|\frac{1}{n}\sum_j \overline{Z}_{\tau_j}\right| > \eta\right) \leq \frac{1}{\eta^2 n}\mathbb{E}_{\vartheta_0}\overline{Z}^2_{\tau_1} = \frac{CN}{\eta^2 n}\mathbb{E}_{\vartheta_0}|Z_{\tau_1}| \to 0, \text{wenn } \frac{N}{n} \to 0.$$

Man braucht nun eine Flankenabschätzung für $\tilde{Z}_{\tau_j}$.

Satz 3.4 (Konsistenzsatz)**:**
Sei $\Theta \subset \mathbb{R}^p$ offen und $\mathcal{P} = \{P_\vartheta, \vartheta \in \Theta\}$ mit einem identifizierbaren Parameter (d.h. $\vartheta \neq \vartheta' \Rightarrow P_\vartheta \neq P_{\vartheta'}$).

(a) *Sei $\vartheta \mapsto p(\vartheta, x)$ stetig auf Θ für alle x. Für $\log p_\vartheta(x)$ existiere für alle $\delta > 0$ und $\vartheta_0 \in \Theta$ eine endliche Überdeckung $(V_i)_i$ von $\{\tau \in \Theta : \|\tau - \vartheta_0\| \geq \delta\} \subset \bigcup_j V_j$, auf der*

$$g_{\vartheta_0, V_j} = \sup_{\tau \in V_j} \log p_\tau - \log p_{\vartheta_0} = \sup_{\tau \in V_j} \log\left(\frac{p_\tau}{p_{\vartheta_0}}\right)$$

lokal gleichmäßig integrierbar bezüglich P_ϑ ist, d.h. es existiert ein $\varepsilon > 0$ mit

$$\lim_{N\to\infty} \sup_{|\vartheta - \vartheta_0| < \varepsilon} \int_{\{|g_{\vartheta_0, V_j}| > N\}} \left|g_{\vartheta_0, V_j}\right| dP_\vartheta = 0 \text{ für } j = 1, \ldots, n$$

und

$$\int g_{\vartheta_0, V_j} dP_{\vartheta_0} < 0$$

(verhindert, dass $\int \log p_\tau dP_\vartheta - \int \log p_\vartheta dP_\vartheta \to 0$ für $|\tau - \vartheta| \to \infty$).
Dann ist jede Folge von Maximum-Likelihood-Schätzern lokal gleichmäßig konsistent.

(b) *Ist $\vartheta \mapsto p(x, \vartheta)$ differenzierbar für alle $x \in X$, so existiert eine Folge von gleichmäßig konsistenten Schätzern auf kompakten Teilmengen $K \subset \Theta$, sodass*

$$\lim_{n\to\infty} \inf_{\vartheta \in K} P^n_\vartheta\left(x \in X^n : \sum_{r=1}^n \frac{\frac{d}{d\vartheta} p(x_r, \widehat{\vartheta}_n(x))}{p(x_r, \widehat{\vartheta}_n(x))} = 0\right) = 1.$$

Beweis. [6], S. 204, Theorem 6.5.3. ff. ∎

3.1 Die asymptotische Verteilung von Schätzfolgen

Seien $x_1, \ldots, x_n$ i.i.d. mit Dichte $f_n(x, \vartheta) = f(x_1, \vartheta) \cdots f(x_n, \vartheta)$ bzgl. μ^n. Im Limes gilt $f_n(x, \vartheta) \approx \delta_{\widehat{\vartheta}_n(x)}$. Also untersucht man das Verhalten von $f_n(x, \vartheta)$ in der Umgebung des wahren Parameters ϑ_0, dort wo $\widehat{\vartheta}_n$ liegt.

Beispiel 3.5:
Sei $\vartheta \mapsto f(x, \vartheta)$ zweimalig stetig partiell differenzierbar, $\Theta \subset \mathbb{R}^p$ offen und $c_n > 0$ eine Folge in $\mathbb{R}$ und $a \in \mathbb{R}^p$. Dann gilt nach dem Satz von Taylor

$$\begin{aligned} \log f_n\left(x, \vartheta + \frac{a}{c_n}\right) &= \log f_n(x, \vartheta) + \frac{1}{c_n} \langle a, \nabla_\vartheta \log f_n(x, \vartheta) \rangle \\ &\quad + \frac{1}{2c_n^2} \langle a, \mathrm{Hess}_\vartheta(\log f_n(x, \vartheta)) a \rangle + \mathcal{O}\left(\frac{n}{c_n^3}\right). \end{aligned}$$

Wähle $c_n \coloneqq \sqrt{n}$, so gilt

$$\begin{aligned} \log\left(\frac{f_n\left(x, \vartheta + \frac{a}{\sqrt{n}}\right)}{f_n(x, \vartheta)}\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{l=1}^p \frac{d}{d\vartheta_l} \log f(x_i, \vartheta) \cdot a_l \\ &+ \frac{1}{2n} \sum_{i=1}^n \sum_{l,k}^p a_l a_k \frac{d}{d\vartheta_l d\vartheta_k} \log f(x_i, \vartheta) + \mathcal{O}\left(n^{-1/2}\right). \end{aligned}$$

Der erste Summand konvergiert schwach gegen eine Normalverteilung nach dem Zentralen Grenzwertsatz. Nach dem Gesetz der großen Zahlen konvergiert der zweite Term gegen eine Konstante.

Definition 3.6 (LAN)**:**
Die Familie $P_\vartheta^n, \vartheta \in \Theta \subset \mathbb{R}^p$ mit Dichten $f_n(x, \vartheta)$ bezüglich μ^n erfüllt die LAN-Bedingung (Local Asymptotic Normality) mit $c_n > 0, c_n \to \infty$, falls für alle $a \in \mathbb{R}^p$ gilt

$$\log \frac{f_n(x, \vartheta + \frac{a}{c_n})}{f_n(x, \vartheta)} = a^T \Delta_n(x, \vartheta) - \frac{1}{2} a^T D(\vartheta) a + R_n(x, \vartheta, a)$$

mit

- (a) D ist positiv definitiv und symmetrisch.
- (b) $x \mapsto \Delta_n(x, \vartheta)$ ist messbar und $P_\vartheta^n \circ \Delta_n(\cdot, \vartheta) \Rightarrow N(0, D(\vartheta))$.
- (c) $R_n(\cdot, \vartheta, a) \to 0$ P_ϑ^n-stochastisch für $n \to \infty$.
- $(c*)$ $\sup_{\|a\| \leq \eta} |R_n(\cdot, \vartheta, a)| \to 0$ P_ϑ^n-stochastisch für alle $\eta > 0$.

Lemma 3.7:
Sei $\vartheta \mapsto l(x,\vartheta) \coloneqq \log f(x,\vartheta), \vartheta \in \Theta$ *zweimal stetig partiell differenzierbar mit partiellen Ableitungen* $l^{(i)}, l^{(ij)}$ *und es gelte*

(*i*) $\int l^{(i)}(x,\vartheta)P_\vartheta(dx) = 0$,

(*ii*) $\int l^{(ij)}(\cdot,\vartheta)dP_\vartheta + \int l^{(i)}(\cdot,\vartheta)l^{(j)}(\cdot,\vartheta)dP_\vartheta = 0$,

(*iii*) $l^i(\cdot,\vartheta), i = 1,\dots,n$ *sind* $P_\vartheta-$*f.s. linear unabhängig,*

(*iv*) *Es existiert eine Umgebung* V *von* ϑ_0, *sodass* $\vartheta \mapsto l^{(ij)}(x,\vartheta)$ *stetig in* V *für* $i,j = 1,\dots,p$ *ist und* $\sup_{\tau\in V}\left|l^{(ij)}(\cdot,\tau)\right|$ *ist* P_ϑ*-integrierbar, d.h. lokal gleichmäßig in* ϑ_0.

Dann gilt die LAN-Bedingung für ϑ_0 *mit* $c_n = \sqrt{n}$ *und*

$$\Delta_n(x,\vartheta) = \frac{1}{\sqrt{n}}\sum_{r=1}^{n} l^{(\cdot)}(x_r,\vartheta),$$

sowie

$$(D(\vartheta))_{i,j} = \int l^{(i)}(\cdot,\vartheta)l^{(j)}(\cdot,\vartheta)dP_\vartheta$$

und $D(\vartheta) > 0$.

Beweis. [6], Proposition 8.1.8, S. 265.
D ist nicht singulär sowie positiv definit, denn falls für ein $u \in \mathbb{R}^p$

$$\langle u, D(\vartheta_0)u\rangle = \mathbb{E}_{\vartheta_0}\left\langle u, l^{(\cdot)}(\cdot,\vartheta_0)\right\rangle^2 = 0 \quad \text{für ein } \vartheta_0 \in \Theta,$$

gilt, gilt auch

$$\left\langle u, l^{(\cdot)}(x,\vartheta_0)\right\rangle = 0 \ P_{\vartheta_0} - \text{f.s.}$$

im Widerspruch zu (*iii*). ∎

Dies bedeutet, dass

$$\tau \mapsto \log \frac{f_n\left(x, \vartheta + \frac{\tau}{\sqrt{n}}\right)}{f_n(x,\vartheta)}$$

lokal approximiert werden kann durch $\mathcal{N}(0, D(\vartheta))$, d.h. lokal durch eine exponentielle Familie, für die man optimale Schätzer und Tests kennt.

Definition 3.8:
Sei $\Theta \subset \mathbb{R}^p$ wie oben und $\mathcal{P}$ erfülle die LAN-Bedingung. Dann heißt eine

Folge von Schätzern $\widehat{\kappa}_n$ für κ[1] : $\Theta \to \mathbb{R}^q$ regulär in ϑ_0, falls es ein W-Maß M_{ϑ_0} auf $\mathcal{B}^q$ gibt mit

$$P^n_{\vartheta_0+\frac{a}{c_n}} \circ \left(c_n \left(\widehat{\kappa}_n - \kappa \left(\vartheta_0 + \frac{a}{c_n} \right) \right) \right) \Rightarrow M_{\vartheta_0} \quad \text{für alle } a \in \mathbb{R}^p \,.$$

Dies bedeutet, dass $\widehat{\kappa}_n$ konsistent für $\kappa(\vartheta_0)$ unter $P^n_{\vartheta_0}$ im Sinne von $\widehat{\kappa}_n - \kappa(\vartheta_0) = \mathcal{O}(c_n^{-1})$ ist, $c_n \uparrow \infty$ und die skalierte Verteilung $c_n(\widehat{\kappa}_n - \kappa(\vartheta))$ konvergiert gegen eine Grenzverteilung unter P^n_ϑ gleichmäßig für $\vartheta = \vartheta_0 + \frac{a}{c_n}$ auf einer kleiner werdenden Umgebung.

Satz 3.9 (Convolution Theorem)**:**
Sei $\kappa : \Theta \to \mathbb{R}^q, q \leq p$ mit stetigen Ableitungen in ϑ_0. Definiere $K_{i,j}(\vartheta_0) := \frac{d}{d\vartheta_i}\kappa_j(\vartheta)$. Sei $K(\vartheta_0)$ die dazugehörige Matrix mit $rang(K) = q$. Sei $\widehat{\kappa}_n$ eine reguläre Schätzfolge für κ. Dann gilt

$$M_{\vartheta_0} = \mathcal{N}(0, \Sigma(\vartheta_0)) * R_{\vartheta_0},$$

wo R_{ϑ_0} ein W-Maß auf $(\mathbb{R}^q, \mathcal{B}^q)$ ist und $\Sigma(\vartheta_0) := K(\vartheta_0)D^{-1}(\vartheta_0)K(\vartheta_0)^T$.

Bemerkung:
Wegen unserer Resultate über Konzentration (Satz von Anderssen, 2.6(ii)) ist somit $M_{\vartheta_0} = \mathcal{N}(0, \Sigma(\vartheta_0))$ die optimal konzentrierte Verteilung einer Schätzfolge für $\kappa(\vartheta)$. Hier heißt $\Sigma(\vartheta_0)$ optimale asymptotische Kovarianz, denn es gilt $\mathcal{N}(0, \Sigma(\vartheta_0))(C) \geq \mathcal{N}(0, \Sigma(\vartheta_0)) * R_{\vartheta_0}(C)$ für alle messbaren, symmetrischen, konvexen Mengen C.

Beweis. Nach Voraussetzung gilt

i) $P^n_{\vartheta_0} \circ \Delta_n(\cdot, \vartheta_0) \Rightarrow \mathcal{N}(0, D(\vartheta_0))$ (LAN).

ii) $P^n_{\vartheta_0} \circ \underbrace{c_n(\widehat{\kappa}_n(\cdot) - \kappa(\vartheta_0))}_{v_n(\cdot,\vartheta_0)} \Rightarrow M_{\vartheta_0}$ (Regularität in ϑ_0).

Hieraus folgt mit Prohorov die Existenz einer Teilfolge ($N_1 \subset \mathbb{N}$) mit

$$P^n_{\vartheta_0}(\Delta_n(\cdot, \vartheta_0), v_n(\cdot, \vartheta_0)) \Rightarrow Q_0 \text{ für } n \in N_1.$$

Sei $Q_a(du, dv) = \exp\left[a^T u - \frac{1}{2}a^T D(\vartheta_0)a\right] Q_0(du, dv)$ (Übung: Zeige, dass dadurch ein W-Maß definiert ist). Man zeigt, dass $P^n_{\vartheta_0+c_n^{-1}a}(\Delta_n(\cdot, \vartheta_0), v_n(\cdot, \vartheta_0)) \Rightarrow Q_a$ (Übung) und damit gilt

$$P^n_{\vartheta_0+c_n^{-1}a}(v_n(\cdot, \vartheta_0) - K(\vartheta_0)a) \Rightarrow Q_a \circ ((u, v) \mapsto v - K(\vartheta_0)a) \,.$$

[1] Abhängigkeit vom Parameter, nicht unbedingt vom Maß.

Da $c_n(\kappa(\vartheta_0 + c_n^{-1}a) - \kappa(\vartheta_0)) \to K(\vartheta_0)a$ folgt

$$P^n_{\vartheta_0 + c_n^{-1}a}\left(c_n\left(\widehat{\kappa}_n - \kappa(\vartheta_0 + c_n^{-1}a\right)\right) \Rightarrow Q_a \circ ((u,v) \mapsto v - K(\vartheta_0)a),$$

sodass $Q_a \circ ((u,v) \mapsto v - K(\vartheta_0)a) = M_{\vartheta_0}$ gilt.
Also sei $\psi_a(t) \coloneqq \int \exp(it^T(v - K(\vartheta_0)a)Q_a(du,dv) = \int \exp(it^T v)M_{\vartheta_0}(dv)$.
$a \mapsto \psi_a(t)$ ist holomorph in $\mathbb{C}^p$. Setze $a \coloneqq i(s - D^{-1}(\vartheta_0)\kappa(\vartheta_0)^T t)$. Dann ist

$$\int \exp\left[is^T u + it^T(v - K(\vartheta_0)D^{-1}(\vartheta_0)u)\right] Q_0(du,dv) = \exp\left[\frac{1}{2}s^T D(\vartheta_0)s\right]\chi(t)$$

mit $\chi(t) = \psi(t)\exp(\frac{1}{2}K(\vartheta_0)D^{-1}(\vartheta_0)K(\vartheta_0)t)$. Für $s = 0$ folgt, dass $\chi(t)$ die charakteristische Funktion von

$$R_{\vartheta_0} \coloneqq Q_0 \circ ((u,v) \mapsto v - K(\vartheta_0)D^{-1}(\vartheta_0)u)$$

ist. Hieraus folgt $Q_0 \circ ((u,v) \mapsto (u, vK(\vartheta_0)D^{-1}(\vartheta_0)u)) = \mathcal{N}(0, D(\vartheta_0)) \otimes R_{\vartheta_0}$, woraus

$$P^n_{\vartheta_0} \circ (\Delta_n(\cdot,\vartheta_0), v_n(\cdot,\vartheta_0) - \kappa(\vartheta_0)^{-1}(\vartheta_0)\Delta_n(\cdot,\vartheta_0)) \Rightarrow \mathcal{N}(0, D(\vartheta_0)) \otimes R_{\vartheta_0}$$

für $n \in N_1$ folgt, wobei R_{ϑ_0} nicht von der Teilfolge abhängt. Also gilt die Behauptung mittels

$$\begin{aligned} P^n_{\vartheta_0} \circ v_n(\cdot,\vartheta_0) &= P^n_{\vartheta_0} \circ (\Delta_n, v_n - K(\vartheta_0)D(\vartheta_0)^{-1}\Delta_n) \circ ((u,v) \mapsto v + K(\vartheta_0)D(\vartheta_0)^{-1}u) \\ &\Rightarrow (\mathcal{N}(0, D(\vartheta_0)) * R_{\vartheta_0}) \circ ((u,v) \mapsto v + K(\vartheta_0)D^{-1}(\vartheta_0)u) \\ &= R_{\vartheta_0} * \mathcal{N}(0, KD^{-1}K^T) \end{aligned}$$

denn $\mathrm{Cov}(KD^{-1}u) = KD^{-1}D(KD^{-1})^T = KD^{-1}K^T$ und da $D(\vartheta_0)$ symmetrisch ist. ∎

Satz 3.10:
Seien die Voraussetzungen von Lemma 3.7 gegeben. Ferner gelte, dass $x \mapsto l^{(j)}(x,\vartheta)^2$ lokal gleichmäßig in ϑ_0 integrierbar ist.
Sei $\widehat{\vartheta}_n : X^n \to \mathbb{R}^p$ eine approximative Lösung der Maximum Likelihood Gleichung, d.h.

$$\frac{1}{\sqrt{n}}\sum_{r=1}^{n} l^{(\cdot)}(x_r, \widehat{\vartheta}_n) \to 0$$

$P^{\mathbb{N}}_{\vartheta}$-stochastisch und lokal gleichmäßig in ϑ_0.
Dann sind $\widehat{\vartheta}_n$ bzw. $\kappa(\widehat{\vartheta}_n)$ reguläre Schätzfolgen für ϑ bzw. $\kappa(\vartheta)$ mit $c_n = \sqrt{n}$ und $M_{\vartheta_0} = \mathcal{N}(0, D(\vartheta)^{-1})$ bzw. $M_{\vartheta_0} = \mathcal{N}(0, K(\vartheta_0)D^{-1}(\vartheta_0)K(\vartheta_0)^T)$.
Es gilt mit $\Lambda(\vartheta) \coloneqq D(\vartheta)^{-1} > 0$:

(a) $\sqrt{n}(\widehat{\vartheta}_n - \vartheta) - \Lambda(\vartheta) n^{-1/2} \sum_{r=1}^n l^{(\cdot)}(x_r, \vartheta) \to 0$ P_ϑ^n – *stochastisch.*

(b) $P_\vartheta^n \circ (\sqrt{n}(\widehat{\vartheta}_n - \vartheta)) \Rightarrow \mathcal{N}(0, \Lambda(\vartheta))$.

lokal gleichmäßig in ϑ_0.

Bemerkung:
Also ist die Folge von Maximum-Likelihood-Schätzern für ϑ_0 bzw. $\tau(\vartheta_0)$ asymptotisch optimal, in dem Sinne, dass M_{ϑ_0} maximal konzentriert ist.

Beispiel (Supereffizienz)**:**
Sei $\mathcal{P}_n \coloneqq \{\mathcal{N}(\mu, 1)^n, \mu \in \mathbb{R}\}$ die Familie der Normalverteilungen mit gegebener Varianz. Sei $\bar{\omega}_n$ der Mittelwertschätzer mit $\mathcal{N}(\mu, 1)^n \circ (\sqrt{n}(\bar{\omega}_n - \mu)) = \mathcal{N}(0, 1)$. Dies ist optimal nach Satz 3.10. Für $n \in \mathbb{N}$ setze

$$\widehat{\mu}_n(\omega_1, \ldots, \omega_n) \coloneqq \begin{cases} \bar{\omega}_n & |\bar{\omega}_n| \geq n^{-1/4} \\ \frac{\bar{\omega}_n}{2} & \text{sonst} \end{cases}$$

Die Schätzerfolge $(\widehat{\mu}_n)_{n\in\mathbb{N}}$ ist nicht regulär.

Beweis von Satz 3.10. [6], Satz 7.5.5, S. 248.

(*i*) Mit dem Satz von Taylor gilt

$$l^{(i)}(x, \vartheta + n^{-1/2}a) = l^{(i)}(x, \vartheta) + n^{-1/2} \sum_{j=1}^p a_j \int_0^1 l^{(ij)}(x, \vartheta + n^{-1/2}au)\, du,$$

d.h. mit $l_n^{(i)}(x, \tau) \coloneqq n^{-1/2} \sum_{r=1}^n l^{(i)}(x_r, \tau)$ gilt

$$l_n^{(i)}(x, \vartheta + n^{-1/2}a) = l_n^{(i)}(x, \vartheta) - \sum_{j=1}^p D_{ij}^{(n)}(x, \vartheta, a) a_j,$$

mit

$$D_{ij}^n(x, \vartheta, a) \coloneqq -\frac{1}{n} \sum_{r=1}^n \int_0^1 l^{(ij)}(x_r, \vartheta + n^{-1/2}au) du.$$

(*ii*) Anwendung des gleichmäßigen Gesetzes der großen Zahlen liefert die Existenz einer Umgebung $U \ni \vartheta_0$ mit der Eigenschaft, dass für alle $\varepsilon > 0$ ein $\delta > 0$ existiert, sodass

$$\sup_{\vartheta \in U} P_\vartheta^n \{ \sup_{\|a\| \leq \sqrt{n}\delta} \left\| D^{(n)}(x, \vartheta, a) - D(\vartheta) \right\| > \varepsilon \} \to 0$$

für $n \to \infty$ (siehe [6], Korollar 6.7.21, S. 221), wo

$$D(\vartheta) = -(\mathbb{E} l^{(ij)}(\cdot, \vartheta))_{i,j=1,\ldots,p} \overset{(b)}{=} L(\vartheta) > 0.$$

(*iii*) Zusammen ergibt sich $l_n^{(\cdot)}(x,\vartheta+n^{-1/2}a) = l_n^{(\cdot)}(x,\vartheta) - L(\vartheta)a + \underbrace{r_n(x,\vartheta,a)}_{D(\vartheta)-D^{(n)}(x,\vartheta,a)} a.$

Es gilt

$$\sup_{\|a\|\leq\sqrt{n}\delta} \|r_n(\cdot,\vartheta,a)a\| \to 0 \; P_\vartheta^n - \text{stochastisch}$$

lokal gleichmäßig in ϑ_0.
Setze $a_n(x) \coloneqq \sqrt{n}\left(\widehat{\vartheta}_n(x) - \vartheta\right)$ in (*iii*), so gilt

$$L(\vartheta)(a_n(x)) - l_n^{(\cdot)}(x,\vartheta) = \underbrace{r_n(x,\vartheta,a_n(x))a_n(x)}_{=:R_n(x,\vartheta,\widehat{\vartheta_n})} - l_n^{(\cdot)}(x,\widehat{\vartheta}_n).$$

wobei der letzte Term gegen 0 P_ϑ^n-stochastisch aufgrund der Maximum-Likelihood-Eigenschaft konvergiert.

(*iv*) Zentraler Grenzwertsatz:

$$l_n^{(\cdot)}(x,\vartheta) \overset{P_\vartheta^n}{\Rightarrow} \mathcal{N}(0, L(\vartheta)).$$

(*v*) Falls $R_n(x,\vartheta,\widehat{\vartheta}_n) \to 0$ P_ϑ^n-stochastisch und lokal gleichmäßig, folgt der Satz, denn

$$L(\vartheta)(a_n(x)) \overset{P_\vartheta^n}{\Longrightarrow} \mathcal{N}(0, D(\vartheta)).$$

oder

$$a_n(x) \overset{P_\vartheta^n}{\Longrightarrow} \mathcal{N}(0, L(\vartheta)^{-1}L(\vartheta)L(\vartheta)^{-1}) = \mathcal{N}(0, \Lambda(\vartheta)).$$

(*vi*) Betrachte nun

$$\begin{aligned}
A_n(\vartheta) &= \{x \in X^n : \overbrace{\sqrt{n}\left\|\widehat{\vartheta}_n(x) - \vartheta\right\|}^{\|a_n(x)\|} \leq M\} \\
B_n(\vartheta) &= \{x \in X^n : M < \sqrt{n}\left\|\widehat{\vartheta}_n(x) - \vartheta\right\| \leq \sqrt{n}\delta\} \\
C_n(\vartheta) &= \{x \in X^n : \left\|\widehat{\vartheta}_n(x) - \vartheta\right\| > \delta\}
\end{aligned}$$

und zeige

$$P_\vartheta^n\{x \in A_n(\vartheta) : \left\|R_n(x,\vartheta,\widehat{\vartheta}_n)\right\| > \varepsilon\} \to 0$$

für alle $\varepsilon > 0$, d.h. (*v*) gilt auf $A_n(\vartheta)$. Sei $x \in B_n(\vartheta)$. Zeige, dass dann mit

$$D_n \coloneqq D^n(x,\vartheta,\sqrt{n}(\widehat{\vartheta}_n(x) - \vartheta))$$

mit $\det(D_n) \neq 0$ mit Wahrscheinlichkeit $\uparrow 1$ für $x \in B_n(\vartheta)$, da $D_n \to D(\vartheta)$ P_ϑ^n-stochastisch und $\det(D(\vartheta)) \neq 0$. Sei $B_n^*(\vartheta)$ die Menge, wo $|\det(D_n)| \geq \delta > 0$. Mit (*iv*) folgt

$$\sqrt{n}(\widehat{\vartheta}_n(x) - \vartheta) - D_n^{-1}l_n^{(\cdot)}(x,\vartheta) = D_n^{-1}l_n^{(\cdot)}(x,\widehat{\vartheta}_n(x)),$$

wo die rechte Seite P_ϑ^n-stochastisch gegen 0 konvergiert, und der zweite Term der linken Seite stochastisch beschränkt ist. Also gilt

$$(D_n^{-1} - D(\vartheta)^{-1}) l_n^{(\cdot)}(x, \vartheta) \to 0$$

P_ϑ^n-stochastisch auf $B_n(\vartheta)^*$, d.h. auf $B_n(\vartheta)^*$ gilt

$$\sqrt{n}(\vartheta^n(x) - \vartheta) D(\vartheta)^{-1} l_n^{(\cdot)}(x, \vartheta) \to 0$$

P_ϑ^n-stochastisch.
Schließlich folgt für $x \in C_n(\vartheta)$ aufgrund der Konsistenz

$$\mathbb{1}_{C_n(\vartheta)}(x) \left(\sqrt{n}(\widehat{\vartheta}_n - \vartheta) - D(\vartheta)^{-1} l_n^{(\cdot)}(x, \vartheta) \right) \to 0$$

P_ϑ^n-stochastisch.

■

3.2 Nichtparametrische Schätzer für Dichten

Sei $H_p \coloneqq \{f : \mathbb{R} \to [0,\infty) : f \text{ } p\text{-fach differenzierbar}, \int f dx = 1, \int (f^{(p)})^2 dx < \infty\}$ sowie $\mathcal{P} \coloneqq \{P \mid (\mathbb{R}, \mathcal{B}^1) : f \in \frac{dP}{d\lambda}, f \in H_p\}, p \geq 1$. Definiere

$$\kappa : \mathcal{P} \to H_p, P \mapsto f \in \frac{dP}{d\lambda}$$

und $(X_1, \ldots, X_n)$ sei P^n-verteilt, d.h. $X_i \sim P$ i.i.d. und $\widehat{f}^x : \mathbb{R} \to [0,\infty)$ sei eine Schätzung für $f \in H_p$.

Definition 3.11:

(*i*) Der mittlere quadratische Fehler ist definiert als

$$MQE_s(\widehat{f}) = \int (\widehat{f}^x(s) - f(s))^2 P^n(dx)$$

für $s \in \mathbb{R}$.

(*ii*)
$$MIQE(\widehat{f}) = \int \int (\widehat{f}^x(s) - f(s))^2 ds P^n(dx) = \int_{\mathbb{R}} MQE_s(\widehat{f}) ds$$

ist der mittlere integrierte quadratische Fehler.

(*iii*) Sei $\mu(\widehat{f})(s) \coloneqq \int \widehat{f}^x(s) P^n(dx)$. Dann gilt

$$MQE_s(\widehat{f}) = \mathrm{Var}_f(\widehat{f}(s)) + \underbrace{(\mu(\widehat{f})(s) - f(s))^2}_{=b_f(\widehat{f})(s)} = \mathrm{Var}_f(\widehat{f}(s)) + b_f(\widehat{f})(s)^2. \tag{3.13}$$

$b_f(\widehat{f})(s)$ heißt Bias. Also gilt

$$MIQE(\widehat{f}) = \int b_f(\widehat{f})(s)^2 ds + \int \mathrm{Var}_f(\widehat{f}(s)) ds. \tag{3.14}$$

Beispiel 3.15 (Kernschätzer)**:**
Betrachte den Schätzer $\widehat{f}$:

$$\widehat{f}^x(s) \coloneqq \int w(\cdot, s) dF_n^x = \frac{1}{n} \sum_{j=1}^{n} w(x_j, s),$$

wobei $w(\cdot, s)$ ein Kern ist, für den i.A. $w \geq 0$ und $\int w(x,s) ds = 1$ für alle x angenommen wird.
Zum Beispiel betrachte w von der Form $w(x,s) = \frac{1}{h} K\left(\frac{s-x}{h}\right)$, wobei K ein Kern ist. Der Parameter $h > 0$ heißt Bandbreite. Es sei $\int K(s) ds = 1$. Definiere

ferner $K_h(s) := h^{-1}K(sh^{-1})$, d.h. $w(x,s) = K_h * \delta_x(s)$. Als Kern wähle z.B. den Gauß'schen Kern $K(s) = \frac{1}{\sqrt{2\pi}} \exp(-s^2/2)$ oder den Viereckskern $K(s) = \mathbb{1}_{[-1/2,1/2]}$.
In *ii*):

$$\widehat{f}^x(s) := \#\{j : 1 \le j \le n : x_j \in [s - h/2, s + h/2]\}/(nh),$$

dies heißt Histogramm-Schätzer. Für Kernschätzer $\widehat{f}$ mit $K(-x) = K(x)$ gilt

$$\mu(\widehat{f})(s) = \int \underbrace{\frac{1}{h} K\left(\frac{s-x}{h}\right)}_{=\widehat{f}^x(s)} \underbrace{f(x)dx}_{P(dx)} = f * K_h(s) \to f(s) \quad \text{für } h \to 0$$

und

$$\mathrm{Var}_f(\widehat{f}(s)) = \frac{1}{n} \left(\int \underbrace{h^{-2} K\left(\frac{s-x}{h}\right)^2}_{=\widehat{f}^x(s)^2} f(x)dx - \mu_f(\widehat{f}(s))^2 \right). \tag{3.16}$$

Beispiel:
Falls $f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ und der Kern gegeben ist durch den Gauß'schen Kern $K(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, gilt

$$MIQE(\widehat{f}) = \frac{1}{2\sqrt{\pi}} \left(\underbrace{\frac{h^{-1} - (\sigma^2 + h^2)^{-1/2}}{n}}_{\to\infty, h\to 0} + \underbrace{\left(\frac{1}{\sigma} + (\sigma^2 + h^2)^{-1/2} - \frac{2\sqrt{2}}{(2\sigma^2 + h^2)^{1/2}} \right)}_{\to 0, h\to 0} \right).$$

Übung: Finde das Minimum in h.
φ_u sei die Dichte der $\mathcal{N}(0,u)$-Verteilung. Dann gilt

$$\mu_f(\widehat{f})(s) = \varphi_{h^2} * \varphi_{\sigma^2}(s) = \varphi_{\sigma^2+h^2}(s),$$

sodass

$$\mu_f(\widehat{f}(s)) - f(s) = \varphi_{\sigma^2+h^2} - \varphi_{\sigma^2}(s) \to 0 \text{ für } h \to 0.$$

Für die Varianz gilt

$$\begin{aligned}
\mathrm{Var}_f(\widehat{f}(s)) &= \mathrm{Var}_f\left(\frac{1}{n} \sum_{j=1}^n K\left(\frac{s-x_j}{h}\right) h^{-1} \right) = \frac{n}{n^2} \mathrm{Var}_f\left(K\left(\frac{s-\cdot}{h}\right) h^{-1} \right) \\
&\overset{(3.16)}{=} \frac{1}{n} \left(\int \underbrace{\varphi_{h^2}^2(s-x)}_{=\varphi_{h/2}(s-x)/(\sqrt{2\pi}h)\cdot 2^{-1/2}} \varphi_{\sigma^2}(x)dx - \varphi_{\sigma^2+h^2}(s)^2 \right) \\
&= \frac{1}{n} \left(\varphi_{\sigma^2+h^2/2}(s)/(\sqrt{2\pi}h) 2^{-1/2} - \varphi_{\sigma^2+h^2}(s)^2 \right).
\end{aligned}$$

Annahmen:
Man wählt einen Kernschätzer $\widehat{f}^x$ für f, wobei der Kern die folgenden Eigenschaften erfüllt:

(i) $\int K(s)ds = 1.$

(ii) $\int sK(s)ds = 0.$

(iii)

$$\int s^2 K(s)ds = k_2 \neq 0 \tag{3.17}$$

(iv)

$$\sum_{j=0}^{4} f^{(j)}(x)^2 \in L(\lambda) \tag{3.18}$$

(v) $\int K(t)^2(1+t^4)dt < \infty.$

Für den Bias gilt mit der Variablentransformation $t = -\frac{s-x}{h}$ und wegen $K(-t) = K(t)$

$$\begin{aligned} b_h(\widehat{f})(s) &= \int h^{-1}K\left(\frac{s-x}{h}\right)f(x)dx - f(s) \\ &= \int K(-t)f(s+th)dt - f(s) \\ &= \int K(t)(f(s+th) - f(s))dt, \end{aligned}$$

was mit dem Satz von Taylor darstellbar ist als

$$\int K(t)\left(f'(s)th + f''(s)\frac{(th)^2}{2} + \int_0^1 f(s+\vartheta th)\frac{(th)^3}{6}(1-\vartheta)^2 d\vartheta\right)dt.$$
$$= 0 + k_2\frac{h^2}{2}f''(s) + \mathcal{O}(h^3 g_1(s))$$

mit $\int g_1(s)^2 ds < \infty$ wegen Cauchy-Schwarz.

$$\int b_h(\widehat{f})(s)^2 ds = \frac{h^4}{4}k_2^2 \int f''(s)^2 ds + \mathcal{O}(h^5).$$

Für die Varianz hingegen gilt nach Formel (3.16)

$$\begin{aligned} \mathrm{Var}_f(\widehat{f}^x(s)) &= n^{-1}\int h^{-2}K\left(\frac{s-x}{h}\right)^2 f(x)dx - n^{-1}\left[f(s) + b_h(\widehat{f})(s)\right]^2 \\ &= \underbrace{n^{-1}h^{-1}\int K(t)^2 f(s+th)dt}_{I_1(s)} - \underbrace{n^{-1}\left[f(s) + \mathcal{O}(h^2 g_2(s))\right]^2}_{I_2(s)}, \end{aligned}$$

wo $\int g_2^2 ds < \infty$, d.h.

$$\begin{aligned} I_1(s) &= (nh)^{-1} \int (f(s) + \mathcal{O}(hg_3(s))) K(t)^2 dt \\ &= (nh)^{-1} f(s) \int K(t)^2 dt + \mathcal{O}(n^{-1}h^{-1}hg_3(s)) \\ &= (nh)^{-1} f(s) \int K(t)^2 dt + \mathcal{O}(n^{-1}g_3(s)) \end{aligned}$$

d.h.

$$\int \mathrm{Var}_f(\widehat{f})(s) ds = \underbrace{(nh)^{-1}}_{h \to 0} \int K(t)^2 dt + \mathcal{O}(n^{-1})$$

und somit

$$MIQE_h(\widehat{f}) = \frac{1}{4} h^4 k_2^2 \int f''(x)^2 dx + (nh)^{-1} \int K(t)^2 dt + \mathcal{O}(n^{-1}) + \mathcal{O}(h^5).$$

Die optimale Bandbreite ergibt sich zu

$$h_{opt} = k_2^{-2/5} \left(\int K(t)^2 dt \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5} n^{-1/5}$$

und der mittlere quadratische integrierte Fehler zu

$$MIQE_{h_{opt}}(\widehat{f}) = \frac{5}{4} C(K) \left(\int f''(x)^2 dx \right)^{-1/5} n^{-4/5} + \mathcal{O}(n^{-1})$$

mit

$$C(K) = k_2^{2/5} \left(\int K(t)^2 dt \right)^{4/5}.$$

Gesucht ist der Kern mit minimalen $C(K)$, d.h. die optimale Lösung zum Problem: Minimiere $\int K(t)^2 dt$ unter $\int K(t) dt = \int t^2 K(t) dt = 1$ (wegen $K \mapsto k_2^{-1/2} K(t k_2^{-1/2}) \Rightarrow k_2 = 1$).
Lösung:

$$K_{opt}(t) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2) & |t| \leq \sqrt{5} \\ 0 & \text{sonst} \end{cases}$$

Dies heißt Epanechnikov-Kern.

Adaptive Wahl von h_{opt}: Kreuzvalidation
Die Idee:

$$IQE_x = \int \left(\widehat{f}^x(s) - f(s)\right)^2 ds = \underbrace{\int \widehat{f}^x(s)^2 ds - 2\int \widehat{f}^x(s) f(s) ds}_{R(\widehat{f}^x)} + \int f^2 ds.$$

Minimiere IQE_x durch Wahl von h in $\widehat{f}^x$, d.h. minimiere $R(\widehat{f}^x)$ in h. Heuristik: Die Wahl

$$\int \widehat{f}^x(s) dP_f(s) \approx \int \widehat{f}^x(s) dF_n^x$$

mit dem empirischem Maß F_n^x ist nicht zielführend, da dies $\frac{1}{n}\sum_j K\left(\frac{s-x_j}{h}\right) h^{-1} = \widehat{f}^x(s)$ ist und das empirische Maß nur Gewichte an den Beobachtungen x_j hat, d.h. wähle stattdessen

$$\int \widehat{f}^x(s) dP_f(s) \approx \sum_{j=1}^{n} \sum_{l \neq j} \frac{1}{n} K\left(h^{-1}(x_l - x_j)\right) h^{-1}/(n-1) =: \widehat{M}_n.$$

Dies nennt sich auch U-Statistik. Ziel: Minimiere $\widehat{R}(\widehat{f}^x) = \int \widehat{f}^x(s)^2 ds - 2\widehat{M}_n$. Ferner gilt $\mathbb{E}_f \widehat{R}(\widehat{f}^x) = \mathbb{E}_f R(\widehat{f}^x)$ und $\widehat{R}(\widehat{f}^x) \to R(\widehat{f}^x)$ für $n \to \infty$ P_f^n-stochastisch. Für Details siehe z.B. [7], S. 50ff.

4 Testtheorie

Sei $\mathcal{P}$ eine Familie von W-Maßen über $(\mathcal{X}, \mathcal{B})$ und $H \subset \mathcal{P}$ eine echte Teilfamilie.
Problem: Gegeben sei eine Stichprobe $x \in \mathcal{X}$. Entscheide, ob die Verteilung, aus der x entstammt, aus H oder aus $A \coloneqq \mathcal{P} \setminus H$ kommt, d.h. ein Entscheidungsproblem mit zwei Möglichkeiten ($D = \{0, 1\}$), wobei 0 einer Verteilung nach einem Maß aus H und 1 nach einem Maß aus A entspricht. Also ist eine Entscheidungsregel $\delta : \mathcal{X} \to D$ gesucht.

Definition 4.1:
Eine Entscheidungsregel für obiges Problem heißt Test. Man testet eine Hypothese $H \subset \mathcal{P}$, welche für $P \in H$ richtig und für $P \in A$ falsch ist.
Die Hypothese wird mit H (bzw. im Fall $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ mit $H = \{P_\vartheta : \vartheta \in \Theta_H\}$ mit Θ_H) identifiziert. Die Menge $A \coloneqq H^c$ heißt Alternativmenge.
Gegeben sei $x \in \mathcal{X}$. Dann sagt man für eine (nicht-randomisierte) Entscheidungsregel δ, dass die Hypothese angenommen (abgelehnt) wird, falls $\delta(x) = 0$ ($\delta(x) = 1$). Die Menge $\{x \in \mathcal{X} : \delta(x) = 0\}$ heißt Annahmeregion, $\{x \in \mathcal{X} : \delta(x) = 1\}$ heißt Ablehnungs- oder kritische Region.
Für eine Zufallsvariable $X \sim P$ mit $P \in H$ ist $\delta(X(\omega)) = 1$ der Fehler erster Art mit Wahrscheinlichkeit $P(\delta(X(\omega)) = 1)$ (Ablehnungswahrscheinlichkeit) und falls $X \sim P, P \in A$ heißt $\delta(X(\omega)) = 0$ Fehler 2. Art mit Wahrscheinlichkeit $P(\delta(y) = 0), P \in A$.
Konvention: $0 < \alpha < 1$ heißt Signifikanzniveau, falls

$$\sup_{P \in H} P(\delta(x) = 1) \leq \alpha. \tag{4.2}$$

Übliche Wahlen sind $\alpha = 0.05, 0.01, 0.001$.

Ziel: Gegeben (4.2) möchte man den Fehler 2. Art durch die Wahl der Entscheidungsregel δ minimieren, d.h.

$$\min_\delta P(\delta(x) = 0) \text{ für alle } P \in A$$

oder

$$\max_\delta P(\delta(x) = 1) \text{ für alle } P \in A.$$

Für jede Entscheidungsregel bzw. Test δ heißt $\inf_{P \in A} P(\delta(x) = 1)$ die Güte auf der Alternative.

$$\mathcal{P} \ni P \mapsto P(\delta(x) = 1) =: \beta(P)$$

heißt Gütefunktion. Nach (4.2) gilt $\beta(P) \leq \alpha$ für alle $P \in H$.

Beispiel:
Sei $\mathcal{P} = \{\mathcal{N}(\mu, 1)^n : \mu \in \mathbb{R}\}$ und $H = \{\mathcal{N}(0, 1)^n\}$. Sei die Entscheidungsregel gegeben durch

$$\delta(x_1, \ldots, x_n) = \mathbb{1}_{\left\{\left|\frac{1}{n}\sum_j x_j\right| > c\right\}}.$$

Dann ist

$$\begin{aligned} \beta(\mu) &= \mathcal{N}(\mu, 1)^n(\delta(x_1, \ldots, x_n) = 1) = \mathcal{N}_{\mu,1/n}(|x| > c) \\ &= \Phi(\sqrt{n}(-c - \mu)) + 1 - \Phi(\sqrt{n}(c - \mu)) \end{aligned}$$

wobei

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy.$$

Wähle für α gegeben c so, dass für $\mu = 0$ gilt: $\beta(0) = \alpha = 2(1 - \Phi(c\sqrt{n}))$. Es sei das Quantil N_α gegeben durch $1 - \alpha/2 = \Phi(N_\alpha)$. Dann ist $c = \frac{N_\alpha}{\sqrt{n}}$.
Umgekehrt sei eine Beobachtung $x_1, \ldots, x_n$ gegeben. Bestimme das kleinste Signifikanzniveau α oder größtes Quantil N_α, sodass $\delta(x_1, \ldots, x_n) = 1$ oder $\left|\frac{1}{n}\sum_j x_j\right| > \frac{N_\alpha}{\sqrt{n}}$ gilt, d.h. dass die Hypothese abgelehnt wird. Dieses α heißt p-Wert des Tests δ und gibt ein Maß für die Sicherheit, mit der man die Hypothese für diese spezielle Stichprobe ablehnen kann.
Randomisierter Schätzer: $\delta : \mathcal{X} \to \{$ W-Maße auf $D = \{0, 1\}\}, x \mapsto \delta(x)$.

Definition 4.3:
Die Funktion $\varphi : \mathcal{X} \to [0, 1], x \mapsto \varphi(x) = \delta(x)(1)$ heißt Test (kritische Funktion). Dann nennt man die Menge $\{x \in X : \varphi(x) = c\}$ für $c = 1$ Ablehnungsbereich, für $c = 0$ Annahmeregion und für $c \in (0, 1)$ Randomisierungsregion.
Fehler 1. Art: $\mathbb{E}_P\varphi = \int \varphi P(dx), P \in H$. Gütefunktion: $P \mapsto \mathbb{E}_P\varphi$.
Fehler 2. Art: $\mathbb{E}_P(1 - \varphi), P \in H^c$.
Problem: Finde eine Funktion $0 \leq \varphi \leq 1$ mit

$$\mathbb{E}_P\varphi = \max \text{ für alle } P \in A \tag{4.4}$$

und

$$\mathbb{E}_P\varphi \leq \alpha \text{ für alle } P \in H, \tag{4.5}$$

die (4.4) unter der Nebenbedingung in (4.5) optimiert. Ein Test mit (4.5) heißt Test für das Problem $H : A$ zum Niveau α.
Gibt es einen Test mit (4.4) und (4.5), so heißt dieser auf A (gleichmäßig) bester bzw. optimaler Test. Hypothesen bzw. Alternativen, die aus mehr als einem W-Maß bestehen, heißen zusammengesetzt.

Satz 4.6 (Neyman-Pearson Lemma)**:**
Seien $\mathcal{P} = \{P_0, P_1\}, H = \{P_0\}$ *W-Maße über* $(\mathcal{X}, \mathcal{B})$ *mit Dichten* $p_0 \in \frac{dP_0}{d\mu}, p_1 \in \frac{dP_1}{d\mu}$*, wobei o.E.* $\mu = P_0 + P_1$*.*
Dann gilt:

(*i*) *Für das Testproblem* $H = \{P_0\}$ *gegen die Alternative* $A = \{P_1\}$ *gibt es zum Niveau* $0 \leq \alpha \leq 1$ *eine Konstante* $k = k(\alpha, p_0, p_1) \geq 0$ *und einen Test* φ *mit*

$$\mathbb{E}_{P_0}\varphi = \int \varphi \cdot p_0 \, d\mu = \alpha \tag{4.7}$$

und

$$\varphi(x) = \begin{cases} 1 & \text{falls } p_1(x) > kp_0(x) \\ 0 & \text{falls } p_1(x) < kp_0(x) \end{cases} \tag{4.8}$$

(*ii*) *Erfüllt* φ *sowohl* (4.7) *als auch* (4.8)*, so ist* φ *der optimale Test für das Problem* $P_0 : P_1$ *zum Niveau* α*, d.h. für einen beliebigen Test* φ^* *für* $P_0 : P_1$ *gilt*

$$\mathbb{E}_{P_0}\varphi^* \leq \alpha \Rightarrow \mathbb{E}_{P_1}\varphi^* \leq \mathbb{E}_{P_1}\varphi.$$

(*iii*) *Ist* φ *ein optimaler Test zum Niveau* α *für* $P_0 : P_1$*, dann existiert ein* $0 \leq k \leq \infty$*, sodass* (4.8) *gilt, d.h. er ist vom NP-Typ. Auch* (4.7) *gilt, falls es keinen Test mit Güte* 1 *und Niveau kleinergleich* α *gibt.*

Bemerkung:

(*i*) Dieser Test fordert keine Spezifikation für $x \in X$ mit $p_1(x) = kp_0(x)$.

(*ii*) Tests der Form (4.8) heißen Neyman-Pearson-Tests.

Beweis. Für $\alpha = 0, 1$ wähle $k = \infty, 0$. Sei nun $0 < \alpha < 1$. Dann

(*i*) Bezeichne mit $f(a_-)$ den linksseitigen und mit $f(a_+)$ den rechtsseitigen Grenzwert von f in a, falls dieser existiert.
Die Funktion $\alpha(c) \coloneqq P_0(p_1(x) > cp_0(x)) = P_0\left(\frac{p_1(x)}{p_0(x)} > c\right)$ ist monoton fallend in c, wobei $P_0(p_0(x) = 0) = 0$. Dann gilt wegen der Maßstetigkeit von P_0

$$\alpha(-\infty) = 1 \text{ und } \alpha(\infty) = 0$$

und $\alpha(c_+) = \lim_{v \downarrow c} P_0\left(\frac{p_1(x)}{p_0(x)} > v\right) = \alpha(c)$, d.h. α ist rechtsstetig. Dann gilt $\alpha(c_-) - \alpha(c_+) = \alpha(c_-) - \alpha(c) = P_0\left(\frac{p_1(x)}{p_0(x)} = c\right)$.
Sei c_0 so gewählt, dass $\alpha(c_0) \leq \alpha \leq \alpha(c_{0,-})$. Dann definiere

$$\varphi(x) = \begin{cases} 1 & p_1(x) > c_0p_0(x) \\ \frac{\alpha - \alpha(c_0)}{\alpha(c_{0,-}) - \alpha(c_0)} & p_1(x) = c_0p_0(x) \\ 0 & p_1(x) < c_0p_0(x) \end{cases} \tag{4.9}$$

Falls $\alpha(c_{0,-}) = \alpha(c_0)$, so ist $P_0\left(\frac{p_1(x)}{p_0(x)} = c_0\right) = 0$, d.h. φ ist μ-f.s. wohldefiniert. Es gilt

$$\begin{aligned}\mathbb{E}_{P_0}\varphi &= P_0\left(\frac{p_1(x)}{p_0(x)} > c_0\right) + \frac{\alpha - \alpha(c_0)}{\alpha(c_{0,-}) - \alpha(c_0)} P_0\left(\frac{p_1(x)}{p_0(x)} = c_0\right) \\ &= \alpha(c_0) + \alpha - \alpha(c_0) = \alpha.\end{aligned}$$

c_0 ist i.A. nicht eindeutig. Wähle nun $k = c_0$.

(ii) Sei φ ein Test mit (4.7) und (4.8) und sei φ^* ein weiterer Test zum Niveau α, d.h. $E_{P_0}\varphi^* \leq \alpha$. Es gilt

$$\int \underbrace{(\varphi - \varphi^*)(p_1 - kp_0)}_{\geq 0} d\mu \geq 0, \qquad (*)$$

denn wegen $0 \leq \varphi^* \leq 1$ gilt

$$\{\varphi(x) - \varphi^*(x) > 0\} \subset \{\varphi(x) > 0\} = \{p_1(x) - kp_0(x) \geq 0\}$$

sowie

$$\{\varphi(x) - \varphi^*(x) < 0\} \subset \{\varphi(x) < 1\} = \{p_1(x) - kp_0(x) \leq 0\}.$$

Also erhalten wir wegen $k \geq 0$

$$\int (\varphi - \varphi^*) p_1 d\mu \geq k \int (\varphi - \varphi^*) p_0 d\mu$$

und damit

$$\mathbb{E}_{P_1}\varphi - \mathbb{E}_{P_1}\varphi^* \geq k(\alpha - \mathbb{E}_{P_0}\varphi^*) \geq 0.$$

(iii) $a)$: Sei φ ein optimaler Test zum Niveau α für $P_0 : P_1$ und erfülle φ^* (4.7) und (4.8). Setze $A \coloneqq \{x \in X : \varphi(x) \neq \varphi^*(x)\} \cap \{x \in X : p_1(x) \neq kp_0(x)\}$. Nehmen wir an, dass $\mu(A) > 0$. Dann gilt mit (ii)

$$h \coloneqq (\varphi^* - \varphi)(p_1 - kp_0) > 0 \text{ auf } A.$$

Da h auf A^c verschwindet, gilt

$$\int h d\mu = \int_A h d\mu \geq \int_A h \mathbb{1}_{h > \varepsilon} d\mu \geq \varepsilon \mu(A \cap \{h \geq \varepsilon\}) > 0$$

für ε klein genug. Wie in (ii) folgt aus $\int h d\mu > 0$ sofort

$$\mathbb{E}_{P_1}\varphi^* > \mathbb{E}_{P_1}\varphi + k(\underbrace{\mathbb{E}_{P_0}\varphi^*}_{=\alpha} - \underbrace{\mathbb{E}_{P_0}\varphi}_{\leq \alpha}) \geq \mathbb{E}_{P_1}\varphi$$

in Widerspruch zur Optimalität von φ. Also ist $\mu(A) = 0$.
b): Sei φ_0 optimal mit $\mathbb{E}_{P_0}\varphi_0 \leq \alpha$ und $\mathbb{E}_{P_1}\varphi_0 < 1$.
Annahme: $\alpha_0 \coloneqq \mathbb{E}_{P_0}\varphi_0 < \alpha$. Definiere einen neuen Test mittels $\Psi_0 \coloneqq (1-\varepsilon)\varphi_0 + \varepsilon$ mit $\varepsilon = \frac{\alpha-\alpha_0}{1-\alpha_0}, 0 < \alpha_0 < 1$. Dann ist $0 \leq \psi_0 \leq 1$ und

$$\mathbb{E}_{P_0}\psi_0 = (1-\varepsilon)\mathbb{E}_{P_1}\varphi_0 + \varepsilon = \left(1 - \frac{\alpha-\alpha_0}{1-\alpha_0}\right)\alpha_0 + \frac{\alpha-\alpha_0}{1-\alpha_0} = \alpha$$

und

$$\mathbb{E}_{P_1}\psi_0 = (1-\varepsilon)\mathbb{E}_{P_1}\varphi_0 + \varepsilon \cdot 1 > \mathbb{E}_{P_1}\varphi_0, \text{ da } \mathbb{E}_P\varphi_0 < 1$$

im Widerspruch zur Optimalität von φ_0. Also gilt $\mathbb{E}_{P_0}\varphi_0 = \alpha$.

■

Korollar 4.10:
Sei β die Güte des besten Tests für $P_0 : P_1$ zum Niveau α ($0 < \alpha < 1$). Wenn $P_0 \neq P_1$, so gilt $\beta > \alpha$.

Beweis. Der Test $\varphi(x) \equiv \alpha$ hat Niveau α. Also gilt

$$\beta \geq \mathbb{E}_{P_1}\varphi = \mathbb{E}_{P_1}\alpha = \alpha.$$

Angenommen es gilt $\beta = \alpha$, so ist $\varphi \equiv \alpha$ mit $1 \geq \alpha = \beta$ bester Test, d.h. er erfüllt die NP-Struktur (4.8). Also gilt $p_1(x) = kp_0(x)$ $\mu-$f.ü., d.h.

$$1 = \int p_1(x)\mu(dx) = k\int p_0(x)\mu(dx) \Longrightarrow k = 1.$$

Also ist $P_0 = P_1$. Widerspruch. ■

Beispiel 4.11 (Stichprobentests)**:**
Es seien N Produkte gegeben, wovon $D \leq N$ defekt sind. Ziehe aus den Produkten eine Stichprobe der Größe $1 \leq n < N$. Seien $x \leq n$ defekte Produkte in der Stichprobe. Zum unbekannten Parameter D erhält man eine hypergeometrische Verteilung, d.h. es gilt

$$P_{D,N,n}(x \text{ Defekte in der Stichprobe}) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}.$$

Betrachte dann die Hypothese $H = \{P_{D_0,N,n}\}$ und die Alternative $A = \{P_{D_1,N,n}\}$, wobei $D_0 < D_1 \leq D$.
Der zugehörige Neyman-Pearson-Test hat die Struktur

$$\varphi(x) = \begin{cases} 1 & \frac{P_{D_1,N,n}}{P_{D_0,N,n}}(x) > k \\ \gamma & \frac{P_{D_1,N,n}}{P_{D_0,N,n}}(x) = k \\ 0 & \frac{P_{D_1,N,n}}{P_{D_0,N,n}}(x) < k \end{cases}.$$

Die Größe

$$\frac{P_{D+1,N,n}}{P_{D,N,n}}(x) = \frac{D+1}{N-D}\frac{N-D-n+x}{D+1-x} \qquad (*)$$

ist monoton steigend in x, d.h. auch

$$q(x) \coloneqq \frac{P_{D_1,N,n}(x)}{P_{D_0,N,n}(x)} = \text{ Produkte von Quotienten in } (*)$$

ist aufsteigend in x. Somit lässt sich die Neyman-Pearson-Struktur umschreiben zu

$$\{0 \leq x \leq D : q(x) > k\} = \{0 \leq x \leq D : x > c_k\}$$

für ein geeignetes c_k. Also ist Neyman-Pearson-Test zum obigen Problem gegeben durch

$$\varphi(x) = \begin{cases} 1 & x > c_k \\ \gamma & x = c_k \\ 0 & x < c_k \end{cases}$$

wobei sich c_k und γ aus $\alpha = P_{D_0,N,n}(x > c_k) + \gamma P_{D_0,N,n}(x = c_k)$ bestimmen lassen.

Zusammengesetzte Hypothesen
Sei $\mathcal{P}$ eine eindimensionale Familie von W-Maßen auf $(\mathcal{X}, \mathcal{B})$ mit einem identifizieren Parameter $\vartheta, \vartheta \in (a, b) \subset \mathbb{R}$.
Teste $H := \{P_\vartheta \in \mathcal{P} : \vartheta \leq \vartheta_0\}$ gegen $A := \{P_\vartheta \in \mathcal{P} : \vartheta > \vartheta_0\}$ für ein gegebenes $\vartheta_0 \in (a, b)$. Gesucht ist ein gleichmäßig bester Test für $H : A$ zum Niveau α (vgl. auch (4.4) und (4.5)). Im Allgemeinen gibt es keinen solchen Test.
Hat $\mathcal{P}$ jedoch zusätzlich monotone Dichtequotienten, so lässt sich solch ein Test finden.

Definition 4.12:
Sei $\mathcal{P} \ll \mu, \mu$ $\sigma-$endlich mit Dichten $p_\vartheta \in \frac{dP_\vartheta}{d\mu}$ und $P_\vartheta \ll P_{\vartheta'}$ für alle ϑ, ϑ'. $\mathcal{P}$ hat monotone Dichtequotienten, falls es eine messbare Abbildung $T : \mathcal{X} \to \mathbb{R}$ gibt mit

$$\frac{p_{\vartheta'}(x)}{p_\vartheta(x)} = h_{\vartheta',\vartheta}(T(x)) \quad \text{für alle } x \in \mathcal{X},$$

wo $t \mapsto h_{\vartheta,\vartheta'}(t)$ strikt monoton wachsend ist für $\vartheta' > \vartheta$.

Lemma 4.13:
Falls $\mathcal{P}$ einen monotonen Dichtequotienten hat, so gibt es für $H : \vartheta \leq \vartheta_0$ gegen $A : \vartheta > \vartheta_0$ einen gleichmäßig besten Test. Dieser ist gegeben durch

$$\varphi(x) = \begin{cases} 1 & T(x) > C \\ \gamma & T(x) = C \\ 0 & T(x) < C \end{cases} \tag{4.14}$$

mit T aus Definition 4.12, und C und γ können durch

$$\mathbb{E}_{\vartheta_0}\varphi = \alpha \tag{4.15}$$

bestimmt werden. Ferner gilt

(*i*) *φ hat Neyman-Pearson-Struktur.*

(*ii*) *$\vartheta \mapsto \beta(\vartheta) = \mathbb{E}_\vartheta\varphi$ ist strikt monoton wachsend, falls $\beta(\vartheta) < 1$.*

(*iii*) *Der Test (4.14) und (4.15) ist auch gleichmäßig bester Test für die Probleme $H' : \vartheta \leq \vartheta'$ gegen $A' : \vartheta > \vartheta'$ zum Niveau $\alpha' = \beta(\vartheta') = \mathbb{E}_{\vartheta'}\varphi$ für alle ϑ'.*

(*iv*) *Für $\vartheta < \vartheta_0$ minimiert der Test (4.14)/(4.15) den Fehler zweiter Art unter allen Neyman-Pearson-Tests.*

Beweis. $(i), (ii)$: Für das Neyman-Pearson-Problem $H = \{P_{\vartheta_0}\}$ gegen $A = \{P_{\vartheta_1}\}$ für $\vartheta_1 > \vartheta_0$ ist der optimale Test nach 4.6 durch

$$\varphi(x) = \begin{cases} 1 & h_{\vartheta_1,\vartheta_0}(T(x)) > k \\ \gamma & h_{\vartheta_1,\vartheta_0}(T(x)) = k \\ 0 & h_{\vartheta_1,\vartheta_0}(T(x)) < k \end{cases}$$

gegeben, was wegen dem monotonen Dichtequotienten äquivalent ist zu

$$\varphi(x) = \begin{cases} 1 & T(x) > C(k) \\ \gamma & T(x) = C(k) \\ 0 & T(x) < C(k) \end{cases}$$

für ein C, γ mit $\mathbb{E}_{\vartheta_0}\varphi = \alpha$ und $C = C(\vartheta_0, \alpha)$.
Wegen Satz 4.6 ist φ auch bester Test für das Problem $\{P'_\vartheta\} : \{P_{\vartheta''}\}$ zum Niveau $\alpha' = \beta(\vartheta')$ mit $\vartheta' < \vartheta''$.
Aus Korollar 4.10 folgt dann $\beta(\vartheta'') > \beta(\vartheta')$, d.h. (ii), und wegen

$$\mathbb{E}_\vartheta\varphi < \mathbb{E}_{\vartheta_0}\varphi = \alpha \text{ für } \vartheta < \vartheta_0$$

folgt schließlich mit Korollar 4.10 auch

$$\mathbb{E}_\vartheta\varphi \leq \alpha \quad \text{für alle } \vartheta \leq \vartheta_0, \tag{4.16}$$

d.h. φ maximiert die Güte $\mathbb{E}_{\vartheta_1}\varphi$ unter allen Tests mit (4.15), also auch unter allen Tests mit (4.16). Da φ unabhängig von $\vartheta_1 > \vartheta_0$ ist, ist φ gleichmäßig bester Test für $H : A$.
(iii): Genauso.
(iv): Folgt aus der Neyman-Pearson-Struktur von φ und Satz 4.6 mit Umkehrung der Ungleichungen (Übung). Ähnliche Lösung erhält man für $H : \vartheta \geq \vartheta_0$ gegen $A : \vartheta < \vartheta_0$. ∎

Korollar 4.17:
Sei $\mathcal{P} = \{P_\vartheta : \vartheta \in \mathbb{R}\}$ eine exponentielle Familie bzgl. μ, d.h.

$$p_\vartheta(x) = C(\vartheta)h(x)\exp\left[Q(\vartheta)T(x)\right],$$

bei der $\vartheta \mapsto Q(\vartheta)$ strikt monoton ist.
Dann gibt es einen gleichmäßig besten Test φ für $H : \vartheta \leq \vartheta_0$ gegen $A : \vartheta > \vartheta_0$. Ist $Q(\vartheta)$ monoton wachsend, so ist der optimale Test durch (4.14)/(4.15) gegeben. Falls $Q(\vartheta)$ monoton fallend ist, so sind die Ungleichungen in der Definition von φ umzukehren.

Beweis. Definiere

$$h_{\vartheta_2,\vartheta_1}(T(x)) = \frac{C(\vartheta_2)}{C(\vartheta_1)} \exp\left[\underbrace{(Q(\vartheta_2) - Q(\vartheta_1))}_{>0} T(x)\right] = \frac{p_{\vartheta_2}(x)}{p_{\vartheta_1}(x)},$$

was monoton in $T(x)$ ist und nutze den vorherigen Satz. ∎

Beispiel 4.18:
Sei $\mathcal{P} = \{\mathcal{N}(\mu, 1)^n : \mu \in \mathbb{R}\}$. Teste $H : \mu \leq \mu_0$ gegen $A : \mu > \mu_0$ zum Niveau α. Dann ist $T(x) = \sum_j x_j$ eine suffiziente Statistik. Mit $z_j \coloneqq x_j - \mu_0$ gilt

$$\begin{aligned} \alpha &= \mathcal{N}(\mu_0, 1)^n(x_1 + \ldots + x_n > c) = \mathcal{N}(0,1)^n \left(\frac{z_1 + \ldots + z_n}{\sqrt{n}} > \frac{c - n\mu_0}{\sqrt{n}}\right) \\ &= \mathcal{N}(0,1)\left(z > \frac{c - n\mu_0}{\sqrt{n}}\right) \end{aligned}$$

d.h. wähle $N_\alpha \coloneqq \frac{c - n\mu_0}{\sqrt{n}}$ und $c = n\mu_0 + \sqrt{n}N_\alpha$. Dann ist

$$\varphi(x_1, \ldots, x_n) = \mathbb{1}\left(\bar{x}_n > \mu_0 + \frac{N_\alpha}{\sqrt{n}}\right)$$

ein gleichmäßig bester Test.

Mehrdimensionale Hypothesen und Störparameter
Teste $H = \{\mathcal{N}(\mu, \sigma^2) : \mu \leq \mu_0, \sigma^2 > 0\}$ gegen $A = \{\mathcal{N}(\mu, \sigma^2) : \mu > \mu_0, \sigma^2 > 0\}$. Hier ist $\sigma^2 > 0$ ein Störparameter.

Definition 4.19:
Ein Test φ hat Neyman-Struktur für $\mathcal{P}_0 \subset \mathcal{P}$, falls es eine Statistik $S : X \to Z$ gibt, die suffizient für $\mathcal{P}_0$ ist und wo die bedingte faktorisierte Erwartung $\mathbb{E}_{P_0}(\varphi \mid S = \cdot)$ (die wegen der Suffizienz unabhängig von $P_0 \in \mathcal{P}_0$ ist) kosntant ist.

Idee: Falls $\mathbb{E}_P\varphi = \mathbb{E}_P(\mathbb{E}_P(\varphi \mid S)) = \alpha$ für alle $P \in \mathcal{P}_0$ ist und falls $\mathcal{P}_0 \circ S$ vollständig ist, impliziert dies $\mathbb{E}_P(\varphi \mid S) = \alpha$ $P-$f.s, d.h. finde im obigen Beispiel $\{\mu \leq \mu_0\{: \{\mu > \mu_0\}$, $\sigma^2 > 0$ beliebig, einen Test mit $\mathbb{E}_{\mu_0,\sigma^2}\varphi = \alpha$ für alle $\sigma^2 > 0$ und schließe daraus auf das Niveau eines konditionierten/bedingten Tests, der unabhängig vom Störparameter σ^2 ist.

Satz 4.20:
Sei $\mathcal{P} = \{P_{\vartheta,\eta} : \vartheta \in \Theta, \eta \in N\}, \Theta = (a, b) \subset \mathbb{R}, N \subset \mathbb{R}^k$ eine exponentielle Familie, d.h.

$$p(x, \vartheta, \eta) = c(\vartheta, \eta)g(x) \exp\left(a_0(\vartheta, \eta)T_0(x) + \sum_{j=1}^{k} a_j(\vartheta, \eta)T_j(x)\right).$$

Ferner existiere ein $\vartheta_0 \in \Theta$ mit

$$a_0(\vartheta_0, \eta) = a_*(\vartheta_0) \text{ für alle } \eta \in N \tag{4.21}$$

sowie

$$a_0(\vartheta, \eta) \lesseqgtr a_*(\vartheta_0) \text{ für alle } \eta \in N, \text{ falls } \vartheta \lesseqgtr \vartheta_0. \tag{4.22}$$

Dann gilt für jede Stichprobengröße n:

(i) *Für $\alpha \in (0, 1)$ gibt es einen Test φ zum Niveau α mit Neyman-Struktur für $H = \{P^n_{\vartheta,\eta} : \vartheta \leq \vartheta_0, \eta \in N\}$ gegen $A = \{P^n_{\vartheta,\eta} : \vartheta > \vartheta_0, \eta \in N\}$, sodass eine Funktion $C_{n,\alpha} : \mathbb{R}^k \to \mathbb{R}$ existiert mit*

$$\varphi(x) = \begin{cases} 1 & \sum_{j=1}^n T_0(x_j) > C_{n,\alpha}\left(\left(\sum_{j=1}^n T_l(x_j)\right)_{l=1,\ldots,k}\right) \\ 0 & \text{sonst} \end{cases} \tag{4.23}$$

(ii) *Dieser Test zum Niveau α (d.h. $\mathbb{E}_{\vartheta_0,\eta}\varphi = \alpha$ $\forall \eta \in N$) für $H : A$ maximiert (minimiert) die Güte $\mathbb{E}_{\vartheta,\eta}\varphi$ gleichzeitig für alle Alternativen $P_{\vartheta_1,\eta_1}, \vartheta_1 > \vartheta_0$ (bzw. $\vartheta_1 < \vartheta_0$) und beliebiges $\eta_1 \in N$ in der Klasse aller Tests, die $\mathbb{E}_{\vartheta_0,\eta}\varphi = \alpha$ für alle $\eta \in N$ erfüllen.*

Beweis.

(i) Sei $T = (T_1, \dots, T_k)$ und OE sei $n = 1$. Dann gibt es eine bedingte faktorisierte Verteilung $M_\vartheta : \mathbb{R}^k \times \mathcal{B} \to [0,1]$ mit $M_{\vartheta_0}(t, A) := P_{\vartheta_0,\eta}(T_0 \in A \mid T = t)$, welche wegen der Suffizienz von T unabhängig von $\eta \in N$ ist.
Wegen der Bemerkung zu Satz 4.6 setze dort $\mu_t(\cdot) = M_{\vartheta_0}(t, \cdot)$ als dominierenden Maß und $p_0(x) = 1, p_1(x) = x$. Dann gibt es nach dem Neyman-Pearson-Lemma 4.6 (i) einen Test

$$\psi(x,t) = \begin{cases} 1 & x > c_\alpha(t) \\ \gamma_\alpha(t) & x = c_\alpha(t) \\ 0 & x < c_\alpha \end{cases} \tag{4.24}$$

mit $\int \psi(x,t) M_{\vartheta_0}(t, dx) = \alpha$ für $t \in \mathbb{R}^k$. Aus der Wahl von $c_\alpha(t), \gamma_\alpha(t)$ folgt, dass $\varphi := \psi(T_0, T) : X \to [0,1]$ messbar ist und es gilt

$$\mathbb{E}_{\vartheta_0,\eta_0}(\varphi \mid T = t) = \int \psi(x,t) M_{\vartheta_0}(t, dx) = \alpha \quad \text{für alle } \eta \in N \tag{4.25}$$

und damit nach Integration

$$\mathbb{E}_{\vartheta_0,\eta_0}\varphi = \alpha \text{ für alle } \eta_0 \in N. \tag{4.26}$$

Also hat φ Neyman-Struktur im Sinne von Defintion 4.19.
(ii): Betrachte $H = \{(\vartheta_0, \eta_0)\} : \{(\vartheta_1, \eta_1)\} = A$ mit $\vartheta_1 > \vartheta_0$. Schreibe die exponentielle Dichte in der Form

$$p(\cdot, \vartheta, \eta) = \underbrace{\exp\left(a_0(\vartheta,\eta) T_0(x)\right)}_{=:H(T_0,\vartheta,\eta)} \cdot \underbrace{c(\vartheta,\eta) \exp\left(\sum_j a_j(\vartheta,\eta) T_j(x)\right)}_{=:G(T,\vartheta,\eta)}.$$

Aus (4.21) folgt nun

$$H(T_0, \vartheta_0, \eta_0) = H_*(T_0, \vartheta_0) \text{ für alle } \eta_0 \in N \tag{4.27a}$$

und

$$s \mapsto \frac{H(s, \vartheta_1, \eta_1)}{H_*(s, \vartheta_0)} \tag{4.27b}$$

ist strikt monoton wachsend für $\vartheta_1 > \vartheta_0$ und alle $\eta_1 \in N$. Betrachte die Größe

$$\widehat{c}_\alpha(t) = \frac{H(c_\alpha(t), \vartheta_1, \eta_1)}{H_*(c_\alpha(t), \vartheta_0)} G(t, \vartheta_1, \eta_1). \tag{4.28}$$

Für $\vartheta_1 > \vartheta_0$ folgt aus (4.27)-(4.28)

$$p(x, \vartheta_1, \eta_1) \gtrless \widehat{c}_\alpha(T(x)) H_*(T_0(x), \vartheta_0) g(x), \tag{4.29}$$

was mittels Division auf beiden Seiten durch $g(x)H_*(T_0(x), \vartheta_0)$ und $G(T(x), \vartheta_1, \eta_1)$ aus $\widehat{c}_\alpha(T(x))$

$$T_0(x) \gtrless c_\alpha(T(x)) \tag{4.30}$$

impliziert, denn

$$\frac{H(T_0, \vartheta_1, \eta_1)}{H_*(T_0, \vartheta_0)} \gtrless \frac{H(c_\alpha(T), \vartheta_1, \eta_1)}{H_*(c_\alpha(T), \vartheta_0)} \Rightarrow T_0 \gtrless c_\alpha(T).$$

Behauptung: Ein Test mit (4.30) von Neyman-Struktur und Niveau α auf $(\vartheta_0, \eta_0), \eta_0 \in N$ ist bester Test für alle Paare $(\vartheta_0, \eta_0) : (\vartheta_1, \eta_1), \vartheta_1 > \vartheta_0$ unter allen Neyman-Pearson-Tests zum Niveau α.
Denn für eine kritische Funktion $\widehat{\varphi}$ wie oben und den Test φ aus (4.30) gilt mit Hilfe von Satz 4.6

$$(\varphi(x) - \widehat{\varphi}(x))(p(x, \vartheta_1, \eta_1) - \widehat{c}_\alpha(T(x)) H_*(T_0(x), \vartheta_0) g(x)) \geq 0 \tag{4.31}$$

für alle $x \in X$. Integration von (4.31) bezüglich μ liefert

$$\mathbb{E}_{\vartheta_1, \eta_1}(\varphi - \widehat{\varphi}) = \int (\varphi - \widehat{\varphi}) p(\cdot, \vartheta_1, \eta_1) d\mu \geq \int (\varphi - \widehat{\varphi}) \widehat{c}_\alpha(T) g H_*(T_0, \vartheta_0) d\mu =: I,$$

wobei das Integral wegen (4.31) und $|\varphi - \widehat{\varphi}| \leq 1$ existiert. Da $\varphi, \widehat{\varphi}$ Tests mit Neyman-Pearson-Struktur zum Niveau α sind, gilt

$$\mathbb{E}_{\vartheta_0, \eta_0}(\varphi - \widehat{\varphi}) f \circ T = \mathbb{E}_{\vartheta_0, \eta_0} \left(\underbrace{\mathbb{E}_{\vartheta_0, \eta_0}(\varphi - \widehat{\varphi} \mid T)}_{=0} \right) f \circ T = 0$$

falls f so, dass die linke Seite existiert. Wähle nun $f(t) := \frac{\widehat{c}_\alpha(t)}{G(t, \vartheta_0, \eta_0)}, t \in \mathbb{R}^k$. Dann gilt für das Integral I von oben

$$I = \int (\varphi - \widehat{\varphi}) f \circ T \underbrace{H_*(T_0, \vartheta_0) G(T, \vartheta_0, \eta_0) g}_{p(\cdot, \vartheta_0, \eta_0)} d\mu = \mathbb{E}_{\vartheta_0, \eta_0}(\varphi - \widehat{\varphi}) f \circ T = 0,$$

d.h. $\mathbb{E}_{\vartheta_1, \eta_1} \varphi \geq \mathbb{E}_{\vartheta_1, \eta} \widehat{\varphi}$ für alle $\vartheta_1 > \vartheta_0$ mit $\eta_0, \eta \in N$, d.h. φ ist optimal in dieser Klasse.

■

Zur Berechnung der kritischen Region CR ist die Kenntnis der bedingten Verteilung von T_0 gegeben T erforderlich. Dies ist einfach, falls T_0 und T unabhängig unter $P_{\vartheta_0,\eta}$ für alle $\eta \in N$ sind. Daher wollen wir die Entscheidungsregel so monoton transformieren, dass das der Fall ist.

Lemma 4.32:
Sei $\mathcal{P} = \{P_{\vartheta,\eta} : \vartheta \in \Theta, \eta \in N\}$ eine exponentielle Familie. Dann gilt:

(*i*) *Sei $q : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}, (t_0, t) \mapsto q(t_0, t)$ messbar und strikt monoton wachsend in t_0 für jedes t.*
Ist $P_{\vartheta_0,\eta} \circ q(T_0, T)$ für $T_0 : \mathcal{X} \to \mathbb{R}, T : \mathcal{X} \to \mathbb{R}^k$ unabhängig von $\eta \in N$, so ist $q(T_0, T)$ stochastisch unabhängig von T unter $P_{\vartheta_0,\eta}$ für alle $\eta \in N$.

(*ii*) *Falls $P_{\vartheta_0,\eta} \circ q(\sum_j T_0(x_j), \sum_j T(x_j))$ eine stetige Verteilungsfunktion hat, so ist die kritische Region CR von Satz 4.20 gegeben durch*

$$CR = \left\{ x \in \mathcal{X}^n : q\left(\sum_j T_0(x_j), \sum_j T(x_j) \right) \geq c_\alpha \right\},$$

wo c_α durch $P_{\vartheta_0,\eta}(CR) = \alpha$ für alle $\eta \in N$ gegeben ist.

Beweis. Sei im Folgenden $n = 1$.

(*i*) Da $\{P_{\vartheta_0,\eta} \circ T, \eta \in N\}$ beschränkt vollständig ist, gilt mit $R \coloneqq q(T_0, T)$, dass

$$a(\vartheta_0, C) \coloneqq \mathbb{E}_{\vartheta_0,\eta} \mathbb{1}_C(R) = \int \mathbb{E}_{\vartheta_0,\eta} \left(\mathbb{1}_C(R) \mid T = t \right) P_{\vartheta_0,\eta} \circ T(dt)$$

unabhängig von η ist, da T suffizient für N ist. Dies impliziert

$$\mathbb{E}_{\vartheta_0,\eta} \left(\mathbb{1}_C(R) \mid T = t \right) = a(\vartheta_0, C) \; P_{\vartheta_0,\eta} \circ T - \text{f.s.}$$

oder

$$\mathbb{E}_{\vartheta_0,\eta} \left(\mathbb{1}_C(R) \mathbb{1}_D(T) \right) = a_0(\vartheta_0, C) \; P_{\vartheta_0,\eta} - \text{f.s.}$$

mit $D \in \sigma(T)$ und $a_0(\vartheta_0, C) = P_{\vartheta_0,\eta}(R \in C)$, falls man $D = \mathbb{R}^k$ wählt. Damit sind R und T unabhängig.

(*ii*) Wegen der Monotonie von $t_0 \mapsto q(t_0, t)$ gilt für die kritische Region CR von Satz 4.20:

$$T_0(x) \lesseqgtr c_\alpha(T(x)) \iff q(T_0(x), T(x)) \lesseqgtr \underbrace{q(c_\alpha(T(x)), T(x))}_{\widehat{q}(T)}$$

und wegen Teil (i) gilt für alle $\eta \in N$

$$\begin{aligned}\alpha &= P_{\vartheta_0,\eta}\left(\underbrace{q(T_0,T)}_{=R} > \widehat{q}(T)\right) = \mathbb{E}_{\vartheta_0,\eta}\mathbb{1}_{\{q(T_0,T)>\widehat{q}(T)\}} \\ &= \int P_{\vartheta_0,\eta}(R > \widehat{q}(T) \mid T = t) P_{\vartheta_0,\eta} \circ T(dt) \\ &= \int \underbrace{P_{\vartheta_0,\eta}(R > \widehat{q}(t))}_{\text{unabhängig von } \eta} P_{\vartheta_0,\eta} \circ T(dt),\end{aligned}$$

wobei eingeht, dass die bedingte faktorisieret Erwartung konstant in $T = t$ ist. Wegen der Vollständigkeit von $\{P_{\vartheta_0,\eta} \circ T, \eta \in N\}$ ist folglich $P_{\vartheta_0,\eta}(R > \widehat{q}(t))$ konstant in t, also ist $P_{\vartheta_0,\eta}(R > c_\alpha)$ unabhängig von η.

■

Beispiel 4.33:

(i) Sei $\mathcal{P} = \{\mathcal{N}(\vartheta, \sigma^2), \vartheta \in \mathbb{R}, \sigma^2 > 0\}$ mit $H = \{\mathcal{N}(\vartheta_0, \sigma^2), \sigma^2 > 0\}$, wobei OE $\vartheta_0 = 0$ (sonst transformiere $x \mapsto x - \vartheta_0$) eine exponentielle Familie mit $a_0(\vartheta, \sigma^2) = \frac{\vartheta}{\sigma^2}$ und $a_1(\vartheta, \sigma^2) = -\frac{1}{2\sigma^2}$.
Ein gleichmäßig bester Test ist nach Satz 4.20 gegeben durch $\mathbb{1}_{\{}(x_1, \ldots, x_n) : \sum_j x_j > C_{n,\alpha}(\sum_j x_j^2)\}$.
Mit $S_n^2 = \frac{1}{n-1}\sum_j (x_j - \bar{x}_n)^2$ gilt, dass $T_0 \mapsto q(T_0, T) = \frac{\bar{x}_n}{S_n}$ strikt monoton wachsend in T_0 ist, denn

$$q(T_0, T) = \frac{\sum_j x_j \cdot \frac{\sqrt{n-1}}{n}}{\sqrt{\sum_j x_j^2 - (\sum_j x_j)^2/n}} = \frac{\sqrt{n-1}}{n} \frac{T_0}{(T - T_0^2/n)^{1/2}}$$

ist für jedes T monoton steigend in T_0 und unabhängig von Skalierung $x \to \sigma \cdot x, \sigma > 0$. Ferner gilt für $x_j \mapsto \lambda x_j, \lambda > 0$, dass

$$\mathcal{N}(0, \sigma^2)^n = \mathcal{N}(0, 1)^n \circ \left(x \mapsto x\sigma^{-1}\right).$$

Also ist $\mathcal{N}(0, \sigma^2)\left(\frac{\bar{x}_n}{S_n} > C_\alpha\right)$ unabhängig von $\sigma^2 > 0$ und somit ist $\varphi(x) = \mathbb{1}_{\{}(x_1, \ldots, x_n) : \frac{\bar{x}_n}{S_n} > c_{n,\alpha}\}$ ein gleichmäßig bester Test für $\vartheta = 0, \sigma^2 > 0$ gegen $\vartheta > 0, \sigma^2 > 0$.
Übung: $\frac{\bar{x}_n}{S_n}$ hat eine sogenannte Studentsche t-Verteilung mit $n - 1$ Freiheitsgraden.

(ii) Teste Unabhängigkeit von zwei Komponenten (X, Y) mittels $(X_j, Y_j)_{j=1,\ldots,n}$, wobei $X \sim \mathcal{N}(\mu_1, \sigma_1^2), Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, d.h. (X_i, Y_i) ist zweidimensional

normalverteilt mit Kovarianzmatrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$, d.h. betrachte $\mathcal{P} = \{\mathcal{N}^n(\mu_j, \sigma_j^2, \rho) : \mu_j \in \mathbb{R}, \sigma_j^2 > 0, |\rho| \leq 1\}$. Die Dichte ist gegeben durch

$$\begin{aligned} & (x,y) \mapsto (2\pi)^{-1}(|\Sigma|)^{-1/2} C(\mu_j, \sigma_j^2, \rho) \exp\left[-1/2 \left\langle x, \sigma^{-1} x \right\rangle\right] \\ = \quad & (2\pi)^{-1}(|\Sigma|)^{-1/2} C(\mu_j, \sigma_j^2, \rho) \exp\left[-\frac{1}{1-\rho^2}\left(\sum_{j=1}^{5} a_j(\mu_j, \sigma_j^2, \rho) T_j(x,y)\right)\right] \end{aligned}$$

mit $a_1 = \rho\frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1}, a_2 = \rho\frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2}, a_3 = \frac{1}{2\sigma_1^2}, a_4 = \frac{1}{2\sigma_2^2}, a_5 = \frac{\rho}{\sigma_1\sigma_2}$ und den Statistiken $T_1 = x, T_2 = y, T_3 = x^2, T_4 = y^2, T_5 = xy$.
Teste nun $H = \{\mathcal{N}^n(\mu_j, \sigma_j^2, 0) : \mu_j \in \mathbb{R}, \sigma_j^2 > 0\}$ gegen $A = \{\mathcal{N}^n(\mu_j, \sigma_j^2, \rho) : \mu_j \in \mathbb{R}, \sigma_j^2 > 0, |\rho| \neq 0\}$. Da unkorrelierte Gauß-verteilte Größen unabhängig sind, sind die Verteilungen in H unkorreliert und daher unabhängig. Hier erfüllt die exp-Dichte (4.19) mit $\vartheta = \rho$ gegeben $\vartheta_0 = 0, \eta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ sowie $T_0(x,y) = xy, T(x,y) = (x, y, x^2, y^2)$ mit $a(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = -\frac{1}{1-\rho^2}(a_1, a_2, a_3, a_4)$ von oben. Also ist der gleichmäßig beste Test nach Satz 4.20 gegeben durch $\varphi(x) = \mathbb{1}_{\mathrm{CR}}$ mit

$$\mathrm{CR} = \left\{((x_1,y_1),\ldots,(x_n,y_n)) : \sum_j x_j y_j > c_{n,\alpha}\left(\sum x_j, \sum y_j, \sum x_j^2, \sum y_j^2\right)\right\}.$$

Benutze die Schätzung

$$\tau_n \coloneqq \frac{\frac{1}{n}\cdot\sum_{j=1}^n (x_j - \bar{x}_n)(y_j - \bar{y}_n)}{\left(\sum (x_j - \bar{x}_n)^2 \sum (y_j - \bar{y}_n)^2 \cdot \frac{1}{n-1}^2\right)^{1/2}}$$

für ρ. Mit $\sum_j x_j y_j \mapsto \tau_n$ ist

$$\tau_n = \tau_n(T_0, T) = \frac{1/n \cdot T_0 - \frac{T_1 T_2}{n^2}}{\sqrt{1/(n-1)\cdot(T_3 - T_1^2/n^2)\cdot 1/(n-1)\cdot(T_4 - T_2^2/n^2)}}$$

monoton steigend in T_0 gegeben T und unabhängig von $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2 > 0$, da dies invariant gegenüber Verschiebung der Mittelwerte und der Änderung der Varianz ist. Also ist

$$\varphi(x) = \mathbb{1}_{\{}((x_1,y_1),\ldots,(x_n,y_n)) : \tau_n > c_{n,\alpha}\}$$

ein gleichmäßig bester Test für $\{\rho = 0\} : \{\rho > 0\}$ in der Klasse aller Tests, die auf der Hypothese $\rho = 0$ gleichmäßigen Fehler α haben.

5 Konfidenzbereiche

Sei $\mathcal{P}$ eine Familie von W-Maßen und $\kappa : \mathcal{P} \to \mathbb{R}$ ein Parameter.

Definition 5.1:
Eine Abbildung

$$K : X \to \mathcal{P}(Y),$$

derart dass $\{x \in X : y \in K(x)\} \in \mathcal{B}^1$ für alle $y \in \mathbb{R}$, heißt Konfidenzbereich (Vertrauensbereich) mit einem Konfidenzniveau von $1-\alpha, 0 < \alpha < 1$, falls

$$\inf_{P \in \mathcal{P}} P(x \in X : \kappa(P) \in K(x)) \geq 1 - \alpha.$$

Interpretation:

(*i*) $K(x)$ enthält den Parameter $\kappa(P)$, wobei x mit P „erzeugt“ wurde, mit Wahrscheinlichkeit größer als oder gleich $1-\alpha$.

(*ii*) Die Aussage $\kappa(P) \in K(x)$ ist mindestens mit W-Keit $1-\alpha$ richtig.

(*iii*) Falls Konfidenzniveau $1-\alpha$, so kann man $1-\alpha : \alpha$ wetten, dass $K(x)$ das richtige $\kappa(P)$ enthält für alle $P \in \mathcal{P}$.

Anwendungen:

(1) Für $\kappa(P) \in \mathbb{R}$ ist $K(x)$ üblicherweise ein Intervall (Konfidenzintervall).

(2) Fehlerschranken für Schätzer: Sei $\widehat{\kappa} : X \to \mathbb{R}$ ein Schätzer für $\kappa(P)$. Angenommen für eine Funktion $\Delta(P) > 0$ gilt

$$P(x \in X : |\widehat{\kappa}(x) - \kappa(P)| \leq \Delta(P)) \geq 1 - \alpha$$

für alle $P \in \mathcal{P}$. Falls für eine Schätzung $\widehat{\Delta}(x)$ von $\Delta(P)$

$$P(x \in X : |\widehat{\kappa}(x) - \kappa(P)| \leq \widehat{\Delta}(x)) \geq 1 - \alpha$$

für alle $P \in \mathcal{P}$ gilt, so ist

$$P(x \in X : \kappa(P) \in [\widehat{\kappa}(x) - \widehat{\Delta}(x), \widehat{\kappa}(x) + \widehat{\Delta}(x)]) \geq 1 - \alpha$$

für alle $P \in \mathcal{P}$. Also sind $x \mapsto K(x) = [\widehat{\kappa}(x) - \widehat{\Delta}(x), \widehat{\kappa}(x) + \widehat{\Delta}(x)]$ Konfidenzintervalle zum Niveau $1-\alpha$.

Beispiel 5.2:
Sei $\mathcal{P} = \{\mathcal{N}^n(\vartheta, \sigma^2), \vartheta \in \mathbb{R}, \sigma^2 > 0\}$ und $CR = \{x : \frac{\bar{x}_n - \vartheta_0}{S_n} > c_{n,\alpha}\}$ der Studentsche t-Test für $H : \{\vartheta \leq \vartheta_0, \sigma^2 > 0\}$ gegen $A : \{\vartheta > \vartheta_0, \sigma^2 > 0\}$. Nun gilt umgeschrieben

$$x \in CR \Leftrightarrow \vartheta_0 \in K(x_1, \ldots, x_n) \coloneqq (-\infty, \bar{x}_n - c_{n,\alpha} S_n).$$

mit $\mathcal{N}^n(\vartheta, \sigma^2)(\vartheta \in K(x_1, \ldots, x_n)) = 1 - \alpha$ für alle $\vartheta \in \mathbb{R}, \sigma^2 > 0$. Dies ist ein einseitiges Konfidenzintervall für ϑ.

Allgemeiner: Zu einem Konfidenzbereich gibt es eine Familie von kritischen Regionen für die Hypothese $\{P \in \mathcal{P} : \kappa(P) = r\}$ definiert durch

$$C(r) \coloneqq \{x \in X : r \notin K(x)\}.$$

Für $r \in \mathbb{R}$ gilt

$$x \in C(r) \Leftrightarrow r \notin K(x) \text{ und } \kappa(P) = r$$

und

$$P(C(r)) = P(x \in X : \kappa(P) \in K(x)) = 1 - (1 - \alpha) = \alpha$$

ist der Fehler erster Art.
Umgekehrt wie im obigen Beispiel: Für eine Familie von kritischen Regionen zum Niveau α ist $C(r), r \in \mathbb{R}$ mit $P(C(r)) \leq \alpha, P \in \mathcal{P}$ mit $r = \kappa(P)$ ist

$$K(x) \coloneqq \{r \in \mathbb{R} : x \notin C(r)\}$$

eine Familie von Konfidenzbereichen mit Niveau $P(x \in X : \kappa(P) \in K(x)) \geq 1 - \alpha$.

Definition 5.3:
$\widehat{K} : X \to \mathbb{R}$ heißt obere Konfidenzschranke für $\kappa : \mathcal{P} \to \mathbb{R}$ mit Konfidenzniveau $1 - \alpha$, falls

$$\inf_{P \in \mathcal{P}} P(x \in X : \kappa(P) \leq \widehat{K}(x)) = 1 - \alpha.$$

Entsprechend definiert man untere Konfidenzschranken.

Definition 5.4 (Güte für Konfidenzbereiche)**:**
Seien $P(\kappa(P) \leq \widehat{K}_i(x)) = 1 - \alpha, i = 0, 1$. Die K-Schranke $\widehat{K}_0$ heißt besser als die K-Schranke von $\widehat{K}_1$, falls

$$P(x \in X : t' \leq \widehat{K}_0(x) \leq t'') \geq P(x \in X : t' \leq \widehat{K}_1(x) \leq t'')$$

für alle $P \in \mathcal{P}$ und alle t', t'' mit $\kappa(P) \in [t', t'']$.
Optimalität heißt hier maximale Konzentration um $\kappa(P)$ unter P.

5.1 Exkurs: Konvergenzgeschwindigkeit im zentralen Grenzwertsatz

Satz (Satz von Berry-Esseen)**:**
Seien $X_1, \dots, X_n$ i.i.d. ZV mit $\mathbb{E}X_i = 0, \operatorname{Var} X_i = \sigma^2, \mathbb{E}\left|X_i\right|^3 < \infty$. Sei F_n die Verteilungsfunktion von $\frac{1}{\sqrt{n}\sigma}\sum_i X_i$ und Φ die Verteilungsfunktion der Standard-Normalverteilung. Dann gilt für alle $n \in \mathbb{N}$ und für alle $x \in \mathbb{R}$

$$|F_n(x) - \Phi(x)| \leq \frac{3\mathbb{E}\left|X_1\right|^3}{\sigma^3\sqrt{n}}.$$

Bemerkung:
1) Die Abschätzung ist gleichmäßig in $x \in \mathbb{R}$, d.h. F_n konvergiert gleichmäßig gegen Φ.
2) Da $\sigma^3 = \left(\mathbb{E}X_1^2\right)^{3/2} \leq \mathbb{E}\left|X_1\right|^3$ (Jensen), ist die Ungleichung trivial für $n \leq 10$.
3) Die Rate $\mathcal{O}\left(n^{-1/2}\right)$ ist optimal. Seien $X_i = \pm 1$ mit Wahrscheinlichkeit $1/2$. Dann folgt mit $S_n = \frac{1}{\sqrt{n}}\sum_{j=1}^n X_j$ und der Stirling-Formel

$$F_{2n}(0) = \frac{1}{2} - \frac{1}{2}P(S_{2n} = 0) = \frac{1}{2} - \frac{1}{2}\frac{\binom{2n}{n}}{2^{2n}} = \frac{1}{2} + \frac{1}{2\sqrt{\pi n}} + o\left(n^{-1/2}\right).$$

4) Der Faktor 3 im Satz ist nicht optimal. Die sogenannte Berry-Esseen-Konstante C erfüllt $0,4097 < C < 0,4748$.

Beweis. Ohne Einschränkung sei $\sigma^2 = 1$.
Idee: Benutze die charakteristische Funktion. Wir benötigen das Resultat, das die Differenz der Verteilungsfunktionen als Differenz der charakteristischen Funktionen ausdrückt. Dazu brauchen wir die „schöne“ charakteristische Funktion mit Hilfe der Glättung durch Faltung.
Sei $h_l(x) \coloneqq \frac{1-\cos(lx)}{\pi l x^2}$ für $l > 0$ die Dichte eines W-Maßes auf $\mathbb{R}$. h_l hat eine charakteristische Funktion $w_l(t) = \left(1 - \frac{|t|}{l}\right)^+$ für $|t| \leq l$. Sei H_l die dazugehörige Verteilungsfunktion. Eine Faltung mit H_l (bzw. h_l) hat dann eine charakteristische Funktion mit kompaktem Support. Zunächst folgende Lemmata. ∎

Lemma 1 (Smoothing Inequalities)**:**
Seien F, G Verteilungsfunktionen mit $\|G'\|_\infty \leq \lambda < \infty$. Definiere $\Delta(x) \coloneqq F(x) - G(x), \eta \coloneqq \sup|\Delta(x)|, \Delta_l \coloneqq \Delta * H_l$ und $\eta_l \coloneqq \sup|\Delta_l(x)|$. Dann gilt

$$\eta_l \geq \frac{\eta}{2} - \frac{12\lambda}{\pi l} \Leftrightarrow \eta \leq 2\eta_l + \frac{24\lambda}{\pi l}.$$

Beweis. Es gilt $\lim_{|x|\to\infty} \Delta(x) = 0$, denn G ist stetig und F eine Verteilungsfunktion, also existiert ein x_0 mit $\Delta(x_0) = \eta$ oder $\Delta(x_0^-) = -\eta$. Ohne Einschränkung sei $\Delta(x_0) = \eta$ (sonst betrachte Verteilungsfunktion von $-X_i$). Da $G'(x) \leq \lambda$ und F aufsteigend ist, gilt

$$\Delta(x_0 + s) \geq \eta - \lambda.$$

Setze $\delta = \frac{\eta}{2\lambda}$ und $t = x_0 + \delta$. Dann gilt

$$\Delta(t - x) \geq \begin{cases} \frac{\eta}{2} + \lambda x & |x| \leq \delta \\ -\eta & \text{sonst} \end{cases}.$$

Ferner gilt wegen $|\cos(lx)| \leq 1$

$$2\int_\delta^\infty h_l(x)dx \leq 2\int_\delta^\infty \frac{2}{\pi l x^2}dx = \frac{4}{\pi l \delta}$$

und

$$\begin{aligned} \Delta_l(t) &= \int_{-\delta}^{\delta} \Delta(t-x)h_l(x)dx + \int_{\mathbb{R}\setminus[-\delta,\delta]} \Delta(t-x)h_l(x)dx \\ &\geq \int_{-\delta}^{\delta} \frac{\eta}{2}h_l(x)dx + \lambda \underbrace{\int_{-\delta}^{\delta} x h_l(x)dx}_{=0} - \eta \int_{\mathbb{R}\setminus[-\delta,\delta]} h_l(x)dx \\ &\geq \frac{\eta}{2}\left(1 - \frac{4}{\pi l \delta}\right) - \eta\frac{4}{\pi l\delta} = \frac{\eta}{2} - \frac{12\lambda}{\pi l}. \end{aligned}$$

■

Lemma 2:
Seien K_1, K_2 Verteilungsfunktionen mit Mittelwert 0 und mit integrierbaren charakteristischen Funktionen κ_i. Dann gilt

$$K_1(x) - K_2(x) = -\frac{1}{2\pi}\int \exp(itx)\,\frac{\kappa_1(t) - \kappa_2(t)}{it}dt.$$

Beweis. Da κ_i integrierbar sind, haben K_i die Dichten k_i mit

$$k_i(y) = \frac{1}{2\pi}\int \exp(ity)\kappa_i(t)dt.$$

Sei $\Delta K = K_1 - K_2$. Dann gilt mit dem Satz von Fubini

$$\begin{aligned} \Delta K(x) - \Delta K(a) &= \frac{1}{2\pi}\int_a^x \int e^{ity}(\kappa_1(t) - \kappa_2(t))dtdy \\ &= \frac{1}{2\pi}\int \left(e^{ita} - e^{itx}\right)\frac{\kappa_1(t) - \kappa_2(t)}{it}dt. \end{aligned}$$

Da $\int x dK_i(x) = 0$, gilt $\lim_{t\to 0} \frac{1-\kappa_i(t)}{t} = 0$ (Übung) und somit ist $\frac{\kappa_1(t)-\kappa_2(t)}{it}$ beschränkt und stetig und daher integrierbar. Für $a \to -\infty$ folgt die Behauptung mit dem Riemann-Lebesgue-Lemma. ∎

Fortsetzung vom Beweis des Satzes von Berry-Esseen. Seien nun φ_F und φ_G die char. Funktionen von F und G. Anwendung von Lemma 2 auf $F_l := F * H_l$ und $G_l := G * H_l$ ergibt

$$\begin{aligned} |F_l(x) - G_l(x)| &\leq \frac{1}{2\pi} \int |\varphi_F(t) w_l(t) - \varphi_G(t) w_l(t)| \, \frac{1}{|t|} dt \\ &\leq \frac{1}{2\pi} \int_{-l}^{l} |\varphi_F(t) - \varphi_G(t)| \, \frac{1}{|t|} dt. \end{aligned}$$

Mit Lemma 1 folgt

$$|F(x) - G(x)| \leq \frac{1}{\pi} \int_{-l}^{l} \frac{|\varphi_F(t) - \varphi_G(t)|}{|t|} dt + \frac{24\lambda}{\pi l}$$

wobei $\lambda = \sup_x G'(x)$. Mit $F = F_1$ und $G = \Phi$ ergibt sich mit φ als charakteristische Funktionen von X_1 und ψ als char. Funktion der Standard-Normalverteilung

$$|F(x) - \Phi(x)| \leq \frac{1}{\pi} \int_{-l}^{l} \left| \varphi^n \left(\frac{t}{\sqrt{n}} \right) - \psi(t) \right| \frac{1}{|t|} dt + \frac{24\lambda}{\pi l}. \qquad (*)$$

Es gilt

$$\lambda = \sup_x \Phi'(x) = \Phi'(0) = \frac{1}{\sqrt{2\pi}} \approx 0.39 < \frac{2}{5}. \qquad (a)$$

Für den ersten Term der rechten Seite von $(*)$ benutzen wir für $|\alpha|, |\beta| < \gamma$

$$|\alpha^n - \beta^n| \leq \sum_{m=0}^{n} \left| \alpha^{n+m} \beta^m - \alpha^{n-m-1} \beta^{m+1} \right| \leq n \, |\alpha - \beta| \, \gamma^{n-1}. \qquad (b)$$

Aus der Ungleichung

$$\left| e^{it} - 1 - it + \frac{t^2}{2} \right| \leq \frac{|t|^3}{6}$$

folgt mit der Ungleichung von Jensen

$$\left| \varphi(t) - 1 + \frac{t^2}{2} \right| \leq \frac{\rho \, |t|^3}{6} \text{ mit } \rho := \mathbb{E} \, |X_1|^3, \qquad (c)$$

also für $t^2 \leq 2$

$$|\varphi(t)| \leq 1 - \frac{t^2}{2} + \frac{\rho \, |t|^3}{6}. \qquad (d)$$

Wähle nun $l \coloneqq \frac{4\sqrt{n}}{3\rho}$. Falls $|t| \le l$, dann gilt wegen (d) und $\frac{\rho|t|}{\sqrt{n}} \le \frac{4}{3}$

$$\left|\varphi\left(\frac{t}{\sqrt{n}}\right)\right| \le 1 - \frac{t^2}{2n} + \rho\frac{|t|^3}{6n^{3/2}} \le 1 - \frac{5t^2}{18n} \le \exp\left(\frac{-5t^2}{18n}\right),$$

da $1 - x \le e^{-x}$. Wende nun (b) an mit

$$\alpha \coloneqq \varphi\left(\frac{t}{\sqrt{n}}\right), \beta \coloneqq \exp\left(-\frac{t^2}{2n}\right), \gamma = \exp\left(-\frac{5t^2}{18n}\right) \qquad (e)$$

Da wir $n \ge 10$ annehmen können, gilt $\gamma^{n-1} \le \exp\left(-\frac{t^2}{4}\right)$. Für die rechte Seite von (b) gilt

$$n\,|\alpha - \beta| \le n\left|\varphi\left(\frac{t}{\sqrt{n}}\right) - 1 + \frac{t^2}{2n}\right| + n\left|1 - \frac{t^2}{2n} - \exp\left(-\frac{t^2}{2n}\right)\right|. \qquad (f)$$

Um dies abzuschätzen, benutze für $0 < x < 1$

$$n\left|\varphi\left(\frac{t}{\sqrt{n}}\right) - 1 + \frac{t^2}{2n}\right| \le \frac{\rho\,|t|^3}{6\sqrt{n}}$$

und

$$\left|e^{-x} - (1-x)\right| \le \left|-x^2/2! + x^3/3! - \ldots\right| \le \frac{x^2}{2}.$$

Mit $x \coloneqq \frac{t^2}{2n}$ folgt für $|t| \le l \le \sqrt{2n}$ unter Beachtung von $\rho \ge 1$

$$n\left|1 - \frac{t^2}{2n} - \exp\left(-\frac{t^2}{2n}\right)\right| \le \frac{t^4}{8n}.$$

Mit $(b) - (f)$ gilt

$$\begin{aligned}\frac{1}{|t|}\left|\varphi^n\left(\frac{t}{\sqrt{n}}\right) - \exp\left(-\frac{t^2}{2}\right)\right| &\le \exp\left(-\frac{t^2}{4}\right)\left(\frac{8t^2}{6\sqrt{n}} + \frac{|t|^3}{8n}\right) \\ &\le \frac{1}{l}\exp\left(-\frac{t^2}{4}\right)\left(\frac{2t^2}{9} + \frac{|t|^3}{18}\right),\end{aligned}$$

da $\frac{\rho}{\sqrt{n}} \le \frac{4}{3t}$ und $\frac{1}{n}\frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}} \le \frac{4}{3l}\frac{1}{3}$. Es folgt mit Lemma 2

$$\pi l\,|F_1(x) - \Phi(x)| \le \int \exp\left(-t^2/4\right)\left(\frac{2t^2}{9} + \frac{|t|^3}{18}\right)dt.$$

Da $l = \frac{4\sqrt{n}}{3\rho}$ gilt also $|F_1(x) - \Phi(x)| \le \frac{C\rho}{\sqrt{n}} + 24\frac{2}{5}$. Um l zu bestimmen, rechne

$$2\int_0^\infty x^3 \exp\left(-\frac{x^2}{4}\right)dx = 2\int_0^\infty \exp\left(-\frac{x^2}{4}\right)dx = -16\exp\left(-x^2/4\right)\,|_0^\infty = 16.$$

Damit gilt

$$|F_1(x) - \Phi(x)| \leq \frac{1}{\pi}\frac{3}{4}\left(\frac{4}{9}\cdot\sqrt{4\pi} + \frac{16}{18} + 9,6\right)\frac{\rho}{\sqrt{n}} < 3\frac{\rho}{\sqrt{n}}.$$

■

Satz 5.5:
Sei $\{P_{\vartheta,\eta} : \vartheta \in (a,b), \eta \in N\}$ *eine exponentielle Familie wie in Satz 4.20 mit*

(*i*) $a_0(\vartheta_\eta) = a_0(\vartheta)$, *d.h.* a_0 *ist unabhängig von* η,

(*ii*) a_0 *ist strikt monoton wachsend und*

(*iii*) $\{(a_1(\vartheta,\eta),\dots,a_k(\vartheta,\eta)) : \eta \in N\}$ *hat nichtleeres Inneres.*

Angenommen eine obere Konfidenzschranke $\widehat{\vartheta}_\alpha : \mathbb{R}\times\mathbb{R}^k \to (a,b)$ *hat die Eigenschaften*

(*i*) $t_0 \mapsto \widehat{\vartheta}_\alpha(t_0,t)$ *ist strikt monoton wachsend für alle* $t \in \mathbb{R}^k$.

(*ii*) $P^n_{\vartheta,\eta}\left(\widehat{\vartheta}_\alpha\left(\sum_j T_0(x_j), \sum_j T(x_j)\right) \geq \vartheta\right) = 1-\alpha$ *für alle* $\eta \in N$ *und* $\vartheta \in (a,b)$.

Dann ist $\widehat{\vartheta}_\alpha$ *optimal konzentriert um* ϑ. *Hierbei ist* $\widehat{\vartheta}_\alpha(t_0,t)$ *gegeben als*

$$\widehat{\vartheta}_\alpha(t_0,t) := \sup\{\vartheta \in (a,b) : P_{\vartheta,\eta}\left(T_0(x) \leq t_0 \mid T = t\right) \geq 1-\alpha\}.$$

Dabei ist dies unabhängig von $\eta \in \mathbb{R}^k$, *falls* $P_{\vartheta,\cdot}(x : T_0(x) \leq t_0 \mid T = t)$ *stetig in* t_0 *für alle* $t \in \mathbb{R}^k$ *ist.*

Beweis. [6], Satz 5.5.9 und 5.5.15, S. 177ff. ∎

Beispiel 5.6:
Gesucht ist ein Konfidenzintervall für σ^2 in $\mathcal{P} = \{\mathcal{N}^n(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$. Dies ist nach Satz 5.5 mit $x \mapsto c(\mu,\sigma^2)\exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right)$ mit $a_0(\sigma^2) = -\frac{1}{2\sigma^2}$. Sei k_α derart, dass $\chi^2_{n-1}(k_\alpha) = 1-\alpha$, wobei χ^2 die Verteilungsfunktion von $\sum_{j=1}^{n-1} Y_j^2, Y_j \sim \mathcal{N}(0,1)$ bezeichnet. Dann gilt

$$\mathcal{N}^n(\mu,\sigma^2)\left((x_1,\dots,x_n) : \sigma^2 \leq \frac{\sum_j (x_j - \bar{x}_n)^2}{k_\alpha}\right) = 1-\alpha.$$

Also ist $\frac{(n-1)S_n^2}{k_{n-1,\alpha}}$ eine obere K-Schranke für σ^2, und da $(n-1)S_n^2 = \sum_j x_j^2 - \frac{1}{n}\left(\sum_j x_j\right)^2 = T_0(x) - \frac{1}{n}T_1(x)^2$ strikt monoton in $T_0(x)$ ist, ist diese K-Schranke maximal konzentriert um σ^2.

5.2 Konstruktion von Konfidenzbereichen mittels Stichprobenverfahren

Beispiel (Student-Statistik)**:**
Problem: Gegeben sei eine Familie $\mathcal{P}$ mit $\mathbb{E}_P x < \infty$ und $\kappa(P) = \int x dP$ und sei $\sup_{P\in\mathcal{P}} \int x^4 dP < \infty$.
Finde eine untere Konfidenzschranke zum K-Niveau $1-\alpha$, d.h. finde ein $c_{n,\alpha} \in \mathbb{R}$ mit

$$P^n\left(\bar{x}_n - c_{n,\alpha} S_n \leq \kappa(P)\right) = P^n\left(\frac{\bar{x}_n - \kappa(P)}{S_n} \leq c_{n,\alpha}\right) = 1-\alpha. \qquad (*)$$

Gesucht wird also ein Funktional

$$c_{\alpha,n} : \mathcal{P} \to \mathbb{R}\,.$$

Betrachte

$$F_{P,n}(a) \coloneqq P^n\left(\frac{\bar{x}_n - \kappa(P)}{S_n} \leq a\right).$$

Diese Funktion erfüllt

(*i*) $\lim_{a\to\pm\infty} F_{P,n}(a) = 1$ bzw. 0 als Verteilungsfunktion eines W-Maßes.

(*ii*) $a \mapsto F_{P,n}(a)$ ist nichtfallend und rechtsseitig stetig.

Dann ist $\tau_{\alpha,n}(P) = \inf\left(a : F_{P,n}(a) \geq 1-\alpha\right)$ das gesuchte Funktional, falls $F_{P,n}(a)$ stetig ist.
Idee: Plugin-Schätzung
Schätze P in $\tau_\alpha(P)$ durch $\widehat{P}_n = \frac{1}{n}\sum_j \delta_{x_j}$, d.h. durch das empirische Maß, und $\tau_{\alpha,n}(P)$ durch $\tau_{\alpha,n}\left(\widehat{P}_n\right)$, vorausgesetzt, dass $P \mapsto \tau_\alpha(P)$ stetig in geeigneter Topologie auf $\mathcal{P}$ ist.

Hierzu versucht man

(*a*) $F_{P,n}(a)$ durch $F_{\widehat{P}_n,n}(a)$ und

(*b*) $\tau_\alpha(P)$ durch $\widehat{\tau}_\alpha \coloneqq \inf\{a : F_{\widehat{P}_n,n}(a) \geq 1-a\}$ zu schätzen.

Beispiel:
Schritt A: Definiere zunächst

$$F_{\widehat{P}_n,n}(a) = \widehat{P}_n^n\left(\frac{\bar{x}_n^* - \kappa\left(\widehat{P}_n\right)}{S_n^*} \leq a\right),$$

wobei $\widehat{P}_n \in \mathcal{P}$ sein sollte, damit $\kappa\left(\widehat{P}_n\right)$ existiert und $\bar{x}_n^* = \frac{1}{n}\left(x_1^* + \ldots + x_n^*\right)$, wo $x_j^*, j = 1, \ldots, n$ Verteilung $\widehat{P}_n$ haben gegeben $(x_1, \ldots, x_n)$ und stochastisch unabhängig sind mit Standard-Abweichung S_n^*.
Nun ist $\widehat{P}_n$ eine bekannte diskrete Verteilung.
Schritt B: Wie berechnet man $F_{\widehat{P}_n,n}(a)$? Mit dem Gesetz der großen Zahlen:

$$F_{\widehat{P}_n,n}(a) = \widehat{P}_n^n\left(\frac{\overline{x}_n^* - \kappa\left(\widehat{P}_n\right)}{S_n^*} \leq a\right) = \lim_{N\to\infty} \underbrace{\frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\left(\frac{\bar{x}_{n,j}^* - \kappa\left(\widehat{P}_n\right)}{S_{n,j}^*} \leq a\right)}_{h_N(x^*,a)},$$

wo $x_j^* = (x_{1,j}^*, \ldots, x_{n,j}^*), j = 1, \ldots, N$ N unabhängige, identisch verteilte N-Tupel mit Verteilung $\widehat{P}_n$ sind und setze $\bar{x}_{n,j}^* = \frac{1}{n}\sum_{l=1}^n x_{l,j}^*$ sowie $S_{n,l}^* = \frac{1}{n-1}\sum_{l=1}^n \left(x_{l,j}^* - \bar{x}_{n,j}^*\right)^2$.
Für endliches N wird $F_{\widehat{P}_n,n}(a)$ durch die relative Häufigkeiten $h_N(x^*, a)$ approximiert mit einem stochastische Fehler $\mathcal{O}_{\widehat{P}_n^\infty}(N^{-1/2})$.
Dieses Verfahren nennt man Stichprobenverfahren oder Resampling-Verfahren oder Bootstrap-Verfahren.
Mit Hilfe von Tschebycheff folgt auch

$$\widehat{P}_n^{nN}\left(\sqrt{N}\left|h_N(x^*, a) - F_{\widehat{P}_n,n}(a)\right| > \varepsilon\right) \leq \frac{c\left(\widehat{P}_n\right)}{\varepsilon^2}.$$

Was ist, wenn $\widehat{P}_n \notin \mathcal{P}$ bzw. $\tau\left(\widehat{P}_n\right)$ nicht definiert ist? Sei τ stetig bezüglich einer Distanz d auf den W-Maßen mit $\mathcal{P} \ni \widehat{Q}_n = \arg\min\{d(\widehat{P}_n, P) : P \in \mathcal{P}\}$, sodass $\lim_n \kappa\left(\widehat{Q}_n\right) = \kappa(P)$ $P^\infty-$f.s.

Definition 5.7:
Sei $\mathcal{P}$ eine Familie von W-Maßen auf $(\mathcal{X}, \mathcal{B})$, $T_n : \mathcal{X}^n \times \mathcal{P} \to \mathbb{R}$ eine Folge von Statistiken. Ein (nichtparametrischer) Bootstrap-Schätzer von $F_{P,n}(a) = P^n\left(T_n(x_1, \ldots, x_n, P) \leq a\right)$ ist gegeben durch zwei Folgen von W-Maßen $(\widehat{P}_n, \widehat{Q}_n)$ für jedes Tupel $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ mit

(*i*) $\widehat{Q}_n \in \mathcal{P}$, sodass mit $x^* = (x_1^*, \ldots, x_n^*) \in \mathcal{X}^n$

(*ii*) $T_n(x_1^*, \ldots, x_n^*, \widehat{Q}_n)$ als Funktion von (x^*, x) messbar ist,

(*iii*) $\lim_{n\to\infty}\left(P^n(T_n(x, P) \leq a) - \widehat{P}_n^n(T_n(x, \widehat{Q}_n) \leq a)\right) = 0$ $P^\infty-$.f.s. gilt.

Dabei wird $\widehat{P}_n^n\left(T_n\left(x^*, \widehat{Q}_n\right) \leq a\right)$ mittels der Monte-Carlo-Approximation durch relative Häufigkeiten h_N für die Werte $\mathbb{1}\left(T_n(x_r, \widehat{Q}_n) \leq a\right), r = 1, \ldots, N$ approximiert.

Beispiel:
Für den Student-Test schätze $F_{P,n}(a) = P^n\left(\frac{\bar{x}_n - \kappa(P)}{S_n} \le a\right)$ wo $\kappa(P)$ das Mittelwertfunktional ist, durch $F_{\widehat{P}_n,n}(a) \coloneqq \widehat{P}_n^n\left(\frac{\bar{x}_n^* - \kappa\left(\widehat{P}_n\right)}{S_n^*} \le a\right)$ auf Basis von $x_1, \ldots, x_n$, wo $x_1^*, \ldots, x_n^*$ i.i.d. nach $\widehat{P}_n$ verteilt sind.
Dann ist eine untere Konfidenzschranke für $\kappa(P)$ gegeben durch

$$\bar{x}_n - F_{\hat{P}_n,n}^{-1}(1-\alpha)/S_n^* \le \kappa(P)$$

zum approximativen Niveau $1-\alpha$ für $n \to \infty$.
Problem: Ziehen von $x^* = (x_1^*, \ldots, x_n^*)$ aus $\widehat{P}_n^n$ ist ein Ziehen mit Zurücklegen aus $\{x_1, \ldots, x_n\}$. Dann hat $Nh_N(a)$ eine Binomialverteilung mit $B\left(N, F_{\hat{P}_n,n}(a)\right)$, wobei $F_{\hat{P}_n,n}(a) = \widehat{P}_n^n\left(T_n(x^*, \widehat{Q}_n) \le a\right)$. Dann gilt nach dem Satz von Berry-Esseen

$$\sup_{a\in\mathbb{R}}\left|\widehat{P}_n^n\left(\sqrt{N}\frac{h_N(a) - F_{\hat{P}_n,n}(a)}{\sqrt{F_{\hat{P}_n,n}(a)(1-F_{\hat{P}_n,n}(a))}} \le a\right) - \Phi(a)\right| \le c(a)N^{-1/2} = \frac{\beta_3}{\mathrm{Var}^{3/2}}.$$

Satz 5.8:
Sei $\mathcal{P} = \{$*W-Maß* P *mit* $\int x^4 P(dx) < M\}$ *und* $\limsup_{|t|\to\infty}|\int \exp(itx)\,P(dx)| < \rho < 1$ *mit* ρ, M *fest. Dies schließt z.B. aus, dass* $P(\mathbb{Z}) = 1$ *oder andere Gitterverteilungen in der Familie vorkommen. Dann gilt*

$$F_{P,n}(a) = P^n(\bar{x}_n - aS_n \le \kappa(P)) = \widehat{F}_n(a) + \mathcal{O}_P\left(n^{-1}\right),$$

d.h.

$$\limsup_n P^n\left(n\left|\widehat{F}_n(a) - F_{P,n}(a)\right| > c\right) \le \varepsilon(c), \tag{5.9}$$

wobei $\varepsilon(c) \downarrow 0$ *für* $c \to \infty$.

Beweis. Peter Hall, Chibisov, Berthus, Götze, van Zwet (1997). ∎

Beispiel 5.10:
Seien $X_1, \ldots, X_n \sim \mathcal{U}(0,1)$ i.i.d. Definiere

$$T_n(x_1, \ldots, x_n, P) = n(1 - \max_i x_i) = n\left|\max\{y : \mathbb{1}_{[0,1]}(y) > 0\} - \max\{y : \widehat{F}_n(y) > 0\}\right|.$$

Dann gilt

$$\mathcal{U}^n(T_n \ge a) = \mathcal{U}^n\left(\max_j X_j \le 1 - \frac{a}{n}\right) = \mathcal{U}\left(X_1 \le 1 - \frac{a}{n}\right)^n = \left(1 - \frac{a}{n}\right)^n \to e^{-a}.$$

Diese Verteilung sollte durch $\max \mathrm{supp}\left(\hat{P}_n\right)$ geschätzt werden:

$$\widehat{P}_n^m(T_m^* \le a),$$

wobei

$$T_m^* = m \left| \max_{j=1,\ldots,m} X_j^* - \max_{j=1,\ldots,n} X_j \right|.$$

Es gilt z.B.

$$\mathcal{U}^n (T_n = 0) = \mathcal{U}^n (\exists i : X_i = 1) \leq n\mathcal{U} (X_1 = 1) = 0.$$

Für $m = m(n) \uparrow \infty$ ist

$$\begin{aligned} \widehat{P}_n^m (T_m^* = 0) &= \widehat{P}_n^m \left(\exists j : X_j^* = \max (X_1, \ldots, X_n) \right) = 1 - \widehat{P}_n^m \left(X_j^* \neq X_{n:n} \right) \\ &= 1 - \widehat{P}_n (X_1 \neq X_{n:n}) = 1 - \left(1 - \frac{1}{n} \right)^m \\ &\to \begin{cases} 0 & \text{falls } \frac{m(n)}{n} \to 0 \\ 1 - e^{-1} & \text{falls } m(n) = n \end{cases} \end{aligned}$$

Allgemeiner gilt $\widehat{P}_n^{m(n)} (T_m^* \geq a) \to e^{-a}$ nur, wenn $\frac{m(n)}{n} \to 0$.

Beispiel 5.11 (Nicht-parametrische Schätzer)**:**
Sei

$$\mathcal{P}_M = \{ f : [a,b] \to \mathbb{R}_+, \int f d\lambda = 1, f \in \mathcal{C}^2, \|f\|_\infty \leq M, \|f''\|_\infty \leq M \}.$$

eine Familie von W-Maßen mit Dichten f. Betrachte den Kernschätzer

$$\hat{f}_n^x(s) = \frac{1}{n} \sum_{j=1}^n \frac{1}{h_n} K \left(\frac{s - x_j}{h_n} \right) \xrightarrow{L^2} f$$

und die Statistik $T_n(x_1, \ldots, x_n, P) = \sup_{s \in [a,b]} \left| \hat{f}_n^x(s) - f(s) \right|$, wobei $P(dx) = f\lambda(dx)$. Finde ein $c_{n,\alpha}$ mit

$$P^n (T_n \leq c_{n,\alpha}) = 1 - \alpha \text{ für alle } P \in \mathcal{P}_M .$$

Nutze die Bootstrap-Methode: Da f bzw. P unbekannt ist, benutze stattdessen zu einer gegebenen Stichprobe $(x_1, \ldots, x_n)$

$$\widehat{Q}_n = \hat{f}_n^x n d\lambda$$

sodass $\widehat{Q}_n \in \mathcal{P}_M$. Hier sollte man für die Bootstrap-Replikationen nur $m(n) < n$ Stichproben resamplen mit $\frac{m(n)}{n} \to 0$ für $n \to \infty$ wählen, um $F_P(c) = P^n (T_n(\cdot, P) \leq c)$ zu schätzen. Dies geschieht mit Hilfe von

$$\widehat{F}_{n,m}(c) \coloneqq \widehat{P}_n^{m(n)} \left(\sup_{s \in [a,b]} \left| \hat{f}_{m(n)}^{x^*}(s) - \hat{f}_n^x(s) \right| \leq c \right) \overset{P^{\mathbb{N}}}{\to} F_P(c)$$

für $n \to \infty$.

Definition 5.12:
Sei $\mathcal{P} = \{P_\vartheta, \vartheta \in \Theta\}$ eine parametrische Familie und $\widehat{\vartheta}_n : \mathcal{X}^n \to \Theta$ ein Schätzer für ϑ. Einseitige Konfidenzintervalle werden mittels $T_n(x_1, \ldots, x_n, \vartheta) \leq a$ konstruiert. Dann ist die parametrische Bootstrap-Approximation von

$$F_\vartheta(a) = P_\vartheta^n \left(T_n(\cdot, \vartheta) \leq a\right)$$

durch

$$P^{m(n)}_{\widehat{\vartheta}_n(x_1,\ldots,x_n)} \left(T_{m(n)} \left(x_1^*, \ldots, x_{m(n)}^*, \widehat{\vartheta}_n(x_1, \ldots, x_n)\right) \leq a\right)$$

gegeben mit $\frac{m(n)}{n} \to 0$ für $n \to \infty$ und $\widehat{\vartheta}_n \to \vartheta$ P^n-f.s.

Das Verfahren „m aus n“ lässt sich wie folgt beschreiben:
Das Ziel ist die Beschreibung der Verteilung von $T_m(x_1^*, \ldots, x_m^*, \widehat{P}_n)$. Seien dazu

1) $h : \mathbb{R} \to \mathbb{R}$ eine messbare, beschränkte Funktion, z.B. $h(x) = \mathbb{1}_{\{x \leq a\}}$ und

2) $(T_n)_n$ eine Folge von Statistiken, welche symmetrisch in den Beobachtungen ist.

Dann sind

$$\tau_n(P) \coloneqq \int h(T_n(x_1, \ldots, x_n, P)) P^n(dx_1, \ldots, dx_n)$$

Funktionalfolgen von P. Dann wollen wir $\tau_n(P)$ für $P \in \mathcal{P}$ schätzen. Hierzu sind verschiedene Schätzerklassen möglich.
Verfahren 1: $B_{m,n}$: Schätze $\tau_n(P)$ durch $\tau_m \left(\widehat{P}_n\right)$ mit $m \leq n$. Für $m = n$ heißt dies Efrons Bootstrap. Schätze allgemeiner $\tau_n(P)$ durch $\tau_m \left(\widehat{P}_n\right)$, wobei auch hier $\frac{m(n)}{n} \to 0$ gelten soll.
Verfahren 2: $E_{m,n}$: Ziehe $m < n$ mal ohne Zurücklegen aus der Stichprobe. Setze dabei

$$\tilde{\tau}_m \left(\hat{P}_n, P\right) \coloneqq \int h\left(T_m(\cdot, P)\right) d\widehat{P}_n^{(m)},$$

wo

$$\int f d\widehat{P}_n^{(m)} = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \ldots < i_m \leq n} f(x_{i_1}, \ldots, x_{i_m})$$

und $(x_1, \ldots, x_n)$ die gegebene Stichprobe aus $\mathcal{P}^n$ ist.

Satz 5.13 (Asymptotische Konsistenz für $E_{m,n}$)**:**
Sei $(x_1, \ldots, x_n)$ eine gegebene Stichprobe aus $P^n, P \in \mathcal{P}$ und h beschränkt und messbar. Dann gilt

(*i*) $\tilde{\tau}_m\left(\widehat{P}_n, P\right) = \tau_m(P) + \mathcal{O}_P\left(\left(\frac{m}{n}\right)^{1/2}\right)$, *d.h. es gilt*

$$\limsup_{m\to\infty, \frac{m}{n}\to 0} P^n\left(\left|\tilde{\tau}_m\left(\widehat{P}_n, P\right) - \tau_m(P)\right| > \kappa\left(\frac{m}{n}\right)^{1/2}\right) \leq \varepsilon(\kappa),$$

mit $\lim_{\kappa\to\infty}\varepsilon(\kappa) = 0$.

(*ii*) *Ist* h *stetig und* $T_m\left(x_1, \ldots, x_m, P\right) = T_m\left(x_1, \ldots, x_m, \widehat{P}_n\right) + o_P(1)$ *für* $m, n \to \infty, \frac{m}{n} \to 0$, *d.h.*

$$\limsup_{m,n\to\infty, \frac{m}{n}\to 0} P^n\left(\left|T_m\left(x_1, \ldots, x_m, P\right)\right) - T_m\left(x_1, \ldots, x_m, \widehat{P}_n\right)\right| > \varepsilon\right) = 0.$$

für alle $\varepsilon > 0$. *Dann gilt*

$$\lim_{m\to\infty, \frac{m}{n}\to 0} P^n\left(\left|\tilde{\tau}_m(\widehat{P}_n, \widehat{P}_n) - \tau_m(P)\right| > \varepsilon\right) = 0$$

für alle $\varepsilon > 0$.

Beweis.
(*i*) Setze

$$\tilde{\tau}_m\left(\widehat{P}_n, P\right) = \frac{1}{\binom{n}{m}} \sum_{1\leq i_1 < \ldots < i_m \leq n} h \circ T_m\left(x_{i_1}, \ldots, x_{i_m}, P\right)$$

Dann ist

$$\int \tilde{\tau}_m\left(\widehat{P}_n, P\right) dP^m = \int h \circ T_m(x_1, \ldots, x_m, P) dP^m = \tau_m(P) \qquad (a)$$

da die Statistiken symmetrisch in den Beobachtungen sind.
Setze $\tilde{\psi}(x_I, P) = \psi(x_I, P) - \tau_m(P)$ für $\psi = h \circ T$ mit $|I| = m, I \subset \{1, \ldots, n\}$. Zerlege $N \coloneqq \left[\frac{n}{m}\right]$ und $\{1, \ldots, n\} = I_1 \cup I_2 \cup \cdots \cup I_N \cup K$, wobei die Vereinigungen disjunkt seien und $|I_j| = m$. Sei $\pi : \{1, \ldots, m\} \to \{1, \ldots, n\}$ eine Permutation. Dann gilt für ein beliebiges $j \in \{1, \ldots, N\}$

$$\{J \subset \{1, \ldots, n\}, |J| = m\} = \{\pi(I_j) : \pi \in S_n\}.$$

Definiere

$$\begin{aligned}
\tilde{\Delta} &\coloneqq \frac{1}{\binom{n}{m}} \sum_{J\subset\{1,\ldots,n\}, |J|=m} \tilde{\psi}(x_J, P) \\
&= \frac{1}{n!} \sum_{\pi\in S_n} \left(\frac{\tilde{\psi}(x_{\pi(I_1)}, P) + \ldots + \tilde{\psi}(x_{\pi(I_N)}, P)}{n}\right) \\
&=: \frac{1}{n!} \sum_{\pi\in S_n} \Delta_{\pi,N}
\end{aligned}$$

weil $\#\{\pi \in S_n : \pi(I_j) = J\} = (n-m)!m!$ für alle $J \subset \{1,\dots,n\}, |J| = m$ für alle $j = 1,\dots,n$. Also ist

$$\int \left|\tilde{\Delta}\right| dP^n \;\leq\; \frac{1}{n!}\sum_{\pi\in S_n}\int |\Delta_{\pi,N}|\, dP^n \leq \frac{1}{n!}\sum_{\pi\in S_n}\left(\int (\Delta_{\pi,N})^2 dP^n\right)^{1/2}$$

Da

$$\int \Delta_{\pi,N}^2 dP^n = \frac{N}{N^2}\int \tilde{\psi}(x_{I_1},P)^2 dP^m \leq \left(\frac{1}{N}\right)\|h\|_\infty^2$$

d.h. $\int \left|\tilde{\Delta}\right| dP^n \leq \frac{1}{n!} n! \frac{1}{\sqrt{N}} \|h\|_\infty$. Für $N = \left[\frac{n}{m}\right]$ geht dies gegen 0, woraus mit Tschebycheff folgt

$$P\left(\left|\tilde{\Delta}\right| > K\right) \leq \frac{2\|h\|_\infty}{K\left(\frac{n}{m}\right)^{\frac{1}{2}}}$$

für $n, m \to \infty$ für alle $K > 0$.

(ii) Es gilt

$$\mathbb{E}_P\left|\tilde{\tau}_m\left(\widehat{P}_n,\widehat{P}_n\right) - \tau_m(P)\right| \leq \mathbb{E}_P\left|\psi(x_1,\dots,x_m,\widehat{P}_n) - \psi(x_1,\dots,x_m,P)\right|.$$

Aus der Voraussetzung und der Stetigkeit von h folgt die Behauptung, d.h.

$$\left|\psi(\cdot,\widehat{P}_n) - \psi(\cdot,P)\right| \to 0$$

P^n-stochastisch und da ψ beschränkt ist.

■

Bemerkung 5.14:
Falls die Funktionalfolge für alle $P \in \mathcal{P}$ konvergiert, d.h. $\lim_n \tau_n(P) =: \tau(P)$, gilt

$$\lim_{m,n\to\infty} |\tau_m(P) - \tau_n(P)| = 0$$

und damit nach Satz 5.13 (ii)

$$\lim_{m,n\to\infty, mn^{-1}\to 0}\left|\tilde{\tau}_m\left(\widehat{P}_n,\widehat{P}_n\right) - \tau_n(P)\right)\right| = 0$$

$P^{\mathbb{N}}$-stochastisch, d.h. $\tilde{\tau}_m\left(\widehat{P}_n,\widehat{P}_n\right)$ ist konsistent für $\tau_n(P)$ und $\tau(P)$.
Z. B. falls $|\tau_n(P) - \tau(P)| \leq cn^{-1/2}$ (z.B. falls Berry-Esseen anwendbar ist), gilt

$$|\tau_n(P) - \tau_m(P)| \leq c_2\left(n^{-1/2} + m^{-1/2}.\right)$$

Daher wähle $m = \sqrt{n}$. Der Gesamtfehler ist also

$$\left|\tilde{\tau}_m\left(\widehat{P}_n,\widehat{P}_n\right) - \tau_n(P)\right| = \mathcal{O}_P\left(n^{-1/4}\right)$$

falls $m = \sqrt{n}$ gewählt wird.

$B_{m,n}-Verfahren$: (m,n)-Bootstrap mit Zurücklegen. Setze $X_j^{(i)} \coloneqq \underbrace{(X_j,\ldots,X_j)}_{i \text{ mal}}$ und für $i=(i_1,\ldots,i_r)$ sei

$$\psi_i = \frac{1}{r!} \sum_{1\le j_1\neq\ldots\neq j_r\le n} h\left(T\left(x_{j_1}^{(i_1)},\ldots,X_{j_r}^{(i_r)},P\right)\right)$$

Das Problem ist, wenn P eine Dichte bezüglich des Lebesgue-Maßes hat und $\int h(x,y)^2P^2(dx,dy)<\infty$ gilt, dass trotzdem $\int h(x,x)^2P(dx)=\infty$.
Für $i\in\Lambda_{r,m}\coloneqq\{(i_1,\ldots,i_r):0\le i_1\le\ldots\le i_r\le m,\sum_j i_j=m\}$ sei

$$\tau_m(P,Q)\coloneqq\int\psi(x_1,\ldots,x_m,Q)P^m(dx_1,\ldots,dx_m).$$

Dann gilt

$$\tau_m\left(\widehat{P}_n,P\right)=\sum_{r=1}^m\sum_{i\in\Lambda_{r,m}}\omega_{m,n}(i)\frac{1}{\binom{n}{r}}\sum_{1\le j_1\le\ldots\le j_r\le m}\psi_i(x_{j_1},\ldots,x_{j_r},P)$$

wo $\omega_{m,n}(i)=\binom{n}{r}\binom{m}{i_1\cdots i_r}\frac{1}{n^m}$.
Setze $\widehat{\tau}_{m,n}(P)=\int\tau_m\left(\widehat{P}_n,P\right)dP^n$ und

$$\delta_m\left(\frac{r}{m}\right)\coloneqq\max\{\left|\int\psi_i(x_1,\ldots,x_r,P)P^n(dx_1,\ldots,dx_r)-\tau_m(P)\right|:i\in\Lambda_{r,m}\}.$$

Die Funktion δ_m wird zu $\delta_m:[0,1]\to\mathbb{R}$ mittels linearer Extrapolation fortgesetzt mit der Konvention $\delta_m(1)\coloneqq 0$.

Satz 5.16 (Asymptotische Konsistenz des (m,n)-Verfahrens)**:**
Falls $\delta_m(1-xm^{-1/2})\to 0$ *für* $m\to\infty$ *gleichmäßig für alle* $0\le x\le 1$ *und* $\frac{m}{n}\to 0$*, so gilt*

$$\tau_m\left(\widehat{P}_n,P\right)=\widehat{\tau}_{m,n}(P)+\mathcal{O}_P\left(\left(\frac{m}{n}\right)^{-1/2}\right),$$

und

$$\widehat{\tau}_{m,n}(P)=\tau_m(P)+o_P(1).$$

Falls $T_m\left(X_1^{i_1},\ldots,X_r^{i_r},P\right)=T_m\left(X_1^{i_1},\ldots,X_r^{i_r},\widehat{P}_n\right)+o_P(1)$ *für alle* $i\in\Lambda_{r,m}$ *und* $\max\{i_1,\ldots,i_r\}=\mathcal{O}(m^{1/2})$ *gilt, so erhält man für* $m\to\infty, mn^{-1}\to 0$

$$\tau_m\left(\widehat{P}_n\right)=\tau_m(P)+o_P(1).$$

Zusammenfassung: Resampling
Problem: Im besten Fall (z.B. bei der Student-Statistik) für $\tau(P) = \int x dP$ ist die Statistik $T_n(x_1, \ldots, x_n, P) = \frac{\bar{x}_n - \tau(P)}{S_n}\sqrt{n} \Rightarrow \mathcal{N}(0,1)$ für $n \to \infty$. Für $n \to \infty$ liefert der Limes fast korrekte Niveaus für T_n. Was ist für $n = 15$ und $P \neq \mathcal{N}(\mu, \sigma^2), \mu, \sigma^2 > 0$?
Frage: Wie erhält man $\tau_{n,a}(P) = P^n\left(T_n(\cdot, P) \leq a\right), a \in \mathbb{R}$?
Antwort: *a*): Simulation von Werten $T_n(x_\nu, P)$ für viele n-Tupel $x_\nu \sim P^n$ mit Monte-Carlo-Methoden. *b*): P ist nicht bekannt. Benutze also eine Schätzung für P - das empirische Maß und verfahre wie in *a*), d.h. man simuliere Werte $T_n\left(x_n^*, \widehat{P}_n\right)$ und hoffe auf Approximation von $\tau_{n,a}(P)$ durch $\tau_{n,a}\left(\widehat{P}_n\right)$ für $n \to \infty$.
Es entstehen allerdings Probleme bei Statistiken, die „sensitiv“ gegenüber Wiederholung der Beobachtungen sind. Es kann sein, dass $\tau_{n,a}\left(\widehat{P}_n\right) \to H_a$ P^n-stochastisch, wobei H_a eine Zufallsvariable ist und $H_a \neq \tau(P, a)$.
Lösung: Unter der Annahme $\tau_{m,a}(P) - \tau_{n,a}(P) \to 0$ für $m, n \to \infty$ (dies ist z.B. erfüllt, falls $\lim_{n\to\infty} \tau_{n,a}(P) =: \tau_a(P)$ für alle a existiert): Schätze für $mn^{-1} \to 0, m, n \to \infty$ die Verteilungsfunktion $\tau_{m,a}(P)$ durch $\tau_{m,a}\left(\widehat{P}_n\right)$ auf eine konsistente Weise. Das Verfahren heißt Subsampling (m, n) mit $m < n$. Falls mn^{-1} hinreichend klein ist, gilt mit

$$p_{m,n} \coloneqq P_n^m\left(x_j^* = x_k^* : 1 \leq j < k \leq m\right) \leq \frac{m(m-1)}{2}\widehat{P}_n^2\left(x_1^* = x_2^*\right) = \frac{m(m-1)}{2n},$$

d.h. für $p_{m,n} \to 0$ muss $\frac{m}{\sqrt{n}} \to 0$ gehen.
Bessere Strategie: Schließe Wiederholungen aus in $(x_1^*, \ldots, x_m^*)$, d.h. ein Ziehen ohne Zurücklegen. Dies ergibt Stichproben aus $\widehat{P}_n^m : (x_1^{**}, \ldots, x_m^{**}) \sim \widehat{P}_n^{(m)}$ (kein Produktmaß). Für $m = n$ gilt $\{x_1^{**}, \ldots, x_n^{**}\} = \{x_1, \ldots, x_n\}$. Also muss auch hier $m < n$ gelten. Für den Schätzer $\tilde{\tau}_{m,n}\left(\widehat{P}_n, P\right)$ von $\tau_{m,a}(P)$ gilt

$$\mathbb{E}_{P^n}\tilde{\tau}_{m,a}\left(\widehat{P}_n, P\right) = \tau_{m,a}(P).$$

Weitere Informationen: [5], [4], [2].

6 Zeitreihenanalyse

Seien $Y_t, t \in \mathbb{N}$ eine Folge von Zufallsvariablen $Y_t : \mathcal{X} \to \mathbb{R}$ auf $(\mathcal{X}, \mathcal{A}, P)$ (Zeitreihe). Modelle der Form für Y_t:

$$Y_t = T_t + S_t + \varepsilon_t \tag{6.1}$$

wobei T_t einen globalen Trend darstellt, S_t einen saisonalen, d.h. periodischen Trend, darstellt (d.h. deterministische Größen) und

$$\varepsilon_t \text{ i.i.d. mit } \int \varepsilon_t dP = 0, \int \varepsilon_t^2 dP = 1. \tag{6.2}$$

Beispiel:
Sei $S_t \equiv 0$, $T_t = \sum_{j=0}^{d} \alpha_j t^j, \alpha_j \in \mathbb{R}$, d.h.

$$Y_t = \sum_{j=0}^{d} \alpha_j t^j + \varepsilon_t.$$

Regressions-Problem: Beobachte Y_t für $t = 1, \ldots, T$ und schätze T_t, d.h. $\alpha_0, \ldots, \alpha_d$. Angenommen die ε_t haben eine Dichte p bezüglich eines dominierenden Maßes μ, so bestimme die Parameter mittels der Maximum-Likelihood-Methode: $(\varepsilon_1, \ldots, \varepsilon_T)$ hat μ^T-Dichte $p(e_1) \cdots p(e_T) d\mu^T$. Also hat $(\varepsilon_1, \ldots, \varepsilon_T)$ die μ^T-Dichte $\prod_{t=1}^{T} p(Y_t - \sum_{j=0}^{d} \alpha_j t^j) =: p^T(Y, \alpha)$. Dies ist ein parametrisches Modell $(\alpha_0, \ldots, \alpha_d)$. Benutze den Maximum-Likelihood-Schätzer

$$\widehat{\alpha} = \arg\max_{\alpha} p^T(Y, \alpha).$$

Zum Beispiel für $p(x) = \varphi(x)$ als Dichte der Standard-Normalverteilung finde den Vektor α mit

$$\sum_{t=1}^{T} \left(\left(Y_t - \sum_{j=0}^{d} \alpha_j t^j \right) t^p \right) = 0 \text{für } p = 0, \ldots, d. \tag{6.3}$$

Dies sind $d + 1$ Gleichungen für α.

6.1 Zyklen in Zeitreihen

Seien ε_t i.i.d. mit Mittelwert 0 und Varianz 1. Benutze das Modell

$$Y_t = \sum_{j=0}^{n} (\alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t))) + \varepsilon_t \tag{6.4}$$

wo $\omega_j = \frac{2\pi j}{T}, j = 0, \ldots, \frac{T}{2}$, $T = 2n, n \in \mathbb{N}$. Dies nennt man ein harmonisches Modell.
Schätze α_j, β_j aus $Y_t, t = 0, \ldots, T-1$. Die Regeln der komplexen Exponentialfunktion liefern

$$0 = \sum_{t=0}^{T-1} \cos(\omega_j t) = \sum_{t=0}^{T-1} \sin(\omega_j t), j \neq lT, l \in \mathbb{Z}\,.$$

und

$$\sum_{t=0}^{T-1} \sin(\omega_j t) \sin(\omega_k t) = \sum_{t=0}^{T-1} \cos(\omega_j t) \cos(\omega_k t) = \delta_{j,k} \frac{T}{2}$$

sowie

$$\sum_{t=0}^{T-1} \sin(\omega_j t) \cos(\omega_k t) = 0.$$

Im Fall $\varepsilon_t \equiv 0$ für alle t ergibt sich

$$\begin{aligned} \alpha_0 &:= \frac{1}{T} \sum_{t=0}^{T-1} Y_t =: \overline{Y} \\ \alpha_j &= \frac{2}{T} \sum_{t=0}^{T-1} Y_t \cos(\omega_j t) \\ \beta_j &= \frac{2}{T} \sum_{t=0}^{T-1} Y_t \sin(\omega_j t) \end{aligned}$$

Dann folgt

$$\sum_{t=0}^{T-1} (Y_t - \bar{Y})^2 = \frac{T}{2} \sum_{j=1}^{n} \left(\alpha_j^2 + \beta_j^2\right) \tag{6.5}$$

da alle Terme des Typs $\alpha_j \alpha_k \sum_{t=0}^{T-1} \cos(\omega_j t) \cos(\omega_k t), j \neq k$ sowie die mit $\beta_j \beta_k \sum_{t=0}^{T-1} \sin(\omega_j t) \sin(\omega_k t)$ und $\alpha_j \beta_k \sum \sin(\omega_j t) \cos(\omega_k t)$ verschwinden. Dann folgt aus obiger Beschreibung von α_j, β_j

$$\sum_{t=0}^{T-1} (Y_t - \bar{Y})^2 = \frac{2}{T^2} \sum_{j=0}^{n} \left(\left(\sum_{j=0}^{T-1} Y_t \cos(\omega_j t) \right)^2 + \left(\sum_{t=0}^{T-1} Y_t \sin(\omega_j t) \right)^2 \right).$$

Die Funktion

$$I(\omega_j) = \frac{T}{2}\left(\alpha_j^2 + \beta_j^2\right) \tag{6.6}$$

für $j = 0, \ldots, T-1$ heißt Periodogramm. Die Verallgemeinerung von (6.6) sind die empirischen Kovarianzen zum „lag“ τ:

$$c_\tau \coloneqq \frac{1}{T}\sum_{t=0}^{T-1}(Y_t - \bar{Y})(Y_{t-\tau} - \bar{Y}) \tag{6.7}$$

(und c_0 entspricht der Varianz). Der Autokorrelationskoeffizient ist definiert als

$$r_\tau = \frac{c_\tau}{c_0}. \tag{6.8}$$

Dann gilt

$$I(\omega_j) = \frac{2}{T}\left(\left(\sum_{t=0}^{T-1}\cos(\omega_j t)(Y_t - \bar{Y})\right)^2 + \left(\sum_{t=0}^{T-1}\sin(\omega_j t)(Y_t - \bar{Y})\right)^2\right) \tag{6.9}$$

Entwickelt man die Quadrate in dieser Formel, so gilt mit den Additionstheoremen der trigonometrischen Funktionen

$$I(\omega_j) = \frac{2}{T}\left(\sum_{t=0}^{T-1}\sum_{s=0}^{T-1}\cos(\omega_j(t-s))(Y_t - \bar{Y})(Y_s - \bar{Y})\right).$$

Setze $\tau \coloneqq t - s$, so gilt

$$I(\omega_j) = 2\sum_{\tau=1-T}^{T-1}\cos(\omega_j\tau)c_\tau.$$

d.h. das Periodogramm ist die diskrete Fouriertransformierte der Autokorrelation c_τ.
Entsprechend definiert man durch Formel (6.8) (bei Götze :-)) Schätzungen für $\alpha_0, \ldots, \alpha_{T-1}$, falls $\varepsilon_t \neq 0, t \in \mathbb{Z}$

$$\widehat{\alpha}_{0,s} = \frac{1}{s}\sum_{t=0}^{s} Y_t \text{ für } s \geq T, s = kT, k \in \mathbb{N}$$

und

$$\widehat{\alpha}_{j,s} = \frac{2}{s}\sum_{t=0}^{s} Y_t\cos(\omega_j t).$$

Überlege, dass $\widehat{\alpha}_{\cdot,s}$ ein konsistenter Schätzer für $\alpha_0, \ldots, \alpha_{T-1}$ ist.

6.2 Modelle abhängiger Zeitreihen

Allgemein: Input besteht aus einem Rauschen $\varepsilon_t, t \in \mathbb{Z}$ i.i.d. $\mathbb{E}\varepsilon_j = 0, \mathbb{E}\varepsilon_j^2 = \sigma^2 > 0$. Beobachtet wird $Y_t = f(Y_{t-\tau}, \tau \geq 1, \varepsilon_s, s \leq t)$ (Rekursion mit Rauschanteil). Y_t ist ein Markov-Prozess.

Beispiel:

(i) Moving-Average-Modell: $Y_t = \sum_{l=0}^{q} \mu_l \varepsilon_{t-l}$, ε_t sind i.i.d. wie oben. Es gilt $\mathbb{E}Y_t = 0$ und

$$\sigma(t,s) \coloneqq \mathbb{E}Y_t Y_s = \begin{cases} \sum_{l=0}^{q-|t-s|} \mu_l \mu_{l+|t-s|} & |t-s| \leq q \\ 0 & |t-s| > q \end{cases}.$$

Die Stichprobenkorrelation ist $c_h = c_{-h} \coloneqq \frac{1}{T} \sum_{t=1}^{T-h} Y_t Y_{t+h}, h = 0, \ldots, T-1$.
Behauptung: Mit $h \geq q$ fest, $n \leq T-1$ hat

$$\left(\sqrt{T} \frac{c_h}{c_0} - \frac{\sigma(h)}{\sigma(0)} \right)$$

im Limes Normalverteilung, wenn $T \to \infty$ der Form $\mathcal{N}(0, \Sigma)$, vgl. [1].

(ii) Ein weiteres Modell ist die einfache Rekursion, genannt Autoregression. Ein einfaches Modell ist das $AR(1)$-Modell, d.h. ein Modell der Autoregression erster Ordnung, d.h.

$$Y_t = \alpha Y_{t-1} + \varepsilon_t \tag{6.10}$$

mit ε_t wie oben, $|\alpha| < 1$. Allgemeiner heißt

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \varepsilon_t$$

ein Autoregressionsmodell p-ter Ordnung, $AR(p)$.

(iii) Sei $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d. Dann ist $Y_t - \alpha Y_{t-1} = \varepsilon_t$ im $AR(1)$-Modell. Dann ist die Dichte von $(Y_t, t = 1, \ldots, T)$ durch

$$\prod_{t=1}^{T} \varphi(Y_t - \alpha Y_{t-1}) \tag{$*$}$$

gegeben, wo φ die Dichte der Standard-Normalverteilung ist. Der Maximum-Likelihood-Schätzer für α ist gegeben durch die Maximierung des Ausdrucks in $(*)$, d.h.

$$\sum_{t=1}^{T} \frac{\varphi'}{\varphi} (Y_t - \alpha Y_{t-1}) \, Y_{t-1} = 0$$

oder

$$0 = \sum_{t=1}^{T} (Y_t - \alpha Y_{t-1}) Y_{t-1} \Leftrightarrow \widehat{\alpha} = \frac{\sum_{t=1}^{T} Y_t Y_{t-1}}{\sum_{t=1}^{T} Y_{t-1}^2}.$$

Hier gilt

$$\widehat{\alpha} = \frac{\sum_{t=1}^{T} Y_{t-1}^2 \alpha}{\sum_{t=1}^{T} Y_{t-1}^2} + \frac{\sum_{t=1}^{T} Y_{t-1} \varepsilon_t}{\sum_{t=1}^{T} Y_{t-1}^2}.$$

Sei $\mathcal{F}_s = \sigma(\varepsilon_t, t \leq s)$. Dann ist $S_t = \sum_{s=1}^{t} Y_{s-1}\varepsilon_s$ ein $\mathcal{F}$-Martingal mit

$$\mathbb{E}S_t^2 = \sum_{s=1}^{T} \mathbb{E}(Y_{s-1}\varepsilon_s)^2 = \sum_{s=1}^{T} \mathbb{E}Y_{s-1}^2 = \frac{T}{1-\alpha^2}$$

wegen

$$Y_t = Y_{t-1}\alpha - \varepsilon_t = Y_{t-2}\alpha^2 + \alpha\varepsilon_{t-1} + \varepsilon_t = \ldots = \sum_{j=0}^{\infty} \varepsilon_{t-j}\alpha^j$$

sodass

$$\mathbb{E}Y_{s-1}^2 = \mathbb{E}\left(\sum_{j=0}^{\infty} \varepsilon_{s-j-1}\alpha^j\right)^2 = \sum_{j=0}^{\infty} \alpha^{2j} = \frac{1}{1-\alpha^2}$$

sowie

$$\frac{1}{T}\sum_{j=1}^{T} Y_{t-1}^2 \to \frac{1}{1-\alpha^2}$$

P-f.s., woraus $\widehat{\alpha} = \alpha + \mathcal{O}_P\left(T^{-1/2}\right)$ (Übung). In (6.15$*$) heißen die ε_j Residuen.
Was passiert, wenn die Residuen abhängig sind?

(iv) Modell: Autoregressives-Moving-Average-Modell (ARMA):

$$Y_t + \sum_{j=1}^{p} \alpha_j Y_{t-j} = \sum_{l=0}^{q} \mu_l \varepsilon_{t-l} \tag{6.11}$$

d.h. ein Modell für abhängige Residuen, wo $\mu_0 = \alpha_0 = 1$ und ε_t i.i.d. zentriert und mit Varianz $\sigma^2 > 0$, kurz: ARMA(p, q).
Unter welchen Bedingungen konvergieren in (6.14) – (6.16) die Verteilungen der Y_t?
Auflösung der Rekursion: Sei $Y = (Y_t, t \in \mathbb{Z})$ die Zeitreihe in (6.16). Definiere auf dem linearen Vektorraum von Funktionen über $\mathbb{R}$ für festes $\omega \in \Omega$ den Lag-Operator

$$(LY)_t := Y_{t-1}$$

und

$$(\nabla Y)_t \coloneqq Y_t - Y_{t-1} = ((\mathrm{Id} - L)Y)_t$$

sowie die Summe

$$(SY)_t \coloneqq \sum_{j=0}^{\infty} Y_{t-j} = \left(\sum_{j=0}^{\infty} (L^j Y)_t \right).$$

Ferner sei $p(x) = \sum_{j=0}^{n} p_j x^j$ ein Polynom und

$$(P(L)Y)_t = p_0 Y_t + p_1 Y_{t-1} + \ldots + p_n Y_{t-n}.$$

Schreibe also (6.16) als Operatorgleichung mit Hilfe von

$$\alpha(z) = \sum j = 0^p \alpha_j z^j \text{ und } \mu(z) = \sum_{l=0}^{q} \mu_l z^l,$$

sodass

$$\alpha(L)Y = \mu(L)\varepsilon \tag{6.12}$$

für $\varepsilon = (\varepsilon_t, t \in \mathbb{Z})$. Man gebe nun der „Lösung“ $Y = \alpha(L)^{-1}\mu(L)\varepsilon$ einen Sinn. Da $\frac{\mu(z)}{\alpha(z)}$ rational ist und über $\mathbb{C}$ gilt

$$\alpha = \alpha_p \prod_{j=1}^{p} (z - \lambda_j)$$

wobei λ_j die komplexen Nullstellen sind, d.h.

$$\alpha = \alpha_0 \prod_{j=1}^{p} \left(1 - \frac{z}{\lambda_j}\right)$$

da $\frac{\alpha_0}{\alpha_p} = \prod_{j=1}^{p} (-\lambda_j)$.
Ferner gilt für

(1) $q = \deg \mu(z) < \deg \alpha(z) = p$,

(2) Alle Wurzeln λ_j sind verschieden,

(3) μ und α haben keine gemeinsamen Nullstellen.

die Partialbruchzerlegung

$$\frac{\mu(z)}{\alpha(z)} = \frac{K_1}{1 - \frac{z}{\lambda_1}} + \ldots + \frac{K_p}{1 - \frac{z}{\lambda_p}}.$$

Falls $|\lambda_j| > 1$ für alle $j = 1, \ldots, p$, dann konvergieren die Reihen $\frac{1}{1-\frac{z}{\lambda_j}} = \sum_{n=0}^{\infty} \frac{z^n}{\lambda_j^n}$ für $|z| < 1$, d.h. auch $\left(Id - \frac{L}{\lambda_j}\right)^{-1}$ konvergiert, da $\|L\|_\infty = 1$.
Folglich würden wir für die Lösung in (6.16∗) für $|\lambda_j| > 1$

$$Y = \sum_{j=1}^{p} \frac{K_j}{1 - \frac{L}{\lambda_j}} \varepsilon \qquad (6.16 * *)$$

oder

$$Y_t = \sum_{j=1}^{p} K_j \left(\sum_{\tau=0}^{\infty} \varepsilon_{t-\tau} \lambda_j^{-\tau} \right) = \sum_{\tau=0}^{\infty} \underbrace{\left(\sum_{j=1}^{p} K_j \lambda_j^{-\tau} \right)}_{=:c_\tau} \varepsilon_{t-\tau}.$$

Da

$$\mathbb{E} Y_t^2 = \sum_{\tau} \left(\operatorname{Var} \left(\operatorname{Re} \left(c_\tau \varepsilon_{t-\tau} \right) \right) \right) + \left(\operatorname{Var} \left(\operatorname{Im} \left(c_\tau \varepsilon_{t-\tau} \right) \right) \right).$$

7 Nachträge

7.1 Einfachster Fall: Sei $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\vartheta, \mathrm{Id}_n), \vartheta \in \mathbb{R}^n$, d.h. x_i sind i.i.d. mit $x_j \sim \mathcal{N}(\vartheta_j, 1)$. Schätze ϑ aus X mit quadratischem Verlust

$$L(x, \vartheta) \coloneqq \|x - \vartheta\|_2^2 .$$

$\tilde{\vartheta}(x) \coloneqq X$ ist ein erwartungstreuer Schätzer für ϑ mit Risiko

$$R(\tilde{\vartheta}, \vartheta) = \mathbb{E}_\vartheta \|x - \vartheta\|_2^2 = \sum_{j=1}^{n} \mathbb{E}_{\vartheta_j} |x_j - \vartheta_j|^2 = n$$

für alle ϑ. Wegen Cramer-Rao ist das Risiko minimal in der Klasse der erwartungstreuen Schätzer.
Frage: Ist die Entscheidungsregel bzw. der Schätzer $T(X) = X$ zulässig für L? Antwort: Nein für $n \geq 3$.

Satz 7.1 (James & Stein)**:**
Der Schätzer $\tilde{\vartheta}_{JS}(X) = \left(1 - \frac{n-2}{\|x\|_2^2}\right) X$ *hat für* $n \geq 3$ *überall kleineres quadratisches Risiko als* $\tilde{\vartheta}$*, nämlich*

$$R(\tilde{\vartheta}_{JS}, \vartheta) = \underbrace{R(\tilde{\vartheta}, \vartheta)}_{=n} - (n-2)\mathbb{E}_\vartheta \left(\frac{1}{\|x\|_2^2}\right) < R(\tilde{\vartheta}, \vartheta) \equiv n$$

für alle ϑ*.*

Beweis. Zunächst sei $\tilde{\vartheta}_{JS} = \left(1 - \frac{\alpha}{\|X\|_2^2}\right) X$, später $\alpha = n - 2$. Dann gilt

$$R(\tilde{\vartheta}_{JS}, \vartheta) = \mathbb{E}_\vartheta \left(\left(X - \vartheta - \frac{\alpha}{\|X\|^2} \right)^2 \right) = R(\tilde{\vartheta}, \vartheta) - \underbrace{2\alpha \mathbb{E}_\vartheta (X - \vartheta) \cdot \frac{X}{\|X\|^2}}_{=:J} + \alpha^2 \mathbb{E}_\vartheta \frac{1}{\|X\|^2}$$

und es gilt mit partieller Integration in jeder Variablen x_j

$$\begin{aligned} J &= \int \frac{(x - \vartheta) \cdot x}{\|x\|^2} \exp\left(-\|x - \vartheta\|^2 / 2\right) \sqrt{2\pi}^{-n} dx \\ &= \int \left(-\nabla_x \exp\left(-\|x - \vartheta\|^2 / 2\right) \cdot \frac{x}{\|x\|^2} \sqrt{2\pi} \right) dx \\ &= \int \exp\left(-\|x - \vartheta\|^2 / 2\right) \nabla_x \cdot \frac{x}{\|x\|^2} dx \end{aligned}$$

Nun gilt

$$\nabla_x \cdot \frac{x}{\|x\|^2} = \frac{n}{\|x\|^2} - \sum_{j=1}^{n} \frac{x_j 2x_j}{\|x\|^4} = \frac{(n-2)}{\|x\|^2}$$

und damit

$$J = (n-2)\mathbb{E}_\vartheta \frac{1}{\|x\|^2}$$

und dies existiert für $n \geq 3$. Zusammen ergibt sich

$$R(\tilde{\vartheta}_{JS}, \vartheta) = R(\tilde{\vartheta}, \vartheta) - 2\alpha(n-2)\mathbb{E}_\vartheta \|x\|^{-2} + \alpha^2 \mathbb{E}_\vartheta \|x\|^{-2},$$

was für $\alpha = n-2$ minimiert wird. Daher gilt

$$R(\tilde{\vartheta}_{JS}, \vartheta) \geq R(\tilde{\vartheta}, \vartheta) - (n-2)^2 \mathbb{E}_\vartheta \|x\|^{-2}.$$

■

Bemerkung 7.2:
a): Gilt auch für Punkte $\vartheta \neq 0$: $\widehat{\vartheta}(x) = \left(1 - \frac{(n-2)\sigma^2}{\|x-\vartheta\|^2}_+ (x-\nu) + \nu\right)$ für ein $\nu \in \mathbb{R}^n$ ist besser als $\tilde{\vartheta}$ für $\mathcal{N}(\vartheta, \sigma^2 \,\mathrm{Id})$.
b): *a*) gilt auch noch, wenn σ^2 durch $\widehat{\vartheta}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$ ersetzt wird.
c): Falls $X = (X_1, \ldots, X_m)$ mit $X_j \in \mathbb{R}^n, m \in \mathbb{N}$ und $\tilde{\vartheta}(X) := \bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ ein Schätzer für ϑ in $\mathcal{N}(\vartheta, \sigma^2 \,\mathrm{Id})$. Dann ist

$$\tilde{\vartheta}_{JS} = \left(1 - \frac{(n-2)\sigma^2}{n\|\bar{x}\|^2}\right)\bar{X}$$

$L-$besser als $\tilde{\vartheta}$.

Beispiel 7.3 (Semiparametrik)**:**
Sei $\mathcal{P} = \{p(x-\vartheta)d\lambda, \vartheta \in \mathbb{R}, p(x) = p(-x), p \in H\}$, wobei $H = \{p \geq 0, I_p := \int \left(\frac{p'}{p}\right)^2 d\lambda < \infty, p \in \mathcal{C}^2, -\int p'' p d\lambda = I_p, \int p' d\lambda = 0, \int x^2 p(x) d\lambda < \infty\}$ und die Bedingungen von Lemma 3.7.
Parameter: $\kappa(P) = \int xp(x)dx = \vartheta, P \in \mathcal{P}$. Beobachte $(x_1, \ldots, x_n) \sim p(x - theta)d\lambda$. Schätze ϑ effektiv, d.h. mit minimaler asymptotischer Varianz, wobei p unbekannt ist.
Frage: Gibt es eine Adaption? Falls p bekannt ist, so haben wir nach Kapitel 3, Lemma 3.7 den Maximum-Likelihood-Schätzer

$$\widehat{\eta}_n = \sum_{j=1}^{n} \frac{p'}{p}(x_j - \widehat{\vartheta}_n) = 0$$

ist konsistent und effektiv, d.h. $\sqrt{n}\left(\widehat{\vartheta}_n - \vartheta\right) \Rightarrow \mathcal{N}(0, I_p^{-1})$ als minimale Varianz.

Definition 7.4 (Stochastische Landau-Notation)**:**
Sei $(X_n)_n$ eine Folge von reellwertigen Zufallsvariablen und $f : \mathbb{N} \to \mathbb{R}$ eine Abbildung. Dann gilt

(1) $X_n = \mathcal{O}_P(f(n)) :\Longleftrightarrow \limsup_n P(|X_n| > f(n) \cdot c) \leq \varepsilon(c)$, wobei $\lim_{c\to\infty} \varepsilon(c) = 0$.

(2) $X_n = o_P(f(n)) :\Longleftrightarrow \lim_n P(|X_n| > f(n)c) = 0$.

(3) Eine reellwertige Zufallsvariable X ist stochastisch beschränkt, falls für $X_n \coloneqq \frac{1}{n}X$ gilt $X_n = o_P(1)$, d.h. es gilt $\lim_n P(|X_n| > 1) = 0 \Longleftrightarrow \lim_K P(|X| > K) = 0$.

(4) Eine Schätzerfolge $(\kappa_n)_{n\in\mathbb{N}}$ ist konsistent für ein Funktional κ, falls $\kappa_n - \kappa(P) = o_{P^{\mathbb{N}}}(1)$.

Literatur

[1] Theodore W. Anderson. *The statistical analysis of time series.* Wiley classics library. Wiley, New York [u.a.], 1994.

[2] Bradley Efron and Robert Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on statistics and applied probability ; 57.* Chapman and Hall, New York [u.a.], 1993.

[3] Herbert Federer. *Geometric measure theory*, volume 153 of *Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen ; 153.* Springer, Berlin [u.a.], 1969.

[4] Peter Hall. *The bootstrap and Edgeworth expansion.* Springer series in statistics. Springer, New York [u.a.], 1992.

[5] Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses.* Springer texts in statistics. Springer, New York [u.a.], 3. ed. edition, 2005.

[6] Johann Pfanzagl. *Parametric statistical theory.* De Gruyter textbook. de Gruyter, Berlin [u.a.], 1994.

[7] Bernard W. Silverman. *Density estimation for statistics and data analysis*, volume [26] of *Monographs on statistics and applied probability ; [26].* Chapman and Hall, London [u.a.], 1986.

[8] Helmut Strasser. *Mathematical theory of statistics : statistical experiments and asymptotic decision theory*, volume 7 of *De Gruyter studies in mathematics ; 7.* de Gruyter, Berlin [u.a.], 1985.