Mathematische Statistik
Skript zur Vorlesung
im Sommersemester 2022

Markus Reiß
Humboldt-Universität zu Berlin
mreiss@math.hu-berlin.de

VORLÄUFIGE FASSUNG: 21. Juli 2022

Inhaltsverzeichnis

1 Entscheidungstheorie

1.1 Formalisierung eines statistischen Problems

1.1 Definition. Ein Messraum $(\mathcal{X}, \mathscr{F})$ versehen mit einer Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ von Wahrscheinlichkeitsmaßen, $\Theta \neq \varnothing$ beliebige Parametermenge, heißt statistisches Experiment oder statistisches Modell. $\mathcal{X}$ heißt Stichprobenraum. Jede $(\mathscr{F}, \mathscr{S})$-messbare Funktion $Y : \mathcal{X} \to S$ heißt Beobachtung oder Statistik mit Werten in $(S, \mathscr{S})$ und induziert das statistische Modell $(S, \mathscr{S}, (\mathbb{P}_\vartheta^Y)_{\vartheta\in\Theta})$. Sind die Beobachtungen $Y_1, \dots, Y_n$ für jedes $\mathbb{P}_\vartheta$ unabhängig und identisch verteilt, so nennt man $Y_1, \dots, Y_n$ eine mathematische Stichprobe.

1.2 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Eine Entscheidungsregel ist eine messbare Abbildung $\rho : \mathcal{X} \to A$, wobei der Messraum $(A, \mathscr{A})$ der sogenannte Aktionsraum ist. Jede Funktion $l : \Theta \times A \to [0, \infty) =: \mathbb{R}^+$, die messbar im zweiten Argument ist, heißt Verlustfunktion. Das Risiko einer Entscheidungsregel ρ bei Vorliegen des Parameters $\vartheta \in \Theta$ ist

$$R(\vartheta, \rho) := \mathbb{E}_\vartheta[l(\vartheta, \rho)] = \int_{\mathcal{X}} l(\vartheta, \rho(x))\, \mathbb{P}_\vartheta(dx).$$

1.3 Beispiele.

(a) Wir formalisieren das Beobachtungsmodell

$$Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, n,$$

mit unabhängigen Fehlervariablen $\varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2)$. Dann ist der Beobachtungsvektor $Y = (Y_1, \dots, Y_n)^\top$ $N(\mu\mathbf{1}_n, \sigma^2 E_n)$-verteilt mit $\mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$ und n-dimensionaler Einheitsmatrix E_n. Als statistisches Modell wählen wir daher $(\mathbb{R}^n, \mathfrak{B}_{\mathbb{R}^n}, (N(\mu\mathbf{1}_n, \sigma^2 E_n))_{\mu\in\mathbb{R},\sigma>0})$. Die Parametermenge ist $\Theta = \mathbb{R} \times (0, \infty)$ mit Parametern $\vartheta = (\mu, \sigma)$. Alternativ können wir sagen, dass $Y_1, \dots, Y_n$ eine $N(\mu, \sigma^2)$-verteilte mathematische Stichprobe ist.

Um das Stichprobenmittel $\bar{Y} := \rho(Y_1, \dots, Y_n) := \frac{1}{n}\sum_{i=1}^n Y_i$ als Entscheidungsregel zu interpretieren und seine Güte bei der Schätzung von μ zu messen, betrachtet man den Aktionsraum $A = \mathbb{R}$ und beispielsweise die quadratische Verlustfunktion $l(\vartheta, a) = l((\mu, \sigma), a) = (\mu - a)^2$. Beim Verlust ist σ irrelevant; da aber die Verteilung $\mathbb{P}_\vartheta$ von σ abhängt, spricht man von einem Störparameter. Das quadratische Risiko (auch MSE: mean squared error) ist $R((\mu, \sigma), \rho) = \mathbb{E}_{\mu,\sigma}[(\mu - \bar{Y})^2] = \sigma^2 n^{-1}$, da ja $\bar{Y} - \mu \sim N(0, \sigma^2 n^{-1})$.

(b) Allgemeiner können wir das Beobachtungsmodell

$$Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, n,$$

mit zentrierten und unkorrelierten Fehlervariablen $\varepsilon_1, \dots, \varepsilon_n$ betrachten. Ist die Art der Verteilung der (ε_i) unbekannt, sollte man auf dem Stichprobenraum $(\mathbb{R}^n, \mathfrak{B}_{\mathbb{R}^n})$ die Familie $\mathcal{P} = \{\mathbb{P}$ W-Maß auf $\mathfrak{B}_{\mathbb{R}^n} \mid \int_{\mathbb{R}^n} x\, \mathbb{P}(dx) =$

$\mu\mathbf{1}_n$, $\int_{\mathbb{R}^n}(x-\mu\mathbf{1}_n)(x-\mu\mathbf{1}_n)^\top\,\mathbb{P}(dx)=\sigma^2E_n$, $\mu\in\mathbb{R}$, $\sigma>0\}$ betrachten. In dieser Betrachtungsweise bleibt von einem unendlich-dimensionalen Parameterraum $\mathcal{P}$ maximal ein zweidimensionaler interessierender Parameter $\vartheta=(\mu,\sigma)$ übrig. Interessanterweise ändert sich das quadratische Risiko des Stichprobenmittels in diesem allgemeineren Modell nicht.

(c) Im Gaußschen multivariaten linearen Modell beobachten wir

$$Y_i=\langle x_i,\beta\rangle+\varepsilon_i,\quad i=1,\dots,n,$$

mit gegebenen Kovariablen $x_1,\dots,x_n\in\mathbb{R}^p$, interessierendem Parameter $\beta\in\mathbb{R}^p$ und $\varepsilon_1,\dots\varepsilon_n\sim N(0,\sigma^2)$ unabhängig. Als statistisches Modell ergibt sich $(\mathbb{R}^n,\mathfrak{B}_{\mathbb{R}^n},(\otimes_{i=1}^n N(\langle x_i,\beta\rangle,\sigma^2))_{\beta\in\mathbb{R}^p,\sigma>0})$. Mit der Designmatrix $X=(x_1^\top,\dots,x_n^\top)^\top\in\mathbb{R}^{n\times p}$ gilt äquivalent $\otimes_{i=1}^n N(\langle x_i,\beta\rangle,\sigma^2)=N(X\beta,\sigma^2E_n)$. Der Kleinste-Quadrate-Schätzer ist $\hat\beta=(X^\top X)^{-1}X^\top Y$, sofern X Rang p besitzt ($x_1,\dots,x_n$ spannen den $\mathbb{R}^p$ auf). Mit Aktionsraum $A=\mathbb{R}^p$ und quadratischem Verlust $l((\beta,\sigma),a)=|\beta-a|^2$ (mit Euklidischer Norm $|\bullet|$) erhalten wir das quadratische Risiko des Kleinste-Quadrate-Schätzers

$$R((\beta,\sigma),\hat\beta)=\mathbb{E}_{\beta,\sigma}[|\beta-\hat\beta|^2]=\mathbb{E}[|(X^\top X)^{-1}X^\top\varepsilon|^2]=\sigma^2\,\mathrm{trace}((X^\top X)^{-1})$$

mit der Spur $\mathrm{trace}(M):=\sum_i M_{i,i}^2$.

(d) Für einen Test auf Wirksamkeit eines neuen Medikaments werden 100 Versuchspersonen mit diesem behandelt. Unter der (stark vereinfachenden) Annahme, dass alle Personen identisch und unabhängig auf das Medikament reagieren, wird in Abhängigkeit von der Anzahl N der erfolgreichen Behandlungen entschieden, ob die Erfolgsquote höher ist als diejenige einer klassischen Behandlung. Als Stichprobenraum wähle $\mathcal{X}=\{0,1,\dots,100\}$ mit der Potenzmenge als σ-Algebra und $\mathbb{P}_p=\mathrm{Bin}(100,p)$, $p\in\Theta=[0,1]$, als mögliche Verteilungen. Die Nullhypothese ist $H_0:p\leqslant p_0$ für den unbekannten Parameter p. Als Aktionsraum dient $A=\{0,1\}$ (H_0 annehmen bzw. verwerfen), und wir wählen den Verlust $l(p,a)=\ell_0\mathbf{1}_{\{p\leqslant p_0,\,a=1\}}+\ell_1\mathbf{1}_{\{p>p_0,\,a=0\}}$ mit Konstanten $\ell_0,\ell_1\geqslant 0$. Dies führt auf das Risiko einer Entscheidungsregel (eines Tests) ρ

$$R(p,\rho)=\begin{cases}\ell_0\,\mathbb{P}_p(\rho=1), & p\leqslant p_0\\ \ell_1\,\mathbb{P}_p(\rho=0), & p>p_0\end{cases}$$

und die Fehlerwahrscheinlichkeit erster Art wird mit ℓ_0, die zweiter Art mit ℓ_1 gewichtet.

1.4 Definition. Die Entscheidungsregel ρ heißt besser als eine Entscheidungsregel ρ', falls $R(\vartheta,\rho)\leqslant R(\vartheta,\rho')$ für alle $\vartheta\in\Theta$ gilt und falls ein $\vartheta_0\in\Theta$ mit $R(\vartheta_0,\rho)<R(\vartheta_0,\rho')$ existiert. Eine Entscheidungsregel heißt zulässig, wenn es keine bessere Entscheidungsregel gibt.

1.5 Bemerkung. Häufig wird die Menge der betrachteten Entscheidungsregeln eingeschränkt. Bei Schätzern wird beispielsweise Erwartungstreue, Linearität

oder allgemeiner Invarianz bezüglich gewisser Gruppenoperationen (vergleiche *Äquivarianz* weiter unten) gefordert. So ist der Kleinste-Quadrate-Schätzer im linearen Modell nach dem Satz von Gauß-Markov zulässig unter quadratischem Verlust in der Klasse der erwartungstreuen und linearen Schätzern.

1.6 Beispiel. Es sei $Y_1, \dots, Y_n$ eine $N(\vartheta, 1)$-verteilte mathematische Stichprobe mit $\vartheta \in \mathbb{R}$. Betrachte $\hat{\vartheta}_1 = \bar{Y}$, $\hat{\vartheta}_2 = \bar{Y} + 0.5$, $\hat{\vartheta}_3 = 6$ unter quadratischem Verlust $l(\vartheta, a) = (\vartheta - a)^2$. Wegen $R(\vartheta, \hat{\vartheta}_1) = 1/n$, $R(\vartheta, \hat{\vartheta}_2) = 0.25 + 1/n$ ist $\hat{\vartheta}_1$ besser als $\hat{\vartheta}_2$, allerdings ist weder $\hat{\vartheta}_1$ besser als $\hat{\vartheta}_3$ noch umgekehrt. In der Tat ist $\hat{\vartheta}_3$ zulässig, weil $R(\vartheta, \hat{\vartheta}_3) = 0$ für $\vartheta = 6$ gilt und jeder Schätzer mit dieser Eigenschaft Lebesgue-fast überall mit $\hat{\vartheta}_3$ übereinstimmt. Später werden wir sehen, dass auch $\hat{\vartheta}_1$ zulässig ist.

1.2 Minimax- und Bayes-Ansatz

1.7 Bemerkung. Da das Risiko $R(\vartheta, \rho)$ einer Entscheidungsregel ρ im Allgemeinen vom unbekannten wahren Parameter ϑ abhängt, werden Entscheidungsregeln üblicherweise gemäß ihrem maximalen Risiko in ϑ oder einem geeignet über ϑ gemittelten Risiko beurteilt.

1.8 Definition. Eine Entscheidungsregel ρ heißt minimax, falls

$$\sup_{\vartheta \in \Theta} R(\vartheta, \rho) = \inf_{\rho'} \sup_{\vartheta \in \Theta} R(\vartheta, \rho'),$$

wobei sich das Infimum über alle Entscheidungsregeln ρ' erstreckt.

1.9 Definition. Der Parameterraum Θ trage die σ-Algebra $\mathscr{F}_\Theta$, die Verlustfunktion l sei produktmessbar und $\vartheta \mapsto \mathbb{P}_\vartheta(B)$ sei messbar für alle $B \in \mathscr{F}$. Die a-priori-Verteilung π des Parameters ϑ ist gegeben durch ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathscr{F}_\Theta)$. Das zu π assoziierte Bayesrisiko einer Entscheidungsregel ρ ist

$$R_\pi(\rho) := \mathbb{E}_\pi[R(T, \rho)] = \int_\Theta R(\vartheta, \rho)\, \pi(d\vartheta) = \int_\Theta \int_{\mathcal{X}} l(\vartheta, \rho(x))\ \mathbb{P}_\vartheta(dx)\, \pi(d\vartheta).$$

ρ heißt Bayesregel oder Bayes-optimal (bezüglich π), falls

$$R_\pi(\rho) = \inf_{\rho'} R_\pi(\rho')$$

gilt, wobei sich das Infimum über alle Entscheidungsregeln ρ' erstreckt.

1.10 Definition. Es sei X eine $(S, \mathscr{S})$-wertige Zufallsvariable auf $(\Omega, \mathscr{F}, \mathbb{P})$. Eine Abbildung $K : S \times \mathscr{F} \to [0, 1]$ heißt reguläre bedingte Wahrscheinlichkeit oder Markovkern bezüglich X, falls

(a) $A \mapsto K(x, A)$ ist Wahrscheinlichkeitsmaß für alle $x \in S$;

(b) $x \mapsto K(x, A)$ ist messbar für alle $A \in \mathscr{F}$;

(c) $K(X, A) = \mathbb{P}(A \,|\, X) := \mathbb{E}[\mathbf{1}_A \,|\, X]$ $\mathbb{P}$-f.s. für alle $A \in \mathscr{F}$.

1.11 Satz. *Es sei (Ω, d) ein vollständiger, separabler Raum mit Metrik d und Borel-σ-Algebra $\mathscr{F}$ (polnischer Raum). Für jede Zufallsvariable X auf $(\Omega, \mathscr{F}, \mathbb{P})$ existiert eine reguläre bedingte Wahrscheinlichkeit K bezüglich X. K ist $\mathbb{P}$-f.s. eindeutig bestimmt, d.h. für eine zweite solche reguläre bedingte Wahrscheinlichkeit K' gilt $\mathbb{P}(\forall A \in \mathscr{F} : \ K(X, A) = K'(X, A)) = 1$.*

Beweis. Siehe z.B. Gänssler, Stute (1977): Wahrscheinlichkeitstheorie, Springer. □

1.12 Bemerkung. Während eine Minimaxregel den maximal zu erwartenden Verlust minimiert, kann das Bayesrisiko als ein (mittels π) gewichtetes Mittel der zu erwartenden Verluste angesehen werden. Alternativ wird π als die subjektive Einschätzung der Verteilung des zugrundeliegenden Parameters interpretiert. Daher wird das Bayesrisiko auch als insgesamt zu erwartender Verlust in folgendem Sinne verstanden.

1.13 Definition. Definiere $\Omega := \mathcal{X} \times \Theta$ und $\tilde{\mathbb{P}}$ auf $(\Omega, \mathscr{F} \otimes \mathscr{F}_\Theta)$ gemäß

$$\tilde{\mathbb{P}}(A \times B) := \iint \mathbf{1}_{A \times B}(x, \vartheta)\, \mathbb{P}_\vartheta(dx)\pi(d\vartheta) = \int_B \mathbb{P}_\vartheta(A)\pi(d\vartheta), \quad A \in \mathscr{F},\ B \in \mathscr{F}_\Theta,$$

und Fortsetzung auf $\mathscr{F} \otimes \mathscr{F}_\Theta$ (gemeinsame Verteilung von Beobachtung und Parameter), wobei π eine a-priori-Verteilung auf $\mathscr{F}_\Theta$ und $(\vartheta, A) \mapsto \mathbb{P}_\vartheta(A)$ ein Markovkern sei. Bezeichne mit X und T die Koordinatenprojektionen von Ω auf $\mathcal{X}$ bzw. Θ. Dann gilt $R_\pi(\rho) = \mathbb{E}_{\tilde{\mathbb{P}}}[l(T, \rho(X))]$.

Die Verteilung von T unter der regulären bedingten Wahrscheinlichkeit $\tilde{\mathbb{P}}(\bullet \,|\, X = x)$ von $\tilde{\mathbb{P}}$ heißt a-posteriori-Verteilung des Parameters gegeben die Beobachtung $X = x$.

1.14 Satz. *(Bayesformel) Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell sowie π eine a-priori-Verteilung auf $(\Theta, \mathscr{F}_\Theta)$, so dass $\mathbb{P}_\vartheta$ für alle $\vartheta \in \Theta$ μ-Dichten $f^{X|T=\vartheta}$ sowie π eine ν-Dichte f^T besitzt mit entsprechenden Maßen μ und ν. Ist $f^{X\,|\,T=\bullet} : \mathcal{X} \times \Theta \to \mathbb{R}^+$ $(\mathscr{F} \otimes \mathscr{F}_\Theta)$-messbar, so besitzt die a-posteriori-Verteilung $\mathbb{P}^{T\,|\,X=x}$ des Parameters für $\tilde{\mathbb{P}}^X$-fast alle $x \in \mathcal{X}$ eine ν-Dichte, nämlich*

$$f^{T\,|\,X=x}(\vartheta) = \frac{f^{X|T=\vartheta}(x) f^T(\vartheta)}{f^X(x)} \ \text{mit}\ f^X(x) := \int_\Theta f^{X|T=\vartheta'}(x) f^T(\vartheta')\nu(d\vartheta').$$

Beweis. Übung! □

1.15 Beispiele.

(a) Für einen Bayestest (oder auch ein Bayes-Klassifikationsproblem) setze $\Theta = \{0, 1\}$ und betrachte eine a-priori-Verteilung π mit $\pi(\{0\}) =: \pi_0$, $\pi(\{1\}) =: \pi_1$. Die Wahrscheinlichkeitsmaße $\mathbb{P}_0, \mathbb{P}_1$ auf $(\mathcal{X}, \mathscr{F})$ mögen die Dichten p_0, p_1 bezüglich einem Maß μ besitzen (z.B. $\mu = \mathbb{P}_0 + \mathbb{P}_1$). Nach der Bayesformel (mit Zählmaß ν) erhalten wir die a-posteriori-Verteilung

$$\tilde{\mathbb{P}}(T = i \,|\, X = x) = \frac{\pi_i p_i(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}, \quad i = 0, 1 \quad \tilde{\mathbb{P}}^X\text{-f.ü.}$$

(b) Es sei $X_1, \dots, X_n$ eine $N(\mu, E_d)$-verteilte mathematische Stichprobe im $\mathbb{R}^d$ und $\pi = N(a, \sigma^2 E_d)$ eine a-priori-Verteilung für $\mu \in \mathbb{R}^d$ mit $a \in \mathbb{R}^d$, $\sigma > 0$. Dann liefert die Bayesformel bezüglich Lebesguemaß und mit $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$:

$$\begin{aligned} f^{T\,|\,X=x}(\mu) &\propto f^{X|T=\mu}(x) f^T(\mu) \\ &\propto \exp\Big(-\frac{|\mu - a|^2}{2\sigma^2} - \frac{1}{2}\sum_{i=1}^n |x_i - \mu|^2 \Big) \\ &\propto \exp\Big(\langle \mu, \sigma^{-2} a + n\bar{x}\rangle - \tfrac{1}{2}|\mu|^2\big(\sigma^{-2} + n\big) \Big) \\ &\propto \exp\Big(-\tfrac{1}{2}\big(\sigma^{-2} + n\big)|\mu - \tfrac{a + n\sigma^2\bar{x}}{1 + n\sigma^2}|^2 \Big). \end{aligned}$$

Die a-posteriori-Verteilung ist also wiederum eine Normalverteilung:

$$\tilde{\mathbb{P}}^{T\,|\,X=x} = N\Big(\frac{a + n\sigma^2\bar{x}}{1 + n\sigma^2}, \frac{\sigma^2}{1 + n\sigma^2} E_d\Big).$$

Beachte, dass für großen Stichprobenumfang n oder große a-priori-Varianz σ^2 sich die a-posteriori-Verteilung um das Stichprobenmittel konzentriert, während sie für sehr kleine a-priori-Varianz und geringen Stichprobenumfang, so dass $n\sigma^2 \ll 1$, nahe bei der a-priori-Verteilung bleibt.

1.16 Satz. *Eine Regel ρ ist Bayes-optimal, falls gilt*

$$\rho(X) \in \operatorname{argmin}_{a \in A} \mathbb{E}_{\tilde{\mathbb{P}}}[l(T, a)\,|\,X] \quad \tilde{\mathbb{P}}\text{-}f.s.,$$

d.h. $\mathbb{E}_{\tilde{\mathbb{P}}}[l(T, \rho(x))\,|\,X = x] \leqslant \mathbb{E}_{\tilde{\mathbb{P}}}[l(T, a)\,|\,X = x]$ für alle $a \in A$ und $\tilde{\mathbb{P}}^X$-fast alle $x \in \mathcal{X}$.

Beweis. Für eine beliebige Entscheidungsregel ρ' gilt

$$R_\pi(\rho') = \mathbb{E}_{\tilde{\mathbb{P}}}[\mathbb{E}_{\tilde{\mathbb{P}}}[l(T, \rho'(X))\,|\,X]] \geqslant \mathbb{E}_{\tilde{\mathbb{P}}}[\mathbb{E}_{\tilde{\mathbb{P}}}[l(T, \rho(X))\,|\,X]] = R_\pi(\rho).$$

□

1.17 Korollar. *Für $\Theta \subseteq \mathbb{R}^d$, $A = \mathbb{R}^d$ und quadratisches Risiko (d.h. $l(\vartheta, a) = |a - \vartheta|^2$) ist die (vektorwertige) bedingte Erwartung $\hat{\vartheta}_\pi := \mathbb{E}_{\tilde{\mathbb{P}}}[T\,|\,X]$ Bayes-optimaler Schätzer von ϑ bezüglich der a-priori-Verteilung π, sofern $T \in L^2(\tilde{P})$ gilt. Für $d = 1$ und den Absolutbetrag $l(\vartheta, a) = |\vartheta - a|$ ist jeder a-posteriori-Median $\hat{\vartheta}_\pi$, d.h. $\tilde{\mathbb{P}}(T \leqslant \hat{\vartheta}_\pi\,|\,X) \geqslant 1/2$ und $\tilde{\mathbb{P}}(T \geqslant \hat{\vartheta}_\pi\,|\,X) \geqslant 1/2$, Bayes-optimaler Schätzer (Annahme: a-posteriori-Verteilung existiert).*

Beweis. Dies folgt aus der L^2-Projektionseigenschaft der bedingten Erwartung bzw. der L^1-Minimierung des Medians, vgl. Stochastik I, II oder Übung. □

1.18 Beispiele. (Fortsetzung)

(a) Nach Satz 1.16 finden wir einen Bayestest $\varphi(x)$ für den 0-1-Verlust $l(\vartheta, a) = \mathbf{1}(a \neq \vartheta)$ als Minimalstelle von

$$a \mapsto \mathbb{E}_{\tilde{\mathbb{P}}}[l(T,a) \,|\, X = x] = \frac{\pi_0 p_0(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)} a + \frac{\pi_1 p_1(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)} (1-a).$$

Daher ist ein Bayestest (Bayesklassifizierer) gegeben durch

$$\varphi(x) = \begin{cases} 0, & \pi_0 p_0(x) > \pi_1 p_1(x) \\ 1, & \pi_1 p_1(x) > \pi_0 p_0(x) \\ \text{beliebig}, & \pi_0 p_0(x) = \pi_1 p_1(x) \end{cases}$$

und wir entscheiden uns für dasjenige $\vartheta \in \{0,1\}$, dessen a-posteriori-Wahrscheinlichkeit am größten ist ("MAP-estimator: maximum a posteriori estimator"). Für später sei bereits auf die Neyman-Pearson-Struktur von φ in Abhängigkeit von $p_1(x)/p_0(x)$ hingewiesen.

(b) Nach Korollar 1.17 ist ein der Bayesschätzer unter quadratischem Risiko für $X_1, \ldots, X_n \sim N(\mu, E_d)$ und $\pi = N(a, \sigma^2 E_d)$ gegeben durch die bedingte Erwartung

$$\hat{\mu}_{a,\sigma^2} = \mathbb{E}_{\tilde{\mathbb{P}}}[T \,|\, X] = \frac{a + n\sigma^2 \bar{X}}{1 + n\sigma^2},$$

wie sofort aus der Normalverteilung der a-posteriori-Verteilung folgt. Man beachte, dass $\hat{\mu}_{a,\sigma^2}$ eine Konvexkombination vom a-priori-Mittelwert a und dem Stichprobenmittel $\bar{X}$ ist.

1.19 Lemma. *Es liege die Situation aus Definition 1.9 vor. Für jede Entscheidungsregel ρ gilt*

$$\sup_{\vartheta \in \Theta} R(\vartheta, \rho) = \sup_{\pi} R_\pi(\rho),$$

wobei sich das zweite Supremum über alle a-priori-Verteilungen π erstreckt. Insbesondere ist das Risiko einer Bayesregel stets kleiner oder gleich dem Minimaxrisiko.

Beweis. Natürlich gilt $R_\pi(\rho) = \int_\Theta R(\vartheta, \rho)\, \pi(d\vartheta) \leqslant \sup_{\vartheta \in \Theta} R(\vartheta, \rho)$. Durch Betrachtung der a-priori-Verteilungen δ_ϑ (Diracmaß im Punkt $\vartheta \in \Theta$) folgt daher die Behauptung. □

1.20 Bemerkung. Man kann dieses Lemma insbesondere dazu verwenden, untere Schranken für das Minimax-Risiko durch das Bayesrisiko abzuschätzen.

1.21 Satz. *Für jede Entscheidungsregel ρ gilt:*

(a) Ist ρ minimax und eindeutig in dem Sinn, dass jede andere Minimax-Regel die gleiche Risikofunktion besitzt, so ist ρ zulässig.

(b) Ist ρ zulässig mit konstanter Risikofunktion, so ist ρ minimax.

(c) Ist ρ eine Bayesregel (bzgl. π) und eindeutig in dem Sinn, dass jede andere Bayesregel (bzgl. π) die gleiche Risikofunktion besitzt, so ist ρ zulässig.

(d) *Die Parametermenge Θ bilde einen metrischen Raum mit Borel-σ-Algebra $\mathscr{F}_\Theta$. Ist ρ eine Bayesregel (bzgl. π), so ist ρ zulässig, falls (i) $R_\pi(\rho) < \infty$; (ii) für jede nichtleere offene Menge U in Θ gilt $\pi(U) > 0$; (iii) für jede Regel ρ' mit $R_\pi(\rho') \leqslant R_\pi(\rho)$ ist $\vartheta \mapsto R(\vartheta, \rho')$ stetig.*

Beweis. Übung! □

1.22 Satz. *Es sei $X_1, \dots, X_n$ eine $N(\mu, E_d)$-verteilte d-dimensionale mathematische Stichprobe mit $\mu \in \mathbb{R}^d$ unbekannt. Bezüglich quadratischem Risiko ist das arithmetische Mittel $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ minimax als Schätzer von μ.*

1.23 Bemerkung. Die Beweisidee ist, dass $\bar{X}$ ein sogenannter "improper Bayes"-Schätzer ist mit dem Lebesguemaß als a-priori-Verteilung. Dies wird mit einem Grenzwertargument formal umgesetzt.

Beweis. Zunächst beachte, dass $\bar{X} - \mu \sim N(0, \frac{1}{n}E_d)$ gilt, so dass

$$R(\mu, \bar{X}) = \sum_{i=1}^d \mathbb{E}_\mu[(\bar{X}_i - \mu_i)^2] = \frac{d}{n}$$

folgt. Betrachte nun die a-priori-Verteilung $\pi = N(0, \sigma^2 E_d)$ für μ. Gemäß Beispiel 1.18 ist der Bayes-optimale Schätzer $\hat{\mu}_{\sigma,n} = \frac{n\sigma^2}{1+n\sigma^2}\overline{X}$. Seine Risikofunktion ist (gemäß Bias-Varianz-Zerlegung)

$$\begin{aligned} R(\mu, \hat{\mu}_{\sigma,n}) &= (\mathbb{E}_\mu[\hat{\mu}_{\sigma,n}] - \mu)^2 + \operatorname{Var}_\mu(\hat{\mu}_{\sigma,n}) \\ &= \Big(\frac{1}{1+n\sigma^2}\Big)^2 |\mu|^2 + \Big(\frac{n\sigma^2}{1+n\sigma^2}\Big)^2 \mathbb{E}[|\bar{X} - \mu|^2] \\ &= \frac{|\mu|^2 + nd\sigma^4}{(1+n\sigma^2)^2}. \end{aligned}$$

Somit können wir das Minimax-Risiko von unten abschätzen:

$$\begin{aligned} \inf_\rho \sup_\mu R(\mu, \rho) &= \inf_\rho \sup_\pi R_\pi(\rho) \\ &\geqslant \inf_\rho \sup_{\sigma>0} R_{N(0,\sigma^2 E_d)}(\rho) \\ &\geqslant \sup_{\sigma>0} \inf_\rho R_{N(0,\sigma^2 E_d)}(\rho) \\ &= \sup_{\sigma>0} \mathbb{E}_\pi \Big[\frac{|\mu|^2 + nd\sigma^4}{(1+n\sigma^2)^2}\Big] \\ &= \sup_{\sigma>0} \frac{d\sigma^2 + nd\sigma^4}{(1+n\sigma^2)^2} = \sup_{\sigma>0} \frac{d\sigma^2}{1+n\sigma^2} = \frac{d}{n}, \end{aligned}$$

wie behauptet.
Anmerkung: da die bedingte Kovarianzmatrix $\operatorname{Var}_{\tilde{P}}(T \mid X) = \frac{\sigma^2}{1+n\sigma^2}E_d$ (s.o.) nicht von X abhängt, ergibt sich das Bayesrisiko alternativ auch direkt aus

$$R_{N(0,\sigma^2 E_d)}(\hat{\mu}_{\sigma,n}) = \mathbb{E}_{\tilde{P}}[|\mathbb{E}[T \mid X] - T|^2] = \sum_{i=1}^d \mathbb{E}_{\tilde{P}}[\operatorname{Var}_{\tilde{P}}(T_i \mid X)] = \frac{d\sigma^2}{1+n\sigma^2}.$$

□

1.24 Satz. *Es sei $X_1, \dots, X_n$ eine $N(\mu, 1)$-verteilte skalare mathematische Stichprobe mit $\mu \in \mathbb{R}$ unbekannt. Bezüglich quadratischem Risiko ist das arithmetische Mittel $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ zulässig als Schätzer von μ.*

Beweis. Gäbe es einen Schätzer $\hat{\mu}$ mit $R(\mu, \hat{\mu}) \leqslant \frac{1}{n}$ und $R(\mu_0, \hat{\mu}) < \frac{1}{n}$ für ein $\mu_0 \in \mathbb{R}$, so wäre wegen Stetigkeit der Risikofunktion $\mu \mapsto R(\mu, \hat{\mu})$ (Übung!) sogar $R(\mu, \hat{\mu}) \leqslant \frac{1}{n} - \varepsilon$ für alle $|\mu - \mu_0| < \delta$ mit $\varepsilon, \delta > 0$ geeignet. Damit hätte $\hat{\mu}$ ein Bayesrisiko $R_{N(0,\sigma^2)}(\hat{\mu}) \leqslant \frac{1}{n} - \varepsilon \int_{\mu_0-\delta}^{\mu_0+\delta} \varphi_{0,\sigma^2}$. Also wäre für $\sigma \to \infty$

$$\frac{1}{n} - R_{N(0,\sigma^2)} \geqslant \frac{2\varepsilon\delta}{\sigma\sqrt{2\pi}} \exp\Big(- \big((\mu_0 - \delta) \vee (\mu_0 + \delta)\big)^2/(2\sigma^2)\Big) \asymp \frac{2\varepsilon\delta}{\sigma\sqrt{2\pi}}$$

größer als ein Vielfaches von σ^{-1}, während für den Bayesschätzer (siehe oben)

$$\frac{1}{n} - R_{N(0,\sigma^2)}(\hat{\mu}_{\sigma,n}) = \frac{1}{n} - \frac{\sigma^2}{1 + n\sigma^2} = \frac{\sigma^{-2}}{n(n + \sigma^{-2})}$$

von der Ordnung σ^{-2} ist. Dies widerspricht der Optimalität des Bayesschätzers bei einer hinreichend großen Wahl von σ. Also ist $\bar{X}$ zulässig. □

1.25 Bemerkung. Liegt eine andere Verteilung mit Erwartungswert μ und Varianz eins vor als die Normalverteilung, so ist $\bar{X}$ weder zulässig noch minimax (sofern $n \geqslant 3$), vergleiche Lehmann/Casella, Seite 153. Für $d = 2$ ist $\bar{X}$ weiterhin zulässig unter Normalverteilungsannahme, allerdings gilt das für $d \geqslant 3$ nicht mehr: Stein-Phänomen s.u.

1.26 Definition. Eine Verteilung π auf $(\Theta, \mathscr{F}_\Theta)$ heißt ungünstigste a-priori-Verteilung zu einer gegebenen Verlustfunktion, falls

$$\inf_\rho R_\pi(\rho) = \sup_{\pi'} \inf_\rho R_{\pi'}(\rho).$$

1.27 Satz. *Es sei eine a-priori-Verteilung π mit zugehöriger Bayesregel ρ_π gegeben. Dann ist die Eigenschaft $R_\pi(\rho_\pi) = \sup_{\vartheta\in\Theta} R(\vartheta, \rho_\pi)$ äquivalent zu folgender Sattelpunktseigenschaft*

$$\forall \pi' \; \forall \rho' : \; R_{\pi'}(\rho_\pi) \leqslant R_\pi(\rho_\pi) \leqslant R_\pi(\rho').$$

Aus jeder dieser Eigenschaften folgt, dass ρ_π minimax und π ungünstigste a-priori-Verteilung ist.

Beweis. Wegen $\sup_\vartheta R(\vartheta, \rho_\pi) = \sup_{\pi'} R_{\pi'}(\rho_\pi)$ folgt aus der Sattelpunktseigenschaft $R_\pi(\rho_\pi) \geqslant \sup_\vartheta R(\vartheta, \rho_\pi)$. Da in jedem Fall '$\leqslant$' gilt, folgt $R_\pi(\rho_\pi) = \sup_\vartheta R(\vartheta, \rho_\pi)$.

Andererseits bedeutet die Eigenschaft von ρ_π, Bayesschätzer zu sein, gerade dass $R_\pi(\rho_\pi) \leqslant R_\pi(\rho')$ für alle ρ' gilt. Mit $R_\pi(\rho_\pi) = \sup_{\vartheta\in\Theta} R(\vartheta, \rho_\pi)$ schließen wir dann auch

$$R_{\pi'}(\rho_\pi) = \int_\Theta R(\vartheta, \rho_\pi)\, \pi'(d\vartheta) \leqslant \int_\Theta R_\pi(\rho_\pi)\, \pi'(d\vartheta) = R_\pi(\rho_\pi).$$

Aus der Sattelpunktseigenschaft folgt direkt die Minimaxeigenschaft:

$$\sup_{\vartheta} R(\vartheta, \rho_\pi) = \sup_{\pi'} R_{\pi'}(\rho_\pi) = \inf_{\rho'} R_\pi(\rho') \leqslant \inf_{\rho'} \sup_{\vartheta} R(\vartheta, \rho').$$

Analog erhalten wir $\inf_{\rho'} R_\pi(\rho') = \sup_{\pi'} R_{\pi'}(\rho_\pi) \geqslant \sup_{\pi'} \inf_\rho R_{\pi'}(\rho)$, so dass π ungünstigste a-priori-Verteilung ist.

□

1.28 Beispiel. Es werde $X \sim \mathrm{Bin}(n,p)$ mit $n \geqslant 1$ bekannt und $p \in [0,1]$ unbekannt beobachtet. Gesucht wird ein Bayesschätzer $\hat{p}_{a,b}$ von p unter quadratischem Risiko für die a-priori-Verteilung $p \sim B(a,b)$, wobei $B(a,b)$ die Beta-Verteilung mit Parametern $a,b > 0$ auf $[0,1]$ bezeichnet. Die a-posteriori-Verteilung berechnet sich zu $p \sim B(a+X, b+n-X)$ und der Bayesschätzer als $\hat{p}_{a,b} = \frac{a+X}{a+b+n}$ (Übung!). Als Risiko ergibt sich $\mathbb{E}_p[(\hat{p}_{a,b}-p)^2] = \frac{(a-ap-bp)^2+np(1-p)}{(a+b+n)^2}$. Im Fall $a^* = b^* = \sqrt{n}/2$ erhält man das Risiko $(2\sqrt{n}+2)^{-2}$ für $\hat{p}_{a^*,b^*} = \frac{X+\sqrt{n}/2}{n+\sqrt{n}} = \frac{X}{n} - \frac{X-\frac{n}{2}}{n(\sqrt{n}+1)}$ (unabhängig von p!), woraus die Sattelpunktseigenschaft folgt:

$$\forall \pi \; \forall \hat{p} : \; R_\pi(\hat{p}_{a^*,b^*}) \leqslant R_{B(a^*,b^*)}(\hat{p}_{a^*,b^*}) \leqslant R_{B(a^*,b^*)}(\hat{p}).$$

Damit ist $B(a^*,b^*)$ ungünstigste a-priori-Verteilung und $\hat{p}_{a^*,b^*}$ Minimax-Schätzer von p. Insbesondere ist der natürliche Schätzer $\hat{p} = X/n$ mit $\mathbb{E}_p[(\hat{p}-p)^2] = p(1-p)/n$ nicht minimax (er ist jedoch zulässig).

1.29 Bemerkung. Erhalten wir bei Wahl einer Klasse von a-priori-Verteilungen für ein statistisches Modell dieselbe Klasse (i.A. mit anderen Parametern) als a-posteriori-Verteilungen zurück, so nennt man die entsprechenden Verteilungsklassen konjugiert. An den Beispielen sehen wir, dass die Beta-Verteilungen zur Binomialverteilung konjugiert sind und die Normalverteilungen zu den Normalverteilungen (genauer müsste man spezifizieren, dass für unbekannten Mittelwert in der Normalverteilung a-priori-Normalverteilungen konjugiert sind). Konjugierte Verteilungen sind die Ausnahme, nicht die Regel, und für komplexere Modelle werden häufig computer-intensive Methoden wie MCMC (Markov Chain Monte Carlo) verwendet, um die a-posteriori-Verteilung zu berechnen (Problem: i.A. hochdimensionale Integration).

1.3 Das Stein-Phänomen

Wir betrachten folgendes grundlegendes Problem: Anhand einer mathematischen Stichprobe $X_1, \ldots, X_n \sim N(\mu, E_d)$ im $\mathbb{R}^d$ soll $\mu \in \mathbb{R}^d$ möglichst gut bezüglich quadratischem Verlust $l(\mu, \hat{\mu}) = |\hat{\mu} - \mu|^2$ geschätzt werden. Intuitiv wegen Unabhängigkeit der Koordinaten ist das (koordinatenweise) arithmetische Mittel $\bar{X}$. Ein anderer, sogenannter empirischer Bayesansatz, beruht auf der Familie der a-priori-Verteilungen $\mu \sim N(0, \sigma^2 E_d)$. In den zugehörigen Bayesschätzern setzen wir dann allerdings statt σ^2 die Schätzung

$$\hat{\sigma}^2 = \frac{|\bar{X}|^2}{d} - n^{-1} \text{ (erwartungstreu wegen } X_i \sim N(0, (\sigma^2 + n^{-1})E_d) \text{ unter } \tilde{\mathbb{P}})$$

ein und erhalten

$$\hat{\mu} = \Big(1 - \frac{1}{1 + n\hat{\sigma}^2}\Big)\bar{X} = \Big(1 - \frac{d}{n|\bar{X}|^2}\Big)\bar{X}.$$

Der Ansatz lässt vermuten, dass $\hat{\mu}$ kleineres Risiko hat als $\bar{X}$, wann immer $|\mu|$ klein ist. Überraschenderweise gilt für Dimension $d \geqslant 3$ sogar, dass $\hat{\mu}$ besser ist als $\bar{X}$. Das folgende Steinsche Lemma ist der Schlüssel für den Beweis.

1.30 Lemma (Stein)**.** *Es sei $f : \mathbb{R}^d \to \mathbb{R}$ eine Funktion, die Lebesgue-f.ü. absolut stetig in jeder Koordinate ist. Dann gilt für $X \sim N(\mu, \sigma^2 E_d)$ mit $\mu \in \mathbb{R}^d$, $\sigma > 0$,*

$$\mathbb{E}[(X - \mu)f(X)] = \sigma^2\, \mathbb{E}[\nabla f(X)],$$

sofern $\mathbb{E}[|\frac{\partial f}{\partial x_i}(X)|] < \infty$ für alle $i = 1, \dots, d$ gilt.

Beweis. Ohne Einschränkung der Allgemeinheit betrachte die Koordinate $i = 1$ sowie $\mu = 0$, $\sigma = 1$; sonst setze $\tilde{f}(x) = f(\sigma x + \mu)$. Es genügt dann,

$$\mathbb{E}[X_1 f(X) \,|\, X_2 = x_2, \dots, X_d = x_d] = \mathbb{E}[\tfrac{\partial f}{\partial x_1}(X) \,|\, X_2 = x_2, \dots, X_d = x_d]$$

zu zeigen für Lebesgue-fast alle $x_2, \dots, x_d \in \mathbb{R}$, was wegen Unabhängigkeit gerade für $f_x(u) := f(u, x_2, \dots, x_d)$ die Identität $\int u f_x(u) e^{-u^2/2} du = \int f_x'(u) e^{-u^2/2} du$ ist. Dies folgt durch partielle Integration, sofern die Randterme verschwinden; ein geschickter Einsatz des Satzes von Fubini zeigt dies jedoch ohne weitere Voraussetzungen:

$$\begin{aligned}
\int_{-\infty}^{\infty} f_x'(u) e^{-u^2/2} du &= \int_0^{\infty} f_x'(u) \int_u^{\infty} z e^{-z^2/2} dz du - \int_{-\infty}^{0} f_x'(u) \int_{-\infty}^{u} z e^{-z^2/2} dz du \\
&= \int_0^{\infty} \Big(\int_0^z f_x' \Big) z e^{-z^2/2} dz - \int_{-\infty}^{0} \Big(\int_z^0 f_x' \Big) z e^{-z^2/2} dz \\
&= \int_{-\infty}^{\infty} z e^{-z^2/2} (f_x(z) - f_x(0))\, dz \\
&= \int_{-\infty}^{\infty} f_x(z) z e^{-z^2/2} dz.
\end{aligned}$$

Die Anwendung von Fubini in der zweiten Zeile wird gerechtfertigt durch dieselbe Rechnung mit $|f_x'|$ statt f_x', da nach Voraussetzung $\iint |f_x'(u)| z e^{-z^2/2} dz du$ endlich ist. □

Betrachten wir nun allgemeine Schätzer der Form $\hat{\mu} = \bar{X} - f(\bar{X})$, so gilt

$$\mathbb{E}_\mu[|\hat{\mu} - \mu|^2] = \mathbb{E}_\mu\Big[|\bar{X} - \mu|^2 + |f(\bar{X})|^2 - 2\langle \bar{X} - \mu, f(\bar{X})\rangle\Big].$$

Kann man nun auf $f = (f_1, \dots, f_d) : \mathbb{R}^d \to \mathbb{R}^d$ das Steinsche Lemma koordinatenweise anwenden, so erhalten wir einen Ausdruck $W(\bar{X})$ unabhängig von μ:

$$\mathbb{E}_\mu[|\hat{\mu} - \mu|^2] = \frac{d}{n} + \mathbb{E}_\mu[W(\bar{X})], \quad W(x) := |f(x)|^2 - \frac{2}{n} \sum_{i=1}^{d} \frac{\partial f_i(x)}{\partial x_i}.$$

Für $f(x) = \frac{cx}{|x|^2}$, $c > 0$ eine Konstante, ist das Steinsche Lemma anwendbar. Wir erhalten

$$\sum_{i=1}^{d} \frac{\partial f_i(x)}{\partial x_i} = c \sum_{i=1}^{d} \frac{|x|^2 - 2x_i^2}{|x|^4} = c(d-2)|x|^{-2}$$

und

$$W(x) = \frac{c^2}{|x|^2} - \frac{2c(d-2)}{n|x|^2} < 0 \text{ falls } c \in (0, 2(d-2)n^{-1}),\ d \geqslant 3.$$

Beachte, dass $f(x) = \frac{2(d-2)x}{n|x|^2}$ gerade $W(x) = 0$ löst, was a posteriori den Ansatz für f plausibel macht. Der minimale Wert $W(x) = -(d-2)^2/(n^2|x|^2)$ wird für $c = (d-2)/n$ erreicht, und wir haben folgendes bemerkenswertes Resultat bewiesen.

1.31 Satz. *Es sei $d \geqslant 3$ und $X_1, \ldots, X_n$ eine $N(\mu, E_d)$-verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}^d$ unbekannt. Dann gilt für den James-Stein-Schätzer*

$$\hat{\mu}_{JS} := \Big(1 - \frac{d-2}{n|\bar{X}|^2}\Big)\bar{X}$$

mit $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$, dass

$$\mathbb{E}_\mu[|\hat{\mu}_{JS} - \mu|^2] = \frac{d}{n} - \mathbb{E}_\mu\Big[\frac{(d-2)^2}{n^2|\bar{X}|^2}\Big] < \frac{d}{n} = \mathbb{E}_\mu[|\bar{X} - \mu|^2].$$

Insbesondere ist $\bar{X}$ bei quadratischem Risiko kein zulässiger Schätzer von μ im Fall $d \geqslant 3$!

1.32 Bemerkungen.

(a) Die Abbildung $\mu \mapsto \mathbb{E}_\mu[|\bar{X}|^{-2}]$ ist monoton fallend in $|\mu|$ und erfüllt $\mathbb{E}_0[|\bar{X}|^{-2}] = n/(d-2)$, $\mathbb{E}_0[|\hat{\mu}_{JS} - \mu|^2] = 2/n$. Daher ist $\hat{\mu}_{JS}$ nur für μ nahe 0, große Dimensionen d und kleine Stichprobenumfänge n eine bedeutende Verbesserung von $\bar{X}$. Der James-Stein-Schätzer heißt auch Shrinkage-Schätzer, weil er die Beobachtungen zur Null hinzieht (wobei auch jeder andere Wert möglich wäre). In aktuellen hochdimensionalen Problemen findet diese Idee breite Anwendung.

(b) Die k-te Koordinate $\hat{\mu}_{JS,k}$ des James-Stein-Schätzers verwendet zur Schätzung von μ_k auch die anderen Koordinaten $X_{i,l}$, $l \neq k$, obwohl diese unabhängig von $X_{i,k}$ sind. Eine Erklärung für diese zunächst paradoxe Situation ist, dass zwar $\sum_{k=1}^{d} \mathbb{E}_\mu[(\hat{\mu}_{JS,k} - \mu_k)^2] < \sum_{k=1}^{d} \mathbb{E}_\mu[(\bar{X}_k - \mu_k)^2]$ gilt, jedoch im Allgemeinen eine Koordinate k_0 existieren wird mit $\mathbb{E}_\mu[(\hat{\mu}_{JS,k_0} - \mu_{k_0})^2] > \mathbb{E}_\mu[(\bar{X}_{k_0} - \mu_{k_0})^2]$. Man beachte auch, dass der stochastische Fehler (die Varianz) von $\bar{X}$ linear mit der Dimension d wächst, so dass es sich auszahlt, diesen Fehler auf Kosten einer Verzerrung (Bias) zu verringern, vgl. Übung.

(c) Selbst der James-Stein-Schätzer (sogar mit positivem Gewicht, s.u.) ist unzulässig. Die Konstruktion eines zulässigen Minimax-Schätzers ist sehr schwierig (gelöst für $d \geqslant 6$, vgl. Lehmann/Casella, S. 358).

1.33 Satz. *Es sei $d \geqslant 3$ und $X_1, \dots, X_n$ eine $N(\mu, E_d)$-verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}^d$ unbekannt. Dann ist der James-Stein-Schätzer mit positivem Gewicht*

$$\hat{\mu}_{JS+} := \Big(1 - \frac{d-2}{n|\bar{X}|^2}\Big)_+ \bar{X}, \qquad a_+ := \max(a, 0),$$

bei quadratischem Risiko besser als der James-Stein-Schätzer $\hat{\mu}_{JS}$.

1.4 Ergänzungen

1.34 Definition. Ein Entscheidungskern oder eine randomisierte Entscheidungsregel $\rho : \mathcal{X} \times \mathscr{A} \to [0,1]$ ist ein Markovkern auf dem Aktionsraum $(A, \mathscr{A})$ mit der Interpretation, dass bei Vorliegen der Beobachtung x gemäß $\rho(x, \bullet)$ eine Entscheidung zufällig ausgewählt wird. Das zugehörige Risiko ist

$$R(\vartheta, \rho) := \mathbb{E}_\vartheta \Big[\int_A l(\vartheta, a)\, \rho(da) \Big] = \int_{\mathcal{X}} \int_A l(\vartheta, a) \rho(x, da)\, \mathbb{P}_\vartheta(dx).$$

1.35 Beispiel. Es sei $\Theta = \Theta_0 \dot\cup \Theta_1$, $A = [0,1]$ und der Verlust $l(\vartheta, a) = l_0 a \mathbf{1}_{\Theta_0}(\vartheta) + l_1(1-a)\mathbf{1}_{\Theta_1}(\vartheta)$ vorgegeben. In diesem Rahmen kann eine Entscheidungsregel ρ als randomisierter Test (oder Entscheidungskern) ρ' von $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$ aufgefasst werden. Dazu setze $A' := \{0,1\}$, $\mathscr{F}_{A'} := \mathcal{P}(A')$, benutze den gleichen Verlust l (eingeschränkt auf A') und definiere die bedingten Wahrscheinlichkeiten $\rho'(x, \{1\}) := \rho(x)$, $\rho'(x, \{0\}) := 1 - \rho'(x, \{1\})$. Dies bedeutet also, dass $\rho(x)$ die Wahrscheinlichkeit angibt, mit der bei der Beobachtung x die Hypothese abgelehnt wird.

1.36 Lemma. *Es sei $A \subseteq \mathbb{R}^d$ konvex sowie $l(\vartheta, a)$ eine im zweiten Argument konvexe Verlustfunktion. Dann gibt es zu jeder randomisierten Entscheidungsregel ρ eine deterministische Entscheidungsregel ρ', deren Risiko nicht größer ist.*

Beweis. Aus der Jensenschen Ungleichung folgt wegen Konvexität von $l(\vartheta, \bullet)$

$$R(\vartheta, \rho) = \int_{\mathcal{X}} \int_A l(\vartheta, a) \rho(x, da)\, \mathbb{P}_\vartheta(dx) \geqslant \int_{\mathcal{X}} l\Big(\vartheta, \int_A a \rho(x, da)\Big)\, \mathbb{P}_\vartheta(dx).$$

Da A konvex ist, gilt $\rho'(x) := \int_A a\rho(x, da) \in A$ und somit $R(\vartheta, \rho) \geqslant R(\vartheta, \rho')$. □

1.37 Definition. Zu vorgegebener Verlustfunktion l heißt eine Entscheidungsregel ρ unverzerrt, falls

$$\forall \vartheta, \vartheta' \in \Theta : \ \mathbb{E}_\vartheta[l(\vartheta', \rho)] \geqslant \mathbb{E}_\vartheta[l(\vartheta, \rho)] =: R(\vartheta, \rho).$$

1.38 Lemma. *Es seien $g : \Theta \to A \subseteq \mathbb{R}$ und $l(\vartheta, \rho) = (\rho - g(\vartheta))^2$ der quadratische Verlust. Dann ist eine Entscheidungsregel (ein Schätzer von $g(\vartheta)$) $\hat{g} : \mathcal{X} \to A$ mit $\mathbb{E}_\vartheta[\hat{g}^2] < \infty$ und $\mathbb{E}_\vartheta[\hat{g}] \in g(\Theta)$ für alle $\vartheta \in \Theta$ genau dann unverzerrt, wenn sie erwartungstreu ist, d.h. $\mathbb{E}_\vartheta[\hat{g}] = g(\vartheta)$ für alle $\vartheta \in \Theta$ gilt.*

Beweis. Es gelte $\mathbb{E}_\vartheta[\hat{g}] = g(\vartheta')$ mit Parametern $\vartheta', \vartheta \in \Theta$. Dann ist

$$\mathbb{E}_\vartheta[(\hat{g} - g(\vartheta'))^2] = \mathrm{Var}_\vartheta(\hat{g}) \leqslant (\mathbb{E}_\vartheta[\hat{g}] - g(\vartheta))^2 + \mathrm{Var}_\vartheta(\hat{g}) = \mathbb{E}_\vartheta[(\hat{g} - g(\vartheta))^2]$$

und Gleichheit gilt genau dann, wenn $g(\vartheta) = \mathbb{E}_\vartheta[\hat{g}]$. Ist $\hat{g}$ unverzerrt, so gilt $\mathbb{E}_\vartheta[(\hat{g} - g(\vartheta'))^2] \geqslant \mathbb{E}_\vartheta[(\hat{g} - g(\vartheta))^2]$, also $\mathbb{E}_\vartheta[\hat{g}] = g(\vartheta)$.

Ist $\hat{g}$ andererseits erwartungstreu, so folgt für alle ϑ, ϑ' analog

$$\mathbb{E}_\vartheta[(\hat{g} - g(\vartheta))^2] = \mathrm{Var}_\vartheta(\hat{g}) \leqslant (\mathbb{E}_\vartheta[\hat{g}] - g(\vartheta'))^2 + \mathrm{Var}_\vartheta(\hat{g}) = \mathbb{E}_\vartheta[(\hat{g} - g(\vartheta'))^2],$$

also Unverzerrtheit. □

1.39 Lemma. *Es sei $\Theta = \Theta_0 \dot\cup \Theta_1$, $A = [0,1]$. Für den Verlust $l(\vartheta, a) = l_0 a \mathbf{1}_{\Theta_0}(\vartheta) + l_1(1-a)\mathbf{1}_{\Theta_1}(\vartheta)$ mit $l_0, l_1 > 0$ ist eine Entscheidungsregel ρ (ein randomisierter Test von $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$) genau dann unverzerrt, wenn sie zum Niveau $\alpha := \frac{l_1}{l_0 + l_1}$ unverfälscht ist, d.h.*

$$\forall \vartheta \in \Theta_0 : \ \mathbb{E}_\vartheta[\rho] \leqslant \alpha, \quad \forall \vartheta \in \Theta_1 : \ \mathbb{E}_\vartheta[\rho] \geqslant \alpha.$$

Beweis. Übung! □

2 Dominierte Modelle und Suffizienz

2.1 Dominierte Modelle

2.1 Bemerkung. Wir sagen, dass ein Maß ν absolutstetig bezüglich einem Maß μ auf $(\Omega, \mathscr{F})$ ist (Notation $\nu \ll \mu$), wenn $\mu(A) = 0 \Rightarrow \nu(A) = 0$ für alle $A \in \mathscr{F}$ gilt. Der Satz von Radon-Nikodym (Stochastik II, Funktionalanalysis) zeigt, dass dann für σ-endliches μ stets eine (μ-f.ü. eindeutige) μ-Dichte f_ν von ν existiert, das heißt eine messbare Funktion $f_\nu : \Omega \to \mathbb{R}^+$ mit $\nu(A) = \int_A f_\nu(x)\mu(dx)$, $A \in \mathscr{F}$. f_ν heißt auch Radon-Nikodym-Dichte von ν bezüglich μ und man schreibt $f_\nu = \frac{d\nu}{d\mu}$.

2.2 Definition. Ein statistisches Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ heißt dominiert (von μ), falls es ein σ-endliches Maß μ auf $\mathscr{F}$ gibt, so dass $\mathbb{P}_\vartheta$ absolutstetig bezüglich μ ist ($\mathbb{P}_\vartheta \ll \mu$) für alle $\vartheta \in \Theta$. Die durch ϑ parametrisierte Radon-Nikodym-Dichte

$$L(\vartheta, x) := \frac{d\,\mathbb{P}_\vartheta}{d\mu}(x), \quad \vartheta \in \Theta,\ x \in \mathcal{X},$$

heißt auch Likelihoodfunktion, wobei diese meist als durch x parametrisierte Funktion in ϑ aufgefasst wird.

2.3 Beispiele.

(a) $\mathcal{X} = \mathbb{R}$, $\mathscr{F} = \mathscr{B}_\mathbb{R}$, $\mathbb{P}_\vartheta$ ist gegeben durch eine Lebesguedichte f_ϑ, beispielsweise $\mathbb{P}_{(\mu,\sigma)} = N(\mu, \sigma^2)$ oder $\mathbb{P}_\vartheta = U([0, \vartheta])$.

(b) Jedes statistische Modell auf dem Stichprobenraum $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ oder allgemeiner auf einem abzählbaren Raum $(\mathcal{X}, \mathcal{P}(\mathcal{X}))$ ist vom Zählmaß dominiert.

(c) Ist $\Theta = \{\vartheta_1, \vartheta_2, \ldots\}$ abzählbar, so ist $\mu = \sum_i c_i \mathbb{P}_{\vartheta_i}$ mit $c_i > 0$, $\sum_i c_i = 1$ ein dominierendes Maß.

(d) $\mathcal{X} = \mathbb{R}$, $\mathscr{F} = \mathscr{B}_{\mathbb{R}}$, $\mathbb{P}_\vartheta = \delta_\vartheta$ für $\vartheta \in \Theta = \mathbb{R}$ (δ_ϑ ist Punktmaß in ϑ) ist nicht dominiert. Ein dominierendes Maß μ müsste nämlich $\mu(\{\vartheta\}) > 0$ für alle $\vartheta \in \Theta$ und damit $\mu(A) = \infty$ für jede überabzählbare Borelmenge $A \subseteq \mathbb{R}$ erfüllen (sonst folgte aus $|\{x \in A \,|\, \mu(\{x\}) \geqslant 1/n\}| \leqslant n\mu(A) < \infty$, dass $A = \bigcup_{n\geqslant 1}\{x \in A \,|\, \mu(\{x\}) \geqslant 1/n\}$ abzählbar ist). Damit kann μ nicht σ-endlich sein.

2.4 Satz. *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein dominiertes Modell. Dann gibt es ein Wahrscheinlichkeitsmaß $\mathbb{Q}$ der Form $\mathbb{Q} = \sum_{i=1}^\infty c_i \mathbb{P}_{\vartheta_i}$ mit $c_i \geqslant 0$, $\sum_i c_i = 1$, $\vartheta_i \in \Theta$, so dass $\mathbb{P}_\vartheta \ll \mathbb{Q}$ für alle $\vartheta \in \Theta$ gilt.*

2.5 Bemerkung. Ein solches Wahrscheinlichkeitsmaß $\mathbb{Q}$ heißt auch privilegiertes dominierendes Maß.

Beweis. Sei zunächst das dominierende Maß μ endlich sowie

$$\mathcal{P}_0 := \Big\{ \sum_i c_i \mathbb{P}_{\vartheta_i} \;\Big|\; \vartheta_i \in \Theta,\, c_i \geqslant 0,\, \sum_i c_i = 1 \Big\} \text{ (konvexe Hülle von } (\mathbb{P}_\vartheta)),$$

$$\mathcal{A} := \Big\{ A \in \mathscr{F} \;\Big|\; \exists\, \mathbb{P} \in \mathcal{P}_0 :\ \mathbb{P}(A) > 0 \text{ und } \frac{d\,\mathbb{P}}{d\mu} > 0 \ \mu\text{-f.ü. auf } A \Big\}.$$

Wähle nun eine Folge (A_n) in $\mathcal{A}$ mit $\mu(A_n) \to \sup_{A\in\mathcal{A}} \mu(A) < \infty$. Setze $A_\infty := \bigcup_n A_n$ und bezeichne $\mathbb{P}_n$ ein Element in $\mathcal{P}_0$ mit $\mathbb{P}_n(A_n) > 0$, $\frac{d\,\mathbb{P}_n}{d\mu} > 0$ μ-f.ü. auf A_n. Für beliebige $c_n > 0$ mit $\sum_n c_n = 1$ setze $\mathbb{Q} := \sum_n c_n \mathbb{P}_n \in \mathcal{P}_0$.

Aus der Wahl von $\mathbb{P}_n$ folgt $\frac{d\,\mathbb{Q}}{d\mu} \geqslant c_n \frac{d\,\mathbb{P}_n}{d\mu} > 0$ μ-f.ü. auf A_n und somit $\frac{d\,\mathbb{Q}}{d\mu} > 0$ μ-f.ü. auf A_∞ und $\mathbb{Q}(A_\infty) > 0$, so dass A_∞ ebenfalls in $\mathcal{A}$ liegt.

Zeige: $\mathbb{P} \ll \mathbb{Q}$ für alle $\mathbb{P} \in \mathcal{P}_0$. Sonst gilt $\mathbb{P}(A) > 0$ und $\mathbb{Q}(A) = 0$ für ein $\mathbb{P}$ und ein $A \in \mathscr{F}$. Dies impliziert $\mathbb{Q}(A\cap A_\infty) = 0 \Rightarrow \mu(A\cap A_\infty) = 0$ (da $\frac{dQ}{d\mu} > 0$ auf A_∞) und weiter $\mathbb{P}(A \cap A_\infty) = 0$ (da $\mathbb{P} \ll \mu$). Für $B := \{\frac{d\,\mathbb{P}}{d\mu} > 0\}$ gilt $\mathbb{P}(B) = 1$, und wir erhalten $\mathbb{P}(A\cap A_\infty^C \cap B) = \mathbb{P}(A) > 0$. Aus $\mathbb{P} \ll \mu$ folgt $\mu(A\cap A_\infty^C \cap B) > 0$ und somit $\mu(A_\infty \dot\cup (A \cap A_\infty^C \cap B)) > \mu(A_\infty)$. Nun ist aber $(\mathbb{P}+\mathbb{Q})/2 \in \mathcal{P}_0$ sowie $\frac{d(\mathbb{P}+\mathbb{Q})}{2d\mu} > 0$ μ-f.ü. auf $A_\infty \dot\cup (A \cap A_\infty^C \cap B)$, was $A_\infty \dot\cup (A \cap A_\infty^C \cap B) \in \mathcal{A}$ zeigt. Dies widerspricht aber der Eigenschaft $\mu(A_\infty) = \sup_{A\in\mathcal{A}} \mu(A)$.

Ist μ σ-endlich, so zerlege $\mathcal{X} := \dot\bigcup_{m\geqslant 1} \mathcal{X}_m$ mit $\mu(\mathcal{X}_m) < \infty$, definiere das Maß $\mathbb{Q}_m$ wie oben $\mathbb{Q}$, wobei im Fall $\mathbb{P}_\vartheta(\mathcal{X}_m) = 0$ für alle $\vartheta \in \Theta$ einfach $\mathbb{Q}_m = \mathbb{P}_\vartheta$ für ein beliebiges $\vartheta \in \Theta$ gesetzt wird. Dann leistet $\sum_{m\geqslant 1} 2^{-m} \mathbb{Q}_m$ das Gewünschte. □

2.2 Exponentialfamilien

2.6 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein von μ dominiertes Modell. Dann heißt $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ Exponentialfamilie (in $\eta(\vartheta)$ und T), wenn $k \in \mathbb{N}$, $\eta : \Theta \to \mathbb{R}^k$,

$C : \Theta \to \mathbb{R}^+$, $T : \mathcal{X} \to \mathbb{R}^k$ messbar und $h : \mathcal{X} \to \mathbb{R}^+$ messbar existieren, so dass

$$\frac{d\,\mathbb{P}_\vartheta}{d\mu}(x) = C(\vartheta)h(x)\exp(\langle \eta(\vartheta), T(x)\rangle_{\mathbb{R}^k}), \quad x \in \mathcal{X},\ \vartheta \in \Theta.$$

T wird natürliche suffiziente Statistik von $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ genannt. Sind $\eta_1, \dots, \eta_k$ linear unabhängige Funktionen und gilt für alle $\vartheta \in \Theta$ die Implikation

$$\lambda_0 + \lambda_1 T_1 + \cdots + \lambda_k T_k = 0 \ \mathbb{P}_\vartheta\text{-f.s.} \Rightarrow \lambda_0 = \lambda_1 = \cdots = \lambda_k = 0$$

$(1, T_1, \dots, T_k$ sind $\mathbb{P}_\vartheta$-f.s. linear unabhängig), so heißt die Exponentialfamilie (strikt) k-parametrisch.

2.7 Bemerkungen.

(a) $C(\vartheta)$ ist nur Normierungskonstante: $C(\vartheta) = (\int h(x)e^{\langle \eta(\vartheta), T(x)\rangle}\mu(dx))^{-1}$.

(b) Die Darstellung ist nicht eindeutig, mit einer invertierbaren Matrix $A \in \mathbb{R}^{k\times k}$ erhält man beispielsweise eine Exponentialfamilie in $\tilde{\eta}(\vartheta) = A\eta(\vartheta)$ und $\tilde{T}(x) = (A^\top)^{-1}T(x)$.

(c) Die Funktion h kann in das dominierende Maß absorbiert werden, indem man $\tilde{\mu}(dx) = h(x)\mu(dx)$ statt μ betrachtet. Da $C(\vartheta) > 0$ gilt, ist dann $\frac{d\,\mathbb{P}_\vartheta}{d\tilde{\mu}} > 0$ $\tilde{\mu}$-f.s. und alle Verteilungen $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ sind untereinander und mit $\tilde{\mu}$ äquivalent (gegenseitig absolut-stetig). Insbesondere bildet für ein ϑ_0 die Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ auch eine Exponetialfamilie bezüglich $\mathbb{P}_{\vartheta_0}$ in $\tilde{\eta}(\vartheta) = \eta(\vartheta) - \eta(\vartheta_0)$ und $T(x)$.

(d) Aus der Identifizierbarkeitsforderung $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$ für alle $\vartheta \neq \vartheta'$ folgt die Injektivität von η. Andererseits impliziert die Injektivität von η bei einer k-parametrischen Exponentialfamilie die Identifizierbarkeitsforderung.

2.8 Definition. Bildet $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine Exponentialfamilie (mit obiger Notation), so heißt

$$\mathcal{Z} := \left\{ u \in \mathbb{R}^k \;\middle|\; \int_{\mathcal{X}} e^{\langle u, T(x)\rangle} h(x)\mu(dx) \in (0,\infty) \right\}$$

ihr natürlicher Parameterraum. Die entsprechend mit $u \in \mathcal{Z}$ parametrisierte Familie wird natürliche Exponentialfamilie in T genannt.

2.9 Beispiele.

(a) $(N(\mu, \sigma^2))_{\mu\in\mathbb{R},\sigma>0}$ ist zweiparametrische Exponentialfamilie in $\eta(\mu, \sigma) = (\mu/\sigma^2, 1/(2\sigma^2))^\top$ und $T(x) = (x, -x^2)^\top$ unter dem Lebesguemaß als dominierendem Maß. Jedes u der Form $u = (\mu/\sigma^2, 1/(2\sigma^2))^\top$ ist natürlicher Parameter, und der natürliche Parameterraum ist gegeben durch $\mathcal{Z} = \mathbb{R} \times (0,\infty)$. Ist $\sigma > 0$ bekannt, so liegt eine einparametrische Exponentialfamilie in $\eta(\mu) = \mu/\sigma^2$ und $T(x) = x$ vor.

(b) $(\mathrm{Bin}(n, p))_{p\in(0,1)}$ bildet eine Exponentialfamilie in $\eta(p) = \log(p/(1-p))$ (auch logit-Funktion genannt) und $T(x) = x$ bezüglich dem Zählmaß μ auf $\{0, 1, \dots, n\}$. Der natürliche Parameterraum ist $\mathbb{R}$. Beachte, dass für den Parameterbereich $p = [0, 1]$ keine Exponentialfamilie vorliegt, da $(\mathrm{Bin}(n, p))_{p\in[0,1]}$ keine äquivalenten Wahrscheinlichkeitsmaße sind.

2.10 Lemma. *Bildet $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine (k-parametrische) Exponentialfamilie in $\eta(\vartheta)$ und $T(x)$, so bilden auch die Produktmaße $(\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta}$ eine (k-parametrische) Exponentialfamilie in $\eta(\vartheta)$ und $\sum_{i=1}^n T(x_i)$ mit*

$$\frac{d\,\mathbb{P}_\vartheta^{\otimes n}}{d\mu^{\otimes n}}(x) = C(\vartheta)^n \Big(\prod_{i=1}^n h(x_i)\Big) \exp\Big(\langle \eta(\vartheta), \textstyle\sum_{i=1}^n T(x_i)\rangle\Big), \quad x\in\mathcal{X}^n,\ \vartheta\in\Theta.$$

Beweis. Dies folgt sofort aus der Produktformel $\frac{d\,\mathbb{P}_\vartheta^{\otimes n}}{d\mu^{\otimes n}}(x) = \prod_{i=1}^n \frac{d\,\mathbb{P}_\vartheta}{d\mu}(x_i)$. □

2.11 Satz. *Es sei $(\mathbb{P}_\vartheta)_{\vartheta\in\mathscr{Z}}$ eine Exponentialfamilie mit natürlichem Parameterraum $\mathscr{Z}\subseteq\mathbb{R}^k$ und Darstellung*

$$\frac{d\,\mathbb{P}_\vartheta}{d\mu}(x) = C(\vartheta)h(x)\exp(\langle\vartheta, T(x)\rangle) = h(x)\exp(\langle\vartheta,T(x)\rangle - A(\vartheta)),$$

wobei $A(\vartheta) = \log\big(\int h(x)\exp(\langle\vartheta,T(x)\rangle)\mu(dx)\big)$. Ist ϑ_0 ein innerer Punkt von $\mathscr{Z}$, so ist die erzeugende Funktion $\psi_{\vartheta_0}(s) = \mathbb{E}_{\vartheta_0}[e^{\langle T,s\rangle}]$ in einer Umgebung der Null wohldefiniert und beliebig oft differenzierbar. Es gilt $\psi_{\vartheta_0}(s) = \exp(A(\vartheta_0+s) - A(\vartheta_0))$ für alle s mit $\vartheta_0+s\in\mathscr{Z}$.

Für $i,j=1,\dots,k$ folgt $\mathbb{E}_{\vartheta_0}[T_i] = \frac{dA}{d\vartheta_i}(\vartheta_0)$ und $\mathrm{Cov}_{\vartheta_0}(T_i,T_j) = \frac{d^2A}{d\vartheta_i d\vartheta_j}(\vartheta_0)$.

Beweis. Für alle $s\in\mathbb{R}^k$ mit $\vartheta_0+s\in\mathscr{Z}$ gilt

$$\psi_{\vartheta_0}(s) = \int e^{\langle T,s\rangle} e^{\langle\vartheta_0,T\rangle - A(\vartheta_0)} h\,d\mu = \int e^{\langle\vartheta_0+s,T\rangle - A(\vartheta_0)} h\,d\mu = e^{A(\vartheta_0+s)-A(\vartheta_0)}.$$

Insbesondere ist ψ_{ϑ_0} in einer Umgebung von $s=0$ endlich und somit wohldefiniert.

Für $v\in\mathbb{R}^k$ und $\varepsilon>0$ hinreichend klein, betrachte den Differenzenquotienten

$$\frac{\psi_{\vartheta_0}(\varepsilon v)-\psi_{\vartheta_0}(0)}{\varepsilon} = \int \frac{e^{\varepsilon\langle T,v\rangle}-1}{\varepsilon} e^{\langle\vartheta_0,T\rangle - A(\vartheta_0)} h\,d\mu.$$

Der Bruch im Integranden konvergiert für $\varepsilon\to 0$ punktweise gegen $\langle T,v\rangle$. Aus der Ungleichung $|\frac{e^{az}-1}{z}| \leqslant \frac{e^{\delta|a|}}{\delta}$ für $|z|\leqslant\delta$, $a\in\mathbb{R}$, ergibt sich $\frac{e^{\varepsilon_0\langle T,v\rangle}+e^{-\varepsilon_0\langle T,v\rangle}}{\varepsilon_0}$ als Majorante des Bruchs für alle $\varepsilon\leqslant\varepsilon_0$. Nach dem ersten Schritt gilt für $\varepsilon_0>0$ mit $\vartheta_0\pm\varepsilon_0 v\in\mathscr{Z}$, dass die Majorante integrierbar ist, und wir schließen mittels dominierter Konvergenz auf die Richtungsableitung

$$\lim_{\varepsilon\to 0}\frac{\psi_{\vartheta_0}(\varepsilon v)-\psi_{\vartheta_0}(0)}{\varepsilon} = \mathbb{E}_{\vartheta_0}[\langle T,v\rangle], \quad v\in\mathbb{R}^k\,.$$

Also ist ψ_{ϑ_0} differenzierbar bei Null mit Gradienten $\mathbb{E}_{\vartheta_0}[T]$. Wegen $A(\vartheta_0+s) = A(\vartheta_0)+\log(\psi_{\vartheta_0}(s))$ für s in einer Nullumgebung ist also auch A differenzierbar bei ϑ_0 mit Gradienten $\mathbb{E}_{\vartheta_0}[T]$ (beachte $\psi_{\vartheta_0}(0)=1$).

Analog ergibt sich, dass ψ_{ϑ_0} beliebig oft differenzierbar ist mit höheren partielle Ableitungen

$$\frac{d^{i_1}}{ds_1^{i_1}}\cdots\frac{d^{i_k}}{ds_k^{i_k}}\psi_{\vartheta_0}(s)\Big|_{s=0} = \int T_1^{i_1}\cdots T_k^{i_k} e^{\langle\vartheta_0,T\rangle - A(\vartheta_0)} d\mu = \mathbb{E}_{\vartheta_0}[T_1^{i_1}\cdots T_k^{i_k}].$$

Wir erhalten insbesondere

$$\frac{d^2A}{d\vartheta_i d\vartheta_j}(\vartheta_0) = \frac{d^2\log(\psi_{\vartheta_0})}{ds_i ds_j}(0) = \Big(\frac{d^2\psi_{\vartheta_0}}{ds_i ds_j} - \frac{d\psi_{\vartheta_0}}{ds_i}\frac{d\psi_{\vartheta_0}}{ds_j}\Big)(0) = \mathrm{Cov}_{\vartheta_0}(T_i, T_j).$$

□

2.12 Beispiel. Für $\mathbb{P}_\vartheta = N(\vartheta,1)^{\otimes n}$ bildet $(\mathbb{P}_\vartheta)_{\vartheta\in\mathbb{R}}$ eine natürliche Exponentialfamilie in $T(x) = \sum_{i=1}^n x_i$, $x \in \mathbb{R}^n$, mit $A(\vartheta) = n\vartheta^2/2$. Wir erhalten $\mathbb{E}_\vartheta[T] = A'(\vartheta)$, d.h. $\mathbb{E}_\vartheta[\sum_{i=1}^n X_i] = n\vartheta$, sowie $\mathrm{Var}_\vartheta(T) = A''(\vartheta)$, d.h. $\mathrm{Var}_\vartheta(\sum_{i=1}^n X_i) = n$.

2.3 Suffizienz

2.13 Beispiel. Es sei $X_1,\dots,X_n$ eine gemäß der Lebesguedichte $f_\vartheta : \mathbb{R} \to \mathbb{R}^+$ verteilte mathematische Stichprobe. Dann liefern die Statistiken $\bar{X}$ oder $\max(X_1,\dots,X_n)$ im Allgemeinen Information über f_ϑ und damit ϑ. Hingegen sind $\mathbf{1}(X_1 < X_2)$ oder $\mathbf{1}(X_1 = \max(X_1,\dots,X_n))$ Statistiken, deren Verteilung nicht von f_ϑ abhängt (sofern die i.i.d.-Annahme gültig ist) und somit keinerlei Informationen über ϑ beinhalten. Allgemein heißt eine Statistik V *ancillary*, wenn ihre Verteilung nicht von ϑ abhängt. Also ist beispielsweise $V = \mathbf{1}(X_1 < X_2)$ ancillary, weil stets $\mathrm{Bin}(1,1/2)$-verteilt. Intuitiv ist alle Information bereits in der Ordnungsstatistik $X_{(1)},\dots,X_{(n)}$ enthalten mit $X_{(1)} = \min\{X_1,\dots,X_n\}$, $X_{(k+1)} := \min\{X_1,\dots,X_n\} \setminus \{X_{(1)},\dots,X_{(k)}\}$ oder äquivalent in der empirischen Verteilungsfunktion $\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \leqslant x)$, $x \in \mathbb{R}$.

2.14 Definition. Eine $(S,\mathscr{S})$-wertige Statistik T auf $(\mathcal{X},\mathscr{F},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt suffizient (für $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$), falls für jedes $\vartheta \in \Theta$ die bedingte Wahrscheinlichkeit von $\mathbb{P}_\vartheta$ gegeben T nicht von ϑ abhängt, d.h. es existiert $k : S \times \mathscr{F} \to [0,1]$, messbar im ersten Argument, so dass

$$\forall\vartheta\in\Theta,\ B\in\mathscr{F}:\ k(T,B) = \mathbb{P}_\vartheta(B\,|\,T) := \mathbb{E}_\vartheta[\mathbf{1}_B\,|\,T] \quad \mathbb{P}_\vartheta\text{-f.s.}$$

Statt $k(t,B)$ schreiben wir $\mathbb{P}_\bullet(B\,|\,T=t)$ bzw. $\mathbb{E}_\bullet[\mathbf{1}_B\,|\,T=t]$.

2.15 Satz (Faktorisierungskriterium von Neyman)**.** *Es sei $(\mathcal{X},\mathscr{F},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein von μ dominiertes Modell mit Likelihoodfunktion L sowie T eine $(S,\mathscr{S})$-wertige Statistik. Dann ist T genau dann suffizient, wenn eine messbare Funktion $h : \mathcal{X} \to \mathbb{R}^+$ existiert, so dass für alle $\vartheta \in \Theta$ eine messbare Funktion $g_\vartheta : S \to \mathbb{R}^+$ existiert mit*

$$L(\vartheta,x) = g_\vartheta(T(x))h(x) \quad \textit{für } \mu\textit{-f.a. } x \in \mathcal{X}.$$

2.16 Lemma. *Es seien $\mathbb{P}$ und μ Wahrscheinlichkeitsmaße mit $\mathbb{P} \ll \mu$ und T eine messbare Abbildung auf $(\mathcal{X},\mathscr{F})$. Dann gilt für alle $B \in \mathscr{F}$*

$$\mathbb{P}(B\,|\,T) = \mathbb{E}_\mathbb{P}[1_B\,|\,T] = \frac{\mathbb{E}_\mu[1_B\frac{d\mathbb{P}}{d\mu}\,|\,T]}{\mathbb{E}_\mu[\frac{d\mathbb{P}}{d\mu}\,|\,T]} \quad \mathbb{P}\textit{-f.s.}$$

Beweis. Für jede beschränkte messbare Funktion φ erfüllt die rechte Seite

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\Big[\frac{\mathbb{E}_\mu[1_B\frac{d\,\mathbb{P}}{d\mu}\,|\,T]}{\mathbb{E}_\mu[\frac{d\,\mathbb{P}}{d\mu}\,|\,T]}\varphi(T)\Big] &= \mathbb{E}_\mu\Big[\frac{\mathbb{E}_\mu[1_B\frac{d\,\mathbb{P}}{d\mu}\,|\,T]}{\mathbb{E}_\mu[\frac{d\,\mathbb{P}}{d\mu}\,|\,T]}\varphi(T)\frac{d\,\mathbb{P}}{d\mu}\Big]\\ &= \mathbb{E}_\mu\Big[\frac{\mathbb{E}_\mu[1_B\frac{d\,\mathbb{P}}{d\mu}\,|\,T]}{\mathbb{E}_\mu[\frac{d\,\mathbb{P}}{d\mu}\,|\,T]}\varphi(T)\,\mathbb{E}_\mu[\tfrac{d\,\mathbb{P}}{d\mu}\,|\,T]\Big]\\ &= \mathbb{E}_\mu[1_B\tfrac{d\,\mathbb{P}}{d\mu}\varphi(T)]\\ &= \mathbb{E}_{\mathbb{P}}[1_B\varphi(T)].\end{aligned}$$

Zusammen mit der $\sigma(T)$-Messbarkeit ist dies genau die Charakterisierung dafür, dass die rechte Seite eine Version der bedingten Erwartung $\mathbb{E}_{\mathbb{P}}[1_B\,|\,T]$ ist. □

2.17 Bemerkung. Mit den üblichen Approximationsargumenten lässt sich dies zu $\mathbb{E}_{\mathbb{P}}[f\,|\,T] = \mathbb{E}_\mu[f\frac{d\,\mathbb{P}}{d\mu}\,|\,T]/\,\mathbb{E}_\mu[\frac{d\,\mathbb{P}}{d\mu}\,|\,T]$ für $f \in L^1(\mathbb{P})$ verallgemeinern.

Beweis des Faktorisierungssatzes. Ohne Einschränkung sei μ ein Wahrscheinlichkeitsmaß, sonst betrachte das äquivalente Wahrscheinlichkeitsmaß $\tilde\mu(dx) = z(x)\mu(dx)$ mit $z = \sum_{m\geqslant 1} 2^{-m}\mu(\mathcal{X}_m)^{-1}\mathbf{1}_{\mathcal{X}_m}$, wobei die Zerlegung $\mathcal{X} := \bigcup_{m\geqslant 1}\mathcal{X}_m$ mit $\mu(\mathcal{X}_m) < \infty$ wegen der σ-Endlichkeit von μ existiert.

Aus dem Lemma und der Form von $L(\vartheta, x)$ folgt daher

$$\mathbb{P}_\vartheta(B\,|\,T) = \frac{g_\vartheta(T)\,\mathbb{E}_\mu[\mathbf{1}_B h\,|\,T]}{g_\vartheta(T)\,\mathbb{E}_\mu[h\,|\,T]} = \frac{\mathbb{E}_\mu[\mathbf{1}_B h\,|\,T]}{\mathbb{E}_\mu[h\,|\,T]} \quad \mathbb{P}_\vartheta\text{-f.s.}$$

Da die rechte Seite unabhängig von ϑ ist, ist T suffizient.

Ist nun andererseits T suffizient, so setze $k(T,B) := \mathbb{P}_\vartheta(B\,|\,T)$, $\vartheta \in \Theta$. Für das privilegierte dominierende Maß $\mathbb{Q}$ gilt dann ebenfalls $\mathbb{Q}(B\,|\,T) = \sum_i c_i\,\mathbb{P}_{\vartheta_i}(B\,|\,T) = k(T,B)$ $\mathbb{Q}$-f.s. Nach dem Satz von Radon-Nikodym gilt auf dem Teilraum $(\mathcal{X}, \sigma(T))$

$$\forall\vartheta\,\exists f_\vartheta : \mathcal{X} \to \mathbb{R}^+ \;\; \sigma(T)\text{-messbar} : \; \frac{d\,\mathbb{P}_\vartheta\,|_{\sigma(T)}}{d\,\mathbb{Q}\,|_{\sigma(T)}} = f_\vartheta.$$

Nach Stochastik II gibt es eine messbare Funktion g_ϑ, so dass $f_\vartheta = g_\vartheta \circ T$, und für beliebiges $B \in \mathscr{F}$ erhalten wir

$$\mathbb{P}_\vartheta(B) = \mathbb{E}_\vartheta[\mathbb{E}_\vartheta[1_B\,|\,T]] = \mathbb{E}_{\mathbb{Q}}[\mathbb{E}_{\mathbb{Q}}[1_B\,|\,T]g_\vartheta(T)] = \mathbb{E}_{\mathbb{Q}}[1_B g_\vartheta(T)],$$

so dass $g_\vartheta \circ T$ auch die Radon-Nikodym-Dichte $\frac{d\,\mathbb{P}_\vartheta}{d\,\mathbb{Q}}$ auf ganz $\mathscr{F}$ ist. Mit $\frac{d\,\mathbb{P}_\vartheta}{d\mu} = \frac{d\,\mathbb{P}_\vartheta}{d\,\mathbb{Q}}\frac{d\,\mathbb{Q}}{d\mu}$ erhalten wir den Ausdruck von $L(\vartheta, x)$, wobei $h(x) = \frac{d\,\mathbb{Q}}{d\mu}(x)$. □

2.18 Beispiele.

(a) Die Identität $T(x) = x$ und allgemein jede bijektive, bi-messbare Transformation T ist suffizient.

(b) Die natürliche suffiziente Statistik T einer Exponentialfamilie ist in der Tat suffizient. Im Normalverteilungsmodell $(N(\mu,\sigma^2)^{\otimes n})_{\mu\in\mathbb{R},\sigma>0}$ ist damit $T_1(x) = (\sum_{i=1}^n x_i, -\sum_{i=1}^n x_i^2)^\top$ suffizient, aber durch Transformation auch $T_2(x) = (\bar{x}, \overline{x^2})$ oder $T_3(x) = (\bar{x}, \bar{s}^2)$ mit der empirischen Varianz $\bar{s}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$. Bei einer Bernoullikette $(\mathrm{Bin}(1,p)^{\otimes n})_{p\in(0,1)}$ ist $T(x) = \sum_{i=1}^n x_i$ (die Anzahl der Erfolge) suffizient.

(c) Ist $X_1, \ldots, X_n$ eine mathematische Stichprobe, wobei X_i gemäß der Lebesguedichte $f_\vartheta : \mathbb{R} \to \mathbb{R}^+$ verteilt ist, so ist die Ordnungsstatistik $(X_{(1)}, \ldots, X_{(n)})$ suffizient. Die Likelihoodfunktion lässt sich nämlich in der Form $L(\vartheta, x) = \prod_{i=1}^n f_\vartheta(x_{(i)})$ schreiben.

(d) Es wird die Realisierung $(N_t,\, t \in [0,T])$ eines Poissonprozesses zum unbekannten Parameter $\lambda > 0$ kontinuierlich auf $[0,T]$ beobachtet (man denke an Geigerzähleraufzeichnungen). Mit $S_k = \inf\{t \geqslant 0 \,|\, N_t = k\}$ werden die Sprungzeiten bezeichnet. In der Wahrscheinlichkeitstheorie wird gezeigt, dass bedingt auf das Ereignis $\{N_T = n\}$ die Sprungzeiten $(S_1, \ldots, S_n)$ dieselbe Verteilung haben wie die Ordnungsstatistik $(X_{(1)}, \ldots, X_{(n)})$ mit unabhängigen $X_i \sim U([0,T])$. Da sich die Beobachtung $(N_t,\, t \in [0,T])$ eindeutig aus den S_k rekonstruieren lässt, ist die Verteilung dieser Beobachtung gegeben $\{N_T = n\}$ unabhängig von λ, und N_T ist somit eine suffiziente Statistik (die Kenntnis der Gesamtzahl der gemessenen radioaktiven Zerfälle liefert bereits die maximal mögliche Information über die Intensität λ).

2.19 Satz (Rao-Blackwell)**.** *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, der Aktionsraum $A \subseteq \mathbb{R}^k$ konvex und die Verlustfunktion $l(\vartheta, a)$ im zweiten Argument konvex. Ist T eine für $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ suffiziente Statistik, so gilt für jede Entscheidungsregel ρ und für $\tilde{\rho} := \mathbb{E}_\bullet[\rho \,|\, T]$ die Risikoabschätzung*

$$\forall \vartheta \in \Theta :\ R(\vartheta, \tilde{\rho}) \leqslant R(\vartheta, \rho).$$

Beweis. Dies folgt aus der Jensenschen Ungleichung für bedingte Erwartungen:

$$R(\vartheta, \tilde{\rho}) = \mathbb{E}_\vartheta[l(\vartheta, \mathbb{E}_\vartheta[\rho \,|\, T])] \leqslant \mathbb{E}_\vartheta[\mathbb{E}_\vartheta[l(\vartheta, \rho) \,|\, T]] = R(\vartheta, \rho).$$

□

2.20 Bemerkung. Ist l sogar strikt konvex sowie $\mathbb{P}_\vartheta(\tilde{\rho} = \rho) < 1$, so gilt in der Jensenschen Ungleichung sogar die strikte Ungleichung und $\tilde{\rho}$ ist besser als ρ.

2.21 Satz. *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und T eine suffiziente Statistik. Dann gibt es zu jedem randomisierten Test φ einen randomisierten Test $\tilde{\varphi}$, der nur von T abhängt und dieselben Fehlerwahrscheinlichkeiten erster und zweiter Art besitzt, nämlich $\tilde{\varphi} = \mathbb{E}_\bullet[\varphi \,|\, T]$.*

Beweis. Dies folgt jeweils aus $\mathbb{E}_\vartheta[\tilde{\varphi}] = \mathbb{E}_\vartheta[\varphi]$. □

2.22 Beispiel. Es sei $X_1, \ldots, X_n$ eine $U([0,\vartheta])$-verteilte mathematische Stichprobe mit $\vartheta > 0$ unbekannt. Dann ist $\bar{X}$ ein erwartungstreuer Schätzer des

Erwartungswerts $\frac{\vartheta}{2}$, so dass $\hat{\vartheta} = 2\bar{X}$ ein plausibler Schätzer von ϑ ist mit quadratischem Risiko $R(\vartheta, \hat{\vartheta}) = 4\,\mathrm{Var}_\vartheta(\bar{X}) = \frac{4\vartheta^2}{12n}$. Nun ist jedoch (bezüglich Lebesguemaß auf $(\mathbb{R}^+)^n$) die Likelihoodfunktion

$$L(\vartheta, x) = \prod_{i=1}^n \big(\vartheta^{-1}\mathbf{1}_{[0,\vartheta]}(x_i)\big) = \vartheta^{-n}\mathbf{1}_{[0,\vartheta]}\big(\max_{i=1,\dots,n} x_i\big).$$

Demnach ist $X_{(n)} = \max_{i=1,\dots,n} X_i$ eine suffiziente Statistik, und wir bilden

$$\tilde{\vartheta} := \mathbb{E}_\bullet[\hat{\vartheta} \,|\, X_{(n)}] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_\bullet[X_i \,|\, X_{(n)}].$$

Aus Symmetriegründen reicht es, $\mathbb{E}_\bullet[X_1 \,|\, X_{(n)}]$ zu bestimmen. Als bedingte Verteilung von X_1 gegeben $\{X_{(n)} = m\}$ vermuten wir $\frac{1}{n}\delta_m + \frac{n-1}{n}U([0,m])$ wegen

$$\begin{aligned}\mathbb{P}_\vartheta(X_1 \leqslant x \,|\, X_{(n)} \in [m, m+h]) &= \frac{(x \wedge (m+h))(m+h)^{n-1} - (x \wedge m)m^{n-1}}{(m+h)^n - m^n} \\ &\xrightarrow{h\to 0} \frac{1}{n}\mathbf{1}(\{m < x\}) + \frac{n-1}{n}\frac{x \wedge m}{m}.\end{aligned}$$

In der Tat gilt für $x \in [0, \vartheta]$:

$$\begin{aligned}&\int_0^\vartheta \Big(\frac{1}{n}\delta_m + \frac{n-1}{n}U([0,m])\Big)([0,x])\ \mathbb{P}_\vartheta^{X_{(n)}}(dm) \\ &\quad= \int_0^\vartheta \Big(\frac{1}{n}\mathbf{1}_{[0,x]}(m) + \frac{n-1}{n}\frac{x \wedge m}{m}\Big) n m^{n-1}\vartheta^{-n}\,dm \\ &\quad= \frac{1}{n}(x/\vartheta)^n + \frac{n-1}{n}\Big((x/\vartheta)^n + \frac{nx(\vartheta^{n-1} - x^{n-1})}{(n-1)\vartheta^n}\Big) \\ &\quad= \frac{x}{\vartheta} = \mathbb{P}_\vartheta(X_1 \leqslant x).\end{aligned}$$

Es folgt $\mathbb{E}[X_1 \,|\, X_{(n)}] = \frac{1}{n}X_{(n)} + \frac{n-1}{n}\frac{X_{(n)}}{2} = \frac{n+1}{2n}X_{(n)}$. Wir erhalten $\tilde{\vartheta} = \frac{n+1}{n}X_{(n)}$. Natürlich ist $\tilde{\vartheta}$ auch erwartungstreu und als quadratisches Risiko ergibt eine kurze Rechnung $R(\vartheta, \tilde{\vartheta}) = \frac{\vartheta^2}{n^2+2n}$. Wir sehen, dass $\tilde{\vartheta}$ bedeutend besser als $\hat{\vartheta}$ ist, für $n \to \infty$ erhalten wir die Ordnung $O(n^{-2})$ anstelle $O(n^{-1})$. Es bleibt, die Frage zu klären, ob auch $\tilde{\vartheta}$ noch weiter verbessert werden kann (s.u.).

2.4 Vollständigkeit

2.23 Definition. Eine $(S, \mathscr{S})$-wertige Statistik T auf $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt vollständig, falls für alle messbaren Funktionen $f : S \to \mathbb{R}$ gilt

$$\forall\vartheta \in \Theta:\ \mathbb{E}_\vartheta[f(T)] = 0 \Longrightarrow \forall\vartheta \in \Theta:\ f(T) = 0 \quad \mathbb{P}_\vartheta\text{-f.s.}$$

2.24 Bemerkung. Ist T vollständig und g messbar, so ist auch $g(T)$ vollständig, wie sofort aus der Definition folgt. Dieses Verhalten ist genau entgegengesetzt zur Suffizienz, wo aus $g(T)$ suffizient folgt, dass T suffizient ist.

Ist T vollständig und $V = f(T)$ integrierbar und ancillary, so hängt die Verteilung von V nicht von ϑ ab und damit ist $\mathbb{E}_\vartheta[V] = c$, c eine Konstante, und wegen Vollständigkeit $V = c\,\mathbb{P}_\vartheta$-f.s. (betrachte $\tilde{f}(x) = f(x) - c$). Vollständigkeit von T impliziert also, dass jede ancillary Statistik der Form $V = f(T)$ trivial (d.h. fast sicher konstant) ist. Es ist keine redundante Information mehr in T enthalten. Da $V = \mathbf{1}(X_1 < X_2)$ in Beispiel 2.13 ancillary und nicht-trivial ist, sind weder $T = (X_1, \ldots, X_n)$ (Identität) noch $T = (X_1, X_2)$ vollständig.

2.25 Satz (Lehmann-Scheffé)**.** *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\gamma(\vartheta) \in \mathbb{R}$, $\vartheta \in \Theta$, der jeweils interessierende Parameter. Es existiere ein erwartungstreuer Schätzer $\hat{\gamma}$ von $\gamma(\vartheta)$ mit endlicher Varianz. Ist T eine suffiziente und vollständige Statistik, so ist $\tilde{\gamma} = \mathbb{E}_\bullet[\hat{\gamma} \,|\, T]$ ein Schätzer von gleichmäßig kleinster Varianz in der Klasse aller erwartungstreuen Schätzer (UMVU: uniformly minimum variance unbiased).*

Beweis. Zunächst ist klar, dass $\tilde{\gamma}$ wiederum erwartungstreu ist. Außerdem ist $\tilde{\gamma}$ der f.s. einzige erwartungstreue Schätzer, der $\sigma(T)$-messbar ist, weil jeder andere solche Schätzer $\bar{\gamma}$ wegen Vollständigkeit $\mathbb{E}[\tilde{\gamma} - \bar{\gamma}] = 0 \Rightarrow \tilde{\gamma} = \bar{\gamma}\ \mathbb{P}_\vartheta$-f.s. erfüllt. Nach dem Satz von Rao-Blackwell besitzt $\tilde{\gamma}$ damit kleineres quadratisches Risiko als jeder andere erwartungstreue Schätzer. Nach der Bias-Varianz-Zerlegung ist das quadratische Risiko bei erwartungstreuen Schätzern gleich der Varianz. □

2.26 Bemerkung. Beachte, dass die Aussage des Satzes von Lehmann-Scheffé sogar analog für das Risiko bei beliebigen im zweiten Argument konvexen Verlustfunktionen gilt, wie sofort aus dem Satz von Rao-Blackwell folgt.

2.27 Satz. *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ eine k-parametrische Exponentialfamilie in T mit natürlichem Parameter $\vartheta \in \Theta \subseteq \mathbb{R}^k$. Besitzt Θ ein nichtleeres Inneres, so ist T suffizient und vollständig.*

Beweis. Es bleibt, die Vollständigkeit zu beweisen. Ohne Einschränkung sei $[-a, a]^k \subseteq \Theta$ für ein $a > 0$ (sonst verschiebe entsprechend) sowie $h(x) = 1$ (sonst betrachte $\tilde{\mu}(dx) = h(x)\mu(dx)$). Für alle $\vartheta \in \Theta$ gelte $\mathbb{E}_\vartheta[f(T)] = 0$ für ein $f \in \bigcap_{\vartheta\in\Theta} L^1(\mathbb{R}^k, \mathbb{P}_\vartheta^T)$. Mit $f^+ = \max(f, 0)$, $f^- = \max(-f, 0)$ sowie mit dem Bildmaß μ^T des dominierenden Maßes μ unter T folgt

$$\forall \vartheta \in [-a, a]^k : \int_{\mathbb{R}^k} \exp(\langle \vartheta, t\rangle) f^+(t)\mu^T(dt) = \int_{\mathbb{R}^k} \exp(\langle \vartheta, t\rangle) f^-(t)\mu^T(dt).$$

Insbesondere gilt $\int f^+(t)\mu^T(dt) = \int f^-(t)\mu^T(dt) =: M \in [0, \infty)$. Ist $M = 0$, so ist $f^+ = f^- = 0\ \mu^T$-f.ü. und somit $f(T) = 0\ \mathbb{P}_\vartheta$-f.s. für alle $\vartheta \in \Theta$. Dann ist die Vollständigkeit von T nachgewiesen.

Betrachte nun den Fall $M > 0$. Dann definieren $\mathbb{P}^+(dt) := M^{-1} f^+(t)\mu^T(dt)$, $\mathbb{P}^-(dt) := M^{-1} f^-(t)\mu^T(dt)$ Wahrscheinlichkeitsmaße auf $(\mathbb{R}^k, \mathfrak{B}_{\mathbb{R}^k})$. Die obige Identität bedeutet gerade, dass die Laplace-Transformierten $\chi^\pm(\vartheta) := \int_{\mathbb{R}^k} \exp(\langle \vartheta, t\rangle)\, \mathbb{P}^\pm(dt)$ für $\vartheta \in [-a, a]^k$ übereinstimmen. χ^+ und χ^- sind darüberhinaus auf dem k-dimensionalen komplexen Streifen $\{\vartheta \in \mathbb{C}^k \,|\, |\mathrm{Re}(\vartheta_j)| < a\}$ wohldefiniert und analytisch (Potenzreihen). Der Eindeutigkeitssatz für analytische Funktionen impliziert daher $\chi^+(iu) = \chi^-(iu)$ für

alle $u \in \mathbb{R}^k$. Also besitzen $\mathbb{P}^+$ und $\mathbb{P}^-$ dieselben charakteristischen Funktionen, so dass $\mathbb{P}^+ = \mathbb{P}^-$ folgt (Eindeutigkeitssatz für charakteristische Funktionen). Dies liefert $f^+ = f^-$ μ^T-f.ü. und somit $f(T) = 0$ $\mathbb{P}_\vartheta$-f.s. für alle $\vartheta \in \Theta$. T ist vollständig. □

2.28 Beispiele.

(a) Das lineare Modell $Y = X\beta + \sigma\varepsilon$ mit Designmatrix $X \in \mathbb{R}^{n\times p}$ vom Rang p, Gaußschen Fehlern $\varepsilon \sim N(0, E_n)$ bildet eine $(p+1)$-parametrische Exponentialfamilie in $\eta(\beta, \sigma) = \sigma^{-2}(\beta, -1/2)^\top \in \mathbb{R}^p \times \mathbb{R}^-$ und $T(Y) = (X^\top Y, |Y|^2)^\top \in \mathbb{R}^p \times \mathbb{R}^+$. Der natürliche Parameterbereich $\mathscr{Z} = \mathbb{R}^p \times \mathbb{R}^-$ besitzt nichtleeres Inneres in $\mathbb{R}^{p+1}$, so dass T suffizient und vollständig ist. Durch bijektive Transformation ergibt sich, dass dies auch für $((X^\top X)^{-1}X^\top Y, |Y|^2) = (\hat\beta, |\Pi_X Y|^2 + (n-p)\hat\sigma^2)$ mit dem Kleinste-Quadrate-Schätzer $\hat\beta$ und $\hat\sigma^2 = \frac{|Y - X\hat\beta|^2}{n-p}$ gilt. Wegen $\Pi_X Y = X\hat\beta$ ist also a fortiori auch $(\hat\beta, \hat\sigma^2)$ suffizient und vollständig. Damit besitzen beide Schätzer jeweils minimale Varianz in der Klasse aller (!) erwartungstreuen Schätzer (von β bzw. σ^2), in Erweiterung des Satzes von Gauß-Markov, der sich auf die Klasse der erwartungstreuen, linearen Schätzer bezieht. Hierfür ist die Normalverteilungsannahme essentiell.

Im Spezialfall $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$ i.i.d. ist also $\hat\mu = \bar Y$ UMVU. Auch wenn wir bereits wussten, dass $\hat\mu$ minimax und zulässig ist, zeigt die UMVU-Eigenschaft, dass es keinen erwartungstreuen Schätzer $\tilde\mu$ von μ geben kann, der $\mathrm{Var}_{\mu_0}(\tilde\mu) < \mathrm{Var}_{\mu_0}(\hat\mu) = \frac{\sigma^2}{n}$ für irgendein $\mu_0 \in \mathbb{R}$ erfüllt.

(b) Es sei $X_1, \dots, X_n \sim U([0, \vartheta])$ eine mathematische Stichprobe mit $\vartheta > 0$ unbekannt. Aus der Form $L(x, \vartheta) = \vartheta^{-n}\mathbf{1}(x_{(n)} \leqslant \vartheta)$ für $x \in (\mathbb{R}^+)^n$ der Likelihoodfunktion folgt, dass das Maximum $X_{(n)}$ der Beobachtungen suffizient ist (s.o.). Gilt für $f : \mathbb{R}^+ \to \mathbb{R}$, integrierbar auf jedem Intervall $[0, \vartheta]$, und für alle $\vartheta > 0$

$$\mathbb{E}_\vartheta[f(X_{(n)})] = \int_0^\vartheta f(t) n \vartheta^{-n} t^{n-1}\, dt = 0,$$

so muss $f = 0$ Lebesgue-fast überall gelten, woraus die Vollständigkeit von $X_{(n)}$ folgt. Andererseits gilt $\mathbb{E}_\vartheta[X_{(n)}] = \frac{n}{n+1}\vartheta$. Also ist $\hat\vartheta = \frac{n+1}{n}X_{(n)}$ erwartungstreuer Schätzer von ϑ mit gleichmäßig kleinster Varianz.

2.5 Cramér-Rao-Schranke

2.29 Lemma (Chapman-Robbins-Ungleichung)**.** *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\hat g$ ein erwartungstreuer Schätzer von $g(\vartheta) \in \mathbb{R}$ und $\vartheta_0 \in \Theta$. Dann gilt für jedes $\vartheta \in \Theta$ mit $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta_0}$, $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta_0}$, $\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} \in L^2(\mathbb{P}_{\vartheta_0})$*

$$\mathrm{Var}_{\vartheta_0}(\hat g) = \mathbb{E}_{\vartheta_0}[(\hat g - g(\vartheta_0))^2] \geqslant \frac{(g(\vartheta) - g(\vartheta_0))^2}{\mathrm{Var}_{\vartheta_0}(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}})}.$$

Beweis. Dies folgt wegen $\mathbb{E}_{\vartheta_0}[\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}] = 1$ aus

$$\begin{aligned} g(\vartheta) - g(\vartheta_0) &= \mathbb{E}_\vartheta[\hat{g} - g(\vartheta_0)] - \mathbb{E}_{\vartheta_0}[\hat{g} - g(\vartheta_0)] \\ &= \mathbb{E}_{\vartheta_0}\left[(\hat{g} - g(\vartheta_0))\Big(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} - 1\Big)\right] \\ &\leqslant \mathbb{E}_{\vartheta_0}[\big(\hat{g} - g(\vartheta_0)\big)^2]^{1/2}\, \mathbb{E}_{\vartheta_0}\left[\Big(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} - 1\Big)^2\right]^{1/2}, \end{aligned}$$

wobei zuletzt die Cauchy-Schwarz-Ungleichung angewendet wurde. □

2.30 Beispiele.

(a) Wir beobachten $X \sim \mathrm{Exp}(\vartheta)$ mit $\vartheta > 0$ unbekannt. Dann ist die Likelihoodfunktion gegeben durch $\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}(x) = (\vartheta/\vartheta_0)e^{-(\vartheta-\vartheta_0)x}$, $x \geqslant 0$. Diese ist in $L^2(\mathbb{P}_{\vartheta_0})$ nur im Fall $\vartheta > \vartheta_0/2$ und besitzt dann die Varianz $\mathrm{Var}_{\vartheta_0}(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}) = \frac{(\vartheta-\vartheta_0)^2}{\vartheta_0(2\vartheta-\vartheta_0)}$.

Im Fall erwartungstreuer Schätzer $\hat{g}$ für $g(\vartheta) = \vartheta$ ergibt die Chapman-Robbins-Gleichung $\mathbb{E}_{\vartheta_0}[(\hat{g}-g(\vartheta_0))^2] \geqslant \sup_{\vartheta>\vartheta_0/2} \vartheta_0(2\vartheta-\vartheta_0) = \infty$. Sofern wir also beliebig große Werte ϑ zulassen, existiert kein erwartungstreuer Schätzer von ϑ mit endlicher Varianz.

Im Fall $g(\vartheta) = \vartheta^{-1}$ hingegen liefert die Chapman-Robbins-Ungleichung $\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geqslant \sup_{\vartheta>\vartheta_0/2} \frac{2\vartheta-\vartheta_0}{\vartheta^2\vartheta_0} = \vartheta_0^{-2}$, und die Identität $\hat{g} = X$ erreicht auch diese Schranke.

(b) Wir beobachten $X \sim N(\vartheta, \frac{1}{n})$ mit $\vartheta \in \mathbb{R}$ unbekannt. Dann ist

$$\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} = \exp\Big(-\frac{n}{2}(X-\vartheta)^2 + \frac{n}{2}(X-\vartheta_0)^2\Big) = \exp\Big(n(\vartheta-\vartheta_0)X - \frac{n}{2}(\vartheta^2-\vartheta_0^2)\Big).$$

Wir berechnen mit $\tilde{\vartheta} := \vartheta_0 + 2(\vartheta - \vartheta_0)^2$ und $\mathbb{E}_{\vartheta_0}[\frac{d\mathbb{P}_{\tilde{\vartheta}}}{d\mathbb{P}_{\vartheta_0}}] = 1$

$$\begin{aligned} \mathbb{E}_{\vartheta_0}\left[\Big(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}\Big)^2\right] &= \mathbb{E}_{\vartheta_0}[\exp(2n(\vartheta-\vartheta_0)X - n(\vartheta^2-\vartheta_0)^2)] \\ &= \mathbb{E}_{\vartheta_0}\left[\frac{d\mathbb{P}_{\tilde{\vartheta}}}{d\mathbb{P}_{\vartheta_0}}\right] \exp\Big(\frac{n}{2}(\tilde{\vartheta}^2 - \vartheta_0^2) - n(\vartheta^2-\vartheta_0^2)\Big) \\ &= \exp\Big(\frac{n}{2}((\vartheta_0 + 2(\vartheta-\vartheta_0))^2 - \vartheta_0^2) - n(\vartheta^2-\vartheta_0^2)\Big) \\ &= \exp(n(\vartheta-\vartheta_0)^2). \end{aligned}$$

Die Chapman-Robbins-Schranke für erwartungstreue Schätzer $\hat{\vartheta}$ von ϑ (also $g(\vartheta) = \vartheta$) ist also

$$\mathrm{Var}_{\vartheta_0}(\hat{\vartheta}) \geqslant \sup_{\vartheta\neq\vartheta_0} \frac{(\vartheta-\vartheta_0)^2}{\exp(n(\vartheta-\vartheta_0)^2) - 1} = \frac{1}{n},$$

wobei das Supremum für $\vartheta \to \vartheta_0$ erhalten wird. Diese untere Schranke wird natürlich von $\hat{\vartheta} = X$ auch erreicht. Beachte, dass $\mathrm{Var}_{\vartheta_0}(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}})$ exponentiell wächst in $(\vartheta - \vartheta_0)^2$, was typisch ist und erklärt, warum es meist reicht, die Chapman-Robbins-Schranke für $\vartheta \to \vartheta_0$ zu betrachten.

2.31 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein von μ dominiertes Modell mit Likelihoodfunktion L. Das Modell heißt Hellinger-differenzierbar bei $\vartheta_0 \in \operatorname{int}(\Theta)$, wenn es einen Zufallsvektor $\dot\ell(\vartheta_0) \in L^2(\mathbb{P}_{\vartheta_0}; \mathbb{R}^k)$ gibt mit

$$\lim_{\vartheta\to\vartheta_0} \int \Big(\frac{\sqrt{L(\vartheta,x)} - \sqrt{L(\vartheta_0,x)} - \frac{1}{2}\langle \dot\ell(\vartheta_0,x), \vartheta-\vartheta_0\rangle \sqrt{L(\vartheta_0,x)}}{|\vartheta-\vartheta_0|} \Big)^2 d\mu(x) = 0.$$

Die Fisher-Informationsmatrix bei $\vartheta_0 \in \operatorname{int}(\Theta)$ ist gegeben durch

$$I(\vartheta_0) = \mathbb{E}_{\vartheta_0}[\dot\ell(\vartheta_0)\dot\ell(\vartheta_0)^\top].$$

Mit $\ell(\vartheta, x) := \log(L(\vartheta, x))$ ($\log 0 := -\infty$) wird die Loglikelihood-Funktion bezeichnet. Man nennt $\vartheta \mapsto \dot\ell(\vartheta)$ auch Score-Funktion.

2.32 Bemerkungen.

(a) Sofern alle folgenden Ausdrücke klassisch differenzierbar sind, gilt

$$\nabla_\vartheta \sqrt{L(\vartheta)} = \frac{\nabla_\vartheta L(\vartheta)}{2\sqrt{L(\vartheta)}} = \tfrac{1}{2}\sqrt{L(\vartheta)}\nabla_\vartheta \log(L(\vartheta)) = \tfrac{1}{2}\sqrt{L(\vartheta)}\dot\ell(\vartheta).$$

Insbesondere ist die Score-Funktion $\dot\ell$ die Ableitung der Loglikelihood-Funktion ℓ.

(b) Die Differenzierbarkeit im quadratischen $L^2(\mu)$-Mittel ist sehr viel allgemeiner und recht natürlich. Wegen $\sqrt{L(\vartheta)} \in L^2(\mu)$, was sofort aus $\int L(\vartheta)\, d\mu = 1 < \infty$ folgt, kann man $\vartheta \mapsto \sqrt{L(\vartheta)}$ als $L^2(\mu)$-wertige Abbildung auffassen, so dass die Verteilungen $(\mathbb{P}_\vartheta)$ im geometrischen Sinne eine Untermannigfaltigkeit des Hilbertraums $L^2(\mu)$ bilden. Insbesondere gilt

$$\mathbb{E}_{\vartheta_0}[|\dot\ell(\vartheta_0)|^2] = \int |\dot\ell(\vartheta_0)(x)|^2 L(\vartheta_0,x)\mu(dx) = \int |\dot\ell(\vartheta_0)(x)\sqrt{L(\vartheta_0,x)}|^2 \mu(dx),$$

so dass $\dot\ell(\vartheta) \in L^2(\mathbb{P}_{\vartheta_0}; \mathbb{R}^k) \iff \dot\ell(\vartheta)\sqrt{L(\vartheta_0)} \in L^2(\mu; \mathbb{R}^k)$ und das Integral bei der Definition von Hellinger-Differenzierbarkeit wohldefiniert ist.

(c) Nach Definition ist die Fisher-Informationsmatrix symmetrisch. Wegen $\langle I(\vartheta_0)v, v\rangle = \mathbb{E}_{\vartheta_0}[\langle \dot\ell(\vartheta_0), v\rangle^2] \geqslant 0$ für beliebige $v \in \mathbb{R}^k$ ist die Fisher-Informationsmatrix auch stets positiv-semidefinit.

(d) Die Score-Funktion und die Fisher-Information sind unabhängig vom dominierenden Maß; denn mit einem privilegierten dominierenden Maß $\mathbb{Q}$ gilt $L(\vartheta) = \frac{d\mathbb{P}_\vartheta}{d\mathbb{Q}}\frac{d\mathbb{Q}}{d\mu}$, so dass in der Definition von $\dot\ell$ der Faktor $\frac{d\mathbb{Q}}{d\mu}$ aus dem Integranden ausgeklammert werden kann und somit $\dot\ell$ ebenso die Definition bezüglich dem dominierenden Maß $\mathbb{Q}$ erfüllt.

2.33 Lemma. *Für alle $\vartheta \in \Theta \subseteq \mathbb{R}^k$ in einer Umgebung von $\vartheta_0 \in \Theta$ gelte $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta_0}$ sowie die $L^2(\mathbb{P}_{\vartheta_0})$-Differenzierbarkeit der Likelihoodfunktion $L_{\vartheta_0}(\vartheta, x) := \frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}(x)$ bei ϑ_0, d.h. mit dem Gradienten $\dot{L}_{\vartheta_0}(\vartheta_0) \in L^2(\mathbb{P}_{\vartheta_0}; \mathbb{R}^k)$ gilt*

$$\lim_{\vartheta \to \vartheta_0} \mathbb{E}_{\vartheta_0} \Big[\Big(\frac{L_{\vartheta_0}(\vartheta) - L_{\vartheta_0}(\vartheta_0) - \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle}{|\vartheta - \vartheta_0|} \Big)^2 \Big] = 0.$$

Dann ist $(\mathbb{P}_\vartheta)$ Hellinger-differenzierbar bei ϑ_0 mit $\dot{\ell}(\vartheta_0) = \dot{L}_{\vartheta_0}(\vartheta_0)$.

Beweis. Aus obiger Bemerkung folgt, dass es genügt, $\mathbb{P}_{\vartheta_0}$ als dominierendes Maß und die Likelihoodfunktion $L_{\vartheta_0}(\vartheta)$ zu betrachten. Wir erhalten mit $L_{\vartheta_0}(\vartheta_0) = 1$ und der Minkowski-Ungleichung in $L^2(\mathbb{P}_{\vartheta_0})$:

$$\begin{aligned}
&\Big\| \sqrt{L_{\vartheta_0}(\vartheta)} - 1 - \tfrac{1}{2} \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle \Big\|_{L^2(\mathbb{P}_{\vartheta_0})} \\
&\leqslant \Big\| \frac{L_{\vartheta_0}(\vartheta) - 1 - \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle}{\sqrt{L_{\vartheta_0}(\vartheta)} + 1} \Big\|_{L^2(\mathbb{P}_{\vartheta_0})} + \Big\| \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle \Big(\frac{1}{\sqrt{L_{\vartheta_0}(\vartheta)} + 1} - \frac{1}{2} \Big) \Big\|_{L^2(\mathbb{P}_{\vartheta_0})} \\
&\leqslant \Big\| L_{\vartheta_0}(\vartheta) - 1 - \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle \Big\|_{L^2(\mathbb{P}_{\vartheta_0})} + \frac{|\vartheta - \vartheta_0|}{2} \Big\| |\dot{L}_{\vartheta_0}(\vartheta_0)| \frac{1 - \sqrt{L_{\vartheta_0}(\vartheta)}}{\sqrt{L_{\vartheta_0}(\vartheta)} + 1} \Big\|_{L^2(\mathbb{P}_{\vartheta_0})}.
\end{aligned}$$

Nach Vorausetzung besitzt der erste Summand die Ordnung $o(|\vartheta - \vartheta_0|)$. Außerdem gilt insbesondere $L_{\vartheta_0}(\vartheta) \to 1$ in $L^2(\mathbb{P}_{\vartheta_0})$ und damit auch in $\mathbb{P}_{\vartheta_0}$-Wahrscheinlichkeit. Weil nun $G(x) := \frac{1-\sqrt{x}}{\sqrt{x}+1}$ für $x \geqslant 0$ im Betrag durch 1 beschränkt ist und $\lim_{x \to 1} G(x) = 0$ gilt, folgt mit dominierter Konvergenz (unter stochastischer Konvergenz), dass die zweite $L^2(\mathbb{P}_{\vartheta_0})$-Norm gegen Null konvergiert. Damit ist der gesamte Ausdruck von der Ordnung $o(|\vartheta - \vartheta_0|)$. □

2.34 Beispiele.

(a) Es sei $X_1, \ldots, X_n$ eine mathematische Stichprobe gemäß der Lebesguedichte $f_\vartheta(x) = \frac{1}{2\sigma} e^{-|x-\vartheta|/\sigma}$, $x \in \mathbb{R}$, $\sigma > 0$ bekannt und $\vartheta \in \mathbb{R}$ unbekannt. Für beliebige $\vartheta_0, \vartheta \in \mathbb{R}$ gilt

$$L_{\vartheta_0}(\vartheta) = \exp\Big(- \sum_{i=1}^n (|X_i - \vartheta| - |X_i - \vartheta_0|)/\sigma \Big)$$

und L_{ϑ_0} ist $L^2(\mathbb{P}_{\vartheta_0})$-differenzierbar (Nachweis!) mit

$$\dot{L}_{\vartheta_0}(\vartheta_0) = \dot{\ell}(\vartheta_0) = \sum_{i=1}^n (\mathbf{1}(X_i - \vartheta_0 > 0) - \mathbf{1}(X_i - \vartheta_0 < 0))/\sigma.$$

Die Fisher-Information ist

$$I(\vartheta_0) = \sum_{i=1}^n \mathrm{Var}_{\vartheta_0} \Big(\mathbf{1}(X_i - \vartheta_0 > 0) - \mathbf{1}(X_i - \vartheta_0 < 0) \Big) \sigma^{-2} = n\sigma^{-2}.$$

Beachte, dass der eher seltene Fall vorliegt, dass die Fisher-Information nicht vom unbekannten Parameter abhängt.

(b) Es sei $f(x) = \frac{1}{2}\Gamma(a)^{-1}|x|^{a-1}e^{-|x|}$, $x \in \mathbb{R}$, eine zweiseitige $\Gamma(a,1)$-Dichte für festes $a > 0$ und $X_1, \dots, X_n$ eine gemäß $f(\bullet - \vartheta)$-verteilte mathematische Stichprobe mit $\vartheta \in \mathbb{R}$ unbekannt. Bemerke, dass sich Beispiel (a) mit $\sigma = 1$ im Spezialfall $a = 1$ ergibt. Dann ist

$$L_{\vartheta_0}(\vartheta) = \Big(\prod_{i=1}^n \frac{|X_i - \vartheta|}{|X_i - \vartheta_0|}\Big)^{a-1} \exp\Big(-\sum_{i=1}^n (|X_i - \vartheta| - |X_i - \vartheta_0|)\Big),$$

$$\dot{L}_{\vartheta_0}(\vartheta_0) = \sum_{i=1}^n \Big(((a-1)|X_i - \vartheta_0|^{-1} - 1)\operatorname{sgn}(\vartheta_0 - X_i)\Big).$$

Wegen der exponentiell abfallenden Dichte gilt $L_{\vartheta_0}(\vartheta) \in L^2(\mathbb{P}_{\vartheta_0})$ genau dann, wenn $\int_{[-K,K]^n} (\prod_{i=1}^n \frac{|x_i-\vartheta|^2}{|x_i-\vartheta_0|})^{a-1} dx < \infty$ für alle $K > 0$, also genau dann, wenn $2(a-1) > -1$ und $a - 1 < 1$, also $a \in (1/2, 2)$. $\dot{L}_{\vartheta_0}(\vartheta_0)$ liegt genau dann in $L^2(\mathbb{P}_{\vartheta_0})$, wenn $a - 3 > -1$ oder $a = 1$, also $a \in \{1\} \cup (2, \infty)$ gilt. Dieses Modell ist demnach im obigen Sinn $L^2(\mathbb{P}_{\vartheta_0})$-differenzierbar genau im Fall $a = 1$, jedoch ist es für alle $a \in \{1\} \cup (2, \infty)$ Hellinger-differenzierbar. Allgemein erfordert Hellinger-Differenzierbarkeit in einem solchen Lokationsmodell, dass jede Nullstelle x_0 von f eine Ordnung größer eins besitzt im Sinn von $\limsup_{x \to x_0} f(x)|x - x_0|^{-\gamma} < \infty$ für ein $\gamma > 1$.

2.35 Satz (Cramér-Rao-Schranke)**.** *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein statistisches Modell, $g : \Theta \to \mathbb{R}$ besitze bei $\vartheta_0 \in \mathrm{int}(\Theta)$ die Ableitung $\dot{g}(\vartheta_0)$ und $\hat{g}$ sei ein erwartungstreuer Schätzer von $g(\vartheta)$. Für alle ϑ in einer Umgebung von ϑ_0 gelte $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta_0}$ sowie die $L^2(\mathbb{P}_{\vartheta_0})$-Differenzierbarkeit der Likelihoodfunktion $L_{\vartheta_0}(\vartheta) := \frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}$ bei ϑ_0. Falls die Fisher-Informationsmatrix $I(\vartheta_0)$ strikt positiv-definit ist, gilt die Cramér-Rao-Ungleichung als untere Schranke für das quadratische Risiko*

$$\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] = \mathrm{Var}_{\vartheta_0}(\hat{g}) \geqslant \langle I(\vartheta_0)^{-1}\dot{g}(\vartheta_0), \dot{g}(\vartheta_0)\rangle.$$

Beweis. Zunächst beachte die Hellinger-Differenzierbarkeit des Modells by ϑ_0 und betrachte $\vartheta_h = \vartheta_0 + hv$ mit $h \downarrow 0$ und $v \in \mathbb{R}^k$, $v \neq 0$. Dann folgt aus der Chapman-Robbins-Ungleichung, $L_{\vartheta_0}(\vartheta_0) = 1 = \mathbb{E}_{\vartheta_0}[L_{\vartheta_0}(\vartheta)]$ und $\dot{\ell}(\vartheta_0) = \dot{L}_{\vartheta_0}(\vartheta_0)$

$$\mathbb{E}_{\vartheta_0}\Big[(\hat{g} - g(\vartheta_0))^2\Big] \geqslant \limsup_{h \downarrow 0} \frac{((g(\vartheta_h) - g(\vartheta_0))/h)^2}{\mathbb{E}_{\vartheta_0}[((L_{\vartheta_0}(\vartheta_h) - L_{\vartheta_0}(\vartheta_0))/h)^2]} = \frac{(\langle \dot{g}(\vartheta_0), v\rangle)^2}{\langle I(\vartheta_0)v, v\rangle}.$$

Das Supremum der rechten Seite über Richtungen v wird bei $v = I(\vartheta_0)^{-1}\dot{g}(\vartheta_0)$ angenommen, was die Behauptung zeigt. □

2.36 Bemerkungen.

(a) Die Cramér-Rao-Schranke zeigt, dass es umso schwieriger ist $g(\vartheta)$ zu schätzen, je stärker g variiert (d.h. bei großem $|\dot{g}(\vartheta)|$) und je kleiner die Fisher-Information ist. Aus der Definition sieht man, dass die Fisher-Information klein ist, wenn die Likelihood-Funktion wenig variiert, also die Verteilung der Beobachtungen $\mathbb{P}_\vartheta$ sehr nahe bei $\mathbb{P}_{\vartheta_0}$ liegt für ϑ nahe bei ϑ_0.

(b) Ist $\hat{g}$ kein erwartungstreuer Schätzer von $g(\vartheta)$, so doch von $\gamma(\vartheta) := \mathbb{E}_\vartheta[\hat{g}]$ (so existent). Mit der Bias-Varianz-Zerlegung liefert die Cramér-Rao-Ungleichung bei Existenz von $\dot{\gamma}(\vartheta_0)$ für diesen Fall

$$\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geqslant (g(\vartheta_0) - \gamma(\vartheta_0))^2 + \langle I(\vartheta_0)^{-1}\dot{\gamma}(\vartheta_0), \dot{\gamma}(\vartheta_0)\rangle.$$

Beachte dazu auch, dass erwartungstreue Schätzer von $g(\vartheta)$ nicht existieren müssen bzw. oftmals keine weiteren erstrebenswerten Eigenschaften besitzen.

2.37 Lemma. *Bildet $(\mathbb{P}_\vartheta)$ eine Exponentialfamilie in T mit natürlichem Parameterbereich Θ, so ist $(\mathbb{P}_\vartheta)$ im Innern von Θ L^2- und Hellinger-differenzierbar mit Fisher-Information $I(\vartheta) = \ddot{A}(\vartheta)$ (Notation aus Satz 2.11).*

Sofern $I(\vartheta_0)$ strikt positiv-definit ist, erreicht T_i, $i = 1, \ldots, k$, als erwartungstreuer Schätzer von $g_i(\vartheta) = \mathbb{E}_\vartheta[T_i]$ die Cramér-Rao-Schranke (ist Cramer-Rao-effizient) bei $\vartheta_0 \in int(\Theta)$.

Beweis. Nach Satz 2.11 gilt $g(\vartheta) = E_\vartheta[T] = \dot{A}(\vartheta)$ und $\mathrm{Cov}_\vartheta(T) = \ddot{A}(\vartheta)$ (Kovarianzmatrix). Andererseits ist die Loglikelihoodfunktion $\ell_{\vartheta_0}(\vartheta) = \langle \vartheta - \vartheta_0, T\rangle - (A(\vartheta) - A(\vartheta_0))$, so dass die Scorefunktion $\dot{\ell}_{\vartheta_0}(\vartheta) = T - \dot{A}(\vartheta)$ im klassischen Sinn existiert und

$$I(\vartheta) = \mathbb{E}_\vartheta[(\dot{l}(\vartheta))(\dot{l}(\vartheta))^\top] = \mathrm{Var}_\vartheta(T) = \ddot{A}(\vartheta)$$

gelten sollte. Da $L_{\vartheta_0}(\vartheta) = \exp(\langle \vartheta - \vartheta_0, T\rangle - A(\vartheta) + A(\vartheta_0))$ klassisch differenzierbar ist, folgt die $L^2(\mathbb{P}_{\vartheta_0})$-Differenzierbarkeit bei ϑ_0 sofern Integration und Grenzwert vertauscht werden dürfen, was wie in Satz 2.11 nachgewiesen wird und die Korrektheit der obigen Rechnungen bestätigt.

Wegen $\dot{g}_i(\vartheta_0) = (\ddot{A}_{ij}(\vartheta_0))_j =: \ddot{A}_{i\bullet}(\vartheta_0)$ ist die Cramér-Rao-Schranke gerade

$$\langle \ddot{A}(\vartheta_0)^{-1}\ddot{A}_{i\bullet}(\vartheta_0), \ddot{A}_{i\bullet}(\vartheta_0)\rangle = \langle e_i, \ddot{A}_{i\bullet}(\vartheta_0)\rangle = \ddot{A}_{ii}(\vartheta_0),$$

was gleich der Varianz von T_i unter $\mathbb{P}_{\vartheta_0}$ ist. □

2.38 Beispiel. Es sei $X_1, \ldots, X_n$ eine $N(\mu, \sigma^2)$-verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}$ unbekannt und $\sigma > 0$ bekannt. Zur erwartungstreuen Schätzung von μ betrachte $\hat{\mu} = \bar{X}$. Dann gilt $\mathrm{Var}_\mu(\hat{\mu}) = \sigma^2/n$ sowie für die Fisher-Information $I(\mu) = n/\sigma^2$ (beachte $A(\mu) = \frac{n\mu^2}{2\sigma^2}$, $\ddot{A}(\mu) = n/\sigma^2$). Also ist $\hat{\mu}$ effizient im Sinne der Cramér-Rao-Ungleichung. Um nun μ^2 zu schätzen, betrachte den erwartungstreuen (!) Schätzer $\widehat{\mu^2} = (\bar{X})^2 - \sigma^2/n$. Es gilt $\mathrm{Var}_\mu(\widehat{\mu^2}) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}$, während die Cramér-Rao-Ungleichung die untere Schranke $\frac{4\mu^2\sigma^2}{n}$ liefert. Damit ist $\widehat{\mu^2}$ nicht Cramér-Rao-effizient. Allerdings ist $\bar{X}$ eine suffiziente und vollständige Statistik, so dass der Satz von Lehmann-Scheffé zeigt, dass $\widehat{\mu^2}$ minimale Varianz unter allen erwartungstreuen Schätzern besitzt. Demnach ist die Cramér-Rao-Schranke hier nicht scharf.

2.39 Bemerkung. In der Tat wird die Cramér-Rao-Schranke nur erreicht, wenn $(\mathbb{P}_\vartheta)$ eine Exponentialfamilie in T bildet und $g(\vartheta) = \mathbb{E}_\vartheta[T]$ oder eine lineare Funktion davon zu schätzen ist. Wegen der Vollständigkeit der Statistik

T könnte man in diesen Fällen alternativ auch mit dem Satz von Lehmann-Scheffé argumentieren. Später werden wir sehen, dass in allgemeineren Modellen immerhin asymptotisch die Cramér-Rao-Schranke erreichbar ist.

2.40 Lemma. *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein bei $\vartheta_0 \in \Theta$ Hellinger-differenzierbares statistisches Modell. Dann ist die Likelihood-Funktion L im $L^1(\mu)$-Sinn differenzierbar mit Ableitung $\dot\ell(\vartheta)L(\vartheta)$, und es gilt $\mathbb{E}_{\vartheta_0}[\dot\ell(\vartheta_0)] = 0$.*

Beweis. Betrachte den Zähler im Kriterium für $L^1(\mu)$-Differenzierbarkeit mit $\dot L(\vartheta_0) = \dot\ell(\vartheta_0)L(\vartheta_0)$:

$$\begin{aligned}
&\Big\|L(\vartheta) - L(\vartheta_0) - \langle\dot\ell(\vartheta_0), \vartheta-\vartheta_0\rangle L(\vartheta_0)\Big\|_{L^1(\mu)}\\
&\leqslant \Big\|\Big(\sqrt{L(\vartheta)} - \sqrt{L(\vartheta_0)} - \tfrac12\langle\dot\ell(\vartheta_0), \vartheta-\vartheta_0\rangle\sqrt{L(\vartheta_0)}\Big)\Big(\sqrt{L(\vartheta)} + \sqrt{L(\vartheta_0)}\Big)\Big\|_{L^1(\mu)}\\
&+ \frac12\Big\|\Big(\langle\dot\ell(\vartheta_0), \vartheta-\vartheta_0\rangle\sqrt{L(\vartheta_0)}\Big)\Big(\sqrt{L(\vartheta)} - \sqrt{L(\vartheta_0)}\Big)\Big\|_{L^1(\mu)}.
\end{aligned}$$

Im ersten Ausdruck konvergiert der erste Faktor nach Voraussetzung in $L^2(\mu)$ mit der Ordnung $o(|\vartheta-\vartheta_0|)$ gegen Null und der zweite Faktor in $L^2(\mu)$ gegen $2\sqrt{L(\vartheta_0)}$. Mit der Cauchy-Schwarz-Ungleichung folgt also, dass dieser Ausdruck von der Ordnung $o(|\vartheta-\vartheta_0|)$ ist. Im zweiten Ausdruck besitzt der erste Faktor eine $L^2(\mu)$-Norm der Ordnung $O(|\vartheta-\vartheta_0|)$, während der zweite Faktor in $L^2(\mu)$ gegen Null konvergiert. Damit ist der gesamte Term von der Ordnung $o(|\vartheta-\vartheta_0|)$ und L somit $L^1(\mu)$-differenzierbar bei ϑ_0.

Aus L^1-Konvergenz folgt Konvergenz der entsprechenden Integrale. Wegen $\int(L(\vartheta,x) - L(\vartheta_0,x))\,d\mu(x) = 1-1 = 0$ schließen wir durch Einsetzen von $\vartheta = \vartheta_0 + he_i$ ($h\to 0$, e_i i-ter Einheitsvektor) $0 = \int\langle\dot\ell(\vartheta_0), e_i\rangle L(\vartheta_0)\,d\mu(x) = \mathbb{E}_{\vartheta_0}[\dot\ell_i(\vartheta_0)]$ für alle $i = 1,\dots,k$. □

2.41 Lemma. *Es seien $X_1,\dots,X_n$ Beobachtungen aus unabhängigen Hellinger-differenzierbaren Modellen $\mathcal{E}_1,\dots,\mathcal{E}_n$ mit derselben Parametermenge $\Theta\subseteq\mathbb{R}^k$. Bezeichnet I_j die entsprechende Fisher-Information, erzeugt von der Beobachtung X_j, so ist das Produktmodell, erzeugt von $X_1,\dots,X_n$, Hellinger-differenzierbar mit Fisher-Information*

$$\forall\vartheta\in\Theta:\ I(\vartheta) = \sum_{j=1}^n I_j(\vartheta).$$

Beweis. Nach Annahme sind die entsprechenden Likelihoodfunktionen $L_1,\dots,L_n$ bezüglich der dominierenden Maße $\mu_1,\dots,\mu_n$ Hellinger-differenzierbar mit Score-Funktionen $\dot\ell_1,\dots,\dot\ell_n$. Also ist auch die gemeinsame Likelihoodfunktion $L(\vartheta,x) = \prod_{j=1}^n L_j(\vartheta,x_j)$ bezüglich $\mu = \mu_1\otimes\cdots\otimes\mu_n$ Hellinger-differenzierbar mit Score-Funktion $\dot\ell(\vartheta,x) = \sum_{j=1}^n \dot\ell_j(\vartheta,x_j)$, wie für

$n = 2$ mit dem Satz von Fubini folgt:

$$\begin{aligned}
&\Big\| \sqrt{L_1(\vartheta)L_2(\vartheta)} - \sqrt{L_1(\vartheta_0)L_2(\vartheta_0)} - \frac{1}{2}\langle \dot\ell_1(\vartheta) + \dot\ell_2(\vartheta), \vartheta - \vartheta_0\rangle \sqrt{L_1(\vartheta_0)L_2(\vartheta_0)} \Big\|_{L^2(\mu)} \\
&\leqslant \|\sqrt{L_1(\vartheta_0)}\|_{L^2(\mu_1)} \Big\| \sqrt{L_2(\vartheta)} - \sqrt{L_2(\vartheta_0)} - \frac{1}{2}\langle \dot\ell_2(\vartheta), \vartheta - \vartheta_0\rangle \sqrt{L_2(\vartheta_0)} \Big\|_{L^2(\mu_2)} \\
&+ \|\sqrt{L_2(\vartheta_0)}\|_{L^2(\mu_2)} \Big\| \sqrt{L_1(\vartheta)} - \sqrt{L_1(\vartheta_0)} - \frac{1}{2}\langle \dot\ell_1(\vartheta), \vartheta - \vartheta_0\rangle \sqrt{L_1(\vartheta_0)} \Big\|_{L^2(\mu_1)} \\
&+ \|\sqrt{L_2(\vartheta)} - \sqrt{L_2(\vartheta_0)}\|_{L^2(\mu_2)} \|\sqrt{L_1(\vartheta)} - \sqrt{L_1(\vartheta_0)}\|_{L^2(\mu_1)} \\
&= o(|\vartheta - \vartheta_0|) + o(|\vartheta - \vartheta_0|) + O(|\vartheta - \vartheta_0|^2).
\end{aligned}$$

Für allgemeine $n \geqslant 2$ verwende vollständige Induktion.

Wegen Unabhängigkeit der $X_1, \dots, X_n$ sowie $\mathbb{E}_\vartheta[\dot\ell_j(\vartheta, X_j)] = 0$ gilt daher

$$\begin{aligned}
&\mathbb{E}_\vartheta \Big[\dot\ell(\vartheta, (X_1, \dots, X_n)) \dot\ell(\vartheta, (X_1, \dots, X_n))^\top \Big] \\
&= \mathbb{E}_\vartheta \Big[\Big(\sum_{j=1}^n \dot\ell_j(\vartheta, X_j) \Big) \Big(\sum_{j=1}^n \dot\ell_j(\vartheta, X_j)^\top \Big) \Big] \\
&= \sum_{j,m=1}^n \mathbb{E}_\vartheta \Big[\dot\ell_j(\vartheta, X_j) \dot\ell_m(\vartheta, X_m)^\top \Big] = \sum_{j=1}^n \mathbb{E}_\vartheta \Big[\dot\ell_j(\vartheta, X_j) \dot\ell_j(\vartheta, X_j)^\top \Big].
\end{aligned}$$

□

2.42 Bemerkung. Das Lemma zeigt also, dass die Fisher-Information unter unabhängigen Beobachtungen additiv ist, so dass sie bei einer mathematischen Stichprobe gerade gleich dem Stichprobenumfang n mal der Fisher-Information bei einer Beobachtung ist.

2.6 Äquivarianz

Mit dem Satz von Lehmann-Scheffé und der Cramér-Rao-Schranke konnten wir beste erwartungstreue Schätzer unter quadratischem Risiko verstehen. Statt Erwartungstreue ist es oft angemessen, gewisse Invarianzeigenschaften von Schätzern zu fordern. Dies führt auf den Begriff der Äquivarianz, der allgemein für Gruppenoperationen existiert, aber hier nur im Fall einer einparametrigen Translationsinvarianz dargestellt wird.

2.43 Definition. Im Lokationsmodell beobachten wir einen $\mathbb{R}^n$-wertigen Zufallsvektor $X = (X_1, \dots, X_n)$, welcher eine gemeinsame Verteilung $\mathbb{P}_\vartheta$ mit Lebesgue-Dichte $f(x_1 - \vartheta, \dots, x_n - \vartheta)$, $(x_1, \dots, x_n) \in \mathbb{R}^n$, besitzt, wobei f bekannt und $\vartheta \in \mathbb{R}$ ein unbekannter Lokationsparameter ist.

2.44 Bemerkung. Gilt $\mathbb{E}_0[X_i] = 0$, so folgt $\mathbb{E}_\vartheta[X_i] = \vartheta$, $i = 1, \dots, n$, und der unbekannte Parameter ϑ ist gerade der Erwartungswert der Beobachtungen.

2.45 Definition. In Lokationsmodell heißt ein Schätzer $\hat\vartheta : \mathbb{R}^n \to \mathbb{R}$ äquivariant, wenn gilt

$$\forall x_1, \dots, x_n \in \mathbb{R}\ \forall \vartheta \in \mathbb{R} :\ \hat\vartheta(x_1 + \vartheta, \dots, x_n + \vartheta) = \hat\vartheta(x_1, \dots, x_n) + \vartheta$$

und die Verlustfunktion l invariant ist, also $l(\vartheta, a) = \ell(\vartheta - a)$ gilt mit $\ell : \mathbb{R} \to [0, \infty)$ messbar.

2.46 Lemma. *Ist $\hat{\vartheta}$ äquivarianter Schätzer und l eine invariante Verlustfunktion, so besitzt $\hat{\vartheta}$ konstantes Risiko: $R(\vartheta, \hat{\vartheta}) = R(0, \hat{\vartheta})$ für alle $\vartheta \in \mathbb{R}$.*

Beweis. Es gilt

$$\begin{aligned} R(\vartheta, \hat{\vartheta}) &= \mathbb{E}_\vartheta[\ell(\hat{\vartheta}(X_1, \ldots, X_n) - \vartheta)] \\ &= \mathbb{E}_0[\ell(\hat{\vartheta}(X_1 + \vartheta, \ldots, X_n + \vartheta) - \vartheta)] \\ &= \mathbb{E}_0[\ell(\hat{\vartheta}(X_1, \ldots, X_n) + \vartheta - \vartheta)] = R(0, \hat{\vartheta}), \end{aligned}$$

wobei wir in der dritten Gleichheit die Äquivarianz von $\hat{\vartheta}$ verwendet haben. □

2.47 Definition. Ein äquivarianter Schätzer $\hat{\vartheta}$ heißt bester äquivarianter Schätzer (MRIE: minimum risk invariant estimator) im Lokationsmodell, falls

$$R(0, \hat{\vartheta}) = \inf_{\tilde{\vartheta} \text{ äquivariant}} R(0, \tilde{\vartheta})$$

gilt, wobei sich das Infimum über alle äquivarianten Schätzer $\tilde{\vartheta}$ erstreckt.

2.48 Satz. *Sei $\hat{\vartheta}_0$ ein äquivarianter Schätzer und sei $T(x_1, \ldots, x_n) := (x_1 - x_n, \ldots, x_{n-1} - x_n) \in \mathbb{R}^{n-1}$, $(x_1, \ldots, x_n) \in \mathbb{R}^n$.*

(a) Ein Schätzer $\hat{\vartheta}$ ist genau dann äquivariant, wenn es eine messbare Funktion $u : \mathbb{R}^{n-1} \to \mathbb{R}$ gibt mit $\hat{\vartheta}(x) = \hat{\vartheta}_0(x) - u(T(x))$, $x \in \mathbb{R}^n$.

(b) Es gelte zusätzlich $\mathbb{E}_0[\hat{\vartheta}_0^2] < \infty$. Dann ist durch $\hat{\vartheta} := \hat{\vartheta}_0 - \mathbb{E}_0[\hat{\vartheta}_0 \,|\, T]$ ein bester äquivarianter Schätzer bei quadratischem Verlust gegeben.

Beweis. Für (a) folgt durch Einsetzen, dass $\hat{\vartheta}_0 + u(T)$ äquivariant ist. Andersherum folgt aus der Äquivarianz von $\hat{\vartheta}$ und $\hat{\vartheta}_0$

$$\forall x_1, \ldots, x_n, \vartheta \in \mathbb{R} : \ (\hat{\vartheta} - \hat{\vartheta}_0)(x_1 + \vartheta, \ldots, x_n + \vartheta) = (\hat{\vartheta} - \hat{\vartheta}_0)(x_1, \ldots, x_n).$$

Mit $\vartheta = -x_n$ und $u(t_1, \ldots, t_{n-1}) = (\hat{\vartheta} - \hat{\vartheta}_0)(t_1, \ldots, t_{n-1}, 0)$ folgt die behauptete Darstellung in (a).

Jeder äquivariante Schätzer $\tilde{\vartheta}$ mit der Darstellung in (a) erfüllt $R(0, \tilde{\vartheta}) = \mathbb{E}_0[(\hat{\vartheta}_0 - u(T))^2]$ für eine messbare Funktion u von T. Nach Charakterisierung der bedingten Erwartung wird dies durch $u(T) = \mathbb{E}_0[\hat{\vartheta}_0 | \, T]$ minimiert und nach (a) ist $\hat{\vartheta} = \hat{\vartheta}_0 - \mathbb{E}_0[\hat{\vartheta}_0 | \, T]$ wiederum äquivariant. □

2.49 Korollar. *Im Lokationsmodell gelte $\mathbb{E}_0[X_n^2] < \infty$. Dann ist der beste äquivariante Schätzer bezüglich quadratischem Risiko gegeben durch den Pitman-Schätzer*

$$\hat{\vartheta} = \frac{\int_{-\infty}^{\infty} z f(X_1 - z, \ldots, X_n - z)\, dz}{\int_{-\infty}^{\infty} f(X_1 - z, \ldots, X_n - z)\, dz}.$$

Beweis. Der Schätzer $\hat{\vartheta}_0 = X_n$ ist äquivariant mit $\mathbb{E}_0[\hat{\vartheta}_0^2] < \infty$ nach Voraussetzung. Aus dem Satz folgt daher, dass $\hat{\vartheta} = X_n - \mathbb{E}_0[X_n \,|\, T]$ bester äquivarianter Schätzer ist. Nach dem Dichtetransformationssatz besitzt (T, X_n) die gemeinsame Lebesguedichte

$$f^{(T,X_n)}(t, x_n) = f(t_1 + x_n, \ldots, t_{n-1} + x_n, x_n), \quad t \in \mathbb{R}^{n-1}, x_n \in \mathbb{R}\,.$$

Nach der Bayesformel erhalten wir die bedingte Dichte

$$f^{X_n|T=t}(x_n) = \frac{f(t_1 + x_n, \ldots, t_{n-1} + x_n, x_n)}{\int_{-\infty}^{\infty} f(t_1 + \xi, \ldots, t_{n-1} + \xi, \xi)\, d\xi}, \quad x_n \in \mathbb{R} \qquad \text{für } \mathbb{P}_0^T\text{-f.a. } t.$$

Wir erhalten die bedingte Erwartung durch Integration und substituieren $z = X_n - x_n$ (für jede Realisierung von X_n):

$$\begin{aligned} X_n - \mathbb{E}_0[X_n \,|\, T] &= \frac{\int_{-\infty}^{\infty} (X_n - x_n) f(T_1 + x_n, \ldots, T_{n-1} + x_n, x_n)\, dx_n}{\int_{-\infty}^{\infty} f(T_1 + x_n, \ldots, T_{n-1} + x_n, x_n)\, dx_n} \\ &= \frac{\int_{-\infty}^{\infty} z f(T_1 + X_n - z, \ldots, T_{n-1} + X_n - z, X_n - z)\, dz}{\int_{-\infty}^{\infty} f(T_1 + X_n - z, \ldots, T_{n-1} + X_n - z, X_n - z)\, dz} \\ &= \frac{\int_{-\infty}^{\infty} z f(X_1 - z, \ldots, X_{n-1} - z, X_n - z)\, dz}{\int_{-\infty}^{\infty} f(X_1 - z, \ldots, X_{n-1} - z, X_n - z)\, dz}. \end{aligned}$$

Dies zeigt die Behauptung. □

2.50 Beispiel. Im Lokationsmodell mit Produktdichte $f(x) = \prod_{i=1}^n f_1(x_i)$ kann man den Pitman-Schätzer $\hat{\vartheta}$ in folgenden Fällen leicht berechnen (Übung!):

(a) $f_1(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, $\hat{\vartheta} = \bar{X}$;

(b) $f_1(x) = a^{-1}\mathbf{1}_{[-\frac{a}{2}, \frac{a}{2}]}(x)$ für $a > 0$, $\hat{\vartheta} = \frac{X_{(1)} + X_{(n)}}{2}$;

(c) $f_1(x) = \lambda e^{-\lambda(x+1)}\mathbf{1}_{[-1,\infty)}(x)$ für $\lambda > 0$, $\hat{\vartheta} = X_{(1)} + 1 - \frac{1}{n\lambda}$.

Für $\sigma = 1$, $a = \sqrt{3}$ und $\lambda = 1$ gilt in allen drei Fällen, dass f_1 Erwartungswert 0 und Varianz 1 besitzt. Weil $\bar{X}$ äquivariant ist jeweils mit $\mathbb{E}_\vartheta[(\bar{X} - \vartheta)^2] = \frac{1}{n}$, muss das quadratische Risiko von $\hat{\vartheta}$ in (b) und (c) kleiner als $\frac{1}{n}$ sein. Allgemein ist unter allen Dichten f_1 mit Erwartungswert μ und endlicher Varianz σ^2 das quadratische Risiko eines besten äquivarianten Schätzers maximal bei der Normalverteilung $N(\mu, \sigma^2)$. Die Normalverteilung ist also ungünstigste Verteilung in dieser Klasse.

2.51 Bemerkungen.

(a) Man kann die Bedingung $\mathbb{E}_0[X_n^2] < \infty$ durch die Existenz eines äquivarianten Schätzers mit endlichem quadratischen Risiko ersetzen, vergleiche das Skript von Martin Wahl.

(b) Der Pitman-Schätzer kann auch als *uneigentlicher Bayes-Schätzer* verstanden werden mit dem translationsinvarianten Lebesguemaß als a-priori-Verteilung für ϑ. Die a-posteriori-Dichte ist dann

$$f^{T|X=x}(\vartheta) = \frac{f(x_1-\vartheta,\dots,x_n-\vartheta)}{\int_{-\infty}^{\infty} f(x_1-\vartheta',\dots,x_n-\vartheta')\,d\vartheta'}, \quad \vartheta\in\mathbb{R},$$

welche nach obiger Herleitung existiert. Der Bayesschätzer bezüglich quadratischem Risiko ergibt sich als bedingte Erwartung

$$\hat{\vartheta} = \frac{\int_{-\infty}^{\infty} \vartheta f(X_1-\vartheta,\dots,X_n-\vartheta)\,d\vartheta}{\int_{-\infty}^{\infty} f(X_1-\vartheta',\dots,X_n-\vartheta')\,d\vartheta'},$$

was gerade der Pitman-Schätzer ist.

(c) Der Pitman-Schätzer ist minimax bezüglich quadratischem Risiko. Wie (b) suggeriert, kann dies über das Bayesrisiko bei a-priori-Verteilung $\pi = U([-R,R])$ und den Grenzübergang $R\to\infty$ gezeigt werden.

3 Asymptotische Schätztheorie

3.1 Momentenschätzer

3.1 Definition. Es seien $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ ein statistisches (Produkt-)Modell und $g(\vartheta)$ mit $g:\Theta\to\mathbb{R}^p$ ein abgeleiteter Parameter. Ferner sei $\psi=(\psi_1,\dots,\psi_q):\mathcal{X}\to\mathbb{R}^q$ derart, dass $\varphi(\vartheta) := \mathbb{E}_\vartheta[\psi]$ für alle $\vartheta\in\Theta$ existiert. Gibt es nun eine Borel-messbare Funktion $G:\varphi(\Theta)\to g(\Theta)$ mit $G\circ\varphi = g$ und liegt $\frac{1}{n}\sum_{i=1}^n \psi(x_i)$ in $\varphi(\Theta)$ für alle $x_1,\dots,x_n\in\mathcal{X}$, so heißt $G(\frac{1}{n}\sum_{i=1}^n\psi(x_i))$ (verallgemeinerter) Momentenschätzer für $g(\vartheta)$ mit Momentenfunktionen $\psi_1,\dots,\psi_q$.

3.2 Beispiele.

(a) Es sei $X_1,\dots,X_n\sim\mathrm{Exp}(\lambda)$ eine mathematische Stichprobe mit $\lambda>0$ unbekannt. Betrachte die klassische Momentenfunktion $\psi(x)=x^k$ für ein $k\in\mathbb{N}$.
Mit $g(\lambda)=\lambda$ und $\varphi(\lambda)=\mathbb{E}_\lambda[X_i^k]=\lambda^{-k}k!$ ergibt sich $G(x)=(k!/x)^{1/k}$ und als Momentenschätzer für λ

$$\hat{\lambda}_{k,n} := \Big(\frac{k!}{\frac{1}{n}\sum_{i=1}^n X_i^k}\Big)^{1/k}.$$

(b) Betrachte einen autoregressiven Prozess der Ordnung 1 (AR(1)-Prozess):

$$X_n = aX_{n-1}+\varepsilon_n, \quad n\geqslant 1,$$

mit (ε_n) i.i.d., $\mathbb{E}[\varepsilon_n]=0$, $\mathrm{Var}(\varepsilon_n)=\sigma^2<\infty$ und $X_0=x_0\in\mathbb{R}$. Um a zu schätzen, betrachte folgende Identität für das bedingte gemeinsame Moment:

$$\mathbb{E}[X_{n-1}X_n \,|\, \varepsilon_1,\dots,\varepsilon_{n-1}] = aX_{n-1}^2.$$

Dies führt auf eine modifizierte Momentenmethode als Schätzidee (Yule-Walker-Schätzer):

$$\hat{a}_n := \frac{\frac{1}{n}\sum_{k=1}^n X_{k-1}X_k}{\frac{1}{n}\sum_{k=1}^n X_{k-1}^2} = a + \frac{\sum_{k=1}^n X_{k-1}\varepsilon_k}{\sum_{k=1}^n X_{k-1}^2}.$$

Im Fall $|a| < 1$ kann man mit Hilfe des Ergodensatzes auf die Konsistenz von $\hat{a}_n$ für $n \to \infty$ schließen. Allgemeiner zeigt man leicht, dass $M_n := \sum_{k=1}^n X_{k-1}\varepsilon_k$ ein Martingal bezüglich $\mathscr{F}_n := \sigma(\varepsilon_1, \ldots, \varepsilon_n)$ ist mit quadratischer Variation $\langle M\rangle_n := \sum_{k=1}^n X_{k-1}^2$. Das starke Gesetz der großen Zahlen für L^2-Martingale liefert daher die Konsistenz

$$\hat{a}_n = a + \frac{M_n}{\langle M\rangle_n} \xrightarrow{\text{f.s.}} a.$$

3.3 Lemma. *Existiert für hinreichend großes n der Momentenschätzer $\hat{g}_n = G(\frac{1}{n}\sum_{i=1}^n \psi(x_i))$ und ist G stetig, so ist $\hat{g}_n$ ein (stark) konsistenter Schätzer von $g(\vartheta)$, d.h. $\lim_{n\to\infty} \hat{g}_n = g(\vartheta)$ $\mathbb{P}_\vartheta^{\otimes\mathbb{N}}$-f.s.*

Beweis. Nach dem starken Gesetz der großen Zahlen gilt wegen der Stetigkeit von G $\mathbb{P}_\vartheta^{\otimes\mathbb{N}}$-fast sicher:

$$\lim_{n\to\infty} G\Big(\frac{1}{n}\sum_{i=1}^n \psi(X_i)\Big) = G\Big(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \psi(X_i)\Big) = G(\varphi(\vartheta)) = g(\vartheta).$$

□

3.4 Satz (Δ-Methode)**.** *Es seien (X_n) eine Folge von Zufallsvektoren im $\mathbb{R}^k$, $\sigma_n > 0$, $\sigma_n \to 0$, $\vartheta_0 \in \mathbb{R}^k$ sowie $\Sigma \in \mathbb{R}^{k\times k}$ positiv semi-definit und es gelte*

$$\sigma_n^{-1}(X_n - \vartheta_0) \xrightarrow{d} N(0, \Sigma).$$

Ist $f : \mathbb{R}^k \to \mathbb{R}$ in einer Umgebung von ϑ_0 stetig differenzierbar mit Gradienten $\dot{f}$, so folgt

$$\sigma_n^{-1}(f(X_n) - f(\vartheta_0)) \xrightarrow{d} N(0, \langle \Sigma\dot{f}(\vartheta_0), \dot{f}(\vartheta_0)\rangle),$$

wobei $N(0,0)$ gegebenenfalls als Punktmaß δ_0 in der Null zu verstehen ist.

Beweis. Nach dem Lemma von Slutsky (vgl. Stochastik II) gilt $X_n - \vartheta_0 = \sigma_n \frac{X_n-\vartheta_0}{\sigma_n} \xrightarrow{d} 0$ und somit (Stochastik I) $X_n \xrightarrow{\mathbb{P}} \vartheta_0$ für $n \to \infty$. Eine Taylorentwicklung ergibt

$$f(X_n) = f(\vartheta_0) + \langle \dot{f}(\vartheta_0), X_n - \vartheta_0\rangle + R_n$$

mit $R_n/|X_n - \vartheta_0| \to 0$ für $X_n \to \vartheta_0$ bezüglich fast sicherer und damit auch stochastischer Konvergenz. Wiederum mittels Slutsky-Lemma folgt

$$\frac{R_n}{\sigma_n} = \frac{|X_n - \vartheta_0|}{\sigma_n}\frac{R_n}{|X_n - \vartheta_0|} \xrightarrow{d} 0$$

und also auch bezüglich stochastischer Konvergenz. Eine dritte Anwendung des Slutsky-Lemmas gibt daher

$$\sigma_n^{-1}(f(X_n)-f(\vartheta_0)) = \langle \dot f(\vartheta_0), \sigma_n^{-1}(X_n-\vartheta_0)\rangle + \sigma_n^{-1}R_n \xrightarrow{d} N(0, \dot f(\vartheta_0)^\top \Sigma \dot f(\vartheta_0));$$

denn es gilt $\langle \dot f(\vartheta_0), \sigma_n^{-1}(X_n-\vartheta_0)\rangle \to \langle \dot f(\vartheta_0), \Sigma^{1/2}Z\rangle \sim N(0, \langle \Sigma \dot f(\vartheta_0), \dot f(\vartheta_0)\rangle)$ mit $Z \sim N(0, E_k)$. □

3.5 Beispiel. Aus einer mathematischen Stichprobe $X_1, \dots, X_n \sim \mathrm{Poiss}(\lambda)$ bestimmt man den UMVU-Schätzer $\hat\lambda_n = \frac{1}{n}\sum_{i=1}^n X_i$. Nach dem zentralen Grenzwertsatz gilt $\sqrt{n}(\hat\lambda_n - \lambda) \xrightarrow{d} N(0,\lambda)$ unter $\mathbb{P}_\lambda^{\otimes\mathbb{N}}$. Um asymptotisch ein Konfidenzintervall herzuleiten, stört es, dass die asymptotische Varianz vom Parameter selbst abhängt. Betrachtet man nun $f(x) = 2x^{1/2}$ mit $\dot f(x) = x^{-1/2}$ in der Δ-Methode, so folgt $\sqrt{n}(2\hat\lambda_n^{1/2} - 2\lambda^{1/2}) \xrightarrow{d} N(0,1)$, so dass $[2\hat\lambda_n^{1/2} - n^{-1/2}q_{1-\alpha/2}, 2\hat\lambda_n^{1/2} + n^{-1/2}q_{1-\alpha/2}]$ mit den $(1-\alpha/2)$-Quantilen von $N(0,1)$ ein asymptotisches $(1-\alpha)$-Konfidenzintervall für $2\lambda^{1/2}$ bildet. Rücktransformation ergibt dann für λ selbst das asymptotische $(1-\alpha)$-Konfidenzintervall $[(\hat\lambda_n^{1/2} - (4n)^{-1/2}q_{1-\alpha/2})_+^2, (\hat\lambda_n^{1/2} + (4n)^{-1/2}q_{1-\alpha/2})^2]$. Die Idee, mittels Δ-Transformation eine asymptotische Varianz unabhängig vom unbekannten zu erhalten, ist in vielen Situationen sehr fruchtbar und nennt sich Varianz-stabilisierende Transformation.

Alternativ kann man die asymptotische Varianz durch $\hat\lambda_n$ konsistent schätzen und mittels Slutsky-Lemma auf $(n/\hat\lambda_n)^{1/2}(\hat\lambda_n - \lambda) \xrightarrow{d} N(0,1)$ schließen. Daraus ergibt sich $[\hat\lambda_n - (\hat\lambda_n/n)^{1/2}q_{1-\alpha/2}, \hat\lambda_n + (\hat\lambda_n/n)^{1/2}q_{1-\alpha/2}]$ als asymptotisches $(1-\alpha)$-Konfidenzintervall.

3.6 Satz. *Es seien $\vartheta_0 \in \Theta$, $g: \Theta \to \mathbb{R}$ und für hinreichend großes n existiere der Momentenschätzer $\hat g_n = G(\frac{1}{n}\sum_{i=1}^n \psi(x_i))$ mit Momentenfunktionen $\psi_j \in L^2(\mathbb{P}_{\vartheta_0})$, $j = 1,\dots,q$. Betrachte $\mathrm{Cov}_{\vartheta_0}(\psi) := (\mathrm{Cov}_{\vartheta_0}(\psi_i,\psi_j))_{i,j=1,\dots,q}$. Sofern G in einer Umgebung von $\varphi(\vartheta_0)$ stetig differenzierbar ist, ist $\hat g_n$ unter $\mathbb{P}_{\vartheta_0}^{\otimes\mathbb{N}}$ asymptotisch normalverteilt mit Rate $n^{-1/2}$, asymptotischem Mittelwert Null und Varianz $\langle \mathrm{Cov}_{\vartheta_0}(\psi)\dot G(\varphi(\vartheta_0)), \dot G(\varphi(\vartheta_0))\rangle$:*

$$\sqrt{n}(\hat g_n - g(\vartheta_0)) \xrightarrow{d} N(0, \langle \mathrm{Cov}_{\vartheta_0}(\psi)\dot G(\varphi(\vartheta_0)), \dot G(\varphi(\vartheta_0))\rangle) \text{ (unter } \mathbb{P}_{\vartheta_0}^{\otimes\mathbb{N}}).$$

3.7 Bemerkung. Die Begriff *asymptotischer Mittelwert* und *asymptotische Varianz* sind leicht irreführend: es gilt nicht notwendigerweise, dass die Momente von $\sqrt{n}(\hat g_n - g(\vartheta_0))$ gegen die entsprechenden Momente von $N(0, \langle \mathrm{Var}_{\vartheta_0}(\psi)\dot G(\varphi(\vartheta_0)), \dot G(\varphi(\vartheta_0))\rangle)$ konvergieren (dafür wird gleichgradige Integrierbarkeit benötigt).

Beweis. Nach dem multivariaten zentralen Grenzwertsatz gilt unter $\mathbb{P}_{\vartheta_0}^{\otimes\mathbb{N}}$

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n \psi(X_i) - \varphi(\vartheta_0)\Big) \xrightarrow{d} N(0, \mathrm{Cov}_{\vartheta_0}(\psi)).$$

Die Behauptung folgt daher unmittelbar mit der Δ-Methode. □

3.8 Beispiel. Im Exponentialverteilungsmodell aus Beispiel 3.2 gilt $G'(x) = -(k!/x)^{1/k}(kx)^{-1}$ und $\Sigma(\lambda_0) = \mathrm{Var}_{\lambda_0}(X_i^k) = ((2k)! - (k!)^2)/\lambda_0^{2k}$. Alle Momentenschätzer $\hat{\lambda}_{k,n}$ sind asymptotisch normalverteilt mit Rate $n^{-1/2}$ und Varianz $\sigma_k^2 = \lambda_0^2 k^{-2}((2k)!/(k!)^2 - 1)$. Da $\hat{\lambda}_{1,n}$ die gleichmäßig kleinste asymptotische Varianz besitzt und auf der suffizienten Statistik $\overline{X}$ basiert, wird dieser Schätzer im Allgemeinen vorgezogen.

3.9 Bemerkung. Die Momentenmethode kann unter folgendem allgemeinen Gesichtspunkt verstanden werden: Ist $X_1, \dots, X_n$ eine mathematische Stichprobe mit Werten in $\mathbb{R}$, so ist die empirische Verteilungsfunktion $F_n(x) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \leqslant x)$ eine suffiziente Statistik und nach dem Satz von Glivenko-Cantelli gilt $\mathbb{P}_\vartheta$-f.s. $F_n(x) \to F_\vartheta(x) = \mathbb{P}_\vartheta(X_i \leqslant x)$ gleichmäßig in $x \in \mathbb{R}$. Ist nun $g(\vartheta)$ als Funktional $G(F_\vartheta(x), x \in \mathbb{R})$ darstellbar, so verwende die empirische Version $G(F_n(x),\, x \in \mathbb{R})$ als Schätzer von $g(\vartheta)$. Falls das Funktional G stetig bezüglich der Supremumsnorm ist, so folgt die Konsistenz.

Der Satz von Donsker für empirische Prozesse zeigt $\sqrt{n}(F_n - F_\vartheta) \xrightarrow{d} \Gamma_\vartheta$ gleichmäßig auf $\mathbb{R}$ mit einem zentrierten Gaußprozess Γ_ϑ von der Kovarianzstruktur $\mathrm{Cov}(\Gamma_\vartheta(x), \Gamma_\vartheta(y)) = F_\vartheta(x \wedge y) - F_\vartheta(x)F_\vartheta(y)$. Ist G ein *Hadamard-differenzierbares* Funktional, so folgt $\sqrt{n}(G(F_n(x),\, x \in \mathbb{R}) - g(\vartheta)) \xrightarrow{d} \dot{G}(F_\vartheta)\Gamma_\vartheta$ unter $\mathbb{P}_\vartheta$, also insbesondere asymptotische Normalverteilung mit Rate $n^{-1/2}$ und explizit bestimmbarer asymptotischer Varianz, siehe z.B. das Buch von van der Vaart für mehr Details.

Als einfaches (lineares) Beispiel sei $g(\vartheta) = \mathbb{E}_\vartheta[\psi(X_i)]$ zu schätzen und $X_i \geqslant 0$ $\mathbb{P}_\vartheta$-f.s. Dann folgt informell $G(F_\vartheta) = \int_0^\infty \psi(x)\, dF_\vartheta(x) = \int_0^\infty \psi'(x)(1 - F_\vartheta(x))\, dx$. Aus der Linearität erhalten wir $\dot{G}(F_\vartheta)\Gamma_\vartheta = \int_0^\infty \psi'(x)(-\Gamma_\vartheta(x))\, dx$. Dies ist normalverteilt mit Erwartungswert Null und Varianz

$$\begin{aligned}
&\int_0^\infty \int_0^\infty \psi'(x)\psi'(y)(F_\vartheta(x \wedge y) - F_\vartheta(x)F_\vartheta(y))\, dx\, dy \\
&= \int_0^\infty \int_0^\infty \psi(x)\psi(y)\partial_{xy}(F_\vartheta(x \wedge y) - F_\vartheta(x)F_\vartheta(y))\, dx\, dy \\
&= \int_0^\infty \psi^2(x)\, dF_\vartheta(x) - \Big(\int_0^\infty \psi(x)\, dF_\vartheta(x)\Big)^2,
\end{aligned}$$

was natürlich gerade der Varianz von $G(F_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i)$ entspricht.

3.2 Maximum-Likelihood- und M-Schätzer

3.10 Beispiele.

(a) Auf dem diskreten Stichprobenraum $\mathcal{X}$ seien Verteilungen $(P_\vartheta)_{\vartheta\in\Theta}$ gegeben. Bezeichnet p_ϑ die zugehörige Zähldichte und ist die Verlustfunktion $l(\vartheta, \rho)$ homogen in $\vartheta \in \Theta$, so ist es für die Schätzung von ϑ plausibel, bei Vorliegen des Versuchsausgangs x für einen Schätzer $\hat{\vartheta}(x)$ denjenigen Parameter $\vartheta \in \Theta$ zu wählen, für den die Wahrscheinlichkeit $p_\vartheta(x)$ des Eintretens von x maximal ist: $\hat{\vartheta}(x) := \mathrm{argmax}_{\vartheta\in\Theta}\, p_\vartheta(x)$. Dieser Schätzer heißt Maximum-Likelihood-Schätzer (MLE). Bereits im vorliegenden Fall

ist weder Existenz noch Eindeutigkeit ohne Weiteres garantiert. Bei Nicht-Eindeutigkeit wählt man einen maximierenden Parameter ϑ nach Belieben aus. Im Fall einer mathematischen Stichprobe $X_1, \dots, X_n \sim \text{Poiss}(\lambda)$ mit $\lambda > 0$ unbekannt, ergibt sich beispielsweise

$$\hat{\lambda} = \text{argmax}_{\lambda>0} \prod_{i=1}^{n} \big(e^{-\lambda} \frac{\lambda^{X_i}}{X_i!}\big) = \bar{X}$$

im Fall $\bar{X} > 0$. Ist $\bar{X} = 0$, d.h. $X_1 = \cdots = X_n = 0$, so wird das Supremum nur asymptotisch für $\lambda \to 0$ erreicht. Hier könnte man sich behelfen, indem man Poiss(0) als Punktmaß in der Null stetig ergänzt.

(b) Besitzen die Verteilungen $\mathbb{P}_\vartheta$ Lebesguedichten f_ϑ, so führt der Maximum-Likelihood-Ansatz analog auf $\hat{\vartheta}(x) = \text{argmax}_{\vartheta\in\Theta} f_\vartheta(x)$. Betrachte die Stichprobe Y der Form $Y = e^X$ mit $X \sim N(\mu, 1)$ mit $\mu \in \mathbb{R}$ unbekannt. Dann ist Y log-normalverteilt, und es gilt

$$\hat{\mu}(Y) = \text{argmax}_{\mu\in\mathbb{R}} \frac{e^{-(\log(Y)-\mu)^2/2}}{\sqrt{2\pi}Y} = \log(Y).$$

Man sieht, dass der MLE invariant unter Parametertransformation ist: bei Beobachtung von $X \sim N(\mu, 1)$ erhält man den MLE $\tilde{\mu}(X) = X$ und Einsetzen von $X = \log(Y)$ führt auf dasselbe Ergebnis. Interessanterweise führt die Momentenmethode unter Benutzung von $\mathbb{E}_\mu[Y] = e^{\mu+1/2}$ auf den Schätzer $\bar{\mu}(Y) = \log(Y) - 1/2$, während $\mathbb{E}_\mu[X] = \mu$ auf $\check{\mu}(X) = X$ führt; Momentenschätzer, beruhend auf demselben Moment, sind also im Allgemeinen nicht transformationsinvariant.

3.11 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein von μ dominiertes Modell mit Likelihoodfunktion $L(\vartheta, x)$. Eine Statistik $\hat{\vartheta} : \mathcal{X} \to \Theta$ (Θ trage eine σ-Algebra $\mathscr{F}_\Theta$) heißt Maximum-Likelihood-Schätzer (MLE) von ϑ, falls $L(\hat{\vartheta}(x), x) = \sup_{\vartheta\in\Theta} L(\vartheta, x)$ für μ-fast alle $x \in \mathcal{X}$ gilt.

3.12 Bemerkung. Der MLE braucht weder zu existieren noch eindeutig zu sein, falls er existiert. Er hängt von der gewählten Version der Radon-Nikodym-Dichte ab; es gibt jedoch häufig eine kanonische Wahl, wie beispielsweise bei stetigen Lebesguedichten. Außerdem ist eine Abänderung auf einer Nullmenge bezüglich aller $\mathbb{P}_\vartheta$ irrelevant, weil der Schätzer vor Realisierung des Experiments festgelegt wird und diese Realisierung damit fast sicher zum selben Schätzwert führen wird.

Bei einer eineindeutigen Parametrisierung $\vartheta \mapsto h(\vartheta)$ ergibt sich $\hat{h} := h(\hat{\vartheta})$ als MLE für $h(\vartheta)$.

3.13 Lemma. *Für eine natürliche Exponentialfamilie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ in $T(x)$ ist der MLE $\hat{\vartheta}$ implizit gegeben durch die Momentengleichung $\mathbb{E}_{\hat{\vartheta}(x)}[T] = T(x)$, vorausgesetzt der MLE existiert und $\vartheta(x) \in \text{int}(\Theta)$.*

Beweis. Schreiben wir die Loglikelihoodfunktion in der Form $\ell(\vartheta, x) = \log(h(x)) + \langle\vartheta, T(x)\rangle - A(\vartheta)$, so folgt (vgl. Satz 2.11) wegen der Differenzierbarkeit im Innern $\dot{\ell}(\hat{\vartheta}(x), x) = T(x) - \dot{A}(\hat{\vartheta}(x)) = 0$ und somit $\mathbb{E}_{\hat{\vartheta}(x)}[T] = T(x)$. □

3.14 Beispiele.

(a) Es sei $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ eine mathematische Stichprobe. Dann ist der MLE für $\vartheta = (\mu/\sigma^2, 1/(2\sigma^2))^\top$ gegeben durch $\mathbb{E}_{\hat\vartheta}[(\bar X, \overline{X^2})^\top] = (\bar X, \overline{X^2})^\top$, also $\hat\mu = \bar X$, $\widehat{\mu^2+\sigma^2} = \overline{X^2}$. Durch Reparametrisierung $(\mu, \mu^2+\sigma^2) \mapsto (\mu, \sigma^2)$ erhalten wir $\hat\sigma^2 = \overline{X^2} - (\bar X)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar X)^2$. Beachte, dass der MLE $\hat\sigma^2$ nicht erwartungstreu ist.

(b) Bei Beobachtung einer Markovkette $(X_0, X_1, \dots, X_n)$ auf dem Zustandsraum $S = \{1, \dots, M\}$ mit parameterunabhängigem Anfangswert $X_0 = x_0$ und unbekannten Übergangswahrscheinlichkeiten $\mathbb{P}(X_{k+1} = j \mid X_k = i) = p_{ij}$ ergibt sich die Likelihoodfunktion (bzgl. Zählmaß) durch

$$L((p_{kl}), X) = \prod_{i=1}^n p_{X_{i-1}, X_i} = \prod_{k,l=1}^M p_{kl}^{N_{kl}(X)},$$

wobei $N_{kl}(X) = |\{i = 1, \dots, n \mid X_{i-1} = k, X_i = l\}|$ die Anzahl der beobachteten Übergänge von Zustand k nach Zustand l angibt. Als MLE ergibt sich nach kurzer Rechnung die relative Häufigkeit $\hat p_{ij} = N_{ij}/(\sum_{m\in S} N_{im})$ der Übergänge.

(c) Beim allgemeinen parametrischen Regressionsmodell mit Beobachtungen

$$Y_i = g_\vartheta(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

ergibt sich unter der Normalverteilungsannahme $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d. als MLE der Kleinste-Quadrate-Schätzer $\hat\vartheta = \operatorname{argmin}_{\vartheta\in\Theta} \sum_{i=1}^n (Y_i - g_\vartheta(x_i))^2$.

3.15 Definition. Für zwei Wahrscheinlichkeitsmaße $\mathbb{P}$ und $\mathbb{Q}$ auf demselben Messraum $(\mathcal{X}, \mathscr{F})$ heißt die Funktion

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}(x)\right) \mathbb{P}(dx), & \text{falls } \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{sonst} \end{cases}$$

Kullback-Leibler-Divergenz (oder auch Kullback-Leibler-Abstand, relative Entropie) von $\mathbb{P}$ bezüglich $\mathbb{Q}$.

3.16 Lemma. *Für die Kullback-Leibler-Divergenz gilt:*

(a) $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) \geqslant 0$ *und* $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$*;*

(b) für Produktmaße ist KL *additiv:*

$$\mathrm{KL}(\mathbb{P}_1 \otimes \mathbb{P}_2 \mid \mathbb{Q}_1 \otimes \mathbb{Q}_2) = \mathrm{KL}(\mathbb{P}_1 \mid \mathbb{Q}_1) + \mathrm{KL}(\mathbb{P}_2 \mid \mathbb{Q}_2);$$

(c) bildet $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ *eine natürliche Exponentialfamilie und ist* ϑ_0 *innerer Punkt von* Θ*, so gilt*

$$\mathrm{KL}(\mathbb{P}_{\vartheta_0} \mid \mathbb{P}_\vartheta) = A(\vartheta) - A(\vartheta_0) + \langle \dot A(\vartheta_0), \vartheta_0 - \vartheta\rangle.$$

3.17 Bemerkung. Im Allgemeinen ist KL nicht symmetrisch und damit keine Metrik. Trotzdem spielt die Kullback-Leibler-Divergenz eine Hauptrolle in der Asymptotik Likelihood-basierter Verfahren.

Beweis. Für (a) können wir o.B.d.A. $\mathbb{P} \ll \mathbb{Q}$ annehmen. Dann folgt aus der strikten Konvexität von $h(x) = x\log(x)$ auf $[0,\infty)$ (mit stetiger Ergänzung $h(0) = 0$) mittels Jensen-Ungleichung

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = \int h\Big(\frac{d\,\mathbb{P}}{d\,\mathbb{Q}}(x)\Big)\ \mathbb{Q}(dx) \geqslant h\Big(\int \frac{d\,\mathbb{P}}{d\,\mathbb{Q}}(x)\ \mathbb{Q}(dx)\Big) = h(1) = 0$$

mit Gleichheit genau dann, wenn $\frac{d\,\mathbb{P}}{d\,\mathbb{Q}}$ $\mathbb{Q}$-f.s. konstant ist. Da eine konstante Dichte zwischen Wahrscheinlichkeitsmaßen notwendigerweise gleich Eins sein muss, folgt Aussage (a) aus $\frac{d\,\mathbb{P}}{d\,\mathbb{Q}} = 1$ $\mathbb{Q}$-f.s. $\iff \mathbb{P} = \mathbb{Q}$.

Für (b) benutze die Produktdichte und Fubini im Fall $\mathrm{KL}(\mathbb{P}_1 \,|\, \mathbb{Q}_1) < \infty$ und $\mathrm{KL}(\mathbb{P}_2, \mathbb{Q}_2) < \infty$:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_1 \otimes \mathbb{P}_2 \mid \mathbb{Q}_1 \otimes \mathbb{Q}_2) &= \int\int \log\Big(\frac{d\,\mathbb{P}_1}{d\,\mathbb{Q}_1}(x_1)\frac{d\,\mathbb{P}_2}{d\,\mathbb{Q}_2}(x_2)\Big)\ \mathbb{P}_1(dx_1)\,\mathbb{P}_2(dx_2)\\ &= \int \log\big(\frac{d\,\mathbb{P}_1}{d\,\mathbb{Q}_1}(x_1)\big)\ \mathbb{P}_1(dx_1) + \int \log\big(\frac{d\,\mathbb{P}_2}{d\,\mathbb{Q}_2}(x_2)\big)\ \mathbb{P}_2(dx_2)\end{aligned}$$

und wir erhalten $\mathrm{KL}(\mathbb{P}_1 \mid \mathbb{Q}_1) + \mathrm{KL}(\mathbb{P}_2 \mid \mathbb{Q}_2)$. Um die Anwendung von Fubini zu begründen, müssen wir noch

$$\int\int \Big|\log\Big(\frac{d\,\mathbb{P}_1}{d\,\mathbb{Q}_1}(x_1)\frac{d\,\mathbb{P}_2}{d\,\mathbb{Q}_2}(x_2)\Big)\Big|\ \mathbb{P}_1(dx_1)\,\mathbb{P}_2(dx_2) < \infty$$

zeigen. Das Doppelintegral kann mittels Dreiecksungleichung und Verwendung der Funktion h abgeschätzt werden durch

$$\int \Big|h\Big(\frac{d\,\mathbb{P}_1}{d\,\mathbb{Q}_1}(x_1)\Big)\Big|\ \mathbb{Q}_1(dx_1) + \int \Big|h\Big(\frac{d\,\mathbb{P}_2}{d\,\mathbb{Q}_2}(x_2)\Big)\Big|\ \mathbb{Q}_2(dx_2).$$

Da h nach unten beschränkt ist ($h(x) \geqslant -e^{-1}$) und $\mathbb{Q}_1, \mathbb{Q}_2$ Wahrscheinlichkeitsmaße sind, sind die Integrale endlich genau dann, wenn die Integrale über die Integranden ohne Absolutwerte endlich sind. Letzteres folgt aus $\mathrm{KL}(\mathbb{P}_1 \,|\, \mathbb{Q}_1) < \infty$ und $\mathrm{KL}(\mathbb{P}_2, \mathbb{Q}_2) < \infty$. Im Fall $\mathrm{KL}(\mathbb{P}_1 \,|\, \mathbb{Q}_1) = \infty$ oder $\mathrm{KL}(\mathbb{P}_2, \mathbb{Q}_2) = \infty$ folgt $\mathrm{KL}(\mathbb{P}_1 \otimes \mathbb{P}_2 \mid \mathbb{Q}_1 \otimes \mathbb{Q}_2) = \infty$ auf ähnliche Weise.

Behauptung (c) folgt durch Einsetzen von $\log(\frac{d\,\mathbb{P}_{\vartheta_0}}{d\,\mathbb{P}_{\vartheta}}(x)) = \langle T(x), \vartheta_0 - \vartheta\rangle + A(\vartheta) - A(\vartheta_0)$ sowie $\mathbb{E}_{\vartheta_0}[T] = \dot{A}(\vartheta_0)$, vergleiche Satz 2.11. □

3.18 Bemerkung. Wegen $\ddot{A}(\vartheta_0) = \mathrm{Cov}_{\vartheta_0}(T)$ in (c) erhalten wir für $\vartheta, \vartheta_0 \in \mathrm{int}(\Theta)$ mit einer Taylorentwicklung $\mathrm{KL}(\mathbb{P}_{\vartheta_0} \mid \mathbb{P}_{\vartheta}) = \frac{1}{2}\langle \mathrm{Cov}_{\bar{\vartheta}}(T)(\vartheta - \vartheta_0), \vartheta - \vartheta_0\rangle$ mit einer Zwischenstelle $\bar{\vartheta}$ zwischen ϑ und ϑ_0. Beachte, dass $\mathrm{Cov}_{\bar{\vartheta}}(T)$ gerade die Fisher-Information bei $\bar{\vartheta}$ angibt. Im Fall der mehrdimensionalen Normalverteilung $N(\mu, \Sigma)$ mit strikt positiv-definiter Kovarianzmatrix folgt aus $A(\mu) = \langle \Sigma^{-1}\mu, \mu\rangle/2$, dass $\ddot{A}(\mu) = \Sigma^{-1}$ unabhängig von μ ist und somit $\mathrm{KL}(N(\vartheta_0, \Sigma) \,|\, N(\vartheta, \Sigma)) = \frac{1}{2}\langle \Sigma^{-1}(\vartheta - \vartheta_0), \vartheta - \vartheta_0\rangle$ gilt.

3.19 Definition. Es sei $(\mathcal{X}_n, \mathscr{F}_n, (\mathbb{P}_\vartheta^n)_{\vartheta\in\Theta})_{n\geqslant 1}$ eine Folge statistischer Modelle sowie $g(\vartheta)$ mit $g : \Theta \to \Gamma$ der interessierende Parameter. Eine Funktion $K : \Theta \times \Gamma \to \mathbb{R}\cup\{+\infty\}$ heißt Kontrastfunktion, falls $\gamma \mapsto K(\vartheta_0, \gamma)$ ein eindeutiges Minimum bei $g(\vartheta_0)$ besitzt für alle $\vartheta_0 \in \Theta$. Eine Folge $K_n : \Gamma\times\mathcal{X}_n \to \mathbb{R}\cup\{+\infty\}$ heißt zugehöriger Kontrastprozess (oder bloß Kontrast), falls folgende Bedingungen gelten:

(a) $K_n(\gamma, \bullet)$ ist $\mathscr{F}_n$-messbar für alle $\gamma \in \Gamma$;

(b) $\forall\gamma \in \Gamma$, $\vartheta_0 \in \Theta$: $K_n(\gamma) \to K(\vartheta_0, \gamma)$ $\mathbb{P}_{\vartheta_0}^n$-stochastisch für $n \to \infty$.

Ein zugehöriger Minimum-Kontrast-Schätzer oder M-Schätzer von $g(\vartheta)$ ist gegeben durch $\hat\gamma_n(x_n) := \operatorname{argmin}_{\gamma\in\Gamma} K_n(\gamma, x_n)$ (sofern existent; nicht notwendigerweise eindeutig).

3.20 Beispiele.

(a) Es sei $g(\vartheta) = \vartheta$, $\Gamma = \Theta$. Beim Produktexperiment $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ mit $\mathbb{P}_\vartheta \sim \mathbb{P}_{\vartheta'}$ für alle $\vartheta, \vartheta' \in \Theta$ ist

$$K_n(\vartheta, x) = -\frac{1}{n}\sum_{i=1}^n \ell(\vartheta, x_i)$$

mit der Loglikelihood-Funktion ℓ bezüglich einem dominierenden Wahrscheinlichkeitsmaß μ ein Kontrastprozess zur Kontrastfunktion

$$\begin{aligned} K(\vartheta_0, \vartheta) &= \mathbb{E}_{\vartheta_0}[-\ell(\vartheta)] = \mathbb{E}_{\vartheta_0}\Big[\log\Big(\frac{L(\vartheta_0)}{L(\vartheta)}\Big)\Big] - \mathbb{E}_{\vartheta_0}[\log(L(\vartheta_0))] \\ &= \mathrm{KL}(\mathbb{P}_{\vartheta_0} \mid \mathbb{P}_\vartheta) - \mathbb{E}_{\vartheta_0}[\ell(\vartheta_0)], \end{aligned}$$

sofern $\ell(\vartheta_0) \in L^1(\mathbb{P}_{\vartheta_0})$ gilt. Der zugehörige M-Schätzer ist der MLE.

(b) Es sei $g(\vartheta) = \vartheta$, $\Gamma = \Theta$. Betrachte das Regressionsmodell aus Beispiel 3.14 mit $f_\vartheta : [0, 1] \to \mathbb{R}$ stetig, äquidistantem Design $x_i = i/n$ und beliebig verteilten Störvariablen (ε_i). Sind die (ε_i) i.i.d. mit $\mathbb{E}[\varepsilon_i] = 0$ und $\mathbb{E}[\varepsilon_i^4] < \infty$, so folgt leicht aus Tschebyschew-Ungleichung und Riemannscher Summen-Approximation, dass $K_n(\vartheta) = \frac{1}{n}\sum_{i=1}^n (Y_i - f_\vartheta(x_i))^2$ einen Kontrastprozess zur Kontrastfunktion $K(\vartheta_0, \vartheta) = \int_0^1 (f_{\vartheta_0}(x) - f_\vartheta(x))^2 dx + \mathbb{E}[\varepsilon_i^2]$ bildet. Dabei muss natürlich die Identifizierbarkeitsbedingung $f_\vartheta \neq f_{\vartheta'}$ für alle $\vartheta \neq \vartheta'$ gelten. Also ist der Kleinste-Quadrate-Schätzer hier ebenfalls M-Schätzer.

(c) Im Regressionsmodell aus (b) liege nun eine Modellmisspezifikation vor in dem Sinne, dass die Beobachtungen gemäß $Y_i = f^0(i/n) + \varepsilon_i$ generiert werden, wobei $f^0 : [0, 1] \to \mathbb{R}$ nicht notwendigerweise gleich einem f_ϑ ist. Nimmt man an, dass die Funktion selbst der Parameter ϑ im Kleinste-Quadrate-Ansatz ist, d.h. $\hat\vartheta_n = \operatorname{argmin}_{\vartheta\in\Theta} \frac{1}{n}\sum_{i=1}^n (Y_i - \vartheta(i/n))^2$ mit $\Theta \subseteq L^2([0, 1])$, so erhalten wir nach obiger Herleitung im Grenzwert die 'Kontrast-Typ-Funktion' $K(f^0, \vartheta) = \int_0^1 (f^0(x) - \vartheta(x))^2 dx + \mathbb{E}[\varepsilon_i^2]$. Für

$f^0 \notin \Theta$ wird das Minimum nun natürlich nicht in f^0 angenommen, so dass in der Kontrasttheorie die Funktion g wesentlich wird.

Dazu nehmen wir an, dass die parametrische Funktionenmenge Γ (vormals Θ) Riemann-integrierbare Funktionen enthält sowie abgeschlossen in $L^2([0,1])$ und konvex ist, so dass für jede Funktion $\vartheta \in L^2([0,1])$ eine eindeutige L^2-Orthogonalprojektion $g(\vartheta)$ auf Γ existiert. Beispielsweise kann Γ die Menge aller Polynome vom Grad $\leqslant d$ sein. Bezeichnet Θ die Menge der quadratisch Riemann-integrierbaren Funktionen in $L^2([0,1])$, so ist $K_n(\gamma) = \frac{1}{n}\sum_{i=1}^n (Y_i - \gamma(i/n))^2$, $\gamma \in \Gamma$, Kontrastprozess zur Kontrastfunktion $K(\vartheta_0, \gamma) = \|\vartheta_0 - \gamma\|_{L^2}^2 + \mathbb{E}[\varepsilon_i^2]$, welche genau bei $\gamma = g(\vartheta_0)$ ihr Minimum in Γ annimmt. Es ist zu erwarten (vgl. Übungen), dass unter geeigneten Bedingungen der Kleinste-Quadrate-Schätzer $\hat{\gamma}_n = \operatorname{argmin}_{\gamma\in\Gamma} \frac{1}{n}\sum_{i=1}^n (Y_i - \gamma(i/n))^2$ unter $\mathbb{P}_{\vartheta_0}^n$ gegen $g(\vartheta_0)$ konvergiert. Im derart misspezifizierten Modell wird also die beste L^2-Approximation an die wahre Funktion ϑ_0 geschätzt, z.B. das best approximierende Polynom vom Grad $\leqslant d$.

3.3 Asymptotik

3.21 Satz. *Es sei $(K_n)_{n\geqslant 1}$ ein Kontrastprozess zur Kontrastfunktion K. Dann ist der zugehörige M-Schätzer $\hat{\gamma}_n$ konsistent für $g(\vartheta_0)$, $\vartheta_0 \in \Theta$, unter folgenden Bedingungen:*

(A1) *Γ ist ein kompakter Raum;*

(A2) *$\gamma \mapsto K(\vartheta_0, \gamma)$ ist stetig und $\gamma \mapsto K_n(\gamma)$ ist $\mathbb{P}_{\vartheta_0}^n$-f.s. stetig für alle $n \geqslant 1$;*

(A3) $\sup_{\gamma\in\Gamma} |K_n(\gamma) - K(\vartheta_0, \gamma)| \to 0$ *$\mathbb{P}_{\vartheta_0}^n$-stochastisch.*

3.22 Bemerkung. Beachte, dass $\hat{\gamma}_n$ als Minimum einer fast sicher stetigen Funktion auf einem Kompaktum stets fast sicher existiert. Es kann außerdem messbar gewählt werden (vgl. Witting, 2. Band, Satz 6.7).

Bedingungen **(A1)** und **(A2)** können ersetzt werden durch die schwächere (wieso?) Bedingung

$$(\mathbf{A1}'): \ \forall \varepsilon > 0 \quad \inf_{d(\gamma, g(\vartheta_0)) \geqslant \varepsilon} K(\vartheta_0, \gamma) > K(\vartheta_0, g(\vartheta_0))$$

mit der Metrik d von Γ, vergleiche Übungen.

Beweis. Zeige, dass die entsprechende Funktion $\operatorname{argmin} : C(\Gamma) \to \Gamma$ stetig bezüglich Maximumsnorm auf $C(\Gamma)$ ist an den Stellen f, wo $m_f := \operatorname{argmin}_\gamma f(\gamma)$ eindeutig ist. Betrachte $f_n \in C(\Gamma)$ mit $\|f_n - f\|_\infty \to 0$. Dann konvergieren auch die Minima $f_n(m_{f_n}) \to f(m_f)$ wegen

$$\begin{aligned} f_n(m_{f_n}) - f(m_f) &\geqslant f(m_{f_n}) - f(m_f) - \|f_n - f\|_\infty \geqslant -\|f_n - f\|_\infty \to 0, \\ f_n(m_{f_n}) - f(m_f) &\leqslant f_n(m_{f_n}) - f_n(m_f) + \|f_n - f\|_\infty \leqslant \|f_n - f\|_\infty \to 0. \end{aligned}$$

Ist nun $m \in \Gamma$ (Γ kompakt) ein Häufungspunkt von (m_{f_n}), so folgt mit gleichmäßiger Konvergenz $f(m) = \lim_{n\to\infty} f_n(m_{f_n}) = f(m_f)$. Eindeutigkeit

des Minimums liefert $m = m_f$, und daher besitzt (m_{f_n}) als einzigen Häufungspunkt notwendigerweise den Grenzwert m_f.

Das *Continuous-Mapping-Theorem* für stochastische Konvergenz liefert mit (A3) die Behauptung, weil argmin stetig ist auf dem deterministischen Grenzwert $K(\vartheta_0, \bullet)$. □

3.23 Satz. *Ist $\Gamma \subseteq \mathbb{R}^k$ kompakt, $(X_n(\gamma),\, \gamma \in \Gamma)_{n\geqslant 1}$ eine Folge stetiger Prozesse mit $X_n(\gamma) \xrightarrow{\mathbb{P}} X(\gamma)$ für alle $\gamma \in \Gamma$ und stetigem Grenzprozess $(X(\gamma),\, \gamma \in \Gamma)$, so gilt $\max_{\gamma\in\Gamma}|X_n(\gamma) - X(\gamma)| \xrightarrow{\mathbb{P}} 0$ genau dann, wenn*

$$\forall \varepsilon > 0 : \lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbb{P}\Big(\sup_{|\gamma_1-\gamma_2|<\delta} |X_n(\gamma_1) - X_n(\gamma_2)| \geqslant \varepsilon \Big) = 0.$$

Beweis. Siehe Stochastik II bzw. Übung. □

3.24 Definition. Für Zufallsvariablen (X_n) und positive Zahlen (a_n) schreiben wir $X_n = O_{\mathbb{P}}(a_n)$, falls $\lim_{K\to\infty} \sup_n \mathbb{P}(|X_n| > Ka_n) = 0$ (X_n/a_n ist stochastisch beschränkt oder straff), sowie $X_n = o_{\mathbb{P}}(a_n)$, falls $X_n/a_n \xrightarrow{\mathbb{P}} 0$.

3.25 Satz. *Der M-Schätzer $\hat\gamma_n$ sei konsistent für $\gamma_0 := g(\vartheta_0)$, z.B. unter Annahmen (A1)-(A3), mit $\Gamma \subseteq \mathbb{R}^k$ und $\gamma_0 \in int(\Gamma)$. Der Kontrastprozess K_n sei zweimal stetig differenzierbar in einer Umgebung von γ_0 ($\mathbb{P}^n_{\vartheta_0}$-f.s.), so dass mit*

$$U_n(\gamma) := \dot K_n(\gamma) \text{ (Score)}, \qquad V_n(\gamma) := \ddot K_n(\gamma)$$

folgende Konvergenzen unter $\mathbb{P}^n_{\vartheta_0}$ gelten:

(B1) *$\sqrt{n}U_n(\gamma_0) \xrightarrow{d} N(0, U(\gamma_0))$ mit $U(\gamma_0) \in \mathbb{R}^{k\times k}$ positiv semi-definit, deterministisch.*

(B2) *Gilt $\gamma_n \xrightarrow{\mathbb{P}^n_{\vartheta_0}} \gamma_0$ für Zufallsvariablen γ_n, so folgt $V_n(\gamma_n) \xrightarrow{\mathbb{P}^n_{\vartheta_0}} V(\gamma_0)$ mit $V(\gamma_0) \in \mathbb{R}^{k\times k}$ regulär, deterministisch.*

Dann gilt für den M-Schätzer $\hat\gamma_n$

$$\sqrt{n}(\hat\gamma_n - \gamma_0) = -V(\gamma_0)^{-1}\sqrt{n}U_n(\gamma_0) + o_{\mathbb{P}^n_{\vartheta_0}}(1).$$

Insbesondere ist $\hat\gamma_n$ unter $\mathbb{P}^n_{\vartheta_0}$ asymptotisch normalverteilt:

$$\sqrt{n}(\hat\gamma_n - \gamma_0) \xrightarrow{d} N(0, V(\gamma_0)^{-1}U(\gamma_0)V(\gamma_0)^{-1}).$$

Beweis. Aus der Konsistenz von $\hat\gamma_n$ folgt mit $\gamma_0 \in \mathrm{int}(\Gamma)$ für $\Omega^1_n := \{[\hat\gamma_n, \gamma_0] \subseteq \mathrm{int}(\Gamma)\}$ (setze $[a,b] := \{ah + b(1-h) \,|\, h \in [0,1]\}$) $\lim_{n\to\infty} \mathbb{P}^n_{\vartheta_0}(\Omega^1_n) = 1$. Auf Ω^1_n gilt somit $\dot K_n(\hat\gamma_n) = 0$ und nach Mittelwertsatz

$$\dot K_n(\hat\gamma_n) - \dot K_n(\gamma_0) = \ddot K_n(\bar\gamma_n)(\hat\gamma_n - \gamma_0), \quad \bar\gamma_n \in [\gamma_0, \hat\gamma_n].$$

Wir erhalten

$$-U_n(\gamma_0) = V_n(\bar\gamma_n)(\hat\gamma_n - \gamma_0).$$

Wegen (B2), und da $V(\gamma_0)$ regulär und die Inversenbildung stetig ist, haben wir $\mathbb{P}^n_{\vartheta_0}(\Omega_n^2) \to 1$ für $\Omega_n^2 := \{V_n(\bar\gamma_n)^{-1} \text{ existiert}\}$. Benutze nun $V_n(\bar\gamma_n)^{-1}\mathbf{1}_{\Omega_n^1\cap\Omega_n^2} \to V(\gamma_0)^{-1}$ in $\mathbb{P}^n_{\vartheta_0}$-Wahrscheinlichkeit, so dass

$$\sqrt{n}(\hat\gamma_n - \gamma_0) = -V(\gamma_0)^{-1}\sqrt{n}U_n(\gamma_0) + o_{\mathbb{P}^n_{\vartheta_0}}(1) \xrightarrow{d} N(0, V(\gamma_0)^{-1}U(\gamma_0)V(\gamma_0)^{-1})$$

aus Slutskys Lemma folgt. □

3.26 Beispiel. Im Beobachtungsmodell $Y_i = \gamma + \varepsilon_i$, $i = 1, \dots, n$, mit $\gamma \in \mathbb{R}$, (ε_i) i.i.d. betrachte den M-Schätzer

$$\hat\gamma_n = \operatorname{argmin}_{\gamma\in\mathbb{R}} \sum_{i=1}^n \rho(Y_i - \gamma)$$

mit einer Funktion $\rho : \mathbb{R} \to [0,\infty)$, so dass $x \mapsto \mathbb{E}[\rho(x + \varepsilon_i)]$ minimal (nur) bei $x = 0$ ist. Mit dem Kontrast $K_n(\gamma) = \frac{1}{n}\sum_{i=1}^n \rho(Y_i - \gamma)$ erhalten wir dann die Kontrastfunktion $K(\vartheta_0, \gamma) = \mathbb{E}[\rho(\varepsilon_i + \gamma_0 - \gamma)]$, wobei $\vartheta_0 = (\gamma_0, \mathbb{P}^{\varepsilon_i})$ allgemeiner Parameter ist. Im Fall $\Gamma = \mathbb{R}$ und symmetrisch verteilter (ε_i), d.h. $\varepsilon_i \overset{d}{=} -\varepsilon_i$ führt $\rho(x) = \frac{1}{2}x^2$ auf das Stichprobenmittel $\hat\gamma_n$ und $\rho(x) = |x|$ auf den Stichprobenmedian $\hat\gamma_n$. Ein Kompromiss zwischen beiden Schätzern ist der Huber-Schätzer für $\kappa > 0$

$$\hat\gamma_n = \operatorname{argmin}_{\gamma\in\mathbb{R}} \sum_{i=1}^n \rho(Y_i - \gamma), \quad \rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{falls } |x| \leqslant \kappa, \\ \kappa|x| - \frac{\kappa^2}{2}, & \text{falls } |x| > \kappa. \end{cases}$$

Setzt man die Regularitätsannahmen im obigen Satz voraus, so erhält man für den M-Schätzer

$$\sqrt{n}(\hat\gamma_n - \gamma_0) \xrightarrow{d} N\Big(0, \frac{\mathbb{E}[\rho'(\varepsilon_i)^2]}{\mathbb{E}[\rho''(\varepsilon_i)]^2}\Big).$$

Im Fall des Stichprobenmittels ist die asymptotische Varianz also gerade $\mathbb{E}[\varepsilon_i^2] = \operatorname{Var}(\varepsilon_i)$. Einsetzen im Fall einer stetigen Dichte f_ε von ε_i liefert heuristisch für den Stichprobenmedian die asymptotische Varianz $\mathbb{E}[\operatorname{sgn}(\varepsilon_i)^2]/\,\mathbb{E}[2\delta_0(\varepsilon_i)]^2 = (4f_\varepsilon(0))^{-1}$ sowie für den Huber-Schätzer $\mathbb{E}[\varepsilon_i^2 \wedge \kappa^2]/\,\mathbb{P}(|\varepsilon_i| \leqslant \kappa)^2$ (rigorose Herleitungen ggf. in den Übungen).

3.27 Satz. *Es sei $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})_{n\geqslant 1}$ mit $\Theta \subseteq \mathbb{R}^k$ eine Folge Hellinger-differenzierbarer Produktexperimente mit Loglikelihoodfunktion (zu einer Beobachtung) $\ell(\vartheta, x) = \log(\frac{d\mathbb{P}_\vartheta}{d\mu}(x))$. Es gelte:*

(a) $\Theta \subseteq \mathbb{R}^k$ ist kompakt und ϑ_0 liegt im Innern $\operatorname{int}(\Theta)$ von Θ.

(b) Es gilt $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta_0}$ für alle $\vartheta \neq \vartheta_0$ (Identifizierbarkeitsbedingung).

(c) $\vartheta \mapsto \ell(\vartheta, x)$ ist stetig auf Θ und zweimal stetig differenzierbar in einer Umgebung U von ϑ_0 für alle $x \in \mathcal{X}$.

(d) Es gibt $H_0, H_2 \in L^1(\mathbb{P}_{\vartheta_0})$ und $H_1 \in L^2(\mathbb{P}_{\vartheta_0})$ mit $\sup_{\vartheta\in\Theta}|\ell(\vartheta, x)| \leqslant H_0(x)$ und $\sup_{\vartheta\in U}|\dot\ell(\vartheta, x)| \leqslant H_1(x)$, $\sup_{\vartheta\in U}|\ddot\ell(\vartheta, x)| \leqslant H_2(x)$, $x \in \mathcal{X}$.

(e) Die Fisher-Informationsmatrix (zu einer Beobachtung) $I(\vartheta_0) = \mathbb{E}_{\vartheta_0}[(\dot\ell(\vartheta_0))(\dot\ell(\vartheta_0))^\top]$ *ist strikt positiv definit (Notation:* $I(\vartheta_0) > 0$*).*

Dann erfüllt der MLE $\hat\vartheta_n$

$$\sqrt{n}(\hat\vartheta_n - \vartheta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n I(\vartheta_0)^{-1} \dot\ell(\vartheta_0) + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(1).$$

Insbesondere ist $\hat\vartheta_n$ *unter* $\mathbb{P}_{\vartheta_0}^{\otimes n}$ *asymptotisch normalverteilt mit Rate* $n^{-1/2}$ *und asymptotischer Kovarianzmatrix* $I(\vartheta_0)^{-1}$*:*

$$\sqrt{n}(\hat\vartheta_n - \vartheta_0) \xrightarrow{d} N(0, I(\vartheta_0)^{-1}).$$

Ferner gilt die Formel $I(\vartheta_0) = -\mathbb{E}_{\vartheta_0}[\ddot\ell(\vartheta_0)]$*.*

Beweis. Setze $g(\vartheta) = \vartheta$, $\Gamma = \Theta$, $K_n(\vartheta, x) := -\frac{1}{n}\sum_{i=1}^n \ell(\vartheta, x_i)$, $x \in \mathcal{X}^n$, sowie $K(\vartheta_0, \vartheta) := -\mathbb{E}_{\vartheta_0}[\ell(\vartheta)]$. Dann ist K_n ein Kontrastprozess zur Kontrastfunktion K, und wir weisen die Bedingungen (A1)-(A3), (B1)-(B2) mit $U(\vartheta_0) = V(\vartheta_0) = I(\vartheta_0)$ nach.

(A1) Dies folgt aus Θ kompakt.

(A2) Wegen $\ell(\bullet, x) \in C(\Theta)$ ist K_n stetig und dominierte Konvergenz mit $|\ell(\vartheta) - \ell(\vartheta')| \leqslant 2H_0$ liefert

$$|K(\vartheta_0, \vartheta) - K(\vartheta_0, \vartheta')| \leqslant \mathbb{E}_{\vartheta_0}[|\ell(\vartheta) - \ell(\vartheta')|] \xrightarrow{\vartheta' \to \vartheta} 0.$$

(A3) Mit dem starken Gesetz der großen Zahlen folgt $\mathbb{P}_{\vartheta_0}$-f.s.:

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{|\vartheta - \vartheta'| < \delta} |K_n(\vartheta, x) - K_n(\vartheta', x)| \\
&\leqslant \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \sup_{|\vartheta - \vartheta'| < \delta} |\ell(\vartheta, x_i) - \ell(\vartheta', x_i)| \\
&= \mathbb{E}_{\vartheta_0}\Big[\sup_{|\vartheta - \vartheta'| < \delta} |\ell(\vartheta) - \ell(\vartheta')| \Big].
\end{aligned}$$

Das Argument im letzten Erwartungswert ist durch $2H_0$ beschränkt und mit dominierter Konvergenz sowie gleichmäßiger Stetigkeit von ℓ auf dem Kompaktum Θ erhalten wir, dass der letzte Erwartungswert für $\delta \to 0$ gegen Null konvergiert. Dies zeigt die Straffheit von K_n (mit f.s.-Konvergenz in Satz 3.23). Insbesondere ist wegen (A1)-(A3) $\hat\vartheta_n$ konsistent.

(B1) Der zentrale Grenzwertsatz liefert wegen $|\dot\ell(\vartheta)| \leqslant H_1 \in L^2$ unter $\mathbb{P}_{\vartheta_0}^{\otimes n}$

$$\sqrt{n}\dot K_n(\vartheta_0) \xrightarrow{d} N(0, \mathrm{Var}_{\vartheta_0}(\dot\ell(\vartheta_0))) = N(0, I(\vartheta_0)).$$

(B2) Mit $\mathbb{E}_\vartheta[\dot\ell(\vartheta)] = 0$, $\vartheta \in U$, aus Lemma 2.40 erhalten wir

$$\mathbb{E}_{\vartheta_0}[\dot\ell(\vartheta)] = \mathbb{E}_{\vartheta_0}\Big[\dot\ell(\vartheta)\frac{L(\vartheta_0) - L(\vartheta)}{L(\vartheta_0)}\Big].$$

Wir verwenden nun $\nabla_\vartheta(\dot\ell(\vartheta)\frac{L(\vartheta_0)-L(\vartheta)}{L(\vartheta_0)})|_{\vartheta=\vartheta_0} = -\dot\ell(\vartheta_0)\dot\ell(\vartheta_0)^\top$ (Produktregel!) und erhalten

$$\mathbb{E}_{\vartheta_0}[\ddot\ell(\vartheta_0)] = \nabla_\vartheta\,\mathbb{E}_{\vartheta_0}[\dot\ell(\vartheta)]|_{\vartheta=\vartheta_0} = \mathbb{E}_{\vartheta_0}[-\dot\ell(\vartheta_0)\dot\ell(\vartheta_0)^\top] = -I(\vartheta_0).$$

Hier haben wir wiederum Integration und Differentiation mittels dominierter Konvergenz vertauscht. Wir haben $\mathbb{P}_{\vartheta_0}$-f.s. mit dem starken Gesetz der großen Zahlen

$$V_n(\vartheta_0, x) := -\frac{1}{n}\sum_{i=1}^n \ddot\ell(\vartheta, x_i) \xrightarrow{n\to\infty} -\,\mathbb{E}[\ddot\ell(\vartheta_0)] = I(\vartheta_0).$$

Weiterhin gilt auf $\Omega_{\delta,n} := \{|\vartheta_n - \vartheta_0| < \delta\}$ (ϑ_n aus (B2)):

$$\mathbb{E}_{\vartheta_0}[|V_n(\vartheta_n) - V_n(\vartheta_0)|\mathbf{1}_{\Omega_{\delta,n}}] \leqslant \mathbb{E}_{\vartheta_0}\Big[\sup_{|\vartheta-\vartheta_0|<\delta}|\ddot\ell(\vartheta) - \ddot\ell(\vartheta_0)|\Big].$$

Der Ausdruck im rechten Erwartungswert ist unabhängig von n, konvergiert für $\delta \to 0$ gegen null (Stetigkeit von $\ddot\ell$) und ist durch $2H_2$ dominiert, so dass

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{E}_{\vartheta_0}[|V_n(\vartheta_n) - V_n(\vartheta_0)|\mathbf{1}_{\Omega_{\delta,n}}] = 0$$

folgt. Mit $\lim_{\delta\to0}\lim_{n\to\infty}\mathbb{P}_{\vartheta_0}^{\otimes n}(\Omega_{\delta,n}) = 1$ und der Konvergenz von $V_n(\vartheta_0)$ erhalten wir daher in $\mathbb{P}_{\vartheta_0}^{\otimes n}$-Wahrscheinlichkeit

$$V_n(\vartheta_n) = V_n(\vartheta_0) + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(1) \xrightarrow{n\to\infty} I(\vartheta_0).$$

□

3.28 Bemerkungen.

(a) Die Fisher-Information $I(\vartheta_0)$ gibt gerade sowohl die asymptotische Varianz der Score-Funktion als auch die lokale Krümmung der Kontrastfunktion $\mathrm{KL}(\vartheta_0\,|\,\bullet)$ beim Minimum ϑ_0 an.

(b) Es ist bemerkenswert, dass unter Regularitätsannahmen in der asymptotischen Verteilung des MLE sowohl Unverzerrtheit als auch Cramér-Rao-Effizienz gilt. Beachte jedoch, dass es weder klar noch im Allgemeinen korrekt ist, dass die Momente ebenfalls konvergieren und dass die Cramér-Rao-Schranke auch asymptotisch gilt.

(c) Oft ist Θ nicht kompakt, aber man kann durch separate Untersuchung die Konsistenz von $\hat\vartheta_n$ nachweisen. Dann gelten die Konvergenzresultate natürlich weiterhin.

(d) Die Regularitätsbedingungen lassen sich in natürlicher Weise abschwächen. Es reicht aus, dass $(\mathbb{P}_\vartheta)$ bei ϑ_0 Hellinger-differenzierbar ist sowie die Loglikelihoodfunktion ℓ in einer Umgebung von ϑ_0 Lipschitzstetig in ϑ ist mit Lipschitzkonstante in $L^2(\mathbb{P}_{\vartheta_0})$. Dies wird in Satz 5.39 bei van der Vaart unter Verwendung von empirischer Prozesstheorie bewiesen.

(e) Im Fall einer Modellmisspezifikation, wo die wahre Verteilung $\mathbb{P}_0$ nicht in $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ enthalten ist (nicht aber die i.i.d.-Annahme verletzt ist), konvergiert der MLE $\hat\vartheta_n$ gegen $\vartheta^* := \operatorname{argmax}_{\vartheta\in\Theta} \int_{\mathcal{X}} \ell(\vartheta, x)\, \mathbb{P}_0(dx)$, sofern ϑ^* existiert und eindeutig ist. Es gilt entsprechend $\vartheta^* = \operatorname{argmin}_{\vartheta\in\Theta} \mathrm{KL}(\mathbb{P}_0 \,|\, \mathbb{P}_\vartheta)$, sofern $\mathbb{P}_0 \ll \mathbb{P}_{\vartheta^*}$, und ϑ^* heißt Kullback-Leiber-Projektion von $\mathbb{P}_0$ auf $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$. Satz 3.25 liefert unter Regularitätsbedingungen, dass asymptotische Normalität $\sqrt{n}(\hat\vartheta_n - \vartheta^*) \to N(0, V^{-1}UV^{-1})$ vorliegt mit $U = \mathbb{E}_0[\dot\ell(\vartheta^*)\dot\ell(\vartheta^*)^\top]$, $V = \mathbb{E}_0[\ddot\ell(\vartheta^*)]$. Im allgemeinen wird dabei $U \neq V$ gelten.

3.29 Beispiel. Bei einer Exponentialfamilie mit natürlichem Parameterraum und natürlicher suffizienter Statistik T erfüllt der MLE (so er existiert und in $\operatorname{int}(\Theta)$ liegt) $\mathbb{E}_{\hat\vartheta(x)}[T] = T(x)$ und die Fisher-Informationsmatrix $I(\vartheta) = \operatorname{Cov}_\vartheta(T)$ (Kovarianzmatrix von T). Es folgt also mit Regularitätsannahmen $\sqrt{n}(\hat\vartheta_n - \vartheta_0) \to N(0, \operatorname{Cov}_{\vartheta_0}(T)^{-1})$ unter $\mathbb{P}_{\vartheta_0}$. Bei einer Bernoullikette $X_1, \dots, X_n$ mit $X_i \sim \mathrm{Bin}(1, p)$ ist $\vartheta = \log(p/(1-p))$ der natürliche Parameter sowie $T(x) = x$. Aus $\sqrt{n}(\hat\vartheta_n - \vartheta_0) \to N(0, p(\vartheta_0)^{-1}(1 - p(\vartheta_0))^{-1})$ folgt mittels Δ-Methode für die p-Parametrisierung $\sqrt{n}(\hat p_n - p_0) \to N(0, p_0(1-p_0))$. Wegen $\hat p_n = \bar X$ ist dieses Resultat natürlich konkret einfach zu überprüfen.

3.30 Definition. Im Rahmen von Satz 3.27 heißt die zufällige Matrix

$$\mathcal{I}_n(x) := -\frac{1}{n}\sum_{i=1}^n \ddot\ell(\hat\vartheta_n(x), x_i)$$

beobachtete Fisher-Informationsmatrix.

3.31 Korollar. *Unter den Voraussetzungen von Satz 3.27 gilt unter* $\mathbb{P}_{\vartheta_0}$

$$\sqrt{n} I(\hat\vartheta_n)^{1/2}(\hat\vartheta_n - \vartheta_0) \xrightarrow{d} N(0, E_k), \quad \sqrt{n}\mathcal{I}_n^{1/2}(\hat\vartheta_n - \vartheta_0) \xrightarrow{d} N(0, E_k).$$

Insbesondere sind für $k = 1$ *und das* $(1-\alpha/2)$*-Quantil* $q_{1-\alpha/2}$ *der Standardnormalverteilung*

$$\big[\hat\vartheta_n - n^{-1/2} I(\hat\vartheta_n)^{-1/2} q_{1-\alpha/2}, \hat\vartheta_n + n^{-1/2} I(\hat\vartheta_n)^{-1/2} q_{1-\alpha/2}\big]$$

und

$$\big[\hat\vartheta_n - n^{-1/2} \mathcal{I}_n^{-1/2} q_{1-\alpha/2}, \hat\vartheta_n + n^{-1/2} \mathcal{I}_n^{-1/2} q_{1-\alpha/2}\big]$$

Konfidenzintervalle für ϑ_0 *zum asymptotischen Vertrauensniveau* $1-\alpha$.

Beweis. Da $\hat\vartheta_n$ konsistent ist und I stetig von ϑ abhängt (benutze Stetigkeit und Dominiertheit von $\ddot\ell$), folgt $I(\hat\vartheta_n) \to I(\vartheta_0)$ in $\mathbb{P}_{\vartheta_0}$-Wahrscheinlichkeit. Ebenso liefert das Argument im obigen Nachweis der Eigenschaft (B2) $\mathcal{I}_n \to I(\vartheta_0)$ $\mathbb{P}_{\vartheta_0}$-f.s. Nach dem Lemma von Slutsky folgt damit die Konvergenzaussage gegen $I(\vartheta_0)^{1/2} N(0, I(\vartheta_0)^{-1}) = N(0, E_k)$. Die Konvergenz in Verteilung impliziert, dass die entsprechenden Konfidenzintervalle asymptotisch das Niveau $1-\alpha$ besitzen (wie im Limesmodell). □

3.32 Beispiel.

(a) Bei natürlichen Exponentialfamilien ist die beobachtete Fisher-Information gerade $\ddot{A}(\hat{\vartheta}_n) = I(\hat{\vartheta}_n)$, also führen beide Ansätze, die Fisher-Informationen zu schätzen, auf dasselbe Verfahren.

(b) Cox (1958) gibt folgendes Beispiel, um die Frage bedingter Inferenz zu klären: es gibt zwei Maschinen, die Messwerte einer interessierenden Größe $\vartheta \in \mathbb{R}$ mit einem $N(0, \sigma_a^2)$-verteilten Fehler, $a = 0, 1$ und $\sigma_0 \neq \sigma_1$, produzieren. In n Versuchen wird zunächst rein zufällig eine Maschine ausgewählt und dann ihr Messwert beobachtet. Wir beobachten also eine mathematische Stichprobe $(Y_i, A_i)_{i=1,\dots,n}$ mit $\mathbb{P}(A_i = 0) = \mathbb{P}(A_i = 1) = 1/2$ und $Y_i \sim N(\vartheta, \sigma_{A_i}^2)$ bedingt auf A_i. Wir erhalten die Loglikelihood-Funktion

$$\ell(\vartheta; y, a) = \text{const.} - \sum_{i=1}^{n} \frac{(y_i - \vartheta)^2}{2\sigma_{a_i}^2}.$$

Der MLE ist also $\hat{\vartheta}_n = (\sum_{i=1}^n Y_i/\sigma_{A_i}^2)/(\sum_{i=1}^n \sigma_{A_i}^{-2})$ und die Fisher-Information $I(\vartheta) = \frac{1}{2\sigma_0^2} + \frac{1}{2\sigma_1^2} =: I$ (unabhängig von ϑ). $\hat{\vartheta}_n$ ist (nach Theorie oder mit direkten Argumenten) asymptotisch normalverteilt mit asymptotischer Varianz $N(0, I^{-1})$, was auf asymptotische Konfidenzintervalle der Form $\hat{\vartheta}_n \pm \frac{1}{\sqrt{n}} I^{1/2} q_{1-\alpha/2}$ führt. Natürlich gilt $I(\hat{\vartheta}_n) = I$, während die beobachtete Fisher-Information

$$\mathcal{I}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma_{A_i}^2} = \frac{\sum_{i=1}^n A_i}{n\sigma_1^2} + \frac{\sum_{i=1}^n (1 - A_i)}{n\sigma_0^2}$$

erfüllt. Damit ist $\mathcal{I}_n^{-1}$ gerade gleich der *bedingten* Varianz $\mathrm{Var}_\vartheta(n^{1/2}\hat{\vartheta}_n \,|\, A_1, \dots, A_n)$. Da wir ja $A_1, \dots, A_n$ beobachten, ist die bedingte Varianz sicherlich ein sinnvolleres Maß für die Güte des Schätzers $\hat{\vartheta}_n$, im konkreten Beispiel der bedingten Normalverteilung können damit sogar einfach nicht-asymptotische bedingte Konfidenzintervalle angegeben werden. Efron und Hinkley (1978) bevorzugen aus diesem Grund auch für allgemeinere Modelle, in denen (approximativ) *ancillary*-Statistiken vorkommen, die Normalisierung mit der beobachteten Fisher-Information $\mathcal{I}_n$ gegenüber der plug-in-Schätzung $I(\hat{\vartheta}_n)$.

4 Testtheorie

4.1 Neyman-Pearson-Theorie

4.1 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit Zerlegung $\Theta = \Theta_0 \dot\cup \Theta_1$. Jede messbare Funktion $\varphi : \mathcal{X} \to [0, 1]$ heißt (randomisierter) Test. φ besitzt Niveau $\alpha \in [0, 1]$, falls $\mathbb{E}_\vartheta[\varphi] \leqslant \alpha$ für alle $\vartheta \in \Theta_0$ gilt. Die Abbildung $\vartheta \mapsto \mathbb{E}_\vartheta[\varphi]$ heißt Gütefunktion von φ. Ein Test φ der Hypothese $H_0 : \vartheta \in \Theta_0$ gegen die Alternative $H_1 : \vartheta \in \Theta_1$ ist ein gleichmäßig bester Test zum Niveau α, falls φ Niveau α besitzt sowie für alle anderen Tests φ' vom Niveau α die *Macht* kleiner (genauer: nicht größer) als die von φ ist:

$$\forall \vartheta \in \Theta_1 : \ \mathbb{E}_\vartheta[\varphi] \geqslant \mathbb{E}_\vartheta[\varphi'].$$

Ein Test φ ist unverfälscht zum Niveau α, falls φ Niveau α besitzt sowie auf der Alternative $\mathbb{E}_\vartheta[\varphi] \geqslant \alpha$, $\vartheta \in \Theta_1$, gilt. φ heißt gleichmäßig bester unverfälschter Test zum Niveau α, falls φ unverfälscht zum Niveau α ist sowie alle anderen unverfälschten Tests φ' zum Niveau α kleinere Macht besitzen.

4.2 Beispiel. Es sei $X_1, \ldots, X_n$ eine $N(\mu, \sigma_0^2)$-verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}$ unbekannt sowie $\sigma_0 > 0$ bekannt. Es soll die einseitige Hypothese $H_0 : \mu \leqslant \mu_0$ gegen $H_1 : \mu > \mu_0$ für ein vorgegebenes $\mu_0 \in \mathbb{R}$ getestet werden. Dies lässt sich durch $\mathcal{X} = \mathbb{R}^n$ mit Borel-σ-Algebra $\mathscr{F}$ und Verteilungen $\mathbb{P}_\mu = N(\mu \mathbf{1}, \sigma_0^2 E_n)$ modellieren, wobei $\Theta = \mathbb{R}$ und $\Theta_0 = (-\infty, \mu_0]$, $\Theta_1 = (\mu_0, \infty)$ gesetzt wird. Der einseitige Gauß-Test beruht auf der unter $N(\mu_0, \sigma_0^2)$ standardnormalverteilten Teststatistik $T(X_1, \ldots, X_n) = \sqrt{n}(\overline{X} - \mu_0)/\sigma_0$. Zu vorgegebenem $\alpha \in (0,1)$ sei K_α das α-Fraktil der Standardnormalverteilung, d.h. $1 - \Phi(K_\alpha) = \alpha$. Dann besitzt der einseitige Gauß-Test $\varphi(X_1, \ldots, X_n) = \mathbf{1}_{\{T(X_1,\ldots,X_n) \geqslant K_\alpha\}}$ das Niveau α; es gilt nämlich nach Konstruktion $\mathbb{P}_\mu(\varphi = 1) = \alpha$ für $\mu = \mu_0$ sowie aus Monotoniegründen $\mathbb{P}_\mu(\varphi = 1) < \alpha$ für $\mu < \mu_0$.

4.3 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein (binäres) statistisches Modell mit $\Theta = \{0, 1\}$. Bezeichnet p_i, $i = 0, 1$, die Dichte von $\mathbb{P}_i$ bezüglich $\mathbb{P}_0 + \mathbb{P}_1$, so heißt ein Test der Form

$$\varphi(x) = \begin{cases} 1, & \text{falls } p_1(x) > k p_0(x) \\ 0, & \text{falls } p_1(x) < k p_0(x) \\ \gamma(x), & \text{falls } p_1(x) = k p_0(x) \end{cases}$$

mit kritischem Wert $k \in \mathbb{R}^+$ und $\gamma(x) \in [0,1]$ Neyman-Pearson-Test.

4.4 Satz (Neyman-Pearson-Lemma)**.**

(a) Jeder Neyman-Pearson-Test φ ist ein (gleichmäßig) bester Test für $H_0 : \vartheta = 0$ gegen $H_1 : \vartheta = 1$ zum Niveau $\mathbb{E}_0[\varphi]$.

(b) Für jedes vorgegebene $\alpha \in (0,1)$ gibt es einen Neyman-Pearson-Test zum Niveau α mit $\gamma(x) = \gamma \in [0,1]$ konstant.

Beweis.

(a) Betrachte einen beliebigen Test φ' vom Niveau $\mathbb{E}_0[\varphi]$. Es gilt $p_1(x) \geqslant k p_0(x)$ für $x \in A := \{\varphi > \varphi'\}$ wegen $\varphi(x) > 0$ sowie $p_1(x) \leqslant k p_0(x)$ für $x \in B := \{\varphi < \varphi'\}$ wegen $\varphi(x) < 1$. Mit der disjunkten Zerlegung $\mathcal{X} = A \cup B \cup \{\varphi = \varphi'\}$ erhalten wir

$$\begin{aligned} \mathbb{E}_1[\varphi] - \mathbb{E}_1[\varphi'] &= \int_{A \cup B} (\varphi - \varphi') p_1 \geqslant \int_A (\varphi - \varphi') k p_0 + \int_B (\varphi - \varphi') k p_0 \\ &= k(\mathbb{E}_0[\varphi] - \mathbb{E}_0[\varphi']) \geqslant 0. \end{aligned}$$

(b) Wir zeigen im Anschluss, dass es ein $k \geqslant 0$ gibt mit $\mathbb{P}_0(p_1 \geqslant k p_0) \geqslant \alpha$ und $\mathbb{P}_0(p_1 > k p_0) \leqslant \alpha$ (k ist $(1-\alpha)$-Quantil von p_1/p_0 unter $\mathbb{P}_0$). Dann besitzt mit $\gamma := (\alpha - \mathbb{P}_0(p_1 > k p_0))/\mathbb{P}_0(p_1 = k p_0)$ bzw. $\gamma \in [0,1]$ beliebig,

falls $\mathbb{P}_0(p_1 = kp_0) = 0$, der entsprechende Neyman-Pearson-Test φ Niveau α: $\mathbb{E}_0[\varphi] = 1_\bullet \mathbb{P}_0(p_1 > kp_0) + \gamma_\bullet \mathbb{P}_0(p_1 = kp_0) = \alpha$.

Es bleibt nachzuweisen, dass $k := \inf\{r \geqslant 0 \,|\, \rho(r) \leqslant \alpha\}$ mit $\rho(r) := \mathbb{P}_0(p_1 > rp_0)$ das gewünschte Quantil ist. Wegen $\mathbb{P}_0(p_0 = 0) = 0$ und σ-Stetigkeit von $\mathbb{P}_0$ gilt $\lim_{r\to\infty} \rho(r) = 0$, und k ist endlich. Weiterhin ist $\rho(r) = 1 - \mathbb{P}_0(p_1/p_0 \leqslant r)$ monoton fallend und rechtsstetig, was aus Eigenschaften der Verteilungsfunktion von p_1/p_0 folgt. Daher gilt $\rho(k) \leqslant \alpha$ und $\rho(r) > \alpha$ für $r < k$, so dass

$$\alpha \leqslant \lim_{r\uparrow k} \rho(r) = \lim_{r\uparrow k} \mathbb{P}_0(p_1 > rp_0) = \mathbb{P}_0(p_1 \geqslant kp_0)$$

aus der σ-Stetigkeit folgt.

□

4.5 Lemma. *Jeder (gleichmäßig) beste Test für $H_0 : \vartheta = 0$ gegen $H_1 : \vartheta = 1$ besitzt $\mathbb{P}_0$-fast sicher und $\mathbb{P}_1$-fast sicher die Form eines Neyman-Pearson-Tests.*

Beweis. Übung! □

4.6 Definition. Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein dominiertes Modell mit $\Theta \subseteq \mathbb{R}$ und Likelihoodfunktion $L(\vartheta, x)$ sowie T eine reellwertige Statistik. Dann besitzt die Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ monotonen Likelihoodquotienten (oder wachsenden Dichtequotienten) in T, falls

(a) $\vartheta \neq \vartheta' \Rightarrow \mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$;

(b) Für alle $\vartheta < \vartheta'$ gibt es eine monoton wachsende Funktion $h(\bullet, \vartheta, \vartheta') : \mathbb{R} \to \mathbb{R}^+ \cup \{+\infty\}$ mit (Konvention $a/0 := +\infty$ für $a > 0$)

$$\frac{L(\vartheta', x)}{L(\vartheta, x)} = h(T(x), \vartheta, \vartheta') \quad \text{für } (\mathbb{P}_\vartheta + \mathbb{P}_{\vartheta'})\text{-f.a. } x \in \mathcal{X}.$$

4.7 Satz. *Ist $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ mit $\Theta \subseteq \mathbb{R}$ eine einparametrische Exponentialfamilie in $\eta(\vartheta)$ und T, so besitzt sie einen monotonen Dichtequotienten, sofern η streng monoton wächst.*

Beweis. Wir können den Likelihood-Quotienten schreiben als

$$\frac{L(\vartheta', x)}{L(\vartheta, x)} = h(T(x), \vartheta, \vartheta') \text{ mit } h(t, \vartheta, \vartheta') = C(\vartheta')C(\vartheta)^{-1} \exp((\eta(\vartheta') - \eta(\vartheta))t).$$

Offensichtlich ist h streng monoton wachsend in t für $\vartheta' > \vartheta$ wegen $\eta(\vartheta') > \eta(\vartheta)$. Die strenge Monotonie impliziert auch, dass $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$ gilt. □

4.8 Beispiel. Beim Binomialmodell $X \sim \mathrm{Bin}(n, p)$ mit $p \in (0, 1)$ liegt eine Exponentialfamilie in $\eta(p) = \log(p/(1-p))$ und $T(x) = x$ vor. η wächst streng monoton, so dass dieses Modell einen monotonen Dichtequotienten in X besitzt. Direkt folgt dies aus der Monotonie bezüglich x des Dichtequotienten:

$$\frac{\binom{n}{x} p^x (1-p)^{n-x}}{\binom{n}{x} r^x (1-r)^{n-x}} = \Big(\frac{p(1-r)}{r(1-p)}\Big)^x \Big(\frac{1-p}{1-r}\Big)^n, \quad x = 0, \ldots, n, \quad p > r.$$

4.9 Satz. *Die Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$, $\Theta \subseteq \mathbb{R}$, besitze monotonen Dichtequotienten in T. Für $\alpha \in (0,1)$ und $\vartheta_0 \in \Theta$ gilt dann:*

(a) Unter allen Tests φ für das einseitige Testproblem $H_0 : \vartheta \leqslant \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$ mit der Eigenschaft $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$ gibt es einen Test φ^*, der die Fehlerwahrscheinlichkeiten erster und zweiter Art gleichmäßig minimiert, nämlich

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } T(x) > k, \\ 0, & \text{falls } T(x) < k, \\ \gamma, & \text{falls } T(x) = k, \end{cases}$$

wobei $k \in \mathbb{R}$, $\gamma \in [0,1]$ gemäß $\mathbb{E}_{\vartheta_0}[\varphi^*] = \alpha$ bestimmt werden.

(b) Dieser Test φ^* ist gleichmäßig bester Test zum Niveau α für $H_0 : \vartheta \leqslant \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$.

4.10 Beispiel. Der einseitige Gauß-Test aus Beispiel 4.2 ist gleichmäßig bester Test, da $N(\mu\mathbf{1}, \sigma_0^2 E_n)$ monotonen Dichtequotienten in $T(x) = \bar{x}$ besitzt.

Beweis.

(a) Die Existenz von k, γ folgt wie im Neyman-Pearson-Lemma. Wähle $\vartheta_2 > \vartheta_1$ beliebig. Wegen des monotonen Likelihoodquotienten und der Bedingung $T(x) \lessgtr k$ in φ^* gilt

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } L(\vartheta_2, x) > h(k, \vartheta_1, \vartheta_2) L(\vartheta_1, x), \\ 0, & \text{falls } L(\vartheta_2, x) < h(k, \vartheta_1, \vartheta_2) L(\vartheta_1, x). \end{cases}$$

Damit ist φ^* gleichmäßig bester Test von $H_0 : \vartheta = \vartheta_1$ gegen $H_1 : \vartheta = \vartheta_2$ zum vorgegebenen Niveau. Insbesondere ist die Fehlerwahrscheinlichkeit zweiter Art $1 - \mathbb{E}_{\vartheta_2}[\varphi^*]$ minimal für $\vartheta_2 > \vartheta_0$ zu vorgegebenen Niveau bei $\vartheta_1 = \vartheta_0$. Für jeden Test φ mit kleinerer Fehlerwahrscheinlichkeit erster Art bei $\vartheta_1 < \vartheta_0$, d.h. $\mathbb{E}_{\vartheta_1}[\varphi] < \mathbb{E}_{\vartheta_1}[\varphi^*]$, gilt $\mathbb{E}_{\vartheta_0}[\varphi] < \mathbb{E}_{\vartheta_0}[\varphi^*]$; denn sonst wäre $\tilde{\varphi} = \kappa\varphi + (1-\kappa)$ mit $\kappa = \frac{1-\mathbb{E}_{\vartheta_1}[\varphi^*]}{1-\mathbb{E}_{\vartheta_1}[\varphi]} \in [0,1)$ ein besserer Test zum Niveau $\mathbb{E}_{\vartheta_1}[\varphi^*]$ von $H_0 : \vartheta = \vartheta_1$ gegen $H_1 : \vartheta = \vartheta_0$ als φ^*. Demnach gilt $\mathbb{E}_{\vartheta_1}[\varphi] \geqslant \mathbb{E}_{\vartheta_1}[\varphi^*]$ für jeden Test φ mit $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$

(b) Da jeder Test φ auf $H_0 : \vartheta = \vartheta_0$ zum Niveau α durch $\tilde{\varphi} = \kappa\varphi + (1-\kappa)$ mit $\kappa = \frac{1-\alpha}{1-\mathbb{E}_{\vartheta_0}[\varphi]}$ zu einem besseren Test mit $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$ gemacht werden kann, bleibt nur noch zu zeigen, dass φ^* das Niveau α für $H_0 : \vartheta \leqslant \vartheta_0$ einhält. In (a) haben wir gesehen, dass φ^* auch bester Test für $H_0 : \vartheta = \vartheta_1$ mit $\vartheta_1 < \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_0$ ist, so dass im Vergleich zum konstanten Test $\varphi = \mathbb{E}_{\vartheta_1}[\varphi^*]$ folgt $\mathbb{E}_{\vartheta_0}[\varphi^*] \geqslant \mathbb{E}_{\vartheta_0}[\varphi] = \mathbb{E}_{\vartheta_1}[\varphi^*]$. Wir schließen $\mathbb{E}_{\vartheta_1}[\varphi] \leqslant \alpha$ für alle $\vartheta_1 < \vartheta_0$.

□

4.11 Bemerkungen.

(a) Die Gütefunktion $G_{\varphi^*}(\vartheta) = \mathbb{E}_\vartheta[\varphi^*]$ ist sogar streng monoton wachsend für alle ϑ mit $G_{\varphi^*}(\vartheta) \in (0,1)$, wie ein ähnlicher Beweis ergibt.

(b) Im Beweis wurde eine Konvexkombination $\tilde{\varphi}$ von Tests betrachtet. Dieses Argument lässt sich gut geometrisch darstellen. Allgemein betrachte bei einem binären Modell mit $(\mathbb{P}_0, \mathbb{P}_1)$ die Menge $C := \{(\mathbb{E}_0[\varphi], \mathbb{E}_1[\varphi]) \mid \varphi \text{ Test}\} \subseteq [0,1]^2$. Diese ist konvex (Menge der Tests ist konvex), abgeschlossen (folgt aus dem Satz von Banach-Alaoglu) und enthält die Diagonale (betrachte konstante Tests). Neyman-Pearson-Tests entsprechen dann gerade der oberen Begrenzungskurve von C.

4.12 Beispiel. Lässt sich die Neyman-Pearson-Theorie genauso auf zweiseitige Testprobleme anwenden? Betrachte dazu das Beispielproblem, anhand einer $\mathrm{Bin}(n,p)$-verteilten Beobachtung die Hypothese $H_0 : p = 1/2$ gegen $H_1 : p \neq 1/2$ zum Niveau α zu testen (Test auf faire Münze). Wäre φ^* ein gleichmäßig bester Test für dieses Problem, so auch für $H_0 : p = 1/2$ gegen $H_1 : p = 3/4$. Dies hieße nach obigem Lemma, dass φ^* Neyman-Pearson-Struktur besitzt und insbesondere fast sicher identisch ist mit dem besten einseitigen Test für $H_0 : p \leqslant 1/2$ gegen $H_1 : p > 1/2$. Damit folgte aber $\mathbb{E}_{1/4}[\varphi^*] < \alpha$ im Widerspruch zur Eigenschaft, dass φ^* auch bester Test für $H_0 : p = 1/2$ gegen $H_1 : p = 1/4$ sein soll (der konstante Test $\tilde{\varphi}(x) = \alpha$ ist dort besser als φ^*). Also gibt es keinen UMP-Test für das beidseitige Testproblem.

Zugrunde liegt das Problem, dass wir auch eher unsinnige Tests zur Konkurrenz zulassen wie einseitige Tests, die auch bei null beobachteten Erfolgen $H_0 : p = 1/2$ (mit sehr großer Wahrscheinlichkeit) akzeptieren. Es zeigt sich, dass eine zufriedenstellende Theorie in der Klasse aller unverfälschten Tests möglich ist. Ist die Gütefunktion eines unverfälschten Tests φ zum Niveau α bei $p = 1/2$ differenzierbar, so folgt aus $\mathbb{E}_0[\varphi] \leqslant \alpha$ und $\mathbb{E}_p[\varphi] \geqslant \alpha$, $p \neq 1/2$, dass $\mathbb{E}_0[\varphi] = \alpha$ gilt und die Ableitung der Gütefunktion bei $p = 1/2$ verschwindet. Diese zusätzliche Bedingung werden wir für Tests in Exponentialfamilien in Verallgemeinerung des obigen Binomialmodells nutzen.

4.13 Satz (Verallgemeinertes NP-Lemma)**.** *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta = \{0,1\}$ ein (binäres) statistisches Modell, p_0, p_1 die entsprechenden Dichten und $T \in L^1(\mathbb{P}_0)$ eine reellwertige Statistik. Ein Test der Form*

$$\varphi(x) = \begin{cases} 1, & \text{falls } p_1(x) > kp_0(x) + lT(x)p_0(x) \\ 0, & \text{falls } p_1(x) < kp_0(x) + lT(x)p_0(x) \\ \gamma, & \text{falls } p_1(x) = kp_0(x) + lT(x)p_0(x) \end{cases}$$

mit $k, l \in \mathbb{R}$ und $\gamma \in [0,1]$, der für $\alpha \in [0,1]$ die Nebenbedingungen

$$\mathbb{E}_0[\varphi] = \alpha \quad \text{und} \quad \mathbb{E}_0[T\varphi] = \alpha\,\mathbb{E}_0[T]$$

erfüllt, maximiert die Güte $\mathbb{E}_1[\varphi]$ in der Menge aller Tests, die diese Nebenbedingungen erfüllen.

Beweis. Übung! □

4.14 Definition. Es sei $\Theta' \subseteq \Theta$. Dann heißt ein Test φ α-ähnlich auf Θ', wenn $\mathbb{E}_\vartheta[\varphi] = \alpha$ für alle $\vartheta \in \Theta'$ gilt.

4.15 Lemma. *Betrachte das Testproblem $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$. Die Parametermenge $\Theta = \Theta_0 \dot\cup \Theta_1$ bilde einen metrischen Raum, $\partial\Theta_0$ bezeichne den topologischen Rand zwischen Hypothese und Alternative. Jeder Test besitze eine stetige Gütefunktion bei allen $\vartheta \in \partial\Theta_0$. Ist dann φ α-ähnlicher Test auf $\partial\Theta_0$ vom Niveau α, der besser ist als alle α-ähnlichen Tests φ' auf $\partial\Theta_0$ (im Sinne von $\forall\vartheta \in \Theta_1 : \mathbb{E}_\vartheta[\varphi] \geqslant \mathbb{E}_\vartheta[\varphi']$), so ist φ gleichmäßig bester unverfälschter Test zum Niveau α.*

Beweis. Aus Stetigkeitsgründen erfüllt jeder unverfälschte Test φ' $G_{\varphi'}(\vartheta) = \alpha$ für $\vartheta \in \partial\Theta_0$, ist also α-ähnlich auf $\partial\Theta_0$. Daher ist φ gleichmäßig bester Test gegenüber allen unverfälschten Tests. φ ist selbst unverfälscht, da φ gleichmäßig besser als der konstante Test $\tilde\varphi(x) = \alpha$ ist. □

4.16 Satz. *$(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ sei eine einparametrische Exponentialfamilie in $\eta(\vartheta)$ und T. $\Theta \subseteq \mathbb{R}$ sei offen, $\vartheta_0 \in \Theta$ und η sei streng monoton (wachsend oder fallend) und stetig differenzierbar um ϑ_0 mit $\eta'(\vartheta_0) \neq 0$. Für $\alpha \in (0,1)$, $k_1 < k_2$ und $\gamma_1, \gamma_2 \in [0,1]$ erfülle der Test*

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } T(x) < k_1 \text{ oder } T(x) > k_2 \\ 0, & \text{falls } T(x) \in (k_1, k_2) \\ \gamma_i, & \text{falls } T(x) = k_i,\ i = 1,2 \end{cases}$$

die Nebenbedingungen

$$\mathbb{E}_{\vartheta_0}[\varphi^*] = \alpha \quad \text{und} \quad \mathbb{E}_{\vartheta_0}[T\varphi^*] = \alpha\,\mathbb{E}_{\vartheta_0}[T].$$

Dann ist φ^* gleichmäßig bester unverfälschter Test zum Niveau α für das zweiseitige Testproblem $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$.

Beweis. Wir zeigen, dass φ^* für $\mathbb{P}_1 := \mathbb{P}_{\vartheta_1}$, $\mathbb{P}_0 := \mathbb{P}_{\vartheta_0}$ für $\vartheta_1 \neq \vartheta_0$ die Form aus dem verallgemeinerten Neyman-Pearson-Lemma besitzt. Mit $a = \eta(\vartheta_1) - \eta(\vartheta_0) \neq 0$, $b = \log(C(\vartheta_1)/C(\vartheta_0))$ gilt

$$L(\vartheta_1, x) > kL(\vartheta_0, x) + lT(x)L(\vartheta_0, x) \iff \exp(aT(x) + b) > lT(x) + k.$$

Wähle nun k, l so, dass die Gerade $t \mapsto lt + k$ die streng konvexe Funktion $t \mapsto \exp(at + b)$ genau bei $t \in \{k_1, k_2\}$ schneidet. Dann gilt

$$L(\vartheta_1, x) > kL(\vartheta_0, x) + lT(x)L(\vartheta_0, x) \iff T(x) \notin [k_1, k_2] \Rightarrow \varphi^*(x) = 1.$$

Analoge Äquivalenzen zeigen, dass φ^* die gewünschte Form besitzt, und für jeden Test φ, der die Nebenbedingungen erfüllt, gilt $\mathbb{E}_{\vartheta_1}[\varphi^*] \geqslant \mathbb{E}_{\vartheta_1}[\varphi]$ für $\vartheta_1 \neq \vartheta_0$. Nach dem vorigen Lemma reicht es nachzuweisen, dass φ^* gleichmäßig bester Test unter allen α-ähnlichen Tests auf $\partial\Theta_0 = \{\vartheta_0\}$ ist; denn wegen dominierter Konvergenz (vergleiche Satz 2.11) ist die Gütefunktion G_φ jedes Tests φ

in ϑ_0 sogar differenzierbar. Für unverfälschte Tests φ besitzt G_φ bei ϑ_0 eine Minimalstelle, so dass

$$G'_\varphi(\vartheta_0) = 0 \Rightarrow \int \varphi(x)\big(C(\vartheta_0)\eta'(\vartheta_0)T(x) + C'(\vartheta_0)\big)\exp(\eta(\vartheta_0)T(x))\,\mu(dx) = 0.$$

Wir erhalten also $\eta'(\vartheta_0)\,\mathbb{E}_{\vartheta_0}[\varphi T] + \alpha C'(\vartheta_0)/C(\vartheta_0) = 0$. Für den konstanten unverfälschten Test $\varphi_\alpha(x) := \alpha$ impliziert dies $\eta'(\vartheta_0)\,\mathbb{E}_{\vartheta_0}[T] + C'(\vartheta_0)/C(\vartheta_0) = 0$, so dass jeder unverfälschte Test φ die angegebenen Nebenbedingungen erfüllt und φ^* gleichmäßig bester unverfälschter Test nach Lemma 4.15 ist. □

4.17 Beispiele.

(a) Es sei $X_1, \dots, X_n \sim N(\vartheta, \sigma_0^2)$ eine mathematische Stichprobe mit $\vartheta \in \mathbb{R}$ unbekannt und $\sigma_0 > 0$ bekannt. Es liegt eine einparametrische Exponentialfamilie in $T(x) = \sum_{i=1}^n x_i$ und $\eta(\vartheta) = \vartheta/\sigma_0^2$ vor. Für alle $\vartheta \in \mathbb{R}$ gilt $\eta'(\vartheta) = \sigma_0^{-2} > 0$, und wir bestimmen einen gleichmäßig besten unverfälschten Test von $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$ gemäß obigem Satz. Aus Symmetriegründen wähle $k_1 = n\vartheta_0 - k$, $k_2 = n\vartheta_0 + k$ und verzichte wegen stetiger Verteilung auf Randomisierung, so dass $\varphi^* = \mathbf{1}(|T(x) - n\vartheta_0| > k)$ gilt. Wir erhalten mit $Z = \sum_{i=1}^n (X_i - \vartheta_0) \sim N(0, n\sigma_0^2)$ unter $\mathbb{P}_{\vartheta_0}$:

$$\mathbb{E}_{\vartheta_0}[\varphi^* T] = \mathbb{E}[(n\vartheta_0 + Z)\mathbf{1}(|Z| > k)] = \mathbb{E}[n\vartheta_0 \mathbf{1}(|Z| > k)] = \mathbb{E}_{\vartheta_0}[T]\,\mathbb{E}_{\vartheta_0}[\varphi^*].$$

Wählt man also $k = \sigma_0\sqrt{n}q_{1-\alpha/2}$ mit dem $(1-\alpha/2)$-Quantil $q_{1-\alpha/2}$ von $N(0,1)$, so gilt $\mathbb{E}_{\vartheta_0}[\varphi^*] = \alpha$, und der beidseitige Gaußtest φ^* ist – wie erwartet – gleichmäßig bester unverfälschter Test.

(b) Ist X eine $\mathrm{Bin}(n,p)$-verteilte Beobachtung mit $p \in (0,1)$ unbekannt, so betrachte das Testproblem $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$ zum Niveau α für ein festes $p_0 \in (0,1)$. Da eine Exponentialfamilie in $T(X) = X$ und $\eta(p) = \log(p/(1-p))$ mit $\eta'(p) > 0$ für alle p vorliegt, betrachte φ^* gemäß obigem Satz. Die Nebenbedingungen sind

$$\mathbb{E}_{p_0}[\varphi^*(X)] = \alpha, \quad \mathbb{E}_{p_0}[\varphi^*(X)(X - np_0)] = 0.$$

Im Fall $p_0 = 1/2$ besitzt $X - np_0$ eine symmetrische Verteilung um Null, und wählt man $k_1 = np_0 - c$, $k_2 = np_0 + c$ sowie $\gamma_1 = \gamma_2 = \gamma$, so gilt $\varphi^*(X) = \mathbf{1}(|X - np_0| > c) + \gamma\mathbf{1}(|X - np_0| = c)$ und aus Symmetriegründen damit $\mathbb{E}_{p_0}[\varphi^*(X)(X - np_0)] = 0$. Wähle daher $c \geqslant 0$ derart, dass $\mathbb{P}_{p_0}(|X - np_0| > c) \leqslant \alpha$, aber $\mathbb{P}_{p_0}(|X - np_0| \geqslant c) \geqslant \alpha$. Der Wert von $\gamma \in [0,1]$ ergibt sich dann aus der Bedingung $\mathbb{P}_{p_0}(|X - np_0| > c) + \gamma\,\mathbb{P}_{p_0}(|X - np_0| = c) = \alpha$.

Im Fall $p_0 \neq 1/2$ ergibt sich für den UMPU-Test φ^* keine(!) Symmetrie um np_0, da die Verteilung von $X - np_0$ nicht mehr symmetrisch um Null ist. Für große n erhält man mittels Normalapproximation asymptotisch wiederum eine symmetrische Form. Für feste n oder im Fall der Poissonapproximation $\mathrm{Bin}(n, p_0) \approx \mathrm{Poiss}(np_0)$ für $np_0 \to \lambda > 0$ ergeben sich jedoch nicht-symmetrische Definitionen. Ein Beispiel ist der Test $\varphi^*(x) = \mathbf{1}(x \notin [1,2])$, der für $n \geqslant 2$

$$\mathbb{E}_{p_0}[\varphi^*(X)(X - np_0)] = 0 \iff 1 - 2p_0 - \tfrac{n^2-3n}{2}p_0^2 = 0$$

erfüllt. Für große n ist die Lösung $p_0 = (\sqrt{2} + o(1))n^{-1}$ und ein UMPU-Binomialtest auf $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$ ist dann nicht symmetrisch um $np_0 = \sqrt{2} + o(1)$, sondern gerade $\varphi^*(x) = \mathbf{1}(x \notin [1, 2])$ (zum Niveau $\mathbb{E}_{p_0}[\varphi^*]$). Der Grund, wieso der Ablehnbereich für X nicht symmetrisch um $\sqrt{2} \approx 1,41$, sondern um $1,5$ liegt, ist, dass wegen der Schiefe der Binomialverteilung (und der Poissonverteilung) ein symmetrischer Ablehnbereich unter $\mathrm{Bin}(n, p)$ für $p < p_0$ nahe bei p_0 zu einer Güte kleiner α führen würde (im Beispiel gilt $\mathbb{E}_{(\sqrt{2}+o(1))n^{-1}}[\varphi^*] < \mathbb{E}_{1,5n^{-1}}[\varphi^*]$, was einen verfälschten Test unter $H_0 : p = 1,5n^{-1}$ ergäbe).

4.2 Likelihood-Quotienten- und χ^2-Test

Inspiriert vom Neyman-Pearson-Test für einfache Hypothesen und Alternativen definieren wir:

4.18 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein dominiertes statistisches Modell mit Likelihoodfunktion L. Likelihood-Quotienten-Test für $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$ heißt jeder Test der Form

$$\varphi(x) = \begin{cases} 1, & \text{falls } \Lambda(x) > k \\ 0, & \text{falls } \Lambda(x) < k \\ \gamma(x), & \text{falls } \Lambda(x) = k \end{cases} \qquad \text{mit } \Lambda(x) := \frac{\sup_{\vartheta\in\Theta_1} L(\vartheta, x)}{\sup_{\vartheta\in\Theta_0} L(\vartheta, x)} \in [0, +\infty]$$

und $k \in \mathbb{R}^+$, $\gamma(x) \in [0, 1]$ geeignet.

4.19 Bemerkung. Im Allgemeinen liegt Θ_1 dicht in Θ und die Likelihood-Funktion ist stetig in ϑ. Dann gilt $\sup_{\vartheta\in\Theta_1} L(\vartheta, x) = \sup_{\vartheta\in\Theta} L(\vartheta, x) = L(\hat{\vartheta}(x), x)$ mit einem Maximum-Likelihood-Schätzer $\hat{\vartheta}$. Dies ist auch Grundlage der asymptotischen Theorie.

4.20 Beispiel. Im Fall einer natürlichen Exponentialfamilie erhalten wir für Θ_1 dicht in Θ:

$$\Lambda(x) = \inf_{\vartheta_0\in\Theta_0} \exp(\langle\hat{\vartheta} - \vartheta_0, T(x)\rangle - A(\hat{\vartheta}) + A(\vartheta_0)).$$

Falls der Maximum-Likelihood-Schätzer $\hat{\vartheta}$ im Innern von Θ liegt, so folgt $\mathbb{E}_{\hat{\vartheta}(x)}[T] = T(x)$ gemäß Satz 2.11 und daher nach Lemma 3.16

$$\log(\Lambda(x)) = \inf_{\vartheta_0\in\Theta_0} \mathrm{KL}(\mathbb{P}_{\hat{\vartheta}} \mid \mathbb{P}_{\vartheta_0}).$$

Die Likelihood-Quotienten-Statistik Λ misst hier also in natürlicher Weise den Abstand der zu $\hat{\vartheta} \in \Theta$ gehörenden Verteilung zur Hypothesenmenge $(\mathbb{P}_{\vartheta_0})_{\vartheta_0\in\Theta_0}$.

4.21 Lemma. *In der Situation vom Satz 4.9 über beste einseitige Tests führt der Likelihood-Quotienten-Test gerade auf den angegebenen besten Test.*

Beweis. Schreibe

$$\begin{aligned} \frac{\sup_{\vartheta\in\Theta_1} L(\vartheta, x)}{\sup_{\vartheta\in\Theta_0} L(\vartheta, x)} &= \sup_{\vartheta>\vartheta_0} \frac{L(\vartheta, x)}{L(\vartheta_0, x)} \inf_{\vartheta'\leqslant\vartheta_0} \frac{L(\vartheta_0, x)}{L(\vartheta', x)} \\ &= \sup_{\vartheta>\vartheta_0} h(T(x), \vartheta_0, \vartheta) \inf_{\vartheta'\leqslant\vartheta_0} h(T(x), \vartheta', \vartheta_0). \end{aligned}$$

Dann sind die beiden Funktionen h monoton wachsend in T und damit auch das Supremum, das Infimum sowie das Produkt. Also gilt für einen Likelihood-Quotienten-Test φ sowohl $\varphi(x) = 1$ für $T(x) > \tilde{k}$ als auch $\varphi(x) = 0$ für $T(x) < \tilde{k}$ mit $\tilde{k} \in \mathbb{R}$ geeignet. □

4.22 Bemerkung. Im Fall des zweiseitigen Testproblems aus Satz 4.16 führt der Likelihood-Quotienten-Test zwar auf einen Test mit Ablehnbereich $\{T(x) \notin [K_1, K_2]\}$, allerdings ist er im Allgemeinen nicht mehr unverfälscht, wie folgendes Gegenbeispiel lehrt: $X \sim \text{Poiss}(\vartheta)$ führt auf einen Ablehnbereich $\{X(\log(X/\vartheta_0) - 1) > \tilde{k}\}$, was für $\tilde{k} > 0$ einem einseitigen Ablehnbereich $\{X > \bar{k}\}$ entspricht. Hingegen sind im Fall der Normalverteilung ein- und zweiseitige Gauß- und t-Tests Likelihood-Quotienten-Tests.

4.23 Satz. *Es mögen die Voraussetzungen aus Satz 3.27 gelten. Es sei* $0 \in \mathit{int}(\Theta)$*, und die Hypothesenmenge* $\Theta_0 := \{(\vartheta_1, \dots, \vartheta_r, 0, \dots, 0) \in \Theta \,|\, \vartheta_1, \dots, \vartheta_r \in \mathbb{R}\}$ *liege in einem* r*-dimensionalen Unterraum,* $0 \leqslant r < k$ *(*$\Theta_0 = \{0\}$ *falls* $r = 0$*). Dann gilt für die Fitted-Loglikelihood-Statistik*

$$\lambda_n(x) := \sup_{\vartheta \in \Theta} \sum_{i=1}^n \ell(\vartheta, x_i) - \sup_{\vartheta \in \Theta_0} \sum_{i=1}^n \ell(\vartheta, x_i)$$

unter jedem $\mathbb{P}_{\vartheta_0}$ *mit* $\vartheta_0 \in \Theta_0 \cap \mathit{int}(\Theta)$ *die Konvergenz*

$$2\lambda_n \xrightarrow{d} \chi^2(k - r).$$

Insbesondere besitzt der Likelihood-Quotienten-Test $\varphi_n(x) = \mathbf{1}(\lambda_n(x) > \frac{1}{2} q_{\chi^2(k-r),1-\alpha})$ *mit dem* $(1-\alpha)$*-Quantil der* $\chi^2(k-r)$*-Verteilung auf* $\Theta_0 \cap \mathit{int}(\Theta)$ *asymptotisch das Niveau* $\alpha \in (0, 1)$*.*

Beweis. Im folgenden sei $\Pi_r : \mathbb{R}^k \to \mathbb{R}^r$ die Koordinatenprojektion auf die ersten r Koordinaten. Im Beweis von Satz 3.27 haben wir für den MLE $\hat{\vartheta}_n$ insbesondere gezeigt, dass mit $K_n(\vartheta, x) = -\frac{1}{n} \sum_{i=1}^n \ell(\vartheta, x_i)$ für n hinreichend groß gilt

$$-\dot{K}_n(\vartheta_0) = \ddot{K}_n(\bar{\vartheta}_n)(\hat{\vartheta}_n - \vartheta_0), \quad K_n(\vartheta_0) - K_n(\hat{\vartheta}_n) = \tfrac{1}{2}\langle \ddot{K}_n(\tilde{\vartheta}_n)(\hat{\vartheta}_n - \vartheta_0), \hat{\vartheta}_n - \vartheta_0 \rangle$$

(beachte $\dot{K}_n(\hat{\vartheta}_n) = 0$ für den Minimierer $\hat{\vartheta}_n$) mit Zwischenstellen $\bar{\vartheta}_n$, $\tilde{\vartheta}_n$.

Nun gilt mit (B2) und der Konsistenz von $\hat{\vartheta}_n$ gerade $\ddot{K}_n(\bar{\vartheta}_n) \to I(\vartheta_0)$, $\ddot{K}_n(\tilde{\vartheta}_n) \to I(\vartheta_0)$ in $\mathbb{P}_{\vartheta_0}^{\otimes n}$-Wahrscheinlichkeit, also $\ddot{K}_n(\bar{\vartheta}_n) - \ddot{K}_n(\tilde{\vartheta}_n) = o_{P_{\vartheta_0}^{\otimes n}}(1)$. Mit der MLE-Entwicklung aus Satz 3.27

$$\hat{\vartheta}_n - \vartheta_0 = \frac{1}{n} \sum_{i=1}^n I(\vartheta_0)^{-1} \dot{\ell}(\vartheta_0) + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(n^{-1/2}) = -I(\vartheta_0)^{-1} \dot{K}_n(\vartheta_0) + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(n^{-1/2})$$

folgt unter Verwendung von $\dot{K}_n(\vartheta_0) = O_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(n^{-1/2})$, $\hat{\vartheta}_n - \vartheta_0 = O_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(n^{-1/2})$:

$$\begin{aligned}
&K_n(\vartheta_0) - K_n(\hat{\vartheta}_n) = \tfrac{1}{2}\langle \ddot{K}_n(\tilde{\vartheta}_n)(\hat{\vartheta}_n - \vartheta_0), \hat{\vartheta}_n - \vartheta_0 \rangle \\
&= \tfrac{1}{2}\Big(\langle -\dot{K}_n(\vartheta_0), \hat{\vartheta}_n - \vartheta_0 \rangle + \langle (\ddot{K}_n(\tilde{\vartheta}_n) - \ddot{K}_n(\bar{\vartheta}_n))(\hat{\vartheta}_n - \vartheta_0), \hat{\vartheta}_n - \vartheta_0 \rangle \Big) \\
&= \tfrac{1}{2}\langle \dot{K}_n(\vartheta_0), I(\vartheta_0)^{-1} \dot{K}_n(\vartheta_0) \rangle + O_{P_{\vartheta_0}^{\otimes n}}(n^{-1/2}) o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(n^{-1/2}) + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(1) O_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(n^{-1/2})^2 \\
&= \tfrac{1}{2}\langle \dot{K}_n(\vartheta_0), I(\vartheta_0)^{-1} \dot{K}_n(\vartheta_0) \rangle + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(n^{-1}).
\end{aligned}$$

Vollkommen analog erhält man für den MLE $\hat{\vartheta}_n^0$ über die kleinere Parametermenge Θ_0, indem man formal $\Theta_0 \subseteq \mathbb{R}^k$ mit $\Pi_r\Theta_0 \subseteq \mathbb{R}^r$ identifiziert,

$$K_n(\vartheta_0) - K_n(\hat{\vartheta}_n^0) = \tfrac{1}{2}\langle I^0(\vartheta_0)^{-1}\dot{K}_n^0(\vartheta_0), \dot{K}_n^0(\vartheta_0)\rangle + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(n^{-1}),$$

wobei $\dot{K}_n^0$ den Gradienten von K_n als Funktion der ersten r Argumente und $I^0(\vartheta_0) = \Pi_r I(\vartheta_0)\Pi_r^\top$ die $r \times r$-Fisher-Informationsmatrix bezüglich dieser r Parameterwerte bezeichne. Im ausgearteten Fall $r = 0$ setze einfach $\hat{\vartheta}_n^0 = 0$. Insgesamt erhalten wir

$$\begin{aligned} 2\lambda_n &= 2n(K_n(\hat{\vartheta}_n^0) - K_n(\hat{\vartheta}_n)) \\ &= \langle (I(\vartheta_0)^{-1} - \Pi_r^\top I^0(\vartheta_0)^{-1}\Pi_r)\sqrt{n}\dot{K}_n(\vartheta_0), \sqrt{n}\dot{K}_n(\vartheta_0)\rangle + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(1). \end{aligned}$$

Nach dem zentralen Grenzwertsatz (wie im Beweis von (B1)) gilt $\sqrt{n}\dot{K}_n(\vartheta_0) \xrightarrow{d} N(0, I(\vartheta_0))$, so dass mit Slutskys Lemma

$$2\lambda_n \xrightarrow{d} \langle (E_k - I(\vartheta_0)^{1/2}\Pi_r^\top I^0(\vartheta_0)^{-1}\Pi_r I(\vartheta_0)^{1/2})Z, Z\rangle$$

mit $Z \sim N(0, E_k)$. Die Matrix $M := I(\vartheta_0)^{1/2}\Pi_r^\top I^0(\vartheta_0)^{-1}\Pi_r I(\vartheta_0)^{1/2}$ ist symmetrisch und beschreibt wegen $M^2 = M$ eine Orthogonalprojektion. Als Spur erhalten wir (benutze $tr(AB) = tr(BA)$)

$$tr(M) = tr(I(\vartheta_0)\Pi_r^\top I^0(\vartheta_0)^{-1}\Pi_r) = tr(\Pi_r I(\vartheta_0)\Pi_r^\top I^0(\vartheta_0)^{-1}) = tr(E_r) = r.$$

Also besitzt M Rang r und $E_k - M$ ist Orthogonalprojektion von Rang $k - r$. Diagonalisierung ergibt $E_k - M = O^\top \Pi_{k-r} O$ mit einer orthogonalen Matrix O. Wegen $OZ \sim N(0, E_k)$ (Standardnormalverteilung ist orthogonal invariant) impliziert dies $\langle (E_k - M)Z, Z\rangle \sim \chi^2(k - r)$.

Schließlich bemerke, dass aus der Stetigkeit von ℓ für die Likelihood-Quotienten-Statistik folgt

$$\log \Lambda_n(x) = \log\Big(\frac{\sup_{\vartheta\in\Theta}\prod_{i=1}^n L(\vartheta, x_i)}{\sup_{\vartheta\in\Theta_0}\prod_{i=1}^n L(\vartheta, x_i)}\Big) = \lambda_n(x)$$

und somit auf Grund der Montonie des Logarithmus der Likelihood-Quotienten-Test allgemein einen Ablehnbereich der Form $\{\lambda_n > k\}$ besitzt. Beachte, dass eine Randomisierung asymptotisch vernachlässigbar ist. □

4.24 Bemerkungen.

(a) Allgemeiner kann man Hypothesenmengen Θ_0 betrachten, die r-dimensionale C^1-Untermannigfaltigkeiten von Θ bilden, vergleiche Satz 6.5 in Shao. Außerdem kann auf die Kompaktheit von Θ verzichtet werden, sofern die Konsistenz des Maximum-Likeliohood-Schätzers garantiert ist. Für offene $\Theta \subseteq \mathbb{R}^k$ gilt dann Konvergenz unter ganz H_0, d.h. unter $\mathbb{P}_{\vartheta_0}$ für alle $\vartheta_0 \in \Theta_0$.

(b) Für Anwendungen äußerst nützlich ist, dass die asymptotische Verteilung von λ_n unabhängig von $\vartheta_0 \in \Theta_0$ ist; der Likelihood-Quotienten-Test ist asymptotisch verteilungsfrei.

(c) Die asymptotische Verteilung von $2\lambda_n$ unter lokalen Alternativen $\vartheta = \vartheta_0 + h/\sqrt{n}$ ist eine nicht-zentrale $\chi^2(k-r)$-Verteilung, vergleiche Satz 16.7 in van der Vaart. Für feste Alternativen $\vartheta \in \Theta_1$ und $n \to \infty$ erhalten wir insbesondere Konsistenz des Likelihood-Quotienten-Tests φ_n, das heißt $\lim_{n\to\infty} \mathbb{E}_\vartheta[\varphi_n] = 1$.

(d) Zwei weitere wichtige asymptotische Likelihood-Tests sind der Wald-Test und der Score-Test. Beim Wald-Test für eine einfache Hypothese $H_0 : \vartheta = \vartheta_0$ wird die Teststatistik $W_n = n\langle I(\hat\vartheta_n)(\hat\vartheta_n - \vartheta_0), \hat\vartheta_n - \vartheta_0\rangle$ mit dem Maximum-Likelihood-Schätzer $\hat\vartheta$ betrachtet, die unter den Bedingungen von Satz 3.27 ebenfalls $\chi^2(k)$-verteilt ist, und der Wald-Test ist von der Form $\mathbf{1}(W_n > k)$. Im selben Modell ist Raos Score-Test gegeben durch $\varphi = \mathbf{1}(R_n > k)$ mit $R_n = \frac{1}{n}\langle I(\vartheta_0)^{-1}(\sum_{i=1}^n \dot\ell(\vartheta_0, X_i)), \sum_{i=1}^n \dot\ell(\vartheta_0, X_i))\rangle$, wobei R_n gerade die Approximation von $2\lambda_n$ aus obigem Beweis ist und somit ebenfalls $\chi^2(k)$-verteilt ist.

4.25 Beispiel. Es werde ein Zufallsvektor $N = (N_1, \dots, N_k)$ beobachtet, der der Multinomialverteilung mit Parametern n und $p = (p_1, \dots, p_k)$ folgt. Beachte, dass N suffiziente Statistik bei n unabhängigen multinomialverteilten Beobachtungen mit Parameter $(1, p)$ ist und somit die obige Asymptotik für $n \to \infty$ greift. Außerdem kann wegen $p_k = 1 - \sum_{i=1}^{k-1} p_i$ als Parametermenge $\Theta = \{p \in [0,1]^{k-1} \mid \sum_i p_i \leqslant 1\} \subseteq \mathbb{R}^{k-1}$ verwendet werden. Wir betrachten das Testproblem $H_0 : p = p^0$ gegen $H_1 : p \neq p^0$. Da $\hat p = N/n$ der Maximum-Likelihood-Schätzer von p ist, erhalten wir

$$\lambda_n = \log\Big(\frac{\binom{n}{N_1\cdots N_k}(N_1/n)^{N_1}\cdots(N_k/n)^{N_k}}{\binom{n}{N_1\cdots N_k}(p_1^0)^{N_1}\cdots(p_n^0)^{N_k}}\Big) = \sum_{i=1}^k N_i \log(N_i/(np_i^0)).$$

Beachte nun, dass $\mathbb{E}_{p^0}[(N_i - np_i^0)^2/(np_i^0)] = 1 - p_i^0 \leqslant 1$ gilt, so dass wegen der Entwicklung $(x+h)\log((x+h)/x) = h + h^2/(2x) + o(h^2/x)$ sowie $\sum_i N_i = n = \sum_i np_i^0$ asymptotisch

$$\begin{aligned} 2\lambda_n &= 2\sum_{i=1}^k \Big((N_i - np_i^0) + \frac{(N_i - np_i^0)^2}{2np_i^0} + o((N_i - np_i^0)^2/(np_i^0))\Big) \\ &= \sum_{i=1}^k \frac{(N_i - np_i^0)^2}{np_i^0} + o_P(1) \end{aligned}$$

gilt. Damit konvergiert also auch $\sum_{i=1}^k \frac{(N_i - np_i^0)^2}{np_i^0}$ unter H_0 in Verteilung gegen $\chi^2(k-1)$.

4.26 Definition. Bei Beobachtung eines Zufallsvektors $N = (N_1, \dots, N_k)$, der der Multinomialverteilung mit Parametern n und $p = (p_1, \dots, p_k)$ folgt, heißt $\chi_n^2 := \sum_{i=1}^k \frac{(N_i - np_i^0)^2}{np_i^0}$ Pearson's χ^2-Statistik für die Hypothese $H_0 : p = p^0$ und $\varphi = \mathbf{1}(\chi_n^2 > q_{\chi^2(k-1),1-\alpha})$ χ^2-Test mit dem $(1-\alpha)$-Quantil $q_{\chi^2(k-1),1-\alpha}$ der $\chi^2(k-1)$-Verteilung.

Wir haben also als Folgerung:

4.27 Korollar. *Der χ^2-Test besitzt unter $H_0 : p = p^0$ asymptotisch das Niveau $\alpha \in (0,1)$.*

4.28 Bemerkung. Es gibt mannigfache Verallgemeinerungen, insbesondere bei Hypothesen H_0 der Dimension $0 < r < k-1$ wird p_0 durch einen MLE $\hat{p}^0$ ersetzt, und es ergibt sich asymptotisch eine $\chi^2(k-r-1)$-Verteilung. Der χ^2-Test dient häufig als Goodness-of-fit-Test, beispielsweise können Zufallszahlen darauf getestet werden, ob jede Ziffer mit gleicher Wahrscheinlichkeit auftritt, was dem Fall $k = 10$ und $p_1^0 = \cdots = p_{10}^0 = 0,1$ mit Ziffernlänge n entspricht.

4.29 Beispiel. Klassische Anwendung des χ^2-Tests ist die Überprüfung von Mendels Erbsendaten. Bei einer Erbsensorte gibt es die Ausprägungen rund (A) oder kantig (a) sowie gelb (B) oder grün (b). Die Merkmale *rund* und *gelb* sind der Theorie nach dominant, so dass die Genotypen AA, Aa, aA zum Phänotyp *rund* und nur der Genotyp aa zum Phänotyp *kantig* führt. Ebenso ist *gelb* dominant. Betrachtet man nun Nachkommen des heterozygoten Genotyps AaBb, so sollten die vier Phänotypen im Verhältnis 9:3:3:1 auftreten. Mendels Daten (1865) waren bei $n = 556$ Erbsen 315 AB, 101 aB, 108 Ab, 32 ab.

Als natürliches Modell ergibt sich unter der Hypothese eine Multinormalverteilung mit Parametern n und $p^0 = (9/16, 3/16, 3/16, 1/16)$. Als χ^2-Statistik erhalten wir

$$\chi_n^2 := \tfrac{(315-312{,}75)^2}{312{,}75} + \tfrac{(101-104{,}25)^2}{104{,}25} + \tfrac{(108-104{,}25)^2}{104{,}25} + \tfrac{(32-34{,}75)^2}{34{,}75} \approx 0{,}47.$$

Der sogenannte p-Wert des χ^2-Tests bei diesen Daten beträgt $0{,}9254$ ($\mathbb{P}(X > 0{,}47) \approx 0{,}9254$ für $X \sim \chi^2(3)$), das heißt, dass der χ^2-Test die Nullhypothese zu jedem Niveau $\alpha \leqslant 0{,}9254$ akzeptiert hätte! Diese beeindruckende Güte der Daten hat andererseits zum Verdacht der Datenmanipulation geführt.

5 Asymptotische Effizienz

5.1 LAN und Kontiguität

5.1 Satz. *Ist $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ Hellinger-differenzierbar bei $\vartheta_0 \in int(\Theta) \subseteq \mathbb{R}^k$ mit Likelihoodfunktion L, score $\dot{\ell}(\vartheta_0)$ und Fisher-Information $I(\vartheta_0)$, so gilt im Produktmodell $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ für $n \to \infty$, $h_n \to h \in \mathbb{R}^k$ die LAN-Entwicklung (lokal-asymptotische Normalität)*

$$\log\Big(\prod_{i=1}^n \frac{L(\vartheta_0 + n^{-1/2}h_n, X_i)}{L(\vartheta_0, X_i)}\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle h, \dot{\ell}(\vartheta_0, X_i)\rangle - \frac{1}{2}\langle I(\vartheta_0)h, h\rangle + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(1).$$

5.2 Bemerkung. Im Gaußschen Shift-Modell $\mathbb{P}_\vartheta = N(\vartheta, I(\vartheta_0)^{-1})$, $\vartheta \in \mathbb{R}^k$, gilt nicht-asymptotisch für $h \in \mathbb{R}^k$, $X = (X_1, \ldots, X_n)$

$$\log\Big(\frac{d\,\mathbb{P}_{\vartheta_0+n^{-1/2}h}^{\otimes n}}{d\,\mathbb{P}_{\vartheta_0}^{\otimes n}}(X)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle h, \dot{\ell}(\vartheta_0, X_i)\rangle - \frac{1}{2}\langle I(\vartheta_0)h, h\rangle$$

mit Score $\dot{\ell}(\vartheta_0, X_i) = I(\vartheta_0)(X_i - \vartheta_0)$. Unter $\mathbb{P}_{\vartheta_0}$ besitzt der Score-Term $\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle h, I(\vartheta_0)(X_i - \vartheta_0)\rangle$ Erwartungswert Null und Varianz $\langle I(\vartheta_0)h, h\rangle$. Die

LAN-Entwicklung zeigt also eine Asymptotik der Likelihood-Funktion im Produktmodell wie im Fall des Normalverteilungsmodells.

Beweis. Wir verwenden $\log(1+x) = x - \frac{1}{2}x^2 + x^2 R(x)$ mit $R(x) \to 0$ für $x \to 0$ und setzen

$$W_{n,i} := \frac{\sqrt{L(\vartheta_0 + n^{-1/2}h_n, X_i)} - \sqrt{L(\vartheta_0, X_i)}}{\sqrt{L(\vartheta_0, X_i)}}.$$

Dann gilt

$$\log\Big(\prod_{i=1}^n \frac{L(\vartheta_0 + n^{-1/2}h_n, X_i)}{L(\vartheta_0, X_i)}\Big) = 2\sum_{i=1}^n \Big(W_{n,i} - \frac{1}{2}W_{n,i}^2 + W_{n,i}^2 R(W_{n,i})\Big).$$

Wir bestimmen zunächst Erwartungswerte. Es gilt

$$\begin{aligned}
\mathbb{E}_{\vartheta_0}[2W_{n,i}] &= 2\int \sqrt{L(\vartheta_0 + n^{-1/2}h_n, x)}\sqrt{L(\vartheta_0, x)}\,\mu(dx) - 2 \\
&= -\int \Big(\sqrt{L(\vartheta_0 + n^{-1/2}h_n, x)} - \sqrt{L(\vartheta_0, x)}\Big)^2 \mu(dx) \\
&= -\frac{1}{4}\langle I(\vartheta_0) n^{-1/2}h_n, n^{-1/2}h_n\rangle + o(n^{-1}h_n^2) \\
&= -\frac{1}{4n}\langle I(\vartheta_0)h, h\rangle + o(n^{-1}),
\end{aligned}$$

wobei in der vorletzten Zeile die Hellinger-Ableitung eingesetzt wurde. Die Hellinger-Differenzierbarkeit zeigt auch

$$\mathbb{E}_{\vartheta_0}\Big[\Big(2W_{n,i} - \frac{1}{\sqrt{n}}\langle h, \dot\ell(\vartheta_0, X_i)\rangle\Big)^2\Big] = o(n^{-1}).$$

Zusammen erhalten wir mittels Bias-Varianz-Zerlegung die L^2-Abschätzung

$$\mathbb{E}_{\vartheta_0}\Big[\Big(\sum_{i=1}^n \Big(2W_{n,i} + \frac{1}{4n}\langle I(\vartheta_0)h, h\rangle - \frac{1}{\sqrt{n}}\langle h, \dot\ell(\vartheta_0, X_i)\rangle\Big)\Big)^2\Big] = o(1),$$

was mittels Tschebyschew-Ungleichung impliziert

$$\sum_{i=1}^n 2W_{n,i} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle h, \dot\ell(\vartheta_0, X_i)\rangle - \frac{1}{4}\langle I(\vartheta_0)h, h\rangle + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(1).$$

Aus $\mathbb{E}_{\vartheta_0}[(2W_{n,i} - \frac{1}{\sqrt{n}}\langle h, \dot\ell(\vartheta_0, X_i)\rangle)^2] = o(n^{-1})$ folgt mit dem Gesetz der großen Zahlen

$$\sum_{i=1}^n W_{n,i}^2 = \frac{1}{n}\sum_{i=1}^n \langle h, \tfrac{1}{2}\dot\ell(\vartheta_0, X_i)\rangle^2 + o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(1) \xrightarrow{\mathbb{P}_{\vartheta_0}^{\otimes n}} \tfrac{1}{4}\langle I(\vartheta_0)h, h\rangle.$$

Die Behauptung folgt, wenn $\max_{1\leqslant i\leqslant n}|R(W_{n,i})| = o_{\mathbb{P}_{\vartheta_0}^{\otimes n}}(1)$ gezeigt ist. Da $W_{n,1}, \dots, W_{n,n}$ identisch verteilt sind, erhalten wir

$$\mathbb{P}_{\vartheta_0}^{\otimes n}\Big(\max_{1\leqslant i\leqslant n}|W_{n,i}| > \varepsilon\Big) \leqslant n\,\mathbb{P}_{\vartheta_0}(|W_{n,1}| > \varepsilon),$$

und es bleibt, $\mathbb{P}_{\vartheta_0}(|W_{n,1}| > \varepsilon) = o(n^{-1})$ für alle $\varepsilon > 0$ zu zeigen. Nun gilt mit Markov-Ungleichung

$$\mathbb{P}_{\vartheta_0}\left(\frac{1}{\sqrt{n}}|\langle h, \dot{\ell}(\vartheta_0, X_1)\rangle| > \varepsilon\right) \leqslant \frac{\mathbb{E}_{\vartheta_0}\left[\langle h, \dot{\ell}(\vartheta_0, X_1)\rangle^2 \mathbf{1}(|\langle h, \dot{\ell}(\vartheta_0, X_1)\rangle| > n^{1/2}\varepsilon)\right]}{n\varepsilon^2}.$$

Wegen $\mathbb{E}_{\vartheta_0}[\langle h, \dot{\ell}(\vartheta_0, X_1)\rangle^2] < \infty$ folgt mit dominierter Konvergenz, dass die rechte Seite $o(n^{-1})$ ist. Mit $\mathbb{E}_{\vartheta_0}[(2W_{n,1} - \frac{1}{\sqrt{n}}\langle h, \dot{\ell}(\vartheta_0, X_1)\rangle)^2] = o(n^{-1})$ und Tschebyshew-Ungleichung folgt daher auch $\mathbb{P}_{\vartheta_0}(|W_{n,1}| > \varepsilon) = o(n^{-1})$ für alle $\varepsilon > 0$. □

5.3 Definition. Eine Folge von Wahrscheinlichkeitsmaßen $(\mathbb{Q}_n)$ auf $(\mathcal{X}_n, \mathscr{F}_n)$ heißt contiguous bezüglich einer anderen Folge $(\mathbb{P}_n)$ von Wahrscheinlichkeitsmaßen auf denselben Messräumen, falls die Implikation $\mathbb{P}_n(A_n) \to 0 \Rightarrow \mathbb{Q}_n(A_n) \to 0$ für alle Folgen von Ereignissen $A_n \in \mathscr{F}_n$ gilt. Notation $(\mathbb{Q}_n) \lhd (\mathbb{P}_n)$.

5.4 Bemerkung. Im Fall $\mathbb{Q}_n = \mathbb{Q}$ und $\mathbb{P}_n = \mathbb{P}$ folgt aus $(\mathbb{Q}_n) \lhd (\mathbb{P}_n)$ sofort die Absolutstetigkeit $\mathbb{Q} \ll \mathbb{P}$. Gilt andererseits $\mathbb{Q} \ll \mathbb{P}$, das heißt $\mathbb{P}(A) = 0 \Rightarrow \mathbb{Q}(A) = 0$ für alle Ereignisse A, so wird in der Maßtheorie (Stochastik II) gezeigt, dass sogar $\mathbb{P}(A_n) \to 0 \Rightarrow \mathbb{Q}(A_n) \to 0$ für Folgen von Ereignissen (A_n) und somit $(\mathbb{Q}_n) \lhd (\mathbb{P}_n)$ gilt. Man kann Kontiguität allgemein als asymptotische Absolutstetigkeit interpretieren.

5.5 Lemma. *Es gilt $(\mathbb{Q}_n) \lhd (\mathbb{P}_n)$ genau dann, wenn für jede Folge von Statistiken $T_n : \mathcal{X}_n \to \mathbb{R}$ gilt $T_n \xrightarrow{\mathbb{P}_n} 0 \Rightarrow T_n \xrightarrow{\mathbb{Q}_n} 0$.*

Beweis. Für '⇒' schließe $T_n \xrightarrow{\mathbb{P}_n} 0 \Rightarrow \forall \varepsilon > 0 : \mathbb{P}_n(|T_n| > \varepsilon) \to 0 \Rightarrow \forall \varepsilon > 0 : \mathbb{Q}_n(|T_n| > \varepsilon) \to 0 \Rightarrow T_n \xrightarrow{\mathbb{Q}_n} 0$. Für '⇐' gelte $\mathbb{P}_n(A_n) \to 0$ und setze $T_n = \mathbf{1}_{A_n}$. Dann folgt $T_n \xrightarrow{\mathbb{P}_n} 0 \Rightarrow T_n \xrightarrow{\mathbb{Q}_n} 0 \Rightarrow \mathbb{Q}_n(A_n) \to 0$. □

5.6 Beispiel. Für $\mathbb{P}_n = U([0,1])$, $\mathbb{Q}_n = U([0, \frac{1}{n}])$ gilt $\mathbb{Q}_n \ll \mathbb{P}_n$, aber nicht $(\mathbb{Q}_n) \lhd (\mathbb{P}_n)$; denn $\mathbb{P}_n([0, \frac{1}{n}]) \to 0$, aber $\mathbb{Q}_n([0, \frac{1}{n}]) = 1 \not\to 0$. Beachte die Likelihood-Asymptotik $\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} = n\mathbf{1}([0, \frac{1}{n}]) \xrightarrow{\mathbb{P}_n} 0$, obwohl $\mathbb{E}_{\mathbb{P}_n}[\frac{d\mathbb{Q}_n}{d\mathbb{P}_n}] = 1$.

5.7 Lemma. *Für Wahrscheinlichkeitsmaße $\mathbb{P}_n$, $\mathbb{P}$ auf einem polnischen Raum $(S, \mathfrak{B}_S)$ gilt $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ genau dann, wenn $\liminf_{n\to\infty} \int f \, d\mathbb{P}_n \geqslant \int f \, d\mathbb{P}$ für alle stetigen $f : S \to [0, \infty)$.*

5.8 Bemerkung. Diese Aussage ist Teil des Portmanteau-Lemmas zur Charakterisierung schwacher Konvergenz (vgl. Stochastik II).

Beweis. Für '⇒' sei $f : S \to [0, \infty)$ stetig. Dann ist $f \wedge R$ stetig und beschränkt für alle $R > 0$ und somit gilt $\int (f \wedge R) \, d\mathbb{P}_n \to \int (f \wedge R) \, d\mathbb{P}$. Damit folgt

$$\liminf_{n\to\infty} \int f \, d\mathbb{P}_n \geqslant \sup_{R>0} \lim_{n\to\infty} \int (f \wedge R) \, d\mathbb{P}_n = \sup_{R>0} \int (f \wedge R) \, d\mathbb{P} = \int f \, d\mathbb{P},$$

wobei die letzte Gleichung mit monotoner Konvergenz folgt.

Für '$\Leftarrow$' sei $f : S \to \mathbb{R}$ stetig und beschränkt. Dann sind $\|f\|_\infty + f, \|f\|_\infty - f$ stetige nicht-negative Funktionen. Somit folgt

$$\liminf_{n\to\infty} \int (\|f\|_\infty + f)\, d\,\mathbb{P}_n \geqslant \int (\|f\|_\infty + f)\, d\,\mathbb{P},$$
$$\liminf_{n\to\infty} \int (\|f\|_\infty - f)\, d\,\mathbb{P}_n \geqslant \int (\|f\|_\infty - f)\, d\,\mathbb{P}\,.$$

Dies zeigt $\lim_{n\to\infty} \int f\, d\,\mathbb{P}_n = \int f\, d\,\mathbb{P}$. □

5.9 Lemma. *Es seien $\mathbb{Q}_n, \mathbb{P}_n$ Wahrscheinlichkeitsmaße auf $(\mathcal{X}_n, \mathscr{F}_n)$ mit Dichten q_n, p_n bezüglich dominierenden Maßen μ_n. Falls $\frac{q_n}{p_n} \xrightarrow{d} L$ unter $\mathbb{P}_n$ gilt mit $\mathbb{E}[L] = 1$, so folgt $(\mathbb{Q}_n) \lhd (\mathbb{P}_n)$.*

5.10 Bemerkung. $\mu_n = \mathbb{P}_n + \mathbb{Q}_n$ ist stets dominierendes Maß. Es gilt

$$\mathbb{P}_n(p_n = 0) = \int_{\{p_n=0\}} p_n\, d\mu_n = 0,$$

so dass $\frac{q_n}{p_n}$ $\mathbb{P}_n$-f.s. wohldefiniert ist. Weiterhin ist für ein Ereignis A

$$\mathbb{E}_{\mathbb{P}_n}[\mathbf{1}_A \frac{q_n}{p_n}] = \int_A \frac{q_n(x)}{p_n(x)} \mathbf{1}(p_n(x) > 0) p_n(x)\, \mu_n(dx) = \mathbb{Q}_n(A \cap \{p_n > 0\}) \leqslant \mathbb{Q}_n(A).$$

Man kann zeigen, dass $\frac{q_n}{p_n}$ die Dichte des absolutstetigen Anteils von $\mathbb{Q}_n$ bezüglich $\mathbb{P}_n$ ist (Lebesgue-Zerlegung).

Beweis. Es gelte $\mathbb{P}_n(A_n) \to 0$, was $\mathbf{1}_{A_n^c} \xrightarrow{\mathbb{P}_n} 1$ impliziert. Nach Slutskys Lemma folgt $\frac{q_n}{p_n}\mathbf{1}_{A_n^c} \xrightarrow{d} L$ unter $\mathbb{P}_n$, und das Portmanteau-Lemma für die Identität auf $\mathbb{R}^+$ zeigt

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{Q}_n(A_n^c) &\geqslant \liminf_{n\to\infty} \int_{\mathcal{X}_n} \mathbf{1}_{A_n^c} \frac{q_n}{p_n}\, d\,\mathbb{P}_n = \liminf_{n\to\infty} \int_{\mathbb{R}^+} x\ \mathbb{P}_n^{\mathbf{1}_{A_n^c}\frac{q_n}{p_n}}(dx) \\ &\geqslant \int_{\mathbb{R}^+} x\ \mathbb{P}^L(dx) = \mathbb{E}[L] = 1,\end{aligned}$$

wobei $\mathbb{P}^L$ die Verteilung von L bezeichnet. Dies beweist $\mathbb{Q}_n(A_n) \to 0$. □

5.11 Bemerkung. Kontiguität $(\mathbb{Q}_n) \lhd (\mathbb{P}_n)$ impliziert andersherum auch, dass, wenn $\frac{q_{n_k}}{p_{n_k}} \xrightarrow{d} L$ unter $\mathbb{P}_{n_k}$ entlang einer Teilfolge (n_k) gilt, dann $\mathbb{E}[L] = 1$ gelten muss. Dies ist Teil des Ersten Lemmas von Le Cam.

5.12 Satz (Le Cams Drittes Lemma)**.** *Sind $\mathbb{Q}_n, \mathbb{P}_n$ Wahrscheinlichkeitsmaße auf $(\mathcal{X}_n, \mathscr{F}_n)$ mit μ_n-Dichten q_n, p_n sowie $T_n : \mathcal{X}_n \to \mathbb{R}^p$ Statistiken, so dass*

$$\left(T_n, \frac{q_n}{p_n}\right) \xrightarrow{d} (T, L) \text{ unter } \mathbb{P}_n \text{ mit } \mathbb{E}[L] = 1,$$

so gilt $T_n \xrightarrow{d} \tilde{\mathbb{P}}$ unter $\mathbb{Q}_n$ mit $\tilde{\mathbb{P}}(B) = \mathbb{E}[\mathbf{1}_B(T)L]$ für $B \in \mathfrak{B}_{\mathbb{R}^p}$.

Insbesondere folgt aus

$$(T_n, \log(q_n/p_n)) \xrightarrow{d} N_{p+1}\Big(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^\top & \sigma^2 \end{pmatrix} \Big) \text{ unter } \mathbb{P}_n$$

mit $\mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p\times p}$, $\sigma > 0$, $\tau \in \mathbb{R}^p$ *die Konvergenz*

$$T_n \xrightarrow{d} N_p(\mu + \tau, \Sigma) \text{ unter } \mathbb{Q}_n\,.$$

5.13 Bemerkung. Aus der Voraussetzung im zweiten Teil folgt die Konvergenz der Randverteilungen $\log(q_n/p_n) \xrightarrow{d} N(-\frac{1}{2}\sigma^2, \sigma^2)$. Beachte dazu

$$\log\Big(\frac{dN(\sigma,1)}{dN(0,1)}(X)\Big) = \sigma X - \tfrac{1}{2}\sigma^2 \sim N(-\tfrac{1}{2}\sigma^2, \sigma^2) \text{ unter } X \sim N(0,1),$$

so dass q_n/p_n asymptotisch die Verteilung des Quotienten zweier Normalverteilungsdichten (jeweils unter dem dominierenden Maß) besitzt.

Beweis. Beachte zunächst, dass $\tilde{\mathbb{P}}$ ein Wahrscheinlichkeitsmaß ist wegen $L \geqslant 0$ und $\mathbb{E}[L] = 1$. Ist $f : \mathbb{R}^p \to [0,\infty)$ stetig und nicht-negativ, so auch $(x,v) \mapsto f(x)v$ auf $\mathbb{R}^p \times [0,\infty)$. Wir schließen mittels Portmanteau-Lemma:

$$\begin{aligned}
\liminf_{n\to\infty} \mathbb{E}_{\mathbb{Q}_n}[f(T_n)] &\geqslant \liminf_{n\to\infty} \mathbb{E}_{\mathbb{P}_n}\Big[f(T_n)\frac{q_n}{p_n}\Big] \\
&= \liminf_{n\to\infty} \int_{\mathbb{R}^p\times[0,\infty)} f(x)v \; \mathbb{P}_n^{(T_n, q_n/p_n)}(dx,dv) \\
&\geqslant \int_{\mathbb{R}^p\times[0,\infty)} f(x)v \; \mathbb{P}^{(T,L)}(dx,dv) \\
&= \mathbb{E}[f(T)L] = \int f\, d\tilde{\mathbb{P}}.
\end{aligned}$$

Dies zeigt wiederum mittels Portmanteau-Lemma $T_n \xrightarrow{d} \tilde{\mathbb{P}}$ unter $\mathbb{Q}_n$.

Für den zweiten Teil sei (T,W) ein $N_{p+1}\Big(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^\top & \sigma^2 \end{pmatrix} \Big)$-verteilter Vektor. Dann folgt mit continuous mapping $(T_n, q_n/p_n) \xrightarrow{d} (T, e^W)$ unter $\mathbb{P}_n$. Nun ist nach obiger Bemerkung $\mathbb{E}[e^W] = 1$ und aus dem ersten Teil folgt $T_n \xrightarrow{d} \tilde{\mathbb{P}}$ unter $\mathbb{Q}_n$ mit $\tilde{\mathbb{P}}(B) = \mathbb{E}[\mathbf{1}_B(T)e^W]$. Die charakteristische Funktion von $\tilde{\mathbb{P}}$ berechnet sich zu

$$\begin{aligned}
\int e^{i\langle u,x\rangle}\tilde{\mathbb{P}}(dx) &= \mathbb{E}[e^{i\langle u,T\rangle}e^W] \\
&= \mathbb{E}[\exp(i\langle (u,-i)^\top, (T,W)^\top\rangle)] \\
&= \exp\Big(i\Big\langle \begin{pmatrix} u \\ -i \end{pmatrix}, \begin{pmatrix} \mu \\ -\sigma^2/2 \end{pmatrix}\Big\rangle - \frac{1}{2}\Big\langle \begin{pmatrix} \Sigma & \tau \\ \tau^\top & \sigma^2 \end{pmatrix}\begin{pmatrix} u \\ -i \end{pmatrix}, \begin{pmatrix} u \\ -i \end{pmatrix}\Big\rangle\Big) \\
&= \exp\big(i\langle \mu+\tau, u\rangle - \tfrac{1}{2}\langle \Sigma u, u\rangle\big).
\end{aligned}$$

Letzteres ist die charakteristische Funktion von $N(\mu+\tau,\Sigma)$, und der Eindeutigkeitssatz für charakteristische Funktionen zeigt $\tilde{\mathbb{P}} = N(\mu+\tau,\Sigma)$. □

5.14 Korollar. *Betrachte die fitted-loglikelihood-Statistik λ_n unter den Voraussetzungen von Satz 4.23. Dann gilt unter den lokalen Alternativen $\mathbb{Q}_n = \mathbb{P}^{\otimes n}_{\vartheta_0+n^{-1/2}h_n}$*

$$2\lambda_n \xrightarrow{d} |(E_k - \tilde{\Pi}_r)(I(\vartheta_0)^{1/2}h + Z)|^2, \quad Z \sim N(0, E_k),$$

mit der Orthogonalprojektion $\tilde{\Pi}_r$ auf $\operatorname{span}(I(\vartheta_0)^{1/2}e_1, \dots, I(\vartheta_0)^{1/2}e_r)$ mit den kanonischen Einheitsvektoren e_i (sowie $\tilde{\Pi}_r = 0$ im Fall $r = 0$).

Die Gütefunktion des Likelihood-Quotienten-Tests unter $\mathbb{Q}_n$ erfüllt asymptotisch

$$\lim_{n\to\infty} \mathbb{Q}_n(\varphi_n = 1) = \mathbb{P}\left(|(E_k - \tilde{\Pi}_r)(I(\vartheta_0)^{1/2}h + Z)|^2 > q_{\chi^2(k-r),1-\alpha}\right).$$

Beweis. Im Beweis von Satz 4.23 haben wir gezeigt, dass

$$2\lambda_n = \left|(E_k - M)I(\vartheta_0)^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}(\vartheta_0, X_i)\right|^2 + o_{\mathbb{P}^{\otimes n}_{\vartheta_0}}(1)$$

mit der Orthogonalprojektion $M = I(\vartheta_0)^{1/2}\Pi_r^\top I^0(\vartheta_0)^{-1}\Pi_r I(\vartheta_0)^{1/2}$. Das Bild von M ist (einsetzen!) $\operatorname{span}(I(\vartheta_0)^{1/2}e_1, \dots, I(\vartheta_0)^{1/2}e_r)$, so dass $M = \tilde{\Pi}_r$ gilt. Setzen wir $\mathbb{P}_n = \mathbb{P}^{\otimes n}_{\vartheta_0}$, $\mathbb{Q}_n = \mathbb{P}^{\otimes n}_{\vartheta_0+n^{-1/2}h_n}$ und $T_n = I(\vartheta_0)^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}(\vartheta_0, X_i)$ in Le Cams drittes Lemma ein, so erhalten wir mit der LAN-Entwicklung aus Satz 5.1 und Slutskys Lemma

$$(T_n, \log(q_n/p_n)) \xrightarrow{d} N_{k+1}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\langle I(\vartheta_0)h, h\rangle \end{pmatrix}, \begin{pmatrix} E_k & I(\vartheta_0)^{1/2}h \\ (I(\vartheta_0)^{1/2}h)^\top & \langle I(\vartheta_0)h, h\rangle \end{pmatrix}\right)$$

unter $\mathbb{P}_n$ und somit

$$T_n \xrightarrow{d} N_k(I(\vartheta_0)^{1/2}h, E_k) \text{ unter } \mathbb{Q}_n.$$

Also folgt mit continuous mapping unter $\mathbb{Q}_n$ mit $Z \sim N(0, E_k)$

$$\left|(E_k - M)I(\vartheta_0)^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}(\vartheta_0, X_i)\right|^2 \xrightarrow{d} |(E_k - \tilde{\Pi}_r)(I(\vartheta_0)^{1/2}h + Z)|^2.$$

Damit folgt auch die Konvergenz von $2\lambda_n$ unter $\mathbb{Q}_n$ gegen diesen Grenzwert, weil $Y_n = o_{\mathbb{P}_n}(1) \Rightarrow Y_n = o_{\mathbb{Q}_n}(1)$ nach Lemma 5.5 aus $(\mathbb{Q}_n) \lhd (\mathbb{P}_n)$, einer Konsequenz von $\log(q_n/p_n) \xrightarrow{d} N(-\sigma^2/2, \sigma^2)$ für $\sigma^2 = \langle I(\vartheta_0)h, h\rangle$, folgt.

Diese Verteilungskonvergenz impliziert direkt die Konvergenz der Güte des Likelihood-Quotienten-Tests. □

5.15 Bemerkung. Wegen $(E_k - \tilde{\Pi}_r)I(\vartheta_0)^{1/2}h = I(\vartheta_0)^{1/2}(E_k - \Pi_r)h$ spielt nur der Anteil von h orthogonal zur Nullypothese H_0 asymptotisch eine Rolle. Die Grenzverteilung $|(E_k - \tilde{\Pi}_r)(I(\vartheta_0)^{1/2}h + Z)|^2$ ist eine sogenannte nicht-zentrale χ^2-Verteilung. Für $|(E_k - \Pi_r)h| \to \infty$ konvergiert der Grenzwert der Likelihood-Quotienten-Test-Güte gegen eins. Man kann in der Tat zeigen, dass für lokale Alternativen ϑ_n mit $n^{1/2}|(E_k - \Pi_r)\vartheta_n| \to \infty$ der Likelihood-Quotienten-Test

φ_n stets konsistent ist, also die Güte gegen eins konvergiert. Andererseits ist für ϑ_n mit $\sqrt{n}|(E_k - \Pi_r)\vartheta_n| \to 0$ die Asymptotik von φ_n dieselbe wie unter H_0 (wähle $h_n \to h := 0$ im Satz), diese lokalen Alternativen sind asymptotisch ununterscheidbar von H_0.

Das Korollar zeigt auch, dass die Güte von φ_n besser ist für h in Richtungen der Eigenvektoren von $I(\vartheta_0)$ zu großen Eigenwerten als in Richtungen der Eigenvektoren zu kleinen Eigenwerten. Dies liegt in der Konstruktion des Likelihood-Quotienten-Tests begründet und für isotrope Alternativen kann er verbessert werden.

5.2 Asymptotische untere Schranken

5.16 Satz. *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ Hellinger-differenzierbar bei $\vartheta_0 \in int(\Theta) \subseteq \mathbb{R}^k$ mit invertierbarer Fisher-Information $I(\vartheta_0)$. Sind $T_n : \mathcal{X}^n \to \mathbb{R}^k$ Statistiken, die unter $\mathbb{P}^{\otimes n}_{\vartheta_0+n^{-1/2}h}$ in Verteilung für alle $h \in \mathbb{R}^k$ konvergieren, so existiert eine randomisierte Statistik T in $(\mathbb{R}^k, \mathfrak{B}_{\mathbb{R}^k}, (\mathbb{P}_h)_{h\in\mathbb{R}^k})$ mit $\mathbb{P}_h = N(h, I(\vartheta_0)^{-1})$, so dass $T_n \xrightarrow{d} T$ für alle h gilt, das heißt $\mathbb{E}_{\vartheta_0+n^{-1/2}h}[f(T_n)] \to \int\int f(y)\, T(x, dy)\, \mathbb{P}_h(dx)$ für alle beschränkten $f \in C(\mathbb{R}^k)$ und $h \in \mathbb{R}^k$.*

5.17 Bemerkungen.

(a) Eine randomisierte Statistik T ist hier ein Markovkern (eine reguläre bedingte Wahrscheinlichkeit) mit $T(x, \bullet)$ Wahrscheinlichkeitsmaß auf $\mathfrak{B}_{\mathbb{R}^p}$ für jedes $x \in \mathbb{R}^k$, sowie $T(\bullet, B)$ messbar für jedes $B \in \mathfrak{B}_{\mathbb{R}^p}$. Dem Statistiker ist also erlaubt, aufgrund der Beobachtung x einen Wert in $\mathbb{R}^p$ durch ein unabhängiges Zufallsexperiment (mit von x abhängigen Wahrscheinlichkeiten) zu gewinnen. Ist $S : \mathbb{R}^k \to \mathbb{R}^p$ eine nicht-randomisierte Statistik, so besitzt sie dieselbe Verteilung wie die randomisierte Statistik $T(x, \bullet) = \delta_{S(x)}$. Im Allgemeinen kann man randomisierte Statistiken auch durch $\tilde{T}(X, U)$ beschreiben, wobei $U \sim U([0, 1])$ eine von X unabhängige Zufallsvariable angibt und $(x, u) \mapsto \tilde{T}(x, u)$ messbar ist, vgl. van der Vaart.

(b) Grob gesagt, existiert für jede Folge statistischer Prozeduren im Produktmodell eine Prozedur im normalverteilten Grenzmodell, die für lokale Parameter um ϑ_0 die asymptotische Verteilung der Prozeduren im Produktmodell besitzt. Wie wir sehen werden, ist dies der Ansatzpunkt, um asymptotische untere Schranken im Produktmodell zu beweisen, indem man sie auf das Normalverteilungsmodell zurückführt.

Beweis. Unter $\mathbb{P}^{\otimes n}_{\vartheta_0}$ konvergieren T_n nach Annahme und $\Delta_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}(\vartheta_0, X_i) \xrightarrow{d} N(0, I(\vartheta_0))$ gemäß zentralem Grenzwertsatz in Verteilung, so dass $(T_n, \Delta_n) = O_{\mathbb{P}^{\otimes n}_{\vartheta_0}}(1)$, also Straffheit gilt. Nach dem Satz von Prohorov (Stochastik II) existieren eine Teilfolge (n_k) und Zufallsvariablen (S, Δ) mit

$$(T_{n_k}, \Delta_{n_k}) \xrightarrow{d} (S, \Delta),$$

wobei $\Delta \sim N(0, I(\vartheta_0))$ aus der Konvergenz der Randverteilung folgt. Die LAN-Entwicklung aus Satz 5.1 in Verbindung mit Slutskys Lemma ergibt daher

$$\Big(T_{n_k}, \log\Big(\prod_{i=1}^{n_k} \frac{L(\vartheta_0 + n_k^{-1/2}h, X_i)}{L(\vartheta_0, X_i)}\Big)\Big) \xrightarrow{d} \Big(S, \langle h, \Delta\rangle - \tfrac{1}{2}\langle I(\vartheta_0)h, h\rangle\Big).$$

Le Cams Drittes Lemma liefert

$$T_{n_k} \to \tilde{\mathbb{P}}_h \text{ unter } \mathbb{P}^{\otimes n_k}_{\vartheta_0 + n_k^{-1/2}h} \text{ mit } \tilde{\mathbb{P}}_h(B) = \mathbb{E}\left[\mathbf{1}_B(S) e^{\langle h, \Delta\rangle - \frac{1}{2}\langle I(\vartheta_0)h, h\rangle}\right].$$

Setze $T(x, B) := \mathbb{P}(S \in B \,|\, \Delta = I(\vartheta_0)x) = \mathbb{E}[\mathbf{1}_B(S) \,|\, \Delta = I(\vartheta_0)x]$ (Existenz als Markovkern auf polnischen Räumen!). Beachte noch, dass $\Delta = I(\vartheta_0)X$ für ein $X \sim N(0, I(\vartheta_0)^{-1})$ gilt. Wir schließen:

$$\begin{aligned}
\tilde{\mathbb{P}}_h(B) &= \mathbb{E}\left[T(I(\vartheta_0)^{-1}\Delta, B) e^{\langle h, \Delta\rangle - \frac{1}{2}\langle I(\vartheta_0)h, h\rangle}\right] \\
&= \mathbb{E}_{\mathbb{P}_0}\left[T(X, B) \frac{dN(h, I(\vartheta_0)^{-1})}{dN(0, I(\vartheta_0)^{-1})}(X)\right] \\
&= \mathbb{E}_{\mathbb{P}_h}[T(X, B)].
\end{aligned}$$

Also ist T unter $\mathbb{P}_h = N(h, I(\vartheta_0)^{-1})$ wie $\tilde{\mathbb{P}}_h$ verteilt (aber T hängt nicht von h ab!). Es folgt $T_{n_k} \xrightarrow{d} T$ für jedes h. Da nach Voraussetzung die gesamte Folge (T_n) für jedes h in Verteilung konvergiert, gilt $T_n \xrightarrow{d} T$ für alle h. □

5.18 Korollar. *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ Hellinger-differenzierbar bei $\vartheta_0 \in \mathrm{int}(\Theta) \subseteq \mathbb{R}^k$ mit $I(\vartheta_0)$ invertierbar. Betrachte eine konvexe Verlustfunktion $l : \mathbb{R}^k \to [0, \infty)$. Sind $\hat{\vartheta}_n : \mathcal{X}^n \to \mathbb{R}^k$ Schätzer und konvergiert $\sqrt{n}(\hat{\vartheta}_n - \vartheta_{n,h})$ in Verteilung unter $\mathbb{P}^{\otimes n}_{\vartheta_{n,h}}$ mit $\vartheta_{n,h} = \vartheta_0 + n^{-1/2}h$ für alle $h \in \mathbb{R}^k$, so existiert ein Schätzer $\hat{h}$ in $(\mathbb{R}^k, \mathfrak{B}_{\mathbb{R}^k}, (\mathbb{P}_h)_{h\in\mathbb{R}^k})$ mit $\mathbb{P}_h = N(h, I(\vartheta_0)^{-1})$, so dass*

$$\forall h \in \mathbb{R}^k : \liminf_{n\to\infty} \mathbb{E}_{\vartheta_{n,h}}[l(\sqrt{n}(\hat{\vartheta}_n - \vartheta_{n,h}))] \geqslant \mathbb{E}_h[l(\hat{h} - h)].$$

Beweis. Vorab sei bemerkt, dass Werkzeuge zur Konvergenz in Verteilung (wie Slutsky-Lemma, continuous mapping, Portmanteau-Lemma) auch für randomisierte Statistiken gelten (mit analogen Beweisen).

Nach Annahme konvergieren die (von h unabhängigen) Statistiken $T_n := \sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = \sqrt{n}(\hat{\vartheta}_n - \vartheta_{n,h}) + h$ in Verteilung unter $\mathbb{P}^{\otimes n}_{\vartheta_{n,h}}$. Nach dem Satz gibt es eine randomisierte Statistik T in $(N(h, I(\vartheta_0)^{-1}))_{h\in\mathbb{R}^k}$ mit $T_n \xrightarrow{d} T$ für alle h. Damit folgt mittels Slutsky-Lemma und continuous mapping (l konvex $\Rightarrow$ l stetig)

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_{n,h}) = T_n - h \xrightarrow{d} T - h, \quad l(\sqrt{n}(\hat{\vartheta}_n - \vartheta_{n,h})) \xrightarrow{d} l(T - h).$$

Nach dem Portmanteau-Lemma folgt für alle $h \in \mathbb{R}^k$

$$\begin{aligned}
\liminf_{n\to\infty} \mathbb{E}_{\vartheta_{n,h}}[l(\sqrt{n}(\hat{\vartheta}_n - \vartheta_{n,h}))] &\geqslant \mathbb{E}_h[l(T - h)] \\
&= \int\int l(y - h)\, T(x, dy)\, N(h, I(\vartheta_0)^{-1})(dx).
\end{aligned}$$

Führe nun den Schätzer $\hat{h}(x) := \int yT(x, dy)$, die bedingte Erwartung von T, ein. Im Fall $\int |y|T(x, dy) = \infty$ setze noch $\hat{h}(x) := 0$. Dann folgt aus der Jensenschen Ungleichung $\mathbb{E}_h[l(T-h)] \geqslant \mathbb{E}_h[l(\hat{h}-h)]$ (vgl. Lemma 1.36), sofern $\mathbb{E}_h[l(T-h)] < \infty$, und $\int yT(x, dy)$ existiert in diesem Fall. Andernfalls folgt $\mathbb{E}_h[l(T-h)] = \infty > l(-h) = \mathbb{E}_h[l(\hat{h}-h)]$ aus $\hat{h} = 0$. Dies zeigt die Behauptung. □

5.19 Korollar (Lokal-asymptotisches Minimax-Theorem, Hájek 1972)**.** *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ Hellinger-differenzierbar bei $\vartheta_0 \in int(\Theta) \subseteq \mathbb{R}^k$ mit $I(\vartheta_0)$ invertierbar. Für beliebige Schätzer $\hat{\vartheta}_n : \mathcal{X}^n \to \mathbb{R}^k$ im Produktmodell $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ gilt mit $\vartheta_{n,h} = \vartheta_0 + n^{-1/2}h$:*

$$\sup_{h\in\mathbb{R}^k} \limsup_{n\to\infty} \mathbb{E}_{\vartheta_{n,h}}[n|\hat{\vartheta}_n - \vartheta_{n,h}|^2] \geqslant \inf_{\hat{h}} \sup_{h\in\mathbb{R}^k} \mathbb{E}_h[|\hat{h}-h|^2] = \operatorname{trace}(I(\vartheta_0)^{-1}),$$

wobei sich das Infimum über Schätzer $\hat{h}$ im Modell $(N(h, I(\vartheta_0)^{-1}))_{h\in\mathbb{R}^k}$ erstreckt.

Beweis. Wir führen

$$\sup_{h\in\mathbb{R}^k} \limsup_{n\to\infty} \mathbb{E}_{\vartheta_{n,h}}[n|\hat{\vartheta}_n - \vartheta_{n,h}|^2] < \operatorname{trace}(I(\vartheta_0)^{-1})$$

zum Widerspruch. Mit $h = 0$ folgt aus der Widerspruchs-Annahme gerade $\limsup_{n\to\infty} \mathbb{E}_{\vartheta_0}[|\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)|^2] < \infty$. Also ist $(\sqrt{n}(\hat{\vartheta}_n - \vartheta_0))_n$ L^2-beschränkt und damit straff unter $(\mathbb{P}_{\vartheta_0}^{\otimes n})$. Nach dem Satz von Prohorov und der LAN-Entwicklung gibt es also eine Teilfolge (n_k), so dass $(\sqrt{n_k}(\hat{\vartheta}_{n_k} - \vartheta_0), \prod_{i=1}^{n_k} \frac{L(\vartheta_{n_k,h}, X_i)}{L(\vartheta_0, X_i)})$ in Verteilung unter $\mathbb{P}_{\vartheta_0}^{\otimes n_k}$ konvergiert. Dabei impliziert die LAN-Entwicklung $\prod_{i=1}^{n_k} \frac{L(\vartheta_{n_k,h}, X_i)}{L(\vartheta_0, X_i)} \xrightarrow{d} Z$ mit $\mathbb{E}[Z] = 1$.

Mit Le Cams Drittem Lemma erhalten wir daher die Konvergenz von $\sqrt{n_k}(\hat{\vartheta}_{n_k} - \vartheta_0)$ in Verteilung unter $\mathbb{P}_{\vartheta_{n_k,h}}^{\otimes n_k}$ für jedes h. Damit konvergiert auch $\sqrt{n_k}(\hat{\vartheta}_{n_k} - \vartheta_{n_k,h})$, und wir erhalten aus Korollar 5.18 mit $l(x-y) = |x-y|^2$

$$\liminf_{n_k\to\infty} \mathbb{E}_{\vartheta_0 + n_k^{-1/2}h}[n_k|\hat{\vartheta}_{n_k} - \vartheta_{n_k,h}|^2] \geqslant \mathbb{E}_h[|\hat{h}-h|^2]$$

für einen Schätzer $\hat{h}$ und alle $h \in \mathbb{R}^k$.

Nun gilt in Verallgemeinerung von Satz 1.22 (und mit analogem Beweis für $N(h, I(\vartheta_0)^{-1})$ anstatt $N(h, E_k)$) im normalverteilten Grenzmodell

$$\inf_{\hat{h}} \sup_{h\in\mathbb{R}^k} \mathbb{E}_h[|\hat{h}-h|^2] = \sup_{h\in\mathbb{R}^k} \mathbb{E}_h[|X-h|^2] = \operatorname{trace}(I(\vartheta_0)^{-1}),$$

d.h. X ist minimax-Schätzer. Wir schließen

$$\begin{aligned}\sup_{h\in\mathbb{R}^k} \limsup_{n\to\infty} \mathbb{E}_{\vartheta_{n,h}}[n|\hat{\vartheta}_n - \vartheta_{n,h}|^2] &\geqslant \sup_{h\in\mathbb{R}^k} \liminf_{n_k\to\infty} \mathbb{E}_{\vartheta_{n_k,h}}[n_k|\hat{\vartheta}_{n_k} - \vartheta_{n_k,h}|^2] \\ &\geqslant \operatorname{trace}(I(\vartheta_0)^{-1}),\end{aligned}$$

was der Annahme widerspricht. □

5.20 Beispiel. Der Hodges-Schätzer für eine mathematische Stichprobe $X_1, \dots, X_n \sim N(\vartheta, 1)$ ist definiert als

$$\hat{\vartheta}_n = \begin{cases} \bar{X}, & \text{falls } |\bar{X}| > n^{-1/4} \\ 0, & \text{falls } |\bar{X}| \leqslant n^{-1/4} \end{cases}$$

und erfüllt $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \xrightarrow{d} N(0,1)$ für alle $\vartheta \neq 0$, während $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \to 0$ in Wahrscheinlichkeit für $\vartheta = 0$ gilt (Übung!). Da die Fisher-Information bei n Beobachtungen $I_n(\vartheta) = n$ für alle $\vartheta \in \mathbb{R}$ erfüllt, sagt man, dass der Hodges-Schätzer bei $\vartheta = 0$ super-effizient ist.

Allerdings gilt $\sqrt{n}|\hat{\vartheta}_n - \vartheta_{n,h}| \geqslant |h|\mathbf{1}(|\bar{X}| \leqslant n^{-1/4})$ für lokale Parameter $\vartheta_{n,h} = n^{-1/2}h$. Aus (verwende Tschebyschew-Ungleichung)

$$\mathbb{P}_{\vartheta_{n,h}}(|\bar{X}| > n^{-1/4}) \leqslant \mathbb{P}_{\vartheta_{n,h}}(|\bar{X} - \vartheta_{n,h}| > n^{-1/4} - |\vartheta_{n,h}|) \leqslant \frac{\operatorname{Var}(\bar{X})}{(n^{-1/4} - |\vartheta_{n,h}|)^2} \to 0$$

folgt die Fehlerabschätzung $\liminf_{n\to\infty} \mathbb{E}_{\vartheta_{n,h}}[n(\hat{\vartheta}_n - \vartheta_{n,h})^2] \geqslant h^2$ und somit $\sup_{h\in\mathbb{R}} \liminf_{n\to\infty} \mathbb{E}_{\vartheta_{n,h}}[n(\hat{\vartheta}_n - \vartheta_{n,h})^2] = \infty$. Insbesondere ist der Hodges-Schätzer nicht lokal-asymptotisch minimax.

5.21 Bemerkungen. (Mehr dazu im Buch von van der Vaart)

(a) Eine Version des lokalen Minimax-Theorems gilt auch mit 'lim inf' statt 'lim sup', falls danach noch ein Supremum in h über endliche Teilmengen von Θ eingefügt wird. Man kann auch Verlustfunktionen zulassen, die sehr viel allgemeiner als der quadratische Verlust sind ("bowl-shaped"). Auch wird das Resultat oft für abgeleitete Parameter $g(\vartheta)$ formuliert. Ein weiteres wichtiges asymptotisches Effizienzresultat ist der sogenannte Hajek-Le Cam-Faltungssatz, der auch die Optimalität der Normalverteilung begründet. Unter Regularitätsbedingungen erfüllt der MLE diese asymptotischen Effizienzkriterien.

(b) Die asymptotische Effizienz des MLE ist keineswegs ein Alleinstellungsmerkmal, auch Bayes-Schätzer erfüllen diese unter ähnlichen Voraussetzungen. Hauptergebnis der Asymptotik des Bayesansatzes ist allerdings:

5.22 Satz (Bernstein-von Mises)**.** *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ Hellinger-differenzierbar bei $\vartheta_0 \in \mathrm{int}(\Theta) \subseteq \mathbb{R}^k$ mit Fisher-Information $I(\vartheta_0) > 0$ und es gebe asymptotisch konsistente Tests für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : |\vartheta - \vartheta_0| \geqslant \varepsilon$ für alle $\varepsilon > 0$. Besitzt die a-priori-Verteilung π eine positive und stetige Dichte bei ϑ_0, so gilt für den Totalvariationsabstand zwischen a-posteriori-Verteilung $\tilde{\mathbb{P}}(\bullet \,|\, X_1, \dots, X_n)$ und einer durch Score und Fisher-Information bestimmten Normalverteilung*

$$\Big\| \tilde{\mathbb{P}}(\bullet \,|\, X_1, \dots, X_n) - N\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^n I(\vartheta_0)^{-1} \dot{\ell}(\vartheta_0, X_i), I(\vartheta_0)^{-1}\Big) \Big\|_{TV} \xrightarrow{\mathbb{P}_{\vartheta_0}^{\otimes n}} 0.$$

Mit der Entwicklung des MLE $\hat{\vartheta}_n$ aus Satz 3.27 kann man dann schließen, dass – unter Regularitätsbedingungen – die a-posteriori-Verteilung

durch $N(\hat{\vartheta}_n, I(\vartheta_0)^{-1})$ unter $\mathbb{P}_{\vartheta_0}^{\otimes n}$ asymptotisch approximiert wird. Damit lässt sich überprüfen, wann Bayessche Kredibilitätsbereiche frequentistische asymptotische Konfidenzbereiche bilden.