Mathematische Statistik
Gliederung zur Vorlesung
im Sommersemester 2015

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

VORLÄUFIGE FASSUNG: 15. Juli 2015

Inhaltsverzeichnis

Einführende Beispiele

- Modellierung
- Modelldiagnostik (QQ-Plot, Boxplot)
- Median, Mittelwert, Ausreißer
- Hypothesentest, Konfidenzintervall

1 Statistik im linearen Modell

1.1 Lineares Modell und kleinste Quadrate

1.1 Beispiel (lineare Regression)**.** Wir beobachten Realisierungen von

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

wobei $a, b \in \mathbb{R}, \sigma > 0$ unbekannte Parameter, (x_i) bekannte Werte (Versuchsplan, Design) sowie (ε_i) zentrierte Zufallsvariablen (d.h. $\mathbb{E}[\varepsilon_i] = 0$) sind mit $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$, die Messfehler modellieren. Man denke z.B. an Messungen der Leitfähigkeit Y_i eines Stoffes in Abhängigkeit der Temperatur x_i.

Gesucht ist eine Regressionsgerade der Form $y = ax+b$, die die Beobachtungen möglichst gut erklärt. Nach der Methode der kleinsten Quadrate erhalten wir Schätzer $\hat{a}$, $\hat{b}$ durch Minimierung der Summe der quadratischen Abstände:

$$(\hat{a}, \hat{b}) := \mathrm{argmin}_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (Y_i - ax_i - b)^2.$$

Differentiation ergibt, dass $\hat{a}$, $\hat{b}$ Lösungen der Normalengleichungen sind:

$$\sum_{i=1}^{n} (Y_i - ax_i - b) = 0 \text{ und } \sum_{i=1}^{n} x_i(Y_i - ax_i - b) = 0.$$

Explizit gilt $\hat{a} = \bar{c}_{xY}/\bar{\sigma}_x^2$, $\hat{b} = \bar{Y} - \hat{a}\bar{x}$ mit $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $\bar{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$, $\bar{c}_{xY} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$.

1.2 Definition. Ein lineares Modell mit n reellwertigen Beobachtungen $Y = (Y_1, \dots, Y_n)^\top$ und p-dimensionalem Parameter $\beta \in \mathbb{R}^p$, $p \leqslant n$, besteht aus einer reellen Matrix $X \in \mathbb{R}^{n\times p}$ von vollem Rang p, der Designmatrix, und einem Zufallsvektor $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$, den Fehler- oder Störgrößen, mit $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Cov}(\varepsilon_i\varepsilon_j) = \Sigma_{ij}$ zur Kovarianzmatrix $\Sigma > 0$. Beobachtet wird eine Realisierung von

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

Der (gewichtete) Kleinste-Quadrate-Schätzer $\hat{\beta}$ von β minimiert den gewichteten Euklidischen Abstand zwischen Beobachtungen und Modellvorhersage:

$$|\Sigma^{-1/2}(X\hat{\beta} - Y)|^2 = \inf_{b\in\mathbb{R}^p} |\Sigma^{-1/2}(Xb - Y)|^2.$$

Im gewöhnlichen Fall $\Sigma = \sigma^2 E_n$ ($E_n \in \mathbb{R}^{n\times n}$: Einheitsmatrix) mit Fehlerniveau $\sigma > 0$, erhalten wir den gewöhnlichen Kleinste-Quadrate-Schätzer (OLS: ordinary least squares), der unabhängig von der Kenntnis von σ^2 ist:

$$|X\hat{\beta} - Y|^2 = \inf_{b\in\mathbb{R}^p} |Xb - Y|^2.$$

1.3 Bemerkung. Wir schreiben $\Sigma > 0$, falls Σ eine symmetrische, strikt positiv-definite Matrix ist. Dann ist Σ diagonalisierbar mit $\Sigma = TDT^\top$, $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ Diagonalmatrix und T orthogonale Matrix, und wir setzen $\Sigma^{-1/2} = TD^{-1/2}T^\top$ mit $D^{-1/2} = \operatorname{diag}(\lambda_1^{-1/2}, \dots, \lambda_n^{-1/2})$. Wie erwartet, gilt $(\Sigma^{-1/2})^2 = \Sigma^{-1}$ und somit $|\Sigma^{-1/2}v|^2 = \langle \Sigma^{-1}v, v\rangle$.

1.4 Beispiele.

(a) Einfaches Shift-Modell: Wir beobachten $Y_i = \mu + \varepsilon_i$, $i = 1, \dots, n$, mit $\mu \in \mathbb{R}$ unbekannt, was auf ein lineares Modell mit $p = 1$, $\beta = \mu$ und $X = (1, \dots, 1)^\top$ führt.

(b) Lineare Regression: $p = 2$, $\beta = (b, a)^\top$, $X = (X_{ij})$ mit $X_{i,1} = 1$, $X_{i,2} = x_i$. Damit X Rang 2 hat, müssen mindestens zwei der (x_i) verschieden sein.

(c) Polynomiale Regression: wir beobachten

$$Y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_{p-1} x_i^{p-1} + \varepsilon_i, \quad i = 1, \dots, n.$$

Damit ergibt sich als Parameter $\beta = (a_0, a_1, \dots, a_{p-1})^\top$ und eine Designmatrix vom Vandermonde-Typ:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix}.$$

Die Matrix X hat vollen Rang, sofern p der Designpunkte (x_i) verschieden sind.

(d) Mehrfache/multiple lineare Regression: bei d-dimensionalem Design mit Punkten $x_i = (x_{i,1}, \dots, x_{i,d})$ beobachtet man

$$Y_i = a_0 + \langle a, x_i\rangle + \varepsilon_i, \quad a = (a_1, \dots, a_d)^\top,\ i = 1, \dots, n.$$

Wir erhalten $p = d + 1$, $\beta = (a_0, a_1, \dots, a_d)^\top$ sowie

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,d} \end{pmatrix}.$$

Die Forderung, dass X vollen Rang besitzt, ist gleichbedeutend damit, dass die Punkte $\begin{pmatrix} 1 \\ x_i \end{pmatrix}$, $i = 1, \dots, n$, den gesamten Raum $\mathbb{R}^{d+1}$ aufspannen.

1.5 Bemerkung. Es gibt wichtige Verallgemeinerungen linearer Modelle (GLM: Generalized Linear Models), die auf exponentiellen Familien beruhen. Als Beispiel sei die logistische Regression genannt, wo Binomial-verteilte $Y_i \sim Bin(n_i, p_i)$ beobachtet werden mit der sogenannten *logit*-Linkfunktion $\log(p_i/(1-p_i))_{i=1,\dots,n} = X\beta$, so dass

$$Y_i/n_i = p_i + \varepsilon_i = \frac{1}{1+\exp(-(X\beta)_i)} + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0.$$

1.2 Der Satz von Gauß-Markov

1.6 Lemma. *Setze $X_\Sigma := \Sigma^{-1/2}X$. Mit Π_{X_Σ} werde die Orthogonalprojektion von $\mathbb{R}^n$ auf den Bildraum $\mathrm{ran}(X_\Sigma)$ bezeichnet. Dann gilt $\Pi_{X_\Sigma} = X_\Sigma(X_\Sigma^\top X_\Sigma)^{-1}X_\Sigma^\top$ und für den Kleinste-Quadrate-Schätzer $\hat\beta = (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}Y$. Insbesondere existiert der Kleinste-Quadrate-Schätzer und ist eindeutig.*

1.7 Bemerkungen.

(a) Im gewöhnlichen linearen Modell gilt $\hat\beta = (X^\top X)^{-1}X^\top Y$, da sich $\sigma > 0$ herauskürzt.

(b) $X_\Sigma^+ := (X_\Sigma^\top X_\Sigma)^{-1}X_\Sigma^\top$ heißt auch Moore-Penrose-Inverse von X_Σ, so dass $\hat\beta = X_\Sigma^+\Sigma^{-1/2}Y$ bzw. $\hat\beta = X^+Y$ im gewöhnlichen linearen Modell gilt.

Beweis. Zunächst beachte, dass $X_\Sigma^\top X_\Sigma = X^\top\Sigma^{-1}X$ invertierbar ist wegen der Invertierbarkeit von Σ und der Rangbedingung an X:

$$X^\top\Sigma^{-1}Xv = 0 \Rightarrow v^\top X^\top\Sigma^{-1}Xv = 0 \Rightarrow |\Sigma^{-1/2}Xv| = 0 \Rightarrow |Xv| = 0 \Rightarrow v = 0.$$

Setze $P_{X_\Sigma} := X_\Sigma(X_\Sigma^\top X_\Sigma)^{-1}X_\Sigma^\top$ und $w = P_{X_\Sigma}v$ für ein $v \in \mathbb{R}^n$. Dann folgt $w \in \mathrm{ran}(X_\Sigma)$ und im Fall $v = X_\Sigma u$ durch Einsetzen $w = P_{X_\Sigma}X_\Sigma u = v$, so dass P_{X_Σ} eine Projektion auf $\mathrm{ran}(X_\Sigma)$ ist. Da P_{X_Σ} selbstadjungiert (symmetrisch) ist, handelt es sich um die Orthogonalprojektion Π_{X_Σ}:

$$\forall u \in \mathbb{R}^n : \langle u - P_{X_\Sigma}u, w\rangle = \langle u, w\rangle - \langle u, P_{X_\Sigma}w\rangle = 0.$$

Aus der Eigenschaft $\hat\beta = \mathrm{argmin}_b|\Sigma^{-1/2}(Y - Xb)|^2$ folgt, dass $\hat\beta$ die beste Approximation im $\mathbb{R}^n$ von $\Sigma^{-1/2}Y$ durch $X_\Sigma b$ liefert. Diese ist durch die Orthogonalprojektionseigenschaft $\Pi_{X_\Sigma}\Sigma^{-1/2}Y = X_\Sigma\hat\beta$ bestimmt. Es folgt

$$X_\Sigma^\top\Pi_{X_\Sigma}\Sigma^{-1/2}Y = (X_\Sigma^\top X_\Sigma)\hat\beta \Rightarrow (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}Y = \hat\beta.$$

□

1.8 Satz. *Im gewöhnlichen linearen Modell mit Fehlerniveau $\sigma > 0$ gelten die folgenden Aussagen:*

(a) Der Kleinste-Quadrate-Schätzer $\hat\beta = (X^\top X)^{-1}X^\top Y$ ist erwartungstreuer Schätzer von β (d.h. $\mathbb{E}[\hat\beta] = \beta$).

(b) *Satz von Gauß-Markov: ist der reelle Parameter $\gamma = \langle \beta, v \rangle$ für ein $v \in \mathbb{R}^p$ zu schätzen, so ist $\hat{\gamma} = \langle \hat{\beta}, v \rangle$ ein (in den Daten Y) linearer erwartungstreuer Schätzer, der unter allen linearen erwartungstreuen Schätzern minimale Varianz besitzt, nämlich* $\mathrm{Var}(\hat{\gamma}) = \sigma^2 |X(X^\top X)^{-1} v|^2$.

(c) *Bezeichnet $R := Y - X\hat{\beta}$ den Vektor der Residuen, so ist die geeignet normalisierte Stichprobenvarianz*

$$\hat{\sigma}^2 := \frac{|R|^2}{n-p} = \frac{|Y - X\hat{\beta}|^2}{n-p}$$

ein erwartungstreuer Schätzer von σ^2.

Beweis.

(a) Aus der Linearität des Erwartungswerts und $\mathbb{E}[\varepsilon] = 0$ folgt

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(X^\top X)^{-1} X^\top (X\beta + \varepsilon)] = \beta + 0 = \beta.$$

(b) Aus (a) folgt, dass $\hat{\gamma}$ linear und erwartungstreu ist. Sei nun $\tilde{\gamma} = \langle Y, w \rangle$ ein beliebiger linearer erwartungstreuer Schätzer von γ. Dies impliziert für alle $\beta \in \mathbb{R}^p$

$$\mathbb{E}[\langle Y, w \rangle] = \gamma \Rightarrow \langle X\beta, w \rangle = \langle \beta, v \rangle \Rightarrow \langle X^\top w - v, \beta \rangle = 0 \Rightarrow X^\top w = v.$$

Nach Pythagoras erhalten wir

$$\mathrm{Var}(\tilde{\gamma}) = \mathbb{E}[\langle \varepsilon, w \rangle^2] = \sigma^2 |w|^2 = \sigma^2 (|\Pi_X w|^2 + |(E_n - \Pi_X) w|^2)$$

und somit $\mathrm{Var}(\tilde{\gamma}) \geqslant \sigma^2 |\Pi_X w|^2 = \sigma^2 |X(X^\top X)^{-1} v|^2 = \mathrm{Var}(\hat{\gamma})$.

(c) Einsetzen zeigt $\mathbb{E}[|Y - X\hat{\beta}|^2] = \mathbb{E}[|Y - \Pi_X Y|^2] = \mathbb{E}[|(E_n - \Pi_X)\varepsilon|^2]$. Ist nun $e_1, \ldots, e_{n-p}$ eine Orthonormalbasis vom $(n-p)$-dimensionalen Bild $\mathrm{ran}(E_n - \Pi_X) \subseteq \mathbb{R}^n$, so folgt

$$\mathbb{E}[|(E_n - \Pi_X)\varepsilon|^2] = \sum_{i=1}^{n-p} \mathbb{E}[\langle \varepsilon, e_i \rangle^2] = \sigma^2 (n-p),$$

was die Behauptung impliziert.

□

1.9 Bemerkungen.

(a) Man sagt, dass der Schätzer $\hat{\gamma}$ im Satz von Gauß-Markov bester linearer erwartungstreuer Schätzer (BLUE: best linear unbiased estimator) ist. Verzichtet man auf die Linearität oder Erwartungstreue des Schätzers, so gibt es im Allgemeinen bessere Schätzer bezüglich quadratischem Fehler (MSE: mean squared error) $\mathbb{E}[(\hat{\gamma} - \gamma)^2]$, zumindest für ausgewählte wahre Parameter β bzw. γ. Ein einfaches nicht-erwartungstreues Beispiel ist $\tilde{\gamma} = 0$. Dies ist ein linearer Schätzer mit MSE γ^2, was für γ in einer Nullumgebung kleiner ist als beim Kleinste-Quadrate-Schätzer.

(b) Oft interessiert auch der Vorhersagefehler $|X\hat{\beta} - X\beta|^2$ (d.h. bei der Regression die quadrierte Differenz der vorhergesagten und wahren Werte an den Designpunkten), so prüft man leicht nach:

$$\mathbb{E}[|X\hat{\beta} - X\beta|^2] = \mathbb{E}[|\Pi_X \varepsilon|^2] = \sigma^2 p.$$

Insbesondere wächst dieser Fehler linear in der Dimension p des Parameterraums.

(c) Eine entsprechende Aussage des Satzes von Gauß-Markov gilt auch im allgemeinen linearen Modell (Übung!).

1.3 Inferenz unter Normalverteilungsannahme

1.10 Beispiel. Sind die Messfehler $(\varepsilon_i) \sim N(0, \sigma^2 E_n)$ gemeinsam normalverteilt, so gilt $\hat{\beta} \sim N(\beta, \sigma^2 (X^\top X)^{-1})$ und $\hat{\gamma} \sim N(\gamma, \sigma^2 v^\top (X^\top X)^{-1} v)$. Ist weiterhin $\sigma > 0$ bekannt, so ist ein Konfidenzintervall zum Niveau 95% für γ gegeben durch

$$I_{0,95}(\gamma) := \Big[\hat{\gamma} - 1,96\sigma\sqrt{v^\top (X^\top X)^{-1} v},\ \hat{\gamma} + 1,96\sigma\sqrt{v^\top (X^\top X)^{-1} v}\Big].$$

Dabei ist der Wert $q_{0,975} = 1,96$ gerade das $0,975$-Quantil bzw. $0,025$-Fraktil der Standardnormalverteilung, d.h. $\Phi(1,96) \approx 0,975$. Analog wird der zweiseitige Gauß-Test der Hypothese $H_0 : \gamma = \gamma_0$ gegen $H_1 : \gamma \neq \gamma_0$ zum Niveau $\alpha \in (0,1)$ konstruiert: H_0 wird akzeptiert, falls $|\hat{\gamma} - \gamma_0| \leqslant q_{1-\alpha/2}\sigma\sqrt{v^\top (X^\top X)^{-1} v}$ gilt mit dem $(1-\alpha/2)$-Quantil $q_{1-\alpha/2}$ von $N(0,1)$, sonst verworfen.

Falls σ unbekannt ist, so ist eine Idee, einfach σ durch einen Schätzer $\hat{\sigma}$ in obigen Formeln zu ersetzen. Allerdings wird dann das vorgegebene Niveau nur noch asymptotisch erreicht für einen konsistenten Schätzer (Slutsky-Lemma!). Im vorliegenden Fall können wir aber sogar die nicht-asymptotische Verteilung exakt bestimmen.

1.11 Definition. Die t-Verteilung (oder Student-t-Verteilung) mit $n \in \mathbb{N}$ Freiheitsgraden auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ ist gegeben durch die Lebesguedichte

$$t_n(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad x \in \mathbb{R}\,.$$

Die F-Verteilung (oder Fisher-Verteilung) mit $(m,n) \in \mathbb{N}^2$ Freiheitsgraden auf $(\mathbb{R}, \mathfrak{B}_\mathbb{R})$ ist gegeben durch die Lebesguedichte

$$f_{m,n}(x) = \frac{m^{m/2} n^{n/2}}{B(m/2, n/2)} \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}} \mathbf{1}_{\mathbb{R}^+}(x), \quad x \in \mathbb{R}\,.$$

Dabei bezeichnet $\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt$ die Gamma-Funktion sowie $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ die Beta-Funktion.

1.12 Lemma. *Es seien $X_1, \dots, X_m, Y_1, \dots, Y_n$ unabhängige $N(0,1)$-verteilte Zufallsvariablen. Dann ist*

$$T_n := \frac{X_1}{\sqrt{\frac{1}{n}\sum_{j=1}^n Y_j^2}}$$

gemäß einer t-Verteilung mit n Freiheitsgraden sowie

$$F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^m X_i^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2}$$

gemäß einer F-Verteilung mit (m,n) Freiheitsgraden verteilt.

Beweis. Beachte zunächst, dass $T_n^2 = F_{1,n}$ gilt, so dass mittels Dichtetransformation $f_{|T_n|}(x) = f_{F_{1,n}}(x^2)2x$, $x \geqslant 0$, gilt. Da T_n symmetrisch (wie $-T_n$) verteilt ist, folgt $f_{T_n}(x) = f_{F_{1,n}}(x^2)|x|$, $x \in \mathbb{R}$, und Einsetzen zeigt die Behauptung für T_n, sofern $F_{1,n}$ $F(1,n)$-verteilt ist.

Dafür benutze, dass $X := \sum_{i=1}^m X_i^2$ $\chi^2(m)$-verteilt und $Y := \sum_{j=1}^n Y_j^2$ $\chi^2(n)$-verteilt sind. Wegen Unabhängigkeit von X und Y gilt für $z > 0$ (setze $w = x/y$)

$$\begin{aligned}\mathbb{P}(X/Y \leqslant z) &= \int\int \mathbf{1}(x/y \leqslant z) f_X(x) f_Y(y)\, dy dx \\ &= \int \mathbf{1}(w \leqslant z)\Big(\int f_X(wy) f_Y(y) y\, dy\Big) dw,\end{aligned}$$

so dass sich die Dichte wie folgt ergibt (setze $w = (x+1)y$)

$$\begin{aligned}f_{X/Y}(x) &= \int f_X(xy) f_Y(y) y\, dy \\ &= \frac{2^{-(m+n)/2}}{\Gamma(m/2)\Gamma(n/2)} \int_0^\infty (xy)^{m/2-1} y^{n/2} e^{-(xy+y)/2} dy \\ &= \frac{2^{-(m+n)/2}}{\Gamma(m/2)\Gamma(n/2)} \int_0^\infty (xw/(x+1))^{m/2-1} (w/(x+1))^{n/2} e^{-w/2} (x+1)^{-1} dw \\ &= \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} x^{m/2-1} (x+1)^{-(m+n)/2}, \quad x > 0.\end{aligned}$$

Dichtetransformation ergibt damit für $F_{m,n} = \frac{n}{m}(X/Y)$ die Dichte $\frac{m}{n} f_{X/Y}(\frac{m}{n}x) = f_{m,n}(x)$. □

1.13 Bemerkung. Für $n = 1$ ist die $t(n)$-Verteilung gerade die Cauchy-Verteilung und für $n \to \infty$ konvergiert sie schwach gegen die Standardnormalverteilung (Slutsky-Lemma!). Für jedes $n \in \mathbb{N}$ besitzt $t(n)$ endliche Momente nur bis zur Ordnung $p < n$ (sie ist *heavy-tailed*). Ähnliches gilt für die F-Verteilung, insbesondere konvergiert $mF(m,n)$ für $F(m,n)$-verteilte Zufallsvariablen $F(m,n)$ und $n \to \infty$ gegen die $\chi^2(m)$-Verteilung.

1.14 Satz. *Im gewöhnlichen linearen Modell unter Normalverteilungsannahme $\varepsilon_i \sim N(0,\sigma^2)$ gelten folgende Konfidenzaussagen für gegebenes $\alpha \in (0,1)$ (Notation wie in Satz 1.8):*

(a) *Konfidenzbereich für β: Ist $q_{F(p,n-p);1-\alpha}$ das $(1-\alpha)$-Quantil der $F(p, n-p)$-Verteilung, so ist*

$$C := \{\beta \in \mathbb{R}^p \mid |X(\beta - \hat{\beta})|^2 < pq_{F(p,n-p);1-\alpha}\hat{\sigma}^2\}$$

ein Konfidenzellipsoid zum Niveau $1-\alpha$ für β.

(b) *Konfidenzbereich für $\gamma = \langle \beta, v\rangle$: Ist $q_{t(n-p);1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der $t(n-p)$-Verteilung, so ist*

$$I := \left[\hat{\gamma} - \hat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}\, q_{t(n-p);1-\alpha/2}, \hat{\gamma} + \hat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}\, q_{t(n-p);1-\alpha/2}\right]$$

ein Konfidenzintervall zum Niveau $1-\alpha$ für γ.

1.15 Korollar. *Im Shiftmodell $Y_i = \mu + \varepsilon_i$, $i = 1, \dots, n$, mit $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d. und $\mu \in \mathbb{R}$, $\sigma > 0$ unbekannt ist*

$$I := [\hat{\mu} - \hat{\sigma} n^{-1/2} q_{t(n-1);1-\alpha/2}, \hat{\mu} + \hat{\sigma} n^{-1/2} q_{t(n-1);1-\alpha/2}]$$

mit $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i$, $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \hat{\mu})^2$ ein Konfidenzintervall zum Irrtumsniveau α für μ.

Beweis. Dies folgt direkt aus Teil (b) des vorigen Satzes mit dem linearen Modell, wo $p = 1$, $X = (1, \dots, 1)^\top$ und $\gamma = \beta$, $v = 1$ ist. □

Beweis des Satzes. Allgemein müssen wir jeweils für einen Konfidenzbereich B für den vom wahren Parameter β abgeleiteten Parameter ϑ_β zum Niveau $1-\alpha$ nachweisen, dass gilt

$$\forall \beta \in \mathbb{R}^p : \mathbb{P}_\beta(\vartheta_\beta \in B) \geqslant 1 - \alpha.$$

Im folgenden werden wir sogar Gleichheit erhalten.

(a) $X(\hat{\beta} - \beta) = \Pi_X \varepsilon$ ist $N(0, \sigma^2 \Pi_X \Pi_X^\top)$-verteilt und somit ist $\sigma^{-2}|X(\hat{\beta} - \beta)|^2$ $\chi^2(p)$-verteilt. Weiterhin gilt ja $\hat{\sigma}^2 = \frac{|Y - \Pi_X Y|^2}{n-p} = \frac{|\varepsilon - \Pi_X \varepsilon|^2}{n-p}$, so dass $X(\hat{\beta} - \beta)$ und $\hat{\sigma}^2$ unabhängig sind, weil $\Pi_X \varepsilon$ und $(E_n - \Pi_x)\varepsilon$ unabhängig sind (da unkorreliert und gemeinsam normalverteilt). Außerdem folgt, dass $\frac{n-p}{\sigma^2}\hat{\sigma}^2$ $\chi^2(n-p)$-verteilt ist. Wie in Lemma 1.12 schließen wir, dass $|X(\hat{\beta} - \beta)|^2/(p\hat{\sigma}^2)$ $F(p, n-p)$-verteilt ist. Damit ist C per Konstruktion ein entsprechender Konfidenzbereich.

(b) Wie in (a) sind $\hat{\gamma}$ und $\hat{\sigma}$ unabhängig. Außerdem gilt $\hat{\gamma} - \gamma \sim N(0, \sigma^2 v^\top (X^\top X)^{-1} v)$, so dass $\frac{\hat{\gamma} - \gamma}{\hat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}}$ wie in Lemma 1.12 $t(n-p)$-verteilt ist und die Behauptung folgt.

□

1.16 Korollar. *Im Beobachtungsmodell $Y_i = \mu + \varepsilon_i$, $i = 1, \dots, n$, mit $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d. und $\mu \in \mathbb{R}$, $\sigma > 0$ unbekannt kann die Hypothese $H_0 : \mu = \mu_0$ gegen die Alternative $\mu \neq \mu_0$ mit dem zweiseitigen t-Test zum Niveau α getestet werden: Falls $|\hat{\mu} - \mu_0| > \hat{\sigma} n^{-1/2} q_{t(n-1);1-\alpha/2}$ gilt, lehne die Hypothese H_0 ab, sonst akzeptiere sie.*

Beweis. Dies folgt aus der Aussage für das Konfidenzintervall I, weil diese insbesondere $\mathbb{P}_{\mu_0}(\mu_0 \notin I) \leqslant \alpha$ impliziert und $\mu_0 \notin I \iff |\hat{\mu} - \mu_0| > \hat{\sigma} n^{-1/2} q_{t(n-1);1-\alpha/2}$ gilt. □

2 Entscheidungstheorie

2.1 Formalisierung eines statistischen Problems

2.1 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.

2.2 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

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

2.3 Beispiele.

(a) Beim gewöhnlichen linearen Modell wähle als Parameterraum $\Theta = \mathbb{R}^p \times \mathbb{R}_+$ mit Parametern $\vartheta = (\beta, \sigma) \in \Theta$. Nun wähle einen Wahrscheinlichkeitsraum $(\Omega, \mathscr{G}, \mathbb{P})$, auf dem der Zufallsvektor $\varepsilon : \Omega \to \mathbb{R}^n$ mit $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[\varepsilon_i \varepsilon_j] = \delta_{i,j}$ definiert ist. Versieht man den Stichprobenraum $\mathcal{X} = \mathbb{R}^n$ mit seiner Borel-σ-Algebra $\mathscr{F} = \mathscr{B}_{\mathbb{R}^n}$ und setzt $Y_\vartheta = Y_{\beta,\sigma} = X\beta + \sigma\varepsilon$, so bilden die Verteilungen $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ der Zufallsvariablen $(Y_\vartheta)_{\vartheta\in\Theta}$ die Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum.

Um den Kleinste-Quadrate-Schätzer $\hat{\beta}$ als Entscheidungsregel zu interpretieren und seine Güte messen, betrachtet man den Aktionsraum $A = \mathbb{R}^p$ und beispielsweise die quadratische Verlustfunktion $l(\vartheta, a) = l((\beta, \sigma), a) = |\beta - a|^2$. Beim Verlust ist σ irrelevant; da aber die Verteilung $\mathbb{P}_\vartheta$ von σ abhängt, spricht man von einem Störparameter.

Beachte, dass bei obiger Modellierung eine feste Verteilung von ε (z.B. Normalverteilung) angenommen wird. Ist realistischerweise auch die Art der Verteilung unbekannt, sollte man statt $(\mathbb{P}_\vartheta)$ die Familie $\mathcal{P} = \{\mathbb{P} \text{ W-Maß auf } \mathscr{F} \mid \mathbb{E}_\mathbb{P}[Y] := \int y \, \mathbb{P}(dy) \in \operatorname{ran}(X), \int (y - \mathbb{E}_\mathbb{P}[Y])(y - \mathbb{E}_\mathbb{P}[Y])^\top \, \mathbb{P}(dy) = \sigma^2 E_n\}$ betrachten. In dieser Betrachtungsweise bleibt von einem unendlich-dimensionalen Parameterraum maximal ein $(p+1)$-dimensionaler interessierender Parameter ϑ übrig (beachte $\beta = X^{-1}(\mathbb{E}_\mathbb{P}[Y])$).

(b) 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 $\mathcal{X} = \{0, 1, \ldots, 100\}$ 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

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

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

2.4 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.

2.5 Bemerkung. Häufig wird für diese Definition die Menge der betrachteten Entscheidungsregeln eingeschränkt. So ist der Kleinste-Quadrate-Schätzer im linearen Modell nach dem Satz 1.8 von Gauß-Markov zulässig unter quadratischem Verlust in der Klasse der erwartungstreuen und linearen Schätzern.

2.6 Beispiel. Es sei $Y_1, \ldots, Y_n$ 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.

2.2 Minimax- und Bayes-Ansatz

2.7 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.

2.8 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

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

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.

2.9 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 $A \in \mathscr{F}$.

2.10 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. □

2.11 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.

2.12 Definition. Definiere $\Omega := \mathcal{X} \times \Theta$ und $\tilde{\mathbb{P}}$ auf $(\Omega, \mathscr{F} \otimes \mathscr{F}_\Theta)$ gemäß $\tilde{\mathbb{P}}(dx, d\vartheta) = \mathbb{P}_\vartheta(dx)\pi(d\vartheta)$ (gemeinsame Verteilung von Beobachtung und Parameter). Falls $\vartheta \mapsto \mathbb{P}_\vartheta(A)$ für alle $A \in \mathscr{F}$ messbar ist (Markovkern, s.o.), so ist $\tilde{\mathbb{P}}$ wohldefiniert (betrachte $\tilde{\mathbb{P}}(A \times B) := \int_B \mathbb{P}_\vartheta(A)\,\pi(d\vartheta)$ und verwende Maßerweiterungssatz). Bezeichne mit X und $\mathbf{\Theta}$ die Koordinatenprojektionen von Ω auf $\mathcal{X}$ bzw. Θ. Dann gilt $R_\pi(\rho) = \mathbb{E}_{\tilde{\mathbb{P}}}[l(\mathbf{\Theta}, \rho(X))]$.

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

2.13 Satz. *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 $\mathbb{P}_\vartheta \ll \mu$ für alle $\vartheta \in \Theta$ sowie $\pi \ll \nu$ gilt mit Maßen μ und ν und Dichten $f_{X|\boldsymbol{\Theta}=\vartheta}$ bzw. $f_{\boldsymbol{\Theta}}$. Ist $f_{X\,|\,\boldsymbol{\Theta}=\bullet} : \mathcal{X} \times \Theta \to \mathbb{R}^+$ $(\mathscr{F} \otimes \mathscr{F}_\Theta)$-messbar, so besitzt die a-posteriori-Verteilung $\mathbb{P}^{\boldsymbol{\Theta}\,|\,X=x}$ des Parameters eine ν-Dichte, nämlich*

$$f_{\boldsymbol{\Theta}\,|\,X=x}(\vartheta) = \frac{f_{X|\boldsymbol{\Theta}=\vartheta} f_{\boldsymbol{\Theta}}(\vartheta)}{\int_\Theta f_{X|\boldsymbol{\Theta}=\vartheta'} f_{\boldsymbol{\Theta}}(\vartheta')\nu(d\vartheta')} \qquad \textit{(Bayesformel).}$$

Beweis. Übung! □

2.14 Beispiel. Für einen Bayestest (oder auch ein Bayes-Klassifikationsproblem) setze $\Theta = \{0,1\}$, $A = \{0,1\}$, $l(\vartheta, a) = |\vartheta - a|$ (0-1-Verlust) 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

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

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

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

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

Beweis. Für eine beliebige Entscheidungsregel ρ' gilt

$$R_\pi(\rho') = \mathbb{E}_{\tilde{\mathbb{P}}}[\mathbb{E}_{\tilde{\mathbb{P}}}[l(\boldsymbol{\Theta}, \rho'(X)) \,|\, X]] \geqslant \mathbb{E}_{\tilde{\mathbb{P}}}[\mathbb{E}_{\tilde{\mathbb{P}}}[l(\boldsymbol{\Theta}, \rho(X)) \,|\, X]] = R_\pi(\rho).$$

□

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

2.17 Beispiel. (Fortsetzung) Nach Satz 2.15 finden wir einen Bayestest $\varphi(x)$ als Minimalstelle von

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

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.

2.18 Satz. *Es liege die Situation aus der vorangegangenen Definition vor.*

(a) 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.

(b) Für eine Minimaxregel ρ gilt $\sup_\pi R_\pi(\rho) = \inf_{\rho'} \sup_\pi R_\pi(\rho')$.

Beweis.

(a) 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 folgt daher die Behauptung.

(b) Nach (a) muss $\sup_{\vartheta\in\Theta} R(\vartheta,\rho) = \inf_{\rho'} \sup_{\vartheta\in\Theta} R(\vartheta,\rho')$ gezeigt werden, was gerade die Minimaxeigenschaft ist.

□

2.19 Bemerkung. Man kann diesen Satz insbesondere dazu verwenden, untere Schranken für das Minimax-Risiko durch das Risiko von Bayesschätzern abzuschätzen.

2.20 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! □

2.21 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.*

Beweis. Betrachte die a-priori-Verteilung $\pi = N(0, \sigma^2 E_d)$ für μ. Dann gilt mit $\eta_1, \dots, \eta_n \sim N(0, E_d)$ i.i.d., unabhängig von μ, die Darstellung $X_i = \mu + \eta_i$, $i = 1, \dots, n$. Als lineare Abbildung von $(\mu, \eta_1, \dots, \eta_n)$ ist $(\mu, X_1, \dots, X_n)$ gemeinsam normalverteilt und die bedingte Erwartung ist linear-affin (vgl. Stochastik II): $\mathbb{E}_{\tilde{\mathbb{P}}}[\mu_j \mid X_1, \dots, X_n] = \sum_{i=1}^n \alpha_{ij}^\top X_i + \beta_j$, $j = 1, \dots, d$. Aus Symmetrie- und Unabhängigkeitsgründen gilt $\alpha_{ij} = \alpha e_j = (0, \dots, 0, \alpha, 0, \dots, 0)^\top$ für ein festes $\alpha \in \mathbb{R}$, und $\mathbb{E}[\mu_j] = 0$ impliziert $\beta_j = 0$. Damit liefert die Orthogonalität $\mathbb{E}[X_{i,j}(\mu_j - \alpha \sum_{l=1}^n X_{l,j})] = 0$ den Wert $\alpha = \frac{1}{n+\sigma^{-2}}$. Der Bayes-optimale Schätzer ist daher $\hat{\mu}_{\sigma,n} = \frac{n}{n+\sigma^{-2}}\overline{X}$ (vektorwertige bedingte Erwartung), seine Risikofunktion ist $R(\mu, \hat{\mu}_{\sigma,n}) = \frac{nd+|\mu|^2\sigma^{-4}}{(n+\sigma^{-2})^2}$.

Somit können wir das Minimax-Risiko von unten abschätzen:

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

wie behauptet, da $R(\mu, \bar{X}) = \frac{d}{n}$. □

2.22 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 $\mu \mapsto R(\mu, \hat{\mu})$ 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 $\frac{1}{n} - R_{N(0,\sigma^2)}$ größer als ein Vielfaches von σ^{-1} für $\sigma \to \infty$, während für den Bayesschätzer $\frac{1}{n} - R_\sigma(\hat{\mu}_{\sigma,n}) = \frac{\sigma^{-2}}{n(n+\sigma^{-2})}$ (s.o.) von der Ordnung σ^{-2} ist. Dies widerspricht der Optimalität des Bayesschätzers bei einer hinreichend großen Wahl von σ. Also ist $\bar{X}$ zulässig. □

2.23 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.

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

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

2.25 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 aus dem gleichen Grund stets '$\leqslant$' folgt, gilt sogar $R_\pi(\rho_\pi) = \sup_\vartheta R(\vartheta, \rho_\pi)$.

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.

□

2.26 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).

2.27 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).

2.3 Das Stein-Phänomen

Wir betrachten folgendes grundlegendes Problem: Anhand einer mathematischen Stichprobe $X_1, \dots, X_n \sim N(\mu, E_d)$ 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

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

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.

2.28 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 $Y \sim N(\mu, \sigma^2 E_d)$ mit $\mu \in \mathbb{R}^d$, $\sigma > 0$,*

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

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

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

$$\mathbb{E}[Y_1 f(Y) \,|\, Y_2 = y_2, \dots, Y_d = y_d] = \mathbb{E}[\tfrac{\partial f}{\partial y_1}(Y) \,|\, Y_2 = y_2, \dots, Y_d = y_d]$$

zu zeigen für Lebesgue-fast alle $y_2, \dots, y_d \in \mathbb{R}$, was wegen Unabhängigkeit gerade für $f_y(u) := f(u, y_2, \dots, y_d)$ die Identität $\int u f_y(u) e^{-u^2/2} du = \int f_y'(u) e^{-u^2/2} du$ ist. Dies folgt durch partielle Integration, sofern die Randterme

verschwinden; ein geschickter Einsatz des Satzes von Fubini zeigt dies jedoch ohne weitere Voraussetzungen:

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

□

Betrachten wir nun allgemeine Schätzer der Form $\hat\mu = g(\bar X)\bar X$, so gilt

$$\begin{aligned}\mathbb{E}_\mu[|\hat\mu-\mu|^2] &= \mathbb{E}_\mu\left[|\bar X-\mu|^2 + |\bar X-\hat\mu|^2 - 2\langle \bar X-\mu, \bar X-\hat\mu\rangle\right] \\ &= \frac{d}{n} + \mathbb{E}_\mu[|(1-g(\bar X))\bar X|^2] - 2\,\mathbb{E}_\mu[\langle \bar X-\mu, (1-g(\bar X))\bar X\rangle].\end{aligned}$$

Kann man nun auf $f(x) = (1-g(x))x : \mathbb{R}^d \to \mathbb{R}^d$ 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

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

und

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

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.

2.29 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 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

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

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

2.30 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).

2.31 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 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 $\hat{\mu}_{JS}$.

2.4 Ergänzungen

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

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

2.33 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.*

2.34 Lemma. *Es sei $\Theta = \Theta_0 \dot{\cup} \Theta_1$, $A = [0, 1]$. Für den Verlust $l(\vartheta, a) = l_0 a \mathbf{1}_{\Theta_0}(\vartheta) + l_1(1-a)\mathbf{1}_{\Theta_1}(\vartheta)$ 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.*

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

2.35 Definition. Ein Entscheidungskern oder 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

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

2.36 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 $A' := \{0,1\}$, $\mathcal{F}_{A'} := \mathcal{P}(A')$, 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.

2.37 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 eine deterministische Entscheidungsregel, deren Risiko nicht größer ist.*

3 Dominierte Modelle und Suffizienz

3.1 Dominierte Modelle

3.1 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.

3.2 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.

3.3 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.*

3.4 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 $\frac{d\mathbb{Q}}{d\mu} > 0$ 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. □

3.2 Exponentialfamilien

3.5 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 $\eta_1, \ldots, \eta_k$ linear unabhängige Funktionen und gilt für alle $\vartheta \in \Theta$ die Implikation

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

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

3.6 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 $\tilde{T}(x) = (A^\top)^{-1} T(x)$. Außerdem kann die Funktion h in das dominierende Maß absorbiert werden: $\tilde{\mu}(dx) := h(x)\mu(dx)$.

(c) 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.

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

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

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

3.8 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 $\mathscr{Z} = \mathbb{R}\times(0,\infty)$. 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.

3.9 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*

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

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

3.10 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 $\tilde{\vartheta}$ ein innerer Punkt von $\mathscr{Z}$, so ist die erzeugende Funktion $\psi_{\tilde{\vartheta}}(s) = \mathbb{E}_{\tilde{\vartheta}}[e^{\langle T,s\rangle}]$ in einer Umgebung der Null

wohldefiniert und beliebig oft differenzierbar. Es gilt $\psi_{\tilde{\vartheta}}(s) = \exp(A(\tilde{\vartheta}+s) - A(\tilde{\vartheta}))$ *für alle* s *mit* $\tilde{\vartheta}+s \in \mathscr{Z}$.

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

Beweis. Übung. □

3.3 Suffizienz

3.11 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 $\mathbb{P}_\vartheta$ und damit ϑ. Hingegen sind $\mathbf{1}(\{X_3 < X_7\})$ oder $\mathbf{1}(\{X_1 = \max(X_1,\dots,X_n)\})$ Statistiken, deren Verteilung nicht von $\mathbb{P}_\vartheta$ abhängt (sofern die i.i.d.-Annahme gültig ist) und somit keinerlei Informationen über ϑ beinhalten (sogenannte *ancillary statistics*). Intuitiv ist alle Information bereits in der Ordnungsstatistik $X_{(1)},\dots,X_{(n)}$ enthalten mit $X_{(1)} = \min\{X_1,\dots,X_n\}$, $X_{(k+1)} := \min\{X_1,\dots,X_n\}\backslash\{X_{(1)},\dots,X_{(k)}\}$. Diese ist in folgendem Sinne suffizient.

3.12 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 reguläre bedingte Wahrscheinlichkeit von $\mathbb{P}_\vartheta$ gegeben T (existiert und) nicht von ϑ abhängt, d.h.

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

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

3.13 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*

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

3.14 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* $B\in\mathscr{F}$

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

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

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

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. □

3.15 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 $h(x) = \frac{d\,\mathbb{Q}}{d\mu}(x)$. □

3.16 Beispiele.

(a) Die Identität $T(x) = x$ und allgemein jede bijektive, bi-messbare Transformation T ist stets 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 $(S_1, \dots, S_n)$ dieselbe Verteilung haben wie die Ordnungsstatistik $(X_{(1)}, \dots, X_{(n)})$ 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).

3.17 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:

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

□

3.18 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.

3.19 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 $\mathbb{E}_\vartheta[\tilde\varphi] = \mathbb{E}_\vartheta[\varphi]$. □

3.20 Beispiel. Es sei $X_1, \dots, X_n$ 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 $R(\vartheta, \hat\vartheta) = 4\operatorname{Var}_\vartheta(\bar X) = \frac{4\vartheta^2}{12n}$. 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.).

3.4 Vollständigkeit

3.21 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

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

3.22 Bemerkung. Wie oben erwähnt, heißt eine Statistik V *ancillary*, wenn ihre Verteilung nicht von ϑ abhängt. Sie heißt *ancillary* erster Ordnung, falls $\mathbb{E}_\vartheta[V]$ unabhängig von ϑ ist. Falls jede Statistik der Form $V = f(T)$, die *ancillary* erster Ordnung ist, $\mathbb{P}_\vartheta$-f.s. konstant ist, so ist keine redundante Information mehr in T enthalten, und T ist vollständig (verwende $\tilde f(T) = f(T) - \mathbb{E}_\bullet[f(T)]$).

3.23 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. □

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

3.25 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 $\tilde{\mu}(dx) = h(x)\mu(dx)$). Es gelte $\mathbb{E}_\vartheta[f(T)] = 0$ für alle $\vartheta \in \Theta$ und ein $f : \mathbb{R}^k \to \mathbb{R}$. 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 $\int f^+(t)\mu^T(dt) = \int f^-(t)\mu^T(dt) =: M$, und $\mathbb{P}^+(dt) := M^{-1}f^+(t)\mu^T(dt)$, $\mathbb{P}^-(dt) := M^{-1}f^-(t)\mu^T(dt)$ definieren 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 $\vartheta \in [-a,a]$ übereinstimmen. χ^+ und χ^- sind darüberhinaus auf dem k-dimensionalen komplexen Streifen $\{\vartheta \in \mathbb{C}^k \mid |\mathrm{Re}(\vartheta_j)| < a\}$ wohldefiniert und analytisch. 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 char. Funktionen). Dies liefert $f^+ = f^-$ μ^T-f.ü. und somit $f(T) = 0$ $\mathbb{P}_\vartheta$-f.s. für alle $\vartheta \in \Theta$. □

3.26 Beispiele.

(a) Das lineare Modell $Y = X\beta + \sigma\varepsilon$ mit 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).
Beachte: hierfür ist die Normalverteilungsannahme essentiell.

(b) Es sei $X_1, \ldots, X_n \sim U([0, \vartheta])$ 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 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.

3.5 Cramér-Rao-Schranke

3.27 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 $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta_0}$, $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta_0}$, $\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} \in L^2(\mathbb{P}_{\vartheta_0})$*

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

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

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

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

3.28 Beispiel. 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.

3.29 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 $\vartheta_0 \in \mathrm{int}(\Theta)$, wenn es einen Zufallsvektor $\dot{\ell}(\vartheta_0) \in \mathbb{R}^k$ 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 $\vartheta_0 \in int(\Theta)$ 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.

3.30 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 notwendigerweise $\langle \dot{\ell}(\vartheta_0, x), \vartheta - \vartheta_0\rangle \sqrt{L(\vartheta_0, x)} \in L^2(\mu)$ und damit $\dot{\ell}(\vartheta_0, x) \in L^2(\mathbb{P}_{\vartheta_0}, \mathbb{R}^k)$, und $I(\vartheta_0)$ ist stets wohldefiniert.

(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.

3.31 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 (einem Zufallsvektor in $\mathbb{R}^k$) $\dot{L}_{\vartheta_0}(\vartheta_0)$ 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 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 $o(|\vartheta - \vartheta_0|)$. □

3.32 Beispiel. Es sei $X_1, \dots, X_n$ 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

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

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

3.33 Satz (Cramér-Rao-Schranke)**.** *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein statistisches Modell, $g : \Theta \to \mathbb{R}$ differenzierbar bei $\vartheta_0 \in \mathrm{int}(\Theta)$ und $\hat{g}$ 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 $\vartheta_h = \vartheta_0 + he$ mit $h \downarrow 0$ und $e \in \mathbb{R}^k$ einem beliebigen Einheitsvektor (*Richtung*). Dann folgt aus der Chapman-Robbins-Ungleichung, $L_{\vartheta_0}(\vartheta_0) = 1 = \mathbb{E}_{\vartheta_0}[L_{\vartheta_0}(\vartheta)]$ und $\dot{\ell}(\vartheta_0) = \dot{L}_{\vartheta_0}(\vartheta_0)$

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

Nehmen wir nun das Supremum der rechten Seite über alle $e \in \mathbb{R}^k$, so folgt die Behauptung. □

3.34 Bemerkung. 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 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.

3.35 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 3.10).*

Sofern $I(\vartheta_0)$ strikt positiv-definit ist, erreicht T_i, $i = 1, \dots, k$, 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 3.10 gilt $g(\vartheta) = E_\vartheta[T] = \dot{A}(\vartheta)$ und $\mathrm{Cov}_\vartheta(T) = \ddot{A}(\vartheta)$ (Kovarianzmatrix). Andererseits ist die Loglikelihoodfunktion $\ell_{\vartheta_0}(\vartheta) = (\vartheta - \vartheta_0)T - (A(\vartheta) - A(\vartheta_0))$, 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 3.10 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. □

3.36 Beispiel. Es sei $X_1, \dots, X_n$ 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.

3.37 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.

3.38 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 $L^1(\mu)$-differenzierbar mit Ableitung $\dot{\ell}(\vartheta)L(\vartheta)$, und es gilt $\mathbb{E}_{\vartheta_0}[\dot{\ell}(\vartheta_0)] = 0$.*

Beweis. Betrachte das Kriterium für $L^1(\mu)$-Differenzierbarkeit mit $\dot{L}(\vartheta_0) = \dot{\ell}(\vartheta_0)L(\vartheta_0)$:

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

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 $i = 1, \ldots, k$. □

3.39 Lemma. *Es seien $X_1, \ldots, X_n$ Beobachtungen aus unabhängigen Hellinger-differenzierbaren Modellen $\mathcal{E}_1, \ldots, \mathcal{E}_n$ 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 $X_1, \ldots, X_n$, 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 $L_1, \ldots, L_n$ bezüglich der dominierenden Maße $\mu_1, \ldots, \mu_n$ Hellinger-differenzierbar mit Score-Funktionen $\ell_1, \ldots, \ell_n$. 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 $\ell(\vartheta, x) = \sum_{j=1}^n \dot{\ell}_j(\vartheta, x_j)$, 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

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

□

4 Allgemeine Schätztheorie

4.1 Momentenschätzer

4.1 Definition. Es seien $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta \in \Theta})$ ein statistisches (Produkt-)Modell mit $\mathcal{X} \subseteq \mathbb{R}$, $\mathscr{F} \subseteq \mathfrak{B}_\mathbb{R}$ 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

$$\varphi(\vartheta) := \mathbb{E}_\vartheta[\psi] = \Big(\int_\mathcal{X} \psi_j(x) \; \mathbb{P}_\vartheta(dx) \Big)_{j=1,\dots,q}$$

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 $\psi_1, \dots, \psi_q$.

4.2 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):

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

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:

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

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

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

4.3 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 $\hat{g}_n$ (stark) konsistent, 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:

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

□

4.4 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, 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 $Z \sim N(0, E_k)$. □

4.5 Beispiel. Aus einer mathematischen Stichprobe $X_1, \dots, X_n \sim \text{Poiss}(\lambda)$ 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.

4.6 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 $\text{Cov}_{\vartheta_0}(\psi) := (\text{Cov}_{\vartheta_0}(\psi_i, \psi_j))_{i,j=1,\dots,q}$. 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 $\langle \text{Cov}_{\vartheta_0}(\psi)\dot{G}(\varphi(\vartheta_0)), \dot{G}(\varphi(\vartheta_0))\rangle$:*

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

4.7 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 $N(0, \langle \text{Var}_{\vartheta_0}(\psi)\dot{G}(\varphi(\vartheta_0)), \dot{G}(\varphi(\vartheta_0))\rangle)$ konvergieren (dafür wird gleichgradige Integrierbarkeit benötigt).

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

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

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

4.8 Beispiel. Im Exponentialverteilungsmodell aus Beispiel 4.2 gilt $G'(x) = -(k!/x)^{1/k}(kx)^{-1}$ und $\Sigma(\lambda_0) = \text{Var}_{\lambda_0}(X_i^k) = ((2k)! - (k!)^2)/\lambda_0^{2k}$. 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.

4.9 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.

4.2 Maximum-Likelihood- und Minimum-Kontrast-Schätzer

4.10 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 $X_1, \dots, X_n \sim \mathrm{Poiss}(\lambda)$ mit $\lambda > 0$ unbekannt, ergibt sich beispielsweise

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

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 $\hat{\vartheta}(x) = \operatorname{argmax}_{\vartheta\in\Theta} f_\vartheta(x)$. 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

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

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.

4.11 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.

4.12 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.

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

4.14 Bemerkung. Bei einer eineindeutigen Parametrisierung $\vartheta \mapsto h(\vartheta)$ ist dann natürlich $\hat{h} := h(\hat{\vartheta})$ der MLE für $h(\vartheta)$.

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 3.10) wegen der Differenzierbarkeit im Innern $\dot{\ell}(\hat{\vartheta}) = T(x) - \dot{A}(\hat{\vartheta}) = 0$ und somit $\mathbb{E}_{\hat{\vartheta}}[T] = T(x)$. □

4.15 Beispiele.

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

4.16 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 $\mathbb{Q}$.

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

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

(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*

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

Beweis. Übung! □

4.18 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 $\mathrm{KL}(N(\vartheta_0, \Sigma) \mid N(\vartheta, \Sigma)) = \frac{1}{2}\langle \Sigma^{-1}(\vartheta - \vartheta_0), \vartheta - \vartheta_0 \rangle$ gilt.

4.19 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 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).

4.20 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

$$K(\vartheta_0, \vartheta) = \mathrm{KL}(\vartheta_0 \,|\, \vartheta) - \mathrm{KL}(\vartheta_0 \,|\, \mu)$$

(schreibe kurz $\mathrm{KL}(\vartheta_0 \,|\, \vartheta) := \mathrm{KL}(\mathbb{P}_{\vartheta_0} \,|\, \mathbb{P}_\vartheta)$ etc.). Der zugehörige Minimum-Kontrast-Schätzer ist der MLE.

(b) Es sei $g(\vartheta) = \vartheta$, $\Gamma = \Theta$. Betrachte das Regressionsmodell aus Beispiel 4.15 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 Minimum-Kontrast-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-Orthogonalfunktion $g(\vartheta)$ 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 $\leqslant d$.

4.3 Asymptotik

4.21 Satz. *Es sei $(K_n)_{n\geqslant 1}$ ein Kontrastprozess zur Kontrastfunktion K. Dann ist der zugehörige Minimum-Kontrast-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.*

4.22 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).

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

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

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 $K(\vartheta_0, \bullet)$. □

4.23 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. □

4.24 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 $X_n/a_n \xrightarrow{\mathbb{P}} 0$.

4.25 Satz. *Der Minimum-Kontrastschätzer $\hat\gamma_n$ 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(Γ). Der Kontrastprozess K_n sei zweimal stetig differenzierbar in einer Umgebung von γ_0 ($\mathbb{P}^n_{\vartheta_0}$-f.s.), so dass mit*

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

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 Minimum-Kontrast-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 $\gamma_0 \in \text{int}(\Gamma)$ für $\Omega^1_n := \{[\hat\gamma_n, \gamma_0] \in \text{int}(\Gamma)\}$ (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 $\mathbb{P}^n_{\vartheta_0}(\Omega^2_n) \to 1$ für $\Omega^2_n := \{V_n(\bar\gamma_n)^{-1} \text{ existiert}\}$. Benutze nun $V_n(\bar\gamma_n)^{-1}\mathbf{1}_{\Omega^1_n\cap\Omega^2_n} \to V(\gamma_0)^{-1}$ 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. □

4.26 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

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

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. $\varepsilon_i \stackrel{d}{=} -\varepsilon_i$ 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 Dichte f_ε von ε_i liefert heuristisch für den Stichprobenmedian die asymptotische Varianz $\mathbb{E}[\operatorname{sgn}(\varepsilon_i)^2] / \mathbb{E}[2\delta_0(\varepsilon_i)]^2 = (4 f_\varepsilon(0))^{-1}$ sowie für den Huber-Schätzer $\mathbb{E}[\varepsilon_i^2 \wedge \kappa^2] / \mathbb{P}(|\varepsilon_i| \leqslant \kappa)^2$, vergleiche Übungen für rigorose Herleitungen.

4.27 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 dominierter Produktexperimente mit eindimensionaler Loglikelihoodfunktion $\ell(\vartheta, x) = \log(\frac{d\mathbb{P}_\vartheta}{d\mu}(x))$. 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 positiv definit.

Dann erfüllt der MLE $\hat{\vartheta}_n$

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

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 $g(\vartheta) = \vartheta$, $\Gamma = \Theta$, $K_n(\vartheta, x) := -\frac{1}{n}\sum_{i=1}^n \ell(\vartheta, x_i)$, $x \in \mathcal{X}$, 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.:

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

Mit dominierter Konvergenz und 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 4.23). 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 $\mathbb{P}_{\vartheta_0}$

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

(B2) Mittels dominierter Konvergenz erhalten wir wie in Lemma 3.38 $\mathbb{E}_{\vartheta_0}[\dot{\ell}(\vartheta_0)] = 0$. Wir verwenden nun die Identität

$$\ddot{\ell}(\vartheta) = \frac{\ddot{L}(\vartheta)}{L(\vartheta)} - \dot{\ell}(\vartheta)\dot{\ell}(\vartheta)^\top,$$

und erhalten mit dominierter Konvergenz

$$\mathbb{E}_{\vartheta_0}[\ddot{\ell}(\vartheta_0)] + I(\vartheta_0) = \mathbb{E}_{\vartheta_0}[\ddot{L}(\vartheta_0)/L(\vartheta_0)] = \int \ddot{L}(\vartheta_0)\, d\mu = \ddot{\mathbf{1}} = 0.$$

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 $\lim_{\delta\to 0}\lim_{n\to\infty}\mathbb{P}_{\vartheta_0}(\Omega_{\delta,n}) = 1$ und der Konvergenz von $V_n(\vartheta_0)$ erhalten wir daher in $\mathbb{P}_{\vartheta_0}^{\otimes n}$-Wahrscheinlichkeit

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

□

4.28 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 $\vartheta^* := \mathrm{argmax}_{\vartheta\in\Theta}\int_\mathcal{X} \ell(\vartheta, x)\,\mathbb{P}_0(dx)$, sofern ϑ^* existiert und eindeutig ist. Es gilt entsprechend $\vartheta^* = \mathrm{argmin}_{\vartheta\in\Theta}\mathrm{KL}(\mathbb{P}_0 \,|\, \mathbb{P}_\vartheta)$, 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 4.25 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.

4.29 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-Information $I(\vartheta) = \operatorname{Var}_\vartheta(T)$ (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 $X_i \sim \operatorname{Bin}(1,p)$ 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.

4.30 Definition. Im Rahmen des vorigen Satzes heißt die zufällige Matrix

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

beobachtete Fisher-Informationsmatrix.

4.31 Korollar. *Unter den Voraussetzungen des vorigen Satzes gilt unter* $\mathbb{P}_{\vartheta_0}$

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

Insbesondere sind für $k = 1$ *und das* $(1-\alpha/2)$*-Quantil* $q_{1-\alpha/2}$ *der Standardnormalverteilung* $[\hat{\vartheta}_n - n^{-1/2} I(\hat{\vartheta}_n)^{-1/2} q_{1-\alpha/2}, \hat{\vartheta}_n + n^{-1/2} I(\hat{\vartheta}_n)^{-1/2} q_{1-\alpha/2}]$ *und* $[\hat{\vartheta}_n - n^{-1/2}\mathfrak{I}_n^{-1/2} q_{1-\alpha/2}, \hat{\vartheta}_n + n^{-1/2}\mathfrak{I}_n^{-1/2} q_{1-\alpha/2}]$ *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 $I(\hat{\vartheta}) \to I(\vartheta_0)$ in $\mathbb{P}_{\vartheta_0}$-Wahrscheinlichkeit. Ebenso liefert das Argument im obigen Nachweis der Eigenschaft (B2) $\mathfrak{I}_n \to I(\vartheta_0)$ $\mathbb{P}_{\vartheta_0}$-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). □

4.32 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 $\mathfrak{I}_n = \frac{\sum_{i=1}^n a_i}{n\sigma_0^2} + \frac{\sum_{i=1}^n (1-a_i)}{n\sigma_1^2}$ erfüllt. Damit ist $\mathfrak{I}_n^{-1}$ gerade gleich der *bedingten* Varianz $\mathrm{Var}_\vartheta(n^{1/2}\hat\vartheta_n \,|\, a)$. Da wir ja a 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 gegenüber der plug-in-Schätzung $I(\hat\vartheta_n)$.

4.4 Allgemeine Schranken (nicht behandelt)

Wir wollen nun Wahrscheinlichkeiten der Form $\mathbb{P}_\vartheta(|\hat\vartheta_n - \vartheta| > \kappa)$ gleichmäßig in $\vartheta \in \Theta$ und für festes n abschätzen, um auch nicht-asymptotische Konfidenzaussagen zu erhalten. Gerade in Anwendungen wird häufig für größere Stichprobenumfänge n auch ein komplexeres Modell gewählt (z.B. mit $\dim(\Theta_n) \to \infty$), um bei gleicher statistischer Güte des Verfahrens den Modellfehler zu verringern (*jedes statistische Modell ist falsch*). Aus mathematischer Sicht ist es für Minimum-Kontrast-Schätzer zunächst natürlich, den Fehler durch $K(\vartheta_0, \hat\vartheta_n) - K(\vartheta_0, \vartheta_0)$ zu messen (bei MLE: *fitted log-likelihood*). Beim MLE für Exponentialfamilien entspricht dies gerade einem gewichteten quadratischen Verlust (s.o.).

4.33 Lemma. *Für einen Minimum-Kontrast-Schätzer $\hat\vartheta_n$ bezüglich einem Kontrastprozess K_n und einer Funktion $K(\vartheta_0, \bullet)$, die ihr Minimum bei $\vartheta_0 \in \Theta_0$ annimmt (z.B. K Kontrastfunktion), gilt*

$$\begin{aligned} 0 \leqslant K(\vartheta_0, \hat\vartheta_n) - K(\vartheta_0, \vartheta_0) &\leqslant \big(K_n(\vartheta_0) - K_n(\hat\vartheta_n)\big) - \big(K(\vartheta_0, \vartheta_0) - K(\vartheta_0, \hat\vartheta_n)\big) \\ &\leqslant \sup_{\vartheta\in\Theta} \Big(\big(K_n(\vartheta_0) - K_n(\vartheta)\big) - \big(K(\vartheta_0, \vartheta_0) - K(\vartheta_0, \vartheta)\big)\Big). \end{aligned}$$

Beweis. Aus den Definitionen erhalten wir, dass $K(\vartheta_0, \hat\vartheta_n) - K(\vartheta_0, \vartheta_0)$ und $K_n(\vartheta_0) - K_n(\hat\vartheta_n)$ nicht-negativ sind, so dass

$$0 \leqslant K(\vartheta_0, \hat\vartheta_n) - K(\vartheta_0, \vartheta_0) \leqslant \big(K_n(\vartheta_0) - K_n(\hat\vartheta_n)\big) - \big(K(\vartheta_0, \vartheta_0) - K(\vartheta_0, \hat\vartheta_n)\big).$$

Das zufällige Argument $\hat\vartheta_n$ liefert stets einen kleineren Wert als das Maximum. □

4.34 Beispiele. In Fortführung von Beispiel 4.20 erhalten wir:

(a) Beim MLE aus einer mathematischen Stichprobe $X_1, \dots, X_n$ gilt

$$\mathrm{KL}(\vartheta_0 \,|\, \hat\vartheta_n) \leqslant \sup_{\vartheta\in\Theta} \Big(\frac{1}{n}\sum_{i=1}^n \big(\ell(\vartheta, X_i) - \ell(\vartheta_0, X_i) + \mathrm{KL}(\vartheta_0 \,|\, \vartheta)\big)\Big).$$

Unter $\mathbb{P}_{\vartheta_0}$ hat der Term in Klammern Erwartungswert Null und eine Varianz von der Ordnung $1/n$. Das Supremum lässt sich allerdings nicht unmittelbar behandeln.

Im Gaußschen Shiftmodell $X_i \sim N(\vartheta, \sigma^2)$ gilt $\mathrm{KL}(\vartheta_0 \,|\, \hat{\vartheta}_n) = \frac{(\vartheta_0 - \hat{\vartheta}_n)^2}{2\sigma^2}$. Wir schätzen hier also einfach den quadratischen Verlust ab.

(b) Beim Kleinste-Quadrate-Schätzer im nichtlinearen Regressionsmodell gilt unter $\mathbb{P}_{\vartheta_0}$

$$\|g_{\hat{\vartheta}_n} - g_{\vartheta_0}\|_{L^2}^2 \leqslant \sup_{\vartheta \in \Theta} \Big(\frac{2}{n} \sum_{i=1}^n \varepsilon_i (g_{\vartheta_0} - g_\vartheta)(x_i) - \frac{1}{n} \sum_{i=1}^n (g_{\vartheta_0} - g_\vartheta)^2(x_i) + \|g_{\vartheta_0} - g_\vartheta\|_{L^2}^2 \Big).$$

Die letzten beiden Terme bilden einen deterministischen Approximationsfehler, der für glatte g_ϑ im Allgemeinen klein ist. Da auf den nichtasymptotischen Fall Wert gelegt werden soll, ist es vernünftig, statt der $L^2([0,1])$-Norm die *empirische L^2-Norm* $\|f\|_n^2 := \frac{1}{n} \sum_{i=1}^n f(x_i)^2$ zu betrachten, wo dieselben Argumente nur den stochastischen Term liefern:

$$\|g_{\hat{\vartheta}_n} - g_{\vartheta_0}\|_n^2 \leqslant \sup_{\vartheta \in \Theta} \Big(\frac{2}{n} \sum_{i=1}^n \varepsilon_i (g_{\vartheta_0} - g_\vartheta)(x_i) \Big).$$

Der innere Term allein ist zentriert und besitzt die Varianz $\frac{\sigma^2}{n} \|g_{\vartheta_0} - g_\vartheta\|_n^2$ (setze $\sigma^2 = \mathrm{Var}(\varepsilon_1)$), ist also von der Ordnung $n^{-1/2}$. Allerdings ist a priori überhaupt nicht klar, ob dasselbe für das Supremum gilt. Wenn beispielsweise $\{((g_{\vartheta_0} - g_\vartheta)(x_i))_{i=1,\dots,n} \,|\, \vartheta \in \Theta\}$ einen Würfel $\{z \in \mathbb{R}^n \mid \max_i |z_i| \leqslant r\}$ mit $r > 0$ enthält, so ist das Supremum größer als $\frac{r}{n} \sum_{i=1}^n |\varepsilon_i|$, was im Erwartungswert $r\, \mathbb{E}[|\varepsilon_1|]$ ergibt und nicht gegen Null konvergiert.

Im folgenden werden wir insbesondere auch den Fall der Modell-Misspezifikation mitbehandeln, das heißt, dass die wahre Verteilung $\mathbb{P}$ nicht notwendigerweise zur parametrischen Familie $(\mathbb{P}_\vartheta)$ gehört. Es ist dann interessant zu sehen, in welchen Fällen der Modellfehler von kleinerer Ordnung als der statistische Fehler ist, so dass ein vereinfachtes Modell weiterhin vernünftige Aussagen erlaubt. Wir benötigen zunächst das Werkzeug der *Konzentrationsungleichungen.*

4.35 Definition. Eine (reellwertige) Zufallsvariable X erfüllt eine Exponentialungleichung mit Parametern $C, D > 0$, $R \in [0, +\infty]$, falls für alle $r \in [0, R]$ gilt

$$\mathbb{P}(|X| > r) \leqslant C e^{-r/D}.$$

4.36 Beispiel. Aus $\mathbb{E}[e^{|X|/D}] \leqslant C < \infty$ folgt mittels verallgemeinerter Markovungleichung die Exponentialungleichung mit $C, D > 0$ für alle $r > 0$ ($R = \infty$).

4.37 Satz (Bernstein-Ungleichung, 1923)**.** *Es seien $X_1, \dots, X_n$ unabhängige Zufallsvariablen mit $\mathbb{E}[X_i] = 0$ und $S_n := \sum_{i=1}^n X_i$. Falls für eine deterministische Konstante $K > 0$ und alle $i = 1, \dots, n$ fast sicher $|X_i| \leqslant K$ gilt, so folgt*

$$\mathbb{P}(|S_n| \geqslant \kappa) \leqslant 2 \exp\Big(- \frac{\kappa^2}{4(\mathrm{Var}(S_n) + \kappa K)} \Big), \quad \kappa > 0.$$

Beweis. Siehe z.B. van der Vaart oder Shiryaev. □

4.38 Beispiele.

(a) Im Fall der Binomialverteilung erhalten wir für $T_n = \sum_{i=1}^n (Y_i - \mathbb{E}[Y_i])$ und eine Bernoullikette $(Y_i)_{i\geqslant 1}$ mit Erfolgswahrscheinlichkeit $p \in (0,1)$:

$$|Y_i - \mathbb{E}[Y_i]| \leqslant \max(p, 1-p), \quad \operatorname{Var}(T_n) = np(1-p).$$

Also impliziert die Bernsteinungleichung $\mathbb{P}(|T_n| \geqslant \kappa) \leqslant 2\exp(-\kappa^2/(4(np(1-p)+\kappa\max(p,1-p))))$. Für $S_n = \sum_{i=1}^n Y_i \sim \operatorname{Bin}(n,p)$ folgt durch Skalierung

$$\mathbb{P}\Big(|S_n - np| \geqslant \kappa\sqrt{np(1-p)}\Big) \leqslant 2\exp\Big(-\frac{\kappa^2}{4 + 4\max(p,1-p)\kappa(np(1-p))^{-1/2}}\Big).$$

Dies bedeutet, dass für große n (bei festem p) die *tails* (*Flanken*) der standardisierten Binomialverteilung fast wie $e^{-\kappa^2/4}$ abfallen, also *Gaußsche tails* vorliegen. Im Fall $np_n \to \lambda > 0$ für $n \to \infty$ ergibt sich nur eine Exponentialungleichung der Ordnung $e^{-\kappa\sqrt{\lambda}/4}$, sogenannte Poissonsche *tails* (vgl. Poissonscher Grenzwertsatz).

(b) Wenn die (X_i) eine Exponentialungleichung mit denselben Parametern $C, D, R > 0$ erfüllen, so folgt für beliebiges $r \in [0, R]$ durch Fallunterscheidung

$$\mathbb{P}(|S_n| \geqslant \kappa) \leqslant \mathbb{P}(\exists i = 1,\dots,n : |X_i| > r) + 2\exp\Big(-\frac{\kappa^2}{4(\operatorname{Var}(S_n) + \kappa r)}\Big).$$

Der erste Term ist kleiner als $nCe^{-r/D}$ und wir wählen $r = D(\log(n) + \kappa^2/(4\operatorname{Var}(S_n)))$ (sofern kleiner R), so dass wir insgesamt die Abschätzung

$$\mathbb{P}(|S_n| \geqslant \kappa) \leqslant (2+C)\exp\Big(-\frac{\kappa^2}{4(\operatorname{Var}(S_n) + \kappa D\log(n) + \kappa^3 D/(4\operatorname{Var}(S_n)))}\Big)$$

erhalten. Im i.i.d.-Fall gilt $\operatorname{Var}(S_n) = n\operatorname{Var}(X_i)$ und Abweichungen von S_n der Größe $x\sqrt{n}$ sind im wesentlichen durch Gaußsche *tails* $e^{-x^2/(4\operatorname{Var}(X_i))}$ beschränkt, sofern $x = o(n^{1/6})$ gilt:

$$\mathbb{P}(|S_n| \geqslant x\sqrt{n}) \leqslant (2+C)\exp\Big(-\frac{x^2}{4\operatorname{Var}(X_i) + \frac{4xD}{\sqrt{n}}(\log(n) + x^2/(4\operatorname{Var}(X_i)))}\Big).$$

Um das Supremum in Lemma 4.33 abzuschätzen, werden wir $A_\vartheta = \sqrt{n}(K_n(\vartheta) - K(\vartheta_0, \vartheta))$ im folgenden Satz betrachten.

4.39 Satz. *Es sei $(A_\vartheta, \vartheta \in \Theta)$ mit $\Theta \subseteq \mathbb{R}^k$ beschränkt ein stetiger Prozess (d.h. $\vartheta \mapsto A_\vartheta$ ist f.s. sicher) und es mögen die Exponentialungleichung*

$$\forall \vartheta, \vartheta' \in \Theta : \mathbb{P}(|A_\vartheta - A_{\vartheta'}| \geqslant r|\vartheta - \vartheta'|) \leqslant Ce^{-r/D}, \quad r \in [0, R],$$

bzw.

$$\forall \vartheta, \vartheta' \in \Theta : \mathbb{P}(|A_\vartheta - A_{\vartheta'}| \geqslant r|\vartheta - \vartheta'|) \leqslant Ce^{-r^2/D^2}, \quad r \in [0, R],$$

mit Konstanten $C, D > 0$, $R \in (0, +\infty]$ gelten. Dann gibt es für beliebige $\vartheta_0 \in \Theta$, $\delta > 0$ eine Konstante $C_{\delta,k}$, so dass mit $\rho := \sup_{\vartheta\in\Theta}|\vartheta - \vartheta_0|$

$$\mathbb{P}\Big(\sup_{\vartheta\in\Theta}|A_\vartheta - A_{\vartheta_0}| \geqslant r\rho\Big) \leqslant C_{\delta,k} \exp\Big(-\frac{r}{(2+\delta)D}\Big), \quad r \in [0, R],$$

bzw.

$$\mathbb{P}\Big(\sup_{\vartheta\in\Theta}|A_\vartheta - A_{\vartheta_0}| \geqslant r\rho\Big) \leqslant C_{\delta,k} \exp\Big(-\frac{r^2}{((2+\delta)D)^2}\Big), \quad r \in [0, R],$$

gilt. $C_{\delta,k}$ wächst dabei mit der Dimension k und mit δ^{-1}.

Beweis. Setze $\Theta_0 = \{\vartheta_0\}$ und wähle endliche Teilmengen $\Theta_{j-1} \subseteq \Theta_j \subseteq \Theta$ für $j \geqslant 1$ mit $\sup_{\vartheta\in\Theta} \inf_{\vartheta_j\in\Theta_j}|\vartheta - \vartheta_j| \leqslant \rho 2^{-j}$ (Θ_j ist $\rho 2^{-j}$-Netz). Man kann stets $|\Theta_j| \leqslant C_1 2^{jk}$ mit einer Konstanten $C_1 = C_1(k) > 0$ erreichen (Übung!).

Für jedes $\vartheta \in \Theta$ konvergiert $\tau_j(\vartheta) := \operatorname{argmin}_{\vartheta_j\in\Theta_j}|\vartheta - \vartheta_j|$ für $j \to \infty$ gegen ϑ. Wegen der Stetigkeit von A gilt $A_\vartheta - A_{\vartheta_0} = \sum_{j=1}^{\infty}(A_{\tau_j(\vartheta)} - A_{\tau_{j-1}(\vartheta)})$. Wir verwenden nun ein *Chaining*-Argument, indem wir $\eta_j \geqslant 0$ mit $\sum_{j\geqslant 1}\eta_j = 1$ wählen:

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\vartheta\in\Theta}|A_\vartheta - A_{\vartheta_0}| \geqslant r\rho\Big) &\leqslant \mathbb{P}\Big(\exists j \geqslant 1 : \sup_{\vartheta_j\in\Theta_j}|A_{\vartheta_j} - A_{\tau_{j-1}(\vartheta_j)}| \geqslant r\rho\eta_j\Big) \\
&\leqslant \sum_{j\geqslant 1}|\Theta_j| \sup_{|\vartheta-\vartheta'|\leqslant\rho 2^{-j+1}} \mathbb{P}(|A_\vartheta - A_{\vartheta'}| \geqslant r\rho\eta_j) \\
&\leqslant \sum_{j\geqslant 1} C_1 2^{jk} C \exp(-(r\eta_j 2^{j-1}/D)^\beta)
\end{aligned}$$

mit $\beta \in \{1, 2\}$ (je nach Voraussetzung). Wähle nun $\eta_j = ((1 + \varepsilon jk)2^{-j+1})/C_2$ mit $C_2 = \sum_{j\geqslant 1}(1 + \varepsilon jk)2^{-j+1} = 2 + 4\varepsilon k$. Dann ist die letzte Zeile kleiner als

$$CC_1 e^{-r^\beta/(DC_2)^\beta} \sum_{j\geqslant 1} 2^{jk} e^{-(r\varepsilon jk/(DC_2))^\beta}.$$

Für $\varepsilon := 2D/(r - 4k)$ und $r > 4k$ ist die Summe maximal $2^k/(e^k - 2^k)$. Da für kleines r eine Exponentialungleichung stets durch Vergrößerung des Vorfaktors C erreicht werden kann, können wir für jedes $\delta > 0$ eine Konstante $C_{\delta,k} \geqslant CC_1 2^k/(e^k - 2^k)$ wählen, so dass die Behauptung gilt. □

4.5 Anwendung auf Regression und Maximum-Likelihood (n.b.)

Um konkrete Resultate für Minimum-Konstrast-Schätzer zu erhalten, wenden wir die Techniken auf den Kleinste-Quadrate-Schätzer in der Regression und später auf den MLE an.

Im parametrischen Regressionsmodell betrachten wir

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

mit $x_i \in D \subseteq \mathbb{R}^d$ und $g_\vartheta \in C(D)$, $\vartheta \in \Theta \subseteq \mathbb{R}^k$, während wir unter dem zugrundeliegenden Wahrscheinlichkeitsmaß $\mathbb{P}$ in Wirklichkeit

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

mit einer beliebigen Funktion $g \in C(D)$ beobachten, wobei (ε_i) i.i.d. mit $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ einer Exponentialungleichung mit Parametern $C_\varepsilon, D_\varepsilon > 0$ und $R = \infty$ genüge. Der (parametrische) Kleinste-Quadrate-Schätzer $\hat\vartheta_n$ minimiert den Kontrastprozess

$$K_n(\vartheta) = \frac{1}{n}\sum_{i=1}^n (Y_i - g_\vartheta(x_i))^2 =: \|Y - g_\vartheta\|_n^2$$

(in Anlehnung an Notation von oben). Wir nehmen an, dass $\{g_\vartheta \,|\, \vartheta \in \Theta\}$ eine bezüglich $\|\bullet\|_n$ abgeschlossene und konvexe Menge ist. Dann existiert $\hat g_n := g_{\hat\vartheta_n}$ und ist sogar eindeutig. Mit $K(\vartheta_0, \vartheta) := \|g_\vartheta - g\|_n^2$ und $\vartheta_0 := \mathrm{argmin}_{\vartheta\in\Theta}\|g_\vartheta - g\|_n^2$ (g_{ϑ_0} ist Projektion von g auf $\{g_\vartheta \,|\, \vartheta \in \Theta\}$) liefert Lemma 4.33

$$\|\hat g_n - g\|_n^2 - \|g_{\vartheta_0} - g\|_n^2 \leqslant \|Y - g_{\vartheta_0}\|_n^2 - \|Y - \hat g_n\|_n^2 - (\|g - g_{\vartheta_0}\|_n^2 - \|g - \hat g_n\|_n^2) = 2\langle \varepsilon, \hat g_n - g_{\vartheta_0}\rangle.$$

Für später notieren wir, dass aus Konvexitätsgründen $\|\hat g_n - g\|_n^2 \geqslant \|\hat g_n - g_{\vartheta_0}\|_n^2 + \|g_{\vartheta_0} - g\|_n^2$ gilt und somit

$$\|\hat g_n - g_{\vartheta_0}\|_n^2 \leqslant 2\langle \varepsilon, \hat g_n - g_{\vartheta_0}\rangle \Rightarrow \|\hat g_n - g_{\vartheta_0}\|_n \leqslant 2\|\varepsilon\|_n.$$

Betrachte nun den stochastischen Prozess $A_\vartheta := \frac{1}{\sqrt n}\sum_{i=1}^n \varepsilon_i g_\vartheta(i/n)$. Es gilt $\mathbb{E}[A_\vartheta - A_{\vartheta'}] = 0$, $\mathrm{Var}(A_\vartheta - A_{\vartheta'}) = \sigma^2\|g_\vartheta - g_{\vartheta'}\|_n^2$ und betrachte $L_2 \geqslant \|g_\vartheta - g_{\vartheta'}\|_n/|\vartheta - \vartheta'|$, $L_\infty \geqslant \|g_\vartheta - g_{\vartheta'}\|_\infty/|\vartheta - \vartheta'|$. Wie im Beispiel 4.38(b) liefert die Bernstein-Ungleichung für $A_\vartheta - A_{\vartheta'}$ eine Konzentrationsungleichung:

$$\begin{aligned}
&\mathbb{P}(|A_\vartheta - A_{\vartheta'}| \geqslant r|\vartheta - \vartheta'|)\\
&\leqslant (2 + C_\varepsilon)\exp\Big(-\frac{r^2}{4\sigma^2 L_2^2 + \frac{4rD_\varepsilon L_\infty}{\sqrt n}(\log(n) + r^2/(4\sigma^2 L_2^2))}\Big)\\
&=: C_\delta \exp\Big(-\frac{r^2}{(4+\delta)\sigma^2 L_2^2}\Big), \quad \delta > 0,\ r \in [0, R_n], \quad R_n = o(n^{1/6}).
\end{aligned}$$

Damit erhalten wir nach Satz 4.39 eine gleichmäßige Abschätzung für $\{\vartheta_0\} \subseteq \Theta_0 \subseteq \Theta$ mit $\rho(\Theta_0) = \sup_{\vartheta\in\Theta_0}|\vartheta - \vartheta_0| < \infty$

$$\mathbb{P}\Big(\sup_{\vartheta\in\Theta_0}|A_\vartheta - A_{\vartheta'}| \geqslant r\rho(\Theta_0)\Big) \leqslant C_{\delta,k}\exp\Big(-\frac{r^2}{(8+\delta)\sigma^2 L_2^2}\Big), \quad r \in [0, R_n],\ R_n = o(n^{1/6}),$$

mit einer Konstante $C_{\delta,k} > 0$ abhängig von $\delta > 0$ und der Dimension k. Wir nehmen dabei an, dass die Lipschitzkonstante $L_\infty := \sup_{\vartheta,\vartheta'\in\Theta}\frac{\|g_\vartheta - g_{\vartheta'}\|_n}{|\vartheta - \vartheta'|}$ endlich ist (und damit auch das entsprechende L_2).

Um sowohl mit (eventuell) nicht-kompaktem Θ als auch mit der doppelten Zufallsabhängigkeit in $\langle \varepsilon, \hat{g}_n - g_{\vartheta_0}\rangle$ umzugehen, verwenden wir nun die sogenannte *Peeling*-Methode in der Form

$$\mathbb{P}(2\langle \varepsilon, \hat{g}_n - g_{\vartheta_0}\rangle \geqslant \kappa/n) = \mathbb{P}\left(\exists j \geqslant 0 :\, 2\langle \varepsilon, \hat{g}_n - g_{\vartheta_0}\rangle \in [2^{2j}\kappa/n, 2^{2(j+1)}\kappa/n]\right).$$

Aus $2\langle \varepsilon, \hat{g}_n - g_{\vartheta_0}\rangle \leqslant 2^{2(j+1)}\kappa/n$ folgt aber (s.o.) $\|\hat{g}_n - g_{\vartheta_0}\|_n^2 \leqslant 2^{2(j+1)}\kappa/n$. Wir definieren

$$\Theta_j := \left\{\vartheta \in \Theta \,\middle|\, \|g_\vartheta - g_{\vartheta_0}\|_n \leqslant 2^{j+1}\sqrt{\kappa/n}\right\}$$

und erhalten so für $M_n > 0$ beliebig

$$\begin{aligned}\mathbb{P}(\|\hat{g}_n - g\|_n^2 - \|g_{\vartheta_0} - g\|_n^2 \geqslant \kappa/n,\, 2\|\varepsilon\|_n \leqslant M_n)\\ \leqslant \mathbb{P}\Big(\exists j = 0, \dots, J_n :\, \sup_{\vartheta\in\Theta_j} 2\langle \varepsilon, g_\vartheta - g_{\vartheta_0}\rangle_n \geqslant 2^{2j}\kappa/n\Big),\end{aligned}$$

wobei nur $2^{J_n+1}\sqrt{\kappa/n} \geqslant M_n$ gelten muss wegen $\|\hat{g}_n - g_{\vartheta_0}\|_n \leqslant 2\|\varepsilon\|_n$ (s.o.). Setzen wir weiterhin $\inf_\vartheta \frac{\|g_\vartheta - g_{\vartheta_0}\|_n}{|\vartheta - \vartheta_0|} \geqslant l_2 > 0$ voraus, so gilt $\rho(\Theta_j) \leqslant l_2^{-1} 2^{j+1}\sqrt{\kappa/n}$ und

$$\begin{aligned}\mathbb{P}(\|\hat{g}_n - g\|_n^2 - \|g_{\vartheta_0} - g\|_n^2 \geqslant \kappa/n,\, \|\varepsilon\|_n \leqslant M_n)\\ \leqslant \sum_{j=0}^{J_n} \mathbb{P}\Big(\sup_{\vartheta\in\Theta_j}(A_\vartheta - A_{\vartheta_0}) \geqslant l_2 2^{j-2}\sqrt{\kappa}\rho(\Theta_j)\Big).\end{aligned}$$

Wir schätzen also mittels gleichmäßiger Schranke für $j = 0, \dots, J_n$ weiter ab:

$$\begin{aligned}&\mathbb{P}(\|\hat{g}_n - g\|_n^2 - \|g_{\vartheta_0} - g\|_n^2 \geqslant \kappa/n,\, 2\|\varepsilon_n\| \leqslant M_n)\\ &\leqslant \sum_{j=0}^{J_n} C_{\delta,k} \exp\Big(-\frac{(\kappa l_2^2 2^{2j-4}) \wedge R_n^2}{(8+\delta)\sigma^2 L_2^2}\Big)\\ &\leqslant \tilde{C}_{\delta,k}\exp\Big(-\frac{\kappa}{(128+\delta)\sigma^2 L_2^2 l_2^{-2}}\Big) + J_n \exp\Big(-\frac{R_n^2}{(8+\delta)\sigma^2 L_2^2 l_2^{-2}}\Big).\end{aligned}$$

Außerdem gilt

$$\mathbb{P}(2\|\varepsilon\|_n > M_n) \leqslant \mathbb{P}(\exists i = 1, \dots, n :\, |\varepsilon_i| \geqslant M_n/2) \leqslant nC_\varepsilon e^{-M_n/(2D_\varepsilon)}.$$

Wähle nun $R_n = n^p$ mit $p = 1/8 < 1/6$ und $M_n = 2D_\varepsilon(\log(n) + R_n^2)$. Dann ist $J_n = O(\log(n))$ und wir erhalten insgesamt mit einer Konstanten $\tilde{c} > 0$.

$$\begin{aligned}&\mathbb{P}(\|\hat{g}_n - g\|_n^2 - \|g_{\vartheta_0} - g\|_n^2 \geqslant \kappa/n)\\ &\leqslant \tilde{C}_{\delta,k}\Big(\exp\Big(-\frac{\kappa}{(128+\delta)\sigma^2 L_2^2 l_2^{-2}}\Big) + e^{-\tilde{c}n^{1/4}}\Big).\end{aligned}$$

Wir haben also folgenden Satz bewiesen.

4.40 Satz. *Es sei $g_\vartheta \in C(D)$, $D \subseteq \mathbb{R}^d$, für $\vartheta \in \Theta \subseteq \mathbb{R}^k$ mit $\{g_\vartheta \,|\, \vartheta \in \Theta\}$ abgeschlossen bezüglich $\|\bullet\|_n$ und konvex gegeben. Betrachte den Kleinste-Quadrate-Schätzer $\hat{\vartheta}_n := \operatorname{argmin}_{\vartheta\in\Theta}\sum_{i=1}^n (Y_i - g_\vartheta(x_i))^2$ und $\hat{g}_n := g_{\hat{\vartheta}_n}$ im (falsch spezifizierten) Modell*

$$Y_i = g(x_i) + \varepsilon_i, \quad i = 1, \dots, n \text{ mit } x_i \in D,\, g \in C(D),$$

wobei (ε_i) *i.i.d. mit* $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ *eine Exponentialungleichung mit* $C_\varepsilon, D_\varepsilon > 0$ *und* $R = \infty$ *erfüllen möge. Dann gilt*

$$\|\hat{g}_n - g\|_n^2 = \inf_{\vartheta\in\Theta} \|g_\vartheta - g\|_n^2 + \frac{Z}{n}$$

mit einer Zufallsvariablen Z*, die die Exponentialungleichung*

$$\mathbb{P}(Z \geqslant \kappa) \leqslant C_k \Big(\exp\Big(- \frac{\kappa}{129\sigma^2 L_2^2 l_2^{-2}} \Big) + e^{-cn^{1/4}} \Big)$$

mit Konstanten $c, C_k > 0$ *erfüllt, sofern* $\frac{\|g_{\vartheta'} - g_\vartheta\|_n}{|\vartheta' - \vartheta|} \leqslant L_2$, $\frac{\|g_{\vartheta_0} - g_\vartheta\|_n}{|\vartheta_0 - \vartheta|} \geqslant l_2 > 0$ *für alle* $\vartheta, \vartheta' \in \Theta$ *gilt und* $\sup_{\vartheta,\vartheta'} \frac{\|g_{\vartheta'} - g_\vartheta\|_\infty}{|\vartheta' - \vartheta|}$ *endlich ist.*

4.41 Bemerkung. Im linearen und Gaußschen Regressionsmodell mit $g_\vartheta(x) = \sum_{j=1}^k \vartheta_j \varphi_j(x)$, $\vartheta \in \mathbb{R}^k$, und (φ_j) orthonormal bezüglich $\|\bullet\|_n$ ergibt eine direkte Rechnung $\hat{g}_n = \sum_{j=1}^k \langle Y, \varphi_j\rangle_n \varphi_j$ und

$$\|\hat{g}_n - g\|_n^2 = \inf_{\vartheta\in\Theta} \|g_\vartheta - g\|_n^2 + \sum_{j=1}^k \langle \varepsilon, \varphi_j\rangle^2.$$

Nun sind $(\langle \varepsilon, \varphi_j\rangle_n)_{j=1,\dots,k}$ unabhängige $N(0, \sigma^2/n)$-verteilte Zufallsvariablen, so dass $U := \frac{n}{\sigma^2} \sum_{j=1}^k \langle \varepsilon, \varphi_j\rangle^2$ $\chi^2(k)$-verteilt ist. Aus dem exponentiellen Moment $\mathbb{E}[e^{\alpha U}] = (1-2\alpha)^{-k/2}$ für $0 \leqslant \alpha = 1/2 - \delta < 1/2$ ergibt sich für $Z = \sigma^2 U$ die Ungleichung $\mathbb{P}(Z \geqslant \kappa) \leqslant (2\delta)^{-k/2} e^{(\frac{1}{2}-\delta)\kappa/\sigma^2}$, $\delta \in (0, 1/2]$. Wir sehen also, dass dies der Struktur unser allgemeinen Ungleichung entspricht, wir aber durch die Chaining- und Peeling-Techniken insbesondere den Faktor $128 + \delta$ im Nenner statt bloß $2 + \delta$ erhalten haben.

Es ergeben sich direkt zwei Korollare.

4.42 Korollar. *Das Exzess-Risiko erfüllt unter den Voraussetzungen des Satzes für jedes* $p > 0$ *mit einer Konstanten* $C_p > 0$

$$\mathbb{E}\Big[\Big| \|\hat{g}_n - g\|_n^2 - \inf_{\vartheta\in\Theta} \|g_\vartheta - g\|_n^2 \Big|^p \Big]^{1/p} \leqslant C_p/n.$$

Im Fall des korrekt spezifizierten Modells ergibt sich die bekannte n^{-1}*-Asymptotik, während im falsch spezifizierten Modell die Projektion* $g_{\vartheta_0} := \mathrm{argmin}_{g_\vartheta} \|g_\vartheta - g\|_n$ *mit dieser Rate geschätzt wird.*

4.43 Korollar. *Im korrekt spezifizierten Fall* $g = g_{\vartheta_0}$ *ist unter den Voraussetzungen des Satzes*

$$\{\vartheta \in \Theta \,|\, \|\hat{g}_n - g_\vartheta\|_n^2 \leqslant z_{1-\alpha}/n\}$$

mit $C_k(\exp(-\frac{z_{1-\alpha}}{129\sigma^2 L_2^2 l_2^{-2}}\Big) + e^{-cn^{1/4}}) \leqslant \alpha$ *ein nicht-asymptotischer Konfidenzbereich zum Niveau* $1 - \alpha$*.*

4.44 Bemerkungen.

(a) Um den Konfidenzbereich konkret auszurechnen, muss man die Konstanten c, C exakt nachvollziehen. Außerdem wird es ein sehr *konservativer* Konfidenzbereich sein, weil einige Abschätzungen eher grob waren wie der Vergleich mit dem linearen Gaußmodell zeigt. Trotzdem kann dies nützlich sein, insbesondere auch weil wir geringere Voraussetzungen (keine Differenzierbarkeit, keine Kompaktheit!) gestellt haben und nicht auf eine hinreichend gute Näherung durch die Asymptotik vertrauen müssen.

(b) Der Konfidenzbereich ist im Allgemeinen kein Intervall bzw. nichtzusammenhängend. Er reflektiert die Tatsache, dass die Parametrisierung prinzipiell unerheblich ist, weil nur die Familie $(g_\vartheta)_{\vartheta\in\Theta}$ für das Modell von Belang ist. Beachte in diesem Zusammenhang auch, dass eine Formulierung des Satzes mit einer konvexen, abgeschlossenen Familie $\mathscr{G} \subseteq C([0,1])$ natürlicher wäre. Identifiziert man $(C([0,1]), \|\bullet\|_n)$ mit dem $\mathbb{R}^n$, so kann man die Identität als Parametrisierung wählen, so dass $L_2 = l_2 = 1$ gilt. Allerdings fließt dann die Überdeckungszahl von Teilmengen $\Theta_j \subseteq \mathscr{G} \subseteq \mathbb{R}^n$ in die Vorfaktoren bei den gleichmäßigen Abschätzungen mit ein und diese sollte unabhängig von n sein (z.B. $\mathscr{G}$ ist k-dimensionale glatte Untermannigfaltigkeit).

Wir betrachten nun die Beobachtung einer mathematischen Stichprobe $X_1, \dots, X_n$, die gemäß einem parametrischen Modell $X_i \sim \mathbb{P}_\vartheta$, $\vartheta \in \Theta$, genügt, während in Wirklichkeit $X_i \sim \mathbb{P}$ i.i.d. gilt. Der Einfachheit halber nehmen wir an, dass $\mathbb{P}$ und $\mathbb{P}_\vartheta$ für alle $\vartheta \in \Theta$ äquivalent sind. Der parametrische Maximum-Likelihood-Schätzer $\hat\vartheta_n$ ist Minimum-Kontrast-Schätzer zum Kontrast $K_n(\vartheta) := -\frac{1}{n}\sum_{i=1}^n \ell(\vartheta, X_i)$, wobei wir als dominierendes Maß $\mathbb{P}$ wählen können, d.h. $\ell(\vartheta, x) := \log(\frac{d\,\mathbb{P}_\vartheta}{d\,\mathbb{P}}(x))$. Außerdem setzen wir $K(\vartheta_0, \vartheta) := \mathbb{E}[-\ell(\vartheta)] = \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta)$ mit $\vartheta_0 \in \Theta$ so, dass $\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) = \inf_{\vartheta\in\Theta} \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta)$ gilt (falls solch ein ϑ_0 nicht existiert, ersetze einfach in allen folgenden Ausdrücken $\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0})$ durch $\inf_{\vartheta\in\Theta} \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta)$).

Wir nehmen an, dass die Loglikelihoodfunktion für alle $\vartheta, \vartheta' \in \Theta$ mit Erwartungswert unter $\mathbb{P}$ die Lipschitz-Bedingung

$$|\ell(\vartheta, x) - \ell(\vartheta', x) - \mathbb{E}[\ell(\vartheta) - \ell(\vartheta')]| \leqslant L(x)|\vartheta - \vartheta'|$$

erfüllt, wobei die Zufallsvariable L folgender Exponentialungleichung genügt:

$$\mathbb{P}(|L| \geqslant r) \leqslant C_L e^{-r/D_L}, \quad r > 0.$$

Mit $A_\vartheta := \sqrt{n}(K_n(\vartheta) - K(\vartheta_0, \vartheta))$ gilt dann $\mathbb{E}[A_\vartheta] = 0$, $\mathrm{Var}(A_\vartheta - A_{\vartheta'}) \leqslant \mathbb{E}[L^2]|\vartheta - \vartheta'|^2$, und die Bernstein-Ungleichung zusammen mit Satz 4.39 liefert für $\{\vartheta_0\} \subseteq \Theta_0 \subseteq \Theta$ mit $\rho(\Theta_0) = \sup_{\vartheta\in\Theta_0}|\vartheta - \vartheta_0| < \infty$

$$\mathbb{P}\Big(\sup_{\vartheta\in\Theta_0} |A_\vartheta - A_{\vartheta'}| \geqslant r\rho(\Theta_0)\Big) \leqslant C_{\delta,k} \exp\Big(-\frac{r^2}{(8+\delta)\,\mathbb{E}[L^2]}\Big), \quad r \in [0, R_n],\ R_n = o(n^{1/6}).$$

Setze nun

$$\Theta_j := \Big\{\vartheta \in \Theta \,\Big|\, \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) \leqslant 2^{2(j+1)}\kappa/n\Big\}.$$

Peeling liefert dann für $M_n > 0$ beliebig

$$\begin{aligned}&\mathbb{P}(\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\hat{\vartheta}_n}) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) \in [\kappa/n, M_n^2])\\ &\leqslant \mathbb{P}\Big(\exists j = 0, \dots, J_n : \sup_{\vartheta \in \Theta_j} n^{-1/2}(A_\vartheta - A_{\vartheta'}) \geqslant 2^{2j}\kappa/n\Big)\end{aligned}$$

mit $2^{J_n+1}\sqrt{\kappa/n} \geqslant M_n$. Setzen wir weiterhin $\inf_\vartheta \frac{\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0})}{|\vartheta - \vartheta_0|^2} \geqslant l^2 > 0$ voraus, so gilt $\rho(\Theta_j) \leqslant l^{-1}2^{j+1}\sqrt{\kappa/n}$, und wir schätzen mittels gleichmäßiger Schranke für $j = 0, \dots, J_n$ weiter ab:

$$\begin{aligned}&\mathbb{P}(\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\hat{\vartheta}_n}) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) \in [\kappa/n, M_n^2])\\ &\leqslant \sum_{j=0}^{J_n} C_\delta \exp\Big(-\frac{(\kappa l^2 2^{2j-4}) \wedge R_n^2}{(8+\delta)\,\mathbb{E}[L^2]}\Big)\\ &\leqslant \tilde{C}_\delta \exp\Big(-\frac{\kappa}{(128+\delta)\,\mathbb{E}[L^2]l^{-2}}\Big) + J_n \exp\Big(-\frac{R_n^2}{(8+\delta)\,\mathbb{E}[L^2]}\Big).\end{aligned}$$

Außerdem gilt wegen der Lipschitztyp-Bedingungen an ℓ und KL:

$$\begin{aligned}\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\hat{\vartheta}_n}) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) &\leqslant \frac{1}{n}\sum_{i=1}^n L(X_i)|\hat{\vartheta}_n - \vartheta_0|\\ &\leqslant \frac{1}{n}\sum_{i=1}^n L(X_i) l^{-1} |\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\hat{\vartheta}_n}) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0})|^{1/2},\end{aligned}$$

so dass $\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\hat{\vartheta}_n}) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) \leqslant (\frac{1}{n}\sum_{i=1}^n L(X_i))^2 l^{-2}$ gilt. Wir erhalten also

$$\mathbb{P}\Big(\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\hat{\vartheta}_n}) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) > M_n^2\Big) \leqslant \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n L(X_i) l^{-1} > M_n\Big) \leqslant n C_L e^{-M_n l / D_L}.$$

Wähle $R_n = n^{1/8}$ und $M_n = l^{-1}D_L(\log(n) + R_n^2)$. Dann ist $J = O(\log(n))$ und wir haben folgenden Satz bewiesen.

4.45 Satz. *Es sei $X_1, \dots, X_n$ eine gemäß $\mathbb{P}$ verteilte mathematische Stichprobe. Betrachte den Maximum-Likelihood-Schätzer $\hat{\vartheta}_n$ unter dem Modell $X_i \sim \mathbb{P}_\vartheta$ i.i.d. mit $\vartheta \in \Theta \subseteq \mathbb{R}^k$, wobei $\mathbb{P}_\vartheta \sim \mathbb{P}$ für alle $\vartheta \in \Theta$ gelte. Weiterhin sei die Lipschitzbedingung*

$$|\ell(\vartheta, x) - \ell(\vartheta', x) - \mathbb{E}[\ell(\vartheta) - \ell(\vartheta')]| \leqslant L(x)|\vartheta - \vartheta'|, \quad \vartheta, \vartheta' \in \Theta,$$

erfüllt, wobei die Zufallsvariable L der Exponentialungleichung $\mathbb{P}(|L| \geqslant r) \leqslant C_L e^{-r/D_L}$, $r > 0$, genügt, und es möge $\inf_\vartheta \frac{\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta) - \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0})}{|\vartheta - \vartheta_0|^2} \geqslant l^2 > 0$ gelten. Dann erhalten wir

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\hat{\vartheta}_n}) = \inf_{\vartheta \in \Theta} \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta) + \frac{Z}{n}$$

mit einer Zufallsvariablen Z, die die Exponentialungleichung

$$\mathbb{P}(Z \geqslant \kappa) \leqslant C_k\Big(\exp\Big(-\frac{\kappa}{129\,\mathbb{E}[L^2]l^{-2}}\Big) + e^{-cn^{1/4}}\Big)$$

mit Konstanten $c, C_k > 0$ erfüllt.

4.46 Bemerkungen.

(a) Im Fall eines korrekt spezifizierten Modells gilt bei differenzierbarer Loglikelikelihood-Funktion die Lipschitzbedingung mit $L(x) := \sup_{\vartheta\in\Theta}|\dot\ell(\vartheta,x) - \mathbb{E}_{\vartheta_0}[\dot\ell(\vartheta)]|$.

(b) Wir haben gesehen, dass unter Regularitätsbedingungen im korrekt spezifizierten Modell $\mathbb{P} = \mathbb{P}_{\vartheta_0}$ der Fehler $\sqrt{n}I(\vartheta_0)^{1/2}(\hat\vartheta_n - \vartheta)$ asymptotisch $N(0, E_k)$-verteilt ist. Ein wichtiger Schritt dabei war ja die Approximation $\mathrm{KL}(\mathbb{P}_{\vartheta_0} \mid \mathbb{P}_{\hat\vartheta_n}) - \langle I(\vartheta_0)(\hat\vartheta_n - \vartheta_0), \hat\vartheta_n - \vartheta_0\rangle \xrightarrow{\mathbb{P}_{\vartheta_0}} 0$. Demnach ist in diesem Fall $Z = n\,\mathrm{KL}(\mathbb{P}_{\vartheta_0} \mid \mathbb{P}_{\hat\vartheta_n})$ asymptotisch $\chi^2(k)$-verteilt, vergleiche den Fall des Kleinste-Quadrate-Schätzers. In unserer jetzigen Exponentialungleichung und im skalaren Fall $k = 1$ gilt näherungsweise $L \approx |\dot\ell(\vartheta_0)|$ und somit $\mathbb{E}_{\vartheta_0}[L^2] \approx I(\vartheta_0)$. Zusammen mit $l^2 \approx I(\vartheta_0)$ aus der Kullback-Leibler-Approximation erhalten wir also näherungsweise eine Exponentialungleichung mit Exponenten $-\kappa/(128 + \delta)$, die wiederum bis auf den Faktor 128 (statt 2) die richtige Struktur besitzt.

(c) Auch diesmal ist der Faktor $\mathbb{E}[L^2]l^{-2}$ nicht ganz natürlich. Definiert man nämlich die Metrik $d(\vartheta,\vartheta') := |\mathrm{KL}(\mathbb{P}_\vartheta \mid \mathbb{P}) - \mathrm{KL}(\mathbb{P}_\vartheta \mid \mathbb{P})|^{1/2}$ und wachsen die Kardinalitäten von ε-Netzen von Θ_j bezüglich dieser Metrik nur polynomiell in ε^{-1} (Konzept allgemeiner Entropien!), so folgt eine entsprechende Ungleichung, allerdings mit $l = 1$ und mit Lipschitzkonstante L bezüglich d anstatt $|\vartheta - \vartheta'|$.

4.47 Beispiel. Es sei $(\mathbb{P}_\vartheta)$ eine natürliche Exponentialfamilie mit $\ell(\vartheta, x) = \langle\vartheta, T(x)\rangle - A(\vartheta)$. Existiert dann $\vartheta_0 := \mathrm{argmin}_{\vartheta\in\Theta}\,\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta)$ und liegt im Innern von Θ, so gilt $\nabla_\vartheta\,\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) = \mathbb{E}[T] - \dot A(\vartheta_0) = 0$, also $\mathbb{E}_{\vartheta_0}[T] = \mathbb{E}[T]$ nach Satz 3.10. Die Kullback-Leibler-Projektion von $\mathbb{P}$ auf die Exponentialfamilie ist also gerade diejenige Verteilung, unter der die Erwartungswerte der Statistiken $T_1, \dots, T_k$ mit denen unter $\mathbb{P}$ übereinstimmen. Weiterhin können wir

$$L(x) := \sup_{\vartheta\in\Theta}|\dot\ell(\vartheta,x) - \mathbb{E}[\dot\ell(\vartheta)]| = |T(x) - \mathbb{E}[T]|$$

wählen. Wir benötigen also unter $\mathbb{P}$ eine Exponentialungleichung für $T - \mathbb{E}[T]$. Beachte, dass diese für $\mathbb{P} = \mathbb{P}_{\vartheta_0}$ mit $\vartheta_0 \in \mathrm{int}(\Theta)$ aus

$$\mathbb{E}_{\vartheta_0}[e^{\langle\alpha,T\rangle}] = \mathbb{E}_{\vartheta_0}[L(\alpha)e^{A(\alpha)}] = e^{A(\alpha)} < \infty, \quad \alpha \in \mathbb{R}^k,\ |\alpha| < \mathrm{dist}(\vartheta_0, \partial\Theta)$$

folgt.

Weiterhin haben wir mit einer Zwischenstelle $\bar\vartheta$ von ϑ_0, ϑ

$$\mathbb{E}[\ell(\vartheta) - \ell(\vartheta_0)] = \langle\vartheta - \vartheta_0, \mathbb{E}[T]\rangle - \langle\dot A(\bar\vartheta), \vartheta - \vartheta\rangle = \langle\vartheta - \vartheta_0, \dot A(\vartheta_0) - \dot A(\bar\vartheta)\rangle.$$

Wir können also mit dem Mittelwertsatz auf $l \geqslant \inf_\vartheta \lambda_{min}(\ddot A(\vartheta))$ (λ_{min} =minimaler Eigenwert) schließen. Im Gaußschen Shift-Modell beispielsweise ist $l \geqslant \lambda_{min}(\Sigma)$.

Eine interessante Perspektive ergibt sich, wenn wir andersherum eine allgemeine Lebesguedichte $f : [0,1] \to \mathbb{R}$ mit $f > 0$ fast überall aus $X_1, \dots, X_n \sim f$

schätzen wollen. Wir wählen beschränkte Funktionen $(\varphi_j)_{j\geqslant 1}$, die eine Orthonormalbasis in $L^2([0,1])$ bilden, wie orthogonale (trigonometrische) Polynome, Splines, Wavelets. Dann gilt natürlich

$$f(x) = \exp\Big(\sum_{j\geqslant 1} f_j\varphi_j(x)\Big) \text{ mit } f_j = \langle \log f, \varphi_j\rangle_{L^2([0,1])}.$$

Machen wir nun den parametrischen Ansatz mit einer natürlichen Exponentialfamilie $(\mathbb{P}_\vartheta)_{\vartheta\in\mathbb{R}^k}$ in $\varphi_1,\dots,\varphi_k$ bezüglich Lebesguemaß, so schätzt der MLE wegen $f_j = \mathbb{E}[\varphi_j] = \mathbb{E}_{\vartheta_0}[\varphi_j]$ gerade das Modell $\mathbb{P}_{\vartheta_0}$ mit Dichte

$$f_k(\vartheta_0, x) = \exp\Big(\sum_{j=1}^k f_j\varphi_j(x) - A_k\Big),$$

indem wir die empirischen Momente bilden:

$$f_k(\hat\vartheta_n) = \exp\Big(\sum_{j=1}^k \Big(\frac{1}{n}\sum_{i=1}^n \varphi_j(X_i)\Big)\varphi_j(x) - \hat A_k\Big),$$

wobei $A_k, \hat A_k \in \mathbb{R}$ Normierungskonstanten sind. Als Kullback-Leibler-Divergenz ergibt sich gerade

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_{\vartheta_0}) = \mathbb{E}[\log(f) - \ell(\vartheta_0)] = \mathbb{E}\Big[\sum_{j>k} f_j\varphi_j\Big] + A_k = A_k + \sum_{j>k} f_j^2.$$

Wegen $\int f = 1$ gilt $A_k = \log(\int \exp(-\sum_{j>k} f_j\varphi_j))$, und der Bias gemessen in Kullback-Leibler-Divergenz ist klein, falls die Koeffizienten f_j für $j > k$ klein sind (z.B. für f glatt). Dieser Ansatz führt auf eine Maximum-Likelihood-Theorie der sogenannten nichtparametrischen Dichteschätzung, wobei die Dimension k des parametrischen Ansatzraums so gewählt wird, dass Approximationsfehler (Bias) und stochastischer Fehler (gemessen in Varianz oder Erwartungswert der KL-Divergenz) gleichgewichtet sind, vgl. Barron and Sheu (1991).

5 Testtheorie

5.1 Neyman-Pearson-Theorie

5.1 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.

5.2 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 $\mu < \mu_0$.

5.3 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.

5.4 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.

5.5 Bemerkung. Es gilt auch umgekehrt, dass jeder gleichmäßig beste Test für eine einfache Hypothese gegen eine einfache Alternative fast sicher die Form eines Neyman-Pearson-Tests besitzt (Übung!).

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 $\mathbb{P}_0(p_1 \geqslant kp_0) \geqslant \alpha$ und $\mathbb{P}_0(p_1 > kp_0) \leqslant \alpha$ (k ist $(1-\alpha)$-Quantil von p_1/p_0 unter $\mathbb{P}_0$). Dann besitzt mit $\gamma := (\alpha - \mathbb{P}_0(p_1 > kp_0))/\,\mathbb{P}_0(p_1 = kp_0)$ 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.

□

5.6 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 $a > 0$)

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

5.7 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. □

5.8 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:

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

5.9 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 $H_1 : \vartheta > \vartheta_0$.

5.10 Beispiel. Der einseitige Gauß-Test aus Beispiel 5.2 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 gilt

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

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 $\kappa = \frac{1-\mathbb{E}_{\vartheta_1}[\varphi^*]}{1-\mathbb{E}_{\vartheta_1}[\varphi]}$ ein besserer Test als φ^* zum Niveau $\mathbb{E}_{\vartheta_1}[\varphi^*]$. 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 $\vartheta_1 < \vartheta_0$.

□

5.11 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 $C := \{(\mathbb{E}_0[\varphi], \mathbb{E}_1[\varphi]) \,|\, \varphi \text{ Test}\} \subseteq [0,1]^2$. 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.

5.12 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.

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

5.14 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. und jeder Test besitze eine stetige Gütefunktion bei allen $\vartheta \in \partial\Theta_0$. Ist dann φ α-ähnlicher Test auf $\partial\Theta_0$, der besser ist als alle unverfälschten, α-ähnlichen Tests auf $\partial\Theta_0$, 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 $\varphi = \alpha$ ist. □

5.15 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 <u>zweiseitige Testproblem</u> $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$.

Beweis. Wir zeigen, dass φ^* für $\mathbb{P}_1 = \mathbb{P}_{\vartheta_1} \neq \mathbb{P}_0 = \mathbb{P}_{\vartheta_0}$ 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 $k, l \in \mathbb{R}$ 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 3.10) 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 obigem Lemma ist. □

5.16 Beispiel. Es sei $X_1, \ldots, X_n \sim N(\vartheta, \sigma^2)$ eine mathematische Stichprobe mit $\vartheta \in \mathbb{R}$ unbekannt und $\sigma > 0$ bekannt. Es liegt eine einparametrische Exponentialfamilie in $T(x) = \sum_{i=1}^n x_i$ und $\eta(\vartheta) = \vartheta/\sigma^2$ vor. Für $\vartheta_0 \in \mathbb{R}$ gilt $\eta'(\vartheta) = \sigma^{-2} > 0$, 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 $Z = \sum_{i=1}^n (X_i - \vartheta_0) \sim N(0, n\sigma^2)$ 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 $k = \sigma\sqrt{n}q_{1-\alpha/2}$ 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.

5.2 Bedingte Tests

5.17 Beispiel. In vielen Fällen sind bestimmte Parameter der Verteilung für einen Test nicht von Interesse, aber sie beeinflussen die Güte eines Tests (sogenannte *Störparameter* oder *nuisance-Parameter*). Als wichtiges Beispiel haben wir bereits den t-Test kennengelernt, wo eine Hypothese über den Mittelwert μ bei unbekannter Varianz im Normalverteilungsmodell $X_1, \ldots, X_n \sim N(\vartheta, \sigma^2)$ getestet wird. Eine weitere wichtige Klasse bilden sogenannte Mehrstichprobentests, wo nur das Verhältnis von Kennwerten (wie Mittelwert) zwischen den Stichproben getestet wird.

5.18 Lemma. *Ist T eine bezüglich Θ' vollständige und suffiziente Statistik und ist φ ein auf Θ' α-ähnlicher Test, so gilt $\mathbb{E}_\bullet[\varphi \,|\, T] = \alpha$ $\mathbb{P}_\vartheta$-f.s. für alle $\vartheta \in \Theta'$.*

Beweis. Aus Suffizienz und Vollständigkeit folgern wir jeweils für alle $\vartheta \in \Theta'$

$$\mathbb{E}_\vartheta[\varphi - \alpha] = 0 \Rightarrow \mathbb{E}_\vartheta[\mathbb{E}_\bullet[\varphi - \alpha \,|\, T]] = 0 \Rightarrow \mathbb{E}_\bullet[\varphi - \alpha \,|\, T] = 0 \quad \mathbb{P}_\vartheta\text{-f.s.}$$

□

5.19 Satz. *Gegeben sei die natürliche Exponentialfamilie*

$$\frac{d\,\mathbb{P}_{\vartheta,\tau}}{d\mu}(x) = C(\vartheta,\tau)\exp\Big(\vartheta U(x) + \langle \tau, T(x)\rangle\Big), \quad x \in \mathcal{X},\ (\vartheta,\tau) = (\vartheta,\tau_1,\dots,\tau_k) \in \Theta,$$

sowie $\alpha \in (0,1)$ und $\vartheta_0 \in \mathbb{R}$ mit $(\vartheta_0,\tau) \in int(\Theta)$ für ein $\tau \in \mathbb{R}^k$. Dann ist

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

mit $K(t) \in \mathbb{R}$, $\gamma(t) \in [0,1]$ derart, dass $\mathbb{E}_{\vartheta_0}[\varphi^* \,|\, T] = \alpha$ $\mathbb{P}_{\vartheta_0}$-f.s., ein gleichmäßig bester unverfälschter Test zum Niveau α von $H_0 : \vartheta \leqslant \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$.

5.20 Bemerkung. Der Beweis wird lehren, dass die bedingte Verteilung von φ^* gegeben T unter $\mathbb{P}_{\vartheta_0}$ nicht von den Störparametern τ abhängt, diese also für die Wahl von $K(t), \gamma(t)$ unerheblich sind.

Beweis. Ohne Einschränkung sei μ Wahrscheinlichkeitsmaß, sonst betrachte $\mu = \mathbb{P}_{\bar\vartheta}$ für ein $\bar\vartheta \in \Theta$. Dann besitzt nach dem Satz vom Bildmaß der Zufallsvektor (U,T) unter $\mathbb{P}_{\vartheta,\tau}$ die Dichte

$$\frac{d\,\mathbb{P}^{U,T}_{\vartheta,\tau}}{d\mu^{U,T}}(u,t) = C(\vartheta,\tau)\exp(\vartheta u + \langle \tau, t\rangle), \quad u \in \mathbb{R},\ t \in \mathbb{R}^k\,.$$

Die bedingte Dichte von $\mathbb{P}^{U|T=t}_{\vartheta,\tau}$ bezüglich $\mu^{U|T=t}$ ist daher gegeben durch

$$\frac{d\,\mathbb{P}^{U|T=t}_{\vartheta,\tau}}{d\mu^{U|T=t}}(u) = \frac{\exp(\vartheta u + \langle \tau, t\rangle)}{\int \exp(\vartheta v + \langle \tau, t\rangle)\mu(dv)} = C(\vartheta)e^{\vartheta u}, \quad u \in \mathbb{R},$$

insbesondere also unabhängig von τ. Unter der Bedingung $\{T = t\}$ ist also

$$\varphi^*(u,t) = \begin{cases} 1, & \text{falls } u > K(t) \\ 0, & \text{falls } u < K(t) \\ \gamma(t), & \text{falls } u = K(t) \end{cases}$$

mit $\mathbb{E}_{\vartheta_0}[\varphi^*(U,T) \,|\, T = t] = \alpha$ (beachte: bedingte Erwartung ist unabhängig von τ) ein gleichmäßig bester Test für $H_0 : \vartheta \leqslant \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$.

Betrachte $\partial\Theta_0 := \{(\vartheta,\tau) \in \Theta \,|\, \vartheta = \vartheta_0\}$, dessen Projektion $\Pi_k \partial\Theta_0$ auf die letzten k Koordinaten nichtleeres Inneres in $\mathbb{R}^k$ besitzt. In Exponentialfamilien

ist die Gütefunktion eines beliebigen Tests im Innern stetig (vergleiche Satz 3.10). Also reicht es nach Lemma 5.14, zu zeigen, dass φ^* gleichmäßig bester Test unter allen α-ähnlichen Tests φ auf $\partial\Theta_0$ ist. Da $\Pi_k \partial\Theta_0$ nichtleeres Inneres im $\mathbb{R}^k$ besitzt, ist T für $(\mathbb{P}_{\vartheta_0,\tau})_\tau$ suffiziente und vollständige Statistik, so dass Lemma 5.18 für diese Tests $\mathbb{E}_{\vartheta_0}[\varphi \,|\, T] = \alpha$ liefert. Da $\varphi^*(\bullet, t)$ die Güte der bedingten Tests mit $\mathbb{E}_{\vartheta_0}[\varphi \,|\, T = t] = \alpha$ nach dem Satz über beste einseitige Tests maximiert, folgt für $\vartheta > \vartheta_0$ und $\tau \in \mathbb{R}^k$ mit $(\vartheta, \tau) \in \Theta$

$$\mathbb{E}_{\vartheta,\tau}[\varphi^*(U,T)] = \mathbb{E}_{\vartheta,\tau}[\mathbb{E}_{\vartheta,\tau}[\varphi^*(U,T) \,|\, T]] \geqslant \mathbb{E}_{\vartheta,\tau}[\mathbb{E}_{\vartheta,\tau}[\varphi \,|\, T]] = \mathbb{E}_{\vartheta,\tau}[\varphi].$$

Als technische Feinheit bleibt die Frage der Produktmessbarkeit von $(u,t) \mapsto \varphi^*(u,t)$, die aus der expliziten Konstruktion von $K(t), \gamma(t)$ mittels rechtsstetiger Verteilungsfunktion folgt, siehe Lehmann/Romano, Seite 149, für Details. □

5.21 Beispiel. Zweistichproben-Poisson-Test: Es seien $X_1, \ldots, X_n \sim \text{Poiss}(a)$ und $Y_1, \ldots, Y_m \sim \text{Poiss}(b)$ zwei unabhängige mathematische Stichproben mit $a, b > 0$ unbekannt. Es soll die Hypothese $H_0 : a \leqslant b$ gegen die Alternative $H_1 : a > b$ getestet werden. Die Beobachtungen folgen einer Exponentialfamilie bezüglich dem Zählmaß μ_0 auf $\mathbb{N}_0^{m+n}$. Es gilt mit $x \in \mathbb{N}_0^n, y \in \mathbb{N}_0^m$

$$\frac{d\,\mathbb{P}_{a,b}}{d\mu_0}(x,y) = \frac{e^{-na-mb}}{(\prod_i x_i!)(\prod_j y_j!)} \exp\Big(\log(a/b) \sum_{i=1}^n x_i + \log(b)\Big(\sum_{i=1}^n x_i + \sum_{j=1}^m y_j\Big)\Big).$$

Wir setzen $\mu(dx,dy) = \frac{1}{(\prod_i x_i!)(\prod_j y_j!)} \mu_0(dx,dy)$ und erhalten mit obiger Notation $\vartheta = \log(a/b)$, $U(x,y) = \sum_i x_i$, $\tau_1 = \log(b)$, $T(x,y) = \sum_i x_i + \sum_j y_j$. Wir können also das reduzierte Testproblem $H_0 : \vartheta \leqslant 0$ gegen $H_1 : \vartheta > 0$ bei Beobachtung der suffizienten Statistiken U, T betrachten. Nach obigem Satz hat ein gleichmäßig bester unverfälschter Test die Form

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

mit $K(t)$ minimal so, dass $\mathbb{P}_0(U > K(t) \,|\, T = t) \leqslant \alpha$ gilt. Als bedingte Verteilung erhalten wir für $u, t \in \mathbb{N}_0$, $u \leqslant t$

$$\begin{aligned} \mathbb{P}_{a,b}(U = u | T = t) &= \frac{\mathbb{P}_{a,b}(U = u, T - U = t - u)}{\mathbb{P}_{a,b}(T = t)} = \frac{\frac{(an)^u}{u!} e^{-an} \frac{(bm)^{t-u}}{(t-u)!} e^{-bm}}{\sum_{i=0}^t \frac{(an)^i}{i!} e^{-an} \frac{(bm)^{t-i}}{(t-i)!} e^{-bm}} \\ &= \frac{\binom{t}{u}(an)^u (bm)^{t-u}}{\sum_{i=0}^t \binom{t}{i}(an)^i (bm)^{t-i}} = \binom{t}{u} \Big(\frac{an}{an+bm}\Big)^u \Big(\frac{bm}{an+bm}\Big)^{t-u}. \end{aligned}$$

Beachte, dass diese Verteilung in der Tat nur von $\vartheta = \log(a/b)$ und nicht von $\tau_1 = \log(b)$ abhängt. Im Fall $\vartheta = 0$, also $a = b$, vereinfacht sich dies zu

$$\mathbb{P}_\vartheta(U = u | T = t) = \binom{t}{u} p^u (1-p)^{t-u} = \text{Bin}_{t,p}(u) \text{ mit } p = \frac{n}{n+m}.$$

Für den möglichen Fall $T = 0$ gilt natürlich $U = 0$, und wir setzen $\mathrm{Bin}_{0,p}(0) := 1$. Ist also $b(1-\alpha, t, p)$ das $(1-\alpha)$-Quantil der $\mathrm{Bin}(t,p)$-Verteilung und $\gamma(t) \in [0,1]$ so gewählt, dass das Niveau ausgeschöpft wird, so ist

$$\varphi^* = \begin{cases} 1, & \text{falls } \sum_{i=1}^n X_i > b\Big(1-\alpha, \sum_i X_i + \sum_j Y_j, n/(n+m)\Big) \\ 0, & \text{falls } \sum_{i=1}^n X_i < b\Big(1-\alpha, \sum_i X_i + \sum_j Y_j, n/(n+m)\Big) \\ \gamma(T(x)), & \text{falls } \sum_{i=1}^n X_i = b\Big(1-\alpha, \sum_i X_i + \sum_j Y_j, n/(n+m)\Big) \end{cases}$$

gleichmäßig bester unverfälschter Test von $H_0 : a \leqslant b$ gegen $H_1 : a > b$ zum Niveau α.

Vergleiche dies mit einem ad-hoc-Test der Form $\varphi = \mathbf{1}(\frac{1}{n}\sum_i X_i > k\frac{1}{m}\sum_j Y_j)$, der zwei suffiziente Statistiken vergleicht. Das Quantil $b(1-\alpha,t,p)$ wächst monoton in t für $\alpha \leqslant 1/2$, so dass für diesen Fall durch monotone Transformationen auch φ^* in der Form $\mathbf{1}(\frac{1}{n}\sum_i X_i > k\frac{1}{m}\sum_j Y_j)$ (plus Randomisierung) dargestellt werden kann.

Mit den gleichen Argumenten ergibt sich auch das folgende Resultat für zweiseitige Tests.

5.22 Satz. *Es liege die Situation des vorigen Satzes vor. Dann ist*

$$\varphi^*(x) = \begin{cases} 1, & \textit{falls } U(x) < K_1(T(x)) \textit{ oder } U(x) > K_2(T(x)) \\ 0, & \textit{falls } U(x) \in (K_1(T(x)), K_2(T(x))) \\ \gamma_i(T(x)), & \textit{falls } U(x) = K_i(T(x)),\ i = 1,2, \end{cases}$$

mit $K_i(t) \in \mathbb{R}$, $\gamma_i(t) \in [0,1]$ derart, dass

$$\mathbb{E}_{\vartheta_0}[\varphi^* \,|\, T] = \alpha \textit{ und } \mathbb{E}_{\vartheta_0}[U\varphi^* \,|\, T] = \alpha\, \mathbb{E}_{\vartheta_0}[U \,|\, T] \quad \mathbb{P}_{\vartheta_0}\textit{-f.s.}$$

ein gleichmäßig bester unverfälschter Test zum Niveau α von $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$.

Für Anwendungen erleichtert das folgende Resultat häufig die Bestimmung der bedingten kritischen Werte.

5.23 Satz (Basu)**.** *Es seien T und V Statistiken auf einem statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$. Ist T suffizient und vollständig sowie V ancillary, d.h. $\mathbb{P}_\vartheta^V$ ist unabhängig von $\vartheta \in \Theta$, so sind T und V unabhängig.*

Beweis. Betrachte $\mathbb{E}_\vartheta[\varphi(V)]$ für messbare $[0,1]$-wertige φ. Dann ist $a := \mathbb{E}_\vartheta[\varphi(V)]$ unabhängig von ϑ wegen V *ancillary* und gemäß Lemma 5.18 folgt $\mathbb{E}_\bullet[\varphi(V) \,|\, T] = a$ $\mathbb{P}_\vartheta$-f.s. für alle $\vartheta \in \Theta$. Dies impliziert Unabhängigkeit mittels $\mathbb{E}_\vartheta[\varphi(V)\psi(T)] = \mathbb{E}_\vartheta[a\psi(T)] = \mathbb{E}_\vartheta[\varphi(V)]\,\mathbb{E}_\vartheta[\psi(T)]$ und Einsetzen von Indikatorfunktion $\varphi = \mathbf{1}_B$, $\psi = \mathbf{1}_C$. □

5.24 Korollar. *In der Situation von Satz 5.19 ist eine Statistik V unabhängig von T unter jedem $\mathbb{P}_{\vartheta_0,\tau}$, wenn die Verteilung von V nicht von τ abhängt.*

Beweis. Dies folgt aus dem Satz von Basu und der Suffizient und Vollständigkeit von T für $\mathbb{P}_{\vartheta_0,\tau}$ auf $\partial\Theta_0$ (vgl. obiger Beweis). □

5.25 Beispiele.

(a) Betrachte eine mathematische Stichprobe $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ mit $\mu \in \mathbb{R}$, $\sigma > 0$ unbekannt. Wir erhalten eine natürliche Exponentialfamilie in $U(x) = \bar{x}$, $T(x) = \sum_{i=1}^n x_i^2$ mit $\vartheta = n\mu/\sigma^2$, $\tau_1 = -1/(2\sigma^2)$. Teste auf den Mittelwert $H_0 : \mu = 0$ gegen $H_1 : \mu \neq 0$ (für allgemeines $H_0 : \mu = \mu_0$ verschiebe entsprechend), d.h. $H_0 : \vartheta = 0$ gegen $H_1 : \vartheta \neq 0$. Betrachte daher $\varphi^*(u,t) := \mathbf{1}(u \notin [K_1(t), K_2(t)])$ mit $\mathbb{E}_{\vartheta=0}[\varphi^*(U,T) \,|\, T] = \alpha$, $\mathbb{E}_{\vartheta=0}[U\varphi^*(U,T) \,|\, T] = \alpha\, \mathbb{E}_{\vartheta=0}[U|T]$. Um diese bedingten Erwartungen auszurechnen, ist es einfacher, $V = U/\sqrt{T}$ zu betrachten. Für $\vartheta = \mu = 0$ ist die Verteilung von V unabhängig von $\tau_1 = -1/(2\sigma^2)$, so dass mit Basus Satz die Unabhängigkeit von V und T folgt und somit

$$\begin{aligned}\alpha &= \mathbb{P}_{\vartheta=0}(U \notin [K_1(T), K_2(T)] \,|\, T = t) \\ &= \mathbb{P}_{\vartheta=0}(V \notin [\sqrt{t}K_1(t), \sqrt{t}K_2(t)] \,|\, T = t) \\ &= \mathbb{P}_{\vartheta=0}(V \notin [\tilde{K}_1, \tilde{K}_2]).\end{aligned}$$

Entsprechend erhalten wir aus der symmetrischen Verteilung von V die Bedingung

$$\mathbb{E}_{\vartheta=0}[V\sqrt{t}\mathbf{1}(V \notin [\tilde{K}_1, \tilde{K}_2])] = \alpha\, \mathbb{E}_{\vartheta=0}[V\sqrt{T}|T = t] = 0.$$

Wegen der Verteilungssymmetrie von V wähle $\tilde{K}_2 = -\tilde{K}_1$, womit die zweite Bedingung erfüllt ist. Für die erste wähle $\tilde{K}$ so, dass $\mathbb{P}_{\vartheta=0}(|V| > \tilde{K}) = \alpha$ gilt. Da $|Z|$ mit $Z = \frac{\sqrt{n(n-1)}V}{\sqrt{1-nV^2}} \sim t_{n-1}$ (vereinfache und vergleiche mit Korollar 1.16) monoton in $|V|$ wächst, ist φ^* gerade der zweiseitige t-Test $\varphi^* = \mathbf{1}(|Z| > q_{t(n-1);1-\alpha/2})$, s.o. Genauso ergibt sich der einseitige t-Test als gleichmäßig bester unverfälschter Test für $H_0 : \mu \leqslant \mu_0$ gegen $H_1 : \mu > \mu_0$.

(b) Betrachte den Fall zweier unabhängiger mathematischer Stichproben $X_1, \dots, X_m \sim N(\mu, \sigma^2)$, $Y_1, \dots, Y_n \sim N(\nu, \tau^2)$ mit $\mu, \nu \in \mathbb{R}$, $\sigma, \tau > 0$ unbekannt. Es soll $H_0 : \mu = \nu$ gegen $H_1 : \mu \neq \nu$ getestet werden. Im Fall $\sigma \neq \tau$ gibt es bislang keine befriedigende Lösung (Behrens-Fisher-Problem, 1935). Hier betrachte daher den Fall $\tau = \sigma$ mit Lebesguedichte

$$f_{\mu,\nu,\sigma}(x,y) = C(\mu,\nu,\sigma) \exp\Big(-\frac{1}{2\sigma^2}\Big(\sum_i x_i^2 + \sum_j y_j^2\Big) + \frac{m\mu}{\sigma^2}\bar{x} + \frac{n\nu}{\sigma^2}\bar{y}\Big),$$

so dass eine Exponentialfamilie in $U(x,y) = \bar{y} - \bar{x}$ mit $\vartheta = (\nu - \mu)/(\sigma^2(m^{-1}+n^{-1}))$, $T_1(x,y) = m\bar{x} + n\bar{y}$ mit $\tau_1 = (m\mu + n\nu)/(\sigma^2(m+n))$ und $T_2(x,y) = \sum_i x_i^2 + \sum_j y_j^2$ mit $\tau_2 = -1/(2\sigma^2)$ vorliegt. Übrigens ist gerade diese Darstellung mit $\vartheta \propto \nu - \mu$ und $U \propto \bar{Y} - \bar{X}$ im Behrens-Fisher-Problem nicht mehr gegeben. Um $H_0 : \vartheta = 0$ gegen $H_1 : \vartheta \neq 0$ zu testen, wähle wieder eine Teststatistik V, unabhängig von T unter $\mathbb{P}_{\vartheta_0,\tau}$. In der Tat hängt die Verteilung von

$$V(X,Y) := \frac{\bar{Y} - \bar{X}}{\sqrt{\sum_i (X_i - \bar{X})^2 + \sum_j (Y_j - \bar{Y})^2}}$$

für $\vartheta = 0$ ($\mu = \nu$) nicht von τ, also von μ und σ^2, ab. Nach dem Satz von Basu sind also V und T unabhängig. Außerdem gilt $V = U/\sqrt{T_2 - T_1^2/(m+n) - mnU^2/(m+n)}$ und $|V|$ wächst streng monoton in $|U|$. Daher ist ein gleichmäßig bester Test $\varphi^* = \mathbf{1}(U \notin [-K(T), K(T)])$ ($K_1(t) = -K_2(t)$ wegen Symmetrie bzw. aus zweiter Nebenbedingung) auch als $\varphi^* = \mathbf{1}(V \notin [-\tilde{K}(T), \tilde{K}(t)])$ darstellbar. Wegen der Unabhängigkeit von T und V für $\vartheta = 0$ ist $\tilde{K}(T) = \tilde{K}$ unabhängig von T zu wählen. Nun ist $Z := \sqrt{(m+n-2)/(m^{-1}+n^{-1})}V$ für $\vartheta = 0$ der Quotient des $N(0,1)$-verteilten Zählers $(\bar{Y} - \bar{X})/\sqrt{\sigma^2(m^{-1}+n^{-1})}$ und des Nenners $\sqrt{(\sum_i (X_i - \bar{X})^2 + \sum_j (Y_j - \bar{Y})^2)/(\sigma^2(m+n-2))}$, dessen Quadrat, mit $m+n-2$ multipliziert, $\chi^2(m+n-2)$-verteilt für $\vartheta = 0$ ist. Wie im Einstichprobenfall sind auch hier Zähler und Nenner unabhängig, so dass Z $t(n+m-2)$-verteilt ist. Als gleichmäßig besten unverfälschten Test erhalten wir also den Zweistichproben-t-Test $\varphi^* = \mathbf{1}(|V(X,Y)| > \sqrt{(m^{-1}+n^{-1})/(m+n-2)}q_{t(n+m-2);1-\alpha/2})$ auf $H_0 : \mu = \nu$ gegen $H_1 : \mu \neq \nu$.

5.3 Likelihood-Quotienten- und χ^2-Test

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

5.26 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.

5.27 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.

5.28 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 3.10 und daher nach Lemma 4.17

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

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

5.29 Lemma. *In der Situation vom Satz 5.9 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. □

5.30 Bemerkung. Im Fall des zweiseitigen Testproblems aus Satz 5.15 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.

5.31 Satz. *Es mögen die Voraussetzungen aus Satz 4.27 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 $\varphi(x)=\mathbf{1}(\lambda_n(x)>\frac{1}{2}q_{\chi^2(k-r),1-\alpha})$ 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 4.27 haben wir für den MLE $\hat{\vartheta}_n$ insbesondere gezeigt, dass mit $K_n(\vartheta,x)=-\frac{1}{n}\sum_{i=1}\ell(\vartheta,x_i)$ für n hinreichend groß gilt

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

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 $\mathbb{P}_{\vartheta_0}$-Wahrscheinlichkeit. Bezeichnet $o_P(1)$ Terme, die stochastisch gegen Null konvergieren, so folgt

$$K_n(\vartheta_0)-K_n(\hat{\vartheta}_n)=\frac{1}{2}\langle I(\vartheta_0)^{-1}\dot{K}_n(\vartheta_0),\dot{K}_n(\vartheta_0)\rangle+o_P(1).$$

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,

$$K_n(\vartheta_0) - K_n(\hat{\vartheta}_n^0) = \frac{1}{2}\langle I^0(\vartheta_0)^{-1}\dot{K}_n^0(\vartheta_0), \dot{K}_n^0(\vartheta_0)\rangle + o_P(1),$$

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

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

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 $k - r$. Dies impliziert (vergleiche die Beweise im linearen Modell), dass $\langle (E_k - M)Z, Z\rangle$ $\chi^2(k-r)$-verteilt ist.

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. □

5.32 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 4.27 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.

5.33 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 $\chi^2(k-1)$.

5.34 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:

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

5.36 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.

5.37 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.