

Mathematische Statistik
Gliederung zur Vorlesung
im Sommersemester 2010

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

VORLÄUFIGE FASSUNG: 3. August 2010

Inhaltsverzeichnis

1	Statistik im linearen Modell	1
1.1	Einführendes Beispiel	1
1.2	Lineares Modell und kleinste Quadrate	1
1.3	Der Satz von Gauß-Markov	3
1.4	Inferenz unter Normalverteilungsannahme	5
2	Entscheidungstheorie	7
2.1	Formalisierung eines statistischen Problems	7
2.2	Minimax- und Bayes-Ansatz	9
2.3	Das Stein-Phänomen	15
2.4	Ergänzungen	17
3	Dominierte Modelle und Suffizienz	18
3.1	Dominierte Modelle	18
3.2	Exponentialfamilien	19
3.3	Suffizienz	21
3.4	Vollständigkeit	24
3.5	Cramér-Rao-Effizienz	25
4	Allgemeine Schätztheorie	28
4.1	Momentenschätzer	28
4.2	Maximum-Likelihood- und Minimum-Kontrast-Schätzer	30
4.3	Asymptotik	33
4.4	Allgemeine Schranken	38
4.5	Anwendung auf Regression und Maximum-Likelihood	41
5	Testtheorie	48
5.1	Neyman-Pearson-Theorie	48
5.2	Bedingte Tests	53
5.3	Likelihood-Quotienten- und χ^2 -Test	58

1 Statistik im linearen Modell

1.1 Einführendes Beispiel

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

1.2 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 $\text{Var}(\varepsilon_i) = \sigma^2 > 0$, 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:

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

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

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

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 k -dimensionalem Parameter $\beta \in \mathbb{R}^k$, $k < n$, besteht aus einer reellen Matrix $X \in \mathbb{R}^{n \times k}$ von vollem Rang k , der Designmatrix, und einem Zufallsvektor $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$, den Fehler- oder Störgrößen, mit $\mathbb{E}[\varepsilon_i] = 0$, $\text{Cov}(\varepsilon_i \varepsilon_j) = \Sigma_{ij}$ 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:

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

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:

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

1.3 Bemerkung. Wir schreiben $\Sigma > 0$, falls Σ eine symmetrische, strikt positiv-definite Matrix ist. Dann ist Σ diagonalisierbar mit $\Sigma = TDT^\top$, $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ Diagonalmatrix und T orthogonale Matrix, und wir setzen $\Sigma^{-1/2} = TD^{-1/2}T^\top$ mit $D^{-1/2} = \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_n^{-1/2})$. 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 $k = 1$, $\beta = \mu$ und $X = (1, \dots, 1)^\top$ führt.
- (b) Lineare Regression: $k = 2$, $\beta = (b, a)^\top$, $X = (X_{ij})$ 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

$$Y_i = a_0 + a_1x_i + a_2x_i^2 + \dots + a_{k-1}x_i^{k-1} + \varepsilon_i, \quad i = 1, \dots, n.$$

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

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

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

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

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

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

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

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 $Y_i \sim \text{Bin}(n_i, p_i)$ beobachtet werden mit der sogenannten *logit*-Linkfunktion $\log(p_i/(1-p_i))_{i=1,\dots,n} = X\beta$, so dass

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

1.3 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 $\text{ran}(X_\Sigma)$ 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 $w \in \text{ran}(X_\Sigma)$ 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 $\text{ran}(X_\Sigma)$ 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 $\hat{\beta} = \text{argmin}_b |\Sigma^{-1/2}(Y - Xb)|^2$ folgt, dass $\hat{\beta}$ die beste Approximation 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 erwartungstreu 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 $v \in \mathbb{R}^k$ 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 $\text{Var}(\hat{\gamma}) = \sigma^2 |X(X^\top X)^{-1}v|^2$.*
- (c) *Bezeichnet $R := Y - X\hat{\beta}$ den Vektor der Residuen, so ist die geeignet normalisierte Stichprobenvarianz*

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

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 $\beta \in \mathbb{R}^k$

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

Nach Pythagoras erhalten wir

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

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

- (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, \dots, e_{n-k} eine Orthonormalbasis vom $(n-k)$ -dimensionalen Bild $\text{ran}(E_n - \Pi_X) \subseteq \mathbb{R}^n$, so folgt

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

was die Behauptung impliziert. □

1.9 Bemerkung. Man sagt, dass der Schätzer $\hat{\gamma}$ im Satz von Gauß-Markov bester linearer erwartungstreuer Schätzer (BLUE: best linear unbiased estimator) ist. Eine entsprechende Aussage des Satzes von Gauß-Markov gilt auch im allgemeinen linearen Modell (Übung!).

1.4 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

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

Dabei ist der Wert 1,96 gerade das 0,975-Quantil bzw. 0,025-Fraktile 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 $|\hat{\gamma} - \gamma_0| \leq k_{\alpha/2} \sigma \sqrt{v^\top (X^\top X)^{-1} v}$ gilt mit dem $(\alpha/2)$ -Fraktile $k_{\alpha/2}$ 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 Lebesguegedichte

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

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 Lebesguegedichte

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

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 $f_{|T_n|}(x) = f_{F_{1,n}}(x^2)2x$, $x \geq 0$, 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 $w = x/y$)

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

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 $\frac{m}{n} f_{X/Y}(\frac{m}{n}x) = f_{m,n}(x)$. \square

1.13 Bemerkung. Für $n = 1$ ist die $t(n)$ -Verteilung gerade die Cauchy-Verteilung und für $n \rightarrow \infty$ konvergiert sie schwach gegen die Standardnormalverteilung (Slutsky-Lemma!). Für jedes $n \in \mathbb{N}$ besitzt $t(n)$ nur Momente 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 $n \rightarrow \infty$ 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 $q_{F(k, n-k); 1-\alpha}$ das $(1-\alpha)$ -Quantil der $F(k, n-k)$ -Verteilung, so ist*

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

ein Konfidenzellipsoid zum Irrtumsniveau α für β .

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

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

ein Konfidenzintervall zum Irrtumsniveau α 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 $k = 1$, $X = (1, \dots, 1)^\top$ und $\gamma = \beta$, $v = 1$ ist. \square

Beweis des Satzes. Allgemein müssen wir jeweils für einen Konfidenzbereich B für den vom wahren Parameter β abgeleiteten Parameter ϑ_β zum Irrtumsniveau α nachweisen, dass gilt

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

Im folgenden werden wir sogar Gleichheit erhalten.

- (a) $X(\hat{\beta} - \beta) = \Pi_X \varepsilon$ ist $N(0, \sigma^2 \Pi_X^\top \Pi_X)$ -verteilt und somit ist $\sigma^{-2} |X(\hat{\beta} - \beta)|^2 \chi^2(k)$ -verteilt. Weiterhin gilt ja $\hat{\sigma}^2 = \frac{|Y - \Pi_X Y|^2}{n-k} = \frac{|\varepsilon - \Pi_X \varepsilon|^2}{n-k}$, 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 $\frac{n-k}{\hat{\sigma}^2} \hat{\sigma}^2 \chi^2(n-k)$ -verteilt ist. Wie in Lemma 1.12 schließen wir, dass $|X(\hat{\beta} - \beta)|^2 / (k \hat{\sigma}^2) F(k, n-k)$ -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 $t(n-k)$ -verteilt ist und die Behauptung folgt. \square

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 $\mathbb{P}_{\mu_0}(\mu_0 \notin I) \leq \alpha$ impliziert und $\mu_0 \notin I \iff |\hat{\mu} - \mu_0| > \hat{\sigma}n^{-1/2}q_{t(n-1);1-\alpha/2}$ gilt. \square

2 Entscheidungstheorie

2.1 Formalisierung eines statistischen Problems

2.1 Definition. Ein Messraum $(\mathcal{X}, \mathcal{F})$ versehen mit einer Familie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ von Wahrscheinlichkeitsmaßen, $\Theta \neq \emptyset$ beliebige Parametermenge, heißt

statistisches Experiment oder statistisches Modell. \mathcal{X} heißt Stichprobenraum. Jede $(\mathcal{F}, \mathcal{S})$ -messbare Funktion $Y : \mathcal{X} \rightarrow S$ heißt Beobachtung oder Statistik mit Werten in (S, \mathcal{S}) und induziert das statistische Modell $(S, \mathcal{S}, (\mathbb{P}_\vartheta^Y)_{\vartheta \in \Theta})$. Sind die Beobachtungen Y_1, \dots, Y_n für jedes \mathbb{P}_ϑ unabhängig und identisch verteilt, so nennt man Y_1, \dots, Y_n eine mathematische Stichprobe.

2.2 Definition. Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell. Eine Entscheidungsregel ist eine messbare Abbildung $\rho : \mathcal{X} \rightarrow A$, wobei der Messraum (A, \mathcal{A}) der sogenannte Aktionsraum ist. Jede Funktion $l : \Theta \times A \rightarrow [0, \infty) =: \mathbb{R}^+$, 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 $\Theta = \mathbb{R}^k \times \mathbb{R}_+$ mit Parametern $\vartheta = (\beta, \sigma) \in \Theta$. Nun wähle einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{G}, \mathbb{P})$, auf dem der Zufallsvektor $\varepsilon : \Omega \rightarrow \mathbb{R}^n$ 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 $\mathcal{F} = \mathcal{B}_{\mathbb{R}^n}$ 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 $A = \mathbb{R}^k$ 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}_ϑ 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}_ϑ) die Familie $\mathcal{P} = \{\mathbb{P} \text{ W-Maß auf } \mathcal{F} \mid \mathbb{E}_{\mathbb{P}}[Y] := \int y \mathbb{P}(dy) \in \text{ran}(X), \int (y - \mathbb{E}_{\mathbb{P}}[Y])(y - \mathbb{E}_{\mathbb{P}}[Y])^\top \mathbb{P}(dy) = \sigma^2 E_n\}$ betrachten. In dieser Betrachtungsweise bleibt von einem unendlich-dimensionalen Parameterraum maximal ein $(k+1)$ -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 $\mathcal{X} = \{0, 1, \dots, 100\}$ mit der Potenzmenge als σ -Algebra und $\mathbb{P}_p = \text{Bin}(100, p)$, $p \in \Theta = [0, 1]$, als mögliche Verteilungen. Die Nullhypothese ist $H_0 : p \leq p_0$ für den unbekannt Parameter p . Als Aktionsraum dient

$A = \{0, 1\}$ (H_0 annehmen bzw. verwerfen), und wir wählen den Verlust $l(p, a) = \ell_0 \mathbf{1}_{\{p \leq p_0, a=1\}} + \ell_1 \mathbf{1}_{\{p > p_0, a=0\}}$ mit Konstanten $\ell_0, \ell_1 \geq 0$. Dies führt auf das Risiko einer Entscheidungsregel (eines Tests) ρ

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

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

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 \mathcal{F}_Θ , die Verlustfunktion l sei produktmessbar und $\vartheta \mapsto \mathbb{P}_\vartheta(B)$ sei messbar für alle $B \in \mathcal{F}$. Die a-priori-Verteilung π des Parameters ϑ ist gegeben durch ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathcal{F}_\Theta)$. 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 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: Definiere $\Omega := \mathcal{X} \times \Theta$ und $\tilde{\mathbb{P}}$ auf $(\Omega, \mathcal{F} \otimes \mathcal{F}_\Theta)$ gemäß $\tilde{\mathbb{P}}(dx, d\vartheta) = P_\vartheta(dx)\pi(d\vartheta)$ (gemeinsame Verteilung von Beobachtung und Parameter). Bezeichne mit X und T die Koordinatenprojektionen von Ω auf \mathcal{X} bzw. Θ . Dann gilt $R_\pi(\rho) = \mathbb{E}_{\tilde{\mathbb{P}}}[l(T, \rho(X))]$.

2.10 Definition. Es sei X eine (S, \mathcal{S}) -wertige Zufallsvariable auf $(\Omega, \mathcal{F}, \mathbb{P})$. Eine Abbildung $K : S \times \mathcal{F} \rightarrow [0, 1]$ 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 $A \in \mathcal{F}$;
- (c) $K(X, A) = \mathbb{P}(A | X) := \mathbb{E}[\mathbf{1}_A | X]$ \mathbb{P} -f.s. für alle $A \in \mathcal{F}$.

2.11 Satz. *Es sei (Ω, d) ein vollständiger, separabler Raum mit Metrik d und Borel- σ -Algebra \mathcal{F} (polnischer Raum). Für jede Zufallsvariable X auf $(\Omega, \mathcal{F}, \mathbb{P})$ 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 $\mathbb{P}(\forall A \in \mathcal{F} : K(X, A) = K'(X, A)) = 1$.*

Beweis. Siehe z.B. Gänsler, Stute (1977): Wahrscheinlichkeitstheorie, Springer. \square

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

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

$$f_{T|X=x}(\vartheta) = \frac{f_{X|T=\vartheta} f_T(\vartheta)}{\int_{\Theta} f_{X|T=\vartheta'} f_T(\vartheta') \nu(d\vartheta')} \quad (\text{Bayesformel}).$$

Beweis. Übung! \square

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 $(\mathcal{X}, \mathcal{F})$ mögen die Dichten p_0, p_1 bezüglich einem Maß μ besitzen (z.B. $\mu = \mathbb{P}_0 + \mathbb{P}_1$). Nach der Bayesformel (mit Zählmaß ν) erhalten wir die a-posteriori-Verteilung

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

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

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

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

Beweis. Für eine beliebige Entscheidungsregel ρ' gilt

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

□

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 $\hat{\vartheta}_\pi := \mathbb{E}_{\tilde{\mathbb{P}}} [T | X]$ 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. $\tilde{\mathbb{P}}(T \leq \hat{\vartheta}_\pi | X) \geq 1/2$ und $\tilde{\mathbb{P}}(T \geq \hat{\vartheta}_\pi | X) \geq 1/2$, 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 oder II. □

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

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

Daher ist ein Bayestest (Bayesklassifizierer) gegeben durch

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

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

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 $R_\pi(\rho) = \int_\Theta R(\vartheta, \rho) \pi(d\vartheta) \leq \sup_{\vartheta \in \Theta} R(\vartheta, \rho)$. 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 \mathcal{F}_Θ . 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 $R_\pi(\rho') \leq R_\pi(\rho)$ 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 (μ, X_1, \dots, X_n) gemeinsam normalverteilt und die bedingte Erwartung ist linear-affin (vgl. Stochastik II): $\mathbb{E}_{\mathbb{P}}[\mu_j | X_1, \dots, X_n] = \sum_{i=1}^n \alpha_{ij} X_i + \beta_j$, $j = 1, \dots, d$. 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 $\hat{\mu}_{\sigma,n} = \frac{n}{n + \sigma^{-2}} \bar{X}$ (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:

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

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

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 $R(\mu, \hat{\mu}) \leq \frac{1}{n}$ 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 $R(\mu, \hat{\mu}) \leq \frac{1}{n} - \varepsilon$ für alle $|\mu - \mu_0| < \delta$ mit $\varepsilon, \delta > 0$ geeignet. Damit hätte $\hat{\mu}$ ein Bayesrisiko $R_{N(0, \sigma^2)}(\hat{\mu}) \leq \frac{1}{n} - \varepsilon \int_{\mu_0 - \delta}^{\mu_0 + \delta} \varphi_{0, \sigma^2}$. Also wäre $\frac{1}{n} - R_{N(0, \sigma^2)}$ größer als ein Vielfaches von σ^{-1} für $\sigma \rightarrow \infty$, 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. \square

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 $n \geq 3$), vergleiche Lehmann/Casella, Seite 153. Für $d = 2$ ist \bar{X} weiterhin zulässig unter Normalverteilungsannahme, allerdings gilt das für $d \geq 3$ nicht mehr: Stein-Phänomen s.u.

2.24 Definition. Eine Verteilung π auf $(\Theta, \mathcal{F}_{\Theta})$ 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 Sattelpunkteigenschaft*

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

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 Sattelpunkteigenschaft $R_{\pi}(\rho_{\pi}) \geq \sup_{\vartheta} R(\vartheta, \rho_{\pi})$. Da aus dem gleichen Grund stets ' \leq ' 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 $R_{\pi}(\rho_{\pi}) \leq R_{\pi}(\rho')$ für alle ρ' gilt. Mit $R_{\pi}(\rho_{\pi}) = \sup_{\vartheta \in \Theta} R(\vartheta, \rho_{\pi})$ schließen wir dann auch

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

Aus der Sattelpunkteigenschaft folgt direkt die Minimaxeigenschaft:

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

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

2.26 Beispiel. Es werde $X \sim \text{Bin}(n, p)$ mit $n \geq 1$ 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 $a, b > 0$ 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 Sattelpunkteigenschaft folgt:

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

Damit ist $B(a^*, b^*)$ ungünstigste a-priori-Verteilung und \hat{p}_{a^*,b^*} Minimax-Schätzer von p . Insbesondere ist der natürliche Schätzer $\hat{p} = X/n$ 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 unbekanntem 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 a-priori-Verteilung $\mu \sim N(0, \sigma^2 E_d)$. In den zugehörigen Bayesschätzern setzen wir dann allerdings statt σ^2 die Schätzung

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

ein und erhalten

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

Der Ansatz lässt vermuten, dass $\hat{\mu}$ kleineres Risiko hat als \bar{X} , wann immer $|\mu|$ klein ist. Überraschenderweise gilt für Dimension $d \geq 3$ 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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,

$$\mathbb{E}[Y_1 f(Y) \mid Y_2 = y_2, \dots, Y_d = y_d] = \mathbb{E}\left[\frac{\partial f}{\partial y_1}(Y) \mid Y_2 = y_2, \dots, Y_d = y_d\right]$$

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:

$$\begin{aligned} \int_{-\infty}^{\infty} f'_y(u) e^{-u^2/2} du &= \int_0^{\infty} f'_y(u) \int_u^{\infty} z e^{-z^2/2} dz du - \int_{-\infty}^0 f'_y(u) \int_{-\infty}^u z e^{-z^2/2} dz du \\ &= \int_0^{\infty} \left(\int_0^z f'_y \right) z e^{-z^2/2} dz - \int_{-\infty}^0 \left(\int_z^0 f'_y \right) z e^{-z^2/2} dz \\ &= \int_{-\infty}^{\infty} z e^{-z^2/2} (f_y(z) - f_y(0)) dz \\ &= \int_{-\infty}^{\infty} f_y(z) z e^{-z^2/2} dz. \end{aligned}$$

□

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 $f(x) = (1 - g(x))x : \mathbb{R}^d \rightarrow \mathbb{R}^d$ das Steinsche Lemma koordinatenweise anwenden, so erhalten wir einen Ausdruck $W(\bar{X})$ unabhängig von μ :

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

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

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

und

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

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 $d \geq 3$ 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*

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

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 $d \geq 3$!

2.30 Bemerkungen.

- (a) Die Abbildung $\mu \mapsto \mathbb{E}_\mu[|\bar{X}|^{-2}]$ ist monoton fallend in $|\mu|$ und erfüllt $\mathbb{E}_0[|\bar{X}|^{-2}] = n/(d-2)$. 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 $d \geq 6$, vgl. Lehmann/Casella, S. 358).

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

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

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

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

2.33 Lemma. *Es seien $g : \Theta \rightarrow A \subseteq \mathbb{R}$ und $l(\vartheta, \rho) = (\rho - g(\vartheta))^2$ der quadratische Verlust. Dann ist eine Entscheidungsregel (ein Schätzer von $g(\vartheta)$) $\hat{g} : \mathcal{X} \rightarrow A$ 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.*

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

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

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

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 $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ heißt dominiert (von μ), falls es ein σ -endliches Maß μ auf \mathcal{F} gibt, so dass \mathbb{P}_ϑ 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) $\mathcal{X} = \mathbb{R}$, $\mathcal{F} = \mathcal{B}_{\mathbb{R}}$, \mathbb{P}_ϑ 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 $\Theta = \{\vartheta_1, \vartheta_2, \dots\}$ 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) $\mathcal{X} = \mathbb{R}$, $\mathcal{F} = \mathcal{B}_{\mathbb{R}}$, $\mathbb{P}_\vartheta = \delta_\vartheta$ 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 $|A \cap \{x \in \mathbb{R} \mid \mu(\{x\}) \geq 1/n\}| \leq n\mu(A) < \infty$, dass $A = A \cap \bigcup_{n \geq 1} \{x \in \mathbb{R} \mid \mu(\{x\}) \geq 1/n\}$ abzählbar ist). Damit kann μ nicht σ -endlich sein.

3.3 Satz. *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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 $c_i \geq 0$, $\sum_i c_i = 1$, $\vartheta_i \in \Theta$, 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

$$\mathcal{P}_0 := \left\{ \sum_i c_i \mathbb{P}_{\vartheta_i} \mid \vartheta_i \in \Theta, c_i \geq 0, \sum_i c_i = 1 \right\},$$

$$\mathcal{A} := \left\{ A \in \mathcal{F} \mid \exists \mathbb{P} \in \mathcal{P}_0 : \mathbb{P}(A) > 0 \text{ und } \frac{d\mathbb{P}}{d\mu} > 0 \mu\text{-f.ü. auf } A \right\}.$$

Wähle nun eine Folge (A_n) in \mathcal{A} mit $\mu(A_n) \rightarrow \sup_{A \in \mathcal{A}} \mu(A) < \infty$. 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 $\frac{d\mathbb{Q}}{d\mu} \geq c_n \frac{d\mathbb{P}_n}{d\mu} > 0$ 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 $A \in \mathcal{F}$. 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 $\mathbb{P}(A) = \mathbb{P}(A \cap A_\infty^C \cap B) > 0$. 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 $\mathcal{X} := \bigcup_{m \geq 1} \mathcal{X}_m$ 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 $\sum_{m \geq 1} 2^{-m} \mathbb{Q}_m$ das Gewünschte. \square

3.2 Exponentialfamilien

3.5 Definition. Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein von μ dominiertes Modell. Dann heißt $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ Exponentialfamilie (in $\eta(\vartheta)$ und T), wenn $k \in \mathbb{N}$, $\eta : \Theta \rightarrow \mathbb{R}^k$, $C : \Theta \rightarrow \mathbb{R}^+$, $T : \mathcal{X} \rightarrow \mathbb{R}^k$ messbar und $h : \mathcal{X} \rightarrow \mathbb{R}^+$ 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 η_1, \dots, η_k linear unabhängige Funktionen und gilt für alle $\vartheta \in \Theta$ die Implikation

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

($1, T_1, \dots, T_k$ sind \mathbb{P}_ϑ -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

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

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

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 $\mathcal{Z} = \mathbb{R} \times (0, \infty)$. Ist $\sigma > 0$ bekannt, so liegt eine einparametrische Exponentialfamilie in $\eta(\mu) = \mu/\sigma^2$ und $T(x) = x$ vor.
- (b) $(\text{Bin}(n, p))_{p \in (0, 1)}$ 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

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

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

3.10 Satz. Es sei $(\mathbb{P}_\vartheta)_{\vartheta \in \mathcal{Z}}$ eine Exponentialfamilie mit natürlichem Parameterraum $\mathcal{Z} \subseteq \mathbb{R}^k$ 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 $A(\vartheta) = \log \left(\int h(x) \exp(\langle \vartheta, T(x) \rangle) \mu(dx) \right)$. Ist $\tilde{\vartheta}$ ein innerer Punkt von \mathcal{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 $\tilde{\vartheta} + s \in \mathcal{Z}$.

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

3.3 Suffizienz

3.11 Beispiel. Es sei X_1, \dots, X_n eine gemäß der Lebesguedichte $f_\vartheta : \mathbb{R} \rightarrow \mathbb{R}^+$ verteilte mathematische Stichprobe. Dann liefern die Statistiken \bar{X} oder $\max(X_1, \dots, X_n)$ im Allgemeinen Information über \mathbb{P}_ϑ 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}_ϑ 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 $X_{(1)} = \min\{X_1, \dots, X_n\}$, $X_{(k+1)} := \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}, \dots, X_{(k)}\}$. Diese ist in folgendem Sinne suffizient.

3.12 Definition. Eine (S, \mathcal{S}) -wertige Statistik T auf $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ heißt suffizient (für $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$), falls für jedes $\vartheta \in \Theta$ die reguläre bedingte Wahrscheinlichkeit von \mathbb{P}_ϑ gegeben T (existiert und) nicht von ϑ abhängt, d.h.

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

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

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

Beweis. Siehe Stochastik II. □

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 $z = \sum_{m \geq 1} 2^{-m} \mu(\mathcal{X}_m)^{-1} \mathbf{1}_{\mathcal{X}_m}$, wobei die Zerlegung $\mathcal{X} := \bigcup_{m \geq 1} \mathcal{X}_m$ 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 $(\mathcal{X}, \sigma(T))$

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

Nach Stochastik II gibt es eine messbare Funktion g_ϑ , so dass $f_\vartheta = g_\vartheta \circ T$, und für beliebige $B \in \mathcal{F}$ 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 \mathcal{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 $h(x) = \frac{d\mathbb{Q}}{d\mu}(x)$. \square

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 $T_2(x) = (\bar{x}, \bar{x}^2)$ oder $T_3(x) = (\bar{x}, \bar{s}^2)$ mit der empirischen Varianz $\bar{s}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. Bei einer Bernoullikette $(\text{Bin}(1, p))^{\otimes n}_{p \in (0,1)}$ ist $T(x) = \sum_{i=1}^n x_i$ (die Anzahl der Erfolge) suffizient.
- (c) Ist X_1, \dots, X_n eine mathematische Stichprobe, wobei X_i gemäß der Lebesgue-dichte $f_\vartheta : \mathbb{R} \rightarrow \mathbb{R}^+$ verteilt ist, so ist die Ordnungsstatistik $(X_{(1)}, \dots, X_{(n)})$ 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 unbekanntem Parameter $\lambda > 0$ kontinuierlich auf $[0, T]$ beobachtet (man denke an Geigerzähleraufzeichnungen). Mit $S_k = \inf\{t \geq 0 \mid N_t = k\}$ 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 $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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*

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

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

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

□

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 $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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 $R(\vartheta, \hat{\vartheta}) = 4 \text{Var}_\vartheta(\bar{X}) = \frac{4\vartheta^2}{12n}$. Nun ist jedoch (bezüglich Lebesguemaß auf $(\mathbb{R}^+)^n$) die Likelihoodfunktion

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

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

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

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

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

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 $n \rightarrow \infty$ 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, \mathcal{F}) -wertige Statistik T auf $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ heißt vollständig, falls für alle messbaren Funktionen $f : S \rightarrow \mathbb{R}$ gilt

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

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}_ϑ -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 $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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 $\mathbb{E}[\tilde{\gamma} - \bar{\gamma}] = 0 \implies \tilde{\gamma} = \bar{\gamma}$ \mathbb{P}_ϑ -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. \square

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 $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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 $f : \mathbb{R}^k \rightarrow \mathbb{R}$. 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 $\{\vartheta \in \mathbb{C}^k \mid |\operatorname{Re}(\vartheta_j)| < a\}$ 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 $f^+ = f^- \mu^T$ -f.ü. und somit $f = 0$ \mathbb{P}_ϑ -f.s. für alle $\vartheta \in \Theta$. \square

3.26 Beispiele.

- (a) Das lineare Modell $Y = X\beta + \sigma\varepsilon$ mit Gaußschen Fehlern $\varepsilon \sim N(0, E_n)$ bildet eine $(k+1)$ -parametrische Exponentialfamilie in $\eta(\beta, \sigma) = \sigma^{-2}(\beta, -1/2)^\top \in \mathbb{R}^k \times \mathbb{R}^-$ und $T(Y) = (X^\top Y, |Y|^2)^\top \in \mathbb{R}^k \times \mathbb{R}^+$. Der natürliche Parameterbereich $\mathcal{L} = \mathbb{R}^k \times \mathbb{R}^-$ besitzt nichtleeres Inneres in \mathbb{R}^{k+1} , so dass T suffizient und vollständig ist. Durch bijektive Transformation ergibt sich, dass dies auch für $((X^\top X)^{-1}X^\top Y, |Y|^2) = (\hat{\beta}, |\Pi_X Y|^2 + (n-k)\hat{\sigma}^2)$ mit dem Kleinste-Quadrate-Schätzer $\hat{\beta}$ und $\hat{\sigma}^2 = \frac{|Y - X\hat{\beta}|^2}{n-k}$ 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 $X_1, \dots, X_n \sim U([0, \vartheta])$ eine mathematische Stichprobe mit $\vartheta > 0$ unbekannt. Aus der Form $L(x, \vartheta) = \vartheta^{-n} \mathbf{1}_{\{x_{(n)} \leq \vartheta\}}$ 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-Effizienz

3.27 Lemma (Chapman-Robbins-Ungleichung). *Es seien $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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 $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta_0}$, $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta_0}$, $\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} \in L^2(\mathbb{P}_{\vartheta_0})$*

$$\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geq \frac{(g(\vartheta) - g(\vartheta_0))^2}{\operatorname{Var}_{\vartheta_0}\left(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}\right)}.$$

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

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

wobei zuletzt die Cauchy-Schwarz-Ungleichung angewendet wurde. \square

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

Im Fall erwartungstreuer Schätzer \hat{g} für $g(\vartheta) = \vartheta$ ergibt die Chapman-Robbins-Gleichung $\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geq \sup_{\vartheta > \vartheta_0/2} \vartheta_0(2\vartheta - \vartheta_0) = \infty$. 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 $\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geq \sup_{\vartheta > \vartheta_0/2} \frac{2\vartheta - \vartheta_0}{\vartheta^2 \vartheta_0} = \vartheta_0^{-2}$, und die Identität $\hat{g} = X$ erreicht auch diese Schranke.

3.29 Definition. Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein von μ dominiertes Modell mit Loglikelihoodfunktion $\ell(\vartheta, x) = \log(L(\vartheta, x))$ ($\log 0 := -\infty$). Die Fisher-Informationsmatrix bei $\vartheta_0 \in \Theta$ ist gegeben durch

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

sofern $\vartheta \mapsto \ell(\vartheta, x)$ für μ -f.a. x differenzierbar bei ϑ_0 mit Gradient $\dot{\ell}(\vartheta_0)$ und der Erwartungswert endlich ist.

3.30 Bemerkung. Die Fisher-Information ist unabhängig vom dominierenden Maß; denn mit dem privilegierten dominierenden Maß \mathbb{Q} gilt

$$\log(L_\mu(\vartheta)) := \log\left(\frac{d\mathbb{P}_\vartheta}{d\mu}\right) = \log\left(\frac{d\mathbb{P}_\vartheta}{d\mathbb{Q}} \frac{d\mathbb{Q}}{d\mu}\right) = \log(L_\mathbb{Q}(\vartheta)) + \log\left(\frac{d\mathbb{Q}}{d\mu}\right),$$

und es folgt $\dot{\ell}_\mu(\vartheta) = \dot{\ell}_\mathbb{Q}(\vartheta)$.

3.31 Satz (Cramér-Rao-Schranke). *Es seien $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein statistisches Modell, $g : \Theta \rightarrow \mathbb{R}$ differenzierbar bei $\vartheta_0 \in \text{int}(\Theta)$ 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 $L(\vartheta, x) := \frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}(x)$ bei ϑ_0 , d.h. mit dem Gradienten (einem Zufallsvektor in \mathbb{R}^k) $\dot{L}(\vartheta_0)$ gilt*

$$\lim_{\vartheta \rightarrow \vartheta_0} \mathbb{E}_{\vartheta_0} \left[\left| L(\vartheta) - L(\vartheta_0) - \langle \dot{L}(\vartheta_0), \vartheta - \vartheta_0 \rangle \right|^2 \right] / |\vartheta - \vartheta_0|^2 = 0.$$

Falls die Fisher-Informationsmatrix $I(\vartheta_0)$ endlich und strikt positiv-definit ist, erhalten wir die Cramér-Rao-Ungleichung als untere Schranke für das quadratische Risiko

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

Beweis. 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, $L(\vartheta_0) = 1 = \mathbb{E}_{\vartheta_0}[L_\vartheta]$ und $\dot{\ell}(\vartheta_0) = \dot{L}(\vartheta_0)$

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

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

3.32 Lemma. *Bildet (\mathbb{P}_ϑ) eine einparametrische Exponentialfamilie in T mit natürlichem Parameter $\vartheta \in \Theta$, so ist T ein erwartungstreuer Schätzer von $g(\vartheta) = \mathbb{E}_\vartheta[T]$, der die Cramér-Rao-Schranke erreicht (Cramer-Rao-effizient), sofern ϑ im Innern von Θ liegt.*

Beweis. Nach Satz 3.10 gilt $g(\vartheta) = E_\vartheta[T] = \dot{A}(\vartheta)$ und $\text{Var}_\vartheta(T) = \ddot{A}(\vartheta)$. Andererseits ist $\ell(\vartheta) = (\vartheta - \vartheta_0)T - (A(\vartheta) - A(\vartheta_0))$, so dass die Loglikelihoodfunktion $\dot{\ell}(\vartheta) = T - \dot{A}(\vartheta)$ existiert und $\mathbb{E}_\vartheta[(\dot{\ell}(\vartheta))^2] = \text{Var}_\vartheta(T) = \ddot{A}(\vartheta)$ erfüllt. Die Cramér-Rao-Schranke ist also gerade $\dot{A}(\vartheta)^2 / \ddot{A}(\vartheta) = \text{Var}_\vartheta(T)$. \square

3.33 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 $\text{Var}_\mu(\hat{\mu}) = \sigma^2/n$ 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 $\text{Var}_\mu(\widehat{\mu^2}) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}$, 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.

3.34 Bemerkung. In der Tat wird die Cramér-Rao-Schranke nur erreicht, wenn (\mathbb{P}_ϑ) 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 genauso gut mit dem Satz von Lehmann-Scheffé argumentieren. Später werden wir sehen, dass in allgemeineren Modellen immerhin asymptotisch die Cramér-Rao-Schranke erreichbar ist.

3.35 Lemma. *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein dominiertes statistisches Modell, und die Loglikelihoodfunktion ℓ sei $L^2(\mathbb{P}_{\vartheta_0})$ -differenzierbar bei $\vartheta_0 \in \Theta$ mit Gradient $\dot{\ell}(\vartheta_0)$. Dann gilt $\mathbb{E}_{\vartheta_0}[\dot{\ell}(\vartheta_0)] = 0$.*

Beweis. L^2 -Konvergenz impliziert L^1 -Konvergenz und insbesondere Konvergenz der Erwartungswerte. Die Behauptung folgt daher aus der Ableitungsregel $\dot{\ell} = \dot{L}/L$ sowie $\mathbb{E}_{\vartheta_0}[(L(\vartheta) - L(\vartheta_0))/L(\vartheta_0)] = \int (L(\vartheta) - L(\vartheta_0))/L(\vartheta_0) d\mu = 1 - 1 = 0$ (μ sei dom. Maß). \square

3.36 Lemma. *Es seien X_1, \dots, X_n Beobachtungen aus unabhängigen statistischen Experimenten $\mathcal{E}_1, \dots, \mathcal{E}_n$ mit derselben Parametermenge $\Theta \subseteq \mathbb{R}^k$. Bezeichnet dann I_j die Fisher-Information, erzeugt von der Beobachtung X_j (die existieren möge), $j = 1, \dots, n$, so existiert die gemeinsame Fisher-Information I , erzeugt von X_1, \dots, X_n , und es gilt*

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

Beweis. Nach Annahme sind die entsprechenden Loglikelihoodfunktionen ℓ_1, \dots, ℓ_n bezüglich der dominierenden Maße μ_1, \dots, μ_n fast überall differenzierbar. Also ist auch die gemeinsame Loglikelihoodfunktion $\ell(\vartheta, x) = \sum_{j=1}^n \ell_j(\vartheta, x_j)$ (bezüglich $\mu_1 \otimes \dots \otimes \mu_n$) fast überall differenzierbar mit Gradienten $\dot{\ell}(\vartheta, x) = \sum_{j=1}^n \dot{\ell}_j(\vartheta, x_j)$. Wegen Unabhängigkeit der X_1, \dots, X_n sowie $\mathbb{E}_{\vartheta}[\dot{\ell}_j(\vartheta, X_j)] = 0$ gilt

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

\square

4 Allgemeine Schätztheorie

4.1 Momentenschätzer

4.1 Definition. Es seien $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}_{\vartheta}^{\otimes n})_{\vartheta \in \Theta})$ ein statistisches (Produkt-)Modell mit $\mathcal{X} \subseteq \mathbb{R}$, $\mathcal{F} \subseteq \mathfrak{B}_{\mathbb{R}}$ und $g(\vartheta)$ mit $g : \Theta \rightarrow \mathbb{R}^p$ ein abgeleiteter Parameter. Ferner sei $\psi = (\psi_1, \dots, \psi_q) : \mathcal{X} \rightarrow \mathbb{R}^q$ derart, dass

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

existiert. Gibt es nun eine Borel-messbare Funktion $G : \varphi(\Theta) \rightarrow g(\Theta)$ 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 ψ_1, \dots, ψ_q .

4.2 Beispiele.

- (a) Es sei $X_1, \dots, X_n \sim \text{Exp}(\lambda)$ 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

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

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

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

mit $(\varepsilon_n) \sim N(0, \sigma^2)$ i.i.d und $X_0 = x_0 \in \mathbb{R}$. Um a zu schätzen, betrachte folgende Identität für das bedingte gemeinsame Moment:

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

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

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

Im Fall $|a| < 1$ kann man mit Hilfe des Ergodensatzes auf die Konsistenz von \hat{a}_n für $n \rightarrow \infty$ schließen. Allgemeiner zeigt man leicht, dass $M_n := \sum_{k=1}^n X_{k-1}\varepsilon_k$ ein Martingal bezüglich $\mathcal{F}_n := \sigma(\varepsilon_1, \dots, \varepsilon_n)$ 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 Satz (Δ -Methode). *Es seien (X_n) eine Folge von Zufallsvektoren im \mathbb{R}^k , $\sigma_n > 0$, $\sigma_n \rightarrow 0$, $\vartheta_0 \in \mathbb{R}^k$ sowie $\Sigma \in \mathbb{R}^{k \times k}$ positiv definit und es gelte*

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

Ist $f : \mathbb{R}^k \rightarrow \mathbb{R}$ in einer Umgebung von ϑ_0 stetig differenzierbar mit $(\nabla f(\vartheta_0))^\top \Sigma \nabla f(\vartheta_0) > 0$, so folgt

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

4.4 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. $\lim_{n \rightarrow \infty} \hat{g}_n = g(\vartheta)$ \mathbb{P}_ϑ -f.s.*

4.5 Satz. *Es seien $\vartheta_0 \in \Theta$, $g : \Theta \rightarrow \mathbb{R}$ und für hinreichend großes n existiere der Momentenschätzer $\hat{g}_n = G(\frac{1}{n} \sum_{i=1}^n \psi(x_i))$ mit Momentenfunktionen $\psi_j \in L^2(\mathbb{P}_{\vartheta_0})$, $j = 1, \dots, q$. Setze $\Sigma(\vartheta_0) := (\text{Cov}_{\vartheta_0}(\psi_i, \psi_j))_{i,j=1, \dots, q}$. Sofern G in einer Umgebung von $\varphi(\vartheta_0)$ stetig differenzierbar ist mit $\sigma^2 := (\nabla G(\varphi(\vartheta_0)))^\top \Sigma(\vartheta_0) \nabla G(\varphi(\vartheta_0)) > 0$, ist \hat{g}_n unter $\mathbb{P}_{\vartheta_0}^{\otimes n}$ asymptotisch normalverteilt mit Rate $n^{-1/2}$ und asymptotischer Varianz σ^2 :*

$$\sqrt{n}(\hat{g}_n - g(\vartheta_0)) \xrightarrow{d} N(0, \sigma^2) \quad (\text{unter } \mathbb{P}_{\vartheta_0}^{\otimes n}).$$

4.6 Bemerkung. Der Begriff *asymptotische Varianz* ist leicht irreführend: es gilt nicht notwendigerweise, dass $n \text{Var}_{\vartheta_0}(\hat{g}_n) \rightarrow \sigma^2$.

4.7 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 \bar{X} basiert, wird dieser Schätzer im Allgemeinen vorgezogen.

4.2 Maximum-Likelihood- und Minimum-Kontrast-Schätzer

4.8 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: $\hat{\vartheta}(x) := \text{argmax}_{\vartheta \in \Theta} p_\vartheta(x)$. Dieser Schätzer heißt Maximum-Likelihood-Schätzer (MLE). Bereits im vorliegenden Fall ist weder Existenz noch Eindeutigkeit ohne Weiteres garantiert. Bei Nicht-Eindeutigkeit wählt man einen maximierenden Parameter ϑ nach Belieben aus. Im Fall einer mathematischen Stichprobe $X_1, \dots, X_n \sim \text{Pois}(\lambda)$ mit $\lambda > 0$ unbekannt, ergibt sich beispielsweise

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

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

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

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

Man sieht, dass der MLE invariant unter Parametertransformation ist: bei Beobachtung von $X \sim N(\mu, 1)$ erhält man den MLE $\hat{\mu}(X) = X$ 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 einen anderen Schätzer: $\bar{\mu}(Y) = \log(Y) - 1/2$.

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

4.10 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}_ϑ irrelevant, weil der Schätzer vor Realisierung des Experiments festgelegt wird und diese Realisierung damit fast sicher zum selben Schätzwert führen wird.

4.11 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 $\text{int}(\Theta)$ von Θ .

4.12 Bemerkung. Bei einer eindeutigen 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 $\mathbb{E}_{\hat{\vartheta}}[T] = T(x)$. \square

4.13 Beispiele.

- (a) Es sei $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ eine mathematische Stichprobe. Dann ist der MLE für $\vartheta = (\mu/\sigma^2, 1/(2\sigma^2))^\top$ gegeben durch $\mathbb{E}_{\hat{\vartheta}}[(\bar{X}, \bar{X}^2)^\top] = (\bar{X}, \bar{X}^2)^\top$, 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 unbekanntem Übergangswahrscheinlichkeiten $\mathbb{P}(X_{k+1} = j | X_k = i) = p_{ij}$ 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 $N_{kl}(X) = |\{i = 1, \dots, n | X_{i-1} = k, X_i = l\}|$ 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 $\hat{\vartheta} = \text{argmin}_{\vartheta \in \Theta} \sum_{i=1}^n (Y_i - g_\vartheta(x_i))^2$.

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

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

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

4.15 Lemma. Für die Kullback-Leibler-Divergenz gilt:

(a) $\text{KL}(\mathbb{P} | \mathbb{Q}) \geq 0$ und $\text{KL}(\mathbb{P} | \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$, aber KL ist nicht symmetrisch;

(b) für Produktmaße ist KL additiv:

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

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

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

Beweis. Übung! □

4.16 Bemerkung. Wegen $\ddot{A}(\vartheta_0) = \text{Cov}_{\vartheta_0}(T)$ in (c) erhalten wir für $\vartheta, \vartheta_0 \in \text{int}(\Theta)$ mit einer Taylorentwicklung $\text{KL}(\mathbb{P}_{\vartheta_0} | \mathbb{P}_\vartheta) = \frac{1}{2} \langle \text{Cov}_{\bar{\vartheta}}(T)(\vartheta - \vartheta_0), \vartheta - \vartheta_0 \rangle$ mit einer Zwischenstelle $\bar{\vartheta}$ zwischen ϑ und ϑ_0 . Beachte, dass $\text{Cov}_{\bar{\vartheta}}(T)$ 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 $\text{KL}(N(\vartheta_0, \Sigma) | N(\vartheta, \Sigma)) = \frac{1}{2} \langle \Sigma^{-1}(\vartheta - \vartheta_0), \vartheta - \vartheta_0 \rangle$ gilt.

4.17 Definition. Es sei $(\mathcal{X}_n, \mathcal{F}_n, (\mathbb{P}_\vartheta^n)_{\vartheta \in \Theta})_{n \geq 1}$ eine Folge statistischer Modelle. Eine Funktion $K : \Theta \times \Theta \rightarrow \mathbb{R} \cup \{+\infty\}$ heißt Kontrastfunktion, falls $\vartheta \mapsto K(\vartheta_0, \vartheta)$ ein eindeutiges Minimum bei ϑ_0 besitzt für alle $\vartheta_0 \in \Theta$. Eine Folge $K_n : \Theta \times \mathcal{X}_n \rightarrow \mathbb{R} \cup \{+\infty\}$ heißt zugehöriger Kontrastprozess (oder bloß Kontrast), falls folgende Bedingungen gelten:

(a) $K_n(\vartheta, \bullet)$ ist \mathcal{F}_n -messbar für alle $\vartheta \in \Theta$;

(b) $\forall \vartheta, \vartheta_0 \in \Theta : K_n(\vartheta) \xrightarrow{d} K(\vartheta_0, \vartheta)$ unter $\mathbb{P}_{\vartheta_0}^n$ für $n \rightarrow \infty$.

Ein zugehöriger Minimum-Kontrast-Schätzer ist gegeben durch $\hat{\vartheta}_n(x_n) := \text{argmin}_{\vartheta \in \Theta} K_n(\vartheta, x_n)$ (sofern existent; nicht notwendigerweise eindeutig).

4.18 Beispiele.

(a) Beim Produktexperiment $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta \in \Theta})$ 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

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

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

(b) Betrachte das Regressionsmodell aus Beispiel 4.13 mit $g_\vartheta : [0, 1] \rightarrow \mathbb{R}$ 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 $K_n(\vartheta) = \frac{1}{n} \sum_{i=1}^n (Y_i - g_\vartheta(x_i))^2$ einen Kontrastprozess zur Kontrastfunktion $K(\vartheta_0, \vartheta) = \int_0^1 (g_{\vartheta_0}(x) - g_\vartheta(x))^2 dx + \mathbb{E}[\varepsilon_i^2]$ bildet. Dabei muss natürlich die Identifizierbarkeitsbedingung $g_\vartheta \neq g_{\vartheta'}$ für alle $\vartheta \neq \vartheta'$ gelten. Also ist der Kleinste-Quadrate-Schätzer hier ebenfalls Minimum-Kontrast-Schätzer.

4.3 Asymptotik

4.19 Satz. *Es sei $(K_n)_{n \geq 1}$ ein Kontrastprozess zur Kontrastfunktion K . Dann ist der zugehörige Minimum-Kontrast-Schätzer $\hat{\vartheta}_n$ konsistent für $\vartheta_0 \in \Theta$ unter folgenden Bedingungen:*

(A1) Θ ist ein kompakter Raum;

(A2) $\vartheta \mapsto K(\vartheta_0, \vartheta)$ ist stetig und $\vartheta \mapsto K_n(\vartheta)$ ist $\mathbb{P}_{\vartheta_0}^n$ -f.s. stetig;

(A3) $\sup_{\vartheta \in \Theta} |K_n(\vartheta) - K(\vartheta_0, \vartheta)| \xrightarrow{d} 0$ unter $\mathbb{P}_{\vartheta_0}^n$.

Beweis. Zunächst beachte, dass $\hat{\vartheta}_n$ als Minimum einer stetigen Funktion auf einem Kompaktum stets existiert. Es kann außerdem messbar gewählt werden (vgl. Witting, 2. Band, Satz 6.7).

Zeige nun, dass die entsprechende Funktion $\text{argmin} : C(\Theta) \rightarrow \Theta$ stetig bezüglich Maximumsnorm auf $C(\Theta)$ ist an den Stellen f , wo $m_f := \text{argmin}_\vartheta f(\vartheta)$ eindeutig ist. Betrachte $f_n \in C(\Theta)$ mit $\|f_n - f\|_\infty \rightarrow 0$. Dann konvergieren auch die Minima: $f_n(m_{f_n}) \rightarrow f(m_f)$. Ist nun $m \in \Theta$ (Θ kompakt) ein Häufungspunkt von (m_{f_n}) , so folgt $f(m) = \lim_{n \rightarrow \infty} f_n(m_{f_n}) = f(m_f)$. Eindeutigkeit des Minimums liefert $m = m_f$, und daher besitzt (m_{f_n}) als einzigen Häufungspunkt den Grenzwert m_f .

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

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

$$\forall \varepsilon > 0 : \lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P} \left(\sup_{|\vartheta_1 - \vartheta_2| < \delta} |X_n(\vartheta_1) - X_n(\vartheta_2)| \geq \varepsilon \right) = 0.$$

Beweis. Siehe Stochastik II. \square

4.21 Satz. *Es mögen die Annahmen (A1)-(A3) sowie $\Theta \subseteq \mathbb{R}^k$ und $\vartheta_0 \in \text{int}(\Theta)$ gelten. Der Kontrastprozess K_n sei zweimal stetig differenzierbar in einer Umgebung von ϑ_0 (\mathbb{P}_{ϑ_0} -f.s.), so dass mit*

$$U_n(\vartheta) := \dot{K}_n(\vartheta) \text{ (Score)}, \quad V_n(\vartheta) := \ddot{K}_n(\vartheta)$$

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

(B1) $\sqrt{n}U_n(\vartheta_0) \xrightarrow{d} N(0, U(\vartheta_0))$ mit $U(\vartheta_0) \in \mathbb{R}^{k \times k}$ positiv definit.

(B2) Aus $\vartheta_n \xrightarrow{\mathbb{P}_{\vartheta_0}^n} \vartheta_0$ folgt $V_n(\vartheta_n) \xrightarrow{\mathbb{P}_{\vartheta_0}^n} V(\vartheta_0)$ mit $V(\vartheta_0) \in \mathbb{R}^{k \times k}$ regulär.

Dann ist der Minimum-Kontrast-Schätzer $\hat{\vartheta}_n$ unter \mathbb{P}_{ϑ_0} asymptotisch normalverteilt:

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \xrightarrow{d} N(0, V(\vartheta_0)^{-1}U(\vartheta_0)V(\vartheta_0)^{-1}).$$

Beweis. Aus (A1)-(A3) folgt die Konsistenz von $\hat{\vartheta}_n$. Wegen $\vartheta_0 \in \text{int}(\Theta)$ gilt für $\Omega_n^1 := \{[\hat{\vartheta}_n, \vartheta_0] \in \text{int}(\Theta)\}$ (setze $[a, b] := \{ah + b(1-h) \mid h \in [0, 1]\}$) $\lim_{n \rightarrow \infty} \mathbb{P}_{\vartheta_0}(\Omega_n^1) = 1$. Auf Ω_n^1 gilt somit $\dot{K}_n(\hat{\vartheta}_n) = 0$ und nach Mittelwertsatz

$$\dot{K}_n(\hat{\vartheta}_n) - \dot{K}_n(\vartheta_0) = \ddot{K}_n(\bar{\vartheta}_n)(\hat{\vartheta}_n - \vartheta_0), \quad \bar{\vartheta}_n \in [\vartheta_0, \hat{\vartheta}_n].$$

Wir erhalten

$$-U_n(\vartheta_0) = V_n(\bar{\vartheta}_n)(\hat{\vartheta}_n - \vartheta_0).$$

Wegen (B2), und da $V(\vartheta_0)$ regulär und die Inversenbildung stetig ist, haben wir $\mathbb{P}(\Omega_n^2) \rightarrow 1$ für $\Omega_n^2 := \{V_n(\bar{\vartheta}_n)^{-1} \text{ existiert}\}$. Benutze nun $V_n(\bar{\vartheta}_n)^{-1} \mathbf{1}_{\Omega_n^1 \cap \Omega_n^2} \rightarrow V(\vartheta_0)^{-1}$ in \mathbb{P}_{ϑ_0} -Wahrscheinlichkeit, so dass

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = -V_n(\bar{\vartheta}_n)^{-1} \sqrt{n}U_n(\vartheta_0) \xrightarrow{d} N(0, V(\vartheta_0)^{-1}I(\vartheta_0)V(\vartheta_0)^{-1})$$

aus Slutskys Lemma folgt. \square

4.22 Satz. Es sei $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}_{\vartheta}^{\otimes n})_{\vartheta \in \Theta})_{n \geq 1}$ 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 $\text{int}(\Theta)$ 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 $\sup_{\vartheta \in \Theta} |\ell(\vartheta, x)| \leq H_0(x)$ und $\sup_{\vartheta \in U} |\dot{\ell}(\vartheta, x)| \leq H_1(x)$, $\sup_{\vartheta \in U} |\ddot{\ell}(\vartheta, x)| \leq H_2(x)$, $x \in \mathcal{X}$.
- (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 ist der MLE $\hat{\vartheta}_n$ unter $\mathbb{P}_{\vartheta_0}^{\otimes n}$ asymptotisch normalverteilt:

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

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

Beweis. Setze $K_n(\vartheta, x) := -\frac{1}{n} \sum_{i=1}^n \ell(\vartheta, x_i)$, $x \in \mathcal{X}$, 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 $|\ell(\vartheta) - \ell(\vartheta')| \leq 2H_0$ liefert

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

(A3) Mit dem starken Gesetz der großen Zahlen folgt \mathbb{P}_{ϑ_0} -f.s.:

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

Mit dominierter Konvergenz und gleichmäßiger Stetigkeit von ℓ auf dem Kompaktum Θ erhalten wir, dass der letzte Erwartungswert für $\delta \rightarrow 0$ gegen Null konvergiert. Dies zeigt die Straffheit von K_n (mit f.s.-Konvergenz in Satz 4.20). Insbesondere ist wegen (A1)-(A3) $\hat{\vartheta}_n$ konsistent.

(B1) Mittels dominierter Konvergenz erhalten wir wie in Lemma 3.35 $\mathbb{E}_{\vartheta_0}[\dot{\ell}(\vartheta_0)] = 0$. 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

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

Der zentrale Grenzwertsatz liefert daher unter \mathbb{P}_{ϑ_0}

$$\sqrt{n}\dot{K}_n(\vartheta_0) \xrightarrow{d} N(0, I(\vartheta_0)).$$

(B2) Wir haben \mathbb{P}_{ϑ_0} -f.s. mit dem starken Gesetz der großen Zahlen

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

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

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

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

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

folgt. Mit $\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} \mathbb{P}_{\vartheta_0}(\Omega_{\delta,n}) = 1$ und der Konvergenz von $V_n(\vartheta_0)$ erhalten wir daher in \mathbb{P}_{ϑ_0} -Wahrscheinlichkeit

$$V_n(\vartheta_n) \xrightarrow{n \rightarrow \infty} I(\vartheta_0).$$

□

4.23 Bemerkungen.

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

4.24 Beispiel. Bei einer Exponentialfamilie mit natürlichem Parameterraum und natürlicher suffizienter Statistik T erfüllt der MLE (so er existiert und in $\text{int}(\Theta)$ liegt) $\mathbb{E}_{\hat{\vartheta}(x)}[T] = T(x)$ und die Fisher-Information $I(\vartheta) = \text{Cov}_{\vartheta}(T)$ (Kovarianzmatrix von T). Es folgt also mit Regularitätsannahmen $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \rightarrow N(0, \text{Cov}_{\vartheta_0}(T)^{-1})$ unter \mathbb{P}_{ϑ_0} . Bei einer Bernoullikette X_1, \dots, X_n mit $X_i \sim \text{Bin}(1, p)$ ist $\vartheta = \log(p/(1-p))$ der natürliche Parameter sowie $T(x) = x$. Aus $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \rightarrow N(0, p(\vartheta_0)^{-1}(1-p(\vartheta_0))^{-1})$ folgt mittels Δ -Methode für die p -Parametrisierung $\sqrt{n}(\hat{p}_n - p_0) \rightarrow N(0, p_0(1-p_0))$. Wegen $\hat{p}_n = \bar{X}$ ist dieses Resultat natürlich konkret einfach zu überprüfen.

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

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

beobachtete Fisher-Informationsmatrix.

4.26 Korollar. *Unter den Voraussetzungen des vorigen Satzes gilt unter \mathbb{P}_{ϑ_0}*

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

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 $[\hat{\vartheta}_n - n^{-1/2}\mathcal{J}_n^{-1/2}q_{1-\alpha/2}, \hat{\vartheta}_n + n^{-1/2}\mathcal{J}_n^{-1/2}q_{1-\alpha/2}]$ 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 $\hat{\ell}$), folgt $I(\hat{\vartheta}) \rightarrow I(\vartheta_0)$ in \mathbb{P}_{ϑ_0} -Wahrscheinlichkeit. Ebenso liefert das Argument im obigen Nachweis der Eigenschaft (B2) $\mathcal{J}_n \rightarrow I(\vartheta_0)$ \mathbb{P}_{ϑ_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). \square

4.27 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 $\mathbb{P}(a_i = 1) = \mathbb{P}(a_i = 2) = 1/2$ und $Y_i \sim N(\vartheta, \sigma_{a_i}^2)$ bedingt auf a_i . Wir erhalten die Loglikelihood-Funktion

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

Der MLE ist also $\hat{\vartheta}_n = (\sum_{i=1}^n Y_i / \sigma_{a_i}^2) / (\sum_{i=1}^n \sigma_{a_i}^{-2})$ und die Fisher-Information $I(\vartheta) = \frac{1}{2\sigma_0^2} + \frac{1}{2\sigma_1^2} =: I$ (unabhängig von ϑ). $\hat{\vartheta}_n$ ist (nach Theorie oder mit direkten Argumenten) asymptotisch normalverteilt mit asymptotischer Varianz $N(0, I^{-1})$, was auf asymptotische Konfidenzintervalle der Form $\hat{\vartheta}_n \pm \frac{1}{\sqrt{n}} I^{1/2} q_{1-\alpha/2}$ führt. Natürlich gilt $I(\hat{\vartheta}_n) = I$, während die beobachtete Fisher-Information $\mathcal{J}_n = \frac{\sum_{i=1}^n a_i}{n\sigma_0^2} + \frac{\sum_{i=1}^n (1-a_i)}{n\sigma_1^2}$ erfüllt. Damit ist \mathcal{J}_n^{-1} gerade gleich der *bedingten* Varianz $\text{Var}_{\vartheta}(n^{1/2}\hat{\vartheta}_n | a)$. Da wir ja a beobachten, ist die bedingte Varianz sicherlich ein sinnvollerer 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

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 $\dim(\Theta_n) \rightarrow \infty$), 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.).

4.28 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*

$$0 \leq K(\vartheta_0, \hat{\vartheta}_n) - K(\vartheta_0, \vartheta_0) \leq (K_n(\vartheta_0) - K_n(\hat{\vartheta}_n)) - (K(\vartheta_0, \vartheta_0) - K(\vartheta_0, \hat{\vartheta}_n)) \\ \leq \sup_{\vartheta \in \Theta} \left((K_n(\vartheta_0) - K_n(\vartheta)) - (K(\vartheta_0, \vartheta_0) - K(\vartheta_0, \vartheta)) \right).$$

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

$$0 \leq K(\vartheta_0, \hat{\vartheta}_n) - K(\vartheta_0, \vartheta_0) \leq (K_n(\vartheta_0) - K_n(\hat{\vartheta}_n)) - (K(\vartheta_0, \vartheta_0) - K(\vartheta_0, \hat{\vartheta}_n)).$$

Das zufällige Argument $\hat{\vartheta}_n$ liefert stets einen kleineren Wert als das Maximum. \square

4.29 Beispiele. In Fortführung von Beispiel 4.18 erhalten wir:

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

$$\text{KL}(\vartheta_0 | \hat{\vartheta}_n) \leq \sup_{\vartheta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n (\ell(\vartheta, X_i) - \ell(\vartheta_0, X_i) + \text{KL}(\vartheta_0 | \vartheta)) \right).$$

Unter \mathbb{P}_{ϑ_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 $\text{KL}(\vartheta_0 | \hat{\vartheta}_n) = \frac{(\vartheta_0 - \hat{\vartheta}_n)^2}{2\sigma^2}$. Wir schätzen hier also einfach den quadratischen Verlust ab.

- (b) Beim Kleinste-Quadrate-Schätzer im nichtlinearen Regressionsmodell gilt unter \mathbb{P}_{ϑ_0}

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

Die letzten beiden Terme bilden einen deterministischen Approximationsfehler, der für glatte g_ϑ im Allgemeinen klein ist. Da auf den nicht-asymptotischen 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:

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

Der innere Term allein ist zentriert und besitzt die Varianz $\frac{\sigma^2}{n} \|g_{\vartheta_0} - g_{\vartheta}\|_n^2$ (setze $\sigma^2 = \text{Var}(\varepsilon_1)$), ist also von der Ordnung $n^{-1/2}$. Allerdings ist a priori überhaupt nicht klar, ob dasselbe für das Supremum gilt. Wenn beispielsweise $\{((g_{\vartheta_0} - g_{\vartheta})(x_i))_{i=1, \dots, n} \mid \vartheta \in \Theta\}$ einen Würfel $\{z \in \mathbb{R}^n \mid \max_i |z_i| \leq r\}$ 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}_{ϑ}) 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*.

4.30 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

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

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

4.32 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 $|X_i| \leq K$ gilt, so folgt*

$$\mathbb{P}(|S_n| \geq \kappa) \leq 2 \exp\left(-\frac{\kappa^2}{4(\text{Var}(S_n) + \kappa K)}\right), \quad \kappa > 0.$$

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

4.33 Beispiele.

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

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

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

$$\mathbb{P}\left(|S_n - np| \geq \kappa \sqrt{np(1 - p)}\right) \leq 2 \exp\left(-\frac{\kappa^2}{4 + 4 \max(p, 1 - p) \kappa (np(1 - p))^{-1/2}}\right).$$

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 $np_n \rightarrow \lambda > 0$ für $n \rightarrow \infty$ 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

$$\mathbb{P}(|S_n| \geq \kappa) \leq \mathbb{P}(\exists i = 1, \dots, n : |X_i| > r) + 2 \exp\left(-\frac{\kappa^2}{4(\text{Var}(S_n) + \kappa r)}\right).$$

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

$$\mathbb{P}(|S_n| \geq \kappa) \leq (2+C) \exp\left(-\frac{\kappa^2}{4(\text{Var}(S_n) + \kappa D \log(n) + \kappa^3 D/(4 \text{Var}(S_n)))}\right)$$

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

$$\mathbb{P}(|S_n| \geq x\sqrt{n}) \leq (2+C) \exp\left(-\frac{x^2}{4 \text{Var}(X_i) + \frac{4xD}{\sqrt{n}}(\log(n) + x^2/(4 \text{Var}(X_i)))}\right).$$

Um das Supremum in Lemma 4.28 abzuschätzen, werden wir $A_\vartheta = \sqrt{n}(K_n(\vartheta) - K(\vartheta_0, \vartheta))$ im folgenden Satz betrachten.

4.34 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*

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

bzw.

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

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 $\rho := \sup_{\vartheta \in \Theta} |\vartheta - \vartheta_0|$

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

bzw.

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

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 $j \geq 1$ mit $\sup_{\vartheta \in \Theta} \inf_{\vartheta_j \in \Theta_j} |\vartheta - \vartheta_j| \leq \rho 2^{-j}$ (Θ_j ist $\rho 2^{-j}$ -Netz). Man kann stets $|\Theta_j| \leq C_1 2^{jk}$ 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 $j \rightarrow \infty$ 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 $\eta_j \geq 0$ mit $\sum_{j \geq 1} \eta_j = 1$ wählen:

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

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

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

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 $C_{\delta,k} \geq CC_1 2^k/(e^k - 2^k)$ wählen, so dass die Behauptung gilt. \square

4.5 Anwendung auf Regression und Maximum-Likelihood

Um konkrete Resultate für Minimum-Kontrast-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 $\mathbb{E}[\varepsilon_i] = 0$, $\operatorname{Var}(\varepsilon_i) = \sigma^2 > 0$ 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 $\{g_\vartheta \mid \vartheta \in \Theta\}$ eine bezüglich $\|\cdot\|_n$ 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 $\vartheta_0 := \operatorname{argmin}_{\vartheta \in \Theta} \|g_\vartheta - g\|_n^2$ (g_{ϑ_0} ist Projektion von g auf $\{g_\vartheta \mid \vartheta \in \Theta\}$) liefert Lemma 4.28

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

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

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

Betrachte nun den stochastischen Prozess $A_\vartheta := \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i g_\vartheta(i/n)$. Es gilt $\mathbb{E}[A_\vartheta - A_{\vartheta'}] = 0$, $\operatorname{Var}(A_\vartheta - A_{\vartheta'}) = \sigma^2 \|g_\vartheta - g_{\vartheta'}\|_n^2$ und betrachte $L_2 \geq \|g_\vartheta - g_{\vartheta'}\|_n / |\vartheta - \vartheta'|$, $L_\infty \geq \|g_\vartheta - g_{\vartheta'}\|_\infty / |\vartheta - \vartheta'|$. Wie im Beispiel 4.33(b) liefert die Bernstein-Ungleichung für $A_\vartheta - A_{\vartheta'}$ eine Konzentrationsungleichung:

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

Damit erhalten wir nach Satz 4.34 eine gleichmäßige Abschätzung für $\{\vartheta_0\} \subseteq \Theta_0 \subseteq \Theta$ mit $\rho(\Theta_0) = \sup_{\vartheta \in \Theta_0} |\vartheta - \vartheta_0| < \infty$

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

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

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

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

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

und erhalten so für $M_n > 0$ beliebig

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

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

und

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

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

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

Außerdem gilt

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

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 $\tilde{c} > 0$.

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

Wir haben also folgenden Satz bewiesen.

4.35 Satz. *Es sei $g_\vartheta \in C(D)$, $D \subseteq \mathbb{R}^d$, für $\vartheta \in \Theta \subseteq \mathbb{R}^k$ mit $\{g_\vartheta \mid \vartheta \in \Theta\}$ 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 $\mathbb{E}[\varepsilon_i] = 0$, $\operatorname{Var}(\varepsilon_i) = \sigma^2 > 0$ 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

$$\mathbb{P}(Z \geq \kappa) \leq C_k \left(\exp\left(-\frac{\kappa}{129\sigma^2 L_2^2 l_2^{-2}}\right) + e^{-cn^{1/4}} \right)$$

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

4.36 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

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

Nun sind $(\langle \varepsilon, \varphi_j \rangle_n)_{j=1, \dots, k}$ unabhängige $N(0, \sigma^2/n)$ -verteilte Zufallsvariablen, so dass $U := \frac{n}{\sigma^2} \sum_{j=1}^k \langle \varepsilon, \varphi_j \rangle_n^2$ $\chi^2(k)$ -verteilt ist. Aus dem exponentiellen Moment $\mathbb{E}[e^{\alpha U}] = (1 - 2\alpha)^{-k/2}$ für $0 \leq \alpha = 1/2 - \delta < 1/2$ ergibt sich für $Z = \sigma^2 U$ die Ungleichung $\mathbb{P}(Z \geq \kappa) \leq (2\delta)^{-k/2} e^{(\frac{1}{2}-\delta)\kappa/\sigma^2}$, $\delta \in (0, 1/2]$. Wir sehen also, dass dies der Struktur unserer 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.

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

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

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

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

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

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

4.39 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. nicht-zusammenhängend. 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 $\mathcal{G} \subseteq C([0, 1])$ 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 $\Theta_j \subseteq \mathcal{G} \subseteq \mathbb{R}^n$ in die Vorfaktoren bei den gleichmäßigen Abschätzungen mit ein und diese sollte unabhängig von n sein (z.B. \mathcal{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}_ϑ 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 $K(\vartheta_0, \vartheta) := \mathbb{E}[-\ell(\vartheta)] = \text{KL}(\mathbb{P} | \mathbb{P}_\vartheta)$ mit $\vartheta_0 \in \Theta$ so, dass $\text{KL}(\mathbb{P} | \mathbb{P}_{\vartheta_0}) = \inf_{\vartheta \in \Theta} \text{KL}(\mathbb{P} | \mathbb{P}_\vartheta)$ gilt (falls solch ein ϑ_0 nicht existiert, ersetze einfach in allen folgenden Ausdrücken $\text{KL}(\mathbb{P} | \mathbb{P}_{\vartheta_0})$ durch $\inf_{\vartheta \in \Theta} \text{KL}(\mathbb{P} | \mathbb{P}_\vartheta)$).

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

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

erfüllt, wobei die Zufallsvariable L folgender Exponentialungleichung genügt:

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

Mit $A_\vartheta := \sqrt{n}(K_n(\vartheta) - K(\vartheta_0, \vartheta))$ gilt dann $\mathbb{E}[A_\vartheta] = 0$, $\text{Var}(A_\vartheta - A_{\vartheta'}) \leq \mathbb{E}[L^2]|\vartheta - \vartheta'|^2$, und die Bernstein-Ungleichung zusammen mit Satz 4.34 liefert für $\{\vartheta_0\} \subseteq \Theta_0 \subseteq \Theta$ mit $\rho(\Theta_0) = \sup_{\vartheta \in \Theta_0} |\vartheta - \vartheta_0| < \infty$

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

Setze nun

$$\Theta_j := \left\{ \vartheta \in \Theta \mid \text{KL}(\mathbb{P} | \mathbb{P}_\vartheta) - \text{KL}(\mathbb{P} | \mathbb{P}_{\vartheta_0}) \leq 2^{2(j+1)}\kappa/n \right\}.$$

Peeling liefert dann für $M_n > 0$ beliebig

$$\begin{aligned} & \mathbb{P}(\text{KL}(\mathbb{P} | \mathbb{P}_{\hat{\vartheta}_n}) - \text{KL}(\mathbb{P} | \mathbb{P}_{\vartheta_0}) \in [\kappa/n, M_n^2]) \\ & \leq \mathbb{P}\left(\exists j = 0, \dots, J_n : \sup_{\vartheta \in \Theta_j} n^{-1/2}(A_\vartheta - A_{\vartheta'}) \geq 2^{2j}\kappa/n\right) \end{aligned}$$

mit $2^{J_n+1}\sqrt{\kappa/n} \geq M_n$. Setzen wir weiterhin $\inf_{\vartheta} \frac{\text{KL}(\mathbb{P} | \mathbb{P}_\vartheta) - \text{KL}(\mathbb{P} | \mathbb{P}_{\vartheta_0})}{|\vartheta - \vartheta_0|^2} \geq l^2 > 0$ voraus, so gilt $\rho(\Theta_j) \leq l^{-1}2^{j+1}\sqrt{\kappa/n}$, und wir schätzen mittels gleichmäßiger

Schranke für $j = 0, \dots, J_n$ weiter ab:

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

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

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

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

$$\mathbb{P}\left(\text{KL}(\mathbb{P} | \mathbb{P}_{\hat{\vartheta}_n}) - \text{KL}(\mathbb{P} | \mathbb{P}_{\vartheta_0}) > M_n^2\right) \leq \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n L(X_i) l^{-1} > M_n\right) \leq n C_L e^{-M_n l / D_L}.$$

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.

4.40 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*

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

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

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

mit einer Zufallsvariablen Z , die die Exponentialungleichung

$$\mathbb{P}(Z \geq \kappa) \leq C_k \left(\exp\left(-\frac{\kappa}{129 \mathbb{E}[L^2] l^{-2}}\right) + e^{-c n^{1/4}} \right)$$

mit Konstanten $c, C_k > 0$ erfüllt.

4.41 Bemerkungen.

- (a) Im Fall eines korrekt spezifizierten Modells gilt bei differenzierbarer Loglikelihood-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 $\sqrt{n}I(\vartheta_0)^{1/2}(\hat{\vartheta}_n - \vartheta_0)$ asymptotisch $N(0, E_k)$ -verteilt ist. Ein wichtiger Schritt dabei war ja die Approximation $\text{KL}(\mathbb{P}_{\vartheta_0} | \mathbb{P}_{\hat{\vartheta}_n}) - \langle I(\vartheta_0)(\hat{\vartheta}_n - \vartheta_0), \hat{\vartheta}_n - \vartheta_0 \rangle \xrightarrow{\mathbb{P}_{\vartheta_0}} 0$. Demnach ist in diesem Fall $Z = n \text{KL}(\mathbb{P}_{\vartheta_0} | \mathbb{P}_{\hat{\vartheta}_n})$ 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 $d(\vartheta, \vartheta') := |\text{KL}(\mathbb{P}_{\vartheta} | \mathbb{P}) - \text{KL}(\mathbb{P}_{\vartheta'} | \mathbb{P})|^{1/2}$ 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 $|\vartheta - \vartheta'|$.

4.42 Beispiel. Es sei (\mathbb{P}_{ϑ}) eine natürliche Exponentialfamilie mit $\ell(\vartheta, x) = \langle \vartheta, T(x) \rangle - A(\vartheta)$. Existiert dann $\vartheta_0 := \text{argmin}_{\vartheta \in \Theta} \text{KL}(\mathbb{P} | \mathbb{P}_{\vartheta})$ und liegt im Inneren von Θ , so gilt $\nabla_{\vartheta} \text{KL}(\mathbb{P} | \mathbb{P}_{\vartheta_0}) = \mathbb{E}[T] - \dot{A}(\vartheta_0) = 0$, 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 $\vartheta_0 \in \text{int}(\Theta)$ aus

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

folgt.

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

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

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

Eine interessante Perspektive ergibt sich, wenn wir andersherum eine allgemeine Lebesguedichte $f : [0, 1] \rightarrow \mathbb{R}$ mit $f > 0$ fast überall aus $X_1, \dots, X_n \sim f$ schätzen wollen. Wir wählen beschränkte Funktionen $(\varphi_j)_{j \geq 1}$, die eine Orthonormalbasis in $L^2([0, 1])$ bilden, wie orthogonale (trigonometrische) Polynome, Splines, Wavelets. Dann gilt natürlich

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

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}_{ϑ_0} mit Dichte

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

indem wir die empirischen Momente bilden:

$$f_k(\hat{\vartheta}_n) = \exp\left(\sum_{j=1}^k \left(\frac{1}{n} \sum_{i=1}^n \varphi_j(X_i)\right) \varphi_j(x) - \hat{A}_k\right),$$

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

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

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 $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell mit Zerlegung $\Theta = \Theta_0 \dot{\cup} \Theta_1$. Jede messbare Funktion $\varphi : \mathcal{X} \rightarrow [0, 1]$ heißt (randomisierter) Test. φ besitzt Niveau $\alpha \in [0, 1]$, falls $\mathbb{E}_\vartheta[\varphi] \leq \alpha$ 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:

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

Ein Test φ ist unverfälscht zum Niveau α , falls φ Niveau α besitzt sowie auf der Alternative $\mathbb{E}_\vartheta[\varphi] \geq \alpha$, $\vartheta \in \Theta_1$, 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, \dots, 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 $H_0 : \mu \leq \mu_0$ 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 \mathcal{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 $T(X_1, \dots, X_n) = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0$. Zu vorgegebenem $\alpha \in (0, 1)$ sei K_α das α -Fraktile der Standardnormalverteilung, d.h. $1 - \Phi(K_\alpha) = \alpha$. Dann besitzt der einseitige Gauß-Test $\varphi(X_1, \dots, X_n) = \mathbf{1}_{\{T(X_1, \dots, X_n) \geq K_\alpha\}}$ 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 $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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

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

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 $p_1(x) \geq kp_0(x)$ für $x \in A := \{\varphi > \varphi'\}$ wegen $\varphi(x) > 0$ sowie $p_1(x) \leq kp_0(x)$ für $x \in B := \{\varphi < \varphi'\}$ wegen $\varphi(x) < 1$. Mit der disjunkten Zerlegung $\mathcal{X} = A \cup B \cup \{\varphi = \varphi'\}$ erhalten wir

$$\begin{aligned} \mathbb{E}_1[\varphi] - \mathbb{E}_1[\varphi'] &= \int_{A \cup B} (\varphi - \varphi') p_1 \geq \int_A (\varphi - \varphi') kp_0 + \int_B (\varphi - \varphi') kp_0 \\ &= k(\mathbb{E}_0[\varphi] - \mathbb{E}_0[\varphi']) \geq 0. \end{aligned}$$

- (b) Wir zeigen im Anschluss, dass es ein $k \geq 0$ gibt mit $\mathbb{P}_0(p_1 \geq kp_0) \geq \alpha$ und $\mathbb{P}_0(p_1 > kp_0) \leq \alpha$ (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 : $\mathbb{E}_0[\varphi] = \mathbf{1} \bullet \mathbb{P}_0(p_1 > kp_0) + \gamma \bullet \mathbb{P}_0(p_1 = kp_0) = \alpha$.

Es bleibt nachzuweisen, dass $k := \inf\{r \geq 0 \mid \rho(r) \leq \alpha\}$ 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 $\lim_{r \rightarrow \infty} \rho(r) = 0$, und k ist endlich. Weiterhin ist $\rho(r) = 1 - \mathbb{P}_0(p_1/p_0 \leq r)$ monoton fallend und rechtsstetig, was aus Eigenschaften der Verteilungsfunktion von p_1/p_0 folgt. Daher gilt $\rho(k) \leq \alpha$ und $\rho(r) > \alpha$ für $r < k$, so dass

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

aus der σ -Stetigkeit folgt. □

5.6 Definition. Es seien $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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 $h(\bullet, \vartheta, \vartheta') : \mathbb{R} \rightarrow \mathbb{R}^+ \cup \{+\infty\}$ 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

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

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 $X \sim \text{Bin}(n, p)$ 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:

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

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 $H_0 : \vartheta \leq \vartheta_0$ 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 $H_0 : \vartheta \leq \vartheta_0$ 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 $\mathbb{E}_{\vartheta_1}[\varphi] \geq \mathbb{E}_{\vartheta_1}[\varphi^*]$ 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 $\mathbb{E}_{\vartheta_0}[\tilde{\varphi}] = \alpha$ gemacht werden kann, bleibt nur noch zu zeigen, dass φ^* das Niveau α für $H_0 : \vartheta \leq \vartheta_0$ 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 $\mathbb{E}_{\vartheta_0}[\varphi^*] \geq \mathbb{E}_{\vartheta_0}[\varphi] = \mathbb{E}_{\vartheta_1}[\varphi^*]$. Wir schließen $\mathbb{E}_{\vartheta_1}[\varphi] \leq \alpha$ 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 $C := \{(\mathbb{E}_0[\varphi], \mathbb{E}_1[\varphi]) \mid \varphi \text{ Test}\} \subseteq [0, 1]^2$. Diese ist konvex (Menge der Tests ist

konvex), abgeschlossen (folgt aus dem Satz von Banach-Alaoglu) und enthält die Diagonale (betrachte konstante Tests). Neyman-Pearson-Tests entsprechen dann gerade der oberen Begrenzungskurve von C .

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

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

die Nebenbedingungen

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

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

Beweis. Wir zeigen, dass φ^* für $\mathbb{P}_1 = \mathbb{P}_{\vartheta_1} \neq \mathbb{P}_0 = \mathbb{P}_{\vartheta_0}$ 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

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

Analoge Äquivalenzen zeigen, dass φ^* die gewünschte Form besitzt, und für jeden Test φ , der die Nebenbedingungen erfüllt, gilt $\mathbb{E}_{\vartheta_1}[\varphi^*] \geq \mathbb{E}_{\vartheta_1}[\varphi]$ für $\vartheta_1 \neq$

ϑ_0 . Jeder unverfälschte Test φ erfüllt wegen dominierter Konvergenz (vergleiche Satz 3.10)

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

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 $\varphi = \alpha$ 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 Test in der Klasse ist. \square

5.14 Beispiel. Es sei $X_1, \dots, X_n \sim N(\vartheta, \sigma^2)$ 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 $Z = \sum_{i=1}^n (X_i - \vartheta) \sim N(0, n\sigma^2)$ unter \mathbb{P}_{ϑ_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

5.15 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 $X_1, \dots, X_n \sim N(\vartheta, \sigma^2)$ getestet wird. Eine weitere wichtige Klasse bilden sogenannte Mehrstichprobentests, wo nur das Verhältnis von Kennwerten (wie Mittelwert) zwischen den Stichproben getestet wird.

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

5.17 Lemma. Die Parametermenge $\Theta = \Theta_0 \dot{\cup} \Theta_1$ bilde einen metrischen Raum und $\partial\Theta_0$ bezeichne den topologischen Rand zwischen Hypothese und Alternative. Gilt für die Verteilungen (\mathbb{P}_ϑ) , dass jeder Test eine stetige Gütefunktion besitzt, und ist φ gleichmäßig bester Test von $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$ unter allen α -ähnlichen Tests auf $\partial\Theta_0$, so ist φ auch gleichmäßig bester unverfälschter Test von $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$.

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

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}_ϑ -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

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

sowie $\alpha \in (0, 1)$ und $\vartheta_0 \in \mathbb{R}$ mit $(\vartheta_0, \tau) \in \text{int}(\Theta)$ 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}_{ϑ_0} -f.s., ein gleichmäßig bester unverfälschter Test zum Niveau α von $H_0 : \vartheta \leq \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$.

5.20 Bemerkung. Der Beweis wird lehren, dass die bedingte Verteilung von φ^* gegeben T unter \mathbb{P}_{ϑ_0} nicht von den Störparametern τ abhängt, diese also für die Wahl von $K(t)$, $\gamma(t)$ unerheblich sind.

Beweis. Die bedingte Dichte von $\mathbb{P}_{\vartheta,\tau}^{U|T=t}$ bezüglich $\mu^{U|T=t}$ (ohne Einschränkung sei μ Wahrscheinlichkeitsmaß, sonst betrachte $\mu = \mathbb{P}_{\bar{\vartheta}}$ für ein $\bar{\vartheta} \in \Theta$) ist gegeben durch

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

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 $H_0 : \vartheta \leq \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$.

Betrachte $\partial\Theta_0 := \{\tau \in \mathbb{R}^k \mid (\vartheta_0, \tau) \in \Theta\}$, das nichtleeres Inneres in \mathbb{R}^k besitzt. In Exponentialfamilien ist die Gütefunktion eines beliebigen Tests stetig, was aus der Stetigkeit von $(\vartheta, \tau) \mapsto \frac{d\mathbb{P}_{\vartheta,\tau}}{d\mu}(x)$ und dem Lemma von Scheffé (eine Anwendung des Lemmas von Fatou, s. van der Vaart, Kor. 2.30) folgt (ohne Details). Also reicht es nach Lemma 5.17, zu zeigen, dass φ^* gleichmäßig bester Test unter allen α -ähnlichen Tests φ auf $\partial\Theta_0$ ist. Da $\partial\Theta_0$ nichtleeres Inneres im \mathbb{R}^k besitzt, ist T suffiziente und bezüglich $\partial\Theta_0$ 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 $(\vartheta, \tau) \in \Theta$

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

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, Seite 149, für Details. \square

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

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

Mit obiger Notation erhalten wir also $\vartheta = \log(a/b)$, $U(x, y) = \sum_i x_i$, $\tau_1 = \log(b)$, $T(x, y) = \sum_i x_i + \sum_j y_j$. Wir können also das reduzierte Testproblem $H_0 : \vartheta \leq 0$ gegen $H_1 : \vartheta > 0$ bei Beobachtung der suffizienten Statistiken U, T betrachten. Nach obigem Satz hat ein gleichmäßig bester unverfälschter Test die Form

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

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

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

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 $\text{Bin}_{0,p}(0) := 1$. Ist also $b(1 - \alpha, t, p)$ das $(1 - \alpha)$ -Quantil der $\text{Bin}(t, p)$ -Verteilung und $\gamma(t) \in [0, 1]$ so gewählt, dass das Niveau ausgeschöpft wird, so ist

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

gleichmäßig bester unverfälschter Test von $H_0 : a \leq b$ 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 $\alpha \leq 1/2$, 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*

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

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

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

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 $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_{\vartheta})_{\vartheta \in \Theta})$. Ist T suffizient und vollständig sowie V ancillary, d.h. \mathbb{P}_{ϑ}^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}_{ϑ} -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 $\varphi = \mathbf{1}_B$, $\psi = \mathbf{1}_C$. \square

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 Suffizienz und Vollständigkeit von T für $\mathbb{P}_{\vartheta_0, \tau}$ auf $\partial\Theta_0$ (vgl. obiger Beweis). \square

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 $\mathbb{E}_{\vartheta=0}[\varphi^*(U, T) | T] =$

α , $\mathbb{E}_{\vartheta=0}[U\varphi^*(U, T) | T] = \alpha \mathbb{E}_{\vartheta=0}[U | T]$. 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 $H_0 : \mu \leq \mu_0$ 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 Lebesguegedichte

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

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 $\varphi^* = \mathbf{1}(|V(X, Y)| > \sqrt{(m^{-1}+n^{-1})/(m+n-2)}t_{t(n+m-2);1-\alpha/2})$ auf $H_0 : \mu = \nu$ gegen $H_1 : \mu \neq \nu$.

5.3 Likelihood-Quotienten- und χ^2 -Test

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

5.26 Definition. Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ 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

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

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 4.15

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

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

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

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. \square

5.30 Bemerkung. Im Fall des zweiseitigen Testproblems aus Satz 5.13 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 4.22 gelten. Es sei $0 \in \text{int}(\Theta)$, und die Hypothesenmenge $\Theta_0 := \{(\vartheta_1, \dots, \vartheta_r, 0, \dots, 0) \in \Theta \mid \vartheta_1, \dots, \vartheta_r \in \mathbb{R}\}$ liege in einem r -dimensionalen Unterraum, $0 \leq r < k$ ($\Theta_0 = \{0\}$ 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}_{ϑ_0} mit $\vartheta_0 \in \Theta_0 \cap \text{int}(\Theta)$ 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 $\Theta_0 \cap \text{int}(\Theta)$ asymptotisch das Niveau $\alpha \in (0, 1)$.

Beweis. Im folgenden sei $\Pi_r : \mathbb{R}^k \rightarrow \mathbb{R}^r$ die Koordinatenprojektion auf die ersten r Koordinaten. Im Beweis von Satz 4.22 haben wir für den MLE $\hat{\vartheta}_n$ insbesondere gezeigt, dass mit $K_n(\vartheta) = -\frac{1}{n} \sum_{i=1}^n \ell(\vartheta, x_i)$ für n hinreichend groß gilt

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

mit Zwischenstellen $\bar{\vartheta}_n, \tilde{\vartheta}_n$. Nun gilt mit (B2) und der Konsistenz von $\hat{\vartheta}_n$ gerade $\ddot{K}_n(\bar{\vartheta}_n) \rightarrow I(\vartheta_0)$, $\ddot{K}_n(\tilde{\vartheta}_n) \rightarrow I(\vartheta_0)$ in \mathbb{P}_{ϑ_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

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

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 $\text{tr}(AB) = \text{tr}(BA)$)

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

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

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

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

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-Likelihood-Schätzers garantiert ist. Für offene $\Theta \subseteq \mathbb{R}^k$ gilt dann Konvergenz unter ganz H_0 , d.h. unter \mathbb{P}_{ϑ_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 $n \rightarrow \infty$ erhalten wir insbesondere Konsistenz des Likelihood-Quotienten-Tests φ_n , das heißt $\lim_{n \rightarrow \infty} \mathbb{E}_\vartheta[\varphi_n] = 1$.

- (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 4.22 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 $n \rightarrow \infty$ greift. Außerdem kann wegen $p_k = 1 - \sum_{i=1}^{k-1} p_i$ als Parametermenge $\Theta = \{p \in [0, 1]^{k-1} \mid \sum_i p_i \leq 1\} \subseteq \mathbb{R}^{k-1}$ 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

$$\lambda_n = \log \left(\frac{\binom{n}{N_1 \dots N_k} (N_1/n)^{N_1} \dots (N_k/n)^{N_k}}{\binom{n}{N_1 \dots N_k} (p_1^0)^{N_1} \dots (p_k^0)^{N_k}} \right) = \sum_{i=1}^k N_i \log(N_i / (np_i^0)).$$

Beachte nun, dass $\mathbb{E}_{p^0}[(N_i - np_i^0)^2 / (np_i^0)] = 1 - p_i^0 \leq 1$ 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

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

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 $p_1^0 = \dots = p_{10}^0 = 0,1$ 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

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

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 $\alpha \leq 0,9254$ akzeptiert hätte! Diese beeindruckende Güte der Daten hat andererseits zum Verdacht der Datenmanipulation geführt.