Mathematische Statistik
Gliederung zur Vorlesung
im Wintersemester 2007/08

Markus Reiß
Universität Heidelberg
reiss@statlab.uni-heidelberg.de

VORLÄUFIGE FASSUNG: 14. Februar 2008

Inhaltsverzeichnis

1 Einführung **1**
1.1 Typische Aufgaben der Statistik 1
1.2 Schätzungen im linearen Modell 1

2 Entscheidungstheorie **3**
2.1 Formalisierung eines statistischen Problems 3
2.2 Minimax- und Bayes-Ansatz 5
2.3 Das Stein-Phänomen 8
2.4 Ergänzungen 9

3 Dominierte Experimente und Suffizienz **9**
3.1 Dominierte Experimente 9
3.2 Exponentialfamilien 10
3.3 Suffizienz 11
3.4 Vollständigkeit 13

4 Testtheorie **14**
4.1 Neyman-Pearson-Theorie 14
4.2 Bedingte Tests 16
4.3 Tests im Normalverteilungsmodell 17

5 Schätztheorie **18**
5.1 Momentenschätzer 18
5.2 Exkurs: Likelihood-Quotienten-Test und χ^2-Test 19
5.3 Maximum-Likelihood- und M-Schätzer 19
5.4 Cramér-Rao-Effizienz 23
5.5 Nichtparametrische Dichteschätzung 24

1 Einführung

1.1 Typische Aufgaben der Statistik

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

1.2 Schätzungen im linearen Modell

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

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

wobei $a, b \in \mathbb{R}$, $\sigma > 0$ unbekannte Parameter, (x_i) bekannte Werte (Versuchsplan, Design) sowie (ε_i) standardisierte Zufallsvariablen (d.h. $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = 1$) sind, die Messfehler modellieren.

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

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

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

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

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

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, dem Fehlerniveau $\sigma > 0$ und einem Vektor $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ von standardisierten Zufallsvariablen, den Fehler- oder Störgrößen. Beobachtet wird eine Realisierung von

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

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

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

1.3 Beispiele.

(a) 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.

(b) Polynomiale Regression: wir beobachten

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

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

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

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

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

$$Y_i = a_0 + a_1 x_{i,1} + \cdots + a_d x_{i,d} + \sigma\varepsilon_i, \quad i = 1, \ldots, n.$$

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

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

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

1.4 Lemma. *Mit Π_X werde die Orthogonalprojektion von $\mathbb{R}^n$ auf den Unterraum $\operatorname{ran}(X)$, das Bild der Designmatrix, bezeichnet. Dann gilt $\Pi_X = X(X^\top X)^{-1}X^\top$ und für den Kleinste-Quadrate-Schätzer $\hat\beta = (X^\top X)^{-1}X^\top Y$. Insbesondere existiert der Kleinste-Quadrate-Schätzer und ist eindeutig.*

1.5 Satz. *Im linearen Modell mit unkorrelierten Fehlergrößen $(\varepsilon_1, \ldots, \varepsilon_n)$ gelten die folgenden Aussagen:*

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

(b) Satz von Gauß-Markov: ist der reelle Parameter $\alpha = \langle \beta, v\rangle$ für ein $v \in \mathbb{R}^k$ zu schätzen, so ist $\hat\alpha = \langle \hat\beta, v\rangle$ ein (in den Daten Y) linearer erwartungstreuer Schätzer, der unter allen linearen erwartungstreuen Schätzern minimale Varianz besitzt.

(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*.*

1.6 Bemerkung. Man sagt, dass der Schätzer $\hat{\alpha}$ im Satz von Gauß-Markov bester linearer erwartungstreuer Schätzer (BLUE: best linear unbiased estimator) ist.

1.7 Beispiel. Sind die Messfehler (ε_i) gemeinsam standardnormalverteilt, so gilt $\hat{\beta} \sim N(\beta, \sigma^2(X^\top X)^{-1})$ und $\hat{\alpha} \sim N(\alpha, \sigma^2 v^\top (X^\top X)^{-1} v)$. Ist weiterhin $\sigma > 0$ bekannt, so ist ein Konfidenzintervall zum Niveau 95% für α gegeben durch

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

Dabei ist der Wert 1,96 gerade das $97,5\%$-Quantil der Standardnormalverteilung, d.h. $\Phi(1,96) = 0,975$. Falls σ unbekannt ist und daher geschätzt werden muss, wird analog zum t-Test vorgegangen, vergleiche Abschnitt 4.3 unten.

2 Entscheidungstheorie

2.1 Formalisierung eines statistischen Problems

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

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

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

2.3 Beispiele.

(a) Beim linearen Modell wähle als Parameterraum $\Theta = \mathbb{R}^k \times \mathbb{R}_+$ mit Parametern $\vartheta = (\beta, \sigma) \in \Theta$. Nun wähle einen Wahrscheinlichkeitsraum $(\Omega, \mathscr{G}, \mathbb{P})$, auf dem der standardisierte Zufallsvektor $\varepsilon : \Omega \to \mathbb{R}^n$ definiert ist. Versieht man den Stichprobenraum $\mathcal{X} = \mathbb{R}^n$ mit seiner Borel-σ-Algebra $\mathscr{F} = \mathscr{B}_{\mathbb{R}^n}$ und setzt $Y_\vartheta = Y_{\beta,\sigma} = \beta X + \sigma\varepsilon$, 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 $\mathbb{P}_\vartheta$ von σ abhängt, spricht man von einem Störparameter.

Beachte, dass bei obiger Modellierung eine feste Verteilung von ε (z.B. Normalverteilung) angenommen wird. Ist realistischerweise auch die Art der Verteilung unbekannt, sollte man statt $(\mathbb{P}_\vartheta)$ die Familie $\mathcal{P} = \{\mathbb{P} \text{ W-Maß auf } \mathscr{F} \mid \int x\,\mathbb{P}(dx) = 0, \int x^2\,\mathbb{P}(dx) = 1\}$ betrachten. In dieser Betrachtungsweise bleibt von einem unendlich-dimensionalen Parameterraum maximal ein $(k+1)$-dimensionaler interessierender Parameter ϑ übrig.

(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 $\mathbb{P}_p = \mathrm{Bin}(100, p)$, $p \in \Theta = [0,1]$, als mögliche Verteilungen. Die Nullhypothese ist $H_0 : p \leqslant p_0$ für den unbekannten Parameter p. Als Aktionsraum dient $A = \{0,1\}$ (H_0 annehmen bzw. verwerfen), und wir wählen den Verlust $l(p,a) = \ell_0 \mathbf{1}_{\{p\leqslant p_0,\, a>p_0\}} + \ell_1 \mathbf{1}_{\{p>p_0,\, a\leqslant p_0\}}$ mit Konstanten $\ell_0, \ell_1 \geqslant 0$. Dies führt auf das Risiko einer Entscheidungsregel (eines Tests) ρ

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

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

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

2.5 Bemerkung. Häufig wird für diese Definition die Menge der betrachteten Entscheidungsregeln eingeschränkt. So ist der Kleinste-Quadrate-Schätzer im linearen Modell nach dem Satz 1.5 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 $\mathscr{F}_\Theta$, die Verlustfunktion l sei produktmessbar und $\vartheta \mapsto \mathbb{P}_\vartheta(B)$ sei messbar für alle $B \in \mathscr{F}$. Die a priori-Verteilung π des Parameters ϑ ist gegeben durch ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathscr{F}_\Theta)$. Das zu π assoziierte Bayesrisiko einer Entscheidungsregel ρ ist

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

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

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

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

2.9 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 $(\Omega, \mathscr{F} \otimes \mathscr{F}_\Theta)$ gemäß $\tilde{\mathbb{P}}(dx, d\vartheta) = P_\vartheta(dx)\pi(d\vartheta)$ (gemeinsame Verteilung von Beobachtung und Parameter). Bezeichne mit X und $\bar{\vartheta}$ die Koordinatenprojektionen von Ω auf $\mathcal{X}$ bzw. Θ. Dann gilt $R_\pi(\rho) = \mathbb{E}_{\tilde{\mathbb{P}}}[l(\bar{\vartheta}, \rho(X))]$.

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

2.11 Satz. *Für jede Entscheidungsregel ρ gilt:*

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

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

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

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

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

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

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

2.13 Korollar. *Für $\Theta \subseteq \mathbb{R}$, $A = \mathbb{R}$ und quadratisches Risiko (d.h. $l(\vartheta, a) = (a - \vartheta)^2$) ist die bedingte Erwartung $\hat{\vartheta}_\pi := \mathbb{E}_{\tilde{\mathbb{P}}}[\bar{\vartheta} \,|\, X]$ Bayes-optimaler Schätzer von ϑ bezüglich der a priori-Verteilung π.*

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

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

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

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

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

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

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

$$p^{\bar{\vartheta} \,|\, X = x}(\vartheta) = \frac{p(x, \vartheta) h(\vartheta)}{\int_\Theta p(x, \vartheta) h(\vartheta) \nu(d\vartheta)} \qquad \textit{(Bayesformel).}$$

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

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

Nach Satz 2.12 finden wir einen Bayestest $\varphi(x)$ als Minimalstelle von

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

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.

2.19 Korollar. *Es sei $X_1, \ldots, X_n$ eine $N(\mu, 1)$-verteilte 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 und minimax als Schätzer von μ.*

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

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

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

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

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

2.22 Beispiel. Es werde $X \sim \mathrm{Bin}(n, p)$ mit $n \geqslant 1$ bekannt und $p \in [0,1]$ unbekannt beobachtet. Gesucht wird ein Bayesschätzer $\hat{p}_{a,b}$ von p unter quadratischem Risiko für die a priori-Verteilung $p \sim B(a, b)$, wobei $B(a, b)$ die Beta-Verteilung mit Parametern $a, b > 0$ 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 $\hat{p}_{a^*,b^*} = \frac{X+\sqrt{n}/2}{n+\sqrt{n}}$ (unabhängig von p!), woraus die Sattelpunktseigenschaft folgt:

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

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

2.3 Das Stein-Phänomen

2.23 Lemma (Stein)**.** *Es sei $f : \mathbb{R}^d \to \mathbb{R}$ eine Funktion, die Lebesgue-f.ü. absolut stetig in jeder Koordinate ist. Dann gilt für $Y \sim N(\mu, \sigma^2 E_d)$ mit $\mu \in \mathbb{R}^d$, $\sigma > 0$, $E_d = \operatorname{diag}(1, \ldots, 1) \in \mathbb{R}^{d\times d}$ und für alle $i = 1, \ldots, d$*

$$\mathbb{E}[(\mu_i - Y_i) f(Y)] = -\sigma^2 \, \mathbb{E}[\tfrac{\partial f}{\partial x_i}(Y)],$$

sofern $\mathbb{E}[|\frac{\partial f}{\partial x_i}(Y)|] < \infty$.

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

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

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

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

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

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

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

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

2.26 Bemerkung. Schätzer wie $\hat{\mu}_{JS}$ und $\hat{\mu}_{JS+}$ heißen Shrinkage-Schätzer, und statt zur Null können sie zu jedem anderen Punkt $x \in \mathbb{R}^d$ hingezogen werden. Der natürliche Schätzer $\bar{X}$ ist stets minimax und für $d \leqslant 2$ auch zulässig. Hingegen ist selbst $\hat{\mu}_{JS+}$ für $d \geqslant 3$ unzulässig (Shao, Strawderman 1994).

2.4 Ergänzungen

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

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

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

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

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

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

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

2.31 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.

2.32 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 Experimente und Suffizienz

3.1 Dominierte Experimente

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

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

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

3.2 Beispiele.

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

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

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

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

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

3.2 Exponentialfamilien

3.5 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein von μ dominiertes Experiment. Dann heißt $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ Exponentialfamilie (in $\eta(\vartheta)$ und T), wenn $k \in \mathbb{N}$, $\eta : \Theta \to \mathbb{R}^k$, $C : \Theta \to \mathbb{R}^+$, $T : \mathcal{X} \to \mathbb{R}^k$ messbar und $h : \mathcal{X} \to \mathbb{R}^+$ messbar existieren, so dass

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

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

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

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

3.6 Bemerkungen.

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

(b) Die Darstellung ist nicht eindeutig, mit einer invertierbaren Matrix $A \in \mathbb{R}^{k\times k}$ erhält man beispielsweise eine Exponentialfamilie in $\tilde\eta(\vartheta) = A\eta(\vartheta)$ und $\tilde T(x) = (A^\top)^{-1} T(x)$. Außerdem kann die Funktion h in das dominierende Maß absorbiert werden: $\tilde\mu(dx) := h(x)\mu(dx)$.

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

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

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

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

3.8 Beispiele.

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

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

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

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

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

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

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

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

3.3 Suffizienz

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

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

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

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

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

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

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

3.14 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.

3.15 Beispiele.

(a) Die Identität $T(x) = x$ ist stets suffizient.

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

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

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

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

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

3.17 Satz. *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Experiment 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 Fehler erster und zweiter Art besitzt, nämlich $\tilde{\varphi} = \mathbb{E}_\bullet[\varphi \,|\, T]$.*

3.4 Vollständigkeit

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

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

3.19 Satz (Lehmann-Scheffé)**.** *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Experiment und $\gamma(\vartheta)$, $\vartheta \in \Theta$, 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.*

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

3.21 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 $\mathscr{Z} = \mathbb{R}^k \times \mathbb{R}^-$ 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 $X_1, \ldots, X_n \sim U([0, \vartheta])$ eine mathematische Stichprobe mit $\vartheta > 0$ unbekannt. Aus der Form $L(x, \vartheta) = \vartheta^{-n} \mathbf{1}_{\{x_{(n)} \leqslant \vartheta\}}$ für $x \in (\mathbb{R}^+)^n$ der Likelihoodfunktion folgt, dass das Maximum $X_{(n)}$ der Beobachtungen suffizient ist. 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.

4 Testtheorie

4.1 Neyman-Pearson-Theorie

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

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

φ heißt gleichmäßig bester unverfälschter Test zum Niveau α, falls φ unverfälscht zum Niveau α ist sowie für alle anderen unverfälschten Tests φ' zum Niveau α obige Ungleichung gilt.

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

4.3 Definition. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein (binäres) statistisches Experiment 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 $k \in \mathbb{R}^+$ und $\gamma(x) \in [0, 1]$ Neyman-Pearson-Test.

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

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

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

4.5 Definition. Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein dominiertes Experiment 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 Dichtequotienten (oder monotonen Likelihoodquotienten) in T, falls

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

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

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

4.6 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.*

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

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

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

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

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

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

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

4.9 Bemerkung. Es gilt darüberhinaus $\mathbb{E}_\vartheta[\varphi^*] \leqslant \mathbb{E}_{\vartheta'}[\varphi^*]$ für alle $\vartheta < \vartheta'$, wobei in den Fällen $\mathbb{E}_\vartheta[\varphi^*] \in (0, 1)$ und $\mathbb{E}_{\vartheta'}[\varphi^*] \in (0, 1)$ sogar die strikte Ungleichung gilt.

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

4.11 Satz (Verallgemeinertes NP-Lemma)**.** *Es seien $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine Exponentialfamilie in $\eta(\vartheta)$ und T, L die zugehörige Likelihoodfunktion sowie $\vartheta_0, \vartheta_1 \in \Theta$ zwei Parameter. Erfüllt ein Test für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_1$ der Form*

$$\varphi(x) = \begin{cases} 1, & \text{falls } L(\vartheta_1, x) > kL(\vartheta_0, x) + lT(x)L(\vartheta_0, x) \\ 0, & \text{falls } L(\vartheta_1, x) < kL(\vartheta_0, x) + lT(x)L(\vartheta_0, x) \\ \gamma, & \text{falls } L(\vartheta_1, x) = kL(\vartheta_0, x) + lT(x)L(\vartheta_0, x) \end{cases}$$

mit $k, l \in \mathbb{R}^+$ und $\gamma \in [0, 1]$ die Nebenbedingungen

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

so maximiert er die Güte $\mathbb{E}_{\vartheta_1}[\varphi]$ in der Menge aller Tests, die diese Nebenbedingungen erfüllen.

4.12 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 $\eta \in C^1(\Theta)$ sei streng monoton (wachsend oder fallend) mit $\eta'(\vartheta_0) \neq 0$. Für $\alpha \in (0, 1)$, $c_1 < c_2$ und $\gamma_1, \gamma_2 \in [0, 1]$ erfülle der Test*

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } T(x) < c_1 \text{ oder } T(x) > c_2 \\ 0, & \text{falls } T(x) \in (c_1, c_2) \\ \gamma_i, & \text{falls } T(x) = c_i,\ i = 1, 2 \end{cases}$$

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 $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$.

4.2 Bedingte Tests

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

$$\forall \vartheta \in \Theta :\ \mathbb{E}_\vartheta[f(T)] = 0 \text{ (und existiert)} \ \Rightarrow \forall \vartheta \in \Theta :\ \mathbb{P}_\vartheta(f(T) = 0) = 1.$$

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

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

4.16 Satz. *Es sei $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine k-parametrische natürliche Exponentialfamilie in T. Enthält $\Theta' \subseteq \Theta$ eine offene Menge im $\mathbb{R}^k$, so ist T suffizient und vollständig bezüglich Θ'.*

4.17 Satz. *Gegeben sei die natürliche Exponentialfamilie*

$$\frac{d\,\mathbb{P}_\vartheta}{d\mu}(x) = C(\vartheta)h(x)\exp\Big(\vartheta^0 U(x) + \sum_{i=1}^{k} \vartheta^i T_i(x)\Big), \quad x \in \mathcal{X},\ \vartheta \in \Theta,$$

sowie $\alpha \in (0,1)$ und ein Punkt ϑ_0 im Innern von Θ. 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 $\mathbb{E}_{\vartheta_0}[\varphi^* \,|\, T] = \mathbb{E}_{\vartheta_0^0}[\varphi^* \,|\, T] = \alpha$ $\mathbb{P}_{\vartheta_0}$-f.s., ein gleichmäßig bester unverfälschter Test zum Niveau α von $H_0 : \vartheta^0 \leqslant \vartheta_0^0$ gegen $H_1 : \vartheta^0 > \vartheta_0^0$ (d.h. $\Theta_0 = \{\vartheta \in \Theta \,|\, \vartheta^0 \leqslant \vartheta_0^0\}$, $\Theta_1 = \{\vartheta \in \Theta \,|\, \vartheta^0 > \vartheta_0^0\}$).

4.18 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

$$\mathbb{E}_{\vartheta_0^0}[\varphi^* \,|\, T] = \alpha \text{ und } \mathbb{E}_{\vartheta_0^0}[U\varphi^* \,|\, T] = \alpha\, \mathbb{E}_{\vartheta_0^0}[U \,|\, T] \quad \mathbb{P}_{\vartheta_0}\text{-f.s.}$$

ein gleichmäßig bester unverfälschter Test zum Niveau α von $H_0 : \vartheta^0 = \vartheta_0^0$ gegen $H_1 : \vartheta^0 \neq \vartheta_0^0$.

4.3 Tests im Normalverteilungsmodell

4.19 Satz. *Es sei $X_1, \ldots, X_n$ eine $N(\mu, \sigma^2)$-verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}$ und $\sigma > 0$ unbekannt. Für $\sigma_0 > 0$ ist ein gleichmäßig bester unverfälschter Test von $H_0 : \sigma \leqslant \sigma_0$ gegen $H_1 : \sigma > \sigma_0$ zum Niveau $\alpha \in (0,1)$ gegeben durch*

$$\varphi^*(X_1, \ldots, X_n) = \begin{cases} 1, & \text{falls } \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \bar{X})^2 > K_\alpha \\ 0, & \text{falls } \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \bar{X})^2 \leqslant K_\alpha \end{cases}$$

mit dem α-Fraktil K_α der $\chi^2(n-1)$-Verteilung:

$$\int_{K_\alpha}^\infty \frac{2^{-(n-1)/2}}{\Gamma((n-1)/2)} z^{(n-1)/2-1} e^{-z/2} dz = \alpha.$$

4.20 Satz. *Es sei $X_1, \ldots, X_n$ eine $N(\mu, \sigma^2)$-verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}$ und $\sigma > 0$ unbekannt. Ein gleichmäßig bester unverfälschter Test von $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$ zum Niveau $\alpha \in (0,1)$ ist gegeben durch den zweiseitigen t-Test*

$$\varphi^*(X) = \mathbf{1}_{\{|t(X)| > K_{\alpha/2}\}}, \quad t(X) := \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}},$$

mit dem $\alpha/2$-Fraktil $K_{\alpha/2}$ der $t(n-1)$-Verteilung :

$$\int_{K_{\alpha/2}}^\infty \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma((n-1)/2)} \Big(1 + \frac{z^2}{n-1}\Big)^{-n/2} dz = \alpha/2.$$

5 Schätztheorie

5.1 Momentenschätzer

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

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

existiert. Gibt es nun eine Borel-messbare Funktion $G : \varphi(\Theta) \to g(\Theta)$ mit $G \circ \varphi = g$ und liegt $\frac{1}{n}\sum_{i=1}^n \psi(x_i)$ in $\varphi(\Theta)$ für alle $x_1, \dots, x_n \in \mathcal{X}$, so heißt $G(\frac{1}{n}\sum_{i=1}^n \psi(x_i))$ (verallgemeinerter) Momentenschätzer für $g(\vartheta)$ mit Momentenfunktionen $\psi_1, \dots, \psi_q$.

5.2 Beispiele.

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

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

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

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

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

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

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

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

Im Fall $|a| < 1$ kann man mit Hilfe des Ergodensatzes auf die Konsistenz von $\hat a_n$ für $n \to \infty$ schließen. Allgemeiner zeigt man leicht, dass $M_n := \sum_{k=1}^n X_{k-1}\varepsilon_k$ ein Martingal bezüglich $\mathscr{F}_n := \sigma(\varepsilon_1, \dots, \varepsilon_n)$ 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

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

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

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

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

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

5.4 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. $\lim_{n\to\infty} \hat{g}_n = g(\vartheta)$ $\mathbb{P}_\vartheta$-f.s.*

5.5 Satz. *Es seien $\vartheta_0 \in \Theta$, $g : \Theta \to \mathbb{R}$ und für hinreichend großes n existiere der Momentenschätzer $\hat{g}_n = G(\frac{1}{n}\sum_{i=1}^n \psi(x_i))$ mit Momentenfunktionen $\psi_j \in L^2(\mathbb{P}_{\vartheta_0})$, $j = 1, \dots, q$. Setze $\Sigma(\vartheta_0) := (\mathrm{Cov}_{\vartheta_0}(\psi_i, \psi_j))_{i,j=1,\dots,q}$. Sofern G in einer Umgebung von $\varphi(\vartheta_0)$ stetig differenzierbar ist 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:*

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

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

5.7 Beispiel. Im Exponentialverteilungsmodell aus Beispiel 5.2 gilt $G'(x) = -(k!/x)^{1/k}(kx)^{-1}$ und $\Sigma(\lambda_0) = \mathrm{Var}_{\lambda_0}(X_i^k) = ((2k)! - (k!)^2)/\lambda_0^{2k}$. Alle Momentenschätzer $\hat{\lambda}_{k,n}$ sind asymptotisch normalverteilt mit Rate $n^{-1/2}$ und Varianz $\sigma_k^2 = \lambda_0^2 k^{-2}((2k)!/(k!)^2 - 1)$. Da $\hat{\lambda}_{1,n}$ die gleichmäßig kleinste asymptotische Varianz besitzt und auf der suffizienten Statistik $\overline{X}$ basiert, wird dieser Schätzer im Allgemeinen vorgezogen.

5.2 Exkurs: Likelihood-Quotienten-Test und χ^2-Test

5.3 Maximum-Likelihood- und M-Schätzer

5.8 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: $\hat{\vartheta}(x) := \mathrm{argmax}_{\vartheta\in\Theta} p_\vartheta(x)$. Dieser Schätzer heißt Maximum-Likelihood-Schätzer (MLE). Bereits im vorliegenden Fall ist weder Existenz noch Eindeutigkeit ohne Weiteres garantiert. Bei Nicht-Eindeutigkeit wählt man einen maximierenden Parameter ϑ nach Belieben

aus. Im Fall einer mathematischen Stichprobe $X_1, \dots, X_n \sim \text{Poiss}(\lambda)$ mit $\lambda > 0$ unbekannt, ergibt sich beispielsweise

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

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

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

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

Man sieht, dass der MLE invariant unter Parametertransformation ist: bei Beobachtung von $X \sim N(\mu, 1)$ erhält man den MLE $\tilde{\mu}(X) = X$ und Einsetzen von $X = \log(Y)$ führt auf dasselbe Ergebnis. Interessanterweise führt die Momentenmethode unter Benutzung von $\mathbb{E}_\mu[Y] = e^{\mu+1/2}$ auf einen anderen Schätzer: $\bar{\mu}(Y) = \log(Y) - 1/2$.

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

Mit $\ell(\vartheta, x) := \log L(\vartheta, x)$ wird die Loglikelihood-Funktion bezeichnet.

5.10 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.

5.11 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 int(Θ) von Θ.*

5.12 Beispiele.

(a) Bei Beobachtung einer Markovkette $(X_0, X_1, \dots, X_n)$ auf dem Zustandsraum $S = \{1, \dots, M\}$ mit parameterunabhängigem Anfangswert $X_0 = x_0$ und unbekannten Übergangswahrscheinlichkeiten $\mathbb{P}(X_{k+1} = j \,|\, X_k = i) = p_{ij}$ 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 das empirische Mittel $\hat{p}_{ij} = N_{ij}/n$.

(b) Beim allgemeinen parametrischen Regressionsmodell mit Beobachtungen

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

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

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

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

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

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

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

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

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

5.15 Beispiele.

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

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

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

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

Der zugehörige M-Schätzer ist der MLE.

(b) Betrachte das Regressionsmodell aus Beispiel 5.12 mit $g_\vartheta : [0,1] \to \mathbb{R}$ stetig, äquidistantem Design $x_i = i/n$ und beliebig verteilten Störvariablen (ε_i). Sind die (ε_i) i.i.d. mit $\mathbb{E}[\varepsilon_i] = 0$ und $\mathbb{E}[\varepsilon_i^4] < \infty$, so folgt leicht aus Tschebyschew-Ungleichung und Riemannscher Summen-Approximation, dass $K_n(\vartheta) = \frac{1}{n} \sum_{i=1}^n (Y_i - g_\vartheta(x_i))^2$ 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 M-Schätzer.

5.16 Satz. *Es sei $(K_n)_{n\geqslant 1}$ ein Kontrastprozess zur Kontrastfunktion K. Dann ist der zugehörige M-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 $\mathbb{P}^n_{\vartheta_0}$-f.s. stetig;*

(A3) $\sup_{\vartheta\in\Theta}|K_n(\vartheta) - K(\vartheta_0, \vartheta)| \xrightarrow{\mathscr{L}} 0$ *unter* $\mathbb{P}^n_{\vartheta_0}$.

5.17 Satz. *Ist $\Theta \subseteq \mathbb{R}^k$ kompakt, $(X_n(\vartheta),\, \vartheta \in \Theta)_{n\geqslant 1}$ 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*

$$\forall \varepsilon, \eta > 0\, \exists \delta > 0 :\ \limsup_{n\to\infty} \mathbb{P}\Big(\sup_{|\vartheta_1 - \vartheta_2|<\delta} |X_n(\vartheta_1) - X_n(\vartheta_2)| \geqslant \varepsilon \Big) \leqslant \eta.$$

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

$$U_n(\vartheta) := \nabla_\vartheta K_n(\vartheta) \ \underline{(Score)}, \qquad V_n(\vartheta) := \nabla^2_\vartheta K_n(\vartheta)$$

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

(a) $\sqrt{n}U_n(\vartheta_0) \xrightarrow{\mathscr{L}} N(0, I(\vartheta_0))$ mit $I(\vartheta_0) \in \mathbb{R}^{k\times k}$ positiv definit.

(b) Aus $\vartheta_n \xrightarrow{\mathscr{L}} \vartheta_0$ folgt $V_n(\vartheta_n) \xrightarrow{\mathscr{L}} V(\vartheta_0)$ mit $V(\vartheta_0) \in \mathbb{R}^{k\times k}$ regulär.

Dann ist der M-Schätzer $\hat{\vartheta}_n$ unter $\mathbb{P}^n_{\vartheta_0}$ asymptotisch normalverteilt:

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

Proof. Aus (A1)-(A3) folgt die Konsistenz von $\hat{\vartheta}_n$. Damit gilt $\hat{\vartheta}_n \in \mathrm{int}(\Theta)$ für hinreichend großes (zufälliges) n wegen $\vartheta_0 \in \mathrm{int}(\Theta)$, woraus folgt $\nabla_\vartheta K_n(\hat{\vartheta}_n) = 0$. Die Taylorentwicklung für n hinreichend groß

$$\begin{aligned}
&\nabla_\vartheta K_n(\hat{\vartheta}_n) - \nabla_\vartheta K_n(\vartheta_0) \\
&= \nabla^2_\vartheta K_n(\vartheta_0)(\hat{\vartheta}_n - \vartheta_0) + \int_0^1 (\nabla^2_\vartheta K_n(\vartheta_0 + (\hat{\vartheta}_n - \vartheta_0)t) - \nabla^2_\vartheta K_n(\vartheta_0))(\hat{\vartheta}_n - \vartheta_0)\, dt
\end{aligned}$$

impliziert mit $H_n := \int_0^1 (\nabla^2_\vartheta K_n(\vartheta_0 + (\hat{\vartheta}_n - \vartheta_0)t) - \nabla^2_\vartheta K_n(\vartheta_0))\, dt$

$$0 - U_n(\vartheta_0) = V_n(\vartheta_0)(\hat{\vartheta}_n - \vartheta_0) + H_n(\hat{\vartheta}_n - \vartheta_0).$$

Da $V(\vartheta_0)$ regulär und die Inversenbildung stetig ist, existiert $(V_n(\vartheta_0) + H_n)^{-1}$ für hinreichend großes n und konvergiert gegen $V(\vartheta_0)^{-1}$ in $\mathbb{P}_{\vartheta_0}$-Wahrscheinlichkeit, so dass

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

aus Slutskys Lemma folgt. □

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

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

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

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

(d) Für $i = 0, 1, 2$ *gibt es* $H_i \in L^1(\mathbb{P}_{\vartheta_0})$ *mit* $\sup_{\vartheta\in\Theta}|\ell(\vartheta, x)| \leqslant H_0(x)$ *und* $\sup_{\vartheta\in U}|\nabla_\vartheta^i \ell(\vartheta, x)| \leqslant H_i(x)$ *für* $i = 1, 2$, $x \in \mathcal{X}$.

(e) Die Fisher-Informationsmatrix $I(\vartheta_0) = \mathbb{E}_{\vartheta_0}[(\nabla_\vartheta \ell(\vartheta_0))(\nabla_\vartheta \ell(\vartheta_0))^\top]$ *ist positiv definit.*

Dann ist der MLE $\hat{\vartheta}_n$ *unter* $\mathbb{P}_{\vartheta_0}^{\otimes n}$ *asymptotisch normalverteilt:*

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

Ferner gilt die Formel $I(\vartheta_0) = -\mathbb{E}_{\vartheta_0}[\nabla_\vartheta^2 \ell(\vartheta_0)]$.

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

5.4 Cramér-Rao-Effizienz

5.21 Satz (Cramér-Rao)**.** *Es sei* $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ *mit* $\Theta \subseteq \mathbb{R}^k$ *ein von* μ *dominiertes Experiment mit Likelihoodfunktion* $L(\vartheta, x)$. *Ferner sei* $g : \Theta \to \mathbb{R}$ *differenzierbar,* $\hat{g}$ *ein erwartungstreuer Schätzer von* $g(\vartheta)$ *sowie*

$$\nabla_\vartheta \int_{\mathcal{X}} h(x) L(\vartheta, x)\, \mu(dx) = \int_{\mathcal{X}} h(x) \nabla_\vartheta L(\vartheta, x)\, \mu(dx), \qquad \vartheta \in \Theta,$$

für $h(x) = 1$ *und* $h(x) = \hat{g}(x)$. *Ist die Fisher-Informationsmatrix* $I(\vartheta)$ *positiv definit, so gilt folgende untere Schranke für das quadratische Risiko von* $\hat{g}$*:*

$$\mathbb{E}_\vartheta[(\hat{g} - g(\vartheta))^2] = \mathrm{Var}_\vartheta(\hat{g}) \geqslant (\nabla_\vartheta g(\vartheta))^\top I(\vartheta)^{-1} \nabla_\vartheta g(\vartheta), \qquad \vartheta \in \Theta.$$

5.22 Bemerkung. Im Allgemeinen wird die Cramér-Rao-Schranke nur erreicht, wenn $(\mathbb{P}_\vartheta)$ eine Exponentialfamilie in T bildet und $g(\vartheta) = \mathbb{E}_\vartheta[T]$ zu schätzen ist.

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

5.5 Nichtparametrische Dichteschätzung

5.24 Definition. Eine Funktion $K : \mathbb{R} \to \mathbb{R}$ heißt Kern (oder Kernfunktion), falls $\int_{-\infty}^{\infty} K(x) = 1$ und $K \in L^2(\mathbb{R})$. Gilt

$$\int_{-\infty}^{\infty} K(x) x^p dx = 0, \quad 1 \leqslant p \leqslant P,$$

sowie $\int |K(x)x^{P+1}|dx < \infty$, so besitzt der Kern K die Ordnung P. Für $h > 0$ setze $K_h(x) := h^{-1}K(h^{-1}x)$. Hierbei wird h als Bandweite bezeichnet.

5.25 Definition. Für reellwertige Beobachtungen $X_1, \dots, X_n$ bezeichnet

$$\hat{f}_{h,n}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \quad x \in \mathbb{R}$$

den Kerndichteschätzer zu gegebenem Kern K mit Bandweite $h > 0$.

5.26 Satz. *Es sei $X_1, \dots, X_n$ eine mathematische Stichprobe gemäß einer Dichte f. Gilt $f \in C^s(\mathbb{R})$ und besitzt der Kern K die Ordnung $P \geqslant s - 1$, so gilt für das quadratische Risiko der Kerndichteschätzung*

$$\forall\, x_0 \in \mathbb{R} : \ \mathbb{E}_f[(\hat{f}_{h,n}(x_0) - f(x_0))^2] \leqslant C(K,s)\|f^{(s)}\|_\infty h^s + \|K\|_{L^2}^2 \|f\|_\infty (nh)^{-1},$$

wobei $C(K, s) > 0$ nur von K und s abhängt.

5.27 Korollar. *Setze für $s \geqslant 1$, $R > 0$*

$$\mathscr{D}(s,R) := \{f : \mathbb{R} \to \mathbb{R}^+ \mid f \in C^s(\mathbb{R}),\ \textstyle\int f(x)dx = 1,\ \max(\|f\|_\infty, \|f^{(s)}\|_\infty) \leqslant R\}.$$

Dann erfüllt der Kerndichteschätzer mit einem Kern der Ordnung $P \geqslant s - 1$ und der Bandweite $h(n) = Cn^{-s/(2s+1)}$, $C > 0$ beliebig, asymptotisch:

$$\forall\, x_0 \in \mathbb{R} : \ \limsup_{n\to\infty} n^{2s/(2s+1)} \sup_{f \in \mathscr{D}(s,R)} \mathbb{E}_f[(\hat{f}_{n,h(n)}(x_0) - f(x_0))^2] < \infty.$$

Insbesondere ergeben sich die Konvergenzraten $n^{-2/3}$ (s=1), $n^{-4/5}$ (s=2) sowie als Grenzwert für $s \to \infty$ die parametrische Rate n^{-1} für das quadratische Risiko.

5.28 Satz. *Es gilt für jedes $x_0 \in \mathbb{R}$ folgende asymptotische untere Schranke:*

$$\liminf_{n\to\infty} n^{2s/(2s+1)} \inf_{\hat{f}_n} \sup_{f\in\mathscr{D}(s,R)} \mathbb{E}_f[(\hat{f}_n(x_0) - f(x_0))^2] > 0,$$

wobei $\hat{f}_n$ einen beliebigen Schätzer basierend auf der mathematischen Stichprobe $X_1, \ldots, X_n$ bezeichnet. Damit ist $n^{-s/(2s+1)}$ die (normalisierte) Minimaxrate im vorliegenden Dichteschätzproblem, und der Kerndichteschätzer ist ratenoptimal.