Ruprecht-Karls-Universität Heidelberg
Institut für Angewandte Mathematik
Prof. Dr. Jan JOHANNES

STATISTIK 1

Gliederung zur Vorlesung
im Wintersemester 2015/16

vorläufige Fassung stand 3. Dezember 2015

Falls Sie **Fehler in der Gliederung** finden, teilen Sie mir diese bitte
per eMail an johannes@math.uni-heidelberg.de mit.

Im Neuenheimer Feld 294, 69120 Heidelberg
Telefon: +49 6221 54.62.76 – Fax: +49 6221 54.53.31
eMail: johannes@math.uni-heidelberg.de
Webseite zur Vorlesung: www.razbaer.eu/jan.johannes/vl/ST1-WS15/

Inhaltsverzeichnis

Kapitel 1

Statistische Inferenz im linearen Modell

1.1 Das lineare Modell

§1.1.1 **Beispiel**. In der folgenden Tabelle ist ein Auszug des „Cars93“ Datensatzes aus dem Statistikpaket R Core Team [2015] (library {MASS}) angegeben. Der Datensatz umfasst unter anderem den Preis, die Anzahl der Zylinder (Zyl.), den Hubraum (Hub.), die Breite sowie das Herkunftsland für 93 in den USA im Jahr 1993 verkauften Autos.

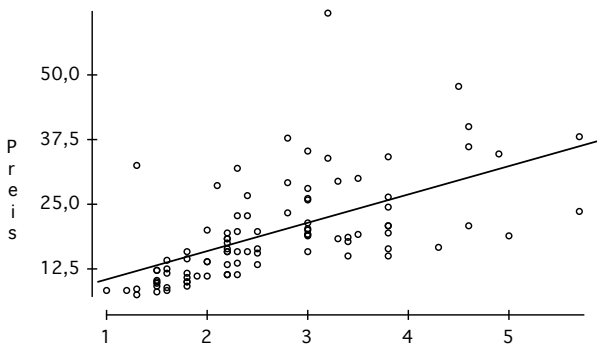
Preis	Zyl.	Hub.	Breite	Herkunft	Preis	Zyl.	Hub.	Breite	Herkunft
15.9	4	1.8	68	non-USA	10	4	1.5	64	non-USA
33.9	6	3.2	71	non-USA	13.9	4	2	69	non-USA
29.1	6	2.8	67	non-USA	47.9	8	4.5	72	non-USA
37.7	6	2.8	70	non-USA	28	6	3	70	non-USA
30	4	3.5	69	non-USA	35.2	6	3	71	non-USA
15.7	4	2.2	69	USA	34.3	6	3.8	73	USA
20.8	6	3.8	74	USA	36.1	8	4.6	77	USA
23.7	6	5.7	78	USA	8.3	4	1.6	66	non-USA
26.3	6	3.8	73	USA	11.6	4	1.8	66	non-USA
34.7	8	4.9	73	USA	16.5	4	2.5	69	non-USA
40.1	8	4.6	74	USA	19.1	6	3	72	non-USA
11.4	4	2.2	68	USA	31.9	4	2.3	67	non-USA
15.1	6	3.4	74	USA	61.9	6	3.2	69	non-USA
15.9	4	2.2	71	USA	14.1	4	1.6	65	USA
16.3	6	3.8	74	USA	14.9	6	3.8	73	USA
16.6	6	4.3	78	USA	10.3	4	1.5	67	non-USA

Preis, Anzahl der Zylinder (Zyl.), Hubraum (Hub.), Breite sowie Herkunftsland von in den USA verkauften Autos.

Sei Y_i der Preis des i-ten Autos mit Hubraum z_{1i} und Breite z_{2i}. Wir nehmen an, die Autos seien austauschbar und es existiert ein linearer Zusammenhang (vgl. nachfolgende Graphik) zwischen dem erwartetem Verhalten des Preises und den erklärenden Variablen Hubraum und Breite:

$$\mathbb{E}Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i}, \quad i = 1, \ldots, 93.$$

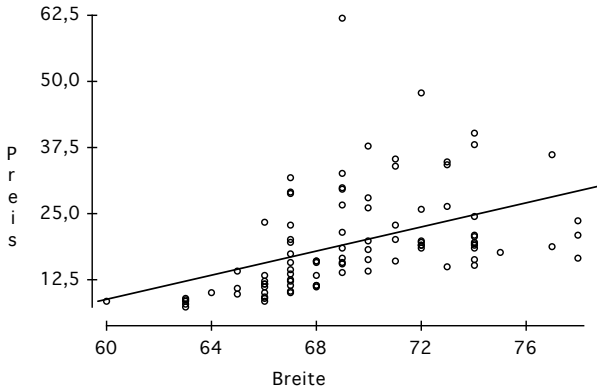
Wir möchten statistische Aussagen über die Parameter β_1 und β_2 treffen, wie zum Beispiel die Werte der Parameter schätzen, Hypothesen der Form $\beta_1 = 0$ oder $\beta_2 = 0$ verifizieren oder den zu Grunde gelegten linearen Zusammenhang überprüfen.



Dependent variable is: **Preis**
No Selector
R squared = 35,7% R squared (adjusted) = 35,0%
s = 7,789 with 93 - 2 = 91 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3063,78	1	3063,78	50,5
Residual	5520,24	91	60,6619	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	4,66919	2,239	2,09	0,0398
Hubraum	5,56294	0,7828	7,11	≤ 0,0001



Dependent variable is: **Preis**
No Selector
R squared = 20,8% R squared (adjusted) = 19,9%
s = 8,644 with 93 - 2 = 91 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1785,15	1	1785,15	23,9
Residual	6798,88	91	74,7129	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-61,3587	16,57	-3,70	0,0004
Breite	1,16565	0,2385	4,89	≤ 0,0001

Dependent variable is: **Preis**
No Selector
R squared = 37,2% R squared (adjusted) = 35,8%
s = 7,737 with 93 - 3 = 90 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3196,80	2	1598,40	26,7
Residual	5387,22	90	59,8581	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	43,6021	26,21	1,66	0,0997
Hubraum	7,58068	1,561	4,86	≤ 0,0001
Breite	-0,638773	0,4285	-1,49	0,1395

Preis in Abhängigkeit des Hubraumes bzw. der Breite des Autos.

□

§1.1.2 **Einfache lineare Regression**. Zu einem vorgegeben (nicht zufälligem) Versuchsplan $z_1, \ldots, z_n \in \mathbb{R}$ beobachten wir Realisierungen der reellwertigen Zufallsvariablen (ZV'en)

$$Y_i = a + bz_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

wobei die zentrierten ZV'en $\{\varepsilon_i\}_{i=1}^n$ (d.h. $\mathbb{E}(\varepsilon_i) = 0$) Messfehler modellieren und $a, b \in \mathbb{R}$ unbekannte Parameter sind. Man denke z.B. an Messungen der Leitfähigkeit Y_i eines Stoffes in Abhängigkeit der Temperatur z_i, eines Effektes Y_i in Abhängigkeit einer Dosierung z_i oder eines Klausurergebnisses Y_i in Abhängigkeit der Klassengröße z_i. Offensichtlich gilt,

$$\mathbb{E}(Y_i) = a + bz_i, \quad i = 1, \ldots, n.$$

so dass ein linearer Zusammenhang nur zwischen der erklärenden Variable x_i und der Erwartung der zu erklärenden zufälligen Größe Y_i zu Grunde gelegt wird. Betrachten wir weiterhin die n-dimensionalen zufälligen Vektoren $Y = (Y_1, \ldots, Y_n)^t$ und $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)^t$, den unbekannten Parametervektor $\beta = (a, b)^t \in \mathbb{R}^2$ sowie die vorgegebene (Design-)Matrix $X = (x_1, \ldots, x_n)^t \in \mathbb{R}^{n \times 2}$ mit Zeilen $x_i^t = (1, z_i)$, $i = 1, \ldots, n$, dann lässt sich die einfache lineare Regression kompakt in der Form $Y = X\beta + \varepsilon$ schreiben. Wir bezeichnen weiterhin mit $\Sigma = \mathbb{C}\mathrm{ov}(\varepsilon) \in \mathbb{R}^{n \times n}$ die Kovarianzmatrix von ε, d.h. für den Eintrag Σ_{ij} in der i-ten Zeile und j-ten Spalte von $\Sigma := (\Sigma_{ij})_{1 \leqslant i,j \leqslant n}$ gilt $\Sigma_{ij} = \mathbb{C}\mathrm{ov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}(\varepsilon_i \varepsilon_j)$. Bezeichnet $\langle v, w \rangle = w^t v$ für $v, w \in \mathbb{R}^2$ das euklidische Skalarprodukt, dann gilt $\mathbb{C}\mathrm{ov}(\langle \varepsilon, v \rangle, \langle \varepsilon, w \rangle) = \langle \Sigma v, w \rangle$. □

§1.1.3 **Bemerkung**. Wir schreiben $\Sigma > 0$, falls Σ eine symmetrische, strikt positiv-definite Matrix ist. Insbesondere, ist dann Σ diagonalisierbar mit $\Sigma = U \Lambda U^t$ für eine Diagonalmatrix

$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ und eine unitäre Matrix U. Für $s \in \mathbb{R}$ setzen wir $\Sigma^s = U\Lambda^s U^t$ mit $\Lambda^s = \operatorname{diag}(\lambda_1^s, \dots, \lambda_n^s)$. Wie erwartet, gilt $(\Sigma^{-1/2})^2 = \Sigma^{-1}$ und somit $\|\Sigma^{-1/2}v\|^2 = \langle \Sigma^{-1}v, v\rangle$. □

§1.1.4 **Definition**. Ein **lineares Modell** beschreibt *adäquat* den Zusammenhang zwischen einem zu erklärenden, zufälligem Vektor (Zielgröße) $Y \in \mathbb{R}^n$ mit $\mathbb{E}\|Y\|^2 < \infty$ und einer erklärenden, vorgegebenen Matrix $X \in \mathbb{R}^{n\times p}$, der *Designmatrix* oder Matrix der Effekte, falls ein Parametervektor $\beta \in \mathbb{R}^p$ existiert, so dass $\mathbb{E}(Y) = X\beta$ gilt. Die Kovarianzmatrix $\Sigma = \mathbb{C}\mathrm{ov}(\varepsilon) \in \mathbb{R}^{n\times n}$ des zentrierten zufälligen Vektors $\varepsilon := Y - X\beta$, den *Fehler- oder Störgrößen*, sowie der Vektor $\beta \in \mathbb{R}^p$ sind unbekannte Parameter in einem linearem Modell. Beobachtet wird eine Realisierung von Y und die Designmatrix X und wir schreiben abkürzend $Y \circlearrowleft \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$. In einem **gewöhnlichen linearen Modell** gilt weiterhin $\Sigma = \sigma^2 \operatorname{Id}_n$ für ein Fehlerniveau $\sigma > 0$, wobei $\operatorname{Id}_n \in \mathbb{R}^{n\times n}$ die Einheitsmatrix bezeichnet. □

§1.1.5 **Beispiele**. (a) Ein zufälliger Vektor $Y \in \mathbb{R}^n$ folgt einem *Lokations-Skalen-Modell*, falls $\mathbb{E}(Y) = \mu\mathbb{1}_n$ mit $\mathbb{1}_n := (1, \dots, 1)^t \in \mathbb{R}^n$ und $\mathbb{C}\mathrm{ov}(Y) = \sigma^2 \operatorname{Id}_n$ gilt. Die unbekannten Parameter sind $\mu \in \mathbb{R}$ als auch $\sigma > 0$. Wir schreiben abkürzend $Y \circlearrowleft \{\mathfrak{L}(\mu\mathbb{1}_n, \sigma^2 \operatorname{Id}_n), \mu \in \mathbb{R}, \sigma > 0\}$. Sind die Koordinaten von Y zusätzlich unabhängige und identisch verteilte (u.i.v.) reellwertige ZV'en, so ist die Verteilung von Y durch das Produkt der eindimensionalen Randverteilungen gegeben und wir schreiben $Y \circlearrowleft \{\mathfrak{L}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$. Wird die Varianz σ_o^2 der Beobachtungen als bekannt vorausgesetzt, so erfüllt der zufällige Vektor Y ein *Lokations-Modell* und wir schreiben abkürzend $Y \circlearrowleft \{\mathfrak{L}(\mu\mathbb{1}_n, \sigma_o^2 \operatorname{Id}_n), \mu \in \mathbb{R}\}$ oder $Y \circlearrowleft \{\mathfrak{L}^{\otimes n}(\mu, \sigma_o^2), \mu \in \mathbb{R}\}$. Wird dagegen der Erwartungswert μ_o als bekannt vorausgesetzt, so folgt der zufällige Vektor Y einem *Skalen-Modell* und wir schreiben abkürzend $Y \circlearrowleft \{\mathfrak{L}(\mu_o\mathbb{1}_n, \sigma^2 \operatorname{Id}_n), \sigma > 0\}$ oder $Y \circlearrowleft \{\mathfrak{L}^{\otimes n}(\mu_o, \sigma^2), \sigma > 0\}$. Setzen wir $\beta = \mu$ und $X = \mathbb{1}_n$ so sind die drei Modelle offensichtlich (gewöhnliche) lineare Modelle.

(b) *Varianzanalyse mit einem Faktor.* Es werden q Proben an p Labore geschickt, wir erhalten zu jeder Probe einen Messwert, die wir als Realisierung von ZV'en

$$Y_{jk} = \mu_j + \varepsilon_{jk}, \quad j = 1, \dots, p, \quad k = 1, \dots, q,$$

auffassen. Ein Anordnen der ZV'en als $n = pq$ dimensionalen Vektor, $Y = (Y_1, \dots, Y_n)^t$ mit $Y_i = Y_{jk}$ und $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$ mit $\varepsilon_i = \varepsilon_{jk}$ für $i = k + (j-1)q$ erlaubt es uns, kompakt $Y = X\beta + \varepsilon$ zu schreiben, wobei $\beta := (\mu_1, \dots, \mu_p)^t$ und $X = \operatorname{Id}_p \otimes \mathbb{1}_q$. Hier bezeichnet $\otimes$ das Kronecker-Produkt, d.h. $A \otimes B := (a_{ij}B)$ für zwei Matrizen A und B. Insbesondere, folgt also der zufällige Vektor Y einem linearen Modell.

(c) Der Zusammenhang zwischen vorgegebenen Designpunkten $z_1, \dots, z_n \in \mathbb{R}$ und einem zufälligem Vektor $Y \in \mathbb{R}^n$ wird durch eine *polynomiale Regression* beschrieben, falls Parameter $a_0, \dots, a_{p-1} \in \mathbb{R}$ existieren, so dass

$$\mathbb{E}(Y_i) = a_0, + a_1 z_i + a_2 z_i^2 + \dots + a_{p-1} z_i^{p-1}, \quad i = 1, \dots, n,$$

gilt. Bezeichnen wir mit $\beta = (a_0, \dots, a_{p-1})^t$ den Vektor der unbekannten Parameter und mit

$$X = \begin{pmatrix} 1 & z_1 & z_1^2 & \cdots & z_1^{p-1} \\ 1 & z_2 & z_2^2 & \cdots & z_2^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_n & z_n^2 & \cdots & z_n^{p-1} \end{pmatrix}$$

die Designmatrix vom Vandermonde-Typ, so gilt $\mathbb{E}(Y) = X\beta$ und es liegt somit ein lineares Modell vor. Für $p > 2$ ist der Zusammenhang zwischen den Designpunkten $\{z_i\}$ und den Beobachtungen $\{Y_i\}$ insbesondere nichtlinear. Auf Grund der linearen Abhängigkeit vom Parametervektor β wird das Modell linear genannt. Eine natürliche Verallgemeinerung der Modellierung eines nichtlinearer Zusammenhang zwischen den Designpunkten $\{z_i\}$ und den Beobachtungen $\{Y_i\}$ ist

$$\mathbb{E}(Y_i) = \beta_1\psi_1(z_i) + \cdots + \beta_p\psi_p(z_i), \quad i = 1, \ldots, n,$$

mit unbekanntem Parametervektor $\beta = (\beta_1, \ldots, \beta_p)^t$ und vorgegebene Basisfunktionen $\{\psi_j\}$, zum Beispiel Splinefunktionen. Setzen wir $X := \big(\psi_k(z_j)\big)_{jk}$ so gilt erneut $\mathbb{E}(Y) = X\beta$ und das zugrunde liegende Modell ist linear. □

§1.1.6 **Definition**. In einem linearen Modell $Y \circlearrowleft \big\{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\big\}$ heißt der Parameter $\beta \in \mathbb{R}^p$ oder allgemeiner der abgeleitete Parameter $\gamma(\beta) \in \mathbb{R}^q$ für eine vorgegebene Funktion $\gamma : \mathbb{R}^p \to \mathbb{R}^q$ *identifizierbar*, falls $\mathbb{E}_{\beta_o}Y = \mathbb{E}_\beta Y$ impliziert $\gamma(\beta_o) = \gamma(\beta)$. □

§1.1.7 **Lemma**. *Sei $Y \circlearrowleft \big\{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\big\}$ und $C \in \mathbb{R}^{q\times p}$ eine vorgegebene Matrix. Der abgeleitete lineare Parameter $\gamma(\beta) := C\beta \in \mathbb{R}^q$ ist genau dann identifizierbar wenn eine Matrix $A \in \mathbb{R}^{q\times n}$ existiert, so dass $C = AX$ gilt.* □

Beweis von Lemma §1.1.7. in der Vorlesung. □

§1.1.8 **Korollar**. *In einem linearen Modell $Y \circlearrowleft \big\{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\big\}$ ist der Parameter $\beta \in \mathbb{R}^p$ genau dann identifizierbar, wenn die Designmatrix X den Rang $\mathrm{rg}[X] = p$ besitzt.* □

Beweis von Korollar §1.1.8. in der Vorlesung. □

§1.1.9 **Bemerkung**. Besitzt in einem linearen Modell die Designmatrix X den Rang $\mathrm{rg}[X] = r < p$, so lässt sich durch eine geeignete Transformation $\gamma = C\beta$ und $\tilde{X} = XU$ für $C \in \mathbb{R}^{r\times p}$ und $U \in \mathbb{R}^{p\times r}$ erreichen, dass γ in dem reparametrisierten linearen Modell $\mathbb{E}Y = \tilde{X}\gamma$ identifizierbar ist. Dies ist genau dann der Fall, wenn $XUC = X$ und $\mathrm{rg}[XU] = r$ gilt. □

§1.1.10 **Beispiele**. (a) (Einfache lineare Regression §1.1.2 fortgesetzt.) Die Parameter a und b sind identifizierbar, falls mindestens zwei Effekte des Versuchsplans $\{z_i\}$ verschieden sind.

(b) (Polynomiale Regression §1.1.5(c) fortgesetzt.) Die Determinante einer Matrix vom Vandermonde-Typ ist im Fall $p = n$ gegeben durch $\prod_{p\geqslant k>j\geqslant 1}(x_k - x_j)$. Damit ist eine hinreichende und notwendige Bedingung für die Identifizierbarkeit des Parameters β, dass mindestens p verschiedene Effekte existieren. □

§1.1.11 **Bemerkung**. Es gibt wichtige *Verallgemeinerungen linearer Modelle* (GLM für *Generalized Linear Model*). Der Zusammenhang zwischen einem zufälligem Vektor $Y \in \mathbb{R}^n$ und einer Designmatrix $X = (x_1, \ldots, x_n)^t \in \mathbb{R}^{n\times p}$ ist durch ein *verallgemeinertes lineares Modell* mit vorgegebener Linkfunktion ℓ beschrieben, falls ein Parametervektor $\beta \in \mathbb{R}^p$ existiert, so dass $\mathbb{E}(Y_i) = \ell(x_i^t\beta)$, $i = 1, \ldots, n$, gilt. Nehmen wir an, dass die ZV Y_i das Auftreten eines positiven oder negativen Effektes nach Verabreichung eines Medikamentes wiedergibt. In diesem Fall ist $Y_i \sim \mathfrak{Bin}(1, \pi_i)$ eine Bernoulli-ZV und die Erfolgswahrscheinlichkeit π_i der unbekanntem Parameter. Eine *logistische Regression* liegt nun vor, falls ein Parametervektor

$\beta \in \mathbb{R}^p$ existiert, so dass $\log(\pi_i/(1-\pi_i)) = x_i^t\beta$ oder äquivalent $\pi_i = \{1+\exp(-x_i^t\beta)\}^{-1}$ für $i = 1, \ldots, n$ gilt. Die Linkfunktion $\ell(x) = \{1+\exp(-x)\}^{-1}$, $x \in \mathbb{R}$, entspricht gerade der logistischen Verteilungsfunktion, so dass wir auch von einem Logitmodell sprechen. Ein weiteres Beispiel, ist das Probitmodell, in dem ℓ der Verteilungsfunktion einer Standardnormalverteilung entspricht. □

1.2 Methode der kleinsten Quadrate

Zur Erinnerung, im Sinne des mittleren quadratischen Fehlers (MSE für *mean squared error*) die beste konstante Approximation einer reellwertigen ZV Z mit $\mathbb{E}(Z^2) < \infty$ ist ihr Erwartungswert $\mu = \mathbb{E}(Z)$, d.h. $\mathbb{E}(Z-\mu)^2 = \min_{a\in\mathbb{R}} \mathbb{E}(Z-a)^2$. Das folgende Lemma verallgemeinert diesen Sachverhalt und motiviert zudem die Methode der kleinsten Quadrate.

§1.2.1 **Lemma**. *In einem linearen Modell* $Y \circlearrowleft \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ *gilt*

$$\beta \in \arg\min_{b\in\mathbb{R}^p} \mathbb{E}\|\Sigma^{-1/2}(Y-Xb)\|^2 \quad :\Leftrightarrow \quad \mathbb{E}\|\Sigma^{-1/2}(Y-X\beta)\|^2 = \min_{b\in\mathbb{R}^p} \mathbb{E}\|\Sigma^{-1/2}(Y-Xb)\|^2. \tag{1.1}$$

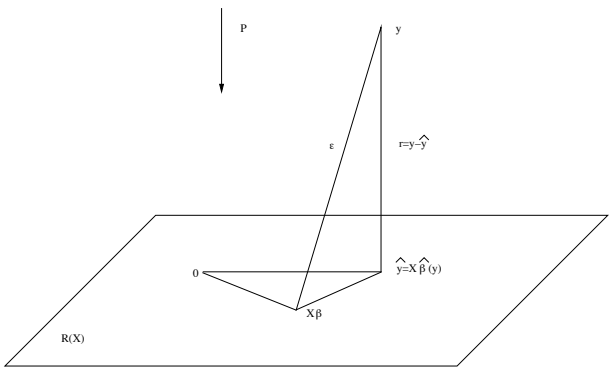
Beweis von Lemma §1.2.1. in der Vorlesung. □

§1.2.2 **Definition**. In einem linearen Modell $Y \circlearrowleft \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ heißt jede (messbare) Wahl von $\widehat{\beta}$, so dass

$$\widehat{\beta} \in \arg\min_{b\in\mathbb{R}^p} \|\Sigma^{-1/2}(Y-Xb)\|^2 \tag{1.2}$$

verallgemeinerter Kleinste-Quadrate-Schätzer (vKQS oder GLSE für *generalized least squares estimator*) des unbekannten Parametervektors β. Im gewöhnlichen Fall ($\Sigma = \sigma^2 \operatorname{Id}_n$) bezeichnen wir $\widehat{\beta}$ als **gewöhnlichen Kleinste-Quadrate-Schätzer** (gKQS oder OLSE für *ordinary least squares estimator*). □

§1.2.3 **Geometrische Interpretation**. Betrachten wir eine Realisierung y der Beobachtung Y als einen Punkt im n-dimensionalen Raum $\mathbb{R}^n$ und variieren wir den Parameter β, so beschreibt $X\beta$ den k-dimensionalen Unterraum $\mathcal{R}(X)$, d.h. eine k-dimensionale Hyperebene durch den Ursprung im $\mathbb{R}^n$. Der gewöhnliche Kleinste-Quadrate-Schätzwert $\hat{\beta}(y)$ gibt uns nun den Punkt $X\hat{\beta}(y)$ auf der Hyperebene, der der Beobachtung y am nächsten liegt. Da die $\mathcal{L}^2$-Norm durch ein Skalarprodukt $\langle\cdot,\cdot\rangle$ induziert ist, bedeutet die Wahl der $\mathcal{L}^2$-Norm als Abstand im $\mathbb{R}^n$, geometrisch, dass wir y orthogonal bzgl. des Skalarproduktes $\langle\cdot,\cdot\rangle$ auf diese Hyperebene projizieren.



□

§1.2.4 **Lemma**. *Setze $\widetilde{X} := \Sigma^{-1/2}X$ sowie $\widetilde{Y} := \Sigma^{-1/2}Y$. Bezeichne mit $\mathcal{R}(\widetilde{X}) := \{\widetilde{X}b : b \in \mathbb{R}^p\}$ den Bildraum der linearen Abbildung $\widetilde{X}$ und mit $\Pi_{\mathcal{R}(\widetilde{X})}$ die orthogonale Projektion von $\mathbb{R}^n$ auf $\mathcal{R}(\widetilde{X})$. Dann sind in einem linearen Modell $Y \circlearrowleft \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ die folgenden Aussagen äquivalent: (i) $\widehat{\beta}$ ist vKQS, d.h. $\widehat{\beta}$ erfüllt (1.2), (ii) $\widetilde{X}\widehat{\beta} = \Pi_{\mathcal{R}(\widetilde{X})}\widetilde{Y}$, (iii) $\widetilde{X}^t\widetilde{X}\widehat{\beta} = \widetilde{X}^t\widetilde{Y}$ („Normalengleichungen"). Insbesondere existiert der vKQS.* □

Beweis von Lemma §1.2.4. in der Vorlesung. □

§1.2.5 **Korollar**. *Sei X eine Designmatrix mit $\operatorname{rg}[X] = p$, dann gilt $\Pi_{\mathcal{R}(\widetilde{X})} = \widetilde{X}(\widetilde{X}^t\widetilde{X})^{-1}\widetilde{X}^t$ und $\widehat{\beta} = (\widetilde{X}^t\widetilde{X})^{-1}\widetilde{X}^t\widetilde{Y} = (X^t\Sigma^{-1}X)^{-1}X^t\Sigma^{-1}Y$ ist der eindeutige vKQS. Weiterhin ist im gewöhnlichen linearen Modell der gKQS $\widehat{\beta} = (X^tX)^{-1}X^tY$ eindeutig und unabhängig von der Kenntnis von σ^2.* □

§1.2.6 **Bemerkung**. Die Matrix $\widetilde{X}^+ := (\widetilde{X}^t\widetilde{X})^{-1}\widetilde{X}^t$ heißt auch *Moore-Penrose-Inverse* von $\widetilde{X}$ und für die vKQS gilt $\widehat{\beta} = \widetilde{X}^+\widetilde{Y}$. □

§1.2.7 **Einfache lineare Regression** (§1.1.2 fortgesetzt). Wir wählen eine alternative Parametrisierung $\beta_1 := a + b\overline{z}$ sowie $\beta_2 := b$ mit $\overline{z} = n^{-1}\sum_{i=1}^n z_i$. Dann gilt

$$Y_i = \beta_1 + \beta_2(z_i - \overline{z}) + \varepsilon_i, \quad i = 1, \dots, n.$$

Setze weiterhin $x_i = (1, z_i - \overline{z})^t$, $i = 1, \dots, n$ und $X = (x_1, \dots, x_n)^t$, so dass $\mathbb{E}(Y) = X\beta$ mit $\beta = (\beta_1, \beta_2)^t$. Wir bestimmen im Folgenden einen gKQS von β, dazu setze $\overline{Y} := n^{-1}\sum_{i=1}^n Y_i$, $S_{zY} := \sum_{i=1}^n (z_i - \overline{z})Y_i = \sum_{i=1}^n (z_i - \overline{z})(Y_i - \overline{Y})$ und $S_{zz} := \sum_{i=1}^n (z_i - \overline{z})^2$, dann gilt

$$X^tY = \sum_{i=1}^n x_iY_i = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n (z_i - \overline{z})Y_i \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ S_{zY} \end{pmatrix}$$
$$X^tX = \sum_{i=1}^n x_ix_i^t = \begin{pmatrix} n & \sum_{i=1}^n (z_i - \overline{z}) \\ \sum_{i=1}^n (z_i - \overline{z}) & \sum_{i=1}^n (z_i - \overline{z})^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & S_{zz} \end{pmatrix}.$$

Somit hat X^tX den vollen Rang falls mindestens zwei $\{z_i\}$ verschieden sind. In dieser Situation ist nach Korollar §1.2.5 der gKQS eindeutig gegeben durch $\widehat{\beta} = (X^tX)^{-1}X^tY = (\overline{Y}, S_{zz}^{-1}S_{zY})^t$ und somit sind $\widehat{a} = \overline{Y} - \widehat{b}\overline{z}$ und $\widehat{b} = S_{zz}^{-1}S_{zY}$ die gKQS von a und b. □

§1.2.8 **Varianzanalyse mit einem Faktor** (§1.1.5 (b) fortgesetzt). Wir bestimmen im Folgenden die gKQS der unbekannten Parameter $\mu_1, \dots, \mu_p$. Bezeichnet $\overline{Y}_{j\bullet} := q^{-1}\sum_{k=1}^q Y_{jk}$, $j = 1, \dots, p$, dann gilt $X^tY = (q\overline{Y}_{1\bullet}, \dots, q\overline{Y}_{p\bullet})$ und $X^tX = q\,\mathrm{Id}_p$. Offensichtlich hat X^tX den vollen Rang so dass $\widehat{\beta} = (X^tX)^{-1}X^tY = (\overline{Y}_{1\bullet}, \dots, \overline{Y}_{p\bullet})^t$ nach Korollar §1.2.5 der eindeutige gKQS von $\beta = (\mu_1, \dots, \mu_p)^t$ ist. □

1.3 Der Satz von Gauß-Markov

§1.3.1 **Satz**. *Besitzt die Designmatrix X den Rang $\operatorname{rg}[X] = p$, so gelten im gewöhnlichen linearen Modell $Y \circlearrowleft \{\mathfrak{L}(X\beta, \sigma^2\,\mathrm{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ die folgenden Aussagen:*

(a) Der gKQS $\widehat{\beta} = (X^tX)^{-1}X^tY$ ist ein erwartungstreuer Schätzer von β (d.h. $\mathbb{E}(\widehat{\beta}) = \beta$).

(b) (Satz von Gauß-Markov) *Unter allen Schätzern des abgeleiteten linearen Parameters $\gamma = \langle \beta, v \rangle$ für ein $v \in \mathbb{R}^p$, die linear (in den Daten Y) und für alle $\beta \in \mathbb{R}^p$ erwartungstreu sind, besitzt der lineare und erwartungstreue Schätzer $\widehat{\gamma} = \langle \widehat{\beta}, v \rangle$ eine minimale Varianz, nämlich $\mathbb{V}\mathrm{ar}(\widehat{\gamma}) = \sigma^2 \|X(X^tX)^{-1}v\|^2$.*

(c) Bezeichnet $R := Y - X\widehat{\beta}$ den Residuenvektor, so ist die geeignet normalisierte Stichprobenvarianz $\widehat{\sigma}^2 := \frac{1}{n-p}\|R\|^2 = \frac{1}{n-p}\|Y - X\widehat{\beta}\|^2$ ein erwartungstreuer Schätzer von σ^2. □

Beweis von Satz §1.3.1. in der Vorlesung. □

§1.3.2 **Bemerkung**. (a) Der Schätzer $\widehat{\gamma}$ im Satz von Gauß-Markov wird bester linearer erwartungstreuer Schätzer (*BLUE* für best linear unbiased estimator) genannt. Verzichtet man auf die Linearität oder Erwartungstreue des Schätzers, so gibt es im Allgemeinen bessere Schätzer im Sinne des mittleren quadratischem Fehlers, zumindest für ausgewählte unbekannte Parameter β bzw. γ. Ein einfacher linearer aber nicht erwartungstreuer Schätzer ist $\widetilde{\gamma} = 0$. Offensichtlich gilt für seinen MSE $\mathbb{E}(\widetilde{\gamma} - \gamma)^2 = \gamma^2$, so dass für alle unbekannten Parameter in einer hinreichend kleine Umgebung um die Null, der MSE von $\widetilde{\gamma}$ strikt kleiner als der MSE des BLUE $\widehat{\gamma}$ ist.

(b) Häufig sind wir nicht am MSE für eine Parameterschätzung im zu Grunde liegenden Modell interessiert, sondern an dem Vorhersagefehler $\|X\widehat{\beta} - X\beta\|^2$. In der Situation einer gewöhnlichen linearen Regression entspricht dies der quadrierten Differenz der vorhergesagten und wahren Werte an den Designpunkten. Der Koordinaten des Vektors $\widehat{Y} = X\widehat{\beta}$ werden angepasste Werte (*fitted values*) genannt. Für den mittleren Vorhersagefehler (MPE für *mean prediction error*) prüft man nun leicht dass

$$\mathbb{E}\|X\widehat{\beta} - X\beta\|^2 = \mathbb{E}\|\Pi_{\mathcal{R}(X)}Y - \Pi_{\mathcal{R}(X)}X\beta\|^2 = \mathbb{E}\|\Pi_{\mathcal{R}(X)}\varepsilon\|^2 = \sigma^2 p.$$

Insbesondere wächst der Vorhersagefehler linear in der Dimension p des Parameterraumes.

(c) Eine entsprechende Aussage des Satzes von Gauß-Markov gilt auch im linearen Modell $Y \circlearrowleft \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma\}$ (Übung!). □

1.4 Die multivariate Normalverteilung

Nicht degenerierte multivariate Normalverteilungen können direkt über ihre Dichte definiert werden. Eine Normalverteilung heißt degeneriert, falls ihre Kovarianzmatrix nicht strikt positiv definit ist (nicht vollen Rang hat). In der Vorlesung werden wir auch Zufallsvariablen mit degenerierten Normalverteilungen betrachten. Beispiele für solche Zufallsvariablen sind Projektionen von nicht degenerierten normalverteilten Zufallsvariablen auf lineare Teilräume. Dies ist etwa der Fall für $X\widehat{\beta}$ im Falle einer deterministischen Designmatrix und unabhängigen normalverteilten Fehlern. Dies wird in der nächsten Sektion behandelt.

§1.4.1 **Lemma**. *Sei $X \in \mathbb{R}^p$ eine ZV mit $\mathbb{E}\|X\|^2 < \infty$. Für alle $b \in \mathbb{R}^q$ und $A \in \mathbb{R}^{q \times p}$ ist dann $Y = AX + b \in \mathbb{R}^q$ eine ZV mit $\mathbb{E}\|Y\|^2 < \infty$. Bezeichnen wir weiterhin mit $\mu := \mathbb{E}(X) \in \mathbb{R}^p$ und $\Sigma := \mathbb{C}\mathrm{ov}(X) \in \mathbb{R}^{p \times p}$ den Erwartungswert und die Kovarianzmatrix von X, dann gilt $\mathbb{E}(Y) = A\mu + b$ und $\mathbb{C}\mathrm{ov}(Y) = A\Sigma A^t$.* □

Beweis von Lemma §1.4.1. in der Vorlesung. □

§1.4.2 **Satz** *(Cramér-Wold). Die Verteilung einer ZV $X \in \mathbb{R}^p$ ist vollständig festgelegt durch die eindimensionalen Verteilungen der linear Formen $\langle X, c\rangle$ für alle $c \in \mathbb{R}^p$.* □

Beweis von Satz §1.4.2. zum Beispiel unter Zuhilfenahme von multivariaten charakteristischen Funktionen, z.Bsp. Theorem 15.55 in Klenke [2008]. □

§1.4.3 **Korollar**. *Die Koordinaten einer ZV $X \in \mathbb{R}^p$ sind genau dann unabhängig und identisch (standardnormal) $\mathfrak{N}(0,1)$-verteilt , wenn für alle $c \in \mathbb{R}^p$ die reellwertige ZV $\langle X, c\rangle$ eine $\mathfrak{N}(0, \langle c, c\rangle)$-Verteilung besitzt, d.h. $\langle X, c\rangle$ ist stetig verteilt mit Dichte*

$$f(x) = \frac{1}{(2\pi\langle c, c\rangle)^{1/2}} \exp\Big(-\frac{x^2}{2\langle c, c\rangle}\Big), \qquad x \in \mathbb{R}.$$ □

Beweis von Korollar §1.4.3. in der Vorlesung. □

§1.4.4 **Definition**. Ein zufälliger Vektor $X \in \mathbb{R}^p$ mit Erwartungswertvektor $\mu \in \mathbb{R}^p$ und Kovarianzmatrix $\Sigma := \mathbb{C}\mathrm{ov}(X) \in \mathbb{R}^{p\times p}$ besitzt eine **multivariate Normalverteilung**, falls für alle $c \in \mathbb{R}^p$ die reellwertige ZV $\langle X, c\rangle$ eine $\mathfrak{N}(\langle \mu, c\rangle, \langle \Sigma c, c\rangle)$-Verteilung besitzt. Wir schreiben dann $X \sim \mathfrak{N}(\mu, \Sigma)$. Die Verteilung $\mathfrak{N}(0, \mathrm{Id}_p) = \mathfrak{N}^{\otimes p}(0,1)$ heißt insbesondere *(p-dimensionale) Standardnormalverteilung*. □

§1.4.5 **Lemma**. *Seien $X \sim \mathfrak{N}(0, \mathrm{Id}_p)$ und $Y \sim \mathfrak{N}(0, \mathrm{Id}_q)$, dann gelten die folgenden Aussagen*

(a) Falls $A \in \mathbb{R}^{m\times p}$ und $B \in \mathbb{R}^{m\times q}$ mit $AA^t = BB^t$ gilt, dann sind die ZV'en $AX \in \mathbb{R}^m$ und $BY \in \mathbb{R}^m$ identisch verteilt.

(b) Falls $U \in \mathbb{R}^{m\times p}$ eine partielle Isometrie ist, dann gilt $UX \sim \mathfrak{N}(0, \Pi_{\mathcal{R}(U)})$.

(c) Falls $A \in \mathbb{R}^{p\times m}$ und $B \in \mathbb{R}^{p\times q}$ mit $A^tB = 0$. Dann sind $\Pi_{\mathcal{R}(A)}X \sim \mathfrak{N}(0, \Pi_{\mathcal{R}(A)})$ und $\Pi_{\mathcal{R}(B)}X \sim \mathfrak{N}(0, \Pi_{\mathcal{R}(B)})$ unabhängig. □

Beweis von Lemma §1.4.5. in der Vorlesung. □

§1.4.6 **Korollar**. *Sei $X \sim \mathfrak{N}(\mu, \Sigma)$, dann gelten die folgenden Aussagen:*

(a) Die i-te Koordinate von X ist $\mathfrak{N}(\mu_i, \Sigma_{ii})$-verteilt.

(b) Die Koordinaten von X sind genau dann unabhängig, wenn sie unkorreliert sind.

(c) Für $A \in \mathbb{R}^{p\times q}$ und $b \in \mathbb{R}^q$ gilt $Y = AX + b \sim \mathfrak{N}(A\mu + \beta, A\Sigma A^t)$.

(d) Ist Σ strikt positiv-definit, dann ist X stetig verteilt mit Lebesgue-Dichte

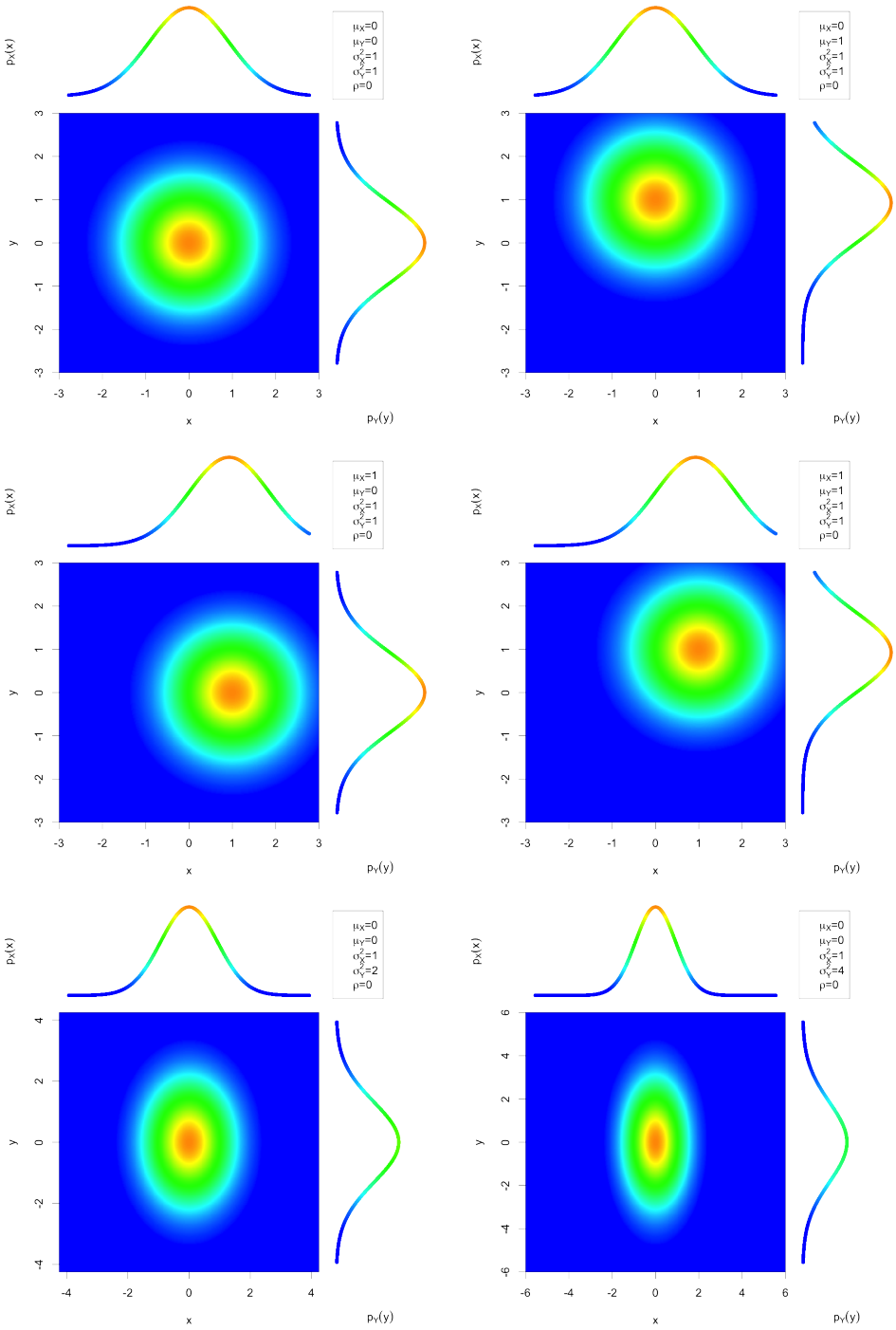
$$f(x) = \frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}} \exp\{-\frac{1}{2}\langle \Sigma^{-1}(x-\mu), (x-\mu)\rangle\}, \qquad x \in \mathbb{R}^p.$$ □

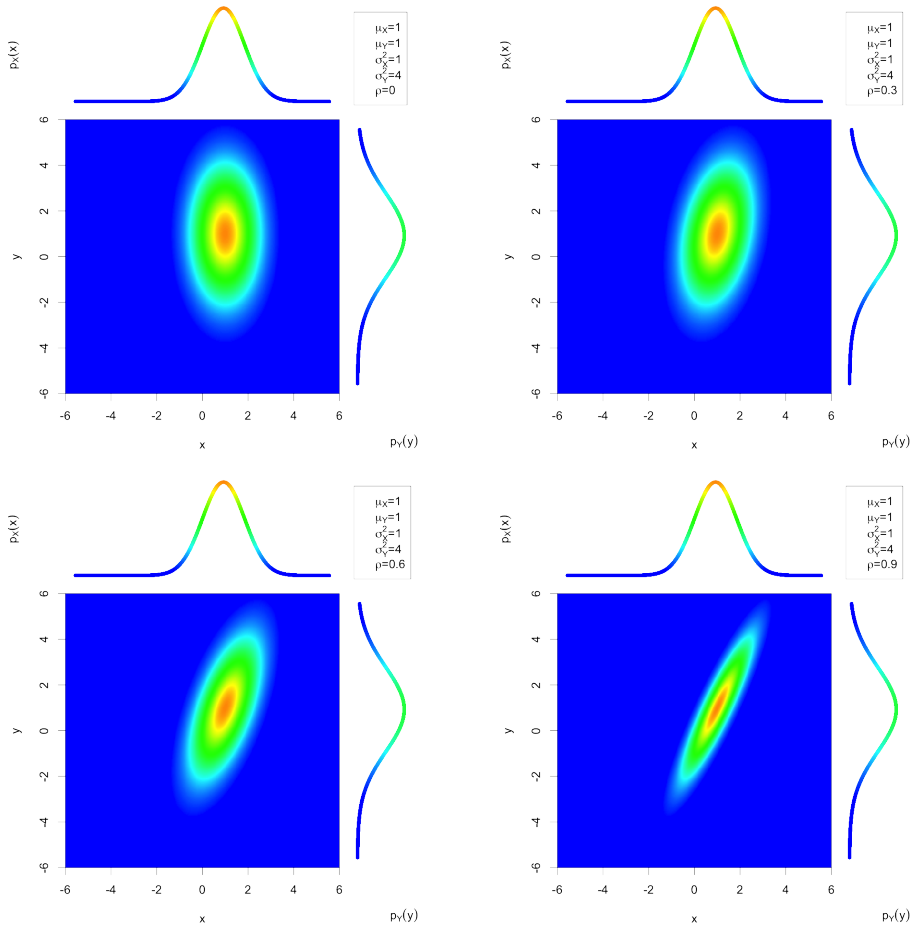
Beweis von Korollar §1.4.6. (Übung). □

§1.4.7 **Beispiel**. Seien X und Y reellwertige ZV'en mit $\mathbb{E}(X^2) < \infty$ und $\mathbb{E}(Y^2) < \infty$. Setze $\mu_X := \mathbb{E}(X)$, $\mu_Y := \mathbb{E}(Y)$, $\sigma_X^2 := \mathbb{V}\mathrm{ar}(X)$, $\sigma_Y^2 := \mathbb{V}\mathrm{ar}(Y)$ und den Korrelationskoeffizienten $\rho := \mathbb{C}\mathrm{orr}(X,Y) = \frac{\mathbb{C}\mathrm{ov}(X,Y)}{\sigma_X\sigma_Y}$. Der zufällige Vektor (X,Y) besitzt eine *bivariate Normalverteilung*, falls für alle Konstanten $a, b \in \mathbb{R}$ die ZV $aX + bY$ eine $\mathfrak{N}\big(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y\big)$-Verteilung besitzt. Die gemeinsame Dichte ist gegeben durch

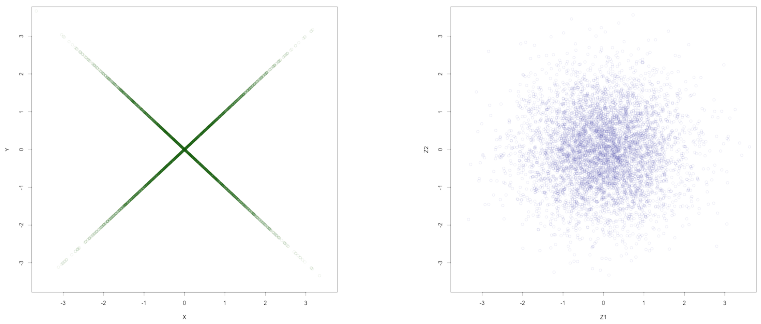
$$p(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\Big(-\frac{(x-\mu_X)^2}{2(1-\rho^2)\sigma_X^2}\Big)$$
$$\times \exp\Big(\frac{2\rho(x-\mu_X)(y-\mu_Y)}{2(1-\rho^2)\sigma_X\sigma_Y}\Big) \times \exp\Big(-\frac{(y-\mu_Y)^2}{2(1-\rho^2)\sigma_Y^2}\Big), \quad x, y \in \mathbb{R}.$$

Die nächsten Graphiken stellen die gemeinsame sowie die marginalen Dichten für verschiedene Werte der Parameter dar:





Besitzt (X, Y) eine *bivariate Normalverteilung* so gilt offensichtlich $X \sim \mathfrak{N}(\mu_x, \sigma_x^2)$ und $Y \sim \mathfrak{N}(\mu_y, \sigma_y^2)$. Sind X und Y weiterhin unkorreliert, d.h. $\rho = 0$, dann sind X und Y unabhängig und es gilt $aX + bY \sim \mathfrak{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$. Insbesondere sind die folgenden beiden Aussagen äquivalent: *(i)* $X \sim \mathfrak{N}(0, \sigma^2)$ und $Y \sim \mathfrak{N}(0, \sigma^2)$ sind unabhängig; *(ii)* $X + Y \sim \mathfrak{N}(0, 2\sigma^2)$ und $X - Y \sim \mathfrak{N}(0, 2\sigma^2)$ sind unabhängig. (Warum?) Es ist natürlich möglich, dass $X \sim \mathfrak{N}(\mu_x, \sigma_x^2)$ und $Y \sim \mathfrak{N}(\mu_y, \sigma_y^2)$ unkorreliert sind, aber der Vektor (X, Y) besitzt keine bivariate Normalverteilung. Betrachte dazu zwei unabhängige ZV'en X und V, wobei $X \sim \mathfrak{N}(0, 1)$ und V ist eine Rademacher-ZV, d.h. $V \in \{-1, 1\}$ mit $P(V = -1) = 1/2 = P(V = 1)$. Es ist nun leicht zu zeigen, dass die ZV'en $Y := VX$ und X unkorreliert sind und dass $Y \sim \mathfrak{N}(0, 1)$ (Übung!). Die ZV'en X und Y sind somit standardnormalverteilt und unkorreliert, aber ihre gemeinsame Verteilung ist keine Normalverteilung (warum?). Die nächsten Graphiken zeigen 5000 Realisierungen von (X, Y) (in grün) und zum Vergleich 5000 Realisierungen einer bivariaten Standardnormalverteilung.

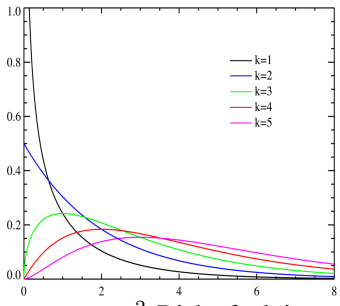


□

§1.4.8 **Definition**. Sei $(Z_1, \dots, Z_k)^t \sim \mathfrak{N}(0, \mathrm{Id}_k)$.
Die Verteilung der ZV

$$Q := \sum_{i=1}^{k} Z_i^2$$

heißt **(zentrale) χ^2-Verteilung** mit k *Freiheitsgraden*. Wir schreiben $Q \sim \chi_k^2$. Für $\alpha \in (0,1)$ bezeichnen wir weiterhin den Wert $\chi_{k,\alpha}^2 \in \mathbb{R}$ als α-Quantil einer (zentralen) χ^2-Verteilung mit k Freiheitsgraden, falls $P(Q \leqslant \chi_{k,\alpha}^2) = \alpha$.



χ_k^2-Dichtefunktionen

Für $\delta \in \mathbb{R}$ heißt die Verteilung der ZV

$$Q := (Z_1 + \delta)^2 + \sum_{i=2}^{k} Z_i^2$$

nichtzentrale χ_k^2-Verteilung mit k *Freiheitsgraden* und *Nichtzentralitätsparameter* δ^2. Wir schreiben $Q \sim \chi_k^2(\delta^2)$ sowie $\chi_{k,\alpha}^2(\delta^2) \in \mathbb{R}$ für das α-Quantil einer nichtzentralen χ^2-Verteilung mit k Freiheitsgraden und Nichtzentralitätsparameter δ^2, d.h. $P\big(Q \leqslant \chi_{k,\alpha}^2(\delta^2)\big) = \alpha$. □

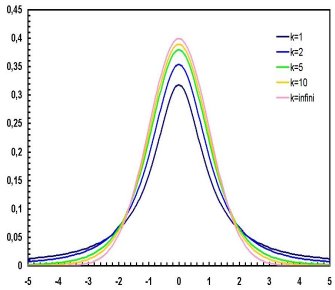
§1.4.9 **Korollar**. *Sei $Q \sim \chi_k^2$ und $W \sim \chi_k^2(\delta^2)$, dann gilt $\mathbb{E}(Q) = k$, $\mathbb{V}\mathrm{ar}(Q) = 2k$ und $\mathbb{E}(W) = \delta^2 + k$. Für $Z \sim \mathfrak{N}(0, \sigma^2 \mathrm{Id}_m)$, $v \in \mathbb{R}^m$ und $A \in \mathbb{R}^{m \times p}$ mit $\mathrm{rg}(A) = p$ gelten außerdem: (i) $\sigma^{-2}\|\Pi_{\mathcal{R}(A)} Z\|^2 \sim \chi_p^2$ und (ii) $\|Z/\sigma + v\|^2 \sim \chi_m^2(\|v\|^2)$.* □

Beweis von Korollar §1.4.9. Übung. □

§1.4.10 **Definition**. Sei $(Z_0, Z_1, \dots, Z_k)^t \sim \mathfrak{N}(0, \mathrm{Id}_{k+1})$.
Die Verteilung der ZV

$$T := \frac{Z_0}{\sqrt{\frac{1}{k} \sum_{i=1}^{k} Z_i^2}}$$

heißt **(Student-) $\mathfrak{t}$-Verteilung** mit k *Freiheitsgraden*. Wir schreiben: $T \sim \mathfrak{t}_k$ und bezeichnen mit $\mathfrak{t}_{k,\alpha}$ das α-Quantil einer Student-$\mathfrak{t}$-Verteilung mit k-Freiheitsgraden, d.h. $P(T \leqslant \mathfrak{t}_{k,\alpha}) = \alpha$.



$\mathfrak{t}_k$-Dichtefunktionen

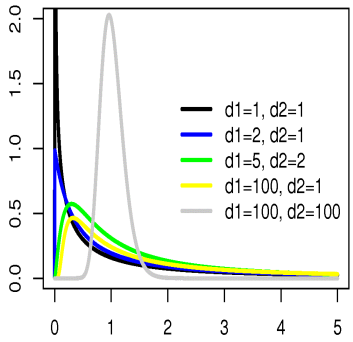
§1.4.11 **Bemerkung**. Die Student-$\mathfrak{t}$-Verteilung mit einem ($k = 1$) Freiheitsgrad entspricht gerade der Cauchy-Verteilung und für $k \to \infty$ konvergiert sie schwach gegen die Standardnormalverteilung (Slutsky-Lemma). Für jedes $k \in \mathbb{N}$ besitzt die $\mathfrak{t}_k$-Verteilung endliche Momente nur bis zur Ordnung $p < k$ (sie ist heavy-tailed). Insbesondere, ist $T \sim \mathfrak{t}_k$ so gilt $\mathbb{E}(T) = 0$ für $k > 1$, sowie $\mathbb{V}\mathrm{ar}(T) = k/(k-2)$ für $k > 2$. □

§1.4.12 **Definition**. Sei $(Z_1, \dots, Z_{m+k})^t \sim \mathfrak{N}(0, \mathrm{Id}_{m+k})$.

Die Verteilung der ZV

$$F := \frac{\frac{1}{m}\sum_{i=1}^{m} Z_i^2}{\frac{1}{k}\sum_{i=m+1}^{m+k} Z_i^2}$$

heißt **zentrale (Fisher-) $\mathfrak{F}$-Verteilung** mit m und k *Freiheitsgraden*. Wir schreiben: $F \sim \mathfrak{F}_{m,k}$ und bezeichnen mit $\mathfrak{F}_{m,k,\alpha}$ das α-Quantil einer zentralen Fisher-$\mathfrak{F}$-Verteilung mit m und k Freiheitsgraden, d.h. $P(F \leqslant \mathfrak{F}_{m,k,\alpha}) = \alpha$.



$\mathfrak{F}_{d1,d2}$-Dichtefunktionen

Für $\delta \in \mathbb{R}$ heißt die Verteilung der ZV

$$F := \frac{\frac{1}{m}\{(Z_1+\delta)^2 + \sum_{i=2}^{m} Z_i^2\}}{\frac{1}{k}\sum_{i=m+1}^{m+k} Z_i^2}$$

nichtzentrale (Fisher-) $\mathfrak{F}$-Verteilung mit m und k *Freiheitsgraden* und *Nichtzentralitätsparameter* δ^2. Wir schreiben $F \sim \mathfrak{F}_{m,k}(\delta^2)$ sowie $\mathfrak{F}_{m,k,\alpha}(\delta^2) \in \mathbb{R}$ für das α-Quantil einer nichtzentralen $\mathfrak{F}$-Verteilung mit m und k Freiheitsgraden und Nichtzentralitätsparameter δ^2, d.h. $P\big(F \leqslant \mathfrak{F}_{m,k,\alpha}(\delta^2)\big) = \alpha$. □

§1.4.13 **Bemerkung**. Sei $F \sim \mathfrak{F}_{m,k}$ mit $k > 1$, dann ist F^{-1} eine $\mathfrak{F}_{k,m}$-verteilte ZV. Für $T \sim \mathfrak{t}_k$ ist T^2 eine $\mathfrak{F}_{1,k}$-verteilte ZV. Weiterhin sei $F_k \sim \mathfrak{F}_{m,k}$, $k \in \mathbb{N}$, dann konvergiert die Folge von ZV'en $(mF_k)_{k\geqslant 1}$ für $k \to \infty$ in Verteilung gegen ein χ^2_m-verteilte ZV. □

1.5 Das normale lineare Modell

§1.5.1 **Definition**. Ein **normales lineares Modell** bezeichnet ein lineares Modell in dem der zu erklärende zufällige Vektor eine multivariate Normalverteilung besitzt. Beobachtet wird eine Realisierung von Y und die Designmatrix X und wir schreiben abkürzend $Y \circledsim \big\{\mathfrak{N}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\big\}$. In einem **gewöhnlichen normalen linearen Modell** gilt weiterhin $\Sigma = \sigma^2 \operatorname{Id}_n$ für ein Fehlerniveau $\sigma > 0$. Im gewöhnlichen Fall sind die Koordinaten des zentrierten Fehlervektors $\varepsilon := Y - X\beta$ unabhängig und identisch $\mathfrak{N}(0, \sigma^2)$-verteilt, d.h. $\varepsilon/\sigma \sim \mathfrak{N}^{\otimes n}(0,1)$. □

§1.5.2 **Satz**. *Besitzt die Designmatrix X den Rang $\operatorname{rg}[X] = p$, so gelten im gewöhnlichen normalen linearen Modell $Y \circledsim \big\{\mathfrak{N}(X\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\big\}$ die folgenden Aussagen:*

(a) Der gKQS ist normalverteilt:

$$\widehat{\beta} \sim \mathfrak{N}(\beta, \sigma^2 (X^t X)^{-1}).$$

(b) Die Stichprobenvarianz $\widehat{\sigma}^2 = \frac{1}{n-p}\|Y - X\widehat{\beta}\|^2$ ist nach geeigneter Normalisierung χ^2-verteilt mit $n-p$ Freiheitsgraden:

$$(n-p)\,\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}.$$

(c) Der gKQS $\widehat{\beta}$ und die Stichprobenvarianz $\widehat{\sigma}^2$ sind unabhängig.

(d) Der zentrierte und geeignet normalisierte gKQS $\widehat{\beta}$ hat eine $\mathfrak{t}$-Verteilung mit $n-p$ Freiheitsgraden: für $v \in \mathbb{R}^p$

$$\frac{\langle \widehat{\beta} - \beta, v\rangle}{\widehat{\sigma}|\langle (X^tX)^{-1}v, v\rangle|^{1/2}} \sim \mathfrak{t}_{n-p}.$$

(e) Der Vorhersagefehler $\|X(\beta - \widehat{\beta})\|^2$ ist nach geeigneter Normalisierung $\mathfrak{F}$-verteilt mit p und $n-p$ Freiheitsgraden:

$$\frac{\|X(\widehat{\beta} - \beta)\|^2}{p\widehat{\sigma}^2} \sim \mathfrak{F}_{p,n-p}.$$

Beweis von Satz §1.5.2. in der Vorlesung. □

§1.5.3 **Korollar**. *Unter den Annahmen und den Notationen des Satzes* §1.5.2 *gelten folgende Konfidenzaussagen für gegebenes $\alpha \in (0,1)$:*

(a) Konfidenzbereich für β: Bezeichnet $\mathfrak{F}_{p,n-p,1-\alpha}$ das $(1-\alpha)$-Quantil einer $\mathfrak{F}$-Verteilung mit p und $n-p$ Freiheitsgraden, so ist

$$C_\alpha = \left\{\beta \in \mathbb{R}^p : \|X(\widehat{\beta} - \beta)\|^2 \leqslant p\widehat{\sigma}^2\mathfrak{F}_{p,n-p,1-\alpha}\right\}$$

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

(b) Konfidenzbereich für $\langle\beta, v\rangle$: Bezeichnet $\mathfrak{t}_\alpha := \mathfrak{t}_{n-p,1-\alpha/2} = -\mathfrak{t}_{n-p,\alpha/2}$ das $(1-\alpha/2)$-Quantil einer $\mathfrak{t}$-Verteilung mit $n-p$ Freiheitsgraden, so ist

$$I_{v,\alpha} = \left[\langle\widehat{\beta}, v\rangle - \mathfrak{t}_\alpha\widehat{\sigma}|\langle (X^tX)^{-1}v, v\rangle|^{1/2}, \langle\widehat{\beta}, v\rangle + \mathfrak{t}_\alpha\widehat{\sigma}|\langle (X^tX)^{-1}v, v\rangle|^{1/2}\right]$$

ein Konfidenzintervall zum Niveau $1-\alpha$ für $\langle\beta, v\rangle$.

§1.5.4 **Beispiel** (§1.1.5 (a) fortgesetzt). In einem *normalen Lokations-Skalen-Modell* $Y \circlearrowleft \left\{\mathfrak{N}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\right\}$ ist

$$I_{v,\alpha} = \left[\overline{Y} - \mathfrak{t}_{n-1,1-\alpha/2}n^{-1/2}\widehat{\sigma}, \overline{Y} + \mathfrak{t}_{n-1,1-\alpha/2}n^{-1/2}\widehat{\sigma}\right]$$

mit $\widehat{\sigma}^2 = \frac{1}{n-1}\|Y - \overline{Y}\mathbb{1}_n\|^2$ ein Konfidenzintervall zum Niveau $1-\alpha$ für den unbekannten Parameter μ. Dies folgt direkt aus Korollar §1.5.3 (b) mit $p = 1$, $v = 1$ und $\gamma = \mu$. □

§1.5.5 **Korollar**. *Unter den Annahmen und den Notationen des Satzes* §1.5.2 *kann für ein $r \in \mathbb{R}$ die lineare Hypothese $H_0 : \langle\beta, v\rangle = r$ gegen die Alternativen (a) $H_A : \langle\beta, v\rangle > r$; (b) $H_A : \langle\beta, v\rangle < r$ sowie (c) $H_A : \langle\beta, v\rangle \neq r$ mit Hilfe der Teststatistik $T := \frac{\langle\widehat{\beta}, v\rangle - r}{\widehat{\sigma}|\langle (X^tX)^{-1}v, v\rangle|^{1/2}}$ und den Entscheidungsregeln*

(a) lehne die Hypothese H_0 ab, falls $T > \mathfrak{t}_{n-p,1-\alpha}$;

(b) lehne die Hypothese H_0 ab, falls $T < -\mathfrak{t}_{n-p,1-\alpha}$;

(c) lehne die Hypothese H_0 ab, falls $|T| > \mathfrak{t}_{n-p,1-\alpha/2}$;

unter Einhaltung des vorgegebenen Niveau $\alpha \in (0,1)$ getestet werden.

§1.5.6 **Beispiel** (§1.5.4 fortgesetzt). In einem *normalen Lokations-Skalen-Modell* $Y \circlearrowleft \left\{\mathfrak{N}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\right\}$ kann die *Hypothese $H_0 : \mu = \mu_o$* gegen die Alternativen (a) $H_A : \mu > \mu_o$; (b) $H_A : \mu < \mu_o$ sowie (c) $H_A : \mu \neq \mu_o$ mit Hilfe der Entscheidungsregeln

(a) lehne die Hypothese H_0 ab, falls $\overline{Y} - \mu_o > \mathfrak{t}_{n-1,1-\alpha} n^{-1/2}\widehat{\sigma}$;

(b) lehne die Hypothese H_0 ab, falls $\overline{Y} - \mu_o < \mathfrak{t}_{n-1,1-\alpha} n^{-1/2}\widehat{\sigma}$;

(c) lehne die Hypothese H_0 ab, falls $|\overline{Y} - \mu_o| > \mathfrak{t}_{n-1,1-\alpha/2} n^{-1/2}\widehat{\sigma}$;

unter Einhaltung des vorgegebenen Niveau $\alpha \in (0,1)$ getestet werden. □

1.6 Asymptotische Theorie

Wir untersuchen nun die Verteilung des Kleinste-Quadrate-Schätzers im Grenzfall, in dem die Anzahl der Beobachtungen gegen unendlich geht. Dazu sei $(Y_n)_{n\in\mathbb{N}}$ eine Folge von zufälligen Zielgrößen und $(x_n)_{n\in\mathbb{N}}$ eine Folge von erklärenden Effekten. Wir nehmen an, dass für alle $n \geqslant n_0$ der Zusammenhang zwischen dem zufälligen Vektor $Y_{(n)} := (Y_1, \ldots, Y_n)^t$ und der Designmatrix $X_{(n)} = (x_1, \ldots, x_n)^t$ adäquat durch ein gewöhnliches lineares Modell beschrieben ist, d.h. $Y_{(n)} \circledcirc \{\mathfrak{L}(X_{(n)}\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$.

§1.6.1 **Satz**. *Sei $Y_{(n)} \circledcirc \{\mathfrak{L}(X_{(n)}\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ mit $\operatorname{rg}[X_{(n)}] = p$ für alle $n \geqslant n_0$. Gelten die folgenden drei Bedingungen:*

(i) $\{Y_n - x_n^t\beta, n \in \mathbb{N}\}$ sind unabhängige und identisch verteilte (u.i.v.) ZV'en.

(ii) Für den kleinsten Eigenwert $\lambda_{(n)}$ der Matrix $X_{(n)}^t X_{(n)}$ gilt $\lim_{n\to\infty} \lambda_{(n)} = \infty$.

(iii) Für die Diagonalelemente der Matrix $P_{(n)} := X_{(n)}(X_{(n)}^t X_{(n)})^{-1} X_{(n)}^t$ gilt $\lim_{n\to\infty} \max_{j=1,\ldots,n} [P_{(n)}]_{jj} = 0$.

Dann ist der Kleinste-Quadrate-Schätzer $\widehat{\beta}_{(n)} := (X_{(n)}^t X_{(n)})^{-1} X_{(n)}^t Y_{(n)}$ konsistent für β und

$$\frac{1}{\sigma}(X_{(n)}^t X_{(n)})^{1/2}(\widehat{\beta}_{(n)} - \beta) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, \operatorname{Id}_p)$$

(konvergiert in Verteilung gegen eine k-dimensionale Standardnormalverteilung) und weiterhin gilt für $v \in \mathbb{R}^p$

$$\frac{\langle \widehat{\beta}_{(n)} - \beta, v\rangle}{\sigma |\langle (X_{(n)}^t X_{(n)})^{-1} v, v\rangle|^{1/2}} \xrightarrow{\mathcal{L}} \mathfrak{N}(0,1).$$

Gilt zusätzlich $\mathbb{E}(Y_1 - x_1^t\beta)^4 < \infty$, dann ist $\widehat{\sigma}^2 = \frac{1}{n-p}\|Y_{(n)} - X\widehat{\beta}_{(n)}\|^2$ ein konsistenter Schätzer für σ^2. □

Beweis von Satz §1.6.1. in der Vorlesung. □

§1.6.2 **Bemerkung**. Die Bedingung §1.6.1 (ii) besagt, dass man mit wachsendem n immer mehr Information bekommt. Weiterhin dominiert kein Vektor von Effekten x_j die anderen unter der Bedingung §1.6.1 (iii). □

§1.6.3 **Korollar**. *Unter den Annahmen und den Notationen des Satzes §1.6.1 gilt folgende asymptotische Konfidenzaussage für gegebenes $\alpha \in (0,1)$. Bezeichnet $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil einer $\mathfrak{N}(0,1)$-Verteilung, so ist*

$$I_{v,\alpha} = \left[\langle \widehat{\beta}_{(n)}, v\rangle - z_{1-\alpha/2}\widehat{\sigma}|\langle (X_{(n)}^t X_{(n)})^{-1} v, v\rangle|^{1/2}, \langle \widehat{\beta}_{(n)}, v\rangle + z_{1-\alpha/2}\widehat{\sigma}|\langle (X_{(n)}^t X_{(n)})^{-1} v, v\rangle|^{1/2}\right]$$

ein Konfidenzintervall zum asymptotischen Niveau $1-\alpha$ für $\langle \beta, v\rangle$.

§1.6.4 **Beispiel** (§1.1.5 (a) fortgesetzt). Sei $Y_{(n)} \circlearrowleft \left\{\mathfrak{L}^{\otimes^n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\right\}$ durch ein *Lokations-Skalen-Modell* mit u.i.v. Koordinaten adäquat beschrieben, dann ist

$$I_{v,\alpha} = \left[\overline{Y}_{(n)} - z_{1-\alpha/2} n^{-1/2} \widehat{\sigma}, \overline{Y}_{(n)} + z_{1-\alpha/2} n^{-1/2} \widehat{\sigma}\right]$$

mit $\overline{Y}_{(n)} = \frac{1}{n} \sum_{i=1}^n Y_i$ ein Konfidenzintervall zum asymptotischen Niveau $1 - \alpha$ für den unbekannten Parameter μ. Dies folgt direkt aus Korollar §1.6.3 mit $v = 1$. □

§1.6.5 **Korollar**. *Unter den Annahmen und den Notationen des Satzes §1.6.1 kann für ein $r \in \mathbb{R}$ die lineare Hypothese $H_0 : \langle \beta, v \rangle = r$ gegen die Alternativen (a) $H_A : \langle \beta, v \rangle > r$; (b) $H_A : \langle \beta, v \rangle < r$ sowie (c) $H_A : \langle \beta, v \rangle \neq r$ mit Hilfe der Teststatistik $T := \frac{\langle \widehat{\beta}, v \rangle - r}{\widehat{\sigma} |\langle (X^t X)^{-1} v, v \rangle|^{1/2}}$ und den Entscheidungsregeln*

(a) lehne die Hypothese H_0 ab, falls $T > z_{1-\alpha}$;

(b) lehne die Hypothese H_0 ab, falls $T < -z_{1-\alpha}$;

(c) lehne die Hypothese H_0 ab, falls $|T| > z_{1-\alpha/2}$;

unter Einhaltung des vorgegebenen asymptotischen Niveau $\alpha \in (0, 1)$ getestet werden.

§1.6.6 **Beispiel** (§1.6.4 fortgesetzt). Sei $Y \circlearrowleft \left\{\mathfrak{L}^{\otimes^n}(\mu, \sigma^2), \mu \in \mathbb{R}\right\}$ durch ein *Lokations-Skalen-Modell* mit u.i.v. Koordinaten adäquat beschrieben, dann kann die *Hypothese $H_0 : \mu = \mu_o$* gegen die Alternativen (a) $H_A : \mu > \mu_o$; (b) $H_A : \mu < \mu_o$ sowie (c) $H_A : \mu \neq \mu_o$ mit Hilfe der Entscheidungsregeln

(a) lehne die Hypothese H_0 ab, falls $\overline{Y}_{(n)} - \mu_o > z_{1-\alpha} n^{-1/2} \sigma$;

(b) lehne die Hypothese H_0 ab, falls $\overline{Y}_{(n)} - \mu_o < z_{1-\alpha} n^{-1/2} \sigma$;

(c) lehne die Hypothese H_0 ab, falls $|\overline{Y}_{(n)} - \mu_o| > z_{1-\alpha/2} n^{-1/2} \sigma$;

unter Einhaltung des vorgegebenen asymptotischen Niveau $\alpha \in (0, 1)$ getestet werden. □

1.7 Residuenanalyse

Wir nehmen im Folgenden an, dass der Zusammenhang zwischen der Zielgröße Y und der Designmatrix durch ein gewöhnliches lineares Modell $Y \circlearrowleft \{\mathfrak{L}(X\beta, \sigma^2 \operatorname{Id}_n)\}$ adäquat dargestellt ist. Bezeichnen wir mit $\overline{Y}$ das arithmetische Mittel der Beobachtung, so ist die totale Quadratsumme $\|Y - \overline{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2$ (SST für *total sum of squares*) ein Maß der Variabilität der Realisierungen der Zielgrößen. Wir wollen nun untersuchen in wie weit diese Variabilität durch die Variabilität der angepassten Schätzwerte $\widehat{Y} = X\widehat{\beta}$ oder der Residuen $Y - \widehat{Y}$ erklärt wird. Eine einfache Zerlegung der totalen Quadratsumme in eine Quadratsumme der Regression bzgl. der angepassten Werte (SSR für *regression sum of squares*) und eine Quadratsumme der Residuen (SSE für *error sum of squares*) ergibt

$$SST := \|Y - \overline{Y}\mathbb{1}_n\|^2 = \|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2 + \|Y - \widehat{Y}\|^2 =: SSR + SSE.$$

Offensichtlich, spricht ein im Verhältnis zum SSR kleiner Wert des SSE für eine gute Anpassung des linearen Modells. Betrachten wir den standardisierten Quotienten

$$F = \frac{\frac{1}{p} SSR}{\frac{1}{n-p} SSE},$$

so sprechen große Werte von F für eine gute Anpassung des linearen Modells. Nehmen wir zusätzlich an, dass die Beobachtung Y normalverteilt ist, so vergleichen wir die Anpassung in dem linearen Modell $Y \circlearrowleft \{\mathfrak{N}(X\beta, \sigma^2 \operatorname{Id}_n)\}$ mit der in einem Lokations-Skalen-Modell $Y \circlearrowleft \left\{\mathfrak{N}^{\otimes^n}(\mu, \sigma^2)\right\}$. Unter der Annahme, dass $\mathbb{1} \in \mathcal{R}(X)$ gilt, hat F^* eine $\mathfrak{F}$-Verteilung mit $(p, n-p)$ Freiheitsgraden.
Alternativ können wir des Verhältnis zwischen der totalen Variabilität und der Variabilität der Schätzwerte betrachten:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Der Wert R^2 wird *Bestimmtheitsmaß* genannt und entspricht im Fall $k = 1$ dem Quadrat des *empirischen Korrelationskoeffizienten*

$$\rho = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2} \cdot \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}.$$

Bezeichnet $\widehat{\beta}_{-i}$ den gewöhnliche Kleinste-Quadrate-Schätzer ohne die i-te Koordinate der Beobachtung Y, dann gilt

$$\widehat{\beta}_{-i} - \hat{\beta} = -\frac{\widehat{Y}_i - Y_i}{1 - [X(X^tX)^{-1}X]_{ii}} (X^tX)^{-1} x_i.$$

Wir sehen also, dass der Einfluss der i-ten Beobachtung sowohl vom i-ten Residuum als auch vom Diagonalelement $[X(X^tX)^{-1}X]_{ii}$, seinem *Leverage-Score*, abhängt. Um einflussreiche Beobachtungen zu entdecken, plottet man daher oft die Residuen R_i gegen die $[X(X^tX)^{-1}X]_{ii}$. Basierend auf der Differenz der geschätzten Parameter ist die *Cook-Distanz* definiert durch

$$\frac{1}{p\hat{\sigma}^2} \|\widehat{\beta}_{-i} - \widehat{\beta}\|_{X^tX} = \frac{1}{k\hat{\sigma}^2} \frac{(\widehat{Y}_i - Y_i)^2}{1 - [X(X^tX)^{-1}X]_{ii}} \frac{[X(X^tX)^{-1}X]_{ii}}{1 - [X(X^tX)^{-1}X]_{ii}}.$$

Sie ist eine einfache Funktion von $[X(X^tX)^{-1}X]_{ii}$ sowie dem Quadrat des studentisierten Residuums $(\hat{Y}_i - Y_i)/\sqrt{\hat{\sigma}^2(1 - [X(X^tX)^{-1}X]_{ii})}$ welche Student-t-verteilt ist unter einer Normalverteilungsannahme. Sie wird häufig als diagnostisches Hilfsmittel verwendet. Diejenigen Beobachtungen, bei denen die Cook-Distanz deutlich größer ist als beim Rest, sollte besonders betrachtet, bzw. in der Analyse weggelassen werden. Analog erhält man als Änderung beim Hinzufügen einer Beobachtung Y_{n+1} zum Effekt x_{n+1}:

$$\frac{1}{1 + x_{n+1}^t (X^tX)^{-1} x_{n+1}} (X^tX)^{-1} x_{n+1} (Y_{n+1} - x_{n+1}^t \widehat{\beta}).$$

Durch Hinzufügen einer einzigen Beobachtung kann somit der Kleinste-Quadrate-Schätzer beliebig verändern werden.

Kapitel 2

Entscheidungstheorie

2.1 Formalisierung eines statistischen Problem

§2.1.1 **Definition**. Sei $\mathcal{P}_\Theta := \{P_\theta, \theta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsmaßen auf einem messbarem Raum $(\mathcal{X}, \mathscr{A})$. Die Indexmenge $\Theta \neq \emptyset$ wird Parametermenge genannt und $\mathcal{X}$ heißt Stichprobenraum. Ist X ein ZV mit Werten in $(\mathcal{X}, \mathscr{A})$ so schreiben wir abkürzend $X \circlearrowleft \mathcal{P}_\Theta$, falls $X \sim P_\theta$ für ein $\theta \in \Theta$ gilt. Wir bezeichnen $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ als *statistisches Experiment* oder *statistisches Modell*. Ein statistisches Experiment $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ heißt *adäquat* für eine ZV X, falls $X \circlearrowleft \mathcal{P}_\Theta$ gilt. Ein *abgeleiteter* oder *interessierender* Parameter $\gamma : \Theta \to \Gamma$ heißt *identifizierbar*, falls für beliebige $\theta, \theta_o \in \Theta$ aus $\gamma(\theta) \neq \gamma(\theta_o)$ folgt $P_\theta \neq P_{\theta_o}$. Jede $(\mathscr{A}, \mathscr{S})$-messbare Funktion $S : \mathcal{X} \to \mathcal{S}$ mit Werten in einem messbarem Raum $(\mathcal{S}, \mathscr{S})$ heißt *Beobachtung* oder *Statistik*. $\mathcal{P}_\Theta^S := \{P_\theta^S, \theta \in \Theta\}$ bezeichnet die induzierte Familie von Wahrscheinlichkeitsmaßen und $(\mathcal{S}, \mathscr{S}, \mathcal{P}_\Theta^S)$ das induzierte statistische Modell. Eine Statistik $\widehat{\gamma}$ mit Werten in Γ heißt *Schätzer* oder *Schätzfunktion* für den abgeleiteten Parameter γ. Eine Statistik φ mit Werten in $\{0, 1\}$ (versehen mit der Potenzmenge $\mathscr{P}$) wird (nicht randomisierter) *Test* für das Testproblem von $H_0 : \gamma \in \Gamma_0$ gegen $H_1 : \gamma \in \Gamma_1$ mit $\Gamma = \Gamma_0 \dot\cup \Gamma_1$ genannt. Nimmt φ den Wert eins an, so wird die Hypothes *H_0 abgelehnt*, und anderenfalls wird die Hypothese *H_0 nicht abgelehnt*. Eine Statistik φ mit Werten in $[0, 1]$ (versehen mit der Borel-σ-Algebra $\mathscr{B}_{[0,1]}$) wird *randomisierter Test* genannt. Dabei wird $\varphi(x)$ als bedingte Wahrscheinlichkeit interpretiert, die Hypothese H_0 abzulehnen, wenn eine Realisierung $X = x$ beobachtet wird. □

§2.1.2 **Definition**. Sei $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein statistisches Experiment. Eine *Entscheidungsregel* ist eine $(\mathscr{A}, \mathscr{E})$-messbare Abbildung $\delta : \mathcal{X} \to \mathcal{E}$ mit Werten in einem messbarem Raum $(\mathcal{E}, \mathscr{E})$, der *Entscheidungsraum* genannt wird. Wir bezeichnen mit Δ eine vorgegebene Menge von Entscheidungsfunktionen. Jede Funktion $\nu : \Theta \times \mathcal{E} \to [0, \infty) =: \mathbb{R}_+$, die messbar im zweiten Argument ist, heißt *Verlustfunktion*. Das Risiko (der mittlere Verlust) einer Entscheidungsregel δ bei Vorliegen des Parameters $\theta \in \Theta$ (P_θ ist die zu Grunde liegende Wahrscheinlichkeitsverteilung und $\mathbb{E}_\theta$ die Erwartung bezüglich P_θ) ist

$$\mathfrak{R}_\nu(\theta, \delta) := \mathbb{E}_\theta[\nu(\theta, \delta)] := \int_{\mathcal{X}} \nu(\theta, \delta(x)) P_\theta(dx).$$

$(\mathcal{E}, \mathscr{E}, \nu)$ wird *statistisches Entscheidungsproblem* genannt. □

§2.1.3 **Beispiele**. (a) In einem *gewöhnlichen linearem Modell* $Y \circlearrowleft \{\mathfrak{L}(X\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ wähle $\Theta := \mathbb{R}^p \times (0, \infty)$ als Parameterraum mit Parametern $\theta = (\beta, \sigma) \in \Theta$, so dass $P_\theta = \mathfrak{L}(X\beta, \sigma^2 \operatorname{Id}_n)$ die Verteilung von Y bei vorliegen des Parameters $\theta = (\beta, \sigma) \in \Theta$ bezeichnet. Versieht man den Stichprobenraum $\mathcal{X} = \mathbb{R}^n$ mit seiner Borel-σ-Algebra $\mathscr{A} = \mathscr{B}_{\mathbb{R}^n}$ so bilden die Verteilungen $\{\mathfrak{L}(X\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum und es liegt zusammenfassend das statistische Experiment $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}, \{\mathfrak{L}(X\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\})$ vor.

Um den (gewöhnlichen) Kleinste-Quadrate-Schätzer $\widehat{\beta}$ als Entscheidungsregel zu interpretieren sowie seine Güte zu messen, betrachtet man den Entscheidungsraum $\mathcal{E} = \mathbb{R}^p$ und beispielsweise die quadratische Verlustfunktion $\nu(\theta, e) = \nu\big((\beta, \sigma), e\big) = \|\beta - e\|^2$. Für diese spezielle Wahl der Verlustfunktion ist der Parameter σ irrelevant. Da aber die Verteilung $P_\theta = \mathfrak{L}(X\beta, \sigma^2 \operatorname{Id}_n)$ von σ abhängt, bezeichnet man σ als einen *Störparameter*.

Beachte, dass bei obiger Modellierung nur das erste und zweite Moment der Verteilung der Beobachtung Y festgelegt werden, d.h. genauer betrachten wir die Familie $\{P \text{ W-maß über } \mathscr{B}_{\mathbb{R}^n} : \mathbb{E}_P(Y) = X\beta \text{ und } \mathbb{C}\mathrm{ov}_P(Y) = \sigma \operatorname{Id}_n \text{ mit } \beta \in \mathbb{R}^p \text{ und } \sigma > 0\}$ so dass vereinfachend die Verteilung der zentrierten und standardisierten Fehler $\sigma^{-1}(Y - X\beta)$ als ein Störparameter aufgefasst werden kann. Dies gilt offensichtlich in einem gewöhnlichen normalen linearen Modell $Y \circlearrowleft \{\mathfrak{N}(X\beta, \sigma \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ nicht, da die multivariate Normalverteilung der Beobachtung Y eindeutig durch das das erste und zweite Moment festgelegt ist.

(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 für jede Person der Erfolg oder Misserfolg der Behandlung notiert, so dass die Anzahl X der erfolgreichen Behandlungen eine Binomial-verteilte ZV mit Erfolgswahrscheinlichkeit $\pi \in (0, 1)$ ist. Zusammenfassend nehmen wir an, dass $X \circlearrowleft \{P_\pi := \mathfrak{Bin}(100, \pi), \pi \in (0, 1)\}$. Wählen wir den Stichprobenraum $\mathcal{X} = \{0, 1, \ldots, 100\}$ versehen mit der Potenzmenge $\mathscr{P}$ als σ-Algebra, so liegt das statistische Experiment $(\mathcal{X}, \mathscr{P}, \{\mathfrak{Bin}(100, \pi), \pi \in (0, 1)\})$ vor. In Abhängigkeit von der Anzahl X der erfolgreichen Behandlungen soll entschieden werden, ob die Erfolgsquote höher ist als diejenige einer klassischen Behandlung mit bekannter Erfolgswahrscheinlichkeit π_o. Die Nullhypothese für den unbekannten Parameter π ist somit $H_0 : \pi \leqslant \pi_o$. Als Entscheidungsraum dient $\mathcal{E} = \{0, 1\}$ (H_o nicht ablehnen bzw. ablehnen), und wir wählen den Verlust $\nu(\pi, e) = \nu_0 e \mathbb{1}_{\{\pi \leqslant \pi_o\}} + \nu_1 (1 - e) \mathbb{1}_{\{\pi > \pi_o\}}$ mit Konstanten $\nu_0, \nu_1 \geqslant 0$. Dies führt auf des Risiko einer Entscheidungsregel (eines Tests) $\delta : \mathcal{X} \to \mathcal{E}$

$$\mathfrak{R}_\nu(\pi, \delta) = \begin{cases} \nu_0 P_\pi(\delta > \pi_0), & \pi \leqslant \pi_0; \\ \nu_0 P_\pi(\delta \leqslant \pi_0), & \pi > \pi_0; \end{cases}$$

so dass die Irrtumswahrscheinlichkeit erster Art $P_\pi(\delta > \pi_0)$ mit ν_0 und die zweiter Art $P_\pi(\delta \leqslant \pi_0)$ mit ν_1 gewichtet wird. □

§2.1.4 **Definition**. Sei $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein statistisches Experiment und $(\mathcal{E}, \mathscr{E}, \nu)$ ein Entscheidungsproblem. Eine Entscheidungsregel $\delta_o \in \Delta$ heißt *(gleichmäßig) besser in* Δ als eine Entscheidungsregel $\delta \in \Delta$, falls $\mathfrak{R}_\nu(\theta, \delta_o) \leqslant \mathfrak{R}_\nu(\theta, \delta)$ für alle $\theta \in \Theta$ gilt und falls ein $\theta_o \in \Theta$ mit $\mathfrak{R}_\nu(\theta_o, \delta_o) < \mathfrak{R}_\nu(\theta_o, \delta)$ existiert. Eine Entscheidungsregel heißt *zulässig* in Δ, wenn es keine (gleichmäßig) bessere Entscheidungsregel in Δ gibt. □

§2.1.5 **Bemerkung**. Häufig schränkt die betrachtete Klasse Δ die möglichen Entscheidungsregeln ein. So ist der gKQS im gewöhnlichen linearen Modell nach dem Satz §1.3.1 von Gauß-Markov zulässig unter quadratischem Verlust in der Klasse der erwartungstreuen und linearen Schätzern. □

§2.1.6 **Beispiel** (§1.1.5 (a) fortgesetzt). Wir vergleichen in einem *normalen Lokations-Modell* $Y \circlearrowleft \{\mathcal{N}(\mu \mathbb{1}_n, \operatorname{Id}_n), \mu \in \mathbb{R}\}$ die Schätzfunktionen $\widehat{\mu}_1 = \overline{Y}$, $\widehat{\mu}_2 = \overline{Y} + 0.5$ sowie $\widehat{\mu}_3 = 6$

unter Verwendung eines quadratischen Verlustes $\nu(\mu,\delta) = (\mu-\delta)^2$. Da $\mathfrak{R}_\nu(\mu,\widehat{\mu}_1) = 1/n$, $\mathfrak{R}_\nu(\mu,\widehat{\mu}_2) = 1/4 + 1/n$ gilt, ist $\widehat{\mu}_1$ besser als $\widehat{\mu}_2$, allerdings ist weder $\widehat{\mu}_1$ besser als $\widehat{\mu}_3$ noch umgekehrt. Insbesondere ist $\widehat{\mu}_3$ zulässig (in der Klasse aller Schätzer!), da $\mathfrak{R}_\nu(6,\widehat{\mu}_3) = 0$ gilt und jeder andere Schätzer $\widetilde{\mu}$ mit $\mathfrak{R}_\nu(6,\widetilde{\mu}) \leqslant \mathfrak{R}_\nu(6,\widehat{\mu}_3) = 0$ mit $\widehat{\mu}_3$ Lebesgue-fast überall übereinstimmt. Später werden wir zeigen dass auch $\widehat{\mu}_1$ zulässig ist. □

§2.1.7 **Definition**. Zu einem vorgegebenen Entscheidungsproblem $(\mathcal{E},\mathscr{E},\nu)$ in einem statistischen Modell $(\mathcal{X},\mathscr{A},\mathcal{P}_\Theta)$ heißt eine Entscheidungsregel δ *unverzerrt*, falls

$$\forall\,\theta,\widetilde{\theta}\in\Theta : \mathbb{E}_\theta[\nu(\widetilde{\theta},\delta)] \geqslant \mathbb{E}_\theta[\nu(\theta,\delta)] = \mathfrak{R}_\nu(\theta,\delta).$$

□

§2.1.8 **Lemma**. *Es seien $(\mathcal{X},\mathscr{A},\mathcal{P}_\Theta)$ ein statistisches Experiment, $\gamma:\Theta\to\mathcal{E}\subset\mathbb{R}$ ein interessierender Parameter und $(\mathcal{E},\mathscr{E},\nu)$ ein statisistisches Entscheidungsproblem mit quadratischem Verlust $\nu(\theta,e) := (\gamma(\theta)-e)^2$. Eine Entscheidungsregel $\widehat{\gamma}:\mathcal{X}\to\mathcal{E}$ ist dann ein Schätzer für den abgeleiteten Parameter γ. Gilt für jedes $\theta\in\Theta$ weiterhin $\mathbb{E}_\theta(\widehat{\gamma}^2)<\infty$ und $\mathbb{E}_\theta(\widehat{\gamma})\in\gamma(\Theta) := \{\gamma(\theta_o),\theta_o\in\Theta\}$, dann ist die Entscheidungsregel $\widehat{\gamma}$ genau dann unverzerrt, wenn sie* erwartungstreu *ist, d.h. $\mathbb{E}_\theta(\widehat{\gamma}) = \gamma(\theta)$ gilt für alle $\theta\in\Theta$.* □

Beweis von Lemma §2.1.8. in der Vorlesung. □

§2.1.9 **Lemma**. *Es seien $(\mathcal{X},\mathscr{A},\mathcal{P}_\Theta)$ ein statistisches Experiment mit $\Theta = \Theta_0\dot\cup\Theta_1$, und $(\mathcal{E},\mathscr{E},\nu)$ ein statisistisches Entscheidungsproblem mit Entscheidungsraum $\mathcal{E}=[0,1]$ und Verlustfunktion $\nu(\theta,e) = \nu_0 e\mathbb{1}_{\Theta_0}(\theta) + \nu_1(1-e)\mathbb{1}_{\Theta_1}(\theta)$ für $\nu_0,\nu_1\in\mathbb{R}_+$. Eine Entscheidungsregel $\varphi:\mathcal{X}\to\mathcal{E}$ (ein randomisierter Test) für das Testproblem $H_0:\theta\in\Theta_0$ gegen $H_1:\theta\in\Theta_1$ ist genau dann unverzerrt, wenn sie zum Niveau $\alpha = \nu_1/(\nu_0+\nu_1)$* unverfälscht *ist, d.h.*

$$\forall\theta\in\Theta_0 : \mathbb{E}_\theta(\varphi)\leqslant\alpha,\quad \forall\theta\in\Theta_1 : \mathbb{E}_\theta(\varphi)\geqslant\alpha.$$

Beweis von Lemma §2.1.9. Übung. □

§2.1.10 **Definition**. Eine Abbildung $K:\mathcal{X}\times\mathscr{S}\to[0,1]$ heißt *Markovkern* von $(\mathcal{X},\mathscr{A})$ nach $(\mathcal{S},\mathscr{S})$, falls

(a) $S\mapsto K(x,S)$ ist eine Wahrscheinlichkeitsmaß auf $(\mathcal{S},\mathscr{S})$ für alle $x\in\mathcal{X}$;

(b) $x\mapsto K(x,S)$ ist messbar für alle $S\in\mathscr{S}$. □

§2.1.11 **Definition**. Zu einem vorgegebenen Entscheidungsproblem $(\mathcal{E},\mathscr{E},\nu)$ in einem statistischen Modell $(\mathcal{X},\mathscr{A},\mathcal{P}_\Theta)$ heißt ein Markovkern D von $(\mathcal{X},\mathscr{A})$ nach $(\mathcal{E},\mathscr{E})$ *Entscheidungskern* oder *randomisierte Entscheidungsregel* mit der Interpretation, dass bei Vorliegen der Beobachtung x gemäß $D(x,\bullet)$ eine Entscheidung zufällig ausgewählt wird. Das zugehörige Risiko ist

$$\mathfrak{R}_\nu(\theta,D) := \mathbb{E}_\theta\Big[\int_{\mathcal{E}}\nu(\theta,e)D(X,de)\Big] = \int_{\mathcal{X}}\int_{\mathcal{E}}\nu(\theta,e)D(x,de)P_\theta(dx).$$

□

§2.1.12 **Beispiele**. (a) Betrachte $\mathcal{E}=\Theta$ versehen mit einer σ-Algebra $\mathscr{B}_\Theta$, ein Markovkern D von $(\mathcal{X},\mathscr{A})$ nach $(\Theta,\mathscr{B}_\Theta)$ ist dann ein „randomisierter“ Schätzer, d.h. bei Vorliegen der Beobachtung x ist $D(x,\bullet)$ eine Wahrscheinlichkeitsverteilung über dem Parameterraum Θ. Falls für jedes $x\in\mathcal{X}$, $D(x,\bullet)$ ein Punktmaß in $\widehat{\theta}(x)\in\mathcal{E}$ ist, d.h. $D(x,\{\widehat{\theta}(x)\}) = 1$, so

dass die Abbildung $\widehat{\theta} : \mathcal{X} \to \mathcal{E}$ messbar ist. Dann ist $\widehat{\theta}$ eine Entscheidungsregel („nicht randomisierter“ Schätzer) und

$$\mathfrak{R}_\nu(\theta, D) = \int_{\mathcal{X}} \int_{\mathcal{E}} \nu(\theta, e) D(x, de) P_\theta(dx) = \int_{\mathcal{X}} \nu(\theta, \widehat{\theta}(x)) P_\theta(dx) = \mathfrak{R}_\nu(\theta, \widehat{\theta}).$$

(b) Betrachte das Testproblem von $H_0 : \theta \in \Theta_0$ gegen $H_1 : \theta \in \Theta_1$ für $\Theta = \Theta_0 \dot\cup \Theta_1$. Es seien $\mathcal{E} = [0, 1]$ und $\nu(\theta, e) := \nu_0 e \mathbb{1}_{\Theta_0}(\theta) + \nu_1(1 - e)\mathbb{1}_{\Theta_1}(\theta)$. Jede (deterministische) Entscheidungsregel φ zum Entscheidungsproblem $([0, 1], \mathscr{B}_{[0,1]}, \nu)$ (randomisierter Test) definiert mit $D(x, \{1\}) := \varphi(x)$ sowie $D(x, \{0\}) := 1 - \varphi(x)$ einen Entscheidungskern D zum Entscheidungsproblem $(\{0, 1\}, \mathscr{P}, \nu)$. Auf der anderen Seite jeder Entscheidungskern D zum Entscheidungsproblem $(\{0, 1\}, \mathscr{P}, \nu)$ definiert eine Entscheidungsregel $\varphi(x) := D(x, \{1\})$ zum Entscheidungsproblem $([0, 1], \mathscr{B}_{[0,1]}, \nu)$. Dies bedeutet also, dass $\varphi(x)$ die Wahrscheinlichkeit angibt, mit der bei Vorliegen der Beobachtung x die Hypothese H_0 abgelehnt wird. Offensichtlich, gilt dann $\mathfrak{R}_\nu(\theta, D) = \mathfrak{R}_\nu(\theta, \varphi)$. □

§2.1.13 **Bemerkung**. Es sei $\mathcal{E} \subset \mathbb{R}^d$ konvex sowie $\nu(\theta, e)$ eine im zweiten Argument konvexe Verlustfunktion. Dann gibt es zu jeder randomisierten Entscheidungsregel eine deterministische Entscheidungsregel, deren Risiko nicht größer ist. □

2.2 Minimax- und Bayes-Ansatz

§2.2.1 **Definition**. Für ein Entscheidungsproblem $(\mathcal{E}, \mathscr{E}, \nu)$ zu einem statistischen Experiment $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ heißt eine Entscheidungsregel δ_o *Δ-minimax*, falls

$$\mathfrak{R}^*_\nu := \sup_{\theta \in \Theta} \mathfrak{R}_\nu(\theta, \delta_o) = \inf_{\delta \in \Delta} \sup_{\theta \in \Theta} \mathfrak{R}_\nu(\theta, \delta)$$

gilt, weiterhin wird $\mathfrak{R}^*_\nu$ *Δ-Minimaxrisiko* genannt. Wir bezeichnen δ_o als *minimax* falls die Menge Δ alle möglichen Entscheidungsregeln (für die das Risiko definiert ist) enthält. □

§2.2.2 **Definition**. Es seien $(\mathcal{X}, \mathscr{A}, P_\Theta)$ ein statistisches Experiment, $\mathscr{B}_\Theta$ eine σ-Algebra über dem Parameterraum Θ, die Verlustfunktion ν $(\mathscr{B}_\Theta \otimes \mathscr{A}, \mathscr{B}_{\mathbb{R}_+})$-messbar, und $\theta \mapsto P_\theta(A)$ messbar für alle $A \in \mathscr{A}$. Sei $\boldsymbol{\vartheta}$ eine ZV mit Werten in $(\Theta, \mathscr{B}_\Theta)$, so dass die Parameter $\theta \in \Theta$ als Realisierung der ZV $\boldsymbol{\vartheta}$ aufgefasst werden können. Die Wahrscheinlichkeitsverteilung $P_{\boldsymbol{\vartheta}}$ von $\boldsymbol{\vartheta}$ auf dem messbaren Raum $(\Theta, \mathscr{B}_\Theta)$ wird *a-priori Verteilung* des Parameters θ genannt und wir bezeichen mit $\mathbb{E}_{\boldsymbol{\vartheta}}$ die Erwartung bezüglich $P_{\boldsymbol{\vartheta}}$. Das mit $P_{\boldsymbol{\vartheta}}$ assoziierte Bayesrisiko einer Entscheidungsregel δ ist

$$\mathfrak{R}^{\boldsymbol{\vartheta}}_\nu(\delta) := \mathbb{E}_{\boldsymbol{\vartheta}} \left[\mathfrak{R}_\nu(\boldsymbol{\vartheta}, \delta)\right] = \int_\Theta \mathfrak{R}_\nu(\theta, \delta) P_{\boldsymbol{\vartheta}}(d\theta) = \int_\Theta \int_{\mathcal{X}} \nu(\theta, \delta(x)) P_\theta(dx) P_{\boldsymbol{\vartheta}}(d\theta).$$

Eine Entscheidungsregel δ_o heißt *Δ-Bayesregel* oder *Δ-Bayes-optimal* (bezüglich $P_{\boldsymbol{\vartheta}}$) falls

$$\mathfrak{R}^{\boldsymbol{\vartheta}}_\nu(\delta_o) = \inf_{\delta \in \Delta} \mathfrak{R}^{\boldsymbol{\vartheta}}_\nu(\delta)$$

gilt. Erstreckt sich das Infimum über alle möglichen Entscheidungsregeln δ so heißt δ_o kurz *Bayesregel* oder *Bayes-optimal*. □

§2.2.3 **Bemerkung**. Während eine Minimaxregel den maximal zu erwartenden Verlust minimiert, kann das Bayesrisiko als ein (mittels P_ϑ) gewichtetes Mittel des zu erwartenden Verlustes angesehen werden. Alternativ wird P_ϑ als die subjektive Einschätzung der Verteilung der zu Grunde liegenden Parameter interpretiert. Daher wird das Bayesrisiko auch als insgesamt zu erwartender Verlust verstanden. □

§2.2.4 **Definition**. Es sei T eine $(\mathcal{S}, \mathscr{S})$-wertige ZV auf dem Wahrscheinlichkeitsraum $(\mathcal{X}, \mathscr{A}, P)$ und $X \sim P$. Ein Markovkern von $(\mathcal{X}, \mathscr{A})$ nach $(\mathcal{S}, \mathscr{S})$ heißt *reguläre bedingte Wahrscheinlichkeitsverteilung* bezüglich T, falls

$$K(T, A) = P_{X|T}(A) := \mathbb{E}_{X|T}(\mathbb{1}_A) := \mathbb{E}(\mathbb{1}_A(X)|\sigma(T)) \quad P - f.s.$$

für alle $A \in \mathscr{A}$ gilt. □

§2.2.5 **Satz**. *Es sei $(\mathcal{X}, d)$ ein vollständiger, separabler Raum mit Metrik d versehen mit der Borel-σ-Algebra $\mathscr{B}$ (polnischer Raum). Für jede ZV T auf $(\mathcal{X}, \mathscr{B}, P)$ existiert eine reguläre bedingte Wahrscheinlichkeitsverteilung K bezüglich T. K ist P-f.s. eindeutig bestimmt, d.h. für eine zweite solche reguläre bedingte Wahrscheinlichkeitsverteilung K_o gilt $P(\forall A \in \mathscr{A} : K(X, A) = K_o(X, A)) = 1$.* □

Beweis von Satz §2.2.5. z.Bsp. in Klenke [2008] Theorem 8.36. □

§2.2.6 **Definition**. Es seien $(\mathcal{X}, \mathscr{A}, P_\Theta)$ ein statistisches Experiment, $X \curvearrowright P_\Theta$ eine Beobachtung, $\boldsymbol{\vartheta} \sim P_\vartheta$ ein ZV mit Werten in $(\Theta, \mathscr{B}_\Theta)$ und $(\theta, A) \mapsto P_\theta(A) = P_{X|\boldsymbol{\vartheta}=\theta}(A)$ eine regulären bedingte Wahrscheinlichkeit (Markovkern) bezüglich $\boldsymbol{\vartheta}$. Bezeichne mit $P_{X,\boldsymbol{\vartheta}}$ die gemeinsame Verteilung des zufälligen Vektors $(X, \boldsymbol{\vartheta})$ mit Werten in dem messbaren Produktraum $(\mathcal{X} \times \Theta, \mathscr{A} \otimes \mathscr{B}_\Theta)$. Die durch $P_{X,\boldsymbol{\vartheta}}$ implizierte reguläre bedingte Wahrscheinlichkeit $(x, B) \mapsto P_{\boldsymbol{\vartheta}|X=x}(B)$ bezüglich X heißt *a-posteriori Verteilung* des zufälligen Parameters $\boldsymbol{\vartheta}$ gegeben die Beobachtung $X = x$. □

§2.2.7 **Bemerkung**. Die gemeinsame Verteilung $P_{X,\boldsymbol{\vartheta}}$ des zufälligen Vektors $(X, \boldsymbol{\vartheta})$ ist wohldefiniert und erfüllt $P_{X,\boldsymbol{\vartheta}}(dx, d\theta) = P_\theta(dx)P_{\boldsymbol{\vartheta}}(d\theta)$ (betrachte $P_{X,\boldsymbol{\vartheta}}(A \times B) = \int_B P_\theta(A)P_{\boldsymbol{\vartheta}}(d\theta)$ und verwende den Maßerweiterungssatz). Wir bezeichnen mit P_X die Randverteilung von X und mit $\mathbb{E}_X$ die assoziierte Erwartung. Insbesondere gilt

$$\begin{aligned}\mathfrak{R}_\nu^{\boldsymbol{\vartheta}}(\delta) = \mathbb{E}_{X,\boldsymbol{\vartheta}}\left[\nu(\boldsymbol{\vartheta}, \delta(X))\right] = \mathbb{E}_X\left[\mathbb{E}_{\boldsymbol{\vartheta}|X}[\nu(\boldsymbol{\vartheta}, \delta(X))]\right] &= \int_{\mathcal{X}} \mathbb{E}_{\boldsymbol{\vartheta}|X=x}\left[\nu(\boldsymbol{\vartheta}, \delta(x))\right]P_X(dx) \\ = \mathbb{E}_{\boldsymbol{\vartheta}}\left[\mathbb{E}_{X|\boldsymbol{\vartheta}}[\nu(\boldsymbol{\vartheta}, \delta(X))]\right] &= \int_\Theta \mathbb{E}_\theta[\nu(\theta, \delta(X))]P_{\boldsymbol{\vartheta}}(d\theta).\end{aligned}$$ □

§2.2.8 **Satz**. *Es seien $(\mathcal{X}, \mathscr{A}, P_\Theta)$ ein statistisches Experiment, $X \curvearrowright P_\Theta$ eine Beobachtung, $\boldsymbol{\vartheta}$ ein ZV mit a-priori Verteilung $P_{\boldsymbol{\vartheta}}$ auf $(\Theta, \mathscr{B}_\Theta)$. Weiterhin sei $f_{\boldsymbol{\vartheta}}$ eine ν-Dichte von $P_{\boldsymbol{\vartheta}}$ bezüglich eines dominierenden Maßes ν ($P_{\boldsymbol{\vartheta}} \ll \nu$) sowie P_Θ eine bezüglich eines Maßes μ dominierte Verteilungsfamilie ($P_\theta \ll \mu$ für alle $\theta \in \Theta$) mit μ-Dichten $\{f_\theta, \theta \in \Theta\}$. Ist $\mathcal{X} \times \Theta \ni (x, \theta) \mapsto f_\theta(x) \in \mathbb{R}_+$ eine $(\mathscr{A} \otimes \mathscr{B}_\Theta)$-messbare Funktion, so besitzt die a-posteriori Verteilung $P_{\boldsymbol{\vartheta}|X=x}$ eine ν-Dichte, nämlich* (Bayesformel)

$$f_{\boldsymbol{\vartheta}|X=x}(\theta) = \frac{f_\theta(x)f_{\boldsymbol{\vartheta}}(\theta)}{\int_\Theta f_\theta(x)f_{\boldsymbol{\vartheta}}(\theta)\nu(d\theta)}.$$

Beweis von Satz §2.2.8. Übung. □

§2.2.9 **Beispiel**. Wir bezeichnen als *Bayestestproblem* (oder Bayes-Klassifikationsproblem) *mit einfachen Hypothesen* ein Entscheidungsproblem $(\mathcal{E}, \mathscr{E}, \nu)$ mit Entscheidungsraum $\mathcal{E} = \{0, 1\}$ sowie 0-1-Verlustes $\nu(\theta, e) = |\theta - e|$ zu einem statistischen Experiment $(\mathcal{X}, \mathscr{A}, P_\Theta)$ mit Parameterraum $\Theta = \{0, 1\}$. Betrachte eine a-priori Verteilung $P_{\boldsymbol{\vartheta}}$ auf $(\Theta, \mathscr{P})$ mit $P_{\boldsymbol{\vartheta}}(\{0\}) =: \pi_0$ und $P_{\boldsymbol{\vartheta}}(\{1\}) =: \pi_1$. Die Familie von Wahrscheinlichkeitsmaße $P_\Theta = \{P_0, P_1\}$ ist dominiert bezüglich eines Maßes μ (z.Bsp $\mu = P_0 + P_1$) und f_0 und f_1 bezeichne die μ-Dichten. Nach der Bayesformel (mit Zählmaß ν) erhalten wir als a-posteriori Verteilung

$$P_{\boldsymbol{\vartheta}\,|X=x}(\{i\}) = \frac{\pi_i f_i(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}, \quad i = 0, 1 \quad (P_X - f.\text{ü}.).$$

□

§2.2.10 **Satz**. *Es gelten die Annahmen und Notationen von Definition* §2.2.6. *Betrachten wir das statistische Entscheidungsproblem* $(\mathcal{E}, \mathscr{E}, \nu)$, *so ist* δ_o *eine* Δ-Bayes-optimale *Entscheidungsregel, falls*

$$\delta_o(X) = \arg\min_{\delta \in \Delta} \mathbb{E}_{\boldsymbol{\vartheta}\,|X}[\nu(\boldsymbol{\vartheta}, \delta(X))] \quad (P_X - f.\ddot{u}.),$$

gilt, d.h. $\mathbb{E}_{\boldsymbol{\vartheta}\,|X=x}[\nu(\boldsymbol{\vartheta}, \delta_o(x))] \leqslant \mathbb{E}_{\boldsymbol{\vartheta}\,|X=x}[\nu(\boldsymbol{\vartheta}, \delta(x))]$ *für alle* $\delta \in \Delta$ *und* P_X*-fast alle* $x \in \mathcal{X}$.

Beweis von Satz §2.2.10. in der Vorlesung. □

§2.2.11 **Korollar**. *Sei* $\Theta \subset \mathbb{R}$. *Unter den Annahmen des Satzes* §2.2.10 *gelten die folgenden Aussagen:*

(a) Für die quadratische Verlustfunktion $\nu(\theta, e) := (e - \theta)^2$ *ist jede Festlegung der bedingten Erwartung* $\widehat{\theta}(x) := \mathbb{E}_{\boldsymbol{\vartheta}\,|X=x}[\boldsymbol{\vartheta}]$ *bezüglich der a-priori Verteilung* $P_{\boldsymbol{\vartheta}}$ *ein Bayes-optimaler Schätzer von* θ *(Bayes-optimale Entscheidungsregel).*

(b) Für den Absolutbetrag $\nu(\theta, e) := |e - \theta|$ *ist jeder a-posteriori Median* $\widehat{\theta}_{med}(x)$, *d.h.* $P_{\boldsymbol{\vartheta}\,|X=x}\big(\boldsymbol{\vartheta} \leqslant \widehat{\theta}_{med}(x)\big) \geqslant 1/2$ *und* $P_{\boldsymbol{\vartheta}\,|X=x}\big(\boldsymbol{\vartheta} \geqslant \widehat{\theta}_{med}(x)\big) \geqslant 1/2$, *bezüglich der a-priori Verteilung* $P_{\boldsymbol{\vartheta}}$ *ein Bayes-optimaler Schätzer von* θ *(Bayes-optimale Entscheidungsregel).*

Beweis von Korollar §2.2.11. Übung. □

§2.2.12 **Beispiel** (§2.2.9 fortgesetzt). Nach Satz §2.2.10 ist ein Bayestest (Bayesklassifizierer) eine Minimalstelle der Abbildung

$$\{0, 1\} \ni e \mapsto \mathbb{E}_{\boldsymbol{\vartheta}\,|X=x}[\nu(\boldsymbol{\vartheta}, e)] = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} e + \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}(1 - e).$$

Eine Lösung des Minimierungsproblems und somit ein Bayestest ist gegeben durch

$$\varphi(x) = \begin{cases} 0, & \pi_0 f_0(x) > \pi_1 f_1(x) \\ 1, & \pi_0 f_0(x) < \pi_1 f_1(x) \\ \text{beliebig}, & \pi_0 f_0(x) = \pi_1 f_1(x) \end{cases}$$

Damit entscheiden wir uns für dasjenige $\varphi \in \{0, 1\}$, dessen a-posteriori Wahrscheinlichkeit am größten ist (MAPE für maximum a posteriori estimator). Insbesondere sei für später auf die Neymann-Pearson-Struktur des Bayestests φ in Abhängigkeit von $f_1(x)/f_0(x)$ hingewiesen. □

§2.2.13 **Satz**. *Es seien die Annahmen und Notationen der Definition §2.2.6 erfüllt. Betrachten wir das statistische Entscheidungsproblem $(\mathcal{E}, \mathscr{E}, \nu)$, so gelten die folgenden Aussagen*

(a) Für jede Entscheidungsregel δ gilt

$$\sup_{\theta\in\Theta} \mathfrak{R}_\nu(\theta, \delta) = \sup_{P_\vartheta} \mathfrak{R}_\nu^\vartheta(\delta),$$

wobei sich das zweite Supremum über alle a-priori Verteilungen P_ϑ erstreckt. Insbesondere ist das Bayes-Risiko einer Δ-Bayesregel stets kleiner oder gleich dem Δ-Minimax-Risiko.

(b) Für eine Δ-Minimaxregel δ_o gilt

$$\sup_{P_\vartheta} \mathfrak{R}_\nu^\vartheta(\delta_o) = \inf_{\delta\in\Delta} \sup_{P_\vartheta} \mathfrak{R}_\nu^\vartheta(\delta).$$

Beweis von Satz §2.2.13. in der Vorlesung. □

§2.2.14 **Bemerkung**. Der letzte Satz wird insbesondere dazu verwendet, untere Schranken für das Minimax-Risiko durch das Bayes-Risiko einer Bayesregel abzuschätzen. □

§2.2.15 **Satz**. *Es seien die Annahmen und Notationen der Definition §2.2.6 erfüllt. Im statistischen Entscheidungsproblem $(\mathcal{E}, \mathscr{E}, \nu)$ gelten für jede Entscheidungsregel δ_o $(\in \Delta)$ die folgenden Aussagen:*

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

(b) Ist δ_o zulässig mit konstanter Risikofunktion, so ist δ_o minimax-optimal.

(c) Ist δ_o eine Bayesregel (bzgl. P_ϑ) und eindeutig (in Δ) in dem Sinne, dass jede andere Bayesregel (bzgl. P_ϑ) die gleiche Risikofunktion besitzt, so ist δ_o zulässig (in Δ).

(d) Die Parametermenge Θ bilde einen metrischen Raum versehen mit der Borel-σ-Algebra $\mathscr{B}_\Theta$. Ist δ_o eine Bayesregel (bzgl. P_ϑ) (in Δ), so ist δ_o zulässig (in Δ), falls (i) $\mathfrak{R}_\nu^\vartheta(\delta_o) < \infty$; (ii) für jede nicht leere offene Menge U in Θ gilt $P_\vartheta(U) > 0$; (iii) für jede Entscheidungsregel δ $(\in \Delta)$ mit $\mathfrak{R}_\nu^\vartheta(\delta) \leqslant \mathfrak{R}_\nu^\vartheta(\delta_o)$ ist die Abbildung $\theta \mapsto \mathfrak{R}_\nu(\theta, \delta)$ stetig.

Beweis von Satz §2.2.15. Übung. □

§2.2.16 **Satz**. *Es seien $X_1, \ldots, X_n$ unabhängig und identisch $\mathcal{N}(\mu, \mathrm{Id}_d)$-verteilte ZV'en mit unbekanntem $\mu \in \mathbb{R}^d$. Bezüglich der quadratischen Verlustfunktion $\nu(\mu, e) = \|\mu - e\|^2$ ist das arithmetische Mittel $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ ein minimax-optimaler Schätzer für μ.*

Beweis von Satz §2.2.15. in der Vorlesung. □

§2.2.17 **Satz**. *Es seien $X_1, \ldots, X_n$ unabhängig und identisch $\mathcal{N}(\mu, 1)$-verteilte ZV'en mit unbekanntem $\mu \in \mathbb{R}$. Bezüglich der quadratischen Verlustfunktion $\nu(\mu, e) = (\mu - e)^2$ ist das arithmetische Mittel $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ ein zulässiger Schätzer für μ.*

Beweis von Satz §2.2.15. in der Vorlesung. □

§2.2.18 **Bemerkung**. Liegt eine andere Verteilung mit Erwartungswert μ und Varianz eins als die Normalverteilung vor, so ist $\overline{X}$ weder zulässig noch minimax (sofern $n \geqslant 3$ gilt), vergleiche Lehmann and Casella [1998], Seite 153. Unter der Normalverteilungsannahme ist $\overline{X}$ für $d = 2$ weiterhin zulässig, allerdings gilt dies für $d = 3$ nicht mehr: Stein-Phänomen in Sektion 2.3. □

§2.2.19 **Definition**. Es seien die Annahmen und Notation von Definition §2.2.6 erfüllt. Für ein Entscheidungsproblem $(\mathcal{E}, \mathscr{E}, \nu)$ heißt eine Verteilung P_{ϑ_o} auf $(\Theta, \mathscr{B}_\Theta)$ *ungünstigste a-priori Verteilung* bzgl. Δ, falls

$$\inf_{\delta\in\Delta} \mathfrak{R}_\nu^{\vartheta_o}(\delta) = \sup_{P_\vartheta} \inf_{\delta\in\Delta} \mathfrak{R}_\nu^{\vartheta}(\delta).$$

§2.2.20 **Satz**. *Für das Entscheidungsproblem $(\mathcal{E}, \mathscr{E}, \nu)$ sei P_{ϑ_o} eine a-priori Verteilung mit zugehöriger Δ-Bayesregel δ_o. Dann sind die Eigenschaften (i) $\mathfrak{R}_\nu^{\vartheta_o}(\delta_o) = \sup_{\theta\in\Theta} \mathfrak{R}_\nu(\theta, \delta_o)$ und (ii) die* Sattelpunkteigenschaft

$$\forall P_\vartheta\ \forall \delta \in \Delta\ :\ \mathfrak{R}_\nu^{\vartheta}(\delta_o) \leqslant \mathfrak{R}_\nu^{\vartheta_o}(\delta_o) \leqslant \mathfrak{R}_\nu^{\vartheta_o}(\delta).$$

äquivalent. Aus jeder dieser Eigenschaften folgt, dass δ_o minimax-optimal in Δ und P_{ϑ_o} ungünstigste a-priori Verteilung bzgl. Δ ist.

Beweis von Satz §2.2.20. in der Vorlesung. □

§2.2.21 **Beispiel**. Sei $X \circlearrowleft \{\mathfrak{Bin}(n, \pi), \pi \in (0, 1)\}$ mit $n \geqslant 1$. Wir bestimmen einen minimax-optimalen Schätzer für π bezüglich der quadratischer Verlustfunktion $\nu(\pi, e) = (e - \pi)^2$ unter Verwendung des Satzes §2.2.20. Dazu betrachten wir die Beta-Verteilung $\mathfrak{Beta}(a, b)$ mit Parametern $a, b > 0$ auf $[0, 1]$ als a-priori Verteilung und bestimmen einen zugehörigen Bayesschätzer $\widehat{\pi}_{a,b}$ für π. Bezeichne mit $\boldsymbol{\pi}_{a,b}$ den zufälligen Parameter mit Werten in $[0, 1]$ und a-priori Verteilung $\mathfrak{Beta}(a, b)$. Die a-posteriori Verteilung $P_{\boldsymbol{\pi}_{a,b}|X}$ ist wieder eine Beta-Veteilung $\mathfrak{Beta}(a + X, b + n - X)$ und der zugehörige Bayesschätzer ist $\widehat{\pi}_{a,b} := \mathbb{E}_{\boldsymbol{\pi}_{a,b}|X}(\boldsymbol{\pi}_{a,b}) = \frac{a+X}{a+b+n}$ (Übung) und für sein Risiko gilt $\mathfrak{R}_\nu(\pi, \widehat{\pi}_{a,b}) = \mathbb{E}_\pi(\widehat{\pi}_{a,b} - \pi)^2 = \frac{(a-a\pi-b\pi)^2+n\pi(1-\pi)}{(a+b+n)^2}$. Im Fall $a^* = b^* = \sqrt{n}/2$ erhält man $\widehat{\pi}_{a^*,b^*} := \mathbb{E}_{\boldsymbol{\pi}_{a^*,b^*}|X}(\boldsymbol{\pi}_{a^*,b^*}) = \frac{X+\sqrt{n}/2}{n+\sqrt{n}} = \frac{X}{n} - \frac{X-n/2}{n(\sqrt{n}+1)}$ mit zugehörigem Risiko $\mathfrak{R}_\nu(\pi, \widehat{\pi}_{a^*,b^*}) = (2\sqrt{n} + 2)^{-2}$ welches unabhängig von π ist, woraus die Sattelpunkteigenschaft folgt:

$$\forall P_{\boldsymbol{\pi}}\ \forall \widehat{\pi} \in [0, 1]\ :\ \mathfrak{R}_\nu^{\boldsymbol{\pi}}(\widehat{\pi}_{a^*,b^*}) \leqslant \mathfrak{R}_\nu^{\boldsymbol{\pi}_{a^*,b^*}}(\widehat{\pi}_{a^*,b^*}) \leqslant \mathfrak{R}_\nu^{\boldsymbol{\pi}_{a^*,b^*}}(\widehat{\pi}).$$

Damit ist $P_{\boldsymbol{\pi}_{a^*,b^*}} = \mathfrak{Beta}(a^*, b^*)$ ungünstigste a-prior Verteilung und $\widehat{\pi}_{a^*,b^*}$ minimax-optimaler Schätzer von π. Insbesondere ist der natürliche Schätzer $\widehat{\pi} = X/n$ mit $\mathfrak{R}_\nu(\widehat{\pi}) = \pi(1 - \pi)/n$ nicht minimax (er ist jedoch zulässig). □

§2.2.22 **Bemerkung**. Gehören für ein statistisches Modell die a-posteriori Verteilungen wieder zur der Klasse von a-priori Verteilungen (i.A. mit geänderten Parametern), so nennt man die entsprechenden Verteilungsklassen *konjugiert*. Zum Beispiel sind Beta-Verteilungen konjugiert zur Binomialverteilung (Beispiel §2.2.21). Konjugierte Verteilungen sind die Ausnahme, nicht die Regel, und für komplexere Modelle werden häufig Rechnen-intensive Methoden wie MCMC (Markov Chain Monte Carlo) verwendet, um die a-posteriori Verteilung zu berechnen. □

2.3 Das Stein-Phänomen

Es seien $X_1, \ldots, X_n$ unabhängig und identisch $\mathfrak{N}(\mu, \mathrm{Id}_d)$-verteilte ZV'en mit unbekanntem $\mu \in \mathbb{R}^d$. Wir betrachten das Entscheidungsproblem, den Parameter μ möglichst gut im Sinne eines quadratischen Verlustes $\nu(\mu, \widehat{\mu}) = \|\widehat{\mu} - \mu\|^2$ zu schätzen. Auf Grund der Unabhängigkeit der

Koordinaten erscheint das (koordiantenweise) arithmetische Mittel $\overline{X}$, eine natürliche Antwort zu sein. Ein alternativer, sogenannter *empirischer Bayesansatz*, beruht auf der Familie der a-priori Verteilungen $\{\mathfrak{N}(0, \sigma^2 \operatorname{Id}_d) : \sigma > 0\}$. Betrachten wir einen zufälligen Parameter $\boldsymbol{\mu}_\sigma \sim \mathfrak{N}(0, \sigma^2 \operatorname{Id}_d)$ so ist der zugehörige Bayesschätzer $\mathbb{E}_{\boldsymbol{\mu}_\sigma|X}(\boldsymbol{\mu}_\sigma) = \frac{n}{n+\sigma^{-2}}\overline{X}$ (vgl. Beweis des Satzes §2.2.16). Der empirische Bayesansatz beruht nun auf der Ersetzung von σ^2 durch die Schätzung $\widehat{\sigma}^2 = \|\overline{X}\|^2/d - n^{-1}$. Da die Randverteilung von $\overline{X}$ bezüglich der gemeinsamen Verteilung $P_{X,\boldsymbol{\mu}_\sigma}$ gerade einer $\mathfrak{N}(0, (\sigma^2 + n^{-1}) \operatorname{Id}_d)$ entspricht, ist $\widehat{\sigma}^2$ ein erwartungstreuer Schätzer von σ^2. Wir erhalten den Schätzer

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

Der Bayessche Ansatz lässt vermuten, dass für kleine Werte von $\|\mu\|$ der Schätzer $\widehat{\mu}$ ein kleineres Risiko als $\overline{X}$ hat. Überraschenderweise gilt für Dimension $d \geqslant 3$ sogar, dass $\widehat{\mu}$ besser als $\overline{X}$ ist. Das folgende Steinsche Lemma liefert das zentrale Argument für den Beweis.

§2.3.1 **Lemma** *(Stein). Es sei $f : \mathbb{R}^d \to \mathbb{R}$ eine in jeder Koordinate Lebesgue-f.ü. absolut stetige Funktion. Dann gilt für $Y \circlearrowleft \{\mathfrak{N}(\mu, \sigma^2 \operatorname{Id}_d), \mu \in \mathbb{R}^d, \sigma > 0\}$*

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

sofern $\mathbb{E}_{\mu,\sigma}[|\frac{\partial f}{\partial y_i}(Y)|] < \infty$ für alle $i = 1, \ldots, n$ gilt.

Beweis von Lemma §2.3.1. in der Vorlesung. □

§2.3.2 **Satz**. *Es sei $d \geqslant 3$ und $X_1, \ldots, X_n$ unabhängig und identisch $\mathfrak{N}(\mu, \operatorname{Id}_d)$-verteilte ZV'en mit unbekanntem $\mu \in \mathbb{R}^d$. Dann gilt für den* James-Stein-Schätzer

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

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

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

Insbesondere ist $\overline{X}$ für eine quadratische Verlustfunktion kein zulässiger Schätzer von μ im Fall $d \geqslant 3$.

Beweis von Satz §2.3.2. in der Vorlesung. □

§2.3.3 **Bemerkungen**. (a) Die Abbildung $\mu \mapsto \mathbb{E}_\mu[\|\overline{X}\|^{-2}]$ ist monoton fallend in $\|\mu\|$ und erfüllt $\mathbb{E}_0[\|\overline{X}\|^{-2}] = n/(d-2)$ und $\mathbb{E}_0\|\widehat{\mu}_{JS} - \mu\|^2 = 2/n$. Damit ist $\widehat{\mu}_{JS}$ für μ nahe 0, große Dimension d und kleine Stichprobenumfänge n eine deutliche Verbesserung von $\overline{X}$. Der James-Stein-Schätzer wird auch *Shrinkage-Schätzer* genannt, weil die Koordinaten des ursprünglichen Schätzers $\overline{X}$ gedämpft (zur Null hingezogen) werden.

(b) Der *James-Stein-Schätzer mit positivem Gewicht*

$$\widehat{\mu}_{JS+} := (1 - \frac{d-2}{n\|\overline{X}\|^2})_+\overline{X}, \quad (a)_+ := \max(a, 0),$$

ist bei quadratischer Verlustfunktion besser als der James-Stein-Schätzer $\widehat{\mu}_{JS}$. Damit ist selbst der James-Stein-Schätzer (sogar mit positivem Gewicht) unzulässig. Die Konstruktion eines zulässigen Minimax-Schätzers ist gelöst für $d \geqslant 6$ (vgl. Lehmann and Casella [1998], S. 385). □

Kapitel 3

Schätztheorie

3.1 Dominierte Modelle

§3.1.1 **Definition**. Ein statistisches Model $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ heißt *dominiert*, falls ein σ-endliches Maß μ auf $\mathscr{A}$ existiert, so dass für jedes $\theta \in \Theta$ das Wahrscheinlichkeitsmaß P_θ absolut stetig bezüglich μ ist ($P_\theta \ll \mu$). Die durch θ paramisierte Radon-Nikodym-Dichte

$$L(\theta, x) := \frac{dP_\theta}{d\mu}(x) \quad \theta \in \Theta,\ x \in \mathcal{X},$$

wird auch *Likelihood-Funktion* genannt, wobei diese meist als eine durch x parametrisierte Funktion in θ aufgefasst wird, d.h. $L(\theta, x) =: L(\theta)$. □

§3.1.2 **Beispiele**. (a) Ein statistische Experiment $(\mathbb{R}, \mathscr{B}_\mathbb{R}, \mathcal{P}_\Theta)$ ist trivialerweise dominiert wenn jedes $P_\theta \in \mathcal{P}_\Theta$ durch eine Lebesguedichte f_θ gegeben ist, beispielsweise $P_{(\mu,\sigma)} = \mathfrak{N}(\mu, \sigma^2)$.

(b) Jedes statistische Modell mit Stichprobenraum $\mathcal{X} = \mathbb{N}$ und Potenzmenge $\mathscr{A} = \mathscr{P}(\mathbb{N})$ oder allgemeiner mit abzählbarem Stichprobenraum $\mathcal{X}$ und Potenzmenge $\mathscr{A} = \mathscr{P}(\mathcal{X})$ ist dominiert bezüglich des Zählmaßes.

(c) Ist die Parametermenge $\Theta = \{\theta_1, \theta_2, \dots\}$ abzählbar, so ist $\mu = \sum_i c_i P_{\theta_i}$ mit $c_i > 0$, $\sum_i c_i = 1$ ein dominierendes Maß.

(d) Sei δ_x das Punktmaß in $x \in \mathbb{R}$. Das statistische Experiment $(\mathbb{R}, \mathscr{B}_\mathbb{R}, \mathcal{P}_\mathbb{R} = \{\delta_\theta, \theta \in \mathbb{R}\})$ ist nicht dominiert. Für ein dominierendes Maß müßte $\mu(\{\theta\}) > 0$ für alle $\theta \in \mathbb{R}$ gelten und damit $\mu(A) = \infty$ für jede überabzählbare Borelmenge $A \subset \mathbb{R}$ erfüllen (sonst folgt aus $|\{x \in A | \mu(\{x\}) \geqslant 1/n\}| \leqslant n\mu(A) < \infty$, dass $A = \cup_{n \geqslant 1}\{x \in A | \mu(\{x\}) \geqslant 1/n\}$ abzählbar ist). Damit kann μ nicht σ-endlich sein. □

§3.1.3 **Satz**. *Es sei $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein dominiertes statistisches Modell. Dann existiert ein Wahrscheinlichkeitsmaß Q der Form $\sum_{i=1}^\infty c_i P_{\theta_i}$ mit $c_\geqslant 0$, $\sum_i c_i = 1$, $\theta_i \in \Theta$, so dass $P_\theta \ll Q$ für alle $\theta \in \Theta$ gilt. Das Wahrscheinlichkeitsmaß Q wird* privilegiertes dominierendes Maß *genannt.*

Beweis von Satz §3.1.3. in der Vorlesung. □

3.2 Erschöpfende Statistik

§3.2.1 **Beispiel**. Es seien $X_1, \dots, X_n$ unabhängige und identisch P_θ-verteilte (u.i.v.) ZV'en mit Werten in $\mathbb{R}$ und jedes $P_\theta \in \mathcal{P}_\Theta$ sei durch eine Lebesguedichte $f_\theta : \mathbb{R} \to \mathbb{R}_+$ gegeben. Allgemeine Informationen über P_θ und somit θ erhalten wir typischerweise mit Hilfe von Statistiken wie $\overline{X}$ oder $\max(X_1, \dots, X_n)$. Intuitiv, enthält die *Ordnungsstatistik* $X_{(1)}, \dots, X_{(n)}$ mit $X_{(1)} = \min\{X_1, \dots, X_n\}$, $X_{(k+1)} := \min\{X_1, \dots, X_n\} \backslash \{X_{(1)}, \dots, X_{(k)}\}$, $k = 2, \dots, n$, wie jede Statistik nicht mehr Informationen über den Parameter θ. Wir werden im Folgenden zeigen, dass die Ordnungsstatistik keine Information verliert. □

§3.2.2 **Definition**. Es seien $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta = \{P_\theta, \theta \in \Theta\})$ und $(\mathcal{S}, \mathscr{S}, \mathcal{Q}_\Theta = \{Q_\theta, \theta \in \Theta\})$ zwei statistische Experimente zum selben Parameterraum Θ. $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ heißt *informativer* als $(\mathcal{S}, \mathscr{S}, \mathcal{Q}_\Theta)$, falls für alle Entscheidungsprobleme $(\mathcal{E}, \mathscr{E}, \nu)$ mit $\|\nu\|_\infty := \sup_{\theta,e} |\nu(\theta, e)| < \infty$ und für alle Entscheidungskerne $D_\mathcal{S}$ von $(\mathcal{S}, \mathscr{S})$ nach $(\mathcal{E}, \mathscr{E})$ ein Entscheidungskern $D_\mathcal{X}$ von $(\mathcal{X}, \mathscr{A})$ nach $(\mathcal{E}, \mathscr{E})$ existiert mit

$$\mathfrak{R}_\nu(\theta, D_\mathcal{X}) = \int_\mathcal{X} \int_\mathcal{E} \nu(\theta, e) D_\mathcal{X}(x, de) P_\theta(dx)$$
$$\leqslant \int_\mathcal{S} \int_\mathcal{E} \nu(\theta, e) D_\mathcal{S}(t, de) Q_\theta(dt) = \mathfrak{R}_\nu(\theta, D_\mathcal{S}).$$

Ist $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ informativer als $(\mathcal{S}, \mathscr{S}, \mathcal{Q}_\Theta)$ und $(\mathcal{S}, \mathscr{S}, \mathcal{Q}_\Theta)$ informativer als $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$, dann heißen $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ und $(\mathcal{S}, \mathscr{S}, \mathcal{Q}_\Theta)$ *äquivalent*. □

§3.2.3 **Lemma**. *Existiert ein Markovkern K von $(\mathcal{X}, \mathscr{A})$ nach $(\mathcal{S}, \mathscr{S})$ mit*

$$KP_\theta = Q_\theta \qquad :\Leftrightarrow \int_\mathcal{X} K(x, S) P_\theta(dx) = Q_\theta(S), \quad \forall S \in \mathcal{S}$$

dann ist $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ informativer als $(\mathcal{S}, \mathscr{S}, \mathcal{Q}_\Theta)$. □

Beweis von Lemma §3.2.3. in der Vorlesung. □

§3.2.4 **Korollar**. *Es seien $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein statistisches Experiment, T eine $(\mathcal{S}, \mathscr{S})$-wertige Statistik auf $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ und $\mathcal{P}_\Theta^T$ die induzierte Verteilungsfamilie auf $(\mathcal{S}, \mathscr{S})$. Dann ist $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ informativer als $(\mathcal{S}, \mathscr{S}, \mathcal{P}_\Theta^T)$.* □

Beweis von Korollar §3.2.4. Übung. □

§3.2.5 **Beispiel** (§1.1.5 (a) fortgesetzt). Betrachte das normale Lokations-Modell $X \circlearrowleft \{\mathfrak{N}(\mu \mathbb{1}_n, \mathrm{Id}_n), \mu \in \mathbb{R} = \Theta\}$ und

$$T : \mathbb{R}^n \ni (x_1, \ldots, x_n)^t \mapsto \overline{x} := n^{-1} \sum_{i=1}^n x_i \in \mathbb{R} = \Theta$$

dann gilt $T(X) = \overline{X} \circlearrowleft \{\mathfrak{N}(\mu, n^{-1}), \mu \in \mathbb{R} = \Theta\}$. Insbesondere folgt aus Korollar §3.2.4, dass das normale Lokations-Modell $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}, \{\mathfrak{N}(\mu \mathbb{1}_n, \mathrm{Id}_n), \mu \in \mathbb{R} = \Theta\})$ informativer ist als das statistische Experiment $(\mathbb{R}, \mathscr{B}_\mathbb{R}, \{\mathfrak{N}(\mu, n^{-1}), \mu \in \mathbb{R} = \Theta\})$. Wir werden im nächsten Abschnitt zeigen dass die statististischen Experimente äquivalent sind. □

§3.2.6 **Definition**. Eine $(\mathcal{S}, \mathscr{S})$-wertige Statistik T auf $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ heißt *erschöpfend* oder *suffizient* (für $\mathcal{P}_\Theta$), falls für jedes $\theta \in \Theta$ die reguläre bedingte Wahrscheinlichkeitsverteilung von $X \sim P_\theta$ gegeben T (existiert und) nicht von θ abhängt, d.h. es existiert ein Markovkern K von $(\mathcal{X}, \mathscr{A})$ nach $(\mathcal{S}, \mathscr{S})$, so dass

$$\forall \theta \in \Theta,\ A \in \mathscr{A} : K(T, A) = P_\theta(A|T) := \mathbb{E}_\theta[\mathbb{1}_A|T] := \mathbb{E}_\theta[\mathbb{1}_A(X)|\sigma(T)] \quad P_\theta - f.s..$$

Statt $K(t, A)$ schreiben wir $P_\bullet(A|T = t)$ bzw. $\mathbb{E}_\bullet[\mathbb{1}_A|T = t]$. □

§3.2.7 **Lemma**. *Ist eine $(\mathcal{S}, \mathscr{S})$-wertige Statistik T auf $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ mit induzierter Verteilungsfamilie $\mathcal{P}_\Theta^T$ auf $(\mathcal{S}, \mathscr{S})$ erschöpfend, dann sind die statistischen Experimente $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ und $(\mathcal{S}, \mathscr{S}, \mathcal{P}_\Theta^T)$ äquivalent.*

Beweis von Lemma §3.2.7. in der Vorlesung. □

§3.2.8 **Bemerkung**. Seien $\mathcal{X}$ polnisch, $\mathscr{A}$ die Borel-σ-Algebra und $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein bzgl. eines σ-endlichen Maßes dominiertes statistisches Experiment. Dann ist eine $(\mathcal{S}, \mathscr{S})$-wertige Statistik T auf $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ mit induzierter Verteilungsfamilie $\mathcal{P}_\Theta^T$ auf $(\mathcal{S}, \mathscr{S})$ genau dann erschöpfend, wenn die statistischen Experimente $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ und $(\mathcal{S}, \mathscr{S}, \mathcal{P}_\Theta^T)$ äquivalent sind. □

§3.2.9 **Satz** *(Faktorisierungskriterium von Neyman)*. *Sei $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein bzgl. eines σ-endlichen Maßes μ dominiertes statistisches Experiment mit Likelihood-Funktion L sowie T eine $(\mathcal{S}, \mathscr{S})$-wertige Statistik. T ist genau dann erschöpfend, wenn eine messbare Funktion $h : \mathcal{X} \to \mathbb{R}_+$ existiert, so dass für jedes $\theta \in \Theta$ eine messbare Funktion $g_\theta : \mathcal{S} \to \mathbb{R}_+$ existiert mit*

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

□

§3.2.10 **Lemma**. *Es seien $X \sim P$ und $Y \sim Q$ ZV'en mit Werten in $(\mathcal{X}, \mathscr{A})$ sowie T eine Statistik auf $(\mathcal{X}, \mathscr{A})$. Ist P absolut stetig bzgl. Q ($P \ll Q$), dann gilt für alle $A \in \mathscr{A}$*

$$\mathbb{E}_{X|T}[\mathbb{1}_A(X)] = \frac{\mathbb{E}_{Y|T}[\mathbb{1}_A(Y)\frac{dP}{dQ}(Y)]}{\mathbb{E}_{Y|T}[\frac{dP}{dQ}(Y)]} \quad P\text{-f.s.}$$

Beweis von Lemma §3.2.10. in der Vorlesung.

§3.2.11 **Bemerkung**. Mit den üblichen Approximationsargumenten lässt sich die Aussage von Lemma §3.2.10 zu $\mathbb{E}_{X|T}[f(X)] = \frac{\mathbb{E}_{Y|T}[f(Y)\frac{dP}{dQ}(Y)]}{\mathbb{E}_{Y|T}[\frac{dP}{dQ}(Y)]}$, P-f.s., für $\mathbb{E}_X(|f(X)|) < \infty$ verallgemeinern. □

Beweis von Satz §3.2.9. in der Vorlesung. □

§3.2.12 **Beispiele**. (a) Die Identität $T(x) = x$ und allgemein jede bijektive, bi-messbare Transformation T ist stets erschöpfend.

(b) Sind $X_1, \ldots, X_n$ unabhängige und identisch P_θ-verteilte ZV'en mit Werten in $\mathbb{R}$ und jedes $P_\theta \in \mathcal{P}_\Theta$ ist durch eine Lebesguedichte $f_\theta : \mathbb{R} \to \mathbb{R}_+$ gegeben, so ist die Ordnungsstatistik $(X_{(1)}, \ldots, X_{(n)})$ erschöpfend, da die Likelihood-Funktion sich in der Form $L(\theta, x) = \prod_{i=1}^n f_\theta(x_{(i)})$ schreiben lässt.

(c) Es wird eine Realisierung $(N_t, t \in [0, T])$ eines Poissonprozesses mit unbekanntem 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 $(U_{(1)}, \ldots, U_{(n)})$ mit unabhängigen und identisch $\mathfrak{U}([0, T])$ verteilten ZV'en $U_1, \ldots, U_n$. Da sich die Beobachtung $(N_t, t \in [0, T])$ eindeutig aus S_k rekonstruieren lässt, ist die Verteilung dieser Beobachtung gegeben $\{N_T = n\}$ unabhängig von λ, und N_T ist somit eine erschöpfende Statistik (die Kenntnis der Gesamtzahl der gemessenen radioaktiven Zerfälle liefert bereits die maximal mögliche Information über die Intensität λ). □

§3.2.13 **Ungleichung von Jensen**. *Es seien $\mathcal{E} \subset \mathbb{R}^k$ konvex, $\psi : \mathcal{E} \to \mathbb{R}$ eine konvexe Funktion und $Z = (Z_1, \ldots, Z_k)^t$ eine $\mathcal{E}$-wertige ZV mit $\mathbb{E}|Z_i| < \infty$, $i = 1, \ldots, k$. Dann gilt $\mathbb{E}[Z] \in \mathcal{E}$ und $\psi(\mathbb{E}[Z]) \leqslant \mathbb{E}[\psi(Z)]$.* □

Beweis der Ungleichung von Jensen §3.2.13. z.Bsp. in Klenke [2008] Theorem 7.11. □

§3.2.14 **Satz** *(Rao-Blackwell Verbesserung). Es seien $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein statistisches Experiment, $(\mathcal{E}, \mathscr{E}, \nu)$ ein Entscheidungsproblem mit konvexem Entscheidungsraum $\mathcal{E} \subset \mathbb{R}^k$ und im zweiten Argument konvexer Verlustfunktion $\nu(\theta, e)$. Ist T eine erschöpfende Statistik für $\mathcal{P}_\Theta$, so gilt für jede Entscheidungsregel $\delta = (\delta_1, \ldots, \delta_k)^t$ mit $\mathbb{E}_\theta|\delta_l(X)| < \infty$, $l = 1, \ldots, k$, für alle $\theta \in \Theta$ und für $\delta_o := \mathbb{E}_\bullet[\delta|T]$ die Risikoabschätzung*

$$\forall\, \theta \in \Theta : \quad \mathfrak{R}_\nu(\theta, \delta_o) \leqslant \mathfrak{R}_\nu(\theta, \delta).$$

□

Beweis von Satz §3.2.14. in der Vorlesung. □

§3.2.15 **Bemerkung**. Ist die Verlustfunktion strikt konvex im zweiten Argument sowie $P_\theta(\delta = \delta_o) < 1$, so ist δ_o besser als δ. Damit gilt im Satz §3.2.14 Gleichheit für die Risiken von δ_o und δ genau dann wenn $\delta_o = \delta$ P_θ-f.s.. □

§3.2.16 **Satz**. *Es sei $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein statistisches Experiment und T eine erschöpfende Statistik. Zu jedem randomisierten Test φ gibt es einen randomisierten Test φ_o, der nur von T abhängt und dieselben Irrtumswahrscheinlichkeiten erster und zweiter Art besitzt, nämlich $\varphi_o = \mathbb{E}_\bullet[\varphi|T]$.* □

Beweis von Satz §3.2.16. in der Vorlesung. □

§3.2.17 **Beispiel**. Für $\theta \in (0, \infty) =: \Theta$ bezeichne $\mathfrak{U}([0, \theta])$ eine Gleichverteilung auf dem Intervall $[0, \theta]$ mit Lebesguedichte $\theta^{-1}\mathbb{1}_{[0,\theta]}(x)$, $x \in \mathbb{R}$. Es seien $X_1, \ldots, X_n$ unabhängig und identisch $\mathfrak{U}([0, \theta])$-verteilte ZV'en mit unbkanntem Parameter $\theta > 0$, d.h. $\mathcal{P}_\Theta = \{P_\theta = \mathfrak{U}([0, \theta]), \theta \in \Theta\}$. Ein erwartungstreuer Schätzer des Erwartungswertes $\theta/2$ ist das arithmetische Mittel $\overline{X}$, so dass $\widehat{\theta} = 2\overline{X}$ ein natürlicher Schätzer für θ ist. Sein Risko bzgl. der quadratischen Verlustfunktion ist $\mathfrak{R}_\nu(\theta, \widehat{\theta}) = 4\,\mathbb{V}\mathrm{ar}_\theta(\overline{X}) = \frac{4\theta^2}{12n}$. Andererseits die Likelihood-Funktion bezüglich des Lebesguemaßes auf $\mathbb{R}^n$ ist

$$L(\theta, x) = \prod_{i=1}^n (\theta^{-1}\mathbb{1}_{[0,\theta]}(x_i)) = \theta^{-n}\mathbb{1}_{[0,\theta]}(\max_{i=1,\ldots,n} x_i).$$

Das Faktorisierungskriterium von Neyman (Satz §3.2.9) anwendend ist $X_{(n)} = \max_{i=1,\ldots,n} X_i$ eine erschöpfende Statistik mit Lebesguedichte $nt^{n-1}\theta^{-n}\mathbb{1}_{[0,\theta]}(t), t \in \mathbb{R}$, und wir bilden

$$\widehat{\theta}_o := \mathbb{E}_\bullet[\widehat{\theta}|X_{(n)}] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_\bullet[X_i|X_{(n)}].$$

Aus Symmetriegründen genügt es, $\mathbb{E}_\bullet[X_1|X_{(n)}]$ zu bestimmen. Da für $x \in [0,\theta]$ gilt:

$$\begin{aligned}\mathbb{E}_\theta\big(\mathbb{1}_{[0,x]}(X_1)\big) &= P_\theta([0,x]) = (x/\theta)\\ &= \frac{1}{n}(x/\theta)^n + \frac{n-1}{n}\Big((x/\theta)^n + \frac{nx(\theta^{n-1}-x^{n-1})}{(n-1)\theta^n}\Big)\\ &= \int_0^\theta \big(\frac{1}{n}\mathbb{1}_{[0,x]}(t) + \frac{n-1}{n}\frac{x\wedge t}{t}\big)nt^{n-1}\theta^{-n}dt\\ &= \int_0^\theta \{\frac{1}{n}\delta_t([0,x]) + \frac{n-1}{n}P_t([0,x])\}P_{X_{(n)}}(dt) = \mathbb{E}_\theta\big(\mathbb{E}_\bullet[\mathbb{1}_{[0,x]}(X_1)|X_{(n)}]\big)\end{aligned}$$

wobei δ_t das Punktmaß in t bezeichnet, erfüllt die bedingte Verteilung $P_{X_1|X_{(n)}=t}$ von X_1 gegeben $X_{(n)} = t$ somit $P_{X_1|X_{(n)}=t}([0,x]) = \frac{1}{n}\delta_t([0,x]) + \frac{n-1}{n}P_t([0,x])$. Damit gilt $\mathbb{E}_\bullet[X_1|X_{(n)}] = \frac{1}{n}X_{(n)} + \frac{n-1}{2n}X_{(n)} = \frac{n+1}{2n}X_{(n)}$, so dass wir $\widehat{\theta}_o = \frac{n+1}{n}X_{(n)}$ erhalten. Der Schätzer $\widehat{\theta}_o$ ist erwartungstreu und sein quadratisches Risiko ist $\mathfrak{R}_\nu(\theta,\widehat{\theta}_o) = \frac{\theta^2}{n^2+2n}$. Für $n > 1$ ist der Schätzer $\widehat{\theta}_o$ offensichtlich besser als $\widehat{\theta}$, für $n \to \infty$ erhalten wir sogar die Ordnung $O(n^{-2})$ gegenüber $O(n^{-1})$. □

3.3 Exponentialfamilien

§3.3.1 **Definition**. Es sei $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein bzgl. eines σ-endlichen Maßes μ dominiertes statistisches Experiment. $\mathcal{P}_\Theta$ wird *Exponentialfamilie* (in $\eta(\theta)$ 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 exisitieren, so dass

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

Die Statistik T wird *natürlich erschöpfend* für $\mathcal{P}_\Theta$ genannt. Sind die Koordinatenfunktionen $\eta_1, \dots, \eta_k$ von η linear unabhängige Funktionen und gilt für die Koordinatenfunktionen $T_1, \dots, T_k$ von T für alle $\theta \in \Theta$ die Implikation

$$\lambda_0 + \lambda_1 T_1 + \cdots + \lambda T_k = 0 \quad P_\theta\text{-f.s.} \quad \Rightarrow \quad \lambda_0 = \lambda_1 = \cdots = \lambda_k = 0$$

d.h. $\mathbb{1}, T_1, \dots, T_k$ sind P_θ-f.s. linear unabhängig. Dann wird die Exponentialfamilie *(strikt) k-parametrisch* genannt. □

§3.3.2 **Bemerkungen**. (i) $C(\theta) = (\int_{\mathcal{X}} h(x)\exp(\langle\eta(\theta), T(x)\rangle)\mu(dx))^{-1}$ ist gerade die Normierungskonstante.

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

(iii) Aus der Identifizierbarkeit des Parameters, d.h. $P_\theta \neq P_{\theta_o}$ für alle $\theta \neq \theta_o$, folgt die Injektivität von η. Andererseits impliziert die Injektivität von η bei einer strikt k-parametrischen Exponetialfamilie die Identifizierbarkeit des Parameters.

(iv) Das Faktorisierungskriterium von Neyman (Satz §3.2.9) anwendend ist die natürliche erschöpfende Statistik T einer Exponentialfamilie $\mathcal{P}_\Theta$ in der Tat erschöpfend für $\mathcal{P}_\Theta$ im Sinne der Definition §3.2.6. □

§3.3.3 **Definition**. Unter den Annahmen und Notationen der Definition §3.3.1 bezeichnet

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

den *natürlichen Parameterraum* einer Exponentialfamilie $\mathcal{P}_\Theta$. Die entsprechend mit $u \in \Theta_{nat}$ parametrisierte Familie wird *natürliche Exponentialfamilie* in T genannt. □

§3.3.4 **Beispiele**. (a) Die Normalverteilungsfamilie $\{\mathfrak{N}(\mu,\sigma^2), \mu \in \mathbb{R}\sigma > 0\}$ ist eine zweiparametrische Exponentialfamilie in $\eta(\mu,\sigma) = (\mu/\sigma^2, 1/(2\sigma^2))^t$ und $T(x) = (x, -x^2)^t$ bzgl. des Lebesguemaßes als dominierierndes Maß. Jedes u der Form $u = (\mu/\sigma^2, 1/(2\sigma^2))^t$ ist ein natürlicher Parameter, und der natürliche Parameterraumm ist gegeben durch $\Theta_{nat} = \mathbb{R} \times (0,\infty)$. Ist entweder $\sigma > 0$ oder $\mu \in \mathbb{R}$ bekannt so liegt eine einparametrische Exponentialfamilie in $\eta(\mu) = \mu/\sigma^2$ bzw. $\eta(\sigma) = 1/(2\sigma^2)$ und $T(x) = x$ bzw. $T(x) = -x^2$ vor.

(b) Die Binomialverteilungsfamilie $\{\mathfrak{Bin}(n,\pi), \pi \in (0,1)\}$ bildet eine Exponentialfamilie in $\eta(\pi) = \log(\pi/(1-\pi))$ (*Logitfunktion* vgl. Bemerkung §1.1.11) und $T(x) = x$ bezüglich dem Zählmaß μ auf $\{0,1,\dots,n\}$. Der natürliche Parameterraum ist $\mathbb{R}$, insbesondere liegt für den Parameterbereich $[0,1]$ keine Exponentialfamilie vor. □

§3.3.5 **Lemma**. *Der natürlichen Parameterraum Θ_{nat} einer Exponentialfamilie ist konvex.* □

Beweis von Lemma §3.3.5. Dies folgt aus der Hölderschen Ungleichung §3.3.6. □

§3.3.6 **Höldersche Ungleichung**. *Für $r > 0$ bezeichne $\mathcal{L}^r_\mu := \{f : \int |f|^r d\mu < \infty\}$ die Menge aller $|f|^r$ μ-integrierbarer Funktionen. Seien $p, q > 1$ mit $1/p + 1/q = 1$, $f \in \mathcal{L}^p_\mu$ und $g \in \mathcal{L}^q_\mu$, dann ist $f \cdot g \in \mathcal{L}^1_\mu$ und es gilt $\big|\int fg\,d\mu\big| \leqslant \big[\int |f|^p d\mu\big]^{1/p} \big[\int |g|^q d\mu\big]^{1/q}$.* □

§3.3.7 **Lemma**. *Bildet $\mathcal{P}_\Theta$ eine (k-parametrische) Exponentialfamilie in $\eta(\theta)$ und $T(x)$, so bildet auch die Familie der Produktmaße $\{P_\theta^{\otimes^n}, \theta \in \Theta\}$ eine (k-parametrische) Exponentialfamilie in $\eta(\theta)$ und $\sum_{i=1}^n T(x_i)$ mit*

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

□

Beweis von Lemma §3.3.7. Dies folgt aus der Produktformel $\frac{dP_\theta^{\otimes^n}}{d\mu^{\otimes^n}}(x) = \prod_{i=1}^n \frac{P_\theta}{d\mu}(x_i)$. □

§3.3.8 **Beispiele**. (a) *(Fortsetzung von §1.1.5 (a))* Betrachte das *normale Lokations-Skalen-Modell* $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}, \{\mathfrak{N}(\mu\mathbb{1}_n, \sigma^2\,\mathrm{Id}_n), \mu \in \mathbb{R}, \sigma > 0\})$, dann entspricht die zu Grunde liegende Normalverteilungsfamilie gerade der Familie der Produktmaße $\{\mathfrak{N}(\mu,\sigma^2)^{\otimes^n}, \mu \in \mathbb{R}, \sigma > 0\}$. Somit ist die natürliche erschöpfende Statistik $T(x) = (\sum_{i=1}^n x_i, -\sum_{i=1}^n x_i^2)^t$. Durch Transformation sind damit auch $(\overline{x}, \overline{x^2})^t$ und $(\overline{x}, S^2)^t$ mit $S^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$ erschöpfende Statistiken.

(b) Sei $\{\mathfrak{Bin}(1,\pi)^{\otimes n}, \pi \in (0,1)\}$ die Verteilungsfamilie einer Bernoullikette, dann ist die Anzahl der Erfolge $T(x) = \sum_{i=1}^n x_i$ erschöpfend. □

§3.3.9 **Satz**. *Sei $\mathcal{P}_{\Theta_{nat}}$ eine Exponentialfamilie mit natürlichem Parametraum $\Theta_{nat} \subset \mathbb{R}^k$ und Darstellung*

$$\frac{dP_\theta}{d\mu}(x) = C(\theta)h(x)\exp\big(\langle\theta, T(x)\rangle\big) = h(x)\exp\big(\langle\theta, T(x)\rangle - A(\theta)\big),$$

mit $A(\theta) = \log\big(\int_{\mathcal{X}} h(x)\exp(\langle\theta, T(x)\rangle)\mu(dx)\big)$. *Ist* θ_o *ein innerer Punkt von* Θ_{nat}, *so ist die erzeugende Funktion* $\psi_{\theta_o}(s) = \mathbb{E}_{\theta_o}[\exp(\langle T, s\rangle)]$ *in einer Umgebung der Null wohldefiniert und beliebig oft differenzierbar. Es gilt weiterhin* $\psi_{\theta_o}(s) = \exp(A(\theta_o + s) - A(\theta_o))$ *für alle* s *mit* $\theta_o + s \in \Theta_{nat}$. *Für* $i, j = 1, \ldots, k$ *folgt außerdem* $\mathbb{E}_{\theta_o}(T_i(X)) = \frac{\partial A}{\partial \theta_i}(\theta_o)$ *und* $\mathbb{C}\mathrm{ov}_{\theta_o}(T_i(X), T_j(X)) = \frac{\partial^2 A}{\partial\theta_i\partial\theta_j}(\theta_o)$. □

Beweis von Satz §3.3.9. Übung. □

§3.3.10 **Satz**. *Gegeben sei eine (strikt)* $(1+k)$*-parametrische natürliche Exponentialfamilie in* (U, T) *mit der Darstellung*

$$\frac{dP_{\theta,\tau}}{d\mu}(x) = C(\theta,\tau)\exp\big(\theta U(x) + \langle\tau, T(x)\rangle\big).$$

Dann bildet die Familie der bedingten Verteilungen $P^{U|T}_{\theta,\tau}$ *von* U *gegeben* T *eine (strikt)* 1*-parametrische natürliche Exponentialfamilie in* U, *die unabhängig von* τ *ist. Insbesondere gilt*

$$\frac{dP^{U|T}_{\theta,\tau}}{d\mu^{U|T}}(u) = \frac{\exp(\theta U)}{\int_{\mathbb{R}}\exp(\theta v)\mu^{U|T}(dv)} \quad \mu^{U|T}\text{-f.s.}.$$

Beweis von Satz §3.3.10. in der Vorlesung. □

3.4 Vollständige Statistik

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

$$\forall\,\theta \in \Theta \;:\; \mathbb{E}_\theta[f(T)] = 0 \quad \Rightarrow \quad \forall\,\theta \in \Theta \;:\; f(T) = 0 \quad P_\theta\text{-f.s.}.$$

□

§3.4.2 **Bemerkung**. Eine Statistik V auf $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ wird *unwesentlich* (ancillary) genannt, wenn ihre Verteilung $P^V_\bullet := P^V_\theta$ nicht vom Parameter θ abhängt. Sie heißt *unwesentlich erster Ordnung*, falls $\mathbb{E}_\bullet[V] := \mathbb{E}_\theta[V]$ unabhängig von θ ist. Falls jede Statistik der Form $V = f(T)$, die ancillary erster Ordnung ist, auch P_θ-f.s. konstant ist, so ist keine redundante Information mehr in T enthalten, und T ist vollständig (verwende $\widetilde{f}(T) = f(T) - \mathbb{E}_\bullet[f(T)]$). □

§3.4.3 **Lemma von Basu**. *Es seien* T *und* V *Statistiken auf* $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$. *Ist* T *erschöpfend und vollständig sowie* V *unwesentlich (ancillary), d.h.* $P^V_\bullet$ *ist unabhängig von* $\theta \in \Theta$, *so sind* T *und* V *unabhängig.* □

Beweis von Lemma §3.4.3. in der Vorlesung. □

§3.4.4 **Satz von Koopman**. *Es sei* $\mathcal{P}_\Theta$ *eine (strikt)* k*-parametrische Exponentialfamilie in* T *mit natürlichem Parameter* $\theta \in \Theta \subseteq \mathbb{R}^k$. *Besitzt* Θ *ein nichtleeres Inneres,* $\mathrm{int}(\Theta)$, *so ist* T *erschöpfend und vollständig.* □

Beweis von Satz §3.4.4. in der Vorlesung. □

§3.4.5 **Bemerkung**. Der natürliche Parameterraum Θ_{nat} einer (strikt) k-parametrischen Exponentialfamilie ist konvex und enthält ein nicht entartetes k-dimensionales Rechteck. □

§3.4.6 **Korollar**. *Gegeben sei eine (strikt) $(1+k)$-parametrische natürliche Exponentialfamilie in (U,T) der Form*

$$\frac{dP_{\theta,\tau}}{d\mu}(x) = C(\theta,\tau)\exp(\theta U(x) + \langle \tau, T(x)\rangle).$$

Des weiteren existiere ein $\theta_o \in \mathbb{R}^1$ und ein $\tau \in \mathbb{R}^k$ so dass $(\theta_o,\tau) \in \operatorname{int}(\Theta)$ gilt. Hängt die Verteilung einer Statistik V nicht von τ ab, so sind die Statistiken V und T unter jedem $P_{\theta_o,\tau}$ unabhängig. □

Beweis von Korollar §3.4.6. in der Vorlesung. □

§3.4.7 **Beispiele**. (a) Besitzt die Designmatrix X einen vollen Spaltenrang ($\operatorname{rg}(X) = k$) so liegt in einem gewöhnlichen normalen linearen Modell $Y \circlearrowleft \{\mathfrak{N}(X\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^k, \sigma > 0\}$ eine (strikt) $(k+1)$-parametrische Exponentialfamilie in $\eta(\beta,\sigma) = \sigma^{-2}(\beta, -1/2)^t \in \mathbb{R}^k \times \mathbb{R}_-$ und $T(Y) = (X^tY, \|Y\|^2)^t \in \mathbb{R}^k \times \mathbb{R}_+$ vor. Der natürliche Parameterraum $\Theta_{nat} = \mathbb{R}^k \times \mathbb{R}_-$ besitzt ein nichtleeres Inneres in $\mathbb{R}^{k+1}$, so dass T erschöpfend und vollständig ist. Mittels einer bijektiven Transformation ergibt sich, dass für den (gewöhnlichen) Kleinste-Quadrate-Schätzer $\widehat{\beta} = (X^tX)^{-1}X^tY$ und $\widehat{\sigma}^2 = (n-k)^{-1}\|Y - X\widehat{\beta}\|^2$ auch die Statistik $(\widehat{\beta}, \|Y\|^2) = (\widehat{\beta}, \|\Pi_{\mathcal{R}(X)}Y\|^2 + (n-k)\widehat{\sigma}^2)$ erschöpfend und vollständig ist. Da weiterhin gilt $\Pi_{\mathcal{R}(X)}Y = X\widehat{\beta}$, ist auch $(\widehat{\beta}, \widehat{\sigma}^2)$ erschöpfend und vollständig. Insbesondere, sind $\widehat{\beta}$ und $\widehat{\sigma}^2$ unabhängig.

(b) Sind $X_1, \ldots, X_n$ unabhängige und identisch $\mathfrak{U}([0,\theta])$-verteilte ZV'en mit unbekanntem Parameter $\theta > 0$ dann folgt aus der Form $L(\theta,x) = \theta^{-n}\mathbb{1}_{\{x_{(n)} \leqslant \theta\}}\mathbb{1}_{\{0 \leqslant x_{(1)}\}}$ der Likelihood-Funktion, dass das Maximum $T(X) := X_{(n)}$ der Beobachtungen eine erschöpfende Statistik ist. Da $L_T(\theta,t) = n\theta^{-n}t^{n-1}\mathbb{1}_{\{0 \leqslant t \leqslant \theta\}}$ die Likelihood-Funktion von T ist, folgt aus

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

für alle $\theta > 0$, dass $f = 0$ Lebesgue-f.ü. gelten muss, woraus die Vollständigkeit für $X_{(n)}$ folgt. □

3.5 Erwartungstreue Schätzer

§3.5.1 **Satz** *(Lehmann-Scheffé). Es seien $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein statistisches Experiment, $\widehat{\gamma}$ ein erwartungstreuer Schätzer des interessierenden Parameters $\gamma : \Theta \to \mathbb{R}$ und T eine erschöpfende und vollständige Statistik für $\mathcal{P}_\Theta$. Dann ist $\widehat{\gamma}_o = \mathbb{E}_{\bullet|T}(\widehat{\gamma})$ der eindeutig bestimmte Schätzer, der in der Klasse aller erwartungstreuen Schätzer gleichmäßig die kleinste Varianz besitzt (KVS für Kleinste-Varianz-Schätzer oder UMVU für uniformly minimum variance unbiased oder BUE für best unbiased estimator). Insbesondere gilt damit:*

(a) (Existenz) *Es gibt einen KVS der Form $\widehat{\gamma}_o(x) = g(T(x))$ für alle $x \in \mathcal{X}$.*

(b) (Eindeutigkeit) *Ist $\widetilde{\gamma}$ ein KVS, dann gilt $P_\theta(\widetilde{\gamma} = \widehat{\gamma}_o) = 1$ für alle $\theta \in \Theta$.*

(c) Ist $\widetilde{\gamma} = h(T)$ erwartungstreu für γ, dann gilt $P_\theta(\widetilde{\gamma} = \widehat{\gamma}_o) = 1$ für alle $\theta \in \Theta$. □

Beweis von Satz §3.5.1. in der Vorlesung. □

§3.5.2 **Bemerkung**. Aus dem Satz §3.2.14 (Rao-Blackwell Verbesserung) folgt die Aussage des Satzes von Lehmann-Scheffé sogar für das Risiko bzgl. einer beliebigen im zweiten Argument strikt konvexen Verlustfunktion. □

§3.5.3 **Beispiele** (§3.4.7 fortgesetzt). (a) Da $\widehat{\beta}$ und $\widehat{\sigma}$ erwartungstreue Schätzer die insbesondere erschöpfend und vollständig sind, besitzen beide Schätzer jeweils minimale Varianz in der Klasse aller erwartungstreuen Schätzer von β und σ^2. Für diese Aussage ist die Normalverteilungsannahme essentiell.

(b) Da $X_{(n)}$ eine erschöpfende und vollständige Statistik mit $\mathbb{E}_\theta(X_{(n)}) = \frac{n}{n+1}\theta$ für alle $\theta \in \Theta$ ist, besitzt der erwartzungstreue Schätzer $\widehat{\theta}_o = \frac{n+1}{n} X_{(n)}$ minimale Varianz in der Klasse aller erwartungstreuen Schätzer von θ. □

§3.5.4 **Bemerkung (Berechnung des Kleinste-Varianz-Schätzer)**. Sei T vollständig und erschöpfend für $\mathcal{P}_\Theta$. Möchte man den Kleinste-Varianz-Schätzer für γ bestimmen so gibt es zwei Möglichkeiten, ihn zu berechnen (die Existenz vorausgesetzt):

(a) *(Direkte Methode, geeigneter für diskrete Verteilungen)* Man sucht einen erwartungstreuen Schätzer der Form $\widehat{\gamma}_o = h(T)$ für γ, dieser ist dann der Kleinste-Varianz-Schätzer. Dies führt zu folgendem Gleichungssystem für die unbekannte Funktion h

$$\forall\, \theta \in \Theta : \quad \gamma(\theta) = \mathbb{E}_\theta[\widehat{\gamma}_o] = \mathbb{E}_\theta[h(T)] = \int h(t) P_\theta^T(dt).$$

Als Übungsaufgabe benutze diese Methode im Fall von Beispiel §3.3.8 (b) für den abgeleiteten Parameter $\gamma(\pi) = \pi(1-\pi)$.

(b) *(Benutze Rao-Blackwell Verbesserung)* Für einen beliebigen erwartungstreuen Schätzer $\widetilde{\gamma}$ ist die Rao-Blackwell Verbesserung $\widehat{\gamma}_o = h(T) = \mathbb{E}_\bullet[\widetilde{\gamma}|T]$ dann der Kleinste-Varianz-Schätzer. Die Berechnung kann entweder direkt mit bedingten Dichten durchgeführt werden (häufig aufwendig), oder man nutzt die Charakterisierung der bedingten Erwartung

$$\forall\, \theta \in \Theta : \forall\, A \in T^{-1}(\mathcal{S}) : \quad \mathbb{E}_\theta\left[\mathbb{1}_A h(T)\right] = \mathbb{E}_\theta\left[\mathbb{1}_A \widetilde{\gamma}\right]$$

was erneut zu einem Gleichungssystem für h führt. Wir haben dieses Verfahren in Beispiel §3.2.17 benutzt. □

§3.5.5 **Bemerkung (Kritik)**. Kleinste-Varianz-Schätzer werden häufig kritisiert, da

(a) Eine Einschränkung auf erwartungstreue Schätze zu viele Schätzer ausschließt. Aus dem Satz von Lehmamn-Scheffé wird deutlich, dass es nur ein von einer erschöpfenden und vollständigen Statistik abhängenden Schätzer existiert.

(b) Die Einschränkung auf erwartungstreue Schätzer schließt häufig interessante Schätzer mit geringerem Risiko aus, da eventuell ein Schätzer mit einer kleinen Verfälschung eine deutlich geringe Varianz besitzen kann (siehe nach folgendes Beispiel §3.5.6(a)).

(c) Es Situationen gibt, in denen erwartungstreue Schätzer und Kleinste-Varianz-Schätzer völlig unsinnig sind (siehe nach folgendes Beispiel §3.5.6(a)).

(d) Es Situationen gibt, in denen erwartungstreue Schätzer und Kleinste-Varianz-Schätzer nicht existieren (Übungsaufgabe). □

§3.5.6 **Beispiele.** (a) *(Fortsetzung von §3.3.8(a))*. Betrachte für σ^2 die folgenden Schätzer

$$\widehat{\sigma}_c^2 := c\sum_{i=1}^n (X_i - \overline{X})^2, \quad \text{wobei für} \quad c = \begin{cases} \frac{1}{n-1} \\ \frac{1}{n} \end{cases} \text{ ist } \quad \widehat{\sigma}_c^2 \quad \begin{cases} \text{KVS;} \\ \text{MLS (im nächsten Kapitel).} \end{cases}$$

Es gilt $\mathbb{E}(\widehat{\sigma}_c^2) = c(n-1)\sigma^2$ und $\mathbb{V}\mathrm{ar}(\widehat{\sigma}_c^2) = c^2(2n-2)\sigma^4$, so dass bzgl. der quadratischen Verlustfunktion für das Risiko von $\widehat{\sigma}_c^2$ gilt

$$\begin{aligned}\mathfrak{R}_\nu(\sigma^2, \widehat{\sigma}_c^2) = \mathbb{V}\mathrm{ar}(\widehat{\sigma}_c^2) + \left[\mathbb{E}(\widehat{\sigma}_c^2) - \sigma^2\right]^2 \\ = \big((n^2-1)c^2 - 2(n-1)c + 1\big)\sigma^4 = h(c)\sigma^4.\end{aligned}$$

Bestimmung der Minimalstelle von h liefert nun $h'(c) = 2c(n^2-1) - 2(n-1) = 0 \Leftrightarrow c = (n+1)^{-1}$. Damit minimiert der Schätzer $(n+1)^{-1}\sum_{i=1}^n (X_i - \overline{X})^2$ das Risiko in der Klasse der Schätzer $\{c\sum_{i=1}^n (X_i - \overline{X})^2, c > 0\}$ und nicht der Kleinste-Varianz-Schätzer. Außerdem sieht man, dass der MLS in diesem Beispiel besser ist als der Kleinste-Varianz-Schätzer ($c = n^{-1}$ ist näher am Scheitelpunkt als $c = (n-1)^{-1}$).

(b) Es sei X eine $\mathfrak{Poi}(\lambda)$-verteilte ZV mit unbekanntem Parameter $\lambda > 0$. Gesucht ist der Kleinste-Varianz-Schätzer für $\gamma(\lambda) = \exp(-2\lambda)$. Es ist bekannt, dass $T(x) = x$ eine vollständige und erschöpfende Statistik für $\lambda > 0$ ist. Daher besitzt der Kleinste-Varianz-Schätzer die Form $\widehat{\gamma}_o = h(T)$ für eine Funktion h, welche Lösung des Gleichungssystems

$$\sum_{k=0}^\infty h(k)\frac{\lambda^k}{k!}e^{-\lambda} = \mathbb{E}_\lambda(\widehat{\gamma}_o(X)) = \gamma(\lambda) = \exp(-2\lambda)$$

ist und somit auch $\sum_{k=0}^\infty h(k)\frac{\lambda^k}{k!} = \exp(-\lambda) = \sum_{k=0}^\infty (-1)^k \frac{\lambda^k}{k!}$ erfüllt, d.h. aber $\widehat{\gamma}_o = (-1)^X$ ist Kleinste-Varianz-Schätzer für $\gamma(\lambda) = \exp(-2\lambda)$, was unsinnig ist. □

3.6 Informationsungleichungen

§3.6.1 **Lemma** *(Chapman-Robins-Ungleichung). Es seien $X \circlearrowleft \mathcal{P}_\Theta$ ein statistisches Experiment, $\widehat{\gamma}$ ein erwartungstreuer Schätzer des interessierenden Parameters $\gamma(\theta) \in \mathbb{R}$ und $\theta_o \in \Theta$. Für jedes $\theta \in \Theta$ mit $P_\theta \neq P_{\theta_o}$, $P_\theta \ll P_{\theta_o}$ und Likelihood-Funktion $L_{\theta_o}(\theta, x) = [dP_\theta/dP_{\theta_o}](x)$ mit $\mathbb{E}_{\theta_o}\big(|L_{\theta_o}(\theta)|^2\big) = \int_{\mathcal{X}} |L_{\theta_o}(\theta, x)|^2 P_{\theta_o}(dx) < \infty$ (kurz $L_{\theta_o}(\theta) \in \mathcal{L}^2_{P_{\theta_o}}$) gilt*

$$\mathbb{V}\mathrm{ar}_{\theta_o}(\widehat{\gamma}) = \mathbb{E}_{\theta_o}\big(|\widehat{\gamma} - \gamma(\theta_o)|^2\big) \geqslant \frac{|\gamma(\theta) - \gamma(\theta_o)|^2}{\mathbb{V}\mathrm{ar}_{\theta_o}\big(L_{\theta_o}(\theta)\big)}.$$

□

Beweis von Lemma §3.6.1. in der Vorlesung. □

§3.6.2 **Beispiel.** Es sei X eine $\mathfrak{Exp}(\theta)$-verteilte ZV mit unbekanntem Parameter $\theta > 0$. Dann ist die Likelihood-Funktion gegeben durch $L_{\theta_o}(\theta, x) = (\theta/\theta_o)\exp(-(\theta - \theta_o)x)$, $x \geqslant 0$. Im Fall $\theta > \theta_o/2$ gilt $L_{\theta_o}(\theta) \in \mathcal{L}^2_{P_{\theta_o}}$ und $\mathbb{V}\mathrm{ar}_{\theta_o}\big(L_{\theta_o}(\theta)\big) = \frac{(\theta-\theta_o)^2}{\theta_o(2\theta-\theta_o)}$.

Sei $\widehat{\theta}$ ein erwartungstreuer Schätzer für θ ist (d.h. $\gamma(\theta) = \theta$). Aus der Chapman-Robins-Ungleichung §3.6.1 folgt dann $\mathbb{V}\mathrm{ar}_{\theta_o}(\widehat{\theta}) \geqslant \sup_{\theta > \theta_o/2} \theta_o(2\theta - \theta_o) = \infty$. Sofern also beliebig große Werte θ zugelassen sind, existiert kein erwartungstreuer Schätzer von θ mit endlicher Varianz.

Betrachten wir den interessierenden Parameter $\gamma(\theta) = \theta^{-1}$ so schließen wir mit Hilfe der Chapman-Robins-Ungleichung §3.6.1, dass $\mathbb{V}\mathrm{ar}_{\theta_o}(\widehat{\gamma}) \geqslant \sup_{\theta>\theta_o/2} \frac{(2\theta-\theta_o)}{\theta^2\theta_o} = \theta_o^{-2}$. Der Schätzer $\widehat{\gamma}_o := X$ ist erwartungstreu für $\gamma(\theta) = \theta^{-1}$ mit Varianz $\mathbb{V}\mathrm{ar}_\theta(\widehat{\gamma}_o) = \theta^{-2}$ und erreicht somit diese Schranke. □

§3.6.3 **Definition**. Es sei $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein bzgl. eines σ-endlichen Maßes μ dominiertes statistisches Modell mit $\Theta \subset \mathbb{R}^k$ und Likelihood-Funktion $L(\theta, x) = [dP_\theta/d\mu](x)$. Weiterhin bezeichnet $\ell(\theta, x) = \log(L(\theta, x))$ (mit der Konvention $\log(0) := -\infty$) die *Loglikelihood-Funktion*. Das statistische Modell wird *Hellinger-differenzierbar* in $\theta_o \in \mathrm{int}(\Theta)$ genannt, falls ein $\mathbb{R}^k$-wertige Funktion $\dot{\ell}(\theta_o, x)$ existiert mit

$$\lim_{\theta\to\theta_o} \int_{\mathcal{X}} \Big(\frac{\sqrt{L(\theta,x)} - \sqrt{L(\theta_o,x)} - \frac{1}{2}\langle \dot{\ell}(\theta_o,x), \theta-\theta_o\rangle \sqrt{L(\theta_o,x)}}{\|\theta-\theta_o\|} \Big)^2 \mu(dx) = 0.$$

Die Abbildung $\theta \to \dot{\ell}(\theta_o)$ heißt auch *Score-Funktion* und

$$I(\theta_o) = \mathbb{E}_{\theta_o}\big(\dot{\ell}(\theta_o)\dot{\ell}(\theta_o)^t\big) = \int_{\mathcal{X}} \big(\dot{\ell}(\theta_o,x)\dot{\ell}(\theta_o,x)^t\big) P_{\theta_o}(dx)$$

wird *Fisher-Informationsmatrix* in $\theta_o \in \mathrm{int}(\Theta)$ genannt. □

§3.6.4 **Bemerkungen**. (a) Sofern alle folgenden Ausdrücke klassisch differenzierbar sind, so gilt

$$\nabla_\theta \sqrt{L(\theta,x)} = \frac{\nabla_\theta L(\theta,x)}{2\sqrt{L(\theta,x)}} = \frac{1}{2}\sqrt{L(\theta,x)}\nabla_\theta \log(L(\theta,x)) = \frac{1}{2}\sqrt{L(\theta,x)}\dot{\ell}(\theta,x)$$

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

(b) Die angenommene Differenzierbarkeit im $\mathcal{L}^2_\mu$-Mittel ist eine natürliche Verallgemeinerung der klassischen Differenzierbarkeit. Da $\int L(\theta,x)\mu(dx) = 1 < \infty$, folgt $\sqrt{L(\theta)} \in \mathcal{L}^2_\mu$, so dass man $\theta \mapsto \sqrt{L(\theta)}$ als $\mathcal{L}^2_\mu$-wertige Abbildung auffassen kann und die Verteilungen $\{P_\theta\}$ im geometrischen Sinne eine Untermannigfaltigkeit des Hilbertraumes $\mathcal{L}^2_\mu$ bilden. Insbesondere gilt notwendigerweise $\langle \dot{\ell}(\theta_o), \theta-\theta_o\rangle\sqrt{L(\theta_o)} \in \mathcal{L}^2_\mu$, d.h. $\int_{\mathcal{X}} |\langle \dot{\ell}(\theta_o,x), \theta-\theta_o\rangle^2 L(\theta_o,x)\mu(dx) = \mathbb{E}_{\theta_o}(|\langle \dot{\ell}(\theta_o), \theta-\theta_o\rangle|^2) < \infty$, und damit $\dot{\ell}(\theta_o) \in L^2_{P_\theta}(\mathbb{R}^k)$, so dass die Matrix $I(\theta_o)$ stets wohldefiniert ist.

(c) Nach Definition ist die Fisher-Informationsmatrix symmetrisch und positiv-semidefinit, da $\langle I(\theta_o)v, v\rangle = \mathbb{E}_{\theta_o}\big(|\langle \dot{\ell}(\theta_o), v\rangle|^2\big) \geqslant 0$ für alle $v \in \mathbb{R}^k$ gilt.

(d) Die Score-Funktion und die Fisher-Information sind unabhängig vom dominierenden Maß. Sei Q ein privilegiertes dominierendes Maß, dann gilt $L(\theta) = \frac{dP_\theta}{dQ}\frac{dQ}{d\mu}$, so dass in der Definition von $\dot{\ell}$ der Faktor $\frac{dQ}{d\mu}$ aus dem Integranden ausgeklammert werden kann und $\dot{\ell}$ ebenso die Definition bezüglich des dominierenden Maßes Q erfüllt. □

§3.6.5 **Lemma**. *Für alle $\theta \in \Theta \subseteq \mathbb{R}^k$ in einer Umgebung von $\theta_o \in \Theta$ gelte $P_\theta \ll P_{\theta_o}$ sowie die $\mathcal{L}^2_{P_{\theta_o}}$-Differenzierbarkeit der Likelihood-Funktion $L_{\theta_o}(\theta,x) := [dP_\theta/dP_{\theta_o}](x)$ in θ_o, d.h. für den Gradienten $\dot{L}_{\theta_o}(\theta_o,x) \in \mathbb{R}^k$ gilt*

$$\lim_{\theta\to\theta_o} \int \Big(\frac{L_{\theta_o}(\theta,x) - L_{\theta_o}(\theta_o,x) - \langle \dot{L}_{\theta_o}(\theta_o,x), \theta-\theta_o\rangle}{\|\theta-\theta_o\|} \Big)^2 P_{\theta_o}(dx) = 0.$$

Dann ist das Modell Hellinger-differenzierbar in θ_o mit $\dot{\ell}(\theta_o, x) = \dot{L}_{\theta_o}(\theta_o, x)$.

Beweis von Lemma §3.6.5. in der Vorlesung. □

§3.6.6 **Beispiel**. Es sei X eine ZV mit Lebesguedichte $f_\theta(x) = \frac{1}{2\sigma}\exp(-|x-\theta|/\sigma)$, $x \in \mathbb{R}$, $\sigma > 0$ bekannt und unbekanntem Parameter $\theta \in \mathbb{R}$. Für beliebige $\theta, \theta_o \in \mathbb{R}$ und $x \in \mathbb{R}$ gilt

$$L_{\theta_o}(\theta, x) = \exp\Big(-(|x-\theta| - |x-\theta_o|)/\sigma\Big).$$

und L_{θ_o} ist $\mathcal{L}^2_{P_{\theta_o}}$-differenzierbar (Nachweis!) mit

$$\dot{L}(\theta_o, x) = (\mathbb{1}_{\{x-\theta_o>0\}} - \mathbb{1}_{\{x-\theta_o<0\}})/\sigma$$

Mit Lemma §3.6.5 gilt $\dot{\ell}(\theta_o) = \dot{L}_{\theta_o}(\theta_o)$ und für die Fisher-Information erhalten wir

$$I(\theta_o) = \mathbb{V}\mathrm{ar}_{\theta_o}\left(\mathbb{1}_{(X-\theta_o>0)} - \mathbb{1}_{(X-\theta_o<0)}\right)/\sigma^2 = 1/\sigma^2.$$

Die Fisher-Information hängt somit nicht vom unbekannten Parameter ab, was nur selten der Fall ist. □

§3.6.7 **Satz** *(Cramér-Rao-Schranke). Es seien $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ mit $\Theta \subseteq \mathbb{R}^k$ ein statistisches Experiment, $\gamma : \Theta \to \mathbb{R}$ differenzierbar in $\theta_o \in \mathrm{int}(\Theta)$ und $\widehat{\gamma}$ ein erwartungstreuer Schätzer für $\gamma(\theta)$. Für alle θ in einer Umgebung von θ_o gelte $P_\theta \ll P_{\theta_o}$ sowie die $\mathcal{L}^2_{P_{\theta_o}}$-Differenzierbarkeit der Likelihood-Funktion $L_{\theta_o}(\theta, x) = [dP_\theta/dP_{\theta_o}](x)$ in θ_o. Falls die Fisher-Informationsmatrix $I(\theta_o)$ strikt positiv-definit ist, gilt die Cramér-Rao-Ungleichung als untere Schranke für das Risiko bezüglich der quadratischen Verlustfunktion*

$$\mathbb{E}_{\theta_o}\Big(\big|\widehat{\gamma} - \gamma(\theta_o)\big|^2\Big) = \mathbb{V}\mathrm{ar}_{\theta_o}(\widehat{\gamma}) \geqslant \langle I(\theta_o)^{-1}\dot{\gamma}(\theta_o), \dot{\gamma}(\theta_o)\rangle.$$

□

Beweis von Satz §3.6.7. in der Vorlesung. □

§3.6.8 **Bemerkung**. Ist $\widehat{\gamma}$ kein erwartungstreuer Schätzer für $\gamma(\theta)$ aber $\widehat{\gamma} \in \mathcal{L}^1_{P_\theta}$ für alle $\theta \in \Theta$, so ist $\widehat{\gamma}$ ein erwartungstreuer Schätzer für $g(\theta) := \mathbb{E}_\theta(\widehat{\gamma})$. In dieser Situation liefert die Cramér-Rao-Ungleichung mit Hilfe der Bias-Varianz-Zerlegung

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

Diese Abschätzung ist insbesondere hilfreich, in Situationen in denen erwartungstreue Schätzer von $\gamma(\theta)$ nicht existieren oder nicht erstrebenswerte Eigenschaften besitzen. □

§3.6.9 **Lemma**. *Es gelten die Annahmen und Notation aus Satz §3.3.9, so dass $\mathcal{P}_{\Theta_{nat}}$ eine (strikt) k-parametrische natürliche Exponentialfamilie in T mit natürlichem Parameterraum Θ_{nat} bildet. Dann ist $\mathcal{P}_{\Theta_{nat}}$ im Innern von Θ_{nat} insbesondere $\mathcal{L}^2_{P_\theta}$ und Hellinger-differenzierbar mit Fisher-Information $I(\theta) = \ddot{A}(\theta)$.*
Sofern $I(\theta_o)$ strikt positiv-definit ist, so erreicht T_i, $i = 1, \ldots, k$, als erwartungstreuer Schätzer von $\gamma_i(\theta) = \mathbb{E}_\theta(T_i)$ die Cramér-Rao-Schranke (ist Cramér-Rao-effizient) in $\theta_o \in \mathrm{int}(\Theta_{nat})$. □

Beweis von Lemma §3.6.9. in der Vorlesung. □

§3.6.10 **Beispiel** (§3.2.5 fortgesetzt). Betrachte das normale Lokations-Modell $X \circlearrowleft \{\mathfrak{N}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}\}$ für bekanntes $\sigma > 0$. Ein erwartungstreuer Schätzer von μ ist $\widehat{\mu} = \overline{X}$. Dann gilt $\mathbb{V}\mathrm{ar}_\mu(\widehat{\mu}) = \sigma^2/n$ sowie für die Fisher-Information $I(\mu) = n/\sigma^2$ (da $A(\mu) = \frac{n\mu^2}{2\sigma^2}$, $\ddot{A}(\mu) = n/\sigma^2$). Also ist $\widehat{\mu}$ effizient im Sinne der Cramér-Rao-Ungleichung. Um nun $\gamma(\mu) := \mu^2$ zu schätzen, betrachte den erwartungstreuen Schätzer $\widehat{\gamma} = (\overline{X})^2 - \sigma^2/n$. Dann gilt $\mathbb{V}\mathrm{ar}_\mu(\widehat{\gamma}) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}$ (Nachweis!), während die Cramér-Rao-Ungleichung die untere Schranke $4\mu^2\sigma^2/n$ liefert. Damit ist $\widehat{\gamma}$ nicht Cramér-Rao-effizient. Allerdings ist $\overline{X}$ eine erschöpfende und vollständige Statistik, so dass der Satz §3.5.1 von Lehmann-Scheffé zeigt, das $\widehat{\gamma}$ minimale Varianz unter allen erwartungstreuen Schätzern besitzt. Demnach ist die Cramér-Rao-Schranke hier nicht scharf. □

§3.6.11 **Bemerkung**. Die Cramér-Rao-Schranke wird nur erreicht wenn $\mathcal{P}_\Theta$ eine Exponentialfamilie in T bildet und $\gamma(\theta) = \mathbb{E}_\theta(T)$ oder eine lineare Funktion davon zu schätzen sind. Wegen der Vollständigkeit der Statistik T könnte man in diesen Fällen auch mit dem Satz §3.5.1 von Lehmann-Scheffé argumentieren. Im nächsten Kapitel betrachten wir allgemeinere Schätzverfahren die zu mindestens asymptotisch die Cramér-Rao-Schranke erreichen. □

§3.6.12 **Lemma**. *Es sei $(\mathcal{X}, \mathscr{A}, \Theta)$ mit $\Theta \subseteq \mathbb{R}^k$ ein in $\theta_o \in \Theta$ Hellinger-differenzierbares statistisches Experiment. Dann ist die Likelihood-Funktion $L(\theta, x) = [dP_\theta/d\mu](x)$ insbesondere $\mathcal{L}^1_\mu$-differenzierbar mit Ableitung $\dot{L}(\theta, x) = \dot{\ell}(\theta, x)L(\theta, x)$, und es gilt $\mathbb{E}_{\theta_o}[\dot{\ell}(\theta_o)] = 0$.*

Beweis von Lemma §3.6.12. in der Vorlesung. □

§3.6.13 **Lemma**. *Es seien $X_1, \ldots, X_n$ unabhängige ZV'en zu Hellinger-differenzierbaren statistischen Modellen über derselben Parametermenge $\Theta \subseteq \mathbb{R}^k$. Im statistischen Modell der ZV X_j bezeichne I_j die Fisher-Informationsmatrix, so ist das Produktmodell, erzeugt von $X_1, \ldots, X_n$, Hellinger-differenzierbar mit Fisher-Informationsmatrix*

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

□

Beweis von Lemma §3.6.13. in der Vorlesung. □

§3.6.14 **Beispiel** (§3.6.6 fortgesetzt). Es seien $X_1, \ldots, X_n$ u.i.v. ZV'en mit Lebesguedichte $f_\theta(x) = \frac{1}{2\sigma}\exp(-|x-\theta|/\sigma)$, $x \in \mathbb{R}$, $\sigma > 0$ bekannt und unbekanntem Parameter $\theta \in \mathbb{R}$. Da $I_j(\theta) = 1/\sigma^2$ die Fisher-Information im statistischen Modell der ZV X_j für $j = 1, \ldots, n$ gilt, erhalten wir die Fisher-Information $I(\theta) = n/\sigma^2$ im Produktmodell. □

3.7 Translations-äquivariante Schätzer

Wir betrachten im Folgenden translations-äquivariante Schätzer für einen unbekannten Parameter $\theta \in \mathbb{R}$, d.h. Schätzer $\widehat{\theta}$ mit der Eigenschaft $\widehat{\theta}(X + a\mathbb{1}_n) = \widehat{\theta}(X) + a$ wobei wir für $x \in \mathbb{R}^n$ und $a \in \mathbb{R}$ schreiben $x + a\mathbb{1}_n = (x_1 + a, \ldots, x_n + a)$. Skalen-äquivariante Schätzer sowie eine allgemeinere Darstellung findet man zum Beispiel in Kapitel 3 in Lehmann and Casella [1998].

§3.7.1 **Definition**. Es seien $(\mathbb{R}^n, \mathscr{B}^n, \mathcal{P}_\Theta)$ ein bzgl. eines translations-invarianten, σ-endlichen Maßes μ dominiertes statistisches Experiment mit Likelihood-Funktion $L(\theta, x) = [dP_\theta/d\mu](x)$,

$\Theta \subseteq \mathbb{R}$ sowie $0 \in \Theta$. Wir bezeichnen $\mathcal{P}_\Theta$ als *Lokations-Familie*, falls

$$L(\theta, x) = L(0, x - \theta \mathbb{1}_n) \quad \text{für alle } x \in \mathbb{R}^n \text{ und } \theta \in \Theta$$

gilt. Ein Schätzer $\widehat{\theta}$ von θ heißt *translations-äquivariant* (TäS), falls $\widehat{\theta}(x + a\mathbb{1}_n) = \widehat{\theta}(x) + a$ für alle $x \in \mathbb{R}^n$, $a \in \mathbb{R}$ gilt. Wir bezeichnen eine Verlustfunktion ν und das entsprechende Risiko $\mathfrak{R}_\nu$ als *translations-invariant*, falls $\nu(\theta, \widehat{\theta}) = \nu(\theta + a, \widehat{\theta} + a)$ für alle $\theta, \widehat{\theta}, a \in \mathbb{R}$ gilt. Ein TäS $\widehat{\theta}_o$ heißt *bester translations-äquivariant Schätzer* bezüglich des translations-invarianten Risikos $\mathfrak{R}_\nu$, falls $\mathfrak{R}_\nu(\theta, \widehat{\theta}_o) \leqslant \mathfrak{R}_\nu(\theta, \widehat{\theta})$ für alle $\widehat{\theta} \in \Delta_{\text{TäS}}$ und $\theta \in \Theta$ gilt.

§3.7.2 **Proposition**. *Es seien $\mathcal{P}_\Theta$ eine Lokations-Familie und $\mathfrak{R}_\nu$ ein translations-invariantes Risiko. Für jeden translations-äquivarianten Schätzer $\widehat{\theta}$ gilt $\mathfrak{R}_\nu(\theta, \widehat{\theta}) = \mathfrak{R}_\nu(0, \widehat{\theta})$, für alle $\theta \in \Theta$, d.h. das Risiko ist konstant in (unabhängig von) θ, und somit ist $\widehat{\theta}_o$ ein bester translations-äquivarianter Schätzer, falls $\mathfrak{R}_\nu(0, \widehat{\theta}_o) \leqslant \mathfrak{R}_\nu(0, \widehat{\theta})$ für alle translations-äquivarianten Schätzer $\widehat{\theta}$ gilt.*

Beweis von Proposition §3.7.2. in der Vorlesung. □

§3.7.3 **Lemma**. *Es seien $n \geqslant 2$, $V(X) := (X_1 - X_n, \ldots, X_{n-1} - X_n)$, $\widehat{\theta}$ ein translations-äquivarianter Schätzer und $\widetilde{\theta}$ ein beliebiger Schätzer für θ. Dann sind die folgenden Aussagen äquivalent:*

(i) $\widetilde{\theta}$ ist ein translations-äquivarianter Schätzer.

(ii) Es existiert eine translations-invariante Funktion $u : \mathbb{R}^n \to \mathbb{R}$, d.h. $u(x + a\mathbb{1}_n) = u(x)$ für alle $x \in \mathbb{R}^n$, $a \in \mathbb{R}$, so dass $\widetilde{\theta} = \widehat{\theta} + u$ gilt.

(iii) Es existiert eine Abbildung $h : \mathbb{R}^{n-1} \to \mathbb{R}$ mit $\widetilde{\theta} = \widehat{\theta} - h(V)$. □

Beweis von Lemma §3.7.3. in der Vorlesung. □

§3.7.4 **Satz**. *Es seien $\mathcal{P}_\Theta$ eine Lokations-Familie mit $n \geqslant 2$, $V(X) := (X_1 - X_n, \ldots, X_{n-1} - X_n)$, $\mathfrak{R}_\nu$ ein translations-invariantes Risiko und $\widehat{\theta}$ ein translations-äquivarianter Schätzer mit $\mathfrak{R}_\nu(0, \widehat{\theta}) < \infty$. Falls eine Funktion $h^* : \mathbb{R}^{n-1} \to \mathbb{R}$ mit*

$$\mathbb{E}_0\big[\nu(0, \widehat{\theta} - h^*(v)) | V = v\big] = \min_{h \in \mathbb{R}} \mathbb{E}_0\big[\nu(0, \widehat{\theta} - h) | V = v\big]$$

existiert, so ist $\widehat{\theta}_o := \widehat{\theta} - h^*(V)$ ein bester translations-äquivarianter Schätzer bezüglich des translations-invarianten Risiko $\mathfrak{R}_\nu$. □

Beweis von Satz §3.7.4. in der Vorlesung. □

§3.7.5 **Korollar**. *Die Voraussetzungen des Satzes §3.7.4 seien erfüllt.*

(i) Ist ν die quadratische Verlustfunktion, d.h. $\nu(\theta, e) = (\theta - e)^2$, so ist der beste translations-äquivariante Schätzer $\widehat{\theta}_o$ eindeutig bestimmt mit $h^*(v) = \mathbb{E}_0(\widehat{\theta} | V = v)$.

(ii) Für den Absolutbetrag $\nu(\theta, e) = |\theta - e|$ ist $\widehat{\theta}_o$ ein bester translations-äquivarianter Schätzer falls h^* ein Median der bedingten $P_0^{\widehat{\theta}|V}$ von $\widehat{\theta}$ gegeben V ist. □

Beweis von Korollar §3.7.5. in der Vorlesung. □

§3.7.6 **Beispiel** (§3.6.10 fortgesetzt). Betrachte das normale Lokations-Modell $X \circlearrowleft \{\mathfrak{N}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}\}$ für bekanntes $\sigma > 0$. Das arithmetische Mittel $\widehat{\mu} = \overline{X}$ ist ein translations-äquivarianter Schätzer. Da $V = (X_1 - X_n, \ldots, X_{n-1} - X_n) \sim \mathfrak{N}^{\otimes(n-1)}(0, 2\sigma^2)$ gilt, ist die Verteilung $P_\bullet^V := P_\mu^V$ unabhängig von μ und somit ist V eine unwesentliche (ancillary) Statistik für μ. Nach dem Lemma §3.4.3 von Basu sind V und $\overline{X}$ somit unabhängig ($\overline{X}$ ist erschöpfend und vollständig) und $h^*(V) = h^*$ im Satz §3.7.4 ist konstant. Betrachten wir eine quadratische Verlustfunktion, so sieht man leicht dass $h^* = 0$ gilt. Damit ist $\widehat{\mu}$ auch ein bester translations-äquivarianter Schätzer bezüglich des quadratischen Risiko. Allgemeiner, ist $\nu(\theta, e) = \rho(\theta - e)$ für eine konvexe und gerade Funktion ρ, so minimiert h^* den Ausdruck $\mathbb{E}_0\rho(\overline{X} - v)$. Man kann leicht zeigen, dass ein Minimum für $h^* = 0$ angenommen wird, d.h. $\widehat{\mu}$ ist auch in dieser Situation ein bester translations-äquivarianter Schätzer. □

§3.7.7 **Satz**. *Für das quadratische Risiko gilt unter den Annahmen von Satz §3.7.4 für den besten translations-äquivarianten Schätzer*

$$\widehat{\theta}_o(x) = \frac{\int_{-\infty}^{\infty} uL(0, x - u\mathbb{1}_n)du}{\int_{-\infty}^{\infty} L(0, x - u\mathbb{1}_n)du}.$$

In dieser Form heißt der Schätzer $\widehat{\theta}_o$ Pitman-Schätzer.

Beweis von Satz §3.7.7. in der Vorlesung. □

§3.7.8 **Beispiel**. Es seien $X_1, \ldots, X_n$ unabhängige und identisch $\mathfrak{U}([\theta - \frac{1}{2b}, \theta + \frac{1}{2b}])$-verteilte ZV'en. Dann gilt für die Likelihood-Funktion

$$L(\theta, x) = L(0, x - \theta\mathbb{1}_n) = b^n \prod_{i=1}^{n} \mathbb{1}_{(-\frac{1}{2b} \leqslant x_i - \theta \leqslant \frac{1}{2b})} = b^n \mathbb{1}_{(x_{(n)} - \frac{1}{2b} \leqslant \theta \leqslant x_{(1)} + \frac{1}{2b})}$$

und $\widehat{\theta}_o = \frac{1}{2}(x_{(1)} + x_{(n)})$ ist der Pitman-Schätzer. □

Kapitel 4

Allgemeine Schätzmethoden

4.1 Momentenschätzer

§4.1.1 **Definition**. Es sei $(\mathcal{X}^n, \mathscr{A}^{\otimes^n}, \mathcal{P}_\Theta^{\otimes^n} = \{P_\theta^{\otimes^n}, \theta \in \Theta\})$ ein statistisches (Produkt-)Modell mit $\mathcal{X} \subseteq \mathbb{R}$, $\mathscr{A} \subseteq \mathscr{B}_\mathbb{R}$ und abgeleiteten Parameter $\gamma : \Theta \to \mathbb{R}^p$. Ferner sei $\psi = (\psi_1, \ldots, \psi_q) : \mathcal{X} \to \mathbb{R}^q$ mit Koordinatenfunktionen $\psi_j \in \mathcal{L}^1_{P_\theta}$, $j = 1, \ldots, k$, für alle $\theta \in \Theta$, und

$$\varphi(\theta) := \mathbb{E}_\theta(\psi) = \Big(\mathbb{E}_\theta(\psi_1), \ldots, \mathbb{E}_\theta(\psi_q)\Big)^t.$$

Existiert weiterhin eine Borel-messbare Funktion $\Gamma : \varphi(\Theta) \to \gamma(\Theta)$ mit $\Gamma \circ \varphi = \gamma$ und liegt $\frac{1}{n}\sum_{i=1}^n \psi(x_i)$ in $\varphi(\Theta)$ für all $x_1, \ldots, x_n \in \mathcal{X}$, so wird $\Gamma\big(\frac{1}{n}\sum_{i=1}^n \psi(x_i)\big)$ *(verallgemeinerter) Momentenschätzer* für $\gamma(\theta)$ mit Momentenfunktionen $\psi_1, \ldots, \psi_q$ genannt. □

§4.1.2 **Beispiele**. (a) Es seien $X_1, \ldots, X_n$ unabhängige und identisch $\mathfrak{Exp}(\lambda)$-verteilte ZV'en mit unbekanntem Parameter $\lambda > 0$. Betrachte die übliche Momentenfunktion $\psi(x) = x^k$ für ein $k \in \mathbb{N}$, dann gilt $\varphi(\lambda) = \mathbb{E}_\lambda(X_i^k) = \lambda^{-k}k!$. Ist $\gamma(\lambda) = \lambda$ der abgeleitete Parameter, so ergibt sich $\Gamma \circ \varphi = \gamma$ für $\Gamma(x) = (k!/x)^{1/k}$. Der k-te Momentenschätzer für λ ist damit

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

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

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

mit $\{\varepsilon_i, i \geqslant 1\}$ u.i.v., $\mathbb{E}(\varepsilon_1) = 0$, $\mathbb{V}\mathrm{ar}(\varepsilon_1) = \sigma^2 < \infty$ und $X_0 = x_0 \in \mathbb{R}$. Insbesondere, motiviert in dieser Situation die Identität $\mathbb{E}[X_{n-1}X_n|\varepsilon_1, \ldots, \varepsilon_{n-1}] = aX_{n-1}^2$, zum *Yule-Walker-Schätzer*

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

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

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

□

§4.1.3 **Lemma**. *Existiert für hinreichend großes n der Momentenschätzer $\widehat{\gamma}_n := \Gamma\big(\frac{1}{n}\sum_{i=1}^n \psi(x_i)\big)$ und ist Γ stetig, so ist $\widehat{\gamma}_n$ (stark) konsistent, d.h. $\lim_{n\to\infty} \widehat{\gamma}_n = \gamma(\theta)$ $P_\theta^{\otimes^\mathbb{N}}$-f.s..* □

Beweis von Lemma §4.1.3. in der Vorlesung. □

§4.1.4 **Satz** *(Δ-Methode). Es seien $(X_n)_{n\geqslant 1}$ eine Folge von zufälligen Vektoren in $\mathbb{R}^k$, $\sigma_n > 0$ mit $\lim_{n\to\infty}\sigma_n = 0$, $\theta_o \in \mathbb{R}^k$ sowie $\Sigma \in \mathbb{R}^{k\times k}$ positiv semi-definit und es gelte*

$$\sigma_n^{-1}(X_n - \theta_o) \xrightarrow{\mathcal{L}} \mathfrak{N}(0,\Sigma).$$

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

$$\sigma_n^{-1}(f(X_n) - f(\theta_o)) \xrightarrow{\mathcal{L}} \mathfrak{N}\big(0, \langle \Sigma \dot{f}(\theta_o), \dot{f}(\theta_o)\rangle\big),$$

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

Beweis von Satz §4.1.4. in der Vorlesung. □

§4.1.5 **Beispiel** (§3.5.6(b) fortgesetzt). Es seien $X_1,\dots,X_n$ unabhängig und identisch $\mathfrak{Poi}(\lambda)$-verteilte ZV'n mit unbekanntem Parameter $\lambda > 0$. Da die vollständige und erschöpfende Statistik $\widehat{\lambda} := \overline{X}$ ein erwartungstreuer Schätzer für λ ist, ist $\widehat{\lambda}$ der KVS. Nach dem zentralen Grenzwertsatz gilt $\sqrt{n}(\widehat{\lambda}_n - \lambda) \xrightarrow{\mathcal{L}} \mathfrak{N}(0,\lambda)$ unter $P_\lambda^{\otimes\mathbb{N}}$. Ist das Ziel nun ein asymptotisches Konfidenzintervall herzuleiten, so stört die Abhängigkeit der asymptotischen Varianz vom unbekannten Parameter. Betrachtet man nun $f(x) = 2x^{1/2}$ mit $\dot{f}(x) = x^{-1/2}$, so folgt mit Hilfe der Δ-Methode, dass $\sqrt{n}(2\widehat{\lambda}_n^{1/2} - 2\lambda^{1/2}) \xrightarrow{\mathcal{L}} \mathfrak{N}(0,1)$, so dass $[2\widehat{\lambda}_n^{1/2} - n^{-1/2}z_{1-\alpha/2}, 2\widehat{\lambda}_n^{1/2} + n^{-1/2}z_{1-\alpha/2}]$ ein asymptotisches $(1-\alpha)$-Konfidenzintervall für $2\lambda^{1/2}$ bildet, wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)$ einer Standardnormalverteilung bezeichnet. Eine Rücktransformation ergibt dann für λ selbst das asymptotische $(1-\alpha)$-Konfidenzintervall $[(\widehat{\lambda}_n^{1/2} - (4n)^{-1/2}z_{1-\alpha/2})^2, (\widehat{\lambda}_n^{1/2} + (4n)^{-1/2}z_{1-\alpha/2})^2]$. Die Idee, mittels Δ-Transformation eine vom unbekannten Parameter unabhängige asymptotische Varianz zu erhalten, ist in vielen Situation sehr erfolgreich und wird *Varianz-stabilisierende Transformation* genannt.

Alternativ kann man die asymptotische Varianz mittels $\widehat{\lambda}_n$ konsistent schätzen und mit Hilfe des Slutsky-Lemma dann auf $(n/\widehat{\lambda}_n)^{1/2}(\widehat{\lambda}_n - \lambda) \xrightarrow{\mathcal{L}} \mathfrak{N}(0,1)$ schließen. Daraus ergibt sich $[\widehat{\lambda}_n - (\widehat{\lambda}_n/n)^{1/2}z_{1-\alpha/2}, \widehat{\lambda}_n + (\widehat{\lambda}_n/n)^{1/2}z_{1-\alpha/2}]$ als asymptotisches Konfidenzintervall. □

§4.1.6 **Satz**. *Es seien $\theta_o \in \Theta$, $\gamma : \Theta \to \mathbb{R}$ und für hinreichend großes n existiere der Momentenschätzer $\widehat{\gamma}_n := \Gamma\big(\frac{1}{n}\sum_{i=1}^n \psi(x_i)\big)$ mit Momentenfunktion $\psi_j \in \mathcal{L}^2_{P_{\theta_o}}$, $j = 1,\dots,q$. Bezeichne mit $\Sigma_{\theta_o}(\psi)\mathbb{R}^{q\times q}$ die Kovarianzmatrix von ψ mit den Einträgen $(\Sigma_{\theta_o}(\psi))_{ij} = \mathbb{C}\mathrm{ov}_{\theta_o}(\psi_i,\psi_j)$ für $i,j = 1\dots,q$. Sofern Γ in einer Umgebung von $\varphi(\theta_o)$ stetig differenzierbar ist, ist $\widehat{\gamma}_n$ unter $P_{\theta_o}^{\otimes\mathbb{N}}$ asymptotische normalverteilt mit Rate $n^{-1/2}$, asymptotischen Erwartungswert Null und asymptotischer Varianz $\langle \Sigma_{\theta_o}(\psi)\dot{\Gamma}(\varphi(\theta_o)), \dot{\Gamma}(\varphi(\theta_o))\rangle$:*

$$\sqrt{n}(\widehat{\gamma}_n - \gamma(\theta_o)) \xrightarrow{\mathcal{L}} \mathfrak{N}\big(0, \langle \Sigma_{\theta_o}(\psi)\dot{\Gamma}(\varphi(\theta_o)), \dot{\Gamma}(\varphi(\theta_o))\rangle\big) \quad (\textit{unter } P_{\theta_o}^{\otimes\mathbb{N}}).$$

Beweis von Satz §4.1.6. in der Vorlesung. □

§4.1.7 **Bemerkung**. Die Begriffe *asymptotischer Erwartungswert* und *asymptotische Varianz* sind leicht irreführend, da nicht notwendigerweise gilt, dass die Momente von $\sqrt{n}(\widehat{\gamma}_n - \gamma(\theta_o))$ gegen die entsprechenden Momente der asymptotischen Verteilung konvergieren (dafür wird gleichgradige Integrierbarkeit benötigt). □

§4.1.8 **Beispiel** (§4.1.2 (a) fortgesetzt). Es gilt $\Sigma_{\lambda_o}(\psi) = \mathbb{V}\mathrm{ar}_{\lambda_o}(X_i^k) = \big((2k)! - (k!)^2\big)/\lambda_o^{2k}$ und $\Gamma'(x) = -(k!/x)^{1/k}(kx)^{-1}$. Alle Momentenschätzer $\widehat{\lambda}_{k,n}$ sind asymptotisch normalverteilt mit Rate $n^{-1/2}$ und asymptotischer Varianz $\sigma_k^2 = \lambda_o^2 k^{-2}\big((2k)!/(k!)^2 - 1\big)$. Da $\widehat{\lambda}_{1,n}$ die gleichmäßig kleinste asymptotische Varianz besitzt und auf der erschöpfenden Statistik $\overline{X}$ basiert, wird dieser Schätzer im Allgemeinen vorgezogen. □

§4.1.9 **Bemerkung**. Die Momentenmethode kann unter folgendem allgemeinem Gesichtspunkt betrachtet werden. Sind $X_1, \ldots, X_n$ unabhängige und identisch P_θ-verteilte ZV'en mit Werten in $\mathbb{R}$, so ist die *empirische Verteilungsfunktion* $\widehat{F}_n(x) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(X_i \leqslant x)}$ eine erschöpfende Statistik und nach dem Satz von Glivenko-Cantelli gilt P_θ-fs $\widehat{F}_n(x) \to F_\theta(x) = P_\theta(X_1 \leqslant x)$ gleichmäßig in $x \in \mathbb{R}$. Ist nun $\gamma(\theta)$ als Funktional $\Gamma(F_\theta(x), x \in \mathbb{R})$ darstellbar, so ist die empirische Version $\Gamma(\widehat{F}_n(x), x \in \mathbb{R})$ ein natürlicher Schätzer für $\gamma(\theta)$. Falls das Funktional Γ stetig bezüglich der Supremumsnorm ist, so folgt die Konsistenz.

Der Satz von Donsker für empirische Prozesse zeigt $\sqrt{n}(\widehat{F}_n - F_\theta) \xrightarrow{\mathcal{L}} G_\theta$ gleichmäßig auf $\mathbb{R}$ für einen zentrierten Gaußprozess G_θ mit Kovarianzstruktur $\mathbb{C}\mathrm{ov}(G_\theta(x), G_\theta(y)) = F_\theta(x \wedge y) - F_\theta(x)F_{Pa}(y)$. Ist Γ ein Hadamard-differenzierbares Funktional, so folgt $\sqrt{n}(\Gamma(\widehat{F}_n(x), x \in \mathbb{R}) - \gamma(\theta)) \xrightarrow{\mathcal{L}} \dot{\Gamma}(F_\theta)G_\theta$ unter P_θ, also insbesondere asymptotische Normalverteilung mit Rate $n^{-1/2}$ und explizit bestimmbarer asymptotischer Varianz. Eine detaillierte Darstellung findet man zum Beispiel in van der Vaart [1998].

Als einfaches (lineares) Beispiel sei $\gamma(\theta) = \mathbb{E}_\theta[\psi(X_1)]$ zu schätzen und $X_i \geqslant 0$ P_θ-f.s.. Dann folgt informell $\Gamma(F_\theta) = \int_0^\infty \psi(x)dF_\theta(x) = \int_0^\infty \psi'(x)(1 - F_\theta(x))dx$. Aus der Linearität folgt weiterhin $\dot{\Gamma}(F_\theta)G_\theta = \int_0^\infty \psi'(x)(-G_\theta(x))dx$, welches normalverteilt ist mit Erwartungswert Null und Varianz

$$\begin{aligned}
&\int_0^\infty \int_0^\infty \psi'(x)\psi'(y)(F_\theta(x \wedge y) - F_\theta(x)F_\theta(y))dxdy \\
&\qquad = \int_0^\infty \int_0^\infty \psi(x)\psi(y)\partial_{xy}(F_\theta(x \wedge y) - F_\theta(x)F_\theta(y))dxdy \\
&\qquad\qquad = \int_0^\infty \psi^2(x)dF_\theta(x) - \Big(\int_0^\infty \psi^2(x)dF_\theta(x)\Big)^2
\end{aligned}$$

was gerade der Varianz von $\Gamma(\widehat{F}_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i)$ entspricht. □

4.2 Maximum-Likelihood-Schätzer

§4.2.1 **Beispiele**. (a) Auf dem Stichprobenraum $\mathcal{X}$ sei jede Verteilung $P_\theta \in \mathcal{P}_\Theta$ durch eine Zähldichte p_θ gegeben. Die Verlustfunktion $\nu(\theta, \delta)$ sei homogen in $\theta \in \Theta$, dann ist eine plausible Schätzmethode für θ, bei Vorliegen einer Realisation x als Schätzwert $\widehat{\theta}(x)$ denjenigen Parameter $\theta \in \Theta$ zu wählen, für den die Wahrscheinlichkeit $p_\theta(x)$ des Eintretens von x maximiert wird, d.h. $\widehat{\theta}(x) := \arg\max_{\theta\in\Theta} p_\theta(x)$. Dieser Schätzer wird *Maximum-Likelihood-Schätzer* (MLS) genannt, Bereits im vorliegenden Fall ist weder Existenz noch Eindeutigkeit ohne Weiteres garantiert. Bei Mehrdeutigkeit wählt man einen maximierenden Parameter θ nach Belieben aus. Im Fall unabhängiger und identisch $\mathfrak{P}\mathrm{oi}(\lambda)$-verteilter ZV'en $X_1 \ldots, X_n$

mit unbekanntem Parameter $\lambda > 0$, ergibt sich beispielsweise

$$\widehat{\lambda} = \arg\max_{\lambda>0} \prod_{i=1}^{n} \left(e^{-\lambda}\frac{\lambda^{X_i}}{X_i!}\right) = \overline{X}$$

falls $\overline{X} > 0$ ist. Für $\overline{X} = 0$, d.h. $X_1 = \cdots = X_n = 0$ wird das Supremum nur asymptotisch für $\lambda \to 0$ erreicht. hier könnte man sich behelfen, indem man die Verteilungsfamilie mit $\mathfrak{Poi}(0)$ als Punktmaß in der Null stetig ergänzt.

(b) Ist jede Verteilung $P_\theta \in \mathcal{P}_\Theta$ durch eine Lebesgue-Dichte f_θ gegeben, so führt der Maximum-Likelihood-Ansatz analog zu dem Schätzwert $\widehat{\theta}(x) := \arg\max_{\theta\in\Theta} f_\theta(x)$. Sei $Y = \exp(X)$ mit $X \circlearrowleft \{\mathfrak{N}(\mu, 1, \mu \in \mathbb{R}\}$. Dann ist Y log-normalverteilt, und es gilt

$$\widehat{\mu}(Y) = \arg\max_{\mu\in\mathbb{R}} \frac{\exp(-(\log(Y)-\mu)^2/2)}{\sqrt{2\pi}Y} = \log(Y).$$

Auf der anderen Seite wird X beobachtet so erhält man den MLS $\widetilde{\mu} = X$. Der MLS ist somit invariant unter Parametertransformation, da Einsetzen von $X = \log(Y)$ auf das selbe Ergebnis führt. Interessanterweise führt die Momentenmethode unter Benutzung von $\mathbb{E}_\mu(Y) = \exp(\mu + 1/2)$ auf den Schätzer $\overline{\mu}(Y) = \log(Y) - 1/2$, während $\mathbb{E}_\mu(X) = \mu$ auch zum Schätzer $\widetilde{\mu} = X$ führt. Momentenschätzer bezüglich der selben Momentenfunktion sind also im Allgemeinen nicht transformationsinvariant. □

§4.2.2 **Definition**. Es sei $(\mathcal{X}, \mathscr{A}, \mathcal{P}_\Theta)$ ein bzgl. eines σ-endlichen Maßes μ dominiertes statistisches Modell mit Likelihood-Funktion $L(\theta, x) = [dP_\theta/d\mu](x)$ für $\theta \in \Theta$ und $x \in \mathcal{X}$. Eine Statistik $\widehat{\theta} : \mathcal{X} \to \Theta$ (Θ sei mit einer σ-Algebra $\mathscr{B}_\Theta$ versehen) wird *Maximum-Likelihood-Schätzer* (MLS) für θ genannt, falls $L(\widehat{\theta}(x), x) = \sup_{\theta\in\Theta} L(\theta, x)$ für μ-f.a. $x \in \mathcal{X}$ gilt. □

§4.2.3 **Bemerkung**. Der MLS 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 im Fall stetiger Lebesguedichten. Außerdem ist eine Abänderung auf einer Nullmenge bezüglich aller P_θ irrelevant, weil der Schätzer vor Realisierung des Experiments festgelegt wird und diese Realisierung damit fast sicher zum selben Schätzwert führen wird. □

§4.2.4 **Lemma**. *Es sei $\mathcal{P}_\Theta$ eine natürliche Exponentialfamilie in $T(x)$, dann ist der MLS $\widehat{\theta}$ implizit durch die Momentengleichung $\mathbb{E}_{\widehat{\theta}(x)}(T) = T(x)$ gegeben, vorausgesetzt der MLS existiert und liegt im Innern* $\operatorname{int}(\Theta)$ *von* Θ. □

Beweis von Lemma §4.2.4. in der Vorlesung. □

§4.2.5 **Bemerkung**. Liegt eine eineindeutige abgeleitete Parametrisierung $\theta \mapsto \gamma(\theta)$ vor, so ist natürlich $\widehat{\gamma} := \gamma(\widehat{\theta})$ der MLS für $\gamma(\theta)$. □

§4.2.6 **Beispiele**. (a) *(Fortsetzung von §3.3.8 (a))* Betrachte das *normale Lokations-Skalen-Modell* $X \circlearrowleft \{\mathfrak{N}(\mu\mathbb{1}_n, \sigma^2 \operatorname{Id}_n), \mu \in \mathbb{R}, \sigma > 0\}$. Dann ist der MLS für $\theta = (\mu/\sigma^2, 1/(2\sigma^2))^t$ durch die Momentengleichung $\mathbb{E}_{\widehat{\theta}(x)}[(\overline{X}, \overline{X}^2)^t] = (\overline{x}, \overline{x}^2)^t$ gegeben, also $\widehat{\mu} = \overline{X}$, $\widehat{\mu^2 + \sigma^2} = \overline{X}^2$. Mittels Reparametrisierung $(\mu, \mu^2 + \sigma^2) \mapsto (\mu, \sigma^2)$ erhalten wir $\widehat{\sigma}^2 = \overline{X}^2 - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$. Damit ist der MLS für σ^2 nicht erwartungstreu.

(b) Sei $(X_0, X_1, \ldots, X_n)$ eine Markovkette auf dem Zustandsraum $S = \{1, \ldots, M\}$ mit vom Parameter unabhängigen Anfangswert $X_0 = x_0$ und unbekannten Übergangswahrscheinlichkeiten $P_{X_{k+1}|X_k=i}(\{j\}) = p_{ij}$ ergibt sich die Likelihood-Funktion (bzgl. des Zählmaßes) durch

$$L\big((p_{kl}), X\big) = \prod_{i=1}^{n} p_{X_{i-2},X_i} = \prod_{k,l=1}^{M} p_{kl}^{N_{kl}(X)},$$

wobei $N_{kl}(X) = |\{i = 1, \ldots, n : X_{i-1} = k, X_i = l\}|$ die Anzahl der beobachteten Übergänge von Zustand k nach l angibt. Als MLS erhalten wir damit die relative Häufigkeit $\widehat{p}_{ij} = N_{ij}/\big(\sum_{l=1}^{m} N_{il}\big)$ der Übergänge.

(c) In einem parametrischen Regressionsmodell mit Beobachtungen

$$Y_i = g_\theta(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

ergibt sich unter der Normalverteilungsannahme $\varepsilon_i \sim \mathfrak{N}(0, \sigma^2)$ u.i.v. als MLS der Kleinste-Quadrate-Schätzer $\widehat{\theta} := \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} (Y_i - g_\theta(x_i))^2$. □

4.3 Minimum-Kontrast-Schätzer

§4.3.1 **Definition**. Für zwei Wahrscheinlichkeitsmaße P und Q auf demselben messbaren Raum $(\mathcal{X}, \mathscr{A})$ wird die Funktion

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

Kullback-Leibler-Divergenz (oder auch *Kullback-Leibler-Abstand*, *relative Entropie*) von P bezüglich Q genannt. □

§4.3.2 **Lemma**. *Für die Kullback-Leibler-Divergenz gilt*

(a) $KL(P|Q) \geqslant 0$ *und* $KL(P|Q) \geqslant 0$ *genau dann wenn* $P = Q$*, aber* KL *ist nicht symmetrisch;*

(b) für Produktmaße ist die Kullback-Leibler-Divergenz additiv:

$$KL(P_1 \otimes P_2|Q_1 \otimes Q_2) = KL(P_1|Q_1) + KL(P_2|Q_2);$$

(c) bildet $\mathcal{P}_\Theta$ *eine natürliche Exponentialfamilie und ist* θ_o *ein innerer Punkt von* Θ*, so gilt*

$$KL(P_{\theta_o}|P_\theta) = A(\theta) - A(\theta_o) + \langle \dot{A}(\theta_o), \theta_o - \theta\rangle.$$

Beweis von Lemma §4.3.2. Übung. □

§4.3.3 **Bemerkung**. Betrachte eine natürliche Exponentialfamilie $\mathcal{P}_\Theta$ in $T(x)$ (Fall §4.3.2 (c)), so gilt $\ddot{A}(\theta_o) = \mathbb{C}\mathrm{ov}_{\theta_o}(T)$ und für $\theta, \theta_o \in \mathrm{int}(\Theta)$ erhalten wir mit Hilfe einer Taylorentwicklung $KL(P_{\theta_o}|P_\theta) = \frac{1}{2}\langle \mathbb{C}\mathrm{ov}_{\theta_*}(T)(\theta - \theta_o), \theta - \theta_o\rangle$ für eine Zwischenstelle θ_* zwischen θ und θ_o. Zur Erinnerung, $\mathbb{C}\mathrm{ov}_{\theta_*}(T)$ gibt gerade die Fisher-Information in θ_* an. Im Fall einer mehrdimensionalen $\mathfrak{N}(\mu, \Sigma)$ mit strikt positiver Kovarianzmatrix folgt nun aus $A(\mu) = \langle \Sigma^{-1}\mu, \mu\rangle/2$, dass $\ddot{A}(\mu) = \Sigma^{-1}$ unabhängig von μ ist und somit $KL(\mathfrak{N}(\mu_o, \Sigma)|\mathfrak{N}(\mu, \Sigma)) = \frac{1}{2}\langle \Sigma^{-1}(\mu - \mu_o), \mu - \mu_o\rangle$ gilt. □

§4.3.4 **Definition**. Es sei $(\mathcal{X}_n, \mathscr{A}_n, \mathcal{P}_\Theta^n = \{P_\theta^n, \theta \in \Theta\})_{n\geqslant 1}$ eine Folge statistische Modelle über demselben Parameterraum Θ sowie $\gamma : \Theta \to \Gamma$ der interessierende Parameter. Eine Funktion $K : \Theta \times \Gamma \to \mathbb{R} \cup \{+\infty\}$ heißt *Kontrastfunktion*, falls für alle $\theta_o \in \Theta$ die Funktion $K(\theta_o, \bullet)$ ein eindeutiges Minimum in $\gamma_o := \gamma(\theta_o)$ besitzt . Eine Folge $K_n : \Gamma \times \mathcal{X}_n \to \mathbb{R} \cup \{+\infty\}$ heißt zugehöriger *Kontrastprozess* (oder kurz Kontrast), falls folgende Bedingungen gelten:

(a) Für alle $\gamma \in \Gamma$ ist $K_n(\gamma) := K_n(\gamma, \bullet)$ eine ZV, d.h. $\mathscr{A}_n$-messbar.

(b) Für alle $\gamma \in \Gamma$, $\theta \in \Theta$ gilt $K_n(\gamma) \xrightarrow{P_{\theta_o}^n} K(\theta_o, \gamma)$ für $n \to \infty$.

Zu einer Beobachtung $X^n \circlearrowleft \mathcal{P}_\Theta^n$ ist ein (nicht notwendigerweise eindeutiger) *Minimum-Kontrast-Schätzer* für $\gamma(\theta)$ (sofern existent) gegeben durch

$$\widehat{\gamma}_n := \widehat{\gamma}_n(X^n) := \arg\min_{\gamma\in\Gamma} K_n(\gamma, X^n).$$

□

§4.3.5 **Beispiele**. (a) Es sei $\mathcal{P}_\Theta$ eine bzgl. eines σ-endlichen Maßes μ dominierte Verteilungsfamilie mit Likelihood-Funktion L und interessierendem Parameter $\gamma(\theta) = \theta$ ($\Gamma = \Theta$). Des weiteren, sei $P_\theta \sim P_{\theta_o}$ ($P_\theta \ll P_{\theta_o}$ und $P_{\theta_o} \ll P_\theta$) für alle $\theta, \theta_o \in \Theta$. Im Produktexperiment $(\mathcal{X}^n, \mathscr{A}^{\otimes^n}, \mathcal{P}_\Theta^{\otimes^n})$ ist

$$K_n(\theta, x) = -\frac{1}{n}\sum_{i=1}^n \ell(\theta, x_i) \quad \text{für } x = (x_1, \ldots, x_n) \in \mathcal{X}^n$$

mit der Loglikelihood-Funktion $\ell(\theta, x) = \log(L(\theta, x))$ ein Kontrastprozess zur Kontrastfunktion

$$K(\theta_o, \theta) = KL(P_{\theta_o}|P_\theta) - KL(P_{\theta_o}|\mu).$$

Der zugehörige Minimum-Kontrast-Schätzer ist der MLS.

(b) *(Fortsetzung von §4.2.6(c))* Zusätzlich seien der interessierende Parameter $\gamma(\theta) = \theta$ ($\Gamma = \Theta$) identifizierbar, d.h. $\theta \neq \theta_o$ impliziert $g_\theta \neq g_{\theta_o}$, die Regressionsfunktionen $g_\theta : [0,1] \to \mathbb{R}$ stetig, das Design $x_i = i/n$ äquidistant und die Fehlerterme $\{\varepsilon_1, \ldots, \varepsilon_n\}$ u.i.v. mit $\mathbb{E}(\varepsilon_1) = 0$ und $\mathbb{E}(\varepsilon_1^4) < \infty$ (nicht notwendigerweise normalverteilt). Mit Hilfe der Tchebyscheff-Ungleichung und der Riemannschen Summen-Approximation zeigt man, dass $K_n(\theta, Y) := \frac{1}{n}\sum_{i=1}^n (Y_i - g_\theta(x_i))^2$ einen Kontrastprozess zur Kontrastfunktion $K(\theta_o, \theta) = \int_0^1 (g_\theta(x) - g_{\theta_o}(x))^2 ds + \mathbb{E}(\varepsilon^2)$ bildet. Der zugehörige Minimum-Kontrast-Schätzer ist der Kleinste-Quadrate-Schätzer.

(c) Betrachte das Regressionsmodell aus (b) im Fall einer Modellmisspezifikation, d.h. das Modell ist nicht adäquat für die Beobachtung $Y = (Y_1, \ldots, Y_n)$, in dem Sinne, dass die Beobachtungen dem Regressionsmodell $Y_i = f(i/n) + \varepsilon_i$, $i = 1, \ldots, n$, genügen, aber die Funktion $f : [0,1] \to \mathbb{R}$ nicht Element der Funktionenklasse $\{g_\theta, \theta \in \Theta\}$ ist. Identifiziert man die Regressionsfunktion mit dem interessierenden Parameter θ, d.h. $\Theta \subset \mathcal{L}^2([0,1])$, im Kleinste-Quadrate-Ansatz, so dass $\widehat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n (Y_i - g_\theta(x_i))^2$. Dann erhält man nach obiger Herleitung im Grenzwert eine „Kontrast-Typ-Funktion“ $K(f, \theta) = \int_0^1 (f(x) - \theta(x))^2 dx + \mathbb{E}(\varepsilon_1^2)$. Für $f \notin \Theta$ wird das Minimum natürlich nicht in f angenommen. Man kann die Kontrasttheorie durch Wahl der Funktion $\gamma(\cdot)$ jedoch trotzdem anwenden. Dazu nehmen wir an, dass die Menge interessierender Parameter Γ Riemann-integrierbare Funktionen enthält sowie abgeschlossen in $\mathcal{L}^2([0,1])$ und konvex ist, so dass für jede Funktion $\theta_o \in \mathcal{L}^2([0,1])$ eine

eindeutige L^2-Orthogonalprojektion $\gamma_o := \gamma(\theta_o)$ von θ_o auf Γ existiert. Beispielsweise kann Γ die Menge aller Polynome vom Grad $\leqslant d$ sein. Bezeichnet Θ die Menge der quadratisch Riemann-integrierbaren Funktionen in $\mathcal{L}^2([0,1])$, so ist $K_n(\gamma, Y) = \frac{1}{n}\sum_{i=1}^n (Y_i - \gamma(i/n))^2$ für $\gamma \in \Gamma$ ein Kontrastprozess zur Kontrastfunktion $K(\theta_o, \gamma) = \|\theta_o - \gamma\|_{\mathcal{L}^2}^2 + \mathbb{E}(\varepsilon_1^2)$, welche genau in $\gamma_o = \gamma(\theta_o)$ ihr Minimum in Γ annimmt. Unter geeigneten Bedingungen konvergiert der KQS $\widehat{\gamma}_n = \arg\min_{\gamma\in\Gamma} \frac{1}{n}\sum_{i=1}^n (Y_i - \gamma(i/n))^2$ unter $P_{\theta_o}^n$ gegen $\gamma_o = \gamma(\theta_o)$ (Übung). Im derart misspezifizierten Modell wird also die beste $\mathcal{L}^2$-Approximation γ_o an die wahre Funktion θ_o geschätzt, z.Bsp. die Beste Approximation durch ein Polynom vom Grad $\leqslant d$. □

§4.3.6 **Satz**. *Es gelten die Annahmen und Notationen aus Definition §4.3.4. Der Minimum-Kontrast-Schätzer $\widehat{\gamma}_n$ ist konsistent für $\gamma_o = \gamma(\theta_o)$, $\theta_o \in \Theta$, unter folgenden Bedingungen:*

(A1) *Γ ist ein kompakter Raum;*

(A2) *Die Funktion $K(\theta_o, \bullet)$ ist stetig und die zufällige Funktion $K_n(\bullet)$ ist $P_{\theta_o}^n$-f.s. stetig für alle $n \geqslant 1$;*

(A3) $\|K_n(\bullet) - K(\theta_o, \bullet)\|_\infty = \sup_{\gamma\in\Gamma} |K_n(\gamma) - K(\theta_o, \gamma)| \xrightarrow{P_{\theta_o}^n} 0$ *für* $n \to \infty$. □

Beweis von Satz §4.3.6. in der Vorlesung. □

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

§4.3.8 **Satz**. *Ist $\Gamma \subset \mathbb{R}^k$ kompakt, $\big(X_n(\bullet) := (X_n(\gamma), \gamma \in \Theta)\big)_{n\geqslant 1}$ eine Folge stetiger Prozesse mit $X_n(\gamma) \xrightarrow{P} X(\gamma)$ für alle $\gamma \in \Gamma$ und stetigem Grenzprozess $X(\bullet) := (X(\gamma), \gamma \in \Gamma)$, so gilt $\|X_n(\bullet) - X(\bullet)\|_\infty = \max_{\gamma\in\Gamma} |X_n(\gamma) - X(\gamma)| \xrightarrow{P} 0$ genau dann, wenn die Folge $(X_n(\bullet))_{n\geqslant 1}$* gleichgradig stetig *(in Wahrscheinlichkeit) ist, d.h. falls*

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

□

Beweis von Satz §4.3.6. in der Vorlesung Stochastik II bzw Übung. □

§4.3.9 **Definition**. Für ZV'en $\{X_n\}$ und positive Zahlen $\{a_n\}$ bezeichnen wir X_n/a_n als *stochastisch beschränkt* oder *straff*, falls $\lim_{K\to\infty} \sup_n P(|X_n| > Ka_n) = 0$ und schreiben kurz $X_n = O_P(a_n)$. Gilt $X_n/a_n \xrightarrow{P} 0$ so schreiben wir $X_n = o_P(a_n)$. □

§4.3.10 **Satz**. *Es gelten die Annahmen und Notationen aus Definition §4.3.4. Der Minimum-Kontrast-Schätzer sei konsistent für $\gamma_o := \gamma(\theta_o)$, z.Bsp. unter den Annahmen §4.3.6 (A1)-(A3), mit $\Gamma \subseteq \mathbb{R}^k$ und $\gamma_o \in \operatorname{int}(\Gamma)$. Der Kontrastprozess K_n sei zweimal stetig differenzierbar in einer Umgebung von γ_o ($P_{\theta_o}^n$-f.s.), so dass mit*

$$U_n(\gamma) := K_n(\gamma) \quad \textit{(Score)}, \qquad V_n(\gamma) := \ddot{K}_n(\gamma)$$

die folgenden Konvergenzen unter $P_{\theta_o}^n$ gelten:

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

(B2) *Gilt $G_n \xrightarrow{P_{\theta_o}^n} G_o$ für ZV'en $\{G_n\}$, so folgt $V_n(G_n) \xrightarrow{P_{\theta_o}^n} V(G_o)$ mit $V(G_o) \in \mathbb{R}^{k\times k}$ regulär, deterministisch.*

Dann gilt für den Minimum-Kontrast-Schätzer $\widehat{\gamma}_n$

$$\sqrt{n}\,(\widehat{\gamma}_n - \gamma_o) = -V(\gamma_o)^{-1}\sqrt{n}\,U_n(\gamma_o) + o_{P^n_{\theta_o}}(1).$$

Insbesondere ist $\widehat{\gamma}_n$ *unter* $P^n_{\theta_o}$ *asymptotisch normalverteilt:*

$$\sqrt{n}\,(\widehat{\gamma}_n - \gamma_o) \xrightarrow{\mathcal{L}} \mathfrak{N}\big(0, V(\gamma_o)^{-1}U(\gamma_o)V(\gamma_o)^{-1}\big).$$

□

Beweis von Satz §4.3.10. in der Vorlesung. □

§4.3.11 **Beispiel** (§1.1.5 (b) fortgesetzt). Im Lokations-Modell $Y \circlearrowleft \{\mathfrak{L}(\gamma\mathbb{1}_n, \mathrm{Id}_n), \mu \in \mathbb{R}\}$, d.h. $Y_i = \gamma + \varepsilon_i$, $i = 1, \ldots, n$, mit u.i.v. $\varepsilon_1, \ldots, \varepsilon$ betrachte den *M-Schätzer*

$$\widehat{\gamma}_n = \arg\min_{\gamma\in\Gamma} \sum_{i=1}^n \rho(Y_i - \gamma)$$

mit einer Funktion $\rho : \mathbb{R} \to [0, \infty)$, so dass die Funktion $x \mapsto \mathbb{E}[\rho(x + \varepsilon_1)]$ minimal (nur) bei $x = 0$ ist. Mit dem Kontrast $K_n(\gamma) := \frac{1}{n}\sum_{i=1}^n \rho(Y - i - \gamma)$ erhalten wir dann die Kontrastfunktion $K(\theta_o, \gamma) = \mathbb{E}[\rho(\gamma_o - \gamma + \varepsilon_1)]$, wobei wir $\theta_o = (\gamma_o, P_{\varepsilon_1})$ als unbekannten Parameter auffassen. Im Fall $\Gamma = \mathbb{R}$ und symmetrisch verteilten Fehlertermen $\{\varepsilon_i\}$, d.h. $\varepsilon_1 \overset{\mathcal{L}}{=} -\varepsilon_1$, führt als zugehörigen Minimum-Kontrast-Schätzer die Funktion $\rho(x) = \frac{1}{2}x^2$ auf das Stichprobenmittel $\widehat{\gamma}_n = \overline{Y}$ und für $\rho(x) = |x|$ auf den Stichprobenmedian $\widehat{\gamma}_n$. Ein Kompromiss zwischen beiden Schätzern ist der Huber-Schätzer für $\kappa > 0$

$$\widehat{\gamma}_n = \arg\min_{\gamma\in\Gamma} \sum_{i=1}^n \rho(Y_i - \gamma), \qquad \rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{falls } |x| \leqslant \kappa, \\ \kappa|x| - \frac{\kappa^2}{2}, & \text{falls } |x| > \kappa. \end{cases}$$

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

$$\sqrt{n}\,(\widehat{\gamma}_n - \gamma_o) \xrightarrow{\mathcal{L}} \mathfrak{N}\Big(0, \frac{\mathbb{E}[\rho'(\varepsilon_1)^2]}{\mathbb{E}[\rho''(\varepsilon_1)^2]}\Big).$$

Im Fall des Stichprobenmittels ist die asymptotische Varianz also gerade $\mathbb{E}(\varepsilon_1^2) = \mathbb{V}\mathrm{ar}(\varepsilon_1)$. einsetzen im Fall einer Dichte f_ε von ε_1 liefert heuristisch für den Stichprobenmedian die asymptotische Varianz $\mathbb{E}[\mathrm{sgn}(\varepsilon_1)^2]/\mathbb{E}[2\delta_0(\varepsilon_1)]^2 = (4f_\varepsilon(0))^{-1}$ sowie für den Huber-Schätzer $\mathbb{E}(\varepsilon^2 \wedge \kappa^2)/P(|\varepsilon_1| \leqslant \kappa)2$. (Übung.) □

§4.3.12 **Satz**. *Es sei* $\big((\mathcal{X}^n, \mathscr{A}^{\otimes^n}, \mathcal{P}_\Theta^{\otimes^n})\big)_{n\geqslant 1}$ *eine Folge* μ*-dominierter Produktexperimente mit eindimensionaler Loglikelihood-Funktion* $\ell(\theta, x) = \log([dP_\theta/d\mu](x))$. *Es gelte:*

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

(b) *Der Parameter* θ *ist identifiziert, d.h.* $\theta \neq \theta_o$ *impliziert* $P_\theta \neq P_{\theta_o}$.

(c) *Für alle* $x \in \mathcal{X}$ *ist* $\ell(\bullet, x)$ *stetig auf* Θ *und zweimal differenzierbar in einer Umgebung* U *von* θ_o.

(d) *Es gibt* $H_0 H_2 \in \mathcal{L}^1_{P_{\theta_o}}$ *und* $H_1 \in \mathcal{L}^2_{P_{\theta_o}}$ *mit* $\|\ell(\bullet, x)\|_\infty = \sup_{\theta\in\Theta} |\ell(\theta, x)| \leqslant H_0(x)$, $\|\dot{\ell}(\bullet, x)\|_\infty \leqslant H_1(x)$ *und* $\|\ddot{\ell}(\bullet, x)\|_\infty \leqslant H_2(x)$ *für alle* $x \in \mathcal{X}$.

(e) *Die Fisher-Informationsmatrix (zu einer Beobachtung)* $I(\theta_o) = \mathbb{E}_{\theta_o}\left[\dot{\ell}(\theta_o)\dot{\ell}(\theta_o)^t\right]$ *ist positiv definit.*

Dann erfüllt der MLS $\widehat{\theta}_n$

$$\sqrt{n}\,(\theta_n - \theta_o) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta_o)^{-1}\dot{\ell}(\theta_o) + o_{P_{\theta_o}}(1).$$

Insbesondere ist $\widehat{\theta}_n$ *unter* $P_{\theta_o}^{\otimes n}$ *asymptotisch normalverteilt mit Rate* $n^{-1/2}$ *und asymptotischer Kovarianzmatrix* $I(\theta_o)^{-1}$*:*

$$\sqrt{n}\,(\theta_n - \theta_o) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, I(\theta_o)^{-1}).$$

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

Beweis von Satz §4.3.12. in der Vorlesung. □

§4.3.13 **Bemerkung**. (a) Die Fisher-Information $I(\theta_o)$ gibt gerade sowohl die asymptotische Varianz der Score-Funktion als auch die lokale Krümmung der Kontrastfunktion $KL(P_{\theta_o}|\bullet)$ im Minimum θ_o an.

(b) Unter Regularitätsbedingungen gilt für die asymptotische Verteilung des MLS sowohl Unverzerrtheit als auch Cramér-Rao-Effizienz. Es ist aber weder klar noch im ALlgemeinen korrekt, dass die Momente ebenfals konvergieren und dass die Cramér-Rao-Schranke auch asymptotisch gilt.

(c) Oft ist Θ nicht kompakt, abaer man kann durch eine separate Untersuchung die Konsistenz von $\widehat{\theta}_n$ nachweisen. Dann gelten die Konvergenzresultate weiterhin.

(d) Die Regularitätsbedingungen lassen sich in natürlicher Weise abschwächen. Es reicht aus, dass $\mathcal{P}_\Theta$ in θ_o Hellinger-differenzierbar ist sowie dass die Loglikelihood-Funktion ℓ in einer Umgebung von θ_o Lipschitzstetig in θ ist mit Lipschitzkonstante in $\mathcal{L}^2_{P_{\theta_o}}$. Einen Beweis unter Verwendung von empirischer Prozesstheorie findet man z.Bsp. in van der Vaart [1998] Satz 5.39.

(e) Im Fall einer Modellmisspezifikation, in der die zu Grunde liegende Verteilung P_o nicht in $\mathcal{P}_\Theta$ enthalten ist (nicht aber die u.i.v. Annahme verletzt ist), konvergiert der MLS $\widehat{\theta}_n$ gegen $\theta^* := \arg\max_{\theta\in\Theta} \int_{\mathcal{X}} \ell(\theta, x)P_o(dx)$, sofern θ^* existiert und eindeutig ist. Es gilt entsprechend $\theta^* = \arg\min_{\theta\in\Theta} KL(P_o|P_\theta)$ sofern $P_o \ll P_{\theta^*}$. θ^* heißt Kullback-Leibler-Projektion von P_o auf $\mathcal{P}_\Theta$. Satz §4.3.10 liefert unter Regularitätsbedingungen diex asymptotische Normalität, $\sqrt{n}\,(\widehat{\theta}_n - \theta^*) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, V^{-1}UV^{-1})$ mit $U = \mathbb{E}_{P_o}[\dot{\ell}(\theta^*)\dot{\ell}(\theta^*)^t]$, $V = \mathbb{E}_{P_o}[\ddot{\ell}(\theta^*)]$. Im Allgemeinen wird dabei $U \neq V$ gelten. □

§4.3.14 **Beispiel**. Bei einer Exponentialfamilie mit natürlichem Parameterraum und natürlicher erschöpfender Statistik T erfüllt der MLS (sofern existent und in $\operatorname{int}(\Theta)$ enthalten) $\mathbb{E}_{\widehat{\theta}(x)}[T] = T(x)$ und die Fisher-Information $I(\theta) = \mathbb{V}\mathrm{ar}_\theta(T)$ (Kovarianzmatrix von T). Es gilt mit Regularitätsannahmen $\sqrt{n}\,(\widehat{\theta}_n - \theta_o)) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, \mathbb{C}\mathrm{ov}_{\theta_o}(T)^{-1})$ unter P_{θ_o}. Im Fall einer Bernoullikette $X_1, \ldots, X_n$ u.i.v. mit $X_i \circlearrowright \{\mathfrak{Bin}(1, \pi), \pi \in (0, 1)\}$ ist $\theta = \log(\pi/(1-\pi))$ der natürliche Parameter sowie $T(x) = x$. Aus $\sqrt{n}\,(\widehat{\theta}_n - \theta_o)) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, \pi(\theta_o)^{-1}(1-\pi(\theta_o))^{-1})$ folgt mittels der Δ-Methode für die π-Parametrisierung $\sqrt{n}\,(\widehat{\pi}_n - \pi_o)) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, \pi_o(1-\pi_o))$. Da $\widehat{\pi}_n = \overline{X}$ gilt, ist dies natürlich einfach direkt nachprüfbar. □

§4.3.15 **Definition**. Im Rahmen des vorigen Satzes heißt die zufällige Matrix

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

swddebeobachtete Fisher-Informationsmatrix. □

§4.3.16 **Korollar**. *Unter den Voraussetzungen des Satzes* §4.3.12 *gilt unter* P_{θ_o}

$$\sqrt{n} I(\widehat{\theta}_n)^{1/2}(\widehat{\theta}_n - \theta_o) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, \mathrm{Id}_k), \quad \mathcal{I}_n^{1/2}(\widehat{\theta}_n - \theta_o) \xrightarrow{\mathcal{L}} \mathfrak{N}(0, \mathrm{Id}_k).$$

Insbesondere sind für $k = 1$ *und das* $(1-\alpha/2)$*-Quantil* $z_{1-\alpha/2}$ *einer Standardnormalverteilung* $[\widehat{\theta}_n - n^{-1/2} I(\widehat{\theta}_n)^{-1/2} z_{1-\alpha/2}, \widehat{\theta}_n + n^{-1/2} I(\widehat{\theta}_n)^{-1/2} z_{1-\alpha/2}]$ *und* $[\widehat{\theta}_n - n^{-1/2}\mathcal{I}^{-1/2} z_{1-\alpha/2}, \widehat{\theta}_n + n^{-1/2}\mathcal{I}^{-1/2} z_{1-\alpha/2}]$ *Konfidenzintervalle für* θ_o *zum asymptotischen Vertrauensniveau* $1-\alpha$. □

Beweis von Korollar §4.3.16. in der Vorlesung. □

§4.3.17 **Beispiel**. (a) Bei natürlichen Exponentialfamilien ist die beobachtete Fisher-Information gerade $\ddot{A}(\widehat{\theta}_n) = I(\theta_n)$, so dass beide Ansätze, die Fisher-Information zu schätzen, auf dasselbe Verfahren führen.

(b) **?** gibt folgendes Beispiel, um die Frage bedingter Inferenz zu klären: es gibt zwei Instrumente, die Messwerte einer interessierenden Größe $\theta \in \mathbb{R}$ mit einem $\mathfrak{N}(0, \sigma_a^2)$-verteilten Fehler, $a = 0, 1$ und $\sigma_1 \neq \sigma_2$, liefern. In n Versuchen wird zunächst zufällig ein Instrument ausgewählt und dann ihr Messwert beobachtet. Es liegen somit u.i.v. ZV'en $\{(Y_i, a_i)\}_{i=1,\dots,n}$ mit $P(a_i = 0) = P(a_i = 1) = 1/2$ und $Y_i | a_i \sim \mathfrak{N}(\theta, \sigma_{a_i}^2)$ vor. Wir erhalten die Loglikelihood-Funktion

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

Der MLS für θ ist dann $\widehat{\theta}_n = (\sum_{=1}^{n} Y_i \sigma_{a_i}^{-2})/(\sum_{i=1}^{n} \sigma_{a_i}^{-2})$ und die Fisher-Information $I(\theta) = \frac{1}{2\sigma_0^2} + \frac{1}{2\sigma_1^2} =: I$ (unabhängig von θ). $\widehat{\theta}_n$ ist (nach vorgestellter Theorie oder mit direkten Argumenten) asymptotisch normalverteilt mit asymptotischer Varianz I^{-1}, was auf asymptotische Konfidenzintervalle der Form $\widehat{\theta}_n \pm \frac{1}{\sqrt{n}} I^{1/2} z_{1-\alpha/2}$ führt. Offensichtlich gilt $I(\widehat{\theta}_n) = I$, während für die beobachtete Fisher-Information $\mathcal{I}_n = \frac{\sum_{i=1}^{n} a_i}{n\sigma_0^2} + \frac{\sum_{i=1}^{n}(1-a_i)}{n\sigma_1^2}$ gilt. Damit ist $\mathcal{I}^{-1}$ gerade gleich der bedingten Varianz $\mathbb{V}\mathrm{ar}_{\theta|a}(n^{1/2}\widehat{\theta}_n)$. Da a beobachtet wird, ist die bedingte Varianz sicherlich ein sinnvolleres Maß für die Güte des Schätzers $\widehat{\theta}_n$. Im konkreten Beispiel der bedingten Normalverteilung können damit sogar einfach nicht-asymptotische bedingte Konfidenzintervalle angegeben werden. **?** bevorzugen aus diesem Grund auch für allgemeinere Modelle, in denen (approximativ) unwesentliche (ancillary) Statistiken vorkommen, die Normalisierung mit der beobachteten Fisher-Information gegenüber der plug-in-Schätzung $I(\widehat{\theta}_n)$. □