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

STATISTIK 1

Gliederung zur Vorlesung
im Wintersemester 2015/16

vorläufige Fassung stand 23. Oktober 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.

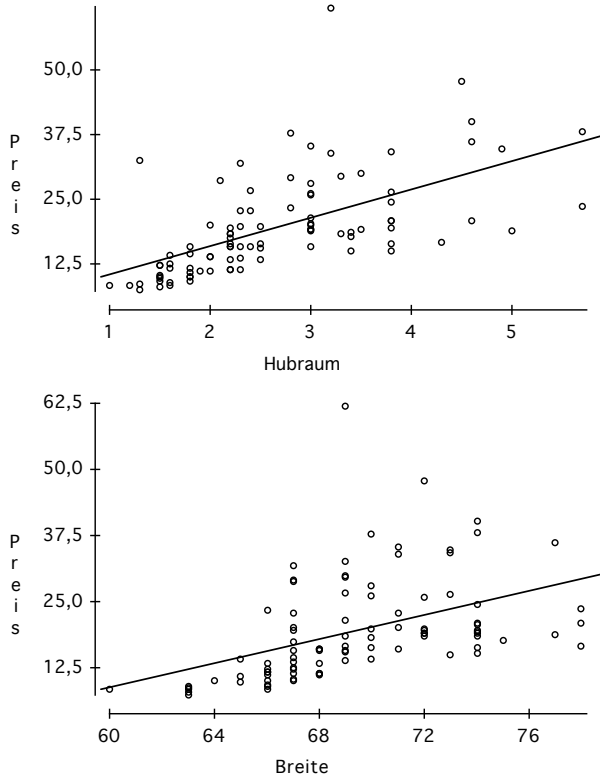
Preis	Zyl.	Hub.	Breite	Herkunft	Preis	Zyl.	Hub.	Breite	Herkunft
15.9	4	1.8	68	non-USA	10	4	1.5	64	non-USA
33.9	6	3.2	71	non-USA	13.9	4	2	69	non-USA
29.1	6	2.8	67	non-USA	47.9	8	4.5	72	non-USA
37.7	6	2.8	70	non-USA	28	6	3	70	non-USA
30	4	3.5	69	non-USA	35.2	6	3	71	non-USA
15.7	4	2.2	69	USA	34.3	6	3.8	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 $z_1, \dots, z_n \in \mathbb{R}$ beobachten wir Realisierungen der reellwertigen Zufallsvariablen (ZV'en)

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

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,

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

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 $Y = (Y_1, \dots, Y_n)^t$ und $\varepsilon := (\varepsilon_1, \dots, \varepsilon_n)^t$, den unbekannten Parametervektor $\beta = (a, b)^t \in \mathbb{R}^2$ sowie die vorgegebene (Design-)Matrix $X = (x_1, \dots, x_n)^t \in \mathbb{R}^{n\times 2}$ mit Zeilen $x_i^t = (1, z_i)$, $i = 1, \dots, n$, 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 $Y \mathbin{\tilde{\in}} \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$. 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 $Y \mathbin{\tilde{\in}} \{\mathfrak{L}(\mu\mathbb{1}_n, \sigma^2 \operatorname{Id}_n), \mu \in \mathbb{R}, \sigma > 0\}$. 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 $Y \mathbin{\tilde{\in}} \{\mathfrak{L}^{\otimes n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$. Wird die Varianz σ_o^2 der Beobachtungen als bekannt vorausgesetzt, so erfüllt der zufällige Vektor Y ein *Lokations-Modell* und wir schreiben abkürzend $Y \mathbin{\tilde{\in}} \{\mathfrak{L}(\mu\mathbb{1}_n, \sigma_o^2 \operatorname{Id}_n), \mu \in \mathbb{R}\}$ oder $Y \mathbin{\tilde{\in}} \{\mathfrak{L}^{\otimes n}(\mu, \sigma_o^2), \mu \in \mathbb{R}\}$. Wird dagegen der Erwartungswert μ_o als bekannt vorausgesetzt, so folgt der zufällige Vektor Y einem *Skalen-Modell* und wir schreiben abkürzend $Y \mathbin{\tilde{\in}} \{\mathfrak{L}(\mu_o\mathbb{1}_n, \sigma^2 \operatorname{Id}_n), \sigma > 0\}$ oder $Y \mathbin{\tilde{\in}} \{\mathfrak{L}^{\otimes n}(\mu_o, \sigma^2), \sigma > 0\}$. 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

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

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 $Y \curlywedge \big\{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\big\}$ 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 $Y \curlywedge \big\{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\big\}$ 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 $Y \curlywedge \big\{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\big\}$ 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 $i = 1, \dots, n$ 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* $Y \mathrel{\wr\!\!\!\sim} \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$ *gilt*

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

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

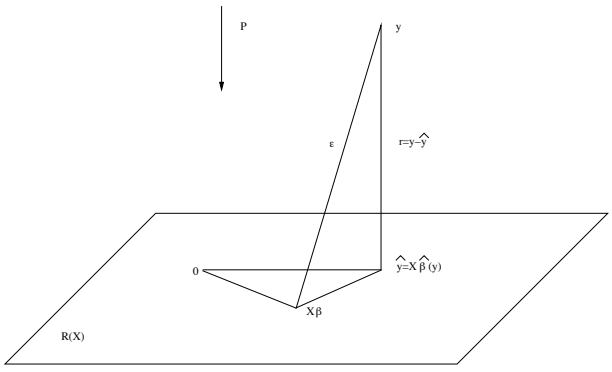
Beweis von Lemma §1.2.1. in der Vorlesung. □

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

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

verallgemeinerter Kleinste-Quadrate-Schätzer (vKQS oder GLSE für *generalized least squares estimator*) des unbekannten Parametervektors β. Im gewöhnlichen Fall ($\Sigma = \sigma^2 \mathrm{Id}_n$) 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* $\mathrm{rg}[X] = p$, *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* $\mathrm{rg}[X] = p$, *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 $Y \circledcirc \{\mathfrak{L}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma\}$ (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

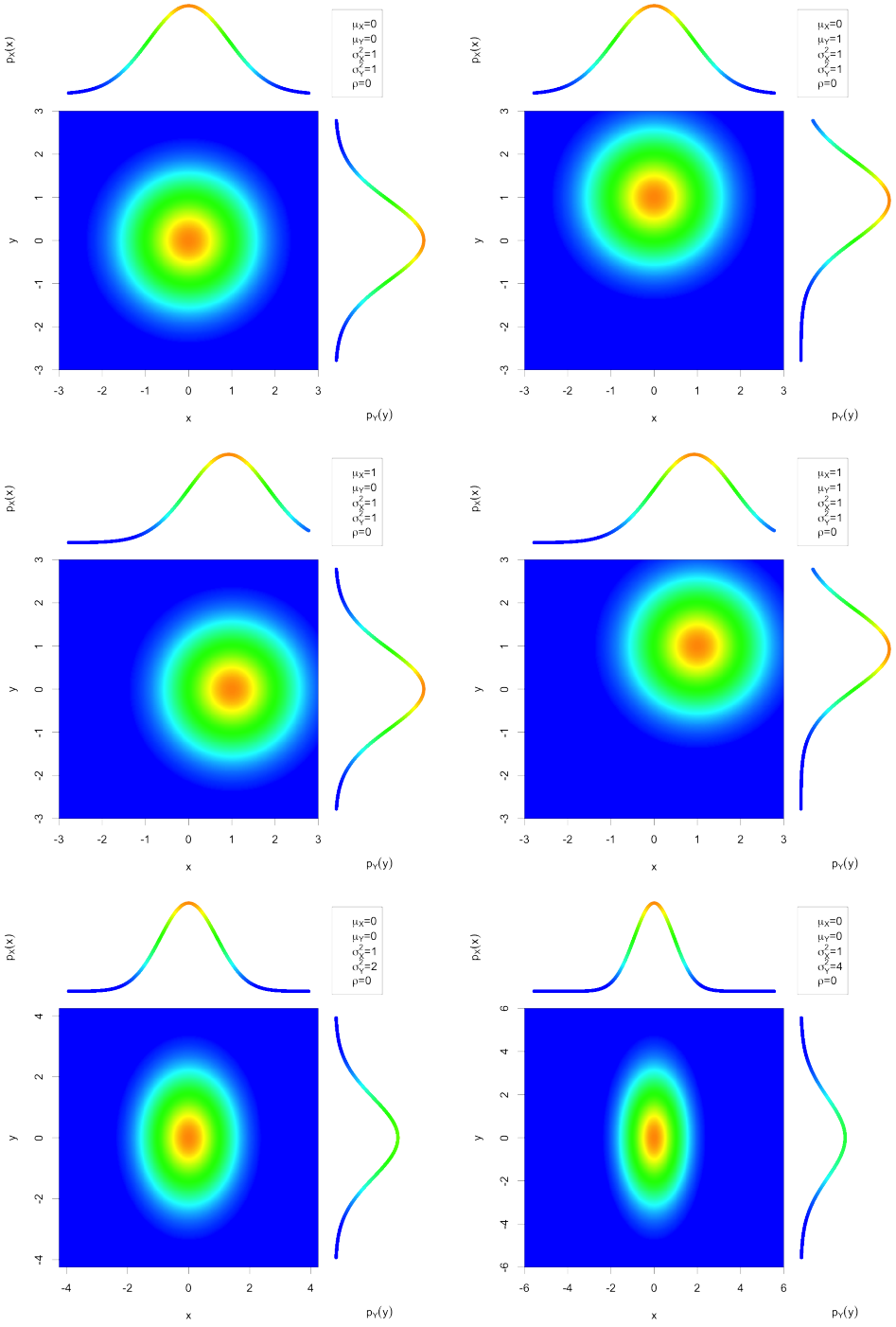
$$f(x) = \frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}} \exp\{-\frac{1}{2}\langle \Sigma^{-1}(x-\mu), (x-\mu)\rangle\}, \qquad x \in \mathbb{R}^p.$$ □

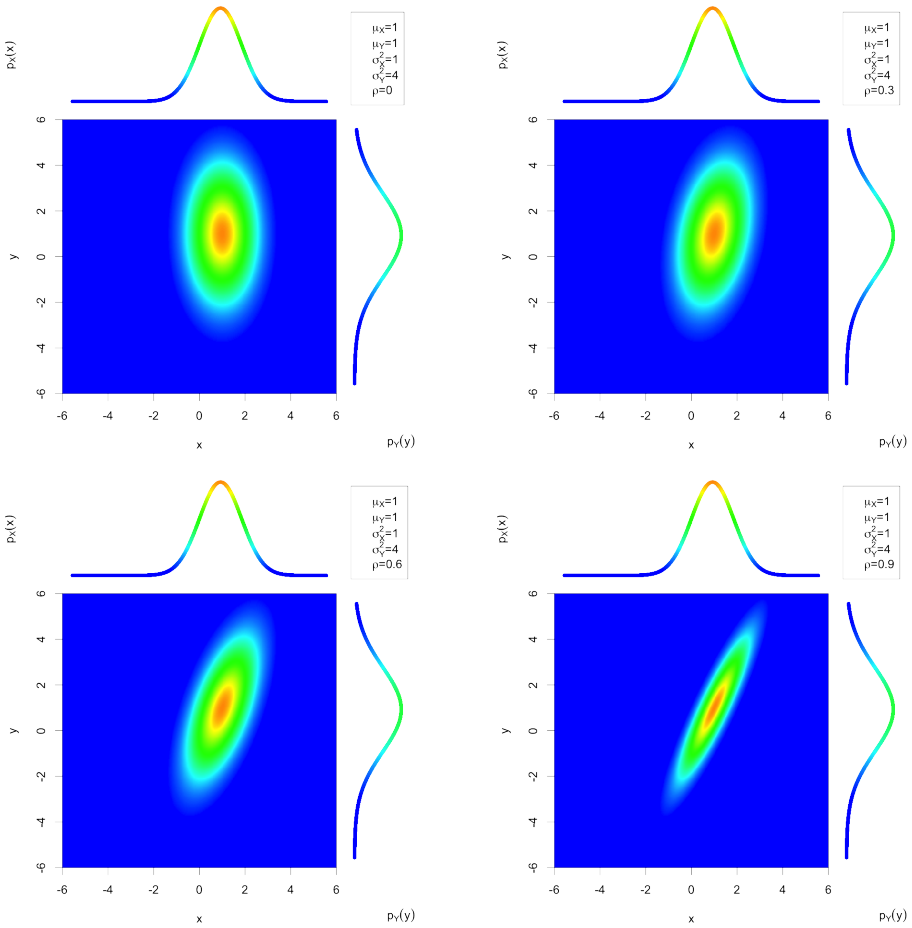
Beweis von Korollar §1.4.6. (Übung). □

§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

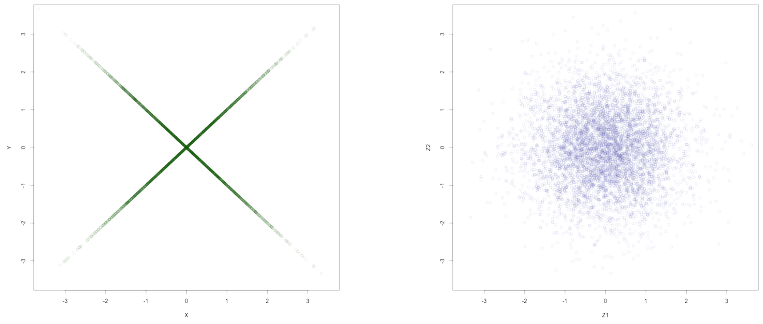
$$p(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left(-\frac{(x-\mu_X)^2}{2(1-\rho^2)\sigma_X^2}\right) \times \exp\left(\frac{2\rho(x-\mu_X)(y-\mu_Y)}{2(1-\rho^2)\sigma_X\sigma_Y}\right) \times \exp\left(-\frac{(y-\mu_Y)^2}{2(1-\rho^2)\sigma_Y^2}\right), \quad x, y \in \mathbb{R}.$$

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





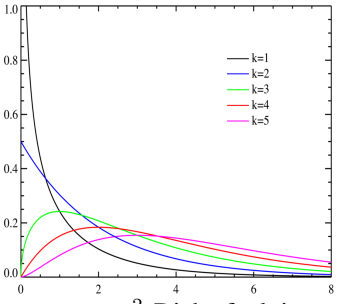
Besitzt (X, Y) eine *bivariate Normalverteilung* so gilt offensichtlich $X \sim \mathfrak{N}(\mu_x, \sigma_x^2)$ und $Y \sim \mathfrak{N}(\mu_y, \sigma_y^2)$. Sind X und Y weiterhin unkorreliert, d.h. $\rho = 0$, dann sind X und Y unabhängig und es gilt $aX + bY \sim \mathfrak{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$. 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

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



χ^2_k-Dichtefunktionen

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

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

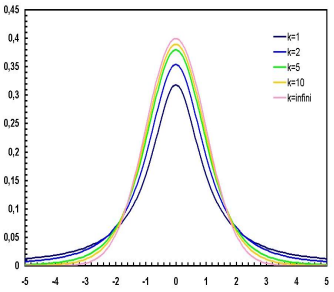
nichtzentrale χ^2_k-Verteilung mit k *Freiheitsgraden* und *Nichtzentralitätsparameter* δ^2. Wir schreiben $Q \sim \chi^2_k(\delta^2)$ sowie $\chi^2_{k,\alpha}(\delta^2) \in \mathbb{R}$ für das α-Quantil einer nichtzentralen χ^2-Verteilung mit k Freiheitsgraden und Nichtzentralitätsparameter δ^2, d.h.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

§1.4.9 **Korollar**. *Sei $Q \sim \chi^2_k$ und $W \sim \chi^2_k(\delta^2)$, 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) $\sigma^{-2}\|\Pi_{\mathcal{R}(A)} Z\|^2 \sim \chi^2_p$ und (ii) $\|Z/\sigma + v\|^2 \sim \chi^2_m(\|v\|^2)$.* □

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

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



$\mathfrak{t}_k$-Dichtefunktionen

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-t-Verteilung mit k-Freiheitsgraden, d.h. $P(T \leqslant \mathfrak{t}_{k,\alpha}) = \alpha$.

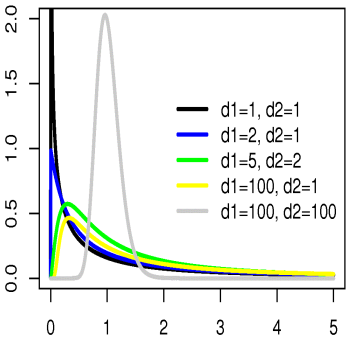
§1.4.11 **Bemerkung**. Die Student-$\mathfrak{t}$-Verteilung mit einem ($k = 1$) Freiheitsgrad entspricht gerade der Cauchy-Verteilung und für $k \to \infty$ konvergiert sie schwach gegen die Standardnormalverteilung (Slutsky-Lemma). Für jedes $k \in \mathbb{N}$ besitzt die $\mathfrak{t}_k$-Verteilung 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. $P(F \leqslant \mathfrak{F}_{m,k,\alpha}) = \alpha$.
Für $\delta \in \mathbb{R}$ heißt die Verteilung der ZV



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

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

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 $Y \mathbin{\wr} \{\mathfrak{N}(X\beta, \Sigma), \beta \in \mathbb{R}^p, \Sigma > 0\}$. 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 $Y \mathbin{\wr} \{\mathfrak{N}(X\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ 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* $Y \bigcirc \left\{\mathfrak{N}^{\otimes^n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\right\}$ 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* $Y \bigcirc \left\{\mathfrak{N}^{\otimes^n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\right\}$ 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. $Y_{(n)} \mathbin{\widetilde{\in}} \{\mathfrak{L}(X_{(n)}\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$.

§1.6.1 **Satz.** *Sei $Y_{(n)} \mathbin{\widetilde{\in}} \{\mathfrak{L}(X_{(n)}\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ 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 $Y_{(n)} \wr \{\mathfrak{L}^{\otimes^n}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$ 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 $T := \frac{\langle \widehat{\beta}, v\rangle - r}{\widehat{\sigma}|\langle (X^tX)^{-1}v, v\rangle|^{1/2}}$ 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 $Y \wr \{\mathfrak{L}^{\otimes^n}(\mu, \sigma^2), \mu \in \mathbb{R}\}$ 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 $Y \wr \{\mathfrak{L}(X\beta, \sigma^2 \mathrm{Id}_n)\}$ 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 $Y \bigcirc\!\!\!\!\sim \{\mathfrak{N}(X\beta, \sigma^2 \operatorname{Id}_n)\}$ mit der in einem Lokations-Skalen-Modell $Y \bigcirc\!\!\!\!\sim \left\{\mathfrak{N}^{\otimes^n}(\mu, \sigma^2)\right\}$. 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 $X \circledcirc \mathcal{P}_\Theta$, 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 $X \circledcirc \mathcal{P}_\Theta$ 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* $Y \circledcirc \{\mathfrak{L}(X\beta, \sigma^2 \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ 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 $Y \mathrel{\text{⫬}} \{\mathfrak{N}(X\beta, \sigma \operatorname{Id}_n), \beta \in \mathbb{R}^p, \sigma > 0\}$ 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 $X \mathrel{\text{⫬}} \{P_\pi := \mathfrak{Bin}(100, \pi), \pi \in (0, 1)\}$. 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* $Y \mathrel{\text{⫬}} \{\mathcal{N}(\mu \mathbb{1}_n, \operatorname{Id}_n), \mu \in \mathbb{R}\}$ 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

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

□

§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. Falls für jedes $x\in\mathcal{X}$, $D(x,\bullet)$ ein Punktmaß in $\delta(x)\in\mathcal{E}$ ist, d.h. $D(x,\{\delta(x)\})=1$, so dass die Abbildung $\delta:\mathcal{X}\to\mathcal{E}$ messbar ist. Dann ist δ eine Entscheidungsregel („nicht randomisierter" Schätzer) und

$$\mathfrak{R}_\nu(\theta,D) = \int_{\mathcal{X}}\int_{\mathcal{E}}\nu(\theta,\varepsilon)D(x,d\varepsilon)P_\theta(dx) = \int_{\mathcal{X}}\nu(\theta,\delta(x))P_\theta(dx) = \mathfrak{R}_\nu(\theta,\delta).$$

(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 δ zum Entscheidungsproblem $([0,1], \mathscr{B}_{[0,1]}, \nu)$ (randomisierter Test) definiert mit $D(x, \{1\}) := \delta(x)$ sowie $D(x, \{0\}) := 1 - \delta(x)$ 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 $\delta(x) := D(x, \{1\})$ zum Entscheidungsproblem $([0,1], \mathscr{B}_{[0,1]}, \nu)$. Dies bedeutet also, dass $\delta(x)$ die Wahrscheinlichkeit angibt, mit der bei Vorliegen der Beobachtung x die Hypothese H_0 abgelehnt wird. Offensichtlich, gilt dann $\mathfrak{R}_\nu(\theta, D) = \mathfrak{R}_\nu(\theta, \delta)$. □

§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

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

gilt, weiterhin wird $\mathfrak{R}^*_\nu$ Δ-*Minimaxrisiko* genannt. Wir bezeichnen δ 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 kurz *Bayesregel* oder *Bayes-optimal*. □

§2.2.3 **Bemerkung**. Während eine Minimaxregel den maximal zu erwartenden Verlust minimiert, kann das Bayesrisiko als ein (mittels $P_{\boldsymbol{\vartheta}}$) gewichtetes Mittel des zu erwartenden Verlustes angesehen werden. Alternativ wird $P_{\boldsymbol{\vartheta}}$ 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 (Ω, 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, $X \wr P_\Theta$ eine Beobachtung, $\boldsymbol{\vartheta} \sim P_{\boldsymbol{\vartheta}}$ 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

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

□

§2.2.8 **Satz**. *Es seien $(\mathcal{X}, \mathscr{A}, P_\Theta)$ ein statistisches Experiment, $X \wr P_\Theta$ 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 $\nu(\theta, \delta) = |\theta - \delta|$ 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 $P_{\vartheta}(\{1\}) =: \pi_1$. 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

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

□

§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 eine Δ-Bayes-optimale Entscheidungsregel, falls*

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

gilt, d.h. $\mathbb{E}_{\vartheta\,|X=x}[\nu(\vartheta, \delta(x))] \leqslant \mathbb{E}_{\vartheta\,|X=x}[\nu(\vartheta, \delta'(x))]$ 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**. *Seien $\Theta \subset \mathbb{R}$ und $\Delta = \mathbb{R}$. Unter den Annahmen des Satzes §2.2.10 gelten die folgenden Aussagen:*

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

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

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

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

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. Betrachten wir das statistische Entscheidungsproblem $(\mathcal{E},\mathscr{E},\nu)$, so gelten für jede Entscheidungsregel $\delta_o \in \Delta$ 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 auf Δ, so ist δ_o minimax-optimal in Δ.

(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 $\delta \in \Delta$ 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 $X_1,\dots,X_n$ unabhängig und identisch $\mathcal{N}(\mu,\mathrm{Id}_d)$-verteilte ZV'en mit unbekanntem $\mu\in\mathbb{R}^d$. Bezüglich der quadratischen Verlustfunktion $\nu(\mu,\delta) = \|\mu-\delta\|^2$ 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 $X_1,\dots,X_n$ unabhängig und identisch $\mathcal{N}(\mu,1)$-verteilte ZV'en mit unbekanntem $\mu\in\mathbb{R}$. Bezüglich der quadratischen Verlustfunktion $\nu(\mu,\delta) = (\mu-\delta)^2$ 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 $X \bigcirc\!\!\!\!\sim \{\mathfrak{Bin}(n, \pi), \pi \in (0, 1)\}$ mit $n \geqslant 1$. Wir bestimmen einen minimax-optimalen Schätzer für π bezüglich der quadratischer Verlustfunktion $\nu(\pi, \delta) = (\delta - \pi)^2$ 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 $\mathbb{E}_{\boldsymbol{\mu}_\sigma|X} = \frac{n}{n+\sigma^{-2}}\overline{X}$ (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 X_i 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 $Y \curvearrowright \{\mathfrak{N}(\mu, \sigma^2 \operatorname{Id}_d), \mu \in \mathbb{R}^d, \sigma > 0\}$*

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

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

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

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). □