Mathematische Statistik

Hajo Holzmann
Fachbereich Mathematik und Informatik, Universität Marburg

Sommersemester 13
(Stand: 1. Juli 2013)

Inhaltsverzeichnis

1 Einführung: Statistik im linearen Modell

1.1 Das lineare Modell

1.1.1 Regression und das lineare Modell

In der Regressionsanalyse geht es darum, den Einfluss einer Reihe von erklärenden Variablen $x_1, \ldots, x_r$, sogenannte *Kovariablen*, auf eine abhängige Variabe Y, die *Zielvariable*, zu modellieren bzw. zu schätzen. Dieser Zusammenhang drückt sich in Form einer Funktion $y = f(x_1, \ldots, x_r)$ aus. Nun wird aber nicht angenommen, dass diese Beziehung exakt gilt. Vielmehr ist sie durch zufällige Störgrößen ϵ überlagert, d.h. es gilt

$$Y = f(x_1, \ldots, x_r) + \epsilon.$$

In der linearen Regressionsanalyse nimmt man an, dass der Einfluss der Kovariablen, zumindest nach geeignter Transformation dieser Variablen, in einer linearen From

$$Y = b_0 + b_1 x_1 + \ldots + b_r x_r + \epsilon.$$

Dabei ist ϵ eine Zufallsvariable (bzw. deren Realisierung) mit Erwartungswert $E\epsilon = 0$ und endlicher Varianz $\operatorname{Var} \epsilon = \sigma^2$, und somit ist auch die Zielgröße Y eine Zufallsvariable (bzw. deren Realisierung). Ziel ist dann die Schätzung der Parameter $b_0, \ldots, b_r$. Diese fasst man in einem Vektor zusammen. Wir schreiben $\boldsymbol{\beta} = (b_0, \ldots, b_r) \in \mathbb{R}^p$, also $p = r + 1$, und für die Komponenten von $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ gilt dann $\beta_i = b_{i-1}$. Weiter schreibt man $\mathbf{x} = (1, x_1, \ldots, x_r)^T \in \mathbb{R}^p$, so dass $Y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$.

Ziel ist nun in einem ersten Schritt die Schätzung der unbekannten Parameter des Modells, insbesondere von $\boldsymbol{\beta}$. Dazu nimmt man an, es werden Daten $(Y_i, x_{i,1}, \ldots, x_{i,r})$, $i = 1, \ldots, n$, beobachtet, so dass

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad \mathbf{x}_i = (1, x_{i,1}, \ldots, x_{i,r})^T.$$

Für die Fehler ϵ_i nimmt man dabei an, dass diese unabhängig oder zumindest unkorreliert sind, also dass $\operatorname{Cov}(\epsilon_i, \epsilon_j) =$, $i \neq j$. Falls darüber hinaus die Varianzen $\sigma_i^2 = \operatorname{Var} \epsilon_i$ alle gleich sind, also $\sigma_1^2 = \ldots = \sigma_n^2$, so sprich man von einer *homoskedastsichen* Fehlerstruktur, ansonsten von einer *heteroskedastischen* Fehlerstruktur.

Die Analyse eines homoskedastischen linearen Regressionsmodells findet nun im Rahmen der Theorie linearer Modelle statt. Dazu schreiben wir das Modell in Vektor- und Matrixform wie folgt.

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n, \qquad X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}, \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n.$$

Es gilt dann $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Für die Kovariablen nimmt man noch an, dass sie dergestalt sind, dass die sogenannte *Designmatrix* X vollen Rang p hat. Der Achsenabschnitt wird meistens, aber nicht immer in das lineare Regressionsmodell mit aufgenommen. Das lineare Regressionsmodell fällt unter die folgende allgemeinere Definition.

Definition 1.1
Das Modell

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

heißt *lineares Modell*, falls $\boldsymbol{\beta} \in \mathbb{R}^p$ ein (konstanter, unbekannter) Parametervektor, $X \in \mathbb{R}^{n\times p}$ eine bekannte Matrix (Designmatrix), $\mathbf{Y}$ ein beobachter Zufallsvektor (Zielgrößen) und $\boldsymbol{\epsilon}$ ein nichtbeobachteter Zufallsvektor (Störgrößen) mit $E\boldsymbol{\epsilon} = \mathbf{0}$ und $\operatorname{Cov}\boldsymbol{\epsilon} = \sigma^2 I_n$ sind. Sind darüber hinaus die Fehler normalverteilt, also $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, dann spricht man von einem linearen Modell mit normalverteilten Fehlern.

1.1.2 Modellierung des Einflusses der Kovariablen

a. Einfluss einer metrischen Kovariable

Angenommen, neben der Zielvariable Y wird noch eine metrische Kovariable x beobachtet. Das einfachste Modell ist nun die direkt lineare Regression von Y auf x, die *einfache lineare Regression*

$$Y = b_0 + b_1 x + \varepsilon,$$

also $\mathbf{x}^T = (1, x)^T$ und $\boldsymbol{\beta}^T = (b_0, b_1)^T$. Manchmal liegt ein linearer Zusammenhang auch erst vor, nachdem x geeignet trasnformiert wurde, etwa zu $\tilde{x} = \log(x)$ (falls $x > 0$) und dann $Y = b_0 + b_1\tilde{x} + \varepsilon$, also $\mathbf{x}^T = (1, \tilde{x})^T$.

Man kann aus einer metrischen Kovariable auch mehrere metrische Kovariablen machen durch Anwendung linear unabhängiger Funktionen $f_1, \ldots, f_r$, also

$$Y = b_0 + b_1 f_1(x) + \ldots + b_r f_r(x) + \varepsilon,$$

also $\mathbf{x} = \big(1, f_1(x), \ldots, f_r(x)\big)^T$. Beispiele sind die polynomiale Regression, bei der $f_i(x) = x^i$ gewählt wird, also

$$Y = b_0 + b_1 x + \ldots + b_r x^r + \epsilon,$$

und $\mathbf{x} = \big(1, x, \ldots, x^r\big)^T$, oder auch für $x \in [0, 1]$ die trigonometrische Regression, bei der $f_{2j-1}(x) = \sin(2j\pi x)$ und $f_{2j}(x) = \cos(2j\pi x)$, $j = 1, \ldots, q$ gewählt wird, also

$$Y = b_0 + \sum_{j=1}^{q} (b_{2j-1}\sin(2j\pi x) + b_{2j}\cos(2j\pi x)),$$

und $\mathbf{x} = \big(1, \sin(2\pi x), \cos(2\pi x)), \sin(2q\pi x), \cos(2q\pi x))\big)^T$, $\boldsymbol{\beta} = (b_0, b_1, b_2, \ldots, b_{2q})^T$, also $p = 2q + 1$.

b. Einfluss einer kategoriellen Kovariable

Bei *kategoriellen Kovariablen* unterscheidet man *nominale* Kovariablen, bei denen die Kategorien nicht geordnet sind (etwa Autotypen), und *ordinale Kovariable*, bei denen die Kategorien in einer natürlichen Reihenfolge vorliegen (etwa Schulnoten).

Wir betrachten zunächst die Modellierung einer nominale Kovariablen mit den Kategorien $i = 1, \ldots, I$. Um nicht zu viele Parameter ins Modell aufzunehmen, damit also die Desigmatrix X noch vollen Rang hat, wählt man eine Referenzkategorie, z.B. $i = 1$, und für künstliche Kovariablen ein, deren Koeffizient den Unterschied zwischen der betrachteten Kategorie

$i = 2, \ldots, I$ und der Referenzkategorie beschreibt. Hier sind insbesondere zwei Kodierungen üblich.

Kodierung durch Dummy Variablen Wird die Kategorie x beobachtet und ist 1 die Referenzkategorie, so setze $\mathbf{x} = (1, 1_{x=2}, \ldots, 1_{x=I})^T \in \mathbb{R}^I$, d.h. falls eine der Kategorien $i = 2, \ldots, I$ vorliegt, kommt eine 1 hinzu, ansonsten gibt es nur den Achsenabschnitt. Im Koeffizientenvektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I)^T$ beschreibt β_i den Unterschied des Einflusses von Kategorie $i \geq 2$ gegenüber der Referenzkategorie, und $\beta_1 + \beta_i$ den Gesamteinfluss von Kategorie $i \geq 2$.

Effektkodierung Wird die Kategorie x beobachtet und ist 1 die Referenzkategorie, so setze $\mathbf{x} = (1, 1_{x=2} - 1_{x=1}, \ldots, 1_{x=I} - 1_{x=1})^T \in \mathbb{R}^I$.

Die Software `R` verwendet standardmäßig die Dummy Kodierung.

Handelt es sich bei x um eine ordinale Kovariable, so kann man versuchen, den geordeneten Kategorien konkrete Zahlen (etwa den Schulnoten die Zahlen $1 - 6$) zuzuordnen, und diese dann wie eine metrische Kovariable zu benutzten. Dies hat den Vorteil, dass in dem Modell wesentlich weniger Parameter (nur ein Parameter β für Kovariable x statt $I - 1$ Parameter) verwendet werden müssen. Dabei müssen die zugeordneten Zahlen (insbesondere das Verhältnis von deren Abständen) aber sogfältig gewählt werden. Falls dies nicht adäquat möglich ist, sollte die Kovariable lieber wie eine nominale Kovariable und mit der Dummay Kodierung behandelt werden.

relevante R Befehle kategorielle Kovariablen müssen bei der Funktion `lm`, die lineare Regression mit kleinsten Quadraten anpasst, als Faktor vorliegen. Dazu kann man den Typ mit `str` erfahren, und gegebenenfall mit `as.factor` zu einem Faktor umwandeln.

c. Interaktionen

Interaktionen zwischen einer kategoriellen und einer stetigen Kovariable

Ist x eine kategorielle (nominale) Kovariable mit den Kategorien $i = 1, \ldots, I$ und t eine stetige Kovariable, die direkt (linear) in die Zielgröße eingeht, so können die Kategorienausprägungen von x auch den Koeffizienten von t beeinflussen. Dies nennt man Interaktionen, man setzt dann bei Referenzkategorie 1 und Dummykodierung von x

$$\mathbf{x} = \big(1, 1_{x=2}, \ldots, 1_{x=I}, t, 1_{x=2}t, \ldots, 1_{x=I}t\big)^T.$$

Im Koeffizientenvektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I, \beta_{I+1}, \ldots, \beta_{2I})^T$ beschreibt dann $\beta_{I+1} + \beta_{2I}$ die Steigung von t bei Vorliegen von Kategorie $i \geq 2$, und β_{I+1} die Steigung bei Vorliegen der Referenzkategorie 1. Man muss dabei natürlich nicht alle Interaktionen in das Modell aufnehmen.

Falls sowohl stetige als auch kategorielle Kovariable auftreten, spricht man manchmal statt von der Regressionsanalyse auch von der *Kovarianzanalyse*.

Interaktionen zwischen zwei kategoriellen Kovariablen

Ist x eine kategorielle (nominale) Kovariable mit den Kategorien $i = 1, \ldots, I$ und t eine kategorielle Kovariable mit Kategorien $j = 1, \ldots, J$, so kann man Interaktionen für gemeinsames Vorliegen von $x = i$ und $t = j$ modellieren. Sind $i = 1$ und $j = 1$ die Referenzkategorien, so bildet man in Dummy Kodierung

$$\mathbf{x} = \big(1, 1_{x=2}, \ldots, 1_{x=I}, 1_{t=2}, \ldots, 1_{t=J}, 1_{x=2}1_{t=2}, \ldots, 1_{x=2}1_{t=J}, \ldots, 1_{x=I}1_{t=J}\big)^T \in \mathbb{R}^{IJ}.$$

Die Terme $1_{x=i}1_{t=j}$, $i = 2, \ldots, I$, $j = 2, \ldots, J$, entsprechen dann den Interaktionen, diese sind wieder als Abweichungen gegenüber den Haupteffekten $1_{x=i}$ und $1_{t=j}$ zu interpretieren.

Interaktionen zwischen zwei metrischen Kovariablen

Interaktionen zwischen zwei metrischen Kovriablen x und t müssen durch Aufnahme bestimmter gemeinsamer nichtlinearer Funktionen, etwa xt oder $e^x e^t$, modelliert werden. Man benutzt häufig gemeinsame Polynome niedrigen Grades.

In welcher Form metrische Kovariablen aufgenommen werden, und welche Interaktionen mit kategoriellen oder anderen metrsichen Kovariablen aufgenommen werden, muss innerhalb der *Modellwahl* und der *Modelldiagnostik* bestimmt werden. Wie nehmen zunächst an, dass ein linearen Modell der Form (1) in seiner korrekten Form gegeben ist.

1.2 Kleinste Quadrate Schätzung

1.2.1 Methode der kleinsten Quadrate

Der bekannteste Schätzer von $\boldsymbol{\beta}$ im linearen Modell (1) ergibt sich aus der Methode der kleinsten Quadrate. Man wählt dabei $\boldsymbol{\beta}$ derart, dass $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 = \sum_{i=1}^n (Y_i - x_i^T\boldsymbol{\beta})^2$ minimal wird, also

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{LS} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2$$

Dabei stehet LS für least squares = kleinste Quadrate. Wir schreiben für die Komponenten von $\hat{\boldsymbol{\beta}}_{LS}$ explizit $\hat{\boldsymbol{\beta}}_{LS} = (\hat{\beta}_{1,LS}, \ldots, \hat{\beta}_{p,LS})$. Wir wollen $\hat{\boldsymbol{\beta}}_{LS}$ in expliziter Form auf zwei Arten herleiten.

Normalengleichungen. Ableiten von $\|Y - X\boldsymbol{\beta}\|^2$ und gleich $\mathbf{0}$ setzten liefert (nach Division durch -2)

$$X^T(\mathbf{Y} - X\boldsymbol{\beta}) = \mathbf{0}.$$

Dies nennt man auch die Normalengleichungen, diese sind eine notwendige Bedingung für ein lokales Extremum. Da X vollen Rang p hat, ist $X^TX \in \mathbb{R}^{p\times p}$ invertierbar und man erhält

$$\hat{\boldsymbol{\beta}}_{LS} = \hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}. \tag{2}$$

Dass $\hat{\boldsymbol{\beta}}_{LS}$ das einzige lokale und somit globale Minimum von $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ ist, sieht man leicht daran, dass die Hessische Matrix (Matrix der zweiten Ableitungen) gleich $2\,X^TX$ und somit positiv definit ist.

Geometrische Herleitung: Ein $\hat{\boldsymbol{\beta}}$ minimiert die Funktion $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ genau dann, wenn $X\hat{\boldsymbol{\beta}}$ die orthogonale Projektion von $\mathbf{Y}$ auf den von den Spaltenvektoren von $X = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$, $\mathbf{v}_i \in \mathbb{R}^n$, erzeugten Unterraum $V = \text{span}\{v_1, \ldots, v_p\}$ im $\mathbb{R}^n$ ist. In der Tat: Für jedes andere $\boldsymbol{\beta}$ gilt nach Pythagoras:

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 = \|\underbrace{\mathbf{Y} - X\hat{\boldsymbol{\beta}}}_{\perp\ \mathbf{v}_1,\ldots,\mathbf{v}_p} + X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 = \|\mathbf{Y} - X\hat{\boldsymbol{\beta}}\|^2 + \|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 \geq \|\mathbf{Y} - X\hat{\boldsymbol{\beta}}\|^2$$

Da X vollen Rang hat, sind $\mathbf{v}_1, \ldots \mathbf{v}_p$ linear unabhängig und somit ist der Koeffizientenvektor $\hat{\boldsymbol{\beta}}_{LS}$ eindeutig bestimmt.
Um den Schätzer $\hat{\boldsymbol{\beta}}_{LS}$ in der expliziten Form (2) zu erhalten, betrachten wir die Matrix $P_X = X(X^TX)^{-1}X^T \in \mathbb{R}^{n\times n}$ (die sogenannte *hat matrix*). Es ist

$$P_X :\ \mathbb{R}^n \to V \qquad \mathbf{z} \mapsto P_X\mathbf{z}$$

die orthogonale Projektion auf V. Dazu zeigt man durch direkte Rechnung:
a. P_X ist orthogonale Projektion: $P_X^2 = P_X$, $P_X^T = P_X$
b. $P_X\mathbb{R}^n = V$
Somit muss gelten:

$$X\hat{\boldsymbol{\beta}} = P_X\mathbf{Y} = X(X^TX)^{-1}X^T\mathbf{Y}.$$

Da X vollen Rang hat, ergibt sich wieder die Form (2).

Bemerkung
Erwartungswert und Varianz von $\mathbf{Y}$ im linearen Modell (1) hängen von den unbekannten Parametern $(\boldsymbol{\beta}, \sigma^2)$, ab, höhere Momente sogar von der unbekannten Verteilung der Störungen $\boldsymbol{\epsilon}$. Daher müsste man diese bei Bildung von Erwartungswert und Varianz mitschreiben, also etwa $E_{\boldsymbol{\beta},\sigma^2}(\cdot)$ und $\mathrm{Cov}_{\boldsymbol{\beta},\sigma^2}(\cdot)$. Wir werden diese Parameter aber in der Notation in diesem Abschnitt unterdrücken, und einfach E und Cov schreiben.

Satz 1.2
Der kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}_{LS}$ im linearen Modell (1) ist unverfälscht, also $E\hat{\boldsymbol{\beta}}_{LS} = \boldsymbol{\beta}$, und es ist

$$\mathrm{Cov}\,\hat{\boldsymbol{\beta}}_{LS} = \sigma^2 (X^T X)^{-1}.$$

Beweis
Mit Satz 1.12 folgt

$$\begin{aligned}
E\hat{\boldsymbol{\beta}}_{LS} &= E(X^T X)^{-1} X^T Y \\
&= E(X^T X)^{-1} X^T (X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= E(X^T X)^{-1} X^T X\boldsymbol{\beta} + E(X^T X)^{-1} X^T \boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + (X^T X)^{-1} X^T \underbrace{E\boldsymbol{\epsilon}}_{=\mathbf{0}}, \\
&= \boldsymbol{\beta} \\
\mathrm{Cov}\,\hat{\boldsymbol{\beta}}_{LS} &= \mathrm{Cov}((X^T X)^{-1} X^T (\underbrace{X\boldsymbol{\beta}}_{konst.} + \boldsymbol{\epsilon})) \\
&= \mathrm{Cov}((X^T X)^{-1} X^T \boldsymbol{\epsilon}) \\
&= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

Man nennt $\sigma\big((X^T X)^{-1}_{ii}\big)^{1/2}$ und für einen Schätzer $\hat{\sigma}^2$ von σ^2 (s.u.) auch $\hat{\sigma}\big((X^T X)^{-1}_{ii}\big)^{1/2}$ den *Standardfehler* von $\hat{\beta}_{i,LS}$.

Im Folgenden zeigen wir, dass $\hat{\boldsymbol{\beta}}_{LS}$ der eindeutig bestimmte, lineare unverfälschte Schätzer mit der kleinsten Varianz ist.

Satz 1.3 (Gauß-Markov-Aitken)

a. Sei $S(\mathbf{Y}) = A\mathbf{Y}$, $A \in \mathbb{R}^{p\times n}$, ein linearer, unverfälschter Schätzer für $\boldsymbol{\beta}$ im linearen Modell (1) (d.h. $ES(\mathbf{Y}) = \boldsymbol{\beta}\ \forall \boldsymbol{\beta} \in \mathbb{R}^p$). Dann gilt

$$\mathrm{Cov}(S(\mathbf{Y})) \geq \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{LS})$$

im Sinne, dass die Differenz $\big(\mathrm{Cov}(S(\mathbf{Y})) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{LS})\big)$ positiv semidefinit ist.

b. Ist $A \neq (X^T X)^{-1} X^T$, so existiert $\mathbf{z} = \mathbf{z}(A) \in \mathbb{R}^p$, so dass

$$\mathbf{z}^T(\mathrm{Cov}(S(\mathbf{Y})) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{LS}))\mathbf{z} > 0$$

Beweis
a. Aus der Unverfälschtheit folgt

$$ES(\mathbf{Y}) = AX\boldsymbol{\beta} \stackrel{!}{=} \boldsymbol{\beta} \qquad \forall \boldsymbol{\beta} \in \mathbb{R}^p,$$

also $AX = I_p$. Damit und mit Satz 1.2 ist

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{LS}) &= \sigma^2 (X^T X)^{-1} = \sigma^2 AX(X^T X)^{-1} X^T A^T = \sigma^2 A P_X A^T, \\ \mathrm{Cov}(S(\mathbf{Y})) &= A\sigma^2 I_n A^T = \sigma^2 A A^T. \end{aligned}$$

Damit erhält man

$$\mathrm{Cov}(S(\mathbf{Y})) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2 A(I_n - P_X)A^T$$

Die Matrix $(I_n - P_X)$ ist idempotent und symmetrisch: $(I_n - P_X)^2 = (I_n - P_X) = (I_n - P_X)^T$. Somit:

$$\sigma^2 \mathbf{z}^T A(I_n - P_X)A^T \mathbf{z} = \sigma^2 \|(I_n - P_X)A^T \mathbf{z}\|^2 \geq 0.$$

b. Angenommen, $(I_n - P_X)A^T \mathbf{z}\|^2 = 0 \ \forall \mathbf{z} \in \mathbb{R}^p$, und somit $(I_n - P_X)A^T = 0$. Sei $A^T = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$, $\mathbf{a}_i \in \mathbb{R}^n$, dann erhält man [1]: $\mathbf{a}_i \in \mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_p)$, also $A^T = XM$ für eine Matrix $M \in \mathbb{R}^{p \times p}$. Wegen $AX = I_p$ folgt $M^T X^T X = I_p$, also $M^T = (X^T X)^{-1} \Rightarrow A = (X^T X)^{-1} X^T$. ∎

Wegen Satz 1.3 heißt der Schätzer $\hat{\boldsymbol{\beta}}_{LS}$ auch der *beste lineare unverfälschte Schätzer* (best linear unbiased estimator, BLUE).

Bemerkung. Satz 1.3, b., impliziert, dass es für einen unverfälschten Schätzer $S(\mathbf{Y}) \neq \hat{\boldsymbol{\beta}}_{LS}$ ein $\mathbf{z} \in \mathbb{R}^p$ gibt, so dass gilt $\mathrm{Var}(\mathbf{z}^T S(\mathbf{Y})) > \mathrm{Var}(\mathbf{z}^T \hat{\boldsymbol{\beta}}_{LS})$. (Schätzer für $\mathbf{z}^T \boldsymbol{\beta}$)

Der kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}_{LS}$ als Maximum-Likelihood-Schätzer. Angenommen, im linearen Modell (1) sind die Fehler normalverteilt, also $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ und somit $Y \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$. Dann ist die Likelihood-Funktion gegeben durch

$$L_n(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right),$$

und die log-Likelihood-Funktion durch

$$\mathcal{L}_n(\boldsymbol{\beta}, \sigma^2) = \log L_n(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{\|\mathbf{Y} - X\boldsymbol{\beta}\|^2}{2\sigma^2}$$

Damit $\mathcal{L}_n(\boldsymbol{\beta}, \sigma^2)$ maximal wird, muss offenbar $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ minimal werden. Daher ist $\hat{\boldsymbol{\beta}}_{LS}$ gleich dem Maximum Likelihood Schätzer $\hat{\boldsymbol{\beta}}_{ML}$ für $\boldsymbol{\beta}$ im linearen Modell mit normalverteilten Fehlern.

Schätzen der Fehlervarianz σ^2.

Zunächste einige Bezeichnungen:

$$\begin{aligned} \hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} = \left(\hat{Y}_1, \ldots, \hat{Y}_n\right)^T \qquad & \textit{angepassten Werte} \text{ (fitted values)} \\ \hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} \qquad & \textit{Residuen} \end{aligned}$$

Idee: Nutze Variation der $\hat{\boldsymbol{\epsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^T$ zum Schätzen von σ^2. Schätze σ^2 durch

$$\hat{\sigma}^2 = \hat{\sigma}^2_{LS} = \frac{1}{n-p} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-p} \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \frac{1}{n-p} \mathbf{Y}^T (I - P_X)\mathbf{Y},$$

[1] $a_i \in \mathrm{Kern}(I_n - P_X) = \mathrm{Bild}(P_X)$; Eigenschaft von Projektionen (hier: P_X)

wobei wir $\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}} = P_X\mathbf{Y}$ und $(I - P_X) = (I - P_X)^2 = (I - P_X)^T$ genutzt haben.

Satz 1.4
Im linearen Modell (1) ist $\hat{\sigma}^2_{LS}$ erwartungstreu für σ^2.

Beweis
Da $E\mathbf{Y} = X\mu$, $\operatorname{Cov}\mathbf{Y} = \sigma^2 I_n$, folgt mit Satz 1.12

$$E(\mathbf{Y}^T(I - P_X)\mathbf{Y}) = \boldsymbol{\mu}^T \underbrace{X^T(I - P_X)X}_{=0}\boldsymbol{\mu} + \operatorname{Spur}(\sigma^2(I - P_X)).$$

Es gilt:

$$\begin{aligned} \operatorname{Spur}(I_n - P_X) &= n - \operatorname{Spur}(X(X^TX)^{-1} \cdot X^T) \\ &= n - \operatorname{Spur}(X^T \cdot X(X^TX)^{-1}) \\ &= n - \operatorname{Spur} I_p \\ &= n - p. \end{aligned} \tag{3}$$

Somit ist

$$E\hat{\sigma}^2_{LS} = \frac{1}{n-p}E(\mathbf{Y}^T(I - P_X)\mathbf{Y}) = \sigma^2.$$

■

Übung Der ML-Schätzer $\hat{\sigma}^2_{ML}$ für σ^2 im linearen Modell mit normalverteilten Fehlern ist gegeben durch $\frac{1}{n}\|Y - X\hat{\boldsymbol{\beta}}_{LS}\|^2$.

Relevante R Befehle `lm` führt eine kleinste Quadrate Schätzung im linearen Modell durch. Auf ein dadurch erzeugtes Objekt kann man `summary` (Schätzwerte, Schätzer für σ^2, Standardfehler und mehr), `fitted.values` (angepasste Werte), `residuals` (Residuen) anwenden.

1.3 Inferenz unter Normalverteilungsannahme

In diesem Abschnitt gehen wir auf die Verteilung der Schätzer sowie relevanter Teststatistiken im linearen Modell (1) mit normalverteilten, homoskedastischen Fehlern, also $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$, ein. Wir beginnen mit dem grundlegenden

Satz 1.5
Im linearen Modell $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, mit normalverteilten, homoskedastischen Fehlern $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$ gilt

$$\hat{\boldsymbol{\beta}}_{LS} \sim N\big(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1}\big), \qquad \frac{n-p}{\sigma^2}\hat{\sigma}^2 \sim \chi^2(n-p),$$

und $\hat{\boldsymbol{\beta}}_{LS}$ und $\hat{\sigma}^2$ sind unabhängig.

Beweis
Da $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$, folgt aus $\hat{\boldsymbol{\beta}}_{LS} = (X^T X)^{-1} X^T \mathbf{Y}$ mit Satz 1.2 direkt die Behauptung für $\hat{\boldsymbol{\beta}}_{LS}$. Weiter ist

$$\hat{\sigma}^2 = \frac{1}{\sigma^2}\, \mathbf{Y}^T (I_n - P_X)\mathbf{Y}.$$

Da $(I_n - P_X)^2 = I_n - P_X$ (die orthogonale Projektion auf das orthogonale Komplement des Spaltenraumes von X), ist Satz 1.13 anwendbar (die Division durch σ^2 standardisiert die $\mathbf{Y}$). Da $(I_n - P_X)X = \mathbf{0}$, ist der Nichtzentralitätsparameter gleich 0. Weiter ist nach (3) $r(I_n - P_X = \text{Spur}(I_n - P_X) = n - p$.
Es bleibt die Unabhängigkeit von $\hat{\boldsymbol{\beta}}_{LS}$ und $\hat{\sigma}^2$ zu zeigen. Diese folgt direkt aus Satz 1.15, da $(I_n - P_X)X = \mathbf{0}$. ∎

Konfidenzintervall für $\mathbf{c}^T\boldsymbol{\beta}$. Wir fixieren einen Vektor $\mathbf{c} \in \mathbb{R}^p$.

Definition. Ein *zweiseitiges Konfidenzintervall* für den Parameter $\mathbf{c}^T\boldsymbol{\beta}$ zum Niveau $1-\alpha$, $\alpha >$ besteht aus zwei messbaren Abbildungen

$$I_l : \mathbb{R}^n \to \mathbb{R}, \qquad I_r : \mathbb{R}^n \to \mathbb{R},$$

so dass für alle $\boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0$ gilt

$$P\big(I_l(\mathbf{Y}) \le \mathbf{c}^T\boldsymbol{\beta} \le I_r(\mathbf{Y})\big) \ge 1 - \alpha.$$

Um ein Konfidenzintervall zu konstruieren, bemerken wir zunächst, dass nach Satz 1.5 gilt

$$\mathbf{c}^T\hat{\boldsymbol{\beta}}_{LS} \sim \mathcal{N}\big(\mathbf{c}^T\boldsymbol{\beta}, \sigma^2 \mathbf{c}^T (X^T X)^{-1}\mathbf{c}\big)$$

und daher

$$\frac{\mathbf{c}^T\hat{\boldsymbol{\beta}}_{LS} - \mathbf{c}^T\boldsymbol{\beta}}{\sigma(\mathbf{c}^T(X^T X)^{-1}\mathbf{c})^{\frac{1}{2}}} \sim \mathcal{N}(0,1).$$

Wir ersetzen nun σ^2 in dem Ausdruck durch den Schätzer. Wegen der Unabhängigkeit von $\hat{\boldsymbol{\beta}}_{LS}$ und $\hat{\sigma}^2$ (und somit auch von $\mathbf{c}^T\hat{\boldsymbol{\beta}}_{LS}$ und $\hat{\sigma}^2$) folgt direkt aus der Definition der t Verteilung

$$PS = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}}_{LS} - \mathbf{c}^T\boldsymbol{\beta}}{\hat{\sigma}\,(\mathbf{c}^T(X^T X)^{-1}\mathbf{c})^{\frac{1}{2}}} \sim t(n-p). \tag{4}$$

PS hängt von den Parametern nur durch den gesuchten Parameter $\mathbf{c}^T\boldsymbol{\beta}$ ab, und seine Verteilung ist unabhängig von den Parametern. Man nennt einen solchen Ausdruck PS dann eine *Pivot-Statistik*.

Hieraus erhält man nun leicht ein Konfidenzintervall zum Niveau $\alpha > 0$: Sei $t_{1-\frac{\alpha}{2}}(n-p)$ das $1-\frac{\alpha}{2}$-Quantil der t-Verteilung mit $n-p$ Freiheitsgraden, wegen der Symmetrie der t-Verteilung ist $t_{\alpha/2} = -t_{1-\frac{\alpha}{2}}(n-p)$. Es ist also für $\alpha < 1/2$

$$P\big(-t_{1-\frac{\alpha}{2}}(n-p) \le PS \le t_{1-\frac{\alpha}{2}}(n-p)\big) = 1-\alpha.$$

Umstellen nach $\mathbf{c}^T\boldsymbol{\beta}$ liefert das Konfidenzintervall

$$I_l = \mathbf{c}^T\hat{\boldsymbol{\beta}}_{LS} - \hat{\sigma}(\mathbf{c}^T(X^TX)^{-1}\mathbf{c})^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p), \qquad I_r = \mathbf{c}^T\hat{\boldsymbol{\beta}}_{LS} + \hat{\sigma}(\mathbf{c}^T(X^TX)^{-1}\mathbf{c})^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p).$$

Man definiert ein *rechtsseitiges Konfidenzintervall* $I_{o,l} : \mathbb{R}^n \to \mathbb{R}$ durch die Bedingung

$$P\big(I_{o,l}(\mathbf{Y}) \le \mathbf{c}^T\boldsymbol{\beta}\big) \ge 1-\alpha,$$

(analog ein *linksseitiges Konfidenzintervall*), und konstruiert diese aus der Pivot Statistik PS.

Aufgabe Konfidenzintervall für σ^2.
relevante R Befehle `confint` liefert für ein Objekt aus `lm` die Konfidenzintervalle der einzelnen Komponenten des KQ Schätzers.

Hypothesen Testen mit dem t Test Für $\mathbf{c} \in \mathbb{R}^p$ und $\delta \in \mathbb{R}$ betrachten wir das *zweiseitige Testproblem*

$$H_{\mathbf{c},\delta} : c^T\boldsymbol{\beta} = \delta \qquad \text{gegen} \qquad K_{\mathbf{c},\delta} : c^T\boldsymbol{\beta} \neq \delta.$$

Definition. Ein *Test* für die Hypothese $H_{\mathbf{c},\delta}$ ist eine messbare Abbildung $T : \mathbb{R}^n \to \{0,1\}$. Der Test T hat *Niveau* $\alpha > 0$, falls gilt:

$$\forall\ \boldsymbol{\beta} \in \mathbb{R}^p \ \text{ mit } c^T\boldsymbol{\beta} = \delta : \qquad P_{\boldsymbol{\beta}}(T(\mathbf{Y}) = 1) \le \alpha.$$

Man interpretiert $T = 0$ als Entscheidung für $H_{\mathbf{c},\delta}$, und $T = 1$ als Entscheidung für $K_{\mathbf{c},\delta}$. Der Test hat also Niveau α, falls man sich nur mit einer Wahrscheinlichkeit von höchstens α fälschlicher Weise für $K_{\mathbf{c},\delta}$ entscheidet.

Um einen Test zu konstruieren, bemerken wir, dass unter $H_{\mathbf{c},\delta}$ gilt

$$S_{\mathbf{c},\delta} = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}}_{LS} - \delta}{\hat{\sigma}(\mathbf{c}^T(X^TX)^{-1}\mathbf{c})^{\frac{1}{2}}} \sim t(n-p).$$

Somit ist

$$T = 1 - 1_{-t_{1-\alpha/2}(n-p) \le S \le t_{1-\alpha/2}(n-p)}$$

ein Test zum Niveau α.

Zusammenhang zum Konfidenzintervall Es gilt

$$T = 0 \quad \Leftrightarrow \quad I_l \le \delta \le I_r.$$

Somit verwirft der Test genau dann, wenn δ nicht im Konfidenzintervall für $\mathbf{c}^T\boldsymbol{\beta}$ liegt. Man zeigt, dass man in dieser Weise aus einem zweiseitigen Konfidenzintervall stets einen zweiseitigen Test für eine Punkthypothese konstruieren kann.

P-Wert Statt der reinen Testentscheidung gibt man den sogenannten P-Wert an. Für das obige Testproblem ist dieser definiert als

$$PV = \begin{cases} 2\,t(n-p)\,(T_{\mathbf{c},\delta}), & T_{\mathbf{c},\delta} < 0, \\ 2\,\big(1 - t(n-p)\,(T_{\mathbf{c},\delta})\big), & T_{\mathbf{c},\delta} > 0, \end{cases} \tag{5}$$

wobei $t(n-p)(x)$ der Wert der Verteilungsfunktion von $t(n-p)$ an der Stelle x ist. Der P-Wert PV ist das beste Niveau, zu dem die Hypothese noch verworfen werden kann. Man verwirft also zu dem (vorher festgelegten) Niveau α, falls $PV > \alpha$.
Unter der Hypothese ist der P-Wert PV auf $(0, 1)$ gleichverteilt.

Für das *einseitige Testproblem*

$$H_{\mathbf{c},\delta} : c^T\boldsymbol{\beta} \leq \delta \qquad \text{gegen} \qquad K_{\mathbf{c},\delta} : c^T\boldsymbol{\beta} > \delta$$

ist

$$S_{\mathbf{c},\delta} = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}}_{LS} - \mathbf{c}^T\boldsymbol{\beta} + \mathbf{c}^T\boldsymbol{\beta} - \delta}{\hat{\sigma}(\mathbf{c}^T(X^TX)^{-1}\mathbf{c})^{\frac{1}{2}}} \sim t\big(n-p; \mathbf{c}^T\boldsymbol{\beta} - \delta\big),$$

wobei unter der Hypothese der Nichtzentralitätsparameter $\mathbf{c}^T\boldsymbol{\beta} - \delta \leq 0$ ist. Man verwirft nun für große Werte von $S_{\mathbf{c},\delta}$, also setzt

$$T = 1 - 1_{S_{\mathbf{c},\delta} \leq t_{1-\alpha}}(n-p),$$

oder äquivalent falls δ nicht in dem linksseitigen Konfidenzintervall für $\mathbf{c}^T\boldsymbol{\beta}$ liegt, bzw. falls der rechtsseitige P-Wert

$$PV = 1 - t(n-p)(S_{\mathbf{c},\delta}) \leq \alpha.$$

Man muss schließlich noch nachweisen, dass der Test T das Niveau überall auf der Hypothese einhält (und nicht nur auf dem Rand). Dies folgt, da die t-Verteilung mit negativem Freiheitsgrad stochastisch kleiner ist als die zentrale t-Verteilung.

relevante R Befehle `summary` liefert für ein Objekt aus `lm` die zweiseitigen P-Werte für die Hypothese H_i.

Vorhersageintervalle Ein *Konfidenzintervall* bezieht sich auf den *Erwartungswert* $c^T\boldsymbol{\beta}$ von $c^T\hat{\boldsymbol{\beta}}$, wobei $\hat{\boldsymbol{\beta}}$ aus dem linearen Modell (1) berechnet wird.

Bei einem *Vorhersageinterall* (Prediction Interval) hingegen ist eine zusätzliche Kovariablenausprägung $\mathbf{x}_{n+1}$ erforderlich, bei der die abhängige Variable Y_{n+1} vorhergesagt werden soll. Das Vorhersageintervall bezieht sich also nicht auf einen Parameter wie das Konfidenzintervall, sondern auf die Zufallsvariable Y_{n+1}.

Sei $\hat{\boldsymbol{\beta}}_{LS}$ der KQ-Schätzer im linearen Modell (1). Als Vorhersage für Y_{n+1} bei $\mathbf{x}_{n+1}$ betrachten man

$$Y^{\text{Pred}} = x_{n+1}^T\hat{\boldsymbol{\beta}}_{LS}.$$

Nach dem linearen Modell würde die Beobachtung Y_{n+1} entstehen durch

$$Y_{n+1} = \mathbf{x}_{n+1}^T \boldsymbol{\beta} + \varepsilon_{n+1},$$

wobei ε_{n+1} und $\boldsymbol{\epsilon}$ unabhängig sind. Somit

$$Y^{\text{Pred}} - Y_{n+1} = \mathbf{x}_{n+1}^T(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}) + \varepsilon_{n+1} \sim \mathcal{N}\big(0, \sigma^2 + \sigma^2 x_{n+1}^T (X^T X)^{-1} x_{n+1}\big),$$

und nach Satz 1.5

$$\frac{Y^{\text{Pred}} - Y_{n+1}}{\hat{\sigma}(1 + \mathbf{x}_{n+1}^T (X^T X)^{-1} \mathbf{x}_{n+1})^{\frac{1}{2}}} \sim t_{n-p}.$$

Als zweiseitigen Vorhersagebereich erhält man

$$\big[Y^{\text{Pred}} - \hat{\sigma}(1+\mathbf{x}_{n+1}^T (X^T X)^{-1} \mathbf{x}_{n+1})^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p), Y^{\text{Pred}} + \hat{\sigma}(1+\mathbf{x}_{n+1}^T (X^T X)^{-1} \mathbf{x}_{n+1})^{\frac{1}{2}} t_{1-\frac{\alpha}{2}}(n-p)\big].$$

Vergleich. Das Vorhersageintervall für Y_{n+1} ist breiter als das Konfidenzintervall für $\mathbf{x}_{n+1}^T \boldsymbol{\beta}$, da der zusätzliche Fehler ε_{n+1} in Y_{n+1} mit berücksichtigt werden muss.

relevante R Befehle `predict.lm` anwenden auf Objekt aus `lm` und zusätzliche Kovariable.

Testen allgemeiner linearer Hypothesen mit dem F-Test Man möchte manchmal allgemeinere lineare Hypothesen, die nicht von der Form $\mathbf{c}^T \boldsymbol{\beta} = \delta$ sind, testen.

Beispiele linearer Hypothesen.

a. H: $\boldsymbol{\beta} = 0$ (alle $\beta_i = 0$)

b. H: $\beta_{i_1} = \ldots = \beta_{i_q} = 0$, $1 \le i_1 < \ldots < i_q \le p$

c. H: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\boldsymbol{\beta}_0 \neq 0$.

Allgemeine lineare Hypothese: Für $A \in \mathbb{R}^{q \times p}$, $q \le p$ mit vollem Rang, $\mathbf{m} \in \mathbb{R}^q$ betrachte

$$H_{A,\mathbf{m}} : \; A\boldsymbol{\beta} = \mathbf{m}.$$

Es gilt

$$A\hat{\boldsymbol{\beta}} - \mathbf{m} \sim \mathcal{N}\big(A\boldsymbol{\beta} - \mathbf{m}, \sigma^2 S\big),$$

wobei wiederum $S = A(X^T X)^{-1} A^T \in \mathbb{R}^{q \times q}$ vollen Rang hat. Dann ist

$$(A\hat{\boldsymbol{\beta}} - \mathbf{m})^T \frac{S^{-1}}{\sigma^2} (A\hat{\boldsymbol{\beta}} - \mathbf{m}) \sim \chi^2(q, \lambda)$$

wobei der Nichtzentralitätsparameter $\lambda = \frac{1}{2}(A\boldsymbol{\beta} - \mathbf{m})^T \frac{S^{-1}}{\sigma^2}(A\boldsymbol{\beta} - \mathbf{m})$, und somit

$$FS = \frac{(A\hat{\boldsymbol{\beta}} - m)^T S^{-1} (A\hat{\boldsymbol{\beta}} - \mathbf{m})}{q\hat{\sigma}^2} \sim F(q, n-p, \lambda)$$

Die Hypothese $H_{A,\mathbf{m}}$ ist äquivalent zu $\lambda = 0$. Daraus bestimmt man den zweiseitigen p-Wert für die lineare Hypothese als $P = 1 - F(q, n-p)(FS)$.

Aufgabe Bestimme den kleinsten Quadrate Schätzer unter der linearen Nebenbedingung $A\boldsymbol{\beta} = m$.

relevante R Befehle `anova` führt den F Test durch, dabei muss das Modell unter $H_{A,\mathbf{m}}$ mit geschätzt worden sein und als Argument übergeben werden. Falls kein zweites Modell übergeben wird, führt `anova` die F-Tests dafür durch, ob bei kategorielle Kovariablen alle Koeffizienten der Dummy Variablen = 0 sind, und gegebenenfalls auch für die Interaktionen.

Konfidenzbereich für $A\boldsymbol{\beta}$ Die Matrix $A \in \mathbb{R}^{q\times p}$, $1 \le q \le p$, habe vollen Rang. Ein wichtiger Spezialfall ensteht, falls A eine Teilmatrix von I_p ist.

Definition. Ein Konfidenzbereich für $A\boldsymbol{\beta}$ zum Niveau $1-\alpha$ ist eine Abbildung

$$K : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^q),$$

so dass für alle $\mathbf{x} \in \mathbb{R}^q$ gilt $\{\mathbf{y} \in \mathbb{R}^n : \ \mathbf{x} \in K(\mathbf{y})\}$ ist messbar, und für alle $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 > 0$ gilt

$$P\big(A\boldsymbol{\beta} \in K(\mathbf{Y})\big) \ge 1 - \alpha.$$

Wir betrachten als (1-α)-Konfidenzbereich für $A\boldsymbol{\beta}$:

$$K(\mathbf{y}) = \Big\{\mathbf{m} \in \mathbb{R}^q : \ \frac{(A\hat{\boldsymbol{\beta}}_{LS} - \mathbf{m})^T S^{-1} (A\hat{\boldsymbol{\beta}}_{LS} - \mathbf{m})}{q\hat{\sigma}^2} \le F_{1-\alpha}(q; n-p)\Big\}.$$

Dann ist

$$\{\mathbf{y} \in \mathbb{R}^n : \ A\boldsymbol{\beta} \in K(\mathbf{y})\} = \{\mathbf{y} \in \mathbb{R}^n : \ \frac{(A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}))^T S^{-1} A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})}{q\hat{\sigma}^2} \le F_{1-\alpha}(q; n-p)\}$$

messbar, und da äÄhnlich wie oben für alle $\boldsymbol{\beta} \in \mathbb{R}^p$ gilt

$$\frac{(A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}))^T S^{-1} A(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})}{q\hat{\sigma}^2} \sim F(q; n-p), \tag{6}$$

erhält man einen $1-\alpha$-Konfidenzbereich für $A\boldsymbol{\beta}$.

Beachte: Der Konfidenzbereich besteht genau aus allen $\mathbf{m} \in \mathbb{R}^q$, für die der F-Test die Hypothese $A\boldsymbol{\beta} = \mathbf{m}$ nicht verwirft. Wir können also umgekehrt auch aus einem Test einen Konfidenzbereich konstruieren.

relevante R Befehle Die library `ellipse` enthält den Befehl `ellipse`, welcher zweidimensionale Konfidenzellipsoide berechnet. Plotten einfach mit `plot`.

1.4 Mittlerer quadratischer Fehler und Ridge-Regression

Für einen Schätzer $\hat{\boldsymbol{\beta}}$ von $\boldsymbol{\beta}$ definiert man den **mittleren quadratischen Fehler** (mean squared error, MSE) durch

$$\mathrm{MSE}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) = E_{\boldsymbol{\beta}}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = E\left(\sum_{i=1}^{p}(\hat{\beta}_i - \beta_i)^2\right)$$

Es gilt:

$$\begin{aligned} E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 &= E\|\hat{\boldsymbol{\beta}} - E\hat{\boldsymbol{\beta}} + E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \\ &= E\|\hat{\boldsymbol{\beta}} - E\hat{\boldsymbol{\beta}}\|^2 + 2\underbrace{E\langle\hat{\boldsymbol{\beta}} - E\hat{\boldsymbol{\beta}}, E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle}_{=0} + \|E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \\ &= E\|\hat{\boldsymbol{\beta}} - E\hat{\boldsymbol{\beta}}\|^2 + \|E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \end{aligned}$$

da

$$E\langle\hat{\boldsymbol{\beta}} - E\hat{\boldsymbol{\beta}}, E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle = \sum_{i=1}^{p} E((\hat{\beta}_i - E\hat{\beta}_i)(E\hat{\beta}_i - \beta_i)) = 0.$$

Also

$$E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \underbrace{E\|\hat{\boldsymbol{\beta}} - E\hat{\boldsymbol{\beta}}\|^2}_{\text{„Varianz-Term“}} + \underbrace{\|E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}_{\text{„Bias-Term“}}.$$

Für *unverfälschte* Schätzer gilt: $\|E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = 0$.
Für *lineare* Schätzer $\hat{\boldsymbol{\beta}} = A\mathbf{Y}$, $A \in \mathbb{R}^{p\times n}$ gilt:

$$\begin{aligned} E\|\hat{\boldsymbol{\beta}} - E\hat{\boldsymbol{\beta}}\|^2 &= E\|A\boldsymbol{\epsilon}\|^2 = E(\boldsymbol{\epsilon}^T A^T A\boldsymbol{\epsilon}) = \sigma^2\,\mathrm{Spur}(A^T A) \\ &= \sigma^2\,\mathrm{Spur}(AA^T) = \mathrm{Spur}(\mathrm{Cov}\,\hat{\boldsymbol{\beta}}). \end{aligned}$$

Für einen linearen, unverzerrten Schätzer $\hat{\boldsymbol{\beta}} = AZ$ ist daher

$$\forall\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \sigma^2 > 0: \qquad \mathrm{MSE}_{\boldsymbol{\beta},\sigma^2}(\hat{\boldsymbol{\beta}}) = \sigma^2\,\mathrm{Spur}(AA^T).$$

Lemma 1.6
Es seien $A, B \in \mathbb{R}^q$ positiv semidefinit. Gilt $A > B$ (d.h. $A - B$ ist positiv semidefinit, aber nicht die Nullmatrix), so folgt $\mathrm{Spur}(A) > \mathrm{Spur}(B)$.

Beweis
Es ist

$$\mathrm{Spur}(A) - \mathrm{Spur}(B) = \mathrm{Spur}(A - B) = \sum_{j=1}^{q}\lambda_j,$$

wobei $\lambda_j \geq 0$ die Eigenwerte von $A - B$ sind, von denen mindestens einer > 0 sein muss, falls $A - B$ nicht die Nullmatrix ist. ∎

Als Folgerung erhalten wir mit dem Satz von Gauß-Markov: Ist $\hat{\boldsymbol{\beta}} = A\mathbf{Y}$ ein linearer unverzerrter Schätzer und ist $A \neq (X^T X)^{-1}X^T$, so gilt

$$\forall\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \sigma^2 > 0: \qquad \mathrm{MSE}_{\boldsymbol{\beta},\sigma^2}(\hat{\boldsymbol{\beta}}) > \mathrm{MSE}_{\boldsymbol{\beta},\sigma^2}(\hat{\boldsymbol{\beta}}_{LS}).$$

Ziel Wir wollen nun einsehen, dass diese Aussage nicht mehr gilt, wenn man innerhalb der Klasse der linearen Schätzer auf die Unverzerrtheit verzichtet.

Betrachte dazu die Spektralzerlegung von X^TX (existiert, da X^TX positiv definit, insbesondere symmetrisch), also

$$X^TX = U\,\mathrm{diag}(\lambda_1,\ldots,\lambda_p)U^T$$

mit U orthogonal, $\lambda_i > 0$. Damit berechnet man den MSE von $\hat{\boldsymbol{\beta}}_{LS}$ als:

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2\,\mathrm{Spur}(X^TX)^{-1} = \sigma^2\sum_{i=1}^{p}\lambda_i^{-1}.$$

Ridge-Regression Für $\alpha > 0$ setze

$$\hat{\boldsymbol{\beta}}_\alpha = (\alpha I_p + X^TX)^{-1}X^T\mathbf{Y}.$$

Berechne Bias- und Varianzterm für $\hat{\boldsymbol{\beta}}_\alpha$:

$$\begin{aligned} E\|\hat{\boldsymbol{\beta}}_\alpha - E\hat{\boldsymbol{\beta}}_\alpha\|^2 &= \sigma^2\,\mathrm{Spur}\left(X(\alpha I_p + X^TX)^{-2}X^T\right) \\ &= \sigma^2\,\mathrm{Spur}\left(X^TX(\alpha I_p + X^TX)^{-2}\right) \end{aligned}$$

Spektralzerlegung

$$X^TX(\alpha I_p + X^TX)^{-2} = U\,\mathrm{diag}\Big(\frac{\lambda_1}{(\alpha+\lambda_1)^2},\ldots,\frac{\lambda_p}{(\alpha+\lambda_p)^2}\Big)U^T$$

Somit:

$$E\|\hat{\boldsymbol{\beta}}_\alpha - E\hat{\boldsymbol{\beta}}_\alpha\|^2 = \sigma^2\cdot\sum_{i=1}^{p}\frac{\lambda_i}{(\alpha+\lambda_i)^2}.$$

Bemerkung Dieser Varianz-Term ist stets kleiner als der Varianz-Term von $\hat{\boldsymbol{\beta}}_{LS}$. Er wird kleiner für wachsendes α.

$$\begin{aligned} \mathrm{Bias}(\alpha) &:= \|E\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}\|^2 \\ &= \|(\alpha I_p + X^TX)^{-1}X^TX\boldsymbol{\beta} - \boldsymbol{\beta}\|^2 \\ &= \|\,\mathrm{diag}\Big(\frac{\lambda_1}{\alpha+\lambda_1} - 1,\ldots,\frac{\lambda_p}{\alpha+\lambda_p} - 1\Big)\cdot U^T\boldsymbol{\beta}\|^2 \\ &= \sum_{i=1}^{p}\frac{\alpha^2}{(\alpha+\lambda_i)^2}(U^T\boldsymbol{\beta})_i^2 \end{aligned}$$

Der Bias-Term wächst mit α.

Satz 1.7
Sind $\boldsymbol{\beta}\in\mathbb{R}^p$, $\sigma^2 > 0$, so gilt für alle $\alpha < \frac{\sigma^2}{\max_i(U^T\boldsymbol{\beta})_i^2}$, dass $\mathrm{MSE}_{\boldsymbol{\beta},\sigma^2}(\hat{\boldsymbol{\beta}}_\alpha) < \mathrm{MSE}_{\boldsymbol{\beta},\sigma^2}(\hat{\boldsymbol{\beta}}_{LS})$.

Beweis
Wir setzen $\mathrm{MSE}_{\boldsymbol{\beta},\sigma^2}(\hat{\boldsymbol{\beta}}_\alpha) =: \mathrm{MSE}(\alpha) = \mathrm{Bias}(\alpha) + \mathrm{Var}(\alpha)$, dann ist $\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{LS}) = \mathrm{MSE}(0)$. Es genügt zu zeigen, dass für $0 \leq \alpha < \frac{\sigma^2}{\max_i (U^T\boldsymbol{\beta})_i^2}$ gilt

$$\frac{d}{d\alpha}(\mathrm{MSE}(\alpha)) < 0.$$

Dies rechnet man einfach nach, den

$$\mathrm{MSE}(\alpha)' = -\sigma^2 \sum_{i=1}^{p} \frac{2\lambda_i}{(\alpha+\lambda_i)^3} + \sum_{i=1}^{p} \frac{2\alpha\lambda_i}{(\alpha+\lambda_i)^3}(U^T\boldsymbol{\beta})_i^2.$$

■

Bemerkungen: 1. Für kleine Werte von $\alpha > 0$ hat der ridge Schätzer einen kleineren MSE als der LS Schätzer. Die Aussage ist aber nicht richtig zufriedenstellend, weil der Wertebereich der zulässigen Parameter α von den unbekannten Parametern des Modells $\boldsymbol{\beta}, \sigma^2$ abhängt. Wir werden später einen Schätzer kennenlerne (James-Stein-Schätzer), bei dem α datengetrieben gewählt wird, und der einen kleineren MSE hat als der LS Schätzer.

2. In der Praxis wird ridge Regression verwendet, wenn die Matrix X^TX schlecht konditioniert ist (also sehr kleine Eigenwerte hat).

3. Enthält X einen Intercept, so wird X meist zunächst standardisiert, und der ridge wird nicht auf den Koeffizienten des Intercept angewendet.

relevante R Befehle Die library `MASS` enthält den Befehl `lm.ridge`. Dabei muss der Ridge-Parameter `lambda` manuell gewählt werden. Man beachte, dass die Matrix X standardisiert wird, und der Ridge nicht auf den Koeffizienten des Intercept angewendet wird.

Ridge als Shrinkage Schätzer

Satz 1.8
a. Ist $\hat{\boldsymbol{\beta}}_{LS} \neq 0$ (äquivalent $X^T\mathbf{Y} \neq 0$), so ist die Funktion $n(\alpha) = \|\hat{\boldsymbol{\beta}}_\alpha\|_p^2$ differenzierbar und streng monoton fallend.

b. Im orthonormalen Design $X^TX = I_p$ gilt speziell $\hat{\boldsymbol{\beta}}_\alpha = \hat{\boldsymbol{\beta}}_{LS}/(1+\alpha)$.

Beweis
b. ist evident, für a. ist $n(\alpha) = \mathbf{Y}^T X(\alpha I_p + X^TX)^{-2}X^T\mathbf{Y}$ differenzierbar mit Ableitung

$$n(\alpha)' = (-2)\,\mathbf{Y}^T X(\alpha I_p + X^TX)^{-3}X^T\mathbf{Y} < 0,$$

da $X^T\mathbf{Y} \neq 0$ nach Voraussetzung und $(\alpha I_p + X^TX)^{-3}$ positiv definit ist. ■

Ridge Schätzer als Lösung von Optimierungsproblemen Man zeigt leicht:

$$\hat{\boldsymbol{\beta}}_\alpha = \mathrm{argmin}_{\boldsymbol{\beta}}(\|Y - X\boldsymbol{\beta}\|_n^2 + \alpha\|\boldsymbol{\beta}\|_p^2).$$

Satz 1.9
Für jedes $\alpha > 0$ existiert ein $t = t(\alpha, X, \mathbf{Y})$, so dass $\hat{\boldsymbol{\beta}}_\alpha$ Lösung des folgenden Optimierungsproblems ist:

$$\text{minimiere} \quad \|\mathbf{Y} - X\boldsymbol{\beta}\|_n^2 \qquad \text{unter Nebenbedingung} \quad \|\boldsymbol{\beta}\|_p^2 \leq t. \tag{7}$$

Bemerkung. Da $\|\mathbf{Y} - X\boldsymbol{\beta}\|_n^2 = \|\mathbf{Y} - X\hat{\boldsymbol{\beta}}_{LS}\|_n^2 + \|X(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})\|_n^2$, kann man in den Optimierungsproblemen $\|\mathbf{Y} - X\boldsymbol{\beta}\|_n^2$ durch $\|X(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})\|_n^2$ ersetzen.

Beweis
Ist $\alpha = 0$, so wähle $t \geq \|\hat{\boldsymbol{\beta}}_{LS}\|_p^2$ beliebig.

Ist $\alpha > 0$, so muss $t < \|\hat{\boldsymbol{\beta}}_{LS}\|_p^2$, und dass Maximum des Optimierungsproblems (7) wird für jedes solche t aus Konvexitätsgründen auf dem Rand angenommen. Ableiten der Lagrange-Funktion

$$L(\boldsymbol{\beta}, \gamma) = \|\mathbf{Y} - X\boldsymbol{\beta}\|_n^2 + \gamma\big(\|\boldsymbol{\beta}\|_p^2 - t\big)$$

nach $\boldsymbol{\beta}$ und $= 0$ setzen liefern als Lösung $\hat{\boldsymbol{\beta}}_\gamma$. Für $t = \|\hat{\boldsymbol{\beta}}_\alpha\|$ liefert die Nebenbedingung dann wegen Satz 1.8 die eindeutige Lösung $\gamma = \alpha$ und somit $\hat{\boldsymbol{\beta}}_\alpha$. ∎

Das Lasso (Tibshirani 1996)
LASSO steht für: least absolute shrinkage and selection operator. Der LASSO Schätzer $\hat{\boldsymbol{\beta}}_t^{\text{Las}}$ ist die Lösung des folgenden Optimierungsproblems. Für $t > 0$ fest

$$\text{minimiere} \quad \|\mathbf{Y} - X\boldsymbol{\beta}\|_n^2 \qquad \text{mit Nebenbedingung} \quad \sum_{k=1}^{p} |\beta_k| \leq t. \tag{8}$$

Analog kann man das Lasso auch über das Optimierungsproblem

$$\hat{\boldsymbol{\beta}}_{Lasso,\alpha} = \text{argmin}_{\boldsymbol{\beta}}(\|Y - X\boldsymbol{\beta}\|_n^2 + \alpha\|\boldsymbol{\beta}\|_1)$$

formulieren.

Hier gilt im orthonormalen Design: $X^T X = I_p$: Lasso = soft thresholdung.

Noch: l_0-Penalisierung = hard thresholding.

relevante R Befehle Die library `lasso2` enthält die Funktion `l1ce`, die den Lasso Schätzer berechnet.

Literatur

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

1.5 Verallgemeinerte kleinste Quadrate

Im linearen Modell (1) habe wir vorausgesetzt, dass die Fehler $\boldsymbol{\epsilon}$ unkorreliert mit gleicher Varianz σ^2 sind. Diese Annahme lassen wir nun fallen und erlauben eine allgemeine Kovarianzstruktur der Fehler. Wir betrachten also das lineare Modell mit allgemeiner Fehlerstruktur

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad E\boldsymbol{\epsilon} = \mathbf{0}, \quad \operatorname{Cov}\boldsymbol{\epsilon} = \boldsymbol{\Sigma}, \tag{9}$$

mit einer positiv definiter Kovarianzmatrix $\Sigma > 0$ für die Fehler $\boldsymbol{\epsilon}$. Man überführt nun das lineare Modell (9) mit allgemeiner Fehlerstruktur in ein Modell mit $\Sigma = I_n$. Dazu setze $\tilde{\mathbf{Y}} = \Sigma^{-\frac{1}{2}}\mathbf{Y}$, $\tilde{X} = \Sigma^{-\frac{1}{2}}X$, $\tilde{\boldsymbol{\epsilon}} = \Sigma^{-\frac{1}{2}}\boldsymbol{\epsilon}$. Dann ergibt (9) mit $\Sigma^{-\frac{1}{2}}$ multipliziert:

$$\tilde{\mathbf{Y}} = \tilde{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}, \tag{10}$$

wobei $\operatorname{Cov}\tilde{\boldsymbol{\epsilon}} = \Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}} = I_n$. Weiter gilt: Genau dann ist $S(\mathbf{Y}) = A\mathbf{Y}$ ein linearer unverfälschter Schätzer im Modell (9), wenn $\tilde{S}(\tilde{\mathbf{Y}}) = A\Sigma^{1/2}\tilde{\mathbf{Y}}$ ein linearer unverfäschter Schätzer im Modell (10) ist. Somit kann man die Resultate im linearen Modell mit unkorrelierten, homoskedastischen Fehler übertragen auf das lineare Modell mit allgemeiner Fehlerstruktur. Wir fassen die wesentlichen Ergebnisse zusammen.

Satz 1.10
Im lineare Modell mit allgemeiner Fehlerstruktur (9) ist der beste lineare, unverfälschte Schätzer für $\boldsymbol{\beta}$ (also der mit kleinster Kovarianzmatrix) gegeben durch

$$\hat{\boldsymbol{\beta}}_{GLS} = (\bar{X}^T\bar{X})^{-1}\bar{X}^T\tilde{\mathbf{Y}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{Y}, \tag{11}$$

dieser hat die Kovarianzmatrix

$$\operatorname{Cov}\hat{\boldsymbol{\beta}}_{GLS} = (X^T\Sigma X)^{-1},$$

und ist bestimmt als Lösung des verallgemeinerten kleinste Quadrate Problems

$$\hat{\boldsymbol{\beta}}_{GLS} = \operatorname{argmin}_{\boldsymbol{\beta}}(\mathbf{Y} - X\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{Y} - X\boldsymbol{\beta}).$$

Der Schätzer $\hat{\boldsymbol{\beta}}_{GLS}$ heißt der *verallgemeinerte Kleinste-Quadrate-Schätzer* (generalized least squares estimator, GLS). Im Modell (9) heißt der Schätzer $\hat{\boldsymbol{\beta}}_{OLS} = (X^TX)^{-1}X^T\mathbf{Y}$ der *gewöhnliche kleinste Quadrate Schätzer* (ordinary least squares, OLS). Dieser ist auch hier unverfälscht und unter allgemeinen Bedingungen konsistent (s. Eicker 1963), hat aber die größere Kovarianzmatrix $\operatorname{Cov}\hat{\boldsymbol{\beta}}_{OLS} = (X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}$. Wir beachten, dass für die Berechnung von $\hat{\boldsymbol{\beta}}_{GLS}$ die Matrix Σ bekannt sein muss.
Falls $\Sigma = \operatorname{diag}(w_1, \ldots, w_n)$, $w_i > 0$, eine Diagonalmatrix ist, spricht man von dem *gewichteten Kleinste-Quadrate-Schätzer*, Notation $\hat{\boldsymbol{\beta}}_{WLS}$ (weighted least squares).

Aufgabe $\hat{\boldsymbol{\beta}}_{GLS}$ als ML-Schätzer, falls der Fehler $\boldsymbol{\epsilon} \sim N(0, \Sigma)$ verteilt ist.

relevante R Befehle Der Befehl `lm` hat die Option `weights`, mit der eine gewichtete kleinste Quadrate Schätzung ausgeführt werden kann.

1.6 Testverteilungen und Verteilung quadratischer Formen

1.6.1 Zufallsvektoren und zufällige quadratische Formen

Sei $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{R}^d$ ein d-variater Zufallsverktor, wobei X_i Zufallsvariable seinen. Der *Erwartunsgwertvektor* von $\mathbf{X}$ ist definiert durch $E\mathbf{X} = (EX_1, \ldots, EX_n)^T$, falls die Erwartungswerte EX_i existieren. Die *Kovarianzmatrix* von $\mathbf{X}$ ist gegeben durch

$$\operatorname{Cov} \mathbf{X} = \big(\operatorname{Cov}(X_i, X_j) \big)_{i,j=1,\ldots,n},$$

falls die X_i endliche Varianzen haben. Für einen Vektor $\mathbf{a} \in \mathbb{R}^d$ gilt

$$\operatorname{Var}(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \operatorname{Cov} \mathbf{X}\, \mathbf{a}.$$

Da die Varianz auf der rechten Seite stets nicht-negativ ist, folgt, dass die Kovarianzmatrix stets positiv semidefinit ist. Weiter ist $\operatorname{Cov} \mathbf{X}$ genau dann degeneriert, falls die X_i (als Abbildungen auf dem zugrundeliegenden W-Raum) linear abhängig sind (fast sicher).

Satz 1.11 (lineare Transformationen)
Sei $\mathbf{X} \in \mathbb{R}^n$ ein Zufallsvektor mit endlichem Erwartungswertvektor $E\mathbf{X}$ und endlicher Kovarianzmatrix $\operatorname{Cov} \mathbf{X}$. Für $A \in \mathbb{R}^{m \times n}$ gilt dann

$$E(A\mathbf{X}) = A\, E\mathbf{X}, \qquad \operatorname{Cov}(A\mathbf{X}) = A \operatorname{Cov} \mathbf{X}\, A^T.$$

Der Beweis ist ein einfaches Nachrechnen. Allgemeiner definieren wir für Zufallsvektoren $X \in \mathbb{R}^d$ und $Y \in \mathbb{R}^q$ die Kovarianzmatrix

$$\operatorname{Cov}(\mathbf{X}, \mathbf{Y}) = \big(\operatorname{Cov}(X_i, Y_j) \big)_{i=1,\ldots,d\, j=1,\ldots,q} \in \mathbb{R}^{d \times q}.$$

Falls $\mathbf{X}$ und $\mathbf{Y}$ unabhängig sind, so gilt (analog zu zwei Zufallsvariable) $\operatorname{Cov}(\mathbf{X}, \mathbf{Y}) = 0$.

Für einen Zufallsvektor $\mathbf{X} \in \mathbb{R}^d$ und eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ betrachten wir die zufällige quadratische Form $\mathbf{X}^T A \mathbf{X}$, eine Zufallsvariable.

Satz 1.12 (Erwartungswert quadratischer Formen)
Sei $\mathbf{X} \in \mathbb{R}^n$ ein Zufallsvektor mit $E\mathbf{X} = \mu \in \mathbb{R}^n$, $\operatorname{Cov} \mathbf{X} = \Sigma \in \mathbb{R}^{n \times n}$, und sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Betrachte die quadratische Form $Q = \mathbf{X}^T A \mathbf{X}$. Dann gilt

$$EQ = \operatorname{Spur}(A \cdot \Sigma) + \mu^T A \mu$$

Beweis

$$Q = \sum_{i,j=1}^{n} a_{ij} Z_i Z_j, \;\; EZ_i Z_j = \Sigma_{ij} + \mu_i \mu_j$$

Somit

$$EQ = \sum_{i,j=1}^{n} a_{ij} (\Sigma_{ij} + \mu_i \mu_j) = \mu^T A \mu + \sum_{i,j=1}^{n} a_{ij} \Sigma_{ij} = \mu^T A \mu + \operatorname{Spur}(A\Sigma)$$

[Beachte: Sowohl A als auch Σ sind symmetrisch.] ∎

1.6.2 Aus der Normalverteilung abgeleitete Verteilungen

a. Ist $\mathbf{X} \sim \mathcal{N}(\mu, I_d)$, so hat $\mathbf{X}^T\mathbf{X} = \sum_{i=1}^d X_i^2$ die **nichtzentrale χ^2-Verteilung** mit d Freiheitsgraden und Nichtzentralitätsparameter $\frac{1}{2}\mu^T\mu$. Schreibweise: $\chi^2(d; \frac{1}{2}\mu^T\mu)$.
In der Tat hängt die Dichte von X^TX,

$$f(u) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{k!} \frac{u^{\frac{1}{2}d+k-1} e^{-\frac{1}{2}u}}{2^{\frac{1}{2}d+k}\Gamma(\frac{1}{2}d+k)}, \qquad \lambda = \frac{1}{2}\mu^T\mu,$$

nur von λ und nicht von ganz μ ab. Für $\lambda = 0$ (bzw. $\mu = 0$) ergibt sich die zentrale χ^2-Verteilung mit d Freiheitsgraden, Bezeichnung $\chi^2(d)$, diese hat die Dichte (Beweis!)

$$f(u) = \frac{u^{d/2-1} e^{-u/2}}{2^{d/2}\Gamma(d/2)}. \tag{12}$$

Man kann zeigen, dass (12) auch für nicht ganzes d eine Dichte definiert, daher kann man die Freiheitsgrade in $(0, \infty)$ variieren lassen.

Weitere Notation:
$\chi^2_\alpha(n)$: Das α-Quantil der zentralen χ^2 Verteilung mit n Freiheitsgraden $(0 < \alpha < 1)$.
$\chi^2(n)(x)$: Wert der Verteilungsfunktion der zentralen χ^2 Verteilung mit n Freiheitsgraden bei x $(x > 0)$.

Relevante R Befehle. `dchisq` (Dichte), `pchisq` (Verteilungsfunktion), `qchisq` (Quantile) und `rchisq` (Zufallszahlen).

b. Ist $U_1 \sim \chi^2(d_1; \lambda)$, $U_2 \sim \chi^2(d_2)$, U_1, U_2 unabhängig, so hat

$$V = \frac{U_1/d_1}{U_2/d_2} \sim F(d_1, d_2; \lambda)$$

die **nichtzentrale F-Verteilung** mit Freiheitsgraden d_1 und d_2 und Nichtzentralitätsparameter λ. Für $\lambda = 0$ erhält man die zentrale F-Verteilung, diese hat die Dichte (Beweis) (DICHTE!!)

Weitere Notation:
$F_\alpha(n, m; \lambda)$: Das α-Quantil der F Verteilung mit n und m Freiheitsgraden $(0 < \alpha < 1)$ und Nichtzentralitätsparameter λ.
$F(n, m; \lambda)(x)$: Wert der Verteilungsfunktion der F Verteilung mit n Freiheitsgraden bei x $(x > 0)$ und Nichtzentralitätsparameter λ.

Relevante R Befehle. `df` (Dichte), `pf` (Verteilungsfunktion), `qf` (Quantile) und `rf` (Zufallszahlen).

c. Ist $X \sim \mathcal{N}(\mu, 1)$, $U \sim \chi^2(d)$, so hat

$$V = \frac{X}{\sqrt{U/d}}$$

die **t-Verteilung** mit d Freiheitsgraden und Nichtzentralitätsparameter μ, Bezeichnung $t(n; \mu)$. Für $\mu = 0$ erhält man die zentrale t-Verteilung, diese hat die Dichte (Beweis) DICHTE!

Weitere Notation:

$t_\alpha(n;\mu)$: Das α-Quantil der t Verteilung mit n und m Freiheitsgraden ($0 < \alpha < 1$) und Nichtzentralitätsparameter μ.
$t(n;\mu)(x)$: Wert der Verteilungsfunktion der t Verteilung mit n Freiheitsgraden und Nichtzentralitätsparameter μ bei x ($x > 0$).

Relevante R Befehle. `dt` (Dichte), `pt` (Verteilungsfunktion), `qt` (Quantile) und `rt` (Zufallszahlen).

Ist bei einer dieser Verteilung der Nichtzentralitätsparameter $= 0$, so lässt man diesen in der Notation einfach weg.

1.6.3 Verteilung quadratischer Formen

Satz 1.13
Sei $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, $A \in \mathbb{R}^{d\times d}$ positiv semidefinit[2]. Ist $A\Sigma$ idempotent, d.h. $(A\Sigma)^2 = A\Sigma$, so gilt

$$\mathbf{X}^T A\mathbf{X} \sim \chi^2(r(A), \frac{1}{2}\mu^T A\mu)$$

(r(A) ist der Rang von A)

Bemerkung
Es gilt auch die Rückrichtung.

Beweis
a. Zunächst sei wieder $\Sigma = I_d$. Wegen $A = A^2$ hat die Spektralzerlegung von A die Form

$$A = Q^T \operatorname{diag}(\underbrace{1,\ldots,1}_{\text{r(A) mal}},0,\ldots,0)Q$$

mit orthogonaler Matrix Q. Somit

$$\mathbf{X}^T A\mathbf{X} = \mathbf{X}^T Q^T \underbrace{\operatorname{diag}(1,\ldots,1,0,\ldots,0)}_{=:D}\underbrace{Q\mathbf{X}}_{=:\mathbf{Y}} = \mathbf{Y}^T D\mathbf{Y} = Y_1^2 + \ldots + Y_{r(A)}^2$$

wobei $Y \sim \mathcal{N}(\underbrace{Q\mu}_{=:v}, I_d)$. Somit gilt:

$$\mathbf{X}^T A\mathbf{X} \sim \chi^2(r(A), \frac{1}{2}\underbrace{(v_1^2 + \ldots + v_{r(A)}^2)}_{=v^T Dv = \mu^T A\mu}) = \chi^2(r(A), \frac{1}{2}\mu^T A\mu)$$

b. Allgemeiner Fall: Ist $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, so gilt $\mathbf{Y} = \Sigma^{-\frac{1}{2}}\mathbf{X} \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}\mu, I_d)$ und $\mathbf{X}^T A\mathbf{X} = \mathbf{Y}^T\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\mathbf{Y}$.
Es ist $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$ idempotent, denn

$$\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}} = \Sigma^{-\frac{1}{2}}\Sigma A\Sigma A\Sigma^{\frac{1}{2}} = \Sigma^{-\frac{1}{2}}\Sigma A\Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}.$$

[2] setzt Symmetrie voraus!

Nach (a) gilt somit

$$\begin{aligned}\mathbf{X}^T A\mathbf{X} &\sim \chi^2(r(\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}), \frac{1}{2}(\Sigma^{-\frac{1}{2}}\mu)^T\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}(\Sigma^{-\frac{1}{2}}\mu)) \\ &= \chi^2(r(A), \frac{1}{2}\mu^T\mu).\end{aligned}$$

da $\Sigma^{\frac{1}{2}}$ vollen Rang hat. ∎

Beispiel 1.14
Es seine $X_1, \ldots, X_n$ unabhängig und $N(\mu, \sigma^2)$ verteilt. Als Schätzer für Erwartungswert und Varianz betrachtet man

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Offenbar ist $\bar{X}_n \sim N(\mu, \sigma^2/n)$. Wir zeigen

$$\frac{n-1}{\sigma^2}S_n^2 \sim \chi^2(n-1). \tag{13}$$

Dazu setze $\mathbf{1}_n = (1, \ldots, 1)^T \in \mathbb{R}^n$ und $P_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ (die Zentrierungsmatrix). Für $\mathbf{X} = (X_1, \ldots, X_n)^T$ ist in der Tat $P_n\mathbf{X} = (X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)^T$. Weiter $P_n^2 = P_n$ (Beweis), also nach Satz 1.13

$$\frac{n-1}{\sigma^2}S_n^2 = \frac{1}{\sigma^2}\mathbf{X}^T P_n\mathbf{X} \sim \chi^2\big(r(P_n), \mu^2\mathbf{1}_n^T P_n\mathbf{1}_n/2\big).$$

Da $P_n^2 = P_n$ und $P_n^T = P_n$, ist $r(P_n) = \text{Spur}\, P_n = n - 1$. Ausserdem ist $P_n\mathbf{1}_n = 0$. Dies zeigt (13).

Satz 1.15 (*Craig und Sakamoto)*
Sei $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$.

a. Ist $A \in \mathbb{R}^{d\times d}$ positiv semidefinit, $B \in \mathbb{R}^{p\times d}$, so gilt

$$B\Sigma A = 0 \;\Rightarrow \mathbf{X}^T A\mathbf{X} \text{ und } B\mathbf{X} \text{ sind unabhängig}$$

b. Ist auch $B \in \mathbb{R}^{d\times d}$ positiv semidefinit, so gilt

$$B\Sigma A = 0 \;\Rightarrow \mathbf{X}^T A\mathbf{X} \text{ und } \mathbf{X}^T B\mathbf{X} \text{ sind unabhängig}$$

Bemerkung Es gilt jeweils auch die Rückrichtung (dies ist der schwerere, aber weniger relevante Teil).

Beweis
a. Spektralzerlegung von A

$$A = Q\,\text{diag}(\lambda_1, \ldots, \lambda_{r(A)}, 0, \ldots, 0)Q^T$$

mit Q orthogonal, $\lambda_i > 0$.
Sei $Q = (q_1, \ldots, q_d)$, $\tilde{Q} = (q_1, \ldots, q_{r(A)}) \in \mathbb{R}^{d \times r(A)}$. Dann

$$A = \tilde{Q} D D \tilde{Q}^T, \; D = \operatorname{diag}(\lambda_1^{\frac{1}{2}}, \ldots, \lambda_{r(A)}^{\frac{1}{2}})$$

Setze $L := \tilde{Q} D$, dann ist $A = LL^T$. Weiter gilt

$$L^T L = D \underbrace{\tilde{Q}^T \tilde{Q}}_{=I_{r(A)}} D = D^2$$

ist invertierbar, also

$$B\Sigma A = B\Sigma L L^T = 0 \; \Rightarrow B\Sigma L (L^T L)(L^T L)^{-1} = B\Sigma L = 0.$$

Nach Satz **??** sind somit die Vektoren $B\mathbf{X}$ und $L^T\mathbf{X}$ unabhängig und somit auch $B\mathbf{X}$ und $\mathbf{X}^T L L^T \mathbf{X} = \mathbf{X}^T A \mathbf{X}$ (ist Funktion von $L^T\mathbf{X}$).

b. Analog. (Zerlege A und B.) ∎

Fortsetzung von Beispiel 13. Da $\bar{X}_n = \mathbf{1}_n^T \mathbf{X}/n$ und $\mathbf{1}_n^T P_n = 0$, sind $\bar{X}_n$ und S_n^2 bei normalverteilten X_i unabhängig.

2 Statistische Modelle

2.1 Formalisierung eines statistischen Modells

Grundlagen

Sei $(\mathcal{X}, \mathcal{F})$ ein messbarer Raum, der sogenannte *Stichprobenraum*.

Definition 2.1
Eine Familie $(P_\theta)_{\theta\in\Theta}$ auf $(\mathcal{X}, \mathcal{F})$, indiziert durch einen Parameter $\theta \in \Theta$, heißt ein *statistisches Modell* (auf $(\mathcal{X}, \mathcal{F})$). Das Tripel $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ heißt *statistischer Raum* (oder auch *statistisches Modell*).
Ein statistisches Modell $(P_\theta)_{\theta\in\Theta}$ heißt *identifizierbar*, falls gilt:

$$\forall\ \theta_1, \theta_2 \in \Theta \text{ gilt: Aus } P_{\theta_1} = P_{\theta_2} \quad \text{folgt} \quad \theta_1 = \theta_2.$$

In etwas loser Terminologie heißt ein statistisches Modell $(P_\theta)_{\theta\in\Theta}$ *parametrisch*, falls $\Theta \subset \mathbb{R}^k$, ansonsten heißt es *nichtparametrisch* (etwa wenn Θ ein Funktionenraum ist oder eine allgemeine Klasse von Wahrscheinlichkeitsmaßen). Hat Θ einen parametrischen (euklidschen) sowie einen nichtparametrischen Anteil, spricht man auch von einem *semiparametrischen Modell*.

Beispiele. 1. Klassische parametrische Modelle.
a. $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, $P_\theta = N(\mu, \sigma^2)$ (Normalverteilung).
b. $(\mathcal{X}, \mathcal{F}) = ((0, \infty), \mathcal{B}(0, \infty))$, $\theta \in (0, \infty)$, $P_\theta = \mathrm{Exp}(\theta)$ (Exponentialverteilung).
c. $\mathcal{X} = \mathbb{N}_0$, $\theta \in (0, \infty)$, $P_\theta = \mathrm{Poi}(\theta)$ (Poissonverteilung).
d. $\mathcal{X} = \{0, \dots, n\}$, $n \in \mathbb{N}$ fest, $\theta \in [0, 1]$, $P_\theta = \mathrm{Bin}(n, \theta)$ (Binomialverteilung).
e. $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}^{n+m}, \mathcal{B}^{n+m})$, $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \in \mathbb{R}^2 \times (0, \infty)^2$, $P_\theta = N(\mu_1, \sigma_1^2)^{\otimes n} \otimes N(\mu_2, \sigma_2^2)^{\otimes m}$ (Zweistichprobenproblem).

2. Nichtparametrische Modelle.
a. $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$, $\Theta = \{\mu \text{ Wahrscheinlichkeitsmaß auf } (\mathbb{R}, \mathcal{B}) \text{ mit } \int_{\mathbb{R}} |x|\, d\mu(x) < \infty\}$, $P_\mu = \mu$.
b. $(\mathcal{X}, \mathcal{F}) = ([a, b], \mathcal{B}[a, b])$, $\Theta = \{f : [a, b] \to [0, \infty) \text{ stetig mit } \int_a^b f(t)\, dt = 1\}$, $dP_f = f d\lambda|_{[a,b]}$.

Die obigen Modelle sind alle identifizierbar.

3. Lineares Modell.
a. Unter Normalverteilungsannahme: $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}^n, \mathcal{B}^n)$, $\Theta = \mathbb{R}^p \times (0, \infty)$, $\theta = (\boldsymbol{\beta}, \sigma^2)$, $P_\theta = N(X\boldsymbol{\beta}, \sigma^2 I_n)$ für eine bekannte (feste) Matrix $X \in \mathbb{R}^{n\times p}$.
b. Allgemeines homoskedastisches Modell: $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}^n, \mathcal{B}^n)$, $\Theta = \mathbb{R}^p \times (0, \infty) \times \mathcal{C}$, $\theta = (\boldsymbol{\beta}, \sigma^2, \mu)$, wobei

$$\mathcal{C} = \Big\{\mu \text{ Wahrscheinlichkeitsmaß auf } (\mathbb{R}^n, \mathcal{B}^n) \text{ mit} \\ \int_{\mathbb{R}^n} x_i\, d\mu(\mathbf{x}) = 0, \int_{\mathbb{R}^n} x_i\, x_j\, d\mu(\mathbf{x}) = \delta_{i,j}, \quad 1 \le i, j \le n,\ i \ne j\Big\},$$

und $P_\theta(B) = \mu\big((B - X\boldsymbol{\beta})/\sigma\big)$.

Beide Modelle sind identifizierbar, falls $p \leq n$ und X vollen Rang p hat. Das allgemeine homoskedastische lineare Modell ist ein typisches Beispiel für ein semiparametrisches Modell.

Definition 2.2
Ist $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ ein statistisches Modell und $(\mathcal{S}, \mathfrak{S})$ ein messbarer Raum, so heißt eine messbare Abbildung $T : \mathcal{X} \to \mathcal{S}$ eine *Statistik* (oder *Stichprobenfunktion*) mit Werten in $\mathcal{S}$. Eine Statistik T induziert auf $(\mathcal{S}, \mathfrak{S})$ ein statistisches Modell $(P_\theta^T)_{\theta\in\Theta}$ über die Verteilungen $P_\theta^T(B) = P_\theta(T^{-1}B)$, $B \in \mathfrak{S}$.

Beispiel. Im linearen Modell unter Normalverteilungsannahme ist der kleineste Quadrateschätzer

$$\hat{\boldsymbol{\beta}}_{LS} : \mathbb{R}^n \to \mathbb{R}^p, \qquad \mathbf{y} \mapsto (X^T X)^{-1} X^T \mathbf{y}$$

eine Statistik mit Verteilungen $P_\theta^{\hat{\boldsymbol{\beta}}_{LS}} = N(\boldsymbol{\beta}, \sigma^2(X^T X)^{-1})$, $\theta = (\boldsymbol{\beta}, \sigma^2)$.

Bemerkung. 1. Häufig ist es nützlich, eine Zufallsvariablen Y oder X mit Werten in $\mathcal{X}$ zu haben, die für jedes θ nach P_θ verteilt ist. Wir können dies leicht durch $Y = id : \mathcal{X} \to \mathcal{X}$ als Statistik konstruieren. Manchmal ist es intuitiver, Y in Abhängigkeit des Parameters θ zu konstruieren, etwa in der Modellgleichung des linearen Modell $\mathbf{Y} = X\boldsymbol{\beta} + \epsilon$. Hier ist $\mathbf{Y}$ ein Zufallsvektor auf dem gleichen Raum wie ϵ, und hängt als Abbildung von $\boldsymbol{\beta}$ ab, diese Abhängigkeit wird in der Notation aber unterdrückt.

2. Ist $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ ein statistisches Modell und $n \in \mathbb{N}$, so können wir das n-fache Produktexperiment $(\mathcal{X}^n, \mathcal{F}^n, (P_\theta^{\otimes n})_{\theta\in\Theta})$ betrachten. Ist $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ wie oben die Identität auf $\mathcal{X}^n$, so sind $Y_1, \ldots, Y_n$ u.i.v.$\sim P_\theta$, $\theta \in \Theta$, wir schreiben oftmals nur diese Annahme.

Dominierte Modelle

Definition 2.3
Ein statistisches Modell $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ heißt *dominiert*, falls ein *dominierendes* σ-endliches Maß μ auf $(\mathcal{X}, \mathcal{F})$ existiert, so dass für alle $\theta \in \Theta$ gilt $P_\theta << \mu$. Die Dichte von P_θ gegeben μ, als Funktion von θ parametrisiert in $x \in \mathcal{X}$,

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

heißt *Likelihood Funktion*.

Bemerkung. Das Produktmodell $(\mathcal{X}^n, \mathcal{F}^, (P_\theta^{\otimes,n})_{\theta\in\Theta})$ eines dominierten statistischen Modells $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ mit dominierenden Maß μ ist dominiert durch das Produktmaß $\mu^{\otimes,n}$ mit Likelihood Funktion

$$L_n(\theta, \mathbf{x}) = \frac{dP_\theta^{\otimes,n}}{d\mu^{\otimes,n}}(\mathbf{x}) = \prod_{i=1}^n \frac{dP_\theta}{d\mu}(x_i) = \prod_{i=1}^n L(\theta, x_i), \qquad \theta \in \Theta,\ \mathbf{x} \in \mathcal{X}^n.$$

Beispiele. 1. $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}^k, \mathcal{B}^k)$, P_θ durch das Lebesguemaß λ^k dominiert.
2. Jedes statistische Modell auf einem endlichen oder abzählbar unendlichen Stichprobenraum ist dominiert durch das Zählmaß.

3. Eine abzählbare Familie $(P_i)_{i\in\mathbb{N}}$ ist durch Wahrscheinlichkeitsmaße der Form $Q = \sum_{i\in\mathbb{N}} c_i P_i$, $c_i > 0$, $\sum_i c_i = 1$ dominiert. Wir werden unten eine weitreichende Verallgemeinerung dieser Tatsache sehen.
4. $(\mathcal{X}, \mathcal{F}) = \big(C[0,1], \mathcal{B}(C[0,1])\big)$. Sei $\mathcal{W}$ das Wiener Maß. Für eine Funktion $f \in L_2[0,1]$ betrachte $F(t) = \int_0^t f(s)\,ds$, $0 \le t \le 1$. Setze

$$\mathcal{W}_f(B) = \mathcal{W}(B - F), \qquad B \in \mathcal{B}(C[0,1]), \quad f \in L_2[0,1].$$

Es ist $\mathcal{W}_f$ die Verteilung des Prozesses $(X_t)_{t\in[0,1]}$ mit

$$dX_t = f(t)\,dt + dB_t, \qquad 0 \le t \le 1,$$

für eine standard Brownsche Bewegung $(B_t)_{t\in[0,1]}$.

Sei $\mathbf{Y} = (Y_t)_{t\in[0,1]}$ der Prozess der Koordinatenauswertung, also $Y_t(f) = f(t)$. Dann besagt der *Satz von Girsanov*: Die Familie

$$\big(\mathcal{W}_f\big)_{f\in L_2[0,1]}$$

wird dominiert durch das Wiener Maß mit Radon-Nikodym Ableitung

$$\frac{d\mathcal{W}_f}{d\mathcal{W}}(\mathbf{Y}) = \exp\Big(\int_0^1 f(t)\,dY_t - \frac{1}{2}\int_0^1 f^2(t)\,dt\Big),$$

wobei $\mathbf{Y}$ unter $\mathcal{W}$ eine standard B.B. ist, und das stochastische Integral $\int_0^1 f(t)\,dY_t$ entsprechend evaluiert wird.

5. $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$, $P_\theta = \delta_\theta$, $\theta \in \Theta$. Diese Familie ist nicht durch ein σ-endliches Maß dominiert: Für jedes dominierende Maß μ muss nämlich gelten $\mu(\{\theta\}) > 0$, $\theta \in \mathbb{R}$. Daraus folgt (s.u.) $\mu(A) = \infty$ für jede überabzählbare Menge $A \in \mathcal{B}$, und daher ist μ nicht σ-endlich. In der Tat: Wegen $\mu(\{\theta\}) > 0$, $\theta \in \mathbb{R}$ ist

$$A = \bigcup_{n\ge 1} \big(A \cap \{x : \mu\{x\} \ge 1/n\}\big).$$

Ist also $\mu(A) < \infty$, so folgt für jede endliche oder abzählbar unendliche Teilmengen $B \subset A \cap \{x : \mu\{x\} \ge 1/n\}$, dass

$$\#B/n \le \mu(A) < \infty,$$

also $\#B < \infty$ und daher $\#A \cap \{x : \mu\{x\} \ge 1/n\} < \infty$. Nach obigem Display ist A selbst daher höchstens abzählbar unendlich.

Satz 2.4
Für ein dominiertes statistisches Modell $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ existieren $c_i \ge 0$, $\sum_i c_i = 1$, $\theta_i \in \Theta$, $i \in \mathbb{N}$, so dass das Wahrscheinlichkeitsmaß $Q = \sum_{i\in\mathbb{N}} c_i P_{\theta_i}$ ein dominierenden Maß für $(P_\theta)_{\theta\in\Theta}$ ist. Ein dominierendes (W-)Maß dieser Gestalt heißt auch *priviligiertes dominierendes Maß*.

Bemerkung. Ist μ ein beliebiges (σ-endliches) dominierendes Maß für $(P_\theta)_{\theta\in\Theta}$ und Q ein priviligiertes dominierendes Maß, so gilt offenbar stets $Q << \mu$, und daher $L_\mu(\theta, x) = L_Q(\theta, x)\,dQ/d\mu$.

Beweis
Wir nehmen zunächst μ als endlich an. Setze

$$\mathcal{P}_0 = \Big\{ \sum_{i=1}^{\infty} c_i P_{\theta_i}, \quad \theta_i \in \Theta,\ c_i \geq 0,\ \sum_i c_i = 1 \Big\},$$
$$\mathcal{A} = \big\{ A \in \mathcal{F} :\ \exists\, P \in \mathcal{P}_0 :\ P(A) > 0,\ dP/d\mu > 0 \quad \mu - \text{f.ü. auf } A \big\}.$$

Wähle $(A_n)_n \subset \mathcal{A}$ mit

$$\mu(A_n) \to \sup_{A \in \mathcal{A}} \mu(A) < \infty, \qquad n \to \infty, \tag{14}$$

wobei das supremum wegen der Endlichkeit von μ endlich ist. Weiter sei $P_n \in \mathcal{P}$ zu $A_n \in \mathcal{A}$ im Sinne der Definition von $\mathcal{A}$ gewählt. Wähle nun $c_n > 0$, $\sum_n c_n = 1$ beliebig und setze

$$Q = \sum_n c_n P_n, \qquad A_\infty = \bigcup_n A_n.$$

Dann gelten: $Q \in \mathcal{P}_0$, und $A_\infty \in \mathcal{A}$ mit zugehörigem Maß $Q \in \mathcal{P}_0$, denn $dQ/d\mu \geq c_n dP_n/d\mu > 0$ μ-f.ü. auf A_n, und daher nach (14)

$$\mu(A_\infty) = \sup_{A \in \mathcal{A}} \mu(A) < \infty. \tag{15}$$

Wir zeigen nun $P << Q$ für alle $P \in \mathcal{P}_0$, da die P_θ inbesondere in $\mathcal{P}_0$ enthalten sind, folgt die Behauptung. Angenommen, dies wäre nicht so, also existiere $P \in \mathcal{P}_0$, $A \in \mathcal{F}$ mit $P(A) > 0$ aber $Q(A) = 0$. Wir führen dies zum Widerspruch zur Maximalität von A_∞ in (15).
Da $dQ/d\mu > 0$ auf A_∞ nach b. sind Q und μ auf A_∞ äquivalent, und da $P << \mu$, folgt

$$Q(A \cap A_\infty) = 0 \quad \Rightarrow \mu(A \cap A_\infty) \quad \Rightarrow P(A \cap A_\infty) = 0.$$

Für $B = \{dP/d\mu > 0\}$ ist $P(B) = 1$, also

$$P(A) = P(A \cap A_\infty^c \cap B) > 0 \quad \Rightarrow \mu(A \cap A_\infty^c \cap B) > 0.$$

Setze nun

$$D = A_\infty \cup \big(A \cap A_\infty^c \cap B\big).$$

Dann
a. $D \in \mathcal{A}$, denn $(P+Q)/2 \in \mathcal{P}_0$, und $d(P+Q)/2d\mu > 0$ μ-f.ü. auf D,
b. $\mu(D) > \mu(A_\infty)$, also Widerspruch zu (15).
μ σ-endlich: Zerlege $\mathcal{X} = \bigcup_n F_n$ mit $\mu(F_n) < \infty$. Ist $\mu(F_n) = 0$, so setze $Q_n = P_\theta$ für $\theta \in \Theta$ beliebig. Ansonsten schränke μ und alle P_θ auf F_n ein, und konstruiere wie oben Q_n (die obige Konstruktion nutzt nicht, dass die P_θ Wahrscheinlichkeitsmaße sind), und setze schließlich $Q = \sum_n 2^{-n} Q_n$. ∎

2.2 Exponentielle Familien

Wir betrachten eine Klasse von dominierten Familien, deren Likelihood Funktionen besonders günstige analytische Eigenschaften haben.

Sei $(\mathcal{X}, \mathcal{F})$ ein messbarer Raum (Stichprobenraum).

Definition 2.5
Ein parametrisches Modell $(P_\theta)_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^p$, auf $(\mathcal{X}, \mathcal{F})$ heißt Exponentialfamilie, falls ein σ-endliches dominierendes Maß μ für $(P_\theta)_{\theta\in\Theta}$ existiert, s.d. die Likelihood Funktion die Form

$$L(\theta, x) = \frac{dP_\theta}{d\mu}(x) = C(\theta)\, \exp\big(Q(\theta)^T\, T(x)\big)\, h(x) \tag{16}$$

hat, wobei

$$\begin{aligned} Q &= (q_1, \dots, q_k)^T : \Theta \to \mathbb{R}^k, \qquad & C &: \Theta \to \mathbb{R}, \\ T &= (T_1, \dots, T_k)^T : \mathcal{X} \to \mathbb{R}^k, \qquad & h &: \mathcal{X} \to [0, \infty) \end{aligned}$$

messbare Abbildungen sind. Die Statistik T heißt die *natürliche suffiziente Statistik* der Exponentialfamilie $(P_\theta)_{\theta\in\Theta}$.

Bemerkungen. 1. Ist ν mit $d\nu = h d\mu$, so gilt offenbar

$$\frac{dP_\theta}{d\nu}(x) = C(\theta)\, \exp\big(Q(\theta)^T\, T(x)\big).$$

Also:
a. Man kann $h = 1$ annehmen.
b. Da $\frac{dP_\theta}{d\nu} > 0$, haben alle P_θ den gleichen Träger (den von ν).

2. Da $\frac{dP_\theta}{d\mu}$ bzgl. μ zu 1 über $\mathcal{X}$ integriert für jedes θ, muss $C(\theta) > 0$ sein, also gilt

$$\int_{\mathcal{X}} \exp\big(Q(\theta)^T\, T(x)\big)\, h(x)\, d\mu(x) = C(\theta)^{-1} < \infty \tag{17}$$

3. Wegen (16) und (17) hängt $\frac{dP_\theta}{d\mu}$ und somit P_θ nur von θ nur duch $Q(\theta)$ ab. Mit $q \in Q(\Theta) \subset \mathbb{R}^k$ und

$$C(q) = \Big(\int_{\mathcal{X}} \exp\big(q^T\, T(x)\big)\, h(x)\, d\mu(x)\Big)^{-1} \tag{18}$$

können wir schreiben

$$L(q, x) = \frac{dP_q}{d\mu}(x) = C(q)\, \exp\big(q^T\, T(x)\big)\, h(x). \tag{19}$$

Dann heißt $q \in Q(\Theta)$ natürlicher Parameter der Exponentialfamilie. Die Menge

$$Q_* = \big\{q \in \mathbb{R}^k : \int_{\mathcal{X}} \exp\big(q^T\, T(x)\big)\, h(x)\, d\mu(x) < \infty\big\}$$

heißt *natürlicher Parameterraum* der Exponentialfamilie. Die Integrale sind dann strikt positiv.

Offenbar ist $Q(\Theta) \subset Q_*$, und Q_* ist der maximale Definitionsbereich für den natürlichen Parameter. Die Exponentialfamilie heißt *komplett*, falls $Q(\Theta) = Q_*$.

Satz 2.6
Der natürliche Parameterraum einer Exponentialfamilie ist konvex.

Beweis
Da $\exp(x)$ konvex ist, also für $\lambda \in [0,1]$ gilt

$$\exp\big(\lambda x + (1-\lambda)y\big) \le \lambda \exp(x) + (1-\lambda)\exp(y), \qquad x, y \in \mathbb{R},$$

folgt für $q, r \in Q_*$

$$\begin{aligned}&\int_{\mathcal{X}} \exp\big(\lambda q^T T(x) + (1-\lambda)\, r^T T(x)\big)\, h(x)\, d\mu(x)\\ \le \lambda &\int_{\mathcal{X}} \exp\big(q^T T(x)\big)\, h(x)\, d\mu(x) + (1-\lambda) \int_{\mathcal{X}} \exp\big(r^T T(x)\big)\, h(x)\, d\mu(x) < \infty\end{aligned}$$ ∎

Beispiele. 1. *Normalverteilung* $N(\mu, \sigma^2)$. $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$, $\mu = \lambda$, $\theta = (\mu, \sigma^2)^T \in \mathbb{R} \times (0, \infty)$. Schreibe

$$f(x,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{\mu^2}{2\sigma^2}\big) \exp\big(\frac{x\,\mu}{\sigma^2} - \frac{x^2}{2\,\sigma^2}\big),$$

daher

$$\begin{aligned} T(x) &= (x, x^2)^T && \text{natürlicher suffiziente Statistik,}\\ q(\theta) &= \big(\frac{\mu}{\sigma^2}, -\frac{1}{2\,\sigma^2}\big)^T && \text{natürlicher Parameter.}\end{aligned}$$

und $Q_* = \mathbb{R} \times (-\infty, 0)$ natürlicher Parameterraum. Die EF ist komplett.

2. *Poissonverteilung* $Poi(\lambda)$. $(\mathcal{X}, \mathcal{F}) = (\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$, $\mu =$ Zählmaß auf $\mathbb{N}_0$, $\theta = \lambda$, $\Theta = (0, \infty)$, und

$$f(x,\theta) = e^{-\theta} \exp(x \log\theta)\, h(x), \qquad x \in \mathbb{N}_0,$$

also $T(x) = x$, $q(\theta) = \log\theta$, $Q_* = \mathbb{R}$, ist komplett.

3. *Binomialverteilung* $Bin(n, p)$. $\mathcal{X} = \{0, \dots, n\}$ mit Potenzmenge, $\mu =$ Zählmaß, $\theta = p \in (0,1)$,

$$f(x,\theta) = (1-\theta)^n \exp\big(x\, \log(\theta/(1-\theta))\big) \binom{n}{x}, \qquad 0 \le x \le n,$$

$T(x) = x$, $q(\theta) = \log(\theta/(1-\theta))$ (log-odds-ratio) natürlicher Parameter, $Q_* = \mathbb{R}$.

4. *Lineares Modell* $N(X\boldsymbol{\beta}, \sigma^2 I_n)$, $\mathcal{X} = \mathbb{R}^n$, $\mu = \lambda^n$, $\theta = (\boldsymbol{\beta}, \sigma^2) \in \Theta = \mathbb{R}^p \times (0, \infty)$, und

$$f(\mathbf{y},\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\big(-\frac{\boldsymbol{\beta}^T X^T X \boldsymbol{\beta}}{2\sigma^2}\big) \exp\big(\frac{\mathbf{y}^T X\boldsymbol{\beta}}{\sigma^2} - \frac{\mathbf{y}^T\mathbf{y}}{2\,\sigma^2}\big),$$

also $T(\mathbf{y}) = (X^T\mathbf{y}, \mathbf{y}^T\mathbf{y})$, $q(\theta) = (\boldsymbol{\beta}/\sigma^2, -1/(2\sigma^2))$.

5. *Normalverteilung* $N(\theta, \theta^2)$, $\theta \in \mathbb{R}$, dann

$$f(x,\theta) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\big(-1/2\big) \exp\big(x/\theta - x^2/(2\,\theta^2)\big),$$

$T(x)(x, x^2)$, $q(\theta) = (1/\theta, -1/(2\theta^2))$ nicht komplett.

6. *Wiener Prozess mit parametrischen Drift* $(\mathcal{X}, \mathcal{F}) = \big(C[0,1], \mathcal{B}(C[0,1])\big)$. Sei $\mathcal{W}$ das Wiener Maß, für $f \in L_2[0,1]$ betrachte $F(t) = \int_0^t f(s)\,ds$, $0 \le t \le 1$ und

$$\mathcal{W}_f(B) = \mathcal{W}(B - F), \qquad B \in \mathcal{B}(C[0,1]), \quad f \in L_2[0,1].$$

Sei $\mathbf{Y} = (Y_t)_{t\in[0,1]}$ der Prozess der Koordinatenauswertung, also $Y_t(f) = f(t)$, so dass $\big(\mathcal{W}_f\big)_{f\in L_2[0,1]}$ dominiert wird durch das Wiener Maß mit Radon-Nikodym Ableitung

$$\frac{d\mathcal{W}_f}{d\mathcal{W}}(\mathbf{Y}) = \exp\Big(\int_0^1 f(t)\,dY_t - \frac{1}{2}\int_0^1 f^2(t)\,dt\Big),$$

wobei $\mathbf{Y}$ unter $\mathcal{W}$ eine standard B.B. ist, und das stochastische Integral $\int_0^1 f(t)\,dY_t$ entsprechend evaluiert wird.

Ist f von beschränkter Variation, so gilt

$$\int_0^1 f(s)\,dY(s) = f(1)Y(1) - \int_0^1 Y(s)\,df(s).$$

Für ein parametrischen Modell $a_0 + a_1 f_1 + \ldots + a_k f_k$, mit festen Funktionen f_j von beschränkter Variation und unbekannten Parametern $a_j \in \mathbb{R}$, erhält man daher eine EF mit natürlichem Parameter

$$\big(a_0 + a_1 f(1) + \ldots + a_k f(k), -a_1, \ldots, -a_k\big)$$

und natürlicher suffizienter Statistik

$$\big(Y(1), \int_0^1 Y(s)\,df_1(s), \ldots, \int_0^1 Y(s)\,df_k(s)\big).$$

7. *Gleichverteilung* $U(0,\theta)$, $\theta > 0$, ist keine Exponentialfamilie, da unterschiedliche Träger.

Satz 2.7
Ist $(P_\theta)_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^p$, eine Exponentialfamilie auf $(\mathcal{X}, \mathcal{F})$ bzgl. μ wie in (16), und sind $X_1, \ldots, X_n$ u.i.v. P_θ, $\theta \in \Theta$, so ist $(P_\theta^n)_{\theta\in\Theta}$ eine Exponentialfamilie auf $(\mathcal{X}^n, \mathcal{F}^n)$ bzgl. μ^n, und es gilt mit $\mathbf{x} = (x_1, \ldots, x_n)^T$

$$\frac{dP_\theta^n}{d\mu^n}(\mathbf{x}) = C(\theta)^n \exp\Big(Q(\theta)^T \sum_{i=1}^n T(x_i)\Big) \prod_{j=1}^n h(x_j).$$

Beweis
Nach der Produktformel gilt

$$\frac{dP_\theta^n}{d\mu^n}(\mathbf{x}) = \prod_{i=1}^n \frac{dP_\theta}{d\mu}(x_i).$$

■

Beispiel. Für $N(\mu 1_d, \sigma^2 I_d)$ auf $(\mathbb{R}^d, \mathcal{B}^d)$, $1_d = (1,\ldots,1)^T$, $\theta = (\mu, \sigma^2)$ ist $T(x) = (\sum_i x_i, \sum_i x_i^2)$ die natürliche suffiziente Statistik.

Beispiel. *Multinomialverteilung* $M(n; p_0, \ldots, p_s)$, also n Wiederholungen, $s+1$ Ausgänge, $p_j > 0$, $p_0 + \ldots + p_s = 1$. Die Dichte bzgl. des Zählmaßes auf $\{0,\ldots,n\}^{s+1}$ ist gegeben durch

$$\begin{aligned} f(x_0,\ldots,x_s; p_0,\ldots,p_s) &= \frac{n!}{x_0!\ldots x_s!}\, p_0^{x_0}\ldots p_s^{x_s}\, 1_{x_0+\ldots+x_s=n} \\ &= \exp\big(x_0\,\log p_0 + \ldots + x_s \log p_s\big)\, \frac{n!}{x_0!\ldots x_s!}\, 1_{x_0+\ldots+x_s=n}. \end{aligned}$$

Beachte: der Träger der Verteilungen ist $\{\mathbf{x} \in \{0,\ldots,n\}^{s+1} :\ x_0 + \ldots + x_s = n\}$. In obiger Darstellung ist $T(\mathbf{x}) = (x_0,\ldots,x_s)$, $q = (\log p_0, \ldots, \log p_s)$. Die Familie ist nicht komplett, denn $Q_* = \mathbb{R}^{s+1}$ und dann

$$C(\mathbf{q}) = \Big(\sum_{x_0+\ldots+x_s=n} \frac{n!}{x_0!\ldots x_s!}\, \exp\big(x_0\, q_0 + \ldots + x_s q_s\big)\Big)^{-1} = \Big(\big(e^{q_0} + \ldots + e^{q_s}\big)^n\Big)^{-1}.$$

Es kommen hierdurch jedoch keine weiteren Verteilungen hinzu, denn es gilt $P_{\mathbf{q}} = P_{\mathbf{q}_{norm}}$, wobei

$$\mathbf{q}_{norm} = (\log p_0, \ldots, \log p_s)^T, \qquad p_j = \frac{e^{q_j}}{e^{q_0} + \ldots + e^{q_s}}.$$

Insbesondere ist die komplette Familie nicht identifizierbar. Dies liegt an $x_0 + \ldots + x_s = n$.

Satz 2.8
Gilt für die Exponentialfamilie in (16) für alle $\lambda_0, \ldots, \lambda_k \in \mathbb{R}$

$$\text{Aus } \lambda_0 + \lambda_1 T_1(x) + \ldots + \lambda_k T_k(x) = 0 \quad h\,d\mu - \text{f.ü.} \qquad \text{folgt} \qquad \lambda_0 = \ldots = \lambda_k = 0,$$

so ist der natürliche Parameter q über den natürlichen Parameterraum Q_* identifizierbar.

Beweis
o.E. $h = 1$. Angenommen, $q_1, q_2 \in Q_*$ mit $P_{q_1} = P_{q_2}$. Dann ist

$$\frac{dP_{q_1}}{d\mu} = \frac{dP_{q_2}}{d\mu} \qquad \mu - \text{f.ü.},$$

also

$$\frac{C_{q_1}}{C_{q_2}} = \exp\big((q_2 - q_1)^T T(x)\big) \qquad \text{für } \mu - \text{f.a. } x \in \mathcal{X},$$

also

$$(q_2 - q_1)^T T(x) - \log\Big(\frac{C_{q_1}}{C_{q_2}}\Big) = 0 \qquad \text{für } \mu - \text{f.a. } x \in \mathcal{X}.$$

Wegen der linearen Unabhängigkeit folgt $q_1 = q_2$. ■

Definition. Unter der Bedingung des Satzes und falls Q_* ein nicht-leeres Inneres hat, heißt die Exponentialfamilie parametrisiert im natürlichen Parameter *strikt k-dimensional oder k-parametrisch.*

Beispiel. *Multinomialverteilung.* Schreibe

$$f(x_0,\ldots,x_s;p_0,\ldots,p_s) = \exp(n\log p_0)\,\exp\big(x_1\,\log(p_1/p_0)+\ldots+x_s\log(p_s/p_0)\big)$$
$$\frac{n!}{x_0!\ldots x_s!}\,1_{x_0+\ldots+x_s=n},$$

also mit $q_i=\log(p_i/p_0)$, $Q_*=\mathbb{R}^s$, so ist die Familie komplett und strikt s-parametrisch mit natürlicher suffizienter Statistik $T(x)=(x_1,\ldots,x_s)$.

Lemma 2.9
Sei $\phi:\mathcal{X}\to\mathbb{R}$ messbar mit

$$\int_{\mathcal{X}} |\phi(x)|\,\exp\big(q^TT(x)\big)\,h(x)\,d\mu(x)<\infty \qquad \forall\, q\in \operatorname{int} Q_*,$$

so ist

$$A(q)=\int_{\mathcal{X}}\phi(x)\,\exp\big(q^TT(x)\big)\,h(x)\,d\mu(x)\in C^\infty$$

und kann unter dem Integral differenziert werden.

Beweis (*Beweisskizze)*
1. Reduktion: $k=1$ (sonst partielle Ableitungen), $\phi(x)h(x)=1$ (sonst zerlege in Positiv- und Negativteil, und ziehe mit ins dominierende Maß μ), nur erste Ableitung (sonst Induktion). Betrachte also:

$$A(q)=\int_{\mathcal{X}}\exp\big(qT(x)\big)d\mu(x).$$

2. Zeige: Um q_0 (innerer Punkt von Q_*) wird der Differenzenquotient durch eine integrierbare Funktion majorisiert. Dann folgt Behauptung (Elstrodt S. ???). Dazu: es gilt für $\delta>0$ und $t,z\in\mathbb{R}$

$$\Big|\frac{\exp(tz)-1}{z}\Big|\le\frac{\exp(\delta|t|)}{\delta},\qquad 0<|z|\le\delta/2, \tag{20}$$

für Beweis s.u. Für $0<|q-q_0|\le\delta/2$ können wir daher abschätzen

$$\begin{aligned}\Big|\frac{\exp\big(qT(x)\big)-\exp\big(q_0T(x)\big)}{q-q_0}\Big| &= \exp\big(q_0T(x)\big)\,\Big|\frac{\exp\big((q-q_0)T(x)\big)-1}{q-q_0}\Big|\\ &\le \frac{1}{\delta}\,\exp\big(q_0T(x)\big)\,\exp\big(\delta|T(x)|\big)\\ &\le \frac{1}{\delta}\Big(\exp\big((q_0+\delta)T(x)\big)+\exp\big((q_0-\delta)T(x)\big)\Big),\end{aligned}$$

eine integrierbare Majorante für $q_0\pm\delta\in Q_*$.

Zu (20): Wir betrachten den Fall $t,z>0$ und müssen zeigen, dass für $0<z<\delta/2$ gilt

$$z\,\exp(\delta t)-\delta\big(\exp(tz)-1\big)\ge 0.$$

■

Für $z=0$ steht links die 0, also genügt es zu zeigen, dass die Ableitung (nach z) nicht-negativ ist. Dies ergibt die Bedingung $\exp(\delta t)-\delta t\exp(tz)\ge 0$, wofür wegen $z<\delta/2$ genügt $\exp(\delta t/2)\ge\delta t$, welches wegen $e^x\ge 2x$ stets erfüllt ist.

Satz 2.10
Sei $(P_q)_{q\in Q}$ eine Exponentialfamilie mit

$$\frac{dP_q}{d\mu}(x) = C(q)\,\exp\left(q^T T(x)\right) h(x),$$

also parametrisiert im natürlichen Parameter. Ist $X \sim P_q$,

$$b(q) = \log\Big(\int_{\mathcal{X}} \exp\left(q^T T(x)\right) h(x)\,d\mu(x)\Big),$$

und q ein innerer Punkt von Q, so gelten

$$\begin{aligned} E_q T(X) &= \frac{d}{dq}\,b(q), \\ \mathrm{Cov}_q T(X) &= \frac{d^2}{dq\,dq^T}\,b(q), \end{aligned} \tag{21}$$

Beweis
Nach dem Lemma ist $b(q) \in C^\infty$, somit nach (18)

$$\begin{aligned} \frac{d}{dq}\,b(q) &= C(q)\,\frac{d}{dq}\Big(\int_{\mathcal{X}} \exp\left(q^T T(x)\right) h(x)\,d\mu(x)\Big) \\ &= C(q)\int_{\mathcal{X}} T(x)\,\exp\left(q^T T(x)\right) h(x)\,d\mu(x) \\ &= E_q T(X), \end{aligned}$$

also der erste Teil von (21). Für den zweiten Teil beachte, dass $C(q) = \exp(-b(q))$, folgt

$$\begin{aligned} \frac{d^2}{dq\,dq^T}\,b(q) &= \frac{d}{dq}\Big(C(q)\int_{\mathcal{X}} T^T(x)\,\exp\left(q^T T(x)\right) h(x)\,d\mu(x)\Big) \\ &= C(q)\int_{\mathcal{X}} T(x)\,T^T(x)\,\exp\left(q^T T(x)\right) h(x)\,d\mu(x)\Big) \\ &\quad - C(q)\,\frac{d}{dq}\,b(q)\int_{\mathcal{X}} T^T(x)\,\exp\left(q^T T(x)\right) h(x)\,d\mu(x)\Big) \\ &= E_q T(X)\,T^T(X) \;-\; E_q T(X)\,E_q T^T(X). \end{aligned}$$

■

2.3 Suffizienz

Reguläre bedingte Verteilungen

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{F})$ und $(\mathcal{S}, \mathfrak{S})$ messbare Räume und $Y : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{S}$ messbar.

Definition Eine *reguläre Version der bedingten Verteilung* $P_{Y|T}(\cdot|\cdot) = P_Y(\cdot|T = \cdot)$ von Y gegeben T ist ein *Markov Kern*

$$P_{Y|T}(\cdot|\cdot): \quad \mathcal{S} \times \mathcal{F} \to [0,1],$$

also
1. $\forall\ t \in \mathcal{S}$ ist $P_{Y|T}(\cdot|t)$ ein W-Maß auf $(\mathcal{X}, \mathcal{F})$,
2. $\forall\ F \in \mathcal{F}$ ist $t \mapsto P_{Y|T}(F|t)$ $(\mathfrak{S} - \mathcal{B})$-messbar,
für den für alle $F \in \mathcal{F}$ gilt

$$P_{Y|T}\big(F|T(\omega)\big) = E\big(1_{Y\in F}|\sigma(T)\big)(\omega) \quad \text{für} P - f.a.\ \omega \in \Omega.$$

Bemerkung. Der Markov-Kern ist also genau dann eine reguläre Version der bedingten Verteilung von Y gegeben T, falls für alle $F \in \mathcal{F}$, $B \in \mathfrak{S}$ gilt

$$\int_{T^{-1}(B)} P_{Y|T}\big(F|T(\omega)\big)\, dP(\omega) = P\big(F \cap T^{-1}(B)\big).$$

Suffizienz: Definition und Beispiele

Sei $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ ein statistisches Modell und $Y \sim P_\theta$.

Wir wollen Statistiken $T : \mathcal{X} \to \mathcal{S}$ charakterisieren, die alle Informationen bzgl. θ enthalten, die auch in Y enthalten sind. Man möchte also die Daten Y ohne Informationsverlust komprimieren.

Beispiel. Sei $Y = (Y_1, \ldots, Y_n)^T$, Y_i u.i.v. $\sim Ber(p)$, $0 < p < 1$. Dann sollte $T(Y) = \sum_{k=1}^n Y_k \sim Bin(n,p)$ alle Informationen über p enthalten, die auch in Y enthalten sind (nur Anzahl der Erfolge, nicht deren Position sind relevant wegen der u.i.v. Annahme).

Definition 2.11
Eine Statistik $T : \mathcal{X} \to \mathcal{S}$ heißt *suffizient* für $(P_\theta)_{\theta\in\Theta}$, falls die reguläre bedingte Verteilung $P_\theta(\cdot|T = \cdot)$ von P_θ gegeben T auf $(\mathcal{X}, \mathcal{F})$ (d.h. von Y gegeben T unter P_θ) (existiert und) nicht von θ abhängt. Notation: $P_\theta(\cdot|T = \cdot) = P_\cdot(\cdot|T = \cdot)$.

Bemerkungen. 1. Man kann also aus den Bildmaßen $(P_\theta^T)_{\theta\in\Theta}$ auf $(\mathcal{S}, \mathfrak{S})$ und aus $P_\cdot(\cdot|T)$ das Modell $(P_\theta)_{\theta\in\Theta}$ rekonstruieren über

$$P_\theta(F) = \int_{\mathcal{S}} P_\cdot(F|T = t)\, dP_\theta^T(t), \qquad F \in \mathcal{F}.$$

2. Ist T suffizient, $T = \psi(S)$, dann S suffizient. Also Suffizienz: Keine zu starke Datenreduktion. Argument s. hinten.

Beispiele. 1. Forsetzung Bernoulli/Binomialverteilung. Für $x \in \{0,1\}^n$, $t \in \{0,\dots,n\}$, $\theta = p$, ist

$$P_\theta(Y = x|T = t) = \frac{P_\theta(Y = x, T = t)}{P_\theta(T = t)} = \begin{cases} 0, & \sum_i x_i \neq t, \\ \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}, & \text{sonst} \end{cases}$$

Man erhält also bedingt auf $T = t$ die Gleichverteilung auf $\{x \in \{0,1\}^n : \sum_i x_i = t\}$.

2. *Ordnungsstatistiken* Für $\mathbf{x} = (x_1,\dots,x_n)^T \in \mathbb{R}^n$ und eine Permutation $\pi \in S_n$ setze $\pi(\mathbf{x}) = \big(x_{\pi(1)},\dots,x_{\pi(n)}\big)$. Wählt man nun zu $\mathbf{x}$ eine Permutation $\pi \in S_n$ mit $x_{\pi(1)} \leq \dots \leq x_{\pi(n)}$, dann setzt man $x_{(i)} = x_{(i,n)} = x_{\pi(i)}$, $i = 1,\dots,n$. Man nennt $x_{(i,n)}$ die i-te Ordnungsstatistik, und den Vektor $\mathbf{x}^{(n)} = \big(x_{(1)},\dots,x_{(n)}\big)^T$ einfach die Ordnungsstatistik.

Wir betrachten eine beliebige Familie $(P_\theta)_{\theta\in\Theta}$ von austauschbaren Verteilungen auf $(\mathbb{R}^n, \mathcal{B}^n)$, d.h. für alle $\theta \in \Theta$ gilt:

$$\forall\ \pi \in S_n,\ B \in \mathcal{B}^n : \quad P_\theta(B) = P_\theta\big(\pi(B)\big).$$

Wir zeigen, dass dann die Ordnungsstatistik $T(\mathbf{x}) = \mathbf{x}^{(n)}$ suffizient für $(P_\theta)_{\theta\in\Theta}$ ist. Dazu betrachten wir den Markov Kern

$$\begin{aligned} P_\cdot : &\quad \mathbb{R}^n \times \mathcal{B}^n \to [0,1], \\ P_\cdot(\mathbf{t}, F) &= \frac{1}{n!}\sum_{\pi\in S_n} 1_F\big(\pi(\mathbf{t})\big), \end{aligned}$$

und zeigen, dass dieser eine reguläre bedingte Verteilung für $\mathbf{Y}|T$ unter jedem P_θ ist

$$\forall F, B \in \mathcal{B}^n : \quad \int_{T^{-1}(B)} P_\cdot\big(T(\mathbf{x}), F\big)\, dP_\theta(\mathbf{x}) = P_\theta\big(T^{-1}(B) \cap F\big).$$

Dazu bemerken wir, dass $\forall\ \mathbf{x} \in \mathbb{R}^n : \sum_{\pi\in S_n} \pi(T(\mathbf{x})) = \sum_{\pi\in S_n} \pi(\mathbf{x})$, daher

$$\begin{aligned} \int_{T^{-1}(B)} P_\cdot\big(T(\mathbf{x}), F\big)\, dP_\theta(\mathbf{x}) &= \frac{1}{n!}\sum_{\pi\in S_n}\int_{\mathbb{R}^n} 1_B(T\mathbf{x})\, 1_F\big(\pi(T(\mathbf{x}))\big)\, dP_\theta(\mathbf{x}) \\ &= \frac{1}{n!}\int_{\mathbb{R}^n} 1_B(T\mathbf{x})\Big(\sum_{\pi\in S_n} 1_F\big(\pi(T(\mathbf{x}))\big)\Big)\, dP_\theta(\mathbf{x}) \\ &= \frac{1}{n!}\int_{\mathbb{R}^n} 1_B(T\mathbf{x})\Big(\sum_{\pi\in S_n} 1_F\big(\pi(\mathbf{x})\big)\Big)\, dP_\theta(\mathbf{x}) \\ &= \frac{1}{n!}\sum_{\pi\in S_n} P_\theta\big(T^{-1}(B) \cap \pi^{-1}(F)\big) \\ &= \frac{1}{n!}\sum_{\pi\in S_n} P_\theta\big(\pi^{-1}\left(T^{-1}(B) \cap F\right)\big) \\ &= P_\theta\big(T^{-1}(B) \cap F\big). \end{aligned}$$

wobei im vorletzten Schritt $\pi^{-1}(T^{-1}B) = T^{-1}B$ benutzt wurde, und im letzten die Austauschbarkeit von P_θ.

3. *Brownsche Bewegung.* Sei $(W_t)_{t\in[0,1]}$ eine standard Brownsche Bewegung. Der Prozess $W_t^0 = W_t - tW_1$ heißt standard Brownsche Brücke, eine stetiger Gauß-Prozess mit Kovarianzfunktion $c(t,s) = t \wedge s - ts$. Da $EW_t^0 W_1 = 0$ für alle $t \in [0,1]$, sind $(W_{t_1}^0, \ldots, W_{t_n}^0)$ und W_1 stets unabhängig, also auch $(W_t^0)_{t\in[0,1]}$ und W_1.

Für einen Drift Parameter $\mu \in \mathbb{R}$ betrachten wir nun die Prozesse $B_t = \mu\, t + W_t$, die zugehörigen Verteilungen sind also um die lineare Funktion $t \mapsto \mu t$ translatierte Wiener Maße. Da für alle μ gilt

$$B_t = (B_t - tB_1) + tB_1 = (W_t - tW_1) + tB_1$$

und für alle μ auch $(W_t - tW_1)$ und tB_1 unabhängig sind, und die bedingte Verteilung von $(B_t)_{t\in[0,1]}$ gegeben $B_1 = x \in \mathbb{R}$ ist für jedes μ die um $t \mapsto xt$ translatierte Verteilung der Brownschen Brücke.

Suffizienz: Dominierte Modelle

Satz 2.12 (*Faktorisierungssatz von Neyman)*
Sei $(P_\theta)_{\theta\in\Theta}$ dominiert durch (das σ-endliche Maß) μ. Dann sind für eine Statistik $T : \mathcal{X} \to \mathcal{S}$ äquivalent:
1. T ist suffizient.
2. Für alle $\theta \in \Theta$ existiert ein messbares $g_\theta : \mathcal{S} \to [0,\infty)$, s.d. für die Likelihoodfunktion folgende Darstellung gilt

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

Beispiele. 1. *Exponentialfamilie* Bei einer Exponentialfamilie

$$\frac{dP_\theta}{d\mu}(x) = C(\theta)\, \exp\big(Q(\theta)^T\, T(x)\big)\, h(x)$$

ist offenbar die natürliche suffiziente Statistik $T(x)$ suffizient für θ bzw. $Q(\theta)$.

Insbesondere: Für $N(\mu 1_d, \sigma^2 I_d)$ auf $(\mathbb{R}^d, \mathcal{B}^d)$, $1_d = (1,\ldots,1)^T$, $\theta = (\mu, \sigma^2)$ ist $T(x) = (\sum_i x_i, \sum_i x_i^2)$ suffizient für $(\mu/\sigma^2, -1/(2\sigma^2))$ und damit auch für (μ, σ^2), und nach Transformation auch für $\tilde{T}(x) = (\bar{x}_n, s^2(x))$, wobei

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i, \qquad s_n^2(x) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{1}{n-1}\big(\sum_i x_i^2 - n\bar{x}_n\big).$$

Im linearen Modell $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$ ist $T(\mathbf{y}) = (\mathbf{y}^T X, \|\mathbf{y}\|_2^2)$ suffizient für $(\boldsymbol{\beta}/\sigma^2, -1/(2\sigma^2))$ und somit auch für $(\boldsymbol{\beta}, \sigma^2)$. Durch Transformation erhält man die Suffizienz von

$$\tilde{T}(\mathbf{y}) = \big(\hat{\boldsymbol{\beta}}_{LS}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})\big), \quad \hat{\boldsymbol{\beta}}_{LS}(\mathbf{y}) = (X^T X)^{-1} X^T \mathbf{y}, \quad \hat{\sigma}^2(\mathbf{y}) = \frac{1}{n-p}\big(\|\mathbf{y}\|^2 - \|P_X \mathbf{y}\|^2\big).$$

2. *Truncation Family* Sei $\phi : \mathbb{R} \to (0, \infty)$ messbar, so dass

$$C(a,b) = \int_a^b \phi(t)\, dt < \infty, \qquad a < b.$$

Für $\theta = (a,b)$ sei

$$f_\theta(x) = \frac{1}{C(a,b)}\, \phi(x) 1_{(a,b)}(x).$$

Ist $Y = (Y_1, \ldots, Y_n)^T$, Y_i u.i.v.$\sim f_\theta$, so gilt

$$g_{Y,\theta}(y) = \prod_{i=1}^n f_\theta(y_i) = \frac{1}{C(a,b)^n}\, 1_{(-\infty,b)}(y_{(n)})\, 1_{(a,\infty)}(y_{(1)}) \prod_{i=1}^n \phi(y_i),$$

also ist $T(y) = (y_{(1)}, y_{(n)})$ suffizient für θ.

Wir kommen zum Beweis des Faktorisierungssatzes von Neyman. Dazu zunächst ein Lemma.

Satz 2.13
Seien P, μ Wahrscheinlichkeitsmaße auf $(\mathcal{X}, \mathcal{F})$ mit $P << \mu$, und sei $T : \mathcal{X} \to \mathcal{S}$ messbar. Dann gilt für alle $F \in \mathcal{F}$

$$E_P\big(1_F|T\big) = \frac{E_\mu\big(1_F \frac{dP}{d\mu}|T\big)}{E_\mu\big(\frac{dP}{d\mu}|T\big)} \qquad P - \text{f.s.}$$

Beweis
Wir zeigen

$$E_P\big(1_F|T\big)\, E_\mu\big(\frac{dP}{d\mu}|T\big) = E_\mu\big(1_F \frac{dP}{d\mu}|T\big) \qquad P - \text{f.s.} \tag{22}$$

Da $\frac{dP}{d\mu} > 0$ P-f.s., so auch $E_\mu\big(\frac{dP}{d\mu}|T\big) > 0$ P-f.s., und die Behauptung folgt.
Beide Seiten von (22) sind $\sigma(T)$-messbar, und es ist

$$E_P\big(1_F|T\big)\, E_\mu\big(\frac{dP}{d\mu}|T\big) = E_\mu\Big(\frac{dP}{d\mu}\, E_P\big(1_F|T\big)|T\Big),$$

da $E_\mu\big(1_F|T\big)$ beschränkt und $\sigma(T)$-messbar ist. Daher

$$\begin{aligned}
\int_{T\in B} E_P\big(1_F|T\big)\, E_\mu\big(\frac{dP}{d\mu}|T\big)\, d\mu &= \int_{T\in B} E_\mu\Big(\frac{dP}{d\mu}\, E_P\big(1_F|T\big)|T\Big)\, d\mu \\
&= \int_{T\in B} \frac{dP}{d\mu}\, E_P\big(1_F|T\big)\, d\mu = \int_{T\in B} E_P\big(1_F|T\big)\, dP \\
&= P\big(\{T \in B\} \cap F\big) = \int_{T\in B} \frac{dP}{d\mu}\, 1_F\, d\mu
\end{aligned}$$

■

Beweis (*des Faktorisierungssatzes 2.12*)
2. $\Rightarrow$ 1.: Wir betrachten für ein festes $\theta_0 \in \Theta$ die bedingte Verteilung $P_{\theta_0}(\cdot|T)$ und zeigen

$$\forall\ \theta \in \Theta, \quad F \in \mathcal{F}: \quad E_\theta(1_{Y\in F}|T) = P_{\theta_0}(F|T).$$

Dazu können wir annehmen, dass das dominierende Maß μ ein Wahrscheinlichkeitsmaß ist, denn ansonsten betrachte für eine disjunkte Zerlegung $\mathcal{X} = \bigcup_n F_n$, $0 < \mu(F_n) < \infty$,

$$\tilde{\mu}(F) = \sum_n \frac{\mu(F \cap F_n)}{2^n \mu(F_n)}.$$

$\tilde{\mu}$ ist ein zu μ äquivalentes Wahrscheinlichkeitsmaß mit

$$\frac{d\mu}{d\tilde{\mu}}(x) = \sum_n 2^n \mu(F_n) 1_{F_n},$$

und wegen

$$\frac{dP_\theta}{d\tilde{\mu}} = \frac{dP_\theta}{d\mu}\,\frac{d\mu}{d\tilde{\mu}}$$

gilt eine Darstellung wie in 2. auch für $\tilde{\mu}$.

Dann ist mit Satz 2.13 und der Darstellung in 2.

$$\begin{aligned} E_\theta(1_{Y\in F}|T) &= \frac{E_\mu(1_{Y\in F}\, g_\theta(T)\, h(Y)|T)}{E_\mu(g_\theta(T)\, h(Y)|T)} \\ &= \frac{E_\mu(1_{Y\in F}\, h(Y)|T)}{E_\mu(h(Y)|T)} \\ &= E_{\theta_0}(1_{Y\in F}|T) = P_{\theta_0}(F|T). \end{aligned}$$

1. $\Rightarrow$ 2. : Sei $Q = \sum_i c_i P_{\theta_i}$ ein priviligiertes dominierendes Maß. Es genügt, die Darstellung in 2. bzgl. Q zu zeigen, für μ folgt sie dann mit dem zusätzlichen Faktor $dQ/d\mu$. Wegen der Suffizienz gilt für die bedingte Verteilung von Y gegeben T unter Q

$$Q(F|T) = \sum_i c_i P_{\theta_i}(F|T) = P_\cdot(F|T), \qquad F \in \mathcal{F}. \tag{23}$$

Auf dem Teilraum $(\mathcal{X}, \sigma(T))$ existiert wegen $P_\theta << Q$ (eingeschränkt auf $(\mathcal{X}, \sigma(T))$) eine $\sigma(T) - \mathcal{B}$-messbare Funktion $f_\theta : \mathcal{X} \to [0, \infty)$ mit

$$\frac{dP_\theta|\sigma(T)}{dQ|\sigma(T)} = f_\theta.$$

Nach dem Faktorisierungslemma existiert ein messbares $g_\theta : \mathcal{S} \to [0, \infty)$ mit $f_\theta = g_\theta \circ T$.

Wir zeigen nun noch, dass $dP_\theta/dQ = g_\theta \circ T$ auf dem gesamten Raum $(\mathcal{X}, \mathcal{F})$ gilt: Für $F \in \mathcal{F}$ ist nämlich mit (23)

$$\begin{aligned} P_\theta(F) &= E_\theta\big(P_\cdot(F|T)\big) = E_\theta\big(Q(F|T)\big) \\ &= E_Q\big(Q_\cdot(F|T) g_\theta \circ T\big) = E_Q\big(E_Q\big(1_F\, g_\theta \circ T|T\big)\big) \\ &= E_Q(1_F\, g_\theta \circ T). \end{aligned}$$

■

Satz 2.14
Sei $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ ein statistisches Modell, und seien $T_1 : \mathcal{X} \to \mathcal{S}_1$ und $T_2 : \mathcal{X} \to \mathcal{S}_2$ Statistiken. Angenommen, es existiert eine messbare Abbildung $\psi : \mathcal{S}_1 \to \mathcal{S}_2$ mit $T_2 = \psi(T_1)$ und T_2 ist suffizient für $(P_\theta)_{\theta\in\Theta}$, so ist auch T_1 suffizient.

Beweis
a. Wir nehmen zunächst an, dass $(P_\theta)_{\theta\in\Theta}$ durch μ dominiert ist. Dann existiert für alle $\theta \in \Theta$ ein $g_\theta : \mathcal{S}_2 \to [0, \infty)$ messbar, s.d.

$$L(\theta, x) = g_\theta \left(T_2(x)\right) h(x) = g_\theta \left(\psi(T_1(x))\right) h(x) \qquad \mu - \text{f.ü},$$

und da $g_\theta \circ \psi : \mathcal{S}_1 \to [0, \infty)$ messbar, folgt dass auch T_1 suffizient für $(P_\theta)_{\theta\in\Theta}$ ist.
b. Allgemein: Angenommen nicht, dann existieren P_{θ_1} und P_{θ_2}, so dass T_2 nicht suffizient für die Familie $\{P_{\theta_i}, i = 1, 2\}$ ist, aber T_1 bleibt für die Teilfamilie suffizient. Diese ist jedoch dominiert (etwa durch $P_{\theta_1} + P_{\theta_2}$), und es ergibt sich ein Widerspruch zu Teil a. ∎

Definition Sei $(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$ ein statistisches Modell, und sei $T : \mathcal{X} \to \mathcal{S}$ eine suffiziente Statistik. Dann heißt T *minimal suffizient*, falls für jede suffziente Statistik $T_1 : \mathcal{X} \to \mathcal{S}_1$ eine messbare Abbildung $\psi : \mathcal{S}_1 \to \mathcal{S}$ existiert mit $T = \psi(T_1)$ P_θ-f.s. für alle $\theta \in \Theta$.

3 Entscheidungstheorie

3.1 Formalisierung eines statistischen Entscheidungsproblems

Sei $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ ein statistisches Modell. Auf Basis des statistischen Modells und einer Stichprobe soll der Statistiker Entscheidungen treffen.

Definition 3.1
Eine *Entscheidungsregel* ist eine messbare Abbildung $\rho : \mathcal{X} \to A$, wobei $(A, \mathcal{A})$ ein messbarer Raum ist, der sogenannte *Aktionsraum*. Jede Funktion $\ell : \Theta \times A \to [0, \infty)$ mit $\ell(\theta, \cdot)$ messbar für alle $\theta \in \Theta$ heißt Verlustfunktion. Das Risiko einer Entscheidungsregel ρ unter der Verlustfunktion ℓ ist die Abbildung

$$R(\cdot, \rho) : \Theta \to [0, \infty], \qquad R(\theta, \rho) = \int_{\mathcal{X}} \ell\big(\theta, \rho(x)\big)\, dP_\theta(x) = E_\theta \ell\big(\theta, \rho\big).$$

Interpretation. 1. $\ell(\theta, a)$: Verlust, falls P_θ, $\theta \in \Theta$ als Verteilung zugrundeliegt, und Entscheidung $a \in A$ getroffen wird.

2. Risiko in θ: mittlerer Verlust, falls für $x \in \mathcal{X}$ Entscheidung $\rho(x)$ getroffen wird und P_θ zugrundeliegt.

Beispiel 3.2
Ist auf Θ eine σ-Algebra $\mathcal{F}_\Theta$ gegeben, und ist $A = \Theta$ gewählt, so heißt eine messbare Abbildung $\rho : \mathcal{X} \to \Theta$ ein Schätzer für $\theta \in \Theta$.
Ist allgemeiner $\gamma : \Theta \to \Gamma$ messbar ein *Parameter* mit Werten in dem messbaren Raum $(\Gamma, \mathcal{G})$, so heißt eine messbare Abbildung $\rho : \mathcal{X} \to \Gamma$ ein Schätzer für $\gamma(\theta) \in \Gamma$.
Ist $\Gamma \subset \mathbb{R}^k$, so betrachtet man als Verlustfunktion häufig den quadratischen Verlust

$$\ell(\theta, a) = \|a - \gamma(\theta)\|_2^2 = \sum_{i=1}^{k} (a_i - \gamma_i(\theta))^2, \qquad a \in \Gamma.$$

Im linearen Modell $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ mit Parameter

$$\theta = (\beta, \sigma^2, F_\epsilon) \subset \mathbb{R}^p \times (0, \infty) \times \mathcal{M}$$

sei $\gamma(\theta) = \boldsymbol{\beta}$. Dann ist für den kleinsten Quadrateschätzer

$$\rho = \hat{\boldsymbol{\beta}}_{LS} : \mathbb{R}^n \to \mathbb{R}^p, \qquad y \mapsto (X^T X)^{-1} X^T y$$

das Risiko bzgl. der quadratischen Verlustfunktion gegeben durch

$$R\big(\theta, \hat{\boldsymbol{\beta}}_{LS}\big) = E_\theta \|\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\|^2 = \sigma^2 \operatorname{tr}\big((X^T X)^{-1}\big).$$

Beispiel 3.3
Ist $A = \{0, 1\}$, so heißt ρ ein Test. Zur Spezialisierung einer Verlustfunktion wird der Parameterbereich disjunkt zerlegt in

$$\Theta = \Theta_0 \cup \Theta_1,$$

$$\Theta_0: \text{ (Null)-Hypothese}, \qquad \Theta_1: \text{ Alternative}.$$

Dann ist

$$\theta \in \Theta_0 \wedge a = 1: \text{ Fehler 1. Art}, \qquad \theta \in \Theta_1 \wedge a = 0: \text{ Fehler 2. Art}.$$

Wir definieren eine Verlustfunktion für $l_0, l_1 > 0$ durch

$$\ell(\theta, a) = l_0\, 1_{\theta\in\Theta_0\wedge a=1} + l_1\, 1_{\theta\in\Theta_1\wedge a=0}.$$

Für das Risiko gilt dann

$$R(\theta, \rho) = \begin{cases} l_0\, P_\theta(\rho = 1), & \theta \in \Theta_0, \\ l_1\, P_\theta(\rho = 0), & \theta \in \Theta_1. \end{cases}$$

Das Risiko setzt sich also zusammen aus den gewichteten Wahrscheinlichkeiten für die Fehler 1. und 2. Art.

Existiert eine suffiziente Statistik für $\theta \in \Theta$, so kann man sich bei Entscheidungsregeln auf solche beschränken, die nur von der suffizienten Statistik abhängen, ohne das Risiko zu erhöhen. Damit dies exakt richtig ist, müssen sogenannte *randomisierte Entscheidungsregeln* zugelassen werden (vg. Abschnitt REF!). Hier stellen wir zunächst folgendes Ergebnis vor.

Überarbeiten: $A \subset \mathbb{R}^p$, Jensen für multivariat, Integrierbarkeit von ρ.

Angenommen, $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}^n, \mathcal{B}^n)$, und $T : \mathbb{R}^n \to \mathbb{R}^k$ sei suffizient für θ. Ist $\rho : \mathbb{R}^n \to A$ eine Entscheidungsregel, so betrachten wir die Entscheidungsregel $\tilde\rho = E.(\rho|T)$, also

$$\tilde\rho(x) = \tilde\rho_0\big(T(x)\big), \qquad \tilde\rho_0(t) = \int_{\mathcal{X}} \rho(x)\, dP.(x|T = t).$$

Satz 3.4 (*von Rao-Blackwell)*
Sei $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}^n, \mathcal{B}^n)$ und $T : \mathbb{R}^n \to \mathbb{R}^k$ suffizient für θ. Sei der Aktionsraum $A \subset \mathbb{R}^p$ konvex und gelte für die Verlustfunktion $\ell : \Theta \times A \to [0, \infty)$: Für jedes $\theta \in \Theta$ ist $\ell(\theta, \cdot) : A \to [0, \infty)$ eine konvexe Funktion. Dann gilt für jede Entscheidungsregel ρ mit zugehöriger Projektion $\tilde\rho$ (s.o.) die Risikoabschätzung

$$R(\theta, \tilde\rho) \le R(\theta, \rho) \qquad \forall\ \theta \in \Theta.$$

Ist $\theta \in \Theta$, so dass $\ell(\theta, \cdot)$ strikt konvex ist und $P_\theta(\tilde\rho = \rho) < 1$, dann gilt sogar $R(\theta, \tilde\rho) < R(\theta, \rho)$.

Für den Beweis erinnern wir an die Jensensche Ungleichung für die bedingte Erwartung: Ist ϕ konvex, Z eine Z.V., so ist $\phi\big(E(Z|\mathcal{F})\big) \le E\big(\phi(Z)|\mathcal{F}\big)$, und daher $E\big(\phi\big(E(Z|\mathcal{F})\big)\big) \le E\big(\phi(Z)\big)$. Ist ϕ strikt konvex und $P(Z = E(Z\mathcal{F})) < 1$, so gilt die strikte Ungleichung für die Erwartungswerte.

Beweis
Wir verwenden die obige Ungleichung mit $\phi = \ell(\theta, \cdot)$, $Z = \rho$, $\sigma(T) = \mathcal{F}$, und erhalten

$$R(\theta, \tilde\rho) = E_\theta\big(\ell(\theta, E.(\rho|T))\big) \le E_\theta\big(\ell(\theta, \rho)\big) = R(\theta, \rho),$$

bzw. die strikte Ungleichung unter den zusätzlichen Voraussetzungen. ∎

Bemerkung. 1. In einem Schätzproblem sind die Voraussetzungen des Satzes ($\Theta \subset \mathbb{R}^p$ konvex, quadratische Verlustfunktion konvex) oft erfüllt, jedoch nicht in Testproblemen.
2. Eine Entscheidungsregel ρ' heißt besser als ρ, falls gilt $R(\theta, \rho') \leq R(\theta, \rho)$ für alle $\theta \in \Theta$, und falls ein $\theta \in \Theta$ existiert mit $R(\theta, \rho') < R(\theta, \rho)$. In der Situation des Satzes ist $\tilde{\rho}$ oftmals besser als ρ.
3. Ist $S = \psi \circ T$ ebenfalls suffizient, wobei $\psi : \mathbb{R}^k \to \mathbb{R}^l$ messbar, dann ist das Risiko von $\tilde{\rho}_S = E(\rho|S) \leq$ Risiko von $\tilde{\rho}_T = E(\rho|T)$ (obiges Argument, da $\sigma(S) \subset \sigma(T)$). Daher sollte man auf mininmal-suffiziente Statistiken bedingen.
4. Anwendung des Satzes von Rao-Blackell: Unverfälschtes Schätzen.

3.2 Optimalität von Entscheidungsregeln

Bemerkung. Bestmöglich wäre eine Entscheidungsregel ρ, die überall ein höchstens ebenso großes Risiko hat wie eine beliebige andere Entscheidungsregel: $\forall\ \theta \in \Theta,\ \rho' :\ R(\theta, \rho) \leq R(\theta, \rho')$.

Solche Entscheidungsregeln existieren jedoch in der Regel nicht. Betrachten wir ein Schätzproblem mit quadratischer Verlustfunktion. Ist $\theta_0 \in \Theta$, und den konstanten Schätzer $\rho_{\theta_0}(x) = \theta_0$ für alle $x \in \mathcal{X}$. Dann ist $R(\theta_0, \rho_{\theta_0}) = E_{P_{\theta_0}} \|\rho_{\theta_0} - \theta_0\|_2^2 = 0$, und für einen Schätzer ρ gilt $R(\theta_0, \rho) = 0$ nur dann, falls $\rho = \theta_0\ P_{\theta_0}$-f.s.

Auswege.

1. Einschränken der Klasse der Entscheidungsregeln. Beispiel: Schätzen von β im linearem Modell bzgl. quadratischer Verlustfunktion. Dann hat $\hat{\beta}_{LS}$ kleinstes Risiko unter allen linearen, unverzerrten Schätzern. Kommen darauf später zurück.

2. Schwächere Optimalitätsbegriffe. Hier weiter verfolgen.

Eine naheliegende Einschränkung ist folgende.

Definition 3.5
Eine Entscheidungsregel ρ heißt zulässig, falls es keine Entscheidungsregel ρ' gibt, die besser ist als ρ.

Beispiel. Konstante Schätzer bei quadratischer Verlustfunktion.

Definition 3.6
Sei $(\Theta, \mathcal{F}_\Theta)$ ein messbarer Raum. Eine a-priori-Verteilung π von θ ist ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathcal{F}_\Theta)$. Angenommen, die Verlustfunktion $\ell : \Theta \times A \to [0, \infty)$ sei $\mathcal{F}_\Theta \otimes \mathcal{A} - \mathcal{B}$-messbar, und für alle $F \in \mathcal{F}$ sei $\theta \mapsto P_\theta(F)$ messbar. Dann ist das zu π-assoziierte *Bayesrisiko* der Entscheidungsregel ρ definiert durch

$$\begin{aligned} R_\pi(\rho) &= \int_\Theta R(\theta, \rho)\, d\pi(\theta) \\ &= \int_\Theta \int_{\mathcal{X}} \ell\big(\theta, \rho(x)\big)\, dP_\theta(x)\, d\pi(\theta). \end{aligned}$$

ρ heißt *Bayesregel* oder *Bayes optimal* bzgl. π, falls gilt

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

wobei das Infimum über alle Entscheidungsregeln genommen wird.

Bemerkung. 1. Nach den Voraussetzungen ist $R(\cdot, \rho)$ messbar (und nicht-negativ), also ist $R_\pi(\rho)$ wohldefiniert.
2. $R_\pi(\rho)$ ist das mit π gemittelte Risiko von ρ.

A-posteriori Verteilungen

Sei $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ ein statistisches Modell, $(\Theta, \mathcal{F}_\Theta)$ ein messbarer Raum, und für alle $F \in \mathcal{F}$ sei $\theta \mapsto P_\theta(F)$ messbar. Sei π eine a-priori-Verteilung von θ. Setze $\Omega = \mathcal{X} \times \Theta$, $\mathcal{F}_\Omega = \mathcal{F} \otimes \mathcal{F}_\Theta$, und für $A \in \mathcal{F}_\Omega$ setze

$$\tilde{P}(A) = \int_\Theta \int_\mathcal{X} 1_A(x,\theta)\, dP_\theta(x)\, d\pi(\theta),$$

dann ist $(\Omega, \mathcal{F}_\Omega, \tilde{P})$ ein Wahrscheinlichkeitsraum. Seien

$$X : \Omega \to \mathcal{X},\ X(x,\theta) = x, \qquad T : \Omega \to \Theta,\ T(x,\theta) = \theta$$

die Koordinatenprojektionen.

Die bedingte Verteilung von T gegeben X, $P_{T|X}$, heißt die *a-posteriori-Verteilung* von θ gegeben X (bei a-priori-Verteilung π).
Interpretation: Ausgehend von a-priori-Verteilung π über θ, was kann man aus Beobachtung $X = x$ über θ lernen bzw. wie verändert sich die Annahme über die Verteilung des Parameters θ.

Die a-posteriori-Verteilung ist das zentrale Objekt der Bayes-Statistik.

Rechenregeln

1. Hat $\tilde{P}$ die Dichte $f(x,\theta)$ bzgl. des Produktmaßes $\mu \otimes \nu$, wobei μ σ-endlich auf $(\mathcal{X}, \mathcal{F})$ und ν σ-endlich auf $(\Theta, \mathcal{F}_\Theta)$, so hat X die Dichte

$$f_X(x) = \int_\Theta f(x,\theta)\, d\nu(\theta)$$

bzgl. μ, und die bedingte Verteilung von T gegeben $X = x$ hat die Dichte

$$f_{T|X}(\theta|x) = \frac{f(x,\theta)}{f_X(x)} \qquad \text{bzgl.}\nu,$$

für $\tilde{P}_X$-f.a. $x \in \mathcal{X}$.

2. Gilt speziell:
(P_θ) wird dominiert bzgl. μ, Dichte $f(x|\theta)$,
π wird dominiert durch ν, Dichte $f_T(\theta)$, so ist $f(x|\theta)f_T(\theta)$ die Dichte von (X,T) bzgl. $\mu \otimes \nu$, und die bedingte Dichte ist gegeben durch

$$f_{T|X}(\theta|x) = \frac{f(x|\theta)f_T(\theta)}{\int_\Theta f(x|\theta)f_T(\theta)\, d\nu(\theta)}.$$

Beispiel. Angenommen,

$$\begin{pmatrix} X \\ T \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_T \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_{XT} \\ \Sigma_{TX} & \Sigma_T \end{pmatrix}\right),$$

wobei $\Sigma'_{XT} = \Sigma_{TX}$. Dann ist

$$\begin{aligned} T|X = x &\sim N\big(\mu_{T|X=x}, \Sigma_{T|X}\big), \\ \mu_{T|X=x} &= \mu_T + \Sigma_{TX}\Sigma_X^{-1}(x - \mu_X), \\ \Sigma_{T|X} &= \Sigma_T - \Sigma_{TX}\Sigma_X^{-1}\Sigma_{XT} \end{aligned}$$

Beispiel. Wir spezialisieren das obige Beispiel. Angenommen, $\mathbf{X} \sim N(\mu, \sigma_0^2 I_d)$, wobei $\mu \in \mathbb{R}^d$ unbekannt und $\sigma_0 > 0$ bekannt. Als priori-Verteilung für μ betrachten wir $\pi \sim N(0, \sigma^2 I_d)$. Dann können wir schreiben

$$X = \mu + \epsilon, \qquad \mu \sim N(0, \sigma^2 I_d), \quad \epsilon \sim N(0, \sigma_0^2 I_d),$$

wobei μ und ϵ unabhängig sind. Dann ist

$$(X, \mu)^T \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (\sigma^2 + \sigma_0^2) I_d & \sigma^2 I_d \\ \sigma^2 I_d & \sigma^2 I_d \end{pmatrix}\right),$$

und daher nach obiger Formel

$$\mu|X = x \sim N\Big(\frac{\sigma^2}{\sigma^2 + \sigma_0^2}\, x, \frac{\sigma^2\, \sigma_0^2}{\sigma^2 + \sigma_0^2}\Big).$$

Beispiel. *(Binomialverteilung/Betaverteilung)* Sei $X \sim Bin(n, p)$,

$$P_p(X = x) = \binom{n}{x} p^x\, (1 - p)^{n-x},$$

und betrachte als a-priori-Verteilung für $p \in (0, 1)$ die Beta-Verteilung $B(a, b)$, $a, b > 0$, also die Dichte

$$f_{a,b}(p) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1 - p)^{b-1}\, 1_{0<p<1}.$$

Die marginale Verteilung von X ist

$$\begin{aligned} P(X = x) &= \int_0^1 \binom{n}{x} p^x\, (1 - p)^{n-x} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1 - p)^{b-1}\, dp \\ &= \binom{n}{x} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a+x-1}(1 - p)^{n-x+b-1}\, dp \\ &= \binom{n}{x} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a + x)\Gamma(n + b - x)}{\Gamma(n + a + b)}, \end{aligned}$$

da das Integral in der zweiten Zeile über die nicht normierte Dichte der Beta-Verteilung geht. Daher ist die a-posteriori-Dichte gegeben durch

$$\begin{aligned} f_{p|X}(p|x) &= \frac{\binom{n}{x} p^x\, (1 - p)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1 - p)^{b-1}\, 1_{0<p<1}}{\binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(n+b-x)}{\Gamma(n+a+b)}} \\ &= \frac{\Gamma(n + a + b)}{\Gamma(a + x)\Gamma(n + b - x)}\, p^{a+x-1}(1 - p)^{n-x+b-1}\, 1_{0<p<1}, \end{aligned}$$

also ist $p|X = x \sim B(a + x, n + b - x)$. Den Nenner hätte man nicht berechnen müssen, da normiert, hätte es der Normierungsfaktor der Beta-Dichte sein müssen.

In den beiden obigen Beispielen sind a-priori und a-posteriori-Verteilung in der gleichen Verteilungsklasse, man spricht auch von konjugierten priors.

Bayes Regeln

Sei nun die Verlustfunktion $\ell : \Theta \times A \to [0, \infty)$ $\mathcal{F}_\Theta \otimes \mathcal{A} - \mathcal{B}$-messbar, und sei ρ eine Entscheidungsregel. Dann ist

$$\begin{aligned} R_\pi(\rho) &= \int_\Theta \int_{\mathcal{X}} \ell\big(\theta, \rho(x)\big)\, dP_\theta(x)\, d\pi(\theta). \\ &= E_{\tilde{P}}\left(\ell(T, \rho(X))\right). \end{aligned}$$

Satz 3.7
Gilt in obiger Situation für die Entscheidungsregel ρ

$$\rho(X) \in \operatorname{argmin}_{a \in A} E_{\tilde{P}}\left(\ell(T, a)|X\right) \qquad \tilde{P} - \text{f.s.},$$

also

$$E_{\tilde{P}}\left(\ell(T, \rho(x))|X = x\right) \leq E_{\tilde{P}}\left(\ell(T, a)|X = x\right) \qquad \forall\ a \in A$$

für $\tilde{P}_X$-f.a. $x \in \mathcal{X}$, dann ist ρ eine Bayesregel (für π).

Bemerkung. In einem Schätzproblem mit $\Theta \subset \mathbb{R}^p$, $A = \mathbb{R}^p$ sowie der quadratischen Verlustfunktion $\ell(T, a) = \|T - a\|_2^2$ ist daher der *a-posteriori-Erwartungswert* (posterior mean) $E_{\tilde{P}}(T|X = x)$ eine Bayesregel.

Beweis
Sei ρ' eine beliebige Entscheidungsregel. Dann ist

$$\begin{aligned} R_\pi(\rho') &= E_{\tilde{P}}\big(\ell(T, \rho'(X))\big) \\ &= E_{\tilde{P}}\Big(E_{\tilde{P}}\big(\ell(T, \rho'(X)|X\big)\Big) \\ &\geq E_{\tilde{P}}\Big(E_{\tilde{P}}\big(\ell(T, \rho(X)|X\big)\Big) \\ &= R_\pi(\rho). \end{aligned}$$

■

Beispiel *(Binomial-Betaverteilung)* Für die Beta-Verteilung $T \sim B(a, b)$ ist $ET = a/(a + b)$. Da $T|X = x \sim B(a + x, n + b - x)$, ist der posterior mean

$$E(T|X = x) = \frac{a + x}{n + b + a},$$

also die Bayesregel bzgl. des quadratischen Risikos.

Beispiel *(normal-normal)* Es ist $X \frac{\sigma^2}{\sigma^2 + \sigma_0^2}$ posterior mean, also Bayesregel bzgl. des quadratischen Risikos.

Sind allgemeiner $X_1, \ldots, X_n$ u.v.i. $\sim N(\mu, I_d)$, so ist $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$ suffizient für μ. Betrachten wir die quadratische Verlustfunktion für μ, so muss nach dem Satz 3.4 von Rao-Blackwell jede Bayesregel (bzgl. einer beliebigen a-posteriori-Verteilung) nur in $\bar{X}_n$ formulierbar sein, also Bayesregel bzgl. der Beobachtung $\bar{X}_n$ sein. Betrachten wir $\pi \sim N(0, \sigma^2 I_d)$, so ist da $\bar{X}_n \sim N(\mu, I_d/n)$, $\bar{X}_n \frac{\sigma^2}{\sigma^2+1/n} = \bar{X}_n \frac{n}{\sigma^{-2}+n}$ Bayes-Schätzer für μ. Beachte die Shrinkage-Struktur des Schätzers gegebenüber $\bar{X}_n$.

Minimax-Regeln

Definition 3.8
Eine Entscheidungsregel ρ heißt minimax, falls gilt

$$\sup_{\theta\in\Theta} R(\theta, \rho) = \inf_{\rho'} \sup_{\theta\in\Theta} R(\theta, \rho').$$

Dabei wird das Infimum über die Klasse aller Entscheidungsregeln genommen. Der Ausdruck auf der rechten Seite heißt das minimax-Risiko des Entscheidungsproblems.

Bemerkung. Eine minimax-Regel minimiert also das maximale Risiko, eine pessimistische Sichtweise. Man kann auch die Klasse der Entscheidungsregeln, über die das Infimum genommen wird, einschränken, dann spricht man von minimax bzgl. der betrachteten Klasse von Entscheidungsregeln.

Lemma 3.9
1. Für eine Entscheidungsregel ρ ist

$$\sup_{\theta\in\Theta} R(\theta, \rho) = \sup_{\pi} R_\pi(\rho).$$

2. Die Entscheidungsregel ρ ist minimax genau dann, wenn gilt

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

Beweis
Zu 1.: $\forall\ \pi$ ist

$$R_\pi(\rho) = \int_\Theta R(\theta, \rho)\, d\pi(\theta) \leq \sup_{\theta\in\Theta} R(\theta, \rho),$$

also " $\geq$". Für " $\leq$" beachte, dass $\pi = \delta_{\theta_0}$, $\theta_0 \in \Theta$, als priori-Verteilungen zugelassen sind.
2. ergibt sich dann unmittelbar aus 1. ∎

Minimax Regeln haben häufig ein konstantes Risiko. Genauer gilt z.B.

Satz 3.10
Sei ρ eine Entscheidungsregel mit $R(\theta, \rho) = C$ für alle $\theta \in \Theta$ für ein $C > 0$. Dann ist ρ minimax, falls (mindestens) eine der beiden folgenden Bedingungen erfüllt ist:
1. ρ ist Bayesregel bzgl. einer beliebigen a-priori-Verteilung π_0,
2. ρ ist zulässig.

Beweis
Zu 1.: Da $R(\theta,\rho) = C$ für alle $\theta \in \Theta$, ist auch $R_\pi(\rho) = C$ für alle a-priori-Verteilungen π. Ist daher ρ' eine weitere Entscheidungsregel, so folgt

$$\begin{aligned}\sup_{\theta\in\Theta} R(\theta,\rho) = C &= R_{\pi_0}(\rho)\\ &\le R_{\pi_0}(\rho') \le \sup_\pi R_\pi(\rho') = \sup_{\theta\in\Theta} R(\theta,\rho')\end{aligned}$$

nach obigem Lemma.
Zu 2.: Gilt für ρ'

$$\sup_{\theta\in\Theta} R(\theta,\rho') \le \sup_{\theta\in\Theta} R(\theta,\rho) = C,$$

also

$$R(\theta,\rho') \le C = R(\theta,\rho) \qquad \forall\ \theta\in\Theta,$$

so gilt wegen der Zulässigkeit von ρ im zweiten Display überall die Gleichheit, und daher auch im ersten Display. ∎

Beispiel. *(Binomialverteilung)* Unter $B(a,b)$ ist der Bayes-Schätzer für p

$$T_{a,b}(x) = \frac{x+a}{n+b+a}.$$

Für das quadratische Risiko gilt die Bias-Varianz-Zerlegung

$$E_p\big(T_{a,b}-p\big)^2 = E_p\big(T_{a,b} - E_p\,T_{a,b} + E_p\,T_{a,b} - p\big)^2 = \operatorname{Var} T_{a,b} + \big(E_p\,T_{a,b} - p\big)^2$$

Da $E_pX = np$, $\operatorname{Var}X = p(1-p)$, so ist

$$\begin{aligned}E_p\,T_{a,b} &= \frac{np + a - p(n+b+a)}{n+b+a},\\ \operatorname{Var} T_{a,b} &= \frac{np\,(1-p)}{(n+b+a)^2}.\end{aligned}$$

Somit folgt

$$R(p,T_{a,b}) = E_p\big(T_{a,b}-p\big)^2 = \frac{np\,(1-p) + \big(a\,(1-p) - pb\big)^2}{(n+b+a)^2}.$$

Im Zähler hat man ein quadratisches Polynom in p, damit Risiko konstant, müssen die Koeffizienten von p^2 und p verschwinden, also

$$(a+b)^2 + n = 0, \qquad n - 2(a+b) = 0,$$

welches man zu $a = b = \sqrt{n}/2$ auflöst. Der sich ergebende Minimax-Schätzer ist

$$T(x) = \frac{x + \sqrt{n}/2}{n+\sqrt{n}}, \qquad R(p,T) = \frac{1}{4n}\,\frac{1}{(1+1/\sqrt{n})^2}.$$

Zum Vergleich: Der Standard-Schätzer X/n hat Risiko

$$R(p,X/n) = \frac{1}{4n}\,4p(1-p),$$

und $p(1-p)$ hat Max. bei $p = 1/2$, mit Wert $1/(4n)$. Allerdings hat X/n für die meisten p ein kleineres Risiko als T.

Beispiel. *(Normalverteilung)* Seien $X_1, \ldots, X_n$ u.i.v. $\sim N(\mu, I_d)$, so gilt für den Schätzer $\bar{X}_n$ bzgl. des quadratischen Risikos

$$R(\mu, \bar{X}_n) = E_\mu \|\bar{X}_n - \mu\|_2^2 = \sum_{i=1}^{d} E_\mu (\bar{X}_{n,i} - \mu_i)^2 = \frac{d}{n}.$$

Wir zeigen, dass $\bar{X}_n$ minimax ist. Wir können zwar $\bar{X}_n$ nicht direkt als Bayesschätzer realisieren, aber durch Bayesschätzer geeignet das Risiko approximieren. Sei dazu $\pi_\sigma \sim N(0, \sigma^2 I_d)$, dann ist $T_\sigma = \bar{X}_n \frac{n}{n+\sigma^{-2}}$ Bayesschätzer bzgl. π_σ. Für das Risiko gilt

$$\begin{aligned} R(\mu, T_\sigma) &= E_\mu \|T_\sigma - \mu\|^2 \\ &= E_\mu \|T_\sigma - E_\mu T_\sigma\|^2 + \|E_\mu T_\sigma - \mu\|^2 \\ &= \frac{n^2}{(n+\sigma^{-2})^2} \frac{d}{n} + \|\frac{n}{n+\sigma^{-2}}\mu - \mu\|_2^2 \\ &= \frac{n^2}{(n+\sigma^{-2})^2} \frac{d}{n} + \frac{\sigma^{-4}}{(n+\sigma^{-2})^2} \|\mu\|_2^2 \end{aligned}$$

und daher für das Bayes-Risiko

$$R_{\pi_\sigma}(T_\sigma) = \frac{n^2}{(n+\sigma^{-2})^2} \frac{d}{n} + \frac{\sigma^{-4}}{(n+\sigma^{-2})^2} d\sigma^2 \to \frac{d}{n}, \quad \sigma \to \infty.$$

Damit folgt die Minimax-Eigenschaft von $\bar{X}_n$ aus folgendem Satz.

Satz 3.11
Sei ρ eine Entscheidungsregel mit konstantem Risiko $R(\theta, \rho) = C$ für alle $\theta \in \Theta$ für ein $C > 0$. Existiert eine Familie $(\pi_n)_n$ von a-priori-Verteilungen für θ mit zugehörigen Bayesregeln $(\rho_n)_n$, für die $\lim_n R_{\pi_n}(\rho_n) = C$ gilt, so ist ρ eine Minimaxregel.

Beweis
Für jede Entscheidungsregel ρ' gilt

$$\begin{aligned} \inf_{\rho'} \sup_{\theta \in \Theta} R(\theta, \rho') &= \inf_{\rho'} \sup_{\pi} R_\pi(\rho') \geq \inf_{\rho'} \sup_{n} R_{\pi_n}(\rho') \\ &\geq \sup_n \inf_{\rho'} R_{\pi_n}(\rho') = \sup_n R_{\pi_n}(\rho_n) \\ &\geq \lim_n R_{\pi_n}(\rho_n) = C. \end{aligned}$$

∎

Wir werden im nächsten Abschnitt die überraschende Tatsache einsehen, dass $\bar{X}_n$ im Beispiel der Normalverteiltung für $d \geq 3$ nicht zulässig ist.

3.3 Zulässigkeit und das Stein Phänomen

Seien $X_1, \ldots, X_n$ u.i.v.$\sim N(\mu, \sigma_0^2 I_d)$, mit $\mu \in \mathbb{R}^d$ unbekannt und $\sigma_0^2 > 0$ bekannt. Wir betrachten die Schätzung von μ bzgl. der quadratischen Verlustfunktion

$$\ell(\mu, \hat{\mu}) = \|\hat{\mu} - \mu\|^2.$$

Wegen der Konvexität und des Satzes von Rao-Blackwell genügt es, Schätzer in der suffizienten Statistik

$$\bar{X}_n = \frac{1}{n} \sim N(\mu, \sigma_0^2 I_d/n)$$

zu betrachten. Für theoretische Überlegungen und Beweise können wir daher stets $n = 1$ annehmen und beobachten nur $X \sim N(\mu, \sigma_0^2 I_d)$.

Wir haben bereits gesehen, dass $\bar{X}_n$ in obiger Situation einige gute Eigenschaften hat: Der Schätzer ist minimax, und Limes von Bayes Schätzern. Wir zeigen hier die erstaunliche Tatsache, dass $\bar{X}_n$ für Dimension $d \geq 3$ nicht zulässig ist. Wir beginnen mit der Zulässigkeit für $d = 1$.

Satz 3.12
Seien $X_1, \ldots, X_n$ u.i.v. Zufallsvariablen mit $X_i \sim\sim N(\mu, \sigma_0^2)$, wobei σ_0^2 bekannt ist. Dann ist $\bar{X}_n$ zulässig für μ bzgl. des quadratischen Risikos.

Beweis
o.E. $n = 1$, sei $\hat{\mu}$ ein quadratintegrierbarer Schätzer. Dann ist das Risiko

$$R(\mu, \hat{\mu}) = \frac{1}{\sqrt{2\pi}\sigma_0} \int_{\mathbb{R}} \big(\hat{\mu}(x) - \mu\big)^2 \, e^{-(x-\mu)^2/(2\sigma_0^2)} \, dx$$

stetig als Funktion von μ. Angenommen, $\hat{\mu}$ ist besser als X, also $R(\mu, \hat{\mu}) \leq \sigma_0^2$ für alle $\mu \in \mathbb{R}$ und es existiert μ_0 mit $R(\mu_0, \hat{\mu}) < \sigma_0^2$. Wegen der Stetigkeit der Risikofunktion existieren $\epsilon, \delta > 0$ mit $R(\mu, \hat{\mu}) \leq \sigma_0^2 - \epsilon$, $|\mu - \mu_0| \leq \delta$. Das Bayesrisiko von $\hat{\mu}$ unter der a-priori-Verteilung $\pi_\sigma = N(0, \sigma^2)$ ist daher

$$R_{\pi_\sigma}(\hat{\mu}) \leq \sigma_0^2 - \epsilon \int_{\mu_0-\delta}^{\mu_0+\delta} \frac{1}{\sqrt{2\pi}\sigma} \, e^{-\mu^2/(2\sigma^2)} \, d\mu,$$

und für $\sigma^2 \geq 1$

$$\sigma_0^2 - R_{\pi_\sigma}(\hat{\mu}) \geq \frac{\epsilon}{\sigma} \int_{\mu_0-\delta}^{\mu_0+\delta} \frac{1}{\sqrt{2\pi}} \, e^{-\mu^2/(2)} \, d\mu = C/\sigma. \tag{24}$$

Andererseits gilt für den Bayes Schätzer $\hat{\mu}_\sigma = \sigma^2/(\sigma^2 + \sigma_0^2)X$ zu π_σ:

$$R(\mu, \hat{\mu}) = \frac{\sigma_0^2 \sigma^2}{\sigma^2 + 2 + \sigma_0^4} + \frac{\sigma_0^4}{(\sigma_0^2 + \sigma^2)^2} \mu^2,$$

also für $\sigma \to \infty$

$$\sigma_0^2 - R_{\pi_\sigma}(\hat{\mu}_\sigma) = \sigma_0^2 - \frac{\sigma_0^2 \sigma^2}{\sigma^2 + 2 + \sigma_0^4} - \frac{\sigma_0^4 \, \sigma^2}{(\sigma_0^2 + \sigma^2)^2} = \frac{2\sigma_0^2 + \sigma_0^6}{\sigma^2 + 2 + \sigma_0^4} - \frac{\sigma_0^4 \, \sigma^2}{(\sigma_0^2 + \sigma^2)^2} = O(\sigma^{-2}).$$

Für große σ ist dies im Widerspruch zu (24) und der Optimalität des Bayes schätzers bzgl. des Bayes Risikos. ∎

Stein Schätzer

Für die a-priori-Verteilung $\pi \sim N(0, \sigma^2 I_d)$ ist $X \sim N(0, (\sigma^2 + \sigma_0^2) I_d)$ und wir erhalten als Bayes-Schätzer $\frac{\sigma^2}{\sigma^2+\sigma_0^2} X$. In einem *empirischen Bayes* Ansatz ersetzt man nun den Hyperparameter σ^2 durch einen Schätzer (unter der Verteilungsannahme $X \sim N(0, (\sigma^2 + \sigma_0^2) I_d)$), ein unverzerrter Schätzer ist dann gegeben durch

$$\hat{\sigma}^2 = \frac{\|X\|^2}{d} - \sigma_0^2.$$

Setzen wir diesen in den Bayesschätzer ein, erhalten wir den *Stein Schätzer*

$$\hat{\mu}_S = X\Big(1 - \frac{\sigma_0^2 d}{\|X\|^2}\Big),$$

bzw. die Stichprobenversion

$$\hat{\mu}_{S,n} = \bar{X}_n\Big(1 - \frac{\sigma_0^2 d}{n\,\|\bar{X}_n\|^2}\Big).$$

Der James-Stein-Schätzer ist für $d \geq 3$ gegeben durch

$$\hat{\mu}_{JS,n} = \bar{X}_n\Big(1 - \frac{\sigma_0^2 (d-2)}{n\,\|\bar{X}_n\|^2}\Big).$$

Es gilt nun folgendes überraschendes Ergebnis.

Satz 3.13
Seien $X_1, \ldots, X_n$ u.i.v.$\sim N(\mu, \sigma_0^2 I_d)$, mit $\mu \in \mathbb{R}^d$ unbekannt und $\sigma_0^2 > 0$ bekannt. Dann gilt für $d \geq 3$

$$\forall\ \mu \in \mathbb{R}^d : \quad E_\mu \|\hat{\mu}_{JS,n} - \mu\|^2 < E_\mu \|\bar{X}_n - \mu\|^2,$$

und für $d \geq 4$ die analoge Aussage für den Stein Schätzer. Insbesondere ist $\bar{X}_n$ für $d \geq 3$ nicht zulässig.

Für den Beweis beginnen wir mit

Lemma 3.14 (*Steinsches Lemma)*
Sei $Y \sim N(\mu, \sigma^2 I_d)$ und sei $f : \mathbb{R}^d \to \mathbb{R}$ messbar mit
i.) $f(\cdot, y_2, \ldots, y_d)$ ist absolut stetig in y_1 für Lebesgue λ^{d-1}-f.a. Werte $(y_2, \ldots, y_d)$,
ii.) Für die Ableitung gelte

$$E\big|\frac{\partial f}{\partial y_1}(Y)\big| < \infty.$$

Dann gilt

$$E\big((\mu_1 - Y_1) f(Y)\big) = -\sigma^2\, E\big(\frac{\partial f}{\partial y_1}(Y)\big).$$

Beweis
1. Schritt: Reduktion auf $\mu = 0$, $\sigma^2 = 1$. Gehen über zu $Z = (Y - \mu)/\sigma \sim N(0, I_d)$, $\bar{f}(z) = f(\sigma z + \mu)$. Dann erfüllt $\bar{f}$ die Bedingung des Satzes bzgl. Z. Eine Substitution zeigt

$$E\big((\mu_1 - Y_1)/\sigma\, f(Y)\big) = -E\big(Z_1\, \bar{f}(Z)\big),$$

sowie wegen $\partial \bar{f}/\partial z_1(z) = \sigma\, \partial f/\partial y_1(\sigma z + \mu)$, und daher mit Substitution

$$E\big(\frac{\partial \bar{f}}{\partial z_1}(Z)\big) = \sigma\, E\big(\frac{\partial f}{\partial y_1}(Y)\big).$$

Gilt die Aussage daher für normierte Zufallsvektoren (wie Z), so folgt sie auch allgemein:

$$E\big((\mu_1 - Y_1)f(Y)\big) = -\sigma E\big(Z_1\, \bar{f}(Z)\big) = -\sigma\, E\big(\frac{\partial \bar{f}}{\partial z_1}(Z)\big) = -\sigma^2\, E\big(\frac{\partial f}{\partial y_1}(Y)\big).$$

2. Schritt: o.E. $Y \sim N(0, I_d)$, zeige

$$E\big(Y_1 f(Y)\big) = E\big(\frac{\partial f}{\partial y_1}(Y)\big).$$

Dazu genügt es zu zeigen, dass für λ^{d-1}-f.a. $y_2, \ldots, y_d$ gilt

$$E\big(Y_1 f(Y) | Y_2 = y_2, \ldots, Y_d = y_d\big) = E\big(\frac{\partial f}{\partial y_1}(Y) | Y_2 = y_2, \ldots, Y_d = y_d\big).$$

Fixiere $y_2, \ldots, y_d$ derart, dass $h(u) = f(u, y_2, \ldots, y_d)$ absolut stetig ist und $\int |h(u)| e^{-u^2/2}\, du < \infty$ gilt (dies ist nach den Voraussetzungen für λ^{d-1}-f.a. $y_2, \ldots, y_d$ erfüllt). Dann bleibt zu zeigen, dass

$$\int_{\mathbb{R}} h'(u)\, e^{-u^2/2}\, du = \int_{\mathbb{R}} u\, h(u)\, e^{-u^2/2}\, du, \tag{25}$$

wobei die Existenz des rechten Integrals Teil der Aussage ist. Dies ist grundsätzlich eine partielle Integration, vgl. Elstrodt, denn es gilt für $a, b \in \mathbb{R}$, $a < b$

$$\int_a^b h'(u)\, e^{-u^2/2}\, du = h(u)\, e^{-u^2/2}\Big|_a^b + \int_a^b u\, h(u)\, e^{-u^2/2}\, du.$$

Wenn man nun voraussetzen würde, dass $h(u)\, e^{-u^2/2} \to 0$, $u \to +/-\infty$, und dass $u\, h(u)\, e^{-u^2/2}$ integrierbar, so würde (25) folgen. Ein Trick zeigt (25) direkt (woraus dann auch $h(u)\, e^{-u^2/2} \to 0$, $u \to +/-\infty$ folgt).
Schritt 3.: Nachweis von (25). Es ist

$$e^{-u^2/2} = \int_u^\infty z\, e^{-z^2/2}\, dz = -\int_{-\infty}^u z\, e^{-z^2/2}\, dz.$$

Wir zerlegen

$$\int_{\mathbb{R}} h'(u)\, e^{-u^2/2}\, du = \int_0^\infty h'(u) \int_u^\infty z\, e^{-z^2/2}\, dz\, du + \int_{-\infty}^0 h'(u) \int_{-\infty}^u (-z)\, e^{-z^2/2}\, dz\, du.$$

In beiden Summanden können wir nun den Satz von Fubini anwenden (die Gleichung gilt analog auch für $|h'(u)|$, und dann sind die Integranden in beiden Termen nicht-negativ), und erhalten

$$\begin{aligned}
\int_{\mathbb{R}} h'(u)\, e^{-u^2/2}\, du &= \int_0^\infty z\, e^{-z^2/2} \int_0^u h'(u)\, du\, dz + \int_{-\infty}^0 (-z)\, e^{-z^2/2} \int_{-\infty}^u h'(u)\, du\, dz \\
&= \int_0^\infty z\, e^{-z^2/2} \big(h(z) - h(0)\big)\, dz - \int_{-\infty}^0 (-z)\, e^{-z^2/2} \big(h(0) - h(z)\big)\, dz \\
&= \int_{\mathbb{R}} u\, h(u)\, e^{-u^2/2}\, du.
\end{aligned}$$

■

Wir betrachten den Fall $n = 1$. Bevor wir in den Beweis von Satz 3.13 einsteigen, noch einige Vorbetrachtungen. Für einen Schätzer der Form

$$\hat{\mu} = g(X)\, X,$$

wobei $g : R^d \to \mathbb{R}$ messbar und $E_\mu\|\hat{\mu}\|^2 < \infty$ für alle $\mu \in \mathbb{R}^d$, gilt

$$E_\mu\|\hat{\mu} - \mu\|^2 = E_\mu\|\hat{\mu} - X\|^2 + 2E_\mu\big\langle X - \hat{\mu}, \mu - X\big\rangle + E_\mu\|X - \mu\|^2.$$

Dabei ist

$$E_\mu\big\langle X - \hat{\mu}, \mu - X\big\rangle = \sum_{i=1}^{d} E_\mu\big((\mu_i - X_i)(1 - g(X))X_i\big).$$

Angenommen, die Funktion $f_i(x) = (1 - g(x))x_i$ erfüllt die Voraussetzung des Steinschen Lemmas für die Koordinate i, $i = 1, \ldots, d$. Wegen $\partial_{x_i} f_i = (1 - g(x)) - \partial_{x_i} g(x)\, x_i$ folgt mit dem Steinschen Lemma

$$E_\mu\big((\mu_i - X_i)(1 - g(X))X_i\big) = -\sigma_0^2\, E_\mu\big(1 - g(X) - \partial_{x_i} g(X) X_i\big),$$

bzw.

$$E_\mu\big\langle X - \hat{\mu}, \mu - X\big\rangle = -d\,\sigma_0^2\, E_\mu\big(1 - g(X)\big) + \sigma_0^2\, E_\mu\big\langle \nabla g(X), X\big\rangle,$$

und daher wegen $E_\mu\|X - \mu\|^2 = d\sigma_0^2$

$$E_\mu\|\hat{\mu} - \mu\|^2 = d\sigma_0^2 - E_\mu W_g(X), \qquad W_g(x) = 2\,\sigma_0^2\,\Big(d\big(1 - g(x)\big) - \big\langle \nabla g(x), x\big\rangle\Big) - \|x\|^2(1 - g(x))^2,$$

und es genügt, g derart zu wählen, dass $W_g(x) > 0$ für alle $x \in \mathbb{R}^d$.

Beweis
Betrachte $g(x) = 1 - c/\|x\|^2$, $c > 0$, wobei sich für $c = \sigma_0^2 d$ der Stein Schätzer und für $c = \sigma_0^2(d - 2)$ der James-Stein-Schätzer ergibt. Für $f_i(x) = c\, x_i/\|x\|^2$ ist $\partial_{x_i} f_i(x) = (\|x\|^2 - 2x_i^2)/\|x\|^4$ mit $|\partial_{x_i} f_i(x)| \le 3/\|x\|^2$ unter der Normalverteilung für $d \ge 3$ integrierbar (s.u.), daher kann Steins Lemma angewendet werden. Da

$$\nabla g(x) = 2cx/\|x\|^4,$$

ergibt sich

$$W_c(x) = \frac{1}{\|x\|^2} c\big(2\sigma_0^2(2 - d) + c\big).$$

Man prüft leicht nach, dass $W_c(x) > 0$ für $0 < c < 2\sigma_0^2(d - 2)$, welches die Behauptungen ergibt. Der Wert $c = \sigma_0^2(d - 2)$ maximiert $W_c(x)$ für jedes x. ∎

Lemma 3.15
Sei $Y \sim N(\mu, \sigma^2 I_d)$. Für $d \ge 3$ gilt

$$E\|Y\|^{-2} < \infty.$$

Speziell gilt für $Y \sim N(0, \sigma^2 I_d)$

$$E\|Y\|^{-2} = \frac{1}{\sigma^2(d - 2)}.$$

Beweis
Sei $f(y;\mu,\sigma^2)$ die Dichte von $N(\mu,\sigma^2 I_d)$. Offenbar ist $|f(y;\mu,\sigma^2)| \leq (2\pi\sigma^2)^{-d/2}$ für alle $y \in \mathbb{R}^d$. Setze $B_1(0) = \{y \in \mathbb{R}^d : \|y\|_2 \leq 1$

$$\begin{aligned} E\|Y\|^{-2} &= \int_{B_1(0)} \|y\|^{-2} f(y;\mu,\sigma^2)\, dy + \int_{B_1(0)^c} \|y\|^{-2} f(y;\mu,\sigma^2)\, dy \\ &\leq (2\pi\sigma^2)^{-d/2} \int_{B_1(0)} \|y\|^{-2}\, dy + \int_{B_1(0)^c} f(y;\mu,\sigma^2)\, dy \\ &\leq 1 + C \int_{B_1(0)} \|y\|^{-2}\, dy. \end{aligned}$$

und falls $C_d\, r^{d-1}$ die Fläche der Spähre $S_d(r) = \{y \in \mathbb{R}^d : \|y\|_2 = r$ bezeichnet,

$$\int_{B_1(0)} \|y\|^{-2}\, dy = \int_0^1 r^{-2} C_d r^{d-1}\, dr = C_d \int_0^1 r^{d-3}\, dr < \infty,$$

da $d \geq 3$. ∎

4 Unverzerrtes Schätzen

4.1 Konzept der Unverzerrtheit

Sei $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ ein statistisches Modell, $(A, \mathcal{A})$ ein Aktionsraum und $\ell : \Theta \times A \to [0, \infty)$ eine Verlustfunktion.

Wir wollen die Klasse der Entscheidungsregeln geeignet einschränken.

Definition 4.1
Eine Entscheidungsregel ρ heißt *unverzerrt* (bzgl. der Verlustfunktion ℓ), falls gilt

$$\forall\ \theta, \theta' \in \Theta : \quad E_\theta\big(\ell(\theta', \rho)\big) \geq E_\theta\big(\ell(\theta, \rho)\big) = R(\theta, \rho).$$

Für jedes $\theta \in \Theta$ minimiert also der Parameter θ die Funktion $\theta' \mapsto E_\theta\big(\ell(\theta', \rho(Y))\big)$.

Definition 4.2
Eine unverzerrte Entscheidungsregel ρ heißt *beste unverzerrte Regel* (bzgl. der Verlustfunktion ℓ), falls für jede unverzerrte Enscheidungsregel ρ' gilt

$$R(\theta, \rho) \leq R(\theta, \rho') \qquad \forall\ \theta \in \Theta.$$

ρ minimiert also gleichmäßig das Risiko unter allen unverzerrten Regeln.

Wir betrachten das Konzept hier im Kontext des Schätzens mit der quadratischen Verlustfunktion. Sei dazu $\gamma : \Theta \to \mathbb{R}^k$ ein k-dimensionaler Parameter. Eine messbare Abbildung $\hat{\gamma} : \mathcal{X} \to \mathbb{R}^k$ heißt dann Schätzer für $\gamma(\theta)$. Angenommen,

$$E_\theta\|\hat{\gamma}(Y)\|_2 = \int_{\mathcal{X}} \|\hat{\gamma}(x)\|_2\, dP_\theta(x) < \infty \qquad \forall\ \theta \in \Theta.$$

Dann heißt

$$\mathrm{bias}_{\hat{\gamma}}(\theta) = E_\theta(\hat{\gamma}) - \gamma(\theta)$$

der *Bias* von $\hat{\gamma}$. $\hat{\gamma}$ heißt *erwartungstreu* (oder *unverzerrt*), falls

$$\mathrm{bias}_{\hat{\gamma}}(\theta) = 0 \qquad \forall\ \theta \in \Theta.$$

Wir betrachten nun die quadratische Verlustfunktion $\ell : \Theta \times \mathbb{R}^k \to [0, \infty)$, $\ell(\theta, t) = \|t - \gamma(\theta)\|^2$. Ist

$$E_\theta\|\hat{\gamma}(Y)\|_2^2 = \int_{\mathcal{X}} \|\hat{\gamma}(x)\|_2^2\, dP_\theta(x) < \infty \qquad \forall\ \theta \in \Theta,$$

so heißt

$$R(\theta, \hat{\gamma}) = E_\theta\|\hat{\gamma} - \gamma(\theta)\|_2^2 =: \mathrm{mse}_{\hat{\gamma}}(\theta)$$

der *mittlere quadratische Fehler* von $\hat{\gamma}$.

Lemma 4.3
Ist $E_\theta\|\hat\gamma(Y)\|_2^2 < \infty \;\forall\, \theta \in \Theta$, so ist

$$\begin{aligned}\mathrm{mse}_{\hat\gamma}(\theta) &= \|\mathrm{bias}_{\hat\gamma}(\theta)\|_2^2 + \mathrm{tr}\ \mathrm{Cov}_\theta\left(\hat\gamma\right)\\ &= \sum_{j=1}^{k}\Big(\big(\mathrm{bias}_{\hat\gamma_j}(\theta)\big)^2 + \mathrm{Var}_\theta\left(\hat\gamma_j\right)\Big),\end{aligned}$$

wobei $\hat\gamma = (\hat\gamma_1, \ldots, \hat\gamma_k)^T$.

Beweis

$$\begin{aligned}\mathrm{mse}_{\hat\gamma}(\theta) &= E_\theta\big\langle \hat\gamma - E_\theta\hat\gamma + E_\theta\hat\gamma - \gamma(\theta), \hat\gamma - E_\theta\hat\gamma + E_\theta\hat\gamma - \gamma(\theta)\big\rangle\\ &= E_\theta\big\|\hat\gamma - E_\theta\hat\gamma\|_2^2 + E_\theta\big\|E_\theta\hat\gamma - \gamma(\theta)\big\|_2^2 + 2\big\langle\hat\gamma - E_\theta\hat\gamma, E_\theta\hat\gamma - \gamma(\theta)\big\rangle\\ &= \mathrm{tr}\ \mathrm{Cov}_\theta\left(\hat\gamma\right) + \|\mathrm{bias}_{\hat\gamma}(\theta)\|_2^2\end{aligned}$$

■

Um also den mse eines Schätzers für einen k-dimensionalen Parameter zu minimieren, müssen alle Koordinaten den 1-dim. mse minimieren.

Daher im folgenden $k = 1$.

Lemma 4.4
Seien $\gamma : \Theta \to \mathbb{R}$ ein ein-dimensionaler Parameter und $\hat\gamma : \mathcal{X} \to \mathbb{R}$ ein Schätzer für $\gamma(\theta)$ mit $E_\theta\hat\gamma^2 < \infty$ und $E_\theta\hat\gamma \in \gamma(\Theta)$ für alle $\theta \in \Theta$. Dann ist $\hat\gamma$ genau dann unverzerrt (als Entscheidungsregel bzgl. der quadratischen Verlustfunktion, wenn $\hat\gamma$ erwartungstreu als Schätzer ist.

Beweis
Wie oben ist für $\theta, \theta' \in \Theta$

$$E_\theta\big(\hat\gamma - \gamma(\theta')\big)^2 = \big(E_\theta\hat\gamma - \gamma(\theta')\big)^2 + \mathrm{Var}_\theta\,\hat\gamma. \tag{26}$$

1. Angenommen, $\hat\gamma$ ist unverzerrt als Entscheidungsregel. Dann wird (26) minimal bei $\theta = \theta'$. Da $E_\theta\hat\gamma \in \gamma(\Theta)$, existiert ein $\theta' \in \Theta$ mit $E_\theta\hat\gamma = \gamma(\theta')$, wegen der Minimalität muss daher $E_\theta\hat\gamma = \gamma(\theta)$.
2. Ist $\hat\gamma$ erwartungstreu als Schätzer, so wird (26) offenbar minimal für $\theta = \theta'$. ■

Ein erwartungstreuer Schätzer $\hat\gamma$ ist somit genau dann bester unverzerrter Schätzer, wenn dieser unter allen erwartungstreuen Schätzern $\tilde\gamma$ gleichmäßig minimale Varianz hat

$$\forall\ \theta \in \Theta: \quad \mathrm{Var}_\theta\,\hat\gamma \le \mathrm{Var}_\theta\,\tilde\gamma.$$

In diesem Fall heißt $\hat\gamma$ uniformly minimum variance unbiased estimator (UMVUE).

Nicht jeder Parameter kann unverfälscht geschätzt werden.

Beispiel Sei $X \sim Bin(n,p)$, $\gamma(p) = 1/p$. Angenommen, $\hat{\gamma} : \{0,\dots,n\} \to (1,\infty)$ ist unverfälscht, also

$$\sum_{j=0}^{n} \hat{\gamma}(j)\binom{n}{j} p^j (1-p)^{n-j} = \frac{1}{p} \qquad \forall\ 0 < p < 1$$

bzw.

$$\sum_{j=0}^{n} \hat{\gamma}(j)\binom{n}{j} p^{j+1} (1-p)^{n-j} = 1 \qquad \forall\ 0 < p < 1.$$

Für $p \to 0$ geht LS $\to 0$, aber RS $= 1$, Widerpsruch. ◇

Sei $\gamma : \Theta \to \mathbb{R}$ ein Parameter und

$$U_\gamma := \big\{\hat{\gamma} : \mathcal{X} \to \mathbb{R} \text{ unverfälscht für } \gamma \text{ mit } E_\theta \hat{\gamma}^2 < \infty\ \forall\ \theta \in \Theta\big\}.$$

Ist $U_\gamma \neq \emptyset$, so heißt γ U-schätzerbar. U_γ ist offenbar konvex, und abgeschlossen in jedem $L_2(P_\theta)$, $\theta \in \Theta$. Sei

$$\mathcal{L}_0 := \big\{T : \mathcal{X} \to \mathbb{R} \text{ messbar},\ E_\theta T^2 < \infty,\ E_\theta T = 0\ \forall\ \theta \in \Theta\big\}.$$

$\mathcal{L}_0$ ist ein abgeschlossener Unterraum von jedem $L_2(P_\theta)$.

Satz 4.5
Sei $\hat{\gamma} \in U_\gamma$. Dann ist $\hat{\gamma}$ UMVUE für γ genau dann, wenn

$$E_\theta\big(\hat{\gamma}\, T\big) = 0 \qquad \forall\ \theta \in \Theta,\ T \in \mathcal{L}_0.$$

Beweis
Zunächst bemerken wir, dass $\hat{\gamma}$ genau dann UMVUE, falls

$$E_\theta \hat{\gamma}^2 = \mathrm{Var}_\theta\, \hat{\gamma} + \gamma(\theta)^2 \le \mathrm{Var}_\theta\, \tilde{\gamma} + \gamma(\theta)^2 = E_\theta \tilde{\gamma}^2 \qquad \forall\ \theta \in \Theta, \tilde{\gamma} \in U_\gamma.$$

Also ist $\hat{\gamma}$ genau dann UMVUE, wenn $\hat{\gamma}$ beste Approximation aus U_γ an 0 ist in jedem $L_2(P_\theta)$. Diese wird charakterisiert durch

$$E_\theta\big((0-\hat{\gamma})(\tilde{\gamma}-\hat{\gamma}\big) \le 0 \qquad \forall\ \theta \in \Theta, \tilde{\gamma} \in U_\gamma,$$

(flacher Winkel), also

$$E_\theta\big(\hat{\gamma}\, T\big) \ge 0 \qquad \forall\ \theta \in \Theta,\ T \in \mathcal{L}_0.$$

Da mit $T \in \mathcal{L}_0$ auch $-T \in \mathcal{L}_0$, folgt die Behauptung des Satzes. ■

Aus der Charakterisierung folgt auch die f.s. Eindeutigkeit eines UMVUE, wir geben unten dafür noch ein einfaches direktes Argument. Zunächst geben wir ein Beispiel, in dem γ U-schätzbar ist, es aber keinen UMVUE gibt.

Beispiel. Sei $X \sim U(\theta - 1/2, \theta + 1/2)$, $\theta \in \mathbb{R}$.

Behauptung. Ist $\gamma : \mathbb{R} \to \mathbb{R}$ nicht konstant, so existiert kein UMVUE für γ

Es gibt aber Parameter, die unverfälscht geschätzt werden können, etwa $\gamma(\theta) = \theta$ durch X.

Beweis der Behauptung. Sei S mit $E_\theta S = 0$ für alle $\theta \in \mathbb{R}$, also

$$\int_{\theta-1/2}^{\theta+1/2} S(x)\,dx = 0 \qquad \forall\ \theta \in \mathbb{R}.$$

Durch Differentiation folgt nach dem Hauptstatz für das Lebesgue Integral $S(\theta + 1/2) = S(\theta - 1/2)$ für f.a. $\theta \in \mathbb{R}$, also $S(x) = S(x+1)$ für f.a. $x \in \mathbb{R}$. Ist $\hat{\gamma}$ UMVUE für γ und $T \in \mathcal{L}_0$, so folgt nach Satz 4.5 auch $E_\theta T\hat{\gamma} = 0$ für alle $\theta \in \mathbb{R}$, also $\hat{\gamma}(x)T(x) = \hat{\gamma}(x+1)T(x+1)$ und, indem wir $T \neq 0$ f.ü. wählen (etwa $T(x) = \cos(2\pi x)$), (da $T(x) = T(x+1)$) auch $\hat{\gamma}(x) = \hat{\gamma}(x+1)$ für f.a. $x \in \mathbb{R}$. Da $\hat{\gamma} \in U_\gamma$, folgt

$$\gamma(\theta) = \int_{\theta-1/2}^{\theta+1/2} \hat{\gamma}(x)\,dx = \text{const.}$$

Im nächsten Abschnitt geben wird jedoch ein allgemeines Kriterium an, wann bei Existenz eines beliebigen unverzerrten Schätzers auch garantiert ein UMVUE existiert.

Wir geben noch ein einfaches direktes Argument, dass ein UMVUE f.s. eindeutig bestimt ist.

Satz 4.6
Sind $\hat{\gamma}_1, \hat{\gamma}_2$ UMVUEs für γ, so folgt $\hat{\gamma}_1 = \hat{\gamma}_2$ P_θ-f.s. für alle $\theta \in \Theta$.

Beweis
Es ist nach Voraussetzung $\mathrm{Var}_\theta\,\hat{\gamma}_1 = \mathrm{Var}_\theta\,\hat{\gamma}_2$, daher für alle $\theta \in \Theta$

$$\begin{aligned}\mathrm{Var}_\theta\left((\hat{\gamma}_1 + \hat{\gamma}_2)/2\right) &= \frac{1}{2}\,\mathrm{Var}_\theta\,\hat{\gamma}_1 + \frac{1}{2}\,\mathrm{Cov}\,(\hat{\gamma}_1, \hat{\gamma}_2),\\ \mathrm{Cov}\,(\hat{\gamma}_1, \hat{\gamma}_2) &\leq \mathrm{Var}_\theta\,\hat{\gamma}_1.\end{aligned}$$

Da $(\hat{\gamma}_1 + \hat{\gamma}_2)/2 \in U_\gamma$ und $\hat{\gamma}_1$ UMVUE, muss in der zweiten Zeile ein $=$ stehen, daher

$$\mathrm{Var}_\theta\left(\hat{\gamma}_1 - \hat{\gamma}_2\right) = 2\mathrm{Var}_\theta\,\hat{\gamma}_1 - 2\mathrm{Cov}\,(\hat{\gamma}_1, \hat{\gamma}_2) = 0.$$

∎

4.2 Vollständigkeit

Sei $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ ein statistisches Modell, und $(S, \mathcal{S})$ ein messbarer Raum.

Definition Eine Statistik $T : \mathcal{X} \to S$ heißt *vollständig* (für $(P_\theta)_{\theta\in\Theta}$), falls für alle $h : S \to \mathbb{R}$ gilt

$$\text{Aus } \forall\ \theta \in \Theta :\ E_\theta h(T) = 0 \quad \text{folgt } \ \forall\ \theta \in \Theta :\ h \circ T = 0 \quad P_\theta - \text{f.s.}$$

Es werden insbesondere nur Funktionen f in der Bedingung zugelassen, für die $h \circ T$ P_θ-integrierbar für alle $\theta \in \Theta$ ist.

Satz 4.7
1. Ist $T_1 : \mathcal{X} \to S_1$ vollständig (für $(P_\theta)_{\theta\in\Theta}$), und ist $T_2 = \psi \circ T_1$ für ein messbares $\psi : S_1 \to S_2$, so ist auch T_2 vollständig.

2. Ist $T : \mathcal{X} \to S$ vollständig und suffizient (für $(P_\theta)_{\theta\in\Theta}$), und existiert eine minimal suffiziente Statistik, so ist auch T minimal suffizient.

Beweis
Zu 1.: Sei $h : S_2 \to \mathbb{R}$. Dann

$$\begin{aligned} & \forall\ \theta \in \Theta :\ E_\theta h(T_2) = 0 \\ \Rightarrow\ & \forall\ \theta \in \Theta :\ E_\theta (h \circ \psi)(T_1) = 0 \\ \Rightarrow\ & \forall\ \theta \in \Theta :\ (h \circ \psi)(T_1) = 0 \quad P_\theta - \text{f.s.} \\ \Rightarrow\ & \forall\ \theta \in \Theta :\ h \circ T_2) = 0 \quad P_\theta - \text{f.s.} \end{aligned}$$

wobei die Vollständigkeit von T_1 für die Funktion $h \circ \psi$ angewendet haben.

Zu 2.: s. ???????? ∎

Satz 4.8 (*von Lehmann und Scheffe)*
Sei $\gamma : \Theta \to \mathbb{R}$ U-schätzbar, etwa $\hat\gamma \in U_\gamma$, und sei $T : \mathcal{X} \to S$ eine suffiziente und vollständige Statistik (für $(P_\theta)_{\theta\in\Theta}$). Dann ist

$$\tilde\gamma = E.(\hat\gamma|T)$$

der UMVUE für γ.

Falls also eine vollständige und suffiziente Statistik existiert, hat jeder U-schätzbare Parameter einen UMVUE.

Bemerkung. Da $E.(\hat\gamma|T)$ $\sigma(T)$-messbar, existiert (nach dem Faktorisierungslemma) $h : S \to \mathbb{R}$ messbar mit $E.(\hat\gamma|T) = h \circ T$. Da UMVUE P_θ-f.s. eindeutig bestimmg ist für alle $\theta \in \Theta$, ist h P_θ^T-f.s. eindeutig für alle $\theta \in \Theta$.

Beweis
a. Zeige: $\tilde\gamma \in U_\gamma$. Dabei ist $E_\theta \tilde\gamma^2 < \infty$ wegen der Jenssen Ungleichung, und

$$E_\theta \tilde\gamma = E_\theta\big(E.(\hat\gamma|T)\big) = E_\gamma \hat\gamma = \gamma(\theta).$$

b. UMVUE: Sei dazu $\hat{\eta} \in U_\gamma$ beliebig, dann ist nach Teil a. auch $\tilde{\eta} = E.(\hat{\eta}|T) \in U_\gamma$. Wir zeigen nun

$$\tilde{\eta} = \tilde{\gamma} \qquad P_\theta - \text{f.s.} \quad \forall\, \theta \in \Theta, \tag{27}$$

dann folgt mit dem Satz von Rao-Blackwell (wegen der Suffizienz von T)

$$\operatorname{Var}_\theta \tilde{\gamma} = \operatorname{Var}_\theta \tilde{\eta} \le \operatorname{Var}_\theta \hat{\eta} \qquad \forall\, \theta \in \Theta.$$

Zu (27): Da $\tilde{\gamma}$ und $\tilde{\eta}$ $\sigma(T)$ messbar, schreibe $\tilde{\gamma} = h \circ T$, $\tilde{\eta} = s \circ T$, wobei $h, s : S \to \mathbb{R}$. Dann ist

$$E_\theta\big((s-h)\circ T\big) = E_\theta\big(\tilde{\eta} - \tilde{\gamma}\big) = \gamma(\theta) - \gamma(\theta) = 0 \qquad \forall\, \theta \in \Theta,$$

also wegen der Vollständigkeit von T: $s = h$ P_θ^T-f.s. für alle $\theta \in \Theta$, wie benötigt. ∎

Beispiel. Seien $X_1, \dots, X_n$ u.i.v. $\sim U(0,\theta)$, $\theta > 0$. Dann ist die gemeinsame Dichte (bzgl. Lebesgue) gegeben durch

$$f_\theta(x_1, \dots, x_n) = \frac{1}{\theta^n}\, 1_{(0,\theta)}(x_{(n)}), \qquad x_i > 0,$$

daher ist $T = X_{(n)}$ (n-te Ordnungsstatistik) suffizient. Die Verteilungsfunktion von $X_{(n)} = \max_{1\le i\le n} X_i$ ist

$$G_\theta(z) = \frac{1}{\theta^n}\, 1_{(0,\theta)}(z) z^n, \qquad z > 0,$$

und die Dichte somit zu

$$g_\theta(z) = \frac{1}{\theta^n}\, n\, 1_{(0,\theta)}(z) z^{n-1}, \qquad z > 0.$$

Gilt daher

$$E_\theta f\big(X_{(n)}\big) = \frac{n}{\theta^n} \int_0^\theta f(t)\, t^{n-1}\, dt = 0 \qquad \forall \theta > 0,$$

so ergibt sich mit dem Hauptsatz nach Lebesgue $f = 0$ Lebesgue-f.ü., also ist $X_{(n)}$ auch vollständig. Da $E_\theta X_{(n)} = \frac{n}{n+1}\theta$, ist z.B. $\frac{n+1}{n} X_{(n)}$ der UMVUE für θ. ◇

Der folgende Satz liefert eine große Klasse von suffizienten und vollständigen Statistiken.

Satz 4.9
Sei $(P_\theta)_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^k$, eine k-parametrische exponentielle Familie mit natürlicher suffizienter Statistik und natürlichem Parameter θ, d.h. bzgl. eines σ-endlichen Maßes μ auf $\mathcal{X}$ gilt

$$L(\theta, x) = \frac{dP_\theta}{d\mu}(x) = Q(\theta)\, \exp\big(\theta^T T(x)\big)\, h(x).$$

Hat Θ nicht-leeres Inneres, so ist T vollständig (und natürlich suffizient) für θ.

Beweis
o.E. $h = 1$, sei $U \subset \Theta$ offen, und $f : \mathbb{R}^k \to \mathbb{R}$ mit

$$\begin{aligned} \forall\, \theta \in U : \quad 0 &= \int_{\mathcal{X}} f(T(x))\, C(\theta)\, \exp\big(\theta^T\, T(x)\big)\, d\mu(x) \\ &= C(\theta) \int_{\mathbb{R}^k} f(t)\, \exp\big(\theta^T\, t\big)\, d\mu_T(t). \end{aligned}$$

also

$$\forall\, \theta \in U: \quad \int_{\mathbb{R}^k} f^+(t)\, \exp\left(\theta^T t\right) d\mu_T(t) = \int_{\mathbb{R}^k} f^-(t)\, \exp\left(\theta^T t\right) d\mu_T(t). \tag{28}$$

Hieraus folgert man (REFERENZ!) die Gleichheit der endlichen Maße

$$f^+ d\mu_T = f^- d\mu_T,$$

also $f^+ = f^-$ μ_T-f.ü., und somit $f \circ T = 0$ μ-f.ü. , und daher auch unter jedem P_θ. ∎

Beispiel. Wir betrachten das lineare Modell $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$, wobei $\boldsymbol{\beta} \in \mathbb{R}^p$. Dann ist

$$T(\mathbf{Y}) = \left(X^T\mathbf{Y}\, \|\mathbf{Y}\|^2\right)$$

suffizient und vollständig, vgl. (????). Daher sind

$$\hat{\boldsymbol{\beta}}_{LS} = (X^TX)^{-1}\, X^T\mathbf{Y} \quad \text{UMVUE (koordinatenweise) für } \boldsymbol{\beta},$$
$$\hat{\sigma}^2 = \frac{1}{n-p}\left(\|\mathbf{Y}\|^2 - \mathbf{Y}^T\, X(X^TX)^{-1}\, X^T\mathbf{Y}\right) \quad \text{UMVUE für } \sigma^2.$$

Beispiel. Für die Binomialverteilung $X \sim \mathrm{Bin}(n,p)$, $p \in (0,1)$, ist

$$L(p,x) = (1-p)^n\, \exp\left(x \log(p/(1-p))\right) \binom{n}{x}, \qquad x = 0, \ldots, n,$$

also X suffizient und vollständig.
Wir betrachten das Schätzen der Varianz $\gamma(p) = np(1-p)$, diese ist für $n \geq 2$ stets unverfälscht schätzbar durch die Stichprobenvarianz $\hat{\gamma}(x) = \left(x - x^2/n\right)/(n-1)$. Um den UMVUE $h(X)$ zu bestimmen, müssen wir $h: \{0, \ldots, n\} \to \mathbb{R}$ finden, für das

$$\sum_{x=0}^{n} \binom{n}{x} h(x)\, p^x (1-p)^{n-x} = np(1-p), \qquad p \in (0,1).$$

Wir setzen $\rho = p/(1-p)$, also $1 + \rho = 1/(1-p)$, dividiere die obige Gleichung durch $(1-p)^n$, dann muss gelten

$$\begin{aligned}\sum_{x=0}^{n} \binom{n}{x} h(x)\, \rho^x &= n\, \frac{p}{1-p}\, \frac{1}{(1-p)^{n-2}} \\ &= n\rho(1+\rho)^{n-2} \\ &= n \sum_{x=1}^{n-1} \binom{n-2}{x-1} \rho^x, \qquad 0 < \rho < \infty.\end{aligned}$$

Also ergibt sich

$$h(x) = \begin{cases} 0, & x = 0, n, \\ n\binom{n-2}{x-1}/\binom{n}{x} = \frac{x(n-x)}{n-1}, & 1 \leq x \leq n-1. \end{cases}$$

Wir beenden den Abschnitt über vollständige Statistiken noch mit einem nützlich Resultat über Unabhängigkeit.

Eine Statistik V heißt *anzillär* (*ancillary*) für $(P_\theta)_{\theta\in\Theta}$, falls $P_\theta^V = P_{\theta'}^V$ für alle $\theta, \theta' \in \Theta$, falls also die Verteilung von V nicht von θ abhängt.

Beispiel. Seien $X_1, \ldots, X_n$ u.i.v. $\sim N(0, \sigma^2)$, so ist für $1 \le i \le n$

$$V_i(x_1, \ldots, x_n) = \frac{x_i^2}{x_1^2 + \ldots + x_n^2} = \frac{(x_i/\sigma)^2}{(x_1/\sigma)^2 + \ldots + (x_n/\sigma)^2}$$

anzillär für σ, da $X_i/\sigma \sim N(0,1)$ unabhängig von σ^2.

Satz 4.10 (*von Basu*)
Ist T vollständig und suffizient für $(P_\theta)_{\theta\in\Theta}$ und V anzillär für $(P_\theta)_{\theta\in\Theta}$, so sind V und T unabhängig unter jedem P_θ, $\theta \in \Theta$.

Beweis
Wir müssen zeigen, dass für messbare A, B, und $\theta \in \Theta$ gilt

$$P_\theta\big(T^{-1}(A) \cap V^{-1}(B)\big) = P_\theta\big(T^{-1}(A)\big)\, P_\theta\big(V^{-1}(B)\big) \tag{29}$$

Nun gilt

$$\begin{aligned} P_\theta\big(T^{-1}(A) \cap V^{-1}(B)\big) &= E_\theta\big(E_\theta(1_A(T)\, 1_B(V)|T)\big) \\ &= E_\theta\big(1_A(T)\, E_\theta(1_B(V)|T)\big), \end{aligned} \tag{30}$$

wobei $E_\theta(1_B(V)|T) = E_\cdot(1_B(V)|T) = h \circ T$ für ein messbares h und alle θ, da T suffizient ist.

Sei μ die Verteilung von V unter einem (dann jedem) P_θ (V ist anzillär), so ist

$$E_\theta\big(h \circ T\big) = E_\theta\big(E_\cdot(1_B(V)|T)\big) = P_\theta(V \in B) = \mu(B) \qquad \forall\ \theta \in \Theta.$$

Da T vollständig ist, folgt

$$E_\cdot(1_B(V)|T) = h \circ T = \mu(B) \qquad \text{f.s. unter } P_\theta \qquad \forall\ \theta \in \Theta.$$

Einsetzen in (30) liefert dann leicht (29). ∎

Beispiel Wir setzen obiges Beispiel fort. Sei

$$T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i^2.$$

Die Dichte von $(X_1, \ldots, X_n)$ ist gegeben durch

$$f\big(x_1, \ldots, x_n; \sigma^2\big) = \frac{1}{(2\pi\,\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^n \frac{x_i^2}{\sigma^2}\Big),$$

also eine EF mit suffizienter Statistik T und natürlichem Parameter $q = -1/(2\sigma^2)$. Daher ist T suffizient und vollständig für q und daher auch für σ^2, und der Satz von Basu liefert die Unabhängigkeit von T und jedem V_i.

4.3 Nichtparametrische Probleme

Wir betrachten nun eine Klasse von UMVUEs für Parameter nicht-parametrischer statistischer Modelle. Sei $\mathcal{X} = \mathbb{R}^k$, und sei $(P_\theta)_{\theta\in\Theta}$ ein statistisches Modell, und $X_1, \dots, X_n$ eine mathematische Stichprobe nach (P_θ). Wir bestimmen, welche Parameter γ aus $n \geq m$ Beobachtungen stets unverfälscht geschätzt werden können.

Satz 4.11
Für einen Parameter $\gamma : \Theta \to \mathbb{R}$ und $m \in \mathbb{N}$ sind äquivalent:

1. Für jedes $n \geq m$ existiert $T_n : \mathcal{X}^n \to \mathbb{R}$, s.d.

$$E_\theta T_n(X_1, \dots, X_n) = \gamma(\theta) \qquad \forall\ \theta \in \Theta.$$

2. Es existiert ein $h : \mathcal{X}^m \to \mathbb{R}$ messbar und symmetrisch mit $h \in L_1(\mathcal{X}^m, P_\theta^m)$ für alle $\theta \in \Theta$, so dass

$$\gamma(\theta) = \int_{\mathcal{X}} \dots \int_{\mathcal{X}} h(x_1, \dots, x_m)\, dP_\theta(x_1) \dots dP_\theta(x_m) \qquad \forall\ \theta \in \Theta. \tag{31}$$

Bemerkung. $h : \mathcal{X}^m \to \mathbb{R}$ heißt symmetrisch, falls für alle Permutationen $\pi \in S_m$ und $x_1, \dots, x_m \in \mathcal{X}$ gilt $h(x_1, \dots, x_m) = h(x_{\pi(1)}, \dots, x_{\pi(m)})$.

Beweis
2. $\Rightarrow$ 1. : Setze

$$T_n(x_1, \dots, x_n) = h(x_1, \dots, x_m).$$

1. $\Rightarrow$ 2. : Setze

$$h(x_1, \dots, x_m) = \frac{1}{m!} \sum_{\pi\in S_m} T_m(x_{\pi(1)}, \dots, x_{\pi(m)}).$$

Dann ist h symmetrisch, und da

$$E_\theta T_m(X_{\pi(1)}, \dots, X_{\pi(m)}) = E_\theta T_m(X_1, \dots, X_m) = \gamma(\theta) \qquad \forall\ \theta \in \Theta$$

folgt die Behauptung. ∎

Definition 4.12
Ist $h : \mathcal{X}^m \to \mathbb{R}$ symmetrisch und $h \in L_1(\mathcal{X}^m, P_\theta^m)$ für alle $\theta \in \Theta$, so heißt

$$U_n(h) = U_n(h)(x_1, \dots, x_n) = \frac{1}{\binom{n}{m}} \sum_{1\leq i_1<\dots<i_m\leq n} h(x_{i_1}, \dots, x_{i_m}),$$

$x_1, \dots, x_n \in \mathcal{X}$, U-Statistik mit Kern h.

Satz 4.13
$U_n(h)$ ist ein unverfälschter Schätzer für γ in (31) für alle $n \geq m$.

Beweis
Klar, da $E_\theta(X_{i_1}, \dots, X_{i_m}) = \gamma(\theta)$ für alle Index Tupel, von denen es $\binom{n}{m}$ gibt. ∎

Beispiele 1. Stichprobenmomente: $h(x) = x^k$, dann

$$U_n(h) = \frac{1}{n}\sum_{j=1}^{n} x_j^k$$

U-Statistik der Ordnung 1.

2. Varianz: Es ist $VarX_1 = E(X_1 - X_2)^2/2$ (wobei X_1, X_2 i.i.d.), mit $h(x_1, x_2) = (x_1 - x_2)^2/2$ ist $U_n(h)$ die Stichprobenvarianz.

3. Gini's mean difference: $h(x_1, x_2) = |x_1 - x_2|$ (robustes Streuungsmaß)

4. Kendell's Korrelationskoeffizient: Für bivariate Beobachtungen $(x, y) \in \mathbb{R}^2$ ist

$$h\big((x_1, y_1), (x_2, y_2)\big) = \operatorname{sign}\big((x_1 - x_2)(y_1 - y_2)\big).$$

Interpretation: $h\big((x_1, y_1), (x_2, y_2) = 1$ falls (x_1, y_1) und (x_2, y_2) konkordant (Gerade hat positive Steigung), $h\big((x_1, y_1), (x_2, y_2) = -11$ falls (x_1, y_1) und (x_2, y_2) diskordant (Gerade hat negative Steigung).

Satz 4.14
Ist $\mathcal{X} = \mathbb{R}$ und ist die Ordnungsstatistik $X^{(n)} = \big(X_{(1)}, \ldots, X_{(n)}\big)$ suffizient und vollständig für $(P_\theta^n)_{\theta\in\Theta}$ (n-fache Produktmaße auf $\mathbb{R}^n$), so ist $U_n(h)$ UMVUE für γ.

Beweis
Nach dem Satz von Lehmann-Scheffe genügt es zu zeigen, dass $U_n(h)$ nur von der Ordnungsstatistik $x^{(n)}$ abhängt. Nun geht $x^{(n)}$ aus $(x_1, \ldots, x_n)$ durch Permutation der Indizes hervor, daher genügt es zu zeigen, dass $U_n(h)$ invariant unter Permutation der Argumente ist: Für $\pi \in S_n$ ist mit der Symmetrie von h

$$\begin{aligned} U_n(h)(x_1, \ldots, x_n) &= \frac{1}{\binom{n}{m}} \sum_{\{i_1,\ldots,i_m\}\subset\{1,\ldots,n\}} h(x_{i_1}, \ldots, x_{i_m}) \\ &= \frac{1}{\binom{n}{m}} \sum_{\{i_1,\ldots,i_m\}\subset\{1,\ldots,n\}} h(x_{\pi(i_1)}, \ldots, x_{\pi(i_m)}) \end{aligned}$$

da mit $\{i_1, \ldots, i_m\}$ auch $\{\pi(i_1), \ldots, \pi(i_m)\}$ alle m-elementigen Teilmengen von $\{1, \ldots, n\}$ durchläuft. ∎

Die Ordnungsstatistik ist nur in großen, nichtparametrischen Familien vollständig. Wir betrachten

$$\Theta = \{f : \mathbb{R} \to [0, \infty) \text{ messbar}, \int_{\mathbb{R}} f(x)\, dx = 1\}$$

die Familie der Lebesgue Dichten und $dP_f = f \circ d\lambda$ die bzgl. des Lebesgue-Maßes absolutstetigen Wahrscheinlichkeitsmaße.

Satz 4.15
Für jedes $n \geq 1$ ist die Ordnungsstatistik $X^{(n)} = \big(X_{(1)}, \ldots, X_{(n)}\big)$ vollständig (und suffizient) für $(P_f^n)_{f\in\Theta}$.

Beweis
Sind T, S Statistiken mit $T = \psi \circ S$, $S = \psi^{-1} \circ T$, wobei ψ und ψ^{-1} messbar, so ist T genau dann vollständig, wenn S vollständig ist. Wir nennen zwei solche Statistiken äquivalent.

Für $x = (x_1, \ldots, x_n)$ betrachten wir die Statistiken

$$\begin{aligned}
T(x) &= \big(x_{(1)}, \ldots, x_{(n)}\big), \\
U(x) &= \big(U_1(x), \ldots, U_n(x)\big), \qquad U_j(x) = \sum_{k=1}^{n} x_k^j, \\
V(x) &= \big(V_1(x), \ldots, V_n(x)\big), \qquad V_1(x) = \sum_{k=1}^{n} x_k,\ V_2(x) = \sum_{1\le j<k\le n} x_j x_k \\
&\qquad V_3(x) = \sum_{1\le j<k<l\le n} x_j x_k x_l, \quad \ldots, \quad V_n(x) = x_1 \ldots x_n.
\end{aligned}$$

Der Satz ergibt sich dann aus den folgenden beiden Behauptungen

Behauptung A: T, U und V sind äquivalent.

Behauptung B: U ist vollständig.

Zu B.: Wir betrachten die parametrische Teilfamilie mit Dichten

$$f(x,\theta) = C(\theta)\, \exp\big(-x^{2n} + \theta_1 x + \ldots + \theta_n x^n\big), \qquad x \in \mathbb{R}, \quad \theta \in \mathbb{R}^n,$$

eine Standard Exponentielle Familie. Die Produktdichte auf $\mathbb{R}^n$ ist dann ebenfalls eine SEF, und deren natürliche suffiziente Statistik ist U, daher ist U vollständig für $(P_{f_\theta})_{\theta\in\mathbb{R}^n}$.

Dann ist U auch vollständig für $(P_f^n)_{f\in\Theta}$. Da $f(x,\theta) \sim \lambda$, folgt für $h : \mathbb{R}^n \to \mathbb{R}$ messbar

$$\begin{aligned}
& E_f\big(h(U)\big) = 0 \qquad \forall\ f \in \Theta \\
\Rightarrow\ & E_{f(\cdot,\theta)}\big(h(U)\big) = 0 \qquad \forall\ \theta \in \mathbb{R}^n \\
\Rightarrow\ & h(U) = 0 \quad \text{f.s. } P^n_{f(\cdot,\theta)} \qquad \forall\ \theta \in \mathbb{R}^n \\
\Rightarrow\ & h(U) = 0 \quad \lambda^n\text{f.ü.} \qquad \forall\ \theta \in \mathbb{R}^n \\
\Rightarrow\ & h(U) = 0 \quad \text{f.ü. } P_f^n \qquad \forall\ f \in \Theta.
\end{aligned}$$

Zu A.: Das $V(x) = \psi(T(x))$ ist klar. Umgekehrt ist

$$p(t) = (t - x_1) \cdot (t - x_n) = t^n - V_1 t^{n-1} + \ldots + (-1)^n V_n.$$

Die Nullstellen von $p(t)$ sind gerade die Koordinaten von T, diese sind stetige Funktionen der Koeffizienten (REFERENZ!), also von den Koordinaten von V, daher $T(x) = \phi(V(x))$.

Die Äquivalenz von $U(X)$ und $V(X)$ folgt mit Induktion aus den *Newtonschen Identitäten* (REFERENZ!)

$$U_k - V_1 U_{k-1} + V_2 U_{k-1} + \ldots + (-1)^{k-1} V_{k-1} U_1 + (-1)^k k V_k = 0, \quad k \ge 1.$$

■

5 Testtheorie

5.1 Grundbegriffe der Testtheorie und entscheidungstheoretische Einbettung

Grundbegriffe

Sei $\left(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\right)$ ein statistisches Modell.

Der Parameterraum $\Theta = \Theta_0 \cup \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0, \Theta_1 \neq \emptyset$, sei disjunkt zerlegt. Dann heißt

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \qquad \text{(Null)-Hypothese,} \\ H_1 &: \theta \in \Theta_1 \qquad \text{Alternativhypothese (Alternative).} \end{aligned}$$

Falls $\#\Theta_j = 1$, so heißt Θ_j *einfach*, sonst *zusammengesetzt*, $j = 1, 2$.

Eine Statistik $\phi : \mathcal{X} \to [0, 1]$ heißt ein *randomisierter Test*.

Interpretation: $\phi(x) = P(H_0 \text{ ablehnen}|X = x)$.

Ist $\phi : \mathcal{X} \to \{0, 1\}$, so heißt ϕ ein *nicht-randomisierter Test* oder einfach ein *Test*. Für einen Test $\phi : \mathcal{X} \to \{0, 1\}$ heißt $\phi^{-1}(1) \subset \mathcal{X}$ sein *kritischer Bereich* (oder Ablehnbereich), $\phi^{-1}(0) \subset \mathcal{X}$ der *Annahmebereich*. Offenbar ist

$$x \in \phi^{-1}(1) \Leftrightarrow H_0 \text{ ablehnen.}$$

Sei $S : \mathcal{X} \to \bar{\mathbb{R}}$ eine Statistik. Ist $\phi(x) = \psi(S(x))$, so heißt S die Prüfgröße für den (randomisierten) Test ϕ.

Ist im nicht-randomisierten Fall für ein $c \in \mathbb{R}$

$$\begin{aligned} \phi(x) = 1_{S(x)>c} &: \qquad \text{Test mit oberem Ablehnbereich,} \\ \phi(x) = 1_{S(x)<c} &: \qquad \text{Test mit unterem Ablehnbereich,} \end{aligned}$$

sowie für $c_1 < c_2$,

$$\phi(x) = 1_{S(x)<c_1} + 1_{S(x)>c_2} : \qquad \text{Test mit zweiseitigem Ablehnbereich.}$$

Fehler 1. Art: Verwerfe H_0, obwohl H_0 richtig,
Fehler 2. Art: Verwerfe H_0 nicht, obwohl H_0 falsch.
Die *Gütefunktion* des randomisierten Tests $\phi : \mathcal{X} \to [0, 1]$ ist

$$G_\phi : \Theta \to [0, 1], \quad G_\phi(\theta) = E_\theta\phi = \int_{\mathcal{X}} \phi(x)\, dP_\theta(x).$$

Ist $\phi : \mathcal{X} \to \{0, 1\}$ nicht randomisiert, so ist

$$G_\phi(\theta) = P_\theta(\phi = 1)$$

die Wahrscheinlichkeit, die Nullhypothese zu verwerfen. Der *Umfang (size)* des Tests ϕ ist

$$\sup_{\theta\in\Theta_0} G_\phi(\theta),$$

der Test hat *Niveau* $\alpha > 0$, falls

$$\sup_{\theta\in\Theta_0} G_\phi(\theta) \le \alpha,$$

dabei muss nicht die Gleichheit gelten.

Bemerkungen zum Randomisieren

- Randomisieren wird benötigt, um ein gegebenes Niveau $\alpha > 0$ voll auszuschöpfen.
- Für $\phi(x) \in (0,1)$ muss ein zusätzliches Bernoulli-Experiment mit Erfolgswahrscheinlichkeit $\phi(x)$ durchgeführt werden, und dann wird bei Erfolg verworfen,
- In der Praxis offenbar unerwünscht, stattdessen Angabe des P-Wertes, s.u.

Entscheidungstheoretische Einbettung

Sei $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ ein statistisches Modell, $(A, \mathcal{A})$ ein Aktionsraum, und $\ell : \Theta \times A \to [0,\infty)$ eine Verlustfunktion.

Definition. Eine randomisierte Entscheidungsregel ist ein stochastischer Kern $\rho : \mathcal{X} \times \mathcal{A} \to [0,1]$, d.h.
1. $\forall\, x \in \mathcal{X}$: $B \mapsto \rho(x,B)$ ein Wahrscheinlichkeitsmaß,
2. $\forall\, B \in \mathcal{A}$: $x \mapsto \rho(x,B)$ ist messbar.
Das Risiko der randomisierten Entscheidungsregel ρ ist

$$R(\theta,\rho) = \int_{\mathcal{X}} \int_A \ell\big(\theta,a\big)\, \rho(x,da)\, dP_\theta(x).$$

Bemerkungen. 1. Ist $x \in \mathcal{X}$ beobachtet, so triff zufällige Entscheidung nach Verteilung $\rho(x,\cdot)$.
2. Ist $\rho : \mathcal{X} \to A$ eine (nicht-randomisierte) Entscheidungsregel, so kann diese im Sinne obiger Definition interpretiert werden als $\rho(x,B) = \delta_{\rho(x)}(B) = 1_B(\rho(x))$.
3. Raum der randomisierte Entscheidungsregeln konvex.

Bemerkung. Ein randomisierter Test $\phi : \mathcal{X} \to [0,1]$ kann als randomisierte Entscheidungsregel

$$\rho : \mathcal{X} \times \{0,1\} \to [0,1]$$

interpretiert werden über $\rho(x,\{1\}) = \phi(x)$ und $\rho(x,\{0\}) = 1 - \phi(x)$.

Satz 5.1
Sei ρ eine randomisierte Entscheidungsregel und sei $T : \mathcal{X} \to S$ suffizient. Setze

$$\tilde\rho(t,B) = \int_{\mathcal{X}} \rho(x,B)\, P_\cdot(dx|T=t),$$

$\rho_T(x,B) = \tilde\rho(T(x),B) = E_\cdot(\rho(\cdot,B)|T)$. Dann ist ρ_T eine randomisierte Entscheidungsregel in der suffizienten Statistik T mit gleicher Risikofunktion wie ρ.

Bemerke, dass das Produkt von zwei stochastischen Kernen wieder ein stochastischer Kern ist.

Beweis
Wir führen den Beweis für eine nicht-randomisierte Entscheidungsregel ρ. In diesem Fall ist

$$\tilde{\rho}(t, B) = P.\big(\rho^{-1}(B)|T = t\big).$$

Daher

$$\int_A \ell(\theta, a)\rho_T(x, da) = \int_{\mathcal{X}} \ell(\theta, \rho(y))\, P.(dy|T = T(x)) \; = E.\big(\ell(\theta, \rho)|T\big)(x),$$

und es folgt

$$\begin{aligned} R(\theta, \rho_T) &= \int_{\mathcal{X}} \int_A \ell\big(\theta, a\big)\, \rho_T(x, da)\, dP_\theta(x) \\ &= E_\theta\big(E.\big(\ell(\theta, \rho)|T\big)\big) = E_\theta\big(\ell(\theta, \rho)\big) = R(\theta, \rho). \end{aligned}$$

■

Satz 5.2
Sei $A \subset \mathbb{R}^k$ konvex, $\ell(\theta, \cdot)$ konvex für jedes θ und sei ρ eine randomisierte Entscheidungsregel mit $\int_A \|a\|\, \rho(x, da) < \infty$ für alle $x \in \mathcal{X}$. Dann hat die nicht-randomisierte Entscheidungsregel $\rho_d(x) = \int_A a\, \rho(x, da)$ gleichmäßig nicht-größeres Risiko:

$$R(\theta, \rho_d) \le R(\theta, \rho) \qquad \forall\ \theta \in \Theta.$$

Beweis
Nach Jensen gilt für alle $x \in \mathcal{X}$, $\theta \in \Theta$:

$$\ell\big(\theta, \int_A a\rho(x, da)\big) \le \int_A \ell\big(\theta, a\big)\rho(x, da),$$

Integration bzgl. P_θ liefert die Behauptung. ■

Schließlich zeigen wir, was die Eigenschaft der Unverzerrtheit für (randomisierte) Tests bedeutet.

Satz 5.3
Im Testproblem $\Theta = \Theta_0 \cup \Theta_1$ betrachten wir die Verlustfunktion $\ell : \Theta \times \{0, 1\} \to [0, \infty)$, $\ell(\theta, a) = l_0 a 1_{\theta \in \Theta_0} + l_1(1 - a)1_{\theta \in \Theta_1}$ für Konstanten $l_0, l_1 > 0$. Ein randomisierter Test $\phi : \mathcal{X} \to [0, 1]$ ist genau dann unverzerrt als Entscheidungsregel bzgl. ℓ, falls ϕ unverzerrter Test zum Niveau $\alpha = l_1/(l_0 + l_1)$ ist, d.h. falls für die Gütefunktion gilt

$$\forall\ \theta \in \Theta_0 : \quad G_\phi(\theta) \le \alpha, \qquad \forall\ \theta \in \Theta_1 : \quad G_\phi(\theta) \ge \alpha. \tag{32}$$

Beweis
Für $\theta, \theta' \in \Theta$ ist

$$E_\theta \ell(\theta', \phi) = \begin{cases} l_0 G_\phi(\theta), & \theta' \in \Theta_0 \\ l_1\big(1 - G_\phi(\theta)\big), & \theta' \in \Theta_1 \end{cases}$$

Somit ist ϕ unverzerrt als Entscheidungsregel genau dann, wenn
1. für $\theta \in \Theta_0$ gilt $l_0 G_\phi(\theta) \le l_1\big(1 - G_\phi(\theta)\big)$,
2. für $\theta \in \Theta_1$ gilt $l_0 G_\phi(\theta) \ge l_1\big(1 - G_\phi(\theta)\big)$,
dies ist aber gerade äquivalent zu (32). ■

Definition 1. Sind $\phi_1, \phi_2 : \mathcal{X} \to [0,1]$ (randomisierte) Tests mit Niveau $\alpha > 0$, so heißt ϕ_1 *gleichmäßig besser* als ϕ_2, falls gilt

$$G_{\phi_1}(\theta) \geq G_{\phi_2}(\theta) \qquad \forall\ \theta \in \Theta_1.$$

2. Ein (randomisierter) Test ϕ^* zum Niveau $\alpha \in [0,1]$ heißt *gleichmäßig bester Test zum Niveau* α (UMP Test zum Niveau α, uniformly most powerful test), falls ϕ^* besser ist als jeder andere (randomisierter) Test zum Niveau α.
3. Ein unverzerrter Test ϕ^* zum Niveau $\alpha \in [0,1]$, d.h. mit der Eigenschaft (**??**) heißt *gleichmäßig bester unverzerrter Test zum Niveau* α (UMPU Test zum Niveau α, uniformly most powerful unbiased test), falls ϕ^* besser ist als jeder andere (randomisierter) unverzerrte Test zum Niveau α.

5.2 Das Neyman-Pearson Lemma

Wir betrachten die Situation mit einfacher Hypothese $\Theta_0 = \{\theta_0\}$ und Alternative $\Theta_1 = \{\theta_1\}$. Setze $P_0 = P_0$ und $P_1 = P_{\theta_1}$, diese werden dominiert etwa durch $\mu = P_0 + P_1$, seien f_j die Dichten bzgl. μ, $j = 1, 2$.

Definition. Ein randomisierter Test $\phi_{NP} : \mathcal{X} \to [0,1]$ heißt *Neyman-Pearson-Test (NP-Test)* für $H_0 : \theta = \theta_0$ gegen $H_1 : \theta = \theta_1$, falls $c \geq 0$, $\gamma : \mathcal{X} \to [0,1]$ messbar existieren mit

$$\phi_{NP}(x) = \begin{cases} 1, & f_1(x) > cf_0(x), \\ \gamma(x), & f_1(x) = cf_0(x) \\ 0, & f_1(x) < cf_0(x) \end{cases} \tag{33}$$

Offenbar ist der Umfang eines NP-Tests

$$E_0(\phi_{NP}) = P_0\big(f_1 > cf_0\big) + \int_{\{f_1 = cf_0\}} \gamma(x) dP_0(x).$$

Ziel des Abschnitts ist der folgende Satz.

Satz 5.4
1. Zu jedem $\alpha \in (0,1)$ existiert (für geeignete Wahl von c, γ in (33)) ein NP-Test mit Umfang α, der auf $\{f_1 = cf_0\}$ konstant ist. Dieser ist UMP-Test zum Niveau α.
2. Ist ϕ ein UMP-Test zum Niveau $\alpha \in (0,1)$, und ist ϕ_{NP} ein NP-Test mit Umfang α und c für ϕ_{NP} wie in (33), so gilt

$$\phi(x) = \phi_{NP}(x) \quad \text{für } \mu - f.a.\ x \in \{y : f_1(y) \neq cf_0(y)\}.$$

3. Ist ϕ UMP-Test zum Niveau $\alpha \in (0,1)$, so ist $E_0\phi = \alpha$, oder es existiert ein Test ϕ' mit Niveau $0 < \alpha' < \alpha$ und Güte $G_{\phi'}(\theta_1) = 1$.

Für den Beweis beginnen wir mit

Lemma 5.5
Sei ϕ_{NP} ein NP-Test und ϕ ein beliebiger Test, so gilt

$$\forall\, x \in \mathcal{X} : \quad \big(\phi_{NP}(x) - \phi(x)\big)\big(f_1(x) - cf_0(x)\big) \geq 0. \tag{34}$$

Beweis
Setze

$$M^+ = \big\{\phi_{NP} > \phi\big\}, \quad M^- = \big\{\phi_{NP} < \phi\big\}, \quad M^= = \big\{\phi_{NP} = \phi\big\}$$

Für $x \in M^=$ ist die Aussage klar. Ist $x \in M^+$, so muss $\phi_{NP}(x) > 0$ und daher muss $f_1(x) \geq cf_0(x)$ nach Definition von ϕ_{NP}. Ist schließlich $x \in M^-$, so muss $\phi_{NP}(x) < 1$ und somit $f_1(x) \leq cf_0(x)$ sein. Insgesamt gilt also stets (34). ∎

Lemma 5.6
Ein NP-Test ϕ_{NP} ist UMP-Test zum Niveau $\alpha := E_0\phi_{NP}$.

Beweis
Ist ϕ ein weiterer Test zum Niveau α, so gilt nach Lemma 5.5

$$\begin{aligned} G_{\phi_{NP}}(\theta_1) - G_\phi(\theta_1) &= \int_{\mathcal{X}} (\phi_{NP} - \phi)\, f_1\, d\mu \\ &\geq c \int_{\mathcal{X}} (\phi_{NP} - \phi)\, f_0\, d\mu \\ &= c\big(E_0\phi_{NP} - E_0\phi\big) \geq 0, \end{aligned} \tag{35}$$

da $E_0\phi_{NP} = \alpha$ und $E_0\phi \leq \alpha$. ∎

Beweis (*von Satz 5.4*)
Zu 1.: Wir betrachten den Likelihood Quotient

$$S(x) = \begin{cases} \frac{f_1(x)}{f_0(x)}, & f_0(x) > 0, \\ \infty, & f_0(x) = 0. \end{cases}$$

Aus $S(x)$ bilde den NP-Test mit Parameter $c \geq 0$, $\gamma \in [0,1]$

$$\phi_{NP}(x) = \begin{cases} 1, & S(x) > c, \\ \gamma, & S(x) = c \\ 0, & S(x) < c \end{cases} \tag{36}$$

Dann ist ϕ_{NP} konstant auf $\{f_1 = cf_0\} \setminus A$ mit $A = \{0 = f_1 = f_0\}$. Auf A können wir ϕ_{NP} aber noch umdefinieren $= \gamma$, ohne Umfang oder Güte zu verändern.

Wir zeigen nun, dass $c \geq 0$, $\gamma \in [0,1]$ existieren, für die der NP-Test (36) Umfang α hat, d.h.

$$G_{\phi_{NP}}(\theta_0) = P_0(S > c) + \gamma P_0(S = c) = \alpha,$$

dann folgt Teil 1. wegen der UMP-Eigenschaft aus Lemma 5.6. Sei $F(t) = P_0(S \leq t)$ die Verteilungsfunktion von S unter P_0. Dann ist obiger Display äquivalent zu

$$\alpha = 1 - F(c) + \gamma\big(F(c) - F(c-)\big),$$

welches mit

$$c = F^{-1}(1-\alpha), \qquad \gamma = \begin{cases} 0, & F(c) = F(c-), \\ \frac{F(c)-(1-\alpha)}{F(c)-F(c-)}, & \text{sonst} \end{cases}$$

gelöst wird.
2. Sei

$$C = \big\{\phi_{NP} \neq \phi\big\} \cap \big\{f_1 \neq cf_0\big\}.$$

Wir zeigen, das $\mu(C) = 0$ ist. Da beide Tests UMP sind, gilt $0 = G_{\phi_{NP}}(\theta_1) - G_\phi(\theta_1)$, also folgt aus (35)

$$\int_{\mathcal{X}} \big(\phi_{NP} - \phi\big)\,\big(f_1 - cf_0\big)\, d\mu = 0.$$

Der Integrand ist nach Lemma 5.5 nicht-negativ und nach Wahl von C daher auf C strikt positiv, also muss $\mu(C) = 0$ sein.
3. Da ϕ auch UMP Test ist, steht in (35) überall ein $=$. Somit folgt $E_0\phi = \alpha$, es sei denn $c = 0$ für den NP-Test zum Niveau α. Ist $c = 0$, so gilt

$$E_1\phi_{NP} = P_1(f_1 > 0) + \gamma P_1(f_1 = 0) = 1,$$

und daher auch $E_1\phi = 1$. ∎

Bemerkung. 1. Die Wahl von c ist i.A. nicht eindeutig, es muss als ein $(1-\alpha)$-Quantil von F gewählt werden. Die Wahl von γ ist dagegen eindeutig.
2. Falls $(1-\alpha) \in F_S(\mathbb{R})$, so gilt $F_S\big(F_S^{-1}(1-\alpha)\big) = 1-\alpha$, und man benötigt keine Randomisierung.
3. Geometrische Veranschaulichung.

Wir folgern noch:

Korollar 5.7
Ist ϕ UMP-Test für $H_0 : \theta = \theta_0$ gegen $H_1 : \theta = \theta_1$ zum Niveau $\alpha = E_{\theta_0} T$ mit $0 < \alpha < 1$, so gilt $E_{\theta_1} > \alpha$ oder $P_0 = P_1$.

Beweis
Der konstante Test $\tilde{\phi} = \alpha$ hat $E_{\theta_1}\tilde{\phi} = \alpha$, da ϕ UMP Test muss $E_{\theta_1}\phi \geq \alpha$. Ist nun $E_{\theta_1}\phi = \alpha$, so ist auch $\tilde{\phi}$ UMP Test, also nach Satz 5.4 2. μ-f.ü. $= 1$ auf $\{f_1 > cf_0\}$, und $= 0$ auf $\{f_1 < cf_0\}$ für ein $c \geq 0$. Da $0 < \alpha < 1$, ist daher offenbar $f_1 = cf_0$ μ-f.ü., und da beides Wahrscheinlichkeitsdichten sind, folgt $c = 1$, also $f_0 = f_1$ μ-f.ü. ∎

5.3 Tests für einparametrische Modelle

Wir betrachten ein durch μ dominiertes Modell $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ mit $\Theta \subset \mathbb{R}$, und wollen UMP-Tests bzw. UMPU Tests für zusammengesetzte Hypothesen und Alternativen konstruieren.

UMP Tests für einseitige Alternativen

Man kann UMP Tests für spezielle Modell für einseitige Hypothesen

$$H_0: \quad \theta \le \theta_0 \qquad \text{gegen} \qquad H_1: \theta > \theta_0 \tag{37}$$

zu gegebenen $\theta_0 \in \Theta$ konstruieren.

Definition Das dominierte Modell $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ mit Dichten $f(x;\theta)$ sei identifizierbar, und sei $T: \mathcal{X} \to \mathbb{R}$ eine Statistik. Dann hat $(P_\theta)_{\theta\in\Theta}$ einen *monotonen Dichtequotienten* in T, falls für alle $\theta, \theta' \in \Theta$, $\theta < \theta'$ eine monotone Funktion $g(\cdot;\theta,\theta'): \mathbb{R} \to [0,\infty]$ existiert, so dass

$$\frac{f(x,\theta')}{f(x,\theta)} = g\big(T(x),\theta,\theta'\big) \qquad \mu - \text{f.ü.} \tag{38}$$

Satz 5.8
Eine univariate Exponentialfamilie mit Dichte

$$f(x,\theta) = C(\theta)\, \exp\big(Q(\theta)\, T(x)\big)\, h(x),$$

wobei $\Theta \subset \mathbb{R}$ und $Q: \Theta \to \mathbb{R}$ streng monoton wachsend sei, sowie $T \neq const\ hd\mu$-f.ü., hat monotonen Dichtequotienten in T.

Beweis
O.E. $h = 1$. Seien $\theta, \theta' \in \Theta$, $\theta < \theta'$, und daher $Q(\theta') - Q(\theta) > 0$.

1. Identifizierbar: Falls $f(x,\theta) = f(x,\theta')$ μ-f.ü., so folgte

$$T(x) = \frac{\log C(\theta') - \log C(\theta)}{Q(\theta) - Q(\theta')} \quad \mu - \text{f.ü.}.$$

2. Die Funktion $g(t;\theta,\theta') = \exp\big(t(Q(\theta') - Q(\theta))\big)$ ist monoton wachsend und

$$\frac{f(x,\theta')}{f(x,\theta)} = \frac{C(\theta')}{C(\theta)}\, \exp\big(T(x)(Q(\theta') - Q(\theta))\big).$$

■

Für $c^* \in \mathbb{R}$, $\gamma^*: \mathcal{X} \to [0,1]$ messbar betrachten wir den Test

$$\phi^*(x) = \begin{cases} 1, & T(x) > c^*, \\ \gamma^*(x), & T(x) = c^* \\ 0, & T(x) < c^* \end{cases} \tag{39}$$

Satz 5.9
Das dominierte Modell $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ mit $\Theta \subset \mathbb{R}$ (sei identifizierbar und) habe monotonen Dichtequotienten in $T : \mathcal{X} \to \mathbb{R}$.
1. Ist für $\theta_0 \in \Theta$ und ϕ^* von der Form (48) mit $\alpha := E_{\theta_0}\phi^* > 0$, so ist ϕ^* UMP-Test für das Testproblem (37) zum Niveau α.
2. Für $\theta_0 \in \Theta$ und $\alpha \in (0, 1)$ existieren $c^* \in \mathbb{R}$ und $\gamma^* \in [0, 1]$ konstant, so dass der Test ϕ^* in (48) Umfang α hat, und daher UMP-Test für das Testproblem (37) ist.
3. Ist ϕ^* von der Form (48), so ist die Gütefunktion $G_{\phi^*}(\theta) = E_\theta(\phi^*)$ monoton wachsend in θ, und auf $G^{-1}_{\phi^*}(0, 1)$ streng monoton wachsend.

Beispiel. *(Einseitiger Gauß-Test)* Seien $X_1, \ldots, X_n$ u.i.v., $N(\mu, \sigma_0^2)$-verteilt, mit $\sigma_0^2 > 0$ bekannt. Diese bilden eine EF mit $Q(\mu) = \mu/\sigma_0^2$ und $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$. Um eine Hypothese $H_0 : \mu \le \mu_0$ gegen $H_1 : \mu > \mu_0$ zu testen, beachten wir dass $T \sim N(n\mu_0, n\sigma_0^2)$ verteilt ist unter P_{μ_0}, und daher wird der UMP Test gegeben durch

$$\phi^*(x) = \begin{cases} 1, & \sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma_0} > q_{1-\alpha} \\ 0, & \sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma_0} \le q_{1-\alpha} \end{cases} \tag{40}$$

mit $q_{1-\alpha}$ das $1-\alpha$-Quantil von $N(0, 1)$.

Beweis (*von Satz 5.9*)
Sei ϕ^* von der Form (48), und $\theta_0 \in \Theta$ mit $\alpha := E_{\theta_0}\phi^* > 0$. Wir fixieren zunächst ein $\theta_1 > \theta_0$ und betrachten das Testproblem

$$H_0' : \quad \theta = \theta_0 \qquad \text{gegen} \qquad H_1' : \quad \theta = \theta_1.$$

Behauptung. ϕ^* ist UMP-Test für H_0' gegen H_1' zum Niveau α.

Beweis der Behauptung. Für g wie in (38) sei $c = g(c^*; \theta_0, \theta_1)$, und $f_j(x) = f(x, \theta_j)$, $j = 0, 1$. Dann gelten

$$\begin{aligned} T(x) \ge c^* \quad &\Rightarrow \quad g\big(T(x); \theta_0, \theta_1\big) = f_1(x)/f_0(x) \ge c, \\ T(x) \le c^* \quad &\Rightarrow \quad g\big(T(x); \theta_0, \theta_1\big) = f_1(x)/f_0(x) \le c. \end{aligned}$$

Daher ist

$$\begin{aligned} \big\{T \ge c^*\big\} \subset \big\{f_1/f_0 \ge c\big\}, \qquad &\text{daher} \quad \big\{T < c^*\big\} \supset \big\{f_1/f_0 < c\big\}, \\ \big\{T \le c^*\big\} \subset \big\{f_1/f_0 \le c\big\}, \qquad &\text{daher} \quad \big\{T > c^*\big\} \supset \big\{f_1/f_0 > c\big\}. \end{aligned}$$

Es folgt

$$\begin{aligned} 0 < \alpha = E_{\theta_0}\phi^* &= P_{\theta_0}(T > c^*) + \int_{T=c^*} \gamma^*(x)\, dP_{\theta_0}(x) \\ &\le P_{\theta_0}(T \ge c^*) \le P_{\theta_0}(f_1/f_0 \ge c). \end{aligned}$$

Da $P_{\theta_0}(f_1/f_0 = \infty) = 0$, ist $c < \infty$. Wir können somit schreiben

$$\phi^*(x) = \begin{cases} 1, & f_1(x) > cf_0(x) (\Rightarrow\ T(x) > c^*), \\ \gamma(x), & f_1(x) = cf_0(x) \\ 0, & f_1(x) < cf_0(x) (\Rightarrow\ T(x) < c^*) \end{cases}$$

wobei $\gamma(x) \in \{0, 1, \gamma^*(x)\}$ ist. Also ist ϕ^* NP-Test, und daher insbesondere UMP zum Niveau α.

Zu 3.: Seien $\theta < \theta'$. Ist $E_\theta \phi^* = 0$, so ist die Monotonie klar, ist $0 < E_\theta \phi^* < 1$, so folgt mit obiger Behauptung, Korollar 5.7 und der Identifizierbarkeit $E_{\theta'} \phi^* > E_\theta \phi^*$, ist $E_\theta \phi^* = 1$, so folgt dies schließlich mit der UMP-Eigenschaft von NP-Test und Vergleich mit $\phi_c = 1$.

Zu 1.: Zunächst ist wegen der Monotonie der Gütefunktion

$$\sup_{\theta \leq \theta_0} E_\theta \phi^* = E_{\theta_0} \phi^* = \alpha.$$

Da ϕ^* UMP-Test für jede Alternative $H_1 : \theta = \theta_1$, für $\theta_1 > \theta_0$ ist, ist ϕ^* auch UMP für die zusammengesetzte Alternative $H_1 : \theta > \theta_1$.

Zu 2.: Man konstruiert sich c^*, γ^* über die Verteilungsfunktion von T unter P_{θ_0} wie im Beweis der Neyman-Pearson-Lemmas. ∎

UMPU Tests für zweiseitige Alternativen

Wir betrachten in diesem Abschnitt die Situation von Satz 5.8, d.h. eine einparametrische EF

$$f(x, \theta) = C(\theta) \exp\big(Q(\theta)\, T(x)\big)\, h(x),$$

wobei $\Theta \subset \mathbb{R}$ und $Q : \Theta \to \mathbb{R}$ streng monoton wachsend sei, sowie $T \neq const$ $hd\mu$-f.ü.

Bemerkung. Die Funktion $g(t; \theta, \theta')$ im Beweis von Satz 5.8 ist stetig und streng monoton. Daher ist nach dem Beweis von Satz 5.9 ein NP-Test für $H_0 : \theta = \theta_0$ gegen $H_0 : \theta = \theta_1$ für ein beliebiges Paar $\theta_0 < \theta_1$ stets von der Form (48), und hat damit eine streng-monoton wachsende Gütefunktion auf dem Urbild von $(0, 1)$.

Wir betrachten nun die zweiseitigen Testprobleme

$$(T1): \quad H_0: \quad \theta = \theta_0 \qquad \text{gegen} \qquad H_1: \quad \theta \neq \theta_0$$

für ein festes $\theta_0 \in \operatorname{Int} \Theta$, sowie

$$(T2): \quad H_0: \quad \theta_1 \leq \theta \leq \theta_2 \qquad \text{gegen} \qquad H_1: \quad \theta < \theta_1 \quad \text{oder} \quad \theta > \theta_2$$

für feste $\theta_1 < \theta_2$, $\theta_i \in \operatorname{Int} \Theta$.

Bemerkung. Wir bemerken zunächst, dass für diese Probleme kein UMP-Test zu einem Niveau $\alpha \in (0, 1)$ existiert. Wäre etwa ϕ UMP-Test für $(T1)$, so auch für $H_0 : \theta = \theta_0$ gegen $H_0 : \theta = \theta_1$ für ein $\theta_1 > \theta_0$, und hätte daher nach obiger Bemerkung eine streng monoton wachsende Gütefunktion, die für $\theta < \theta_0$ unter der des konstanten Tests $\phi_c = \alpha$ läge, Widerspruch.

Wir wollen daher für diese Testprobleme UMPU Tests konstruieren. Wegen der strengen Monotonie von Q können wir äquivalent zum natürlichen Parameter übergehen, und nehmen daher direkt $Q(\theta) = \theta$, also

$$f(x,\theta) = C(\theta)\, \exp\big(\theta\, T(x)\big)\, h(x),$$

an.

Nach Lemma 2.9 ist die Gütefunktion eines jeden Tests $\in C^\infty$. Daher muss für einen unverzerrten Test gelten

$$\text{Für } (T2): \quad E_{\theta_1}\phi = \alpha, \quad E_{\theta_2}\phi = \alpha. \tag{41}$$

Für $(T1)$ ist analog $E_{\theta_0}\phi = \alpha$, und die Gütefunktion hat bei θ_0 ihr globales Minimum, also $G'_\phi(\theta_0) = 0$. Mit

$$b(\theta) = \log\Big(\int_{\mathcal{X}} \exp\big(\theta\, T(x)\big)\, h(x)\, d\mu(x)\Big)$$

ist

$$G_\phi(\theta) = \int_{\mathcal{X}} \phi(x)\, \exp\big(\theta\, T(x) - b(\theta)\big)\, h(x)\, d\mu(x)$$

und wegen $b'(\theta) = E_\theta T$

$$\begin{aligned} G'_\phi(\theta) &= \int_{\mathcal{X}} \phi(x)\, T(x)\, \exp\big(\theta\, T(x) - b(\theta)\big)\, h(x)\, d\mu(x) \\ &\quad - \int_{\mathcal{X}} \phi(x)\, b'(\theta)\, \exp\big(\theta\, T(x) - b(\theta)\big)\, h(x)\, d\mu(x) \\ &= E_\theta\big(\phi T\big) - E_\theta T\, E_\theta \phi. \end{aligned}$$

Es ergibt sich für einen unverzerrten Test:

$$\text{Für } (T1): \quad E_{\theta_0}\phi = \alpha, \quad E_{\theta_0}\big(\phi T\big) = \alpha\, E_{\theta_0} T. \tag{42}$$

Bemerkung. Das *NP-Lemma* besagt insbesondere: Ist $0 < \alpha < 1$, so hat das Optimierungsproblem

$$\max \int_{\mathcal{X}} \phi f_1\, d\mu(x)$$

über

$$\phi : \mathcal{X} \to [0,1] \quad \text{Test mit} \qquad \int_{\mathcal{X}} \phi f_0\, d\mu(x) = \alpha$$

eine Lösung von der Form

$$\phi_{NP}(x) = \begin{cases} 1, & f_1(x) > c f_0(x), \\ 0, & f_1(x) < c f_0(x) \end{cases}$$

mit $c \geq 0$, und jede Lösung ist μ-f.ü. von dieser Gestalt.

Um mit den zwei Restriktionen in (42) und (41) umgehen zu können, benötigen wir folgende Verallgemeinerung.

Lemma 5.10 (*Verallgemeinertes NP-Lemma)*
Sei μ σ-endlich, $g_i : \mathcal{X} \to \mathbb{R}$ μ-integrierbar, $i = 1, 2, 3$. Betrachte das Optimierungsproblem

$$\max \int_{\mathcal{X}} \phi \, g_3 \, d\mu \tag{43}$$

über

$$\phi : \mathcal{X} \to [0, 1] \quad \text{Test mit} \qquad \int_{\mathcal{X}} \phi \, g_i \, d\mu = k_i, \ i = 1, 2, \tag{44}$$

für Konstante $k_1, k_2 \in \mathbb{R}$.
1. Ist für Konstante $c_1, c_2 \in \mathbb{R}$

$$\phi^*(x) = \begin{cases} 1, & g_3(x) > c_1 g_1(x) + c_2 g_2(x), \\ 0, & g_3(x) < c_1 g_1(x) + c_2 g_2(x) \end{cases} \tag{45}$$

und erfüllt ϕ^* (44), so ist ϕ^* eine Lösung von (43).
2. *Existenz.* Ist (k_1, k_2) im Inneren $\left\{(\int g_1\phi, \int g_2\phi), \ \phi : \mathcal{X} \to [0, 1]\right\}$, so existiert ein Test ϕ^* der Form (45), welcher (44) erfüllt.
3. *Eindeutigkeit.* Jede Lösung ϕ des Optimierungsproblems (43) unter (45), erfüllt für geeignete Konstante c_i (45) μ-f.ü.

Man zeigt dabei Teil 1. wie Lemmas 5.5 und 5.6, für die Teile 2. und 3. s. Lehmann.

Für die Testprobleme $(T1)$ und $(T2)$ setzen wir nun wie folgt an: Wähle $\gamma_i \in [0, 1]$, $c_i^* \in \mathbb{R}$, $i = 1, 2$, und setze

$$\phi^*(x) = \begin{cases} 1, & T(x) < c_1^* \text{ oder } T(x) > c_2^*, \\ \gamma_i, & T(x) = c_i^*, \ i = 1, 2, \\ 0, & c_1^* < T(x) < c_2^* \end{cases} \tag{46}$$

Satz 5.11
Sei $\alpha \in (0, 1)$.

1. a. Ist ϕ^* von der Form (46) und erfüllt ϕ^* die Bedingungen (42), so ist ϕ^* gleichmäßig bester Test in der Klasse der Tests, die die Bedingungen (42) erfüllen, und daher insbesondere UMPU-Test für $(T1)$.
b. Ist ϕ^* von der Form (46) und erfüllt ϕ^* die Bedingungen (41), so ist ϕ^* gleichmäßig bester Test in der Klasse der Tests, die die Bedingungen (41) erfüllen, und daher insbesondere UMPU-Test für $(T2)$.

2. Es existiert ein Test der Form (46), der (42) erfüllt, und analog existiert ein Test der Form (46), der (41) erfüllt. Die Konstanten c_i^* und γ_i werden dabei als Lösungen von (42) bzw. (41) bestimmt.

3. Jeder UMPU-Test ϕ für $(T1)$ bzw. $(T2)$ erfüllt für Konstante c_i^* μ-f.ü.

$$\phi(x) = \begin{cases} 1, & T(x) < c_1^* \text{ oder } T(x) > c_2^*, \\ 0, & c_1^* < T(x) < c_2^* \end{cases}$$

Beweis
Wir zeigen Teil 1. für $(T1)$, und weisen dafür nach, dass der Test (46) für jedes $\theta_1 \neq \theta_0$ für geeignet gewählte c_1, c_2 (abhängig von θ_1) die Gestalt des verallgemeinerten NP-Lemmas hat, wobei $g_3 = f_1$, $g_1 = f_0$ $g_2 = f_0 T$, $k_1 = \alpha$, $k_2 = \alpha E_0 T$.
Für $b = \log(C(\theta_1) - C(\theta_0))$ ist $f_1(x) > c_1 f_0(x) + c_2 f_0(x) T(x)$ äquivalent zu

$$\exp\big((\theta_1 - \theta_0)T(x) + b\big) > c_1 + c_2 T(x).$$

Wir müssen daher c_1, c_2 derart wählen, dass gilt

$$t \notin [c_1^*, c_2^*] \Leftrightarrow \exp\big((\theta_1 - \theta_0)t + b\big) > c_1 + c_2 t.$$

Wegen der strikten Konvexität von $\exp\big((\theta_1 - \theta_0)t + b\big)$ gilt dies, falls c_1, c_2 derart gewählt werden, dass die Gerade $c_1 + c_2 t$ die Funktion $\exp\big((\theta_1 - \theta_0)t + b\big)$ genau an den Stellen c_1^*, c_2^* schneidet. ∎

Beispiel. *(Zweiseitiger Poisson Test)* Für $X_1, \ldots, X_n$ u.i.v. Poi(λ) betrachten wir die zweiseitige Hypothese $H_0 : \lambda = \lambda_0$ gegen $H_1 : \lambda \neq \lambda_0$. Da $X_1 + \ldots + X_n \sim \text{Poi}(n\lambda)$ suffizient für λ, genügt es $n = 1$ zu betrachten, es werde also $X \sim \text{Poi}(\lambda)$ beobachtet. Es ist $T(x) = x$, und daher sind $C_1 < C_2$, $C_i \in \mathbb{N}_0$ sowie $\gamma_i \in [0,1]$, $i = 1, 2$, derart zu wählen, dass gelten

$$\sum_{k=0}^{C_1 - 1} \frac{\lambda_0^k}{k!} + \sum_{k=C_2+1}^{\infty} \frac{\lambda_0^k}{k!} + \sum_{i=1,2} \gamma_i \frac{\lambda_0^{C_i}}{C_i!} = \alpha e^{\lambda_0},$$

sowie

$$\sum_{k=0}^{C_1 - 1} k\, \frac{\lambda_0^k}{k!} + \sum_{k=C_2+1}^{\infty} k\, \frac{\lambda_0^k}{k!} + \sum_{i=1,2} \gamma_i\, C_i\, \frac{\lambda_0^{C_i}}{C_i!} = \alpha \lambda_0\, e^{\lambda_0}$$

erfüllt sind.

Bemerkung. Ist $h : \mathbb{R} \to \mathbb{R}$ streng monoton wachsend, so kann ϕ^* in (46) offenbar auch in $h(T)$ formuliert werden.
Ist speziell $\tilde{T} = aT + b$, $a > 0$, so kann für $E_{\theta_0}\phi = \alpha$ die Bedingung $E_{\theta_0}\big(\phi T\big) = \alpha\, E_{\theta_0} T$ auch äquivalent für $\tilde{T}$ nachgeprüft werden, denn

$$E_{\theta_0}\big(\phi \tilde{T}\big) = b\,\alpha + a E_{\theta_0}\phi T = b\alpha + a\alpha E_{\theta_0} T = \alpha E_{\theta_0} \tilde{T}.$$

Bemerkung. Ist für $(T1)$ die Statistik T unter P_{θ_0} symmetrisch um t_0 verteilt, also

$$P_{\theta_0}(T - t_0 \leq -t) = P_{\theta_0}(T - t_0 \geq t), \qquad t \in \mathbb{R},$$

und $0 < \alpha < 1$, so setzen wir

$$\tilde{\phi}^*(x) = \begin{cases} 1, & |T(x) - t_0| > c^*, \\ \gamma^*, & |T(x) - t_0| = c^*, \\ 0, & |T(x) - t_0| < c^* \end{cases} \tag{47}$$

wobei $c^* > 0$, $\gamma^* \in [0,1]$ als Lösung der Gleichung

$$P_{\theta_0}(T - t_0 > c^*) + \gamma^*\, P_{\theta_0}(T - t_0 = c^*) = \frac{\alpha}{2}$$

gewählt sind (beachte $P_{\theta_0}(T \geq 0) \geq 1/2$). Wir zeigen, dass $\tilde{\phi}^*$ (42) erfüllt. Zunächst ist

$$E_{\theta_0}\tilde{\phi}^* = P_{\theta_0}(|T - t_0| > c^*) + \gamma^* P_{\theta_0}(|T - t_0| = c^*) = \alpha.$$

Weiter ist

$$E_{\theta_0}\tilde{\phi}^* T = E_{\theta_0}\tilde{\phi}^*(T - t_0) + \alpha\, t_0 = \alpha E_{\theta_0} T.$$

Hierfür bemerken wir zunächst $E_{\theta_0} T = t_0$ wegen der Symmetrie. Mit $\tilde{\varphi}^*(T(x)) = \tilde{\phi}^*(x)$ ist $g(t) = (t - t_0)\tilde{\varphi}^*(t)$ schiefsymmetrisch um t_0: $g(t_0 + \delta) = -g(t_0 - \delta)$, $\delta > 0$, und daher

$$E_{\theta_0}\tilde{\phi}^*(T - t_0) = \int_{\mathbb{R}} g(t)\, dP^T_{\theta_0}(t) = 0,$$

da $P^T_{\theta_0}$ eine um t_0 symmetrische Verteilung ist.

Beispiel. *(Zweiseitiger Gauß-Test)* Seien $X_1, \ldots, X_n$ u.i.v., $N(\mu, \sigma_0^2)$-verteilt, mit $\sigma_0^2 > 0$ bekannt. Diese bilden eine EF mit $Q(\mu) = \mu/\sigma_0^2$ und $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$. Um eine Hypothese $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$ zu testen, beachten wir dass $T \sim N(n\mu_0, n\sigma_0^2)$ um $n\mu_0$-symmetrisch verteilt ist unter P_{μ_0}, und daher wird der UMPU Test gegeben durch

$$\phi^*(x) = \begin{cases} 1, & \sqrt{n}\,\frac{|\bar{X}_n - \mu_0|}{\sigma_0} > q_{1-\alpha/2} \\ 0, & \sqrt{n}\,\frac{|\bar{X}_n - \mu_0|}{\sigma_0} \leq q_{1-\alpha/2} \end{cases} \tag{48}$$

5.4 Bedingte Tests in mehrparametrigen Exponentialfamilien

Mehrparametrische Exponentialfamilien und bedingte Verteilungen

Wir betrachten eine $r+s$-parametrische Exponentialfamilie mit Verteilungen

$$dP_{(\theta,\xi)}(x) = C(\theta,\xi)\,\exp\left(U(x)^T\theta + T(x)^T\xi\right)\,d\mu(x).$$

Für die Verteilungen von (U,T) auf $\mathbb{R}^{r+s}$ gilt dann

$$dP^{(U,T)}_{(\theta,\xi)}(u,t) = C(\theta,\xi)\,\exp\left(u^T\theta + t^T\xi\right)\,d\mu^{(U,T)}(u,t).$$

Satz 5.12
1. Für alle θ existiert ein σ-endliches Maß λ_θ (auf $(\mathbb{R}^s,\mathcal{B}^s)$), s.d. für die marginale Verteilung von T unter $P_{(\theta,\xi)}$ gilt

$$dP^{(T)}_{(\theta,\xi)}(t) = C(\theta,\xi)\,\exp\left(t^T\xi\right)\,d\lambda_\theta(t).$$

2. Für jedes t existiert ein σ-endliches Maß ν_t auf (auf $(\mathbb{R}^r,\mathcal{B}^r)$), so dass für die bedingte Verteilung von U gegeben $T=t$ unter $P_{(\theta,\xi)}$ gilt

$$dP^{(U|T=t)}_{(\theta,\xi)}(u) = C_t(\theta)\,\exp\left(u^T\theta\right)\,d\nu_t(u),$$

wobei $C_t(\theta)$ eine geeignete Normierungskonstante ist.

Beweis
Wir fixieren (θ_0,ξ_0), und können schreiben

$$dP^{(U,T)}_{(\theta,\xi)}(u,t) = \frac{C(\theta,\xi)}{C(\theta_0,\xi_0)}\,\exp\left(u^T(\theta-\theta_0)\right)\,\exp\left(t^T(\xi-\xi_0)\right)\,dP^{(U,T)}_{(\theta_0,\xi_0)}(u,t).$$

Die Behauptung ergibt sich dann aus folgendem Lemma.

Lemma 5.13
Sind P_0, P_1, Verteilungen auf $\mathbb{R}^{r+s}$ mit

$$\frac{dP_1}{dP_0}(u,t) = a(u)\,b(t),$$

wobei $a(u)>0$ für alle u gelte, und sei (U,T) die Identität auf $\mathbb{R}^{r+s}$, so gilt
1. für die marginale Verteilung von T

$$\frac{dP_1^T(t)}{dP_0^T(t)} = b(t)\,\int_{\mathbb{R}^r} a(u)\,dP_0^{U|T=t}(u),$$

2. für die bedingte Verteilung von U gegeben $T=t$

$$\frac{dP_1^{U|T=t}(u)}{dP_0^{U|T=t}(u)} = \frac{a(u)}{\int_{\mathbb{R}^r} a(u)\,dP_0^{U|T=t}(u)}.$$

Beweis
Zu 1.: Für $B \in \mathcal{B}^s$ gilt

$$\begin{aligned} P_1(T \in B) &= E_1 1_B(T) = E_0\big(1_B(T)a(U)b(T)\big) \\ &= E_0\big(E_0\big(1_B(T)a(U)b(T)|T\big)\big) \\ &= E_0\big(1_B(T)\, b(T)\, E_0\big(a(U)|T\big)\big) \\ &= \int_B b(t)\, \Big(\int_{\mathbb{R}^r} a(u)\, dP_0^{U|T=t}(u)\Big)\, dP_0^T(t) \end{aligned}$$

Zu 2.: Für jedes t liegt offenbar eine Verteilung vor, es genügt daher, die definierende Eigenschaft für $E_1(1_A(U)|T)$ nachzuprüfen, wobei $A \in \mathcal{B}^r$. Sei dazu $B \in \mathcal{B}^s$, dann ist nach Teil 1.

$$\begin{aligned} E_1\big(1_A(U)\, 1_B(T)\big) &= E_0\big(1_B(T)\, b(T)\, E_0\big(a(U) 1_A(U)\, |T\big)\big) \\ &= \int_B b(t)\, \Big(\int_A a(u)\, dP_0^{U|T=t}(u)\Big)\, dP_0^T(t) \\ &= \int_B \frac{\int_A a(u)\, dP_0^{U|T=t}(u)}{\int_{\mathbb{R}^r} a(u)\, dP_0^{U|T=t}(u)}\, dP_1^T(t) \end{aligned}$$

und wegen

$$E_1\big(1_A(U)\, 1_B(T)\big) = \int_B P_1^{U|T=t}(A)\, dP_1^T(t)$$

folgt 2. ∎

Die Aussage des Satzes ergibt sich nun mit

$$\begin{aligned} d\lambda_\theta(t) &= \frac{1}{C(\theta_0, \xi_0)}\, \exp\big(t^T \xi_0\big) \int_{\mathbb{R}^r} \exp\big(u^T(\theta - \theta_0)\big)\, dP_{(\theta_0,\xi_0)}^{U|T=t}(u)\, dP_{(\theta_0,\xi_0)}^T(t), \\ d\nu_t(u) &= \exp\big(-\, u^T\theta_0\big)\, dP_{(\theta_0,\xi_0)}^{U|T=t}(u). \end{aligned}$$ ∎

Neyman-Struktur und Ähnlichkeit von Tests

Sei $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$, $\Theta \subset \mathbb{R}^k$ messbar, ein statistisches Modell.

Sei $\Theta' \subset \Theta$, $\alpha \in [0,1]$. Ein Test $\phi : \mathcal{X} \to [0,1]$ heißt α-ähnlich auf Θ', falls $E_\theta\phi = \alpha$ für alle $\theta \in \Theta'$ gilt.

Lemma 5.14
Angenommen, das Modell ist derart, dass jeder Test $\phi : \mathcal{X} \to [0,1]$ eine stetige Gütefunktion habe. Sei $\Theta = \Theta_0 \cup \Theta_1$ und $\Theta' = \bar{\Theta}_0 \cap \bar{\Theta}_1$ (Abschlüsse in Θ der Rand zwischen Hypothese und Alternative. Dann ist jeder unverfälschte Test ϕ zum Niveau α auch α-ähnlich auf Θ'.

Lemma 5.15
Sei $\Theta' \subset \Theta$, $\alpha \in [0,1]$, und sei T eine für das Teilmodell $\big(\mathcal{X}, \mathcal{F}, (P_\theta)_{\theta\in\Theta'}\big)$ suffiziente und vollständige Statistik. Ist $\phi : \mathcal{X} \to [0,1]$ α-ähnlich auf Θ', so gilt

$$E.(\phi|T) = \alpha \qquad P_\theta - \text{f.s.} \quad \forall \quad \theta \in \Theta'. \tag{49}$$

Beweis
Es gilt

$$\begin{aligned} \forall\ \theta \in \Theta' : &\quad E_\theta(\phi - \alpha) = 0 && \alpha - (\text{Ähnlichkeit}) \\ \Rightarrow \forall\ \theta \in \Theta' : &\quad E_\theta(E_\cdot(\phi - \alpha)|T) = 0 && (\text{Suffizienz}) \\ \Rightarrow \forall\ \theta \in \Theta' : &\quad E_\cdot\big((\phi - \alpha)|T\big) = 0 \quad P_\theta - \text{f.s.} && (\text{Vollständigkeit}) \end{aligned}$$

■

Ein Test, der (49) erfüllt, hat *α-Neyman-Struktur* auf Θ' bzgl. T. Dann ist der Test offenbar α-ähnlich auf Θ'.

Bedingte Tests

Wir betrachten nun eine $k+1$-parametrische EF im natürlichen Parameter:

$$dP_{(\theta,\xi)}(x) = C(\theta,\xi)\, \exp\big(\theta\, U(x) + < \xi, T(x) > \big)\, d\mu(x), \tag{50}$$

$(\theta, \xi) \in \Theta \subset \mathbb{R}^{k+1}$, und wollen Hypothesen an den eindimensionalen Parameter θ der Form

$$\begin{aligned} (OS): &\quad H_0 : \theta \le \theta_0 \quad \text{gegen} \quad H_0 : \theta > \theta_0, \\ (TS): &\quad H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_0 : \theta \neq \theta_0, \end{aligned}$$

testen, dabei sind die ξ sogenannte Störparameter (Nuisance Parameter), die weder unter Hypothese noch Alternative restringiert sind.

Wir betrachten nun zunächst (OS). Es stellt sich heraus, dass es im Falle eines Nuisance Parameters auch im einseitigen Fall im allgemeinen keinen UMP Test mehr gibt (REFERENZ LEHMANN PAPER). Man kann aber UMPU Tests wie folgt konstruieren.

Die bedingte Verteilung von U gegeben $T = t$ ist eine einparametrische EF in θ (und unabhängig von ξ), daher kann bedingt auf $T = t$ ein UMP-Test für (OS), basierend auf U, der Form

$$\varphi^*(u;t) = \begin{cases} 1, & u > c(t), \\ \gamma(t), & u = c(t), \\ 0, & u < c(t) \end{cases} \tag{51}$$

konstruiert werden, wobei $c(t), \gamma(t)$ bestimmt sind durch

$$E_{\theta_0}\big(\varphi^*(U;T)|T = t\big) = \alpha \qquad \forall\ t. \tag{52}$$

Satz 5.16
Für $\alpha \in (0,1)$ existiert im Modell (50) ein Test $\varphi^*(u;t)$ der Form (51), der (52) für alle t erfüllt und für den $c(t)$ sowie $\gamma(t)$ messbare Funktionen sind, und $\phi^*(x) = \varphi^*(U(x);T(x))$ ist UMPU-Test für das Testproblem (OS).

Beweis
Für die messbare Wahl der Funktionen $c(t)$ und $\gamma(t)$ s. Lehmann.

Sei $\phi : \mathcal{X} \to [0,1]$ ein Test für (OS). Durch Bedingen auf (U,T) (Suffizienz) können wir annehmen, dass $\phi(x) = \varphi(U(x),T(x))$. Dann gilt

$$E_{(\theta,\xi)}\phi = \int \Big(\int \varphi(u,t)\, dP_\theta^{U|T=t}(u) \Big)\, dP^T_{(\theta,\xi)}(t). \tag{53}$$

Da ein UMP-Test stets unverfälscht ist (Vergleich mit konstantem Test), erhält man unmittelbar, das ϕ^* ein unverfälschter Test zum Niveau α ist.

Setze nun $\Theta' = \{(\theta_0,\xi) \in \Theta\} = \bar\Theta_0 \cap \bar\Theta_1$ und sei $\varphi(u,t)$ ein beliebiger unverfälschter Test. Da die Gütefunktion stetig ist, ist $\varphi(u,t)$ α-ähnlich auf Θ', und da T suffizient und vollständig für Θ', hat $\varphi(u,t)$ Neyman-Struktur auf Θ', d.h. $\varphi(u,t)$ erfüllt

$$E_{\theta_0}\big(\varphi(U;T)|T=t\big) = \alpha \qquad \forall\, t,$$

und hat somit bedingt auf $T=t$ Niveau α. Nach Konstruktion von φ^* als bedingter UMP-Test folgt für alle $\theta \neq \theta_0$

$$E_\theta\big(\varphi(U;T)|T=t\big) \le E_\theta\big(\varphi^*(U;T)|T=t\big),$$

und aus (53) daher die UMPU-Eigenschaft von φ^*. ∎

NOCH: ANNAHME ÜBER DAS INNERE!

Analog kann bedingt auf $T=t$ ein UMPU-Test für (TS), basierend auf U, der Form

$$\varphi^{\#}(u;t) = \begin{cases} 1, & u < c_1(t) \text{ oder } c_2(t) < u, \\ \gamma_i(t), & u = c_i(t),\ i=1,2, \\ 0, & c_1(t) < u < c_2(t) \end{cases} \tag{54}$$

konstruiert werden, wobei $c_i(t), \gamma_i(t)$ bestimmt sind durch

$$E_{\theta_0}\big(\varphi^{\#}(U;T)|T=t\big) = \alpha, \qquad E_{\theta_0}\big(U\,\varphi^{\#}(U;T)|T=t\big) = \alpha E_{\theta_0}\big(U\,|T=t\big)\ \forall\, t. \tag{55}$$

Satz 5.17
Für $\alpha \in (0,1)$ existiert im Modell (50) ein Test $\varphi^{\#}(u;t)$ der Form (51), der (52) für alle t erfüllt und für den $c_i(t)$ sowie $\gamma_i(t)$, $i=1,2$, messbare Funktionen sind, und $\phi^{\#}(x) = \varphi^{\#}(U(x);T(x))$ ist UMPU-Test für das Testproblem (TS).

Beweis
Sei ϕ ein unverzerrter Test für (TS), wir können wieder annehmen, dass $\phi(x) = \varphi(U(x),T(x))$. Wie oben gilt dann

$$E_{\theta_0}\big(\varphi(U;T)|T=t\big) = \alpha \qquad \forall\, t.$$

Hält man ξ fest und variiert θ, so ergibt sich aus der Unverzerrtheit, dass $\theta \mapsto E_{(\theta,\xi)}\varphi$ bei $\theta = \theta_0$ ein Minimum hat. Ableiten nach θ ergibt

$$E_{(\theta_0,\xi)}\big(U\,\varphi(U,T)\big) = \alpha\, E_{(\theta_0,\xi)}U.$$

Da T wieder suffizient und vollständig für das Teilmodell $\{(\theta_0, \xi) \in \Theta\}$ ist, ergibt sich wie oben

$$E_{(\theta_0)}\big(U\,\varphi(U,T)|T=t\big) = \alpha\, E_{(\theta_0,\xi)}(U|T=t).$$

Nach Konstruktion von $\varphi^{\#}$ ist für alle $\theta \neq \theta_0$

$$E_{(\theta_0)}\big(\varphi^{\#}(U,T)|T=t\big) \geq E_{(\theta_0)}\big(\varphi(U,T)|T=t\big),$$

und daher

$$\begin{aligned} E_{(\theta_0,\xi)}\big(\varphi^{\#}(U,T)\big) &= E_{(\theta_0,\xi)}\big(E_{\theta_0}\big(\varphi^{\#}(U,T)|T=t\big)\big) \\ &\geq E_{(\theta_0,\xi)}\big(E_{\theta_0}\big(\varphi(U,T)|T=t\big)\big) \\ &= E_{(\theta_0,\xi)}\big(\varphi(U,T)\big) \end{aligned}$$

■

Beispiel. Seien $X \sim Poi(\lambda)$, $Y \sim Poi(\mu)$, X und Y unabhängig. Dann ist

$$\begin{aligned} P_{\lambda,\mu}(X=x, Y=y) &= \frac{e^{-(\lambda+\mu)}}{x!\,y!}\,\exp\big(x\log\lambda + y\log\mu\big) \\ &= \frac{e^{-(\lambda+\mu)}}{x!\,y!}\,\exp\big(y\log\frac{\mu}{\lambda} + (x+y)\,\log\lambda\big) \end{aligned}$$

Nach dem Satz existiert ein UMPU Test für einseitige Hypothesen bzgl. $\log\frac{\mu}{\lambda}$ bzw. auch $\theta = \frac{\mu}{\lambda}$, insbesondere für

$$H_0: \quad \theta \leq 1 \quad \Leftrightarrow \quad \mu \leq \lambda.$$

Die bedingte Verteilung von Y gegeben $X+Y=t$ unter $P_{\lambda,\mu}$ ist $Bin(t,p)$ wobei $p = \mu/(\lambda+\mu) = (\theta+1)^{-1}$. Da H_0 äquivalent zu $p \leq 1/2$ ist, ist der UMPU Test der Binomialtest unter der bedingten Verteilung für diese Hypothese. ◇

Wir untersuchen nun, wann man die Tests in (51) und (54) als unbedingte Tests schreiben kann.

Sei dazu in dem Modell (50) $V = h(U,T)$ eine unter jedem $P_{\theta_0,\xi}$ von T unabhängige Statistik, somit ist die bedingte Verteilung von V gegeben $T=t$ gleich der unbedingten Verteilung von V für jedes $P_{\theta_0,\xi}$. Da wegen der Suffizienz von T die bedingte Verteilung von V gegeben $T=t$ nicht von $P_{\theta_0,\xi}$ abhängt, hängt also die Verteilung von V selbst nicht von ξ ab.

Satz 5.18
Sei in dem Modell (50) $V = h(U,T)$ eine unter jedem $P_{\theta_0,\xi}$ von T unabhängige Statistik, und $\alpha \in (0,1)$.
1. Ist $h(\cdot,t)$ streng monoton wachsend für jedes feste t, so ist

$$\varphi_1^*(v) = \begin{cases} 1, & v > \tilde{c}, \\ \tilde{\gamma}, & v = c, \\ 0, & v < c \end{cases} \tag{56}$$

mit $E_{\theta_0}\varphi_1^*(V) = \alpha$ UMPU-Test zum Niveau α für (OS).
2. Ist $h(u,t) = a(t)u + b(t)$, $a(t) > 0$ für alle t, so ist

$$\varphi_1^{\#}(v) = \begin{cases} 1, & v < \tilde{c}_1 \text{ oder } v > \tilde{c}_2, \\ \tilde{\gamma}_i, & v = \tilde{c}_i,\ i = 1,2, \\ 0, & \tilde{c}_1 < v < \tilde{c}_2 \end{cases} \tag{57}$$

mit

$$E_{\theta_0}\varphi_1^{\#}(V) = \alpha, \qquad E_{\theta_0} V\, \varphi_1^{\#}(V) = \alpha\, E_{\theta_0} V$$

UMPU-Test zum Niveau α für (TS).

Beweis
Zu 1.: Der UMPU Test ist gegeben durch (51). Da $h(\cdot,t)$ streng monoton wachsend ist, kann dieser Test auch umgeschrieben werden in v zu einem Test φ_1^* mit zugehörigen Funktionen $\tilde{c}(t)$, $\tilde{\gamma}(t)$, die bestimmt werden durch $E_{\theta_0}\big(\varphi_1^*(V,T)|T=t\big) = \alpha$. Wegen der Unabhängigkeit können die Konstanten $\tilde{c}(t) = \tilde{c}$ und $\tilde{\gamma}(t) = \tilde{\gamma}$ global bestimmt werden.
Zu 2.: Der UMPU Test ist gegeben durch (54). Da $h(u,t)$ lineare und streng monoton wachsend in u, kann der Test umgeschrieben werden (vgl. Bemerkung ???) zu einem Test $\varphi_1^{\#}$ in v, dessen Konstanten durch

$$E_{\theta_0}\big(\varphi_1^{\#}(V,T)|T=t\big) = \alpha, \qquad E_{\theta_0}\big(V\,\varphi_1^{\#}(V,T)|T=t\big) = \alpha\, E_{\theta_0}(V|T=t)$$

bestimmt werden, welches wegen der Unabhängigkeit wieder global, also unabhängig von T möglich ist. ∎

Beispiel: Ein-Stichproben t-test Seien $X_1,\ldots,X_n$ u.i.v. $N(\mu,\sigma^2)$, $(\theta,\sigma^2) \in \mathbb{R}\times(0,\infty)$. Es hat $(X_1,\ldots,X_n)$ eine EF-Verteilung mit natürlichen Parametern $\theta = \mu/\sigma^2$ und $\xi = -1/(2\sigma^2)$, und natürlichen suffizienten Statistiken $U(x) = x_1+\ldots+x_n$, $T(x) = x_1^2+\ldots+x_n^2$. Betrachte die Testprobleme

$$\begin{aligned} (OS): &\quad H_0: \mu \le \mu_0 \quad \text{gegen} \quad H_0: \mu > \mu_0, \\ (TS): &\quad H_0: \mu = \mu_0 \quad \text{gegen} \quad H_0: \mu \ne \mu_0. \end{aligned}$$

Wir können o.E. annehmen $\mu_0 = 0$, sonst Übergang zu $X_i - \mu_0$. Dann sind die Testprobleme äquivalent zu

$$\begin{aligned} (OS): &\quad H_0: \theta \le 0 \quad \text{gegen} \quad H_0: \theta > 0, \\ (TS): &\quad H_0: \theta = 0 \quad \text{gegen} \quad H_0: \theta \ne 0. \end{aligned}$$

Betrachte die t-Statistik

$$V = \sqrt{n}\,\frac{\bar{X}_n}{S_n} = \frac{1}{\sqrt{n}}\,\frac{U}{\left(\frac{T-U^2/n}{n-1}\right)^{1/2}} =: h(U,T).$$

$V \sim t_{n-1}$ und damit insbesondere anzillär für $(P_{0,\xi})$, und T ist suffizient und vollständig für $(P_{0,\xi})$, daher sind nach dem Satz von Basu T und V unabhängig unter allen $(P_{0,\xi})$.

MONOTONIE VON $h(u,t)$ NACHRECHNEN!

Wir erhalten somit als UMPU-Test für (OS) den einseitigen t-Test.

Für den zweiseitigen Test ist zunächst $h(u,t)$ nicht linear in u. Wir betrachten daher $\tilde{v} = u/\sqrt{t}$, dabei haben wir schon gesehen, dass $\tilde{V}$ und T unabhängig unter jedem $P_{0,\xi}$ sind. Hier ist $\tilde{h}$ offenbar linear in u, und da die Verteilung von $\tilde{V}$ unter $P_{0,\xi}$ symmetrisch um 0 ist, erhalten wir den UMPU Test von der Form $\varphi_1^{\#}(\tilde{v}) = 1_{\tilde{v}>c}$, wobei c als $1-\alpha/2$ Quantil der Verteilung von $\tilde{V}$ (unter $N(0,1)$) gewählt wird.

Da $v = g(\tilde{v})$, wobei

$$g(t) = \sqrt{\frac{n-1}{n}}\,\frac{t}{(1-t^2/n)^{1/2}},$$

wobei g schiefsymmetrisch und streng monoton wachsend auf $[-\sqrt{n},\sqrt{n}]$ ist, und da wegen Cauchy-Schwarz $|\tilde{V}| \leq \sqrt{n}$, kann der Test $\varphi_1^{\#}$ auch in V, also als zweiseitiger t-Test, formuliert werden.