Notizen zur Vorlesung Statistik
WS 2019/20, JGU Mainz

Matthias Birkner

Vorläufige Version, Stand 16. Februar 2020

Kommentare, Korrekturvorschläge, Hinweise auf (Tipp- und sonstige)
Fehler gerne per Email an `birkner@mathematik.uni-mainz.de` senden

Inhaltsverzeichnis

Kapitel 0

Auftakt

Ein Beispiel. 48 Teilnehmern eines Management-Kurses wurde je eine (fiktive) Personalakte vorgelegt, und sie sollten anhand der Aktenlage entscheiden, ob sie die betreffende Person befördern oder die Akte zunächst ablegen und weitere Kandidaten begutachten würden. Die Akten waren identisch bis auf die Geschlechtsangabe — 24 waren als „weiblich“ und 24 als „männlich“ gekennzeichnet — und wurden rein zufällig an die Teilnehmer verteilt.

Es kam zu folgendem Ergebnis[1]:

	Weiblich	Männlich
Befördern	14	21
Ablegen	10	3

Kann diese Aufteilung durch „reinen Zufall“ erklärt werden?

Betrachten wir folgendes „Ersatzexperiment“:

Die 48 Gutachter, 35 „wohlgesonnene“ und 13 „strenge“, ziehen ohne Zurücklegen je eine Akte aus einer Urne mit 24 „weiblichen“ und 24 „männlichen“ Akten, dann ist

$$X := \text{Anz. beförderter „männlicher“ Akten } \sim \mathrm{Hyp}_{24,24,35}$$

und die W'keit, eine so große Abweichung wie die tatsächlich gesehene im Ersatzexperiment zu beobachten, wäre

$$P(X \geq 21) + P(X \leq 14) = \mathrm{Hyp}_{24,24,35}\big(\{11,12,13,14\} \cup \{21,22,23,24\}\big) \approxeq 0{,}049$$

(dies ist der sogenannte p-Wert des Tests).

Somit: Die Hypothese „die Aufteilung entsteht durch reinen Zufall“ kommt uns unplausibel vor.

(Im Statistik-Jargon: „Wir werfen diese (Null-)Hypothese.“)

Bemerkung. Das obige Testverfahren heißt „Fishers exakter Test“.

[1] Aus Benson Rosen, Thomas H. Jerdee, Influence of sex role stereotypes on personnel decisions, J. Appl. Psych. **59**, 9–14, 1974; siehe Table 1 dort (nur der Teil “simple job”)

Kapitel 1

Rekapitulation (und Ergänzung) grundlegender Konzepte aus „Einführung in die Stochastik“

Definition 1. Ein *statistisches Modell* ist ein Tripel $(\mathcal{M} =) \left(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}\right)$, wo $\Omega \neq \varnothing$ Menge („Beobachtungs- oder Stichprobenraum“), $\mathscr{F} \subset 2^\Omega$ eine σ-Algebra, Θ eine Menge (mit $|\Theta| > 1$) und für jedes $\vartheta \in \Theta$ ist $\mathbb{P}_\vartheta$ ein W'maß auf $(\Omega, \mathscr{A})$.

Das Modell $\mathcal{M}$ heißt *parametrisch*, wenn $\Theta \subset \mathbb{R}^d$ für ein $d \in \mathbb{N}$, speziell *einparametrisch*, wenn $d = 1$.

$\mathcal{M}$ heißt *diskret*, wenn Ω abzählbar ist, $\mathcal{M}$ heißt *stetig*, wenn $\Omega \subset \mathbb{R}^n$ und jedes $\mathbb{P}_\vartheta$ eine Dichte $\rho_\vartheta : \Omega \to [0, \infty]$ besitzt.

Ein diskretes oder stetiges Modell heißt ein *Standardmodell*.

Definition 2. $\left(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}\right)$ statistisches Modell, $(S, \mathscr{A})$ messbarer Raum.

1. Eine Zufallsvariable X (definiert auf $(\Omega, \mathscr{F})$ mit Werten in S, d.h. $X : \Omega \to S$ ist $\mathscr{F}$-$\mathscr{A}$-messbar) heißt eine *Statistik* (manchmal auch: „Stichprobe“).

2. Sei $\tau : \Theta \to \mathbb{R}$ eine reelle Kenngröße (oder „Parametermerkmal“), eine Statistik $T : \Omega \to \mathbb{R}$ heißt ein *Schätzer* (genauer: „Punktschätzer“) für τ.

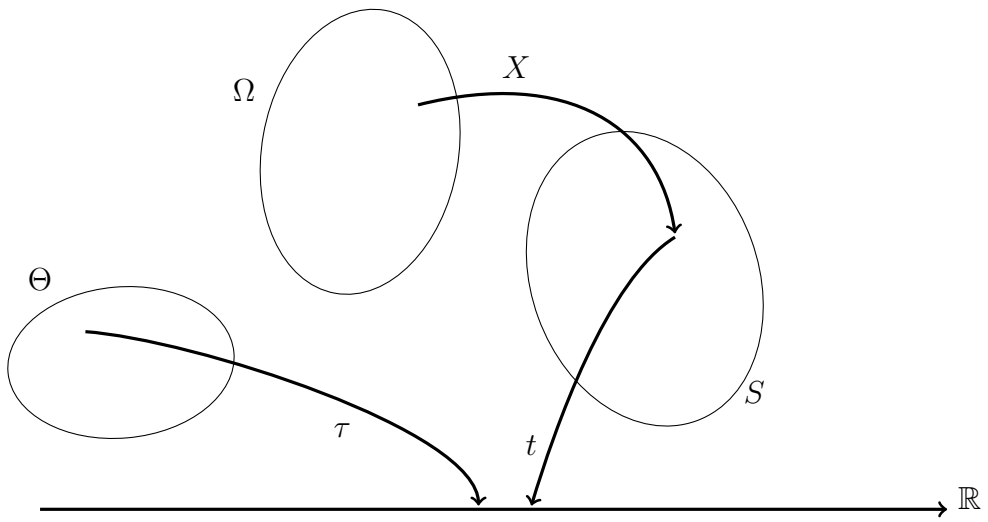
3. Ein Schätzer T für τ heißt *erwartungstreu* (oder „unverzerrt“), wenn gilt

$$\forall\, \vartheta \in \Theta : \quad \mathbb{E}_\vartheta[T] = \vartheta.$$

   $b_\vartheta(T) := \mathbb{E}_\vartheta[T] - \vartheta$ heißt die *Verzerrung* (englisch: bias) von T.

Die typische Konstruktion / Situation eines Schätzers ist $T = t(X)$ für eine Funktion $t : S \to \mathbb{R}$.

Man schreibt / benennt einen Schätzer für τ oft $\widehat{\tau}$.



Schematische Darstellung eines Schätzers $T = t(X)$ für τ

Beispiel.

Beobachtung 3 (Erwartungstreue Schätzer für Mittelwert und Varianz im Produktmodell)**.** Für $\vartheta \in \Theta$ sei Q_ϑ ein W'maß auf $\mathbb{R}$ mit endlichem Mittelwert

$$m(\vartheta) := \int_{\mathbb{R}} x\, Q_\vartheta(dx)$$

und endlicher Varianz

$$v(\vartheta) := \int_{\mathbb{R}} \big(x - m(\vartheta)\big)^2\, Q_\vartheta(dx).$$

Unter $\mathbb{P}_\vartheta$ seien $X_1, \dots, X_n$ u.i.v., $X_i \sim Q_\vartheta$.

(In der Formalisierung von Definition 1 und 2 könnten wir wählen: $\mathcal{M} = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (P_\vartheta)_{\vartheta\in\Theta})$ mit $\mathbb{P}_\vartheta = Q_\vartheta^{\otimes n}$ für $\vartheta \in \Theta$, als Statistik betrachten wir $X = (X_1, \dots, X_n)$ mit $X_i : \mathbb{R}^n \to \mathbb{R}$ die Projektion auf die i-te Koordinate.

Bemerke: dies ist u.U. kein parametrisches Modell, man könnte z.B.

$$\Theta := \big\{Q : Q \text{ ist W'maß auf } \mathbb{R} \text{ mit } \textstyle\int_{\mathbb{R}} x^2\, Q(dx) < \infty\big\}$$

wählen.)

Dann ist

$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i \quad \text{ein erwartungstreuer Schätzer für } m(\vartheta),$$
$$S^2 := \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}\big)^2 \quad \text{ein erwartungstreuer Schätzer für } v(\vartheta).$$

(In diesem Kontext heißt $\overline{X}$ auch der empirische Mittelwert oder Stichprobenmittelwert , S^2 die korrigierte Stichprobenvarianz)

Für $\vartheta \in \Theta$ gilt nämlich

$$\mathbb{E}_\vartheta[\overline{X}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\vartheta[X_i] = \frac{1}{n}\cdot n\, m(\vartheta) = m(\vartheta),$$

$$\begin{aligned}\mathbb{E}_\vartheta\Big[\sum_{i=1}^n \big(X_i - \overline{X}\big)^2\Big] &= n\mathbb{E}_\vartheta\big[(X_i - \overline{X})^2\big] = n\mathrm{Var}_\vartheta\big[X_i - \overline{X}\big] \\ &= n\mathrm{Var}_\vartheta\Big[\frac{n-1}{n}X_1 - \frac{1}{n}\sum_{i=2}^n X_i\Big] = n\Big(\big(\tfrac{n-1}{n}\big)^2\mathrm{Var}_\vartheta[X_1] + \tfrac{n-1}{n^2}\mathrm{Var}_\vartheta[X_1]\Big) \\ &= (n-1)\mathrm{Var}_\vartheta[X_1],\end{aligned}$$

also

$$\mathbb{E}_\vartheta\big[S^2\big] = \frac{1}{n-1}\mathbb{E}_\vartheta\Big[\sum_{i=1}^n \big(X_i - \overline{X}\big)^2\Big] = v(\vartheta).$$

Definition und Beobachtung 4 (Konsistenz)**.** Betrachten wir in der Situation von Beispiel 3 die Stichprobengröße n als variabel (formal: wir gehen zum unendlichen Produktmodell $\mathcal{M} = (\mathbb{R}^\infty, \mathcal{B}^{\otimes\infty}, (Q_\vartheta^{\otimes\infty})_{\vartheta\in\Theta})$ über) mit $X_i : \mathbb{R}^\infty \to \mathbb{R}$ Projektion auf i-te Koordinate).

Dann gilt für jedes $\vartheta \in \Theta$

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \underset{n\to\infty}{\longrightarrow} m(\vartheta) \quad \text{stochastisch bzgl. } P_\vartheta \quad \text{und}$$

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2 \underset{n\to\infty}{\longrightarrow} v(\vartheta) \quad \text{stochastisch bzgl. } P_\vartheta.$$

Man sagt: Diese (Folgen von) Schätzer(n) sind *konsistent.*

Für $\overline{X}_n$ folgt dies direkt aus dem Gesetz der großen Zahlen, weiterhin ist

$$\begin{aligned}\frac{n-1}{n}S_n^2 = \frac{1}{n}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2 &= \Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) - 2\Big(\frac{1}{n}\sum_{i=1}^n X_i\overline{X}_n\Big) + \big(\overline{X}_n\big)^2 = \Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big) - \big(\overline{X}_n\big)^2 \\ &\underset{n\to\infty}{\overset{P_\vartheta}{\longrightarrow}} \int_{\mathbb{R}} x^2\, Q_\vartheta(dx) - \big(m(\vartheta)\big)^2 = v(\vartheta)\end{aligned}$$

gemäß dem Gesetz der großen Zahlen zusammen mit $\frac{n-1}{n} \to 1$ folgt die Behauptung.

Definition und Satz 5 (Cramér-Rao-Schranke)**.** *Sei $(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Standardmodell, $\rho(\vartheta, x)$ die Likelihoodfunktion.*

Ein erwartungstreuer Schätzer T für ein reelles Parametermerkmal $\tau(\vartheta)$ heißt varianzminimierend*, falls für jeden anderen erwartungstreuen Schätzer $\widetilde{T}$ für τ gilt*

$$\mathrm{Var}_\vartheta[T] \le \mathrm{Var}_\vartheta[\widetilde{T}] \quad \textit{für alle } \vartheta \in \Theta.$$

Ein einparametriges Standardmodell (d.h. $\Theta \subset \mathbb{R}$) heißt regulär*, falls gilt:*

(i) $\Theta \subset \mathbb{R}$ *ist ein offenes Intervall.*

(ii) Likelihood-Funktion $\rho(\vartheta, x)$ *ist strikt positiv auf* $\Theta \times \Omega$ *und für jedes* x *ist* $\vartheta \mapsto \rho(\vartheta, x)$ *stetig diff'bar.*

(iii) $U_\vartheta(x) := \frac{d}{d\vartheta} \log \rho(\vartheta, x)$ *erfüllt* $I_\vartheta := \mathrm{Var}_\vartheta[U_\vartheta] \in (0, \infty)$

(U_ϑ *heißt die* Scorefunktion *und* I_ϑ *heißt die* Fisher-Information *(im Modell)) und es gilt*

$$\int_\Omega \tfrac{d}{d\vartheta} \rho(\vartheta, x)\, dx = \tfrac{d}{d\vartheta} \int_\Omega \rho(\vartheta, x)\, dx \ \ (= 0).$$

Weiter heißt ein Schätzer T *regulär, wenn für jedes* $\vartheta \in \Theta$ *gilt*

$$\frac{d}{d\vartheta} \int_\Omega T(x) \rho(\vartheta, x)\, dx = \int_\Omega T(x) \frac{d}{d\vartheta} \rho(\vartheta, x)\, dx.$$

Cramér-Rao-Schranke *Sei* $\tau : \Theta \to \mathbb{R}$ *ein stetig differenzierbares Parametermerkmal,* T *ein regulärer, erwartungstreuer Schätzer für* τ *in einem regulären Standardmodell.*

Dann gilt die Cramér-Rao-Schranke:

$$\mathrm{Var}_\vartheta[T] \geq \frac{(\tau'(\vartheta))^2}{I(\vartheta)} \quad \forall\, \vartheta \in \Theta,$$

wobei Gleichheit genau dann gilt, wenn

$$T(x) - \tau(\vartheta) = \frac{\tau'(\vartheta) U_\vartheta(x)}{I(\vartheta)}.$$

(Für $\tau = \mathrm{Id}$ *insbesondere:*
Jeder erwartungstreue Schätzer für ϑ *hat Varianz* $\geq 1/I(\vartheta)$*.)*

Definition 6 (Exponentielle Familie)**.** Sei

$$\rho(\vartheta, x) = h(x) \cdot \exp\big(a(\vartheta) \cdot T(x) - b(\vartheta)\big)$$

für gewisse (geeignete) Funktionen $a, b : \Theta \to \mathbb{R}$ und $h : \Omega \to \mathbb{R}$

Dann ist

$$U_\vartheta(x) = a'(\vartheta) T(x) - b'(\vartheta), \quad \text{also} \quad \mathbb{E}_\vartheta[T] = \frac{b'(\vartheta)}{a'(\vartheta)} =: \tau(\vartheta),$$

$[\mathbb{E}_\vartheta[U_\vartheta] = \int \rho(\vartheta, x) \frac{\partial}{\partial\vartheta} \log \rho(\vartheta, x)\, dx = \int \frac{\partial}{\partial\vartheta} \rho(\vartheta, x)\, dx = \frac{\partial}{\partial\vartheta} \int \rho(\vartheta, x)\, dx = \frac{\partial}{\partial\vartheta} 1 = 0.]$ Man kann zeigen, dass

$$I(\vartheta) = \mathrm{Var}_\vartheta[U_\vartheta] = a'(\vartheta) \cdot \tau'(\vartheta),$$

d.h. es gilt

$$T(x) = \frac{b'(\vartheta)}{a'(\vartheta)} + \frac{U_\vartheta(x)}{a'(\vartheta)} = \tau(\vartheta) + \frac{\tau'(\vartheta) U_\vartheta(x)}{I(\vartheta)}$$

und T ist varianzminimierender erwartungstreuer Schätzer für τ.

Beispiel. 1. Binomialverteilungen: $P_\vartheta = \mathrm{Bin}_{n,\vartheta}$, $\vartheta \in [0,1]$

$$\begin{aligned}\rho(\vartheta, x) &= \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x} \\ &= \binom{n}{x}\exp\Big(\underbrace{\frac{x}{n}}_{T(x)}\underbrace{n\log(\frac{\vartheta}{1-\vartheta})}_{=a(\vartheta)} + \underbrace{n\log(1-\vartheta)}_{=-b(\vartheta)}\Big), \quad x \in \mathbb{N}_0\end{aligned}$$

$T(x) = \dfrac{x}{n}$ ist varianzminimierender erwartungstreuer Schätzer für $\tau(\vartheta) = \vartheta$.

2. Poissonverteilungen: $P_\vartheta = \mathrm{Poi}_\vartheta$, $\vartheta \in (0,\infty)$

$$\rho(\vartheta, x) = e^{-\vartheta}\frac{\vartheta^x}{x!} = \underbrace{\frac{1}{x!}}_{=h(x)}\ e^{\overbrace{x}^{T(x)}\cdot\overbrace{\log\vartheta}^{=a(\vartheta)} - \overbrace{\vartheta}^{b(\vartheta)}}$$

Es ist $\tau(\vartheta) = \frac{1}{1/\vartheta} = \vartheta$, $T(x) = x$ ist varianzminimierender erwartungstreuer Schätzer für ϑ, seine Varianz ist $\frac{(\tau'(\vartheta))^2}{a'(\vartheta)\tau'(\vartheta)} = \frac{1^2}{\frac{1}{\vartheta}\cdot 1} = \vartheta$.

3. Normalverteilungen bei bekannter Varianz: $P_\vartheta = \mathcal{N}_{\vartheta,\sigma^2}$ mit festem $\sigma^2 > 0$

$$\begin{aligned}\rho(\vartheta, x) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{1}{2\sigma^2}(x-\vartheta)^2\Big) \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/(2\sigma^2)}}_{=h(x)}\cdot\exp\Big(-\underbrace{x}_{=T(x)}\cdot\underbrace{\frac{\vartheta}{\sigma^2}}_{=a(\vartheta)} - \underbrace{\frac{\vartheta^2}{2\sigma^2}}_{=b(\vartheta)}\Big),\end{aligned}$$

also: $T(x) = x$ ist varianzminimierender erwartungstreuer Schätzer für $\vartheta = \tau(\vartheta) = \frac{b'(\vartheta)}{a'(\vartheta)}$, seine Varianz ist $\sigma^2 = \frac{1}{I(\vartheta)} = 1^2/(a'(\vartheta)\tau'(\vartheta))$.

Bemerkung (Produktmodell)). Das n-fache Produktmodell $\mathscr{M}^{\otimes n}$ eines regulären Modells ist wiederum regulär, seine Fisher-Information erfüllt $I^{(n)}(\vartheta) = n \cdot I(\vartheta)$.

Ist $\mathscr{M}$ exponentielles Modells bzgl. der Statistik T, so ist $\mathscr{M}^{\otimes n}$ ebenfalls ein exponentielles Modell, und die zugrundeliegende Statistik ist

$$T_n(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^n T(x_i),$$

denn

$$\begin{aligned}\rho^{\otimes n}(\vartheta,(x_1,\dots,x_n)) &= \prod_{i=1}^n \rho(\vartheta, x_i) \\ &= \prod_{i=1}^n h(x_i)\exp\left(na(\vartheta)\cdot\frac{1}{n}\cdot\sum_{i=1}^n(T(x_i)) - nb(\vartheta)\right).\end{aligned}$$

1.1 Suffizienz, Vollständigkeit und Verteilungsfreiheit

Notationsvereinbarung. Wir fassen $X : \Omega \to \Omega$, $X = \mathrm{Id}_\Omega$ als ZV auf und interpretieren X als die „beobachteten Daten“.

Definition 7. Eine Statistik T heißt *suffizient*, wenn die bedingte Verteilung $\mathbb{P}_\vartheta(\cdot \mid T)$ der Beobachtungen nicht von $\vartheta \in \Theta$ abhängt.

Intuition: T enthält bereits sämtliche Informationen über ϑ, die „in den Beobachtungsdaten stecken."

Beispiel.

- n-facher Münzwurf: $\Theta = [0,1]$, unter $\mathbb{P}_\vartheta$ sei $X = (X_1, \dots, X_n) \sim \mathrm{Ber}_\vartheta^{\otimes n}$, dann ist $T := X_1 + \cdots + X_n$ suffizient:

  Gegeben $T = t \in \{0, 1, \dots, n\}$ ist X uniform verteilt auf

  $$\big\{(x_1, \dots, x_n) \in \{0,1\}^n : x_1 + \cdots + x_n = t\big\}$$

- Sei $\Theta = \mathbb{N}$, $\Omega = \mathbb{Z}_+^2$, $\mathbb{P}_\vartheta = \mathrm{Unif}(\{0, 1, \dots, \vartheta\})^{\otimes 2}$ (aufgefasst als Wahrscheinlichkeitsmaß auf $\mathbb{Z}_+^2$) für $\vartheta \in \Theta$, d.h. unter $\mathbb{P}_\vartheta$ sind X_1 und x_2 unabhängig und jeweils uniform verteilt auf $\{0, 1, \dots, \vartheta\}$. Hier ist $T := X_1 + X_2$ nicht suffizient:

  T hat Werte in $\{0, 1, \dots, 2\vartheta\}$ und

  $$\mathbb{P}_\vartheta(T = t) = \frac{1 + \vartheta - |t - \vartheta|}{\vartheta^2} \quad \text{für } t = 0, 1, \dots, 2\vartheta$$

  somit ist für $x_1 \in \{0, 1, \dots, \vartheta\} \ni x_2 := t - x_1$

  $$\mathbb{P}_\vartheta\big((X_1, X_2) = (x_1, x_2) \,\big|\, T = t\big) = \mathbf{1}_{\{x_1 \le \vartheta, x_2 \ge 0\}} \frac{1/\vartheta^2}{(1 + \vartheta - |t - \vartheta|)/\vartheta^2} = \mathbf{1}_{\{x_1 \le \vartheta, x_2 \ge 0\}} \frac{1}{1 + \vartheta - |t - \vartheta|}$$

  was von ϑ abhängt.

Bemerkung. Offenbar ist stets $X = \mathrm{Id}_\Omega$ suffizient, aber diese Beobachtung nützt i.A. überhaupt nichts.

Satz 8 (Faktorisierungssatz von Fisher-Neyman)**.** *Gegeben sei ein statistisches Standardmodell. Eine Statistik $T = t(X)$ mit Werten in E ist suffizient, genau dann wenn die Dichte-/Gewichtsfunktion $\rho(\vartheta, x)$ die Gestalt*

$$\rho(\vartheta, x) = h(x) g_\vartheta\big(t(x)\big), \quad x \in \Omega, \vartheta \in \Theta \tag{1.1}$$

hat (für eine Funktion $h : \Omega \to \mathbb{R}_+$ und Funktionen $g_\vartheta : E \to \mathbb{R}_+$, $\vartheta \in \Theta$).

Der *Beweis von Satz 8 im diskreten* Fall ist elementar: Falls Ω (und damit X und o.E. auch E) diskret ist, so ist für $t \in E$ mit $\mathbb{P}_\vartheta(T = t) > 0$

$$\mathbb{P}_\vartheta\big(X = x \,\big|\, T = t\big) = \frac{\mathbb{P}_\vartheta(X = x, T = t)}{\mathbb{P}_\vartheta(T = t)} = \begin{cases} \dfrac{\mathbb{P}_\vartheta(X = x)}{\sum_{y : t(y) = t} \mathbb{P}_\vartheta(X = y)}, & \text{falls } t(x) = t, \\ 0, & \text{sonst} \end{cases}$$

Falls $\rho(\vartheta, x)$ die Gestalt (1.1) hat, so ist demnach

$$\mathbb{P}_\vartheta\big(X = x \,\big|\, T = t\big) = \begin{cases} \dfrac{h(x)g_\vartheta(t)}{\sum_{y:t(y)=t} h(y)g_\vartheta(t)} = \dfrac{h(x)}{\sum_{y:t(y)=t} h(y)}, & \text{falls } t(x) = t, \\ 0, & \text{sonst} \end{cases}$$

was offenbar nicht von ϑ abhängt.

Falls andererseits $\mathbb{P}_\vartheta\big(X = x \,\big|\, T = t\big)$ nicht von ϑ abhängt, so kann man schreiben

$$\mathbb{P}_\vartheta(X = x) = \underbrace{\mathbb{P}_\vartheta\big(T = t(x)\big)}_{=:\, g_\vartheta\left(t(x)\right)} \cdot \underbrace{\mathbb{P}_\vartheta\big(X = x \,\big|\, T = t(x)\big)}_{=:\, h(x)}$$

□

Im stetigen Fall brauchen wir das folgende Lemma 9.

Lemma 9. *Sei $(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein stetiges statistisches Standardmodell mit $\mathbb{P}_\vartheta(dx) = \rho(\vartheta, x)dx$, $t : \Omega \to E$ messbar, $T = t \circ X$ eine Statistik.*

1. *(gemeinsame Dominierung) Es gibt $c_1, c_2, \dots \in [0, 1]$ mit $\sum_{i=1}^\infty c_i = 1$ und $\vartheta_1, \vartheta_2, \dots \in \Theta$, so dass*

$$\mathbb{P}_\vartheta \ll \nu^* := \sum_{i=1}^\infty c_i \mathbb{P}_{\vartheta_i} \quad \text{für alle } \vartheta \in \Theta \tag{1.2}$$

2. *Ist T suffizient, so ist für $\vartheta \in \Theta$*

$$\nu^*(A \mid T) = \mathbb{P}_\vartheta(A \mid T) \ \ \nu^*\text{-fast sicher}$$

3. *Falls $\rho(\vartheta, x) = h(x)g_\vartheta(t(x))$ für $\mathbb{P}_\vartheta$-fast alle x, $\vartheta \in \Theta$ gilt, so ist*

$$\frac{d\mathbb{P}_\vartheta}{d\nu^*}(x) = \frac{g_\vartheta(t(x))}{\sum_{i=1}^\infty c_i g_{\vartheta_i}(t(x))} \tag{1.3}$$

   eine Version der Dichte, die nur von $t(x)$ abhängt.

4. *Falls für jedes $\vartheta \in \Theta$ die Dichte $\frac{d\mathbb{P}_\vartheta}{d\nu^*}(x)$ nur von $t(x)$ abhängt, so ist T suffizient.*

Beweis. 1. $\Omega \subset \mathbb{R}^n$, $\lambda =$ Lebesgue-Maß auf $\mathbb{R}^n$. Da $\mathscr{L}^1(\Omega, \lambda)$ separabel ist (beachte: $\mathcal{B}(\Omega)$ ist abzählbar erzeugt), enthält $\{\rho(\vartheta, \cdot) : \vartheta \in \Theta\} \subset \mathscr{L}^1(\Omega, \lambda)$ eine abzählbare, (bzgl. der $\mathscr{L}^1$-Norm) dichte Teilmenge $\{\rho(\vartheta_n, \cdot) : n \in \mathbb{N}\}$, d.h. für jedes $\vartheta \in \Theta$ gibt es eine Folge $(n_j)_j \subset \mathbb{N}$ mit

$$\lim_{j\to\infty} \int_\Omega \left|\rho(\vartheta, x) - \rho(\vartheta_{n_j}, x)\right| \lambda(dx) = 0 \tag{1.4}$$

Wir setzen

$$\nu^* := \sum_{n=1}^\infty 2^{-n} \mathbb{P}_{\vartheta_n}$$

dies leistet das Gewünschte: Sei $A \in \mathcal{B}(\Omega)$ mit $\nu^*(A) = 0$ und $\vartheta \in \Theta$ gegeben, so ist mit der Folge $(n_j)_j$ aus (1.4) auch

$$\mathbb{P}_\vartheta(A) = \int_A \rho(\vartheta, x)\, \lambda(dx) = \lim_{j\to\infty} \int_A \rho(\vartheta_{n_j}, x)\, \lambda(dx) = \lim_{j\to\infty} 0 = 0$$

d.h. $\mathbb{P}_\vartheta \ll \nu^*$.

2. Nach Voraussetzung gibt es eine Funktion $\kappa : E \times \mathcal{B}(\Omega) \to [0,1]$, so dass für jedes $\vartheta \in \Theta$ und $A \in \mathcal{B}(\Omega)$ gilt

$$\mathbb{P}_\vartheta(A \mid T = t) = \kappa(t, A) \quad \mathbb{P}_\vartheta\text{-fast sicher.}$$

Damit

$$\nu^*(A \mid T = t) = \sum_{j=1}^{\infty} 2^{-j} \mathbb{P}_{\vartheta_j}(A \mid T = t) \overset{\nu^*\text{-f.s.}}{=} \kappa(t, A) \overset{\mathbb{P}_\vartheta\text{-f.s.}}{=} \kappa(t, A)$$

3. Es ist nach Voraussetzung

$$\frac{d\nu^*}{d\lambda}(x) = \sum_{j=1}^{\infty} 2^{-j} \frac{d\mathbb{P}_{\vartheta_j}}{d\lambda}(x) = \sum_{j=1}^{\infty} 2^{-j} \rho(\vartheta_j, x) = \sum_{j=1}^{\infty} 2^{-j} h(x) g_{\vartheta_j}\big(t(x)\big)$$

und somit

$$\frac{d\mathbb{P}_\vartheta}{d\nu^*}(x) = \frac{d\mathbb{P}_\vartheta}{d\lambda}\Big(\frac{d\nu^*}{d\lambda}\Big)^{-1}(x) = \frac{g_\vartheta(t(x))}{\sum_{i=1}^{\infty} c_i g_{\vartheta_i}(t(x))}$$

wie gefordert.

4. Wir zeigen für $\vartheta \in \Theta$

$$\mathbb{E}_\vartheta\big[\mathbb{P}_\vartheta(A \mid T)\mathbf{1}_B\big] = \mathbb{E}_\vartheta\big[\nu^*(A \mid T)\mathbf{1}_B\big] \quad \text{für } A \in \mathcal{B}(\Omega),\ B \in \sigma(T)$$

Dann ist nämlich $\nu^*(\cdot \mid T)$, das nicht von ϑ abhängt, eine Version von $\mathbb{P}_\vartheta(\cdot \mid T)$ und somit T suffizient. Tatsächlich ist

$$\begin{aligned}\mathbb{E}_\vartheta\big[\mathbb{P}_\vartheta(A \mid T)\mathbf{1}_B\big] = \mathbb{E}_\vartheta\big[\mathbf{1}_A\mathbf{1}_B\big] = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\mathbf{1}_A\mathbf{1}_B\Big] = \mathbb{E}_{\nu^*}\Big[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\mathbf{1}_A\mathbf{1}_B\,\Big|\,T\Big]\Big] \\ = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\mathbf{1}_B\mathbb{P}_{\nu^*}(A \mid T)\Big] = \mathbb{E}_\vartheta\big[\mathbb{P}_{\nu^*}(A \mid T)\mathbf{1}_B\big]\end{aligned}$$

(wir verwenden in der ersten Gleichung der zweiten Zeile die Tatsache, dass $\frac{d\mathbb{P}_\vartheta}{d\nu^*}$ messbar bzgl. $\sigma(T)$ ist). □

Beweis von Satz 8, stetiger Fall. Falls $\rho(\vartheta, x)$ die Gestalt (1.1) hat, so folgt aus Lemma 9, 3. und 4., dass T suffizient ist.

Sei umgekehrt T suffizient, ν^* wie in Lemma 9. Nach Lemma 9, 2. gilt für $\vartheta \in \Theta$

$$\mathbb{P}_\vartheta(\cdot \mid T) = \nu^*(\cdot \mid T)$$

Es ist für beliebiges $A \in \mathcal{B}(\Omega)$

$$\begin{aligned}\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\mathbf{1}_A\Big] &= \mathbb{P}_\vartheta(A) = \mathbb{E}_\vartheta\big[\mathbb{P}_\vartheta(A \mid T)\big] = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\nu^*(A \mid T)\Big] \\ &= \mathbb{E}_{\nu^*}\bigg[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\nu^*(A \mid T)\,\Big|\,T\Big]\bigg] = \mathbb{E}_{\nu^*}\bigg[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\,\Big|\,T\Big]\nu^*(A \mid T)\bigg] \\ &= \mathbb{E}_{\nu^*}\bigg[\mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\,\Big|\,T\Big]\mathbf{1}_A\bigg]\end{aligned}$$

und daher mit $T = t(X)$

$$\frac{d\mathbb{P}_\vartheta}{d\nu^*}(X) = \mathbb{E}_{\nu^*}\Big[\frac{d\mathbb{P}_\vartheta}{d\nu^*}\,\Big|\,t(X)\Big] := g_\vartheta\big(t(X)\big)$$

Setze $h(x) := \frac{d\nu^*}{d\lambda}(x)$, so ist

$$\frac{d\mathbb{P}_\vartheta}{d\lambda}(x) = \Big(\frac{d\mathbb{P}_\vartheta}{d\nu^*}\frac{d\nu^*}{d\lambda}\Big)(x) = g_\vartheta\big(t(x)\big)h(x)$$

□

Beispiel. Normalmodell: $\Theta = \mathbb{R} \times (0,\infty) \ni \vartheta = (\mu, \sigma^2)$, $\Omega = \mathbb{R}^n$, $\mathbb{P}_{(\mu,\sigma^2)} = \mathcal{N}(\mu,\sigma^2)^{\otimes n}$. Es ist

$$\begin{aligned}\rho\big((\mu,\sigma^2),(x_1,\dots,x_n)\big) &= \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\big(-(x_j-\mu)^2/(2\sigma^2)\big) \\ &= (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{j=1}^n(x_j-\mu)^2\Big) \\ &= (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{n\mu^2}{2\sigma^2}+\frac{\mu}{\sigma^2}\sum_{j=1}^n x_j - \frac{1}{2\sigma^2}\sum_{j=1}^n x_j^2\Big)\end{aligned}$$

Daher ist $\big(\sum_{j=1}^n X_j, \sum_{j=1}^n X_j^2\big)$ (minimal-)suffizient und (da es sich um eine deterministische Umparametrisierung handelt) ebenso

$$\big(\overline{X}, S^2\big) := \left(\frac{1}{n}\sum_{j=1}^n X_j\,,\ \frac{1}{n-1}\sum_{j=1}^n (X_j-\overline{X})^2\right)$$

$\Theta = (0,\infty) \ni \vartheta$, $\Omega = [0,\infty)^n$, $\mathbb{P}_\vartheta = \mathrm{Unif}([0,\vartheta])^{\otimes n}$, so ist $T := \max\{X_1,\dots,X_n\}$ suffizient: Es ist

$$\rho\big(\vartheta,(x_1,\dots,x_n)\big) = \prod_{j=1}^n \big(\mathbf{1}_{0\le x_j\le\vartheta}/\vartheta\big) = \vartheta^{-n}\mathbf{1}_{0\le\max\{x_j:j\le n\}\le\vartheta}$$

Definition 10. Eine suffiziente Statistik T heißt *minimalsuffizient*, wenn es für jede suffiziente Statistik U eine Funktion φ gibt, so dass für alle $\vartheta \in \Theta$ gilt

$$T = \varphi(U) \quad \mathbb{P}_\vartheta\text{-fast sicher.}$$

Satz 11 (Kriterium für Minimalsuffizienz)**.** *Gegeben ein statistisches Standardmodell und $t : \Omega \to E$ messbar, so dass gilt*

$$t(x) = t(y) \quad \Longleftrightarrow \quad \text{es gibt } 0 < \ell(x,y) < \infty \text{ mit } \rho(\vartheta, y) = \ell(x,y)\rho(\vartheta, x) \text{ für } \vartheta \in \Theta. \tag{1.5}$$

Dann ist $T = t(X)$ minimalsuffizient.

Bemerkung. Wir sehen anhand von Satz 11, dass in obigen Beispielen $(\overline{X}, S^2)$ im Normalmodell bzw. $\max\{X_1, \dots, X_n\}$ im $\mathrm{Unif}([0,\vartheta])^{\otimes n}$-Beispiel tatsächlich minimalsuffizient sind.

Beweis. Sei ν^* wie in Lemma 9, $x, y \in \Omega$ mit $t(x) = t(y)$, so ist

$$\begin{aligned}\frac{d\mathbb{P}_\vartheta}{d\nu^*}(x) &= \frac{\frac{d\mathbb{P}_\vartheta}{d\lambda}(x)}{\frac{d\nu^*}{d\lambda}(x)} = \frac{\rho(\vartheta, x)}{\sum_{i=1}^\infty c_i \rho(\vartheta_i, x)} = \frac{\rho(\vartheta, x)\ell(x,y)}{\sum_{i=1}^\infty c_i \rho(\vartheta_i, x)\ell(x,y)} \\ &= \frac{\rho(\vartheta, y)}{\sum_{i=1}^\infty c_i \rho(\vartheta_i, y)} = \frac{d\mathbb{P}_\vartheta}{d\nu^*}(y)\end{aligned}$$

d.h. $\frac{d\mathbb{P}_\vartheta}{d\nu^*}$ hängt nur von $t(x)$ ab und gemäß Lemma 9, 4. ist $T = t(X)$ suffizient.

Sei $U = u(X)$ eine weitere suffiziente Statistik.

Nach Satz 8 ist $\rho(\vartheta, x) = h(x) g_\vartheta(u(x))$ (und o.E. $h > 0$, sonst schränke Ω auf $\{h > 0\}$ ein).

Seien $x, y \in \Omega$ mit $u(x) = u(y)$ gegeben, so ist

$$\frac{\rho(\vartheta, y)}{\rho(\vartheta, x)} = \frac{h(y) g_\vartheta(u(y))}{h(x) g_\vartheta(u(x))} = \frac{h(y)}{h(x)}$$

(was nicht von ϑ abhängt), mit (1.5) folgt $t(x) = t(y)$.

Somit

$$u(x) = u(y) \quad \Longrightarrow \quad t(x) = t(y)$$

d.h. es gibt eine Funktion f mit $t(x) = f(u(x))$. Da dieses Argument für jedes suffiziente U greift, ist T minimalsuffizient. □

Definition 12. Eine Statistik T mit Werten in E heißt *vollständig*, wenn für alle Funktionen (messbaren) $g : E \to \mathbb{R}$ gilt

$$\forall\, \vartheta \in \Theta : \ \mathbb{E}_\vartheta\big[g(T)\big] = 0 \qquad \Longrightarrow \qquad \forall\, \vartheta \in \Theta : \ g(T) = 0 \ \ \mathbb{P}_\vartheta\text{-f.s.} \tag{1.6}$$

T heißt *beschränkt vollständig*, wenn dies nur für beschränkte Testfunktionen g gefordert wird.

Beispiel.

- $X \sim \mathrm{Poi}(\vartheta)$, $\vartheta \in \Theta = (0, \infty)$ ist beschränkt vollständig: Sei $g : \mathbb{N}_0 \to \mathbb{R}$ beschränkt, so ist die Funktion

$$\phi : \Theta \ni \vartheta \mapsto \mathbb{E}_\vartheta\big[g(X)\big] = \sum_{x=0}^\infty e^{-\vartheta} \frac{\vartheta^x}{x!} g(x)$$

(als Potenzreihe mit Konvergenzradius ∞) insbesondere analytisch. Wenn nun $\phi(\vartheta) = 0$ für alle $\vartheta > 0$ gilt, so muss $\phi \equiv 0$ gelten, d.h. $g(x) = 0$ für alle $x \in \mathbb{N}_0$.

- In einem Produktmodell (auf einem kompakten Wertebereich, sagen wir) ist die Beobachtung $X = (X_1, \dots, X_n)$ selbst nicht (beschränkt) vollständig: Offenbar ist $\mathbb{E}_\vartheta[X_1 - X_2] = 0$ für jedes ϑ, aber die Funktion $(x_1, \dots, x_n) \mapsto x_1 - x_2$ ist nicht konstant.

- (Nach [LR06, Problem 4.12 (S. 141)]) Betrachte $\Theta = [0,1]$, $\Omega = \{-1, 0\} \cup \mathbb{N}$,

$$\mathbb{P}_\vartheta(x) = \begin{cases} \vartheta, & x = -1 \\ (1-\vartheta)^2 \vartheta^x, & x = 0, 1, 2, \dots \end{cases}$$

  (d.h. unter $\mathbb{P}_\vartheta$ ist $X = -1$ mit W'keit ϑ und mit der Gegenw'keit $1-\vartheta$ ist $X \sim \mathrm{Geom}(1-\vartheta)$). Dann gilt $X \in \mathscr{L}^1(\mathbb{P}_\vartheta)$ und

$$\mathbb{E}_\vartheta[X] = -\vartheta + (1-\vartheta)\Big(\frac{1}{1-\vartheta} - 1\Big) = -\vartheta + \vartheta = 0$$

  für jedes $\vartheta \in \Theta$. Weiter ist $|\sum_{x=0}^\infty \vartheta^x g(x)| \le \|g\|_\infty \sum_{x=0}^\infty \vartheta^x = \|g\|_\infty/(1-\vartheta)$ und $|g(-1)| = \vartheta^{-1}(1-\vartheta)^2 \, |\sum_{x=0}^\infty \vartheta^x g(x)| \le (1-\vartheta)\|g\|_\infty/\vartheta$, mit $\vartheta \uparrow 1$ folgt $g(-1) = 0$ und somit $\sum_{x=0}^\infty \vartheta^x g(x) \equiv 0$, was $g(x) = 0$ für $x \in \mathbb{N}_0$ erzwingt. (Denn $\phi(\vartheta) := \sum_{x=0}^\infty \vartheta^x g(x)$ ist analytisch in $(-1,1)$ und $\equiv 0$ in $(0,1)$, somit muss $\phi(\cdot) \equiv 0$ sein.)

  Hier ist X beschränkt vollständig, aber nicht vollständig.

Satz 13 (Satz von Bahadur)**.** *Eine Statistik $T = t(X)$ mit Werten in $\mathbb{R}^k$ für ein k, die vollständig und suffizient ist, ist minimalsuffizient.*

Beweis. Wir notieren $t(x) = \big(t_1(x), \dots, t_k(x)\big)$. Sei $u : \Omega \to E$ messbar und $U = u(X)$ eine (weitere) suffiziente Statistik, zu zeigen: $T = f(U)$ für eine Funktion $f : E \to \mathbb{R}^k$.

$\varphi : \mathbb{R} \ni z \mapsto 1/(1+e^z) \in (0,1)$ ist bijektiv (und bi-messbar), für $t = (t_1, \dots, t_k) \in \mathbb{R}^k$ sei $s = s(t) = (s_1, \dots, s_k) \in (0,1)^k$ mit $s_i = \varphi(t_i)$.

Setze $S := s(T) = (\varphi(T_1), \dots, \varphi(T_k))$ und für $i \in \{1, \dots, k\}$

$$H_i(U) := \mathbb{E}_\vartheta[S_i \mid U] \ \Big(= \mathbb{E}_{\nu^*}[S_i \mid U]\Big)$$
$$J_i(T) := \mathbb{E}_\vartheta[H_i(U) \mid T] \ \Big(= \mathbb{E}_{\nu^*}[H_i(U) \mid T]\Big)$$

(mit ν^* wie in Lemma 9, die bedingten Erwartungswerte hängen nicht von ϑ ab), da S beschränkt ist, gilt dies auch für $H_i(U)$ und $J_i(T)$.

Weiter ist für $\vartheta \in \Theta$

$$\mathbb{E}_\vartheta\big[\varphi(T_i) - J_i(T)\big] = \mathbb{E}_\vartheta\Big[\varphi(T_i) - \mathbb{E}_\vartheta\big[\mathbb{E}_\vartheta[\varphi(T_i) \mid U] \,\big|\, T\big]\Big] = \mathbb{E}_\vartheta\big[\varphi(T_i)\big] - \mathbb{E}_\vartheta\big[\varphi(T_i)\big] = 0,$$

d.h. (da T beschränkt vollständig ist) es gilt für jedes $\vartheta \in \Theta$

$$J_i(T) = \varphi(T_i) \quad \mathbb{P}_\vartheta\text{-f.s.} \tag{1.7}$$

$$\mathrm{Var}_\vartheta\big[J_i(T)\big] = \mathbb{E}_\vartheta\Big[\mathrm{Var}_\vartheta\big[J_i(T) \,\big|\, U\big]\Big] + \mathrm{Var}_\vartheta\big[\mathbb{E}_\vartheta[J_i(T) \mid U]\big] = \mathbb{E}_\vartheta\Big[\mathrm{Var}_\vartheta\big[J_i(T) \,\big|\, U\big]\Big] + \mathrm{Var}_\vartheta\big[H_i(U)\big]$$
$$\mathrm{Var}_\vartheta\big[H_i(U)\big] = \mathbb{E}_\vartheta\Big[\mathrm{Var}_\vartheta\big[H_i(U) \,\big|\, T\big]\Big] + \mathrm{Var}_\vartheta\big[J_i(T)\big]$$

und da (stets) $\mathrm{Var}_\vartheta\big[J_i(T)\big] \le \mathrm{Var}_\vartheta\big[H_i(U)\big]$ gilt, folgt

$$\mathrm{Var}_\vartheta\big[J_i(T)\,\big|\,U\big] = 0 \;\; \mathbb{P}_\vartheta\text{-f.s.} \quad \text{und damit dann auch} \quad \mathrm{Var}_\vartheta\big[H_i(U)\,\big|\,T\big] = 0 \;\; \mathbb{P}_\vartheta\text{-f.s.}$$

Somit

$$\varphi(T_i) = \mathbb{E}_\vartheta[\varphi(T_i) \mid U] = H_i(U) \;\; \mathbb{P}_\vartheta\text{-f.s.}$$

und $T = (T_1, \dots, T_k)$ mit $T_i = \varphi^{-1}(H_i(U))$ ist (ν^*-f.s.) eine Funktion von U. □

Satz 14 (Rao-Blackwell)**.** *Sei S ein Schätzer für ein Parametermerkmal $\tau(\vartheta)$ mit $\mathbb{E}_\vartheta[|S|] < \infty$ für alle $\vartheta \in \Theta$ und sei T eine suffiziente Statistik. Dann erfüllt die „Rao-Blackwellisierung“*

$$S^* := \mathbb{E}\big[S\,\big|\,T\big] \tag{1.8}$$

von S (beachte, dass obige bedingte Erwartung nicht von ϑ abhängt, da T suffizient ist)

$$\mathbb{E}_\vartheta[S^*] = \mathbb{E}_\vartheta[S] \quad \textit{für alle } \vartheta \in \Theta \quad \textit{und} \tag{1.9}$$

$$\mathbb{E}_\vartheta\left[(S^* - \tau(\vartheta))^2\right] \le \mathbb{E}_\vartheta\left[(S - \tau(\vartheta))^2\right] \quad \textit{für alle } \vartheta \in \Theta. \tag{1.10}$$

Falls $\mathrm{Var}_\vartheta[S] < \infty$ für alle ϑ gilt, so ist die Ungleichung strikt, sofern $S \neq S^*$.

Beweis. (1.9) folgt aus der Turmeigenschaft der bedingten Erwartung.

Zu (1.10): Stets ist

$$\mathbb{E}_\vartheta\left[(S - \tau(\vartheta))^2\right] = \mathbb{E}_\vartheta\left[\big(S - \mathbb{E}_\vartheta[S]\big)^2\right] + \big(\mathbb{E}_\vartheta[S] - \tau(\vartheta)\big)^2$$

d.h. erw. quadr. Abw. = Varianz + (Bias)^2, denn

$$\begin{aligned}
\mathbb{E}_\vartheta\left[(S - \tau(\vartheta))^2\right] &= \mathbb{E}_\vartheta\left[S^2\right] - 2\tau(\vartheta)\mathbb{E}_\vartheta\left[S\right] + \tau(\vartheta)^2 \\
&= \mathbb{E}_\vartheta\left[S^2\right] - \big(\mathbb{E}_\vartheta\left[S\right]\big)^2 + \big(\mathbb{E}_\vartheta\left[S\right]\big)^2 - 2\tau(\vartheta)\mathbb{E}_\vartheta\left[S\right] + \tau(\vartheta)^2 \\
&= \mathrm{Var}_\vartheta\left[S\right] + \big(\mathbb{E}_\vartheta[S] - \tau(\vartheta)\big)^2
\end{aligned}$$

Somit folgt (1.10) aus (1.9) mit bedingter Varianzzerlegung

$$\mathrm{Var}_\vartheta[S] = \mathbb{E}_\vartheta\big[\mathrm{Var}_\vartheta[S \mid T]\big] + \mathrm{Var}_\vartheta\big[\mathbb{E}_\vartheta[S \mid T]\big] \le \mathrm{Var}_\vartheta\big[\mathbb{E}_\vartheta[S \mid T]\big] = \mathrm{Var}_\vartheta[S^*]$$

□

Korollar 15 (Satz von Lehmann-Scheffé)**.** *T eine suffiziente und vollständige Statistik und S ein erwartungstreuer Schätzer für $\tau(\vartheta)$. Dann ist S^* aus Satz 14 varianzminimierend.*

Beweis. Nach Konstruktion ist $S^* = f_1(T)$ für eine gewisse Funktion f_1. Sei $\widetilde{S}$ ein weiterer erwartungstreuer Schätzer für $\tau(\vartheta)$, $\widetilde{S}^* = \mathbb{E}_\vartheta[\widetilde{S} \mid T] =: f_2(T)$ dessen „Rao-Blackwellisierung“. Dann gilt (da beide erwartungstreu sind)

$$\mathbb{E}_\vartheta[f_1(T) - f_2(T)] = \tau(\vartheta) - \tau(\vartheta) = 0$$

mit Vollständigkeit von T folgt, dass $g(t) := f_1(t) - f_2(t) \equiv 0$ ist. □

Beispiel. $X_1, \dots, X_n$ u.i.v. $\sim \mathrm{Unif}([0,\vartheta])$, $S := \frac{2}{n}(X_1 + \cdots + X_n)$, $T = \max\{X_1, \dots, X_n\}$; es ist

$$\mathbb{E}_\vartheta\big[X_i \,\big|\, T\big] = \frac{1}{n}T + \frac{n-1}{n}\frac{1}{2}T$$

und damit ist

$$S^* = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_\vartheta\big[X_i \,\big|\, T\big] = \frac{2}{n}T + \frac{n-1}{n}T = \frac{n+1}{n}T$$

in diesem Modell ein erwartungstreuer Schätzer für ϑ mit kleinstmöglicher Varianz.

Definition 16. Eine Statistik U heißt *verteilungsfrei*, wenn $\mathscr{L}_\vartheta(U)$ nicht von ϑ abhängt. U heißt *maximal verteilungsfrei*, wenn es für jede andere verteilungsfreie Statistik V eine Funktion g gibt mit $V = g(U)$.

Beispiel.

- Im Normalmodell ist $\big(X_1 - \overline{X}, \dots, X_n - \overline{X}\big)/\sqrt{S^2}$ verteilungsfrei.
- Für $(X_1, \dots, X_n) \sim \mathrm{Unif}([0,\vartheta])^{\otimes n}$ ist mit $T := \max\{X_1, \dots, X_n\}$ $\big(X_1/T, \dots, X_n/T\big)$ verteilungsfrei.

Satz 17 (Satz von Basu)**.** *Sei $(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell., $T = t(X)$ mit $t : \Omega \to E$ messbar und $U = u(X)$ mit $u : \Omega \to E'$ messbar Statistiken.*

1. Falls T (beschränkt) vollständig und suffizient ist und U verteilungsfrei, so sind T und U unter jedem $\mathbb{P}_\vartheta$ unabhängig.

2. Falls T suffizient ist und T und U unter jedem $\mathbb{P}_\vartheta$ unabhängig sind, so ist U verteilungsfrei.

3. Seien T und U unter jedem $\mathbb{P}_\vartheta$ unabhängig und U verteilungsfrei, es gelte $\sigma(T,U) = \mathscr{F}$. Dann ist T suffizient.

Beweis. 1. $A \in \mathcal{B}(E')$, $\vartheta \in \Theta$, stets ist

$$\mathbb{P}_\vartheta(U \in A) = \mathbb{E}_\vartheta\big[\mathbb{P}_\vartheta(U \in A \mid T)\big]$$

Da U verteilungsfrei ist, hängt $\mathbb{P}_\vartheta(U \in A)$ nicht von ϑ ab, und da T suffizient ist, hängt auch $\mathbb{P}_\vartheta(U \in A \mid T)$ nicht von ϑ ab. Somit ist

$$g : E \ni t \mapsto \mathbb{P}_\vartheta(U \in A \mid T = t) - \mathbb{P}_\vartheta(U \in A) \in [-1, 1]$$

beschränkt mit $\mathbb{E}_\vartheta[g(T)] = 0$ für alle ϑ. Vollständigkeit von T erzwingt $g(t) \equiv 0$ ($\mathbb{P}_\vartheta$-f.s. für jedes ϑ), d.h.

$$\mathbb{P}_\vartheta(U \in A) = P_\vartheta(U \in A \mid T) \quad \mathbb{P}_\vartheta\text{f.s.}$$

Dies bedeutet gerade, dass T und U unter $\mathbb{P}_\vartheta$ unabhängig sind.

2. Für $A \in \mathscr{F}$ hängt $\mathbb{P}_\vartheta(U \in A \mid T)$ nicht von ϑ ab, d.h. es gibt ein $\nu(A) \in [0,1]$, so dass für jedes $\vartheta \in \Theta$ gilt

$$\mathbb{P}_\vartheta(U \in A \mid T) = \nu(A) \quad \mathbb{P}_\vartheta\text{-f.s.}$$

Unabhängigkeit von U und T unter $\mathbb{P}_\vartheta$ impliziert

$$\mathbb{P}_\vartheta(U \in A \mid T) = \mathbb{P}_\vartheta(U \in A) \quad (\mathbb{P}_\vartheta\text{-f.s.})$$

und somit ist $\mathbb{P}_\vartheta(U \in A) = \nu(A)$; da dies nicht von ϑ abhängt, ist U verteilungsfrei.

3. Zu zeigen ist, dass für $A \in \mathscr{F}$

$$\mathbb{P}_\vartheta(A \mid T) \quad \text{nicht von } \vartheta \text{ abhängt}$$

Es genügt, dies für A vom Typ $A = \{T \in B\} \cap \{U \in B'\}$ nachzuweisen (denn solche Ereignisse bilden einen $\cap$-stabilen Erzeuger von $\sigma(T,U)$ und nach Voraussetzung ist $\mathscr{F} = \sigma(T,U)$). Nun ist

$$\mathbb{P}_\vartheta\big(\{T \in B\} \cap \{U \in B'\} \,\big|\, T\big) = \mathbb{E}_\vartheta\big[\mathbf{1}_{\{T\in B\}}\mathbf{1}_{\{U\in B'\}} \,\big|\, T\big] = \mathbf{1}_{\{T\in B\}}\mathbb{P}_\vartheta(U \in B')$$

(wobei wir für die 2. Gleichheit verwenden, dass T und U unabhängig sind). Da U verteilungsfrei ist, hängt dies nicht von ϑ ab, d.h. T ist suffizient. □

1.2 Konfidenzintervalle (und Konfidenzbereiche)

Definition 18. Sei $\mathscr{M} = (\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\tau(\vartheta)$ reelles Parametermerkmal, L, R Statistiken mit $L \le R$, $\alpha \in (0,1)$.

Das (zufällige) Intervall $I := [L,R]$ heißt ein *Konfidenzintervall* (manchmal auch „Vertrauensintervall“) für τ zum (Sicherheits-)Niveau $1-\alpha$ (bzw. Irrtumsniveau α), wenn gilt

$$\forall\, \vartheta \in \Theta : \quad \mathbb{P}_\vartheta\big(\tau(\vartheta) \in I\big) \ge 1-\alpha.$$

Beachte: I ist zufällig, nicht aber ϑ (zumindest in unserer hier verwendeten (sogenannten frequentistischen) Interpretation).

Allgemeiner heißt eine in Abhängigkeit von den Beobachtungen $x \in \Omega$ konstruierte Menge $C(x) \subset \Theta$ ein *Konfidenzbereich* für τ zum (Sicherheits-)Niveau $1-\alpha$, wenn gilt

$$\forall\, \vartheta \in \Theta : \quad \mathbb{P}_\vartheta\big(\{x \in \Omega : C(x) \ni \tau(\vartheta)\}\big) \ge 1-\alpha.$$

Offenbar möchte man i.A. I so kurz wie möglich wählen (soweit verträglich mit dem geforderten Niveau).

Beispiel (Konfidenzintervall für den Mittelwert im normalen Modell bei bekannter Varianz)**.** Unter $\mathbb{P}_\vartheta$ seien $X_1, X_2, \dots, X_n$ u.i.v. $\sim \mathcal{N}_{\vartheta,\sigma^2}$ mit $\vartheta \in \Theta := \mathbb{R}$, $\sigma^2 > 0$ sei bekannt (und fest).

Sei $q := \Phi^{-1}(1-\frac{\alpha}{2})$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung, $\overline{X} = \dfrac{1}{n}\sum_{i=1}^n X_i$.

$$I := \Big[\overline{X} - q\cdot\frac{\sigma}{\sqrt{n}}, \overline{X} + q\cdot\frac{\sigma}{\sqrt{n}}\Big]$$

ist ein Konfidenzintervall für ϑ zum (Sicherheits-)Niveau $1-\alpha$, denn unter $\mathbb{P}_\vartheta$ ist $\overline{X} \sim \mathcal{N}_{\vartheta,\sigma^2/n}$,

$$\begin{aligned}\mathbb{P}_\vartheta\Big(\overline{X} - q\cdot\frac{\sigma}{\sqrt{n}} \le \vartheta \le \overline{X} + q\cdot\frac{\sigma}{\sqrt{n}}\Big) &= \mathbb{P}_\vartheta\Big(q \ge \frac{\overline{X}-\vartheta}{\sigma/\sqrt{n}} \ge -q\Big)\\ &= \mathbb{P}(-q \le Z \le q) = \mathbb{P}(Z \le q) - \mathbb{P}(Z \ge -q) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha\end{aligned}$$

(mit $Z \sim \mathcal{N}(0,1)$).

Beispiel (Student-Konfidenzintervall für den Erwartungswert im normalen Modell)**.** Unter $\mathbb{P}_\vartheta$, $\vartheta = (\mu,\sigma^2) \in \Theta := \mathbb{R}\times(0,\infty)$ seien

$$X_1, X_2, \ldots, X_n \text{ u.i.v. } \sim \mathcal{N}_{\mu,\sigma^2}.$$

Sei $\alpha \in (0,1)$, $q = q_{n-1,1-\alpha/2}$ das $1-\frac{\alpha}{2}$-Quantil der Student-Verteilung mit $n-1$ Freiheitsgraden,

$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Dann ist

$$I := \Big[\overline{X} - q\sqrt{\frac{S^2}{n}}, \overline{X} + q\sqrt{\frac{S^2}{n}}\,\Big]$$

ein Konfidenzintervall für μ zum Irrtumsniveau α.

Beweis. $T := \frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{S^2}}$ ist Student-verteilt mit $n-1$ Freiheitsgraden (für jede Wahl von μ und σ^2, siehe z.B. Proposition A.2.4 in Anhang A.2), somit

$$\begin{aligned}&\mathbb{P}_{(\mu,\sigma^2)}\Big(\overline{X} - q\sqrt{\tfrac{S^2}{n}} \le \mu \le \overline{X} + q\sqrt{\tfrac{S^2}{n}}\Big)\\ &= \mathbb{P}(-q \le T \le q) = \mathbb{P}(T \le q) - \mathbb{P}(T \le -q) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.\end{aligned}$$

□

Beispiel. Zwei Schlafmittel sollen verglichen werden, 10 Patienten erhielten in aufeinanderfolgenden Nächten Medikament A und B. Die Daten[1] (x_i = Anz. Stunden Schlaf mit Mittel A - Anz. Stunden Schlaf mit Mittel B bei Patient Nr. i):

i	1	2	3	4	5	6	7	8	9	10
x_i	1,2	2,4	1,3	1,3	0,0	1,0	1,8	0,8	4,6	1,4

Es ist

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i \approxeq 1{,}58, \quad s = \Big(\frac{1}{9}\sum_{i=1}^{10}(x_i-\overline{x})^2\Big)^{1/2} \approxeq 1{,}23.$$

Nehmen wir an, die Daten stammen aus einer Normalverteilung mit unbekanntem Mittelwert μ und unbekannter Varianz σ^2 (und die Ergebnisse der verschiedenen Patienten sind unabhängig).

[1] nach [Geo07, Bsp. 8.6], dies sind die Daten aus Sect. IX, Illustration 1 aus dem Originalpaper von Student (=W.S. Gossett), The Probable Error of a Mean, Biometrika, Vol. 6, No. 1 (Mar., 1908), pp. 1-25

Es ist $q_{9,0{,}995} \approxeq 3{,}25$ (aus einer Quantiltabelle oder beispielsweise mit R berechnet), demnach ist

$$\left[\overline{x} \pm q\frac{s}{\sqrt{n}}\right] \approxeq [0{,}31,\ 2{,}85]$$

ein Konfidenzintervall für μ (die mittlere zusätzliche Anzahl Stunden Schlaf, die Medikament A mehr bringt als Medikament B) zum Sicherheitsniveau $0{,}99 = 1 - 0{,}01$.

(Beachte: (Sinnlos) genaue Werte mit Rechnergenauigkeit sind $\overline{x} - q\frac{s}{\sqrt{n}} \approxeq 0{,}3159481$, $\overline{x} + q\frac{s}{\sqrt{n}} \approxeq 2{,}8440519$, man sollte allerdings die Grenzen eines Konfidenzintervalls stets „konservativ“, d.h. nach außen, runden.)

Beispiel (Approximatives Konfidenzintervall im Binomialmodell mittels Normalapproximation)**.** $X \sim \mathrm{Bin}_{n,\vartheta}$, $\theta \in \Theta = [0,1]$,
$\widehat{\vartheta} := \frac{X}{n}$, $\widehat{\sigma} := \sqrt{\widehat{\vartheta}(1-\widehat{\vartheta})}$, $\alpha \in (0,1)$, $q := \Phi^{-1}(1-\frac{\alpha}{2})$, dann ist

$$I := \left[\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

ein (approximatives) Konfidenzintervall für ϑ zum Sicherheitsniveau $1-\alpha$, denn unter $\mathbb{P}_\vartheta$ gilt für $n \to \infty$

$$\widehat{\vartheta} \xrightarrow{d} \vartheta, \quad \widehat{\sigma} \xrightarrow{d} \sqrt{\vartheta(1-\vartheta)}, \quad \text{und} \quad \frac{\widehat{\vartheta} - \vartheta}{\widehat{\sigma}/\sqrt{n}} = \frac{X - n\vartheta}{\widehat{\sigma}\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

(mit Satz von de Moivre-Laplace)

$$\mathbb{P}_\vartheta\Big(\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}} \le \vartheta \le \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\Big) = \mathbb{P}_\vartheta\Big(-q \le \frac{\widehat{\vartheta} - \vartheta}{\widehat{\sigma}/\sqrt{n}} \le q\Big) \approx \mathbb{P}(-q \le Z \le q) = 1 - \alpha \tag{1.11}$$

Gelegentlich hört man die „Faustregel“, dass $np(1-p) \ge 9$ sein sollte, damit die Approximation in (1.11) brauchbar ist. Siehe auch Anhang A.3 für exakte Konfidenzintervalle im Binomialmodell.

Ein Konfidenzintervall für den Median: Ein Mini-Ausflug in die nicht-parametrische Statistik

Beispiel (Ein Konfidenzintervall für den Median)**.** Seien $X_1, \ldots, X_n$ u.i.v. reellwertig, mit (unbekannter) Verteilung Q, die eine stetige Verteilungsfunktion besitzt (d.h. Q hat keine Atome). (Im Formalismus: $\Theta = \{\vartheta : \vartheta \text{ nicht-atomares W'maß auf } \mathbb{R}\}$, $\Omega = \mathbb{R}^n$, $\mathbb{P}_\vartheta = \vartheta^{\otimes n}$)

$m(Q)$ sei „der“ Median von Q (d.h. $Q\big((-\infty, m(Q)]\big) = \frac{1}{2} = Q\big([m(Q), \infty)\big)$; falls mehrere Werte in Frage kommen, nehmen wir das arithmetische Mittel aus dem kleinsten und dem größten möglichen Wert).

Die zugehörige Ordnungsstatistik ist

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

Zu $\alpha \in (0,1)$ wähle k maximal, so dass $\mathrm{Bin}_{n,1/2}(\{0, \ldots, k-1\}) \le \frac{\alpha}{2}$, dann ist

$$\left[X_{(k)}, X_{(n-k+1)}\right]$$

ein Konfidenzintervall für den Median $m(Q)$ zum Sicherheitsniveau $1 - \alpha$.

Beweis. Es ist

$$Q^{\otimes n}\big(X_{(k)} > m(Q)\big) = Q^{\otimes n}\big(|\{1 \le i \le n : X_i \le m(Q)\}| \le k-1\big)$$
$$= \mathrm{Bin}_{n,1/2}(\{0,\dots,k-1\}) \le \frac{\alpha}{2},$$

analog ist

$$Q^{\otimes n}\big(X_{(n-k-1)} < m(Q)\big) = Q^{\otimes n}\big(|\{1 \le i \le n : X_i \ge m(Q)\}| \le k-1\big) \le \frac{\alpha}{2},$$

somit

$$Q^{\otimes n}\Big(\big[X_{(k)}, X_{(n-k+1)}\big] \not\ni m(Q)\Big) \le Q^{\otimes n}\big(X_{(k)} > m(Q)\big) + Q^{\otimes n}\big(X_{(n-k-1)} < m(Q)\big) \le \alpha.$$

□

Im „Schlafmittel-Vergleich"-Beispiel oben mit $n = 10$ ergäbe sich für $\alpha = 0{,}01$: Man muss in $k = 1$ wählen (es ist $\mathrm{Bin}_{10,1/2}(\{0\}) \approxeq 0{,}001$, aber $\mathrm{Bin}_{10,1/2}(\{0,1\}) \approxeq 0{,}012$), d.h. ein Konfidenzintervall für den Median (der Differenz der Schlafdauer unter Mittel A versus Mittel B) zum Sicherheitsniveau 99% ist $[X_{(1)}, X_{(10)}] = [0,\ 4{,}6]$.

(Für $\alpha = 0{,}05$ könnte man $k = 2$ wählen und erhielte $[X_{(2)}, X_{(9)}] = [0{,}8,\ 2{,}4]$ als Konfidenzintervall zum Sicherheitsniveau 95%.)

1.3 Statistische Tests

Beispiel (für einen einseitiger Binomialtest)**.** Herr A behauptet, (mit W'keit $\vartheta > 1/2$) vorhersagen zu können, ob die oberste Karte eines verdeckten, gut gemischten Skatblatts rot oder schwarz ist.

Frau B ist skeptisch (und verdächtigt, dass A einfach rät, d.h. $\vartheta = 1/2$) und schlägt vor, $n = 20$ Versuche durchzuführen.

Sei

$$X := \text{Anzahl richtige Vorhersagen von A,}$$

wir modellieren $X \sim \mathrm{Bin}_{n,\vartheta}$ (mit uns unbekanntem $\vartheta \in [0,1]$).

B wählt $\alpha = 0{,}05$, sagen wir, und k (möglichst klein) mit (und hier $\vartheta_0 := 1/2$)

$$\mathrm{Bin}_{n,\vartheta_0}\big(\{k, k+1, \dots, n\}\big) \le \alpha$$

(hier $k = 15$, denn $\mathrm{Bin}_{n,\vartheta_0}(\{15,16,\dots,20\}) \approxeq 0{,}021$, $\mathrm{Bin}_{n,\vartheta_0}(\{14,15,\dots,20\}) \approxeq 0{,}058$) und wird (auf dem Signifikanzniveau α) die

$$\text{Nullhypothese} : \vartheta \le \frac{1}{2}$$

verwerfen zugunsten der

$$\text{Alternative} : \vartheta > \frac{1}{2},$$

wenn das Ereignis $\{X \geq k\}$ eintritt, ansonsten die Nullhypothese beibehalten.

Demnach: Falls A tatsächlich rät (also in Wirklichkeit $\vartheta = 1/2$ gilt), ist die W'keit, ihm versehentlich hellseherische Fähigkeiten zuzuschreiben (dies wäre dann ein sogenannter „Fehler 1. Art“: die Nullhypothese abzulehnen, obwohl sie zutrifft), höchsten α.

Nehmen wir an, die 20 Versuche werden durchgeführt und A erzielt 13 „Treffer“. B wird also die Nullhypothese beibehalten (denn $\{X \geq k\}$ tritt nicht ein; quantitativer: der p-Wert ist $P_{1/2}(X \geq 13) \approxeq 0{,}132 > 0{,}05$; d.h., wenn A einfach nur rät, würde er in ca. 13,2% der Fälle mindestens ebensoviele „Treffer“ erzielen wie beobachtet) und etwa sagen:

„Die Beobachtungen zeigen (auf dem Niveau $\alpha = 5\%$) keine keine signifikante Abweichung von der Nullhypothese.“

Der formale Rahmen statistischer Tests

Definition 19. Sei $(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell, $\Theta = \Theta_0 \dot\cup \Theta_1$ disjunkte Zerlegung in „Nullhypothese“ und „Alternative“ (auch „Gegenhypothese“).

Eine Statistik $\varphi : \Omega \to [0,1]$ heißt ein *Test* von Θ_0 gegen Θ_1.

Der Test heißt *randomisiert*, wenn $\varphi(X) \notin \{0,1\}$, sonst *nicht-randomisiert*,

für einen nicht randomisierten Test φ heißt

$$\{x : \varphi(x) = 1\} \text{ der Ablehnungs- oder Verwerfungsbereich (von } \Theta_0).$$

$$G_\varphi : \Theta \to [0,1], \quad G_\varphi(\vartheta) = \mathbb{E}_\vartheta[\varphi]$$

heißt die *Gütefunktion* von φ ($1 - G_\varphi$ heißt Operationscharakteristik), φ heißt ein Test zum Niveau (auch: Signifikanzniveau) $\alpha \in (0,1)$, wenn gilt

$$\sup_{\vartheta \in \Theta_0} G_\varphi(\vartheta) \leq \alpha.$$

$\sup_{\vartheta \in \Theta_0} G_\varphi(\vartheta)$ heißt effektives Niveau (auch: Umfang) von φ.

Für $\vartheta \in \Theta_1$ heißt $G_\varphi(\vartheta)$ die Macht (auch: Schärfe, englisch: power) des Tests φ bei ϑ.

Sei $(\varphi_\alpha)_{\alpha \in (0,1)}$ eine Familie von nicht-randomisierten Tests mit $\varphi_\alpha \leq \varphi_{\alpha'}$ für $\alpha \leq \alpha'$, φ_α habe effektives Niveau α. Dann heißt für $x \in \Omega$

$$p\big(\,= p(x)\big) = \inf\{\alpha \in (0,1) : \varphi_\alpha(x) = 1\}$$

der *p-Wert* (bei Beobachtung x).

Interpretation.

1. Man interpretiert φ als Entscheidungsregel: Bei gegebener Beobachtung x

$\varphi(x) = 0$	:	behalte Nullhypothese bei
$\varphi(x) = 1$	:	verwirf Nullhypothese, entscheide für die Alternative
$\varphi(x) \in (0,1)$	:	wirf eine Münze, die mit W'keit $\varphi(x)$ für die Alternative entscheidet

2. Niveau α bedeutet, dass die W'keit für einen „Fehler 1. Art" (die Nullhypothese fälschlicherweise zu verwerfen) $\leq \alpha$ ist (uniform in $\vartheta \in \Theta_0$).

3. Für $\vartheta \in \Theta_1$ ist $1 - G_\varphi(\vartheta)$ die Wahrscheinlichkeit, einen „Fehler 2. Art" zu begehen (die Nullhypothese fälschlicherweise zu akzeptieren).

4. Viele „praktische" Tests haben die folgende Form, z.B. der z-Test, die t-Tests, der χ^2-Test: Berechne eine gewisse (Test-)Statistik Y (aus den Beobachtungen), verwirf die Nullhypothese, wenn $Y > q$ für einen gewissen Wert $q = q(\alpha)$, der in Abhängigkeit von den Parametern des Tests (insbesondere dem gewünschten Niveau α) gewählt wird. In der Sprache von Definition 19 also: $\varphi_\alpha(x) = \mathbf{1}(Y(x) > q(\alpha))$ und $\sup_{\vartheta\in\Theta_0} \mathbb{E}_\vartheta[\varphi] = \alpha$.

   Dann kann man den p-Wert des Test(ergebnisses) interpretieren als die Wahrscheinlichkeit, bei Gültigkeit der Nullhypothese einen mindestens so „extremen" Wert der Teststatistik zu finden wie den tatsächlich anhand der Daten beobachteten.

Demnach sind für φ wünschenswert:

$$G_\varphi \text{ sollte auf } \Theta_1 \text{ möglichst groß sein}$$

(solange mit dem gewünschten Signifikanzniveau verträglich), zudem sollte für einen Test zum Niveau α gelten

$$\sup_{\vartheta\in\Theta_0} G_\varphi(\vartheta) \leq \alpha \leq \inf_{\vartheta\in\Theta_1} G_\varphi(\vartheta) \qquad \text{(dann heißt } \varphi \text{ „unverfälscht").}$$

Beispiel (Binomialtest, allgemein)**.** $\Theta = [0,1]$, unter $\mathbb{P}_\vartheta$ sei die Beobachtung $X \sim \mathrm{Bin}_{n,\vartheta}$ (oder aber Beobachtungen $X_1, \dots, X_n$ sind unter $\mathbb{P}_\vartheta$ u.i.v. $\sim \mathrm{Ber}_\vartheta$ und wir bilden $X := X_1 + \cdots + X_n$). Wähle $\alpha \in (0, 1/2)$.

1. Zweiseitiger Binomialtest: $\Theta_0 = \{\vartheta_0\}$ für ein $\vartheta \in [0,1]$, $\Theta_1 = \Theta \setminus \Theta_0$. Setze

$$\begin{aligned} c_\ell &:= \max\left\{x \in \{0,1,2,\dots,n\} : \mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\}) \leq \alpha/2\right\}, \\ c_r &:= \min\left\{x \in \{0,1,2,\dots,n\} : \mathrm{Bin}_{n,\vartheta_0}(\{x,x+1,\dots,n\}) \leq \alpha/2\right\}, \\ \varphi(x) &:= \mathbf{1}_{\{0,1,\dots,c_\ell\}}(x) + \mathbf{1}_{\{c_r,c_r+1,\dots,n\}}(x). \end{aligned}$$

Dann gilt

$$\begin{aligned} \mathbb{E}_{\vartheta_0}[\varphi(X)] &= \mathbb{P}_{\vartheta_0}(X \leq c_\ell) + \mathbb{P}_{\vartheta_0}(X \geq c_r) \\ &= \mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,c_\ell\}) + \mathrm{Bin}_{n,\vartheta_0}(\{c_r, c_r+1,\dots,n\}) \leq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha \end{aligned}$$

nach Konstruktion, d.h. der Test hält Niveau α ein. (Wegen der Diskretheit der möglichen Beobachtungen ist das tatsächliche Niveau i.A. etwas kleiner.)

Bei gegebener Beobachtung x ist der p-Wert dann $2\mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\})$ falls $x < n\vartheta_0$ und $2\mathrm{Bin}_{n,\vartheta_0}(\{x,x+1,\dots,n\})$ falls $x > n\vartheta_0$.

2. a) Einseitiger Binomialtest (linksseitige Alternative): $\Theta_0 = [\vartheta_0, 1]$ für ein $\vartheta \in (0,1]$, $\Theta_1 = [0,\vartheta_0) = \Theta \setminus \Theta_0$. Setze

$$\begin{aligned} c &:= \max\left\{x \in \{0,1,2,\dots,n\} : \mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\}) \leq \alpha\right\}, \\ \varphi(x) &:= \mathbf{1}_{\{0,1,\dots,c\}}(x). \end{aligned}$$

Nach Konstruktion ist $\mathbb{E}_{\vartheta_0}[\varphi(X)] = \mathbb{P}_{\vartheta_0}(X \le c) \le \alpha$ und man kann (leicht) zeigen, dass für $\vartheta > \vartheta_0$ gilt $P_\vartheta(X \le c) \le P_{\vartheta_0}(X \le c)$ $(\le \alpha)$, d.h. der Test hält Niveau α ein.

Bei gegebener Beobachtung x ist der p-Wert dann $\mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\})$.

2. b) Einseitiger Binomialtest (rechtssseitige Alternative): $\Theta_0 = [0,\vartheta_0]$ für ein $\vartheta \in [0,1)$, $\Theta_1 = (\vartheta_0,1] = \Theta \setminus \Theta_0$. Analog setze

$$\begin{aligned} C &:= \max\big\{x \in \{0,1,2,\dots,n\} : \mathrm{Bin}_{n,\vartheta_0}(\{x,x,+1,\dots,n\}) \le \alpha\big\}, \\ \varphi(x) &:= \mathbf{1}_{\{C,C+1,\dots,n\}}(x). \end{aligned}$$

Der Test hält Niveau α ein, bei gegebener Beobachtung x ist der p-Wert dann $\mathrm{Bin}_{n,\vartheta_0}(\{x,x,+1,\dots,n\})$.

Bemerkung. Offenbar benötigt man die Verteilungsfunktion der Binomialverteilung, um die „kritischen Werte“ c_ℓ, c_r bzw. c, C für den Binomialtest bei vorgegebenem n und α zu bestimmen. Für kleine Werte von n kann man diese „von Hand“ bestimmen, für größere Werte konsultiert man entweder ein Computerprogramm oder eine entsprechende Tabelle oder man verwendet die Normalapproximation der Binomialverteilung: Mit dem Satz von deMoivre-Laplace (oder auch mit dem zentralen Grenzwertsatz) ist

$$\mathrm{Bin}_{n,\vartheta_0}(\{0,1,\dots,x\}) \approxeq \Phi\left(\frac{x - n\vartheta_0}{\sqrt{n\vartheta_0(1-\vartheta_0)}}\right)$$

wobei $\Phi = F_{\mathcal{N}_{0,1}}$ die Verteilungsfunktion der Standard-Normalverteilung ist.

Beispiel (z-Test oder Gauß-Test)**.** $\Theta = \mathbb{R}$, unter $\mathbb{P}_\vartheta$ seien die Beobachtungen $X_1,\dots,X_n$ u.i.v. $\sim \mathcal{N}_{\vartheta,\sigma^2}$ mit bekanntem, festem $\sigma^2 > 0$. Wähle $\alpha \in (0,1)$.

1. Zweiseitiger z-Test: $\Theta_0 = \{\vartheta_0\}$ für ein $\vartheta \in \mathbb{R}$, $\Theta_1 = \Theta \setminus \Theta_0$

$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i, \quad Z := \frac{\overline{X} - \vartheta_0}{\sqrt{\sigma^2/n}}$$

mit $q := \Phi^{-1}(1-\alpha/2)$ (das $(1-\alpha/2)$-Quantil der Standardnormalverteilung) ist

$$\varphi(X_1,\dots,X_n) := \mathbf{1}_{\{|Z|>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α (denn unter $\mathbb{P}_{\vartheta_0}$ ist $\overline{X} \sim \mathcal{N}_{\vartheta_0,\sigma^2/n}$).

Der p-Wert ist dann $2\big(1-\Phi(|Z|)\big)$, wobei Φ die Verteilungsfunktion der Standard-Normalverteilung ist.

2. Einseitiger Test: $\Theta_0 = \{\vartheta : \vartheta \le \vartheta_0\}$ für ein $\vartheta \in \mathbb{R}$, $\Theta_1 = \Theta \setminus \Theta_0 = \{\vartheta : \vartheta > \vartheta_0\}$.

Mit $q := \Phi^{-1}(1-\alpha)$ ist

$$\varphi(X_1,\dots,X_n) := \mathbf{1}_{\{Z>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α. Als p-Wert ergibt sich $1-\Phi(Z)$

(Je nach Anwendungssituation kann man auch $\Theta_0 = \{\vartheta : \vartheta \ge \vartheta_0\}$ betrachten, dann ist $\varphi(X_1,\dots,X_n) := \mathbf{1}_{\{Z<-q\}}$ zu wählen und der p-Wert wäre $\Phi(Z) = 1-\Phi(-Z)$).

Beispiel ((ein-Stichproben- oder gepaarter) t-Test)**.** $\Theta = \mathbb{R} \times (0,\infty) \ni \vartheta = (\mu, \sigma^2)$, unter $\mathbb{P}_\vartheta$ seien die Beobachtungen $X_1, \dots, X_n$ u.i.v. $\sim \mathcal{N}_{\mu,\sigma^2}$ (mit unbekanntem $\mu \in \mathbb{R}$ und unbekanntem $\sigma^2 > 0$). Wähle $\alpha \in (0,1)$.

Sei $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 := \frac{1}{n-1}\sum_{i=1}^{n-1}(X_i - \overline{X})^2$.

1. Zweiseitiger (ein-Stichproben) t-Test: $\Theta_0 = \{\vartheta = (\mu, \sigma^2) \in \Theta : \mu = \mu_0\}$ für ein $\mu_0 \in \mathbb{R}$ (man schreibt dies oft knapp als „$\Theta_0 : \mu = \mu_0$“), $\Theta_1 = \Theta \setminus \Theta_0$.

$$T := \sqrt{n}\frac{\overline{X} - \mu_0}{\sqrt{S^2}}$$

Mit $q := q_{n-1,1-\alpha/2} = (1-\alpha/2)$-Quantil der Student-$(n-1)$-Verteilung ist

$$\varphi(X_1, \dots, X_n) := \mathbf{1}_{\{|T|>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α (denn nach Satz A.2.4 ist T für jedes $\vartheta \in \Theta_0$ unter $\mathbb{P}_\vartheta$ Student-$(n-1)$-verteilt).

Der p-Wert ist $2\big(1 - F_{T_{n-1}}(|T|)\big)$ mit $F_{T_{n-1}}$ der Verteilungsfunktion der Student-$(n-1)$-Verteilung.

2. Einseitiger Test: $\Theta_0 = \{\vartheta = (\mu, \sigma^2) \in \Theta : \mu \le \mu_0\}$ für ein $\mu_0 \in \mathbb{R}$ (oft knapp geschrieben als „$\Theta_0 : \mu \le \mu_0$“), $\Theta_1 = \Theta \setminus \Theta_0$. Mit $q := q_{n-1,1-\alpha} = (1-\alpha)$-Quantil der Student-$(n-1)$-Verteilung ist

$$\varphi(X_1, \dots, X_n) := \mathbf{1}_{\{T>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α (und analog $\varphi(X_1, \dots, X_n) := \mathbf{1}_{\{T<-q\}}$ ein Test für $\Theta_0 = \{\mu \ge \mu_0\}$, beachte auch: $-q$ ist das α-Quantil der Student-$(n-1)$-Verteilung).

Der p-Wert ist $1 - F_{T_{n-1}}(T)$.

(Je nach Anwendungssituation kann man auch $\Theta_0 = \{\vartheta = (\mu, \sigma^2) \in \Theta : \mu \ge \mu_0\}$ betrachten, dann ist $\varphi(X_1, \dots, X_n) := \mathbf{1}_{\{T<-q\}}$ zu wählen und der p-Wert wäre $F_{T_{n-1}}(T) = 1 - F_{T_{n-1}}(-T)$).

Ein **Anwendungsbeispiel.** a) Die Wirksamkeit eines gewissen Schlafmittels soll geprüft werden. 10 Patienten erhalten das Schlafmittel, die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet.

Wir nehmen an, die Beob. sind u.i.v. $\sim \mathcal{N}_{\mu,\sigma^2}$ und wir möchten die Nullhypothese $\mu = 0$, sagen wir, zum Niveau $\alpha = 0{,}05$ testen.

Die Daten[2]:

Patient i	1	2	3	4	5	6	7	8	9	10
zus. Schl.	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0

Es ist $n = 10$, $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 0{,}75$, $s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 \approxeq 1{,}79$, $t = \frac{\overline{x}-0}{s/\sqrt{10}} \approxeq 1{,}326$

Das 0,975-Quantil der Student-9-Verteilung ist $\approxeq 2{,}262$, demnach können wir die Nullhypothese nicht ablehnen.

[2]Widerum aus Student (William S. Gosset), The Probable Error of a Mean, Biometrika 6:1–25 (1908)

(Für ein Student-9-verteiltes T ist $P(|T| \geq 1{,}326) \cong 0{,}2176$, dies ist der p-Wert des Tests.)

Man kann diesen Befund folgendermaßen formulieren:

„Die Beobachtungen sind mit der Nullhypothese $\mu = 0$ (im statistischen Sinne) verträglich." oder

„Die beobachtete Abweichung $\overline{x} = 0{,}75$ ist nicht signifikant von 0 verschieden (t-Test, $\alpha = 0{,}05$)."

b) Die Wirksamkeit eines Schlafmittels soll mit der eines anderen verglichen werden. 10 Patienten erhalten Schlafmittel A, die Anzahl zusätzlicher Stunden Schlaf wird in einer Nacht beobachtet. Dann erhalten dieselben 10 Patienten Schlafmittel B, wieder wird die Anzahl zusätzlicher Stunden Schlaf in einer Nacht beobachtet.

Da dieselben Patienten untersucht werden, können (und sollten) wir die Messungen paaren: Wir interessieren uns bei jedem Patienten für die Differenz des (zusätzlichen) Schlafs bei Mittel 2 und bei Mittel 1.

Wir nehmen an, die beobachteten Differenzen sind Realisierungen von u.i.v. ZVn mit Vert. $\mathcal{N}_{\mu,\sigma^2}$ und wir möchten die Nullhypothese $\mu \leq 0$ gegen die Alternative $\mu > 0$, sagen wir, zum Niveau $\alpha = 0{,}05$ testen.

(Dies wäre beispielsweise in folgender Situation angemessen: Wir möchten darlegen, dass Mittel B wirksamer ist als Mittel A, indem wir die Nullhypothese „$\mu \leq 0$" entkräften.)

Die Daten (wiederum aus Student, a.a.O.) :

Patient i	1	2	3	4	5	6	7	8	9	10
Mittel A	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
Mittel B	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4
Diff.	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

Es ist $n = 10$, $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 1{,}58$, $s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 \cong 1{,}23$, $t = \frac{\overline{x}-0}{s/\sqrt{10}} \cong 4{,}062$

Das 0,95-Quantil der Student-9-Verteilung ist $\cong 1{,}833$, demnach können wir die Nullhypothese ablehnen.

(Für ein Student-9-verteiltes T ist $P(T > 4{,}062) \cong 0{,}0014$, dies ist der p-Wert des Tests.)

Mögliche knappe Formulierung dieses Befunds:

„Die beobachtete Differenz $\overline{x} = 1{,}58$ ist signifikant größer als 0 (einseitiger t-Test, $\alpha = 0{,}05$)."

Beispiel (Test für die Varianz im normalen Modell)**.** In der Situation von Beispiel 1.3 sei $\Theta_0 = \{\vartheta = (\mu, \sigma^2) \in \Theta : \sigma^2 \leq v_0\}$ für ein $v_0 > 0$, $\Theta_1 = \Theta \setminus \Theta_0$. Wähle $\alpha \in (0, 1)$.

Mit $q :=$ $(1-\alpha)$-Quantil der χ^2_{n-1}-Verteilung ist

$$\varphi(X_1, \ldots, X_n) := \mathbf{1}_{\{S^2 > q v_0/(n-1)\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α (vgl. Satz A.2.4).

Beispiel (zwei-Stichproben oder ungepaarter t-Test [mit Annahme gleicher Varianzen])**.** $\Theta = \mathbb{R} \times \mathbb{R} \times (0,\infty) \ni \vartheta = (\mu_1, \mu_2, \sigma^2)$, unter $\mathbb{P}_\vartheta$ sind $X_1, \dots, X_m$ u.i.v. und davon unabhängig $Y_1, \dots, Y_n$ u.i.v. $(m, n \in \mathbb{N})$, $X_i \sim \mathcal{N}_{\mu_1,\sigma^2}$, $Y_j \sim \mathcal{N}_{\mu_2,\sigma^2}$. Seien

$$\overline{X} := \frac{1}{m}\sum_{i=1}^{m} X_i,\ \overline{Y} := \frac{1}{n}\sum_{j=1}^{n} Y_j$$

die jeweiligen Stichprobenmittelwerte,

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2,\ S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2,$$

die (korrigierten) Stichprobenvarianzen,

$$S^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \quad \left(= \frac{1}{m+n-2}\Big(\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2 \Big) \right),$$

(die „gepoolte Stichprobenvarianz“),

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

(Beachte: Stets gilt $\mathbb{E}_{(\mu_1,\mu_2,\sigma^2)}[S^2] = \sigma^2$ [S^2 ist ein erwartungstreuer Schätzer für σ] und T ist unter $P_{(\mu_1,\mu_2,\sigma^2)}$ Student-$(n+m-2)$-verteilt, Argument analog zum Beweis von Proposition A.2.4 in Anhang A.2).

1. Zweiseitiger ungepaarter t-Test : $\Theta_0 = \{(\mu_1, \mu_2, \sigma^2) \in \Theta : \mu_1 = \mu_2\}$ (oft knapp geschrieben als „$\Theta_0 : \mu_1 = \mu_2$“), $\Theta_1 = \Theta \setminus \Theta_0$.

Wähle $\alpha \in (0,1)$, mit $q := q_{m+n-2,1-\alpha/2} = (1-\alpha/2)$-Quantil der Student-$(m+n-2)$-Verteilung ist

$$\varphi(X_1, \dots, X_m, Y_1, \dots, Y_n) := \mathbf{1}_{\{|T|>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α.

1. Einseitiger Test : $\Theta_0 = \{(\mu_1, \mu_2, \sigma^2) \in \Theta : \mu_1 \le \mu_2\}$ (oft knapp geschrieben als „$\Theta_0 : \mu_1 \le \mu_2$“), $\Theta_1 = \Theta \setminus \Theta_0$. Mit $q := q_{m+n-2,1-\alpha} = (1-\alpha)$-Quantil der Student-$(m+n-2)$-Verteilung ist

$$\varphi(X_1, \dots, X_m, Y_1, \dots, Y_n) := \mathbf{1}_{\{T>q\}}$$

ein Test von Θ_0 gegen Θ_1 zum Niveau α.

(Analog ist $\varphi(X_1, \dots, X_m, Y_1, \dots, Y_n) := \mathbf{1}_{\{T<-q\}}$ ein Test von $\Theta_0 = \{(\mu_1, \mu_2, \sigma^2) \in \Theta : \mu_1 \ge \mu_2\}$.)

(p-Werte werden analog zum ein-Stichproben-Fall (Bsp. 1.3) berechnet, wobei $F_{T_{n-1}}$ durch $F_{T_{m+n-2}}$ ersetzt wird.)

Beispiel (zwei-Stichproben-t-Test ohne Annahme gleicher Varianz, Welchs t-Test[3])**.** Es gibt auch eine Version des zwei-Stichproben-t-Tests, der die Annahme gleicher Varianzen nicht trifft (wir werden ihn im Verlauf der Vorlesung allerdings nicht verwenden):

Man schätzt die Streuung von $\overline{X} - \overline{Y}$ durch

$$\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} \quad \text{und bildet} \quad T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}.$$

Unter $P_{(\mu_0,\mu_0,\sigma_1^2,\sigma_2^2)}$ ist T „approximativ Student-verteilt mit g Freiheitsgraden“, wobei

$$g = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{s_X^4}{n_X^2(n_X-1)} + \frac{s_Y^4}{n_Y^2(n_Y-1)}} \tag{1.12}$$

aus den Daten geschätzt wird.

Seien die Werte $T = t$ und g beobachtet worden, man verwirft die Nullhypothese „$\mu_1 = \mu_2$“ (zum Niveau α), wenn $1 - F_{T_g}(t) \leq \alpha/2$, wobei F_{T_g} die Verteilungsfunktion der Student-Verteilung mit g Freiheitsgraden, d.h. wenn die Wahrscheinlichkeit, dass eine Student-verteilte Zufallsgröße mit g Freiheitsgraden einen betragsmäßig mindestens so großen Wert wie den beobachteten t-Wert annimmt, $\leq \alpha$ ist.

(Wir hatten in Korollar A.2.3 die Student-Verteilung nur für ganzzahlige Werte von n definiert, aber man kann dort allgemeine Werte $n > 0$ zulassen).

Dieser Test hat approximativ Niveau α und wird in der Praxis häufig verwendet. Beispielsweise führt der Befehl `t.test` in dem Statistikprogramm `R` automatisch diese Version des zwei-Stichproben-t-Tests durch, wenn man zwei Stichproben übergibt und keine weiteren Zusatzparameter setzt. Siehe Anhang A.4 zur Motivation der Setzung (1.12).

Bemerkung (Zur „reinen Lehre“ des statistischen Testens)**.** Nehmen wir an, wir möchten eine gewisse Aussage anhand experimenteller oder empirischer Daten statistisch prüfen. Das korrekte („lehrbuchmäßige“) Vorgehen sieht folgendermaßen aus:

1. Statistisches Modell formulieren, Nullhypothese und Alternative angeben (was die Nullhypothese ist, hängt von der konkreten Anwendungsfrage ab, oft ernennt man „das Gegenteil dessen, was man erhärten möchte“ zur Nullhypothese).

2. Dann einen Test (einschließlich gewünschtem Niveau) festlegen.

3. Dann erst: Daten erheben (bzw. Daten anschauen), Test-Entscheidung fällen.

Die Kontrolle der Fehlerwahrscheinlichkeiten, die die Theorie des statistischen Testens liefert, bezieht sich auf dieses Vorgehen. Wenn man die Reihenfolge herumdreht, also zuerst die Daten anschaut und dann einen Test wählt, verfälscht man strenggenommen zumindest das Signifikanzniveau, möglicherweise bis ins Unsinnige (Beispiel: zuerst den empirischen Mittelwert bestimmen, dann je nachdem, ob er links oder rechts von ϑ_0 liegt, entscheiden, ob man eine rechts- oder eine linksseitige Alternative wählt, ist offenbar „geschummelt“.)

[3] B. L. Welch, The Significance of the Difference between Two Means When the Population Variances Are Unequal, Biometrika 29:350–362, (1938)

Man sollte dieselben Daten nicht für explorative Statistik (d.h. Beobachtungen, die zu neuen Hypothesen führen [sollen]) und schließende Statistik (d.h. Beobachtungen, anhand denen eine Hypothese getestet werden soll) zugleich verwenden.

Beobachtung 20 (Zur Äquivalenz von Konfidenzbereichen und Tests)**.** Sei $C(X)$ ein Konfidenzbereich für das Parametermerkmal $\tau(\vartheta)$ zum Sicherheitsniveau $1-\alpha$, d.h.

$$\mathbb{P}_\vartheta\big(C(X) \ni \tau(\vartheta)\big) \geq 1-\alpha \quad \text{für alle } \vartheta \in \Theta$$

Betrachte die *Hypothese* $H_0 : \tau(\vartheta) = \tau_0$ und die *Alternative* $H_1 : \tau(\vartheta) \neq \tau_0$ und folgenden Test: Lehne H_0 ab, wenn $C(X) \ni \tau_0$, sonst nehme H_0 an.

Dies ist ein valider Test zum Signifikanzniveau α, denn

$$\sup_{\vartheta:\tau(\vartheta)=\tau_0} \mathbb{P}_\vartheta\big(\text{„lehne } H_0 \text{ ab“}\big) = \sup_{\vartheta:\tau(\vartheta)=\tau_0} \mathbb{P}_\vartheta\big(C(X) \not\ni \tau_0\big) \sup_{\vartheta:\tau(\vartheta)=\tau_0} \mathbb{P}_\vartheta\big(C(X) \not\ni \tau(\vartheta)\big) \leq \alpha$$

Korollar (Vorzeichentest für den Median)**.** Wir können unsere Konfidenzintervalle für den Median (vgl. Seite 18) somit folgendermaßen in einen Test verwandeln:

Seien $X_1, \dots, X_n$ u.i.v. reellwertig, $\sim \vartheta$ unter $\mathbb{P}_\vartheta$ (und ϑ habe stetige Verteilungsfunktion), $\alpha \in (0,1)$, wähle k maximal mit $Bin_{n,\frac{1}{2}}(\{0,\dots,k-1\}) \leq \frac{\alpha}{2}$, teste $H_0 : m(\vartheta) = m_0$ gegen $H_1 : m(\vartheta) \neq m_0$ folgendermaßen:

Wenn $[X_{(k)}, X_{(n-k+1)}] \ni m_0$, so nehme H_0 an, sonst lehne H_0 ab und nehme H_1 an. Der Fehler 1. Art dieses Tests ist $\leq \alpha$.

Analog für $H_0 : m(\vartheta) \leq m_0$ gegen $H_1 : m(\vartheta) > m_0$:

Wenn $X_{(k')} > m_0$, so lehne H_0 ab, wo k' so gewählt ist, dass $Bin_{n,\frac{1}{2}}(\{0,\dots,k'-1\}) \leq \alpha$ (und k' maximal mit dieser Eigenschaft).

1.3.1 Alternativtests und das Lemma von Neyman-Pearson

Wir betrachten ein Standardmodell (vgl. Def. 1) mit jeweils einpunktiger Nullhypothese und Alternative, d.h. $\Theta = \{0,1\}$, $\Omega \subset \mathbb{R}^n$ oder Ω diskret und $\mathbb{P}_i$ besitzt Dichte bzw. $\rho(i,x)$ auf Ω für $i = 0,1$.

Setze

$$R(x) := \begin{cases} \dfrac{\rho(1,x)}{\rho(0,x)} & \text{wenn } \rho(0,x) > 0, \\ \infty & \text{sonst.} \end{cases}$$

R heißt der *Likelihood-Quotient*.

Ein Test von $\mathbb{P}_0$ gegen $\mathbb{P}_1$ (formal hier wörtlich $\Theta_0 = \{0\}$ gegen $\Theta_1 = \{1\}$) der Form

$$\varphi(x) = \begin{cases} 1 & \text{für } R(x) > c, \\ 0 & \text{für } R(x) < c, \end{cases}$$

für ein $c \geq 0$ heißt ein Neyman-Pearson-Test[4]. (Im Fall $R(x) = c$ kann Randomisierung notwendig sein.)

[4]nach Jerzy Neyman, 1894–1981 und Egon Pearson, 1895–1980

Satz 21 (Neyman-Pearson-Lemma)**.** *Betrachte Standardmodell mit einpunktiger Nullhypothese und einpunktiger Alternative,* $\alpha \in (0,1)$.

1. Es gibt einen Neyman-Pearson-Test φ *mit* $\mathbb{E}_0[\varphi] = \alpha$.

2. Sei φ *ein Neyman-Pearson-Test* φ *mit* $\mathbb{E}_0[\varphi] = \alpha$, $\widetilde{\varphi}$ *irgendein Test von* $\mathbb{P}_0$ *gegen* $\mathbb{P}_1$ *zum Niveau* α. *Dann gilt* $\mathbb{E}_1[\varphi] \geq \mathbb{E}_1[\widetilde{\varphi}]$, *d.h. die Macht von* φ *ist mindestens so groß wie die von* $\widetilde{\varphi}$.

Man sagt: φ *ist (in dieser Situation) ein* gleichmäßig bester *Test.*

Beweis. 1. Wähle c mit $\mathbb{P}_0(R \geq c) \geq \alpha$ und $\mathbb{P}_0(R \leq c) \geq 1 - \alpha$.

Falls $\mathbb{P}_0(R = c) = 0$, so ist $\varphi(x) := \mathbf{1}_{\{R(x)>c\}}$ ein Neyman-Pearson-Test mit

$$\mathbb{E}_0[\varphi] = \mathbb{P}_0(R > c) = \mathbb{P}_0(R \geq c) = \alpha$$

Falls $\mathbb{P}_0(R = c) > 0$, so setze $\gamma := \dfrac{\alpha - \mathbb{P}_0(R > c)}{\mathbb{P}_0(R = c)}$ $(\in [0,1])$ und

$$\varphi(x) := \begin{cases} 1 & \text{wenn } R(x) > c, \\ \gamma & \text{wenn } R(x) = c, \\ 0 & \text{wenn } R(x) < c. \end{cases}$$

Dies ist ein Neyman-Pearson-Test mit $\mathbb{E}_0[\varphi] = 1 \cdot \mathbb{P}_0(R > c) + \gamma \mathbb{P}_0(R = c) + 0 \cdot \mathbb{P}_0(R < c) = \alpha$.

2. Sei φ ein Neyman-Pearson-Test (mit Schwellenwert c) mit $\mathbb{E}_0[\varphi] = \alpha$, $\widetilde{\varphi}$ irgendein Test mit $\mathbb{E}_0[\widetilde{\varphi}] \leq \alpha$. Es gilt

$$\text{für alle } x \in \Omega: \quad \big(\varphi(x) - \widetilde{\varphi}(x)\big)\big(\rho(1,x) - c\rho(0,x)\big) \geq 0$$

denn die beiden Faktoren haben dasselbe Vorzeichen (sofern der zweite $\neq 0$), da $\varphi(x) \geq \mathbf{1}(\rho(1,x) > c\rho(0,x))$. Somit

$$f_1(x) := \big(\varphi(x) - \widetilde{\varphi}(x)\big)\rho(1,x) \geq c\big(\varphi(x) - \widetilde{\varphi}(x)\big)\rho(0,x) =: c f_0(x) \tag{1.13}$$

und folglich

$$\mathbb{E}_1[\varphi] - \mathbb{E}_1[\widetilde{\varphi}] = \int_\Omega f_1(x)\,dx \geq c \int_\Omega f_0(x)\,dx = c(\alpha - \mathbb{E}_0[\widetilde{\varphi}]) \geq 0. \tag{1.14}$$

(Wenn diskret ist, muss das Integral natürlich durch eine Summe ersetzt werden.) □

Bemerkung. Wir sehen aus dem Beweis auch: Wenn $\widetilde{\varphi}$ ebenfalls ein gleichmäßig bester Test von $\mathbb{P}_0$ gegen $\mathbb{P}_1$ mit $\mathbb{E}_0[\widetilde{\varphi}] = \alpha$ ist, so gilt Gleichheit in (1.14) und daher auch Gleichheit in (1.13) (möglicherweise mit Ausnahme einer Menge von [Lebesgue-]Maß 0). In diesem Sinne ist also hier ein gleichmäßig bester Test „identisch“ mit einem Neyman-Pearson-Test.

Beispiel. Beobachtungen $X_1, \dots, X_n$ seien unter $\mathbb{P}_0$ u.i.v. mit $X_i \sim \mathcal{N}(\mu_0, \sigma^2)$, unter $\mathbb{P}_1$ u.i.v. mit $X_i \sim \mathcal{N}(\mu_1, \sigma^2)$, wobei $\sigma^2 > 0$ und $\mu_0 < \mu_1$ bekannt (und fest) sind.

Mit $x = (x_1, \dots, x_n)$ und $\overline{x} := \frac{1}{n}\sum_{i=1}^n x_i$ ist

$$\begin{aligned} R(x) &= \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2 \Big) \\ &= \exp\Big(-\frac{n}{2\sigma^2}\big(2(\mu_1 - \mu_0)\overline{x} + \mu_1^2 - \mu_0^2\big)\Big). \end{aligned}$$

Offenbar ist $R(x)$ eine montone (fallende) Funktion von $\overline{x}$ und unter $\mathbb{P}_0$ ist $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}_{\mu_0, \sigma^2/n}$, also ist

$$\varphi(x) := \mathbf{1}_{\{\overline{x} > c\}}$$

mit der Wahl $c := \mu_0 + \sqrt{\sigma^2/n}\,\Phi^{-1}(1-\alpha)$ ein Neyman-Pearson-Test zum Niveau $\alpha \in (0,1)$ (denn dann ist $\mathbb{E}_0[\varphi] = \mathbb{P}_0\big(\overline{X} > c\big) = \alpha$).

Satz 22 (Lemma von Stein)**.** *Im unendlichen Produktmodell* $(\Omega^{\mathbb{N}}, \mathscr{F}^{\otimes\mathbb{N}}, \mathbb{P}_\theta^{\otimes\mathbb{N}} : \theta \in \{0,1\})$*, sei* φ_n *ein Neyman-Pearson-Test mit* $\mathbb{E}_0[\varphi_n] = \alpha$*, der nur von den Beobachtungen* $X_1, \dots, X_n$ *abhängt. Dann gilt:*

$$\lim_{n\to\infty} \frac{1}{n}\log(1 - \mathbb{E}_1[\varphi_n]) = -H(\mathbb{P}_0 \mid \mathbb{P}_1),$$

d.h. $\mathbb{E}[\varphi_n] \approx 1 - e^{-nH(\mathbb{P}_0|\mathbb{P}_1)}$*, die Macht von* φ *konvergiert exponentiell schnell gegen* 1*.*

Für Maße $\mathbb{P}, \mathbb{Q}$ mit Dichten $\rho(x)$ bzw. $\sigma(x)$ bzgl. des Lebesgue-Maßes (oder bzgl. des Zählmaßes, wenn ΩX abzählbar ist) heißt

$$H(\mathbb{P} \mid \mathbb{Q}) := \int_\Omega \rho(x) \log \frac{\rho(x)}{\sigma(x)}\, dx$$

die *relative Entropie von* $\mathbb{P}$ *bzgl.* $\mathbb{Q}$, es gilt: $H(\mathbb{P} \mid \mathbb{Q}) \in [0, \infty]$ und $H(\mathbb{P} \mid \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$.

Die Funktion $\psi(s) := 1 - s + s\log s$ ist konvex, setze $f(x) := \begin{cases} 1 & \text{wenn } \sigma(x) = 0, \\ \frac{\rho(x)}{\sigma(x)} & \text{sonst.} \end{cases}$

Es gilt

$$0 \le \int \psi(f(x))\sigma(x)\, dx, \int (1 - f(x))\sigma(x)\, dx = \int \sigma(x) - \rho(x)\, dx = 0,$$

folglich

$$\int \psi(f(x))\sigma(x)\, dx = \int (1 - f(x))\sigma(x)\, dx + \int \frac{\rho(x)}{\sigma(x)} \log \frac{\rho(x)}{\sigma(x)} \sigma(x)\, dx,$$

d.h. $H(\mathbb{P} \mid \mathbb{Q}) \ge 0$ und $H(\mathbb{Q} \mid \mathbb{Q}) = 0$.

Umgekehrt: Sei $H(\mathbb{P} \mid \mathbb{Q}) = 0 = \int \psi(f(x))\sigma(x)dx \iff \mathbb{Q}(\underbrace{\{x : \psi(f(x)) = 0\}}_{=\{x : f(x) = 1 = \frac{\rho(x)}{\sigma(x)}\}}) = 1.$

Beweis von Satz 22. Sei

$$R_n := \frac{\rho_1^{\otimes n}(X_1, \dots, X_n)}{\rho_0^{\otimes n}(X_1, \dots, X_n)} = \prod_{i=1}^n \frac{\rho(1, X_i)}{\rho(0, X_i)},$$

$$H_n := -\frac{1}{n}\log R_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

mit $h(x) = \log\left(\frac{\rho(0,x)}{\rho(1,x)}\right)$. Beachte

$$\mathbb{E}_0[h(X_1)] = H(\mathbb{P}_0 \mid \mathbb{P}_1).$$

Dann ist (nach Konstruktion)

$$\varphi_n = \begin{cases} 1 & \text{wenn } H_n < a_n \\ 0 & \text{wenn } H_n > a_n \end{cases}$$

für ein a_n. Sei $a < \mathbb{E}_0[h]$,

$$\mathbb{P}_0^{\otimes\mathbb{N}}\big(H_n \le a\big) \underset{n\to\infty}{\longrightarrow} 0 \text{ mit dem Gesetz großer Zahlen,}$$

also $a_n > a$ für n genügend groß,

$$\{1-\varphi_n > 0\} \subseteq \{H_n \ge a_n\} = \{\rho_0^{\otimes n} \ge e^{na_n}\rho_1^{\otimes n}\}$$

Somit

$$1 \ge \mathbb{E}_0[1-\varphi_n] \ge e^{na_n}\mathbb{E}_1[1-\varphi_n],$$

d.h.

$$\limsup_{n\to\infty} \frac{1}{n}\log(1-\mathbb{E}_1[\varphi_n]) \le -a,$$

also auch

$$\limsup_{n\to\infty} \frac{1}{n}\log(1-\mathbb{E}_1[\varphi_n]) \le -\mathbb{E}_0[h].$$

Sei $a > \mathbb{E}_0[h]$ (und ohne Einschränkung $\mathbb{E}_0[h] < \infty$), dann ist

$$\mathbb{P}_0^{\otimes\mathbb{N}}(\rho_1^{\otimes n} \ge e^{-na}\rho_0^{\otimes n}) = \mathbb{P}_0^{\otimes\mathbb{N}}(H_n \le a) \ge \frac{1+\alpha}{2}$$

für n genügend groß.

$$\begin{aligned}\mathbb{E}_1[1-\varphi_n] &= \mathbb{E}_0\Big[(1-\varphi_n)\frac{\rho_1^{\otimes n}}{\rho_0^{\otimes n}}\Big] \ge e^{-na}\mathbb{E}_0[(1-\varphi_n)\mathbf{1}_{H_n\le a}] \\ &\ge e^{-na}\big(\mathbb{P}_0^{\otimes\mathbb{N}}(H_n \le a) - \underbrace{\mathbb{E}_0[\varphi_n]}_{=\alpha}\big) \ge e^{-na}\tfrac{1-\alpha}{2}\end{aligned}$$

Somit

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{E}_1[1-\varphi_n] \ge -a,$$

mit $a \searrow \mathbb{E}_0[h]$ folgt die Behauptung. □

1.3.2 Zum Fall monotoner Likelihood-Quotienten

Definition 23 (Likelihood-Quotient, etc.)**.** Ein Standardmodell $(\Omega, \mathscr{F}, \mathbb{P}_\vartheta : \vartheta \in \Theta)$ mit $\Theta \subseteq \mathbb{R}$ hat wachsende Likelihood-Quotienten (bzgl. einer reellwertigen Statistik T), wenn für alle $\vartheta < \vartheta'$ der Likelihood-Quotient

$$R_{\vartheta',\vartheta}(x) := \frac{\rho(\vartheta', x)}{\rho(\vartheta, x)}$$

eine wachsende Funktion von T ist, d.h. $R_{\vartheta',\vartheta}(x) = f_{\vartheta',\vartheta}(T(x)) = (f_{\vartheta',\vartheta} \circ T)(x)$ für eine wachsende Funktion $f_{\vartheta',\vartheta}$.

Beobachtung. Ein einparametriges exponentielles Modell (mit $\vartheta \mapsto a(\vartheta)$ strikt wachsend oder strikt fallend) hat wachsende Likelihood-Quotienten:

$$\rho(\vartheta, x) = h(x)e^{a(\vartheta)T(x) - b(\vartheta)}$$

demnach ist für $\vartheta < \vartheta'$

$$R_{\vartheta',\vartheta}(x) = \frac{\rho(\vartheta', x)}{\rho(\vartheta, x)} = \exp\Big(\big(a(\vartheta') - a(\vartheta)\big)T(x) - b(\vartheta') + b(\vartheta)\Big).$$

Klar, falls a streng monoton wachsend, sonst gehe zu $-T$ über.

Satz 24. *Sei $(\Omega, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, $\Theta \subseteq \mathbb{R}$ ein Standardmodell mit wachsenden Likelihoodquotienten bzgl. T, $\vartheta_0 \in \Theta$, $\alpha \in (0,1)$. Ein gleichmäßig bester Test von $H_0 : \{\vartheta \le \vartheta_0\}$ gegen $H_1 : \{\vartheta > \vartheta_0\}$ zum Niveau α hat die Form*

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

für gewisse $c \in \mathbb{R}, \gamma \in [0,1]$, die durch die Bedingung $G_\varphi(\vartheta_0) = \alpha$ festgelegt sind. Die Gütefunktion ist monoton wachsend in ϑ.

Bemerkung. Um $H_0 : \{\vartheta \ge \vartheta_0\}$ gegen $H_1 : \{\vartheta < \vartheta_0\}$ zu testen, vertausche „<“ und „>“ der Definition von φ (formal: multipliziere ϑ und T jeweils mit -1).

Beweis von Satz 24. Setze $\Theta_0 := \Theta \cap (-\infty, \vartheta_0], \Theta_1 := \Theta \cap (\vartheta_0, \infty)$. Sei ψ irgendein Test von H_0 gegen H_1 der das Niveau α einhält, d.h.

$$\sup_{\vartheta\in\Theta_0} G_\psi(\vartheta) \le \alpha$$

Fixiere $\vartheta' \in \Theta_1$ und φ und ψ auf als Tests von $\{\vartheta_0\}$ gegen $\{\vartheta'\}$; in diesem Szenario ist dann φ ein Neyman-Pearson-Test mit $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$.

Demnach (gemäß dem Neyman-Pearson-Lemma, Satz 21) ist

$$G_\varphi(\vartheta') \ge G_\psi(\vartheta') \quad \forall \vartheta' \in \Theta_1.$$

Zur Monotonie: Sei $\vartheta < \vartheta'$, fasse φ auf als einen (Neyman-Pearson-)Test von $\{\vartheta\}$ gegen $\{\vartheta'\}$ mit Niveau $\beta := G_\varphi(\vartheta)$.

Vergleiche mit dem (trivialen) Test $\psi_\beta(x) \equiv \beta$, der ebenfalls Niveau β hat.
Gemäß dem Neyman-Pearson-Lemma (Satz 21) ist

$$G_\varphi(\vartheta') \geq G_{\psi_\beta}(\vartheta') = \beta = G_\varphi(\vartheta)$$

insbesondere ist $\vartheta \mapsto G_\varphi(\vartheta)$ monton wachsend, also

$$\sup_{\vartheta \in \Theta_0} G_\varphi(\vartheta) \overset{(<)}{=} \alpha.$$

□

Beispiel. 1. Unter $\mathbb{P}_\vartheta$ seien $X_1, \ldots, X_n$ u.i.v $\sim \mathcal{N}(\vartheta, \sigma^2)$, $\sigma^2 > 0$ fest, $\vartheta \in \mathbb{R} := \Theta$. Teste $\{\vartheta \leq \vartheta_0\}$ gegen $\{\vartheta > \vartheta_0\}$ folgendermaßen:

Dies ist ein exponentielles Modell bzgl. $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Der beste Test zum Niveau $\alpha \in (0,1)$ hat Ablehnungsbereich

$$\left\{\overline{X} > \vartheta_0 + \sqrt{\frac{v}{n}}\Phi^{-1}(1-\alpha)\right\}$$

(d.h. ist der einseitige Gauß-Test).

2. Unter $\mathbb{P}_\vartheta$ seien $X_1, \ldots, X_n$ u.i.v $\sim \mathcal{N}(\mu, \vartheta)$ mit $\mu \in \mathbb{R}$ fest, $\vartheta \in (0, \infty) =: \Theta$. Teste $\{\vartheta \geq \vartheta_0\}$ gegen $\{\vartheta < \vartheta_0\}$:

   Dies ist eine exponentielle Familie bzgl. $T := \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$, der beste Test zum Niveau $\alpha \in (0,1)$ hat dann den Ablehnungsbereich

$$\left\{T < \frac{1}{n}\vartheta_0 \cdot \chi^2_{n,\alpha}\right\}$$

   wobei $\chi^2_{n,\alpha}$ das α-Quantil von χ^2_n ist.

3. Poissonmodell (ist eine einparametrige exponentielle Familie, vgl. die Beispiele nach Def. 6)

4. Binomialmodell (ist ebenfalls eine einparametrige exponentielle Familie, vgl. die Beispiele nach Def. 6)

Bemerkung (Monotone Likelihood-Quotienten und stochastische Ordnung)**.** $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$ mit zugehörigen Verteilungsfunktionen $F_\mu(x) = \mu((-\infty, x])$, $F_\nu(x) = \nu((-\infty, x])$. μ heißt *stochastisch kleiner* als ν, geschrieben $\mu \preceq \nu$, falls gilt

$$\forall\, x \in \mathbb{R}:\ F_\mu(x) \geq F_\nu(x) \tag{1.15}$$

Zu (1.15) äquivalent sind

$$\int_{\mathbb{R}} f(x)\,\mu(dx) \leq \int_{\mathbb{R}} f(x)\,\nu(dx) \quad \text{für } \vartheta < \vartheta' \text{ und } f : \mathbb{R} \to \mathbb{R} \text{ monoton wachsend und beschränkt} \tag{1.16}$$

und

Es gibt eine *Kopplung* von μ und ν, d.h. ein Paar (X,Y) von Zufallsvariablen mit $X =^d \mu, Y =^d \nu$, so dass gilt

$$X \le Y \text{ fast sicher} \tag{1.17}$$

Für (1.16) $\Rightarrow$ (1.15) betrachte $f(y) = \mathbf{1}_{(x,\infty)}$ und beachte $\int f\,d\mu = 1 - F_\mu(x)$; für (1.15) $\Rightarrow$ (1.16) approximiere eine monotone Funktion f mittels Funktionen des Typs $\sum_{i=1}^m a_i \mathbf{1}_{(x_i,\infty)}$ mit $a_i > 0$, $x_1 < \cdots < x_m$.

Offenbar folgt (1.15) aus (1.17): Es ist $\mathbf{1}_{(-\infty,x]}(X) \ge \mathbf{1}_{(-\infty,x]}(Y)$ fast sicher und daher $F_\mu(x) = \mathbb{E}\big[\mathbf{1}_{(-\infty,x]}(X)\big] \ge \mathbb{E}\big[\mathbf{1}_{(-\infty,x]}(Y)\big] = F_\nu(x)$. Sei umgekehrt (1.15) erfüllt, dann gilt für die (verallgemeinerten) Inversen $F_\mu^{-1}(u) := \inf\{x : F_\mu(x) \ge u\} \le F_\nu^{-1}(u)$, $u \in (0,1)$ und mit $U \sim \mathrm{Unif}([0,1])$ erfüllen $X := F_\mu^{-1}(U), Y := F_\nu^{-1}(U)$ die Forderung (1.17).

(Man kann dies etwas allgemeiner für Maße auf einer partiell geordneten Menge betrachten, vgl. z.B. [Kle06, Bsp. 18.7].)

Falls die Familie $\mathbb{P}_\vartheta$ bezüglich der Statistik T monotone Likelihood-Quotienten hat, so gilt

$$\mathscr{L}_\vartheta(T) \le \mathscr{L}_{\vartheta'}(T) \quad \text{für } \vartheta < \vartheta'$$

Sei dazu $f : \mathbb{R} \to \mathbb{R}$ monoton wachsend (und beschränkt), $\vartheta < \vartheta'$:

$$\begin{aligned}
&\mathbb{E}_{\vartheta'}[f(T)] - \mathbb{E}_\vartheta[f(T)] = \int_\Omega \int_\Omega \big(f(T(x)) - f(T(y))\big)\rho(\vartheta',x)\rho(\vartheta,y)\,dx\,dy \\
&= \int_\Omega \int_\Omega \Big(\mathbf{1}_{\{T(x)>T(y)\}}\big(f(T(x)) - f(T(y))\big)\rho(\vartheta',x)\rho(\vartheta,y) \\
&\qquad + \mathbf{1}_{\{T(x)<T(y)\}}\big(f(T(x)) - f(T(y))\big)\rho(\vartheta',x)\rho(\vartheta,y)\Big)\,dx\,dy \\
&= \int_\Omega \int_\Omega \mathbf{1}_{\{T(x)>T(y)\}}\big(f(T(x)) - f(T(y))\big)\big(\rho(\vartheta',x)\rho(\vartheta,y) - \rho(\vartheta',y)\rho(\vartheta,x)\big)\,dx\,dy \ge 0
\end{aligned}$$

denn auf $\{T(x) > T(y)\}$ ist $f(T(x)) - f(T(y)) \ge 0$ und $\frac{\rho(\vartheta',x)}{\rho(\vartheta,x)} \ge \frac{\rho(\vartheta',y)}{\rho(\vartheta,y)}$ und also auch $\rho(\vartheta',x)\rho(\vartheta,y) - \rho(\vartheta',y)\rho(\vartheta,x) \ge 0$

(Nochmal zu) Tests im (2-parametrigen) Gauß'schen Modell

Szenario: Unter $\mathbb{P}_\vartheta$, $\vartheta = (\mu,\sigma^2) \in \mathbb{R}\times(0,\infty)$ seien $X_1,\dots,X_n$ u.i.v. $\sim \mathcal{N}(\mu,\sigma^2)$; wir notieren (wieder)

$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}\big)^2 \tag{1.18}$$

Zweiseitiger t-Test für den Mittelwert Wir betrachten die Nullhypothese $\Theta_0 = \mu_0 \times (0,\infty)$ und die Alternative $\Theta_1 = (\mathbb{R}\setminus\{\mu_0\})\times(0,\infty)$ mit einem gegebenen μ_0 im Gauß'schen Produktmodell. Die t-Statistik lautet

$$T = \sqrt{n}\frac{\overline{X} - \mu_0}{\sqrt{S^2}} \tag{1.19}$$

Sei $\alpha \in (0,1)$. Der zweiseitige t-Test

$$\varphi = \mathbf{1}_{\{|T| > q_{\text{Student}(n-1),1-\alpha/2}\}} \tag{1.20}$$

hat unter allen unverfälschten Tests von Θ_0 gegen Θ_1 (punktweise) die größte Macht (er ist also der *beste unverfälschte Test* in dieser Situation).

Beweis. Betrachte o.E. $\mu_0 = 0$. Es ist

$$\begin{aligned}\rho\big((\mu,\sigma^2),x\big) &= (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\Big)\\ &= (2\pi\sigma^2)^{-n/2}e^{-n\mu/(2\sigma^2)}\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sqrt{n}\frac{\sum_{i=1}^n x_i}{\sqrt{n}}\Big)\end{aligned}$$

Re-parametrisiere $\alpha = \mu\sqrt{n}/\sigma^2$, $\beta = 1/(2\sigma^2)$, damit ist

$$\widetilde{\rho}\big((\alpha,\beta),x\big) = (\pi/\beta)^{-n/2}e^{-\sqrt{n}\alpha/2}\exp\big(-\beta\widetilde{s}+\alpha\widetilde{x}\big) = c(\alpha,\beta)\exp\big(-\beta\widetilde{s}+\alpha\widetilde{x}\big)$$

mit $\widetilde{x} := n^{-1/2}\sum_{i=1}^n x_i$, $\widetilde{s} := \sum_{i=1}^n x_i^2$, $c(\alpha,\beta) := (\pi/\beta)^{-n/2}e^{-\sqrt{n}\alpha/2}$.

Mit $\overline{X}$, S^2 aus (1.18) und $\widetilde{X} := \sum_{i=1}^n X_i/\sqrt{n} = \sqrt{n}\overline{X}$, $\widetilde{S} := \sum_{i=1}^n X_i^2$ ist

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}(\overline{X})^2 = \frac{1}{n-1}\big(\widetilde{S}-\widetilde{X}\big)$$

ist (für $\mu_0 = 0$)

$$T = \frac{\widetilde{X}}{\sqrt{(\widetilde{S}-\widetilde{X})/(n-1)}} = \sqrt{n-1}\frac{\widetilde{X}}{\sqrt{\widetilde{S}-\widetilde{X}}}$$

und der Ablehnungsbereich also

$$\Big|\frac{\widetilde{X}}{\sqrt{\widetilde{S}-\widetilde{X}}}\Big| > r := \frac{1}{\sqrt{n-1}}q_{\text{Student}(n-1),1-\alpha/2}$$

Der zweiseitige t-Test lehnt die Nullhypothese also genau dann ab, wenn

$$\begin{aligned}&\widetilde{X}^2 > r^2\big(\widetilde{S}-\widetilde{X}\big) \quad \Longleftrightarrow \quad \widetilde{X}^2 > \frac{r^2\widetilde{S}}{1+r^2}\\ &\Longleftrightarrow \quad |\widetilde{X}| > r\sqrt{\widetilde{S}/(1+r^2)} := f(\widetilde{S})\end{aligned}$$

Sei ψ ein unverfälschter Test von Θ_0 gegen Θ_1 zum Niveau α. Dann gilt

$$\mathbb{E}_{(0,\sigma^2)}[\psi] = \alpha \quad \forall\sigma^2 > 0$$

(denn $\mathbb{R} \ni \mu \mapsto \mathbb{E}_{(\mu,\sigma^2)}[\psi]$ ist stetig mit $\mathbb{E}_{(0,\sigma^2)}[\psi] \le \alpha$ und $\mathbb{E}_{(\mu,\sigma^2)}[\psi] \ge \alpha$ für $\mu \neq 0$) und

$$\mathbb{R} \ni \mu \mapsto \mathbb{E}_{(\mu,\sigma^2)}[\psi] \quad \text{hat in } \mu = 0 \text{ ein globales Minimum} \tag{1.21}$$

Da auch der t-Test φ das Niveau α (exakt) einhält, d.h. $\mathbb{E}_{(0,\sigma^2)}[\varphi] = \alpha$, gilt insbesondere

$$\mathbb{E}_{(0,\sigma^2)}[\psi-\varphi] = 0 \quad \forall\sigma^2 > 0$$

bzw. in der Parametrisierung mit (α, β):

$$\widetilde{\mathbb{E}}_{(0,\beta)}[\psi - \varphi] = 0 \quad \forall \beta > 0$$

Für $\beta = \gamma + k$ mit $\gamma > 0$, $k \in \mathbb{N}_0$ bedeutet dies insbesondere

$$\begin{aligned} 0 = \widetilde{\mathbb{E}}_{(0,\gamma+k)}[\psi - \varphi] &= \int_{R^n} \widetilde{\rho}((0,\gamma+k),x)\big(\psi(x) - \varphi(x)\big)\,dx \\ &= c(0,\gamma+k)\int_{R^n} e^{-\gamma\widetilde{s}(x) - k\widetilde{s}(x)}\big(\psi(x) - \varphi(x)\big)\,dx = \frac{c(0,\gamma+k)}{c(0,\gamma)}\widetilde{\mathbb{E}}_{(0,\gamma)}\big[(e^{-\widetilde{S}})^k\big(\psi - \varphi\big)\big] \end{aligned}$$

d.h.

$$\mathbb{E}_{(0,\sigma^2)}\big[(e^{-\widetilde{S}})^k\big(\psi - \varphi\big)\big] = 0 \quad \forall\, \sigma^2 > 0$$

für jedes Polynom g und via Weierstraß'schem Approximationssatz somit für jedes stetige $g : [0,1] \to \mathbb{R}$.

Dies impliziert

$$\mathbb{E}_{(0,\sigma^2)}\big[h(\widetilde{S})\big(\psi - \varphi\big)\big] = 0 \tag{1.22}$$

für jede subexponentiell wachsende Funktion $h : (0,\infty) \to \mathbb{R}$ (d.h. h erfüllt $\lim_{u\to\infty} e^{-\delta u}h(u) = 0$ für alle $\delta > 0$).

Sei nämlich ein solches h gegeben, wähle $\delta > 0$ und setze

$$g_\delta(x) := \begin{cases} h(\log(1/x))x^\delta, & 0 < x \le 1, \\ 0, & x = 0, \end{cases}$$

g_δ ist stetig, und $g_\delta(e^{-w}) = h(w)e^{-\delta w}$. Damit ist

$$\mathbb{E}_{(0,\sigma^2)}\big[h(\widetilde{S})e^{-\delta\widetilde{S}}\big(\psi - \varphi\big)\big] = 0$$

für jedes $\delta > 0$, mit $\delta \downarrow 0$ folgt (1.22).

Wegen (1.21) ist

$$0 = \frac{d}{d\alpha}\widetilde{E}_{(\alpha,\beta)}[\psi]\Big|_{\alpha=0} = \widetilde{E}_{(0,\beta)}\big[\widetilde{X}\,\psi\big] \quad \text{für jedes } \beta > 0$$

Analog erhalte

$$0 = \widetilde{E}_{(0,\beta)}\big[h(\widetilde{S})\widetilde{X}\,(\psi - \varphi)\big] \quad \text{für jedes } \beta > 0$$

Sei nun $\alpha \neq 0$, $\beta > 0$. Der Likelihood-Quotient

$$\widetilde{R}_{(\alpha,\beta),(0,\beta)}(x) = \frac{\widetilde{\rho}((\alpha,\beta),x)}{\widetilde{\rho}((0,\beta),x)} = \frac{c(\alpha,\beta)}{c(0,\beta)}e^{\alpha\widetilde{x}} =: \widetilde{c}(\alpha,\beta)e^{\alpha\widetilde{x}}$$

ist eine strikt konvexe Funktion von $\widetilde{x}$. Wir wählen $a(\widetilde{s}), b(\widetilde{s}) \in \mathbb{R}$, so dass die lineare Funktion $\widetilde{x} \mapsto a(\widetilde{s}) + b(\widetilde{s})\,\widetilde{x}$ im Inneren des Ablehnungsbereichs $-f(\widetilde{s}) < \widetilde{x} < f(\widetilde{s})$ oberhalb des Graphen von $\widetilde{x} \mapsto \widetilde{c}(\alpha,\beta)e^{\alpha\widetilde{x}}$ verläuft und

$$a(\widetilde{s}) + b(\widetilde{s})\,\widetilde{x} = \widetilde{c}(\alpha,\beta)e^{\alpha\widetilde{x}} \quad \text{für } \widetilde{x} = \pm f(\widetilde{s})$$

erfüllt (d.h. $a(\widetilde{s}) = \widetilde{c}(\alpha,\beta)\cosh\big(\alpha f(\widetilde{s})\big)$, $b(\widetilde{s}) = \widetilde{c}(\alpha,\beta)\sinh\big(\alpha f(\widetilde{s})\big)/f(\widetilde{s})$; beachte $|a(\widetilde{s})|, |b(\widetilde{s})| \le c_1 e^{c_2\sqrt{\widetilde{s}}}$ wachsen subexponentiell).

Es ist (nach obigem)

$$0 = \widetilde{E}_{(0,\beta)}\Big[\big(a(\widetilde{S}) + b(\widetilde{S})\widetilde{X}\big)\,(\psi - \varphi)\Big] \quad \text{für jedes } \beta > 0$$

und damit ergibt sich

$$\begin{aligned}\widetilde{E}_{(\alpha,\beta)}[\varphi - \psi] &= \widetilde{E}_{(0,\beta)}\big[R_{(\alpha,\beta),(0,\beta)}(X)(\varphi(X) - \psi(X))\big] \\ &= \widetilde{E}_{(0,\beta)}\Big[\big(R_{(\alpha,\beta),(0,\beta)}(X) - a(\widetilde{S}) - b(\widetilde{S})\widetilde{X}\big)(\varphi(X) - \psi(X))\Big] \ge 0\end{aligned}$$

(denn die beiden Terme in dem Produkt innerhalb des Erwartungswerts in der zweiten Zeile haben stets dasselbe Vorzeichen). Folglich ist die Macht von φ nicht kleiner als die von ψ.

□

Bemerkung. 1. Prinzipiell kann man die Gütefunktion $\mathbb{E}_{(\mu,\sigma^2)}[\varphi]$ des t-Tests explizit berechnen: Unter $\mathbb{P}_{(\mu,\sigma^2)}$ hat T eine nicht-zentrale Student-verteilung (und beispielsweise `R`s `[d|p|q|r]t`-Befehle haben den Parameter `ncp`).

2. Die Zusatzforderung der Unverfälschtheit kann nicht weggelassen werden. Es gibt Tests ψ, die Niveau α ($< 1/2$) einhalten, d.h. $\sup_{\vartheta\in\Theta_0}\mathbb{E}_\vartheta[\psi(X)] \le \alpha$, für die für gewisse $\vartheta' \in \Theta_1$ gilt $\mathbb{E}_{\vartheta'}[\psi(X)] > \mathbb{E}_{\vartheta'}[\varphi(X)]$ (allerdings gilt für viele $\vartheta' \in \Theta_1$ auch $\mathbb{E}_{\vartheta'}[\psi(X)] < \alpha < \mathbb{E}_{\vartheta'}[\varphi(X)]$). Siehe Anhang A.5 für Beispiele.

1.4 Zur Bayes-Statistik

Wir folgten bisher dem klassischen, sogenannten frequentistischen Ansatz der Statistik:

Wir fassen eine Menge Θ von „Parametern“ ins Auge, für $\Theta \ni \vartheta$ beschreibt (in einem statistischen Modell) $\mathbb{P}_\vartheta$ die Verteilung der Beobachungen, wenn dieses ϑ der tatsächlich zutreffende (sozusagen der „wahre“) Parameter ist. In der konkreten Anwendungssituation kennen wir dieses „wahre“ ϑ natürlich i.A. nicht, wir fassen es als zwar unbekannte, aber prinzipiell feste Größe auf. Wahrscheinlichkeitsaussagen beziehen sich *nicht* auf ϑ, sondern auf zufällige Beobachtungen unter $\mathbb{P}_\vartheta$.

Dies ist anders in der *Bayes-Statistik*: Man wählt eine Wahrscheinlichkeitsverteilung α auf Θ, die *a priori*-Verteilung (auch *Vorbewertung*) und stellt sich vor, dass die Daten einem zweistufigen Experiment entstammen:

- Zunächst wird der Parameter ϑ gemäß der a priori-Verteilung α erzeugt (insbesondere ist ϑ jetzt selbst eine Zufallsvariable),
- dann werden die Beobachtungen X zufällig erzeugt mit einer Verteilung, die vom gewählten ϑ abhängt.

Insbesondere besitzt in dieser Formulierung das Paar (X,ϑ) eine *gemeinsame Verteilung*.

Es gelte: Die a priori-Verteilung auf Θ hat Dichte bzw. Gewichte $\alpha(\vartheta)$

(je nachdem, ob ϑ kontinuierlich oder diskret verteilt ist; wir betrachten im Folgenden nur den Fall, dass $\Theta \subset \mathbb{R}$ ein Intervall ist und ϑ eine Dichte besitzt)

Interpretation: Ohne Kenntnis der Beobachtungen nehmen wir an, dass ϑ die a priori-Verteilung besitzt (z.B. aus „Erfahrung“ oder aus „Expertenwissen“).

Wir interpretieren die Likelihood-Funktion $\rho : S \times \Theta \to [0, \infty)$ als

$$\rho(p, x) = \mathbb{P}(X = x \mid \vartheta = p)$$

(bzw. $\mathbb{P}(X \in dx \mid \vartheta = p) = \rho(p, x)\, dx$ falls X eine Dichte besitzt)

Mit Formel von der totalen Wahrscheinlichkeit

$$\mathbb{P}(X = x) = \int_\Theta \rho(t, x)\, \alpha(t) dt$$

(bzw. $\mathbb{P}(X \in dx) = \int_\Theta \rho(t, x)\, \alpha(t) dt\, dx$, d.h. $\mathbb{P}(X \le x) = \int_\Theta \int_{-\infty}^{x} \rho(t, y)\, dy\, \alpha(t) dt$, wenn X eine Dichte besitzt), und mit der Formel von Bayes ist

$$\mathbb{P}(\vartheta \in dp \mid X = x) = \frac{\rho(p, x)\alpha(p)\, dp}{\int_\Theta \alpha(\vartheta')\rho(\vartheta', x)\, d\vartheta'}$$

Die *a posteriori-Dichte* (bzw. die *a posteriori-Gewichte*, wenn ϑ diskret ist) bei Beobachtung x,

$$\pi_x(\vartheta) = \frac{\alpha(\vartheta)\rho(\vartheta, x)}{\int_\Theta \alpha(\vartheta')\rho(\vartheta', x)\, d\vartheta'},$$

ist die Dichte von ϑ, bedingt auf Beobachtung $X = x$, d.h.

$$\mathbb{P}(\vartheta \le u \mid X = x) = \int_{\Theta \cap (-\infty, u]} \pi_x(p)\, dp$$

Der *Bayes-Schätzer* (zur a priori-Verteilung α) ist

$$\widehat{\vartheta}_{\mathrm{B}} = \widehat{\vartheta}_{\mathrm{B}}(x) := \mathbb{E}_{\pi_x}[\vartheta] = \int_\Theta \vartheta\, \pi_x(\vartheta)\, d\vartheta$$

d.h. der Erwartungswert von ϑ bedingt auf $X = x$ (Wir betrachten hier nur den Fall, dass Θ ein Intervall ist).

Definition. Für einen Schätzer $Y = Y(X)$ (für ϑ) ist

$$F_\alpha(Y) := \int_\Theta \mathbb{E}\big[(Y - \vartheta)^2 \,\big|\, \vartheta = p\big]\alpha(p)\, dp$$

der erwartete quadratische Fehler (zur Vorbewertung α).

Der Bayes-Schätzer minimiert den erwarteten quadratischen Fehler (zur Vorbewertung α) :

Beobachtung. Für jeden Schätzer Y gilt

$$F_\alpha(Y) \geq F_\alpha(\widehat{\vartheta}_{\mathrm{B}}(X))$$

Beweis.

$$\begin{aligned}
&F_\alpha(Y(X)) - F_\alpha(\widehat{\vartheta}_{\mathrm{B}}(X)) \\
&= \int_\Theta \mathbb{E}\big[(Y(X)-\vartheta)^2 - (\widehat{\vartheta}_{\mathrm{B}}(X)-\vartheta)^2 \,\big|\, \vartheta = p\big]\alpha(p)\,dp \\
&= \int_\Theta \mathbb{E}\big[Y(X)^2 - 2Y(X)\vartheta - \widehat{\vartheta}_{\mathrm{B}}(X)^2 + 2\vartheta\widehat{\vartheta}_{\mathrm{B}}(X) \,\big|\, \vartheta = p\big]\alpha(p)\,dp \\
&= \mathbb{E}\big[Y(X)^2 - 2Y(X)\vartheta - \widehat{\vartheta}_{\mathrm{B}}(X)^2 + 2\vartheta\widehat{\vartheta}_{\mathrm{B}}(X)\big] \\
&= \mathbb{E}\big[Y(X)^2 - \widehat{\vartheta}_{\mathrm{B}}(X)^2\big] - 2\mathbb{E}\big[Y(X)\vartheta\big] + 2\mathbb{E}\big[\vartheta\widehat{\vartheta}_{\mathrm{B}}(X)\big]
\end{aligned}$$

Weiter ist

$$\begin{aligned}
&\mathbb{E}\big[\vartheta\widehat{\vartheta}_{\mathrm{B}}(X)\big] \\
&= \sum_{x\in S} \mathbb{E}\big[\vartheta\widehat{\vartheta}_{\mathrm{B}}(X) I_{\{X=x\}}\big] = \sum_{x\in S} \mathbb{P}(X=x)\widehat{\vartheta}_{\mathrm{B}}(x)\mathbb{E}[\vartheta \mid X=x] \\
&= \sum_{x\in S} \mathbb{P}(X=x)\widehat{\vartheta}_{\mathrm{B}}(x)\mathbb{E}_{\pi_x}[\vartheta] = \sum_{x\in S} \mathbb{P}(X=x)\big(\widehat{\vartheta}_{\mathrm{B}}(x)\big)^2 \\
&= \mathbb{E}\big[\big(\widehat{\vartheta}_{\mathrm{B}}(X)\big)^2\big]
\end{aligned}$$

und analog

$$\mathbb{E}\big[\vartheta Y(X)\big] = \mathbb{E}\big[Y(X)\widehat{\vartheta}_{\mathrm{B}}(X)\big]$$

Insgesamt:

$$\begin{aligned}
&F_\alpha(Y(X)) - F_\alpha(\widehat{\vartheta}_{\mathrm{B}}(X)) \\
&= \mathbb{E}\big[Y(X)^2 - \widehat{\vartheta}_{\mathrm{B}}(X)^2 - 2Y(X)\widehat{\vartheta}_{\mathrm{B}}(X) + 2\widehat{\vartheta}_{\mathrm{B}}(X)^2\big] \\
&= \mathbb{E}\big[\big(Y(X) - \widehat{\vartheta}_{\mathrm{B}}(X)\big)^2\big] \geq 0
\end{aligned}$$

□

Beispiel (Münzwurf mit zufälliger Erfolgswahrscheinlichkeit)**.** ϑ wird gemäß einer Verteilung α auf $\Theta := [0,1]$ „ausgewürfelt“, dann:

$n \in \mathbb{N}$, gegeben $\vartheta = u \in [0,1]$ seinen $X_1, X_2, \ldots, X_n$ unabhängig und jeweils $\sim \mathrm{Ber}_u$ (d.h. $\mathbb{P}(X_i = 1 \mid \vartheta = u) = u = 1 - \mathbb{P}(X_i = 0 \mid \vartheta = u)$).

Somit: Für $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ ist

$$\rho(\vartheta, x) = \vartheta^{\#\{i\leq n : x_i = 1\}}(1-\vartheta)^{\#\{i\leq n : x_i = 0\}}.$$

Eine Situation, in der dieses Modell sinnvoll ist, könnte folgende sein: Nehmen wir an, ein Versicherungsnehmer hat jedes Jahr mit einer gewissen (zu ihm „gehörigen“) Wahrscheinlichkeit ϑ einen Schadensfall (unabhängig über die Jahre), und $\alpha(\vartheta)\,d\vartheta$ beschreibt die Verteilung der Schadenswahrscheinlichkeiten aller Kunden dieser Versicherung (diese Verteilung sei der Versicherung aus Erfahrungswerten bekannt).

Mit $\rho(\vartheta, x) = \mathrm{Bin}_{n,\vartheta}(x)$ ist dann die Wahrscheinlichkeit, dass ein „typischer Kunde“ in n Jahren k Schadensfälle verursacht $\int_0^1 \alpha(\vartheta)\rho(k,\vartheta)\,d\vartheta$, und $\pi_k(\vartheta)$ ist die Verteilung der Schadenswahrscheinlichkeit pro Jahr eines Kunden, bedingt darauf, dass er in den letzten n Jahren k Schäden hatte. Diese Information kann die Versicherung beispielsweise für Vertragsverlängerung, Tarifanpassung, etc. benutzen.

Wir betrachten hier (nur) den Fall $\alpha = \mathrm{unif}_{[0,1]}$: Gehe über zu $Y = X_1 + \cdots + X_n$, dann ist gegeben $\vartheta = u$, $Y \sim \mathrm{Bin}_{n,u}$ und für $k \in \{0, 1, \dots, n\}$

$$\begin{aligned}\mathbb{P}(Y = k) &= \int_0^1 \binom{n}{k} u^k (1-u)^{n-k}\,du \\ &= \frac{n!}{k!(n-k)!}\frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}\end{aligned}$$

(d.h. Y ist uniform auf $\{0, 1, \dots, n\}$).

Für das Weitere benötigen wir die Beta-Funktionen: $a, b \in (0, \infty)$ ist die Beta-Funktion gegeben durch

$$B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

wobei $\Gamma(a) = \int_0^\infty y^{a-1}e^{-y}\,dy$ die Gamma-Funktion ist (beachte: $\Gamma(a+1) = a\Gamma(a)$, speziell für $a \in \mathbb{N}$ ist $\Gamma(a) = (a-1)!$, wie man mit partieller Integration nachrechnen kann).

Für $a, b \in \mathbb{N}$ kann man explizit rechnen: Es ist dann $B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$, denn $B(a,1) = \int_0^1 u^{a-1}\,du = \left[\frac{1}{a}u^a\right]_{u=0}^{u=1} = \frac{1}{a}$ und für $b \in \{2, 3, \dots\}$ ist (mit partieller Integration)

$$\begin{aligned}\int_0^1 u^{a-1}(1-u)^{b-1}\,du &= \left[\tfrac{1}{a}u^a(1-u)^{b-1}\right]_{u=0}^{u=1} - \int_0^1 \tfrac{1}{a}u^a \cdot (b-1)(1-u)^{b-2}(-1)\,du \\ &= \frac{b-1}{a}\int_0^1 u^a(1-u)^{b-2}\,du\end{aligned}$$

also

$$\begin{aligned}B(a,b) &= \frac{b-1}{a}B(a+1,b-1) = \frac{(b-1)\cdot(b-2)\cdots 2\cdot 1\cdot B(a+b-1,1)}{a\cdot(a+1)\cdots(a+b-3)\cdots(a+b-2)} \\ &= \frac{(b-1)\cdot(b-2)\cdots 2\cdot 1}{a\cdot(a+1)\cdots(a+b-2)\cdots(a+b-1)} = \frac{(a-1)!(b-1)!}{(a+b-1)!}\end{aligned}$$

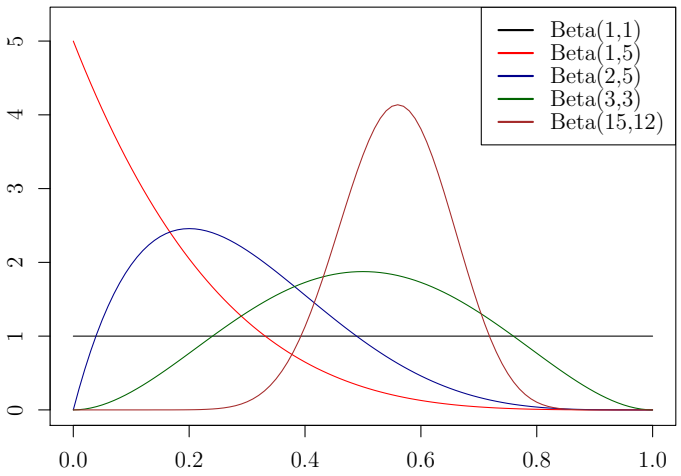
Erinnerung (Beta-Verteilungen)**.** $r, s > 0$. Eine ZV V mit Werten in $[0,1]$ ist *Beta-verteilt* mit Parametern $r, s > 0$, in Formeln $V \sim \beta_{r,s}$, wenn V die Dichte

$$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}v^{r-1}(1-v)^{s-1}\mathbf{1}_{(0,1)}(v)$$

besitzt. Es gilt dann

$$\begin{aligned}\mathbb{E}[V] &= \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^1 v\cdot v^{r-1}(1-v)^{s-1}\,dv \\ &= \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\frac{\Gamma(r+1)\Gamma(s)}{\Gamma(r+s+1)} = \frac{\Gamma(r+s)}{\Gamma(r+s+1)}\frac{\Gamma(r+1)}{\Gamma(r)} = \frac{r}{r+s}\end{aligned}$$

Einige Beta-Dichten



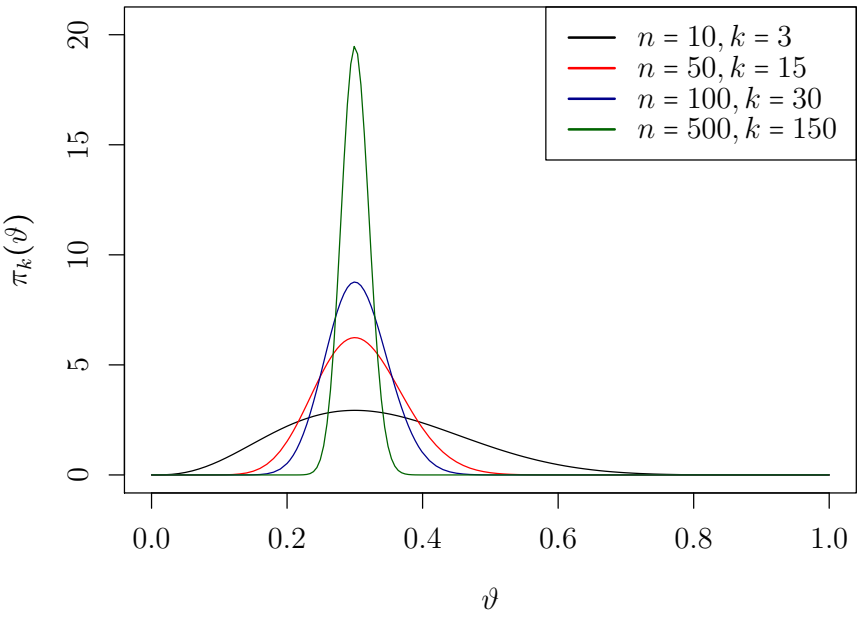
Die a posteriori-Verteilung ist $\mathcal{L}(\vartheta \mid Y = k) = \beta_{k+1,n-k+1}$:

$$\mathbb{P}(\vartheta \in dp \mid Y = k) = \frac{\mathbb{P}(\vartheta \in dp, Y = k)}{\mathbb{P}(Y = k)} = \frac{1}{1/(n+1)} \binom{n}{k} p^k (1-p)^{n-k}\, dp = \frac{(n+1)!}{k!(n-k)!} p^k (1-p)^{n-k}\, dp$$

Demnach (mit obigem zur Beta-Verteilung) ist

$$\widehat{\vartheta}_{\mathrm{B}} = \widehat{\vartheta}_{\mathrm{B}}(Y) = \frac{Y+1}{n+2}$$

A posteriori-Dichte $\pi_k(\vartheta)$
für verschiedene n und k mit $k/n = 0.3$



Wir sehen, dass für großes n die a posteriori-Verteilung recht eng um $\frac{Y+1}{n+2} \approx \frac{Y}{n}$ konzentriert ist. Zudem: die frequentistische und die Bayes'sche „Antwort" stimmen für große n „nahezu" überein.

Beispiel. Laplace[5] antwortete auf die von ihm (vielleicht mit einem Augenzwinkern) gestellte Frage:

„Angenommen die Sonne ist bis heute n-mal aufgegangen. Mit welcher Wahrscheinlichkeit geht sie morgen auf?“

$$\frac{n+1}{n+2}$$

Dies passt zur Antwort des Bayes-Schätzers in obigem Beispiel.

Konjugierte Verteilungen / Vorbewertungen

In dem (für explizite Rechnungen besonders angenehmen) Fall, dass $\alpha(\vartheta)d\vartheta$ und $\pi_x(\vartheta)d\vartheta$ (für jede mögliche Beobachtung x) zu derselben parametrisierten Familie von Verteilungen gehören, nennt man die Verteilungen *konjugiert*, $\alpha(\vartheta)d\vartheta$ ist dann eine konjugierte Vorbewertung / ein konjugierter Prior.

Es gibt viele „klassische“ Beispiele dazu, für einen kleinen Eindruck siehe die folgende Tabelle:

$\rho(\vartheta, x)$	$\alpha(\vartheta)$	$\pi_x(\vartheta)$
$\mathrm{Bin}(n, \vartheta)$	$\mathrm{Beta}(u, v)$	$\mathrm{Beta}(u+x, v+n-x)$
$\mathrm{Poi}(\vartheta)$	$\mathrm{Gamma}(\alpha, \nu)$	$\mathrm{Gamma}(\alpha+1, \nu+x)$
$\mathcal{N}(\vartheta, 1)$	$\mathcal{N}(\mu_0, \sigma_0^2)$	$\mathcal{N}\Big(\frac{1}{\frac{1}{1}+\frac{1}{\sigma_0^2}}\big(\frac{\mu_0}{\sigma_0^2}+\frac{x}{1}\big), \big(\frac{1}{\sigma_0^2}+\frac{1}{1}\big)^{-1}\Big)$

Ein „allgemeines“ Bauprinzip hierfür stellen exponentielle Familien dar: Sei

$$t(x) = \big(t_1(x), \dots, t_d(x)\big)$$

eine d-dimensionale Statistik und $\vartheta = (\vartheta_1, \dots, \vartheta_d)$ ein d-dimensionaler Parameter, wir betrachten eine exponentielle Familie bzgl. der Statistik $T = t(X)$,

$$\rho(\vartheta, x) = h(x) \exp\big(\vartheta \cdot t(x)^{\mathrm{T}} - b(\vartheta)\big) = h(x) \exp\Big(\sum_{i=1}^{d} \vartheta_i t_i(x) - b(\vartheta)\Big)$$

mit dem Ansatz

$$\alpha(\vartheta) = \alpha_{\eta,\nu}(\vartheta) = Z(\eta, \nu) \exp\big(\eta \cdot \vartheta^{\mathrm{T}} - \nu b(\vartheta)\big)$$

(wobei $= \eta = (\eta_1, \dots, \eta_d)$, $\nu > 0$ „Hyperparameter“ sind und $Z(\eta, \nu)$ eine geeignete Normierungskonstante, so dass $\int_\Theta \alpha(\vartheta)\, d\vartheta = 1$) findet man

$$\begin{aligned}
\pi_x(\vartheta) &= \frac{\alpha(\vartheta)\rho(\vartheta, x)}{\int_\Theta \alpha(\vartheta')\rho(\vartheta', x)\, d\vartheta'} = \frac{\exp\big(\eta \cdot \vartheta^{\mathrm{T}} - \nu b(\vartheta)\big) \exp\big(\vartheta \cdot t(x)^{\mathrm{T}} - b(\vartheta)\big)}{\int_\Theta \exp\big(\eta \cdot (\vartheta')^{\mathrm{T}} - \nu b(\vartheta')\big) \exp\big(\vartheta' \cdot t(x)^{\mathrm{T}} - b(\vartheta')\big)\, d\vartheta'} \\
&= \frac{\exp\big(\vartheta \cdot (\eta + t(x))^{\mathrm{T}} - (\nu+1) b(\vartheta)\big)}{\int_\Theta \exp\big(\vartheta' \cdot (\eta + t(x))^{\mathrm{T}} - (\nu+1) b(\vartheta')\big)\, d\vartheta'} \\
&= Z\big(\eta + t(x), \nu + 1\big) \exp\big(\vartheta \cdot (\eta + t(x))^{\mathrm{T}} - (\nu+1) b(\vartheta)\big) = \alpha_{\eta + t(x), \nu+1}(\vartheta)
\end{aligned}$$

Beachte: Zu einer exponentiellen Familie bezüglich einer d-dimensionalen Statistik gehört eine konjugierte Familie mit $d+1$ Hyperparametern.

[5] Pierre-Simon Laplace, 1749–1827; zitiert nach Kersting & Wakolbinger, *Elementare Stochastik*, 2. Aufl., Birkhäuser 2010, S. 127

Zu Jeffreys-Priors

Eine prinzipielle Schwierigkeit (oder je nach Standpunkt auch: Chance) der Bayes-Statistik ist die Rechtfertigung der a priori-Verteilung. Gelegentlich wünscht man, eine möglichst „neutrale“ oder „uninformative“ a priori-Verteilung zu wählen, um die Daten „möglichst für sich selbst sprechen zu lassen“. Ein naheliegender Gedanke dazu könnte sein, $\alpha(\vartheta) \equiv \text{const}$ zu wählen (ein „flacher“ Prior). Eine Schwierigkeit dabei ist allerdings, dass die „Flachheit“ von der Parametrisierung abhängt (beispielsweise kann man in exponentialverteilten Beobachtungen entweder die Rate λ oder den Erwartungswert $1/\lambda$ als Parameter auffassen, und diese – logisch äquivalenten – Wahlen sind durch eine nicht-lineare Transformation der Parametermenge(n) verknüpft).

Ein systematischer Vorschlag geht auf H. Jeffreys[6] zurück: Wir betrachten den 1-dimensionalen Fall $\Theta \subseteq \mathbb{R}$ (ein Intervall), das a priori-Maß sei $\mu(d\vartheta) = \alpha(\vartheta)\, d\vartheta$ mit

$$\alpha(\vartheta) = c\big(I(\vartheta)\big)^{1/2} \tag{1.23}$$

(mit einer geeigneten Normierungskonstanten c), wobei

$$\begin{aligned} I(\vartheta) &= \mathbb{E}_\vartheta\left[\Big(\frac{\partial}{\partial\vartheta}\log\rho(\vartheta,X)\Big)^2\right] = \int_\Omega \frac{\big(\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\big)^2}{\rho(\vartheta,x)}\,dx = \int_\Omega \Big(\frac{\partial}{\partial\vartheta}\log\rho(\vartheta,x)\Big)\Big(\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\Big)\,dx \\ &= -\int_\Omega \Big(\frac{\partial^2}{\partial\vartheta^2}\log\rho(\vartheta,x)\Big)\rho(\vartheta,x)\,dx + \underbrace{\text{Randterme}}_{\text{nehme an }=0} = -\mathbb{E}_\vartheta\left[\frac{\partial^2}{\partial\vartheta^2}\log\rho(\vartheta,X)\right] \end{aligned}$$

die Fisher-Information [an der Stelle ϑ] ist (wir hatten sie bereits in Satz 5 betrachtet).

Sei eine (bijektive und genügend glatte) Umparametrisierung $\varphi : \Theta \to \Xi \subseteq \mathbb{R}$ gegeben und $\xi = \varphi(\vartheta)$ der neue Parameter. Dann hat ξ die a priori-Verteilung $\mu \circ \varphi^{-1}$ mit Dichte

$$\widetilde{\alpha}(\xi) = \frac{\alpha\big(\varphi^{-1}(\xi)\big)}{\varphi'\big(\varphi^{-1}(\xi)\big)} = \alpha\big(\varphi^{-1}(\xi)\big)(\varphi^{-1})'(\xi)$$

(denn mit Substitution $\xi = \varphi(\vartheta) \Leftrightarrow \vartheta = \varphi^{-1}(\xi)$ ist $\widetilde{\mu}(B) = \mu\big(\varphi^{-1}(B)\big) = \int_{\varphi^{-1}(B)} \alpha(\vartheta)\,d\vartheta = \int_B \widetilde{\alpha}(\xi)\,d\xi$ für $B \subset \Xi$).

Dann ist $\widetilde{\alpha}(\xi) = \widetilde{c}\big(\widetilde{I}(\xi)\big)^{1/2}$ mit einer geeigneten Normierungskonstante $\widetilde{c}$ und

$$\widetilde{I}(\xi) = -\widetilde{\mathbb{E}}_\xi\left[\frac{\partial^2}{\partial\xi^2}\log\widetilde{\rho}(\xi,X)\right]$$

(der Fisher-Information [an der Stelle ξ in den neuen Koordinaten]).

Man nennt die a priori-Verteilung (1.23) den Jeffreys-Prior. Es kann vorkommen, dass (1.23) ein unendliches Maß auf Θ ist (ein sogenannte uneigentliche a priori-Verteilung), aber trotzdem $\pi_x(\vartheta)\,d\vartheta$ zum Wahrscheinlichkeitsmaß normierbar ist.

[6] Sir Harold Jeffreys, FRS, 1891–1989; publiziert in H. Jeffreys (1946), An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. 186(1007):453–461

Beweis. Es ist

$$\begin{aligned}\widetilde{I}(\xi) &= \int_\Omega \Big(\frac{\partial}{\partial\xi}\log\widetilde{\rho}(\xi,x)\Big)^2\widetilde{\rho}(\xi,x)\,dx = \int_\Omega \Big(\frac{\frac{\partial}{\partial\xi}\widetilde{\rho}(\xi,x)}{\widetilde{\rho}(\xi,x)}\Big)^2\widetilde{\rho}(\xi,x)\,dx\\ &= \int_\Omega \Big(\frac{\partial}{\partial\vartheta}\rho(\vartheta,x)\big|_{\vartheta=\varphi^{-1}(\xi)}\cdot(\varphi^{-1})'(\xi)\Big)^2\frac{1}{\rho(\varphi^{-1}(\xi),x)}dx\\ &= \big((\varphi^{-1})'(\xi)\big)^2\int_\Omega\Big(\frac{\partial}{\partial\vartheta}\log\rho(\vartheta,x)\big|_{\vartheta=\varphi^{-1}(\xi)}\Big)^2\rho\big(\varphi^{-1}(\xi),x\big)\,dx = \big((\varphi^{-1})'(\xi)\big)^2 I\big(\varphi^{-1}(\xi)\big)\end{aligned}$$

und damit

$$\widetilde{\alpha}(\xi) = c\alpha\big(\varphi^{-1}(\xi)\big)\cdot(\varphi^{-1})'(\xi) = c\big(I(\varphi^{-1}(\xi))\big)^{1/2}\cdot(\varphi^{-1})'(\xi) = c\big(\widetilde{I}(\xi)\big)^{1/2}$$

□

Im multi-dimensionalen Fall $\vartheta = (\vartheta_1,\dots,\vartheta_d)\in\Theta\subseteq\mathbb{R}^d$ ist die Fisher-Informationsmatrix $I(\vartheta) = (I_{ij}(\vartheta))_{i,j=1}^d$ gegeben durch

$$I_{ij}(\vartheta) = \mathbb{E}_\vartheta\left[\Big(\frac{\partial}{\partial\vartheta_i}\log\rho(\vartheta,X)\Big)\Big(\frac{\partial}{\partial\vartheta_j}\log\rho(\vartheta,X)\Big)\right] = -\mathbb{E}_\vartheta\left[\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\log\rho(\vartheta,X)\right]$$

und die Dichte des Jeffreys-Priors durch

$$\alpha(\vartheta) = c\big(\det I(\vartheta)\big)^{1/2}$$

(wie man mittels analoger Rechnungen mit d-dimensionaler Transformationsformel zeigen kann).

Bayes-Analoga zu Konfidenzintervallen und Tests

Das Bayes-Analogon zum (frequentistischen) Konfidenzintervall ist ein „Kredibilitätsintervall“ (auch „Glaubwürdigkeitsintervall“)

$$\widetilde{C}(x) := \{\vartheta : \pi_x(\vartheta)\geq c\}$$

bei Beobachtungen x (wobei wir c so wählen, dass $\pi_x(\widetilde{C}(x))\geq\gamma$ für ein gegebenes Niveau $\gamma\in(0,1)$ ist.

Analog kann man für gegebene Zerlegung $\Theta_0\dot\cup\Theta_1=\Theta$ die a posteriori-Verteilung „befragen“, welche „Hypothese“ plausibler ist:

$$\text{entscheide für } \Theta_i, \text{ falls } \pi_x(\Theta_i) > \pi_x(\Theta_{1-i})$$

Kapitel 2

Lineares Modell

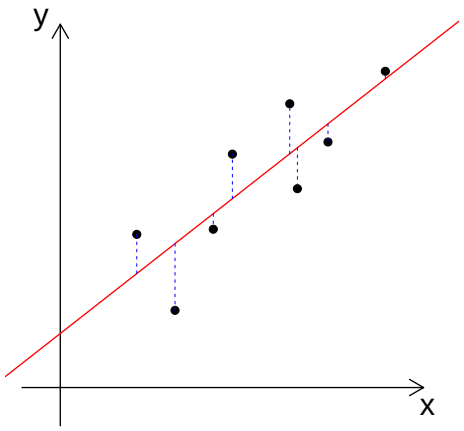
Erinnerung (lineare Regression)**.** Nehmen wir an, die Beobachtungen bestehen aus n Messwertpaaren (x_i, y_i), $i = 1, \dots, n$ (Werte in $\mathbb{R}^2$) und wir vermuten aus theoretischen Gründen einen zumindest „ungefähren“ (affin-)linearen Zusammenhang, d.h. bei „perfekter“ Messung und „perfektem“ Zusammenhang gälte

$$y_i = \beta_0 + \beta_1 x_i$$

für gewisse (uns unbekannte) Zahlen β_0 und β_1.

(Ein „Lehrbuchbeispiel“: y_i ist die Länge einer Stahlfeder bei Zugbelastung mit Gewicht x_i innerhalb des Gültigkeitsbereich des Hooke'schen Gesetzes.)

Aufgrund beispielsweise von Messungenauigkeiten (oder womöglich auch weil der lineare Zusammenhang in Wirklichkeit nur approximativ gilt) werden die realen Datenpunkte typischerweise nicht auf einer Geraden liegen.



Formulierung als statistisches Modell: $x_1, \dots, x_n$ seien feste (bekannte) Werte (x ist die „erklärende Variable“), für $\vartheta = (\beta_0, \beta_1) \in \Theta = \mathbb{R}^2$ sei unter P_ϑ

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \dots, n \quad \text{mit } \varepsilon_i \text{ u.i.v. mit } \mathbb{E}[\varepsilon_i] = 0,\ \mathrm{Var}[\varepsilon_i] = \sigma^2$$

und wir fassen die beobachteten y_i-Werte als Realisierungen der Y_i auf (y ist die „abhängige Variable“ oder „Zielgröße“).

Ein naheliegender Ansatz, $\vartheta = (\beta_0, \beta_1)$ zu schätzen, ist der *kleinste-Quadrate-Schätzer*: Finde $\widehat{\beta}_0, \widehat{\beta}_1$ so, dass

$$\sum_{i=1}^{n} \big(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\big)^2 = \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \tag{2.1}$$

Die Gerade $x \mapsto \widehat{\beta}_0 + \widehat{\beta}_1 x$ heißt dann auch die *Ausgleichsgerade* und man nennt

$$r_i := y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$$

das *Residuum* zum i-ten Beobachtungswert (der „Rest“ der Abweichung, die von dem Modell (nur) durch den „Rauschterm“ erklärt wird).

Mit

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i,$$
$$\sigma_x^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad \sigma_y^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2, \quad \mathrm{cov}_{x,y} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

($\mathrm{cov}_{x,y}$ ist die „empirische Kovarianz“ der x- und der y-Werte) schreibt sich die Lösung als

$$\widehat{\beta}_1 = \frac{\mathrm{cov}_{x,y}}{\sigma_x^2}, \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{2.2}$$

(man sieht daran insbesondere, dass die Ausgleichsgerade durch den Schwerpunkt $(\overline{x}, \overline{y})$ der Datenpunkte geht).

Zur Formel (2.2): Es ist

$$\frac{1}{n}\sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 = \frac{1}{n}\sum_{i=1}^{n} \big((y_i - \overline{y}) - \beta_1(x_i - \overline{x}) + (\overline{y} - \beta_0 - \beta_1\overline{x})\big)^2$$
$$\frac{1}{n}\sum_{i=1}^{n} \big((y_i - \overline{y}) - \beta_1(x_i - \overline{x})\big)^2 + 2\cdot 0 + \frac{1}{n}\sum_{i=1}^{n} \big(\overline{y} - \beta_0 - \beta_1\overline{x}\big)^2$$
$$= \sigma_y^2 - 2\beta_1 \mathrm{cov}_{x,y} + \beta_1^2\sigma_x^2 + \big(\overline{y} - \beta_0 - \beta_1\overline{x}\big)^2 = \sigma_y^2 - \frac{(\mathrm{cov}_{x,y})^2}{\sigma_x^2} + \Big(\beta_1\sigma_x - \frac{\mathrm{cov}_{x,y}}{\sigma_x}\Big)^2 + \big(\overline{y} - \beta_0 - \beta_1\overline{x}\big)^2$$

was offensichtlich für die Wahlen $\beta_1 = \widehat{\beta}_1$, $\beta_0 = \widehat{\beta}_0$ aus (2.2) minimal wird.

Alternativ kann man $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ nach β_1 und β_0 ableiten und jeweils die Ableitung $= 0$ setzen.

Der Wert des Minimierungsproblems in (2.1) ist $\sigma_y^2(1 - \kappa_{x,y}^2)$ (und dies zeigt: man kann Y anhand einer affin-linearen Funktion von X mit um den Faktor $1 - \kappa_{x,y}^2$ kleinerer Varianz „vorhersagen“ als mit der „besten“ Konstante $\mathbb{E}[Y]$).

$$\kappa_{x,y} := \frac{\mathrm{cov}_{x,y}}{\sigma_x \sigma_y}$$

ist der (empirische) Korrelationskoeffizient, auch *Pearsons Korrelationskoeffizient*[1] (stets gilt $-1 \le \kappa_{x,y} \le 1$, nach Cauchy-Schwarz-Ungleichung).

[1] nach Karl Pearson, 1858–1936

$$R = \kappa_{x,y}^2$$

nennt man auch das *Bestimmtheitsmaß*. Je näher R^2 an 1 liegt, um so besser passt die lineare Approximation der y-Werte durch die x-Werte. Das sieht man auch gut an der alternativen Formel

$$R = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^n r_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

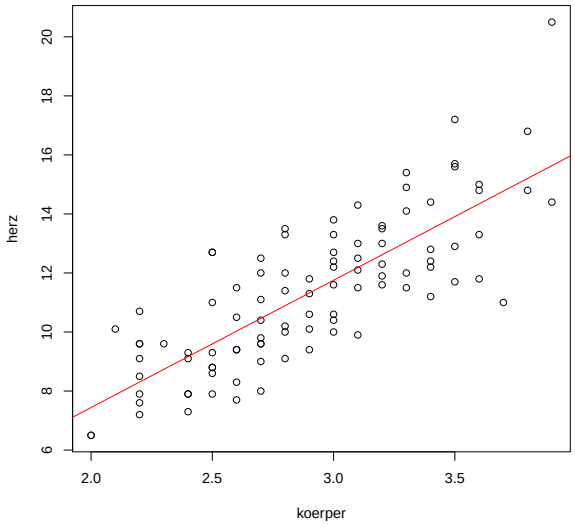
Übrigens: Wenn man zusätzlich annimmt, dass die ε_i u.i.v. $\sim \mathcal{N}(0, \sigma^2)$ sind, so ist der kleinste-Quadrate-Schätzer hier auch zugleich der Maximum-Likelihood-Schätzer (siehe Unterkapitel „Das normale lineare Modell“, Seite 53 unten).

Beispiel (Lineare Regression mit R)**.** Rs cats-Datensatz (aus R.F. Fisher, Biometrics, 3, 65-68 (1947); wir betrachten nur die 97 Kater):

```
> data(cats, package="MASS"); attach(cats)
> koerper <- Bwt[Sex=="M"]; herz <- Hwt[Sex=="M"]
> modell <- lm(herz ~ koerper)
> summary(modell)

Call:
lm(formula = herz ~ koerper)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7728 -1.0478 -0.2976  0.9835  4.8646

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1841     0.9983  -1.186    0.239
koerper       4.3127     0.3399  12.688   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.557 on 95 degrees of freedom
Multiple R-squared: 0.6289,Adjusted R-squared: 0.625
F-statistic:   161 on 1 and 95 DF,  p-value: < 2.2e-16
```

Wir werden die Bedeutung dieser Ausgaben unten genauer anschauen. Für den Moment betrachten wir das Ergebnis grafisch:

```
> plot(koerper, herz)
> abline(modell$coeff, col="red")
```



Der allgemeine Rahmen

Modellvorstellung: In einer „Population“ besitzt jedes Individuum/Objekt einen (reellen) y-Wert („Zielgröße“) und p (reelle) x-Werte $x_1, \dots, x_p$ (p „erklärende Variablen“).

Seien $\big((X_1, \dots, X_p), Y\big)$ die Merkmale eines (zufällig) aus der Population gezogenen Individuums.

Wir nehmen an, dass für die Verteilung von Y, gegeben beobachtete (oder je nach Situation auch von uns vorgegebene) Werte $(X_1, \dots, X_p) = (x_1, \dots, x_p)$ gilt

$$\mathbb{E}_{\boldsymbol{\beta}}[Y \,|\, (X_1, \dots, X_p) = (x_1, \dots, x_p)] = \beta_1 x_1 + \cdots + \beta_p x_p$$

und

$$\mathrm{Var}_{\boldsymbol{\beta}}[Y \,|\, (X_1, \dots, X_p) = (x_1, \dots, x_p)] = \sigma^2$$

für einen Parametervektor $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ und ein $\sigma^2 > 0$.

Der Einfachheit der Formulierung halber integrieren wir hier den Parameter für einen konstanten Wert („Achsenabschnitt“) in das Modell, indem wir annehmen, dass für eine der erklärenden Variablen $\equiv 1$ erfüllt.

Gegeben seien n Beobachtungen, zur i-ten Beobachtung gehört y_i (beobachteter Wert der Zielgröße) und $x_{i,1}, x_{i,2}, \dots, x_{i,p}$ (Werte der p erklärenden Variablen), wir interpretieren diese als n unabhängige Realisierungen von $\big((X_1, \dots, X_p), Y\big)$ und schreiben

$$y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

mit ε_i (Realisierungen von) unabhängige(n) Zufallsvariablen mit $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2$, wobei wir das „wahre“ $\boldsymbol{\beta}$ und das „wahre“ σ^2 nicht kennen.

Lineares Modell in Matrix-Schreibweise n Beobachtungen, zur i-ten Beobachtung gehört y_i und $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$ (und wir denken $n \geq p$),

$$y_i = \sum_{j=1}^{p} x_{i,j}\beta_j + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.3}$$

zusammengefasst

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.4}$$

mit

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, wobei

$$\mathbb{E}[\varepsilon_i] = 0, \quad \mathrm{Var}[\varepsilon_i] = \sigma^2 \tag{2.5}$$

und die $\varepsilon_1, \ldots, \varepsilon_n$ unabhängig (bzw. für L^2-Theorie genügt: paarweise unkorreliert).

$\boldsymbol{X}$ ist die „Designmatrix“ (die i-te Zeile gehört zur i-ten Beobachtung, die j-te Spalte beschreibt den Einfluß des Parameters β_j).

Für Identifizierbarkeit nehmen wir an, dass die Spalten von $\boldsymbol{X}$ linear unabhängig sind (d.h. $\boldsymbol{X}$ hat Spaltenrang p und somit insgesamt Rang p). Andernfalls gäbe es ein k_*, so dass Spalte $x_{\cdot,k_*} = \sum_{k\neq k_*}^{p} \lambda_k x_{\cdot,k}$ für gewisse λ_k mit $\sum_{k\neq k_*}^{p} |\lambda_k| > 0$ erfüllt. Die Parametervektoren $\boldsymbol{\beta}$ und $\boldsymbol{\beta}' = \boldsymbol{\beta} + \boldsymbol{e}_{k_*} - \sum_{k\neq k_*}^{p} \lambda_k \boldsymbol{e}_k$ führten dann auf dieselben Beobachtungen:

$$y_i' = \sum_{j=1}^{p} x_{i,j}\beta_j' + \varepsilon_i = \sum_{j=1}^{p} x_{i,j}\beta_j \; + \; x_{i,k_*} - \sum_{k\neq k_*}^{p} \lambda_k x_{i,k} + \varepsilon_i = \sum_{j=1}^{p} x_{i,j}\beta_j + 0 + \varepsilon_i = y_i, \quad i = 1, \ldots, n$$

Kleinste-Quadrate-Ansatz

Sei $n \geq p$ (wir denken an $n \gg p$), gegeben beobachtetes $\boldsymbol{y}$ und bekannte/beobachtete Designmatrix $\boldsymbol{X}$. Typischerweise besitzt das lineare Gleichungssystem

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$$

(aufgefasst als Bestimmungsgleichungen für die β_i) keine Lösung.

Kleinste-Quadrate-Schätzer (“least squares”)

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} := \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \right\} \tag{2.6}$$

Beobachtung 25 (Berechnung des kleinste-Quadrate-Schätzers)**.** $\boldsymbol{X}$ habe (maximalen) Rang p, d.h. die p Spalten von $\boldsymbol{X}$ sind linear unabhängig. Dann hat die $p \times p$-Matrix $\boldsymbol{X}^T\boldsymbol{X}$ vollen Rang p, ist insbesondere invertierbar, und der kleinste-Quadrate-Schätzer ist

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.7}$$

Beweis. $\boldsymbol{X}^T\boldsymbol{X}$ hat vollen Rang: Falls

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0} \text{ für ein } \boldsymbol{c} \in \mathbb{R}^p \setminus \{\boldsymbol{0}\}$$

gälte, so wäre $\|\boldsymbol{X}\boldsymbol{c}\|_2^2 = \boldsymbol{c}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c} = 0$ im Widerspruch zur Ann., dass $\boldsymbol{X}$ Rang p hat.

Zur Form von $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$: Sei

$$L = \{\boldsymbol{X}\boldsymbol{\eta} : \boldsymbol{\eta} \in \mathbb{R}^p\} \subset \mathbb{R}^n$$

der Spaltenraum von $\boldsymbol{X}$, $\Pi_L : \mathbb{R}^n \to L$ die orthogonale Projektion auf L, für $\boldsymbol{y} \in \mathbb{R}^n$ ist $\Pi_L\boldsymbol{y}$ ist durch die äquivalenten Eigenschaften (i), (ii) und (iii) charakterisiert

(i) $\Pi_L\boldsymbol{y} \in L$ und $\|\boldsymbol{y} - \Pi_L\boldsymbol{y}\|_2^2 \le \min_{\boldsymbol{u}\in L}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2$

(ii) $\Pi_L\boldsymbol{y} \in L$ und $(\boldsymbol{y} - \Pi_L\boldsymbol{y}) \perp L$

(iii) $\Pi_L\boldsymbol{y} \in L$ und $\boldsymbol{u}^T(\boldsymbol{y} - \Pi_L\boldsymbol{y}) = 0$ für alle $\boldsymbol{u}$ aus einer Basis von $L \iff \boldsymbol{X}^T(\boldsymbol{y} - \Pi_L\boldsymbol{y}) = \boldsymbol{0}$

Nun ist für $\boldsymbol{y} \in \mathbb{R}^n$:

$$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \in L \text{ und } \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{0}$$

also $\Pi_L\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ und somit erfüllt $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ aus (2.7) die definierende Eigenschaft (2.6). □

Bemerkung. Man kann Beob. 25 auch analytisch einsehen: Die „Zielfunktion“

$$\boldsymbol{\beta} \mapsto f(\boldsymbol{\beta}) := \sum_{i=1}^n \Big(y_i - \textstyle\sum_{j=1}^p x_{i,j}\beta_j\Big)^2$$

ist quadratisch, das Gleichungssystem

$$\frac{\partial}{\partial\beta_k}\sum_{i=1}^n \Big(y_i - \textstyle\sum_{j=1}^p x_{i,j}\beta_j\Big)^2 = -2\sum_{i=1}^n \Big(y_i - \textstyle\sum_{j=1}^p x_{i,j}\beta_j\Big)x_{i,k} \stackrel{!}{=} 0 \quad \text{für } k = 1,\dots,p$$

ist in Matrixform geschrieben äquivalent zu den „Normalgleichungen“

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

Diese bestimmen die Lösungsmenge $\{\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\}$ auch in dem Fall, dass $\boldsymbol{X}^T\boldsymbol{X}$ nicht invertierbar ist.

(Beachte: Die Hesse-Matrix $\Big(\frac{\partial^2}{\partial\beta_k\partial\beta_\ell}f(\boldsymbol{\beta})\Big)_{k,\ell=1}^p = \big(\sum_{i=1}^n x_{i,\ell}x_{i,k}\big)_{k,\ell=1}^p = \boldsymbol{X}^T\boldsymbol{X}$ [hängt nicht von $\boldsymbol{\beta}$ ab und] ist [überall] positiv [semi-]definit).

$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ der kleinste-Quadrate-Schätzer aus (2.7),

$$\widehat{\boldsymbol{y}} := \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

die angepassten Werte der y-Beobachtungen (“gefittete” / angepasste Werte), $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ die „Hut-Matrix“,

$$\boldsymbol{r} := \boldsymbol{y} - \widehat{\boldsymbol{y}} \ \Big(= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\Big)$$

sind die / ist der Vektor der „Residuen“ (die Abweichungen der Beobachtungen vom Modell, sozusagen der „unerklärte Rest“).

Satz 26 (Satz von Gauß-Markov)**.** *Im linearen Modell (mit unkorrelierten, zentrierten ε_i mit derselben Varianz σ^2) gilt*

1. *Der kleinste-Quadrate-Schätzer $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ ist erwartungstreu für $\boldsymbol{\beta}$, d.h. $\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\big] = \boldsymbol{\beta}$ (für jede Wahl des „wahren“ $\boldsymbol{\beta}$).*

2. *$\tau : \mathbb{R}^p \to \mathbb{R}$ eine lineare Kenngröße von $\boldsymbol{\beta}$ (d.h. $\tau(\boldsymbol{\beta}) = \boldsymbol{c}^T\boldsymbol{\beta}$ für ein $\boldsymbol{c} \in \mathbb{R}^p$, z.B. $\tau(\boldsymbol{\beta}) = \beta_1$), dann ist $\widehat{\tau} := \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ erwartungstreuer Schätzer für $\tau(\boldsymbol{\beta})$ und hat unter allen erwartungstreuen linearen Schätzern für $\tau(\boldsymbol{\beta})$ die kleinste Varianz.*

3. *Weiterhin ist*

$$\widehat{\sigma}^2 := \frac{\|\boldsymbol{r}\|_2^2}{n-p} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\|_2^2}{n-p} = \frac{\|\boldsymbol{y} - \Pi_L\boldsymbol{y}\|_2^2}{n-p} = \frac{\|\boldsymbol{y}\|_2^2 - \|\Pi_L\boldsymbol{y}\|_2^2}{n-p}$$

ein erwartungstreuer Schätzer für die Varianz σ^2 (der ε_i).

Beweis. Sei $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma > 0$.
1. Es ist

$$\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{y}\big] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\big] = \boldsymbol{X}\boldsymbol{\beta}$$

somit nach (2.7)

$$\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\big] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\big] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

d.h. $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ ist erwartungstreu.

2. Nach 1. ist $\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\big] = \boldsymbol{c}^T\boldsymbol{\beta} = \tau(\boldsymbol{\beta})$, d.h.

$$\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c})^T\boldsymbol{y} =: \boldsymbol{a}^T\boldsymbol{y}$$

ist erwartungstreu.

Sei $S := \boldsymbol{b}^T\boldsymbol{y}$ ein weiterer erwartungstreuer Schätzer für $\tau(\boldsymbol{\beta})$, d.h.

$$\tau(\boldsymbol{\beta}) = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[S\big] = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta} \quad \text{für jede Wahl von } \boldsymbol{\beta} \in \mathbb{R}^p$$

Folglich gilt für jedes $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\boldsymbol{0} = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[S\big] - \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\big] = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{b}-\boldsymbol{a})^T\boldsymbol{X}\boldsymbol{\beta}$$

d.h. $(\boldsymbol{b}-\boldsymbol{a}) \perp L$, wegen $\boldsymbol{a} \in L$ daher

$$\boldsymbol{a} = \Pi_L\boldsymbol{b}, \quad \text{insbesondere } \|\boldsymbol{a}\|_2^2 \le \|\boldsymbol{b}\|_2^2$$

Somit

$$\begin{aligned}
&\mathrm{Var}_{(\boldsymbol{\beta},\sigma^2)}\big[S\big] - \mathrm{Var}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\big] \\
&= \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\Big[\big(\boldsymbol{b}^T\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta}\big)^2 - \big(\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{\beta}\big)^2\Big] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\Big[\big(\boldsymbol{b}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big)^2 - \big(\boldsymbol{a}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big)^2\Big] \\
&= \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[(\boldsymbol{b}^T\boldsymbol{\varepsilon})^2 - (\boldsymbol{a}^T\boldsymbol{\varepsilon})^2\big] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{b}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{b} - \boldsymbol{a}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{a}\big] = \boldsymbol{b}^T\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\big]\boldsymbol{b} - \boldsymbol{a}^T\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\big]\boldsymbol{a} \\
&= \sigma^2(\boldsymbol{b}^T\boldsymbol{b} - \boldsymbol{a}^T\boldsymbol{a}) = \sigma^2\left(\|\boldsymbol{b}\|_2^2 - \|\boldsymbol{a}\|_2^2\right) \ge 0
\end{aligned}$$

3. Sei $u_1, \dots, u_p$ eine Orthonormalbasis von L, ergänze zu ONB $u_1, \dots, u_n$ von $\mathbb{R}^n$. Mit

$$\boldsymbol{O} := \begin{pmatrix} u_1 & u_2 & \dots & u_n \\ | & | & & | \end{pmatrix} \quad \text{ist} \quad \Pi_L = \boldsymbol{O} \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & \ddots \\ 0 & & & & 0 \end{pmatrix} \boldsymbol{O}^T \tag{2.8}$$

(wobei die Diagonalmatrix p Einsen auf der Diagonale besitzt) die Darstellung der Projektion auf L in Koordinaten. Setze

$$\widetilde{\boldsymbol{\varepsilon}} := \boldsymbol{O}^T \boldsymbol{\varepsilon}$$

Damit ist

$$\begin{aligned} (n-p)\widehat{\sigma}^2 &= \left\| \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \Pi_L(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \right\|_2^2 = \left\| \boldsymbol{\varepsilon} - \Pi_L \boldsymbol{\varepsilon} \right\|_2^2 \\ &= \left\| \boldsymbol{O}\widetilde{\boldsymbol{\varepsilon}} - \boldsymbol{O}\boldsymbol{D}\boldsymbol{O}^T\boldsymbol{O}\widetilde{\boldsymbol{\varepsilon}} \right\|_2^2 = \left\| \boldsymbol{O}\widetilde{\boldsymbol{\varepsilon}} - \boldsymbol{O}\boldsymbol{D}\widetilde{\boldsymbol{\varepsilon}} \right\|_2^2 = \left\| \widetilde{\boldsymbol{\varepsilon}} - \boldsymbol{D}\widetilde{\boldsymbol{\varepsilon}} \right\|_2^2 = \sum_{i=p+1}^{n} \widetilde{\varepsilon}_i^2 \end{aligned}$$

und mit

$$\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\left[\widetilde{\varepsilon}_k\right] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\left[\sum_{i,j=1}^{n} O_{ik} O_{jk} \varepsilon_i \varepsilon_j \right] = \sigma^2 \sum_{i=1}^{n} O_{ik} O_{ik} = \sigma^2 \sum_{i=}^{n} O_{ik} O_{ki}^T = \sigma^2 (\boldsymbol{O}\boldsymbol{O}^T)_{ii} = \sigma^2$$

folgt die Behauptung. □

Beobachtung. Die Kovarianzmatrix von $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\widehat{\boldsymbol{\beta}}_{\mathrm{LS},1}, \dots, \widehat{\boldsymbol{\beta}}_{\mathrm{LS},p})^T$ ist

$$\mathrm{Cov}_{\boldsymbol{\beta},\sigma^2}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right] = \sigma^2 (\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2.9}$$

Insbesondere ist die Varianz von $\widehat{\boldsymbol{\beta}}_{\mathrm{LS},j}$ gleich $\sigma^2((\boldsymbol{X}^T\boldsymbol{X})^{-1})_{jj}$, der „Standardfehler“ von $\widehat{\boldsymbol{\beta}}_{\mathrm{LS},j}$ ist gegeben durch $\sqrt{\widehat{\sigma^2}}\left(((\boldsymbol{X}^T\boldsymbol{X})^{-1})_{jj}\right)^{1/2}$.

Weiter ist die Kovarianzmatrix des Residuenvektors (mit

$$\boldsymbol{H} = (h_{i,j})_{i,j=1}^n = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

der „Hut-Matrix“)

$$\begin{aligned} \mathrm{Cov}_{\boldsymbol{\beta},\sigma^2}\left[\boldsymbol{r}\right] &= \mathrm{Cov}_{\boldsymbol{\beta},\sigma^2}\left[\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}\right] = (\boldsymbol{I} - \boldsymbol{H})\mathrm{Cov}_{\boldsymbol{\beta},\sigma^2}\left[\boldsymbol{y}\right](\boldsymbol{I} - \boldsymbol{H})^T = \sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^T \\ &= \sigma^2\left(\boldsymbol{I} - \boldsymbol{H} - \boldsymbol{H}^T + \boldsymbol{H}\boldsymbol{H}^T\right) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}) \end{aligned}$$

Man betrachtet daher gelegentlich auch die *standardisierten Residuen* $\boldsymbol{r'} = (r'_1, \dots, r'_n)^T$ mit

$$r'_i := \frac{r_i}{\widehat{\sigma}\sqrt{1 - h_{i,i}}}, \quad i = 1, \dots, n$$

(diese sollten Streuung ≈ 1 haben, $h_{i,i}$ heißt auch die “leverage” oder der „Hebelwert“ von Beobachtung i).

Beweis. Allgemein: Es ist

$$\mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[\boldsymbol{y}] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\big] = \sigma^2\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\big] = \sigma^2\boldsymbol{I}$$

und für eine $n\times n$-Matrix $\boldsymbol{M}$ ist

$$\mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[\boldsymbol{M}\boldsymbol{y}] = \mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[\boldsymbol{M}\boldsymbol{\varepsilon}] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[(\boldsymbol{M}\boldsymbol{\varepsilon})(\boldsymbol{M}\boldsymbol{\varepsilon})^T\big] = \boldsymbol{M}\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}^T\big]\boldsymbol{M}^T = \sigma^2\boldsymbol{M}\boldsymbol{M}^T$$

Mit $\boldsymbol{M} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ in obigem ergibt sich

$$\begin{aligned}\mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}\big[\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\big] &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big)^T\\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\big)^T = \sigma^2\big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\big)^T = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

□

Bemerkung. Man kann mittels (2.9) auch die Konsistenz des LS-Schätzers zeigen (für $n\to\infty$ bei festem p) Sei $\boldsymbol{X}^{(n)}$ und $\boldsymbol{y}^{(n)}$ Designmatrix und Beobachtungsvektor zu n Beobachtungen (jeweils mit u.i.v. Störgrößen ε_i) und es gelte $\mathrm{tr}\big((\boldsymbol{X}^{(n)})^T\boldsymbol{X}^{(n)}\big)\to 0$ für $n\to\infty$, $\widehat{\boldsymbol{\beta}}^{(n)}_{\mathrm{LS}}$ der zugehörige LS-Schätzer. Dann gilt $\widehat{\boldsymbol{\beta}}^{(n)}_{\mathrm{LS}} \xrightarrow{\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}} \boldsymbol{\beta}$ für $n\to\infty$, denn $\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\big[\widehat{\boldsymbol{\beta}}^{(n)}_{\mathrm{LS},j}\big] = \beta_j$ und $\mathrm{Var}_{(\boldsymbol{\beta},\sigma^2)}\big[\widehat{\boldsymbol{\beta}}^{(n)}_{\mathrm{LS},j}\big]\to 0$.

Definition und Beobachtung (RSS und Bestimmtheitsmaß)**.** Man nennt

$$\mathrm{RSS} := \big\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\big\|_2^2 = \big\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\big\|_2^2 = \sum_{i=1}^n r_i^2$$

die Residuenquadratsumme oder "residual sum of squares".

Mit $\overline{y} := \frac{1}{n}\sum_{i=1}^n y_i$ nennt man

$$R^2 := \frac{\big(\sum_{i=1}^n (y_i-\overline{y})(\widehat{y}_i-\overline{y})\big)^2}{\sum_{i=1}^n (y_i-\overline{y})^2\cdot\sum_{i=1}^n(\widehat{y}_i-\overline{y})^2} \tag{2.10}$$

das *Bestimmtheitsmaß*.

Es gilt

$$R^2 = \frac{\sum_{i=1}^n(\widehat{y}_i-\overline{y})^2}{\sum_{i=1}^n(y_i-\overline{y})^2} = 1-\frac{\mathrm{RSS}}{\sum_{i=1}^n(y_i-\overline{y})^2} \tag{2.11}$$

(2.11) zeigt, dass die durch die Regression angepassten Werte den Anteil R^2 der Variabilität [Varianz] der Beobachtungswerte $(y_i)_{i=1,\dots,n}$ „erklären".

Zum Beweis von (2.11): Es ist $(\boldsymbol{y}-\widehat{\boldsymbol{y}})\perp L$, und da eine der Spalten von $\boldsymbol{X}\equiv 1$ ist, gilt

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \widehat{y}_i$$

Wegen $\widehat{\boldsymbol{y}}\in L$ gilt auch $(\boldsymbol{y}-\widehat{\boldsymbol{y}})^T\widehat{\boldsymbol{y}} = 0$, also

$$\widehat{\boldsymbol{y}}^T\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{y}}^T\big(\widehat{\boldsymbol{y}}^T-\boldsymbol{y}+\boldsymbol{y}\big) = \widehat{\boldsymbol{y}}^T\boldsymbol{y}$$

und

$$\sum_{i=1}^{n}\big((y_i-\overline{y})^2-(\widehat{y}_i-\overline{y})^2\big)=\sum_{i=1}^{n}\big(y_i^2-\widehat{y}_i^2\big)=\boldsymbol{y}^T\boldsymbol{y}-\widehat{\boldsymbol{y}}^T\widehat{\boldsymbol{y}}=\boldsymbol{y}^T(\boldsymbol{y}-\widehat{\boldsymbol{y}})=(\boldsymbol{y}-\widehat{\boldsymbol{y}})^T(\boldsymbol{y}-\widehat{\boldsymbol{y}})=\mathrm{RSS}$$

Andererseits ist auch

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y})(\widehat{y}_i-\overline{y})&=(\boldsymbol{y}-\overline{y}\mathbf{1})^T(\widehat{\boldsymbol{y}}-\overline{y}\mathbf{1})=(\boldsymbol{y}-\widehat{\boldsymbol{y}}+\widehat{\boldsymbol{y}}-\overline{y}\mathbf{1})^T(\widehat{\boldsymbol{y}}-\overline{y}\mathbf{1})=(\widehat{\boldsymbol{y}}-\overline{y}\mathbf{1})^T(\widehat{\boldsymbol{y}}-\overline{y}\mathbf{1})\\&=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2\end{aligned}$$

Insgesamt folgt

$$R^2=\frac{\big(\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2\big)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2\cdot\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}=\frac{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=1-\frac{\mathrm{RSS}}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

Bemerkung. Man betrachtet auch das angepasste / adjustierte Bestimmtheitsmaß

$$\bar{R}^2=1-\frac{\big(\sum_{i=1}^{n}r_i^2\big)/(n-p)}{\big(\sum_{i=1}^{n}(y_i-\overline{y})^2\big)/(n-1)}=1-\frac{n-1}{n-p}\frac{\mathrm{RSS}}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

(bei dem in Zähler und Nenner jeweils durch die „korrekte Anzahl Freiheitsgrade“ normiert wird).

Das normale lineare Modell

Wir modellieren die Beobachtungen $\boldsymbol{y}=(y_1,\dots,y_n)^T$ wie in (2.3–2.4) via

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

Wir nehmen nun zusätzlich an, dass die ε_i u.i.v., $\sim\mathcal{N}(0,\sigma^2)$ sind.

Gegeben $\boldsymbol{X}$ (und festes $\boldsymbol{\beta}\in\mathbb{R}^p$, $\sigma^2>0$) hat $\boldsymbol{y}=(y_1,\dots,y_n)^T$ dann die Dichte

$$\prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\big(-(y_i-(\boldsymbol{X}\boldsymbol{\beta})_i)^2/(2\sigma^2)\big)=(2\pi\sigma^2)^{-n/2}\exp\big(-\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2\sigma^2)\big)$$

Demnach ist $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ auch der Maximum-Likelihood-Schätzer für $\boldsymbol{\beta}$ und es ist

$$\widehat{\sigma}^2_{\mathrm{ML}}=\big\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\big\|_2^2/n$$

(denn $\frac{\partial}{\partial v}\log\rho\big((\boldsymbol{\beta},v);\boldsymbol{y}\big)=\frac{\partial}{\partial v}\big(-\frac{n}{2}\log v-\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2/(2v)\big)=-\frac{n}{2}\frac{1}{v}+\frac{1}{v^2}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2=0\iff v=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2/n$).

Beobachtung 27 (Verteilungseigenschaften im normalen linearen Modell)**.** $\boldsymbol{\beta}\in\mathbb{R}^p,\sigma>0$, unter $\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}$ gilt:

1. $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\sim\mathcal{N}\big(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big)$
2. $\frac{n-p}{\sigma^2}\widehat{\sigma}^2$ ist χ^2_{n-p}-verteilt und unabhängig von $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$.

3. $\frac{1}{\sigma^2}\left\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}-\boldsymbol{\beta})\right\|_2^2 = \frac{1}{\sigma^2}\left\|\Pi_L\boldsymbol{y}-\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[\boldsymbol{y}]\right\|_2^2$ ist χ^2_p-verteilt und unabhängig von $\widehat{\sigma}^2$, insbesondere

$$\frac{\left\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}-\boldsymbol{\beta})\right\|_2^2}{p\widehat{\sigma}^2} \sim \mathrm{Fisher}(p, n-p)$$

4. $H \subset L$ linearer Teilraum von L, $\dim H = r < p$ und $\boldsymbol{X}\boldsymbol{\beta} \in H$, dann ist $\frac{1}{\sigma^2}\left\|\Pi_L\boldsymbol{y}-\Pi_H\boldsymbol{y}\right\|_2^2 \sim \chi^2_{p-r}$ und unabhängig von $\widehat{\sigma}^2$, insbesondere ist

$$F_{H,L} := \frac{\left\|\Pi_L\boldsymbol{y}-\Pi_H\boldsymbol{y}\right\|_2^2/(p-r)}{\left\|\boldsymbol{y}-\Pi_L\boldsymbol{y}\right\|_2^2/(n-p)} = \frac{\left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}-\Pi_H\boldsymbol{y}\right\|_2^2}{(p-r)\widehat{\sigma}^2} \sim \mathrm{Fisher}(p-r, n-p)$$

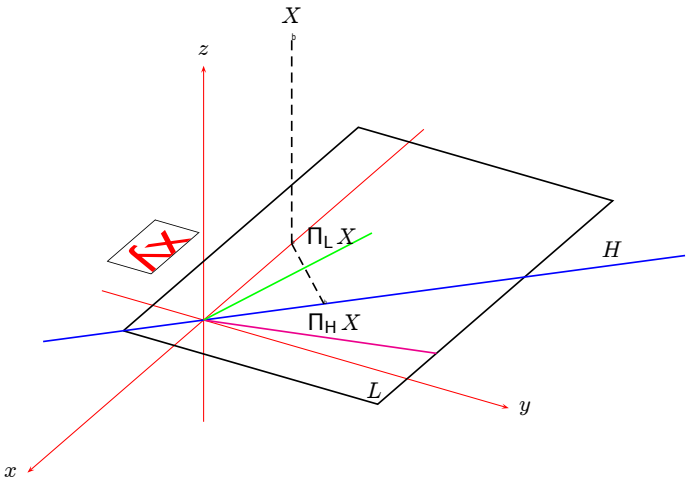


Abbildung 2.1: Darstellung der Projektionen eines Punktes X im $\mathbb{R}^3$ auf $H \subset L \subset \mathbb{R}^3$, L ist dabei die Ebene, die durch das Rechteck L dargestellt werden soll und H der Unterraum, der von der blauen Geraden aufgespannt wird.

Beweis von Beob. 27. Es gilt (mit dem Satz von Pythagoras)

$$\left\|\Pi_L\boldsymbol{y}-\Pi_H\boldsymbol{y}\right\|_2^2 = \left\|\Pi_L\boldsymbol{y}\right\|_2^2 - \left\|\Pi_H\boldsymbol{y}\right\|_2^2 = \left\|\boldsymbol{y}-\Pi_H\boldsymbol{y}\right\|_2^2 - \left\|\boldsymbol{y}-\Pi_L\boldsymbol{y}\right\|_2^2$$

und

$$\left\|\boldsymbol{y}-\Pi_L\boldsymbol{y}\right\|_2^2 = \left\|\boldsymbol{y}\right\|_2^2 - \left\|\Pi_L\boldsymbol{y}\right\|_2^2$$

1. $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ ist p-dimensional normalverteilt (als affin-lineares Bild des multivariat normalverteilten $\boldsymbol{\varepsilon}$), wie oben berechnet ist

$$\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right] = \boldsymbol{\beta}, \quad \mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

2. Sei o.B.d.A. $u_1, \dots, u_n$ eine ONB von $\mathbb{R}^n$ mit $H = \mathrm{span}(u_1, \dots, u_r)$, $L = \mathrm{span}(u_1, \dots, u_s)$; setze (wie im Beweis des Satzes von Gauß-Markov, Satz 26)

$$\boldsymbol{O} := \begin{pmatrix} u_1 & u_2 & \dots & u_n \\ | & | & & | \\ | & | & \dots & | \end{pmatrix}, \quad \widetilde{\boldsymbol{\varepsilon}} := \boldsymbol{O}^T\boldsymbol{\varepsilon}$$

dann ist $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\sigma^2)^{\otimes n}$.

$$\frac{n-p}{\sigma^2}\widehat{\sigma}^2 = \frac{1}{\sigma^2}\left\|\boldsymbol{y}-\Pi_L\boldsymbol{y}\right\|_2^2 = \frac{1}{\sigma^2}\left\|\boldsymbol{\varepsilon}-\Pi_L\boldsymbol{\varepsilon}\right\|_2^2 = \frac{1}{\sigma^2}\sum_{i=p+1}^{n}\widetilde{\varepsilon}_i^2 \sim \chi^2_{n-p}$$

3. $\frac{1}{\sigma^2}\left\|\Pi_L \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 = \sum_{i=1}^{p} p\widehat{\varepsilon}_i^2 \sim \chi^2_p$ und ist unabhängig von $\widehat{\sigma}^2$.

4. $\frac{1}{\sigma^2}\left\|\Pi_L \boldsymbol{y} - \Pi_H \boldsymbol{y}\right\|_2^2 = \frac{1}{\sigma^2}\left\|\Pi_L \boldsymbol{\varepsilon} - \Pi_H \boldsymbol{\varepsilon}\right\|_2^2 = \frac{1}{\sigma^2}\sum_{i=r+1}^{p} \widetilde{\varepsilon}_i^2 \sim \chi^2_{p-r}$ und unabhängig von $\widehat{\sigma}^2$.

□

Korollar 28 (Konfidenzbereiche). *Sei $\alpha \in (0,1)$.*

1. $C_{\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}} := \{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p : \|\boldsymbol{X}(\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\mathrm{LS}})\|_2^2 < p\widehat{\sigma}^2 f_{p,n-p;1-\alpha}\}$ *ist ein Konfidenzbereich für $\boldsymbol{\beta}$ zum Sicherheitsniveau $1-\alpha$, wo $f_{p,n-p;1-\alpha}$ = $(1-\alpha)$-Quantil der Fisher$_{p,n-p}$-Verteilung.*

2. $\tau(\boldsymbol{\beta}) = \boldsymbol{c}^T\boldsymbol{\beta}$ *(mit einem $\boldsymbol{c} \in \mathbb{R}^p$) ein lineares Parametermerkmal.*

$$C_{\tau(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}})} = \left(\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \delta\sqrt{\widehat{\sigma}^2}, \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} + \delta\sqrt{\widehat{\sigma}^2}\right),$$

mit $\delta = \sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}}\, t_{n-p,1-\frac{\alpha}{2}}$ und $t_{n-p,1-\frac{\alpha}{2}}$ = $(1-\frac{\alpha}{2})$-Quantil der Student-T-Verteilung mit $n-p$-Freiheitsgraden, ist ein Konfidenzintervall für $\tau(\boldsymbol{\beta})$ zum Sicherheitsniveau $1-\alpha$.

3. $C_{\widehat{\sigma}^2} := \left(\frac{n-p}{\chi^2_{n-p,1-\frac{\alpha}{2}}}\widehat{\sigma}^2, \frac{n-p}{\chi^2_{n-p,\frac{\alpha}{2}}}\widehat{\sigma}^2\right)$ *ist ein Konfidenzintervall für σ^2 zum Sicherheitsniveau $1-\alpha$, wobei $\chi^2_{n-p,\frac{\alpha}{2}}$ das $\alpha/2$- und $\chi^2_{n-p,1-\frac{\alpha}{2}}$ das $(1-\alpha/2)$-Quantil der χ^2-Verteilung mit $n-p$ Freiheitsgraden ist.*

Beweis. 1. Unter $\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}$ ist $\dfrac{\|\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathrm{LS}})\|_2^2}{p\widehat{\sigma}^2} \sim \mathrm{Fisher}(p, n-p)$ ✓

2. $Z := \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} \sim \mathcal{N}(\boldsymbol{c}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c})$, also $Z^* := \dfrac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim \mathcal{N}(0,1)$, also $T := \dfrac{Z^*}{\sqrt{\widehat{\sigma}^2/\sigma^2}} \sim \mathrm{Student}(n-p)$, folglich

$$\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\left(C_{\tau(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}})} \not\ni \boldsymbol{c}^T\boldsymbol{\beta}\right) = \mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\left(|T| > t_{n-p,1-\frac{\alpha}{2}}\right) = \alpha$$

3. $$\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\left(\sigma^2 \in \left(\frac{n-p}{\chi^2_{n-p,1-\frac{\alpha}{2}}}\widehat{\sigma}^2, \frac{n-p}{\chi^2_{n-p,\frac{\alpha}{2}}}\widehat{\sigma}^2\right)\right) = \mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\left(\chi^2_{n-p,\frac{\alpha}{2}} < \frac{n-p}{\sigma^2}\widehat{\sigma}^2 < \chi^2_{n-p,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

□

Korollar 29 (Tests im normalverteilten linearen Modell). *Sei $\alpha \in (0,1)$.*

1. *(t-Test für $\boldsymbol{c}^T\boldsymbol{\beta} = \tau_0$, bzw. $\leq$ oder $\geq$)*
   Der Ablehnungsbereich

$$\left\{\left|\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \tau_0\right| > t_{n-p,1-\frac{\alpha}{2}}\sqrt{\widehat{\sigma}^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}}\right\}$$

definiert einen Test von $H_0 : \boldsymbol{c}^T\boldsymbol{\beta} = \tau_0$ gegen $H_1 : \boldsymbol{c}^T\boldsymbol{\beta} \neq \tau_0$ zum Irrtumsniveau α.

Analog definiert

$$\left\{ \boldsymbol{c}^T \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} - \tau_0 \begin{matrix} > \\ < \end{matrix} t_{n-s,1-\alpha} \sqrt{\widehat{\sigma}^2 \boldsymbol{c}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{c}} \right\}$$

einen (einseitigen) Test von $H_0 : \boldsymbol{c}^T \boldsymbol{\beta} \begin{matrix} \leq \\ \geq \end{matrix} \tau_0$ *gegen* $H_1 : \boldsymbol{c}^T \boldsymbol{\beta} \begin{matrix} > \\ < \end{matrix} \tau_0$

2. *Sei* $H \subseteq L$ *linearer Unterraum,* $\dim H = r < p$,

$$F_{H,L} := \frac{n-p}{p-r} \frac{\left\| \Pi_L \boldsymbol{y} - \Pi_H \boldsymbol{y} \right\|_2^2}{\left\| \boldsymbol{y} - \Pi_L \boldsymbol{y} \right\|_2^2}.$$

Der Ablehnungsbereich

$$\{ F_{H,L} > f_{s-r,n-s;1-\alpha} \}$$

mit $f_{p-r,n-p;1-\alpha} = (1-\alpha)$*-Quantil der Fisher*$_{p-r,n-p}$*-Verteilung definiert einen Test von* $H_0 : \boldsymbol{X}\boldsymbol{\beta} \in H$ *gegen* $H_1 : \boldsymbol{X}\boldsymbol{\beta} \notin H$ *zum Niveau* α.

3. *Sei* $v_0 > 0$. *Der Ablehnungsbereich*

$$\left\{ (n-p)\widehat{\sigma}^2 \begin{matrix} > \\ < \end{matrix} v_0 \chi^2_{n-p,1-\alpha} \right\}$$

definiert einen Test von $H_0 : \sigma^2 \begin{matrix} \leq \\ \geq \end{matrix} v_0$ *gegen* $H_1 : \sigma^2 \begin{matrix} > \\ < \end{matrix} v_0$ *zum Niveau* α.

Beweis. Wir betrachten exemplarisch (ii): Für jedes $(\boldsymbol{\beta}, \sigma^2) \in H_0$ ist $F_{H,L}$ unter $\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}$ Fisher$_{p-r,n-p}$-verteilt, demnach gilt

$$\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)} \left(F_{H,L} > f_{s-r,n-s,1-\alpha} \right) = \alpha$$

□

Beispiel: Größen von Vätern und Söhnen

Den Begriff „Regression“ hat Francis Galton[2] geprägt, wir betrachten dazu ein Beispiel, siehe die Folien `Pearsons_Vaeter_Soehne.pdf`.

Der (klassische) Datensatz ist in einem `R`-Paket aufbereitet:

```
# K. Pearsons Daten ueber die Groesse von 1078
# Vater-Sohn-Paaren (jew. in Inches), zitiert nach dem
# R-Paket UsingR, data(father.son)
> ?father.son

father.son                package:UsingR                R Documentation

Pearson's data set on heights of fathers and their sons
```

[2] Francis Galton (1822–1911, engl. Wissenschaftler); siehe auch den Artikel von Robert Langkjaer-Bain, The troubling legacy of Francis Galton, Significance 16, Issue 3, June 2019

```
Description:

     1078 measurements of a father's height and his son's height.

Usage:

     data(father.son)

Format:

     A data frame with 1078 observations on the following 2 variables.

     fheight Father's height in inches

     sheight Son's height in inches

Details:

     Data set used by Pearson to investigate regression. See data set
     'galton' for data set used by Galton.

Source:

     Read into R by the command

     'read.table("http://stat-www.berkeley.edu/users/juliab/141C/pearson.dat",sep="
     ")[,-1]',

     as mentioned by Chuck Cleland on the r-help mailing list.
```

Bemerkung („Rückkehr zum Mittelwert“ im Licht der zweidimensionalen Normalverteilung)**.** Sei (X,Y) bivariat normal mit $\mathbb{E}[X] = \mu_X$, $\mathrm{Var}[X] = \sigma_X^2$, $\mathbb{E}[Y] = \mu_Y$, $\mathrm{Var}[Y] = \sigma_Y^2$, $\mathrm{Cov}[X,Y] = \rho\sigma_X\sigma_Y$ (mit Korrelationskoeffizient $\rho \in (-1,1)$), dann hat (X,Y) die gemeinsame Dichte

$$\varphi_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\Big(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\Big)\right)$$

und X hat Dichte

$$\varphi_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\Big)$$

also hat Y, bedingt auf $X = x$ die Dichte (vgl. auch Bem. A.1.8)

$$\frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}} \exp\Big(-\frac{\big(y-\mu_Y-\rho(\sigma_Y/\sigma_X)(x-\mu_X)\big)^2}{2\sigma_Y^2(1-\rho^2)}\Big)$$

d.h. gegeben $X = x$ ist Y normalverteilt mit

$$\mathbb{E}[Y \mid X = x] = \mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X), \quad \mathrm{Var}[Y \mid X = x] = \sigma_Y^2(1 - \rho^2)$$

Insbesondere ist $\mathbb{E}[Y \mid X = x] - \mu_Y = \rho(\sigma_Y/\sigma_X)(x - \mu_X)$, für $0 < \rho < 1$ und $\sigma_X \approx \sigma_Y$ sehen wir also: $\mathbb{E}[Y \mid X = x] - \mu_Y$ hat dasselbe Vorzeichen wie $x - \mu_X$, aber um den Faktor ρ kleineren Betrag (was die von Galton beobachtete „Rückkehr zum Mittelwert“ zeigt).

Ergänzung (weitere Korrelationsmaße)**.** 1. Gegeben n Beobachtungs-/Wertepaare (x_i, y_i), $r_{x,i}$ der Rang des i-ten x-Werts (wir „teilen“ die Ränge im Fall von Bindungen, d.h. $r_{x,i} = \#\{j \neq i : x_j < x_i\} + 1/\big(\#\{j \neq i : x_j = x_i\}\big)$), analog $r_{y,i}$ der Rang des i-ten y-Werts. Dann ist *Spearmans Rangkorrelationskoeffizient*

$$\rho_S = \frac{\sum_{i=1}^n (r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)}{\big(\sum_{i=1}^n (r_{x,i} - \overline{r}_x)^2\big)^{1/2}\big(\sum_{i=1}^n (r_{y,i} - \overline{r}_y)^2\big)^{1/2}}$$

($\overline{r}_x = n^{-1}\sum_{i=1}^n r_{x,i} = (n+1)/2$ ist der „mittlere x-Rang“ und ebenso $\overline{r}_x = (n+1)/2$). Zumindest im Fall ohne Bindungen ist $\frac{1}{n}\sum_{i=1}^n (r_{x,i} - \overline{r}_x)^2 = \frac{1}{n}\sum_{i=1}^n r_{x,i}^2 - \big(\frac{n+1}{2}\big)^2 = \frac{1}{n}\sum_{i=1}^n i^2 - \big(\frac{n+1}{2}\big)^2 = \frac{n^2-1}{12}$ [dies ist die Varianz der uniformen Verteilung auf $\{1, 2, \ldots, n\}$] und damit

$$\begin{aligned}\rho_S &= \frac{\frac{1}{n}\sum_{i=1}^n (r_{x,i} - \overline{r}_x)(r_{y,i} - \overline{r}_y)}{\big(\frac{1}{n}\sum_{i=1}^n (r_{x,i} - \overline{r}_x)^2\big)^{1/2}\big(\frac{1}{n}\sum_{i=1}^n (r_{y,i} - \overline{r}_y)^2\big)^{1/2}} \\ &= \frac{12}{n(n^2-1)}\sum_{i=1}^n r_{x,i} r_{y,i} - \frac{12}{(n^2-1)}\Big(\frac{n+1}{2}\Big)^2 = \frac{12}{n(n^2-1)}\sum_{i=1}^n r_{x,i} r_{y,i} - \frac{3(n+1)}{n-1}\end{aligned}$$

Demnach ist für Testzwecke ρ_S äquivalent zu $\sum_{i=1}^n r_{x,i} r_{y,i}$.

Wenn wir annehmen, dass die x-Werte und die y-Werte n Realisierungen des zufälligen Paars (X, Y), dessen Koordinaten unabhängig sind mit stetiger Verteilung, so ist

$$\sum_{i=1}^n r_{x,i} r_{y,i} \overset{d}{=} \sum_{i=1}^n i S_i$$

mit $(S_1, S_2, \ldots, S_n) \sim \mathrm{Unif}(\mathcal{S}_n)$ einer uniformen Permutation von $1, 2, \ldots, n$: Sortiere im Geiste die Paare gemäß wachsendem x-Rang, dann bilden unter der Unabhängigkeitsannahme die zugehörigen y-Ränge eine rein zufällige Permutation; man kann daraus kritische Werte konstruieren, für kleines n durch explizites Abzählen, für größere n durch Approximation (vgl. `R`-Hilfe zu `cor.test`).

`R` führt dies mittels `cor.test(x,y,method='spearman')` aus.

```
# Aus Hartung-Elpelt, Kap. III.2.1, Tabelle 11:
x <- c(19,12,18,16,26,15,27,23,20,21,19,15,17,15,21,16,23,17,14,18,17,19)
y <- c(103,119,124,133,155,112,108,103,90,114,120,100,109,112,157,118,113,94,106,109,141
cbind(x,y)
plot(x,y)
> cor.test(x,y,method='spearman')

Spearman's rank correlation rho
```

```
data:  x and y
S = 1610.1, p-value = 0.6877
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0908323

Warnmeldung:
In cor.test.default(x, y, method = "spearman") :
  Kann exakten p-Wert bei Bindungen nicht berechnen
```

2. Ähnlich definiert man Kendalls τ (Kendall'scher Rangkorrelationskoeffizient)

$$\tau = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \operatorname{sign}(x_j - x_i)\operatorname{sign}(y_j - y_i) = \frac{\text{Anz. konkordante Paare} - \text{Anz. diskordante Paare}}{\binom{n}{2}}$$

In `R`: `cor.test(x,y,method='kendall')` . Zu Variationen, Umgang mit Bindungen, etc., siehe auch [HE07, Kap. III.2.2].

2.1 Beispiel Varianzanalyse

Gegeben $n = n_1 + n_2 + \cdots + n_s$ Beobachtungswerte in s Gruppen der Größen $n_1, n_2, \ldots, n_s$, wir nehmen an

$$y_{i,k} = \beta_i + \varepsilon_{i,k}, \quad i = k, \ldots, n_i, \;\; i = 1, \ldots, s$$

mit $\beta_i \in \mathbb{R}$ (dem mittleren Effekt in der k-ten Gruppe) und $\varepsilon_{i,k}$ u.i.v. $\sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. Formulierung im Rahmen des linearen Modells:

$$\boldsymbol{Y} = (Y_{1,1}, Y_{1,2}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots, Y_{2,n_2}, \ldots, Y_{s,1}, \ldots, Y_{s,n_s})^T, \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_2)^T \in \mathbb{R}^s$$

mit Designmatrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \ldots & \vdots \\ 0 & 1 & \ldots & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & 1 & & \vdots \\ \vdots & 0 & \ldots & \vdots \\ \vdots & \vdots & & 0 \\ \vdots & \vdots & & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 1 \end{pmatrix}$$

wobei in der ersten Spalte n_1 viele Einsen stehen, in Spalte 2 n_2 viele Einsen stehen, usw.; damit ist

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$$

mit $\varepsilon \sim \mathcal{N}(0,1)^{\otimes n}$. Der Spaltenraum $L \subset \mathbb{R}^n$ von $\boldsymbol{X}$, ist offenbar s-dimensional.

Man nennt die s verschiedenen „Gruppenindizes“ $1, 2, \ldots, s$ hier auch „Faktorstufen“ oder „Levels“ – die Gruppenzugehörigkeit ist eine „kategorielle Eigenschaft“ oder „Faktor“.

Es ist

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & n_s \end{pmatrix}$$

und

$$\boldsymbol{X}^T\boldsymbol{Y} = (n_1\overline{Y}_{1,\bullet}, \ldots, n_s\overline{Y}_{s,\bullet})^T$$

mit $\overline{Y}_{i,\bullet} = \frac{1}{n_i}\sum_{k=1}^{n_i}\overline{Y}_{i,k}$ dem Mittelwert in der i-ten Gruppe und somit

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} \overline{Y}_{1,\bullet} \\ \overline{Y}_{2,\bullet} \\ \vdots \\ \overline{Y}_{s,\bullet} \end{pmatrix}$$

sowie

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n-s}\left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right\|_2^2 = \frac{1}{n-s}\Big\|\boldsymbol{Y} - (\underbrace{\overline{Y}_{1,\bullet}, \ldots, \overline{Y}_{1,\bullet}}_{n_1}, \underbrace{\overline{Y}_{2,\bullet}, \ldots, \overline{Y}_{2,\bullet}}_{n_2}, \ldots, \underbrace{\overline{Y}_{s,\bullet}, \ldots, \overline{Y}_{s,\bullet}}_{n_s})^T\Big\|_2^2 \\ &= \frac{1}{n-s}\sum_{i=1}^{s}\sum_{k=1}^{n_i}(X_{i,k} - \overline{Y}_{i,\bullet})^2 = \sum_{i=1}^{s}\frac{n_i - 1}{n-s}\widehat{\sigma}_i^2\end{aligned}$$

mit $\widehat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(X_{i,k} - \overline{Y}_{i,\bullet})^2$ der (korrigierten Stichproben-)Varianz innerhalb der i-ten Gruppe.

Bemerkung (Varianzzerlegung im Kontext der einfaktoriellen Varianzanalyse)**.** Mit $\overline{Y}_{\bullet,\bullet} := \frac{1}{n}\sum_{i=1}^{s}\sum_{k=1}^{n_i}X_{i,k}$ (dem „Globalmittelwert“) ist

$$\begin{aligned}\widehat{\sigma}_{\mathrm{tot}}^2 &:= \frac{1}{n-1}\sum_{i=1}^{s}\sum_{k=1}^{n_i}(Y_{i,k} - \overline{Y}_{\bullet,\bullet})^2 = \frac{1}{n-1}\left\|\boldsymbol{Y} - \overline{Y}_{\bullet,\bullet}\cdot\mathbf{1}\right\|_2^2 \\ &= \frac{1}{n-1}\Big(\left\|\boldsymbol{Y} - \Pi_L\boldsymbol{Y}\right\|_2^2 + \left\|\Pi_L\boldsymbol{Y} - \overline{Y}_{\bullet,\bullet}\cdot\mathbf{1}\right\|_2^2\Big),\end{aligned}$$

also

$$(n-1)\widehat{\sigma}_{\mathrm{tot}}^2 = (n-s)\widehat{\sigma}_{\mathrm{iG}}^2 + (s-1)\widehat{\sigma}_{\mathrm{zG}}^2$$

mit $\widehat{\sigma}_{\mathrm{zG}}^2 := \frac{1}{s-1}\sum_{i=1}^{s}n_i\big(\overline{Y}_{i,\bullet} - \overline{Y}_{\bullet,\bullet}\big)^2 = \frac{1}{s-1}\left\|\Pi_L\boldsymbol{Y} - \overline{Y}_{\bullet,\bullet}\cdot\mathbf{1}\right\|_2^2$ („Varianz zwischen den Gruppen“) und $\widehat{\sigma}_{\mathrm{iG}}^2 := \widehat{\sigma}^2$ der „Varianz innerhalb der Gruppen“.

Beobachtung (Konfidenzbereich für $\boldsymbol{\beta} = (\beta_1, \dots, \beta_s)^T$)**.**] Für $\alpha \in (0,1)$ ist

$$C := \left\{ \tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \dots, \tilde{\beta}_s)^T : \frac{1}{s} \sum_{i=1}^{s} n_i (\tilde{\beta}_i - \overline{Y}_{i,\cdot})^2 < \widehat{\sigma}^2 \cdot f_{s,n-s;1-\alpha} \right\}$$

(wo $f_{s,n-s;1-\alpha} = (1-\alpha)$-Quantil der Fisher$_{s,n-s}$-Verteilung) ein Konfidenzbereich (oder Konfidenzellipsoid) zum Irrtumsniveau α.

Beweis.

$$\begin{aligned} \mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\big(C \ni \boldsymbol{\beta}\big) &= \mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\Big(\frac{1}{s} \sum_{i=1}^{s} n_i (\beta_i - \overline{Y}_{i,\cdot})^2 < \widehat{\sigma}^2 \cdot f_{s,n-s;1-\alpha}\Big) \\ &= \mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\left(\frac{\frac{1}{s\sigma^2} \sum_{i=1}^{s} n_i (\beta_i - \overline{Y}_{i,\cdot})^2}{\frac{1}{\sigma^2}\widehat{\sigma}^2} < f_{s,n-s;1-\alpha} \right) = 1 - \alpha \end{aligned}$$

□

Beobachtung (F-Test auf Gleichheit der Gruppenmittelwerte)**.** $H = \{m \cdot \mathbf{1} : m \in \mathbb{R}\} \subset L(\subset \mathbb{R}^n)$, (d.h. $\boldsymbol{X}\boldsymbol{\beta} \in H \iff \beta_1 = \beta_2 = \dots = \beta_s$). Der Ablehnungsbereich

$$\left\{ \widehat{\sigma}^2_{\mathrm{zG}} > f_{s-1,n-s;1-\alpha} \widehat{\sigma}^2_{\mathrm{iG}} \right\}$$

definiert einen Test von $H_0 : \beta_1 = \dots = \beta_s$ gegen H_1 : "nicht alle β_i sind gleich" zum Niveau α.

Beweis. Sei $\sigma^2 > 0$, $\boldsymbol{\beta} \in \mathbb{R}^s$ mit $\in \boldsymbol{X}\boldsymbol{\beta} \in H$. Dann ist unter $\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}$

$$\frac{\widehat{\sigma}^2_{\mathrm{zG}}}{\widehat{\sigma}^2_{\mathrm{iG}}} = \frac{\frac{1}{\sigma^2}\widehat{\sigma}^2_{\mathrm{zG}}}{\frac{1}{\sigma^2}\widehat{\sigma}^2_{\mathrm{iG}}} \sim \mathrm{Fisher}_{s-1,n-s}$$

□

Beispiel. Für ein Beispiel zur Varianzanalyse siehe Folien `Beispiel_zur_Varianzanalyse.pdf` und R-Code `R-Befehle_ANOVA.R`.

2.2 Zum Problem des multiplen Testens

Beobachtung (Simultane Konfidenzintervalle für die paarweisen Differenzen $\beta_j - \beta_i$ der Gruppenmittelwerte im Kontext der Varianzanalyse)**.** Im Fall gleicher Gruppengrößen $n_1 = n_2 = \dots = n_s$ kann man folgendermaßen für alle $s(s-1)/2$ Differenzen $\beta_i - \beta_j$ gleichzeitig gültige Konfidenzintervalle konstruieren:

Seien $V_1, \dots, V_k$ u.i.v. $\sim \mathcal{N}(0,1)$, davon u.a. $mW^2 \sim \chi^2(m)$, so ist die „studentisierte Spannweitenverteilung" mit Parametern k und m (engl. "studentized range distribution") die Verteilung von

$$\frac{\max_{1 \le i \le k} V_i - \min_{1 \le i \le k} V_i}{W}$$

[man kann prinzipiell deren Dichte bestimmen, `R` kennt `TukeyHSD` und `[p|q]tukey`].

Mit $q = 1-\alpha$-Quantil der studentisierten Spannweitenverteilung mit Parametern s und $n-s$ $(= s(n_1-1))$ ist für beliebiges $\boldsymbol{\beta} \in \mathbb{R}^s$, $\sigma^2 > 0$

$$\mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\left(\overline{Y}_{j,\bullet} - \overline{Y}_{i,\bullet} - q\sqrt{\frac{\widehat{\sigma}^2}{n_1}} \le \beta_j - \beta_i \le \overline{Y}_{j,\bullet} - \overline{Y}_{i,\bullet} + q\sqrt{\frac{\widehat{\sigma}^2}{n_1}} \text{ für } 1 \le i < j \le s\right)$$
$$= \ldots = \mathbb{P}_{(\boldsymbol{\beta},\sigma^2)}\Big(\sigma \max_{1\le i<j} |V_j - V_i|/\sqrt{\widehat{\sigma}^2} \le q\Big) = 1-\alpha$$

mit $V_i = \sqrt{n_1}(\overline{Y}_{i,\bullet} - \beta_i)/\sigma$.

Angesichts der (abstrakten) Äquivalenz von Konfidenzintervallen und zweiseitigen Tests hat man damit auch eine Familie von Tests für die $s(s-1)/2$ Nullhypothesen $H_{0;(i,j)} : \beta_i = \beta_j$, $1 \le i < j \le s$ gegen $H_{1;(i,j)} : \beta_i \ne \beta_j$:

$$\varphi_{(i,j)} = \mathbf{1}_{\left\{|\overline{Y}_{j,\bullet} - \overline{Y}_{i,\bullet}| > q\sqrt{\frac{\widehat{\sigma}^2}{n_1}}\right\}}$$

und diese Familie hält das *simultane Signifikanzniveau* α ein (engl.: family-wise error rate α): Die Wahrscheinlichkeit, irgendeine der Nullhypothesen zu Unrecht abzulehnen, ist $\le \alpha$.

Zwei allgemeine Korrekturverfahren für multiple Tests

Eine ganz allgemeine Korrektur für multiples Testen ist die Bonferroni[3]-Methode:

Beobachtung (Bonferroni-Korrektur)**.** Wenn m Tests zum *multiplen Signifikanzniveau* $\alpha \in (0,1)$ durchgeführt werden sollen,

so führe jeden Test für sich zum *lokalen* Signifikanzniveau $\dfrac{\alpha}{m}$ durch.

Dann gilt: Die Wahrscheinlichkeit, dass *irgendeine zutreffende* Nullhypothese zu Unrecht ablehnt wird, beträgt höchstens α.

Alternativ bedeutet dies: Multipliziere jeden (individuellen) p-Wert mit der Anzahl m der durchgeführten Tests (denn wenn die jeweilige Nullhypthese zutrifft, so ist der p-Wert uniform verteilt in $[0,1]$).

Beweis. Sei $\alpha \in (0,1)$, es seien m Nullhypothesen $H_{0,1}, H_{0,2}, \ldots, H_{0,m}$ und m Tests $\mathcal{T}_1, \ldots, \mathcal{T}_m$ gegeben (formal: betrachte ein statistisches Modell $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, $\mathcal{T}_i$ sei ein Test von $H_{0,i} : \vartheta \in \Theta_{0,i}$ gegen $H_{0,i} : \vartheta \in \Theta \setminus \Theta_{0,i}$) mit

$$\forall\, \vartheta \in \Theta_{0,i} : \ \mathbb{P}_\vartheta(\mathcal{T}_i \text{ lehnt } H_{0,i} \text{ ab}) \le \frac{\alpha}{m} \quad i = 1, \ldots, m$$

(d.h. wenn $H_{0,i}$ [und ggfs. noch irgendwelche anderen $H_{0,j}$] zutrifft, so wird sie von $\mathcal{T}_i$ nur mit W'keit $\le \alpha/m$ zu Unrecht abgelehnt).

Eine gewisse Teilmenge $W \subset \{1, 2, \ldots, m\}$ der Nullhypothesen sei wahr. Dann ist für $\vartheta \in \cap_{j\in W}\Theta_{0,j}$

$$\mathbb{P}_\vartheta\Big(\text{es gibt ein } j \in W, \text{ so dass } H_{0,j} \text{ von } \mathcal{T}_j \text{ abgelehnt wird}\Big)$$
$$\le \sum_{j\in W} \mathbb{P}_\vartheta(\mathcal{T}_j \text{ lehnt } H_{0,j} \text{ ab}) \le \sum_{j\in W} \frac{\alpha}{m} = |W|\frac{\alpha}{m} \le \alpha.$$

□

[3] Carlo Emilio Bonferroni, 1892–1960

Die Bonferroni-Methode ist sehr *konservativ*, d.h. um auf der sicheren Seite zu sein, lässt man sich lieber die eine oder andere Signifikanz entgehen. Eine Verbesserung der Bonferroni-Methode ist die Bonferroni-Holm-Methode:

Beobachtung. Ist m die Anzahl der Tests, so multipliziere den kleinsten p-Wert mit m, den zweitkleinsten mit $m-1$, den drittkleinsten mit $m-2$ usw., lehne all die Nullhypothesen ab,

$$\text{deren so korrigierter } p\text{-Wert} < \alpha \text{ ist.}$$

Dies ist ein Test aller m Nullhypothesen gleichzeitig zum multiplen Signifikanzniveau α.

Beweis. Gegeben m Nullhypothesen $H_{0,1}, H_{0,2}, \ldots, H_{0,m}$ und m Tests $\mathcal{T}_1, \ldots, \mathcal{T}_m$, P_i sei der p-Wert aus dem i-ten Test.

Nach Voraussetzung ist $\mathcal{T}_i$ ein gültiger Test für $H_{0,i} : \vartheta \in \Theta_{0,i}$, d.h. für $\vartheta \in \Theta_{0,i}$ gilt $\mathbb{P}_\vartheta(P_i \le u) \le u$ für $u \in [0,1]$.

Seien

$$P_{(1)} < P_{(2)} < \cdots < P_{(m)}$$

die der Größe nach sortierten p-Werte und $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(m)}$ die entsprechend umsortierten Hypothesen, $\alpha \in (0,1)$.

Wenn

$$mP_{(1)}, (m-1)P_{(2)}, \ldots, (m-\ell-1)P_{(\ell)} < \alpha \le (m-\ell)P_{(\ell+1)}$$

gilt, so lehne $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(\ell)}$ (zum multiplen Niveau α) ab und behalte $H_{0,(\ell+1)}, \ldots, H_{0,(m)}$ bei.

Sei $W \subset \{1, \ldots, m\}$ (mit $|W| = k$, sagen wir) und die Nullhypothesen $H_{0,i}$, $i \in W$ seinen wahr.

Es gilt

$$\bigcap_{i \in W} \left\{P_i > \tfrac{\alpha}{k}\right\} \subset \left\{P_{(\,m-(k-1)\,)} > \tfrac{\alpha}{k}\right\}$$

(denn dann kann es höchstens $m-k$ viele p-Werte $\le \alpha/k$ geben), insbesondere stoppt das Verfahren dann in Schritt ℓ mit $\ell \le m-k+1$ und alle $H_{0,i}$, $i \in W$ werden akzeptiert.

Weiter ist für $\vartheta \in \cap_{j \in W} \Theta_{0,j}$

$$\begin{aligned}\mathbb{P}_\vartheta\Big(\bigcap_{i \in W} \left\{P_i > \tfrac{\alpha}{k}\right\}\Big) &= 1 - \mathbb{P}_\vartheta\Big(\bigcup_{i \in W} \left\{P_i \le \tfrac{\alpha}{k}\right\}\Big) \\ &\ge 1 - \sum_{i \in W} \mathbb{P}_\vartheta\big(P_i \le \tfrac{\alpha}{k}\big) \ge 1 - \sum_{i \in W} \frac{\alpha}{k} = 1 - \alpha.\end{aligned}$$

□

Übrigens: In `R` gibt des den Befehl `p.adjust`, der p-Werte für multiples Testen korrigiert und dabei defaultmäßig die Bonferroni-Holm-Korrektur verwendet.

```
> pwerte <- c(0.01470, 0.00024, 0.16689, 1.00000, 0.00509, 0.00010)
> pwerte
[1] 0.01470 0.00024 0.16689 1.00000 0.00509 0.00010
```

```
> p.adjust(pwerte)
[1] 0.04410 0.00120 0.33378 1.00000 0.02036 0.00060
> p.adjust(pwerte, method="bonferroni")
[1] 0.08820 0.00144 1.00000 1.00000 0.03054 0.00060
```

Bemerkung. Für ein Beispiel, wie man durch Verschweigen eines multiplen Testproblems „Unsinn“ erzeugen kann, siehe die Folien `ein_Stoppproblem.pdf`.

2.3 Zur Hauptkomponentenanalyse

Hauptkomponentenanalyse ist eigentlich, so wie wir es hier ansehen, ein Verfahren der deskriptiven Statistik, sie passt aber thematisch hierher (Bezug zu Varianzzerlegung und zur multivariaten Normalverteilung).

Sei $\boldsymbol{Y} = (y_{i,j}) \in \mathbb{R}^{n\times p}$, wir fassen die n Zeilen als n Beobachtungspunkte $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ im $\mathbb{R}^p$ auf.

Frage: In welcher p-dimensionalen Richtung variieren die $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ (die Zeilen von $\boldsymbol{Y}$) am stärksten?

Sei dazu $\overline{y}_j := \frac{1}{n}\sum_{i=1}^{n} y_{i,j}$ das empirische Mittel der j-ten Spalte und

$$\widetilde{\boldsymbol{Y}} = (\widetilde{y}_{i,j})_{i=1,\dots,n;\, j=1,\dots,p} \quad \text{mit} \quad \widetilde{y}_{i,j} = y_{i,j} - \overline{y}_j$$

die Matrix der (spaltenweise) zentrierten Beobachtungungen.

Für $\boldsymbol{v} = (v_1, \dots, v_p)^T \in \mathbb{R}^p$ mit $\|v\|_2 = 1$ ist

$$\widetilde{\boldsymbol{Y}} \cdot \boldsymbol{v} = \Big(\sum_{j=1}^{p} \widetilde{y}_{i,j} v_j \Big)_{i=1}^{n} \in \mathbb{R}^n$$

der (n-dimensionale) Vektor der Projektionen von $\widetilde{\boldsymbol{y}}_1, \dots, \widetilde{\boldsymbol{y}}_n$ in Richtung $\boldsymbol{v}$ und dessen (empirische oder Stichproben-)Varianz ist

$$\frac{1}{n-1}\sum_{i=1}^{n} \Big(\sum_{j=1}^{p} \widetilde{y}_{i,j} v_j \Big)^2 = \frac{1}{n-1} \|\widetilde{\boldsymbol{Y}} \cdot \boldsymbol{v}\|_2^2$$

Demnach:

Gesucht: $\boldsymbol{v} \in \mathbb{R}^p$ mit $\|\boldsymbol{v}\|_2 = 1$, so dass

$$\|\widetilde{\boldsymbol{Y}} \cdot \boldsymbol{v}\|_2^2 = (\widetilde{\boldsymbol{Y}} \cdot \boldsymbol{v})^T \cdot (\widetilde{\boldsymbol{Y}} \cdot \boldsymbol{v}) = \boldsymbol{v}^T (\widetilde{\boldsymbol{Y}}^T \widetilde{\boldsymbol{Y}}) \boldsymbol{v} \stackrel{!}{=} \max \tag{2.12}$$

$$\boldsymbol{S} := \frac{1}{n-1} \widetilde{\boldsymbol{Y}}^T \widetilde{\boldsymbol{Y}}$$

ist die (empirische) Kovarianzmatrix von $\boldsymbol{Y}$, $\boldsymbol{S}$ ist symmetrisch und positiv (semi-)definit (denn für $\boldsymbol{v} \in \mathbb{R}^p$ ist $\boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v} = (n-1)^{-1} \boldsymbol{v}^T \widetilde{\boldsymbol{Y}}^T \widetilde{\boldsymbol{Y}} \boldsymbol{v} = \|\widetilde{\boldsymbol{Y}} \boldsymbol{v}\|_2^2/(n-1) \geq 0$, demnach hat $\boldsymbol{S}$ ist diagonalisierbar mit reellen Eigenwerten $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$:

$$\boldsymbol{S} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^T$$

mit

$$\boldsymbol{U} = \begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_p \\ | & | & & | \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & \cdots & & 0 & \lambda_p \end{pmatrix}$$

mit $\boldsymbol{U}$ orthogonal (die Spalten $\boldsymbol{u}_j$ von $\boldsymbol{U}$ sind die Eigenvektoren von $\boldsymbol{S}$, $\boldsymbol{S}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$.

Schreibe $\boldsymbol{v} = \sum_{j=1}^p a_j \boldsymbol{u}_j$ mit $\sum_{i=1}^p a_j^2 = 1$, somit ist

$$\frac{1}{n-1}\left\|\widetilde{\boldsymbol{Y}} \cdot \boldsymbol{v}\right\|_2^2 = \boldsymbol{v}^T \boldsymbol{S} \boldsymbol{v} = (\boldsymbol{U}^T \boldsymbol{v})^T \boldsymbol{\Lambda} (\boldsymbol{U}^T \boldsymbol{v}) = \Big(\sum_{j=1}^p a_j \boldsymbol{u}_j\Big)^T \boldsymbol{\Lambda} \Big(\sum_{k=1}^p a_j \boldsymbol{u}_k\Big)$$
$$= \sum_{j,k=1}^p a_j a_k \lambda_k \boldsymbol{u}_j^T \boldsymbol{u}_k = \sum_{k=1}^n a_k^2 \lambda_k$$

und demnach löst $\boldsymbol{v} = \boldsymbol{u}_1$ die Optimierungsaufgabe (2.12).

Iterieren wir:

Gesucht: $\boldsymbol{w} \in \mathbb{R}^p$ mit $\|\boldsymbol{w}\|_2 = 1$, so dass

$$\|\widetilde{\boldsymbol{Y}}\boldsymbol{w}\|_2^2 \stackrel{!}{=} \max \quad \text{und} \quad \boldsymbol{w} \perp \boldsymbol{u}_1$$

Analoge Rechnung zeigt: Die Wahl $\boldsymbol{w} = \boldsymbol{u}_2$ ist optimal, etc.

Man nennt für $j = 1, \dots, p$

$$\boldsymbol{z}_j = \boldsymbol{Y}\boldsymbol{u}_j \ (\in \mathbb{R}^n)$$

die j-te Hauptkomponente (von $\boldsymbol{Y}$ bzw. von $\boldsymbol{S}$) und $\boldsymbol{u}_j$ den Vektor der Gewichte der j-ten Hauptkomponente.

Beispiel. (Siehe auch R-Code `Beispiel_zur_PCA.R`).

```
# ein Beispiel zur Hauptkomponentenanalyse

data("iris")
?iris
iris
# Bem.: sepal = Kelchblatt, petal = Kronblatt/Bluetenblatt

Y <- iris[,1:4] # wir lassen die Artklassifikation weg (ist nicht-numerisch)
Y
pairs(Y)
pairs(Y,col=iris$Species)
hka <- prcomp(Y, center=TRUE)
hka
plot(hka, type='l')
summary(hka)
biplot(hka)
```

```
# Fuer bessere Lesbarkeit nochmal, mit Zentrierung und Standardisierung:
hka <- prcomp(Y, center=TRUE, scale=TRUE)
hka
plot(hka, type='l')
summary(hka)
biplot(hka)

plot(hka$x[,1],hka$x[,2],col=iris$Species, pch=20, xlab="Hauptkomponente 1", ylab="Haupt
```

Kapitel 3

Etwas nicht-parametrische Statistik

3.1 Der Wilcoxon(-Mann-Whitney)-Rangsummentest

Wilcoxons Rangsummen-Test Sei $n = k + l, X_1, \ldots, X_k$ u.i.v. $\sim F_1$ und $X_{k+1}, \ldots, X_{k+l}$ u.i.v. $\sim F_1$ mit F_2, F_2 Wahrscheinlichkeitsmaßen auf $\mathbb{R}$ mit stetiger Verteilungsfunktion.

Wie möchten prüfen, ob unter F_1 „typischerweise kleinere Werte angenommen werden“ als unter F_2.

Sei für $1 \leq i \leq n$

$$R_i = \left|\{1 \leq j \leq n : X_j \leq X_i\}\right| \quad \text{der Rang der } i\text{-ten Beobachtung}$$

und

$$W := \sum_{i=1}^{k} R_i \quad \text{die Wilcoxon-Rangsummen-Statistik}$$

Beobachtung. Es gilt

1. $\sum_{i=1}^{n} R_i = \sum_{i1}^{n} i = \frac{n(n+1)}{2}$
2. $W = U + \frac{k(k+1)}{2}$ mit

$$U := \sum_{i=1}^{k} \sum_{j=k+1}^{k+l} \mathbf{1}_{\{X_i > X_j\}} \quad \text{der Mann-Whitney-}U\text{-Statistik} \tag{3.1}$$

denn wir können O.E. annehmen, dass $X_1 < \ldots < X_k$ (da U, W invariant unter Permutation von $X_1, \ldots, X_k$), und dann ist $R_i = i + \sum_{j=k+1}^{k+l} \mathbf{1}_{\{X_i > X_j\}}$.

Der Wilcoxon-Test heißt auch Mann-Whitney-Test[1]. In der Literatur sind verschiedene Zentrierungen der Rangsumme gebräuchlich, verwendete Notationskonvention ggfs. prüfen, bevor man eine Formel verwendet.

[1] siehe Frank Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1(6):80–83, (1945) und Henry Mann, Donald Whitney, On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics. 18:50–60, (1947)

Beobachtung. Sei F_1 stetige Verteilung auf $\mathbb{R}$, $n = k + l, X_1, \ldots, X_n$ u.i.v. gemäß F_1, U wie in (3.1) oben. Dann gilt für $m \in \{0, 1, \ldots, k \cdot l\}$:

$$\mathbb{P}(U = m) = \frac{1}{\binom{n}{k}} \cdot N(m; k, l) \tag{3.2}$$

mit

$$N(m; k, l) = \#\Big\{(m_1, \ldots, m_k) \in \{0, 1, \ldots, l\}^k : m_1 \le \ldots \le m_k, \textstyle\sum_{i=1}^k m_i = m\Big\}$$

U ist also „verteilungsfrei“ in dem Sinn, dass seine Verteilung nicht von F_1 abhängt (die Gewichte sind „rein kombinatorische“ Terme). Man kann die kombinatorischen Größe

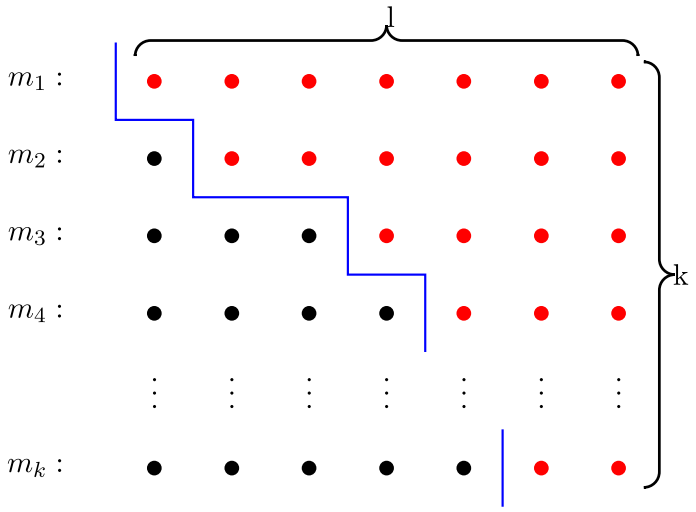


Abbildung 3.1: Geometrische Interpretation von $N(m; k, l)$

$N(m; k, l)$ interpretieren (und visualiseren) als die Anzahl der Wege (von „links oben“ nach „rechts unten“) durch ein $k \times l$-Gitter, so dass „unter“ dem Weg genau $m = \sum m_i$ viele Punkte liegen (siehe Abb. 3.1).

Es gilt

$$N(m; k, l) = N(m; l, k) = N(kl - m; k, l) \text{ und } N(m; k, l) = \sum_{j=0}^{k} N(m - j; j, l - 1)$$

Für die erste Gleichung vertausche Zeilen und Spalten in Abb. 3.1, für die zweite Gleichung vertausche die Rollen von „rot“ und „schwarz“ (und dann Zeilen und Spalten) in Abb. 3.1, die dritte Gleichung ergibt sich durch Zerlegung nach der Anzahl schwarzer Punkte in der ersten Spalte.

Beweis von (3.2). $(R_1, \ldots, R_n)$ ist unter $\mathbb{P}^{\otimes n}$ uniform verteilt auf allen Permutationen von $\{1, \ldots, n\}$, denn

$$\mathbb{P}\Big(X_{\sigma(1)} < X_{\sigma(2)} < \ldots < X_{\sigma(n)}\Big) = \frac{1}{n!}$$

Demnach ist $\{R_1, \ldots, R_k\}$ uniform verteilt auf den k-elementigen Teilmengen von $\{1, \ldots, n\}$. Parametrisiere $A \subseteq \{1, \ldots, n\}$ mit $|A| = k$ als $(r_1, \ldots, r_k), 1 \le r_1 < r_2 < \ldots < r_k \le n$. Dann ist

$$(m_1, \ldots, m_k)(A) := (r_1 - 1, r_2 - 2, \ldots, r_k - k) \in \{0, 1, \ldots, l\}^k,$$

$$m_i(A) = |\{s \in \{1,\dots,n\} \setminus A : s < r_i\}|.$$

Somit

$$\mathbb{P}(U = m) = \sum_{\substack{A\subseteq\{1,\dots,n\}\\ |A|=k\\ m_1(A)+\dots+m_k(A)=m}} \underbrace{\mathbb{P}^{\otimes n}[\{R_1,\dots,R_k\} = A]}_{=\frac{1}{\binom{n}{k}}}$$

□

Ohne Einschränkung seien $X_1,\dots,X_n$ unabhängig $\sim \mathrm{Unif}([0,1])$ $(n = k+l)$,

$$U_{k,l} := \sum_{i=1}^{k}\sum_{j=k+1}^{k+l} \mathbf{1}_{X_i>X_j}.$$

Es gilt $\mathbb{E}[U_{k,l}] = \frac{1}{2}kl$,

$$\mathrm{Cov}(\mathbf{1}_{(X_i>X_j)}, \mathbf{1}_{(X_{i'}<X_{j'})}) = \begin{cases} 0 & \text{wenn } i \neq i' \text{ und } j \neq j' \\ \mathbb{P}[X_1 > X_2, X_1 > X_3] - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} & \text{wenn } i = i' \text{ und } j \neq j' \text{ oder } i \neq i' \text{ und } j = \\ \frac{1}{2} - \frac{1}{4} = \frac{1}{4} & \text{wenn } i = i' \text{ und } j = j' \end{cases}$$

und somit

$$\begin{aligned} \mathrm{Var}(U_{k,l}) &= \mathrm{Cov}(U_{k,l}, U_{k,l}) = \sum_{i=1}^{k}\sum_{j=k+1}^{k+l}\sum_{i'=1}^{k}\sum_{j'=k+1}^{k+l} \mathrm{Cov}(\mathbf{1}_{(X_i>X_j)}, \mathbf{1}_{(X_{i'}<X_{j'})}) \\ &= \frac{1}{4}\cdot kl + \frac{1}{12}kl(l-1) + \frac{1}{12}lk(k-1) \\ &= \frac{kl(k+l+1)}{12} =: V_{k,l}. \end{aligned}$$

Proposition.

$$U^*_{k,l} := \frac{U_{k,l} - \frac{1}{2}kl}{\sqrt{V_{k,l}}} \xrightarrow[k,l\to\infty]{\text{in Vert.}} \mathcal{N}(0,1)$$

Beweis. Idee: Approximiere $\mathbf{1}_{X_i>X_j} - \frac{1}{2}$ durch $X_i - X_j$. Definiere weiter

$$Z_{k,l} := \sum_{i=1}^{k}\sum_{j=k+1}^{k+l}(X_i - X_j) = l\sum_{i=1}^{k} X_i - k\sum_{j=k+1}^{k+l} X_j.$$

Dann ist

$$\mathbb{E}[Z_{k,l}] = 0 \text{ und } \mathrm{Var}(Z_{k,l}) = l^2 k\frac{1}{12} + k^2 l\frac{1}{12} = \frac{kl(k+l)}{12}.$$

Sei $Z^*_{k,l} := \dfrac{Z_{k,l}}{\sqrt{V_{k,l}}}$. Wir möchten zeigen, dass $\mathrm{Var}(U^*_{k,l} - Z^*_{k,l}) \underset{k,l\to\infty}{\longrightarrow} 0$:

$$\begin{aligned}
\mathrm{Cov}(U_{k,l}, Z_{k,l}) &= l\sum_{i=1}^{k}\mathrm{Cov}(U_{k,l}, X_i) - k\sum_{j=k+1}^{k+l}\mathrm{Cov}(U_{k,l}, X_j),\\
\mathrm{Cov}(U_{k,l}, X_i) &= \sum_{i'=1}^{k}\sum_{j'=k+1}^{k+l}\mathrm{Cov}(\mathbf{1}_{X_{i'}>X_{j'}}, X_i) = l\frac{1}{12},
\end{aligned}$$

$$\text{denn}\quad \mathrm{Cov}(\mathbf{1}_{X_{i'}>X_{j'}}, X_i) = \begin{cases} 0 & \text{für } i \neq i' \\ \mathrm{Cov}(\mathbf{1}_{X_1>X_2}, X_1) & \\ \quad = \displaystyle\int_0^1 \underbrace{\int_0^1 \mathbf{1}_{x_1>x_2} x_1\, dx_2}_{x_1}\, dx_1 - \frac{1}{4} = \tfrac{1}{12} & \text{falls } i = i' \end{cases}$$

$$\mathrm{Cov}(U_{k,l}, X_i) = -\frac{k}{12} \text{ analog.}$$

Also $\mathrm{Cov}(U_{k,l}, Z_{k,l}) = \dfrac{kl(k+l)}{12}$, somit

$$\begin{aligned}
\mathrm{Var}(U_{k,l} - Z_{k,l}) &= \mathrm{Var}(U_{k,l}) - 2\mathrm{Cov}(U_{k,l}, Z_{k,l}) + \mathrm{Var}(Z_{k,l}) = \frac{kl}{12}\\
\text{und}\quad \mathrm{Var}(U^*_{k,l} - Z^*_{k,l}) &= \frac{1}{k+l+1} \underset{k,l\to\infty}{\longrightarrow} 0.
\end{aligned}$$

$$Z^*_{k,l} = \sqrt{\underbrace{\frac{l}{k+l+1}}_{=:a_{k,l}}}\cdot S^*_k - \sqrt{\underbrace{\frac{k}{k+l+1}}_{=:b_{k,l}}} T^*_l$$

mit $S^*_k = \dfrac{1}{\sqrt{\frac{k}{12}}}\sum_{i=1}^{k}(X_i - \frac{1}{2}), T^*_l = \dfrac{1}{\sqrt{\frac{l}{12}}}\sum_{j=k+1}^{k+l}(X_j - \frac{1}{2})$.

Mit dem zentralen Grenzwertsatz gilt

$$(S^*_k, T^*_l) \underset{k,l\to\infty}{\longrightarrow}{}^{d} (S, T), \text{ mit } S, T \text{ unabhängig und } \mathcal{N}(0,1) \text{ verteilt.}$$

Sei $(k_i, l_i)_{i\in\mathbb{N}}$ mit $k_i \to \infty, l_i \to \infty$, dann gibt es eine Teilfolge (k_{i_j}, l_{i_j}), sodass $a_{k_{i_j},l_{i_j}} \to a$ und $b_{k_{i_j},l_{i_j}} \to b$ und $a + b = 1$.

Demnach $Z^*_{k_{i_j},l_{i_j}} \overset{\text{in Vert.}}{\underset{j\to\infty}{\to}} \underbrace{\sqrt{a}S}_{\sim\mathcal{N}(0,a)} + \underbrace{\sqrt{b}T}_{\sim\mathcal{N}(0,b)} \sim \mathcal{N}(0, \underbrace{a+b}_{=1})$. □

Zweiseitiger Wilcoxon-Rangsummen-Test Sei $n = k+l$, $X_1, \ldots, X_k$ u.i.v. $\sim \vartheta_1$, $X_{k+1}, \ldots, X_{k+l}$ u.i.v. $\sim \vartheta_2$ mit $\vartheta_1, \vartheta_2 \in \mathcal{M}_1(\mathbb{R})$ stetige Verteilungen, sei $u_{k,l,1-\alpha/2}$ mit $\mathbb{P}_{H_0}(U_{k,l} \le u_{k,l,1-\alpha/2}) = 1 - \alpha/2$ (aus Tabelle oder mit `R`).

Lehne $H_0 : \vartheta_1 = \vartheta_2$ ab, falls

$$U_{k,l} > u_{k,l,1-\alpha/2} \quad \text{oder} \quad U_{k,l} < k\cdot l - u_{k,l,1-\alpha/2}$$

Erinnerung (Stochastische Dominierung, vgl. auch Seite 32)**.** Seien $\mathbb{P}, \mathbb{Q}$ zwei Wahrscheinlichkeitsmaße auf $\mathbb{R}$, so heißt $\mathbb{P}$ *stochastisch kleiner* als $\mathbb{Q}$, $\mathbb{P} \preccurlyeq \mathbb{Q}$, wenn

$$\mathbb{P}[[c,\infty)] \le \mathbb{Q}[[c,\infty)] \quad \forall c \in \mathbb{R},$$

d.h. für die zugehörigen Verteilungsfunktionen gilt $F_{\mathbb{P}} \ge F_{\mathbb{Q}}$.

$\mathbb{P} \preccurlyeq \mathbb{Q} \iff \exists$ Zufallsvariablen X, Y (auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, \mathbb{P})$) mit $X \le Y$ $\mathbb{P}$-f.s., $\mathbb{P} \circ X^{-1} = \mathbb{P}$ und $\mathbb{P} \circ Y^{-1} = \mathbb{Q}$.

„ $\Leftarrow$ “: $\mathbb{P}[[c,\infty)] = \mathbb{P}[X \ge c] \le \mathbb{P}[Y \ge c] = \mathbb{Q}[[c,\infty)]$
„ $\Rightarrow$ “: Sei $F_{\mathbb{P}}^{-1}(u) := \inf\{x : F_{\mathbb{P}}(x) \ge u\}$ und analog $F_{\mathbb{Q}}^{-1}$. Sei weiter U uniform auf $[0,1]$, $X := F_{\mathbb{P}}^{-1}(U) \le Y := F_{\mathbb{Q}}^{-1}(U)$, denn $F_{\mathbb{P}}^{-1}(u) \le F_{\mathbb{Q}}^{-1}(u)$.

$X \sim \mathbb{P}, c > 0, X + c \sim \mathbb{Q}$, dann ist $\mathbb{P} \preccurlyeq \mathbb{Q}$, insbesondere ist $\mathcal{N}(\mu_1, \sigma^2) \preccurlyeq \mathcal{N}(\mu_2, \sigma^2)$, wenn $\mu_1 \le \mu_2$.

Einseitiger Wilcoxon-Rangsummen-Test Sei $n = k+l$, $X_1, \dots, X_k$ u.i.v. $\sim \vartheta_1$, $X_{k+1}, \dots, X_{k+l}$ u.i.v. $\sim \vartheta_2$ mit $\vartheta_1, \vartheta_2 \in \mathcal{M}_1(\mathbb{R})$ stetige Verteilungen, sei $u_{k,l,1-\alpha/2}$ mit $\mathbb{P}_{H_0}(U_{k,l} \le u_{k,l,1-\alpha}) = 1-\alpha$ (aus Tabelle oder mit `R`).

Lehne $H_0 : \vartheta_1 \preccurlyeq \vartheta_2$ ab, falls

$$U_{k,l} > u_{k,l,1-\alpha}$$

Beispiel (Verkehrstote in Großbritannien 1969–1984)**.** Der mit `R` mitgelieferte Datensatz `Seatbelts`[2] enthält (u.A.) für jeden Monat im Zeitraum Januar 1969 bis Dezember 1984 die Anzahl bei Verkehrsunfällen in Großbritannien getöteter Autofahrer.

```
# Beispiel fuer Rangtest mit R

data(Seatbelts)
Seatbelts
?Seatbelts
plot(Seatbelts[,1], xlab='Zeit', ylab='Tote/Monat',
  main='Anzahl bei Verkehrsunfällen in Großbritannien getöteter Autofahrer')

# Am 31. Januar 1983 wurde in Großbritannien die Gurtpflicht eingeführt,
# wir spalten die Daten entsprechend:

A<-Seatbelts[Seatbelts[,8]==0,1]
B<-Seatbelts[Seatbelts[,8]==1,1]

boxplot(list('Ohne'=A,'Mit'=B))

# einseitiger Test:
```

[2]Basierend auf A. C. Harvey and J. Durbin, The effects of seat belt legislation on British road casualties: A case study in structural time series modelling, *J. Roy. Stat. Soc. B* **149**, 187–227 (1986).

```
wilcox.test(B,A,alt="less")

# beachte: Wegen Bindungen kann R den
# "kombinatorisch exakten" p-Wert nicht bestimmen:
wilcox.test(B,A,alt="less", exact=TRUE)

# und zweiseitig:
wilcox.test(B,A)
```

3.2 Der Kruskal-Wallis-Test

Die einfaktorielle Varianzanalyse basiert auf der Annahme, dass die gemessenen Werte unabhängig und normalverteilt sind. Die Gruppenmittelwerte $\beta_1, \beta_2, \ldots, \beta_s$ können verschieden sein (das herauszufinden ist Ziel des Tests), aber die Varianzen innerhalb der verschiedenen Gruppen müssen gleich sein.

In Formeln: Ist $Y_{i,j}$ die j-te Messung in der i-ten Gruppe, so muss gelten

$$Y_{i,j} = \beta_i + \varepsilon_{i,j},$$

wobei alle $\varepsilon_{i,j}$ unabhängig $\mathcal{N}(0, \sigma^2)$-verteilt sind, mit demselben σ^2 für alle Gruppen.

Die zu testende Nullhypothese ist $\beta_1 = \beta_2 = \cdots = \beta_s$. Falls man Normalitätsannahmen nicht machen will oder kann, so kann man den *Kruskal-Wallis-Test* verwenden, der wie der Wilcoxon-Test die *Ränge* statt der tatsächlichen Werte verwendet. Es handelt sich also um einen *nicht-parameterischen Test*, d.h. es wird keine bestimmte Wahrscheinlichkeitsverteilung vorausgesetzt.

Szenario: $n = n_1 + \cdots + n_s$ Beobachtungen in s Gruppen der Größen $n_1, n_2, \ldots, n_s$,

$$Y_{k,i} = i\text{-te Beabachtung in der } k\text{-ten Gruppe}$$

Nullhypothese des Kruskal-Wallis-Tests: alle Werte $Y_{i,j}$ kommen aus derselben (und wörtlich: stetigen) Verteilung, unabhängig von der Gruppe. Grundvoraussetzung ist auch beim Kruskal-Wallis-Test, dass die Werte unabhängig voneinander sind.

- Sei $R_{i,j}$ der Rang von Y_{ij} innerhalb der Gesamtstichprobe.
- Sei
$$\overline{R}_{i,\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{i,j}$$
der durchschnittliche Rang in Gruppe i, wobei n_i die Anzahl der Messungen in Gruppe i ist.
- Der mittlere Rang der Gesamtstichprobe ist
$$\overline{R}_{\bullet,\bullet} = \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} R_{i,j} = \frac{1}{n} \sum_{k=1}^{n} k = \frac{n+1}{2},$$
wobei s die Anzahl der Gruppen ist und n der Umfang der Gesamtstichprobe.

- Unter der Nullhypothese haben die mittleren Ränge der Gruppen denselben Erwartungswert $\overline{R}_{\bullet,\bullet}$.
- Die Abweichung von dieser Erwartung kann man messen mit der Teststatistik

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{s} n_i \cdot (\overline{R}_{i,\bullet} - \overline{R}_{\bullet,\bullet})^2.$$

  (die Normierung mit $12/(n(n+1))$ wird gewählt, damit H unter der Nullhypothese approximativ $\chi^2(s-1)$-verteilt ist, gelegentlich hört man die „Faustregel“, dass die Approximation „ausreichend“ ist, falls $s \geq 3$ und $n_i \geq 5$ oder falls $s \geq 4$ und $n_i \geq 4$.

  Eine alternative Darstellung ist

$$H = \frac{12}{n \cdot (n+1)} \cdot \left(\sum_{i=1}^{s} n_i \cdot \overline{R}_{i,\bullet}^2 \right) - 3 \cdot (n+1)$$

- Um aus H einen p-Wert zu erhalten, muss man die Verteilung von H unter der Nullhypothese kennen. Diese kann man für verschiedene s und n_i in Tabellen finden.
- Es gibt auch eine „Korrekturformel“ für den Fall von Bindungen (vgl. [BD77, Problem 9.3.2, S. 397]):

$$H^* = \frac{12}{n(n+1)} \sum_{i=1}^{s} n_i \cdot \left(\overline{R}_{i,\bullet}^* - \overline{R}_{\bullet,\bullet}\right)^2 \Big/ \left(1 - \frac{\sum_{i=1}^{d} (b_i^3 - b_i)}{n^3 - n}\right)^2$$

  wobei $\overline{R}_{i,\bullet}^*$ = “average midrank in i-th group”, d = number of different observations in the sample, b_i = number of observations tied with the i-th largest observation value
- `R` kennt den Kruskal-Wallis-Test mittels `kruskal.test`.

3.3 Empirische Verteilungsfunktion und Kolmogorov-Smirnov-Test

Seien $X_1, X_2, \ldots, X_n, \ldots$ u.i.v. reelle ZVn mit Verteilung ϑ ($\in \mathcal{M}_1(\mathbb{R})$), notiere die Verteilungsfunktion als $F_\vartheta(x) = \vartheta((-\infty, x])$. Wir fassen $X_1, \ldots, X_n$ als n Beobachtungen aus einer (uns möglicherweise unbekannten) Verteilung ϑ auf.

Beobachtung (Satz von Glivenko-Cantelli)**.** Die empirische Verteilungsfunktion (von n u.i.v. Beobachtungen)

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}, \quad t \in \mathbb{R} \tag{3.3}$$

erfüllt für jedes $t \in \mathbb{R}$ (gemäß dem Gesetz der großen Zahlen)

$$\lim_{n\to\infty} \widehat{F}_n(t) = \mathbb{E}_\vartheta\left[\mathbf{1}_{\{X_1 \leq t\}}\right] = F_\vartheta(t) \quad \text{fast sicher}$$

Da die $\widehat{F}_n(\cdot)$ monoton wachsend (und beschränkt) sind (und F_ϑ stetig ist), gilt sogar

$$D_n := \sup_{t\in\mathbb{R}} \left|\widehat{F}_n(t) - F_\vartheta(t)\right| \xrightarrow[n\to\infty]{} 0 \quad \text{fast sicher} \tag{3.4}$$

(diese Konvergenzaussage nennt man den Satz von Glivenko-Cantelli).

Zum Beweis von (3.4). Zu $\varepsilon > 0$ wähle $m \in \mathbb{N}$ und $t_0 < t_1 < \cdots < t_m$ mit

$$F_\vartheta(t_0) < \frac{\varepsilon}{2},\ F_\vartheta(t_i) - F_\vartheta(t_{i-1}) < \frac{\varepsilon}{2} \text{ für } i = 1, 2, \ldots, m,\ \ F_\vartheta(t_m) > 1 - \frac{\varepsilon}{2} \tag{3.5}$$

Für genügend großes n gilt

$$\max_{i=0,1,\ldots,m} \left|F_\vartheta(t_i) - \widehat{F}_n(t_i)\right| < \frac{\varepsilon}{2}$$

und somit

$$\sup_{t_{i-1} \le t < t_i} \left|F_\vartheta(t) - \widehat{F}_n(t)\right| \le F_\vartheta(t_i) - \widehat{F}_n(t_{i-1}) = \left(F_\vartheta(t_i) - F_\vartheta(t_{i-1})\right) + \left(F_\vartheta(t_{i-1}) - \widehat{F}_n(t_{i-1})\right) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

Analog zeigt man $\sup_{-\infty < t < t_0} \left|F_\vartheta(t) - \widehat{F}_n(t)\right| < \varepsilon$ und $\sup_{t_m \le t < \infty} \left|F_\vartheta(t) - \widehat{F}_n(t)\right| < \varepsilon$.

(In (3.5) verwenden wir in unserem Argument die Stetigkeit von F_ϑ. Die Aussage (3.4) gilt auch, falls F_ϑ Sprungstellen hat, diese muss man dann im Beweis explizit berücksichtigen.) □

Bemerkung (Effiziente Berechnung von D_n)**.** Es ist

$$D_n = \max\{d_1, d_2, ..., d_n\}$$

mit $d_i = \max\left\{|\widehat{F}_n(X_{(i)}) - F_\vartheta(X_{(i)})|, |\widehat{F}_n(X_{(i)}) - F_\vartheta(X_{(i+1)}-)|\right\}$ für $i = 1, 2, \ldots, n-1$ und $d_n = 1 - F_\vartheta(X_{(n)})$, wobei $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ die Ordnungsstatistik von $X_1, \ldots, X_n$ ist (d.h. die der Größe nach sortierten Werte): $\widehat{F}_n(\cdot)$ ist konstant auf $[X_{(i)}, X_{(i+1)})$ mit Wert $\widehat{F}_n(X_{(i)})$, $F(\cdot)$ wächst dort von $F_\vartheta(X_{(i)})$ auf $F_\vartheta(X_{(i+1)}-)$ an.

Beobachtung. Falls $\vartheta \in \mathcal{M}_1(\mathbb{R})$ stetig ist (d.h. F_ϑ hat keine Sprünge), so hängt die Verteilung von D_n aus (3.4) nicht von ϑ ab.

Beweis. Sei

$$F_\vartheta^{-1}(u) := \inf\{x \in \mathbb{R} : F_\vartheta(x) \ge u\},\ \ u \in [0,1]$$

die (hier: linksstetige) inverse Verteilungsfunktion (oder „Quantilfunktion“) von ϑ (mit Interpretation $\inf \varnothing = +\infty$, $\inf \mathbb{R} = -\infty$). Es gilt

$$F_\vartheta\left(F_\vartheta^{-1}(u)\right) = u,\ \ u \in [0,1]$$

und

$$F_\vartheta^{-1}\big((0,1)\big) = \{F_\vartheta^{-1}(u) : u \in (0,1)\} = \operatorname{supp}(\vartheta)$$

Somit

$$D_n := \sup_{t \in \mathbb{R}} \left|\widehat{F}_n(t) - F_\vartheta(t)\right| = \sup_{u \in (0,1)} \left|\widehat{F}_n(F_\vartheta^{-1}(u)) - F_\vartheta(F_\vartheta^{-1}(u))\right| = \sup_{u \in (0,1)} \left|\widehat{F}_n(F_\vartheta^{-1}(u)) - u\right|$$

Weiterhin ist

$$\widehat{F}_n(F_\vartheta^{-1}(u)) = \frac{1}{n}\#\{1 \le i \le n : X_i \le F_\vartheta^{-1}(u)\} = \frac{1}{n}\#\{1 \le i \le n : F_\vartheta(X_i) \le u\} = \widehat{F}_n^*(u),\ \ u \in (0,1)$$

mit $U_i := F_\vartheta(X_i)$ und $\widehat{F}_n^*(u) = \frac{1}{n}\{1 \le i \le n : U_i \le u\}$ und $U_1, \ldots, U_n$ sind u.i.v $\sim \mathrm{Unif}([0,1])$. □

Satz. *ϑ stetige Verteilung, so gilt*

$$\lim_{n\to\infty} \mathbb{P}_\vartheta(\sqrt{n}D_n > x) = \mathbb{P}\Big(\sup_{0\le t\le 1} |B_t| > x\Big) \tag{3.6}$$
$$= 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2x^2) = \sqrt{2\pi}x\sum_{k=1}^{\infty}\exp\big(-(2k-1)^2\pi^2/(8x^2)\big)$$

Hierbei ist $(B_t)_{0\le t\le 1}$ eine Brownschen Brücke. (Man kann diese gewinnen als $B_t = W_t - tW_1$, $0 \le t \le 1$ mit $(W_t)_{t\ge 0}$ Standard-Brownbewegung.)

Beweisgedanken. Dazu: $U_1, U_2, \dots$ u.i.v. $\sim \mathrm{Unif}([0,1])$, $\widehat{F}_n^*(u) = \frac{1}{n}\#\{1 \le i \le n : U_i \le u\} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{[0,u]}(U_i)$ wie oben und $\widehat{B}_n(u) = \sqrt{n}(\widehat{F}_n^*(u) - u) = n^{-1/2}\sum_{i=1}^n(\mathbf{1}_{[0,u]}(U_i) - u)$ für $0 \le u \le 1$.

Es ist $\mathbb{E}[\widehat{B}_n(u)] = 0$ und für $0 \le u_1 \le u_2 \le 1$ ist

$$\begin{aligned}
\mathrm{Cov}[\widehat{B}_n(u_1), \widehat{B}_n(u_2)] &= \mathbb{E}[\widehat{B}_n(u_1)\widehat{B}_n(u_2)] = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\big[(\mathbf{1}_{[0,u_1]}(U_i) - u_1)(\mathbf{1}_{[0,u_2]}(U_j) - u_2)\big] \\
&= \mathbb{E}\big[(\mathbf{1}_{[0,u_1]}(U_1) - u_1)(\mathbf{1}_{[0,u_2]}(U_1) - u_2)\big] \\
&= \mathbb{E}\big[\mathbf{1}_{[0,u_1]}(U_1)\mathbf{1}_{[0,u_2]}(U_1)\big] - u_1\mathbb{E}\big[\mathbf{1}_{[0,u_2]}(U_1)\big] - u_2\mathbb{E}\big[\mathbf{1}_{[0,u_1]}(U_1)\big] + u_1u_2 \\
&= \mathbb{E}\big[\mathbf{1}_{[0,u_1]}(U_1)\big] - u_1u_2 - u_2u_1 + u_1u_2 = u_1 - u_1u_2 = u_1(1-u_2) \\
&= (u_1 \wedge u_2)(1 - u_1 \vee u_2)
\end{aligned}$$

(und dies *ist* die Kovarianzstruktur der Brownschen Brücke). Weiterhin ist für $0 \le u_1 < u_2 < \dots < u_k \le 1$, $k \in \mathbb{N}$ der Vektor

$$\big(n\widehat{F}_n^*(u_1), n(\widehat{F}_n^*(u_2) - \widehat{F}_n^*(u_1)), \dots, n(\widehat{F}_n^*(u_k) - \widehat{F}_n^*(u_{k-1})), n(1 - \widehat{F}_n^*(u_{k-1}))\big)$$

$\mathrm{Multinom}(n; u_1, u_2 - u_1, \dots, u_k - u_{k-1}, 1 - u_k)$-verteilt, ein (multivariater) zentraler Grenzwertsatz für die Multinomialverteilung zeigt dann, dass $(\widehat{B}_n(u))_{0\le u\le 1}$ für $n \to \infty$ gegen einen Gauß'schen Prozess (im f.d.d.-Sinne) mit der Kovarianzstruktur der Brownschen Brücke konvergiert; verwende schließlich z.B. den Satz von Kolmogorov-Chentsov (siehe z.B. [Kle06, Satz 21.6]) für Straffheit auf dem Pfadraum.

Alternativ ein sehr schönes Argument aus [Bre92, Ch. 13.6]:

Die Funktion $\widehat{F}_n^*(\cdot)$ springt jeweils an den Stellen $U_{(1)} < U_{(2)} < \dots < U_{(n)}$ um $1/n$ und $\widehat{F}_n^*(U_{(k)}) = k/n$, also ist (mit Setzungen $U_{(0)} := 0$, $U_{(n+1)} := 1$)

$$\begin{aligned}
D_n &= \sup_{0<u<1}\big|\widehat{F}_n^*(u) - u\big| = \max_{k=0,\dots,n}\sup\Big\{\big|\tfrac{k}{n} - u\big| : U_{(k)} \le u < U_{(k+1)}\Big\} \\
&= \max_{k=0,\dots,n}\max\big\{\big|\tfrac{k}{n} - U_{(k)}\big|, \big|\tfrac{k}{n} - U_{(k+1)}\big|\big\}
\end{aligned}$$

d.h.

$$|D_n - \widetilde{D}_n| \le \frac{1}{n} \quad \text{mit} \quad \widetilde{D}_n = \max_{k=1,\dots,n}\Big|\frac{k}{n} - U_{(k)}\Big|$$

Für die Ordnungsstatistik gilt

$$\big(U_{(1)}, U_{(2)}, \dots, U_{(n)}\big) \overset{d}{=} \Big(\frac{Z_1}{Z_{n+1}}, \frac{Z_1}{Z_{n+1}}, \dots, \frac{Z_n}{Z_{n+1}}\Big)$$

mit $Z_k = \sum_{i=1}^k Y_i$ und $Y_1, Y_2, \ldots$ u.i.v. $\sim \mathrm{Exp}(1)$. (Man kann dazu explizit mit Dichtetransformationsformel rechnen oder induktiv die Rechnungen rund um die „Gamma-Beta-Algebra“, siehe Prop. A.2.2, verwenden.)

Somit

$$\begin{aligned}\sqrt{n}\widetilde{D}_n &\stackrel{d}{=} \sqrt{n}\max_{k=1,\ldots,n}\Big|\frac{Z_k}{Z_{n+1}} - \frac{k}{n}\Big| \\ &= \frac{n}{Z_{n+1}}\max_{k=1,\ldots,n}\Big|\frac{Z_k - k}{\sqrt{n}} - \frac{k}{n}\frac{Z_{n+1}-n}{\sqrt{n}}\Big| = \frac{n}{Z_{n+1}}\max_{k=1,\ldots,n}\Big|S^{(n)}_{k/n} - \frac{k}{n}S^{(n)}_1\Big|\end{aligned}$$

mit

$$S^{(n)}_t := n^{-1/2}\sum_{i=1}^{\lfloor tn\rfloor}(V_i - 1), \quad 0 \le t \le 1$$

(und mit einem „Körnchen Salz“ am Ende, wörtlich müsste es $S^{(n)}_{(n+1)/n} + n^{-1/2}$ statt $S^{(n)}_1$ lauten). Wegen $n/Z_{n+1} \to 1$ f.s. (präziser: $|1-n/Z_{n+1}| = O_{\mathbb{P}}(n^{-1/2})$ und $\limsup_{n\to\infty}(n/\log\log n)^{1/2}|1-n/Z_{n+1}| < \infty$ f.s.) gilt mit Donskers Invarianzprinzip (siehe z.B. [Kle06, Satz 21.43])

$$\sqrt{n}\widetilde{D}_n \xrightarrow[n\to\infty]{d} \sup_{0\le t\le 1}\big|B_t - tB_1\big|$$

□

Kolmogorov-Smirnov-Test Seien $X_1, \ldots, X_n$ u.i.v $\sim \vartheta$ ($\in \mathcal{M}_1(\mathbb{R})$), die emprische Verteilungsfunktion $\widehat{F}_n(\cdot)$ wie in (3.3).

Lehne $H_0 : \vartheta = \vartheta_0$ zum Niveau α ab, wenn

$$\sqrt{n}D_n = \sup_{t\in\mathbb{R}}\big|\widehat{F}_n(t) - F_\vartheta(t)\big| > q_\alpha$$

(wobei q_α so gewählt, dass $2\sum_{k=1}^\infty (-1)^{k-1}\exp(-2k^2q_\alpha^2) = \alpha$, vgl. (3.6); Tabelle z.B. in [SH06])

Konfidenzbereich für F_ϑ: $\widehat{F}_n(\cdot) \pm q_\alpha/\sqrt{n}$, denn nach obigem ist

$$\lim_{n\to\infty}\mathbb{P}_\vartheta\Big(\sqrt{n}\big\|\widehat{F}_n - F_\vartheta\big\|_\infty > q_\alpha\Big) = \alpha$$

Einseitige Kolmogorov-Smirnov-Tests

$$D_n^+ := \sup_{t\in\mathbb{R}}\big(\widehat{F}_n(t) - F_{\vartheta_0}(t)\big), \quad D_n^- := \inf_{t\in\mathbb{R}}\big(\widehat{F}_n(t) - F_{\vartheta_0}t\big)$$

Es gilt für $x > 0$

$$\lim_{n\to\infty}\mathbb{P}_{\vartheta_0}\big(\sqrt{n}D_n^+ > x\big) = \lim_{n\to\infty}\mathbb{P}_{\vartheta_0}\big(\sqrt{n}D_n^- < -x\big) = e^{-2x^2}$$

Lehne $H_0 : \vartheta \le \vartheta_0$ ($\Leftrightarrow F_\vartheta(\cdot) \ge F_{\vartheta_0}(\cdot)$ ab, wenn

$$\sqrt{n}D_n^- < -\sqrt{\log(1/\alpha)/2}$$

Lehne $H_0 : \vartheta \ge \vartheta_0$ ($\Leftrightarrow F_\vartheta(\cdot) \le F_{\vartheta_0}(\cdot)$) ab, wenn

$$\sqrt{n}D_n^+ > \sqrt{\log(1/\alpha)/2}$$

Bericht. Tatsächlich gilt eine quantitativ verschärfte Version von (3.4) allgemein auch für endliches $n \in \mathbb{N}$ (Dvoretzky-Kiefer-Wolfowitz-Ungleichung, z.B. [LR06, Thm. 11.2.18]):

$$\mathbb{P}_\vartheta(D_n > t) \le Ce^{-2nt^2} \text{ für } t \ge 0,\ n \in \mathbb{N}$$

mit einer universellen Konstante $C < \infty$ (man kann $C = 2$ wählen, [Mas90]).

3.4 Zu Kernschätzern für Dichten

$X_1, \dots, X_n$ u.i.v. reellwertige Zufallsvariablen mit Dichte f, wir wollen anhand von n Beobachtungswerten die Dichtefunktion $f : \mathbb{R} \to [0, \infty)$ schätzen.

Wir wählen einen Kern $K : \mathbb{R} \to [0, \infty)$ mit

$$\int_{-\infty}^{\infty} K(y)\,dy = 1, \quad \int_{-\infty}^{\infty} yK(y)\,dy = 0, \quad \int_{-\infty}^{\infty} y^2K(y)\,dy < \infty \tag{3.7}$$

zum Beispiel $K(y) = (2\pi)^{-1/2}\exp(-y^2/2)$, der Gauß-Kern, oder $K(y) = \frac{3}{4}\mathbf{1}_{[-\sqrt{5},\sqrt{5}]}(y)(1 - x^2/5)/\sqrt{5}$, der Epanechnikov-Kern.

Wir bilden (für eine Wahl der *Bandbreite* $h > 0$)

$$\widehat{f}_n(x) := \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big),\ \ x \in \mathbb{R} \tag{3.8}$$

Beispiel. Für ein Daten-Beispiel (mit Rs `faithful`-Datensatz zu Eruptionsdauern eines Geysirs im Yellowstone-Nationalpark) siehe den R-Code `Beispiel_zu_Kernschaetzern.R`:

```
# ein Beispiel zur Dichteschaetzung
data(faithful)
?faithful

X <- faithful$eruptions
X

hist(X)

stripchart(X, method='jitter')
F <- ecdf(X)
plot(F)
1/sqrt(272)

hist(X, breaks=20, prob=T)
hist(X, breaks=40, prob=T)
rug(X)

h <- 0.5

fhut <- function(x) sum(sapply(1:272, function(i) dnorm(x,mean=X[i],sd=h)))/272
fhut(1.7)
```

```
fhv <- Vectorize(fhut)

curve(fhv, add=T, col='red')

h <- 0.1
curve(fhv, add=T, col='blue')

rug(X)

h <-0.01
curve(fhv, add=T, col='brown')

h <- 1.0
curve(fhv, add=T, col='green')

# zur Wahl der Bandbreite via Kreuzvalidierung:
n <- length(X)

curve(fhv, xlim=c(0,6))
fhutquad <- function(x) fhut(x)^2

integrate(Vectorize(fhutquad), lower=0, upper=6)

s <- 0
for (i in 1:n)
  for (j in (1:n)[-i])
    s <- s+dnorm(X[i]-X[j],sd=h)
2*s/(n*(n-1))

integrate(Vectorize(fhutquad), lower=0, upper=6,subdivisions = 1000)$value-2*s/(n*(n-
1))

hs <- seq(from=0.01, to=1.5, by=0.01)
Js <- numeric(length(hs))

for (k in 1:length(hs)) {
  h <- hs[k]

  s <- 0
  for (i in 1:n)
    for (j in (1:n)[-i])
      s <- s+dnorm(X[i]-X[j],sd=h)

  Js[k] <- integrate(Vectorize(fhutquad), lower=0, upper=6,subdivisions = 1000)$value-
2*s/(n*(n-1))
}
```

```
plot(hs,Js)

which.min(Js)
# welcher Wert von h optimiert?
hs[which.min(Js)]
```

Der mittlere integrierte quadratische Fehler von $\widehat{f}_n$ ist

$$\begin{aligned}\mathrm{MISE}_f(\widehat{f}_n) &= \int_{-\infty}^{\infty} \mathbb{E}_f\Big[\big(\widehat{f}_n(x) - f(x)\big)^2\Big]\,dx \\ &= \int_{-\infty}^{\infty} \mathrm{Var}_f\big[\widehat{f}_n(x)\big]\,dx + \int_{-\infty}^{\infty} \Big(\mathbb{E}_f\big[\widehat{f}_n(x) - f(x)\big]\Big)^2\,dx \qquad (3.9)\end{aligned}$$

Wir sehen hier eine Instanz des Verzerrung-Varianz-Dilemmas (engl.: variance-bias trade off): Wenn wir h klein machen, sinkt die Verzerrung, dafür steigt aber die Varianz des Schätzers (es stellt sich heraus: die Terme sind von der Ordnung $\frac{1}{nh}$ bzw. h^4, siehe den Beweis von Satz 30 unten).

Für ein intuitives Argument beachte

$$\mathbb{E}_f\big[\widehat{f}_n(x)\big] = \frac{1}{h}\mathbb{E}_f\Big[K\Big(\frac{x - X_1}{h}\Big)\Big] = \int_{\mathbb{R}} \frac{1}{h} K\Big(\frac{x-y}{h} f(y)\,dy = \big(K_h * f\big)(x) \neq f(x)$$

und

$$\begin{aligned}\mathbb{E}_f\big[\widehat{f}_n(x) - f(x)\big] &= \int_{-\infty}^{\infty} \frac{1}{h} K\big(\frac{x-z}{h}\big)\, f(z) dz - f(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\big(\frac{x-z}{h}\big)\big(f(z) - f(x)\big)\,dz \\ &= \int_{-\infty}^{\infty} K(y)\big(f(x - hy) - f(x)\big)\,dy \\ &\approx \int_{-\infty}^{\infty} K(y)\big(-hyf(x) + \frac{1}{2}y^2h^2f''(x) + o(h^2)\big)\,dy \\ &\approx -h\cdot 0 + \frac{1}{2}h^2 f''(x)\int_{-\infty}^{\infty} y^2 K(y)\,dy + o(h^2)\end{aligned}$$

(mit Substitution $(x-z)/h = y$, also $z = x - hy$ und $(1/h)dz = dy$ und Taylor-Entwicklung von f bis zur 2. Ordnung; wir verwenden zudem $\int K(y)\,dy = 1$, $\int yK(y)\,dy = 0$ und $\int y^2K(y)\,dy < \infty$), dies macht zumindest plausibel, dass (Verzerrung)$^2 = O(h^4)$.

Satz 30. *Es gelte $\int_{-\infty}^{\infty} |f''(x)|^2\,dx < \infty$, K erfülle (3.7) und zudem $\int_{-\infty}^{\infty} K(y)^2\,dy < \infty$. Dann gilt*

$$\int_{-\infty}^{\infty} \mathbb{E}_f\Big[\big(\widehat{f}_n(x) - f(x)\big)^2\Big]\,dx \leq C_f\Big(\frac{1}{nh} + h^4\Big) \qquad (3.10)$$

Insbesondere ist mit Wahl $h_n \sim n^{-1/5}$: $\mathrm{MISE}_f(\widehat{f}_n) = O(n^{-4/5})$

Bei einem parametrischen Problem könnte man (in relativ allgemeinen Situationen) MISE $= O(n^{-1})$ erwarten, insofern zeigt (3.10), dass man durch die „Nicht-Parametrizität“ einen Faktor $n^{1/5}$ „verliert“. Die Ordnung $n^{-4/5}$ ist in dieser Allgemeinheit optimal, siehe [vdV98, Ch. 24.3].

Beweis von Satz 30. Nach Konstruktion (3.8) ist

$$\begin{aligned}\mathrm{Var}_f\big[\widehat{f}_n(x)\big] &= \frac{1}{n}\mathrm{Var}_f\Big[\frac{1}{h}K\big(\frac{x-X_1}{h}\big)\Big]\\ &\le \frac{1}{nh^2}\mathbb{E}_f\Big[K\big(\frac{x-X_1}{h}\big)^2\Big] = \frac{1}{nh^2}\int_{-\infty}^{\infty} K\big(\frac{x-z}{h}\big)^2 f(z)\,dz = \frac{1}{nh}\int_{-\infty}^{\infty} K(y)^2 f(x-hy)\,dy\end{aligned}$$

(mit Substitution $(x-z)/h = y$, also $z = x - hy$ und $(1/h)dz = dy$) somit

$$\begin{aligned}\int_{-\infty}^{\infty}\mathrm{Var}_f\big[\widehat{f}_n(x)\big]\,dx &\le \frac{1}{nh}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(y)^2 f(x-hy)\,dy\,dx\\ &= \frac{1}{nh}\int_{-\infty}^{\infty} K(y)^2\int_{-\infty}^{\infty} f(x-hy)\,dx\,dy = \frac{1}{nh}\int_{-\infty}^{\infty} K(y)^2\,dy\end{aligned}$$

Für den zweiten Term in (3.9) verwenden wir eine Taylor-Entwicklung von f in x bis zur 2. Ordnung (mit Restglied in Integralform):

$$f(x+h) = f(x) + hf'(x) + h^2\int_0^1 f''(x+uh)(1-u)\,du$$

und damit

$$\begin{aligned}\mathbb{E}_f\big[\widehat{f}_n(x) - f(x)\big] &= \int_{-\infty}^{\infty}\frac{1}{h}K\big(\frac{x-z}{h}\big)\,f(z)dz - f(x) = \int_{-\infty}^{\infty}\frac{1}{h}K\big(\frac{x-z}{h}\big)\big(f(z)-f(x)\big)\,dz\\ &= \int_{-\infty}^{\infty} K(y)\big(f(x-hy)-f(x)\big)\,dy\\ &= \int_{-\infty}^{\infty} K(y)\Big(-hyf'(x) + h^2y^2\int_0^1 f''(x-uhy)(1-u)\,du\Big)\,dy\\ &= 0 + h^2\int_{-\infty}^{\infty}\int_0^1 y^2 f''(x-uhy)(1-u)\,du\,K(y)dy\\ &= h^2\mathbb{E}\big[Y^2 f''(x-hYU)(1-U)\big]\\ &\le h^2\Big(\mathbb{E}\big[Y^2\big]\cdot\mathbb{E}\big[Y^2|f''(x-hYU)|^2(1-U)^2\big]\Big)^{1/2}\end{aligned}$$

(wobei Y eine Zufallsvariable mit Dichte $K(\cdot)$ ist; verwende Cauchy-Schwarz-Ungleichung in der letzten Zeile) also

$$\begin{aligned}&\int_{-\infty}^{\infty}\Big(\mathbb{E}_f\big[\widehat{f}_n(x)-f(x)\big]\Big)^2\,dx\\ &\le \int_{-\infty}^{\infty}\Big(h^4\int_{-\infty}^{\infty} y^2K(y)\,dy\int_{-\infty}^{\infty}\int_0^1 y^2|f''(x-hyu)|^2(1-u)^2\,du\,K(y)dy\Big)\,dx\\ &= h^4\int_{-\infty}^{\infty} y^2K(y)\,dy\int_{-\infty}^{\infty}\int_0^1 y^2\int_{-\infty}^{\infty}|f''(x-hyu)|^2\,dx\,(1-u)^2\,du\,K(y)dy\\ &= h^4\Big(\int_{-\infty}^{\infty} y^2K(y)\,dy\Big)^2\int_{-\infty}^{\infty}|f''(x)|^2\,dx\int_0^1(1-u)^2\,du\\ &= h^4\frac{1}{3}\Big(\int_{-\infty}^{\infty} y^2K(y)\,dy\Big)^2\int_{-\infty}^{\infty}|f''(x)|^2\,dx\end{aligned}$$

und wir können

$$\int_{-\infty}^{\infty}K(y)^2\,dy + \frac{1}{3}\Big(\int_{-\infty}^{\infty} y^2K(y)\,dy\Big)^2\int_{-\infty}^{\infty}|f''(x)|^2\,dx$$

wählen. □

Bericht. Wenn man annimmt, dass f m-mal stetig differenzierbar ist mit $\int_{-\infty}^{\infty} |f^{(m)}(x)|^2\,dx < \infty$, so kann man einen Kern K wählen (mit $\int_{-\infty}^{\infty} K(y)\,dy = 1$, $\int_{-\infty}^{\infty} yK(y)\,dy = 0$, $\int_{-\infty}^{\infty} y^2K(y)\,dy = 0$, …, $\int_{-\infty}^{\infty} y^{(m-1)}K(y)\,dy = 0$, $\int_{-\infty}^{\infty} |y|^m K(y)\,dy < \infty$, so dass die rechte Seite von (3.10) ersetzt werden kann durch $C_f\Big(\frac{1}{nh} + h^{2m}\Big)$ für h genügend klein; man erhält dann für die (asymptotisch optimale) Wahl $h = h_n \sim n^{-1/(2m+1)}$, dass $\mathrm{MISE}_f(\widehat{f_n}) = O(n^{-2m/(2m+1)})$ (man kann also der "parametrischen Fehlerrate" $1/n$ in diesem Sinne "beliebig nahe kommen"), siehe [vdV98, Thm. 24.2].

Bandbreite und Kreuzvalidierung Wir entnehmen dem Beweis von Satz 30, dass

$$\int_{-\infty}^{\infty} \mathbb{E}_f\Big[\big(\widehat{f_n}(x) - f(x)\big)^2\Big]\,dx \approx \frac{\int_{-\infty}^{\infty} K(x)^2\,dx}{nh} + \frac{1}{3}\sigma_K^4 h^4 \int_{-\infty}^{\infty} |f''(x)|^2\,dx$$

Um für gegebenes n das (im Sinne der Asymptotik des MISE) optimale h zu wählen, müssten wir daher $\int_{-\infty}^{\infty} |f''(x)|^2\,dx$ kennen (was in der Praxis unrealistisch ist, da wir ja gerade $f(\cdot)$ schätzen möchten).

Der L^2-Abstand zwischen f und $\widehat{f_n}$ ist

$$\int_{-\infty}^{\infty} \big(\widehat{f_n}(x) - f(x)\big)^2\,dx = \int_{-\infty}^{\infty} \widehat{f_n}(x)^2\,dx - 2\int_{-\infty}^{\infty} \widehat{f_n}(x)f(x)\,dx + \int_{-\infty}^{\infty} f(x)^2\,dx$$

da der letzte Term nicht von der Wahl von $K(\cdot)$ und h abhängt, ist die Minimierung dieses Abstands also äquivalent zur Minimierung von

$$J_n(h) := \int_{-\infty}^{\infty} \widehat{f_n}(x)^2\,dx - 2\int_{-\infty}^{\infty} \widehat{f_n}(x)f(x)\,dx \tag{3.11}$$

(Beachte: $\widehat{f_n}$ hängt von h ab, auch wenn die Notation das „verschweigt".)

Da wir f nicht kennen: Bilde

$$\widehat{f}_{n,-i}(x) := \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j\neq i}}^{n} K\Big(\frac{x - X_j}{h}\Big), \quad x \in \mathbb{R}, \quad i = 1, 2, \ldots, n$$

(dies entspricht sozusagen "leave-one-out cross-validation") und damit

$$\widehat{J_n}(h) := \int_{-\infty}^{\infty} \widehat{f_n}(x)^2\,dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{n,-i}(X_i)$$

als Schätzwert für $J_n(h)$. (Immerhin ist $\mathbb{E}_f\big[\widehat{f}_{n,-i}(X_i)\big] = \mathbb{E}_f\Big[\int_{-\infty}^{\infty} \widehat{f}_{n-1}(x)f(x)\,dx\Big]$ und somit $\mathbb{E}_f[\widehat{J_n}(h)] \approx \mathbb{E}_f[J_n(h)]$.) Wähle dann

$$h_n \in \operatorname{argmin}_{h>0} \widehat{J_n}(h) \tag{3.12}$$

Bericht (Asymptotische Optimalität der Kreuzvalidierung)**.** f beschränkt, h_n aus (3.12), so gilt

$$\frac{\int \big(f(x) - \widehat{f}_{n,h_n}(x)\big)^2\,dx}{\inf_{h>0} \int \big(f(x) - \widehat{f}_{n,h}(x)\big)^2\,dx} \xrightarrow[n\to\infty]{\mathbb{P}} 1$$

(siehe z.B. [Was04, Thm. 20.16]; die Aussage wurde von Charles J. Stone [Sto84, Thm. 1] bewiesen).

Kapitel 4

Tests für kategorielle Beobachtungen (zum χ^2-Test)

χ^2-Anpassungstest

Ein Experiment mit s möglichen Ausgängen werde n mal (unabhängig) wiederholt, Ausgang i habe die (unbekannte) Wahrscheinlichkeit ϑ_i, $i = 1, \dots, s$.

Angenommen, wir beobachten h_i-mal Ausgang i für $i = 1, \dots, s$. Passt dies zur (Null-) Hypothese, dass

$$\vartheta = (\vartheta_1, \dots, \vartheta_s) = (\rho_1, \dots, \rho_s) = \rho$$

gilt für einen vorgegebenen Vektor ρ von Wahrscheinlichkeitsgewichten (auf $\{1, \dots, s\}$)?

Beispiel. Wir vermuten, dass ein gegebener sechsseitiger Würfel unfair ist und möchten dies auf dem 5%-Niveau testen. Bei 120-maligem Würfeln finden wir folgende Häufigkeiten:

i	1	2	3	4	5	6
h_i	13	12	20	18	26	31

Es ist

$$\widehat{\vartheta}^{(\mathrm{ML})} = \big(\widehat{\vartheta}_1^{(\mathrm{ML})}, \dots, \widehat{\vartheta}_s^{(\mathrm{ML})}\big) = \Big(\frac{h_1}{n}, \dots, \frac{h_s}{n}\Big)$$

denn $\rho\big(\vartheta; (h_1, \dots, h_s)\big) = \binom{n}{h_1 \cdots h_s} \prod_{j=1}^s \vartheta_j^{h_j}$; wir maximieren

$$\vartheta \mapsto \log \rho\big(\vartheta; (h_1, \dots, h_s)\big) = C_h + \sum_{j=1}^s h_j \log(\vartheta_j)$$

unter der Nebenbedingung $g(\vartheta) = \vartheta_1 + \cdots + \vartheta_s = 1$, also

$$\frac{\partial}{\partial \vartheta_j} \sum_{j=1}^s h_j \log(\vartheta_j) = \frac{h_j}{\vartheta_j} = \lambda \frac{\partial}{\partial \vartheta_j} g(\vartheta) = \lambda, \quad j = 1, \dots, s$$

mit einem Lagrange-Multiplikator $\lambda \in \mathbb{R}$. Also $\widehat{\vartheta}_j^{(\mathrm{ML})} = \frac{h_j}{\lambda}$ und wegen $g\big(\widehat{\vartheta}^{(\mathrm{ML})}\big) = 1$ folgt $\lambda = n$.

Somit

$$R_n := \frac{\sup_{\vartheta \in \mathcal{M}_1(\{1,\dots,s\})} \rho\big(\vartheta; (h_1, \dots, h_s)\big)}{\rho\big(\rho; (h_1, \dots, h_s)\big)} = \prod_{j=1}^s \Big(\frac{h_j/n}{\rho_j}\Big)^{h_j} = \exp\Big(n \sum_{j=1}^s \frac{h_j}{n} \log\big(\frac{h_j/n}{\rho_j}\big)\Big)$$

Ein Likelihood-Quotienten-Test (in Anlehnung an die Neyman-Pearson-Philosphie) wäre also

$$\text{lehne } H_0 \text{ ab, wenn } \sum_{j=1}^{s} \frac{h_j}{n} \log\Big(\frac{h_j/n}{\rho_j}\Big) > c$$

wobei $c = c(\rho, n)$ so gewählt wird, dass das gewünschte Niveau α eingehalten wird. (Dies ist Hoeffdings Entropietest (1965), siehe [Geo07, Satz 11.15]; den kritischen Wert explizit zu bestimmen ist allerdings rechnerisch aufwendig.)

Man approximiert: Setze $a_j := \frac{h_j}{n\rho_j} - 1$, es ist

$$n \sum_{j=1}^{s} \frac{h_j}{n} \log\Big(\frac{h_j/n}{\rho_j}\Big) = n \sum_{j=1}^{s} \rho_j \psi(1 + a_j)$$

mit $\psi(u) = 1 - u + u \log u$. Taylor-Approximation von ψ in $u = 1$ liefert ($\psi(1) = \psi'(1) = 0$)

$$\psi(u) = \frac{(u-1)^2}{2} + O\big(|u-1|^3\big)$$

d.h.

$$\begin{aligned} n \sum_{j=1}^{s} \frac{h_j}{n} \log\Big(\frac{h_j/n}{\rho_j}\Big) &= n \sum_{j=1}^{s} \rho_j \Big(\frac{h_j}{n\rho_j} - 1\Big)^2 + O\bigg(n \sum_{j=1}^{s} \rho_j \Big|\frac{h_j}{n\rho_j} - 1\Big|^3\bigg) \\ &= \sum_{j=1}^{s} \frac{(h_j - n\rho_j)^2}{n\rho_j} + O_{\mathbb{P}_\rho}(n^{-1/2}) \end{aligned}$$

Offenbar ist unter H_0 der Vektor der beobachteten Häufigkeiten multinomialverteilt, $(H_1^{(n)}, \dots, H_s^{(n)}) \sim \mathrm{Mult}(n; \rho_1, \dots, \rho_s)$; wir benötigen daher eine multivariate Version des Satzes von de Moivre-Laplace:

Satz. *Sei* $\rho \in \Delta_s := \{(\vartheta_1, \dots, \vartheta_s) \in [0,1]^s : \vartheta_1 + \cdots + \vartheta_s = 1\}$,

$$(H_1^{(n)}, \dots, H_s^{(n)}) \sim \mathrm{Mult}(n; \rho_1, \dots, \rho_s),$$

dann gilt mit $u_\rho := (\sqrt{\rho_1}, \dots, \sqrt{\rho_s}) \in \mathbb{R}^s$ *(beachte:* $\|u_\rho\| = 1$*),* $\mathbb{H}_\rho := \{x \in \mathbb{R}^s : x \cdot u_\rho = 0\}$ *(der Hyperebene durch den Ursprung mit Normale* u_ρ*) und* $\Pi_\rho : \mathbb{R}^s \to \mathbb{H}_\rho$ *der orthogonalen Projektion auf* $\mathbb{H}_\rho$

$$\Big(\frac{H_i^{(n)} - n\rho_i}{\sqrt{n\rho_i}}\Big)_{i=1,\dots,s} \xrightarrow[n\to\infty]{d} \mathcal{N}_{0,1}^{\otimes s} \circ (\Pi_\rho)^{-1}$$

und

$$\sum_{i=1}^{s} \frac{(H_i^{(n)} - n\rho_i)^2}{n\rho_i} \xrightarrow[n\to\infty]{d} \chi^2_{s-1}$$

Übrigens (alternative Berechnung der χ^2-Statistik): Es ist

$$\sum_{i=1}^{s} \frac{(H_i^{(n)} - n\rho_i)^2}{n\rho_i} = n \sum_{i=1}^{s} \rho_i \cdot \Big(\frac{H_i^{(n)}}{n\rho_i} - 1\Big)^2 = n \sum_{i=1}^{s} \rho_i \Big(\frac{(H_i^{(n)})^2}{n^2\rho_i^2} - 2\frac{H_i^{(n)}}{n\rho_i} + 1\Big) = n\Big(\sum_{i=1}^{s} \frac{(H_i^{(n)}/n)^2}{\rho_i}\Big) - n$$

was zum konkreten Berechnen angenehmer sein kann.

Beweisskizze. Zentrale Idee: Aufheben der Abhängigkeiten durch Poissonisierung. Seien $X_1^{(n)}, \dots, X_s^{(n)}$ u.a., $X_i^{(n)} \sim \mathrm{Poi}_{n\rho_i}$, dann ist

$$N_n := X_1^{(n)} + \cdots + X_s^{(n)} \sim \mathrm{Poi}_n$$

Beachte: Für $m \in \mathbb{N}$, $h_1, \dots, h_s \in \mathbb{N}_0$ mit $h_1 + \cdots + h_s = m$ ist

$$\begin{aligned} &P\big(X_1^{(n)} = h_1, \dots, X_s^{(n)} = h_s \,\big|\, N_n = m\big) \\ &= \big(e^{-n}\frac{n^m}{m!}\big)^{-1} \prod_{i=1}^{s} e^{-n\rho_i}\frac{(n\rho_i)^{h_i}}{h_i!} = \binom{m}{h_1, h_2, \dots, h_s}\rho_1^{h_1}\cdots\rho_s^{h_s}, \end{aligned}$$

d.h. bedingt auf $\{N_n = m\}$ ist $(X_1^{(n)}, \dots, X_s^{(n)}) \sim \mathrm{Mult}_{m;\rho_1,\dots,\rho_s}$.

Sei

$$\widetilde{X}_i^{(n)} := \frac{X_i^{(n)} - n\rho_i}{\sqrt{n\rho_i}}, \quad \widetilde{H}_i^{(n)} := \frac{H_i^{(n)} - n\rho_i}{\sqrt{n\rho_i}}$$

(beachte: $\mathbb{E}[\widetilde{X}_i^{(n)}] = 0$, $\mathrm{Var}[\widetilde{X}_i^{(n)}] = 1$),

$$\widetilde{N}_n := \frac{N_n - n}{\sqrt{n}} = \sum_{i=1}^{s} \sqrt{\rho_i}\,\widetilde{X}_i^{(n)}$$

(beachte: $N_n = n \iff \widetilde{N}_n = 0 \iff (\widetilde{X}_1^{(n)}, \dots, \widetilde{X}_s^{(n)})^T \in \mathbb{H}_\rho$)

Der zentrale Grenzwertsatz, angewendet auf jede der (unabhängigen) Koordinaten, liefert

$$\widetilde{X}^{(n)} := (\widetilde{X}_1^{(n)}, \dots, \widetilde{X}_s^{(n)})^T \xrightarrow[n\to\infty]{d} \mathcal{N}_{0,1}^{\otimes s}$$

und somit gilt auch

$$\sum_{i=1}^{s} \big(\widetilde{X}_1^{(n)}\big)^2 \xrightarrow[n\to\infty]{d} \chi_s^2.$$

Das macht zumindest plausibel, dass auch

$$\mathscr{L}\Big(\sum_{i=1}^{s} \big(\widetilde{X}_1^{(n)}\big)^2 \,\Big|\, \widetilde{N}_n = 0\Big) \xrightarrow[n\to\infty]{d} \chi_{s-1}^2$$

gilt.

Hier sind etwas mehr Details: Sei O eine orthogonale $s \times s$-Matrix, deren letzte Spalte u_ρ ist (ergänze u_ρ zur ONB),

$$\Pi_\rho = O \cdot \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ & & 1 \\ 0 & & 0 \end{pmatrix} O^T$$

ist die (orthogonale) Projektion(smatrix) auf $\mathbb{H}_\rho$.

Sei $\widetilde{\mathcal{N}}_\rho := \mathcal{N}_{0,1}^{\otimes s} \circ (\Pi_\rho)^{-1}$,
$a_1, \dots, a_s \in \mathbb{R}$, $A := (-\infty, a_1] \times \times (-\infty, a_s]$,

$$q_{m,n} := P\big(\widetilde{X}^{(n)} \in A \,\big|\, N_n = m\big)$$

Wir möchten zeigen: $q_{n,n} = P\big(\widetilde{H}^{(n)} \in A\big) \to \widetilde{\mathcal{N}}_\rho(A)$.

Es gilt

$$q_{m,n} \geq q_{m+1,n} \quad \text{für } m, n \in \mathbb{N}$$

(verwende die „natürliche“ Kopplung der Multinomialverteilungen und die spezielle Form von A), somit für $\varepsilon > 0$

$$P\big(\widetilde{X}^{(n)} \in A \,\big|\, \widetilde{N}_n \in [0,\varepsilon]\big) \leq P\big(\widetilde{H}^{(n)} \in A\big) \leq P\big(\widetilde{X}^{(n)} \in A \,\big|\, \widetilde{N}_n \in [-\varepsilon,0]\big)$$

(denn

$$q_{n,n} \geq \sum_{m=n}^{\lceil n+\varepsilon\sqrt{n}\rceil} q_{m,n} P\big(N_n = m \,\big|\, \widetilde{N}_n \in [0,\varepsilon]\big) = P\big(\widetilde{X}^{(n)} \in A \,\big|\, \widetilde{N}_n \in [0,\varepsilon]\big)$$

und analog für die andere Schranke).

Setze $\widetilde{Y}^{(n)} := O^T \widetilde{X}^{(n)}$, $U_\varepsilon := \{(x_1,\dots,x_s)^T \in \mathbb{R}^s : 0 \leq x_s \leq \varepsilon\} = \mathbb{R}^{s-1} \times [0,\varepsilon]$

$$\begin{aligned}
&P\big(\widetilde{X}^{(n)} \in A \,\big|\, \widetilde{N}_n \in [0,\varepsilon]\big) = P\big(\widetilde{Y}^{(n)} \in O^T A \,\big|\, \widetilde{Y}_n \in U_\varepsilon\big) \\
&= \frac{P\big(\widetilde{Y}_n \in O^T A \cap U_\varepsilon\big)}{P\big(\widetilde{Y}_n \in U_\varepsilon\big)} \\
&\xrightarrow[n\to\infty]{} \frac{\mathcal{N}_{0,1}^{\otimes s}(O^T A \cap U_\varepsilon)}{\mathcal{N}_{0,1}^{\otimes s}(U_\varepsilon)} \\
&\qquad = \frac{1}{\mathcal{N}_{0,1}([0,\varepsilon])} \int_0^\varepsilon \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathcal{N}_{0,1}^{\otimes(s-1)}\big(\{x \in \mathbb{R}^{s-1} : (x,t) \in O^T A\}\big)\, dt
\end{aligned}$$

denn mit zentralem Grenzwertsatz und Rotationssymmetrie der s-dim. Normalverteilung (Beispiel A.2.8) folgt $\widetilde{Y}^{(n)} \longrightarrow^d_{n\to\infty} \mathcal{N}_{0,1}^{\otimes s}$. Mit $\varepsilon \downarrow 0$ konvergiert die rechte Seite gegen $\widetilde{\mathcal{N}}_\rho(A)$. □

Korollar (χ^2-Anpassungstest[1])**.** *Sei* $\vartheta \in \Theta = \Delta_s := \{(\vartheta_1,\dots,\vartheta_s) \in [0,1]^s : \vartheta_1 + \dots + \vartheta_s = 1\}$, *unter* P_ϑ *sei* $(H_1,\dots,H_s) \sim \mathrm{Mult}_{n;\vartheta_1,\dots,\vartheta_s}$.

Sei $\rho \in \Delta_s$,

$$D := \sum_{i=1}^s \frac{(H_i - n\rho_i)^2}{n\rho_i},$$

$\alpha \in (0,1)$, q *das* $(1-\alpha)$*-Quantil der* χ^2_{s-1}*-Verteilung.*

Der Test von $H_0 : \{\vartheta = \rho\}$ *gegen* $H_1 : \{\vartheta \neq \rho\}$ *mit Ablehnungsbereich* $\{D > q\}$ *hat (asymptotisches) Niveau* α.

Dies folgt aus Satz 4. Satz 4 macht allerdings keine Aussage darüber, wie groß n sein sollte, damit die Approximation plausibel ist. Eine oft zitierte Faustregel (für die Gültigkeit der χ^2-Approximation) ist $n\rho_i \geq 5$ für alle i.

[1] von Karl Pearson (1857–1936) im Jahr 1900 vorgeschlagen

Beispiel (Mendels Erbsenexperimente[2])**.** Betrachte zwei Merkmale: Farbe: grün (rezessiv) vs. gelb (dominant), Form: rund (dominant) vs. runzlig (rezessiv)

Beim Kreuzen von Doppelhybriden erwarten wir folgende Phänotypwahrscheinlichkeiten unter Mendel'scher Segregation („rund" und „gelb" sind jeweils dominant, $n = 556$ Versuche):

Typ	rund/gelb	rund/grün	kantig/gelb	rund/gelb
Anteil	9/16	3/16	3/16	1/16
Erwartete Anzahl	315	104,25	104,25	34,75
beobachtet	315	108	101	32

Wir finden $D \cong 0{,}47$, $\chi^2_3([0, 0.47]) \cong 0{,}075$, ein χ^2-Test zum 1%-Niveau lehnt H_0 nicht ab (und auch zu „nahezu egal welchem Niveau" nicht). Insoweit passen die Daten sehr gut zu den theoretischen Häufigkeiten.

```
> ## G.Mendel, Versuche ueber Pflanzenhybriden,
> ## Verhandlungen des naturforschenden Vereines in Bruenn,
> ## Bd. IV fuer 1865, Abhandlungen: 3-47, (1866)
> erbsen <- c(315, 108, 101, 32)
> names(erbsen) <- c('rund-gelb', 'rund-gruen','kantig-gelb','kantig-gruen')
> erbsen
   rund-gelb   rund-gruen  kantig-gelb kantig-gruen
         315          108          101           32
> sum(erbsen)
[1] 556

> erbsen/sum(erbsen) # Anteile
   rund-gelb   rund-gruen  kantig-gelb kantig-gruen
  0.56654676   0.19424460   0.18165468   0.05755396

> # gemaess Mendel'schen Regeln erwartete Anteile
> # (bei Kreuzung von "Doppelhybriden"),
> # 'rund' und 'gelb' sind dominant, 'kantig' und 'gruen' rezessiv
> theoret <- c(9/16,3/16,3/16,1/16)
> chisq.test(erbsen, p=theoret)

Chi-squared test for given probabilities

data:  erbsen
X-squared = 0.47002, df = 3, p-value = 0.9254

> # falls wir uns nicht auf die Asymptotik verlassen wollen:
> chisq.test(erbsen, p=theoret, simulate.p.value=TRUE)
```

[2]Gregor Mendel, 1822–1884; G. Mendel, Versuche über Pflanzenhybriden, Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen: 3–47, (1866).

```
Chi-squared test for given probabilities with simulated p-value (based
on 2000 replicates)

data:  erbsen
X-squared = 0.47002, df = NA, p-value = 0.9355
```

Beispiel. Wir vermuten, dass ein gegebener sechsseitiger Würfel unfair ist und möchten dies auf dem 5%-Niveau testen. Bei 120-maligem Würfeln finden wir folgende Häufigkeiten:

i	1	2	3	4	5	6
h_i	13	12	20	18	26	31

Es ist $D \cong 13{,}7$, das 95%-Quantil der χ^2_5-Verteilung ist $\cong 11{,}07$, wir können die Nullhypothese „$\vartheta = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$“ also auf dem 5%-Niveau ablehnen.

```
> w <- c(13,12,20,18,26,31)
> chisq.test(w,p=c(1/6,1/6,1/6,1/6,1/6,1/6))

Chi-squared test for given probabilities

data:  w
X-squared = 13.7, df = 5, p-value = 0.01763
```

Zum χ^2-Test auf Homogenität (auch: „auf Unabhängigkeit“)

In einem Experiment werden zwei „Merkmale“ beobachtet, wobei das erste Merkmal a und das zweite Merkmal b viele Ausprägungen besitzt (also insgesamt $s = a \cdot b$ mögliche Ausgänge).

Unter n u.a. Wiederholungen werde h_{ij} mal Ausgang (i, j) beobachtet ($i \in \{1, 2, \ldots, a\}$, $j \in \{1, 2, \ldots, b\}$), man fasst die Beobachtungen in einer $a \times b$-Kontingenztafel zusammen:

i \ j	1	2	3	
1	h_{11}	h_{12}	h_{13}	$h_{1\cdot}$
2	h_{21}	h_{22}	h_{23}	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot 3}$	$h_{\cdot\cdot} = n$

mit Zeilensummen $h_{i\cdot} = \sum_{j=1}^{b} h_{ij}$, Spaltensummen $h_{\cdot j} = \sum_{i=1}^{a} h_{ij}$ und Gesamtsumme $h_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{b} h_{ij} = n$.

Wir fassen die beobachteten Häufigkeiten als Realisierungen einer

$$\text{multinomial}(n, (\vartheta_{ij})_{i=1,\ldots,a;j=1,\ldots,b})\text{-verteilten ZV } (H_{ij})_{i=1,\ldots,a;j=1,\ldots,b}$$

auf, wobei

$(\vartheta_{ij})_{i=1,\dots,a;j=1,\dots,b}$ ein $a \cdot b$-dimensionaler Vektor von Wahrscheinlichkeitsgewichten ist.

Passen die Beobachtungen zur Nullhypothese, dass

$$\vartheta_{ij} = \eta_i \cdot \rho_j, \quad \text{für } i = 1, \dots, a,\ j = 1, \dots, b$$

mit $(\eta_i)_{i=1,\dots,a}$, $(\rho_j)_{j=1,\dots,b}$ gewissen a- bzw. b-dimensionalen Vektoren von Wahrscheinlichkeitsgewichten?

Wir bilden

$$\widehat{\vartheta}_{i\cdot} = \frac{H_{i\cdot}}{n}, \quad \widehat{\vartheta}_{\cdot j} = \frac{H_{\cdot j}}{n}$$

(dies *sind* hier die ML-Schätzer) und die Teststatistik

$$D = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(H_{ij} - n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j})^2}{n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j}}$$

Bericht 31. Unter H_0 : „$(\vartheta_{ij})_{i=1,\dots,a;j=1,\dots,b}$ hat Produktform“ ist D (approximativ) $\chi^2_{(a-1)(b-1)}$-verteilt.

Wir würden also H_0 zum Niveau α ablehnen, falls der beobachtete Wert größer ist als das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $(a-1)(b-1)$ Freiheitsgraden. Siehe z.B. [Geo07, Kap. 11.3, S. 308 unten bis S. 311] für einen Beweis.

Exkurs: χ^2-Test für Modelle mit angepassten Parametern

Wir betrachten hier nur ein Beispiel, siehe dazu die Folien `Exkurs_chiquadrat-Test_mit_angepassten_Parametern_kompakt.pdf`

Zum Yule-Simpson-Paradoxon

Durch Zusammenfassen von Gruppen können sich (scheinbare) statistische Trends in ihr Gegenteil verkehren. Dieses Phänomen heißt Simpson-Paradoxon oder Yule-Simpson-Effekt[3].

Beispiel (Zulassungsstatistik der UC Berkeley 1973)**.** Im Herbst 1973 haben sich an der Universität Berkeley 12763 Kandidaten für ein Studium beworben, davon 8442 Männer und 4321 Frauen. Es kam zu folgenden Zulassungszahlen:

	Aufgenommen	Abgelehnt
Männer	3738	4704
Frauen	1494	2827

Demnach betrug die Zulassungsquote
bei den Männern $\frac{3738}{8442} \approx 44\%$, bei den Frauen nur $\frac{1494}{4321} \approx 35\%$.

Ein χ^2-Test auf Homogenität (z.B. mit `R`) zeigt, dass eine solche Unverhältnismäßigkeit nur mit verschwindend kleiner Wahrscheinlichkeit durch „reinen Zufall“ entsteht:

[3] nach Edward H. Simpson, *1922 und George Udny Yule, 1871–1951

```
> berkeley <- matrix(c(3738,1494,4704,2827),nrow=2)
> berkeley
     [,1] [,2]
[1,] 3738 4704
[2,] 1494 2827
> chisq.test(berkeley,correct=FALSE)

Pearson's Chi-squared test

data:  berkeley
X-squared = 111.2497, df = 1, p-value < 2.2e-16
```

Dieser Fall hat einiges Aufsehen erregt, s.a. P.J. Bickel, E.A. Hammel, J.W. O'Connell, Sex Bias in Graduate Admissions: Data from Berkeley, *Science* 187, no. 4175, 398–404, (1975).

Das Ungleichgewicht verschwindet, wenn man die Zulassungszahlen nach Departments aufspaltet:

Es stellt sich heraus, dass innerhalb der Departments die Aufnahmewahrscheinlichkeiten nicht signifikant vom Geschlecht abhängen, aber sich Frauen häufiger bei Departments mit (absolut) niedriger Aufnahmequote beworben haben als Männer – dies ist ein Beispiel für das *Simpson-Paradox*.

Die genauen nach Departments aufgeschlüsselten Bewerber- und Zulassungszahlen sind leider nicht öffentlich zugänglich (siehe aber Abb. 1 in Bickel et. al, loc. cit., für eine grafische Aufbereitung der Daten, die den Simpson-Effekt zeigt).

Bickel et. al demonstrieren das Phänomen mittels eines hypothetischen Beispiels:

	Aufgenommen	Abgelehnt
	Department of machismathics	
Männer	200	200
Frauen	100	100
	Department of social warfare	
Männer	50	100
Frauen	150	300
	Gesamt	
Männer	250	300
Frauen	250	400

Literaturverzeichnis

[AS64] Abramowitz, Milton und Irene A. Stegun: *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Band 55 der Reihe *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[BD77] Bickel, Peter J. und Kjell A. Doksum: *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, 1977.

[Bir17] Birkner, Matthias: *Stochastik I*, Sommersemester 2017. `https://www.staff.uni-mainz.de/birkner/StochI_17/`.

[Bre92] Breiman, Leo: *Probability*. Classics in applied mathematics ; 7. SIAM, Philadelphia, Pa., unabridged, corr. republ. Auflage, 1992.

[Geo07] Georgii, Hans Otto: *Stochastik*. de Gruyter Lehrbuch. [de Gruyter Textbook]. Walter de Gruyter & Co., Berlin, expanded Auflage, 2007, ISBN 978-3-11-019349-7. `https://doi.org/10.1515/9783110206777`, Einführung in die Wahrscheinlichkeitstheorie und Statistik. [Introduction to probability theory and statistics].

[HE07] Hartung, Joachim und Bärbel Elpelt: *Multivariate Statistik : Lehr- und Handbuch der angewandten Statistik*. Oldenbourg, München [u.a.], 7., unveränd. aufl. Auflage, 2007.

[Kle06] Klenke, Achim: *Wahrscheinlichkeitstheorie*. Springer, 2006.

[LR06] Lehmann, E.L. und J.P. Romano: *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 2006, ISBN 9780387276052.

[LS48] Lehmann, Erich L. und Charles Stein: *Most powerful tests of composite hypotheses. I. Normal distributions*. Ann. Math. Statistics, 19:495–516, 1948, ISSN 0003-4851. `https://doi.org/10.1214/aoms/1177730147`.

[Mas90] Massart, Pascal: *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality*. Ann. Probab., 18(3):1269–1283, 1990, ISSN 0091-1798.

[SH06] Sachs, L. und J. Hedderich: *Angewandte Statistik: Methodensammlung mit R*. Springer Berlin Heidelberg, 2006, ISBN 9783540321613. `https://books.google.de/books?id=MLElBAAAQBAJ`.

[Sto84] Stone, Charles J.: *An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates*. Ann. Statist., 12(4):1285–1297, 1984.

[vdV98] Vaart, Aad W. van der: *Asymptotic statistics*. Cambridge studies in statistical and probabilistic mathematics. Cambridge Univ. Press, Cambridge [u.a.], 1998, ISBN 0-521-49603-9. Literaturverz. S. 433 - 438.

[Was04] Wasserman, Larry: *All of Statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2004, ISBN 0-387-40272-1. https://doi.org/10.1007/978-0-387-21736-9, A concise course in statistical inference.

Anhang A

Ergänzungen / Hintergrundmaterial

A.1 Ein Steilkurs zur bedingten Verteilung / bedingten Erwartung

Siehe auch [Kle06, Kap. 8]

Erinnerung A.1.1. Sei $(\Omega, \mathscr{F}, \mathbb{P})$ W'raum, $B \in \mathscr{F}$ mit $\mathbb{P}(B) > 0$.

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad A \in \mathscr{F}$$

heißt die *bedingte Wahrscheinlichkeit von A, gegeben B*.

$\mathbb{P}(\cdot \mid B)$ definiert durch diese Formel ein W'Maß auf $(\Omega, \mathscr{F})$ mit $\mathbb{P}(B \mid B) = 1$, für $Y \in \mathcal{L}^1(\mathbb{P})$ ist

$$\mathbb{E}[Y \mid B] = \int Y(\omega)\, \mathbb{P}(d\omega \mid B) = \frac{\mathbb{E}[\mathbf{1}_B Y]}{\mathbb{P}(B)}.$$

Bemerkung A.1.2 (Diskreter Fall der bedingten Erwartung)**.** X und Y ZVn (auf einem W'raum $(\Omega, \mathscr{F}, \mathbb{P})$), $Y \in \mathcal{L}^1(\mathbb{P})$, X nehme höchstens abzählbar viele Werte $x_1, x_2, \dots$ an. Setze

$$f(x) := \mathbb{E}[Y \mid \{X = x\}] \quad \text{für } x \in \{x_1, x_2, \dots\}$$

und $\mathbb{E}[Y \mid X] := f(X)$.

Offenbar ist $\mathbb{E}[Y \mid X]$ $\sigma(X)$-messbar (es ist eine Funktion von X) und für $A \in \sigma(X)$ gilt

$$\mathbb{E}[Y \mathbf{1}_A] = \mathbb{E}\big[\mathbb{E}[Y \mid X] \mathbf{1}_A\big]$$

(klar für $A = \{X = x_i\}$ und dann auch für $A = \{X \in B\}$ mit $B \subset \{x_1, x_2, \dots\}$; jedes $A \in \sigma(X)$ hat diese Form).

Wir betrachten im folgenden einen W'raum $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition A.1.3. $X \in \mathcal{L}^1(\mathbb{P})$, $\mathscr{G} \subset \mathscr{F}$ Teil-σ-Algebra. Eine reellwertige ZV Y heißt (eine Version der) bedingte(n) Erwartung von X gegeben $\mathscr{G}$ (geschrieben $\mathbb{E}[X \mid \mathscr{G}]$), wenn gilt

i) Y ist $\mathscr{G}$-messbar,

ii) $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$ für alle $A \in \mathscr{G}$.

Bemerkung. Äquivalent kann ii) durch ii') ersetzt werden:

ii') $\mathbb{E}[Y \cdot H] = \mathbb{E}[X \cdot H]$ für alle reellwertigen, beschr., $\mathscr{G}$-messbaren ZV H.

Falls $\mathscr{G} = \sigma(Z)$ für eine Zufallsvariable Z, so schreibt man oft auch $\mathbb{E}[X \mid Z] := \mathbb{E}[X \mid \sigma(Z)]$.

Satz A.1.4. *Für $X \in \mathscr{L}^1(\mathbb{P})$ existiert $\mathbb{E}[X \mid \mathscr{G}]$ und ist eindeutig (bis auf $\mathbb{P}$-f.s.-Gleichheit).*

Beweis von Satz A.1.4 (Eindeutigkeit). Seien $Y, \widetilde{Y}$ bedingte Erwartungen.

$$\mathbb{E}[Y \cdot \mathbf{1}_{\{Y>\widetilde{Y}\}}] = \mathbb{E}[X \cdot 1_{\{Y>\widetilde{Y}\}}] = \mathbb{E}[\widetilde{Y} \cdot \mathbf{1}_{\{Y>\widetilde{Y}\}}],$$

also $\mathbb{E}\big[(Y-\widetilde{Y}) \cdot \mathbf{1}_{\{Y>\widetilde{Y}\}}\big] = 0 \Rightarrow \mathbb{P}(Y > \widetilde{Y}) = 0$; analog gilt $\mathbb{P}(\widetilde{Y} > Y) = 0$, also $Y = \widetilde{Y}$ $\mathbb{P}$-f.s. □

Satz A.1.5. *$\mathscr{H} \subset \mathcal{L}^2(\mathbb{P})$ abgeschlossener Unterraum. Zu $X \in \mathcal{L}^2(\mathbb{P})$ gibt es (bis auf $\mathbb{P}$-f.s. Gleichheit) genau ein $Y \in \mathscr{H}$ mit*

i) $\|Y - X\|_2 = \inf\{\|W - X\|_2 : W \in \mathscr{H}\}$

ii) $\mathbb{E}\big[(X-Y)Z\big] = 0$ *für alle* $Z \in \mathscr{H}$

Y ist (die Äquivalenzklasse) der orthogonalen Projektion von X auf $\mathscr{H}$, auch $\mathrm{Proj}_{\mathscr{H}}(X)$ *geschrieben.*

Beweis. Wähle $Y_n \in \mathscr{H}$ mit

$$\|X - Y_n\|_2 \underset{n\to\infty}{\longrightarrow} \alpha := \inf\{\|W - X\|_2 : W \in \mathscr{H}\}$$

Es ist

$$\|X - Y_n\|_2^2 + \|X - Y_m\|_2^2 = 2\left\|X - \frac{1}{2}(Y_n + Y_m)\right\|^2 + 2\left\|\frac{1}{2}(Y_n - Y_m)\right\|^2$$

wegen $\lim\limits_{n\to\infty} \|X - Y_n\|_2^2 = \lim\limits_{m\to\infty} \|X - Y_m\|_2^2 = \alpha^2 \le \liminf\limits_{n,m\to\infty} \left\|X - \frac{1}{2}(Y_n + Y_m)\right\|^2$ folgt

$$\limsup_{n,m\to\infty} \|Y_n - Y_m\|_2 = 0,$$

d.h. $(Y_n)_n \subset \mathcal{L}^2(\mathbb{P})$ ist Cauchy-Folge.

Demnach gibt es $Y \in \mathcal{L}^2(\mathbb{P})$ mit $Y_n \overset{\mathcal{L}^2(\mathbb{P})}{\underset{n\to\infty}{\longrightarrow}} Y$ ($\mathcal{L}^2(\mathbb{P})$ ist vollständig). Wähle Teilfolge $(n_k)_k$ mit $Y = \lim\limits_{k\to\infty} Y_{n_k}$ $\mathbb{P}$-f.s., dann ist $Y \in \mathscr{H}$ mit

$$\alpha^2 \le \mathbb{E}\big[(X-Y)^2\big] \le \liminf_{n\to\infty} \mathbb{E}\big[(X-Y_n)^2\big] = \alpha^2$$

Sei $Z \in \mathscr{H}$, für $t \in \mathbb{R}$ ist

$$\mathbb{E}\big[(X - (Y - tZ))^2\big] \ge \mathbb{E}\big[(X-Y)^2\big]$$

(nach Wahl von Y), also

$$2t\,\mathbb{E}\big[(X-Y)Z\big] + t^2\mathbb{E}[Z^2] \ge 0 \quad \text{für alle } t \in \mathbb{R},$$

was $\mathbb{E}\big[(X-Y)Z\big]=0$ erzwingt.

(Übrigens: Für $Y\in\mathscr{H}$ gilt *i)* $\Leftrightarrow$ *ii).*)

Zur Eindeutigkeit: $\widetilde{Y}$ erfülle ebenfalls *i)*, es ist

$$\begin{aligned}\Big(X-\frac{1}{2}(Y+\widetilde{Y})\Big)^2 &= \Big(\frac{1}{2}(X-Y)+\frac{1}{2}(X-\widetilde{Y})\Big)^2\\ &\le \frac{1}{2}(X-Y)^2+\frac{1}{2}(X-\widetilde{Y})^2,\end{aligned}$$

die Ungleichung ist strikt auf $\{Y\neq\widetilde{Y}\}$.

Wäre $\mathbb{P}(Y\neq\widetilde{Y})>0$, so wäre

$$\mathbb{E}\Big[\Big(X-\frac{1}{2}(Y+\widetilde{Y})\Big)^2\Big]<\frac{1}{2}\mathbb{E}\big[(X-Y)^2\big]+\frac{1}{2}\mathbb{E}\big[(X-\widetilde{Y})^2\big]=\frac{1}{2}\alpha+\frac{1}{2}\alpha=\alpha,$$

ein Widerspruch, da $(Y+\widetilde{Y})/2\in\mathscr{H}$. □

Beweis von Satz A.1.4 (Existenz). Sei zunächst $X\in\mathcal{L}^2(\mathbb{P})$,

$$\mathscr{H}:=\big\{Y\in\mathcal{L}^2(\mathbb{P}):Y\text{ ist }\mathscr{G}\text{-messbar}\big\}\subset\mathcal{L}^2(\mathbb{P})$$

ist ein abgeschlossener Unterraum, $Y:=\mathrm{Proj}_{\mathscr{H}}(X)$ (nach Satz A.1.5) leistet das Gewünschte.

Beachte:

$$X\ge 0\ \Rightarrow\ Y:=\mathbb{E}[X\,|\,\mathscr{G}]\ge 0\quad\text{f.s.},$$

denn dann ist $0\le\mathbb{E}[X\mathbf{1}_{\{Y<0\}}]=\mathbb{E}[Y\mathbf{1}_{\{Y<0\}}]$.

Sei nun $X\in\mathcal{L}^1(\mathbb{P})$, $X\ge 0$:

$$Y_n:=\mathbb{E}\big[X\wedge n\,\big|\,\mathscr{G}\big]\underset{n\to\infty}{\nearrow}Y\ (\ge 0)\quad\text{(f.s.)}$$

($X\wedge n\in\mathcal{L}^1(\mathbb{P})$, nutze dann obige Monotonie-Eigenschaft, um $\mathbb{E}[X\wedge n\,|\,\mathscr{G}]\le\mathbb{E}[X\wedge(n+1)\,|\,\mathscr{G}]$ f.s. zu sehen), für $A\in\mathscr{G}$ gilt

$$\mathbb{E}[X\mathbf{1}_A]=\lim_{n\to\infty}\mathbb{E}\big[(X\wedge n)\mathbf{1}_A\big]=\lim_{n\to\infty}\mathbb{E}[Y_n\mathbf{1}_A]=\mathbb{E}[Y\mathbf{1}_A]$$

(mit monotoner Konvergenz).

Für allgemeines $X\in\mathcal{L}^1(\mathbb{P})$ leistet

$$Y:=\mathbb{E}[X^+\,|\,\mathscr{G}]-\mathbb{E}[X^-\,|\,\mathscr{G}]$$

das Gewünschte. □

Satz A.1.6. *$X,Y\in\mathcal{L}^1(\mathbb{P})$, $\mathscr{G}\subset\mathscr{F}$ Teil-σ-Algebra.*

i) $\mathbb{E}[aX+bY\,|\,\mathscr{G}]=a\mathbb{E}[X\,|\,\mathscr{G}]+b\mathbb{E}[Y\,|\,\mathscr{G}]$ *f.s. für* $a,b\in\mathbb{R}$ *(Linearität)*

ii) $X\le Y$ *f.s.* $\Rightarrow$ $\mathbb{E}[X\,|\,\mathscr{G}]\le\mathbb{E}[Y\,|\,\mathscr{G}]$ *f.s. (Monotonie)*

iii) $\big|\mathbb{E}[X \mid \mathscr{G}]\big| \le \mathbb{E}\big[|X| \big| \mathscr{G}\big]$ *f.s.* *(Dreiecksungleichung)*

iv) Es gelte $\mathbb{E}\big[|XY|\big] < \infty$ *und* Y *sei* $\mathscr{G}$*-messbar, dann ist*

$$\mathbb{E}[XY \mid \mathscr{G}] = Y \cdot \mathbb{E}[X \mid \mathscr{G}] \text{ f.s.},$$

insbesondere ist $\mathbb{E}[Y|\mathscr{G}] = Y$ *f.s.* *(„Herausziehen von Bekanntem“)*

v) Sind $\sigma(X)$ *und* $\mathscr{G}$ *unabhängig, so ist* $\mathbb{E}[X \mid \mathscr{G}] = \mathbb{E}[X]$ *f.s.* *(Verhalten bei Unabhängigkeit)*

vi) $\mathscr{G}' \subset \mathscr{G}$ *Teil-*σ*-Algebra, so ist*

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathscr{G}] \big| \mathscr{G}'\big] = \mathbb{E}[X \mid \mathscr{G}'] \text{ f.s.},$$

(„Turmeigenschaft“) insbesondere ist

$$\mathbb{E}[\mathbb{E}[X \mid \mathscr{G}]] = \mathbb{E}[X]$$

vii) $0 \le X_n \nearrow X$ *f.s. für* $n \to \infty$*, so gilt*

$$\mathbb{E}[X_n \mid \mathscr{G}] \underset{n\to\infty}{\nearrow} \mathbb{E}[X \mid \mathscr{G}] \quad \text{f.s. und in } \mathcal{L}^1(\mathbb{P})$$

(monotone Konvergenz)

viii) X_n *reelle ZVn mit* $|X_n| \le Y \ \forall\, n$ *und* $X_n \to X$ *f.s., so gilt*

$$\mathbb{E}[X_n \mid \mathscr{G}] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[X \mid \mathscr{G}] \quad \text{f.s. und in } \mathcal{L}^1(\mathbb{P})$$

(dominierte Konvergenz)

Beweis. (Die folgenden Gleichungen, etc. gelten jeweils f.s., auch wenn wir dies in der Notation nicht explizit machen.)

i) $a\mathbb{E}[X \mid \mathscr{G}] + b\mathbb{E}[Y \mid \mathscr{G}]$ ist $\mathscr{G}$-messb., und für $A \in \mathscr{G}$ gilt

$$\begin{aligned}\mathbb{E}\Big[\mathbf{1}_A\big(a\mathbb{E}[X \mid \mathscr{G}] + b\mathbb{E}[Y \mid \mathscr{G}]\big)\Big] &= a\mathbb{E}\big[\mathbf{1}_A\mathbb{E}[X \mid \mathscr{G}]\big] + b\mathbb{E}\big[\mathbf{1}_A\mathbb{E}[Y \mid \mathscr{G}]\big] \\ &= a\mathbb{E}[\mathbf{1}_A X] + b\mathbb{E}[\mathbf{1}_A Y] = \mathbb{E}\big[\mathbf{1}_A(aX + bY)\big]\end{aligned}$$

ii) $A := \big\{\mathbb{E}[X \mid \mathscr{G}] > \mathbb{E}[Y \mid \mathscr{G}]\big\} \in \mathscr{G}$ und

$$0 \le \mathbb{E}\big[\mathbf{1}_A \underbrace{(Y - X)}_{\ge 0 \text{ f.s.}}\big] = \mathbb{E}\Big[\mathbf{1}_A\big(\mathbb{E}[Y \mid \mathscr{G}] - \mathbb{E}[X \mid \mathscr{G}]\big)\Big] \le 0,$$

also $\mathbb{P}(A) = 0$.

iii) $\big|\mathbb{E}[X \mid \mathscr{G}]\big| \overset{i)}{=} \big|\mathbb{E}[X^+ \mid \mathscr{G}] - \mathbb{E}[X^- \mid \mathscr{G}]\big| \le \mathbb{E}[X^+ \mid \mathscr{G}] + \mathbb{E}[X^- \mid \mathscr{G}] \overset{i)}{=} \mathbb{E}\big[|X| \big| \mathscr{G}\big]$

iv) Seien zunächst $X, Y \ge 0$, $Y_n := 2^{-n}\lfloor 2^n Y \rfloor \wedge n$ $(\nearrow Y)$.

Für $A \in \mathscr{G}$ ist

$$\begin{aligned}\mathbb{E}\big[\mathbf{1}_A Y \mathbb{E}[X \mid \mathscr{G}]\big] &= \lim_{n\to\infty} \mathbb{E}\big[\mathbf{1}_A Y_n \mathbb{E}[X \mid \mathscr{G}]\big] \\ &= \lim_{n\to\infty} \mathbb{E}\Big[\mathbf{1}_A \sum_{k=0}^{n2^n} \frac{k}{2^n} \mathbf{1}_{\{Y=k/2^n\}} \mathbb{E}[X \mid \mathscr{G}]\Big] \\ &= \lim_{n\to\infty} \sum_{k=0}^{n2^n} \frac{k}{2^n} \mathbb{E}\big[\mathbf{1}_{A\cap\{Y=k/2^n\}} X\big] \\ &= \lim_{n\to\infty} \mathbb{E}\big[\mathbf{1}_A Y_n X\big] = \mathbb{E}[\mathbf{1}_A Y X]\end{aligned}$$

Für den allg. Fall zerlege $X = X^+ - X^-$, $Y = Y^+ - Y^-$, verwende *i)*.

v) Für $A \in \mathscr{G}$ gilt $\mathbb{E}[\mathbf{1}_A X] = \mathbb{E}[\mathbf{1}_A]\mathbb{E}[X] = \mathbb{E}\big[\mathbf{1}_A \mathbb{E}[X]\big]$

vi) Sei $A \in \mathscr{G}'$

$$\mathbb{E}\big[\mathbf{1}_A \mathbb{E}[X \mid \mathscr{G}]\big] = \mathbb{E}[\mathbf{1}_A X] = \mathbb{E}\big[\mathbf{1}_A \mathbb{E}[X \mid \mathscr{G}']\big]$$

d.h. $\mathbb{E}[X \mid \mathscr{G}']$ erfüllt die Def. von $\mathbb{E}\big[\mathbb{E}[X \mid \mathscr{G}] \,\big|\, \mathscr{G}'\big]$.

Zum Zusatz: Stets ist $\mathbb{E}[X \mid \{\varnothing, \Omega\}] = \mathbb{E}[X]$

vii) $\mathbb{E}[X_n \mid \mathscr{G}]$ ist (f.s.) nicht-fallend in n nach *ii)*,

$$\mathbb{E}\Big[\big|\mathbb{E}[X \mid \mathscr{G}] - \mathbb{E}[X_n \mid \mathscr{G}]\big|\Big] \overset{iii)}{\leq} \mathbb{E}\Big[\mathbb{E}\big[|X - X_n| \,\big|\, \mathscr{G}\big]\Big] \overset{iv)}{=} \mathbb{E}\big[|X - X_n|\big] \underset{n\to\infty}{\longrightarrow} 0$$

also $\mathbb{E}[X_n \mid \mathscr{G}] \to \mathbb{E}[X \mid \mathscr{G}]$ in $\mathcal{L}^1(\mathbb{P})$.

Weiterhin: Für monotone Folgen von ZVn impliziert $\mathcal{L}^1(\mathbb{P})$-Konvergenz f.s.-Konvergenz: Seien $Z_n \nearrow Z$ und $Z \to Z_n$ in $\mathcal{L}^1(\mathbb{P})$, dann auch $Z_n \nearrow Z = \sup_m Z_m$ stochastisch, $\mathbb{P}\Big(\bigcap_{m\geq n} \big\{|Z_m - Z| < \varepsilon\big\}\Big) = \mathbb{P}(Z - \varepsilon < Z_n) \longrightarrow_{n\to\infty} 1$, somit

$$\mathbb{P}\Big(\bigcup_n \bigcap_{m\geq n} \big\{|Z_m - Z| < \varepsilon\big\}\Big) = 1 \quad \text{für jedes } \varepsilon > 0.$$

viii) $0 \leq Z_n := \sup_{m\geq n} |X_m - X|$ $(\leq 2Y)$, $Z_n \searrow 0$ f.s. für $n \to \infty$.

$$\mathbb{E}\Big[\big|\mathbb{E}[X \mid \mathscr{G}] - \mathbb{E}[X_n \mid \mathscr{G}]\big|\Big] \overset{iii)}{\leq} \mathbb{E}\Big[\mathbb{E}\big[|X - X_n| \,\big|\, \mathscr{G}\big]\Big] \overset{iv)}{=} \mathbb{E}\big[|X - X_n|\big] \leq \mathbb{E}[Z_n] \underset{n\to\infty}{\longrightarrow} 0,$$

also $\mathbb{E}[X_n \mid \mathscr{G}] \to \mathbb{E}[X \mid \mathscr{G}]$ in $\mathcal{L}^1(\mathbb{P})$. Weiter ist

$$\mathbb{E}\Big[\lim_{n\to\infty} \mathbb{E}[Z_n \mid \mathscr{G}]\Big] \leq \liminf_{n\to\infty} \mathbb{E}[Z_n] = 0$$

d.h. $\mathbb{E}[Z_n \mid \mathscr{G}] \to 0$ f.s., somit

$$\big|\mathbb{E}[X \mid \mathscr{G}] - \mathbb{E}[X_n \mid \mathscr{G}]\big| \leq \mathbb{E}\big[|X - X_n| \,\big|\, \mathscr{G}\big] \leq \mathbb{E}[Z_n \mid \mathscr{G}]\Big] \underset{n\to\infty}{\longrightarrow} 0 \quad \text{f.s.}$$

□

Satz A.1.7 (Jensen'sche Ungleichung für die bedingte Erwartung)**.** *$X \in \mathcal{L}^1(\mathbb{P})$, $k : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ konvex, dann gilt*

$$\mathbb{E}\big[k(X) \,\big|\, \mathscr{G}\big] \geq k\big(\mathbb{E}[X \mid \mathscr{G}]\big) \quad \textit{f.s.}$$

Beweis. Die Behauptung ist offenbar erfüllt (mit Satz A.1.6, *i)*), wenn k affin-linear ist, d.h. $k(x) = ax + b$ mit gewissen $a, b \in \mathbb{R}$.

Sei nun k konvex und nicht von dieser Form. Man kann die Funktion k als das Supremum ihrer Stützgeraden schreiben, d.h.

$$k(x) = \sup\{ax + b : (a,b) \in S\} \quad \text{mit } S := \{(a,b) \in \mathbb{R}^2 : ay + b \le k(y)\ \forall\, y \in \mathbb{R}\}$$

und es gilt auch ($(a,b) \mapsto ax + b$ ist stetig für jedes $x \in \mathbb{R}$ und da k keine Gerade ist, kann man zu $(a,b) \in S$ gewisse $(a_m, b_m) \in S \cap \mathbb{Q}^2$ finden mit $a_m \to a, b_m \to b$ für $m \to \infty$)

$$k(x) = \sup\{ax + b : (a,b) \in S \cap \mathbb{Q}^2\}.$$

Sei $(a_n, b_n)_{n\in\mathbb{N}}$ eine Aufzählung von $S \cap \mathbb{Q}^2$, für jedes n gilt

$$\mathbb{E}\big[k(X)\,\big|\,\mathscr{G}\big] \ge \mathbb{E}\big[a_n X + b_n\,\big|\,\mathscr{G}\big] = a_n \mathbb{E}[X \mid \mathscr{G}] + b_n \quad \text{f.s.}$$

(mit Satz A.1.6, *ii)* und *i)*), daher auch (man muss oben nur abzählbar viele Ausnahmemengen betrachten)

$$\mathbb{E}\big[k(X)\,\big|\,\mathscr{G}\big] \ge \sup\big\{a_n \mathbb{E}[X \mid \mathscr{G}] + b_n : n \in \mathbb{N}\big\} = k\big(\mathbb{E}[X \mid \mathscr{G}]\big) \quad \text{f.s.}$$

□

Bemerkung A.1.8. X, Y reelle ZVn mit gemeinsamer Dichte $f_{X,Y}$, d.h.

$$\mathbb{P}\big((X,Y) \in A\big) = \int_A f_{X,Y}(x,y)\, \lambda^{\otimes 2}\big(d(x,y)\big) \quad \text{für } A \in \mathcal{B}(\mathbb{R}^2),$$

$Y \in \mathcal{L}^1(\mathbb{P})$.

Sei für $x \in \mathbb{R}$

$$f_X(x) := \int_{\mathbb{R}} f_{X,Y}(x,y)\, \lambda(dy) \quad \text{die Marginaldichte von } X,$$
$$\varphi(x) := \int_{\mathbb{R}} y \frac{f_{X,Y}(x,y)}{f_X(x)}\, \lambda(dy) \mathbf{1}_{f_X(x)>0}.$$

Dann gilt

$$\mathbb{E}[Y|\sigma(X)] = \varphi(X) \ \text{ f.s.}$$

denn für $B \in \mathcal{B}(\mathbb{R})$ ist

$$\begin{aligned}
\mathbb{E}\big[\mathbf{1}_{\{X\in B\}}\varphi(X)\big] &= \int_{\mathbb{R}} \mathbf{1}_B(x)\varphi(x) f_X(x)\, \lambda(dx) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbf{1}_B(x) y f_{X,Y}(x,y) \lambda(dy)\ \lambda(dx) \\
&= \int_{\mathbb{R}^2} \mathbf{1}_B(x) y f_{X,Y}(x,y)\, \lambda^{\otimes 2}\big(d(x,y)\big) = \mathbb{E}\big[\mathbf{1}_{\{X\in B\}} Y\big].
\end{aligned}$$

Bericht A.1.9. (Zu regulären Versionen bedingter Verteilungen) Wenn $Y = 1_B$ für ein Ereignis $B \in \mathscr{A}$, so schreibt man gelegentlich auch $\mathbb{P}(B \mid \mathscr{G}) = \mathbb{E}[Y \mid \mathscr{G}]$. Man muss allerdings etwas vorsichtig sein bei der Interpretation von $\mathbb{P}(\cdot \mid \mathscr{G})$ als ein (zufälliges) Maß, da i.A. überabzählbar viele B in Frage kommen und damit die Kompatibilität der in der Definition der bedingten Erwartung implizit vorkommenden Nullmengen (vgl. Def. A.1.3) wenigstens a priori unklar bleibt.

In „gutartigen" Fällen ist eine konsistente Wahl möglich, das Stichwort dazu lautet „reguläre bedingte Verteilung einer Zufallsvariable". Wir skizzieren hier knapp den reellwertigen Fall:

Sei X reellwertige ZV (auf einem W'raum $(\Omega, \mathscr{F}, \mathbb{P})$), $\mathscr{G} \subset \mathscr{F}$ Teil-σ-Algebra. Dann gibt es einen stochastischen Kern $\kappa_{X|\mathscr{G}}$ von $(\Omega, \mathscr{G})$ nach $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit

$$\kappa_{X|\mathscr{G}}(\omega, B) = \mathbb{E}\big[\mathbf{1}_{\{X \in B\}} \,\big|\, \mathscr{G}\big](\omega) \text{ f.s. für alle } B \in \mathcal{B}(\mathbb{R}),$$

d.h. (vgl. [Bir17, Def. 5.4] oder [Kle06, Def. 8.24])

> für jedes $B \in \mathcal{B}(\mathbb{R})$ ist $\omega \mapsto \kappa_{X|\mathscr{G}}(\omega, B)$ eine Version von $\mathbb{E}\big[\mathbf{1}_{\{X \in B\}} \,\big|\, \mathscr{G}\big]$ und
> für jedes ω ist $\kappa_{X|\mathscr{G}}(\omega, \cdot)$ ein W'maß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Die Hauptidee besteht darin, das gewünschte Maß $\kappa_{X|\mathscr{G}}(\omega, \cdot)$ auf $\mathbb{R}$ anhand seiner Verteilungsfunktion zu charakterisieren (vgl. [Bir17, Satz 1.27] oder [Kle06, Bsp. 1.44]) die zielführende Beobachtung ist dann, dass eine Verteilungsfunktion (wegen der Monotonie) bereits durch ihre Werte an abzählbar vielen Stellen festgelegt ist. Man betrachtet also $B = (-\infty, r], r \in \mathbb{Q}$ und setzt

$$F_r := \mathbb{E}[\mathbf{1}_{(-\infty, r]}(X) \mid \mathscr{G}].$$

Dann gilt (mit den Eigenschaften der bedingten Erwartung aus Satz A.1.6) wie gewünscht $\mathbb{P}$-f.s.: $F_r \le F_{r'}$, für $r < r', (r, r' \in \mathbb{Q})$, $\lim_{n\to\infty} F_{r+\frac{1}{n}} = F_r$ für $r \in \mathbb{Q}$, $\lim_{n\to\infty} F_n = 1$, $\lim_{n\to-\infty} F_n = 0$.

Wegen der Abzählbarkeit von $\mathbb{Q}$ gibt es $N \in \mathscr{F}$ mit $\mathbb{P}(N) = 0$, so dass obiges für $\omega \in \Omega \setminus N$ und alle $r, r' \in \mathbb{Q}$ gilt. Dann definiert

$$\widetilde{F}_s := \begin{cases} \inf\{F_r : r \ge s, r \in \mathbb{Q}\}, & \omega \in \Omega \setminus N, \\ \overline{F}_s, & \omega \in N, \end{cases}$$

wobei $\overline{F}$ irgendeine Verteilungsfunktion ist, die (zufällige) Verteilungsfunktion von $k_{X,\mathscr{G}}$. Details finden sich beispielsweise in [Kle06, Kap. 8.3, speziell Satz 8.28].

Man kann dieses Argument relativ leicht erweitern auf die Situation, dass der Wertebereich $(E, \mathscr{B})$ von X ein sogenannter Standard-Borel-Raum ist (auch der Name Borel'scher Raum ist üblich), d.h. wenn es ein $A \in \mathcal{B}(\mathbb{R})$ und eine Bijektion $\phi : E \to A$ gibt, so dass ϕ und ϕ^{-1} jeweils messbar sind (dann sind $(E, \mathscr{B})$ und $(A, \mathcal{B}(A))$ isomorph als messbare Räume). Dann ist nämlich $X' := \phi \circ X$ eine reellwertige ZV, und die Argumentation oben greift (vgl. auch [Kle06, Satz 8.36]).

Schließlich kann man zeigen, dass jeder separable und vollständige metrische Raum E, versehen mit seiner Borel-σ-Algebra, ein Standard-Borel-Raum ist (siehe z.B. L.C.G. Rogers, D. Williams, *Diffusions, Markov processes and martingales*, Bd. 1, Ch. II.82; L. Breiman,

Probability, Appendix 7). Solche Wertebereiche heißen *polnische Räume*, sie spielen in der allgemeinen Theorie der Stochastik eine wichtige Rolle (beispielsweise sind $\mathbb{R}^d$ oder $C([0,1])$ mit Supremumsnorm polnisch).

A.2 Rund um die multivariate Normalverteilung

Wir benötigen einige Eigenschaften der multivariaten Normalverteilung.

Proposition A.2.1. *$n \in \mathbb{N}$, $X_1, X_2, \ldots, X_n$ u.i.v. $\sim \mathcal{N}_{0,1}$, so hat*

$$X := X_1^2 + X_2^2 + \cdots + X_n^2 \quad \textit{die Dichte } \tfrac{1}{\Gamma(n/2)} 2^{-n/2} x^{\frac{n}{2}-1} e^{-x/2} \mathbf{1}_{[0,\infty)}(x),$$

$\chi_n^2 := \mathscr{L}(X)$ *heißt* Chiquadrat-Verteilung *mit n Freiheitsgraden.*

Beachte: $\chi_n^2 = \Gamma_{1/2,n/2}$, wo die Gamma-Verteilung $\Gamma_{\alpha,\nu}$ die Dichte

$$\frac{1}{\Gamma(\nu)} \alpha^\nu x^{\nu-1} e^{-\alpha x} \mathbf{1}_{(0,\infty)}(x)$$

besitzt (α = Skalen-, ν = Formparameter)

Proposition A.2.2. *Seien $\alpha, r, s > 0$, $X \sim \Gamma_{\alpha,r}, Y \sim \Gamma_{\alpha,s}$ unabhängig. Dann sind*

$$X + Y \;\; \textit{und} \;\; V := \frac{X}{X+Y} \;\; \textit{unabhängig}$$

und $X + Y \sim \Gamma_{\alpha,r+s}$, $V \sim \beta_{r,s}$, wobei die Beta-Verteilung $\beta_{r,s}$ die Dichte

$$\tfrac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} v^{r-1}(1-v)^{s-1} \mathbf{1}_{(0,1)}(v)$$

besitzt.

Insbesondere bilden die Gamma-Verteilungen eine Faltungsfamilie (bezüglich des zweiten, des sogenannten Formparameters): $\Gamma_{\alpha,r} * \Gamma_{\alpha,s} = \Gamma_{\alpha,r+s}$.

Beweis. (X,Y) hat Dichte

$$f_{(X,Y)}(x,y) = \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} x^{r-1} y^{s-1} e^{-\alpha(x+y)} \text{ auf } (0,\infty)^2.$$

Sei $\varphi(x,y) = \begin{pmatrix} x+y \\ \frac{x}{x+y} \end{pmatrix}$, so ist

$$\varphi^{-1}(z,v) = \begin{pmatrix} zv \\ z(1-v) \end{pmatrix}, \quad D\varphi(x,y) = \begin{pmatrix} 1 & \frac{y}{(x+y)^2} \\ 1 & -\frac{x}{(x+y)^2} \end{pmatrix}, \quad |\det D\varphi(x,y)| = \frac{|x+y|}{(x+y)^2} = \frac{1}{|x+y|}$$

Schreibe $Z := X + Y$, $V := \frac{X}{X+Y}$. Gemäß 2-dimensionaler Dichtetransformation (siehe Bericht A.2.7) ist die Dichte von (Z,V):

$$\begin{aligned} f_{(Z,V)}(z,v) &= \frac{f_{(X,Y)}(\varphi^{-1}(z,v))}{|\det D\varphi(\varphi^{-1}(z,v))|} \\ &= z \cdot \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} (zv)^{r-1} (z(1-v))^{s-1} e^{-\alpha z} \\ &= \underbrace{\frac{\alpha^{r+s}}{\Gamma(r+s)} z^{r+s-1} e^{-\alpha z}}_{\text{Dichte von } \Gamma_{\alpha,r+s}} \underbrace{\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} v^{r-1}(1-v)^{s-1}}_{\text{Dichte von } \beta_{r,s}} \end{aligned}$$

□

Beweis von Beob. A.2.1. $X \sim \mathcal{N}_{0,1}$, so ist $X^2 \sim \Gamma_{\frac{1}{2},\frac{1}{2}}$ $(= \chi_1^2)$:

$|X|$ hat Dichte $\frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ auf $(0,\infty)$; sei $\varphi : (0,\infty) \to (0,\infty), x \mapsto x^2$, $\varphi^{-1}(y) = \sqrt{y}$, $\frac{d}{dy}\varphi^{-1}(y) = 2y$, also hat X^2 die Dichte $\frac{1}{2\sqrt{y}} \frac{2}{\sqrt{2\pi}} e^{-\frac{y}{2}} = \left(\frac{1}{2}\right)^{\frac{1}{2}} \frac{1}{\Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} e^{-\frac{1}{2}y}$ (siehe Beob. A.2.5).

Dies zeigt die Behauptung für $n = 1$, der allgemeine Fall folgt daraus induktiv unter Verwendung von Proposition A.2.2. □

Korollar und Definition A.2.3. Seien $m, n \in \mathbb{N}$, $X_1, \dots, X_m, Y_1, \dots, Y_n$ unabhängig, $\sim \mathcal{N}_{0,1}$.

1. $F_{m,n} := \dfrac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2}$ hat Dichte $f_{m,n}(x) = \dfrac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \dfrac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{(m+n)}{2}}} \mathbf{1}_{(0,\infty)}(x)$.

$\mathscr{L}(F_{m,n})$ heißt *Fisher-Verteilung*[1] mit m und n Freiheitsgraden (präziser: mit m Zähler- und n Nenner-Freiheitsgraden).

2. $T_n := \dfrac{X}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} Y_j^2}}$ hat Dichte $t_n(x) = \dfrac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})} \dfrac{1}{\sqrt{n}} \left(1 + \dfrac{x^2}{n}\right)^{-\frac{n+1}{2}}$.

$\mathscr{L}(T_n)$ heißt *Student-Verteilung*[2] mit n Freiheitsgraden (auch Student'sche T-Verteilung genannt).

Bemerke: Die Student-Verteilung mit einem Freiheitsgrad ist die Cauchy-Verteilung.

Bemerkung. Sei T_n Student-verteilt mit n Freiheitsgraden, so ist $T_n \xrightarrow[n\to\infty]{d} \mathcal{N}_{0,1}$.
(denn es gilt $t_n(x) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ lokal gleichmäßig).

Beweis von Korollar A.2.3. 1. $X := \sum_{i=1}^{m} X_i^2 \sim \Gamma_{\frac{1}{2},\frac{m}{2}}$, $Y := \sum_{j=1}^{n} Y_j^2 \sim \Gamma_{\frac{1}{2},\frac{n}{2}}$ sind unabhängig, also ist $V := \dfrac{X}{X+Y} \sim \beta_{\frac{m}{2},\frac{n}{2}}$ nach Proposition A.2.2.

Dann ist

$$F_{m,n} = \frac{nX}{mY} = \frac{n}{m} \frac{\frac{X}{X+Y}}{\frac{Y}{X+Y}} = \frac{n}{m} \frac{V}{1-V},$$

mit

$$\varphi : (0,1) \to (0,\infty), v \mapsto \frac{n}{m} \frac{v}{1-v}, \quad \text{also } \varphi^{-1}(z) = \frac{mz}{n+mz}, \quad \frac{d}{dv}\varphi(v) = \frac{n}{m} \frac{1}{(1-v)^2}$$

ist $F_{m,n} = \varphi(V)$, hat also die Dichte

$$\begin{aligned} f_{m,n}(z) &= \frac{mnz}{(n+mz)^2} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{mz}{n+mz}\right)^{\frac{m}{2}-1} \left(\frac{n}{n+mz}\right)^{\frac{n}{2}-1} \\ &= \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{z^{\frac{m}{2}-1}}{(n+mz)^{\frac{(m+n)}{2}}} \end{aligned}$$

[1] Nach Ronald Aylmer Fisher, 1890–1962

[2] Nach William Sealy Gosset, 1876–1937, der sie 1908 unter dem Pseudonym "Student" veröffentlichte.

2. T_n^2 hat (nach 1.) Dichte $f_{1,n}$, also hat $|T_n|$ Dichte $2tf_{1,n}(t^2)\mathbf{1}_{[0,\infty)}(t)$.

Da T_n symmetrisch um 0 verteilt ist (klar aus der Symmetrie von X_1), hat T_n die Dichte

$$\begin{aligned}|t|f_{1,n}(t^2) &= |t|\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}n^{n/2}\frac{(t^2)^{\frac{1}{2}-1}}{(n+t^2)^{(n+1)/2}}\\ &= \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}\frac{1}{\sqrt{n}}\frac{1}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}\end{aligned}$$

□

Proposition A.2.4. *$X_1, \dots, X_n$ u.i.v. $\sim \mathcal{N}_{\mu,\sigma^2}$ mit $\mu \in \mathbb{R}$, $\sigma > 0$,*

$$M := \frac{1}{n}\sum_{i=1}^n X_i, \quad S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - M)^2.$$

Es gilt

1. M und S^2 sind unabhängig, $M \sim \mathcal{N}_{\mu,\sigma^2/n}$, $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

2. $T := \dfrac{\sqrt{n}(M-\mu)}{\sqrt{S^2}}$ ist Student-verteilt mit $n-1$ Freiheitsgraden.

Beweis. Sei o.E. $\mu = 0$, $\sigma^2 = 1$, sonst betrachte $X_i' := (X_i - \mu)/\sqrt{\sigma^2}$.

1. Sei O orthogonale $n \times n$-Matrix, deren erste Zeile $z_1 = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$ ist, d.h. ergänze z_1 zu einer Orthonormalbasis $z_1, \dots, z_n$ von $\mathbb{R}^n$, setze

$$O = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}.$$

Dann ist

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} := O \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

n-dimensional Standardnormalverteilt (nach Beispiel A.2.8, Invarianz der n-dim. Normalverteilung unter orthogonalen Transformationen).

Somit

$$\begin{aligned} Y_1 &= \sum_{i=1}^n \frac{1}{\sqrt{n}} X_i = \sqrt{n}M, \qquad \text{also } M \sim \mathcal{N}_{0,1/n},\\ (n-1)S^2 &= \sum_{i=1}^n (X_i - M)^2 = \sum_{i=1}^n X_i^2 - nM^2\\ &= \|(X_1, \dots, X_n)^T\|^2 - Y_1^2 = \|(Y_1, \dots, Y_n)^T\|^2 - Y_1^2 = \sum_{i=2}^n Y_i^2, \end{aligned}$$

also $(n-1)S^2 \sim \chi^2_{n-1}$ und unabhängig von M.

(Geometrisch ausgedrückt: Zerlege $\mathbb{R}^n = D \oplus D^\perp$ in die Diagonale $D = \{(x, x, \dots, x) : x \in \mathbb{R}\} \subset \mathbb{R}^n$ und ihr orthogonales Komplement $D^\perp = \{(x_1, x_2, \dots, x_n) : x_1 + \cdots + x_n = 0\}$, seien $\mathcal{P}_D : \mathbb{R}^n \to D$, $\mathcal{P}_{D^\perp} = \mathrm{Id}_{\mathbb{R}^n} - \mathcal{P}_D : \mathbb{R}^n \to D^\perp$ die orthogonalen Projektionen auf D bzw. auf $D^\perp$, dann ist $\sqrt{n}M$ die (signierte) Länge von $\mathcal{P}_D X$ und $(n-1)S^2 = \|\mathcal{P}_{D^\perp} X\|^2$.)

2. Dies folgt aus 1. und der Definition (vgl. Korollar und Definition A.2.3, 2.) □

Zur Dichtetransformationsformel

Beobachtung A.2.5 (Dichtetransformation im Fall $\mathbb{R}^1$)**.** X reelle ZV mit Dichte f_X, d.h. $F_X(x) = \int_{-\infty}^{x} f_X(z)\,dz$, $I \subset \mathbb{R}$ (möglicherweise unbeschränktes) offenes Intervall mit $P(X \in I) = 1$, $J \subset I$, $\varphi : I \to J$ stetig differenzierbar, bijektiv.

Dann hat $Y := \varphi(X)$ die Dichte

$$f_Y(y) = \begin{cases} \dfrac{f_X\big(\varphi^{-1}(y)\big)}{\big|\varphi'\big(\varphi^{-1}(y)\big)\big|}, & y \in J, \\ 0, & y \notin J. \end{cases}$$

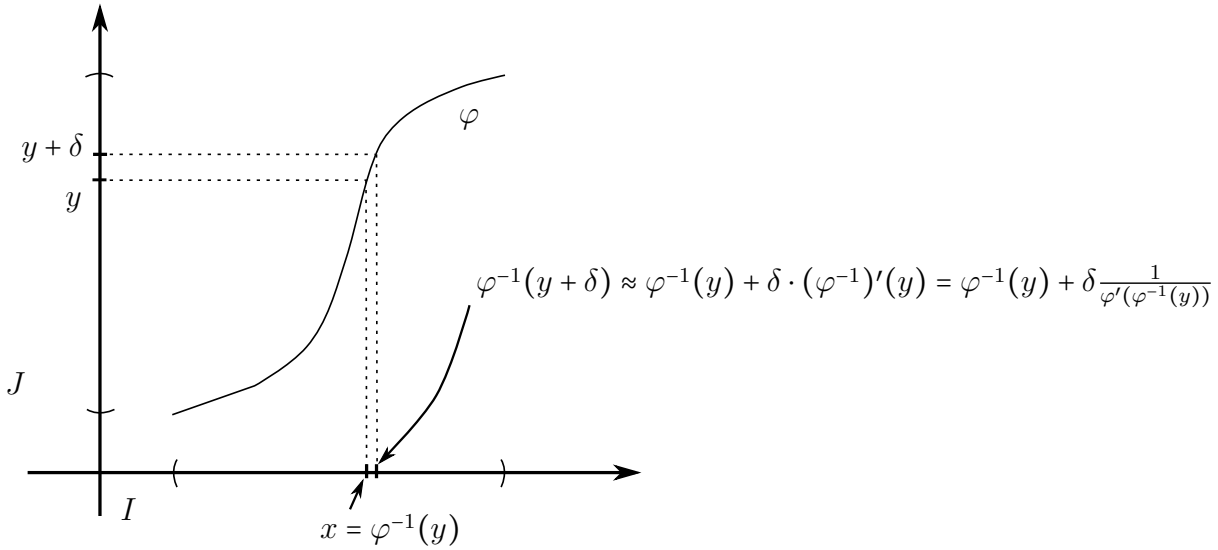
Beweis. φ muss offenbar strikt wachsend oder strikt fallend sein, wir betrachten den wachsenden Fall.

Für $z < \inf J$ ist $P(Y \le z) = 0$, für $z > \sup J$ ist $P(Y \le z) = 1$.

Sei $z \in J$:

$$\begin{aligned} P(Y \le z) &= P\big(\varphi(X) \le z\big) = P\big(X \le \varphi^{-1}(z)\big) \\ &= \int_{-\infty}^{\varphi^{-1}(z)} f_X(x)\,dx = \int_{-\infty}^{z} f_X\big(\varphi^{-1}(y)\big) \frac{1}{\big|\varphi'\big(\varphi^{-1}(y)\big)\big|}\,dy, \end{aligned}$$

wobei wir $x = \varphi^{-1}(y)$ (und somit $\frac{dx}{dy} = \frac{1}{\varphi'(\varphi^{-1}(y))}$ substituiert haben). Siehe auch die Skizze unten. □



Beispiel A.2.6. $X \sim \mathcal{N}_{0,1}$, $\mu \in \mathbb{R}$, $\sigma > 0$, so ist $Y := \sigma X + \mu \sim \mathcal{N}_{\mu,\sigma^2}$ (Übung).

Bericht A.2.7 (Dichtetransformation im $\mathbb{R}^d$)**.** X $\mathbb{R}^d$-wertige ZV mit Dichte f_X, $I \subset \mathbb{R}^d$ offen mit $P(X \in I) = 1$, $J \subset \mathbb{R}^d$ offen, $\varphi : I \to J$ bijektiv, stetig differenzierbar mit Ableitung

$$\varphi'(x) = \left(\frac{\partial \varphi_i}{\partial x_j}(x) \right)_{i,j=1}^{d} \qquad (\text{„Jacobi-Matrix“}),$$

dann hat $Y := \varphi(X)$ die Dichte

$$f_Y(y) = \begin{cases} \dfrac{f_X\big(\varphi^{-1}(y)\big)}{\big|\det \varphi'\big(\varphi^{-1}(y)\big)\big|}, & y \in J, \\ 0, & y \notin J. \end{cases}$$

Beweise finden sich in Analysis-Lehrbüchern, z.B. G. Kersting und M. Brokate, *Maß und Integral*, S. 107, H. Heuser, *Analysis, Teil 2*, Satz 205.2 (“Substitutions-Regel”), O. Forster, *Analysis 3*, Kap. 9, Satz 1 (“Transformationsformel”).

Wir betrachten hier nur folgende Heuristik (im Fall $d = 2$): Lokal sieht

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \varphi(x) = \begin{pmatrix} \varphi_1(x) \\ \varphi_2(x) \end{pmatrix}$$

„aus wie“

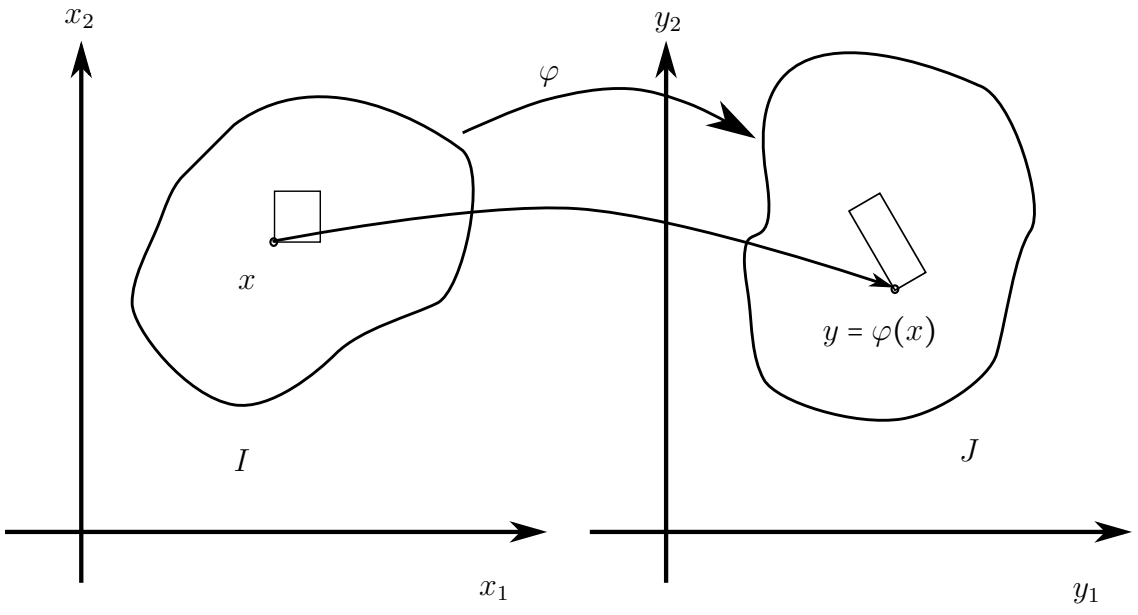
$$\begin{aligned} \varphi(x') &\approx \varphi(x) + \varphi'(x) \cdot (x' - x) \\ &= \varphi(x) + \begin{pmatrix} \frac{\partial}{\partial x_1}\varphi_1(x) & \frac{\partial}{\partial x_2}\varphi_1(x) \\ \frac{\partial}{\partial x_1}\varphi_2(x) & \frac{\partial}{\partial x_2}\varphi_2(x) \end{pmatrix} \cdot \begin{pmatrix} x_1' - x_1 \\ x_2' - x_2 \end{pmatrix} \end{aligned}$$

(plus Terme, die $O(\|x' - x\|^2)$ sind), also:

die Fläche der Größe $h_1 \cdot h_2$ „rund um x“

wird auf

$\approx$ Fläche $h_1 \cdot h_2 \cdot |\det \varphi'(x)|$ „rund um y“ abgebildet.



Wenden wir dies auf $Y = \varphi(X)$ an, so bedeutet das anschaulich: Für $y = \varphi(x) \in J$ (und sehr kleines $h > 0$) ist

$$\begin{aligned} f_Y(y)h^2 &\approx \mathbb{P}\big(Y \text{ nimmt Wert in einem Quadrat der Fläche } h^2 \text{ mit „Aufpunkt“ } y \text{ an}\big) \\ &\approx \mathbb{P}\big(X \text{ nimmt Wert in einem Quader der Fläche } h^2/|\det \varphi'(x)| \text{ mit „Aufpunkt“ } x \text{ an}\big) \\ &\approx f_X(x)\frac{h^2}{|\det \varphi'(x)|} = \frac{f_X(\varphi^{-1}(y))}{|\det \varphi'(\varphi^{-1}(y))|}h^2. \end{aligned}$$

Beispiel A.2.8. 1. $X_1, \dots, X_n$ u.a., $X_i \sim \mathcal{N}_{0,1}$, dann hat $X := (X_1, \dots, X_n)$ Dichte

$$f_X(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2}\|x\|^2\big),\ x = (x_1, \dots, x_n) \in \mathbb{R}^n$$

(mit $\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$, der euklidischen Norm).

Sei $M = (m_{ij})_{i,j=1}^n$ orthogonale $n \times n$-Matrix (d.h. $M^T M = I$, die $n \times n$-Identitätsmatrix),

$$Y^T := MX^T \quad \text{d.h. } Y = (Y_1, \dots, Y_n) \text{ mit } Y_i = \sum_{j=1}^n m_{ij} X_j,$$

dann sind $Y_1, \dots, Y_n$ u.a., $Y_i \sim \mathcal{N}_{0,1}$.

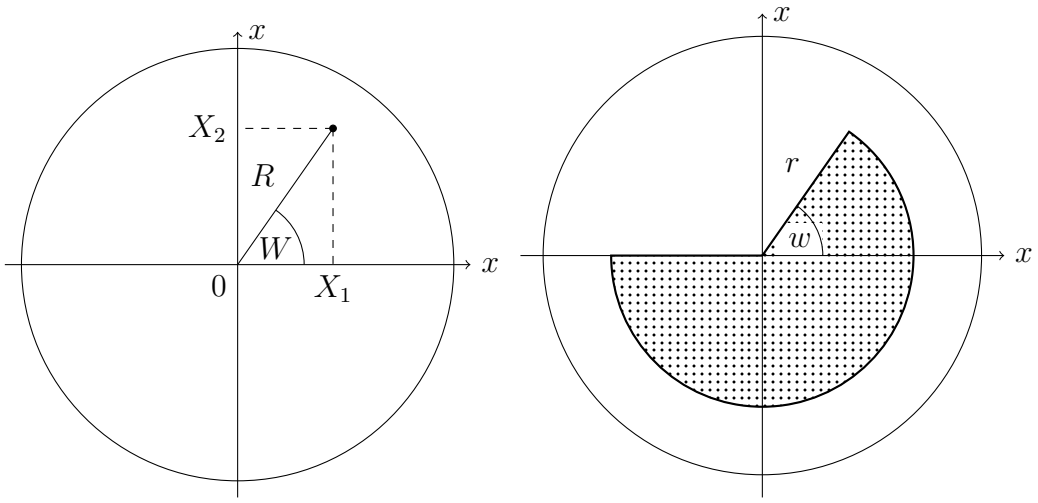
$\varphi : \mathbb{R}^n \to \mathbb{R}^n$, $\varphi(x) = Mx$ ist bijektiv und differenzierbar mit $\varphi^{-1}(y) = M^T y$, $\varphi' = M$, die Dichtetransformationsformel (Bericht A.2.7) zeigt: Y hat Dichte

$$\begin{aligned} f_Y(y) &= \frac{f_X(\varphi^{-1}(y))}{|\det \varphi'(\varphi^{-1}(y))|} = \frac{f_X(M^T y)}{|\det M|} = f_X(M^T y) \\ &= (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2}\underbrace{\|M^T y\|^2}_{\|y\|^2}\big) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2}. \end{aligned}$$

2. Sei X ein uniform im Einheitskreis $\{x \in \mathbb{R}^2 : \|x\| \le 1\}$ verteilter Punkt (in kartesischen Koordinaten $X = (X_1, X_2)$), R der Radius, W der Winkel von X (in Polarkoordinaten), also

$$R = \sqrt{X_1^2 + X_2^2}, \quad W = \begin{cases} \arcsin\big(\frac{X_2}{R}\big), & X_1 \ge 0, \\ \pi - \arcsin\big(\frac{X_2}{R}\big), & X_1 < 0, X_2 \ge 0, \\ -\pi - \arcsin\big(\frac{X_2}{R}\big), & X_1 < 0, X_2 < 0, \end{cases}$$

(siehe auch die Skizze unten).



Links: Punkt (X_1, X_2) und seine Polarkoordinten (R, W). Rechts: Schraffiert ist $B(r,w) := \{\text{Punkte mit Radius} \le r \text{ und Winkel} \le w\}$.

Dann sind R und W unabhängig,

$$R \text{ hat Dichte } f_R(r) = 2r\mathbf{1}_{[0,1]}(r),$$
$$W \text{ hat Dichte } f_W(w) = \frac{1}{2\pi}\mathbf{1}_{[-\pi,\pi)}(w),$$

denn (für $0 \le r \le 1$, $-\pi \le w < \pi$)

$$P(R \le r, W \le w) = P\Big(X \in B(r,w)\Big)$$
$$= \frac{\pi r^2 \frac{w+\pi}{2\pi}}{\pi 1^2} = r^2\frac{w+\pi}{2\pi} = \int_0^r 2s\,ds \cdot \int_{-\pi}^w \frac{1}{2\pi}\,dv.$$

Zur allgemeinen mehrdimensionalen Normalverteilung

Beobachtung A.2.9. i) Sei $X = (X_1, \dots, X_d)^t$ eine $\mathbb{R}^d$-wertige Zufallsvariable. Die Kovarianzmatrix $C = (c_{ij})_{i,j=1,\dots,d}$ mit $c_{ij} = \mathrm{Cov}[X_i, X_j]$ ist symmetrisch und positiv definit, denn $c_{ij} = \mathrm{Cov}[X_i, X_j] = \mathrm{Cov}[X_j, X_i] = c_{ji}$ und für $a = (a_1, \dots, a_d)^t \in \mathbb{R}^d$ ist

$$a^tCa = \sum_{i,j=1}^d a_i c_{ij} a_j = \sum_{i,j=1}^d a_i a_j \mathrm{Cov}[X_i, X_j] = \mathrm{Cov}\left[\sum_{i=1}^d a_i X_i, \sum_{j=1}^d a_j X_j\right] = \mathrm{Var}[\langle a, X\rangle] \ge 0.$$

ii) Sind $Z_1, \dots, Z_d$ unabhängig und standardnormalverteilt, so hat $Z = (Z_1, \dots, Z_d)^t$ die Dichte

$$f_Z(z) = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}\left(z_1^2 + \cdots + z_d^2\right)\right) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|z\|^2}, \quad z \in \mathbb{R}^d.$$

$\mathcal{L}(Z)$ heißt die d-dimensionale Standardnormalverteilung.

iii) Sei $\mu \in \mathbb{R}^d$ und $A = (a_{ij}) \in \mathbb{R}^{d\times d}$. Dann hat $X := \mu + AZ$ den Erwartungswert(-vektor) $\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d]) = \mu$ und die Kovarianzmatrix $C := AA^t$, denn

$$\mathrm{Cov}[X_k, X_l] = \mathrm{Cov}\left[\mu_k + \sum_{i=1}^d a_{k,i} Z_i,\ \mu_l + \sum_{j=1}^d a_{l,j} Z_l\right] = \sum_{i,j=1}^d a_{ki} a_{lj} \mathrm{Cov}[Z_i, Z_j]$$
$$= \sum_{i,j=1}^d a_{ki} a_{lj} \delta_{ij} = \sum_{i=1}^d a_{ki} a_{li} = (AA^t)_{kl}.$$

Falls A vollen Rang d hat, so hat X die Dichte

$$f_{\mu,C}(x)\frac{1}{\sqrt{(2\pi)^d \det C}} \exp\left(-\frac{1}{2}\langle x-\mu, C^{-1}(x-\mu)\rangle\right), \quad x \in \mathbb{R}^d,$$

denn für $g(z) := \mu + Az$ gilt $g^{-1}(x) = A^{-1}(x-\mu)$ und $\left(\frac{\partial}{\partial x_i} g_j(z)\right)_{i,j} = \mathrm{D}g(z) = A$. Also folgt mit der Dichtetransformationsformel und mit $\det C = \det\left(AA^{\mathrm{t}}\right) = \left(\det A\right)^2$:

$$f_{\mu,C}(x) = f_Z\left(g^{-1}(x)\right) \frac{1}{|\det \mathrm{D}g^{-1}(x))|}.$$

Falls A nicht vollen Rang hat, so besitzt X keine Dichte bezüglich λ^d.

Was ist jedoch in dem Fall, in dem A (und damit auch C) nicht vollen Rang haben? Betrachte für $u \in \mathbb{R}^d$:

$$\begin{aligned}
\mathbb{E}\left[e^{i\langle u,X\rangle}\right] &= \mathbb{E}\left[e^{i\langle u,\mu\rangle} \cdot e^{i\langle u,AZ\rangle}\right] = e^{i\langle u,\mu\rangle}\mathbb{E}\left[e^{i\sum_{k,l=1}^d u_k a_{kl} Z_l}\right] = e^{i\langle u,\mu\rangle}\prod_{l=1}^d \mathbb{E}\left[e^{i\sum_{k=1}^d u_k a_{kl} Z_l}\right] \\
&= e^{i\langle u,\mu\rangle}\prod_{l=1}^d e^{-\frac{1}{2}\left(\sum_{k=1}^d u_k a_{kl}\right)^2} = e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\sum_{l=1}^d \left((u^{\mathrm{t}}A)_l\right)^2} = e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\langle u^{\mathrm{t}}A, u^{\mathrm{t}}A\rangle} \\
&= e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\langle u^{\mathrm{t}}, u^{\mathrm{t}}AA^{\mathrm{t}}\rangle} = e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\langle u, Cu\rangle}.
\end{aligned}$$

Dies legt folgende Definition nahe.

Definition A.2.10. Sei $\mu \in \mathbb{R}^d$, $C \in \mathbb{R}^{d\times d}$ symmetrisch und positiv definit. X heißt *d-dimensional normalverteilt mit Erwartungswert μ und Kovarianzmatrix C*, falls

$$\varphi_X(u) = e^{i\langle u,\mu\rangle} e^{-\frac{1}{2}\langle u, Cu\rangle}.$$

Man schreibt auch $\mathcal{L}(X) =: \mathcal{N}(\mu, C)$.

Bemerkung A.2.11. Sei $X \sim \mathcal{N}(\mu, C)$, $A \in \mathbb{R}^{d\times d}$ und $Y := AX$. Dann ist $Y \sim \mathcal{N}(A\mu, ACA^{\mathrm{t}})$, denn

$$\mathbb{E}\left[e^{i\langle u,Y\rangle}\right] = \mathbb{E}\left[e^{i\langle u,AX\rangle}\right] = \mathbb{E}\left[e^{i\langle A^{\mathrm{t}}u,X\rangle}\right] = e^{i\langle A^{\mathrm{t}}u,\mu\rangle} e^{-\frac{1}{2}\langle A^{\mathrm{t}}u, CA^{\mathrm{t}}u\rangle} = e^{i\langle u,A\mu\rangle} e^{-\frac{1}{2}\langle u, ACA^{\mathrm{t}}u\rangle}.$$

A.3 Exakte Konfidenzintervalle für den Erfolgsparameter in der Binomialverteilung

Unter n unabhängigen Versuchen seien x Erfolge beobachtet worden, wir fassen x als Realisierung einer $\mathrm{Bin}_{n,\vartheta}$-verteilten ZV auf und wollen anhand der Beobachtung auf ϑ schließen.

Wir hatten in der Vorlesung das auf asymptotischer Normalität fußende (approximative) Konfidenzintervall für ϑ zum Niveau $1-\alpha$ betrachtet:

$$\left[\widehat{\vartheta} - q\frac{\widehat{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widehat{\sigma}}{\sqrt{n}}\right]$$

mit $\widehat{\vartheta} = \dfrac{x}{n}$, $\widehat{\sigma} = \sqrt{\widehat{\vartheta}(1-\widehat{\vartheta})}$, q das $1-\frac{\alpha}{2}$-Quantil von $\mathcal{N}_{0,1}$

Frage: Wie können wir vorgehen, wenn wir uns nicht auf die Asymptotik verlassen möchten? Beobachte $X \sim \mathbb{P}_\vartheta := \mathrm{Bin}_{n,\vartheta}$, Konfidenzintervall für $\vartheta \in \Theta = [0,1]$?

Idee: Zu $\vartheta \in \Theta := [0,1]$ wähle $c_\vartheta \in (0,1)$, so dass für

$$C_\vartheta := \big\{x \in \{0,1,\dots,n\} : \mathrm{Bin}_{n,\vartheta}(\{x\}) \geq c_\vartheta\big\}$$

gilt $\mathrm{Bin}_{n,\vartheta}(C_\vartheta) \geq 1-\alpha$ (und c_ϑ möglichst groß, so dass C_ϑ möglichst klein).

Setze $C(x) := \{\vartheta \in \Theta : x \in C_\vartheta\}$ für $x \in \Omega := \{0,1,\dots,n\}$,
dann gilt

$$\forall\, \vartheta \in \Theta :\ \mathbb{P}_\vartheta(\vartheta \in C(X)) = \mathbb{P}_\vartheta(X \in C_\vartheta) \geq 1-\alpha$$

nach Konstruktion.

Es gilt

1. Für $\vartheta \in (0,1)$ ist $\{0,\dots,n\} \ni x \mapsto \mathrm{Bin}_{n,\vartheta}(\{x\})$ strikt wachsend auf $\{0,1,\dots,\lceil (n+1)\vartheta - 1\rceil\}$, strikt fallend auf $\{\lfloor (n+1)\vartheta\rfloor,\dots,n\}$, also maximal auf $x = \lfloor (n+1)\vartheta\rfloor$ (und auf $(n+1)\vartheta - 1$, wenn $(n+1)\vartheta \in \mathbb{Z}$).

2. Für $x \in \{1,\dots,n\}$ ist $[0,1] \ni \vartheta \mapsto Bin_{n,\vartheta}(\{x,x+1,\dots,n\})$ stetig, strikt monoton wachsend mit

   $$\mathrm{Bin}_{n,\vartheta}(\{x,x+1,\dots,n\}) = \mathrm{Beta}_{x,n-x+1}([0,\vartheta]),$$

   wo $\mathrm{Beta}_{a,b}$ die Dichte

   $$f_{\mathrm{Beta}_{a,b}}(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}$$

   auf $(0,1)$ hat.

Beweis. 1.

$$\begin{aligned}
\frac{\mathrm{Bin}_{n,\vartheta}(\{x\})}{\mathrm{Bin}_{n,\vartheta}(\{x-1\})} &= \frac{\binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}}{\binom{n}{x-1}\vartheta^{x-1}(1-\vartheta)^{n-x+1}} \\
&= \frac{(n-x+1)\vartheta}{x(1-\vartheta)} \\
&> 1 \\
\iff x &< (n+1)\vartheta
\end{aligned}$$

2. $U_1,\dots,U_n$ unabhängig und uniform auf $[0,1]$, $S_\vartheta := \sum_{i=1}^n \mathbf{1}_{[0,\vartheta]}(U_i)$ ist $\mathrm{Bin}_{n,\vartheta}$-verteilt.

   Sei $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ die „Ordnungsstatistik“.

$$\begin{aligned}
&\mathrm{Bin}_{n,\vartheta}(\{x,\dots,n\}) = \mathbb{P}(S_\vartheta \geq x) = \mathbb{P}(U_{(x)} \leq \vartheta) \\
&= \sum_{k=1}^{n} \sum_{\substack{B \subseteq \{1,\dots,n\}\setminus\{k\} \\ |B| = x-1}} \underbrace{\mathbb{P}\Big(\begin{array}{l} U_k \leq \vartheta, U_m \leq U_k \text{ für } m \in B, \\ U_l > U_k \text{ für } l \in \{1,\dots,n\}\setminus(\{k\}\cup B)\end{array}\Big)}_{= \int_0^\vartheta u^{|B|}(1-u)^{n-|B|-1}\,du = \int_0^\vartheta u^{x-1}(1-u)^{n-x}\,du} \\
&= \underbrace{n\binom{n-1}{x-1}}_{= \frac{n!}{(x-1)!(n-x)!} = \frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)}} \int_0^\vartheta u^{x-1}(1-u)^{n-x}\,du
\end{aligned}$$

□

Wähle $C_\vartheta := \{x_-(\vartheta), x_-(\vartheta)+1, \dots, x_+(\vartheta)\}$ mit $x_-(\vartheta) = \max\{x : \mathrm{Bin}_{n,\vartheta}(\{0,\dots,x-1\}) \le \frac{\alpha}{2}\}$ und $x_+(\vartheta) = \min\{x : \mathrm{Bin}_{n,\vartheta}(\{x+1,\dots,n\}) \le \frac{\alpha}{2}\}$.

Dann gilt:

- $x \le x_+(\vartheta) \iff Bin_{n,\vartheta}(\{x,\dots,n\}) = Beta_{x,n-x+1}([0,\vartheta]) > \frac{\alpha}{2}$
  $\iff \vartheta > p_-(x) := \frac{\alpha}{2}$-Quantil von $\mathrm{Beta}_{x,n-x+1}$.
- $x \ge x_+(\vartheta) \iff \mathrm{Bin}_{n,\vartheta}(\{0,\dots,x\}) = 1 - \mathrm{Bin}(\{x+1,\dots,n\}) = \mathrm{Beta}_{x+1,n-x}([\vartheta,1]) \ge \frac{\alpha}{2}$.
  $\iff \vartheta < p_+(x) := 1 - \frac{\alpha}{2}$-Quantil von $\mathrm{Beta}_{x+1,n-x}$.

Fazit (exaktes Konfidenzintervall im Binomialmodell)**.** Mit

$$p_-(x) := \tfrac{\alpha}{2}\text{-Quantil von } \mathrm{Beta}_{x,n-x+1},$$
$$p_+(x) := 1 - \tfrac{\alpha}{2}\text{-Quantil von } \mathrm{Beta}_{x+1,n-x}$$

ist $x \mapsto [p_-(x), p_+(x)]$ ist ein Konfidenzintervall für ϑ zum Sicherheitsniveau $1-\alpha$.

Bemerkung.

- Quantile der Beta-Verteilungen sind tabelliert, gelegentlich kann beim Nachschlagen in Tabellen die Symmetrieeigenschaft

  $$\mathrm{Beta}_{a,b}([0,x]) = \mathrm{Beta}_{b,a}([1-x,1]) = 1 - \mathrm{Beta}_{b,a}([0,1-x])$$

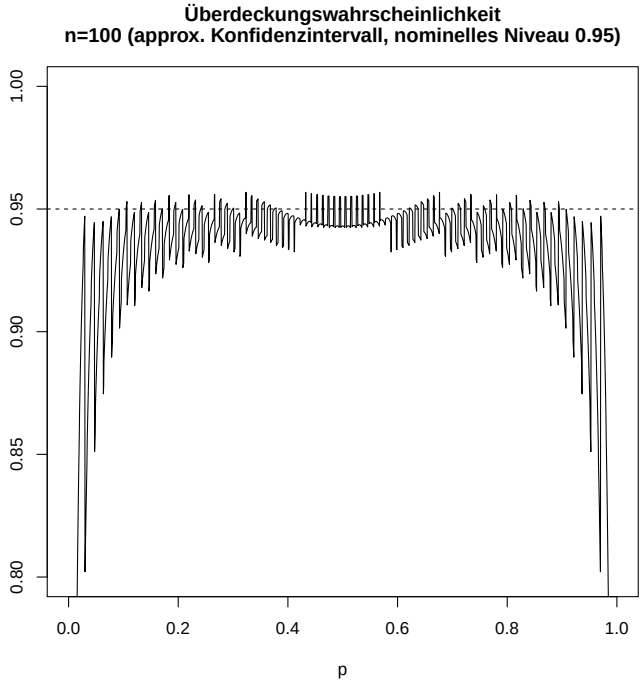
  nützlich sein.
- `R` kennt die Beta-Verteilungen, ihre Verteilungsfunktionen `pbeta(x, a, b)` und ihre Quantile `qbeta(p, a, b)`

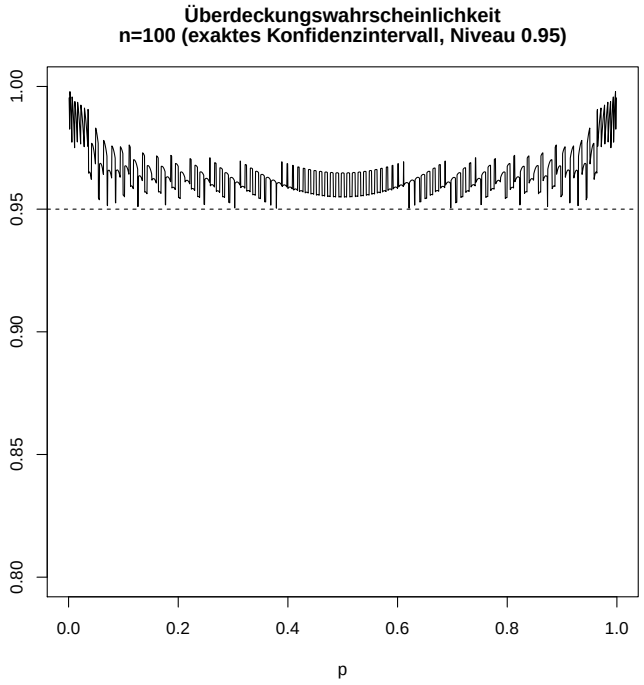
Beispiel. $n = 53$, $x = 23$, wähle $\alpha = 0{,}05$

$\widehat{\vartheta} = \frac{23}{53} \approx 0{,}434$, $\widehat{\sigma} \approx 0{,}496$, $q_{0{,}975} \approx 1{,}96$
$\left[\widehat{\vartheta} \pm q\frac{\widehat{\sigma}}{\sqrt{53}}\right] \approx [0{,}30, 0{,}57]$

$p_-(23) = 0{,}025$-Quantil von $\mathrm{Beta}_{23,31} \approx 0{,}30$, $p_+(23) = 0{,}975$-Quantil von $\mathrm{Beta}_{24,30} \approx 0{,}57$

(Absurd präzise Werte wären: $\left[\widehat{\vartheta} \pm q\frac{\widehat{\sigma}}{\sqrt{53}}\right] \approx [0{,}3005306, 0{,}5673939]$
$[p_-(23), p_+(23)] \approx [0{,}2983921, 0{,}5771742]$)





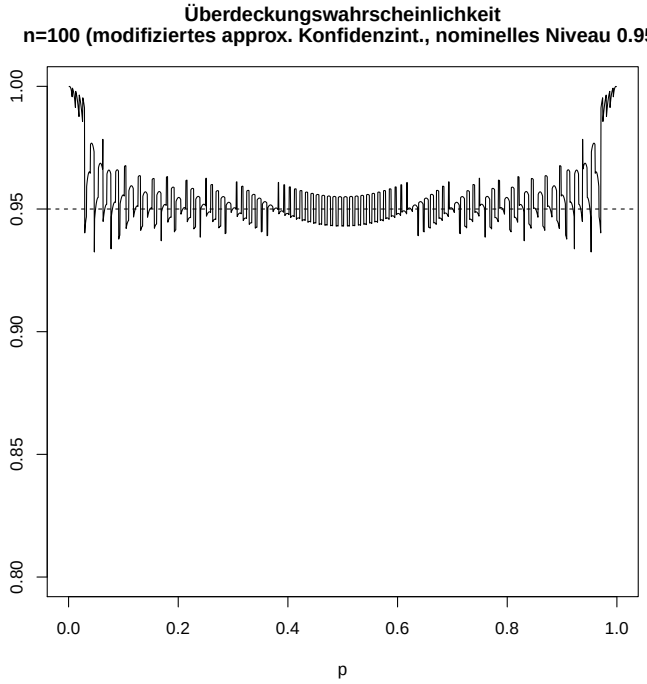
Beispiel. Manchmal betrachtet man auch den Schätzer

$$\widetilde{\vartheta} = \frac{x+1}{n+2}$$

für ϑ und bildet als (approximatives) Konfidenzintervall zum Niveau $1-\alpha$

$$\left[\widetilde{\vartheta} - q\frac{\widetilde{\sigma}}{\sqrt{n}}, \widehat{\vartheta} + q\frac{\widetilde{\sigma}}{\sqrt{n}}\right]$$

mit $\widetilde{\sigma} = \sqrt{\widetilde{\vartheta}(1-\widetilde{\vartheta})}$, q das $1-\frac{\alpha}{2}$-Quantil von $\mathcal{N}_{0,1}$



A.4 Zu Welchs t-Test

Die Gamma-Verteilung $\Gamma_{\alpha,\nu}$ hat Dichte $\frac{1}{\Gamma(\nu)}\alpha^\nu x^{\nu-1}e^{-\alpha x}\mathbf{1}_{(0,\infty)}(x)$ (α = Skalen-, ν = Formparameter) mit Erwartungswert ν/α und Varianz ν/α^2. Nun ist $\Gamma_{\alpha,\nu_1} * \Gamma_{\alpha,\nu_2} = \Gamma_{\alpha,\nu_1+\nu_2}$, aber für $\alpha_1 \neq \alpha_2$ ist $\Gamma_{\alpha_1,\nu_1} * \Gamma_{\alpha_2,\nu_2}$ keine Gamma-Verteilung.

Beobachtung. Seien $\nu_1, \nu_2 > 0$, $\alpha_1 \neq \alpha_2 > 0$, so hat $\mu := \Gamma_{\alpha_1,\nu_1} * \Gamma_{\alpha_2,\nu_2}$ für $x > 0$ die Dichte

$$
\begin{aligned}
&\int_0^x \frac{1}{\Gamma(\nu_1)}\alpha_1^{\nu_1}(x-y)^{\nu_1-1}e^{-\alpha_1(x-y)}\frac{1}{\Gamma(\nu_2)}\alpha_2^{\nu_2}y^{\nu_2-1}e^{-\alpha_2 y}\,dy\\
&= \frac{\alpha_1^{\nu_1}\alpha_2^{\nu_2}e^{-\alpha_1 x}}{\Gamma(\nu_1)\Gamma(\nu_2)}\int_0^x y^{\nu_2-1}(x-y)^{\nu_1-1}e^{(\alpha_1-\alpha_2)y}\,dy\\
&= \frac{\alpha_1^{\nu_1}\alpha_2^{\nu_2}e^{-\alpha_1 x}}{\Gamma(\nu_1)\Gamma(\nu_2)}x^{\nu_1+\nu_2-1}\int_0^1 u^{\nu_2-1}(1-u)^{\nu_1-1}e^{(\alpha_1-\alpha_2)x\,u}\,du\\
&= \frac{\alpha_1^{\nu_1}\alpha_2^{\nu_2}x^{\nu_1+\nu_2-1}e^{-\alpha_1 x}}{\Gamma(\nu_1)\Gamma(\nu_2)}\cdot\frac{\Gamma(\nu_1)\Gamma(\nu_2)}{\Gamma(\nu_1+\nu_2)}M\big(\nu_2,\nu_1+\nu_2,(\alpha_1-\alpha_2)x\big)\\
&= \frac{\alpha_1^{\nu_1}\alpha_2^{\nu_2}x^{\nu_1+\nu_2-1}e^{-\alpha_1 x}}{\Gamma(\nu_1+\nu_2)}M\big(\nu_2,\nu_1+\nu_2,(\alpha_1-\alpha_2)x\big) = \frac{\alpha_1^{\nu_1}\alpha_2^{\nu_2}x^{\nu_1+\nu_2-1}e^{-\alpha_2 x}}{\Gamma(\nu_1+\nu_2)}M\big(\nu_1,\nu_1+\nu_2,(\alpha_2-\alpha_1)x\big)
\end{aligned}
$$

(mit Substitution $y = xu$, also $dy = x\,du$ für das 2. Gleichheitszeichen); siehe [AS64, Ch. 13: Confluent hypergeometric functions], speziell die Integraldarstellung in [AS64, Formula 13.2.1]: $\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)}M(a,b,z) = \int_0^t e^{zt}t^{a-1}(1-t)^{b-a-1}\,dt$; [AS64, Formula 13.1.27 ("Kummer Transformations"]: $M(a,b,z) = e^z M(b-a,b,-z)$. [AS64, Formula 13.1.2] liefert

$$M(a,b,z) = {}_1F_1(a,b;z) = \sum_{k=0}^{\infty}\frac{(a)_{k\uparrow}z^k}{(b)_{k\uparrow}k!}$$

(mit $(a)_{k\uparrow} = a(a+1)\cdots(a+k-1)$, $(a)_{0\uparrow} = 1$)

μ hat Erwartungswert $\mathbb{E}[\mu] = \int_0^\infty x\,\mu(dx) = \frac{\nu_1}{\alpha_1} + \frac{\nu_2}{\alpha_2}$ und Varianz $\mathrm{Var}[\mu] = \int_0^\infty (x - \mathbb{E}[\mu])^2\,\mu(dx) = \frac{\nu_1}{\alpha_1^2} + \frac{\nu_2}{\alpha_2^2}$, d.h. die beiden ersten Momente entsprechen denen von

$$\Gamma_{\alpha',\nu'} \quad \text{mit} \quad \alpha' = \frac{\mathbb{E}[\mu]}{\mathrm{Var}[\mu]} = \frac{\frac{\nu_1}{\alpha_1} + \frac{\nu_2}{\alpha_2}}{\frac{\nu_1}{\alpha_1^2} + \frac{\nu_2}{\alpha_2^2}}, \quad \nu' = \frac{(\mathbb{E}[\mu])^2}{\mathrm{Var}[\mu]} = \frac{\left(\frac{\nu_1}{\alpha_1} + \frac{\nu_2}{\alpha_2}\right)^2}{\frac{\nu_1}{\alpha_1^2} + \frac{\nu_2}{\alpha_2^2}}$$

Unter P_ϑ sind $X_1, \dots, X_{n_1}$ u.i.v. und davon unabhängig $Y_1, \dots, Y_{n_2}$ u.i.v. ($n_1, n_2 \in \mathbb{N}$), $X_i \sim \mathcal{N}_{\mu_1,\sigma_1^2}$, $Y_j \sim \mathcal{N}_{\mu_2,\sigma_2^2}$. Seien

$$\overline{X} := \frac{1}{n_1}\sum_{i=1}^{n_1} X_i,\ \overline{Y} := \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$$

die jeweiligen Stichprobenmittelwerte,

$$S_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2,\ S_Y^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_j - \overline{Y})^2,$$

die (korrigierten) Stichprobenvarianzen. Man schätzt die Streuung von $\overline{X} - \overline{Y}$ durch

$$\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}} \quad \text{und bildet} \quad T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}.$$

Unter $P_{(\mu_0,\mu_0,\sigma_1^2,\sigma_2^2)}$ ist T „approximativ Student-verteilt mit g Freiheitsgraden“, wobei

$$g = \frac{\left(\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}\right)^2}{\frac{s_X^4}{n_1^2(n_1-1)} + \frac{s_Y^4}{n_2^2(n_2-1)}}$$

(“Satterthwaite's formula”) aus den Daten geschätzt wird (siehe F. E. Satterthwaite, An Approximate Distribution of Estimates of Variance Components, Biometrics Bulletin Vol. 2, No. 6, Dec., 1946 (pp. 110-114), Eq. (7)).

Nun ist $\frac{n_1-1}{\sigma_1^2}S_X^2 \sim \chi^2_{n_1-1} = \Gamma_{1/2,(n_1-1)/2}$, $\frac{n_2-1}{\sigma_2^2}S_Y^2 \sim \chi^2_{n_2-1} = \Gamma_{1/2,(n_2-1)/2}$, also

$$S_X^2 \sim \Gamma_{(n_1-1)/(2\sigma_1^2),(n_1-1)/2},\ S_Y^2 \sim \Gamma_{(n_2-1)/(2\sigma_2^2),(n_2-1)/2}$$
$$\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2} \sim \Gamma_{n_1(n_1-1)/(2\sigma_1^2),(n_1-1)/2} * \Gamma_{n_2(n_2-1)/(2\sigma_2^2),(n_2-1)/2}$$

A.5 Verfälschte Tests, die den (zweiseitigen 1-Stichproben-) t-Test „lokal schlagen“

Wir diskutieren mittels numerischer Beispiele verfälschte Tests von $\mu = 0$ gegen $\mu \neq 0$ im Normalmodell aus [LS48], die an gewissen Stellen der Alternative eine höhere Macht haben als der t-Test.

Seien $\alpha \in (0, 1/2)$, $n \in \mathbb{N}$, $Z_1, \dots, Z_n$ u.i.v. $\sim \mathcal{N}(0,1)$. Wir wählen $\eta \in (0, n)$ und $\xi_* > 0$ so, dass

$$\sup_{\xi\in\mathbb{R}} \mathbb{P}\Big(\sum_{i=1}^n (Z_i - \xi)^2 \le (n-\eta)\xi^2\Big) = \mathbb{P}\Big(\sum_{i=1}^n (Z_i - \xi_*)^2 \le (n-\eta)\xi_*^2\Big) = \alpha \tag{A.1}$$

Zu (A.1): Mit $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ ist

$$\sum_{i=1}^n (Z_i - \xi)^2 = \sum_{i=1}^n \big((Z_i - \overline{Z}) + (\overline{Z} - \xi)\big)^2 = \sum_{i=1}^n (Z_i - \overline{Z})^2 + \sum_{i=1}^n (\overline{Z} - \xi)^2 \stackrel{d}{=} Y + n\big(n^{-1/2}Z - \xi\big)^2$$

mit unabhängigen $Y \sim \chi^2_{n-1}$, $Z \sim \mathcal{N}(0,1)$.

$$\begin{aligned} Y + n\big(n^{-1/2}Z - \xi\big)^2 &< (n-\eta)\xi^2 \\ \iff \quad Y &< (n-\eta)\xi^2 - n\big(n^{-1/2}Z - \xi\big)^2 = -\eta\xi^2 + 2\sqrt{n}\xi Z - Z^2 \\ &= -\eta\Big(\xi - \frac{\sqrt{n}}{\eta}Z\Big)^2 + \Big(\frac{n}{\eta} - 1\Big)Z^2 \end{aligned}$$

Da $Y \geq 0$ stets, muss für gegebenes $\xi > 0$ gelten

$$\begin{aligned} & n\big(n^{-1/2}Z - \xi\big)^2 < (n-\eta)\xi^2 \\ \iff \quad & |Z - \sqrt{n}\xi| < (1 - \eta/n)^{1/2}|\xi| \end{aligned}$$

Somit ist die Wahrscheinlichkeit innerhalb des sup auf der linken Seite von (A.1)

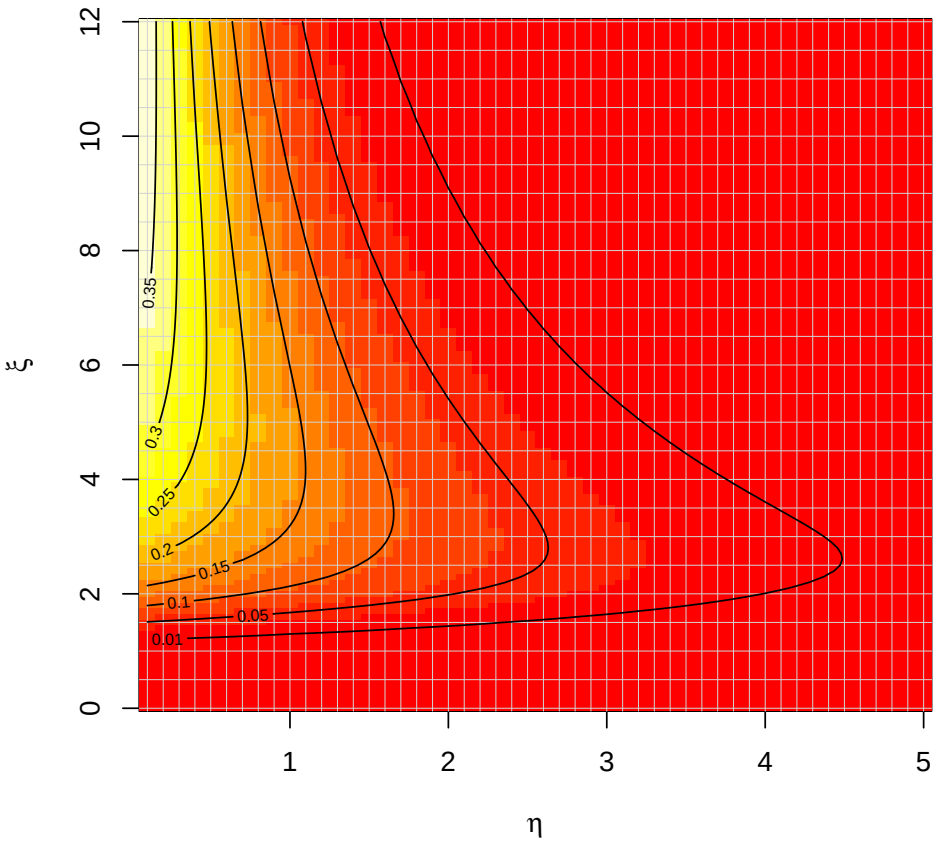
$$\int_{n^{1/2}\xi - (1-\eta/n)^{1/2}|\xi|}^{n^{1/2}\xi + (1-\eta/n)^{1/2}|\xi|} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, \chi^2_{n-1}\Big(\big[0, -\eta\xi^2 + 2\sqrt{n}\xi Z - Z^2\big]\Big) dz \qquad \text{(A.2)}$$

Wir implementieren (A.2) in R:

```
Niveaufkt <- function(n,eta,xi) {
  f <- Vectorize(function (z) dnorm(z)*pchisq(-eta*xi^2+2*sqrt(n)*xi*z-z^2,df=n-1))
  integrate(f, lower=xi/sqrt(n)-sqrt(1-eta/n)*abs(xi),
            upper=xi/sqrt(n)+sqrt(1-eta/n)*abs(xi), subdivisions=500L)
}
```

Schauen wir die Werte grafisch an:

```
n<- 20
eta.max <- n/4
xi.max <- 12
etas <- seq(from=0.1,to=eta.max,by=0.1) ## etas <- seq(from=0.1,to=n/5,by=0.05)
xis <- seq(from=0, to=xi.max, by=0.1) ## xis <- seq(from=0, to=15, by=0.05)
niv <- matrix(0,nrow=length(etas),ncol=length(xis))
for (i in 1:length(etas))
  for (j in 1:length(xis))
    niv[i,j] <- Niveaufkt(n,etas[i],xis[j])$value

image(etas,xis,niv,xlab=expression(eta),ylab=expression(xi))
abline(h=seq(from=0,to=xi.max,by=0.5),col='lightgray',lwd=0.5)
abline(v=seq(from=0,to=eta.max,by=0.1),col='lightgray',lwd=0.5)
contour(etas,xis,niv,levels=c(0.01,seq(from=0.05,to=0.4,by=0.05)),add=TRUE)
```



Sei $\mu_1 > 0$, $\sigma_1 > 0$. Betrachten wir nun den Test ψ mit Ablehnungsbereich

$$\Big\{x = (x_1, \dots, x_n) \in \mathbb{R}^n : \Big(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\Big)\sum_{i=1}^n x_i^2 - 2\frac{\mu_1}{\sigma_1^2}\sum_{i=1}^n x_i \le c\Big\} \tag{A.3}$$

(mit gewissen Wahlen von $\sigma_0 > \sigma_1$ und $c \in \mathbb{R}$). Die Idee ist, dass dies der Ablehnungsbereich eines Neyman-Pearson-Tests von $\{(0, \sigma_0^2)\}$ gegen $\{(\mu_1, \sigma_1^2)\}$ vom Niveau α ist, wenn wir c richtig einstellen. Wir werden mittels (A.1) σ_0 und c so wählen können, dass er zugleich für jedes $\sigma > 0$ ein Test von $\{(0, \sigma^2)\}$ gegen $\{(\mu_1, \sigma_1^2)\}$ ist, der effektives Niveau $\le \alpha$ hat.

(A.3) ist äquivalent zu

$$\Big\{x = (x_1, \dots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n \Big(x_i - \frac{\mu_1}{1 - \sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{c}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}} + n\frac{\mu_1^2}{(1 - \sigma_1^2/\sigma_0^2)^2}\Big\} \tag{A.4}$$

Setze

$$\sigma_0 = \frac{\mu_1 + \sqrt{\mu_1^2 + 4\sigma_1^2\xi_*^2}}{2\xi_*} \quad (> \sigma_1) \tag{A.5}$$

(dies ist die positive Lösung $\sigma = \sigma_0$ von $\sigma^2 - \frac{\mu_1}{\xi_*}\sigma - \sigma_1^2 = 0$, d.h. es erfüllt $\xi_* = \frac{\mu_1}{\sigma_0(1-\sigma_1^2/\sigma_0^2)} = \frac{\mu_1}{\sigma_0 - \sigma_1^2/\sigma_0}$) sowie

$$c = -\frac{\eta\mu_1^2}{(1 - \sigma_1^2/\sigma_0^2)^2}\Big(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\Big) \tag{A.6}$$

damit schreibt sich (A.4) zunächst als

$$\Big\{x = (x_1, \dots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n \Big(x_i - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{\mu_1^2}{(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big\} \tag{A.7}$$

Für $\vartheta = (0, \sigma^2) \in \Theta_0$ ist (mit Z_i wie oben in (A.1))

$$\begin{aligned}
&\mathbb{P}_\vartheta\Big(\sum_{i=1}^n \Big(X_i - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{\mu_1^2}{(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big)\\
&= \mathbb{P}\Big(\sum_{i=1}^n \Big(\sigma Z_i - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{\mu_1^2}{(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big)\\
&= \mathbb{P}\Big(\sum_{i=1}^n \Big(Z_i - \frac{\mu_1}{\sigma(1-\sigma_1^2/\sigma_0^2)}\Big)^2 \le \frac{\mu_1^2}{\sigma^2(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big) \le \alpha
\end{aligned}$$

(denn mit $\xi = \frac{\mu_1}{\sigma(1-\sigma_1^2/\sigma_0^2)}$ ist $\xi = \xi_*$ genau dann, wenn $\sigma = \sigma_0$).

Demnach ist ψ ein Test von $\Theta_0 = \{0\} \times (0,\infty)$ gegen $\Theta_1 = (\mathbb{R} \times (0,\infty)) \setminus \Theta_0$, der Niveau α einhält.

Betrachten wir andererseits $\vartheta' = (\mu_1, \sigma_1^2) \in \Theta_1$, so ist die Wahrscheinlichkeit, dass ψ Θ_0 ablehnt, gleich

$$\begin{aligned}
G_\psi(\vartheta') &= \mathbb{P}_{\vartheta'}\Big(\sum_{i=1}^n \Big(X_i - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{\mu_1^2}{(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big)\\
&= \mathbb{P}\Big(\sum_{i=1}^n \Big(\sigma_1 Z_i + \mu_1 - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{\mu_1^2}{(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big)
\end{aligned}$$

mit $\mu_1 - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2} = -\mu_1 \frac{\sigma_1^2/\sigma_0^2}{1-\sigma_1^2/\sigma_0^2} = -\frac{\mu_1\sigma_1^2}{\sigma_0^2-\sigma_1^2}$ also

$$G_\psi(\vartheta') = \mathbb{P}\Big(\sum_{i=1}^n \Big(Z_i - \frac{\mu_1\sigma_1^2}{\sigma_0^2-\sigma_1^2}\Big)^2 \le \frac{\mu_1^2}{\sigma_1^2(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big)$$

d.h. dies ist die Wahrscheinlichkeit, dass eine nicht-zentral χ^2-verteilte Zufallsgröße[3] mit n Freiheitsgraden und Nichtzentralitätsparameter $\frac{n\mu_1^2\sigma_1^4}{(\sigma_0^2-\sigma_1)^2}$ einen Wert $\le \frac{\mu_1^2}{\sigma_1^2(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)$ annimmt.

Für allgemeines $\vartheta' = (\mu, \sigma^2) \in \Theta_1$ (d.h. $\mu \ne 0, \sigma^2 > 0$) ist

$$\begin{aligned}
G_\psi(\vartheta') &= \mathbb{P}_{\vartheta'}\Big(\sum_{i=1}^n \Big(X_i - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{\mu_1^2}{(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big)\\
&= \mathbb{P}\Big(\sum_{i=1}^n \Big(\sigma Z_i + \mu - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{\mu_1^2}{(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big)\\
&= \mathbb{P}\Big(\sum_{i=1}^n \Big(Z_i + \frac{\mu}{\sigma} - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2 \le \frac{\mu_1^2}{\sigma^2(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)\Big)
\end{aligned}$$

d.h. dies ist die Wahrscheinlichkeit, dass eine nicht-zentrale χ^2-Verteilung mit k Freiheitsgraden und Nichtzentralitätsparameter $n\Big(\frac{\mu}{\sigma} - \frac{\mu_1}{1-\sigma_1^2/\sigma_0^2}\Big)^2$ einen Wert $\le \frac{\mu_1^2}{\sigma^2(1-\sigma_1^2/\sigma_0^2)^2}(n-\eta)$ annimmt.

[3] Für $Z_1, \dots, Z_k$ u.i.v. $\sim \mathcal{N}(0,1)$, $\mu_1, \dots, \mu_k \in \mathbb{R}$ heißt die Verteilung von $\sum_{i=1}^k (Z_i+\mu_i)^2$ die nicht-zentrale χ^2-Verteilung mit k Freiheitsgraden und Nichtzentralitätsparameter $\lambda = \sum_{i=1}^k \mu_i^2$. Die Verteilung von tatsächlich nur von k und λ ab, siehe dazu z.B. [LR06, Problem 7.2], wo auch ein Ausdruck für die Dichte angegeben wird.

```
Gpsi <- function(mw, sigm) {
pchisq((n-eta)*mw^2/(sigm^2*(1-sigma1^2/sigma0^2)^2),
       df=n,ncp=n*(mw/sigm - mu1/(1-sigma1^2/sigma0^2))^2)
}
```

Unter $\mathbb{P}_{\vartheta'}$ ist T nicht-zentral Student-verteilt[4] mit $n-1$ Freiheitsgraden und Nichtzentralitätsparameter $\sqrt{n}\mu/\sigma$, d.h. für den t-Test φ zum Niveau α ist

$$G_\varphi(\vartheta') = t_{n-1,\sqrt{n}\mu/\sigma}\big((-\infty, -q_{\text{Student}(n-1),1-\alpha/2}]\big) + t_{n-1,\sqrt{n}\mu/\sigma}\big([q_{\text{Student}(n-1),1-\alpha/2}, \infty)]\big)$$

```
Gt.test <- function(mw, sigm) {
  (pt(-q,df=n-1,ncp=sqrt(n)*mw/sigm)+
    pt(q,df=n-1,ncp=sqrt(n)*mw/sigm,lower.tail=FALSE)) }
```

Vergleichen wir im Beispiel:

```
n<- 20
alpha <- 0.05; q <- qt(1-alpha/2,df=n-1)
eta <- 2.62; xi.stern <- 2.75
mu1 <- 0.5; sigma1 <- 1.0
sigma0 <- (mu1 + sqrt(mu1^2+4*sigma1^2*xi.stern^2))/(2*xi.stern)
sigma0
```

```
[1] 1.095033
```

```
# Wir vergleichen die Macht bei theta'=(mu_1,sigma_1^2):
Gt.test(mu1,sigma1)
```

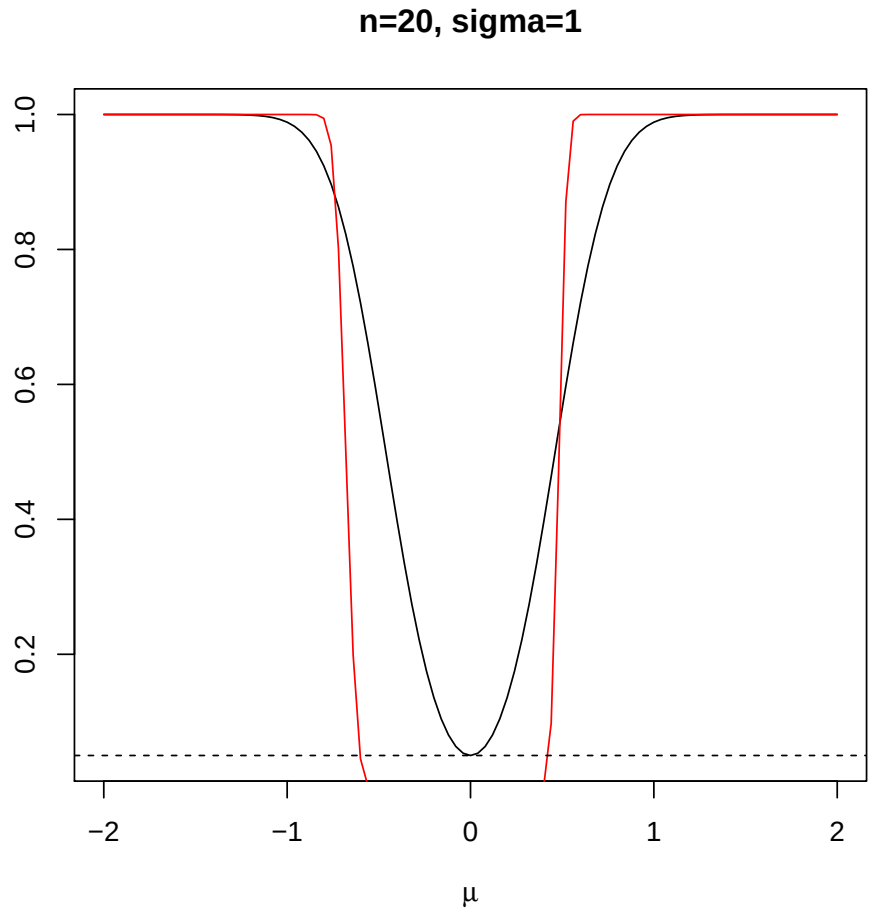
```
[1] 0.5645044
```

```
Gpsi(mu1,sigma1)
```

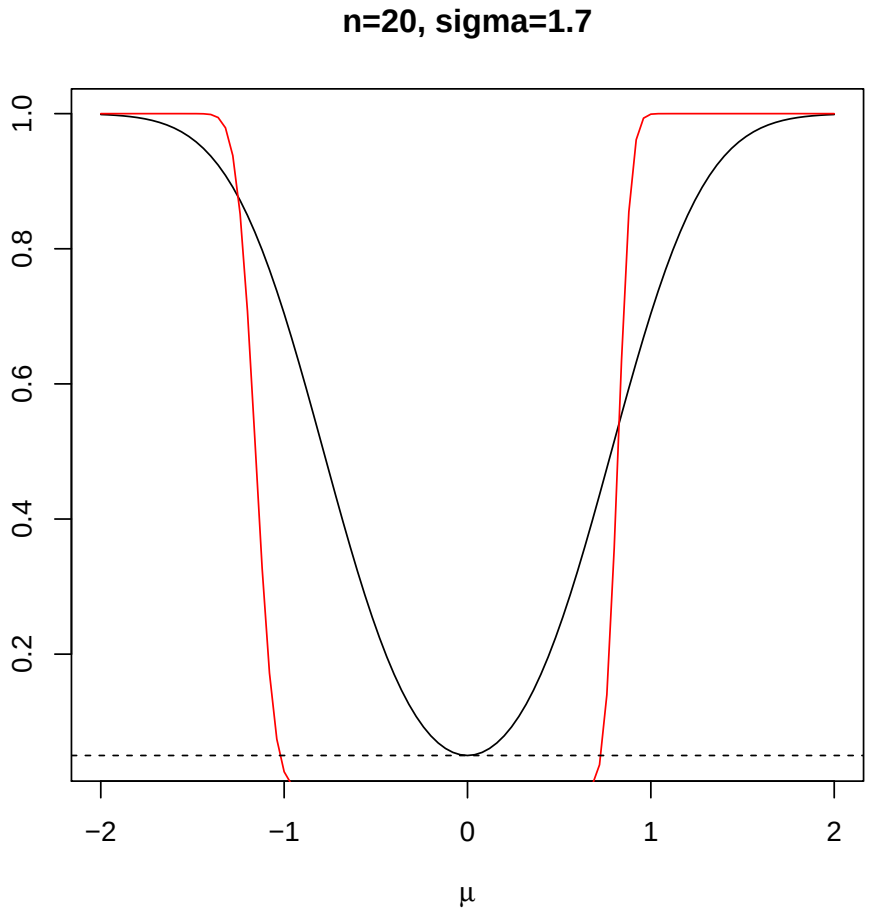
```
[1] 0.6994541
```

```
# Schauen wir den Vergleich als Funktion von mu an
# (mit sigma=sigma_1, dem fuer psi "massgeschneiderten" sigma):
plot.function(function(x) Gt.test(x,sigma1), xlim=c(-2,2),
              xlab=expression(mu), ylab='',main=paste('n=',n,', sigma=',sigma1,sep='',
plot.function(Vectorize(function(x) Gpsi(x,sigma1)), xlim=c(-2,2),col='red',add=TRUE)
abline(h=alpha,lty=2)
```

[4] Für $Z \sim \mathcal{N}(0,1)$, $Y \sim \chi^2_k$ u.a. und $a \in \mathbb{R}$ heißt die Verteilung von $(Z+a)/\sqrt{Y/k}$ die nicht-zentrale t-Verteilung (oder nicht-zentrale Student-Verteilung) mit k Freiheitsgraden und Nichtzentralitätsparameter a; siehe z.B. [LR06, Problem 5.3] für einen Ausdruck für die Dichte.




```
# Wie schaut es bei einem anderen sigma aus?
sigma2 <- 1.7
plot.function(function(x) Gt.test(x,sigma2), xlim=c(-2,2),
              xlab=expression(mu), ylab='',
              main=paste('n=',n,', sigma=',sigma2,sep=''))
plot.function(Vectorize(function(x) Gpsi(x,sigma2)), xlim=c(-2,2),col='red',add=TRUE)
abline(h=alpha,lty=2)
```


n=20, sigma=1.7
1.0
0.8
0.6
0.4
0.2
−2
−1
0
1
2
μ