Mathematische Statistik

Sommersemester 2017

Mathias Trabs*
Universität Hamburg

9. Juni 2017

Inhaltsverzeichnis

1 Grundbegriffe der Statistik **2**
1.1 Drei grundlegende Fragestellungen . . . 3
1.1.1 Schätzprobleme . . . 3
1.1.2 Hypothesentests . . . 6
1.1.3 Konfidenzmengen (Bereichsschätzung) . . . 8
1.2 Minimax- und Bayesansatz . . . 9
1.3 *Ergänzungen: Quantile . . . 13

2 Lineares Modell **14**
2.1 Regression und kleinste Quadrate . . . 14
2.2 Inferenz unter Normalverteilungsannahme . . . 18

3 Exponentialfamilien und Optimalität **23**
3.1 Exponentialfamilien . . . 23
3.2 Suffizienz und Vollständigkeit . . . 24
3.3 Cramér-Rao-Effizienz . . . 28
3.4 Asymptotik für den Maximum-Likelihood-Schätzer . . . 33
3.5 *Ergänzung: Verallgemeinertes lineares Modell . . . 36

4 Testtheorie **39**
4.1 Neyman-Pearson-Tests . . . 39
4.2 Likelihood-Quotienten- und χ^2-Test . . . 43

*Email: mathias.trabs@uni-hamburg.de

1 Grundbegriffe der Statistik

Während die Wahrscheinlichkeitstheorie anhand eines gegebenen Modells die Eigenschaften der (zufälligen) Ereignisse untersucht, ist das Ziel der Statistik genau andersherum: Wie kann man aus den gegebenen Beobachtungen Rückschlüsse auf das Modell ziehen?

Beispiel 1.1 (Werbung)**.** Wir verwenden den "Advertising"-Datensatz aus James et al. (2013). Für 200 Märkte haben wir die Anzahl der verkauften Produkte Y sowie das jeweilige Budget für Fernsehwerbung X^F, für Radiowerbung X^R und für Zeitungsannoncen X^Z gegeben.

Betrachten wir das *Modell*

$$Y_i = aX_i^F + b + \varepsilon_i, \quad i = 1, \dots, 200,$$

wobei die zufälligen Störgrößen ε_i Marktunsicherheiten, externe Einflüsse etc. modellieren. Plausible Annahmen an das Modell sind

(i) (ε_i) sind unabhängig (näherungsweise),

(ii) (ε_i) sind identisch verteilt,

(iii) $\mathbb{E}[\varepsilon_i] = 0$ (kein systematischer Fehler)

(iv) ε_i normalverteilt (wegen ZGWS).

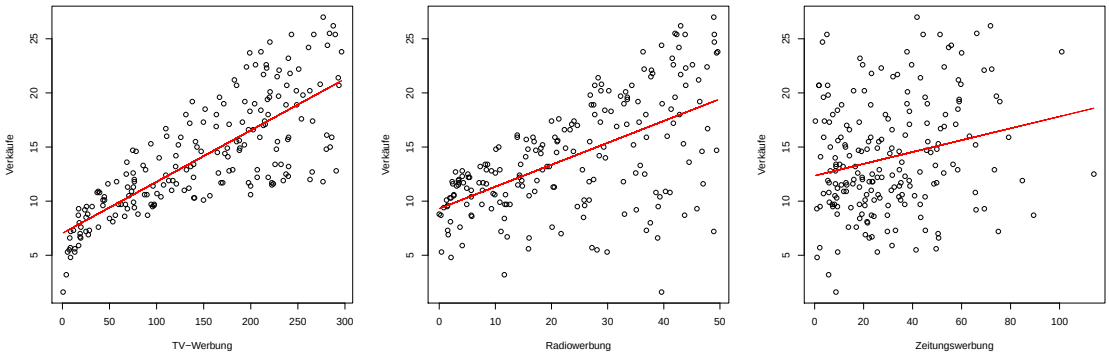

Abbildung 1: Budget für Fernseh-, Radio bzw. Zeitungswerbung sowie die jeweilen Verkaufszahlen mit den resultierenden Regressionsgraden aus Beispiel 1.1.

Naheliegende *Ziele/Fragestellungen:*

(i) Es sollen a, b anhand der Daten ermittelt werden. Ein mögliches Schätzverfahren ist der *Kleinste-Quadrate-Schätzer*

$$(\widehat{a}, \widehat{b}) := \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - aX_i - b)^2$$

(wir minimieren die Summe der quadrierten Residuen). Mit $\widehat{a}, \widehat{b}$ erhalten wir die *Regressionsgrade*

$$y = \widehat{a}x^F + \widehat{b}.$$

(ii) Sind die Modellannahmen erfüllt? Histogramm, Boxplot und QQ-Plot (Quantil-Quantil-Plot) der Residuen.

(iii) Wenn wir die Verteilung von $\widehat{a}$ kennen (Verteilungsannahme an ε nötig!), können wir Intervalle der Form $I = [\widehat{a} - c, \widehat{a} + c]$ für $c > 0$ konstruieren, sodass der tatsächliche Parameter a mit vorgegebener Wahrscheinlichkeit in I liegt.

(iv) Wir wollen *testen,* ob es einen Effekt gibt, d.h. gilt die Hypothese $H_0 : a = 0$ oder kann sie verworfen werden? Beispielsweise kann man die Hypothese verwerfen, falls $|\widehat{a}| > c$ für einen kritischen Wert $c > 0$. Um einen sinnvollen Wert zu bestimmen, benötigen wir wieder Verteilungsannahmen an die Fehler (ε_i).

Definition 1.2. Ein messbarer Raum $(\mathcal{X}, \mathscr{F})$ versehen mit einer Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ von Wahrscheinlichkeitsmaßen mit einer beliebigen Parametermenge $\Theta \neq \emptyset$ heißt statistisches Experiment oder statistisches Modell. $\mathcal{X}$ heißt Stichprobenraum. Jede $(\mathscr{F}, \mathscr{S})$-messbare Funktion $Y\colon \mathcal{X} \to \mathcal{S}$ heißt Beobachtung oder Statistik mit Werten in $(\mathcal{S}, \mathscr{S})$ und induziert das statistische Modell $(\mathcal{S}, \mathscr{S}, (\mathbb{P}_\vartheta^Y)_{\vartheta\in\Theta})$. Sind die Beobachtungen $Y_1, \dots, Y_n$ für jedes $\mathbb{P}_\vartheta$ unabhängig und identisch verteilt (i.i.d.), so nennt man $Y_1, \dots, Y_n$ eine mathematische Stichprobe.

Bemerkung 1.3.

(i) Gilt $\Theta \subseteq \mathbb{R}^k$, so heißt das statistische Modell *parametrisch* und sonst *nichtparametrisch.* Wir werden in dieser Vorlesung (fast) ausschließlich parametrische Modelle betrachten.

(ii) Für $n \in \mathbb{N}$ sei $X_1, \dots, X_n$ eine mathematische Stichprobe mit Werten in $\mathcal{X}$ und Randverteilung $X_1 \sim \mathbb{P}_\vartheta$ mit Parameter $\vartheta \in \Theta$. Dann ist der Stichprobenvektor $(X_1, \dots, X_n)$ gemäß dem Produktmaß $\mathbb{P}_\vartheta^n(\mathrm{d}x) = \prod_{i=1}^n \mathbb{P}_\vartheta(\mathrm{d}x_i)$ auf $(\mathcal{X}^n, \mathscr{F}^{\otimes n})$ verteilt.

Beispiel 1.4.

(i) *Parametrisch*: $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ mit unbekannten Parametern μ, σ^2. Wir betrachten also den Parameterraum $(\mu, \sigma) \in \Theta = \mathbb{R} \times (0, \infty)$. Äquivalent können wir fordern/annehmen, dass für die Verteilungsfunktion F der X_i gilt: $F \in \{\Phi((x - \mu)/\sigma) : \mu \in \mathbb{R}, \sigma > 0\}$ mit der Verteilungsfunktion Φ der $\mathcal{N}(0, 1)$-Verteilung.

(ii) *Nichtparametrisch*: $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$ für eine Verteilungsfunktion F, wobei wir die ganze Funktion F als Parameter betrachten. Der Parameterraum ist damit

$$\Theta := \big\{G\colon \mathbb{R} \to [0, 1] : G \text{ monoton wachsend, càdlàg und } \lim_{x\to-\infty} G(x) = 0, \lim_{x\to+\infty} G(x) = 1\big\}.$$

Merke: „Alle Modelle sind falsch, doch manche sind nützlich." (George Box).

1.1 Drei grundlegende Fragestellungen

Die meisten statistischen Fragestellungen kann man einer der drei Grundprobleme *Schätzen*, *Testen* und *Konfidenzintervalle* zuordnen. Diese werden im Folgenden kurz umrissen und im Laufe der Vorlesung weiter vertieft.

1.1.1 Schätzprobleme

Ziel ist es, aufgrund der vorhandenen Beobachtungen den unbekannten Parameter im statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ zu bestimmen, also einen einzelnen (bestmöglichen) Wert anzugeben (*Punktschätzung*). Damit ist ein Schätzer eine Abbildung, die nur von den Beobachtungen abhängt.

Definition 1.5. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\rho : \Theta \to \mathbb{R}^d$ ein (abgeleiteter) d-dimensionaler Parameter, $d \in \mathbb{N}$. Ein Schätzer ist eine messbare Abbildung $\widehat{\rho}\colon \mathcal{X} \to \mathbb{R}^d$. Gilt $\mathbb{E}_\vartheta[\widehat{\rho}] = \rho(\vartheta)$ so heißt $\widehat{\rho}$ unverzerrt oder erwartungstreu (engl.: unbiased).

Beispiel 1.6. Seien $X_1, \dots, X_n$ eine Bernoulli-verteilte mathematische Stichprobe mit Parameter $p \in (0, 1)$. Betrachte den Schätzer $\widehat{p}_n := n^{-1} \sum_{i=1}^n X_i$. Dann gilt $\mathbb{E}_p[\widehat{p}_n] = n^{-1} \sum_{i=1}^n \mathbb{E}_p[X_i] = p$. Also ist $\widehat{p}_n$ erwartungstreu. Um die Streuung des Schätzers um den wahren Parameter p zu messen, berechnen wir

$$\mathrm{Var}_p(\widehat{p}_n) = n^{-2} \sum_{i=1}^n \mathrm{Var}_p(X_i) = \frac{p(1 - p)}{n}.$$

Für größer werdenden Stichprobenumfang konzentriert sich also $\widehat{p}_n$ um p.

Wie gut ein Schätzer ist, wird mithilfe einer Verlustfunktion bestimmt. Diese misst den Abstand zwischen geschätztem und wahrem Parameter.

Definition 1.7. Eine Funktion $\ell\colon \Theta \times \mathbb{R}^d \to \mathbb{R}_+$ heißt Verlustfunktion, falls $\ell(\vartheta, \cdot)$ für jedes $\vartheta \in \Theta$ messbar ist. Der erwartete Verlust $R(\vartheta, \widehat{\rho}) := \mathbb{E}_\vartheta[\ell(\vartheta, \widehat{\rho})]$ eines Schätzers $\widehat{\rho}$ heißt Risiko. Typische Verlustfunktionen sind

(i) der 0-1-Verlust $\ell(\vartheta, r) = \mathbb{1}_{\{r \neq \rho(\vartheta)\}}$,

(ii) der absolute Verlust $\ell(\vartheta, r) = |r - \rho(\vartheta)|$ (euklidischer Abstand im $\mathbb{R}^d$) sowie

(iii) der quadratische Verlust $\ell(\vartheta, r) = |r - \rho(\vartheta)|^2$.

Lemma 1.8 (Bias-Varianz-Zerlegung)**.** *Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell und $\widehat{\rho}\colon \mathcal{X} \to \mathbb{R}^d$ ein Schätzer des Parameters $\rho(\vartheta)$ mit $\mathbb{E}_\vartheta[|\widehat{\rho}|^2] < \infty$ für alle $\vartheta \in \Theta$. Dann gilt für den quadratischen Verlust*

$$\mathbb{E}_\vartheta\big[|\widehat{\rho} - \rho(\vartheta)|^2\big] = \mathrm{Var}_\vartheta(\widehat{\rho}) + |\underbrace{\mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta)}_{Bias}|^2 \quad \textit{für alle } \vartheta \in \Theta.$$

Beweis. Es gilt

$$\begin{aligned}\mathbb{E}_\vartheta\big[|\widehat{\rho} - \rho(\vartheta)|^2\big] &= \mathbb{E}_\vartheta\big[|\widehat{\rho} - \mathbb{E}_\vartheta[\widehat{\rho}] + \mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta)|^2\big] \\ &= \mathbb{E}_\vartheta[|\widehat{\rho} - \mathbb{E}_\vartheta[\widehat{\rho}]|^2] + 2\mathbb{E}_\vartheta\big[(\widehat{\rho} - \mathbb{E}_\vartheta[\widehat{\rho}])^\top(\mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta))\big] + |\mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta)|^2 \\ &= \mathrm{Var}_\vartheta(\widehat{\rho}) + |\mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta)|^2. \end{aligned}$$

□

Beispiel 1.6 (Fortsetzung). *In der Situation von Beispiel 1.6 betrachten wir den Schätzer $\widetilde{p}_n := (\sum_{i=1}^n X_i + 1)/(n+2)$. Dieser hat den Bias*

$$\mathbb{E}_p[\widetilde{p}_n] - p = \frac{1-2p}{n+2}$$

und die Varianz

$$\mathrm{Var}_p(\widetilde{p}_n) = \frac{np(1-p)}{(n+2)^2}.$$

Damit hat $\widetilde{p}_n$ einen kleineren quadratischen Fehler als $\widehat{p}_n$, wenn $|p - 1/2| \leqslant 1/\sqrt{8}$.

Obwohl wir in dieser Vorlesung nur wenig Asymptotik behandeln, also das Verhalten der Schätzer bei Stichprobenumfängen $n \to \infty$, seien noch zwei weitere wichtige Grundbegriffe erwähnt.

Definition 1.9. Sei $X_1, \ldots, X_n \overset{iid.}{\sim} \mathbb{P}_\vartheta$ eine mathematische Stichprobe. Dann heißt ein Schätzer $\widehat{\rho}_n$ für den abgeleiteten Parameter $\rho(\vartheta)$ konsistent, falls

$$\widehat{\rho}_n \overset{\mathbb{P}_\vartheta}{\to} \rho(\vartheta) \quad \text{für} \quad n \to \infty.$$

Der Schätzer $\widehat{\rho}_n$ heißt asymptotisch normalverteilt, falls $\mathbb{E}[|\widehat{\rho}_n|^2] < \infty$ und

$$\frac{\widehat{\rho}_n - \mathbb{E}_\vartheta[\widehat{\rho}_n]}{\sqrt{\mathrm{Var}_\vartheta(\widehat{\rho}_n)}} \overset{d}{\to} \mathcal{N}(0,1) \quad \text{unter} \quad \mathbb{P}_\vartheta.$$

Aufgrund des zentralen Grenzwertsatzes sind viele Schätzer asymptotisch normalverteilt, so auch in Beispiel 1.6 (Übung ☑). Daher kommt der Untersuchung von statistischen Modellen unter Normalverteilungsannahme eine besondere Bedeutung zu.

Zwei wichtige Konstruktionsprinzipien von Schätzern sind die Momentenmethode und Maximum-Likelihood-Schätzer:

Momentenmethode: Sei $X_1, \dots, X_n$ eine mathematische Stichprobe reeller Zufallsvariablen mit $\mathbb{E}[|X_1|^d] < \infty$. Offensichtlich hängen i.A. die Momente einer Verteilung $m_k = m_k(\vartheta) := \mathbb{E}_\vartheta[X_1^k], k \in \mathbb{N}$, von ihrem Parameter $\vartheta \in \mathbb{R}^d$ ab. Aufgrund des Gesetzes der großen Zahlen ist der kanonische Schätzer von m_k gegeben durch das Stichprobenmoment $\widehat{m}_k := \frac{1}{n}\sum_{j=1}^n X_j^k$. Der Momentenschätzer $\widehat{\vartheta}$ von ϑ ist definiert als die Lösung der d-Gleichungen

$$\begin{aligned} m_1(\widehat{\vartheta}) &= \widehat{m}_1, \\ m_2(\widehat{\vartheta}) &= \widehat{m}_2, \\ \vdots \quad &\vdots \quad \vdots \\ m_d(\widehat{\vartheta}) &= \widehat{m}_d. \end{aligned}$$

Beispiel 1.10. Sei $X_1, \dots, X_n \overset{iid.}{\sim} \mathcal{N}(\mu, \sigma^2)$. Dann ist $m_1 = \mathbb{E}_{\mu,\sigma^2}[X_1] = \mu$ und $m_2 = \mathbb{E}_{\mu,\sigma^2}[X_1^2] = \mathrm{Var}_{\mu,\sigma^2}(X_1) + \mathbb{E}_{\mu,\sigma^2}[X_1]^2 = \sigma^2 + \mu^2$. Folglich müssen wir die Gleichungen

$$\widehat{\mu} = \frac{1}{n}\sum_{j=1}^n X_j \quad \text{und} \quad \widehat{\sigma}^2 + \widehat{\mu}^2 = \frac{1}{n}\sum_{j=1}^n X_j^2$$

lösen. Bezeichnen wir das Stichprobenmittel mit $\overline{X}_n := \frac{1}{n}\sum_{j=1}^n X_j$, erhalten wir die Lösung

$$\widehat{\mu} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \overline{X}_n)^2.$$

Die Momentenmethode kann auf die Erwartungswerte allgemeinerer Funktionale verallgemeinert werden (siehe Übung ☑). Für die zweite Methode benötigen wir etwas mehr Struktur, die wir auch im weiteren Verlauf der Vorlesung immer wieder aufgreifen.

Definition 1.11. Ein statistisches Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt dominiert, falls es ein σ-endliches Maß μ gibt, sodass für alle $\vartheta \in \Theta$ das Maß $\mathbb{P}_\vartheta$ eine Dichte $L(\vartheta, \cdot)$ bzgl. μ besitzt:

$$\forall A \in \mathscr{F}, \vartheta \in \Theta: \quad \mathbb{P}_\vartheta(A) = \int_A L(\vartheta, x)\mu(\mathrm{d}x).$$

$L(\vartheta, x)$ heißt Likelihoodfunktion, wobei diese meist als durch x parametrisierte Funktion in ϑ aufgefasst wird.

Bemerkung 1.12. Wie in der Vorlesung "Maßtheoretische Konzepte der Stochastik" gezeigt wird, folgt aus der *Absolutstetigkeit von $\mathbb{P}_\vartheta$ bzgl. μ, d.h.* $\mathbb{P}_\vartheta(A) = 0$ für alle $A \in \mathscr{F}$ mit $\mu(A) = 0$, die Existenz einer Dichte $L(\vartheta, \cdot) = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}$ (Satz von Radon-Nikodym).

Beispiel 1.13.

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

(ii) Jedes statistische Modell auf dem Stichprobenraum $(\mathbb{N}, \mathscr{P}(\mathbb{N}))$ oder allgemeiner auf einem abzählbaren Raum $(\mathcal{X}, \mathscr{P}(\mathcal{X}))$ ist vom Zählmaß dominiert. Die Likelihoodfunktion ist durch die Zähldichte gegeben.

(iii) Ist $\Theta = \{\vartheta_1, \vartheta_2, \dots\}$ abzählbar, so ist $\mu = \sum_i c_i \mathbb{P}_{\vartheta_i}$ mit $c_i > 0$ und $\sum_i c_i = 1$ ein dominierendes Maß.

Maximum-Likelihood-Prinzip: Für ein dominiertes statistisches Modell mit Likelihoodfunktion $L(\vartheta, x)$ heißt eine Statistik $\widehat{\vartheta}\colon \mathcal{X} \to \Theta$ (Θ trage eine σ-Algebra) Maximum-Likelihood-Schätzer (MLE: maximum likelihood estimator), falls

$$L(\widehat{\vartheta}, x) = \sup_{\vartheta\in\Theta} L(\vartheta, x) \quad \text{für } \mathbb{P}_\vartheta\text{-f.a. } x \in \mathcal{X} \text{ und alle } \vartheta \in \Theta.$$

Beispiel 1.14. Betrachten wir wieder eine mathematische Stichprobe $X_1, \dots, X_n$ normalverteilter Zufallsvariablen. Dann ist $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathbb{P}^n_{\mu,\sigma^2})$ mit $\mathbb{P}_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2)$ ein vom Lebesguemaß auf $\mathbb{R}^n$ dominiertes Modell mit Likelihoodfunktion, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$,

$$L(\mu, \sigma^2; x) = (2\pi\sigma^2)^{-n/2} \prod_{j=1}^n \exp\Big(-\frac{(x_j-\mu)^2}{2\sigma^2}\Big).$$

Um den Maximum-Likelihood-Schätzer zu berechnen, nutzen wir die Monotonie des Logarithmus und betrachten

$$\log L(\mu, \sigma^2; x) = -\frac{n}{2}(\log(2\pi) + \log\sigma^2) - \sum_{j=1}^n \frac{(x_j-\mu)^2}{2\sigma^2} \to \max_{\mu,\sigma^2}.$$

Ableiten nach μ und σ^2 führt auf die Gleichungen

$$0 = \sigma^{-2}\sum_{j=1}^n (x_j - \mu), \quad \frac{n}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{j=1}^n (x_j-\mu)^2.$$

Umstellen der ersten Gleichung nach μ liefert $\widehat{\mu} = \overline{X}_n$ und Einsetzen in die zweite Gleichung ergibt $\widehat{\sigma}^2 = n^{-1}\sum_j (X_j - \overline{X}_n)^2$. Es ist leicht nachzuprüfen, dass $\widehat{\mu}$ und $\widehat{\sigma}^2$ tatsächlich das Maximierungsproblem lösen (und messbar sind). In diesem Fall stimmt der Maximum-Likelihood-Schätzer also mit dem Momentenschätzer überein.

1.1.2 Hypothesentests

Häufig interessiert man sich weniger für die gesamte zugrunde liegende Verteilung als die Frage ob eine bestimmte Eigenschaft erfüllt ist, oder nicht. Beispielsweise möchte man wissen, ob eine neue Behandlungsmethode I besser ist als die alte bisher genutzte Methode II. Aufgrund einer Beobachtung soll entschieden werden, ob die Hypothese "I ist besser als II" akzeptiert werden kann oder verworfen werden sollte.

Um derartige Fragestellungen in einem statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ zu formalisieren, wird die Parametermenge in zwei disjunkte Teilmengen Θ_0 und Θ_1 zerlegt, d.h. $\Theta = \Theta_0 \cup \Theta_1$ und $\emptyset = \Theta_0 \cap \Theta_1$. Das *Testproblem* liest sich dann als

$$H_0 : \vartheta \in \Theta_0 \quad \text{versus} \quad H_1 : \vartheta \in \Theta_1.$$

Dabei werden H_0, H_1 als *Hypothesen* bezeichnet, genauer heißt H_0 *Nullhypothese* und H_1 *Alternativhypothese* oder *Alternative*. Ein statistischer Test entscheidet nun zwischen H_0 und H_1 aufgrund einer Beobachtung $x \in \mathcal{X}$.

Definition 1.15. Ein (nicht-randomisierter) statistischer Test ist eine messbare Abbildung $\varphi\colon (\mathcal{X}, \mathscr{F}) \to (\{0,1\}, \mathscr{P}(\{0,1\}))$, wobei $\varphi(x) = 1$ heißt, dass die Nullhypothese verworfen/ die Alternative angenommen wird und $\varphi(x) = 0$ bedeutet, dass die Nullhypothese nicht verworfen wird/ akzeptiert wird. Die Menge $\{\varphi = 1\} = \{x \in \mathcal{X} : \varphi(x) = 1\}$ heißt Ablehnbereich von φ.

Allgemeiner ist ein randomisierter statistischer Test eine messbare Abbildung $\varphi\colon (\mathcal{X}, \mathscr{F}) \to ([0,1], \mathscr{B}([0,1]))$. Im Fall $\varphi(x) \in (0,1)$ entscheidet ein unabhängiges Bernoulli-Zufallsexperiment mit Erfolgswahrscheinlichkeit $p = \varphi(x)$, ob die Hypothese verworfen wird.

Testen beinhaltet mögliche Fehlerentscheidungen:

(i) Fehler 1. Art (α-Fehler, type I error): Entscheidung für H_1, obwohl H_0 wahr ist,

(ii) Fehler 2. Art (β-Fehler, type II error): Entscheidung für H_0, obwohl H_1 wahr ist.

Definition 1.16. Sei φ ein Test der Hypothese $H_0 : \vartheta \in \Theta_0$ gegen die Alternative $H_1 : \vartheta \in \Theta_1$ im statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$. Die Gütefunktion von φ ist definiert als

$$\beta_\varphi \colon \Theta \to \mathbb{R}_+, \vartheta \mapsto \mathbb{E}_\vartheta[\varphi]$$

Ein Test φ erfüllt das Signifikanzniveau $\alpha \in [0,1]$ (oder φ ist Test zum Niveau α), falls $\beta_\varphi(\vartheta) \leqslant \alpha$ für alle $\vartheta \in \Theta_0$. Ein Test φ zum Niveau α heißt unverfälscht, falls $\beta_\varphi(\vartheta) \geqslant \alpha$ für alle $\vartheta \in \Theta_1$.

Somit hat ein nicht-randomisierten Test das Niveau $\alpha \in (0,1)$, falls

$$\mathbb{P}_\vartheta(\varphi = 1) \leqslant \alpha, \quad \text{für alle } \vartheta \in \Theta_0,$$

beschränkt also die Wahrscheinlichkeit des Fehlers 1. Art mit der vorgegeben oberen Schranke α. In der Regel ist es nicht möglich, die Wahrscheinlichkeiten für die Fehler 1. und 2. Art gleichzeitig zu minimieren. Daher werden diese typischerweise asymmetrisch betrachtet:

(i) Begrenzung der Fehlerwahrscheinlichkeit 1. Art durch ein vorgegebenes Signifikanzniveau α.

(ii) Unter der Maßgabe (i) wird die Wahrscheinlichkeit für Fehler 2. Art minimiert.

Eine zum Niveau α statistisch abgesicherte Entscheidung kann also immer nur zu Gunsten von H_1 getroffen werden. Daraus folgt die Merkregel "Was nachzuweisen ist, stets als Alternative H_1 formulieren".

Ein allgemeines Konstruktionsprinzip von Tests verwendet **Teststatistik***:* Betrachten wir das Testproblem einer Hypothese $H_0 : \vartheta \in \Theta_0$ vs. $H_1 : \vartheta \in \Theta_1$ mit $\Theta_0 \neq \emptyset$ und $\Theta_1 = \Theta \setminus \Theta_0$. Für Ablehnbereiche $(\Gamma_\alpha)_{\alpha\in(0,1)} \subseteq \mathscr{B}(\mathbb{R})$ und eine Teststatistik $T \colon (\mathcal{X}, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ sei ein Test gegeben durch

$$\varphi(x) = \mathbb{1}_{\{T(x)\in\Gamma_\alpha\}}, \quad x \in \mathcal{X}. \tag{1.1}$$

Oft werden die Ablehnbereiche als Intervalle $\Gamma_\alpha = (c_\alpha, \infty)$ konstruiert für kritische Werte

$$c_\alpha = \inf\Big\{c \in \mathbb{R} : \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) > c) \leqslant \alpha\Big\}, \quad \alpha \in (0,1). \tag{1.2}$$

Ist $\Theta_0 = \{\vartheta_0\}$ einelementig, dann ist der kritische Wert genau das $(1-\alpha)$-*Quantil* der Verteilung von T unter $\mathbb{P}_{\vartheta_0}$.

Beispiel 1.17 (Binomialtests)**.**

(i) Von den 13 Todesfällen unter 55- bis 65-jährigen Arbeitern eines Kernkraftwerkes im Jahr 1995 waren 5 auf einen Tumor zurückzuführen. Die Todesursachenstatistik 1995 weist aus, dass Tumore bei etwa $1/5$ aller Todesfälle die Ursache in der betreffenden Altersklasse (in der Gesamtbevölkerung) darstellen. Ist die beobachtete Häufung von tumorbedingten Todesfällen signifikant zum Niveau 5%?
Bezeichne X die Anzahl der Tumortoten unter $n = 13$ Todesfällen. Dann ist das statistische Modell gegeben durch $\mathcal{X} = \{0, \dots n\}, \mathscr{F} = \mathscr{P}(\mathcal{X})$ und $\mathbb{P}_p = Bin(13, p)$ mit Parameter $p \in [0,1]$ und das Testproblem ist gegeben durch

$$H_0 : p \leqslant 1/5 \quad \text{versus} \quad H_1 : p > 1/5.$$

Ziel ist ein nicht-randomisierter Test zum Niveau $\alpha = 0,05$. Wir verwenden den *einseitigen Binomialtest* $\varphi(x) = \mathbb{1}_{\{x>c\}}$ mit Teststatistik X und kritischem Wert $c > 0$. Für jedes c mit $\sup_{p\leqslant 1/5} \mathbb{P}_p(X > c) \leqslant \alpha$ besitzt φ das Niveau α. Um eine möglichst große Güte zu erreichen, sollte c unter dieser Nebenbedingung möglichst klein gewählt werden. Für $k \in \mathcal{X}$ gilt

$$\mathbb{P}_p(X \leqslant k) = \sum_{l=0}^{k} \binom{13}{l} p^l (1-p)^{13-l}.$$

Da $p \mapsto \mathbb{P}_p(X \leqslant k)$ für alle $k \in \mathcal{X}$ monoton fallend auf $[0,1]$ ist (ableiten), folgt $\sup_{p \leqslant 1/5} \mathbb{P}_p(X > c) = \mathbb{P}_{1/5}(X > c)$. Wegen

$$\mathbb{P}_{1/5}(X \leqslant 4) \approx 0,901 \quad \text{und} \quad \mathbb{P}_{1/5}(X \leqslant 5) \approx 0,970,$$

wählen wir $c = 5$. Somit kann die Hypothese zum Niveau $0,05$ nicht verworfen werden. Die Gütefunktion von φ

$$\beta_\varphi(p) = \mathbb{P}_p(X > 5) = \sum_{l=6}^{13} \binom{13}{l} p^l (1-p)^{13-l}, \quad p \in [0,1],$$

ist monoton wachsend und somit ist φ auch unverfälscht.

(ii) Wir wollen die Hypothese "Es werden genauso viele Jungen wie Mädchen geboren." testen. Sind von $n \in \mathbb{N}$ Geburten $w \leqslant n$ Mädchen zur Welt gekommen, ist das statistische Modell gegeben durch den Stichprobenraum $\mathcal{X} = \{0, \dots, n\}$ und somit $(\mathcal{X}, \mathscr{P}(\mathcal{X}), (\mathbb{P}_\vartheta)_{\vartheta \in [0,1]})$ mit Binomialverteilungen $\mathbb{P}_\vartheta = Bin(n, \vartheta)$. Die Hypothese führt auf das zweiseitige Testproblem

$$H_0 : \vartheta = \vartheta_0 \quad \text{versus} \quad H_1 : \vartheta \neq \vartheta_0$$

mit $\vartheta_0 = 1/2$, wobei $w \in \mathcal{X}$ beobachtet wird. Wir legen das Niveau $\alpha = 0,05$ fest. Die Teststatistik $T(w) = w$ führt auf den *zweiseitigen Binomialtest*

$$\varphi(w) = 1 - \mathbb{1}_{\{\underline{c}_{\alpha,\vartheta_0} \leqslant w \leqslant \overline{c}_{\alpha,\vartheta_0}\}}$$

mit kritischen Werten

$$\underline{c}_{\alpha,\vartheta_0} = \max\big\{k \in \mathbb{N} : \mathbb{P}_{\vartheta_0}(X < k) \leqslant \alpha/2\big\} \quad \text{und} \quad \overline{c}_{\alpha,\vartheta_0} = \min\big\{k \in \mathbb{N} : \mathbb{P}_{\vartheta_0}(X > k) \leqslant \alpha/2\big\}.$$

Für $\vartheta_0 = 1/2$ erhalten wir den Test $\varphi(w) = \mathbb{1}_{\{|\frac{w}{n} - \frac{1}{2}| > c\}}$ für ein geeignetes c (Übung ☑).

Bemerkung 1.18. Bei großen Stichprobenumfängen ist es sinnvoll, einen *Gauß-Test* für eine geeignet normalisierte Teststatistik zu verwenden, um den Binomialtest zu approximieren: Für $\vartheta \in (0,1)$ normalisieren wir die Beobachtung $X \sim Bin(n, \vartheta)$ durch $Y := \frac{X - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}$. Aus dem Zentralen Grenzwertsatz folgt dann für eine standardnormalverteilte Zufallsvariable $Z \sim \mathcal{N}(0,1)$, dass

$$\begin{aligned}
\mathbb{P}_\vartheta(T(X) > c_\alpha) &= \mathbb{P}_\vartheta\Big(\frac{|X - n\vartheta|}{\sqrt{n\vartheta(1-\vartheta)}} > \sqrt{\frac{n}{\vartheta(1-\vartheta)}} c_\alpha\Big) \\
&\overset{n\to\infty}{\longrightarrow} \mathbb{P}\Big(|Z| > \sqrt{\frac{n}{\vartheta(1-\vartheta)}} c_\alpha\Big) \\
&= 2\Big(1 - \Phi\Big(\sqrt{\frac{n}{\vartheta(1-\vartheta)}} c_\alpha\Big)\Big) \overset{!}{=} \alpha,
\end{aligned}$$

Mit der Verteilungsfunktion $\Phi(x) = \mathbb{P}(Z \leqslant x)$. Folglich wählen wir $c_\alpha = \sqrt{\frac{\vartheta_0(1-\vartheta_0)}{n}} q_{1-\alpha/2} = \sqrt{\frac{\vartheta_0(1-\vartheta_0)}{n}} \Phi^{-1}(1-\alpha/2)$ mit $\vartheta = \vartheta_0$ unter H_0.

1.1.3 Konfidenzmengen (Bereichsschätzung)

Während ein (Punkt-)Schätzer einen einzelnen Wert angibt, möglichst in der Nähe des wahren Parameters, um Rückschlüsse auf das zugrunde liegende Modell zu ziehen, geben Konfidenzbereiche ein Intervall an, in dem der Parameter mit gegebener Wahrscheinlichkeit liegt.

Definition 1.19. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^d$. Eine mengenwertige Abbildung $C\colon \mathcal{X} \to \mathscr{P}(\mathbb{R}^d)$ heißt Konfidenzmenge zum

Konfidenzniveau $1-\alpha$ (oder zum Irrtumsniveau α) für $\alpha \in (0,1)$, falls die Messbarkeitsbedingung $\{x \in \mathcal{X} : \rho(\vartheta) \in C(x)\} \in \mathscr{F}$ für alle $\vartheta \in \Theta$ erfüllt ist und es gilt

$$\mathbb{P}_\vartheta(\rho(\vartheta) \in C) = \mathbb{P}_\vartheta\big(\{x \in \mathcal{X} : \rho(\vartheta) \in C(x)\}\big) \geqslant 1-\alpha \quad \text{für alle } \vartheta \in \Theta.$$

Im Fall $d = 1$ und falls $C(x)$ für jedes $x \in \mathcal{X}$ ein Intervall ist, heißt C Konfidenzintervall.

Beachte, dass $\rho(\vartheta)$ fix ist, während C zufällig ist. Man muss Konfidenzmengen also wie folgt *interpretieren*: Werden in m unabhängigen Experimenten für (verschiedene) Parameter Konfidenzmengen zum Niveau $0,95$ konstruiert, dann liegt der unbekannte Parameter in 95% der Fälle im der jeweiligen Konfidenzmenge (für m groß genug; starkes Gesetz der großen Zahlen).

Ein verbreitetes Konstruktionsprinzip für die Konfidenzintervalle ist die Verwendung eines Schätzers und dessen Verteilung, wie im nächsten Beispiel illustriert.

Beispiel 1.20. Im Bernoulli-Experiment von Beispiel 1.6 gilt für $C_n := [\widehat{p}_n - \varepsilon_n, \widehat{p}_n + \varepsilon_n]$

$$\mathbb{P}_p(p \in C_n) = \mathbb{P}_p(|\widehat{p}_n - p| < \varepsilon_n) = \mathbb{P}_p\Big(\Big|\sum_{i=1}^n (X_i - p)\Big| < n\varepsilon_n\Big) \overset{!}{\geqslant} 1-\alpha.$$

Da $\sum_{i=1}^n X_i \sim Bin(n,p)$ können wir ε_n mithilfe der Quantile der Binomialverteilung bestimmen. Für große n könnte man wieder eine Normalapproximation verwenden. Das resultierende Konfidenzintervall besitzt dann aber nur asymptotisch das Niveau $1-\alpha$.

Eine alternative Konstruktion von Konfidenzmengen bietet folgender Korrespondenzsatz:

Satz 1.21. *Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\alpha \in (0,1)$. Dann gilt:*

(i) Liegt für jedes $\vartheta_0 \in \Theta$ ein Test φ_{ϑ_0} der Hypothese $H_0 : \vartheta = \vartheta_0$ zum Signifikanzniveau α vor, so definiert $C(x) = \{\vartheta \in \Theta : \varphi_\vartheta(x) = 0\}$ eine Konfidenzmenge zum Konfidenzniveau $1-\alpha$.

(ii) Ist C eine Konfidenzmenge zum Niveau $1-\alpha$, dann ist $\varphi_{\vartheta_0}(x) = 1 - \mathbb{1}_{C(x)}(\vartheta_0)$ ein Niveau-α-Test der Hypothese $H_0 : \vartheta = \vartheta_0$.

Beweis. Nach Konstruktion erhält man in beiden Fällen,

$$\forall \vartheta \in \Theta : \forall x \in \mathcal{X} : \varphi_\vartheta(x) = 0 \quad \Longleftrightarrow \quad \vartheta \in C(x).$$

Damit ist φ_ϑ ein Test zum Niveau α für alle ϑ genau dann, wenn

$$1-\alpha \leqslant \mathbb{P}_\vartheta(\varphi = 0) = \mathbb{P}_\vartheta(\{x : \vartheta \in C(x)\})$$

und somit ist C eine Konfidenzmenge zum Niveau α. □

Beispiel 1.22. Mit Hilfe des Korrespondenzsatzes können wir ein Konfidenzintervall zum Niveau $0,95$ für die Geburtswahrscheinlichkeit von Mädchen berechnen. Im Modell aus Beispiel 1.17(ii) ist ein Konfidenzbereich gegeben durch

$$C(w) = \{\vartheta \in [0,1] : \underline{c}_{0,05,\vartheta} \leqslant w \leqslant \overline{c}_{0,05,\vartheta}\},$$

C ist sogar ein Konfidenzintervall (Übung ☑) und wird *Clopper-Pearson-Intervall* genannt.

1.2 Minimax- und Bayesansatz

Wir haben bereits verschiedene Schätzmethoden, wie den Maximum-Likelihood-Schätzer oder die Momentenmethode kennen gelernt. Natürlich gibt es noch viel mehr Konstruktionen. Wie sollte eine Methode anhand des gegeben Schätzproblems ausgewählt werden? Sei also $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^d$ und Verlustfunktion L. Als mögliches Vergleichskriterium käme die Risikofunktion $R(\vartheta, \widehat{\rho}) = \mathbb{E}_\vartheta[L(\vartheta, \widehat{\rho})]$ eines Schätzers $\widehat{\rho}$ in Frage. Beachte jedoch folgendes Beispiel:

Beispiel 1.23. Sei $X \sim \mathcal{N}(\mu, 1), \mu \in \mathbb{R}$, und $L(\mu, \widehat{\mu}) = (\widehat{\mu} - \mu)^2$. Betrachte die zwei Schätzer $\widehat{\mu}_1 = X$ und $\widehat{\mu}_2 = 5$. Die Risiken sind dann gegeben durch

$$R(\mu, \widehat{\mu}_1) = \mathbb{E}_\mu[(X - \mu)^2] = 1 \quad \text{und} \quad R(\mu, \widehat{\mu}_2) = (5 - \mu)^2.$$

Damit hat $\widehat{\mu}_1$ kleineres Risiko als $\widehat{\mu}_2$ genau dann, wenn $\mu \notin [4, 6]$.

Definition 1.24. Im statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^d$ und Verlustfunktion L, heißt ein Schätzer $\widehat{\rho}$ minimax, falls

$$\sup_{\vartheta \in \Theta} R(\vartheta, \widehat{\rho}) = \inf_{\widetilde{\rho}} \sup_{\vartheta \in \Theta} R(\vartheta, \widetilde{\rho}),$$

wobei sich das Infimum über alle Schätzer (d.h. messbaren Funktionen) $\widetilde{\rho}\colon \mathcal{X} \to \mathbb{R}^d$ erstreckt.

Definition 1.25. Der Parameterraum Θ trage eine σ-Algebra $\mathscr{F}_\Theta$, die Verlustfunktion L sei produktmessbar und $\vartheta \mapsto \mathbb{P}_\vartheta(B)$ sei messbar für alle $B \in \mathscr{F}$. Die a-priori-Verteilung π des Parameters ϑ ist gegeben durch ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathscr{F}_\Theta)$. Das zu π assoziierte Bayesrisiko eines Schätzers $\widehat{\rho}$ ist

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

Der Schätzer $\widehat{\rho}$ heißt Bayesschätzer oder Bayes-optimal (bezüglich π), falls

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

wobei sich das Infimum über alle Schätzer (d.h. messbaren Funktionen) $\widetilde{\rho}\colon \mathcal{X} \to \mathbb{R}^d$ erstreckt.

Während ein Minimaxschätzer den maximal zu erwartenden Verlust minimiert, kann das Bayesrisiko als ein (mittels π) gewichtetes Mittel der zu erwartenden Verluste angesehen werden. Alternativ wird π als die subjektive Einschätzung der Verteilung des zugrundeliegenden Parameters interpretiert.

Beispiel 1.23 (Fortsetzung). Offensichtlich kann $\widehat{\mu}_2$ kein Minimaxschätzer sein. Zunächst ist es aber nicht klar, ob es einen besseren Schätzer als $\widehat{\mu}_1$ gibt. Tatsächlich werden wir später beweisen, dass $\widehat{\mu}_1$ minimax ist. Unter der a-priori-Verteilung $\mu \sim \pi = \mathcal{U}([4, 6])$ hat jedoch $\widehat{\mu}_2$ das kleinere Bayesrisiko $R_\pi(\widehat{\mu}_2) = \frac{1}{3} < 1 = R_\pi(\widehat{\mu}_1)$.

Das Bayesrisiko kann auch als insgesamt zu erwartender Verlust in folgendem Sinne verstanden werden: Definiere $\Omega := \mathcal{X} \times \Theta$ und die gemeinsame Verteilung von Beobachtung und Parameter $\widetilde{\mathbb{P}}$ auf $(\mathcal{X} \times \Theta, \mathscr{F} \otimes \mathscr{F}_\Theta)$ gemäß $\widetilde{\mathbb{P}}(\mathrm{d}x, \mathrm{d}\vartheta) = \mathbb{P}_\vartheta(\mathrm{d}x)\pi(\mathrm{d}\vartheta)$. Bezeichnen X und T die Koordinatenprojektionen von Ω auf $\mathcal{X}$ bzw. Θ, dann gilt $R_\pi(\widehat{\rho}) = \mathbb{E}_{\widetilde{\mathbb{P}}}[L(T, \widehat{\rho}(X))]$.

Wiederholung: Auf einem Wahrscheinlichkeitsraum $(\Omega, \mathscr{F}, \mathbb{P})$ ist die bedingte Wahrscheinlichkeit eines Ereignisses $A \in \mathscr{F}$ gegeben $B \in \mathscr{F}$ mit $\mathbb{P}(B) > 0$ definiert als $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ (ist $\mathbb{P}(B) = 0$ setzen wir $\mathbb{P}(A|B) = 0$). Sei $\Omega = \bigcup_{i \in I} B_i$ eine abzählbare Zerlegung in paarweise disjunkte Ereignisse $B_i \in \mathscr{F}$, dann besagt die **Bayesformel** für jedes $A \in \mathscr{F}$ mit $\mathbb{P}(A) > 0$ und alle $k \in I$

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(B_k)\mathbb{P}(A|B_k)}{\sum_{i \in I} \mathbb{P}(B_i)\mathbb{P}(A|B_i)}.$$

Mittels bedingten Erwartungswerten (Maßtheorie) kann diese Formel auf Dichten ausgedehnt werden.

Definition 1.26. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein von μ dominiertes statistisches Modell mit Dichten $f_{X|T=\vartheta} = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}$. Sei π eine a-priori-Verteilung auf $(\Theta, \mathscr{F}_\Theta)$ mit Dichte f_T bzgl. eines Maßes ν. Ist

$f_{X|T=\cdot}\colon \mathcal{X}\times\Theta\to\mathbb{R}_+$ $(\mathscr{F}\otimes\mathscr{F}_\Theta)$-messbar, dann ist die a-posteriori-Verteilung des Parameters gegeben der Beobachtung $X=x$ definiert durch die ν-Dichte

$$f_{T|X=x}(\vartheta)=\frac{f_{X|T=\vartheta}(x)f_T(\vartheta)}{\int_\Theta f_{X|T=t}(x)f_T(t)\nu(\mathrm{d}t)},\quad \vartheta\in\Theta\quad (\widetilde{\mathbb{P}}^X\text{-f.ü.}). \tag{1.3}$$

Das a-posteriori-Risiko eines Schätzers $\widehat{\rho}$ gegeben $X=x$ ist definiert durch

$$R_\pi(\widehat{\rho}|x)=\int_\Theta L(\vartheta,\widehat{\rho}(x))f_{T|X=x}(\vartheta)\nu(\mathrm{d}\vartheta).$$

Beachte, dass im Nenner in (1.3) die Randdichte $f_X=\int_\Theta f_{X|T=t}(\cdot)f_T(t)\nu(\mathrm{d}t)$ bzgl. μ von X in $(\mathcal{X}\times\Theta,\mathscr{F}\otimes\mathscr{F}_\Theta,\widetilde{\mathbb{P}})$ steht, sodass der Nenner in (1.3) für $\widetilde{\mathbb{P}}^X$-f.a. $x\in\mathcal{X}$ größer als null ist.

Beispiel 1.27. Setze $\Theta=\{0,1\}, L(\vartheta,r)=|\vartheta-r|$ (0-1-Verlust) und betrachte eine a-priori-Verteilung π mit $\pi(\{0\})=:\pi_0$ und $\pi(\{1\})=:\pi_1=1-\pi_0$. Die Wahrscheinlichkeitsmaße $\mathbb{P}_0$ und $\mathbb{P}_1$ mögen Dichten p_0 und p_1 bzgl. einem Maß μ besitzen (z.B. $\mu=\mathbb{P}_0+\mathbb{P}_1$). Dann ist die a-posteriori-Verteilung durch die Zähldichte

$$f_{T|X=x}(i)=\frac{\pi_i p_i(x)}{\pi_0p_0(x)+\pi_1p_1(x)},\quad i=0,1\quad (\widetilde{\mathbb{P}}^X\text{-f.ü.})$$

gegeben. Damit ist das a-posteriori-Risiko eines Schätzers $\widehat{\vartheta}\colon\mathcal{X}\to\{0,1\}$ gegeben durch

$$R_\pi(\widehat{\vartheta}|x)=\frac{\widehat{\vartheta}(x)\pi_0p_0(x)+(1-\widehat{\vartheta}(x))\pi_1p_1(x)}{\pi_0p_0(x)+\pi_1p_1(x)}.$$

Satz 1.28. *Es gelten die Bedingungen der vorangegangenen Definition. Für das Bayesrisiko eines Schätzers $\widehat{\rho}$ gilt*

$$R_\pi(\widehat{\rho})=\int R_\pi(\widehat{\rho}|x)f_X(x)\mu(\mathrm{d}x).$$

Minimiert $\widehat{\rho}(x)$ für $\widetilde{\mathbb{P}}^X$-f.a. $x\in\mathcal{X}$ das a-posteriori-Risiko $\min_{t\in\mathrm{ran}(\rho)}R_\pi(t|x)$, dann ist $\widehat{\rho}$ Bayesschätzer.

Beweis. Aus (1.3) folgt $f_{T|X=x}(\vartheta)f_X(x)=f_{X|T=\vartheta}(x)f_T(\vartheta)$. Der Satz von Fubini ergibt

$$\begin{aligned}R_\pi(\widehat{\rho})&=\int_\Theta\int_\mathcal{X}L(\vartheta,\widehat{\rho}(x))\mathbb{P}_\vartheta(\mathrm{d}x)\pi(\mathrm{d}\vartheta)\\&=\int_\Theta\int_\mathcal{X}L(\vartheta,\widehat{\rho}(x))f_{T|X=x}(\vartheta)f_X(x)\mu(\mathrm{d}x)\nu(\mathrm{d}\vartheta)=\int_\mathcal{X}R_\pi(\widehat{\rho}|x)f_X(x)\mu(\mathrm{d}x).\end{aligned}$$

□

Korollar 1.29. *Unter quadratischem Verlust ist der Bayesschätzer gegeben durch*

$$\widehat{\rho}(x)=\int_\Theta\rho(\vartheta)f_{T|X=x}(\vartheta)\nu(\mathrm{d}\vartheta)=:\mathbb{E}[\rho(\vartheta)|X=x].$$

Der Bayesschätzer bzgl. absolutem Verlust ist gegeben durch den Median der a-posteriori-Verteilung. Für den 0-1-Verlust ist der Bayesschätzer der Modus der a-posteriori-Verteilung.

Beweis. Übung ☑. □

Durch die Wahl einer Verlustfunktion und einer a-priori-Verteilung im statistischen Modell erhalten wir nach Berechnung der a-posteriori-Verteilung und durch das vorangegangene Korollar einen expliziten **Bayesschätzer**.

Beispiel 1.30. Sei $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ eine mathematische Stichprobe mit bekanntem $\sigma^2 > 0$ und a-priori-Verteilung $\mu \sim \mathcal{N}(a, b^2)$. Mittels Bayesformel kann die a-posteriori-Verteilung für eine Realisierung $x = (x_1, \dots, x_n)$ berechnet werden:

$$\begin{aligned} f_{T|X=x}(\mu) \sim & f_{X|T=\mu}(x) f_T(\mu) \\ & \sim \exp\Big(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}\Big) \exp\Big(-\frac{(\mu-a)^2}{2b^2}\Big) \\ & \sim \exp\Big(-\frac{\mu^2 - 2\mu\overline{x}_n}{2\sigma^2/n} - \frac{\mu^2 - 2a\mu}{2b^2}\Big) \\ & \sim \exp\Big(-\frac{(b^2+\sigma^2/n)\mu^2 - 2\mu(b^2\overline{x}_n + a\sigma^2/n)}{2b^2\sigma^2/n}\Big) \\ & \sim \exp\Big(-\frac{1}{2}\Big(\frac{n}{\sigma^2}+\frac{1}{b^2}\Big)\Big(\mu - \frac{b^2\overline{x}_n}{b^2+\sigma^2/n} - \frac{a\sigma^2/n}{b^2+\sigma^2/n}\Big)^2\Big). \end{aligned}$$

Gegeben der Beobachtung X ist ϑ also a-posteriori verteilt gemäß

$$\mathcal{N}\Big(\frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X}_n + \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a, \Big(\frac{n}{\sigma^2}+\frac{1}{b^2}\Big)^{-1}\Big).$$

Der Bayesschätzer bzgl. des quadratischen Verlustes, gegeben durch den a-posteriori Mittelwert, ist damit

$$\widehat{\vartheta}_n = \frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X}_n - \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a.$$

Bemerkung 1.31. Erhalten wir bei Wahl einer Klasse von a-priori-Verteilungen für ein statistisches Modell dieselbe Klasse (i.A. mit anderen Parametern) als a-posteriori-Verteilung zurück, so nennt man die entsprechenden Verteilungsklassen konjugiert. Im obigen Beispiel haben wir gesehen, dass die Normalverteilungen zur den Normalverteilungen konjugiert sind (genauer müsste man sagen, dass für unbekannten Mittelwert in der Normalverteilung a-priori Normalverteilungen konjugiert sind). Als weiteres Beispiel sind die Beta-Verteilungen zur Binomialverteilung konjugiert (siehe Übung ☑). In diesen (Einzel-)Fällen ist es besonders einfach, die Bayesschätzer zu konstruieren. Für komplexere Modelle werden häufig computer-intensive Methoden wie MCMC (Markov Chain Monte Carlo) verwendet, um die a-posteriori-Verteilung zu berechnen (Problem: i.A. hochdimensionale Integration).

Lemma 1.32. *Unter den Bedingungen der vorangegangen Definition gilt für jeden Schätzer* $\widehat{\rho}$

$$\sup_{\vartheta\in\Theta} R(\vartheta, \widehat{\rho}) = \sup_{\pi} R_\pi(\widehat{\rho}),$$

wobei sich das zweite Supremum über alle a-priori-Verteilungen π *erstreckt. Insbesondere ist das Risiko eines Bayesschätzers stets kleiner oder gleich dem Minimaxrisiko.*

Beweis. Natürlich gilt $R_\pi(\widehat{\rho}) = \int_\Theta R(\vartheta, \widehat{\rho})\pi(\mathrm{d}\vartheta) \leqslant \sup_{\vartheta\in\Theta} R(\vartheta, \widehat{\rho})$. Durch Betrachtung der a-priori-Verteilung δ_ϑ folgt daher die Behauptung. □

Durch dieses Lemma können wir untere Schranken für das Minimaxrisiko durch das Risiko von Bayesschätzern abschätzen. Mögliche Anwendungen illustriert folgender Satz.

Satz 1.33. *Sei* $X_1, \dots, X_n$ *eine* $\mathcal{N}(\mu, \sigma^2)$*-verteilte mathematische Stichprobe mit unbekanntem* $\mu \in \mathbb{R}$ *und bekanntem* $\sigma^2 > 0$*. Bezüglich des quadratischen Risikos ist das arithmetische Mittel* $\overline{X}_n$ *ein Minimaxschätzer von* μ*.*

Beweis. Wir betrachten a-priori-Verteilungen $\mu \sim \pi = \mathcal{N}(0, b^2)$. Nach Beispiel 1.30 ist die a-posteriori-Verteilung

$$\mathcal{N}\Big(\frac{b^2\overline{X}_n}{b^2+\frac{\sigma^2}{n}}, \Big(\frac{n}{\sigma^2}+b^{-2}\Big)^{-1}\Big),$$

der Bayesschätzer bzgl. quadratischen Risikos ist gegeben durch den a-posteriori-Erwartungswert $\widehat{\mu}_n = b^2\overline{X}_n/(b^2+\sigma^2 n^{-1})$ und dessen a-posteriori-Risiko ist gegeben durch die Varianz der a-posteriori-Verteilung. Ist f_X die Randdichte von X von $\widetilde{\mathbb{P}}$, folgt aus Satz 1.28

$$\begin{aligned} R_\pi(\widehat{\mu}_n) &= \int_{\mathbb{R}^n} \mathrm{Var}_{T|X=x}(\mu) f_X(x)\mathrm{d}x \\ &= \int_{\mathbb{R}^n} \left(n\sigma^{-2}+b^{-2}\right)^{-1} f_X(x)\mathrm{d}x = \left(n\sigma^{-2}+b^{-2}\right)^{-1}. \end{aligned}$$

Somit können wir das Minimaxrisiko nach unten abschätzen:

$$\begin{aligned} \inf_{\widetilde{\mu}} \sup_{\mu\in\mathbb{R}} R(\mu,\widetilde{\mu}) = \inf_{\widetilde{\mu}} \sup_{\pi} R_\pi(\widetilde{\mu}) &\geqslant \inf_{\widetilde{\mu}} \sup_{b>0} R_{\mathcal{N}(0,b^2)}(\widetilde{\mu}) \\ &\geqslant \sup_{b>0} \inf_{\widetilde{\mu}} R_{\mathcal{N}(0,b^2)}(\widetilde{\mu}) = \sup_{b>0} \left(n\sigma^2+b^{-2}\right)^{-1} = \frac{\sigma^2}{n}, \end{aligned}$$

wie behauptet, da $R(\mu,\overline{X}_n) = \sigma^2/n$. □

1.3 *Ergänzungen: Quantile

Definition. Sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ mit Verteilungsfunktion $F(x) = \mathbb{P}((-\infty,x])$. Für $\alpha\in(0,1)$ ist das α-Quantil $q_\alpha\in\mathbb{R}$ von $\mathbb{P}$ definiert durch

$$\mathbb{P}((-\infty,q_\alpha)) \leqslant \alpha \leqslant \mathbb{P}((-\infty,q_\alpha]).$$

Die Quantilfunktion ist definiert als verallgemeinertes Inverses von F:

$$F^{-1}(p) := \inf\{x\in\mathbb{R} : F(x)\geqslant p\}, \quad p\in[0,1].$$

α-Quantile sind nicht eindeutig, falls F auf dem Niveau α irgendwo konstant ist. Es gilt aber

Lemma. *$F^{-1}(\alpha)$ ist ein α-Quantil.*

Beweis. Aufgrund der Rechtsstetigkeit von F gilt $F(F^{-1}(\alpha))\geqslant\alpha$. Für alle $x<F^{-1}(\alpha)$ gilt $F(x)<\alpha$ und wegen der linken Grenzwerte von F

$$\alpha \geqslant \lim_{r\uparrow F^{-1}(\alpha)} F(x) = \lim_{r\uparrow F^{-1}(\alpha)} \mathbb{P}((-\infty,r]) = \mathbb{P}((-\infty,r)).$$ □

Das verallgemeinerte Inverse hat folgende **Eigenschaften**:

(i) $F^{-1}(p)\leqslant x \Leftrightarrow p\leqslant F(x)$;

(ii) $F\circ F^{-1}(p)\geqslant p$ und Gleichheit gilt genau dann, wenn $p\in\operatorname{ran}F$. Die Gleichheit kann nur dann nicht gelten, wenn F unstetig bei $F^{-1}(p)$ ist;

(iii) $F^{-1}\circ F(x)\leqslant x$, wobei Gleichheit genau dann nicht gilt wenn x im Inneren oder am rechten Rand einer "Ebene" (kein Anstieg) von F liegt.

Damit gilt $F\circ F^{-1}(p)=p$ auf $(0,1)$ genau dann, wenn F stetig ist (d.h. $\operatorname{ran}F=[0,1]$) und $F^{-1}\circ F(x)=x$ gilt auf $\mathbb{R}$ genau dann, wenn F strikt monoton wachsend ist. Folglich ist F^{-1} ein echtes Inverses genau dann, wenn F stetig und streng monoton wachsend ist.

Satz. *Ist $U\sim Uni([0,1])$, dann besitzt die Zufallsvariable $F^{-1}(U)$ die Verteilungsfunktion F (Quantilstransformation). Besitzt X die Verteilungsfunktion F, dann gilt $F(X)\sim Uni([0,1])$ genau dann, wenn F stetig ist.*

Beweis. Aus (i) folgt $\mathbb{P}(F^{-1}(U)\leqslant x) = \mathbb{P}(U\leqslant F(x)) = F(x)$ für alle $x\in\mathbb{R}$. Anderseits gilt für $p\in(0,1)$ wegen (i) und (ii)

$$\mathbb{P}(F(X)\leqslant p) = \mathbb{P}(X\leqslant F^{-1}(p)) = F(F^{-1}(p)) = p \quad\Longleftrightarrow\quad p\in\operatorname{ran}F.$$ □

Schließlich wollen wir noch den **QQ-Plot** (Qunatil-Quantil-Plot) verstehen: Die empirische Verteilungsfunktion einer mathematischen Stichprobe $X_1, \dots, X_n$ ist gegeben durch $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant x\}}$. Die Verteilungsfunktion der Standardnormalverteilung ist $\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} e^{-y^2/2} \mathrm{d}y$. Für große n approximiert F_n die wahre Verteilungsfunktion F, da nach dem starken Gesetz der großen Zahlen $F_n(x) \to \mathbb{E}[\mathbb{1}_{\{X_1 \leqslant x\}}] = F(x)$ $\mathbb{P}$-f.s. für alle $x \in \mathbb{R}$ gilt (tatsächlich gilt diese Konvergenz sogar gleichmäßig auf $\mathbb{R}$ nach dem Satz von Borel-Cantelli). Falls $X_i \sim \mathcal{N}(\mu, \sigma^2)$, so gilt $F(x) = \Phi(\frac{x-m}{\sigma})$. Für die Quantilfunktion gilt also

$$F^{-1}(\Phi(x)) = \Phi^{-1}(\Phi(x)) \cdot \sigma + m = \sigma \cdot x + m,$$

d.h. $F^{-1} \circ \Phi$ ist eine Gerade. Im QQ-Plot wird F_n^{-1} (die empirischen Quantile) gegen Φ^{-1} aufgetragen und unter einer $\mathcal{N}(\mu, \sigma^2)$-Annahme sollten die Werte in etwa auf einer Geraden liegen.

2 Lineares Modell

2.1 Regression und kleinste Quadrate

Regression ist eine Methode, um den Zusammenhang zwischen einer *Zielgröße* (*Response-Variable*) Y und einem Vektor von erklärenden Variablen (*Kovariablen, Regressoren*) $X = (x_1, \dots, x_k)$ zu analysieren. Beginnen wir mit dem *einfachen linearen Modell*

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

mit Zufallsvariablen $\varepsilon_1, \dots, \varepsilon_n$, die zentriert sind ($\mathbb{E}_i[\varepsilon_i] = 0$) und endliche Varianz $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ haben. Die Parameter $a, b \in \mathbb{R}, \sigma > 0$ sind unbekannt. Gesucht ist eine *Regressionsgerade* der Form $y = ax + b$, die die Beobachtungen möglichst gut erklärt. Der Parameter σ ist typischerweise nicht das Ziel der statistischen Inferenz und somit ein *Störparameter*.

Beispiel 2.1. Y_i ist das Wachstum von Deutschlands Bruttoinlandsprodukt im Jahr i. Die Kovariable x_i ist die Veränderung der Arbeitslosenquote im Vergleich zum Vorjahr. Unter Verwendung der Daten von 1992 bis 2012 aus den "World Development Indicators" der Weltbank erhalten wir als Regressionsgrade $y = -1,080 \cdot x + 1,338$. Betrachten wir alle sechs Gründungsmitglieder der EU im gleichen Zeitraum, ergibt sich ganz ähnlich $y = -1,075 \cdot x + 1,819$ Der lineare Zusammenhang beider Größen ist als *Okuns Gesetz* bekannt.

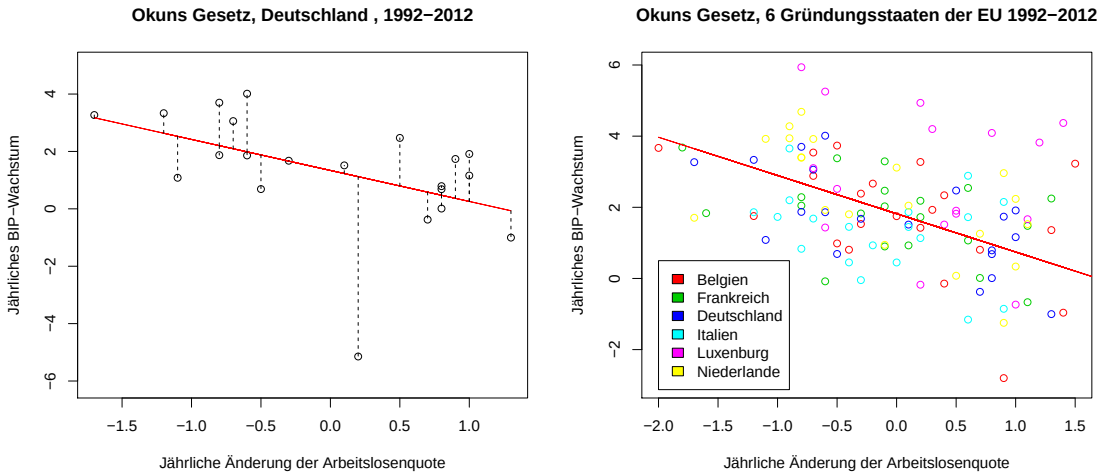

Abbildung 2: Jährliche Veränderung der Arbeitlosenquote und jährliches Wachstum des Bruttoinlandprodukts für Deutschland bzw. die 6 Gründungsstaaten der EU sowie jeweilige Regressionsgrade.

Um die Situation weiter zu vereinfachen, nehmen wir zunächst an, dass $\varepsilon_1, \dots, \varepsilon_n$ unabhängig und $\mathcal{N}(0, \sigma^2)$-verteilt sind. Nun können wir den Maximum-Likelihood-Schätzer bestimmen: Der Beobachtungsvektor ist verteilt gemäß der Lebesguedichte

$$\begin{aligned} L(a, b, \sigma; y) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\Big(- \frac{(y_i - a x_i - b)^2}{2\sigma^2} \Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a x_i - b)^2 \Big), \quad y \in \mathbb{R}^n. \end{aligned}$$

Somit ist die Loglikelihoodfunktion

$$l(a, b, \sigma; y) := \log L(a, b, \sigma; y) = -\frac{n}{2}(\log \sigma^2 + \log(2\pi)) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a x_i - b)^2.$$

Das Maximieren der Likelihood über a, b ist also äquivalent zum Minimieren der Summe der quadrierten Residuen (RSS: residual sum of squares). Wir erhalten also die **Kleinste-Quadrate-Schätzer**:

$$(\widehat{a}, \widehat{b}) := \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - a x_i - b)^2$$

Auch wenn die Fehler nicht normalverteilt sind, kann diese Methode gute Ergebnisse erzielen.

Satz 2.2. *Im einfachen linearen Modell mit unabhängigen und $\mathcal{N}(0, \sigma^2)$-verteilten Fehlern ist der Maximum-Likelihood-Schätzer gleich dem Kleinste-Quadrate-Schätzer und es gilt*

$$\widehat{a} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2} \quad \textit{und} \quad \widehat{b} = \overline{Y}_n - \widehat{a}\overline{x}_n,$$

wobei $\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ und $\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Beweis. Es bleibt festzustellen, dass wir durch Differentiation folgende Normalgleichungen erhalten:

$$0 = \sum_{i=1}^{n} x_i (Y_i - a x_i - b) \quad \text{und} \quad 0 = \sum_{i=1}^{n} (Y_i - a x_i - b),$$

die leicht gelöst werden können. □

Bemerkung 2.3. Bei der Wahl anderer Fehlerverteilungen ergibt das Maximum-Likelihood-Prinzip andere (nicht weniger sinnvolle) Schätzer (Übung ☑), die aber im Allgemeinen nicht in geschlossener Form darstellbar sind. Populäre nicht gaußsche Fehlerverteilungen sind Laplace- und Exponentialverteilungen.

Dies führt uns zur allgemeinen Definition des *linearen Modells:*

Definition 2.4. Ein lineares Modell mit n reellwertigen Beobachtungen $Y = (Y_1, \dots, Y_n)^\top$ und k-dimensionalem Parameter $\beta \in \mathbb{R}^k, k < n$, besteht aus einer reellen Matrix $X \in \mathbb{R}^{n \times k}$ von vollem Rang k, der Designmatrix, und einem Zufallsvektor $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$, den Fehler- oder Störgrößen, mit $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \Sigma_{i,j}$ für eine Kovarianzmatrix $\Sigma > 0$. Beobachtet wird eine Realisierung von

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

Der (gewichtete) Kleinste-Quadrate-Schätzer $\widehat{\beta}$ von β minimiert den gewichteten Euklidischen Abstand zwischen Beobachtungen und Modellvorhersage:

$$|\Sigma^{-1/2}(X\widehat{\beta} - Y)|^2 = \inf_{b \in \mathbb{R}^k} |\Sigma^{-1/2}(Xb - Y)|^2.$$

Im gewöhnlichen Fall $\Sigma = \sigma^2 I_n$ mit Fehlerniveau $\sigma > 0$ erhalten wir den gewöhnlichen Kleinste-Quadrate-Schätzer (OLS: ordinary least squares)

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

der unabhängig von der Kenntnis von σ^2 ist.

Bemerkung 2.5. Wir schreiben $\Sigma > 0$, falls Σ eine symmetrische, strikt positiv-definite Matrix ist. Dann ist Σ diagonalisierbar mit $\Sigma = TDT^\top, D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ Diagonalmatrix und T orthogonale Matrix, und wir setzen $\Sigma^{-1/2} := TD^{-1/2}T^\top$ mit $D^{-1/2} := \operatorname{diag}(\lambda_1^{-1/2}, \dots, \lambda_n^{-1/2})$. Wie erwartet, gilt $(\Sigma^{-1/2})^2 = \Sigma^{-1}$ und somit $|\Sigma^{-1/2}v|^2 = \langle \Sigma^{-1}v, v\rangle$ für alle $v \in \mathbb{R}^n$.

Beispiel 2.6.

(i) *Polynomiale Regression:* Wir beobachten

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

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

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

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

(ii) *Multiple Regression*: Haben wir $k \geqslant 2$ Kovariablen und n Beobachtungen Y_i, führt das zur *multiplen linearen Regression*

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i, \quad i = 1, \dots, n,$$

wobei die Fehlerterme (ε_i) iid. und zentriert sind mit $0 < \operatorname{Var}(\varepsilon_i) =: \sigma^2 < \infty$. In Vektorschreibweise erhalten wir wieder $Y = X\beta + \varepsilon$ mit Designmatrix

$$X := \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n \times (k+1)}.$$

Lemma 2.7. *Setze $X_\Sigma := \Sigma^{-1/2}X$. Mit Π_{X_Σ} werde die Orthogonalprojektion von $\mathbb{R}^n$ auf den Bildraum* $\operatorname{ran}(X_\Sigma)$ *bezeichnet. Dann gilt*

$$\Pi_{X_\Sigma} = X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top$$

und für den Kleinste-Quadrate-Schätzer

$$\widehat{\beta} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Y.$$

Insbesondere existiert der Kleinste-Quadrate-Schätzer, ist eindeutig und erwartungstreu.

Beweis. Zunächst beachte, dass $X_\Sigma^\top X_\Sigma = X^\top \Sigma^{-1} X$ invertierbar ist wegen der Invertierbarkeit von Σ und der Rangbedingung an X:

$$X^\top \Sigma^{-1} X v = 0 \Rightarrow v^\top X^\top \Sigma^{-1} X v = 0 \Rightarrow |\Sigma^{-1/2} X v| = 0 \Rightarrow |Xv| = 0 \Rightarrow v = 0.$$

Setze $P_{X_\Sigma} := X_\Sigma(X_\Sigma^\top X_\Sigma)^{-1}X_\Sigma^\top$ und $w = P_{X_\Sigma}v$ für ein $v \in \mathbb{R}^n$. Dann folgt $w \in \operatorname{ran}(X_\Sigma)$ und im Fall $v = X_\Sigma u$ durch Einsetzen $w = P_{X_\Sigma}X_\Sigma u = v$, sodass P_{X_Σ} eine Projektion auf $\operatorname{ran}(X_\Sigma)$ ist. Da P_{X_Σ} selbstadjungiert (symmetrisch) ist, handelt es sich um die Orthogonalprojektion Π_{X_Σ}:

$$\forall u \in \mathbb{R}^n, \forall w \in \operatorname{ran} X_\Sigma : \langle u - P_{X_\Sigma}u, w\rangle = \langle u, w\rangle - \langle u, P_{X_\Sigma}w\rangle = 0.$$

Aus der Eigenschaft $\widehat{\beta} = \arg\min_b |\Sigma^{-1/2}(Y - Xb)|^2$ folgt, dass $\widehat{\beta}$ die beste Approximation von $\Sigma^{-1/2}Y$ durch $X_\Sigma b$ liefert. Diese ist durch die Orthogonalprojektionseigenschaft $\Pi_{X_\Sigma}\Sigma^{-1/2}Y = X_\Sigma\widehat{\beta}$ bestimmt. Es folgt

$$X_\Sigma^\top \Pi_{X_\Sigma}\Sigma^{-1/2}Y = (X_\Sigma^\top X_\Sigma)\widehat{\beta} \Rightarrow (X_\Sigma^\top X_\Sigma)^{-1}X^\top\Sigma^{-1}Y = \widehat{\beta}.$$

Schließlich folgt aus der Linearität des Erwartungswertes und $\mathbb{E}[\varepsilon] = 0$:

$$\mathbb{E}[\widehat{\beta}] = \mathbb{E}[(X_\Sigma^\top X_\Sigma)^{-1}X^\top\Sigma^{-1}(X\beta + \varepsilon)] = \beta + 0 = \beta. \qquad \square$$

Bemerkung 2.8.

- Im gewöhnlichen linearen Modell gilt $\widehat{\beta} = (X^\top X)^{-1}X^\top Y$, unabhängig vom unbekannten Parameter $\sigma > 0$.
- $X_\Sigma^\dagger := (X_\Sigma^\top X_\Sigma)^{-1}X_\Sigma^\top$ heißt auch Moore-Penrose-(Pseudo-)Inverse von X_Σ, sodass $\widehat{\beta} = X_\Sigma^\dagger\Sigma^{-1/2}Y$ bzw. $\widehat{\beta} = X^\dagger Y$ im gewöhnlichen linearen Modell gilt.

Wir kommen zum zentralen Satz in der Regressionsanalyse:

Satz 2.9 (Gauß-Markov)**.** *Ist der Parameter $\rho = \langle\beta, v\rangle$ für ein $v \in \mathbb{R}^k$ im linearen Modell zu schätzen, so ist $\widehat{\rho} = \langle\widehat{\beta}, v\rangle$ ein (in den Daten Y) linearer erwartungstreuer Schätzer, der unter allen linearen erwartungstreuen Schätzern die minimale Varianz $\operatorname{Var}(\widehat{\rho}) = |X_\Sigma(X_\Sigma^\top X_\Sigma)^{-1}v|^2$ besitzt.*

Beweis. Die Linearität ist klar und aus dem vorangegangen Lemma folgt, dass $\widehat{\rho}$ erwartungstreu ist. Sei nun $\widetilde{\rho} = \langle Y, w\rangle$ ein beliebiger linearer erwartungstreuer Schätzer von ρ. Dies impliziert für alle $\beta \in \mathbb{R}^k$

$$\mathbb{E}[\langle Y, w\rangle] = \rho \Rightarrow \langle X\beta, w\rangle = \langle\beta, v\rangle \Rightarrow \langle X^\top w - v, \beta\rangle = 0$$

und somit $v = X^\top w = X_\Sigma^\top\Sigma^{1/2}w$. Nach Pythagoras erhalten wir

$$\begin{aligned}\operatorname{Var}(\widetilde{\rho}) &= \mathbb{E}[\langle\varepsilon, w\rangle^2] = \mathbb{E}[w^\top\varepsilon\varepsilon^\top w] \\ &= w^\top\Sigma w = |\Sigma^{1/2}w|^2 = |\Pi_{X_\Sigma}(\Sigma^{1/2}w)|^2 + |(I_n - \Pi_{X_\Sigma})(\Sigma^{1/2}w)|^2.\end{aligned}$$

Damit gilt $\operatorname{Var}(\widetilde{\rho}) \geqslant |\Pi_{X_\Sigma}(\Sigma^{1/2}w)|^2 = |X_\Sigma(X_\Sigma^\top X_\Sigma)^{-1}X^\top w|^2 = |X_\Sigma(X_\Sigma^\top X_\Sigma)^{-1}v|^2 = \operatorname{Var}(\widehat{\rho})$. $\square$

Bemerkung 2.10. Man sagt, dass der Schätzer $\widehat{\rho}$ im Satz von Gauß-Markov bester linearer erwartungstreuer Schätzer (blue: best linear unbiased estimator) ist. Eingeschränkt auf lineare Schätzer ist der Kleinste-Quadrate-Schätzer damit minimax. Ob es einen besseren nichtlinearen Schätzer geben kann, werden wir in Kapitel 3 beantworten.

Im gewöhnlichen linearen Modell ist die optimale Varianz insbesondere $\sigma^2|X(X^\top X)^{-1}v|^2$. In diesem Spezialfall ist es auch von Interesse das Rauschniveau σ^2 zu schätzen. Dies ermöglicht es insbesondere Tests und Konfidenzbereiche zu konstruieren.

Lemma 2.11. *Im gewöhnlichen linearen Modell mit $\sigma > 0$ und Kleinste-Quadrate-Schätzer $\widehat{\beta}$ gilt $X\widehat{\beta} = \Pi_X Y$ und $R := Y - X\widehat{\beta}$ bezeichne den Vektor der Residuen. Die geeignet normalisierte Stichprobenvarianz*

$$\widehat{\sigma}^2 := \frac{|R|^2}{n-k} = \frac{|Y - X\widehat{\beta}|^2}{n-k}$$

ist erwartungstreuer Schätzer von σ^2.

Beweis. $X\widehat{\beta} = \Pi_X Y$ folgt aus Lemma 2.7. Einsetzen zeigt $\mathbb{E}[|Y - X\widehat{\beta}|^2] = \mathbb{E}[|Y - \Pi_X Y|^2] = \mathbb{E}[|(I_n - \Pi_X)\varepsilon|^2]$. Ist nun $e_1, \dots, e_{n-k}$ eine Orthonormalbasis vom $(n-k)$-dimensionalen Bild $\mathrm{ran}(I_n - \Pi_X) \subseteq \mathbb{R}^n$, so folgt

$$\mathbb{E}[|(I_n - \Pi_X)\varepsilon|^2] = \sum_{i=1}^{n-k} \mathbb{E}[\langle \varepsilon, e_i\rangle^2] = \sigma^2(n-k),$$

was die Behauptung impliziert. □

Beachte, dass der Maximum-Likelihood-Schätzer von σ^2 gegeben ist durch $\widehat{\sigma}^2_{ML} = n^{-1}|R|^2 \neq \widehat{\sigma}^2$ (Übung ☑). Der erwartungstreue Schätzer $\widehat{\sigma}^2$ wird in der Praxis bevorzugt, hat jedoch größere Varianz als $\widehat{\sigma}^2_{ML}$.

2.2 Inferenz unter Normalverteilungsannahme

Im Folgenden werden wir das gewöhnliche lineare Modell unter der Normalverteilungsannahme $(\varepsilon_i) \sim \mathcal{N}(0, \sigma^2 I_n)$ betrachten.

Beispiel 2.12 (Gauß-Test)**.** Sind die Messfehler $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ gemeinsam normalverteilt und $\rho = \langle v, \beta\rangle$ für $v \in \mathbb{R}^k$, so gilt

$$\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1}) \quad \text{und} \quad \widehat{\rho} = \langle v, \widehat{\beta}\rangle \sim \mathcal{N}(\rho, \sigma^2 v^\top (X^\top X)^{-1} v).$$

Ist $\sigma > 0$ bekannt, so ist ein Konfidenzintervall zum Niveau 95% für ρ gegeben durch

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

Dabei ist der Wert $1,96$ gerade das 0,975-Qunatil der Standardnormalverteilung. Analog (Korrespondenzsatz) wird der zweiseitige *Gauß-Test* der Hypothese $H_0 : \rho = \rho_0$ gegen $H_1 : \rho \neq \rho_0$ zum Niveau $\alpha \in (0,1)$ konstruiert: Wähle die Teststatistik $|\widehat{\rho} - \rho_0|$ und den kritischen Wert $q_{1-\alpha/2}\sigma\sqrt{v^\top (X^\top X)^{-1} v}$ mit dem $(1-\alpha/2)$-Quantil von $\mathcal{N}(0,1)$.

Ist σ unbekannt, so ist eine Idee, einfach σ durch den Schätzer $\widehat{\sigma}$ in obiger Formel zu ersetzen. Allerdings wird dann das vorgegebene Niveau nur noch asymptotisch erreicht für einen konsistenten Schätzer (Slutsky-Lemma). Im vorliegenden Fall können wir aber sogar die Verteilung für endliche Stichprobenumfänge exakt bestimmen.

Definition 2.13. Die t-Verteilung $t(n)$ (oder Student-t-Verteilung) mit $n \in \mathbb{N}$ Freiheitsgraden auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ ist gegeben durch die Lebesguedichte

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}} \Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad x \in \mathbb{R}.$$

Die F-Verteilung $F(m,n)$ (oder Fisher-Verteilung) mit $(m,n) \in \mathbb{N}^2$ Freiheitsgraden auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ ist gegeben durch die Lebesguedichte

$$f_{m,n}(x) = \frac{m^{m/2} n^{n/2}}{B(\frac{m}{2}, \frac{n}{2})} \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}} \mathbb{1}_{\mathbb{R}^+}(x), \quad x \in \mathbb{R}.$$

Dabei bezeichnet $\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} \mathrm{d}t$ die Gamma-Funktion und $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ die Beta-Funktion.

Erinnerung: Für $X_1, \dots, X_m \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ ist $X := \sum_{i=1}^m X_i^2 \sim \chi^2(m)$ verteilt mit Lebesguedichte $f_X(x) = (2^{m/2}\Gamma(\frac{m}{2}))^{-1} x^{m/2-1} e^{-x/2} \mathbb{1}_{\mathbb{R}^+}(x)$.

Lemma 2.14. *Es seien $X_1, \dots, X_m, Y_1, \dots, Y_n$ unabhängige $\mathcal{N}(0,1)$-verteilte Zufallsvariablen. Dann gilt*

$$T_n := \frac{X_1}{\sqrt{\frac{1}{n}\sum_{j=1}^n Y_j^2}} \sim t(n) \quad \textit{und} \quad F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^m X_i^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2} \sim F(m,n).$$

Beweis. Es gilt $T_n^2 = F_{1,n}$, sodass mittels Dichtetransformation $f_{|T_n|}(x) = f_{F_{1,n}}(x^2)2x, x \geqslant 0$, gilt. Da T_n symmetrisch (wie $-T_n$) verteilt ist, folgt $f_{T_n} = f_{F_{1,n}}(x^2)|x|, x \in \mathbb{R}$, und Einsetzen zeigt die Behauptung für T_n, sofern $F_{1,n}$ $F(1,n)$-verteilt ist.

Um die Behauptung für $F_{m,n}$ nachzuweisen, benutze, dass $X := \sum_{i=1}^m X_i^2$ $\chi^2(m)$-verteilt und $Y := \sum_{j=1}^n Y_j^2$ $\chi^2(n)$-verteilt sind. Wegen Unabhängigkeit von X und Y gilt für $z > 0$ (setze $w = x/y$)

$$\begin{aligned}\mathbb{P}(X/Y \leqslant z) &= \int\int \mathbb{1}_{\{x/y \leqslant z\}} f_X(x) f_Y(y) \mathrm{d}x \mathrm{d}y \\ &= \int \mathbb{1}_{\{w \leqslant z\}} \Big(\int f_X(wy) f_Y(y) y \mathrm{d}y \Big) \mathrm{d}w,\end{aligned}$$

sodass sich die Dichte wie folgt ergibt (setze $w = (z+1)y$)

$$\begin{aligned}f_{X/Y}(z) =& \int f_X(zy) f_Y(y) y \mathrm{d}y \\ =& \frac{2^{-(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty (zy)^{m/2-1} y^{n/2} e^{-(zy+y)/2} \mathrm{d}y \\ =& \frac{2^{-(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty (zw/(z+1))^{m/2-1} (w/(z+1))^{n/2} e^{-w/2} (z+1)^{-1} \mathrm{d}w \\ =& \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} z^{m/2-1} (z+1)^{-(m+n)/2}, \quad z > 0.\end{aligned}$$

Dichtetransformation ergibt damit für $F_{m,n} = \frac{n}{m}\frac{X}{Y}$ die Dichte $\frac{m}{n} f_{X/Y}(\frac{m}{n} z) = f_{m,n}(z)$. □

Bemerkung 2.15. Es gilt $T_n^2 = F_{1,n}$. Für $n = 1$ ist die $t(n)$-Verteilung gerade die Cauchy-Verteilung und für $n \to \infty$ konvergiert sie schwach gegen die Standardnormalverteilung. Für jedes $n \in \mathbb{N}$ besitzt $t(n)$ nur Momente bis zur Ordnung $p < n$ (sie ist *heavy-tailed*). Ähnliches gilt für die F-Verteilung, insbesondere konvergiert die Verteilung von $mF_{m,n}$ für $n \to \infty$ gegen die $\chi^2(m)$-Verteilung.

Damit erhalten wir Konfidenzbereiche für die Schätzung von β und linearen Funktionalen im gewöhnlichen linearen Modell unter der Normalverteilungsannahme.

Satz 2.16. *Im gewöhnlichen linearen Modell unter der Normalverteilungsannahme $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ für $\sigma > 0$ gelten folgende Konfidenzaussagen für gegebenes Niveau $\alpha \in (0,1)$:*

(i) Ist $q_{t(n-k);1-\alpha/2}$ das $(1-\frac{\alpha}{2})$-Quantil der $t(n-k)$-Verteilung, so ist

$$I := \Big[\widehat{\rho} - \widehat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v} q_{t(n-k);1-\alpha/2}, \widehat{\rho} + \widehat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v} q_{t(n-k);1-\alpha/2}\Big]$$

ein Konfidenzintervall zum Konfidenzniveau $1-\alpha$ für $\rho = \langle v, \beta \rangle$.

(ii) Ist $q_{F(k,n-k);1-\alpha}$ das $(1-\alpha)$-Quantil der $F(k,n-k)$-Verteilung, so ist

$$C := \big\{\beta \in \mathbb{R}^k \big| |X(\beta - \widehat{\beta})|^2 < k\widehat{\sigma}^2 q_{F(k,n-k);1-\alpha}\big\}$$

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

Beweis. Schritt 1: Wir zeigen, dass die Zufallsgrößen $X\widehat{\beta} = XX^\dagger Y = \Pi_X Y = X\beta + \Pi_X\varepsilon$ und $\widehat{\sigma}^2 = |(I_n - \Pi_X)\varepsilon|^2/(n-k)$ unabhängig sind. Dies folgt wiederum aus der Unabhängigkeit von $\Pi_X\varepsilon$ und $(I_n - \Pi_X)\varepsilon$. Da beide Vektoren gemeinsam normalverteilt sind, folgt die Unabhängig aus der Unkorreliertheit

$$\mathbb{E}\big[(\Pi_X\varepsilon)^\top (I_n - \Pi_X)\varepsilon\big] = \mathbb{E}\big[\varepsilon^\top \underbrace{\Pi_X(I_n - \Pi_X)}_{=0}\varepsilon\big] = 0.$$

Schritt 2: Wir zeigen für jede Orthognalprojektion Π vom Rang r, dass $\sigma^{-2}\varepsilon^\top\Pi\varepsilon$ $\chi^2(r)$-verteilt ist. Nach Voraussetzung existiert eine Orthogonalmatrix P mit $\Pi = PD_rP^\top$, wobei $D_r = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$. Da P orthogonal ist und $\sigma^{-1}\varepsilon$ standardnormalverteilt, gilt $W := P^\top(\sigma^{-1}\varepsilon) \sim \mathcal{N}(0, I_n)$. Dann folgt die Behauptung aus

$$\sigma^{-2}\varepsilon^\top\Pi\varepsilon = \sigma^{-2}\varepsilon^\top\Pi^2\varepsilon = (\sigma^{-1}\Pi\varepsilon)^\top(\sigma^{-1}\Pi\varepsilon) = (PD_rW)^\top(PD_rW) = W^\top D_r W = \sum_{i=1}^{r} W_i^2.$$

Schritt 3: Wir zeigen (i). Wegen $\widehat{\rho} = \langle v, \widehat{\beta}\rangle \sim \mathcal{N}(\rho, \sigma^2 v^\top(X^\top X)^{-1}v)$ nach dem Satz von Gauß-Markov ist

$$\frac{\rho - \widehat{\rho}}{\sigma\sqrt{v^\top(X^\top X)^{-1}v}} \sim \mathcal{N}(0,1).$$

Da X vollen Rang hat, existiert ein $w \in \mathbb{R}^n$ mit $v = X^\top w$. Also gilt $\widehat{\rho} = \langle w, X\widehat{\beta}\rangle$ und aus Schritt 1 folgt die Unabhängigkeit von $\widehat{\rho}$ und $\widehat{\sigma}^2$. Da $(I_n - \Pi_X)$ eine Orthogonalprojektion auf $(\operatorname{ran} X)^\perp$ ist und den Rang $(n-k)$ hat, gilt weiterhin $\widehat{\sigma}^2 = \sigma^2 Z/(n-k)$ für eine Zufallsvariable $Z \sim \chi^2(n-k)$. Damit ist

$$\frac{\rho - \widehat{\rho}}{\sqrt{\widehat{\sigma}^2 v^\top(X^\top X)^{-1}v}} \sim t(n-k)$$

und (i) folgt aus der Wahl des kritischen Wertes.

Schritt 4: Wir zeigen (ii). Durch die Wahl des kritischen Wertes genügt es zu zeigen:

$$\frac{|X(\beta - \widehat{\beta})|^2}{k\widehat{\sigma}^2} = \frac{(n-k)}{k}\frac{\varepsilon^\top\Pi_X\varepsilon}{\varepsilon^\top(I_n - \Pi_X)\varepsilon} \sim F(k, n-k).$$

Aus Schritt 1 folgt die Unabhängigkeit von Zähler und Nenner. Aus Schritt 2 zusammen mit $\operatorname{rank}\Pi_X = k$ und $\operatorname{rank}(I_n - \Pi_X) = n - k$ folgt $\sigma^{-2}\varepsilon^\top\Pi_X\varepsilon \sim \chi^2(k)$ und $\sigma^{-2}\varepsilon^\top(I_n - \Pi_X)\varepsilon \sim \chi^2(n-k)$. Mit Lemma 2.14 ergibt sich die behauptete F-Verteilung. □

Bemerkung 2.17. Ebenso kann man ein Konfidenzintervall für die Varianz konstruieren (Übung ☑).

Zusammen mit dem Korrespondenzsatz liefert dieses Resultat:

Korollar 2.18 (t-Test und F-Test)**.** *Wir betrachten das gewöhnliche lineare Modell unter Normalverteilungsannahme $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Sei $\alpha \in (0,1)$.*

(i) Für $\rho_0 = \langle v, \beta_0\rangle$ ist ein Niveau-α-Test der Hypothese $H_0 : \rho = \rho_0$ gegen die Alternative $H_1 : \rho \neq \rho_0$ ist gegeben durch den (zweiseitigen) t-Test

$$\varphi_{\rho_0}(Y) = \mathbb{1}_{\{|T_{n-k}(Y)| > q_{t(n-k);1-\alpha/2}\}} \quad \textit{mit} \quad T_{n-k}(Y) := \frac{\rho_0 - \widehat{\rho}}{\widehat{\sigma}\sqrt{v^\top(X^\top X)^{-1}v}},$$

wobei $q_{t(n-k);1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der $t(n-k)$-Verteilung ist.

(ii) Ein Niveau-α-Test der Hypothese $H_0 : \beta = \beta_0$ gegen die Alternative $H_1 : \beta \neq \beta_0$ ist gegeben durch den F-Test

$$\varphi_{\beta_0}(Y) = \mathbb{1}_{\{F_{k,n-k}(Y) > q_{F(k,n-k);1-\alpha}\}} \quad \textit{mit} \quad F_{k,n-k}(Y) := \frac{|X(\beta_0 - \widehat{\beta})|^2}{k\widehat{\sigma}^2},$$

wobei $q_{F(k,n-k);1-\alpha}$ das $(1-\alpha)$-Quantil der $F(k, n-k)$-Verteilung ist.

Beispiel 2.19 (Klimaentwicklung)**.** Wir folgen Beispiel 12.24 von Georgii (2007) und betrachten die mittleren Augusttemperaturen von 1799 bis 2008 in Karlsruhe (Quelle: `http://www.klimadiagramme.de/Europa/special01.htm`). Für die Jahre 1854 und 1945 liegen keine Daten vor, so dass wir $n = 208$ Beobachtungen haben. Eine polynomielle Regression in der Zeit t (in Jahrhunderten beginnend bei 1799) mit Graden $d = 1, \ldots, 4$ liefert

$$\begin{aligned} p_1(t) &= 18,7 + 0,1t, \\ p_2(t) &= 20,0 - 3,5t + 1,7t^2, \\ p_3(t) &= 19,5 - 0,6t - 1,7t^2 + 1,1t^3, \\ p_4(t) &= 19,4 + 0,5t - 4,1t^2 + 2,9t^3 - 0,4t^4. \end{aligned}$$

Zunächst ist es plausibel, dass die zufälligen Schwankungen unabhängig von einander sind und als näherungsweise normalverteilt angenommen werden können (QQ-Plot). Um statistisch verwertbare Aussagen zu treffen, setzen wir noch das Niveau $\alpha = 0,05$ fest. Der Parametervektor ist $\beta = (\beta_0, \ldots, \beta_d)^\top$. Welcher Grad des Regressionspolynoms ist sinnvoll?

Frage 1: Ist der positive Trend von p_1 signifikant? $H_0 : \beta_1 \leqslant 0$ vs. $H_1 : \beta_1 > 0$. Die zugehörige t-Statistik $T = \frac{\widehat{\beta}_1}{\widehat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}} \approx 0,62$ $(v = (0,1)^\top)$liegt deutlich unter dem kritischen Wert $q_{t(n-2),1-\alpha} \approx 1,65$ (einseitiger T-Test), so dass die Hypothese nicht verworfen werden kann.

Frage 2: Genügt eine lineare Regression? Wir testen im Modell mit $d = 2$) die Hypothes $H_0 : \beta_2 = 0$. Die zugehörige t-Statistik hat den Wert $6,02$ was wesentlich größer als das Quantil $q_{t(n-3),0.975} \approx 3,18$ ist (zweiseitiger t-Test). Folglich kann die Hypothese abgelehnt werden und wir schlussfolgern, dass eine Regressionsgerade unzureichend ist.

Frage 3: Benötigen wir ein Polynom dritten Grades ? $H_0 : \beta_3 = 0$ (im Modell mit $d = 3$). Die zugehörige t-Statistik hat den Wert $2,05 > 1,97 \approx q_{t(n-4),0.975}$. Die Hypothese kann also abgelehnt werden und der kubische Anteil im Regressionspolynom ist signifikant, d.h. p_3 ist signifikant besser geeignet die Beobachtungen zu beschreiben als p_2.

Frage 4: Benötigen wir ein Polynom vierten Grades? $H_0 : \beta_4 = 0$. Die zugehörige t-Statistik hat den Wert $-0,41$ dessen Absolutbetrag kleiner als das Quantil $q_{t(n-5),0.975} \approx 1,97$ ist (zweiseitiger t-Test). Diese Nullhypothese kann also akzeptiert werden.

p_3 zeigt einen deutlichen Anstieg der Temperaturen im 19. Jahrhundert. Es sei bemerkt, dass wir hier nur eine Zeitreihe betrachtet haben und somit nicht auf einen allgemeinen Zusammenhang schließen können (Aufgabe der Klimatologen).

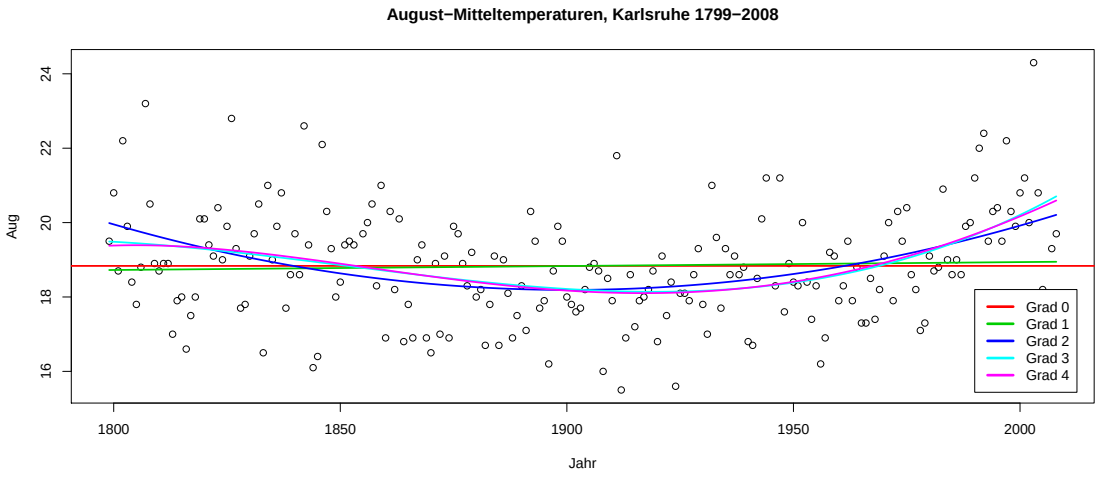

Abbildung 3: Streudiagramm der August-Mittelthemperaturen in Karlsruhe von 1799 bis 2008 mit Rregressionspolynomen vom Grad 0 bis 4.

Zum Abschluss dieses Kapitels noch ein in Anwendungen sehr wichtiger Spezialfall des linearen Modells:

Definition 2.20. Das Modell der (einfaktoriellen) Varianzanalyse (ANOVA: analysis of variance) ist gegeben durch Beobachtungen

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, k, j = 1, \dots, n_i,$$

mit iid.-verteilten Störgrößen $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Wir bezeichnen die erste Koordinate als den Faktor und den Wert $i = 1, \dots, k$ als die Faktorstufe. Folglich geben $(n_i)_{i=1,\dots,k}$ die Anzahl der unabhängigen Versuchswiederholungen pro Faktor an und $n := \sum_{i=1}^k n_i$ ist der Gesamtstichprobenumfang.

Damit ist das ANOVA-Modell ein Spezialfall des gewöhnlichen linearen Modells der Form

$$\mathbb{R}^n \ni Y := \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}}_{=:X \in \mathbb{R}^{n \times k}} \cdot \underbrace{\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}}_{=:\mu \in \mathbb{R}^k} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix}.$$

Beachte, dass $\operatorname{rank} X = k$. Die klassische Fragestellung der Varianzanalyse lautet: „Existieren Unterschiede in den faktorstufenspezifischen Mittelwerten μ_i?". Intuitiv spricht es gegen die Nullhypothese $H_0 : \mu_1 = \cdots = \mu_k$, wenn die Streuung zwischen den Gruppen größer ist als die Streuung innerhalb der Gruppen (d.h. $F(Y)$ ist groß). Wir erhalten folgendes Testverfahren:

Satz 2.21 (ANOVA)**.** *Im (einfaktoriellen) Varianzanalysemodell sei*

$$\overline{Y}_{i\bullet} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad \textit{bzw.} \quad \overline{Y}_{\bullet\bullet} := \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, \dots, k,$$

sowie

$$SSB := \sum_{i=1}^{k} n_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \quad \textit{und} \quad SSW := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2$$

(SSB: sum of squares between groups*; SSW:* sum of squares within groups *). Dann gilt*

(i) Der Kleinste-Quadrate-Schätzer von $\mu = (\mu_1, \dots, \mu_k)^\top$ ist gegeben durch $\widehat{\mu} = (\overline{Y}_{1\bullet}, \dots, \overline{Y}_{k,\bullet})^\top$.

(ii) SSW und SSB sind unabhängig und es gilt $F(Y) := \frac{n-k}{k-1} \frac{SSB}{SSW} \overset{H_0}{\sim} F(k-1, n-k)$.

(iii) Ein Test zum Niveau $\alpha \in (0,1)$ für das Testproblem

$$H_0 : \mu_1 = \cdots = \mu_k \quad \textit{versus} \quad H_1 : \exists i, l \in \{1, \dots, k\} : \mu_i \neq \mu_l$$

ist gegeben durch den F-Test $\varphi(Y) = \mathbb{1}_{\{F(Y) > q_{F(k-1,n-k);1-\alpha}\}}$, wobei $q_{F(k-1,n-k);1-\alpha}$ das $(1-\alpha)$-Quantil der $F(k-1, n-k)$-Verteilung ist.

Beweis. Übung ⧄. □

3 Exponentialfamilien und Optimalität

3.1 Exponentialfamilien

Definition 3.1. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein von μ dominiertes statistisches Modell und $k \in \mathbb{N}$. Dann heißt $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ Exponentialfamilie in $\eta(\vartheta)$ und T, wenn messbare Funktionen $\eta\colon \Theta \to \mathbb{R}^k, T\colon \mathcal{X} \to \mathbb{R}^k$ und $h\colon \mathcal{X} \to [0,\infty)$ existieren, sodass

$$\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x) = h(x)\exp\big(\langle\eta(\vartheta), T(x)\rangle - \zeta(\vartheta)\big), \quad x \in \mathcal{X}, \vartheta \in \Theta,$$

wobei $\zeta(\vartheta) := \log(\int h(x)e^{\langle\eta(\vartheta),T(x)\rangle}\mu(\mathrm{d}x))$. T wird natürliche suffiziente Statistik von $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ genannt. Sind $\eta_1, \dots, \eta_k$ linear unabhängige Funktionen und gilt für alle $\vartheta \in \Theta$ die Implikation

$$\forall\lambda \in \mathbb{R}^{k+1}: \quad \lambda_0 + \lambda_1 T_1 + \cdots + \lambda_k T_k = 0 \ \mathbb{P}_\vartheta\text{-f.s.} \quad \Rightarrow \quad \lambda_0 = \lambda_1 = \cdots = \lambda_k = 0$$

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

$$\Xi := \Big\{\eta \in \mathbb{R}^k : \int_{\mathcal{X}} h(x)e^{\langle\eta,T(x)\rangle}\mu(\mathrm{d}x) \in (0,\infty)\Big\}$$

heißt natürlicher Parameterraum. Ist die Exponentialfamilie durch $\eta \in \Xi$ parametrisiert, wird sie als natürliche Exponentialfamilie in T bezeichnet.

Bemerkung 3.2.

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

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

Beispiel 3.3.

(i) $(Bin(n,p))_{p\in(0,1)}$ bildet eine Exponentialfamilie in $\eta(p) = \log(p/(1-p))$ (auch logit-Funktion genannt) und $T(x) = x$ bzgl. dem Zählmaß μ auf $\{0,1,\dots,n\}$: Für $k = 0,\dots,n$ gilt

$$L(p,k) = \binom{n}{k}p^k(1-p)^{n-k} = \binom{n}{k}\exp\big(k\log p + (n-k)\log(1-p)\big) = \binom{n}{k}\exp(k\eta(p) + n\log(1-p)).$$

Der natürliche Parameterraum ist $\mathbb{R}$. Beachte, dass für den Parameterraum $p = [0,1]$ keine Exponentialfamilie vorliegt.

(ii) $(\mathcal{N}(\mu,\sigma^2))_{\mu\in\mathbb{R},\sigma>0}$ ist eine zweiparametrische Exponentialfamilie in $\eta(\mu,\sigma) = (\mu/\sigma^2, 1/(2\sigma^2))^\top$ und $T(x) = (x, -x^2)^\top$ unter dem Lebesguemaß als dominierendes Maß:

$$L(\eta,x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)} = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x^2-2\mu x+\mu^2)/(2\sigma^2)}, \quad x \in \mathbb{R}.$$

Der natürliche Parameterraum ist $\Xi = \mathbb{R}\times(0,\infty)$. Ist $\sigma > 0$, bekannt, so liegt eine einparametrische Exponentialfamilie in $\eta(\mu) = \mu/\sigma^2$ und $T(x) = x$ vor.

Lemma 3.4. *Bildet $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine (k-parametrische) Exponentialfamilie in $\eta(\vartheta)$ und $T(x)$, so bilden auch die Produktmaße $(\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta}$ (also die Verteilung einer entsprechenden mathematische Stichprobe) eine (k-parametrische) Exponentialfamilie in $\eta(\vartheta)$ und $\sum_{i=1}^n T(x_i)$ mit*

$$\frac{\mathrm{d}\mathbb{P}_\vartheta^{\otimes n}}{\mathrm{d}\mu^{\otimes n}}(x) = \Big(\prod_{i=1}^n h(x_i)\Big)\exp\big(\big\langle\eta(\vartheta), \sum_{i=1}^n T(x_i)\big\rangle - n\zeta(\vartheta)\big), \quad x \in \mathcal{X}^n, \vartheta \in \Theta.$$

Beweis. Folgt sofort aus der Produktformel $\frac{\mathrm{d}\mathbb{P}_\vartheta^{\otimes n}}{\mathrm{d}\mu^{\otimes n}}(x) = \prod_{i=1}^n \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x_i)$ für $x \in \mathcal{X}^n$. □

Satz 3.5. *Es sei $(\mathbb{P}_\eta)_{\eta\in\Xi}$ eine natürliche Exponentialfamilie mit $\Xi \subseteq \mathbb{R}^k$ und Darstellung*

$$\frac{\mathrm{d}\mathbb{P}_\eta}{\mathrm{d}\mu}(x) = h(x)\exp\big(\langle\eta, T(x)\rangle - \zeta(\eta)\big), \quad x \in \mathcal{X}, \eta \in \Xi,$$

wobei $\zeta(\eta) := \log(\int h(x)e^{\langle\eta,T(x)\rangle}\mu(\mathrm{d}x))$. Ist η_0 ein innerer Punkt von Ξ, so gilt für die momentenerzeugende Funktion

$$\psi_{\eta_0}(s) := \mathbb{E}_{\eta_0}[e^{\langle T,s\rangle}] = \exp(\zeta(\eta_0 + s) - \zeta(\eta_0))$$

für alle s mit $\eta_0 + s \in \Xi$. Insbesondere ist ψ_{η_0} in einer Umgebung von Null wohldefiniert und beliebig oft differenzierbar. Für $i, j = 1, \dots, k$ folgt $\mathbb{E}_{\eta_0}[T_i] = \frac{\partial\zeta}{\partial\eta_i}(\eta_0)$ und $\mathrm{Cov}(T_i, T_j) = \frac{\partial^2\zeta}{\partial\eta_i\partial\eta_j}(\eta_0)$.

Beweis. Übung ☑ □

Abschließend klären wir noch die Frage was das Maximum-Likelihood-Prinzip für natürliche Exponentialfamilien ergibt.

Lemma 3.6. *Ist $(\mathbb{P}_\eta)_{\eta\in\Xi}$ auf $(\mathcal{X}, \mathscr{F})$ eine natürliche Exponentialfamilie in $\eta \in \Xi$ und $T\colon \mathcal{X} \to \mathbb{R}$, dann ist T auf dem Ereignis $\{T(X) \in \mathrm{ran}(\zeta')\}$ der eindeutige Maximum-Likelihood-Schätzer des Parameters $\rho(\eta) := \mathbb{E}_\eta[T]$. Ferner ist $\zeta'\colon \Theta \to \mathbb{R}$ invertierbar und der eindeutige Maximum-Likelihood-Schätzer des natürlichen Parameters η ist gegeben durch*

$$\widehat{\eta} = (\zeta')^{-1}(T(X)).$$

Beweis. Um die Maximalstelle der Likelihood-Funktion zu finden, setzen wir die Ableitung der Likelihood gleich null. Auf $\{T(X) \in \mathrm{ran}(\zeta')\}$ gilt

$$\partial_\eta \log L(\eta, x) = T(x) - \zeta'(\eta) = 0 \quad \Leftrightarrow \quad T(x) = \zeta'(\eta).$$

Da $\partial_\eta^2 \log L(\eta, x) = -\zeta''(\eta) = -\mathrm{Var}_\eta(T) < 0$, ist $\eta \mapsto -\log L(\eta, x)$ konvex und somit T der eindeutige Maximum-Likelihood-Schätzer des Parameters $\rho(\eta) = \zeta'(\eta)$. Aus $\zeta'' > 0$ folgt außerdem, dass ζ' invertierbar ist, sodass der Maximum-Likelihood-Schätzer des natürlichen Parameters gegeben ist durch $(\zeta')^{-1} \circ T$. □

3.2 Suffizienz und Vollständigkeit

Beispiel 3.7. Es sei $X_1, \dots, X_n$ eine gemäß der Lebesguedichte $f_\vartheta\colon \mathbb{R} \to \mathbb{R}_+$ verteilte mathematische Stichprobe. Dann liefern die Statistiken $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ oder $\max(X_1, \dots, X_n)$ im Allgemeinen Informationen über $\mathbb{P}_\vartheta$ und damit ϑ. Hingegen sind $\mathbb{1}_{\{X_3<X_7\}}$ oder $\mathbb{1}_{\{X_1=\max\{X_1,\dots,X_n\}\}}$ Statistiken, deren Verteilung nicht von $\mathbb{P}_\vartheta$ abhängt (aufgrund der i.i.d. Annahme) und somit keinerlei Information über ϑ beinhalten (sogenannte *ancillary statistics*). Intuitiv ist alle Information bereits in der Ordungsstatistik $X_{(1)}, \dots, X_{(n)}$enthalten, wobei $X_{(1)} := \min\{X_1, \dots, X_n\}$ und $X_{(k)} = \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}, \dots, X_{(k-1)}\}, k = 2, \dots, n$. Diese sind also in einem geeigneten Sinn *suffizient.*

Um diese Idee im Allgemeinen zu formalisieren, werden bedingte Erwartungswerte (Maßtheorie) benötigt. Daher beschränken wir uns in diesem Abschnitt auf die Fälle in denen das Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ entweder vom Zählmaß oder vom Lebesguemaß dominiert wird.

Definition 3.8. Im statistischen Experiment auf $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ sei T eine $(S, \mathscr{S})$-wertige Statistik. Im stetigen Fall nehmen wir an, dass die Randverteilung $\mathbb{P}_\vartheta^T$ für jedes $\vartheta \in \Theta$ eine echt positive Lebesguedichte besitzt. Dann heißt T suffizient (für $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$), falls für jedes $\vartheta \in \Theta$ die bedingte Wahrscheinlichkeit $\mathbb{P}_\vartheta(X = \cdot|T)$ gegeben T nicht von ϑ abhängt.

Bemerkung 3.9. Existiert im Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ die Lebesguedichte $f_\vartheta^{(X,T(X))}$ des Bildmaßes $\mathbb{P}_\vartheta^{(X,T(X))}$ und bezeichne $f_\vartheta^T(t) := \int_{\mathbb{R}} f_\vartheta^{(X,T(X))}(x,t)\mathrm{d}x$ die Randdichte von T, dann kann man für alle $t \in \mathbb{R}$ mit $f_\vartheta^T(t) > 0$, die bedingte Wahrscheinlichkeit $\mathbb{P}_\vartheta(X = \cdot|T(X) = t)$ definieren als

$$\mathbb{P}_\vartheta(X \in A|T = t) = \int_A \frac{f^{(X,T(X))}(x,t)}{f_\vartheta^T(t)}\mathrm{d}x, \quad \text{für alle } A \in \mathscr{F}.$$

Anschaulich bedeutet Suffizienz also, dass man bei Kenntnis von $T(X)$ keine zusätzlichen Informationen über den Parameter ϑ durch Kenntnis von X gewinnen kann.

Satz 3.10 (Faktorisierungskriterium von Neyman)**.** *Im statistischen Experiment $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ist eine $(S, \mathscr{S})$-wertige Statistik T genau dann suffizient, wenn eine messbare Funktion $h\colon \mathcal{X} \to \mathbb{R}_+$ existiert, sodass es für alle $\vartheta \in \Theta$ eine messbare Funktion $g_\vartheta\colon S \to \mathbb{R}_+$ gibt mit*

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

Beweis. Wir führen den Beweis hier nur im diskreten Fall. Für den allgemeinen Fall sei auf Shao (2003, Theorem 2.2) verwiesen.

„⇒“ T sei suffizient. Für $x \in \mathcal{X}$ setze $t = T(x)$. Für alle x mit $L(\vartheta, x) = 0$ für alle ϑ wählen wir $h(x) = 0$ und g_ϑ beliebig. Existiert ein ϑ_0 mit $L(\vartheta_0, x) > 0$, so gilt auch $\mathbb{P}_{\vartheta_0}(T(X) = t) > 0$ und damit

$$L(\vartheta_0, x) = \mathbb{P}_{\vartheta_0}(T(X) = t) \cdot \underbrace{\mathbb{P}_{\vartheta_0}(X = x|T(X) = t)}_{=:h(x)}.$$

Beachte, dass nach Voraussetzung $h(x)$ unabhängig von ϑ ist (und als 0 definiert, falls $\mathbb{P}_\vartheta(T = t) = 0$). Wir wählen also $g_\vartheta(t) := \mathbb{P}_\vartheta(T(X) = t)$ für alle $\vartheta \in \Theta$. Ist $g_\vartheta(t) = 0$, dann ist auch $L(\vartheta, x) = \mathbb{P}_\vartheta(X = x) = 0$.

„⇐“ Es gelte nun $L(\vartheta, x) = g_\vartheta(T(x))h(x)$. Für jedes $t \in S$ mit $\mathbb{P}_\vartheta(T(X) = t) > 0$ folgt

$$\mathbb{P}_\vartheta(T(X) = t) = \sum_{x:T(x)=t} \mathbb{P}_\vartheta(X = x) = \sum_{x:T(x)=t} g_\vartheta(T(x))h(x) = g_\vartheta(t) \sum_{x:T(x)=t} h(x).$$

Damit ist die bedinge Wahrscheinlichkeit für jedes $x \in T^{-1}(\{t\})$

$$\mathbb{P}_\vartheta(X = x|T(X) = t) = \frac{P_\vartheta(X = x)}{\mathbb{P}_\vartheta(T(X) = t)} = \frac{g_\vartheta(t)h(x)}{g_\vartheta(t)\sum_{x:T(x)=t} h(x)} = \frac{h(x)}{\sum_{x:T(x)=t} h(x)}$$

unabhängig von ϑ, also T suffizient. □

Bemerkung 3.11. Ist T eine suffiziente Statistik für $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$, so folgt aus der Darstellung der Likelihoodfunktion, dass ein Maximum-Likelihood-Schätzer existiert, der eine Funktion von T ist.

Beispiel 3.12.

(i) Die Identität $T(x) = x$ und allgemeiner jede bijektive, bi-messbare Transformation T ist stets suffizient.

(ii) Die natürliche suffiziente Statistik T einer Exponentialfamilie ist tatsächlich suffizient. Im Normalverteilungsmodell $(\mathcal{N}(\mu, \sigma^2)^{\otimes n})_{\mu\in\mathbb{R},\sigma>0}$ ist damit

$$T_1(x) = \Big(\sum_{i=1}^n x_i, -\sum_{i=1}^n x_i^2\Big)^\top$$

suffizient, aber auch durch Transformation $T_2(x) = (\overline{x}, \overline{x^2})$ oder $T_3(\overline{x}, \overline{s}^2)$ mit der empirischen Varianz $\overline{s}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$. Bei einer Bernoullikette $(Bin(1, p)^{\otimes n})_{p\in(0,1)}$ ist die Anzahl der Erfolge $T(x) = \sum_{i=1}^n x_i$ suffizient.

(iii) Ist $X_1, \dots, X_n$ eine mathematische Stichprobe, wobei X_i gemäß einer Lebesguedichte $f_\vartheta \colon \mathbb{R} \to \mathbb{R}_+$ verteilt ist, so ist die Ordnungsstatistik $(X_{(1)}, \dots, X_{(n)})$ suffizient, denn

$$L(\vartheta, x) = \prod_{i=1}^{n} f_\vartheta(x_{(i)}).$$

Bedingen wir einen Schätzer auf eine suffiziente Statistik, so hat der resultierende Schätzer gleichmäßig über alle $\vartheta \in \Theta$ ein kleineres (ggf. gleich großes) Risiko:

Satz 3.13 (Rao-Blackwell)**.** *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\rho \colon \Theta \to \mathbb{R}^d$ ein Parameter und $\ell \colon \Theta \times \mathbb{R}^d \to \mathbb{R}_+$ eine Verlustfunktion, die im zweiten Argument konvex ist. Ist T eine für $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ suffiziente Statistik, so gilt für jeden Schätzer $\widehat{\rho}$ und für $\widetilde{\rho} := \mathbb{E}_\cdot[\widehat{\rho}(X)|T]$ die Risikoabschätzung*

$$\forall \vartheta \in \Theta : \quad R(\vartheta, \widetilde{\rho}) \leqslant R(\vartheta, \widehat{\rho}).$$

Beweis. Folgt aus Jensens Ungleichung (für bedingte Erwartungen). Hier wieder nur der diskrete Fall:

$$\begin{aligned}
\mathbb{E}_\vartheta\big[\ell(\vartheta, \widetilde{\rho})\big] &= \sum_{t\in S} \ell\big(\vartheta, \mathbb{E}[\widehat{\rho}(X)|T = t]\big)\mathbb{P}_\vartheta(T = t) \\
&\leqslant \sum_{t\in S} \mathbb{E}\big[\ell(\vartheta, \widehat{\rho}(X))|T = t\big]\mathbb{P}_\vartheta(T = t) \\
&= \sum_{x\in\mathcal{X}} \ell(\vartheta, \widehat{\rho}(x))\mathbb{P}_\vartheta(X = x) = R(\vartheta, \widehat{\rho}).
\end{aligned}$$

□

Auch für Testprobleme sind suffiziente Statistiken nützlich:

Satz 3.14. *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und T eine suffiziente Statistik. Dann gibt es zu jedem randomisierten Test φ einen randomisierten Test $\widetilde{\varphi}$, der nur von T abhängt und dieselben Fehlerwahrscheinlichkeiten erster und zweiter Art besitzt, nämlich $\widetilde{\varphi} = \mathbb{E}_\cdot[\varphi|T]$.*

Beweis. Folgt jeweils aus $\mathbb{E}_\vartheta[\widetilde{\varphi}] = \mathbb{E}_\vartheta[\mathbb{E}_\cdot[\varphi|T]] = \mathbb{E}_\vartheta[\varphi]$. □

Der Satz von Gauß-Markov hat uns bereits ein Optimalitätsresultat geliefert, das allerdings auf lineare Schätzer im linearen Modell eingeschränkt ist. Wir suchen nun allgemeiner nach unverzerrten Schätzern, deren Schätzwerte möglichst wenig um den korrekten Wert streuen.

Definition 3.15. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Ein erwartungstreuer Schätzer T eines abgeleiteten Parameters $\rho(\vartheta)$ heißt varianzminimierend bzw. (gleichmäßig) bester Schätzer (UMVU: uniformly minimum variance unbiased estimator), wenn für jeden weiteren erwartungstreuen Schätzer S gilt:

$$\mathrm{Var}_\vartheta(T) \leqslant \mathrm{Var}_\vartheta(S) \quad \text{für alle } \vartheta \in \Theta.$$

Nach dem Satz von Rao-Blackwell (angewendet auf den quadratischen Verlust) sind suffiziente Statistiken gute Kandidaten, beste Schätzer zu sein. Suffizienz allein reicht allerdings nicht aus.

Definition 3.16. Eine $(S, \mathscr{S})$-wertige Statistik T im Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt vollständig, falls für alle messbaren Funktionen $f \colon S \to \mathbb{R}$ gilt:

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

Bemerkung 3.17. Wie oben erwähnt, heißt eine Statistik V *ancillary*, wenn ihre Verteilung nicht von ϑ abhängt. Sie heißt *ancillary erster Ordung*, falls $\mathbb{E}_\vartheta[V]$ unabhängig von ϑ ist. Falls jede Statistik der Form $V = f(T)$, die ancillary erster Ordnung ist, $\mathbb{P}_\vartheta$-f.s. konstant ist, so ist keine redundante Information mehr in T enthalten und T ist vollständig (verwende $\widetilde{f}(T) := f(T) - \mathbb{E}_\cdot[f(T)]$).

Satz 3.18 (Lehmann-Scheffé). *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\rho(\vartheta), \vartheta \in \Theta$, der interessierende Parameter. Es existiere ein erwartungstreuer Schätzer $\widehat{\rho}$ von $\rho(\vartheta)$ mit endlicher Varianz. Ist T eine suffiziente und vollständige Statistik, so ist $\widetilde{\rho} = \mathbb{E}_\cdot[\widehat{\rho}|T]$ der f.s. eindeutige varianzminimierende, erwartungstreue Schätzer.*

Beweis. $\widetilde{\rho}$ ist offensichtlich wieder erwartungstreu. Zudem ist $\widetilde{\rho}$ der f.s. einzige erwartungstreue Schätzer, der $\sigma(T)$-messbar ist (nach dem Faktorisierungslemma aus der Maßtheorie ist die $\sigma(T)$-messbarkeit äquivalent dazu, dass eine Funktion g existiert mit $\widetilde{\rho} = g(T)$): Für jeden anderen derartigen Schätzer $\overline{\rho}$ folgt wegen der Vollständigkeit aus $\mathbb{E}[\widetilde{\rho} - \overline{\rho}] = 0$, dass $\widetilde{\rho} - \overline{\rho} = 0$ $\mathbb{P}_\vartheta$-f.s. gilt. Nach dem Satz von Rao-Blackwell besitzt $\widetilde{\rho}$ also kleineres quadratisches Risiko als jeder andere erwartungstreue Schätzer. Aus der Bias-Varianz-Zerlegung folgt nun, dass das quadratische Risiko gleich der Varianz ist. □

Bemerkung 3.19. Die Aussage von Lehmann-Scheffé gilt sogar für das Risiko bei beliebigen im zweiten Argument konvexen Verlustfunktionen, wie sofort aus dem Satz von Rao-Blackwell folgt.

Beispiel 3.20.

(i) Es sei $X_1, \dots, X_n \sim \mathcal{U}([0, \vartheta])$ eine mathematische Stichprobe mit unbekanntem Parameter ϑ. Aus der Form der Likelihoodfunktion

$$L(\vartheta, x) = \prod_{i=1}^{n} \frac{\mathbb{1}_{[0,\vartheta]}(x_i)}{\vartheta} = \vartheta^{-n}\mathbb{1}_{\{x_{(n)}\leqslant\vartheta\}}, \quad x \in \mathbb{R}_+^n,$$

folgt, dass $X_{(n)}$ eine suffiziente Statistik ist. Wegen $\mathbb{P}_\vartheta(X_{(n)} \leqslant r) = (\frac{r}{\vartheta})^n \mathbb{1}_{\{r\leqslant\vartheta\}} + \mathbb{1}_{\{r>\vartheta\}}, r \in \mathbb{R}_+$, besitzt $\mathbb{P}^{X_{(n)}}$ die Dichte $f^{X_{(n)}}(t) = n\vartheta^{-n}t^{n-1}\mathbb{1}_{[0,\vartheta]}(t)$. Gilt also für eine messbare Funktion f

$$\mathbb{E}_\vartheta[f(X_{(n)})] = \int_0^\vartheta f(t)n\vartheta^{-n}t^{n-1}\mathrm{d}t = 0, \quad \forall\vartheta > 0,$$

muss $f = 0$ Lebesgue-fast überall gelten. Damit ist $X_{(n)}$ auch vollständig. Andererseits gilt $\mathbb{E}[X_{(n)}] = \frac{n}{n+1}\vartheta$. Damit ist $\widehat{\vartheta} = \frac{n+1}{n}X_{(n)}$ ein erwartungstreuer Schätzer von ϑ mit gleichmäßig kleinster Varianz.

(ii) Betrachten wir die nichtparametrische Schätzung der Verteilungsfunktion F einer mathematischen Stichprobe $X_1, \dots, X_n \in \mathbb{R}$. Wir haben bereits in Beispiel 3.12 gesehen, dass die Ordnungsstatistiken $X_{(1)}, \dots, X_{(n)}$ suffizient sind und es ist leicht zu sehen, dass diese auch vollständig sind. Für jedes $x \in \mathbb{R}$ ist $\widehat{\rho}(X_1, \dots, X_n) := \mathbb{1}_{\{X_1\leqslant x\}}$ ein erwartungstreuer Schätzer des Parameters $\rho(F) := F(x)$. Dann folgt aus dem Satz von Lehmann-Scheffé, dass

$$\widetilde{\rho} = \mathbb{E}[\widehat{\rho}|X_{(1)}, \dots, X_{(n)}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_{(i)}\leqslant x\}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\leqslant x\}} =: \widehat{F}_n(x)$$

(die empirische Verteilungsfunktion) ein UMVU-Schätzer ist.

Satz 3.21. *Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ eine k-parametrische (vom Lebesguemaß dominierte) Exponentialfamilie in T mit natürlichem Parameter $\vartheta \in \Theta \subseteq \mathbb{R}^k$. Besitzt Θ ein nichtleeres Inneres, so ist T suffizient und vollständig.*

Beweis. Es bleibt die Vollständigkeit zu beweisen. Ohne Einschränkung sei $[-a, a]^k \subseteq \Theta$ für ein $a > 0$ (sonst verschiebe entsprechend) sowie $h(x) = 1$ (sonst betrachte $\widetilde{\mu}(\mathrm{d}x) := h(x)\mu(\mathrm{d}x)$). Es gelte $\mathbb{E}_\vartheta[f(T)] = 0$ für alle $\vartheta \in \Theta$ und ein $f\colon \mathbb{R}^k \to \mathbb{R}$. Mit $f_+(x) := \max\{f(x), 0\}$ und $f_-(x) = \max\{-f(x), 0\}$ folgt

$$\begin{aligned}\mathbb{E}_\vartheta[f_\pm(T)] &= \int_{\mathcal{X}} f_\pm(T(x))\exp\big(\langle\vartheta, T(x)\rangle - \zeta(\vartheta)\big)\widetilde{\mu}(\mathrm{d}x)\\ &= e^{\zeta(0)-\zeta(\vartheta)}\int_{\mathcal{X}} f_\pm(T(x))\exp\big(\langle\vartheta, T(x)\rangle\big)e^{-\zeta(0)}\widetilde{\mu}(\mathrm{d}x) = e^{\zeta(0)-\zeta(\vartheta)}\mathbb{E}_0\big[f_\pm(T)e^{\langle\vartheta,T\rangle}\big].\end{aligned}$$

Wegen $f = f_+ - f_-$ gilt damit $\mathbb{E}_0\big[f_+(T)e^{\langle\vartheta,T\rangle}\big] = \mathbb{E}_0\big[f_-(T)e^{\langle\vartheta,T\rangle}\big]$ für alle $\vartheta \in [-a,a]^k$. Wir betrachten nun die Wahrscheinlichkeitsmaße

$$\mathbb{P}_\pm(\mathrm{d}t) := \frac{1}{C} f_\pm(t)\mathbb{P}_0^T(\mathrm{d}t) \quad \text{mit} \quad C := \mathbb{E}_0\big[f_+(T)\big] = \mathbb{E}_0\big[f_-(T)\big].$$

Wir haben bereits gezeigt, dass die Laplace-Transformierten $\chi_\pm(\vartheta) := \int e^{\langle\vartheta,t\rangle}\mathrm{d}\mathbb{P}_\pm(\mathrm{d}t)$ für $\vartheta \in [-a,a]^k$ übereinstimmen. Diese charakterisieren aber die Verteilungen $\mathbb{P}_+$ bzw. $\mathbb{P}_-$ eindeutig (analog zur charakteristischen Funktion, vgl. Klenke (2006, Satz 15.6), sodass $\mathbb{P}_+ = \mathbb{P}_-$. Dies liefert $f_+ = f_-$ $\mathbb{P}_0^T$-f.ü. und somit $f = 0$ $\mathbb{P}_\vartheta^T$-f.s. für alle $\vartheta \in \Theta$ aufgrund der Vollständigkeit von T.

[Genauer: χ_+ und χ_- sind darüber hinaus auf dem k-dimensionalen komplexen Streifen $\{\vartheta \in \mathbb{C}^k : |\operatorname{Re}(\vartheta)| < a\}$ wohldefiniert und analytisch. Der Eindeutigkeitssatz für analytische Funktionen impliziert daher $\chi_+(iu) = \chi_-(iu)$ für $u \in \mathbb{R}^k$. Also besitzen $\mathbb{P}_+$ und $\mathbb{P}_-$ dieselben charakteristischen Funktionen.] □

Beispiel 3.22. Das lineare Modell $Y = X\beta + \sigma\varepsilon$ mit Gaußschen Fehlern $\varepsilon \sim \mathcal{N}(0, I_n)$ bildet eine $(k+1)$-parametrische Exponentialfamilie in $\eta(\beta,\sigma) = \sigma^{-2}(\beta,-1/2)^\top \in \mathbb{R}^k \times \mathbb{R}_-$ und $T(Y) = (X^\top Y, |Y|^2)^\top \in \mathbb{R}^k \times \mathbb{R}_+$. Der natürliche Parameterraum ist $\Xi = \mathbb{R}^k \times \mathbb{R}_-$ besitzt nichtleeres Inneres in $\mathbb{R}^{k+1}$, sodass T suffizient und vollständig ist. Durch bijektive Transformation ergibt sich, dass dies auch für

$$\big((X^\top X)^{-1}X^\top Y, |Y|^2\big) = \big(\widehat{\beta}, |\Pi_X Y|^2 + (n-k)\widehat{\sigma}^2\big)$$

mit Kleinste-Quadrate-Schätzern $\widehat{\beta}$ und $\widehat{\sigma}^2 = \frac{|Y - X\widehat{\beta}|}{n-k}$ gilt. Wegen $\Pi_X Y = X\widehat{\beta}$ ist also auch $(\widehat{\beta}, \widehat{\sigma}^2)$ suffizient und vollständig. Damit besitzen beide Schätzer jeweils minimale Varianz in der Klasse aller (!) erwartungstreuen Schätzer von β bzw. σ^2. Beachte, dass hierfür die Normalverteilungsannahme essentiell ist, während der Satz von Gauß-Markov keine Verteilungsannahme benötigt.

3.3 Cramér-Rao-Effizienz

Statt wie im Satz von Lehmann-Scheffé direkt die Optimalität eines Schätzers zu zeigen, werden wir nun zunächst eine untere Schranke für die Varianz beweisen und anschließend untersuchen, für welche Schätzer diese erreicht wird.

Definition 3.23. Ein vom Maß μ dominiertes, statistisches Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt regulär, wenn die folgenden Eigenschaften erfüllt sind:

(i) Θ ist eine offene Menge in $\mathbb{R}^d, d \geqslant 1$.

(ii) Die Likelihood-Funktion $L(\vartheta, x)$ ist auf $\Theta \times \mathcal{X}$ strikt positiv und nach ϑ stetig differenzierbar. Bezeichnen wir den Gradienten in ϑ mit $\nabla_\vartheta = (\frac{\partial}{\partial\vartheta_1}, \dots, \frac{\partial}{\partial\vartheta_d})^\top$, existiert insbesondere die Scorefunktion

$$U_\vartheta(x) := \nabla_\vartheta \log L(\vartheta, x) = \frac{\nabla_\vartheta L(\vartheta, x)}{L(\vartheta, x)}.$$

(iii) Für jedes $\vartheta \in \Theta$ existiert die Fisher-Information

$$I(\vartheta) := \mathbb{E}_\vartheta\Big[U_\vartheta(X)U_\vartheta(X)^\top\Big]$$

und ist positiv definit.

(iv) Es gilt die Vertauschungsrelation

$$\int h(x)\nabla_\vartheta L(\vartheta, x)\mu(\mathrm{d}x) = \nabla_\vartheta \int h(x)L(\vartheta, x)\mu(\mathrm{d}x) \tag{3.1}$$

für $h(x) = 1$.

Ein Schätzer $T\colon \mathcal{X} \to \mathbb{R}^d$ heißt regulär, falls $\mathbb{E}[|T(X)|^2] < \infty$ und (3.1) auch für $h(x) = T(x)$ gilt.

Bemerkung 3.24.

(i) Der Satz von Lebesgue (dominierte Konvergenz, vgl. Maßtheorie-Vorlesung) liefert eine hinreichende Bedingung für die Vertauschungsrelation (3.1): Sie gilt, falls für jedes $\vartheta_0 \in \Theta$ eine Umgebung $V_{\vartheta_0} \subseteq \Theta$ existiert, so dass

$$\int_{\mathcal{X}} \sup_{\vartheta \in V_{\vartheta_0}} \Big|\nabla_\vartheta L(\vartheta, x)\Big| \mu(\mathrm{d}x) < \infty.$$

Außerdem kann man (3.1) für jedes gegebene Modell (und jeden Schätzer) explizit nachprüfen.

(ii) Als Konsequenz von (3.1) ergibt sich

$$\mathbb{E}_\vartheta[U_\vartheta] = \int \nabla_\vartheta L(\vartheta, x)\mu(\mathrm{d}x) = \nabla_\vartheta \int L(\vartheta, x)\mu(\mathrm{d}x) = \nabla_\vartheta 1 = 0$$

und damit $\mathrm{Var}_\vartheta(U_\vartheta) = I(\vartheta)$.

(iii) Ist $L(\vartheta, x)$ in ϑ zweimal stetig differenzierbar und gilt (3.1) mit $h(x) = 1$ und L ersetzt mit $\frac{\partial L}{\partial \vartheta_i}$ für alle $i \in \{1, \dots, d\}$, dann gilt $I(\vartheta) = -\mathbb{E}_\vartheta[H_\vartheta(X)]$ für die Hesse-Matrix $H_\vartheta(x)$ der Loglikelihoodfunktion $\vartheta \mapsto \log L(\vartheta, x)$ (Übung ☑).

(iv) Warum heißt $I(\vartheta)$ Information? Erstens: $I(\vartheta) = 0$ gilt auf einer Umgebung $\Theta_0 \subseteq \Theta$ genau dann, wenn $U_\vartheta(x) = 0$ für alle $\vartheta \in \Theta_0$ und μ-f.a. $x \in \mathcal{X}$, also wenn $L(\vartheta, x)$ für μ-f.a. x konstant in ϑ ist und somit keine Beobachtung die Parameter in Θ_0 unterscheiden kann (dieser Fall ist daher in der Definition ausgeschlossen). Zweitens, verhält sich die Fisher-Information bei unabhängigen Beobachtungen additiv: Ist $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein reguläres Modell mit Fisher-Information $I = I_1$, so hat das Produktmodell $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ die Fisher-Information $I_n = nI_1$ (Beweis als Übung ☑).

Satz 3.25 (Cramér-Rao-Ungleichung, Informationsschranke)**.** *Gegeben seien ein reguläres statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, eine zu schätzende stetig differenzierbare Funktion $\rho\colon \Theta \to \mathbb{R}$ und ein regulärer erwartungstreuer Schätzer T von ρ. Dann gilt*

$$\mathrm{Var}_\vartheta(T) \geqslant \Big(\nabla\rho(\vartheta)\Big)^\top I(\vartheta)^{-1}\nabla\rho(\vartheta) \quad \textit{für alle } \vartheta \in \Theta. \tag{3.2}$$

Beweis. Aus der Zentriertheit von U_ϑ und der Regularität und Erwartungstreue von T erhalten wir

$$\begin{aligned}\mathrm{Cov}_\vartheta(U_\vartheta, T) = \mathbb{E}_\vartheta[TU_\vartheta] &= \int_{\mathcal{X}} T(x)\nabla_\vartheta L(\vartheta, x)\mu(\mathrm{d}x) \\ &= \nabla \int_{\mathcal{X}} T(x)L(\vartheta, x)\mu(\mathrm{d}x) = \nabla\mathbb{E}_\vartheta[T] = \nabla\rho\end{aligned}$$

für alle $\vartheta \in \Theta$. Für jeden Vektor $e \in \mathbb{R}^d$ ergibt die Cauchy-Schwarz-Ungleichung somit

$$(\langle e, \nabla\rho\rangle)^2 = \mathrm{Cov}_\vartheta(\langle e, U_\vartheta\rangle, T)^2 \leqslant \mathrm{Var}_\vartheta(\langle e, U_\vartheta\rangle)\,\mathrm{Var}_\vartheta(T) = \langle I(\vartheta)e, e\rangle\,\mathrm{Var}_\vartheta(T),$$

also

$$\mathrm{Var}_\vartheta(T) \geqslant \frac{(\langle\nabla\rho, e\rangle)^2}{\langle I(\vartheta)e, e\rangle}.$$

Maximieren über $e \in \mathbb{R}^d$ ergibt mit $e = I(\vartheta)^{-1}\nabla\rho(\vartheta)$ die Behauptung. □

Definition 3.26. Ein regulärer erwartungstreuer Schätzer für den Gleichheit in (3.2) gilt, heißt Cramér-Rao-effizient oder kurz effizient.

Bemerkung 3.27. In einem regulären Modell sind erwartungstreue, effiziente Schätzer also UMVU-Schätzer, wenn alle erwartungstreuen Schätzer regulär sind. Nicht jeder UMVU-Schätzer ist effizient.

Beispiel 3.28.

(i) Betrachte $X_1, \dots, X_n \overset{i.i.d.}{\sim} Bin(1,p)$ mit Parameter $p \in (0,1)$. Dann gilt

$$\begin{aligned} I_1(p) &= \mathbb{E}_p\Big[\Big(\frac{\partial}{\partial p}\log\big(p^{X_1}(1-p)^{1-X_1}\big)\Big)^2\Big] \\ &= \mathbb{E}_p\Big[\Big(\frac{\partial}{\partial p}\big(X_1\log p + (1-X_1)\log(1-p)\big)\Big)^2\Big] \\ &= \mathbb{E}_p\Big[\Big(\frac{X_1}{p} - \frac{1-X_1}{1-p}\Big)^2\Big] \\ &= \frac{\mathbb{E}_p[(X_1-p)^2]}{p^2(1-p)^2} = \frac{\mathrm{Var}_p(X_1)}{p^2(1-p)^2} = \frac{1}{p(1-p)}. \end{aligned}$$

Also $I_n(p) = nI_1(p) = \frac{n}{p(1-p)}$. Für alle regulären erwartungstreuen Schätzer $\widehat{p}$ von p gilt somit $\mathrm{Var}_p(\widehat{p}) \geqslant \frac{p(1-p)}{n}$. Dieser untere Schranke wird vom Stichprobenmittel $\overline{X}_n$ erreicht. Da auch $\mathbb{E}_p[\overline{X}_n] = p$ gilt, ist $\overline{X}_n$ ein effizienter Schätzer von p.

(ii) Betrachte $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ mit bekanntem σ^2 und unbekanntem Parameter $\mu \in \mathbb{R}$. Dann gilt

$$\begin{aligned} I_1(p) &= \mathbb{E}_\mu\Big[\Big(\frac{\partial}{\partial \mu}\log\big((2\pi\sigma^2)^{-1/2}e^{-(X_1-\mu)^2/(2\sigma^2)}\big)\Big)^2\Big] \\ &= \mathbb{E}_\mu\Big[\Big(\frac{\partial}{\partial \mu}\big(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X_1-\mu)^2}{2\sigma^2}\big)\Big)^2\Big] \\ &= \mathbb{E}_\mu\Big[\Big(\frac{2(X_1-\mu)}{2\sigma^2}\Big)^2\Big] = \frac{\mathrm{Var}(X_1)}{\sigma^4} = \frac{1}{\sigma^2}. \end{aligned}$$

Wir erhalten die Informationsschranke $\mathrm{Var}_\mu(\widehat{\mu}) \geqslant \frac{1}{nI_1(\mu)} = \frac{\sigma^2}{n}$, die wieder vom Stichprobenmittel $\overline{X}_n$ erreicht wird.

Tatsächlich wird die Cramér-Rao-Schranke nur in wenigen Modellen erreicht. Im Folgenden beschränken wir uns auf einparametrische ($d = 1$) Modelle.

Satz 3.29. *Unter den Bedingungen von Satz 3.25 mit $\Theta \subseteq \mathbb{R}$ erreicht der Schätzer T die untere Schranke für alle $\vartheta \in \Theta$ genau dann, wenn μ-f.ü. gilt:*

$$T - \rho(\vartheta) = \rho'(\vartheta)I(\vartheta)^{-1}U_\vartheta \quad \textit{für alle } \vartheta \in \Theta.$$

Falls $\rho' \neq 0$ ist dies äquivalent dazu, dass $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ eine einparametrische Exponentialfamilie in $\eta(\vartheta)$ und T ist, wobei $\eta\colon \Theta \to \mathbb{R}$ eine Stammfunktion von I/ρ' ist.

Beweis. Definieren wir $v(\vartheta) := \rho'(\vartheta)I^{-1}(\vartheta)$ (konstant in x) erhalten wir wegen $\mathrm{Cov}_\vartheta(U_\vartheta, T) = \rho'(\vartheta)$

$$\begin{aligned} 0 &\leqslant \mathrm{Var}_\vartheta\big(T - v(\vartheta)U_\vartheta\big) \\ &= \mathrm{Var}_\vartheta(T) + v(\vartheta)^2\,\mathrm{Var}_\vartheta(U_\vartheta) - 2v(\vartheta)\,\mathrm{Cov}_\vartheta(U_\vartheta, T) = \mathrm{Var}_\vartheta(T) - \rho'(\vartheta)^2 I^{-1}(\vartheta), \end{aligned}$$

also wieder die Informationsungleichung. Gleichheit gilt genau dann, wenn $T - v(\vartheta)U_\vartheta$ $\mathbb{P}_\vartheta$-f.s. konstant, also gleich seinem Erwartungswert $\rho(\vartheta)$ ist. Da $\mathbb{P}_\vartheta$ eine strikt positive μ-Dichte hat, gilt $\mu(T - \rho(\vartheta) \neq v(\vartheta)U_\vartheta) = 0$. Wenn dies nun für alle $\vartheta \in \Theta$ gilt, so folgt sogar

$$\mu\big(T - \rho(\vartheta) \neq v(\vartheta)U_\vartheta \text{ für ein } \vartheta \in \Theta\big) = 0,$$

denn aus Stetigkeitsgründen kann man sich auf rationale ϑ beschränken und die abzählbare Vereinigung von Nullmengen ist wieder eine Nullmenge. Die explizite Form der Likelihoodfunktion folgt durch unbestimmte Integration bzgl. ϑ. □

Lemma 3.30. *Ist Θ offen und ein statistisches Modell durch eine Exponentialfamilie in $\eta\colon \Theta \to \mathbb{R}$ und $T\colon \mathcal{X} \to \mathbb{R}$ mit differenzierbarem η und $\eta'(\vartheta) \neq 0$, für alle $\vartheta \in \Theta$, gegeben, so ist dieses regulär. Ferner gilt*

(i) Jede Statistik $S\colon \mathcal{X} \to \mathbb{R}$ mit existierendem Erwartungswert ist regulär. $\rho(\vartheta) := \mathbb{E}_\vartheta[T]$ ist stetig differenzierbar mit $\rho'(\vartheta) = \eta'(\vartheta)\operatorname{Var}_\vartheta(T) \neq 0, \vartheta \in \Theta$.

(ii) Die Normierungsfunktion ζ ist auf $\Theta \subseteq \mathbb{R}$ stetig differenzierbar mit $\zeta'(\vartheta) = \eta'(\vartheta)\mathbb{E}_\vartheta[T]$ für $\vartheta \in \Theta$. Die Scorefunktion ist $U_\vartheta = \eta'(\vartheta)T - \zeta'(\vartheta)$.

(iii) Für die Fisher-Information gilt $I(\vartheta) = \eta'(\vartheta)\rho'(\vartheta)$ für alle $\vartheta \in \Theta$.

Beweis. O.B.d.A. ist $\eta(\vartheta) = \vartheta$ und somit $\eta' = 1$ für alle $\vartheta \in \Theta$. Der allgemeine Fall ergibt sich durch Reparametrisierung und Anwendung der Kettenregel.

Schritt 1: Sei S eine beliebige reelle Statistik mit $S \in \mathcal{L}^1(\mathbb{P}_\vartheta)$ für alle $\vartheta \in \Theta$. Dann ist die Funktion

$$u_S(\vartheta) := e^{\zeta(\vartheta)}\mathbb{E}_\vartheta[S] = \int_{\mathcal{X}} S(x)e^{\vartheta T(x)}h(x)\mu(\mathrm{d}x)$$

auf Θ wohl definiert. Wir zeigen nun, dass u_s beliebig oft differenzierbar ist.

Ist $\vartheta \in \Theta$ und $t \in \mathbb{R}$ so klein, dass auch $\vartheta \pm t \in \Theta$, so gilt mittels monotoner Konvergenz

$$\begin{aligned}\sum_{k\geqslant 0}\frac{|t|^k}{k!}\int_{\mathcal{X}}|S(x)||T(x)|^k e^{\vartheta T(x)}h(x)\mu(\mathrm{d}x) &= \int_{\mathcal{X}}|S(x)|e^{\vartheta T(x)+|tT(x)|}h(x)\mathrm{d}x\\ &\leqslant \int_{\mathcal{X}}|S(x)|\big(e^{(\vartheta+t)T(x)}+e^{(\vartheta-t)T(x)}\big)h(x)\mathrm{d}x<\infty.\end{aligned}$$

Also ist $ST^k \in \mathcal{L}^1(\mathbb{P}_\vartheta)$ für alle $\vartheta \in \Theta$ und insbesondere $T \in \mathcal{L}^2(\mathbb{P}_\vartheta)$ für alle ϑ. Ferner ist die Reihe

$$\sum_{k\geqslant 0}\frac{t^k}{k!}\int_{\mathcal{X}} S(x)T(x)^k e^{\vartheta T(x)}h(x)\mu(\mathrm{d}x)$$

absolut konvergent und Summation und Integration können vertauscht werden. Die Reihe nimmt daher den Wert $u_S(\vartheta + t)$ an. Damit ist u_S sogar analytisch.

Schritt 2: Es folgt $u_S'(\vartheta) = e^{\zeta(\vartheta)}\mathbb{E}_\vartheta[ST]$ und insbesondere $u_1'(\vartheta) = u_1(\vartheta)\mathbb{E}_\vartheta[T]$ sowie $u_1''(\vartheta) = u_1(\vartheta)\mathbb{E}_\vartheta[T^2]$. Für $\zeta(\vartheta) = \log u_1(\vartheta)$ bekommen wir also $\zeta'(\vartheta) = \mathbb{E}_\vartheta[T] =: \rho(\vartheta)$ und

$$\rho'(\vartheta) = \zeta''(\vartheta) = u_1''(\vartheta)/u_1(\vartheta) - (u_1'(\vartheta)/u_1(\vartheta))^2 = \operatorname{Var}_\vartheta(T).$$

Aus der Differenzierbarkeit von ζ folgt

$$U_\vartheta = \frac{\partial}{\partial\vartheta}\log L(\vartheta, x) = T - \zeta'(\vartheta), \quad \vartheta \in \Theta$$

und somit $I(\vartheta) = \operatorname{Var}_\vartheta(U_\vartheta) = \operatorname{Var}_\vartheta(T) > 0$. Weiter können wir schreiben

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\vartheta}\mathbb{E}_\vartheta[S] =& \big(u_S(\vartheta)e^{-\zeta(\vartheta)}\big)' = \big(u_S'(\vartheta) - u_S(\vartheta)\zeta'(\vartheta)\big)e^{-\zeta(\vartheta)}\\ =& \mathbb{E}_\vartheta[ST] - \mathbb{E}_\vartheta[S]\zeta'(\vartheta) = \mathbb{E}_\vartheta[SU_\vartheta]\\ =& \int_{\mathcal{X}} S(x)\frac{\partial}{\partial\vartheta}L(\vartheta, x)\mu(\mathrm{d}x).\end{aligned}$$

Daher gilt einerseits (3.1) für alle $h \in \mathcal{L}^1(\mathbb{P}_\vartheta)$ und anderseits folgt die Regularität des Modells. □

Korollar 3.31 (Existenz von besten Schätzern)**.** *Für jedes statistische Modell gegeben durch eine Exponentialfamlie mit differenzierbarem η und $\eta' \neq 0$ ist die zugrunde liegende Statistik T ein bester und Cramér-Rao-effizienter Schätzer für $\rho(\vartheta) := \mathbb{E}_\vartheta[T] = \zeta'(\vartheta)/\eta'(\vartheta)$. In dem Fall gilt*

$$\mathrm{Var}_\vartheta(T) = \rho'(\vartheta)/\eta'(\vartheta) \quad \textit{und} \quad I(\vartheta) = \eta'(\vartheta)\rho'(\vartheta) \quad \textit{für alle } \vartheta \in \Theta.$$

Für natürliche Exponentialfamilien gilt also insbesondere $\mathrm{Var}_\eta(T) = I(\eta)$.

Beweis. Folgt unmittelbar aus Satz 3.25 und Lemma 3.30. Für natürliche Exponentialfamilien gilt also $\mathrm{Var}_\eta(T) = \rho'(\eta) = I(\eta)$ und die Informationsschranke ist gegeben durch $\rho'(\eta)^2/I(\eta) = I(\eta)$. □

Beispiel 3.32.

(i) $(Bin(1,p)^{\otimes n})_{p\in(0,1)}$ ist eine Exponentialfamilie in $\eta(p) = n\log\frac{p}{1-p}$ und $T(x) = \overline{x}_n$ (Beispiel 3.3). Also ist η differenzierbar mit $\eta'(p) = \frac{n}{p(1-p)} > 0$, sodass $\overline{X}_n$ nicht nur effizient (Beispiel 3.28), sondern auch varianzminimierend unter allen unverzerrten Schätzern (UMVU) ist.

(ii) $(\mathcal{N}(\mu,\sigma^2)^{\otimes n})_{\mu\in\mathbb{R}}$ mit bekanntem $\sigma > 0$ ist eine Exponentialfamilie in $\eta(\mu) = n\mu/\sigma^2$ und $T(x) = \overline{x}_n$. Aus der Effizienz von $\overline{X}_n$ (Beispiel 3.28) folgt wieder, dass $\overline{X}_n$ gleichmäßig bester Schätzer unter allen erwartungstreuen Schätzern ist. Da $\overline{X}_n$ nicht von σ abhängt, ist das Stichprobenmittel sogar bester Schätzer für den Erwartungswert für alle Normalverteilungen.

Obwohl die Cramér-Rao-Schranke häufig nicht erreicht wird, dient sie doch als hilfreiche Bezugsgröße. In vielen Modellen kann man für große Stichprobenumfänge beliebig nahe an die Informationsschranke herankommen und sie asymptotisch für $n \to \infty$ erreichen.

Beispiel 3.33. Wir betrachten die mathematische Stichprobe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu,\sigma^2)$ mit unbekanntem Parametervektor $\vartheta = (\mu,\sigma^2)^\top$. Dann ist für $n = 1$ die Likelihoodfunktion $L(\mu,\sigma^2;x) = (2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/(2\sigma^2)}$ und die Scorefunktion ist gegeben durch

$$U_\vartheta(x) = \begin{pmatrix} \frac{\partial}{\partial\mu}L(\mu,\sigma^2;x) \\ \frac{\partial}{\partial\sigma^2}L(\mu,\sigma^2;x) \end{pmatrix} = \begin{pmatrix} \frac{x-\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4} \end{pmatrix}.$$

Wir erhalten

$$I_n(\mu,\sigma^2) = n\mathbb{E}_\vartheta\left[U_\vartheta(X_1)\cdot U_\vartheta(X_1)^\top\right] = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Folglich gilt für alle regulären, erwartungstreuen Schätzer $\widehat{\mu}$ und $\widehat{\sigma}^2$ von μ bzw. σ^2:

$$\mathrm{Var}_{\mu,\sigma}(\widehat{\mu}) \geqslant \frac{\sigma^2}{n}, \qquad \mathrm{Var}_{\mu,\sigma^2}(\widehat{\sigma}^2) \geqslant \frac{2\sigma^4}{n}.$$

Nun ist die Statistik $T(X_1,\ldots,X_n) = (\overline{X}_n, \overline{S}_n^2)^\top$ mit $\overline{S}_{n-1}^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ suffizient, vollständig (Beispiel 3.12 und Satz 3.21), ein erwartungstreuer Schätzer von $(\mu,\sigma^2)^\top$ und somit auch UMVU. Allerdings gilt für die Kovarianzmatrix

$$\mathrm{Cov}_{\mu,\sigma}(T) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix},$$

sodass $\overline{S}_n^2$ nicht effizient ist. Immerhin gilt $\mathrm{Var}_{\mu,\sigma^2}(\overline{S}_n^2) = \frac{2\sigma^4}{n}(1+o(1))$ für $n \to \infty$.

3.4 Asymptotik für den Maximum-Likelihood-Schätzer

Die Idee aus dem vorangegangen Beispiel wollen wir nun vertiefen. Unser Ziel ist, nachzuweisen, dass in einer mathematischen Stichprobe der Maximum-Likelihood-Schätzer asymptotisch für Stichprobenumfänge $n \to \infty$ die Cramér-Rao-Schranke erreicht. Im Folgenden betrachten wir stets eine Folge dominierter Produktmodelle $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ mit $\Theta \subseteq \mathbb{R}^d$ und (eindimensionaler) Likelihoodfunktion $L(\vartheta, x) = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x)$. Die multivariate Likelihood schreiben wir als $L_n(\vartheta, x) = \prod_{i=1}^n L(\vartheta, x_i)$, $x \in \mathcal{X}^n$.

Zunächst zeigen wir Konsistenz. Als Kandidat für einen Maximum-Likelihood-Schätzer kommen Lösungen der Likelihoodgleichungen (RLE: root of the likelihood equation)

$$\forall i = 1, \dots, d: \qquad \frac{\partial}{\partial\vartheta_i} L_n(\vartheta, X) = 0$$

infrage. Äquivalent formuliert, suchen wir ein $\widehat{\vartheta}_n$, sodass $U^n_{\widehat{\vartheta}_n}(X) = 0 \in \mathbb{R}^d$ für die Scorefunktion $U^n_\vartheta = \nabla_\vartheta \log L_n(\vartheta, X)$. Man beachte, dass es sich hierbei ggf. um lokale Maxima handeln kann.

Satz 3.34. *Es sei $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ eine Folge dominierter Produktmodelle mit (eindimensionaler) Likelihoodfunktion $L(\vartheta, x) = \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x)$ und einer σ-kompakten Teilmenge $\Theta \subseteq \mathbb{R}^d$. Es gelte weiterhin:*

(i) *$(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ist regulär mit strikt positiver Likelihood $L(\vartheta, X)$, Scorefunktion $U_\vartheta(x) := \frac{\nabla_\vartheta L(\vartheta,x)}{L(\vartheta,x)}$ und positiv definiter Fisher-Information $I(\vartheta) := \mathbb{E}_\vartheta[U_\vartheta(X)U_\vartheta(X)^\top]$.*

(ii) *Es gilt die Identifizierbarkeitsbedingung $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$ für $\vartheta \neq \vartheta'$.*

(iii) *$\Theta \ni \vartheta \mapsto L(\vartheta, x)$ ist μ-f.s. zwei mal stetig differenzierbar und es gilt*

$$\nabla_\vartheta \int_{\mathcal{X}} \psi(\vartheta, x)\mu(\mathrm{d}x) = \int_{\mathcal{X}} \nabla_\vartheta \psi(\vartheta, x)\mu(\mathrm{d}x)$$

für $\psi(\vartheta, x) = L(\vartheta, x)$ und $\psi(\vartheta, x) = \frac{\partial}{\partial\vartheta_j} L(\vartheta, x), j = 1, \dots, d$.

(iv) *Für $j, k \in \{1, \dots, d\}$ existieren Zufallsvariablen $M_{jk} \in \mathcal{L}^1(\mathbb{P}_{\vartheta_0})$ mit*

$$\sup_{\vartheta:|\vartheta-\vartheta_0|<c} \Big|\frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k} \log L(\vartheta, x)\Big| \leqslant M_{jk}(x)$$

für $\vartheta_0 \in \Theta$ und ein $c > 0$.

Dann existiert eine Folge $\widehat{\vartheta}_n = \widehat{\vartheta}_n(X_1, \dots, X_n)$, sodass für $\vartheta_0 \in \Theta$

$$\mathbb{P}_{\vartheta_0}\Big(\forall i = 1, \dots, d: \frac{\partial}{\partial\vartheta_i} L_n(\widehat{\vartheta}_n, X) = 0\Big) \to 1 \quad \text{und} \quad \widehat{\vartheta}_n \overset{\mathbb{P}^{\otimes n}_{\vartheta_0}}{\to} \vartheta_0. \tag{3.3}$$

Beweis. Wir schreiben kurz $\ell_n(\vartheta, x) := \log L_n(\vartheta, x) = \sum_{i=1}^n \log L(\vartheta, x_i)$. Für $S_r(\vartheta_0) := \{\vartheta \in \mathbb{R}^d : \sqrt{n}|\vartheta - \vartheta_0| = r\} \subseteq \Theta$ werden wir zeigen, dass für jedes $\varepsilon > 0$ ein $r > 0$ und ein $N \in \mathbb{N}$ existieren, sodass gilt:

$$\mathbb{P}_{\vartheta_0}\big(\{\forall\vartheta \in S_r(\vartheta_0) : \ell_n(\vartheta, X) < \ell_n(\vartheta_0, X)\}\big) \geqslant 1 - \varepsilon \quad \text{für alle } n \geqslant N. \tag{3.4}$$

Auf diesem Ereignis muss $\ell_n(\cdot, X)$ im Ball $\{\vartheta \in \Theta : |\vartheta - \vartheta_0| \leqslant r/\sqrt{n}\}$ ein lokales Maximum $\widehat{\vartheta}_n$ haben für das notwendigerweise die Likelihoodgleichungen erfüllt sind (die Messbarkeit folgt aus der σ-Kompaktheit von Θ, vgl. Witting and Müller-Funk (1995, Hilfssatz 6.7)). Da für jedes $c > 0$ $\lim_{n\to\infty} \mathbb{P}_{\vartheta_0}(|\widehat{\vartheta}_n - \vartheta_0| > c) \leqslant \lim_{n\to\infty} \mathbb{P}_{\vartheta_0}(|\widehat{\vartheta}_n - \vartheta_0| > r/\sqrt{n}) = 0$ für alle $c > 0$ gilt, folgt die Behauptung.

Es bleibt also noch (3.4) zu zeigen. Beachte, dass im Produktmodell die Scorefunktion und die Hessematrix (bzw. die Jakobimatrix der Scorefunktion) durch

$$U_\vartheta^n := \nabla_\vartheta \ell_n(\vartheta, X) = \sum_{i=1}^n \nabla_\vartheta \log L(\vartheta, X_i) = \sum_{i=1}^n U_\vartheta(x_i),$$
$$V_\vartheta^n := \Big(\frac{\partial}{\partial_j \partial_k} \ell_n(\vartheta, X)\Big)_{j,k=1,\dots,d} = \sum_{i=1}^n V_\vartheta(X_i) \qquad \text{mit} \qquad V_\vartheta(X_i) = \Big(\frac{\partial}{\partial_j \partial_k} \log L(\vartheta, X_i)\Big)_{j,k=1,\dots,d},$$

gegeben sind. Aus einer (multivariaten) Taylorentwicklung um ϑ_0 folgt für alle $\vartheta \in S_r(\vartheta_0)$ und ein ξ zwischen ϑ und ϑ_0

$$\begin{aligned} \ell_n(\vartheta, X) - \ell_n(\vartheta_0, X) =& \langle U_{\vartheta_0}^n, \vartheta - \vartheta_0\rangle + \frac{1}{2}\langle \vartheta - \vartheta_0, V_\xi^n(\vartheta - \vartheta_0)\rangle \\ =& -\frac{1}{2}\underbrace{\langle (\vartheta - \vartheta_0), nI(\vartheta_0)(\vartheta - \vartheta_0)\rangle}_{>0} + \underbrace{\langle U_{\vartheta_0}^n, \vartheta - \vartheta_0\rangle}_{=:L_{\vartheta_0}(X)} \\ &+ \frac{1}{2}\underbrace{\big\langle \vartheta - \vartheta_0, \big(V_\xi^n + nI(\vartheta_0)\big)(\vartheta - \vartheta_0)\big\rangle}_{=:Q_{\vartheta_0}(X)}. \end{aligned} \tag{3.5}$$

Da $I(\vartheta_0)$ symmetrisch und positiv definit ist, existieren eine Orthogonalmatrix $P \in \mathbb{R}^{d\times d}$ und Eigenwerte $0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_d$, sodass $I(\vartheta) = P\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)P^\top$ und

$$\langle (\vartheta - \vartheta_0), nI(\vartheta_0)(\vartheta - \vartheta_0)\rangle = n\langle P^\top(\vartheta - \vartheta_0), \mathrm{diag}(\lambda_1, \ldots, \lambda_d)P^\top(\vartheta - \vartheta_0)\rangle \geqslant \lambda_1 n |P^\top(\vartheta - \vartheta_0)|^2 = \lambda_1 r^2$$

für alle $\vartheta \in S_r(\vartheta_0)$. Folglich ist der erste Term echt negativ. Wir werden nun zeigen, dass die übrigen beiden Terme in (3.5) mit hoher Wahrscheinlichkeit kleiner als $\lambda_1 r^2/2$ sind.

Den linearen Term schätzen wir mit Markovs Ungleichung ab: Für jedes $c > 0$ erhalten wir mit der Spektralnorm $\|\cdot\|_2$

$$\begin{aligned} \mathbb{P}_{\vartheta_0}^{\otimes n}(|L_{\vartheta_0}(X)| > \lambda_1 r^2/4) &= \mathbb{P}_{\vartheta_0}^{\otimes n}\Big(\langle \vartheta - \vartheta_0, U_{\vartheta_0}^n\rangle \cdot \langle U_{\vartheta_0}^n, \vartheta - \vartheta_0\rangle > \lambda_1^2 r^4/16\Big) \\ &= \mathbb{P}_{\vartheta_0}^{\otimes n}\Big((\vartheta - \vartheta_0)^\top U_{\vartheta_0}^n (U_{\vartheta_0}^n)^\top (\vartheta - \vartheta_0) > \lambda_1^2 r^4/16\Big) \\ &\leqslant \frac{16}{\lambda_1^2 r^4}\mathbb{E}_{\vartheta_0}\Big[(\vartheta - \vartheta_0)^\top U_{\vartheta_0}^n (U_{\vartheta_0}^n)^\top (\vartheta - \vartheta_0)\Big] \\ &= \frac{16}{\lambda_1^2 r^4}(\vartheta - \vartheta_0)^\top (nI(\vartheta_0))(\vartheta - \vartheta_0) \\ &\leqslant \frac{16}{\lambda_1^2 r^4}|\vartheta - \vartheta_0|^2 \|nI(\vartheta_0)\|_2 \\ &= \frac{16}{\lambda_1^2 r^2}\|I(\vartheta_0)\|_2, \end{aligned}$$

was kleiner als $\varepsilon/2$ ist, wenn r groß genug gewählt wird. Den quadratischen Term in (3.5) zerlegen wir in

$$\begin{aligned} Q_{\vartheta_0}(X) &= \big\langle \vartheta - \vartheta_0, \big(V_{\vartheta_0}^n + nI(\vartheta_0)\big)(\vartheta - \vartheta_0)\big\rangle + \big\langle \vartheta - \vartheta_0, \big(V_\xi^n - V_{\vartheta_0}^n\big)(\vartheta - \vartheta_0)\big\rangle \\ &\leqslant |\vartheta - \vartheta_0|^2 d\|V_{\vartheta_0}^n + nI(\vartheta_0)\|_{max} + |\vartheta - \vartheta_0|^2 d\|V_\xi^n - V_{\vartheta_0}^n\|_{max}, \end{aligned}$$

wobei wir die Spektralnorm $\|A\|_2 \leqslant d\|A\|_{max} = d\max_{j,k}|a_{j,k}|$ mit der Maximumsnorm für eine Matrix $A = (a_{j,k}) \in \mathbb{R}^{d\times d}$ abgeschätzt haben. Nach dem (schwachen) Gesetz der großen Zahlen gilt zudem $\frac{1}{n}V_{\vartheta_0}^n = \frac{1}{n}\sum_{i=1}^n V_{\vartheta_0}(X_i) \overset{\mathbb{P}_{\vartheta_0}^{\otimes n}}{\to} -I(\vartheta_0)$ (koordinatenweise) und wir erhalten für $c > 0$ und

$n \to \infty$

$$\begin{aligned}\mathbb{P}_{\vartheta_0}^{\otimes n}\Big(|\vartheta-\vartheta_0|^2 d\Big\|V_{\vartheta_0}^n + nI(\vartheta_0)\Big\|_{max} > r^2 c\Big) &= \mathbb{P}_{\vartheta_0}^{\otimes n}\Big(dr^2\frac{1}{n}\Big\|V_{\vartheta_0}^n + nI(\vartheta_0)\Big\|_{max} > r^2 c\Big)\\ &= \mathbb{P}_{\vartheta_0}^{\otimes n}\Big(\max_{j,k}\Big|\big(\frac{1}{n}V_{\vartheta_0}^n + I(\vartheta_0)\big)_{j,k}\Big| > \frac{c}{d}\Big)\\ &\leqslant \sum_{j,k}\mathbb{P}_{\vartheta_0}^{\otimes n}\Big(\Big|\Big(\frac{1}{n}V_{\vartheta_0}^n + I(\vartheta_0)\Big)_{j,k}\Big| > \frac{c}{d}\Big) \to 0.\end{aligned}$$

Für den zweiten Term erhalten wir für hinreichend große n und jedes $c > 0$

$$\begin{aligned}&\mathbb{P}_{\vartheta_0}^{\otimes n}\Big(|\vartheta-\vartheta_0|^2 d\|V_\xi^n - V_{\vartheta_0}^n\|_{max} > r^2 c\Big)\\ &\quad\leqslant \frac{d}{r^2 c}|\vartheta-\vartheta_0|^2\mathbb{E}_{\vartheta_0}\big[\|V_\xi^n - V_{\vartheta_0}^n\|_{max}\big]\\ &\quad\leqslant \frac{d}{c}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\vartheta_0}\Big[\max_{j,k}\sup_{\vartheta:|\vartheta-\vartheta_0|<r/\sqrt{n}}\big|\big(V_\vartheta(X_i) - V_{\vartheta_0}(X_i)\big)_{j,k}\big|\Big]\\ &\quad= \frac{d}{c}\sum_{j,k}\mathbb{E}_{\vartheta_0}\Big[\sup_{\vartheta:|\vartheta-\vartheta_0|<r/\sqrt{n}}\Big|\frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k}\log L(\vartheta, X_1) - \frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k}\log L(\vartheta_0, X_1)\Big|\Big] \to 0 \qquad (3.6)\end{aligned}$$

aufgrund von dominierter Konvergenz (unter Voraussetzung (iv)) und der Stetigkeit von $\frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k}\log L(\xi, x)$. Bezeichnet $o_p(1)$ Terme, die stochastisch gegen Null konvergieren, erhalten wir insgesamt

$$\ell_n(\vartheta, X) \leqslant \ell_n(\vartheta_0, X) - r^2\big(\frac{\lambda_1}{2} - \frac{\lambda_1}{4} - o_p(1)\big) < \ell_n(\vartheta_0, X)$$

mit Wahrscheinlichkeit größer als $1-\varepsilon$ für hinreichend große n. Somit ist (3.4) gezeigt. □

Dieser Satz zeigt also die (asymptotische) Existenz einer konsistenten Folge von Lösungen der Likelihoodgleichungen. Für eine gegebene Folge solcher Lösungen muss die Konsistenz jedoch überprüft werden, es sei denn, für hinreichend große n besitzen die Likelihoodgleichungen eine eindeutige Lösung. Zudem muss ein RLE kein MLE sein (selbst dann nicht, wenn der RLE eindeutig ist). Hinreichende Bedingungen hierfür können im jeweiligen Einzelfall nachgeprüft werden.

Beispiel 3.35. Es sei $X_1, \dots, X_n \in \mathcal{X}$ eine mathematische Stichprobe, wobei X_i gemäß einer natürlichen Exponentialfamilie $(\mathbb{P}_\eta)_{\eta\in\Xi}$ in $T(X_i)$ verteilt ist, d.h.

$$L(\eta, x) = \exp(\eta^\top T(x) - \zeta(\eta))h(x), \qquad x \in \mathcal{X}.$$

In diesem Fall können die Bedingungen (i)-(iii) aus Satz 3.34 leicht nachgeprüft werden (analog zu Lemma 3.30) und (iv) folgt aus $\frac{\partial^2}{\partial\eta_j\partial\eta_k}\log L(\eta, x) = -\frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k}\zeta(\eta)$. Es gilt $U_\eta^n = \sum_{i=1}^n (T(X_i) - \nabla_\eta\zeta)$. Ist also $\widehat{\vartheta}_n = \frac{1}{n}\sum_i T(X_i) \in \Theta := \operatorname{ran}(\nabla_\eta\zeta)$, dann ist $\widehat{\vartheta}_n$ der eindeutige RLE von $\vartheta := \nabla_\eta\zeta(\eta)$ und auch der eindeutige MLE von ϑ, da $(\frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k}\zeta(\eta))_{j,k} = \operatorname{Var}_\eta(T(X_i))$ positiv definit ist. Ist $g = \nabla_\eta\zeta$ invertierbar, so ist der eindeutige RLE (und MLE) von η gegeben durch $\widehat{\eta}_n = g^{-1}(\widehat{\vartheta}_n)$.

Im Fall $\widehat{\vartheta}_n \notin \Theta$ sagt uns obiger Satz, dass zumindest $\lim_{n\to\infty}\mathbb{P}_\eta(\widehat{\vartheta}_n \in \Theta) = 1$ gilt. Letzteres kann man in diesem speziellen Fall auch direkt mit dem starken Gesetz der großen Zahlen folgern: $\widehat{\vartheta}_n \to \mathbb{E}_\eta[T(X_1)] = \nabla_\eta\zeta(\eta)$ f.s. für $n \to \infty$.

Mithilfe der Techniken aus dem Beweis von Satz 3.34 können wir auch asymptotische Normalität für jede Folge konsistenter RLEs schlussfolgern.

Satz 3.36. *Es sei $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ eine Folge dominierter Produktmodelle, die die Bedingungen (i)-(iv) aus Satz (3.34) erfüllt. Für jede Folge $(\widehat{\vartheta}_n)$ mit den Eigenschaften (3.3) gilt*

$$\sqrt{n}(\widehat{\vartheta}_n - \vartheta_0) \xrightarrow{d} \mathcal{N}\big(0, I(\vartheta_0)^{-1}\big).$$

Beweis. Wir betrachten den Ball $B_r := \{\vartheta \in \mathbb{R}^d : |\vartheta - \vartheta_0| \leqslant r\} \subseteq \Theta$ für jedes hinreichend kleine $r > 0$. Für jede Folge $(\widehat{\vartheta}_n)$ mit (3.3) gilt dann $\mathbb{P}_{\vartheta_0}^{\otimes n}(\{\nabla_\vartheta L_n(\widehat{\vartheta}_n, X) = 0\} \cap \{\widehat{\vartheta}_n \in B_r\}) \to 1$. Wir können uns also auf dieses Ereignis beschränken. Mit dem Mittelwertsatz für vektorwertige Funktionen in mehreren Variablen erhalten wir

$$-U_{\vartheta_0}^n = U_{\widehat{\vartheta}_n}^n - U_{\vartheta_0}^n = \Big(\int_0^1 V_{\vartheta_0 + t(\widehat{\vartheta}_n - \vartheta)}^n \mathrm{d}t\Big)(\widehat{\vartheta}_n - \vartheta).$$

Weiter gilt mit der Abschätzung aus (3.6)

$$\frac{1}{n}\Big\| \int_0^1 V_{\vartheta_0 + t(\widehat{\vartheta}_n - \vartheta)}^n \mathrm{d}t - V_{\vartheta_0}^n \Big\|_{max} \leqslant \frac{1}{n} \sup_{\vartheta \in B_r} \|V_\vartheta^n - V_{\vartheta_0}^n\|_{max} \xrightarrow{\mathbb{P}_{\vartheta_0}^{\otimes n}} 0.$$

Zusammen mit dem Gesetz der großen Zahlen $\frac{1}{n} V_{\vartheta_0}^n = \frac{1}{n}\sum_{i=1}^n V_{\vartheta_0}(X_i) \xrightarrow{\mathbb{P}_{\vartheta_0}^{\otimes n}} -I(\vartheta_0)$ folgt

$$-\frac{1}{n} U_{\vartheta_0}^n = -I(\vartheta_0)(\widehat{\vartheta}_n - \vartheta) + o_p(1)|\widehat{\vartheta}_n - \vartheta|. \tag{3.7}$$

Da $\mathbb{E}_{\vartheta_0}[U_{\vartheta_0}(X_i)] = 0$ und $\mathrm{Var}_{\vartheta_0}(U_{\vartheta_0}(X_i)) = I(\vartheta_0)$, folgt aus dem zentralen Grenzwertsatz

$$\frac{1}{\sqrt{n}} U_{\vartheta_0}^n = \frac{1}{\sqrt{n}} \sum_{i=1}^n U_{\vartheta_0}(X_i) \xrightarrow{d} \mathcal{N}(0, I(\vartheta_0))$$

und somit $n^{-1/2} I(\vartheta_0)^{-1} U_{\vartheta_0}^n \xrightarrow{d} \mathcal{N}(0, I(\vartheta_0)^{-1})$. Wenden wir diese Verteilungskonvergenz und Slutskys Lemma auf (3.7) an, erhalten wir

$$\sqrt{n}(\widehat{\vartheta}_n - \vartheta) \xrightarrow{d} \mathcal{N}(0, I(\vartheta_0)^{-1})$$

für $n \to \infty$ unter $\mathbb{P}_{\vartheta_0}$. □

Bemerkung 3.37.

(i) Als asymptotische Kovarianzmatrix im zentralen Grenzwertsatz für eine Folge $(\widehat{\vartheta}_n)$ von konsistenten RLEs erhalten wir also gerade die Cramér-Rao-Schranke. In diesem Fall sprechen wir von *asymptotischer Effizienz*.

(ii) In Modellen, in denen die Verteilung von Schätzern für endliche Stichprobengrößen nicht einfach beschrieben werden kann, sind solche zentralen Grenzwertsätze nützlich, um Konfidenzbereiche zu konstruieren, welche asymptotisch für $n \to \infty$ das gegebene Niveau $\alpha \in (0, 1)$ erreichen. Dabei ist zu beachten, dass die Varianz $I(\vartheta_0)^{-1}$ vom unbekannten ϑ_0 abhängt und wiederum geschätzt werden muss (Übung ☑).

3.5 *Ergänzung: Verallgemeinertes lineares Modell

Mit Hilfe von Exponentialfamilien wollen wir nun lineare Modelle verallgemeinern. Wie in Beispiel 3.3 gesehen bildet $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}}$ eine Exponentialfamilie mit natürlichem Parameter $\eta(\mu) = \mu/\sigma^2$ und Statistik $T(x) = x$, die ein effizienter Schätzer des Parameters $\rho(\mu) = \mathbb{E}_\mu[T] = \mu$ ist. Im gewöhnlichen linearen Modell sind nun die Beobachtungen gegeben durch

$$\mathbb{R}^n \ni Y = X\beta + \varepsilon,$$

mit Parametervektor $\beta \in \mathbb{R}^k$, Designmatrix $X \in \mathbb{R}^{n \times k}$ und iid. Fehlervariablen $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ mit Varianz $\sigma > 0$. Schreiben wir die Designmatrix als

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{mit Zeilenvektoren } x_1, \ldots, x_n \in \mathbb{R}^k,$$

ist Beobachtung Y_i gemäß $\mathcal{N}(x_i\beta, \sigma^2)$ verteilt, folgt also einer Exponentialfamilie mit $\eta_i(\beta) = x_i\beta/\sigma^2$ und $\rho_i(\beta) = x_i\beta$, $i = 1, \dots, n$. Lassen wir nun andere Exponentialfamilien zu, können wir sowohl Situationen modellieren, in denen der Zusammenhang zwischen $\mathbb{E}[Y_i]$ und den Kovariablen (codiert in der Designmatrix X) nichtlinear ist, als auch diskrete Beobachtungen Y_i zulassen.

Definition 3.38. Auf einem Produktmodell $(\mathcal{X}^n, \mathscr{F}^{\otimes n})$ liegt ein verallgemeinertes lineares Modell (GLM: generalized linear model) mit n unabhängigen Beobachtungen $Y_1, \dots, Y_n$ vor, falls die Randverteilungen von Y_i durch natürliche Exponentialfamilien gegeben sind mit Dichten

$$\frac{\mathrm{d}\mathbb{P}_{\eta_i}^{Y_i}}{\mathrm{d}\mu}(y_i) = \exp\Big(\frac{\eta_i y_i - \zeta(\eta_i)}{\varphi}\Big) h(y_i, \varphi), \quad i = 1, \dots, n,$$

bzgl. einem dominierenden Maß μ, mit unbekanntem Dispersionsparameter $\varphi > 0$,

$$\eta_i \in \Xi = \Big\{\eta \in \mathbb{R} \colon \int_{\mathcal{X}} e^{\eta y/\varphi} h(y, \varphi)\mu(\mathrm{d}y) \in (0, \infty)\Big\} \subseteq \mathbb{R}$$

für alle i und bekannten Funktionen $\zeta \colon \Xi \to \mathbb{R}$ und $h \colon \mathcal{X} \times (0, \infty) \to \mathbb{R}_+$ mit $\zeta''(\eta) > 0$ für alle inneren Punkte $\eta \in \Xi^\circ$. Setze $\rho(\eta_i) := \mathbb{E}_\eta[Y_i]$. Für einen unbekannten Parametervektor $\beta \in \mathbb{R}^k$, eine Designmatrix $X \in \mathbb{R}^{n\times k}$ und eine bijektive, stetig differenzierbare Funktion $g \colon \mathbb{R} \to \mathbb{R}$ gelte weiter

$$\begin{pmatrix} g(\rho(\eta_1)) \\ \vdots \\ g(\rho(\eta_n)) \end{pmatrix} = X\beta.$$

g heißt Linkfunktion. Falls $\rho = g^{-1}$, gilt $(\eta_1, \dots, \eta_n)^\top = X\beta$ und g heißt kanonische Linkfunktion (oder kanonischer Link).

Während β der interessierende Parameter ist, wird φ als Störparameter angesehen. Für fixiertes φ ist Y_i also gemäß einer natürlichen Exponentialfamilie in $T(y) = y/\varphi$ verteilt. Aus den Eigenschaften natürlicher Exponentialfamilien folgt

$$\mathbb{E}_{\beta,\varphi}[Y_i] = \zeta'(\eta_i) \quad \text{und} \quad \mathrm{Var}_{\beta,\varphi}(Y_i) = \varphi\zeta''(\eta_i), \quad i = 1, \dots, n.$$

Beispiel 3.39. Das gewöhnliche lineare Modell ist ein GLM mit kanonischer Linkfunktion $g(x) = x$, $\zeta(\eta) = \eta^2/2$ und Dispersionsparameter $\varphi = \sigma^2$. Lassen wir allgemeinere Linkfunktionen zu, erhalten wir *nicht-lineare Regressionsmodelle* (mit normalverteilten Fehlern) gegeben durch Beobachtungen $Y_i \sim \mathcal{N}(g^{-1}((X\beta)_i), \varphi)$.

Der Dispersionsparameter wird dazu verwendet eine Unterschätzung der (empirisch beobachteten) Varianz durch das Modell auszugleichen (siehe Übung ☑).

Um den unbekannten Parametervektor β in einem verallgemeinerten linearen Modell zu schätzen, verwenden wir den Maximum-Likelihood-Ansatz. Da ζ' streng monoton wachsend und die Linkfunktion g invertierbar sind, existiert die Funktion $\psi := (g \circ \rho)^{-1}$. Ist $x_i \in \mathbb{R}^k$ wieder die i-te Zeile von X, kann die Loglikelihood-Funktion geschrieben werden als

$$\log L(\beta, \varphi; y) = \sum_{i=1}^n \Big(\frac{\psi(x_i\beta)y_i - \zeta(\psi(x_i\beta))}{\varphi} + \log(c(y_i, \varphi))\Big).$$

Als notwendige Bedingung an einen Maximum-Likelihood-Schätzer $\widehat{\beta}$ erhalten wir durch Ableiten

$$\nabla_\beta \log L(\widehat{\beta}, \varphi; y) = \frac{1}{\varphi}\sum_{i=1}^n \big(y_i - \rho(\psi(x_i\widehat{\beta}))\big)\psi'(x_i\widehat{\beta})x_i^\top = 0. \tag{3.8}$$

Lemma 3.40. *In einem verallgemeinerten linearen Modell mit kanonischer Linkfunktion ist die Fisher-Information gegeben durch*

$$I(\beta) = \frac{1}{\varphi}\sum_{i=1}^n \zeta''(x_i\beta)x_i^\top x_i \in \mathbb{R}^{k\times k}.$$

Ist $I(\beta)$ positiv definit für alle β und existiert eine Lösung $\widehat{\beta}$ von (3.8), so ist $\widehat{\beta}$ der eindeutige Maximum-Likelihood-Schätzer von β.

Beweis. Aus Lemma 3.30 folgt, dass die Fisher-Information im natürlichen Parameter $(\eta_1, \dots, \eta_n)^\top$ gegeben ist durch $\frac{1}{\varphi} \sum_{i=1}^n \zeta''(\eta_i)$. Die Reparametrisierung $\eta_i = x_i\beta$ zusammen mit der Kettenregel ergibt die Darstellung von $I(\beta)$.

Der kanonische Link ist gegeben durch $g = \rho^{-1}$, sodass ψ in (3.8) die Identität ist. Wegen $\rho = \zeta'$, gilt also

$$\frac{\partial^2 \log L(\beta, \varphi; y)}{\partial\beta\partial\beta^\top} = -\frac{1}{\varphi} \sum_{i=1}^n \zeta''(x_i\beta) x_i^\top x_i = -I(\beta).$$

Da $I(\beta) > 0$, ist $\beta \mapsto -\log L(\beta, \varphi; y)$ streng konvex und somit $\widehat{\beta}$ der eindeutige Maximum-Likelihood-Schätzer. □

Bemerkung 3.41.

(i) Typischerweise besitzt $\widehat{\beta}$ keine geschlossene Form mehr und muss durch numerische Verfahren bestimmt werden. *Fishers Scoring-Methode* verwendet hierfür das iterative Verfahren

$$\widehat{\beta}^{(t+1)} = \widehat{\beta}^{(t)} + I(\beta)^{-1} \nabla_\beta \log L(\widehat{\beta}^{(t)}, \varphi; y), \quad t = 0, 1, \dots$$

(Beachte, dass sich der unbekannte Dispersionsparameter φ gerade rauskürzt). Für den kanonischen Link ist dieses Verfahren äquivalent zur *Newton-Raphson-Methode.*

(ii) Ist g nicht der kanonische Link, ist eine Lösung von (3.8) nicht notwendigerweise ein Maximum-Likelihood-Schätzer.

Zwei wichtige Beispielklassen für verallgemeinerte lineare Modelle sind die *Poisson-Regression* und die *logistische Regression*, die abschließend eingeführt werden.

Die Poisson-Regression modelliert unabhängige Poisson-verteilte Beobachtungen, deren Intensitätsparameter von Kovariablen abhängen. Sie eignet sich also für Beobachtungen, die Zähldatenstruktur haben. Wir hatten bereits gesehen, dass die Familie $(Poiss(\lambda))_{\lambda>0}$ eine Exponentialfamilie in $\eta(\lambda) = \log\lambda$ und $T(x) = x$ ist: Bezüglich des Zählmaßes ist die Likelihood-Funktion gegeben durch

$$L(\lambda, x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{x\log\lambda - \lambda}, \quad x \in \mathbb{Z}_+,$$

und es gilt $\rho(\lambda) = \mathbb{E}_\lambda[T] = \lambda$.

Definition 3.42. Ein verallgemeinertes lineares Modell auf $(\mathbb{Z}_+^n, \mathscr{P}(\mathbb{Z}_+^n))$ heißt Poisson-Regression, falls die unabhängigen Beobachtungen Y_i $Poiss(\lambda_i)$-verteilt sind, wobei $\lambda_i = e^{\eta_i}$ mit natürlichen Parametern $\eta_i \in \mathbb{R}, i = 1, \dots, n$, und

$$\eta = (\eta_1, \dots, \eta_n)^\top = X\beta$$

mit unbekanntem $\beta \in \mathbb{R}^k$ und Designmatrix $X \in \mathbb{R}^{n\times k}$.

Bemerkung 3.43. Wir verwenden hier also den kanonischen Link $g(\lambda) = \log\lambda$. In der Praxis wird oft das erweiterte Modell $Y_i \sim Poiss(\lambda_i \cdot s_i)$ verwendet für einen so genannten *Zählrahmen* $s_i > 0$, $i = 1, \dots, n$. Dann gilt $\mathbb{E}_\beta[Y_i] = \exp(x_i\beta + \log(s_i))$ mit den Zeilen x_i von X. Der Term $\log(s_i)$ wird als *Offset* bezeichnet, da er jeder Beobachtung einen individuellen "Intercept" zuweist.

Definition 3.44. Ein verallgemeinertes lineares Modell auf $(\{0,1\}^n, \mathscr{P}(\{0,1\}^n))$ heißt logistische Regression, falls die unabhängigen Beobachtungen Y_i $Bin(1, p_i)$-verteilt sind, $i = 1, \dots, n$, mit natürlichem Parameterraum $\mathbb{R}$, der kanonischen Link-Funktion $g\colon (0,1) \to \mathbb{R}, g(p) = \log(p/(1-p))$ und

$$\eta = (g(p_1), \dots, g(p_n))^\top = X\beta$$

mit unbekanntem $\beta \in \mathbb{R}^k$ und Designmatrix $X \in \mathbb{R}^{n\times k}$. Die Funktion g heißt Logit-Funktion und ihre Umkehrfunktion $g^{-1}\colon \mathbb{R} \to (0,1), g^{-1}(x) = (1 + e^{-x})^{-1}$ heißt logistische Funktion.

Bemerkung 3.45. Es gilt also $\mathbb{E}[Y_i] = g^{-1}(\eta_i) = e^{\eta_i}/(1+e^{\eta_i})$, wobei die Funktion g^{-1} gerade die Verteilungsfunktion der standardisierten logistischen Verteilung ist (welche im Allgmeinen einen Mittelwerts- und einen Streuungsparameter besitzt). Das motiviert ein populäres Beispiel für eine nicht kanonische Linkfunktion: die Probit-Funktion $g(\lambda) = \Phi^{-1}(\lambda)$ mit der Verteilungsfunktion der Standardnormalverteilung Φ.

4 Testtheorie

4.1 Neyman-Pearson-Tests

Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit einer Zerlegung $\Theta = \Theta_0 \dot\cup \Theta_1$. Wir erinnern uns, dass ein (randomisierter) Test $\varphi\colon \mathcal{X} \to [0,1]$ der Hypothese $H_0 : \vartheta \in \Theta_0$ gegen die Alternative $H_1 : \vartheta \in \Theta_1$ das Niveau $\alpha \in [0,1]$ besitzt, falls $\mathbb{E}_\vartheta[\varphi] \leqslant \alpha$ für alle $\vartheta \in \Theta_0$. Der Test φ heißt unverfälscht, falls $\mathbb{E}_\vartheta[\varphi] \geqslant \alpha$ für alle $\vartheta \in \Theta_1$.

Beispiel 4.1. Es sei $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ eine mathematische Stichprobe mit unbekanntem $\mu \in \mathbb{R}$ und bekanntem $\sigma > 0$. Es soll die einseitige Hypothese $H_0 : \mu \leqslant \mu_0$ gegen $H_1 : \mu > \mu_0$ für ein vorgegebenes $\mu_0 \in \mathbb{R}$ getestet werden. Wir modellieren also $\mathcal{X} = \mathbb{R}^n, \mathcal{F} = \mathscr{B}_{\mathbb{R}^n}$, $\mathbb{P}_\mu = \mathcal{N}((\mu,\dots,\mu)^\top, \sigma^2 I_n)$ und $\Theta = \mathbb{R} = \Theta_0 \dot\cup \Theta_1$ mit $\Theta_0 = (0, \mu_0]$ und $\Theta_1 = (\mu_0, \infty)$. Zur Konstruktion eines einseitigen Gaußtests verwenden wir die Teststatistik

$$T(X_1,\dots,X_n) := \sqrt{n}(\overline{X}_n - \mu_0)/\sigma \overset{\mathbb{P}_\mu}{\sim} \mathcal{N}\Big(\frac{\sqrt{n}}{\sigma}(\mu-\mu_0), 1\Big).$$

Für ein vorgegebenes $\alpha > 0$ wählen wir das $(1-\alpha)$-Quantil $q_{1-\alpha}$ der Standardnormalverteilung und erhalten den Niveau-α-Test $\varphi(X_1,\dots,X_n) = \mathbb{1}_{\{T(X_1,\dots,X_n) > q_{1-\alpha}\}}$, da aus Monotoniegründen für alle $\mu \leqslant \mu_0$ gilt

$$\mathbb{P}_\mu(\varphi = 1) \leqslant \mathbb{P}_{\mu_0}(\varphi = 1) = \alpha.$$

Wie verhält sich φ unter der Alternative? Gibt es einen Niveau-α-Test φ' mit kleinerer Fehlerwahrscheinlichkeit 2.Art, d.h. $1 - \mathbb{E}_\mu[\varphi'] < 1 - \mathbb{E}_\mu[\varphi]$ für ein $\mu > \mu_0$?

Definition 4.2. Für das Testproblem $H_0 : \vartheta \in \Theta_0$ vs. $H_1 : \vartheta \in \Theta_1$ heißt $\varphi\colon \mathcal{X} \to [0,1]$ gleichmäßig bester Test zum Niveau $\alpha \in [0,1]$, falls φ das Niveau α besitzt sowie für alle anderen Niveau-α-Tests φ' die *Macht* bzw. *Teststärke* nicht größer als die von φ ist:

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

φ heißt gleichmäßig bester unverfälschter Test zum Niveau α, falls φ unverfälscht zum Niveau α ist sowie alle anderen unverfälschten Niveau-α-Tests φ' keine größere Macht besitzen.

Um gleichmäßig beste Tests zu studieren betrachten wir zunächst binäre statistische Modelle, d.h. $\Theta = \{0,1\}$ und die entsprechenden Testprobleme $H_0 : \vartheta = 0$ vs. $H_1 : \vartheta = 1$.

Bemerkung 4.3. Ein binäres statistisches Modell ist stets dominiert vom Maß $\mu := \mathbb{P}_0 + \mathbb{P}_1$, da aus $\mu(A) = 0$ auch $\mathbb{P}_0(A) = \mathbb{P}_1(A) = 0$ für alle $A \in \mathscr{F}$ folgt. Besitzen $\mathbb{P}_0, \mathbb{P}_1$ Zähl- oder Lebesguedichten f_0 bzw. f_1, dann gilt $\frac{\mathrm{d}\mathbb{P}_i}{\mathrm{d}\mu}(x) = \frac{f_i(x)}{f_0(x)+f_1(x)}, i = 0,1, x \in \mathcal{X}$. Beachte, dass hier der Nenner nur an Stellen 0 sein kann, an denen auch der Zähler 0 ist; in diesem Fall setzten wir den Quotienten auf 0.

Definition 4.4. Im statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta = \{0,1\}$ sei die Dichte $\mathbb{P}_i$ bzgl. $\mathbb{P}_0 + \mathbb{P}_1$ mit p_i für $i = 0,1$ bezeichnet. Ein Test der Form

$$\varphi(x) = \begin{cases} 1, & \text{falls } p_1(x) > k p_0(x), \\ 0, & \text{falls } p_1(x) < k p_0(x), \\ \gamma(x), & \text{falls } p_1(x) = k p_0(x) \end{cases}$$

mit kritischem Wert $k \in \mathbb{R}_+$ und $\gamma(x) \in [0,1]$ heißt Neyman-Pearson-Test.

Satz 4.5 (Neyman-Pearson-Lemma)**.**

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

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

Beweis. (i) Betrachte einen beliebigen Test φ' vom Niveau $\mathbb{E}_0[\varphi]$. Es gilt $p_1(x) \geqslant kp_0(x)$ für $x \in A := \{\varphi > \varphi_1\}$ wegen $\varphi(x) > 0$ sowie $p_1(x) \leqslant kp_0(x)$ für $x \in B := \{\varphi < \varphi'\}$ wegen $\varphi(x) < 1$. Mit der disjunkten Zerlegung $\mathcal{X} = A \dot\cup B \dot\cup \{\varphi = \varphi'\}$ erhalten wir

$$\begin{aligned}\mathbb{E}_1[\varphi] - \mathbb{E}_1[\varphi'] &= \int_{A\cup B} (\varphi - \varphi')p_1 \geqslant \int_A (\varphi - \varphi')kp_0 + \int_B (\varphi - \varphi')kp_0 \\ &= k\big(\mathbb{E}_0[\varphi] - \mathbb{E}_0[\varphi']\big) \geqslant 0.\end{aligned}$$

(ii) Wir zeigen zunächst, dass für $k := \inf\{r \geqslant 0 : \rho(r) \leqslant \alpha\}$ mit $\rho(r) := \mathbb{P}_0(p_1 > rp_0)$ gilt:

$$\mathbb{P}_0(p_1 \geqslant kp_0) \geqslant \alpha \quad \text{und} \quad \mathbb{P}_0(p_1 > kp_0) \leqslant \alpha \tag{4.1}$$

(k ist also das $(1-\alpha)$-Quantil von p_1/p_0 unter $\mathbb{P}_0$). Wegen $\mathbb{P}_0(p_0 = 0) = 0$ und der σ-Stetigkeit von $\mathbb{P}_0$ gilt $\lim_{r\to\infty} \rho(r) = 0$, sodass k endlich ist. Weiterhin ist $\rho(r) = 1 - \mathbb{P}_0(p_1/p_0 \leqslant r)$ monoton fallend und rechtsstetig, was aus den Eigenschaften der Verteilungsfunktion von p_1/p_0 folgt. Daher gilt $\rho(k) \leqslant \alpha$ und $\rho(r) > \alpha$ für $r < k$. Somit folgt aus der σ-Stetigkeit:

$$\alpha \leqslant \lim_{r\uparrow k} \rho(r) = \lim_{r\uparrow k} \mathbb{P}_0(p_1 > rp_0) = \mathbb{P}_0(p_1 \geqslant kp_0).$$

Wir haben also (4.1) gezeigt.

Wir setzen nun $\gamma := (\alpha - \mathbb{P}_0(p_1 > kp_0))/\mathbb{P}_0(p_1 = kp_0)$ bzw. $\gamma \in [0,1]$ beliebig, falls $\mathbb{P}_0(p_1 = kp_0) = 0$. Dann besitzt der resultierende Neyman-Pearson-Test φ das Niveau α:

$$\mathbb{E}_0[\varphi] = 1 \cdot \mathbb{P}_0(p_1(x) > kp_0(x)) + \gamma \cdot \mathbb{P}_0(p_1(x) = kp_0(x)) = \alpha. \qquad \square$$

Bemerkung 4.6. Es gilt auch umgekehrt, dass jeder gleichmäßig beste Test für eine einfach Hypothese gegen eine einfache Alternative f.s. die Form eines Neyman-Pearson-Tests besitzt (Übung ☑).

Die Neyman-Pearson Idee wollen wir nun auf zusammengesetzte Hypothesen ausweiten. Hierzu benötigen wir eine Strukturannahme.

Definition 4.7. Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein dominiertes Modell mit $\Theta \subseteq \mathbb{R}$ und Likelihoodfunktion $L(\vartheta, x)$ sowie T eine reellwertige Statistik. Dann besitzt die Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ einen monotonen Likelihoodquotienten (oder wachsenden Dichtequotienten) in T, falls:

(i) $\vartheta \neq \vartheta'$ impliziert $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$.

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

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

Lemma 4.8. *Ist $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ mit $\Theta \subseteq \mathbb{R}$ eine einparametrische Exponentialfamilie in $\eta(\vartheta)$ und T, so besitzt sie einen monotonen Likelihoodquotienten, sofern η streng monoton wächst.*

Beweis. Wir können den Dichtequotienten als $\frac{L(\vartheta',x)}{L(\vartheta,x)} = h(T(x), \vartheta, \vartheta')$ schreiben, wobei

$$h(t, \vartheta, \vartheta') = \exp\big((\eta(\vartheta') - \eta(\vartheta))t - \zeta(\vartheta') + \zeta(\vartheta)\big).$$

Offensichtlich ist h streng monoton wachsend in t für $\vartheta' > \vartheta$ wegen $\eta(\vartheta') > \eta(\vartheta)$. Die strenge Monotonie impliziert auch $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$ für $\vartheta \neq \vartheta'$. $\square$

Beispiel 4.9. Im Binomialmodell $X \sim Bin(n,p)$ mit $p \in (0,1)$ liegt eine Exponentialfamilie in $\eta(p) = \log(p/(1-p))$ und $T(x) = x$ vor. η wächst streng monoton, sodass dieses Modell einen monotonen Dichtequotienten in X besitzt. Man sieht dies auch direkt aus der Monotonie des Dichtequotienten in $x \in \{0,\dots,n\}$

$$\frac{\binom{n}{x}p^x(1-p)^{n-x}}{\binom{n}{x}q^x(1-q)^{n-x}} = \Big(\frac{p(1-q)}{q(1-p)}\Big)^x\Big(\frac{1-p}{1-q}\Big)^n \qquad \text{für } p > q.$$

Satz 4.10. *Die Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \Theta \subseteq \mathbb{R}$, besitze einen monotonen Likelihoodquotienten in T. Für $\alpha \in (0,1)$ und $\vartheta_0 \in \Theta$ gilt dann:*

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

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

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

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

Beweis. (i) Die Existenz von k, γ folgt wie im Neyman-Pearson-Lemma. Für beliebige $\vartheta_2 > \vartheta_1$ gilt wegen des monotonen Dichtequotienten

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } L(\vartheta_2,x) > h(k,\vartheta_1,\vartheta_2)L(\vartheta_1,x), \\ 0, & \text{falls } L(\vartheta_2,x) < h(k,\vartheta_1,\vartheta_2)L(\vartheta_1,x), \\ \gamma & \text{falls } L(\vartheta_2,x) = h(k,\vartheta_1,\vartheta_2)L(\vartheta_1,x). \end{cases}$$

Damit ist φ^* gleichmäßig bester Test von $H_0 : \vartheta = \vartheta_1$ vs. $H_1 : \vartheta = \vartheta_2$ zum vorgegeben Niveau $\mathbb{E}_{\vartheta_1}[\varphi^*]$. Insbesondere ist die Fehlerwahrscheinlichkeit zweiter Art $1 - \mathbb{E}_{\vartheta_2}(\varphi^*)$ minimal für alle $\vartheta_2 > \vartheta_0$ zum vorgegeben Niveau bei $\vartheta_1 = \vartheta_0$.

Für jeden Test φ mit kleinerer Fehlerwahrscheinlichkeit erster Art bei $\vartheta_1 < \vartheta_0$, d.h. $\mathbb{E}_{\vartheta_1}[\varphi] < \mathbb{E}_{\vartheta_1}[\varphi^*]$, gilt $\mathbb{E}_{\vartheta_0}[\varphi] < \mathbb{E}_{\vartheta_0}[\varphi^*]$: Sonst wäre $\widetilde{\varphi} = \kappa\varphi + (1-\kappa)$ mit $\kappa = \frac{1-\mathbb{E}_{\vartheta_1}[\varphi^*]}{1-\mathbb{E}_{\vartheta_1}[\varphi]} \in (0,1)$ ebenfalls Test für $H_0 : \vartheta = \vartheta_1$ vs. $H_1 : \vartheta = \vartheta_0$ zum Niveau

$$\mathbb{E}_{\vartheta_1}[\widetilde{\varphi}] = \frac{1-\mathbb{E}_{\vartheta_1}[\varphi^*]}{1-\mathbb{E}_{\vartheta_1}[\varphi]} \cdot \mathbb{E}_{\vartheta_1}[\varphi] + \frac{\mathbb{E}_{\vartheta_1}[\varphi^*] - \mathbb{E}_{\vartheta_1}[\varphi]}{1-\mathbb{E}_{\vartheta_1}[\varphi]} = \mathbb{E}_{\vartheta_1}[\varphi^*]$$

und $\widetilde{\varphi}$ wäre besser als φ^* wegen $\mathbb{E}_{\vartheta_0}[\widetilde{\varphi}] > \kappa\mathbb{E}_{\vartheta_0}[\varphi] + (1-\kappa)\mathbb{E}_{\vartheta_0}[\varphi] \geqslant \mathbb{E}_{\vartheta_0}[\varphi^*]$. Demnach gilt $\mathbb{E}_{\vartheta_1}[\varphi] \geqslant \mathbb{E}_{\vartheta_1}[\varphi^*]$ für jeden Test φ mit $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$.

(ii) Da jeder Test φ auf $H_0 : \vartheta = \vartheta_0$ zum Niveau α durch $\widetilde{\varphi} = \kappa\varphi + (1-\kappa)$ mit $\kappa = \frac{1-\alpha}{1-\mathbb{E}_{\vartheta_0}[\varphi]}$ zu einem besseren Test mit $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$ gemacht werden kann, bleibt nur noch zu zeigen, dass φ^* das Niveau α für $H_0 : \vartheta \leqslant \vartheta_0$ einhält. In (i) haben wir gesehen, dass φ^* auch bester Test für $H_0 : \vartheta = \vartheta_1$ mit $\vartheta_1 < \vartheta_0$ gegen $H_1 : \vartheta = \vartheta_0$ ist, sodass im Vergleich zum konstanten Test $\varphi = \mathbb{E}_{\vartheta_1}[\varphi^*]$ folgt, dass $\mathbb{E}_{\vartheta_0}[\varphi^*] \geqslant \mathbb{E}_{\vartheta_0}[\varphi] = \mathbb{E}_{\vartheta_1}[\varphi^*]$ gilt. Wir schließen $\mathbb{E}_{\vartheta_1}[\varphi^*] \leqslant \alpha$ für alle $\vartheta_1 < \vartheta_0$. □

Beispiel 4.11. Der einseitige Gaußtest aus Beispiel 4.1 ist ein gleichmäßig bester Test, da $(\mathcal{N}((\mu,\dots,\mu)^\top, \sigma^2 I_n))_\mu$ eine einparametrische Exponentialfamilie in $\eta(\mu) = \mu/\sigma^2$ und $T(x) = \sum_{i=1}^n x_i$ bildet und damit einen monotonen Dichtequotienten in besitzt.

Bemerkung 4.12.

(i) Die Gütefunktion $\beta_{\varphi^*}(\vartheta) = \mathbb{E}_\vartheta[\varphi^*]$ ist sogar streng monoton wachsend für alle ϑ mit $\beta_{\varphi^*}(\vartheta) \in (0,1)$, wie ein ähnlicher Beweis ergibt.

(ii) Im Beweis wurde eine Konvexkombination $\widetilde{\varphi}$ von Tests betrachtet. Dieses Argument lässt sich gut geometrisch darstellen. Allgemein betrachte bei einem binären Modell mit $(\mathbb{P}_0, \mathbb{P}_1)$ die Menge $C := \{(\mathbb{E}_0[\varphi], \mathbb{E}_1[\varphi]) : \varphi \text{ Test}\} \subseteq [0,1]^2$. Diese ist konvex (Menge der Tests ist konvex), abgeschlossen (siehe Lehmann and Romano (2005, Seite 62)) und enthält die Diagonale (betrachte konstante Tests). Neyman-Pearson-Tests entsprechen dann gerade der oben Begrenzungskurve von C.

Schließlich wollen wir noch klären, was uns die Neyman-Pearson-Theorie für zweiseitige Testprobleme lehrt.

Satz 4.13 (Verallgemeinertes Neyman-Pearson-Lemma)**.** *Es seien $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta = \{0,1\}$ ein (binäres) statistisches Modell, p_0, p_1 die entsprechenden Dichten und $T \in \mathcal{L}^1(\mathbb{P}_0)$ eine reellwertige Statistik. Ein Test der Form*

$$\varphi(x) = \begin{cases} 1, & \text{falls } p_1(x) > kp_0(x) + lT(x)p_0(x), \\ 0, & \text{falls } p_1(x) < kp_0(x) + lT(x)p_0(x), \\ \gamma, & \text{falls } p_1(x) = kp_0(x) + lT(x)p_0(x) \end{cases}$$

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

$$\mathbb{E}_0[\varphi] = \alpha \qquad \text{und} \qquad \mathbb{E}_0[T\varphi] = \alpha\mathbb{E}_0[T]$$

erfüllt, maximiert die Güte $\mathbb{E}_1[\varphi]$ in der Menge aller Test, die diese Nebenbedingung erfüllen.

Beweis. Übung ⍁ □

Satz 4.14. *$(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ sei eine einparametrische Exponentialfamilie in $\eta(\vartheta)$ und T. $\Theta \subseteq \mathbb{R}$ sei offen, $\vartheta_0 \in \Theta$ und η sei streng monoton (wachsend oder fallend) und stetig differenzierbar um ϑ_0 mit $\eta'(\vartheta_0) \neq 0$.*

(i) Für jeden Test φ ist die Gütefunktion $\beta_\varphi(\vartheta) = \mathbb{E}_\vartheta[\varphi]$ in ϑ_0 differenzierbar und es gilt

$$\beta'_\varphi(\vartheta_0) = 0 \qquad \Longleftrightarrow \qquad \mathbb{E}_{\vartheta_0}[\varphi T] = \mathbb{E}_{\vartheta_0}[\varphi]\mathbb{E}_{\vartheta_0}[T]. \tag{4.2}$$

Insbesondere erfüllt jeder unverfälschte Niveau-α-Test φ für das zweiseitige Testproblem $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta \neq \vartheta_0$ die Bedingungen $\mathbb{E}_{\vartheta_0}[\varphi] = \alpha$ und $\mathbb{E}_{\vartheta_0}[T\varphi] = \alpha\mathbb{E}_{\vartheta_0}[T]$.

(ii) Für $\alpha \in (0,1), k_1 < k_2$ und $\gamma_1, \gamma_2 \in [0,1]$ erfülle der Test

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } T(x) > k_1 \text{ oder } T(x) > k_2, \\ 0, & \text{falls } T(x) \in (k_1, k_2), \\ \gamma, & \text{falls } T(x) \in \{k_1, k_2\} \end{cases}$$

die Nebenbedingungen

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

Dann ist φ^* gleichmäßig bester unverfälschter Test zum Niveau α für das zweiseitige Testproblem $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta \neq \vartheta_0$.

Beweis. (i) Es gilt $\beta_\varphi(\vartheta) = \int_\mathcal{X} \varphi^*(x)h(x)e^{\eta(\vartheta)T(x)-\zeta(\vartheta)}\mu(\mathrm{d}x)$ und mit dominierter Konvergenz und $\zeta'(\vartheta) = \eta'(\vartheta)\mathbb{E}_\vartheta[T]$ (siehe Lemma 3.30) folgt in einer ϑ_0-Umgebung

$$\begin{aligned} \beta'_\varphi(\vartheta) &= \int_\mathcal{X} \varphi(x)h(x)\big(\eta'(\vartheta)T(x) - \zeta'(\vartheta)\big)e^{\eta(\vartheta)T(x)-\zeta(\vartheta)}\mu(\mathrm{d}x) \\ &= \eta'(\vartheta)\int_\mathcal{X} \varphi(x)h(x)\big(T(x) - \mathbb{E}_\vartheta[T]\big)e^{\eta(\vartheta)T(x)-\zeta(\vartheta)}\mu(\mathrm{d}x) \\ &= \eta'(\vartheta)\big(\mathbb{E}_\vartheta[\varphi T] - \mathbb{E}_\vartheta[T]\mathbb{E}_\vartheta[\varphi]\big). \end{aligned}$$

Aus $\eta'(\vartheta_0) \neq 0$ ergibt sich (4.2). Da für einen unverfälschten Test φ die Gütefunktion β_φ in ϑ_0 eine Minimalstelle haben muss, ergibt sich der Zusatz.

(ii) Wir zeigen, dass φ^* für $\mathbb{P}_1 = \mathbb{P}_{\vartheta_1} \neq \mathbb{P}_{\vartheta_0} = \mathbb{P}_0$ die Form aus dem verallgemeinerten Neyman-Pearson-Lemma besitzt. Mit $a = \eta(\vartheta_1) - \eta(\vartheta_0) \neq 0$ und $b = \zeta(\vartheta_0) - \zeta(\vartheta_1)$ gilt

$$L(\vartheta_1, x) > kL(\vartheta_0, x) + lT(x)L(\vartheta_0, x) \quad \iff \quad \exp\big(aT(x) + b\big) > lT(x) + k.$$

Wähle nun $k, l \in \mathbb{R}$ so, dass die Gerade $t \mapsto lt + k$ die streng konvexe Funktion $t \mapsto \exp(at + b)$ genau bei $t \in \{k_1, k_2\}$ schneidet. Dann gilt

$$L(\vartheta_1, x) > kL(\vartheta_0, x) + lT(x)L(\vartheta_0, x) \quad \iff \quad T(x) \notin [k_1, k_2] \quad \iff \quad \varphi^*(x) = 1.$$

Analoge Äquivalenzen zeigen, dass φ^* die gewünscht Form besitzt. Für jeden Test φ, der die Nebenbedingungen erfüllt, gilt also $\mathbb{E}_{\vartheta_1}[\varphi^*] \geqslant \mathbb{E}_{\vartheta_1}[\varphi]$ für $\vartheta_1 \neq \vartheta_0$ nach Satz 4.13. Zusammen mit (i) ergibt sich, dass φ^* gleichmäßig bester Test in der Klasse ist. □

Beispiel 4.15. Wir betrachten wieder eine mathematische Stichprobe $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ mit unbekanntem $\mu \in \mathbb{R}$ und bekanntem $\sigma > 0$ und nun das zweiseitige Testproblem $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. Da die zugrundeliegende Exponentialfamilie $(\mathcal{N}((\mu, \dots, \mu)^\top, \sigma^2 I_n))_\mu$ in $\eta(\mu) = \mu/\sigma^2$ und $T(x) = \sum_{i=1}^n x_i$ die Bedingung $\eta'(\mu) = \sigma^{-2} > 0$ erfüllt, können wir einen gleichmäßig besten unverfälschten Test gemäß des vorangegangenen Satzes bestimmen.

Aus Symmetriegründen wählen wir $k_1 = n\mu_0 - k, k_2 = n\mu_0 + k$ und verzichten wegen der stetigen Verteilung auf eine Randomisierung, sodass $\varphi^*(x) = \mathbb{1}_{\{|T(x) - n\mu_o| > k\}}$ gilt. Wir erhalten mit $Z = \sum_{i=1}^n (X_i - \mu_0) \sim \mathcal{N}(0, n\sigma^2)$ unter $\mathbb{P}_{\mu_0}$:

$$\mathbb{E}_{\mu_0}[\varphi^* T] = \mathbb{E}\big[(n\mu_0 + Z)\mathbb{1}_{\{|Z| > k\}}\big] = \mathbb{E}\big[n\mu_0 \mathbb{1}_{\{|Z| > k\}}\big] = \mathbb{E}_{\mu_0}[T]\mathbb{E}_{\mu_0}[\varphi^*].$$

Wählt man also $k = \sigma\sqrt{n} q_{1-\alpha/2}$ mit dem $(1 - \alpha/2)$-Quantil von $\mathcal{N}(0, 1)$, so gilt $\mathbb{E}_{\vartheta_0}[\varphi^*] = \alpha$ und der zweiseitige Gaußtest φ^* ist tatsächlich ein gleichmäßig bester unverfälschter Test.

4.2 Likelihood-Quotienten- und χ^2-Test

Inspiriert vom Neyman-Pearson-Test für einfache Hypothesen und Alternativen definieren wir:

Definition 4.16. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein dominiertes statistisches Modell mit Likelihoodfunktion L. Ein Test für die Hypothese $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$ von der Form

$$\varphi(x) = \begin{cases} 1, & \text{falls } \Lambda(x) > k, \\ 0, & \text{falls } \Lambda(x) < k, \\ \gamma(x), & \text{falls } \Lambda(x) = k \end{cases} \qquad \text{mit} \quad \Lambda(x) := \frac{\sup_{\vartheta \in \Theta_1} L(\vartheta, x)}{\sup_{\vartheta \in \Theta_0} L(\vartheta, x)} \in [0, \infty]$$

und $k \in \mathbb{R}_+, \gamma(x) \in [0, 1]$ heißt Likelihood-Quotienten-Test.

Bemerkung 4.17. Häufig liegt Θ_1 dicht in Θ und die Likelihoodfunktion ist stetig in ϑ. Dann gilt $\sup_{\vartheta \in \Theta_1} L(\vartheta, x) = \sup_{\vartheta \in \Theta} L(\vartheta, x) = L(\widehat{\vartheta}, x)$ mit einem Maximum-Likelihood-Schätzer $\widehat{\vartheta}$. Das ist auch Grundlage der asymptotischen Theorie.

Beispiel 4.18. Im Fall einer natürlichen Exponentialfamilie erhalten wir für Θ_1 dicht in Θ:

$$\Lambda(x) = \inf_{\vartheta_0 \in \Theta_0} \exp\big(\langle \widehat{\vartheta} - \vartheta_0, T(x) \rangle - \zeta(\widehat{\vartheta}) + \zeta(\vartheta_0)\big).$$

Falls der ML-Schätzer $\widehat{\vartheta}(x)$ im Inneren von Θ liegt, so folgt aus Satz 3.5 $\mathbb{E}_{\widehat{\vartheta}(x)}[T] = \nabla_\vartheta \zeta(\widehat{\vartheta}(x)) = T(x)$ (wobei die zweite Gleichheit gerade die notwendige Bedingung an den MLE bei der Maximierung ist). Damit folgt für den Kullback-Leibler Abstand $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} \log(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x))\mathbb{P}(\mathrm{d}x)$ für $\mathbb{P} \ll \mathbb{Q}$ (und sonst gleich $+\infty$)

$$\log(\Lambda(x)) = \inf_{\vartheta_0 \in \Theta_0} \mathrm{KL}(\mathbb{P}_{\widehat{\vartheta}}, \mathbb{P}_{\vartheta_0}).$$

Die Likelihood-Quotienten-Statistik Λ misst hier also in natürlicher Weise den Abstand der Hypothesenmenge $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta_0}$ zu der zu $\widehat{\vartheta} \in \Theta$ gehörenden Verteilung.

Lemma 4.19. *In der Situation von Satz 4.10 über beste einseitige Tests existiert ein Likelihood-Quotienten-Test, der mit dem dort angegebenen besten Test übereinstimmt.*

Beweis. Wir schreiben

$$\begin{aligned}\frac{\sup_{\vartheta\in\Theta_1} L(\vartheta,x)}{\sup_{\vartheta\in\Theta_0} L(\vartheta,x)} &= \sup_{\vartheta>\vartheta_0} \frac{L(\vartheta,x)}{L(\vartheta_0,x)} \inf_{\vartheta'\leqslant\vartheta_0} \frac{L(\vartheta_0,x)}{L(\vartheta',x)} \\ &= \sup_{\vartheta>\vartheta_0} h(T(x),\vartheta_0,\vartheta) \inf_{\vartheta'\leqslant\vartheta_0} h(T(x),\vartheta',\vartheta_0).\end{aligned}$$

Die beiden Funktionen h sind wachsend in T und damit auch das Supremum, das Infimum und deren Produkt. Also gilt für einen Likelihood-Quotienten-Test φ sowohl $\varphi(x) = 1$ für $T(x) > k'$ als auch $\varphi(x) = 0$ für $T(x) < k'$ für ein geeignetes $k' \in \mathbb{R}$. □

Bemerkung 4.20. Im Fall des zweiseitigen Testproblems aus Satz 4.14 führt der Likelihood-Quotiententest zwar auf einen Test mit Ablehnbereich $\{T(x) \notin [k_1, k_2]\}$, allerdings ist er im Allgemeinen nicht mehr unverfälscht, wie folgendes Gegenbeispiel lehrt: $X \sim Poiss(\vartheta)$ führt auf einen Ablehnbereich $\{X(\log(X/\vartheta_0) - 1) > k'\}$, was für $k' > 0$ einem einseitigen Ablehnbereich $\{X > k''\}$ entspricht. Hingegen sind im Fall der Normalverteilung ein- und zweiseitige Gaußtests Likelihood-Quotienten-Test.

Satz 4.21. *Es mögen die Voraussetzung aus Satz 3.34 gelten. Es sei $0 \in \Theta^\circ$ ein innerer Punkt von Θ und die Hypothesenmenge $\Theta_0 := \{(\vartheta_1,\dots,\vartheta_r,0,\dots,0) \in \Theta : \vartheta_1,\dots,\vartheta_r \in \mathbb{R}\}$ sei ein r-dimensionaler Unterraum, $0 \leqslant r < d$, mit $\Theta_0 = \{0\}$, falls $r = 0$. Weiterhin nehmen wir an, dass Maximumlikelihood-Schätzer $\widehat{\vartheta}_n$ und $\widehat{\vartheta}_n^0$ in den Parametermengen Θ bzw. Θ_0 existieren. Dann gilt für die Fitted-Loglikelihood-Statistik*

$$\lambda_n(x) := \sup_{\vartheta\in\Theta} \sum_{i=1}^n \log L(\vartheta, x_i) - \sup_{\vartheta\in\Theta_0} \sum_{i=1}^n \log L(\vartheta, x_i)$$

unter jedem $\mathbb{P}_{\vartheta_0}$ mit $\vartheta_0 \in \Theta_0 \cap \Theta^\circ$ die Konvergenz $2\lambda_n \xrightarrow{d} \chi^2(d-r)$. Insbesondere besitzt der Likelihood-Quotienten-Test

$$\varphi(x) = \mathbb{1}_{\{\lambda_n(x) > q_{k-r,1-\alpha}/2\}} = \mathbb{1}_{\{\Lambda_n(x) > \exp(q_{k-r,1-\alpha}/2)\}}, \qquad \Lambda_n(x) := \frac{\sup_{\vartheta\in\Theta_1} \prod_{i=1}^n L(\vartheta, x_i)}{\sup_{\vartheta\in\Theta_0} \prod_{i=1}^n L(\vartheta, x_i)},$$

mit dem $(1-\alpha)$-Quantil $q_{d-r,1-\alpha}$ der $\chi^2(d-r)$-Verteilung auf $\Theta_0 \cap \Theta^\circ$ asymptotisch das Niveau $\alpha \in (0,1)$.

Beweis. Im Folgenden sei $\Pi_r \colon \mathbb{R}^d \to \mathbb{R}^r$ die Koordinatenprojektion auf die ersten r Koordinaten. Aus den Beweisen der Sätze 3.34 und 3.36 folgt insbesondere (siehe (3.5) und (3.7)) für die Loglikelihood $\ell_n(\vartheta, x) = \sum_{i=1}^n \log L(\vartheta, x_i)$ und Scorefunktion $U_\vartheta^n = U_\vartheta^n(X) = \nabla_\vartheta \ell_n(\vartheta, X)$

$$n^{1/2} I_1(\vartheta_0)(\widehat{\vartheta}_n - \vartheta_0) = n^{-1/2} U_{\vartheta_0}^n + o_p(1)$$

und

$$2(\ell_n(\widehat{\vartheta}_n, X) - \ell_n(\vartheta_0, X)) = -n\langle \widehat{\vartheta}_n - \vartheta_0, I(\vartheta_0)(\widehat{\vartheta}_n - \vartheta_0)\rangle + o_p(1).$$

Wir erhalten

$$2(\ell_n(\widehat{\vartheta}_n, X) - \ell_n(\vartheta_0, X)) = -n^{-1}\langle I(\vartheta_0)^{-1} U_{\vartheta_0}^n, U_{\vartheta_0}^n\rangle + o_p(1).$$

Ganz analog erhalten wir für die kleinere Parametermenge Θ_0,h3 indem man formal $\Theta_0 \subseteq \mathbb{R}^d$ mit $\Pi_r\Theta_0 \subseteq \mathbb{R}^r$ identifiziert:

$$2(\ell_n(\widehat{\vartheta}_n^0, X) - \ell_n(\vartheta_0, X)) = -n^{-1}\langle I_0(\vartheta_0)^{-1} \widetilde{U}_{\vartheta_0}^n, \widetilde{U}_{\vartheta_0}^n\rangle + o_p(1),$$

wobei $\widetilde{U}^n_{\vartheta_0} = \Pi_r U^n_{\vartheta_0}$ den Gradienten von $\ell_n(\vartheta, X)$ als Funktion in den ersten r Argumenten und $I_0(\vartheta_0) = \Pi_r I(\vartheta_0)\Pi_r^\top$ die $r\times r$-Fischer-Informationsmatrix bzgl. dieser r Parameterwerte bezeichne. Im ausgearteten Fall $r = 0$ setzen wir einfach $\widehat{\vartheta}_0 = 0$. Insgesamt erhalten wir

$$\begin{aligned}2\lambda_n(x) &= 2(\ell_n(\widehat{\vartheta}_n) - \ell_n(\widehat{\vartheta}^0_n))\\ &= n^{-1}\langle I(\vartheta_0)^{-1}U^n_{\vartheta_0}, U^n_{\vartheta_0}\rangle - n^{-1}\langle I_0(\vartheta_0)^{-1}\widetilde{U}^n_{\vartheta_0}, \widetilde{U}^n_{\vartheta_0}\rangle + o_p(1)\\ &= n^{-1}\big\langle \big(I(\vartheta_0)^{-1} - \Pi_r^\top I_0(\vartheta_0)^{-1}\Pi_r\big)U^n_{\vartheta_0}, U^n_{\vartheta_0}\big\rangle + o_p(1).\end{aligned}$$

Wie wir im Beweis von Satz 3.36 gesehen haben, gilt $n^{-1/2}U^n_{\vartheta_0} \xrightarrow{d} \mathcal{N}(0, I(\vartheta_0))$, sodass Slutskys Lemma für ein $Z \sim \mathcal{N}(0, I_d)$ ergibt:

$$2\lambda_n \xrightarrow{d} \big\langle \big(I_d - I(\vartheta_0)^{1/2}\Pi_r^\top I_0(\vartheta_0)^{-1}\Pi_r I(\vartheta_0)^{1/2}\big)Z, Z\big\rangle.$$

Die Matrix $M := I(\vartheta_0)^{1/2}\Pi_r^\top I_0(\vartheta_0)^{-1}\Pi_r I(\vartheta_0)^{1/2}$ ist symmetrisch und beschreibt wegen $M^2 = M$ eine Orthogonalprojektion. Als Spur erhalten wir

$$\operatorname{tr}(M) = \operatorname{tr}\big(I(\vartheta_0)\Pi_r^\top I_0(\vartheta_0)^{-1}\Pi_r\big) = \operatorname{tr}\big(\Pi_r I(\vartheta_0)\Pi_r^\top I_0(\vartheta_0)^{-1}\big) = \operatorname{tr}(I_r) = r.$$

Damit besitzt M den Rang r und $I_d - M$ ist eine Orthogonalprojektion vom Rang $d - r$. Daraus folgt, dass $\langle (I_r - M)Z, Z\rangle \sim \chi^2(d-r)$ (siehe Satz 2.16).

Abschließend bemerken wir, dass aus der Stetigkeit von ℓ_n für die Likelihood-Quotienten-Statistik folgt

$$\log \Lambda_n(x) = \log\Big(\frac{\sup_{\vartheta\in\Theta}\prod_{i=1}^n L(\vartheta, x_i)}{\sup_{\vartheta\in\Theta_0}\prod_{i=1}^n L(\vartheta, x_i)}\Big) = \lambda_n(x)$$

und somit aus der Monotonie des Logarithmus der Likelihood-Quotienten-Test einen Ablehnbereich der Form $\{\lambda_n > k\}$ besitzt. Beachte, dass eine Randomisierung asymptotisch vernachlässigbar ist. □

Bemerkung 4.22.

(i) Für Anwendungen äußert nützlich ist, dass die asymptotische Verteilung von λ_n unabhängig von $\vartheta_0 \in \Theta_0$ ist. Der Likelihood-Quotienten-Test ist damit asymptotisch verteilungsfrei.

(ii) Zwei weitere wichtige asymptotische Likelihood-Tests einer einfache Hypothese $H_0 : \vartheta = \vartheta_0$ sind der Wald-Test und der Score-Test. Ersterer verwendet die Teststatistik $W_n = n\langle I(\widehat{\vartheta}_n)(\widehat{\vartheta}_n - \vartheta_0), \widehat{\vartheta}_n - \vartheta_0\rangle$ mit dem MLE $\widehat{\vartheta}_n$. Unter den Bedingungen von Satz 3.34 ist W_n ebenfalls asymptotisch $\chi^2(d)$ verteilt und der Wald-Test ist von der Form $\varphi = \mathbb{1}_{\{W_n > k\}}$. Raos Score-Test ist gegeben durch $\varphi = \mathbb{1}_{\{R_n > k\}}$ mit $R_n = n^{-1}\langle I(\vartheta_0)^{-1}U^n_{\vartheta_0}, U^n_{\vartheta_0}\rangle$, wobei R_n gerade die Approximation von $2\lambda_n$ aus obigem Beweis ist und somit auch $\chi^2(d)$-verteilt.

Beispiel 4.23. Wir beobachten einen Zufallsvektor $N = (N_1, \dots, N_d)$, der gemäß einer Multinomialverteilung mit Parametern n und $p = (p_1, \dots, p_d)$ verteilt ist. Dann ist N eine suffiziente Statistik für n unabhängige Beobachtungen einer Multinomialverteilung mit Parametern $(1, p)$, sodass obige Asymptotik für $n \to \infty$ angewendet werden kann. Da $\sum_{i=1}^d p_i = 1$ gilt, betrachten wir die Parametermenge $\Theta = \{p \in [0,1]^{d-1} : \sum_{i=1}^{d-1} p_i \leqslant 1\} \subseteq \mathbb{R}^{d-1}$. Für ein $p^0 \in \Theta$ betrachten wir das Testproblem

$$H_0 : p = p^0 \qquad \text{vs.} \qquad H_1 : p \neq p^0.$$

Da $\widehat{p}_n = N/n$ der MLE von p ist, erhalten wir

$$\lambda_n = \log\Big(\frac{\binom{n}{N_1\cdots N_d}(N_1/n)^{N_1}\cdots(N_d/n)^{N_d}}{\binom{n}{N_1\cdots N_d}(p^0_1)^{N_1}\cdots(p^0_d)^{N_d}}\Big) = \sum_{i=1}^d N_i \log\Big(\frac{N_i}{p^0_i n}\Big).$$

Wir wenden nun die Entwicklung für $\frac{h}{x} \to 0$

$$(x+h)\log\big(\frac{x+h}{x}\big) = (x+h)\big(\frac{h}{x} - \frac{h}{2x}(1+o(1))\big) = h + \frac{h^2}{2x} + o\big(\frac{h^2}{x}\big)$$

auf $x = p_i^0 n$ und $h = N_i - np_i^0$ an. Aus $\sum_{i=1}^d N_i = n = \sum_{i=1}^d np_i^0$ und $\mathbb{E}_{p^0}[(N_i - np_i^0)^2/np_i^0] = (1-p_i^0) \leqslant 1$ folgt damit

$$\begin{aligned} 2\lambda_n &= 2\sum_{i=1}^d \Big(N_i - np_i^0 + \frac{(N_i - np_i^0)^2}{2p_i^0 n} + o\Big(\frac{(N_i - np_i^0)^2}{2p_i^0 n}\Big)\Big) \\ &= \sum_{i=1}^d \frac{(N_i - np_i^0)^2}{p_i^0 n} + o_p(1). \end{aligned}$$

Somit konvergiert auch $\sum_{i=1}^d \frac{(N_i - np_i^0)^2}{p_i^0 n}$ unter H_0 gegen eine $\chi^2(d-1)$-Verteilung.

In diesem konkreten Beispiel erhalten wir also einen weiteren, in Praxis wichtigen, Test.

Definition 4.24. Bei Beobachtungen eines Zufallsvektors $N = (N_1, \dots, N_d)$, der gemäß einer Multinomialverteilung mit Parametern n und $p = (p_1, \dots, p_d)$ verteilt ist, heißt

$$\chi_n^2 := \sum_{i=1}^d \frac{(N_i - np_i^0)^2}{p_i^0 n}$$

Pearsons χ^2-Statistik für die Hypothese $H_0 : p = p^0$ und $\varphi = \mathbb{1}_{\{\chi_n^2 > q_{k-1,1-\alpha}\}}$ χ^2-Anpassungstest oder kurz χ^2-Test mit dem $(1-\alpha)$-Quantil $q_{k-1,1-\alpha}$ der $\chi^2(k-1)$-Verteilung.

Korollar 4.25. *Der χ^2-Test besitzt unter $H_0 : p = p^0$ asymptotisch das Niveau $\alpha \in (0,1)$.*

Bemerkung 4.26. Es gibt vielfältige Verallgemeinerungen dieses Tests. Insbesondere bei Hypothesen H_0 der Dimension $0 < r < d-1$ wird p^0 durch einen MLE $\widehat{p}^0$ ersetzt und es ergibt sich eine $\chi^2(d-r-1)$-Verteilung. Der χ^2-Test dient häufig als *Goodness-of-fit-Test*, bspw. können Zufallszahlen darauf getestet werden, ob jede Ziffer mit gleicher Wahrscheinlichkeit auftritt, was dem Fall $k = 10$ und $p_1^0 = \cdots = p_{10}^0 = 0,1$ mit Ziffernlänge n entspricht. Diese Idee wird insbesondere in der Betriebsprüfung angewendet: Bucht ein Unternehmer fingierte Zahlen statt der tatsächlichen, wird er vermutlich häufiger seine „Lieblingsziffer“ verwenden. Diese Abweichung kann durch einen χ^2-Test aufgedeckt werden und zum Besuch der Steuerfahndung führen.

Beispiel 4.27. Eine klassische Anwendung des χ^2-Tests ist die Überprüfung von Mendels Erbsendaten. Bei einer Erbsensorte gibt es die Ausprägungen rund (A) oder kantig (a) sowie gelb (B) oder grün (b). Die Merkmale „rund“ und „gelb“ sind der Theorie nach dominant, sodass die Genotypen AA, Aa, aA zum Phänotyp „rund“ und der Genotyp aa zum Phänotyp „kantig“ führt. Ebsnso ist gelb dominant. Betrachtet man nun Nachkommen des heterozygoten Genotyps AaBb, so sollten die vier Phänotypen im Verhältnis 9:3:3:1 auftreten. In Mendels Daten (von 1865) waren von $n = 556$ Erbsen 315 AB, 101 aB, 108 Ab, 21 ab.

Als natürliches Modell ergibt sich unter der Hypothese eine Multinomialverteilung mit Parametern n und $p^0 = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{6})$. Als χ^2-Statistik erhalten wir

$$\chi_n^2 = \frac{(315 - 312,75)^2}{312,75} + \frac{(101 - 104,25)^2}{104,25} + \frac{(108 - 104,25)^2}{104,25} + \frac{(21 - 34,75)^2}{34,75} \approx 0,47.$$

Der sogenannte p-Wert des χ^2-Tests bei diesen Daten beträgt $0,9254$ ($\mathbb{P}(X > 0,47) \approx 0,9254$) für $X \sim \chi^2(3)$), d.h. dass der χ^2-Test die Nullhypothese zu jedem Niveau $\alpha \leqslant 0,9254$ akzeptiert hätte! Diese beeindruckende Güte der Daten hat andererseits zum Verdacht der Datenmanipulation geführt.

Literatur

Georgii, H.-O. (2007). *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik.* de Gruyter, Berlin.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning (with Applications in R).* Springer, New York.

Klenke, A. (2006). *Wahrscheinlichkeitstheorie.* Springer.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation.* Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses.* Springer Texts in Statistics. Springer, New York, third edition.

Shao, J. (2003). *Mathematical Statistics.* Springer, New York.

Wasserman, L. (2004). *All of statistics: a concise course in statistical inference.* Springer, New York.

Witting, H. (1985). *Mathematische Statistik. I.* B. G. Teubner, Stuttgart. Parametrische Verfahren bei festem Stichprobenumfang.

Witting, H. and Müller-Funk, U. (1995). *Mathematische Statistik. II.* B. G. Teubner, Stuttgart. Asymptotische Statistik: parametrische Modelle und nichtparametrische Funktionale.