Inhalt

2 Klassische Schätz- und Testtheorie

Grundmodell:

Die Stichprobe $\mathbf{X} = (X_1, \ldots, X_n)$ besitzt die Verteilung $\mathbb{P} \in \mathcal{P} = \{\mathbb{P}_\theta : \boldsymbol{\theta} \in \Theta\}, \Theta \subseteq \mathbb{R}^k$, wobei

- $\boldsymbol{\theta}$: k-dimensionaler Parameter
- Θ: Parameterraum
- $k < n$, oft $k \ll n$, mit $\dim(\boldsymbol{\theta}) = k$ fest für asymptotische ($n \to \infty$)-Betrachtungen.
- In der Regel vorausgesetzt: Es existiert Dichte

  $$f(\mathbf{x}|\boldsymbol{\theta}) = f(x_1, \ldots, x_n|\boldsymbol{\theta}) \text{ zu } \mathbb{P}_\theta,$$

  so dass man analog schreiben kann:

  $$\mathcal{P} = \{f(\mathbf{x}|\boldsymbol{\theta}) : \ \boldsymbol{\theta} \in \Theta\}.$$

2 Klassische Schätz- und Testtheorie

- Klassische Schätz- und Testtheorie für finite (d.h. für festen Stichprobenumfang n) i.i.d.-Stichprobe von besonderer Relevanz; es gilt:

$$f(\mathbf{x}|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta}) \cdot \ldots \cdot f(x_n|\boldsymbol{\theta}).$$

- Viele Begriffe, insbesondere der Schätztheorie, jedoch von genereller Bedeutung.
- Literatur: Lehmann & Casella (1998), Lehmann & Romano (2005), Rüger (1999, 2002) Band I+II

Definition 2.1 (Statistik)

Eine Statistik ist eine messbare Funktion

$$\mathbf{T} : \begin{cases} \mathcal{X} & \longrightarrow & \mathbb{R}^l \\ \mathbf{x} & \longmapsto & \mathbf{T}(\mathbf{x}). \end{cases}$$

Normalerweise ist $l < n$, da mit der Statistik $\mathbf{T}$ eine Dimensionsreduktion erzielt werden soll.

Beispiel 2.1

- $\mathbf{T}$ Schätzfunktion
- T Teststatistik

Gesucht: Punkt- oder Bereichsschätzung für $\boldsymbol{\theta}$ oder einen transformierten Parametervektor $\boldsymbol{\tau}(\boldsymbol{\theta})$.

Beispiel 2.2
$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ mit $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. Hier könnte $\tau(\boldsymbol{\theta}) = \mu$ sein (d.h. σ^2 ist Nuisance-Parameter) oder $\tau(\boldsymbol{\theta}) = 1/\sigma^2$ (d.h. die Präzision ist von Interesse).

Definition 2.2 (Punktschätzung, Schätzer, Schätzfunktion)

Sei

$$\mathbf{T} : \begin{cases} \mathcal{X} & \longrightarrow & \Theta \subseteq \mathbb{R}^k \\ \mathbf{x} & \longmapsto & \mathbf{T}(\mathbf{x}) \end{cases}$$

eine messbare Abbildung, die Schätzfunktion.
Man bezeichnet mit $\mathbf{T}(\mathbf{x})$ den Schätzwert oder die Punktschätzung (zu konkreter Realisation $\mathbf{x}$) und mit $\mathbf{T}(\mathbf{X})$ den Punktschätzer von $\boldsymbol{\theta}$, der eine Zufallsvariable ist (auch gebräuchlich: $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ oder kurz $\widehat{\boldsymbol{\theta}}$, d.h. notationell wird nicht zwischen Schätzwert und Schätzfunktion unterschieden).

2.1 Klassische Schätztheorie

Überblick

Der Begriff der Suffizienz ist von grundlegender Bedeutung in der klassischen parametrischen Inferenz; darüber hinaus ist die Bedeutung (stark) abgeschwächt, vgl. auch Statistik IV.

Definition 2.3

Eine Statistik $\mathbf{T}$ *heißt suffizient für* $\boldsymbol{\theta}$ *(oder auch für* $\mathcal{P}$*)* $\overset{def}{\Leftrightarrow}$ *die bedingte Verteilung bzw. Dichte von* $\mathbf{X}$ *gegeben* $\mathbf{T}(\mathbf{X}) = \mathbf{t}$ *hängt für alle Werte von* $\mathbf{T}(\mathbf{X}) = \mathbf{t}$ *nicht von* $\boldsymbol{\theta}$ *ab, d.h.*

$$f_{X|T}(\mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{t}, \boldsymbol{\theta}) = f_{X|T}(\mathbf{x}|\mathbf{T}(\mathbf{X}) = \mathbf{t})$$

hängt nicht von $\boldsymbol{\theta}$ *ab.*

Idee: Zusätzliche Information in $\mathbf{X}$, die nicht in $\mathbf{T}$ enthalten ist, ist durch $f_{X|T}$ gegeben. Falls $f_{X|T}$ von $\boldsymbol{\theta}$ unabhängig ist, dann enthält die Stichprobe $\mathbf{x}$ nicht mehr Information über $\boldsymbol{\theta}$ als $\mathbf{T}(\mathbf{x})$.

Im Folgenden nehmen wir die Existenz einer Dichte für $\mathbf{X}$ an. Das Kriterium im folgenden Satz ist äquivalent und konstruktiv:

Satz 2.4 (Faktorisierungssatz, Neyman-Kriterium)

Eine Statistik $\mathbf{T}$ *ist suffizient für* $\boldsymbol{\theta}$ *genau dann wenn*

$$f(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\mathbf{T}(\mathbf{x})|\boldsymbol{\theta})$$

für fast alle $\mathbf{x}$.

D.h. die Dichte lässt sich in zwei Teile faktorisieren, von denen ein Teil von $\mathbf{x}$, aber nicht von $\boldsymbol{\theta}$, und der andere nur von $\boldsymbol{\theta}$ und $\mathbf{T}(\mathbf{x})$ abhängt.

Beweis.

„⇒": Falls $\mathbf{T}$ suffizient ist, gilt:

$$f_{X|T}(\mathbf{x}|\mathbf{T}(\mathbf{X})=\mathbf{t},\boldsymbol{\theta}) = \frac{f_{X,T}(\mathbf{x},\mathbf{t}|\boldsymbol{\theta})}{f_T(\mathbf{t}|\boldsymbol{\theta})} = f_{X|T}(\mathbf{x}|\mathbf{T}(\mathbf{X})=\mathbf{t}).$$

Weiterhin ist

$$f_{X,T}(\mathbf{x},\mathbf{t}|\boldsymbol{\theta}) = \begin{cases} f_X(\mathbf{x}|\boldsymbol{\theta}) & \text{für } \mathbf{T}(\mathbf{x})=\mathbf{t} \\ 0 & \text{sonst,} \end{cases}$$

d.h. insgesamt $f_{X|T}(\mathbf{x}|\mathbf{T}(\mathbf{X})=\mathbf{t}) \cdot f_T(\mathbf{t}|\boldsymbol{\theta}) = f_X(\mathbf{x}|\boldsymbol{\theta}) \cdot 1_{\mathbf{T}(\mathbf{x})=\mathbf{t}}$
und damit

$$\underbrace{f_{X|T}(\mathbf{x}|\mathbf{T}(\mathbf{X})=\mathbf{T}(\mathbf{x}))}_{h(\mathbf{x})} \cdot \underbrace{f_T(\mathbf{T}(\mathbf{x})|\boldsymbol{\theta})}_{g(\mathbf{T}(\mathbf{x})|\boldsymbol{\theta})} = f_X(\mathbf{x}|\boldsymbol{\theta}).$$

Beweis fortgeführt

„$\Leftarrow$": Man erhält die Dichte von $\mathbf{T}$, ausgewertet an $\mathbf{t}$, indem man im obigen Faktorisierungskriterium über die $\mathbf{x}$, für die $\mathbf{T}(\mathbf{x}) = \mathbf{t}$ gilt, summiert (bzw. integriert). Im diskreten Fall also:

$$f_T(\mathbf{t}|\boldsymbol{\theta}) = \sum_{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}} h(\mathbf{x})g(\mathbf{T}(\mathbf{x})|\boldsymbol{\theta}) = g(\mathbf{t}|\boldsymbol{\theta}) \sum_{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}} h(\mathbf{x}).$$

Die bedingte Dichte von $\mathbf{X}$ gegeben $\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}) = \mathbf{t}$,

$$\frac{f_X(\mathbf{x}|\boldsymbol{\theta})}{f_T(\mathbf{t}|\boldsymbol{\theta})} = \frac{h(\mathbf{x})g(\mathbf{T}(\mathbf{x})|\boldsymbol{\theta})}{\sum_{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}} h(\mathbf{x})g(\mathbf{t}|\boldsymbol{\theta})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}} h(\mathbf{x})},$$

hängt damit nicht von $\boldsymbol{\theta}$ ab. Im stetigen Fall werden Summen durch Integrale ersetzt (und es sind Messbarkeitsbedingungen zu beachten).

Beispiel 2.3 (Bernoulli-Experiment)

Seien $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, \pi)$ und $Z = \sum_{i=1}^n X_i$ die Anzahl der Erfolge.
Dann ist Z suffizient für π, denn

$$\begin{aligned} f_{X|Z}(\mathbf{x}|z, \pi) &= \mathbb{P}_\pi(\mathbf{X} = \mathbf{x}|Z = z) \\ &= \frac{\prod_{i=1}^n \pi^{x_i}(1-\pi)^{1-x_i}}{\binom{n}{z}\pi^z(1-\pi)^{n-z}}, \quad \text{wobei } \sum_{i=1}^n x_i = z \\ &= \binom{n}{z}^{-1} \end{aligned}$$

hängt nicht von π ab.

Beispiel 2.3 (Bernoulli-Experiment) fortgeführt

Gemäß Faktorisierungssatz ist

$$f_X(\mathbf{x}|\pi) = \underbrace{\frac{1}{\binom{n}{z}}}_{=h(\mathbf{x})} \underbrace{\binom{n}{z}\pi^z(1-\pi)^{n-z}}_{=g(z|\pi)} = \underbrace{1}_{=h^*(\mathbf{x})} \underbrace{\pi^z(1-\pi)^{n-z}}_{=g^*(z|\pi)}.$$

Beispiel 2.4 (Normalverteilung)

Sei $\mathbf{X} = (X_1, \ldots, X_n)$, $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ und $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$.

$$f_X(\mathbf{x}|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right)$$

$$= \underbrace{(2\pi)^{-n/2}}_{h(x)} \underbrace{(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right)\right)}_{g((\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)|\theta)},$$

d.h. $\mathbf{T}(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ ist suffizient für $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$.
Aber: Die bijektive Transformation $\widetilde{\mathbf{T}}(\mathbf{X}) = (\bar{X}, S^2)$ ist auch suffizient für $\boldsymbol{\theta}$, wobei S^2 die Stichprobenvarianz bezeichnet.

Beispiel 2.5 (Exponentialverteilung)

Sei $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, dann

$$f_X(\mathbf{x}|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \underbrace{1}_{h(\mathbf{x})} \cdot \underbrace{\lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)}_{g(T(\mathbf{x})|\lambda)}$$

mit $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$. Nach der ursprünglichen Definition ist

$$\frac{f_{X,T}(\mathbf{x}, t|\lambda)}{f_T(t|\lambda)} = \frac{\lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)}{\frac{\lambda^n}{\Gamma(n)} \left(\sum_{i=1}^{n} x_i\right)^{n-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)} = \frac{\Gamma(n)}{\left(\sum_{i=1}^{n} x_i\right)^{n-1}},$$

und hängt nicht von λ ab. Dabei wird benutzt, dass die Summe von n unabhängigen und identisch exponentialverteilten Zufallsvariablen mit Parameter λ gammaverteilt ist mit Parametern n und λ.

Beispiel 2.6 (Ordnungsstatistik)
Sei $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\boldsymbol{\theta})$ (wobei f Dichte einer stetigen Verteilung ist) und $\mathbf{T}(\mathbf{X}) = \mathbf{X}_{(\cdot)} = (X_{(1)}, \ldots, X_{(n)})$ die Ordnungsstatistik.

Dann gilt
$$f_{X|T}(\mathbf{x}|\mathbf{T} = \mathbf{x}_{(\cdot)}, \boldsymbol{\theta}) = \frac{1}{n!}\,.$$
Die Gleichheit folgt aus der Stetigkeit, denn $x_i \neq x_j \; \forall i \neq j$ (mit Wahrscheinlichkeit 1). $\mathbf{X}_{(\cdot)}$ ist suffizient für $\boldsymbol{\theta}$. Wir haben also bei i.i.d.-Beobachtungen keinen Informationsverlust durch Ordnen der Daten.

Bemerkung.

- Offensichtlich ist $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, d.h. die Stichprobe selbst, suffizient.
- Ebenso ist jede eineindeutige Transformation von $\mathbf{X}$ oder von einer suffizienten Statistik $\mathbf{T}(\mathbf{X})$ suffizient.
- Ist $\mathbf{T}$ suffizient, dann auch $(\mathbf{T}, \mathbf{T}^*)$, wobei $\mathbf{T}^*$ eine beliebige weitere Statistik darstellt.

Dies zeigt: Die Dimension einer suffizienten Statistik sollte soweit wie möglich reduziert werden.

Definition 2.5 (Minimalsuffizienz)

Eine Statistik $\mathbf{T}$ *heißt minimalsuffizient für* $\boldsymbol{\theta} \overset{def}{\Leftrightarrow}$ $\mathbf{T}$ *ist suffizient, und zu jeder anderen suffizienten Statistik* $\mathbf{V}$ *existiert eine Funktion* $\mathbf{H}$ *mit*

$$\mathbf{T}(\mathbf{X}) = \mathbf{H}(\mathbf{V}(\mathbf{X})) \;\; \mathcal{P} - \textit{fast überall.}$$

Frage: Existieren minimalsuffiziente Statistiken? Wenn ja, sind sie eindeutig?

Beispiel 2.7 (Normalverteilung)
Seien $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

1. $T(\mathbf{X}) = \bar{X}$ ist minimalsuffizient für μ bei bekanntem σ^2.
2. $T(\mathbf{X}) = \sum_{i=1}^n (X_i - \mu)^2$ ist minimalsuffizient für σ^2 bei bekanntem μ.
3. $\mathbf{T}(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ ist minimalsuffizient für μ und σ^2.
4. $T(X) = |X|$ ist minimalsuffizient für σ^2 wenn $X \sim N(0, \sigma^2)$ $(n = 1, \mu = 0)$. X ist ebenfalls suffizient, aber nicht minimalsuffizient (trotz gleicher Dimension).

Lemma 2.6

Sind $\mathbf{T}$ *und* $\mathbf{S}$ *minimalsuffiziente Statistiken, dann existieren injektive Funktionen* $\mathbf{g}_1$, $\mathbf{g}_2$, *so dass* $\mathbf{T} = \mathbf{g}_1(\mathbf{S})$ *und* $\mathbf{S} = \mathbf{g}_2(\mathbf{T})$.

Satz 2.7 (Charakterisierung von Minimalsuffizienz durch Likelihood-Quotienten)

Definiere den Likelihood-Quotienten

$$\Lambda_x(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{f(\mathbf{x}|\boldsymbol{\theta}_1)}{f(\mathbf{x}|\boldsymbol{\theta}_2)}.$$

Eine notwendige und hinreichende Bedingung für die Minimalsuffizienz einer Statistik $\mathbf{T}$ *für* $\boldsymbol{\theta}$ *ist, dass gilt:*

$$\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{x}') \;\Leftrightarrow\; \Lambda_x(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \Lambda_{x'}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\; \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2.$$

Beispiel 2.8 (Suffizienz in Exponentialfamilien)

Die Dichte einer r-parametrischen Exponentialfamilie hat die Form

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\theta}) &= h(\mathbf{x}) \cdot c(\boldsymbol{\theta}) \cdot \exp(\gamma_1(\boldsymbol{\theta}) T_1(\mathbf{x}) + \ldots + \gamma_r(\boldsymbol{\theta}) T_r(\mathbf{x})) \\ &= h(\mathbf{x}) \cdot \exp(b(\boldsymbol{\theta}) + \boldsymbol{\gamma}(\boldsymbol{\theta})^\top \mathbf{T}(\mathbf{x})), \end{aligned}$$

d.h. $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_r(\mathbf{X}))^\top$ ist suffizient für $\boldsymbol{\theta}$ nach dem Faktorisierungssatz. Mit Satz 2.7 folgt, dass $\mathbf{T}(\mathbf{X})$ auch minimalsuffizient ist, falls $\boldsymbol{\gamma}(\Theta)$ ein offenes Rechteck in $\mathbb{R}^r$ enthält.

Es folgt nun die Charakterisierung der Minimalsuffizienz nach Lehmann-Scheffé. Dazu wird der Begriff der Vollständigkeit benötigt.

Definition 2.8

Eine Statistik $\mathbf{T}$ *ist vollständig* $\overset{def}{\Leftrightarrow}$ *für jede reelle (messbare) Funktion* g *gilt:*

$$\mathbb{E}_\theta[g(\mathbf{T})] = 0 \;\; \forall \boldsymbol{\theta} \;\Rightarrow\; \mathbb{P}_\theta(g(\mathbf{T}) = 0) = 1 \; \forall \boldsymbol{\theta}.$$

Beispiel Vollständigkeit

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} Po(\lambda)$ und $T = \sum_{i=1}^n X_i$. Nach Beispiel 2.8 ist T suffizient.

Sei g messbar. Da $T = \sum_i X_i \sim Po(n\lambda)$, gilt für

$$E_\lambda[g(T)] = \sum_{t=0}^{\infty} g(t) \frac{(n\lambda)^t}{t!} \exp(-n\lambda) = \exp(-n\lambda) \sum_{t=0}^{\infty} \underbrace{g(t) \frac{n^t}{t!}}_{c_t} \lambda^t = 0 \;\; \forall \lambda$$

$$\Rightarrow \sum_{t=0}^{\infty} c_t \lambda^t = \sum_{t=0}^{\infty} 0 \lambda^t \;\; \forall \lambda$$

(zwei konvergierende Potenzreihen). Daraus folgt $c_t = 0$ für alle t und wegen $\frac{n^t}{t!} \neq 0$ also $g(t) = 0$ für alle t. Damit ist T vollständig (und mit dem folgenden Satz auch minimalsuffizient).

Aus der Definiton wird nicht unmittelbar klar, warum „Vollständigkeit" eine wünschenswerte Eigenschaft eines Schätzers sein sollte. Einen möglichen Grund liefert der folgende Satz.

Satz 2.9 (Lehmann-Scheffé)

Angenommen, $\mathbf{X}$ *besitzt eine Dichte* $f(\mathbf{x}|\boldsymbol{\theta})$ *und* $\mathbf{T}(\mathbf{X})$ *ist suffizient und vollständig für* $\boldsymbol{\theta}$*. Dann ist* $\mathbf{T}(\mathbf{X})$ *minimalsuffizient für* $\boldsymbol{\theta}$*.*

Beweis des Satzes von Lehmann-Scheffé:

Vorausgesetzt wird, dass eine minimalsuffiziente Statistik existiert - bewiesen wurde dies von Lehmann und Scheffé (1950). Ist dies der Fall, so ist diese bis auf bijektive Transformationen eindeutig. Bezeichne $\mathbf{S} = \mathbf{g}_1(\mathbf{T})$ eine solche minimalsuffiziente Statistik für eine Funktion $\mathbf{g}_1$. Definiere $\mathbf{g}_2(\mathbf{S}) = \mathbb{E}[\mathbf{T}|\mathbf{S}]$. Da $\mathbf{S}$ suffizient für $\boldsymbol{\theta}$ ist, hängt $\mathbf{g}_2(\mathbf{S})$ nicht von $\boldsymbol{\theta}$ ab. Betrachte nun

$$\mathbf{g}(\mathbf{T}) = \mathbf{T} - \mathbf{g}_2(\mathbf{S}) = \mathbf{T} - \mathbf{g}_2(\mathbf{g}_1(\mathbf{T})).$$

Anwendung des Satzes von der iterierten Erwartung liefert:

$$\mathbb{E}_\theta[\mathbf{g}(\mathbf{T})] = \mathbb{E}_\theta[\mathbf{T}] - \mathbb{E}_\theta[\mathbb{E}[\mathbf{T}|\mathbf{S}]] = \mathbf{0}.$$

Da $\mathbf{T}$ vollständig ist, ist $\mathbf{g}(\mathbf{T}) = \mathbf{0}$ bzw. $\mathbf{g}_2(\mathbf{S}) = \mathbf{T}$ mit Wahrscheinlichkeit 1, d.h. $\mathbf{T}$ ist eine Funktion von $\mathbf{S}$. Da $\mathbf{S}$ eine Funktion jeder suffizienten Statistik ist, gilt dies damit auch für $\mathbf{T}$ und $\mathbf{T}$ ist also minimalsuffizient. ($\mathbf{S}$ und $\mathbf{T}$ sind äquivalent.) $\square$

Bemerkung (Ancillarity einer Statistik)
Eine Statistik $\mathbf{V}(\mathbf{X})$ heißt *ancillary* („Hilfsstatistik", „ankillar") für $\mathcal{P}$, wenn ihre Verteilung nicht von $\boldsymbol{\theta}$ abhängt (also bekannt ist).
Häufiger Sachverhalt: $\mathbf{T} = (\mathbf{U}, \mathbf{V})$ ist suffizient für $\boldsymbol{\theta}$, $\mathbf{V}$ ancillary, $\mathbf{U}$ nicht suffizient.

Beispiel 2.9 (Ancillary Statistik)

$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathit{Unif}\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$.

Man kann dann zeigen (Davison, 2004, Ex. 12.3), dass mit

$$
\begin{aligned}
U &= \frac{1}{2}(X_{(1)} + X_{(n)}) \\
V &= X_{(n)} - X_{(1)}
\end{aligned}
$$

$T = (U, V)$ minimalsuffizient (aber nicht vollständig) für θ ist.
Ferner ist U alleine nicht suffizient und V ancillary.

- Fehler eines Schätzers $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$ ist $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$.
- Messung des Fehlers durch Verlustfunktion, zum Beispiel

$$
\begin{aligned}
L(\widehat{\theta}, \theta) &= |\widehat{\theta} - \theta| && \text{Abstand } (\theta \text{ skalar}), \\
L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 && \text{quadratischer Fehler}, \\
L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \frac{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2}{\|\boldsymbol{\theta}\|^2} && \text{relativer quadratischer Fehler}, \\
L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathbf{D} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) && \text{gewichteter quadratischer Fehler} \\
& && (\mathbf{D} \text{ positiv definit}).
\end{aligned}
$$

- Risikofunktion $R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}_\theta[L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})]$.
- Hier wird (hauptsächlich) quadratischer Verlust betrachtet.

Definition 2.10 (Erwartungstreue, Bias, Varianz eines Schätzers)

- $\widehat{\boldsymbol{\theta}}$ *heißt erwartungstreu* $\overset{def}{\Leftrightarrow}$ $\mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$.
- $\mathrm{Bias}_\theta(\widehat{\boldsymbol{\theta}}) = \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}$.
- $\mathrm{Var}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta} - \mathbb{E}_\theta[\widehat{\theta}])^2]$, θ *skalar.*
  $\mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}}) = \mathbb{E}_\theta[(\widehat{\boldsymbol{\theta}} - \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}])(\widehat{\boldsymbol{\theta}} - \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}])^\top]$

Definition 2.11 (MSE)

Der mittlere quadratische Fehler (mean squared error) ist definiert als

$$\mathrm{MSE}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] = \mathrm{Var}_\theta(\widehat{\theta}) + (\mathrm{Bias}_\theta(\widehat{\theta}))^2.$$

Der Gesamtfehler lässt sich also aufteilen in einen zufälligen Fehler (Varianz) und einen systematischen (quadrierter Bias).

Vergleicht man zwei Schätzer bezüglich ihres MSE, kann für einen Teilbereich von Θ der MSE des einen, für andere Teilbereiche der MSE des zweiten Schätzers kleiner sein:

Beispiel 2.10 (siehe Übung 2, Aufgabe 3)
$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(1, \pi)$.

1. MSE von $\widehat{\pi} = \bar{X}$:

$$\mathbb{E}_\pi[(\bar{X} - \pi)^2] = \mathrm{Var}_\pi(\bar{X}) = \frac{\pi(1-\pi)}{n}.$$

Beispiel 2.10 fortgeführt

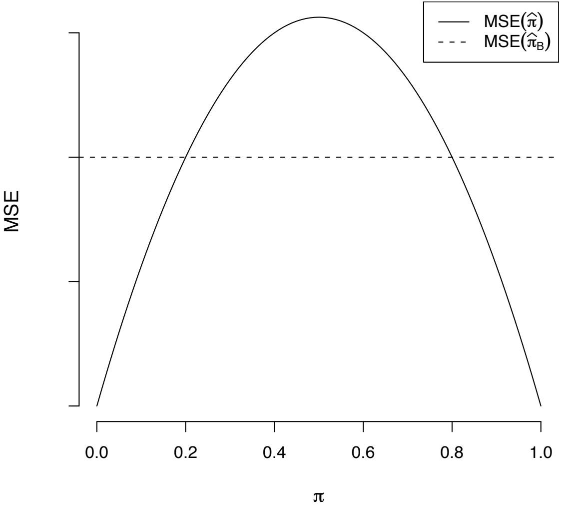
2. MSE des Bayes-Schätzers (Posteriori-Erwartungswert) bei einer Priori $p(\pi) \sim \text{Be}(a, b)$:

$$\widehat{\pi}_B = \frac{n\bar{X} + a}{a + b + n},$$

$$\mathrm{MSE}(\widehat{\pi}_B) = \frac{n\pi(1-\pi) + (a - (a+b)\pi)^2}{(a+b+n)^2}.$$

Für $a = b = \sqrt{n}/2$ ergibt sich

$$\mathrm{MSE}_\pi(\widehat{\pi}_B) = \frac{1}{4(1+\sqrt{n})^2} = \text{const bezüglich } \pi.$$


MSE($\hat{\pi}$)
MSE($\hat{\pi}_B$)
MSE
0.0
0.2
0.4
0.6
0.8
1.0
π

Fazit: In der Regel wird man keinen „MSE-optimalen" Schätzer $\widehat{\theta}^{\text{opt}}$ finden in dem Sinne, dass $\mathrm{MSE}_\theta(\widehat{\theta}^{\text{opt}}) \leq \mathrm{MSE}_\theta(\widehat{\theta})$ für alle θ und alle konkurrierenden $\widehat{\theta}$. Bei Einschränkung auf erwartungstreue Schätzer ist dies öfter möglich.
Deshalb die Forderung:

Definition 2.12 (zulässiger („admissible") Schätzer)

Ein Schätzer $\widehat{\theta}$ *heißt zulässig* $\overset{def}{\Leftrightarrow}$ *es gibt keinen Schätzer* $\widetilde{\theta}$ *mit* $\mathrm{MSE}_\theta(\widetilde{\theta}) \leq \mathrm{MSE}_\theta(\widehat{\theta})$ *für alle* θ *und* $\mathrm{MSE}_\theta(\widetilde{\theta}) < \mathrm{MSE}_\theta(\widehat{\theta})$ *für mindestens ein* θ, *d.h. es gibt keinen Schätzer* $\widetilde{\theta}$, *der* $\widehat{\theta}$ *gleichmäßig/strikt „dominiert".*

Definition 2.13 (Verallgemeinerungen des MSE auf $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$, $k > 1$)

Üblich sind die folgenden zwei Alternativen:

1. *MSE (skalar):*

$$\begin{aligned} \mathrm{MSE}_\theta^{(1)}(\widehat{\boldsymbol{\theta}}) &= \mathbb{E}_\theta[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2] \\ &= \sum_{j=1}^{k} \mathbb{E}_\theta[(\widehat{\theta}_j - \theta_j)^2] \\ &= \sum_{j=1}^{k} \mathrm{MSE}_\theta(\widehat{\theta}_j) \end{aligned}$$

2. *MSE-Matrix:*

$$\begin{aligned} \mathrm{MSE}_\theta^{(2)}(\widehat{\boldsymbol{\theta}}) &= \mathbb{E}_\theta[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top] \\ &= \mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}}) + (\mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta})(\mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta})^\top \end{aligned}$$

Diese Variante wird häufig bei linearen Modellen betrachtet.

Bemerkung. Das j-te Diagonalelement der MSE-Matrix ist $\mathrm{MSE}_\theta(\widehat{\theta}_j)$. Vergleich von MSE-Matrizen gemäß *„Löwner"-Ordnung*:

$$\mathrm{MSE}_\theta(\widetilde{\boldsymbol{\theta}}) \overset{(\leq)}{<} \mathrm{MSE}_\theta(\widehat{\boldsymbol{\theta}})$$

bedeutet, dass die Differenz $\mathrm{MSE}_\theta(\widehat{\boldsymbol{\theta}}) - \mathrm{MSE}_\theta(\widetilde{\boldsymbol{\theta}})$ positiv (semi-)definit ist.
Man definiert allgemein für symmetrische $(m \times m)$-Matrizen $\mathbf{A}, \mathbf{B}$:

$$\begin{aligned} \mathbf{A} \leq \mathbf{B} \quad &\overset{\text{def}}{\Leftrightarrow} \quad \mathbf{B} - \mathbf{A} \text{ ist positiv semidefinit,} \\ \mathbf{A} < \mathbf{B} \quad &\overset{\text{def}}{\Leftrightarrow} \quad \mathbf{B} - \mathbf{A} \text{ ist positiv definit.} \end{aligned}$$

Beispiel 2.11 (Gauß-Experiment) Seien $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$.

- σ^2 bekannt, μ unbekannt: MSE-Vergleich von $\bar{X}$ und $T = a\bar{X} + b$.
- σ^2 unbekannt, μ bekannt:
  - Eine Möglichkeit:

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2,\ \mathbb{E}_{\sigma^2}(S_\mu^2) = \sigma^2$$

  - Weitere Möglichkeit:

$$V_\mu^2 = \frac{1}{n+2}\sum_{i=1}^{n}(X_i - \mu)^2, \mathbb{E}_{\sigma^2}(V_\mu^2) = \frac{n}{n+2}\sigma^2$$

Es stellt sich heraus, dass $\mathrm{MSE}_{\sigma^2}(V_\mu^2) < \mathrm{MSE}_{\sigma^2}(S_\mu^2)$ ist.

Beispiel 2.11 (Gauß-Experiment) fortgeführt

- μ und σ^2 unbekannt:
  - Eine Möglichkeit:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

$$\mathbb{E}_{\sigma^2}(S^2) = \sigma^2, \;\; \mathrm{MSE}_{\sigma^2}(S^2) = \mathrm{Var}_{\sigma^2}(S^2) = \frac{2}{n-1}\,\sigma^4.$$

  - Weitere Möglichkeit:

$$V^2 = \frac{1}{n+1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

$$\mathbb{E}_{\sigma^2}(V^2) = \frac{n-1}{n+1}\,\sigma^2, \; \mathrm{MSE}_{\sigma^2}(V^2) = \frac{2}{n+1}\,\sigma^4,$$

d.h. V^2 dominiert S^2.

Beispiel 2.11 (Gauß-Experiment) fortgeführt

- μ und σ^2 (weiterhin) unbekannt:
  - Der sogenannte *Stein-Schätzer*

$$T = \min\left\{V^2, \frac{1}{n+2}\sum_{i=1}^{n} X_i^2\right\}$$

dominiert V^2 (und damit S^2).

Plausibilitätsbetrachtung: Ist $\mu = 0$, so ist $\sum_{i=1}^{n} X_i^2/(n+2)$ besserer Schätzer als V^2. Ist $\mu \neq 0$, so ist V^2 ein besserer Schätzer als $\sum_{i=1}^{n} X_i^2/(n+2)$. Beim Stein-Schätzer wird fallweise mit hoher Wahrscheinlichkeit der jeweils bessere Schätzer benutzt.

Beispiel 2.12 (Steins Paradoxon)
Seien $(X_1, \ldots, X_m)^\top \sim N_m(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_m)$ multivariat normalverteilt mit $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ und σ^2 bekannt. Es sollen simultan die Erwartungswerte $\mu_1, \ldots, \mu_m$ geschätzt werden. Man beachte dabei, dass die einzelnen Komponenten als unabhängig angenommen werden. Die Stichprobe hat die Form

$$X_{11}, \ldots, X_{1n_1}, \ldots, X_{m1}, \ldots, X_{mn_m}$$

(i.i.d. Stichproben aus „Gruppen" $1, \ldots, m$). Übliche Schätzer:

$$T_j = \bar{X}_j, \quad j = 1, \ldots, m, \qquad \mathbf{T} = (T_1, \ldots, T_m)^\top = (\bar{X}_1, \ldots, \bar{X}_m)^\top.$$

Der (skalare) MSE ist:

$$\mathbb{E}_\mu[\|\mathbf{T} - \boldsymbol{\mu}\|^2] = \sum_{j=1}^{m} \mathbb{E}_\mu[(\bar{X}_j - \mu_j)^2] = \sum_{j=1}^{m} \frac{\sigma^2}{n_j}.$$

Beispiel 2.12 (Steins Paradoxon) fortgeführt
Paradoxerweise gilt:

1. Für $m \leq 2$ ist $\mathbf{T}$ zulässig.
2. Für $m \geq 3$ ist $\mathbf{T}$ *nicht* zulässig und wird dominiert durch den (James-)Stein-Schätzer

$$\mathbf{T}^* = \left(1 - \frac{(m-2)\sigma^2}{\sum_{j=1}^m n_j \bar{X}_j^2}\right)\mathbf{T}.$$

Es lässt sich zeigen: $\mathbf{T}^*$ ist selbst unzulässig. Der Stein-Schätzer ist ein sogenannter *Shrinkage-Schätzer*.

Beispiel 2.13 (Lineares Modell)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \;\; \boldsymbol{\varepsilon} \;\sim\; (N_n)(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

$$\begin{aligned} &\text{KQ-Schätzer:} && \widehat{\boldsymbol{\beta}}_{\text{KQ}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &\text{Ridge-Schätzer:} && \widehat{\boldsymbol{\beta}}_{\text{Ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{D})^{-1} \mathbf{X}^\top \mathbf{y}, \end{aligned}$$

wobei $\mathbf{D}$ eine Diagonalmatrix mit positiven Diagonalelementen ist. Für einen MSE-Vergleich siehe Übung.

Fazit: Bereits im einfachen Beispiel der Schätzung von π in $B(1, \pi)$ (siehe Beispiel 2.10) zeigt sich, dass es im Allgemeinen keine MSE-optimalen Schätzer gibt.

Auswege:

1. Einschränkung auf Teilklasse von Schätzern, zum Beispiel erwartungstreue (und lineare) Schätzer, äquivariante Schätzer, ...
2. MSE-Kriterium verändern:
   - Ersetze $\mathrm{MSE}_\theta(\widehat{\boldsymbol{\theta}})$ durch Minimierung von $\max_{\boldsymbol{\theta} \in \Theta} \mathrm{MSE}_\theta(\widehat{\boldsymbol{\theta}})$ (Minimax-Kriterium)
   - oder ersetzte $\mathrm{MSE}_\theta(\widehat{\boldsymbol{\theta}})$ durch $\mathbb{E}_{p(\boldsymbol{\theta})}[\mathrm{MSE}_\theta(\widehat{\boldsymbol{\theta}})]$ bei einer Priori-Verteilung $p(\boldsymbol{\theta})$ (Bayes-Schätzer).

Hier: Strategie 1 mit erwartungstreuen Schätzern, vgl. 2.1.4.

Definition 2.14 (Fisher-reguläre Verteilungsfamilien)

Eine Familie von Verteilungen $\mathcal{P}$ mit Dichte $f(\mathbf{x}|\boldsymbol{\theta}) = f(x_1, \ldots, x_n|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, heißt Fisher-regulär, wenn Folgendes gilt:

1. *Der Träger $\{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}|\boldsymbol{\theta}) > 0\}$ hängt nicht von $\boldsymbol{\theta}$ ab (dies ist zum Beispiel bei $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Unif[0;\theta]$ oder bei der Pareto-Verteilung verletzt).*
2. *Θ ist offen in $\mathbb{R}^k$ (verletzt zum Beispiel bei $\sigma^2 \geq 0$).*
3. *Die ersten und zweiten Ableitungen von $f(\mathbf{x}|\boldsymbol{\theta})$ bzgl. $\boldsymbol{\theta}$ existieren und sind für jedes $\boldsymbol{\theta}$ endliche Funktionen von $\mathbf{x}$.*
4. *Vertauschbarkeit: Sowohl für $f(\mathbf{x}|\boldsymbol{\theta})$ als auch für $\log(f(\mathbf{x}|\boldsymbol{\theta}))$ kann erstes und zweites Differenzieren nach $\boldsymbol{\theta}$ und Integration über $\mathbf{x}$ vertauscht werden.*

Definition 2.15 (Log-Likelihood, Scorefunktion und Information)

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \log f(\mathbf{x}|\boldsymbol{\theta}) \quad \textit{(Log-Likelihood von } \boldsymbol{\theta} \textit{ bzgl. der Stichprobe } \mathbf{x}\textit{)}$$

$$\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta}|\mathbf{x}) = \left(\frac{\partial}{\partial\theta_1}\ell(\boldsymbol{\theta}|\mathbf{x}), \ldots, \frac{\partial}{\partial\theta_k}\ell(\boldsymbol{\theta}|\mathbf{x})\right)^\top$$

(Score-Funktion)

$$\mathbf{J}(\boldsymbol{\theta}|\mathbf{x}) = -\frac{\partial^2\ell(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \quad \textit{(beobachtete Informationsmatrix der}$$

$$\textit{Stichprobe mit Elementen} \quad (\mathbf{J}(\boldsymbol{\theta}|\mathbf{x}))_{ij} = -\frac{\partial^2 \log f(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\textit{)}$$

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathrm{E}_\theta[\mathbf{J}(\boldsymbol{\theta}|\mathbf{X})] \quad \textit{(erwartete oder Fisher-Informationsmatrix)}$$

Satz 2.16

Ist $\mathcal{P}$ Fisher-regulär, so gilt:

1. $\mathbb{E}_\theta\left[\mathbf{s}(\boldsymbol{\theta}|\mathbf{X})\right] = \mathbf{0}$
2. $\mathrm{Cov}_\theta[\mathbf{s}(\boldsymbol{\theta}|\mathbf{X})] = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$.

Beweis

Zu 1.:

$$
\begin{aligned}
\mathbb{E}_\theta[\mathbf{s}(\boldsymbol{\theta}|\mathbf{X})] &= \int \mathbf{s}(\boldsymbol{\theta}|\mathbf{x}) f(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} \log(f(\mathbf{x}|\boldsymbol{\theta})) f(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} \\
&= \int \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}|\boldsymbol{\theta})}{f(\mathbf{x}|\boldsymbol{\theta})} f(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = \mathbf{0}
\end{aligned}
$$

Zu 2.:

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathbb{E}_\theta\left[-\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right] = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\frac{\partial}{\partial\boldsymbol{\theta}^\top} f(\mathbf{X}|\boldsymbol{\theta})}{f(\mathbf{X}|\boldsymbol{\theta})}\right)\right]$$

$$= -\mathbb{E}_\theta\left[\frac{f(\mathbf{X}|\boldsymbol{\theta})\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} f(\mathbf{X}|\boldsymbol{\theta}) - (\frac{\partial}{\partial\boldsymbol{\theta}} f(\mathbf{X}|\boldsymbol{\theta}))(\frac{\partial}{\partial\boldsymbol{\theta}^\top} f(\mathbf{X}|\boldsymbol{\theta}))}{f(\mathbf{X}|\boldsymbol{\theta})^2}\right]$$

unter Verwendung der Quotientenregel der Differentiation. Dies ist

$$\begin{aligned} = \quad - \quad & \mathbb{E}_\theta\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} f(\mathbf{X}|\boldsymbol{\theta})}{f(\mathbf{X}|\boldsymbol{\theta})}\right] + \mathbb{E}_\theta\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}} f(\mathbf{X}|\boldsymbol{\theta})}{f(\mathbf{X}|\boldsymbol{\theta})} \cdot \frac{\frac{\partial f(\mathbf{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top}}{f(\mathbf{X}|\boldsymbol{\theta})}\right] \\ = \quad - \quad & \int \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} f(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} + \mathbb{E}_\theta[\mathbf{s}(\boldsymbol{\theta};\mathbf{X})\mathbf{s}(\boldsymbol{\theta}|\mathbf{X})^\top] \end{aligned}$$

Zu 2. (fortgeführt):
Der erste Summand ist unter Vertauschung von Differentiation und Integration gleich null. Für den zweiten Teil ergibt sich mit Teil 1.

$$\mathbb{E}_\theta[\mathbf{s}(\boldsymbol{\theta}|\mathbf{X})\mathbf{s}(\boldsymbol{\theta}|\mathbf{X})^\top] = \mathrm{Cov}_\theta(s(\boldsymbol{\theta}|\mathbf{X})).$$

□

Weitere Eigenschaften:

- Sind $X_1, \ldots, X_n$ unabhängig und gemäß $X_i \;\sim\; f_i(x|\boldsymbol{\theta}),\; i = 1, \ldots, n$, verteilt, so gilt:

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}|x_i) \quad , \quad \ell_i(\boldsymbol{\theta}|x_i) = \log f_i(x_i|\boldsymbol{\theta})$$

$$\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \mathbf{s}_i(\boldsymbol{\theta}|x_i) \quad , \quad s_i(\boldsymbol{\theta}|x_i) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_i(x_i|\boldsymbol{\theta})$$

$$\mathbf{J}(\boldsymbol{\theta}|\mathbf{x}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \;=\; \sum_{i=1}^{n} -\left(\frac{\partial^2 \log f_i(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right)$$

Weitere Eigenschaften (fortgeführt):

- Für $X_1, \ldots, X_n$ i.i.d. wie $X_1 \sim f_1(x|\boldsymbol{\theta})$ folgt

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathbb{E}_\theta[\mathbf{J}(\boldsymbol{\theta}|\mathbf{X})] = n \cdot \mathbf{i}(\boldsymbol{\theta}),$$

wobei

$$\mathbf{i}(\boldsymbol{\theta}) = \mathbb{E}_\theta\left[-\frac{\partial^2 \ell_1(\boldsymbol{\theta}|X_1)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right] = \mathrm{Cov}_\theta\left(\frac{\partial \log f_1(X_1|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)$$

die erwartete Information einer Einzelbeobachtung ist, d.h. die erwartete Informationsmatrix der Stichprobe $X_1, \ldots, X_n$ ist die n-fache erwartete Information einer (typischen) Stichprobenvariable X_1.

Weitere Eigenschaften (fortgeführt):

- Für eine Statistik $\mathbf{T} = \mathbf{T}(\mathbf{X}), \mathbf{X} = (X_1, \ldots, X_n)^\top$ mit $\mathbf{T} \sim f_T(\mathbf{t}|\boldsymbol{\theta})$ kann man die Begriffe Score-Funktion und Fisher-Information völlig analog definieren. Insbesondere ist

$$\boldsymbol{\mathcal{I}}_T(\boldsymbol{\theta}) = \mathbb{E}_\theta \left[- \frac{\partial^2 \log f_T(\mathbf{t}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right].$$

Satz 2.17 (Suffizienz und Fisher-Information)

Sei $\mathcal{I}(\boldsymbol{\theta})$ die Fisher-Information für $\mathbf{X}$*. Dann gilt unter Fisher-Regularität für jede Statistik* $\mathbf{T}$*:*

1. $\mathcal{I}_T(\boldsymbol{\theta}) \leq \mathcal{I}(\boldsymbol{\theta})$.
2. $\mathcal{I}_T(\boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta}) \ \Leftrightarrow\ \mathbf{T}$ *ist suffizient für* $\boldsymbol{\theta}$.

Also: Bei einer suffizienten Statistik $\mathbf{T}$ wird keine (erwartete) Information „verschenkt".

- „Schöne" Resultate für finites n, aber für vergleichsweise einfache statistische Modelle.

- Problem: Für komplexere Modelle existieren keine „vernünftigen" erwartungstreuen Schätzer.

- Aber: Etliche Resultate lassen sich übertragen auf asymptotische Betrachtungen für $n \to \infty$.

Informationsungleichungen

I. $\theta \in \mathbb{R}$ (skalar). Neben θ werden auch transformierte Parameter $\tau(\theta)$ betrachtet. Wenn Ableitungen benötigt werden, nehmen wir stillschweigend an, dass sie existieren.

Satz 2.18

Sei $f(\mathbf{x}|\theta)$ Fisher-regulär.

1. *Ist $\widehat{\theta}$ erwartungstreu für θ, so gilt:*

$$\mathrm{Var}_\theta(\widehat{\theta}) \geq \frac{1}{\mathcal{I}(\theta)} \qquad \text{(Cramer-Rao-Ungleichung)}.$$

2. *Ist $T = T(\mathbf{X})$ erwartungstreu für $\tau(\theta)$, so gilt:*

$$\mathrm{Var}_\theta(T) \geq \frac{(\tau'(\theta))^2}{\mathcal{I}(\theta)}.$$

$\frac{(\tau'(\theta))^2}{\mathcal{I}(\theta)}$ *heißt Cramer-Rao-Schranke.*

3. *Besitzt $\widehat{\theta}$ den Bias $B(\theta) = \mathbb{E}_\theta[\widehat{\theta}] - \theta$, so gilt*

$$\mathrm{MSE}_\theta(\widehat{\theta}) \geq B^2(\theta) + \frac{(1 + B'(\theta))^2}{\mathcal{I}(\theta)}.$$

Beweis.
Gezeigt wird 2. Daraus folgt 1. für $\tau(\theta) = \theta$ und 3. für $\tau(\theta) = \theta + B(\theta)$.
Differentiation von

$$\tau(\theta) = \mathbb{E}_\theta[T] = \int T(\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x}$$

bezüglich θ, und Verwendung der Fisher-Regularität liefert:

$$\begin{aligned}\tau'(\theta) &= \int T(\mathbf{x}) \frac{d}{d\theta} f(\mathbf{x}|\theta)\, d\mathbf{x} \\ &= \int T(\mathbf{x}) s(\theta|\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x} \\ &= \mathrm{Cov}_\theta(T(\mathbf{X}), s(\theta|\mathbf{X})).\end{aligned}$$

Beweis fortgeführt.
Unter Verwendung der Cauchy-Schwarz-Ungleichung

$$|\mathrm{Cov}(U, V)| \leq \sqrt{\mathrm{Var}(U)}\sqrt{\mathrm{Var}(V)}$$

folgt

$$\begin{aligned} (\tau'(\theta))^2 &\leq \mathrm{Var}_\theta(T(\mathbf{X}))\mathrm{Var}_\theta(s(\theta|\mathbf{X})) \\ &= \mathrm{Var}_\theta(T(\mathbf{X}))\mathcal{I}(\theta). \end{aligned}$$

Also:

$$\mathrm{Var}_\theta(T(\mathbf{X})) \geq \frac{(\tau'(\theta))^2}{\mathcal{I}(\theta)}.$$

□

Bemerkung.
Die Gleichheit wird genau dann angenommen, wenn eine einparametrische Exponentialfamilie $f(\mathbf{x}|\theta) = h(\mathbf{x}) \exp(b(\theta) + \gamma(\theta) T(\mathbf{x}))$ vorliegt. In diesem Fall ist $T(\mathbf{x})$ ein effizienter Schätzer für seinen Erwartungswert $\tau(\theta) = -b'(\theta)/\gamma'(\theta)$. Also: eher eine kleine Modellklasse.

II. **Informationsungleichungen** für $\boldsymbol{\theta}$ bzw. $\boldsymbol{\tau}(\boldsymbol{\theta})$ mehrdimensional.

Satz 2.19

Sei $f(\mathbf{x}|\boldsymbol{\theta})$ Fisher-regulär.

1. *Ist $\widehat{\boldsymbol{\theta}}$ erwartungstreu für $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, so gilt:*

$$\mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}}) \geq \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}),$$

   wobei sich „$\geq$" auf die Löwner-Ordnung bezieht (vgl. F. 117). Daraus folgt insbesondere $\mathrm{Var}_\theta(\widehat{\theta}_j) \geq f_{jj}$, $j = 1, \ldots, k$, wobei f_{jj} das j-te Diagonalelement von $\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ bezeichnet.

2. *Ist $\mathbf{T}$ erwartungstreu für $\boldsymbol{\tau}(\boldsymbol{\theta})$, so gilt*

$$\mathrm{Cov}_\theta(\mathbf{T}) \geq \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})^\top$$

   mit der Funktionalmatrix $(\mathbf{H}(\boldsymbol{\theta}))_{ij} = \frac{\partial}{\partial\theta_j}\tau_i(\boldsymbol{\theta})$. Die Matrix $\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})^\top$ ist die Cramer-Rao-Schranke.

Bemerkung. Obige Bemerkung für skalares θ gilt analog für

$$f(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp(b(\boldsymbol{\theta}) + \boldsymbol{\gamma}^\top(\boldsymbol{\theta})\mathbf{T}(\mathbf{x})),$$

d.h. für mehrparametrische Exponentialfamilien.

Beispiel 2.14 (Cramer-Rao-Schranke bei $X \sim N(\mu, \sigma^2)$)
$X_1, \ldots, X_n$ i.i.d. wie $X \sim N(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2)$. Dann gilt für die Informationsmatrix

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \quad \text{bzw.} \quad \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

Beste erwartungstreue Schätzer

Erwartungstreue Schätzer minimaler Varianz innerhalb einer vorgegebenen Klasse nennt man *effizient*. Die Informationsungleichungen motivieren:

Definition 2.20 (Gleichmäßig bester erwartungstreuer (UMVU) Schätzer)

1. *θ skalar:*
   Der Schätzer $\widehat{\theta}_{\mathit{eff}}$ für θ heißt gleichmäßig bester erwartungstreuer oder UMVU („uniformly minimum variance unbiased")-Schätzer $\overset{def}{\Leftrightarrow}$ $\widehat{\theta}_{\mathit{eff}}$ ist erwartungstreu, und es gilt $\mathrm{Var}_\theta(\widehat{\theta}_{\mathit{eff}}) \leq \mathrm{Var}_\theta(\widehat{\theta})$ für alle θ und jeden erwartungstreuen Schätzer $\widehat{\theta}$.
2. *$\boldsymbol{\theta}$ mehrdimensional:*
   Ersetze $\mathrm{Var}_\theta(\widehat{\theta}_{\mathit{eff}}) \leq \mathrm{Var}_\theta(\widehat{\theta})$ durch $\mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}}_{\mathit{eff}}) \leq \mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}})$.

Satz 2.21 (Effizienz und Informationsungleichungen)

Sei $f(\mathbf{x}|\boldsymbol{\theta})$ Fisher-regulär und $\widehat{\boldsymbol{\theta}}$ erwartungstreu für $\boldsymbol{\theta}$. Falls $\mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ für alle $\boldsymbol{\theta}$, so ist $\widehat{\boldsymbol{\theta}}$ ein UMVU-Schätzer.

Beweis. Die Aussage folgt direkt aus der Informationsungleichung und obiger Definition. □

Beispiel 2.15 (Gauß-Experiment)
Seien $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ mit μ, σ^2 unbekannt.
Aus Beispiel 2.14 wissen wir, dass $\mathcal{I}(\mu) = n/\sigma^2$ und somit $\mathcal{I}^{-1}(\mu) = \sigma^2/n = \mathrm{Var}(\bar{X})$. Dann ist $\bar{X}$ UMVU für μ.
Aber

$$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n} = \mathcal{I}^{-1}(\sigma^2).$$

Die Cramer-Rao-Schranke wird also nicht erreicht, somit kann nicht gefolgert werden, dass S^2 UMVU für σ^2 ist.

Beispiel 2.16 (Lineares Modell)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim\ N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n) \quad \text{bzw.} \quad \mathbf{y}\ \sim\ N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\mathrm{KQ}} &= \widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \text{ ist effizient für } \boldsymbol{\beta}, \\ \widehat{\sigma}^2 &= \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \text{ ist UMVU-Schätzer für } \sigma^2. \end{aligned}$$

Bemerkung. Zu unterscheiden sind folgende Situationen:

1. Es existiert ein UMVU-Schätzer, dessen Varianz gleich der Cramer-Rao-Schranke ist.
2. Es existiert ein UMVU-Schätzer, dessen Varianz größer als die Cramer-Rao-Schranke ist (findet man mit dem Satz von Lehmann-Scheffé, siehe Satz 2.23).
3. Der häufigste Fall: Es existiert (für finiten Stichprobenumfang) kein UMVU-Schätzer.

Fazit: Finite Theorie erwartungstreuer Schätzer ist von eingeschränkter Anwendungsrelevanz.

Aber: Es existiert eine analoge asymptotische Theorie mit breiter Anwendungsrelevanz, die sich an finiter Theorie orientiert (siehe Abschnitt 2.1.5).

Zur Konstruktion von UMVU-Schätzern sind folgende zwei Aussagen nützlich:

Satz 2.22 (Rao-Blackwell)

Sei $\mathbf{T} = \mathbf{T}(\mathbf{X})$ *suffizient für* $\boldsymbol{\theta}$ *bzw.* $\mathcal{P}$ *und* $\widehat{\boldsymbol{\theta}}$ *erwartungstreu für* $\boldsymbol{\theta}$*.*
Für den Schätzer

$$\widehat{\boldsymbol{\theta}}_{RB} = \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}|\mathbf{T}] \quad (\text{„Rao-Blackwellization''})$$

gilt:

1. $\widehat{\boldsymbol{\theta}}_{RB}$ *ist erwartungstreu für* $\boldsymbol{\theta}$*.*
2. $\mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}}_{RB}) \leq \mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}})$.
3. *In 2. gilt die Gleichheit, wenn* $\widehat{\boldsymbol{\theta}}$ *nur von* $\mathbf{T}$ *abhängt, d.h.* $\widehat{\boldsymbol{\theta}}_{RB} = \widehat{\boldsymbol{\theta}}$ *mit Wahrscheinlichkeit 1.*

Satz 2.23 (Lehmann-Scheffé)

Ist $\mathbf{T} = \mathbf{T}(\mathbf{X})$ *suffizient und vollständig (also minimalsuffizient) und* $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$ *ein erwartungstreuer Schätzer, so ist*

$$\widehat{\boldsymbol{\theta}}^* = \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}|\mathbf{T}]$$

der mit Wahrscheinlichkeit 1 eindeutig bestimmte UMVU-Schätzer für $\boldsymbol{\theta}$.

Beispiele Satz von Lehmann-Scheffé

- Normalverteilung: Seien $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $\boldsymbol{\theta} = (\mu, \sigma^2)$. $\mathbf{T} = (\sum X_i, \sum X_i^2)$ ist suffizient und vollständig. Äquivalent ist $\widetilde{\mathbf{T}} = (\bar{X}, S^2)$ suffizient und vollständig. Damit gilt:
  - $\bar{X}$ als Schätzer für μ ist UMVU.
  - S^2 als Schätzer für σ^2 ist UMVU, auch wenn die Cramer-Rao-Schranke nicht angenommen wird.
  - $\frac{\bar{X}}{S^2}$ als Schätzer für $\mathbb{E}(\frac{\bar{X}}{S^2})$ is UMVU.

- In Beispiel 2.16 (lineares Modell) ist $\hat{\sigma}^2$ erwartungstreu für σ^2. Außerdem lässt sich ebenfalls zeigen, dass $\hat{\sigma}^2$ eine Funktion einer vollständigen suffizienten Statistik ist.
  Damit ist $\hat{\sigma}^2$ UMVU-Schätzer für σ^2.

Wichtige Schätzer (Momentenschätzer, Shrinkage-Schätzer, ML- und Quasi-ML-Schätzer etc.) sind im Allgemeinen nicht erwartungstreu, besitzen aber günstige asymptotische ($n \to \infty$) Eigenschaften.

Im Folgenden sei

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}(X_1, \ldots, X_n)$$

Schätzer für $\boldsymbol{\theta}$.

Definition 2.24 (Asymptotische Erwartungstreue)

$\widehat{\boldsymbol{\theta}}_n$ *heißt asymptotisch erwartungstreu* $\overset{def}{\Leftrightarrow}$

$$\lim_{n \to \infty} \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}_n] = \boldsymbol{\theta} \quad \textit{für alle } \boldsymbol{\theta}.$$

Definition 2.25 (Konsistenz)

1. $\widehat{\boldsymbol{\theta}}_n$ *ist (schwach) konsistent für* $\boldsymbol{\theta}$ *(in Zeichen:* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}$ *(für alle* $\boldsymbol{\theta}$*))* $\overset{def}{\Leftrightarrow}$

$$\lim_{n \to \infty} \mathbb{P}_\theta(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \leq \varepsilon) = 1 \quad \textit{für alle } \varepsilon > 0 \textit{ und alle } \boldsymbol{\theta}.$$

2. $\widehat{\boldsymbol{\theta}}_n$ *heißt MSE-konsistent für* $\boldsymbol{\theta}$ $\overset{def}{\Leftrightarrow}$

$$\lim_{n \to \infty} \mathrm{MSE}_\theta(\widehat{\boldsymbol{\theta}}_n) = 0 \quad \textit{für alle } \boldsymbol{\theta}.$$

3. $\widehat{\boldsymbol{\theta}}_n$ *ist stark konsistent für* $\boldsymbol{\theta}$ $\overset{def}{\Leftrightarrow}$

$$\mathbb{P}_\theta\left(\lim_{n \to \infty} \widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}\right) = 1 \quad \textit{für alle } \boldsymbol{\theta}.$$

Bemerkung.

1. Aus der (verallgemeinerten) Tschebyscheff-Ungleichung folgt

$$\widehat{\boldsymbol{\theta}}_n \text{ MSE-konsistent } \Rightarrow \widehat{\boldsymbol{\theta}}_n \text{ schwach konsistent.}$$

2. Wegen $\mathrm{MSE}_\theta(\widehat{\theta}_n) = \mathrm{Var}_\theta(\widehat{\theta}_n) + (\mathrm{Bias}_\theta(\widehat{\theta}_n))^2$ folgt:

$$\widehat{\theta}_n \text{ ist MSE-konsistent}$$

$$\Leftrightarrow \mathrm{Var}_\theta(\widehat{\theta}_n) \rightarrow 0 \text{ und } \mathrm{Bias}_\theta(\widehat{\theta}_n) \rightarrow 0 \text{ für alle } \theta$$

(bzw. komponentenweise für Vektoren $\boldsymbol{\theta}$).

Bemerkung fortgeführt.

3. Ist $\widehat{\boldsymbol{\theta}}_n$ konsistent für $\boldsymbol{\theta}$ und g eine stetige Abbildung, so ist auch $g(\widehat{\boldsymbol{\theta}}_n)$ konsistent für $g(\boldsymbol{\theta})$ (Continuous Mapping Theorem/Stetigkeitssatz).
4. Konsistenznachweise bestehen in der Regel in der Anwendung (schwacher) Gesetze großer Zahlen (für i.i.d. Variablen; i.n.i.d. Variablen; abhängige Variablen, z.B. Martingale, Markov-Prozesse, ...).

Beispiel 2.17

1. $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist wegen $\mathbb{E}(\bar{X}_n) = \mu$ und $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \to 0$ für $n \to \infty$ (MSE-)konsistent.
2. $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ und $\widetilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ sind (MSE-)konsistent für σ^2.
3. Mit $g(x) = \sqrt{x}$ folgt, dass

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} \quad \text{und} \quad \widetilde{S}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

konsistent sind für σ.

Beispiel 2.17 fortgeführt

4. $S_n^2/\bar{X}_n$ ist konsistent für σ^2/μ für $\mu > 0$, da mit $\boldsymbol{\theta} = (\mu, \sigma)$ und $g(\boldsymbol{\theta}) = \sigma^2/\mu$ wieder der Stetigkeitssatz benutzt werden kann.

5. $\widehat{\pi}_n$ ist konsistent für π (im Bernoulli-Experiment).

6. $\widehat{\boldsymbol{\beta}}_{KQ}, \widehat{\boldsymbol{\beta}}_{Ridge}$ sind konsistent für $\boldsymbol{\beta}$ im linearen Modell unter gewissen schwachen Annahmen an $\mathbf{X}$, siehe Beispiel 2.19.

Asymptotische Normalität

Viele Schätzer (KQ-, Momenten-, ML-, Quasi-ML-, Bayes-Schätzer) sind unter Regularitätsannahmen asymptotisch normalverteilt. Informell ausgedrückt heißt das: Für große n ist $\widehat{\boldsymbol{\theta}}_n$ nicht nur approximativ erwartungstreu, sondern zusätzlich approximativ normalverteilt, kurz

$$\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N_k(\boldsymbol{\theta}, \mathbf{V}(\boldsymbol{\theta}))$$

mit (approximativer) Kovarianzmatrix

$$\mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}}_n) \overset{a}{\sim} \mathbf{V}(\boldsymbol{\theta}),$$

die durch

$$\widehat{\mathrm{Cov}}_\theta(\widehat{\boldsymbol{\theta}}_n) := \mathbf{V}(\widehat{\boldsymbol{\theta}}_n)$$

geschätzt wird.

In der Diagonalen von $\mathbf{V}(\widehat{\boldsymbol{\theta}}_n)$ stehen dann die (geschätzten) Varianzen

$$\widehat{\mathrm{Var}}(\widehat{\theta}_j) = v_{jj}(\widehat{\boldsymbol{\theta}}_n)$$

der Komponenten $\widehat{\theta}_j, j = 1, \ldots, k$, von $\widehat{\boldsymbol{\theta}}_n$.

$\Rightarrow$ "Üblicher" Output statistischer Software ist

$$\underbrace{\widehat{\theta}_j}_{\text{Schätzer}} \qquad \underbrace{\sqrt{v_{jj}(\widehat{\boldsymbol{\theta}})}}_{\text{Standardfehler}} \qquad \underbrace{\text{t}}_{\text{t-Statistik}} \qquad \underbrace{\text{p}}_{\text{p-Wert}}$$

Beispiel 2.18.

Seien $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F(x|\boldsymbol{\theta})$ mit $\mathbb{E}(X_i) = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$. Aber F sei *nicht* gleich Φ (Normalverteilung), sondern z.B. die Verteilungsfunktion von $B(\pi) = Bin(1, \pi)$ oder $Po(\lambda)$. Für $\bar{X}_n$ gilt

$$\mathbb{E}(\bar{X}_n) = \mu \text{ und } \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Aufgrund des zentralen Grenzwertsatzes folgt

$$\bar{X}_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right),$$

zum Beispiel

$$\bar{X}_n \overset{a}{\sim} N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \quad \text{bei } B(\pi).$$

Beispiel 2.18 fortgeführt.
Genauere Formulierung:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \quad \text{für } n \to \infty,$$

im Beispiel also

$$\sqrt{n}(\bar{X}_n - \pi) \xrightarrow{d} N(0, \pi(1-\pi)) \quad \text{für } n \to \infty$$

bzw.

$$\left.\begin{array}{rcl} \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} & \xrightarrow{d} & N(0,1), \\ \frac{\bar{X}_n - \pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} & \xrightarrow{d} & N(0,1). \end{array}\right\} \begin{array}{c}\text{zentraler}\\ \text{Grenzwertsatz}\end{array}$$

Die $\sqrt{n}$-Normierung ist vor allem bei i.i.d. Stichprobenvariablen geeignet.

Für nicht identisch verteilte Stichprobenvariablen wie zum Beispiel $y_1|\mathbf{x}_1, \ldots, y_n|\mathbf{x}_n$ in Regressionssituationen benötigt man bei $\sqrt{n}$-Normierung Voraussetzungen, die (teilweise) unnötig restriktiv sind.

Besser ist dann eine „Matrix-Normierung" mit Hilfe einer „Wurzel" $\boldsymbol{\mathcal{I}}^{\frac{1}{2}}(\boldsymbol{\theta})$ der Informationsmatrix.

Einschub: Wurzel einer positiv definiten Matrix

- $\mathbf{A} \in \mathbb{R}^{m \times m}$ ist positiv definit, wenn $\mathbf{A}$ symmetrisch ist und $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ für alle $\mathbf{x} \neq \mathbf{0}$ gilt.
- Dann heißt eine Matrix $\mathbf{A}^{\frac{1}{2}}$ *(linke) Wurzel* von $\mathbf{A}$ $\overset{\text{def}}{\Leftrightarrow}$

$$\mathbf{A}^{\frac{1}{2}} \underbrace{(\mathbf{A}^{\frac{1}{2}})^\top}_{=\mathbf{A}^{\frac{\top}{2}}\text{, rechte Wurzel}} = \mathbf{A}.$$

Allerdings ist $\mathbf{A}^{\frac{1}{2}}$ nicht eindeutig, da für eine beliebige orthogonale Matrix $\mathbf{Q}$ auch $\mathbf{A}^{\frac{1}{2}}\mathbf{Q}$ eine linke Wurzel ist:

$$\mathbf{A}^{\frac{1}{2}}\mathbf{Q}(\mathbf{A}^{\frac{1}{2}}\mathbf{Q})^\top = \mathbf{A}^{\frac{1}{2}} \underbrace{\mathbf{Q}\mathbf{Q}^\top}_{=\mathbf{I}_m} \mathbf{A}^{\frac{\top}{2}} = \mathbf{A}.$$

- Zwei gebräuchliche Wurzeln sind:

  1. **Symmetrische Wurzel:** Betrachte die Spektralzerlegung von $\mathbf{A} \in \mathbb{R}^{m \times m}$. Mit der Matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$ der orthonormalen Eigenvektoren als Spalten ist

$$\mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix},$$

     wobei für alle i die $\lambda_i > 0$ die Eigenwerte von $\mathbf{A}$ sind. (Diese Zerlegung ist numerisch aufwändig!) Dann gilt auch

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top = \underbrace{\mathbf{P}\mathbf{\Lambda}^{\frac{1}{2}}}_{=\mathbf{A}^{\frac{1}{2}}} \underbrace{(\mathbf{\Lambda}^{\frac{1}{2}})^\top \mathbf{P}^\top}_{=\mathbf{A}^{\frac{\top}{2}}},$$

     und $\mathbf{A}^{\frac{1}{2}}$ heißt *symmetrische Wurzel von* $\mathbf{A}$.

2. **Cholesky-Wurzel:** Sei $\mathbf{A}^{\frac{1}{2}} := \mathbf{C}$ untere Dreiecksmatrix mit positiven Diagonalelementen und $\mathbf{C}\mathbf{C}^\top = \mathbf{A}$. Dann ist $\mathbf{C}$ die *eindeutig* bestimmte *Cholesky-Wurzel* von $\mathbf{A}$.
   (Diese ist numerisch vergleichsweise einfach zu erhalten!)

- **Anwendungen in der Statistik**
  1. Erzeugen von $N_m(\mathbf{0}, \boldsymbol{\Sigma})$-verteilten Zufallszahlen ($\boldsymbol{\Sigma}$ vorgegeben): Falls $\mathbf{Z} \sim N_m(\mathbf{0}, \mathbf{I}_m)$, ist einfache Simulation möglich, indem m unabhängige $N(0,1)$-verteilte Zufallsvariablen $Z_1, \ldots, Z_m$ simuliert werden.
     Dann gilt auch

     $$\boldsymbol{\Sigma}^{1/2}\mathbf{Z} \ \sim \ N_m(\mathbf{0}, \underbrace{\boldsymbol{\Sigma}^{1/2}\mathbf{I}_m\boldsymbol{\Sigma}^{\top/2}}_{\boldsymbol{\Sigma}}).$$

     Also: Berechne Cholesky-Wurzel von $\boldsymbol{\Sigma}$, ziehe m $N(0,1)$-verteilte Zufallsvariablen $\mathbf{Z} = (Z_1, \ldots, Z_m)^\top$, berechne $\mathbf{Y} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z}$. Dann ist $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$ ein $N_m(\mathbf{0}, \boldsymbol{\Sigma})$-verteilter Zufallsvektor.

2. Matrixnormierung bei asymptotischer Normalverteilung:

Beispiel 2.19 (Asymptotische Normalität des KQ-Schätzers im linearen Modell)
Seien $y_1|\mathbf{x}_1, \ldots, y_n|\mathbf{x}_n$ unabhängig. Dann gilt

$$\mathbb{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad \mathrm{Var}(y_i|\mathbf{x}_i) = \sigma^2, \quad i = 1, \ldots, n,$$

$$\Leftrightarrow \ \mathbf{y}_n = \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \ \ \mathbb{E}[\boldsymbol{\varepsilon}_n] = \mathbf{0}, \ \mathrm{Cov}(\boldsymbol{\varepsilon}_n) = \sigma^2\mathbf{I}_n,$$

wobei hier $\mathbf{X}_n$ die Design-Matrix mit Zeilen $\mathbf{x}_1, \ldots, \mathbf{x}_n$ bezeichnet. Der KQ-Schätzer ist

$$\widehat{\boldsymbol{\beta}}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}\mathbf{X}_n^\top \mathbf{y}_n, \ \ \mathbb{E}[\widehat{\boldsymbol{\beta}}_n] = \boldsymbol{\beta}, \ \ \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_n) = \sigma^2(\mathbf{X}_n^\top \mathbf{X}_n)^{-1}.$$

Die Informationsmatrix unter der Normalverteilungsannahme ist

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \frac{\mathbf{X}_n^\top \mathbf{X}_n}{\sigma^2} = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_n)^{-1}.$$

Beispiel 2.19 fortgeführt

Zentrale Grenzwertsätze (für unabhängige, nicht identisch verteilte Zufallsvariablen, kurz: i.n.i.d.) liefern unter geeigneten Voraussetzungen (informell):

$$\widehat{\boldsymbol{\beta}}_n \overset{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}_n^\top \mathbf{X}_n)^{-1}).$$

Genauere Formulierungen nehmen an, dass

$$\text{(p-)}\lim_{n \to \infty} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n =: \mathbf{A} > 0 \tag{1}$$

existiert (also: $\mathbf{X}_n^\top \mathbf{X}_n \approx n\mathbf{A} \;\Leftrightarrow\; (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \approx \mathbf{A}^{-1}/n$ für große n).

Beispiel 2.19 fortgeführt

Anwendung des (multivariaten) zentralen Grenzwertsatzes liefert dann:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \sigma^2 \mathbf{A}^{-1})$$

bzw.

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_n \quad &\overset{a}{\sim} \quad N_p(\boldsymbol{\beta}, \sigma^2 \mathbf{A}^{-1}/n) \\ \widehat{\boldsymbol{\beta}}_n \quad &\overset{a}{\sim} \quad N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}). \end{aligned}$$

Die Annahme (1) ist zum Beispiel erfüllt, wenn $\mathbf{x}_i,\ i = 1, \dots, n,$ i.i.d. Realisierungen stochastischer Kovariablen $\mathbf{X} = (X_1, \dots, X_p)^\top$ sind.

Beispiel 2.19 fortgeführt

Dann gilt nach dem Gesetz der großen Zahlen:

$$\frac{1}{n}\mathbf{X}_n^\top \mathbf{X}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \ \xrightarrow[n \to \infty]{\mathbb{P}} \ \mathbb{E}[\mathbf{X}\mathbf{X}^\top] =: \mathbf{A}.$$

Typischerweise ist die Annahme (1) *nicht* erfüllt bei deterministischen Regressoren mit Trend. Das einfachste Beispiel hierfür ist ein linearer Trend: $x_i = i$ für $i = 1, \ldots, n$ und $y_i = \beta_1 i + \varepsilon_i$. Dann ist

$$\mathbf{X}_n^\top \mathbf{X}_n = \sum_{i=1}^{n} i^2$$

und daher

$$\frac{1}{n}\mathbf{X}_n^\top \mathbf{X}_n = \frac{\sum_{i=1}^{n} i^2}{n} \geq n \overset{n \to \infty}{\to} \infty.$$

Beispiel 2.19 fortgeführt

In diesem Fall ist eine andere Normierung nötig, zum Beispiel eine Matrixnormierung mit

$$\mathbf{C}_n = (\mathbf{X}_n^\top \mathbf{X}_n).$$

Dann lässt sich die asymptotische Normalität des KQ-Schätzers

$$\mathbf{C}_n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \sigma^2 \mathbf{I}_p)$$

bzw.

$$\tilde{\mathbf{C}}_n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) := \frac{\mathbf{C}_n^{1/2}}{\sigma}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{I}_p)$$

unter folgenden, sehr schwachen Bedingungen zeigen:

Beispiel 2.19 fortgeführt

(D) Divergenzbedingung: Für $n \to \infty$ gilt:

$$(\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \xrightarrow{(\mathbb{P})} \mathbf{0}_{\mathbf{p}\times\mathbf{p}}.$$

Eine äquivalente Forderung ist:

$$\lambda_{\mathsf{min}}(\mathbf{X}_n^\top \mathbf{X}_n) \xrightarrow{(\mathbb{P})} \infty,$$

wobei λ_{min} den kleinsten Eigenwert von $\mathbf{X}_n^\top \mathbf{X}_n$ bezeichnet.
Die Divergenzbedingung sichert, dass die „Informationsmatrix"

$$\mathbf{X}_n^\top \mathbf{X}_n = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$$

für $n \to \infty$ gegen ∞ divergiert, die Information mit $n \to \infty$ also laufend wächst.
Es gilt: (D) ist hinreichend und notwendig für die (schwache und starke) Konsistenz des KQ-Schätzers $\widehat{\boldsymbol{\beta}}_n$.

Beispiel 2.19 fortgeführt

(N) Normalitätsbedingung:

$$\max_{i=1,\dots,n} \mathbf{x}_i^\top (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{x}_i \xrightarrow{(\mathbb{P})} 0 \quad \text{für } n \to \infty.$$

(N) sichert, dass die Information jeder Beobachtung i asymptotisch gegenüber der Gesamtinformation $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$ vernachlässigbar ist.

Beispiel 2.19 fortgeführt

Unter (D) und (N) gilt

$$(\mathbf{X}_n^\top \mathbf{X}_n)^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \overset{d}{\to} N_p(\mathbf{0}, \sigma^2 \mathbf{I}_p)$$

(Beweis mit Grenzwertsätzen für unabhängige, nicht identisch verteilte Zufallsvariablen), d.h. für praktische Zwecke:

$$\widehat{\boldsymbol{\beta}}_n \overset{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}_n^\top \mathbf{X}_n)^{-1})$$

für genügend großen Stichprobenumfang n. Dabei darf zusätzlich σ^2 durch einen konsistenten Schätzer $\widehat{\sigma}^2$ ersetzt werden.

Definition 2.26 (Asymptotische Normalität)

$\widehat{\boldsymbol{\theta}}_n$ *heißt asymptotisch normalverteilt für* $\boldsymbol{\theta} \overset{def}{\Leftrightarrow}$ *es existiert eine Folge von Matrizen* $\mathbf{A}_n$ *mit* $\lambda_{\min}(\mathbf{A}_n) \xrightarrow{(\mathbb{P})} \infty$*, so dass*

$$\mathbf{A}_n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_k(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta})) \quad \textit{für } n \to \infty$$

mit nicht-negativ definiter (in der Regel positiv definiter) Matrix $\mathbf{V}(\boldsymbol{\theta})$.

Für den Spezialfall $\mathbf{A}_n = n\mathbf{I}_k$ *entspricht dies der* $\sqrt{n}$*-Normierung*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_k(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta})) \quad \textit{für } n \to \infty.$$

Bemerkung.

1. Praxisformulierung:

$$\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N_k(\boldsymbol{\theta}, \mathbf{V}(\boldsymbol{\theta})/n)$$

   bzw.

$$\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N_k(\boldsymbol{\theta}, (\mathbf{A}_n^{1/2})^{-1} V(\boldsymbol{\theta})(\mathbf{A}_n^{1/2})^{-\top}).$$

   Dabei darf $\boldsymbol{\theta}$ in $\mathbf{V}(\boldsymbol{\theta})$ durch $\widehat{\boldsymbol{\theta}}_n$ ersetzt werden.

2. Oft: $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{I}_k$ möglich, wenn geeignet normiert wird, zum Beispiel bei ML-Schätzung.

Beispiel 2.20.

Seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit (bekanntem) Erwartungswert μ und Varianz σ^2.

$$S_\mu^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

ist asymptotisch normalverteilt (für σ^2) mit $V(\sigma^2) = \mu_4 - \sigma^4$, $\mu_4 = \mathbb{E}[(X_i - \mu)^4] < \infty$. S_μ^2 ist erwartungstreu.

Beispiel 2.20 fortgeführt.

Für die Varianz erhält man:

$$\begin{aligned}
\mathrm{Var}(S_\mu^2) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) \\
&= \frac{1}{n^2}\cdot n\cdot \mathrm{Var}\left[(X_1-\mu)^2\right] \\
&= \frac{1}{n}\Big(\mathbb{E}[(X_1-\mu)^4] - \big(\mathbb{E}[(X_1-\mu)^2]\big)^2\Big) \\
&= \frac{1}{n}(\mu_4-\sigma^4).
\end{aligned}$$

Es liegen die Voraussetzungen zur Anwendung des zentralen Grenzwertsatzes vor. Aus ihm folgt:

$$S_\mu^2 \overset{a}{\sim} N(\sigma^2, (\mu_4-\sigma^4)/n) \quad \text{bzw.} \quad \sqrt{n}(S_\mu^2-\sigma^2) \xrightarrow{d} N(0, \mu_4-\sigma^4).$$

Die Delta-Methode

$\widehat{\boldsymbol{\theta}}_n$ sei asymptotisch normalverteilter Schätzer für $\boldsymbol{\theta}$.

Frage: Wie ist für eine gegebene Abbildung

$$\mathbf{h}: \mathbb{R}^k \to \mathbb{R}^r, r \leq k$$

der Schätzer $\mathbf{h}(\widehat{\boldsymbol{\theta}})$ für $\mathbf{h}(\boldsymbol{\theta})$ verteilt?

Satz 2.27 (Delta-Methode)

Sei h wie oben.

1. *θ skalar: Für alle θ, für die h stetig differenzierbar ist mit $h'(\theta) \neq 0$, gilt:*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta))$$

$$\Rightarrow \sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, [h'(\theta)]^2 V(\theta))$$

Satz 2.27 (fortgeführt)

2. $\boldsymbol{\theta}$ *vektoriell: Sei* $\mathbf{h}$ *gegeben durch*

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \mapsto \mathbf{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \ldots, h_r(\boldsymbol{\theta}))^\top$$

mit Funktionalmatrix

$$(\mathbf{H}(\boldsymbol{\theta}))_{ij} = \frac{\partial h_i(\boldsymbol{\theta})}{\partial \theta_j}$$

mit vollem Rang r. *Für alle* $\boldsymbol{\theta}$, *für die* $\mathbf{h}(\boldsymbol{\theta})$ *komponentenweise stetig partiell differenzierbar ist und jede Zeile von* $H(\boldsymbol{\theta})$ *ungleich dem Nullvektor ist, gilt:*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_k(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}))$$

$$\Rightarrow \sqrt{n}(\mathbf{h}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{h}(\boldsymbol{\theta})) \xrightarrow{d} N_r(\mathbf{0}, \mathbf{H}(\theta)\mathbf{V}(\theta)\mathbf{H}(\theta)^\top).$$

Beweisskizze für skalares θ

Taylorentwicklung von $h(\widehat{\theta}_n)$ um θ liefert:

$$h(\widehat{\theta}_n) = h(\theta) + (\widehat{\theta}_n - \theta)h'(\theta) + o(\widehat{\theta}_n - \theta)^2.$$

Dabei ist für eine Folge von Zufallsvariablen Z_n

$$Z_n = o(a_n) \quad \text{falls } Z_n/a_n \xrightarrow{\mathbb{P}} 0 \text{ für } n \to \infty.$$

Also:

$$h(\widehat{\theta}_n) \approx h(\theta) + (\widehat{\theta}_n - \theta)h'(\theta)$$

bzw.

$$\sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \approx \sqrt{n}(\widehat{\theta}_n - \theta)h'(\theta)$$

Aus $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta))$ folgt dann, dass
$\sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, h'(\theta)^2 V(\theta))$. $\square$

Beispiel Delta-Methode

Seien $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} B(\pi)$. Für den Schätzer $\widehat{\pi} = \frac{1}{n}\sum_{i=1}^n X_i$ gilt:

$$\sqrt{n}(\widehat{\pi} - \pi) \xrightarrow{d} N(0, \pi(1-\pi))$$

(vgl. Bsp. 2.18). Betrachte nun die Odds $h(\pi) = \frac{\pi}{1-\pi}$ und den Schätzer $h(\widehat{\pi}) = \frac{\widehat{\pi}}{1-\widehat{\pi}}$. Nach der Delta-Regel gilt dann

$$\sqrt{n}\left(\frac{\widehat{\pi}}{1-\widehat{\pi}} - \frac{\pi}{1-\pi}\right) \xrightarrow{d} N\left(0, \frac{\pi}{(1-\pi)^3}\right),$$

und daher $\frac{\widehat{\pi}}{1-\widehat{\pi}} \overset{a}{\sim} N(\frac{\pi}{1-\pi}, \frac{\pi}{n(1-\pi)^3})$, da

$$[h'(\pi)]^2\pi(1-\pi) = \left[\frac{1}{(1-\pi)^2}\right]^2 \pi(1-\pi) = \frac{\pi}{(1-\pi)^3}.$$

Asymptotische Cramer-Rao Schranke und asymptotische Effizienz

Seien $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\boldsymbol{\theta})$ und

$$\mathbf{i}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]$$

die erwartete Fisher-Information einer Beobachtung X_i.
Die Information der gesamten Stichprobe $X_1, \ldots, X_n$ ist dann

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = n \cdot \mathbf{i}(\boldsymbol{\theta}).$$

Satz 2.28 (Asymptotische Cramer-Rao Ungleichung)

Unter Fisher-Regularität sowie leichten Zusatzannahmen gilt:

1. *Aus* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_k(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}))$ *folgt* $\mathbf{V}(\boldsymbol{\theta}) \geq \mathbf{i}^{-1}(\boldsymbol{\theta})$.
2. *Aus* $\sqrt{n}(\mathbf{h}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{h}(\boldsymbol{\theta})) \xrightarrow{d} N_r(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta}))$ *folgt*

$$\mathbf{D}(\boldsymbol{\theta}) \geq \mathbf{H}(\boldsymbol{\theta})\mathbf{i}^{-1}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})^\top$$

mit "$\geq$" die Löwner-Ordnung (und den Bezeichnungen aus der Delta-Regel, Satz 2.27).

Definition 2.29 (Bester asymptotisch normaler (BAN)-Schätzer)

$\widehat{\boldsymbol{\theta}}_n$ heißt BAN-Schätzer, falls in 1. oben gilt:

$$\mathbf{V}(\boldsymbol{\theta}) = \mathbf{i}^{-1}(\boldsymbol{\theta}).$$

Mit der Delta-Regel folgt unmittelbar:

Satz 2.30 (Transformation von BAN-Schätzern)

Ist $\widehat{\boldsymbol{\theta}}_n$ BAN-Schätzer für $\boldsymbol{\theta}$, so ist $\mathbf{h}(\widehat{\boldsymbol{\theta}}_n)$ BAN-Schätzer für $\mathbf{h}(\boldsymbol{\theta})$.

Bemerkung. Das Konzept der asymptotischen Effizienz lässt sich auf die Matrix-Normierung übertragen: $\widehat{\boldsymbol{\theta}}$ ist BAN-Schätzer für $\boldsymbol{\theta}$ genau dann, wenn

$$\boldsymbol{\mathcal{I}}^{1/2}(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_k(\mathbf{0}, \mathbf{I}_k)$$

bzw. $\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N_k(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}}_n))$, mit $\boldsymbol{\mathcal{I}}^{1/2}(\boldsymbol{\theta})$ Wurzel der Fisher-Information $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ der Stichprobe $X_1, \ldots, X_n$. Anstelle der erwarteten kann auch die beobachtete Fisher-Information $\mathbf{J}(\boldsymbol{\theta})$ verwendet werden.