Schätzen und Testen I

Wintersemester 2015/16

Folien zur Vorlesung von

Sonja Greven
Christian Heumann

nach einer Vorlage von

Christiane Fuchs
Ludwig Fahrmeir
Volker Schmid

Worum geht es in dieser Vorlesung?

Statistik umfasst die

- deskriptive Statistik
- explorative Statistik
- induktive Statistik

Hier: Statistische Inferenz

Verschiedene Inferenzkonzepte. In dieser Vorlesung (ST1) Fokus auf

- klassischer Inferenz
- Likelihood-Inferenz
- Bayes-Inferenz

Inhalt von Teil 1 (Sonja Greven)

1. Einführung in statistische Modelle und Inferenzkonzepte
   - Statistische Modelle
   - Konzepte der statistischen Inferenz

2. Klassische Schätz- und Testtheorie
   - Klassische Schätztheorie
   - Klassische Testtheorie
   - Bereichsschätzung und Konfidenzintervalle
   - Multiples Testen

3. Likelihood-Inferenz
   - Parametrische Likelihood-Inferenz
   - Maximum-Likelihood-Schätzung
   - Testen linearer Hypothesen und Konfidenzintervalle
   - Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Inhalt

1 Einführung in statistische Modelle und Inferenzkonzepte

Ziele:

1.1 Statistische Modelle im Überblick, von einfachen hin zu komplexeren Modellen. Auswahl orientiert an Datenstrukturen, Modellklassen und Fragestellungen aus dem Bachelorprogramm und darüber hinaus.

1.1 Problemstellungen der zugehörigen statistischen Inferenz.

1.2 Konzepte statistischer Inferenz im Überblick.

1.1 Statistische Modelle

Überblick

1.1.1 unabhängig identisch verteilter (i.i.d.) Fall

1.1.2-4 bedingt unabhängiger Fall: Regression
Verallgemeinerungen vom einfachen linearen Modell zu

- 1.1.2 größerer Klasse an Verteilungsannahmen (GLM)
- 1.1.3 nichtlinearen Kovariableneffekten (additive Modelle)
- 1.1.4 verteilungsfreien Ansätzen (Quantilsregression)

1.1.5 abhängiger Fall: Beispiel Longitudinaldaten

1.1.6 fehlende/unvollständige Daten

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

Ein-Stichproben-Fall: Seien $x_1, \ldots, x_n$ die Daten als Realisierungen von Stichprobenvariablen $X_1, \ldots, X_n$ und diese i.i.d. wie Zufallsvariable X mit Verteilungsfunktion $F(x)$ bzw. (stetiger, diskreter bzw. allgemeiner „Radon-Nikodym"-) Dichte $f(x)$.

1.1.1. a) Parametrische Modelle

$$X \sim f(x|\boldsymbol{\theta}), \quad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \Theta \subseteq \mathbb{R}^k$$

In der Regel ist k fest und klein im Verhältnis zu n.

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

Beispiel 1.1

1. $X \sim N(\mu, \sigma^2)$; Schätzen/Testen von μ, zum Beispiel Gauß-Test, t-Test; Schätzen/Testen von σ^2.
2. $\mathbf{X} = (X_1, \ldots, X_p)^\top$ mehrdimensional, zum Beispiel $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
3. Analoge Problemstellungen für $X \sim \text{Bin}(N, \pi)$, $X \sim \text{Po}(\lambda), \ldots$ bzw. allgemein $X \sim$ einfache lineare Exponentialfamilie mit natürlichem (skalarem) Parameter θ.

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

Definition 1.1 (Exponentialfamilien)

Eine Verteilungsfamilie heißt Exponentialfamilie $\overset{def}{\Leftrightarrow}$

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

mit $h(\mathbf{x}) \geq 0$ *und*

$$\begin{aligned} b(\boldsymbol{\theta}) &= \log(c(\boldsymbol{\theta})) \\ \mathbf{T}(\mathbf{x}) &= (T_1(\mathbf{x}), \ldots, T_r(\mathbf{x}))^\top \\ \boldsymbol{\gamma}(\boldsymbol{\theta}) &= (\gamma_1(\boldsymbol{\theta}), \ldots, \gamma_r(\boldsymbol{\theta}))^\top. \end{aligned}$$

$\gamma_1, \ldots, \gamma_r$ *heißen die natürlichen oder kanonischen Parameter der Exponentialfamilie (nach Reparametrisierung von* $\boldsymbol{\theta}$ *mit* $\boldsymbol{\gamma}$*).*
Annahme: $1, \gamma_1, \ldots, \gamma_r$ *und* $1, T_1(\mathbf{x}), \ldots, T_r(\mathbf{x})$ *sind linear unabhängig, d.h.* f *ist strikt* r*-parametrisch.*

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

Beispiel Bernoulli-Experiment:

$$\mathbf{X} = (X_1, \ldots, X_n),\ X_i \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, \pi).$$

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

Beispiel Bernoulli-Experiment fortgeführt:

d.h. es liegt eine einparametrische Exponentialfamilie vor mit

$$\begin{aligned} T(\mathbf{x}) &= \sum_{i=1}^{n} x_i \\ \gamma &= \log\left(\frac{\pi}{1-\pi}\right) =: \operatorname{logit}(\pi). \end{aligned}$$

Bemerkung: Eine Verteilungsfamilie heißt *einfache lineare Exponentialfamilie*, falls

$$f(x|\theta) \propto \exp(b(\theta) + \theta x)$$

bzw. (mit Dispersionsparameter ϕ) falls

$$f(x|\theta) \propto \exp(\frac{b(\theta) + \theta x}{\phi}).$$

Beispiel 1.1 fortgeführt

4. **Lokations- und Skalenmodelle:**

$$X \sim F_0\left(\frac{x-a}{b}\right)$$

mit gegebener Verteilungsfunktion $F_0(z)$.
$a \in \mathbb{R}$ heißt Lokationsparameter, $b > 0$ Skalenparameter.
Dichten im stetigen Fall:

$$X \sim \frac{1}{b} f_0\left(\frac{x-a}{b}\right)$$

mit gegebener Dichte $f_0(z)$.

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

Beispiele für Lokations- und Skalenmodelle:

- $X \sim N(a, b^2)$ (Normalverteilung), $f_0(z) = \phi(z)$:

$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{1}{\sqrt{2\pi}b} \exp\left(-\frac{1}{2}\frac{(x-a)^2}{b^2}\right)$$

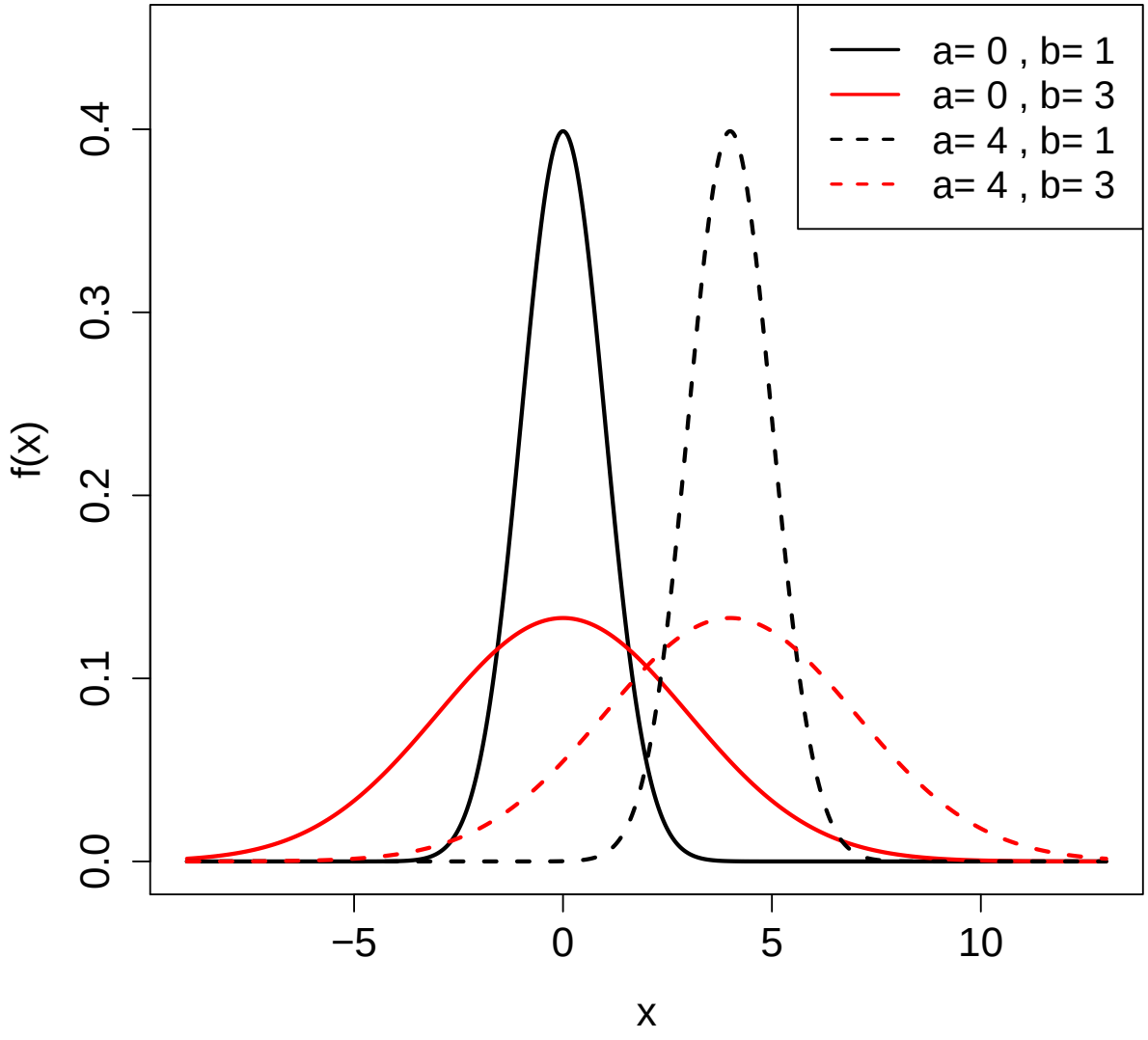
- $X \sim \mathrm{DE}(a, b)$ (Laplace- oder Doppelexponentialverteilung):

$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{1}{2b} \exp\left(-\frac{|x-a|}{b}\right)$$

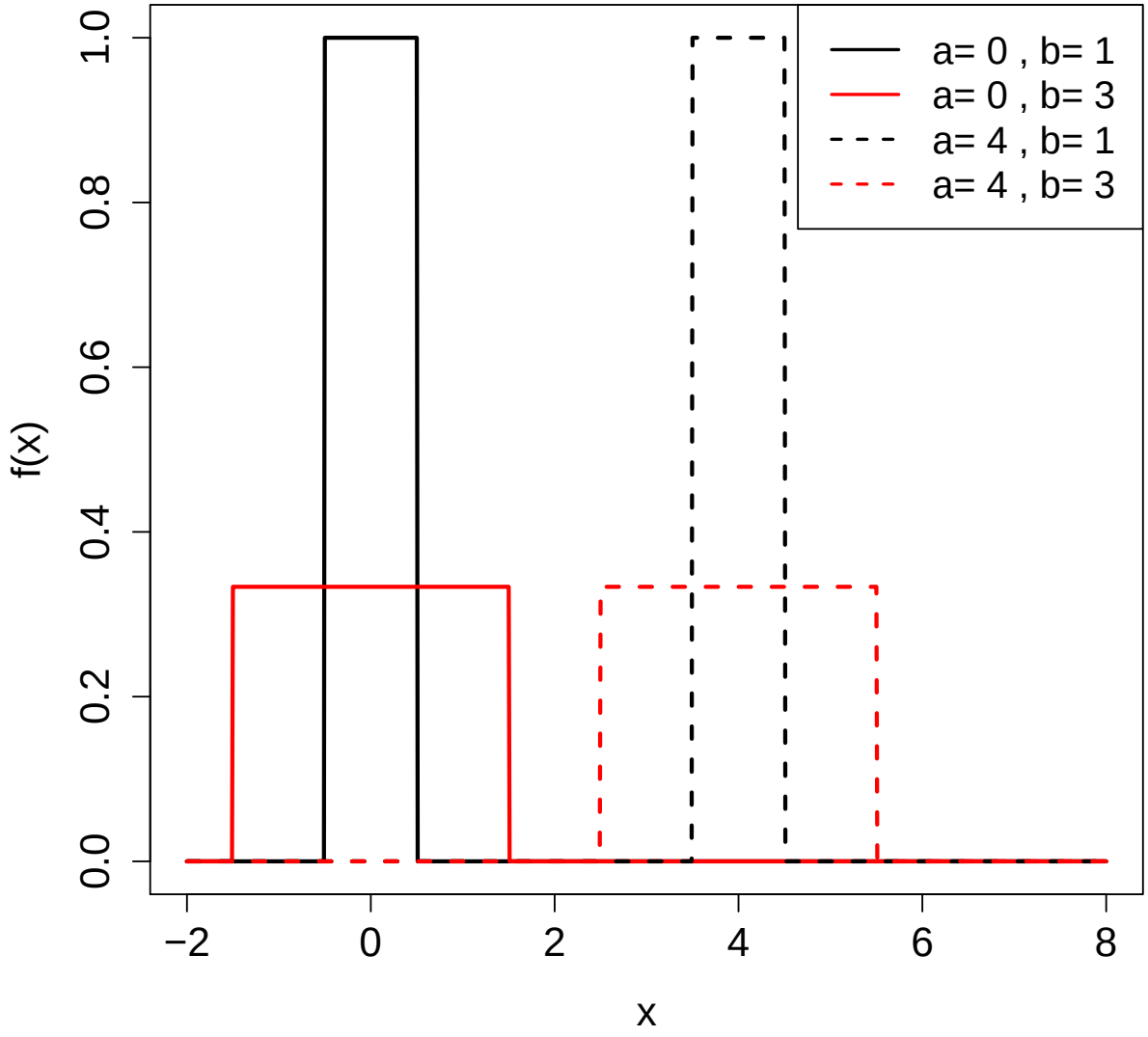
- $X \sim U(a, b)$ (Gleichverteilung):

$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{1}{b} I_{(a-\frac{b}{2}, a+\frac{b}{2})}(x)$$

Der Träger ist abgeschlossen und hängt von den Parametern ab.



Lokations- und Skalenmodelle: Normalverteilung



Lokations- und Skalenmodelle: Gleichverteilung

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

Weitere Beispiele für Lokations- und Skalenmodelle:

- $X \sim C(a, b)$ (Cauchy-Verteilung):

$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{b}{\pi} \cdot \frac{1}{b^2 + (x-a)^2}$$

- $X \sim L(a, b)$ (logistische Verteilung):

$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{1}{b} \cdot \frac{\exp\left(-\frac{x-a}{b}\right)}{\left(1 + \exp\left(-\frac{x-a}{b}\right)\right)^2}$$

- $X \sim E(a, b)$ (Exponentialverteilung):

$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{1}{b} \exp\left(-\frac{x-a}{b}\right) I_{[a,\infty)}(x)$$

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

Beispiel 1.1 fortgeführt

5. **Mischverteilungen**:

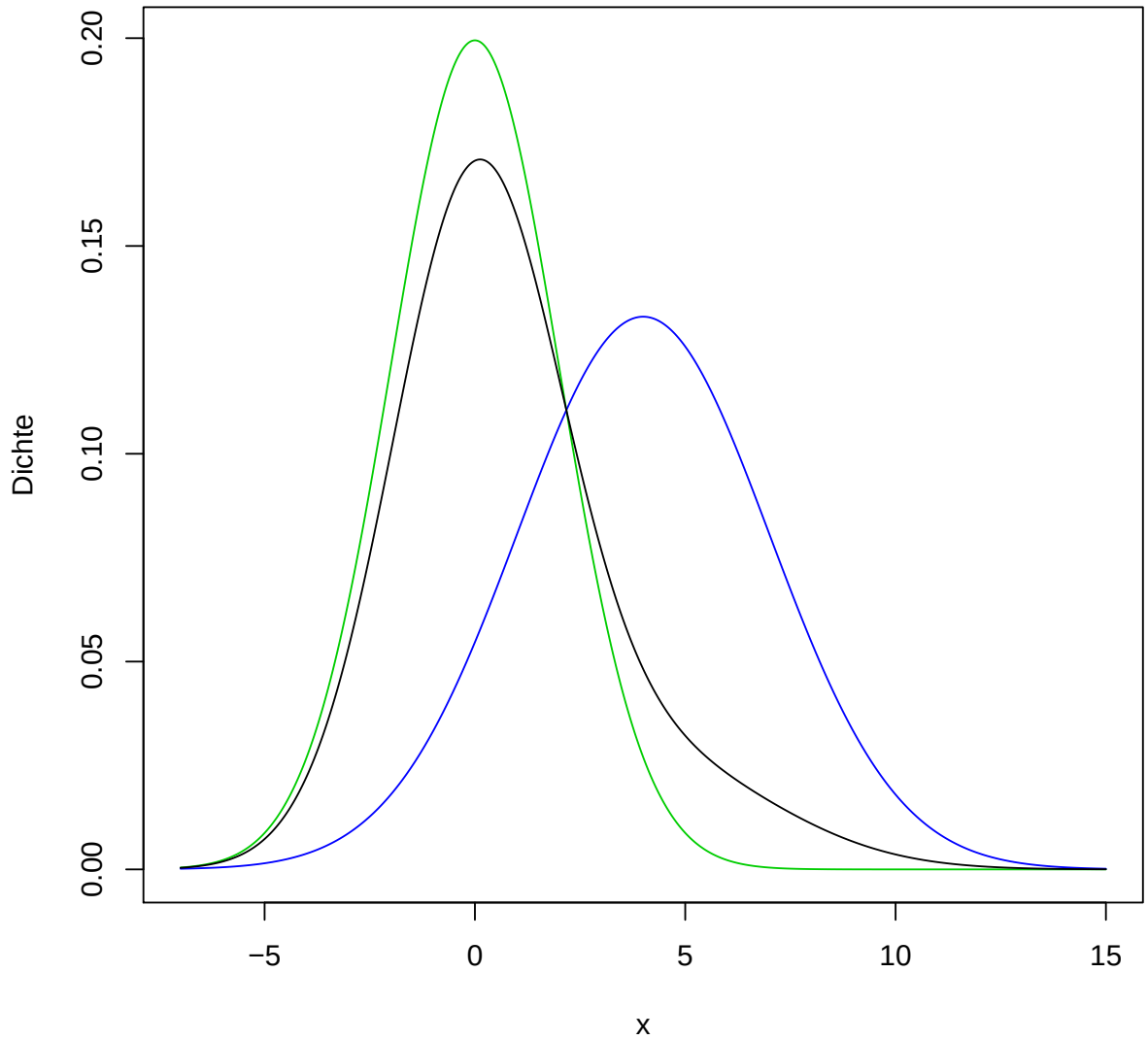
$$X \sim \pi_1 f_1(x|\boldsymbol{\vartheta_1}) + \ldots + \pi_J f_J(x|\boldsymbol{\vartheta_J})$$

mit $\pi_1 + \ldots + \pi_J = 1$, wobei die π_j als *Mischungsanteile* und die $f_j(x|\boldsymbol{\vartheta_j})$ als *Mischungskomponenten* bezeichnet werden. Genauer spricht man von *diskreter Mischung*.

Beispiel Normalverteilungsmischung:

$$X \sim \pi_1 \phi(x|\mu_1, \sigma_1^2) + \ldots + \pi_J \phi(x|\mu_J, \sigma_J^2).$$

Unbekannt sind meistens $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta_1}, \ldots, \boldsymbol{\vartheta_J})$ und $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$. Das Schätzen von $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\pi})$ erfolgt mit ML-Schätzung, meist mit Hilfe des EM-Algorithmus. Auch gewünscht: Testen auf Anzahl J der Mischungskomponenten.



Mischung mit $\pi_1 = 0.8, f_1(x) = \phi(x|0, 2^2), \pi_2 = 0.2, f_2(x) = \phi(x|4, 3^2)$.

1.1.1 b) Nichtparametrische Modelle/Inferenz

- $X \sim F(x)$, X stetige Zufallsvariable, F stetige Verteilung
  $\triangleright$ Kolmogorov-Smirnov-Test auf $H_0 : F(x) = F_0(x)$
- $X \sim F(x)$, X diskret bzw. gruppiert
  $\triangleright$ χ^2-Anpassungstest auf $H_0 : F(x) = F_0(x)$
- $X \sim f(x)$, X stetige Zufallsvariable, f bis auf endlich viele Punkte stetig, differenzierbar etc.
  $\triangleright$ nichtparametrische Dichteschätzung, zum Beispiel Kerndichteschätzung

Der Zwei-und Mehr-Stichprobenfall kann analog behandelt werden; vgl. Statistik II.

Daten $(y_i, \mathbf{x}_i),\ i = 1, \ldots, n$, sind gegeben, mit $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. $y_1|\mathbf{x}_1, \ldots, y_n|\mathbf{x}_n$ sind (bedingt) unabhängig, aber *nicht* identisch verteilt.

1.1.2 a) Klassisches lineares Modell (LM)

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} [N](0, \sigma^2) \Leftrightarrow y_i|\mathbf{x}_i \sim [N](\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$$

- Annahme: $p = \dim(\boldsymbol{\beta}) < n$ und n fest.
- Schätzen von $\boldsymbol{\beta}$ und σ^2, Tests für $\boldsymbol{\beta}$ mit oder ohne Normalverteilungsannahme.
- Variablenselektion und Modellwahl.
  Spezialfall: Varianzanalyse.

1.1.2 b) Generalisierte lineare Modelle (GLM)

$y_i|\mathbf{x}_i,\ i = 1, \ldots, n$, besitzen Dichte aus einfacher linearer Exponentialfamilie, zum Beispiel Normal-, Binomial-, Poisson- oder Gammaverteilung, und sind bedingt unabhängig.

$$\mathbb{E}[y_i|\mathbf{x}_i] = \mu_i = h(\mathbf{x}_i^\top \boldsymbol{\beta})$$

Dabei ist h die *inverse Linkfunktion* (oder *Responsefunktion*).

Die Inferenzprobleme im GLM sind wie im linearen Modell.
Es ist likelihoodbasierte oder bayesianische Inferenz möglich.
Beachte: Die $y_i|\mathbf{x}_i$ sind nicht identisch verteilt.

Beispiel 1.2

Sei $y_i|\mathbf{x}_i \in \{0, 1\}$ und

$$\mu_i = \pi_i = \mathbb{P}(y_i = 1|\mathbf{x}_i), \qquad \pi_i = h(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

Beispiele für h sind die Verteilungsfunktion der logistischen Verteilung ($\rightarrow$ Logit-Modell) oder die Verteilungsfunktion der Normalverteilung ($\rightarrow$ Probit-Modell).

1.1.3 a) Nichtparametrische Einfachregression

Daten wie im linearen Modell, x_i skalar.

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \ \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

Regressionsfunktion $f(x_i) = \mathbb{E}[y_i | x_i]$ *nicht* parametrisch spezifiziert.

- Nicht- oder semiparametrisches Schätzen von f
- Testen von

$$H_0: \quad f(x) = \beta_0 + x\beta_1 \text{ vs.}$$
$$H_1: \quad f \text{ nichtlinear.}$$

1.1.3 b) Additive Modelle (AM)

$$y_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \mathbf{z}_i^\top \boldsymbol{\beta} + \varepsilon_i, \ \varepsilon_i \text{ wie bisher},$$

$$\mu_i = \mathbb{E}[y_i | \mathbf{x}_i, \mathbf{z}_i] = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \mathbf{z}_i^\top \boldsymbol{\beta}$$

mit Kovariablenvektoren $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ und $\mathbf{z}_i$.

- Schätzen, Testen von $f_1, \ldots, f_p, \boldsymbol{\beta}$
- Variablenselektion und Modellwahl (zum Beispiel Einfluss einer bestimmten Kovariable linear oder nichtlinear)

1.1.3 c) Generalisierte Additive Modelle (GAM)

$y_i|\mathbf{x}_i$ wie bei GLM; analog zu additiven Modellen lässt man aber

$$\mu_i = \mathbb{E}[y_i|\mathbf{x}_i, \mathbf{z}_i] = h\left(f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \mathbf{z}_i^\top \boldsymbol{\beta}\right)$$

zu.

1.1 Statistische Modelle

1.1.4 Quantil-Regression/Robuste Regression

Daten wie im linearer Modell ($y_i|\mathbf{x}_i$ bedingt unabhängig). Keine **Annahmen** an die Fehlerverteilung, „verteilungsfreier Ansatz"

Ziel: Schätze nicht (nur) $\mathbb{E}[y_i|\mathbf{x}_i]$, sondern den bedingten Median ($\tau = 0.5$) oder allgemeiner die (bedingten) Quantile $Q_\tau(y_i|\mathbf{x}_i)$. Statt $\widehat{\boldsymbol{\beta}}_{\text{KQ}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$ mit $\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{\text{KQ}} = \widehat{\mathbb{E}}(y|\mathbf{x})$ suche z.B. (für $\tau = 0.5$)

$$\widehat{\boldsymbol{\beta}}_{\text{med}} := \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^n |y_i - \mathbf{x}_i^\top \boldsymbol{\beta}|$$

$\Rightarrow \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{\text{med}} = \widehat{\text{med}}(y|\mathbf{x})$.

Frage: Welche Konzepte zum Schätzen und Testen verwenden?
$\rightarrow$ Quasi-Likelihood-Methoden.

1.1 Statistische Modelle

1.1.5 Abhängige Daten: Beispiel Longitudinaldaten

- **Longitudinaldaten** (Längsschnittdaten): (y_{ij}, x_{ij}) für $i = 1, \ldots, m$ und $j = 1, \ldots, n_i$ als Beobachtungen von Zielvariablen y_{ij} und Kovariablen x_{ij} zu Zeitpunkten $t_{i1} < \ldots < t_{ij} < \ldots < t_{in_i}$. Speziallfall $m = 1$: Zeitreihen.

- **1.1.5 a) Autoregressive bzw. Markov-Modelle**: Bedingte Verteilung von $y_{ij}|y_{i,j-1}, y_{i,j-2}, \ldots, y_{i1}, \mathbf{x}_{ij}$ ist (bei Markov-Modell 1. Ordnung) $y_{ij}|y_{i,j-1}, \mathbf{x}_{ij}$, zum Beispiel

$$y_{ij} = \alpha y_{i,j-1} + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \underbrace{\varepsilon_{ij}}_{\text{i.i.d.}} .$$

  Likelihood-Inferenz: algorithmisch simpel, asymptotische Theorie schwieriger (da y_{ij} abhängig).

1.1 Statistische Modelle

1.1.5 Abhängige Daten: Beispiel Longitudinaldaten

1.1.5 b) Lineares gemischtes Modell (LMM): Z.B.

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}$$

- $\beta_0, \beta_1, \boldsymbol{\beta}$: feste Populationseffekte, z.B. $\beta_0 + \beta_1 t$ fester (linearer) Populationstrend
- b_{0i}, b_{1i}: individuenspezifische Effekte
  $\Rightarrow$ Anzahl der Parameter von der Ordnung des Stichprobenumfangs
- Annahme:

$$b_{0i} \overset{\text{i.i.d.}}{\sim} N(0, \tau_0^2),$$
$$b_{1i} \overset{\text{i.i.d.}}{\sim} N(0, \tau_1^2)$$

  d.h. die b-Parameter sind „zufällige" Parameter.
- Inferenz: algorithmisch/methodisch variierte Likelihood-Inferenz oder Bayes-Inferenz mit MCMC-Simulationsmethoden. Für GLMM deutlich komplexer als für LMM.

1.1.5 c) Marginale Modelle

→ Kapitel 6.2 und 6.4 bzw. kurze Einführung in 3.4 (Quasi-Likelihood-Inferenz/GEEs)

1.1 Statistische Modelle

1.1.6 Fehlende/unvollständige Daten

- Daten: „beliebig" (Querschnitts-, Survival-, Längsschnittdaten)
- Beispiele:
  - Nicht-Antworter bei statistischen Befragungen
  - „Drop-out" bei klinischen Studien
  - zensierte Daten (wie in Survivalanalyse)
  - Modelle mit latenten Variablen
- Übliche Modelle und statistische Methodik setzen vollständige Daten voraus.

1.1 Statistische Modelle

1.1.6 a) Verweildaueranalyse: Cox-Modell

Rechtszensierte Survivaldaten: Evtl. rechtszensierte Beobachtungen $t_1, \ldots, t_n$ von unabhängigen stetigen Lebensdauern $T_1, \ldots, T_n \geq 0$, Zensierungsindikatoren $\delta_1, \ldots, \delta_n$ und zugehörige Kovariablen $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Ziel: Schätze $\lambda(t; \mathbf{x})$ bzw. zumindest den Einfluss der Kovariablen auf die Hazardrate

$$\lambda(t; \mathbf{x}) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t; \mathbf{x})}{\Delta t}.$$

Interpretation: $\lambda(t; \mathbf{x})\Delta t \approx$ bedingte Wahrscheinlichkeit für Ausfall in $[t, t + \Delta t]$ gegeben „Überleben“ bis zum Zeitpunkt t bei „kleinem” Δt.

1.1 Statistische Modelle

1.1.6 a) Verweildaueranalyse: Cox-Modell

Cox-Modell (auch: *Proportional Hazards-Modell*)

$$\begin{aligned}\lambda(t; \mathbf{x}_i) &= \lambda_0(t) \cdot \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\ &= \lambda_0(t) \cdot \exp(x_{i1}\beta_1) \cdot \ldots \cdot \exp(x_{ip}\beta_p).\end{aligned}$$

Primäres Interesse: Schätzen/Testen von $\boldsymbol{\beta}$.
Die von i unabhängige „Baseline"-Hazardrate $\lambda_0(t)$ wird als Nuisanceparameter (bzw. -funktion) betrachtet.

$\Rightarrow$ Die Likelihood faktorisiert sich in

$$L(\boldsymbol{\beta}; \lambda_0(t)) = L_1(\boldsymbol{\beta}) \cdot L_2(\boldsymbol{\beta}; \lambda_0(t)).$$

Die partielle Likelihood $L_1(\boldsymbol{\beta})$ wird bzgl. $\boldsymbol{\beta}$ maximiert.
Erstaunlicherweise ist der Informationsverlust gering.

1.1 Statistische Modelle

1.1.6 b) Modellbasierte Clusteranalyse

- Idee: $\mathbf{x} = (x_1, \ldots, x_p)^\top$ stammt aus multivariater Mischverteilung

$$f(\mathbf{x}) = \sum_{j=1}^{J} \pi_j f_j(\mathbf{x}|\boldsymbol{\vartheta_j}),$$

  z.B. f_j Dichte der multivariaten Normalverteilung.

- Gesucht:
  1. Schätzungen für $\boldsymbol{\vartheta_j}, \pi_j,\ j = 1, \ldots, J$.
  2. Schätzungen für unbekannte Klassenzugehörigkeit j eines Objekts mit Merkmalsvektor $\mathbf{x}$. Formel von Bayes liefert:

$$\widehat{\pi}(j|\mathbf{x}) = \frac{\widehat{\pi}_j f_j(\mathbf{x}|\widehat{\boldsymbol{\vartheta}_\mathbf{j}})}{\widehat{f}(\mathbf{x})}.$$

- Likelihood-Maximierung: mit EM-Algorithmus
  Bayes: mit MCMC-Algorithmus

1.2 Konzepte der statistischen Inferenz

- $\mathbf{x} = (x_1, \ldots, x_n)^\top$ oder $\mathbf{y} = (y_1, \ldots, y_n)^\top$ sind Realisierungen von Stichprobenvariablen (Zufallsvariablen)
  $\mathbf{X} = (X_1, \ldots, X_n)^\top$ oder $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$.
  Die Komponenten $X_1, \ldots, X_n$ können auch selbst wieder mehrdimensional sein.

- Weitere Annahmen:
  - $X_1, \ldots, X_n$ i.i.d. wie $X \rightarrow$ einfache Zufallsstichprobe (vgl. Abschnitt 1.1.1).
  - $Y_1, \ldots, Y_n$ (bzw. $Y_1|X_1, \ldots, Y_n|X_n$ im Regressionsmodell) sind (bedingt) unabhängig aber *nicht* identisch verteilt (vgl. Abschnitte 1.1.2-1.1.4).
  - $Y_1, \ldots, Y_n$ sind abhängig, zum Beispiel zeitlich oder räumlich korreliert (vgl. 1.1.5).

1.2 Konzepte der statistischen Inferenz

- In allen Fällen gilt: $\mathbf{x} \in \mathcal{X}$ bzw. $\mathbf{y} \in \mathcal{Y}$, wobei $\mathcal{X}$ bzw. $\mathcal{Y}$ der entsprechende Stichprobenraum ist.

  $\mathbf{X} = (X_1, \ldots, X_n)^\top$ bzw. $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ sind auf dem Stichprobenraum nach einer gemeinsamen Verteilung $\mathbb{P}$ bzw. Verteilungsfunktion $F(\mathbf{x}) = F(x_1, \ldots, x_n)$ bzw. $F(\mathbf{y})$ verteilt.

  $\mathbb{P}$ (bzw. F) gehört einer Menge (oder Klasse oder Familie) von Verteilungen $\mathcal{P}_\theta = \{\mathbb{P}_\theta : \boldsymbol{\theta} \in \Theta\}$ an. Zugehörige Verteilungsfunktionen sind $F(\mathbf{x}|\boldsymbol{\theta})$ bzw. (falls existent) Dichten $f(\mathbf{x}|\boldsymbol{\theta}) = f(x_1, \ldots, x_n|\boldsymbol{\theta})$ (analog für $\mathbf{y}$).

1.2 Konzepte der statistischen Inferenz

Gemeinsame Dichten $f(\mathbf{x}|\boldsymbol{\theta})$:

- i.i.d. Fall:

$$f(\mathbf{x}|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta}) \cdot \ldots \cdot f(x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

- Unabhängige Zufallsvariablen $Y_1, \ldots, Y_n$:

$$f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}),$$

  die Dichten hängen also vom Index i ab.

- Bei potentiell abhängigen $Y_1, \ldots, Y_n$ ist $f(\mathbf{y}|\boldsymbol{\theta})$ nicht immer faktorisierbar und teils auch analytisch schwer oder nicht darstellbar.

1.2 Konzepte der statistischen Inferenz

- (Übliche) **parametrische** Inferenz:
  $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \Theta \subseteq \mathbb{R}^k$, k fest mit $k < n$.

- **Nichtparametrische/verteilungsfreie** Inferenz:
  Θ ist Funktionenraum, θ eine bestimmte Funktion. Zum Beispiel ist Θ der Raum der stetigen oder differenzierbaren Funktionen.

  Beispiele für Methoden: (Kern-)Dichteschätzung, nichtparametrische Regression.

1.2 Konzepte der statistischen Inferenz

- **Semiparametrische** Inferenz (vgl. Kapitel 7):
  Begriff wird verwendet für
  1. Θ hat eine endlich-dimensionale und eine unedlich-dimensionale Komponente.
     Beispiel: Cox-Proportional-Hazard-Modell.

  2. Parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ hochdimensional und k wächst mit n (unter Umständen $k \sim n$), zum Beispiel bei der semiparametrischen Regression mit Glättungssplines.

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

$\mathbf{X} = (X_1, \ldots, X_n)$ besitzt Verteilung
$\mathbb{P} \in \mathcal{P} = \{\mathbb{P}_\theta : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \Theta\}$ mit $\Theta \subseteq \mathbb{R}^k$ und $k < n$ fest, oft $k \ll n$.

In der Regel existiert zur Verteilung $\mathbb{P}_\theta$ eine (diskrete oder stetige bzw. Radon-Nikodym-) Dichte

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

Anmerkung: Allgemein ist dies die Radon-Nikodym-Dichte bezüglich eines dominierenden Maßes, vgl. Maß- und Wahrscheinlichkeitstheorie-Vorlesung.

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **Punktschätzung**:
  Geschätzt werden soll $\boldsymbol{\theta}$. Eine messbare Abbildung

$$\mathbf{T} : \begin{cases} \mathcal{X} & \longrightarrow & \Theta \\ \mathbf{x} & \longmapsto & \mathbf{T}(\mathbf{x}) =: \widehat{\boldsymbol{\theta}} \end{cases}$$

  heißt *Schätzfunktion* oder *Schätzer*.

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **Punktschätzung** fortgeführt:
  Eine Beurteilung der Güte/Optimalität kann z. B. durch
  - $\mathrm{Bias}_\theta(\mathbf{T}) = \mathbb{E}_\theta[\mathbf{T}] - \boldsymbol{\theta}$,
  - $\mathrm{Var}_\theta(\mathbf{T}) = \mathbb{E}_\theta[(\mathbf{T} - \mathbb{E}_\theta[\mathbf{T}])(\mathbf{T} - \mathbb{E}_\theta[\mathbf{T}])^\top]$,
  - $\mathrm{MSE}_\theta(\mathbf{T}) = \mathbb{E}_\theta[\|\mathbf{T} - \boldsymbol{\theta}\|^2] = \mathrm{Spur}(\mathrm{Var}_\theta(\mathbf{T})) + \|\mathrm{Bias}_\theta(\mathbf{T})\|^2$

  erfolgen.

  Das Konzept der „Güte" ist frequentistisch, da beurteilt wird, wie „gut" $\mathbf{T} = \mathbf{T}(\mathbf{X})$ bei „allen" denkbaren wiederholten Stichproben $\mathbf{x}$ als Realisierung von $\mathbf{X}$ „im Schnitt" funktioniert. Anders ausgedrückt: Beurteilt wird nicht die konkret vorliegende Stichprobe, sondern (in der Häufigkeitsinterpretation) das „Verfahren" $\mathbf{T} = \mathbf{T}(\mathbf{X})$.

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **Bereichsschätzung / Intervallschätzung:**

$$C : \begin{cases} \mathcal{X} & \longrightarrow & \mathcal{P}(\Theta) \\ \mathbf{x} & \longmapsto & C(\mathbf{x}) \subseteq \Theta \end{cases}$$

so dass $\mathbb{P}_\theta(C(\mathbf{X}) \ni \boldsymbol{\theta}) \geq 1 - \alpha$ für alle $\boldsymbol{\theta} \in \Theta$.

Dabei ist $1 - \alpha$ der *Vertrauensgrad* (auch: Konfidenzniveau oder Überdeckungswahrscheinlichkeit) des *Konfidenzbereiches*.

Man beachte die frequentistische/Häufigkeitsinterpretation: $C(\mathbf{X})$ ist ein *zufälliger* Bereich.

Ist $\Theta \subseteq \mathbb{R}$ und $C(\mathbf{x})$ für alle $\mathbf{x}$ ein Intervall, dann heißt C *Konfidenzintervall*.

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **Testen**: Mit einem Test ϕ soll eine Hypothese H_0 gegen eine Alternativhypothese H_1 geprüft werden:

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \in \Theta_1,$$

wobei $\Theta_0 \cap \Theta_1 = \emptyset$.
Es gilt oft, aber nicht notwendigerweise, $\Theta = \Theta_0 \cup \Theta_1$.

Ergebnisse/Aktionen:

$A_0 : H_0$ wird nicht abgelehnt,

$A_1 : H_1$ wird bestätigt, das Ergebnis „ist signifikant".

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **Testen** fortgeführt:
  Der Test ist eine Abbildung

$$\phi : \mathcal{X} \rightarrow \{A_0, A_1\} = \{0, 1\}.$$

Ein nicht-randomisierter Test hat die Form

$$\phi(\mathbf{x}) = \begin{cases} 1, \text{ falls } \mathbf{x} \in K, \\ 0, \text{ falls } \mathbf{x} \notin K. \end{cases}$$

Dabei ist $K \subset \mathcal{X}$ der sogenannte *kritische Bereich* und als eine Teilmenge aller möglichen Stichproben zu verstehen. Oft formuliert man dies über eine Teststatistik $T(\mathbf{x})$:

$$\phi(\mathbf{x}) = \begin{cases} 1, \text{ falls } T(\mathbf{x}) \in C, \\ 0, \text{ falls } T(\mathbf{x}) \notin C. \end{cases}$$

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **Testen** fortgeführt:
  Test zum Niveau („size") α, wobei α „klein":

$$\mathbb{P}_\theta(A_1) \leq \alpha \text{ für alle } \boldsymbol{\theta} \in \Theta_0.$$

Dabei ist die Wahrscheinlichkeit für den *Fehler 1. Art* kleiner als α. Die Funktion

$$g_\phi(\boldsymbol{\theta}) = \mathbb{P}_\theta(A_1) = \mathbb{E}_\theta[\phi(\mathbf{X})]$$

heißt *Gütefunktion* von ϕ. Synonym zum Begriff Güte werden auch die Begriffe *Power* oder *Macht* gebraucht. Die Forderung für den Fehler formuliert über die Gütefunktion lautet

$$g_\phi(\boldsymbol{\theta}) \leq \alpha \text{ für } \boldsymbol{\theta} \in \Theta_0.$$

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **Testen** fortgeführt:
  „Programm" der klassischen parametrischen Schätztheorie (siehe Kapitel 2):
  Finde Test ϕ zum Niveau α mit „optimaler" Power bzw. minimaler Wahrscheinlichkeit für den *Fehler 2. Art*, $1 - g_\phi(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta_1$. Das Konzept ist wiederum frequentistisch.

  Das „Programm" ist dabei hauptsächlich für spezielle Verteilungsfamilien (zum Beispiel für Exponentialfamilien) und spezielle Testprobleme im i.i.d. Fall durchführbar. Weniger tauglich ist es für (etwas) komplexere Modelle, zum Beispiel bereits für GLMs. Dann:
  - $\rightarrow$ Likelihood-Inferenz
  - $\rightarrow$ Bayes-Inferenz
  - $\rightarrow$ Nicht- und semiparametrische Inferenz

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **Testen** fortgeführt:
  Im einfachsten Fall von zwei Punkthypothesen

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \qquad H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$$

  für $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$ hat der „beste" Test
  Likelihood-Quotienten-Struktur: H_0 wird abgelehnt, falls

$$\frac{f(\mathbf{x}|\boldsymbol{\theta}_1)}{f(\mathbf{x}|\boldsymbol{\theta}_0)} > k_\alpha$$

  (vgl. Neyman-Pearson Theorem, Abschnitt 2.2 oder Wahrscheinlichkeitstheorie und Inferenz II).

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

- **p-Werte (p-values):**

Beispiel 1.3 (Gauß-Test)
$X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$, σ^2 bekannt. Betrachte

$$H_0 : \mu \leq \mu_0 \quad , \quad H_1 : \mu > \mu_0.$$

Teststatistik ist

$$T(\mathbf{X}) = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \overset{\mu=\mu_0}{\sim} N(0,1).$$

H_0 wird abgelehnt, wenn $T(\mathbf{x}) > z_{1-\alpha}$. Der p-Wert ist
$p = \mathbb{P}(T(\mathbf{X}) > T(\mathbf{x})|\mu = \mu_0) = \sup_\mu \mathbb{P}(T(\mathbf{X}) > T(\mathbf{x})|H_0)$.
Offensichtlich gilt:

$$T(\mathbf{x}) > z_{1-\alpha} \Leftrightarrow p < \alpha.$$

Der p-Wert liefert mehr Information (nämlich wie nahe man an der Entscheidungsgrenze ist) als die reine „Bekanntgabe" der Entscheidung.

Definition 1.2 (p-Wert)

Gegeben sei ein Test bzw. eine Teststatistik $T(\mathbf{X})$ *für* H_0 *vs.* H_1 *mit*

1. $\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{P}_\theta(T(\mathbf{X}) \in C_\alpha | H_0) \leq \alpha$,
2. *für* $\alpha \leq \alpha'$ *gilt* $C_\alpha \subseteq C_{\alpha'}$.

Dann gilt $p = \inf\{\widetilde{\alpha} : T(\mathbf{x}) = t \in C_{\widetilde{\alpha}}\}$, *und* H_0 *wird abgelehnt, falls* $p < \alpha$.

1.2 Konzepte der statistischen Inferenz

1.2.1 Klassische parametrische Inferenz

Beispiel 1.3 (Gauß-Test) fortgeführt:

1.2 Konzepte der statistischen Inferenz

1.2.2 (Parametrische) Likelihood-Inferenz

- Sei $\mathcal{P} = \{f(\mathbf{x}|\boldsymbol{\theta})|\boldsymbol{\theta} \in \Theta\}$, d.h. es existieren Dichten zu der vorgegebenen parametrisierten Verteilungsfamile $\mathcal{P}$. Nach der Beobachtung von $\mathbf{X} = \mathbf{x}$ heißt

$$L(\boldsymbol{\theta}|\mathbf{x}) := f(\mathbf{x}|\boldsymbol{\theta})$$

Likelihoodfunktion von $\boldsymbol{\theta}$ zur Beobachtung $\mathbf{x}$.

- Likelihoodprinzip: Besitzen zwei Beobachtungen $\mathbf{x}$ und $\tilde{\mathbf{x}}$ zueinander proportionale Likelihoodfunktionen, sollen sie zu denselben Schlüssen über $\boldsymbol{\theta}$ führen.

  Beispiel 1.4: $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$, σ^2 bekannt.

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

  Zwei Beobachtungen $\mathbf{x}$ und $\mathbf{y}$ mit $\bar{x} = \bar{y}$ führen nach dem Likelihood-Prinzip zu den gleichen Schlüssen über μ.

1.2 Konzepte der statistischen Inferenz

1.2.2 (Parametrische) Likelihood-Inferenz

- Punktschätzung: Maximum-Likelihood- (ML-) Schätzung

$$\mathbf{T}(\mathbf{x}) = \widehat{\boldsymbol{\theta}}_{\text{ML}} \text{ mit } f(\mathbf{x}|\widehat{\boldsymbol{\theta}}_{\text{ML}}) = \max_{\boldsymbol{\theta}} f(\mathbf{x}|\boldsymbol{\theta})$$

  bzw.

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} = \operatorname*{argmax}_{\boldsymbol{\theta}} f(\mathbf{x}|\boldsymbol{\theta}).$$

- In der Regel existieren keine finiten Optimalitätseigenschaften, jedoch asymptotische.

- Testen: Likelihood-Quotienten-Test, Wald-Test, Score-Test.

- Quasi-Likelihood-Inferenz (Kapitel 3.4 bzw. 6)
- penalisierte Likelihood-Inferenz (u.a. Kapitel 7.4),
- semiparametrische Modelle (Kapitel 7).

1.2 Konzepte der statistischen Inferenz

1.2.4 Bayes-Inferenz

Wir betrachten wieder $\mathcal{P} = \{f(\mathbf{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, zusätzlich wird aber die Unsicherheit über $\boldsymbol{\theta}$ durch die *Prioridichte* $p(\boldsymbol{\theta})$ auf Θ bewertet. Dabei kann Θ auch sehr hochdimensional sein.

- Prinzip: Nach Beobachtung von $\mathbf{x}$ ist sämtliche Information über $\boldsymbol{\theta}$ enthalten in der *Posterioridichte*

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int f(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \overset{\substack{\text{proportional bzgl.} \\ \text{Parameter } \boldsymbol{\theta}}}{\propto} f(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})$$

$$= L(\boldsymbol{\theta}|\mathbf{x}) \cdot p(\boldsymbol{\theta}).$$

1.2 Konzepte der statistischen Inferenz

1.2.4 Bayes-Inferenz

- Bayes-Schätzung:

  - Posteriori-Erwartungswert:

$$\mathbf{T}_{\mathbb{E}}(\mathbf{x}) = \widehat{\boldsymbol{\theta}}_{\text{post-EW}} = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}(\boldsymbol{\theta}|\mathbf{x}) = \int_{\Theta} \boldsymbol{\theta}\ p(\boldsymbol{\theta}|\mathbf{x})\ d\boldsymbol{\theta}$$

  - Posteriori-Median:

$$\mathbf{T}_{\text{med}}(\mathbf{x}) = \widehat{\boldsymbol{\theta}}_{\text{post-Med}} = \text{med}_{\boldsymbol{\theta}|\mathbf{x}}(\boldsymbol{\theta}|\mathbf{x})$$

  - Posteriori-Modus:

$$\mathbf{T}_{\text{mod}}(\mathbf{x}) = \widehat{\boldsymbol{\theta}}_{\text{post-Mod}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\boldsymbol{\theta}|\mathbf{x}) = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})$$

1.2 Konzepte der statistischen Inferenz

1.2.4 Bayes-Inferenz

- Es sind auch *uneigentliche Prioriverteilungen* zulässig, d.h. Dichten mit

$$\int_{\Theta} p(\boldsymbol{\theta}) d\boldsymbol{\theta} = +\infty,$$

die sich somit nicht normieren lassen. Allerdings muss die Posterioridichte eigentlich sein!

Ein Spezialfall ist $p(\boldsymbol{\theta}) \propto 1$ („Gleichverteilungs-Priori"), bei deren Verwendung

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, L(\boldsymbol{\theta}|\mathbf{x}) = \widehat{\boldsymbol{\theta}}_{\text{post-Mod}}$$

gilt, d.h. der ML-Schätzwert und der Posteriori-Modus-Schätzwert identisch sind.

- Die Verteilung zu $p(\boldsymbol{\theta})$ heißt die *konjugierte Verteilung* für $f(\mathbf{x}|\boldsymbol{\theta})$, wenn $f(\boldsymbol{\theta}|\mathbf{x})$ (posteriori) und $f(\boldsymbol{\theta})$ (priori) dieselbe Form haben, d.h. wenn Priori- und Posterioverteilung zur selben Verteilungsfamilie gehören.

1.2 Konzepte der statistischen Inferenz

1.2.4 Bayes-Inferenz

- Bayes-Bereichsschätzung: Wähle *Kredibilitätsbereiche/-intervalle* $C(\mathbf{x})$ so, dass

$$\int_{C(\mathbf{x})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = \mathbb{P}_{\boldsymbol{\theta}|\mathbf{x}}\Big(\underbrace{\boldsymbol{\theta}}_{\text{zufällig}} \in \underbrace{C(\mathbf{x})}_{\substack{\text{nicht zufällig,}\\ \text{deterministisch}}}\Big) \geq 1-\alpha.$$

  Es ist also eine Wahrscheinlichkeitsaussage für eine konkrete Stichprobe möglich und keine Häufigkeitsinterpretation notwendig!

- Bei Bayes-Inferenz wird keine Häufigkeitsinterpretation *benötigt*. Allerdings kann sie trotzdem gemacht werden. ($\rightarrow$ Asymptotik der Bayes-Schätzer)

1.2 Konzepte der statistischen Inferenz

1.2.5 Statistische Entscheidungstheorie

Sichtweise in der Entscheidungstheorie: Schätzen und Testen als Entscheidung unter Unsicherheit.

Wie bisher betrachten wir $\mathbb{P} \in \mathcal{P}_\theta = \{\mathbb{P}_\theta : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \Theta\}$ als statistisches Modell; $\mathbf{x}$ bezeichne eine Stichprobe / konkrete Beobachtung von $\mathbf{X}$.

Zusätzlich werden folgende Funktionen betrachtet:

Definition 1.3 (Entscheidungsfunktion)

Als Entscheidungsfunktion bezeichnet man eine Funktion

$$\mathbf{d} : \begin{cases} \mathcal{X} & \longrightarrow & D \\ \mathbf{x} & \longmapsto & \mathbf{d}(\mathbf{x}). \end{cases}$$

Mit D wird der Entscheidungs- oder Aktionenraum bezeichnet.

Definition 1.4 (Verlustfunktion)

Eine Verlustfunktion (oft auch stattdessen Gewinnfunktion)

$$L : \begin{cases} D \times \Theta & \longrightarrow & \mathbb{R} \\ (\mathbf{d}, \boldsymbol{\theta}) & \longmapsto & L(\mathbf{d}, \boldsymbol{\theta}) \end{cases}$$

ordnet einer Entscheidung $\mathbf{d}(\mathbf{x})$ *(„decision") einen Verlust („loss") zu. Im Allgemeinen ist L so gewählt, dass der Verlust bei richtiger Entscheidung null ist, also L eine nicht-negative Funktion ist.*

Beispiel 1.5

1. **Test**: Betrachte

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

(zum Beispiel Gauß-Test).
Der Entscheidungsraum sei $D = \{d_0, d_1\}$ mit

d_0: Entscheidung für H_0,

d_1: Entscheidung für H_1.

Eine mögliche Verlustfunktion ist:

$$L(d_0, \theta) = \begin{cases} 0, & \text{falls } \theta \leq \theta_0 \quad \text{(Entscheidung richtig)} \\ a \in \mathbb{R}_+, & \text{falls } \theta > \theta_0 \quad \text{(Fehler 2. Art)} \end{cases}$$

$$L(d_1, \theta) = \begin{cases} 0, & \text{falls } \theta > \theta_0 \quad \text{(Entscheidung richtig)} \\ b \in \mathbb{R}_+, & \text{falls } \theta \leq \theta_0 \quad \text{(Fehler 1. Art)} \end{cases}$$

Beispiel 1.5 fortgeführt

2. **Schätzung**: „Entscheidung" ist reelle Zahl:

$$d(\mathbf{x}) = T(\mathbf{x}) = \widehat{\theta} \in \Theta, \text{ d.h. } D = \Theta.$$

Mögliche Verlustfunktionen:

$$L(d,\theta) = (d-\theta)^2 \qquad \text{quadratischer Verlust,}$$
$$L(d,\theta) = |d-\theta| \qquad \text{absoluter Verlust,}$$
$$L(d,\theta) = w(\theta)(d-\theta)^2 \qquad \text{gewichteter quadratischer Verlust,}$$

wobei w eine feste Gewichtsfunktion ist.

Beispiel 1.5 fortgeführt

3. **Mehrentscheidungsverfahren**, zum Beispiel Wahl zwischen drei Alternativen

$$d_0 : \theta \leq \theta_0, \quad d_1 : \theta > \theta_1, \quad d_2 : \theta_0 < \theta \leq \theta_1.$$

4. Analog: Modellwahl, Variablenselektion

Definition 1.5 (Risikofunktion)

Eine Risikofunktion ist definiert als

$$R(\mathbf{d}, \boldsymbol{\theta}) = \mathbb{E}_\theta[L(\mathbf{d}(\mathbf{X}), \boldsymbol{\theta})] = \int_{\mathcal{X}} L(\mathbf{d}(\mathbf{x}), \boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) \, d\mathbf{x}$$

(„Verlust im Mittel"). Sie ist unabhängig von $\mathbf{x}$*. Dabei wird* $\mathbf{d}(\mathbf{X})$ *rausintegriert, d.h.* $R(\mathbf{d}; \boldsymbol{\theta})$ *ist bei gegebenem* $\mathbf{d}$ *nur noch eine Funktion von* $\boldsymbol{\theta}$*.*

Beispiel 1.6

1. **Schätzen**, d.h.

$$\mathbf{d}(\mathbf{x}) = \mathbf{T}(\mathbf{x}) \quad \text{Schätzwert}, \quad \mathbf{d}(\mathbf{X}) = \mathbf{T}(\mathbf{X}) \quad \text{Punktschätzer}.$$

Bei quadratischer Verlustfunktion ist

$$L(\mathbf{T}(\mathbf{X}), \boldsymbol{\theta}) = \|\mathbf{T}(\mathbf{X}) - \boldsymbol{\theta}\|^2$$

mit Risikofunktion

$$R(\mathbf{T}, \boldsymbol{\theta}) = \mathbb{E}_\theta[\|\mathbf{T}(\mathbf{X}) - \boldsymbol{\theta}\|^2] = \mathrm{MSE}_\theta(\mathbf{T}(\mathbf{X})).$$

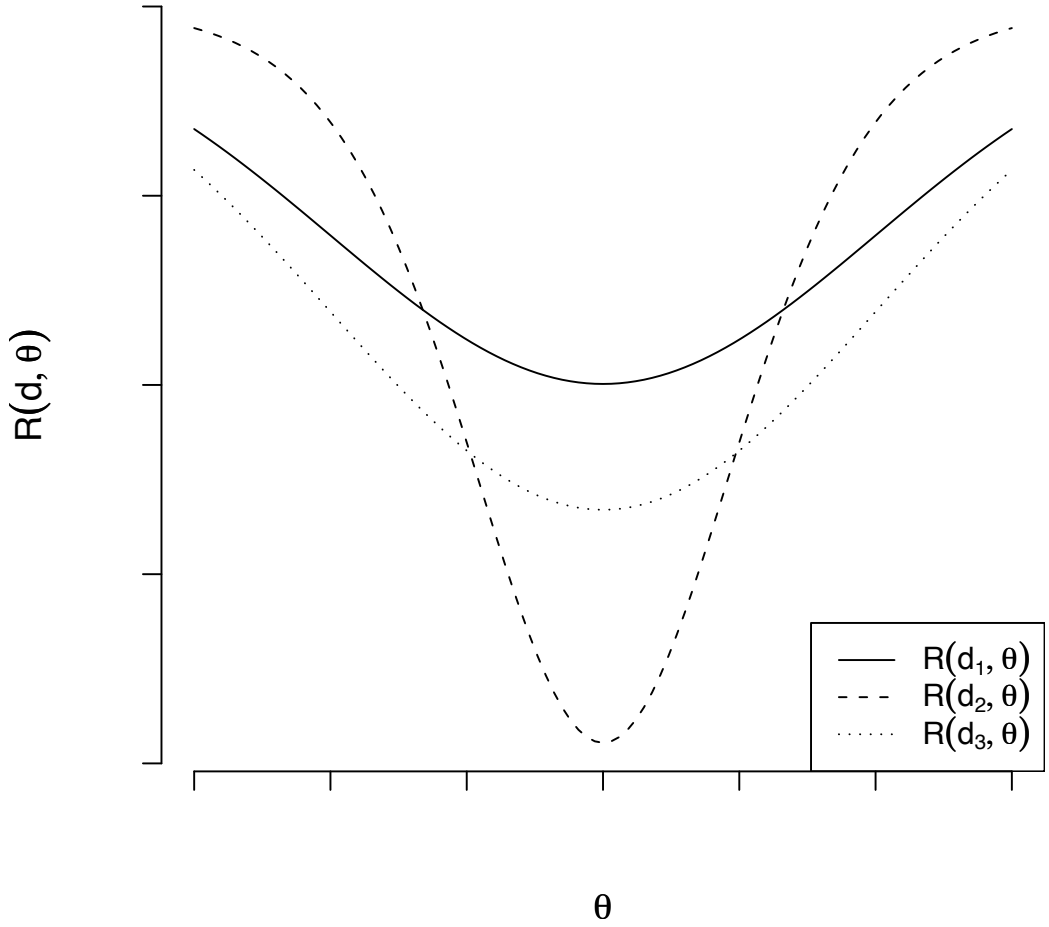
Man beachte, dass das Argument $\mathbf{T}$ in $R(\mathbf{T}, \boldsymbol{\theta})$ den Schätzer und nicht den konkreten Schätzwert bezeichnet.

2. **Testen**: vgl. Übung.

1.2 Konzepte der statistischen Inferenz

1.2.5 Statistische Entscheidungstheorie

Vergleich von Entscheidungsregeln mittels der Risikofunktion



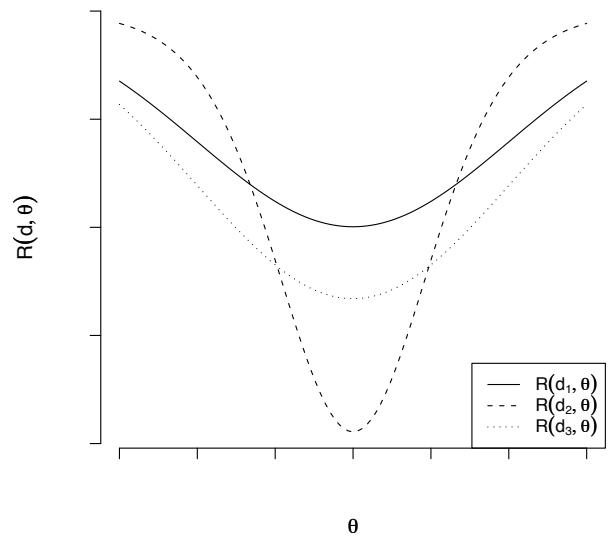
Aus der Abbildung geht hervor, dass d_3 besser als d_1 ist für alle $\theta \in \Theta$, d.h. d_3 dominiert d_1 gleichmäßig.

1.2 Konzepte der statistischen Inferenz

1.2.5 Statistische Entscheidungstheorie

Ziel: Finde Regel d^*, die alle „konkurrierenden" Regeln d dominiert.

Problem: Diese Idee funktioniert im Allgemeinen nicht, in der Regel überschneiden sich die Risikofunktionen, zum Beispiel ist in obiger Abbildung d_2 nur in einem gewissen Bereich besser als d_1 und d_3.



$\longrightarrow$ „Optimale" Entscheidungsregeln nur möglich durch:

- Einschränkung auf spezielle Klassen von Verlustfunktionen,
- Einschränkung auf spezielle Klassen von Entscheidungsregeln, zum Beispiel unverzerrter Schätzer oder unverfälschter Test,
- oder zusätzliches Kriterium.

Kriterien für "Optimale" Entscheidungsregeln

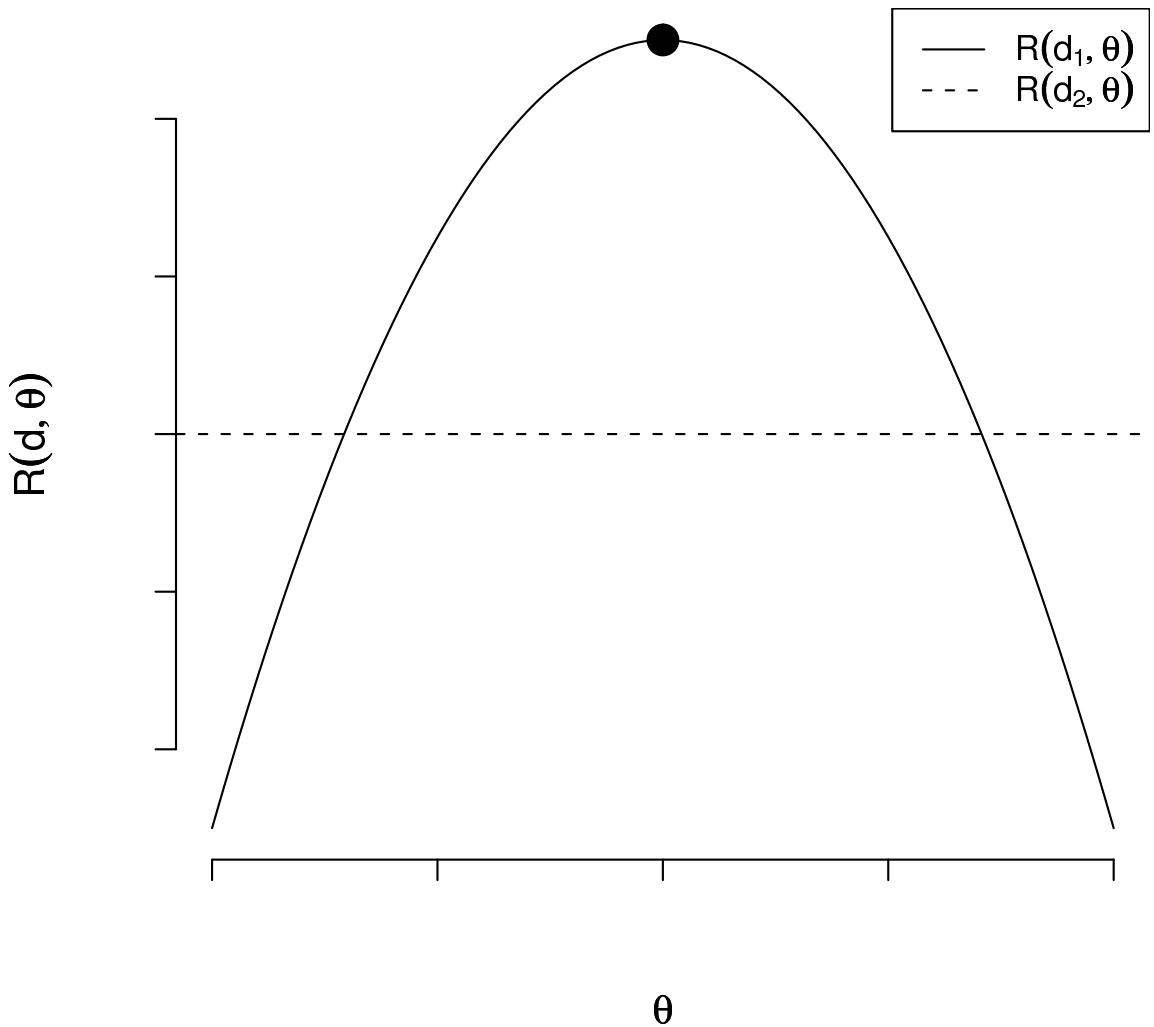
1. **Minimax-Kriterium**
   Idee: Betrachte Maximum der Risikofunktion, d.h. präferiere in der folgenden Abbildung d_2, da

$$\max_{\theta} R(d_2, \theta) < \max_{\theta} R(d_1, \theta).$$

1.2 Konzepte der statistischen Inferenz

1.2.5 Statistische Entscheidungstheorie

$$\max_{\theta} R(d_2, \theta) < \max_{\theta} R(d_1, \theta).$$



Definition 1.6 (Minimax-Entscheidungsregel)

Sei $\mathbf{d}^* : \mathcal{X} \to D$ *eine Entscheidungsregel.* $\mathbf{d}^*$ *heißt Minimax, falls es das supremale Risiko minimiert:*

$$\sup_{\boldsymbol{\theta}\in\Theta} R(\mathbf{d}^*, \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta}\in\Theta} R(\mathbf{d}, \boldsymbol{\theta}) \ \ \forall \mathbf{d} \in D \ \Leftrightarrow \ \mathbf{d}^* = \operatorname*{arginf}_{\mathbf{d}\in D} \sup_{\boldsymbol{\theta}\in\Theta} R(\mathbf{d}, \boldsymbol{\theta}).$$

Bemerkung. In vielen Fällen werden Supremum und Infimum auch angenommen, so dass tatsächlich

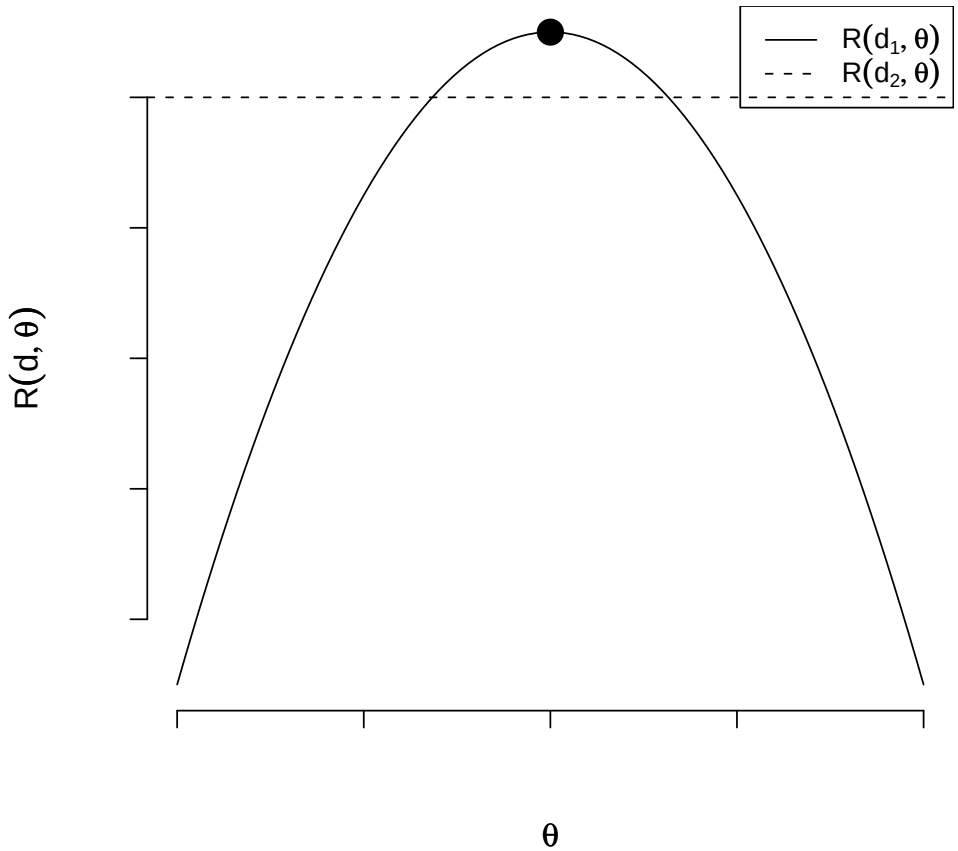
$$\mathbf{d}^* = \operatorname*{argmin}_{\mathbf{d}\in D} \max_{\boldsymbol{\theta}\in\Theta} R(\mathbf{d}, \boldsymbol{\theta})$$

gilt, daher auch der Name Minimax.

Beim Minimax-Kriterium schützt man sich gegen den schlimmsten Fall, was aber nicht unbedingt immer vernünftig ist, wie die folgende Abbildung zeigt:

1.2 Konzepte der statistischen Inferenz

1.2.5 Statistische Entscheidungstheorie



Hier wäre d^* nur dann vernünftig, wenn θ-Werte in der Mitte „besonders wahrscheinlich" sind.

Kriterien für "Optimale" Entscheidungsregeln

2. **Bayes-Kriterium**
   Wie in der Bayes-Inferenz nehmen wir für $\boldsymbol{\theta}$ eine Prioridichte $p(\boldsymbol{\theta})$ an (aus frequentistischer Sichtweise ist $p(\boldsymbol{\theta})$ eine – nicht notwendigerweise normierte – Gewichtsfunktion).
   Das *Bayes-Risiko* ist

$$\begin{aligned} r(\mathbf{d}, p) &= \int_{\Theta} R(\mathbf{d}, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \mathbb{E}_p[R(\mathbf{d}, \boldsymbol{\theta})] \\ &= \mathbb{E}_p \mathbb{E}_\theta[L(\mathbf{d}(\mathbf{X}), \boldsymbol{\theta})] \\ &= \int_{\Theta} \int_{\mathcal{X}} L(\mathbf{d}(\mathbf{x}), \boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}\; p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \end{aligned}$$

2. **Bayes-Kriterium** fortgeführt:
   Das *Bayes-Risiko* wird durch die *Bayes-optimale Regel* $\mathbf{d}^*$ minimiert:

$$r(\mathbf{d}^*, p) = \inf_{\mathbf{d} \in D} r(\mathbf{d}, p).$$

   Sei $p(\boldsymbol{\theta}|\mathbf{x})$ (eigentliche) Posterioridichte. Dann heißt

$$\int_{\Theta} L(\mathbf{d}(\mathbf{x}), \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) \, d\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}[L(\mathbf{d}(\mathbf{x}), \boldsymbol{\theta})]$$

   das *Posteriori-Bayes-Risiko*.

Es gilt folgendes praktisches Resultat:

Satz 1.7

Eine Regel $\mathbf{d}^*$ *ist genau dann Bayes-optimal, wenn* $\mathbf{d}^*(\mathbf{x})$ *für jede Beobachtung/Stichprobe* $\mathbf{x}$ *das Posteriori-Bayes-Risiko minimiert.*

Anmerkungen:

- Satz 1.8. erleichtert die Bestimmung der Bayes-optimalen Entscheidungsregel in konkreten Fällen.
- Er zeigt eine intuitive Eigenschaft des Bayesianischen Vorgehens: Um eine Entscheidung geg. eine Beobachtung $\mathbf{x}$ zu treffen, reicht es, den Verlust für $\mathbf{d}(\mathbf{x})$ zu betrachten. Es ist nicht nötig, Verluste für $\mathbf{d}(\mathbf{X})$ für andere mögliche aber nicht beobachtete $\mathbf{X}$ zu berücksichtigen.
- Bayes-optimale Regeln $\mathbf{d}^*$ sind *zulässig*, d.h. sie werden von keiner anderen Regel $\mathbf{d} \neq \mathbf{d}^*$ dominiert.
- Eine enge Beziehung zur Minimax-Regel ist durch die Wahl einer „ungünstigsten" Prioridichte $p^*(\boldsymbol{\theta})$ herstellbar.

Optimalität von Bayes-Schätzern:

- $$\widehat{\theta} = \mathbb{E}[\theta|\mathbf{x}] = \int_{\Theta} \theta \; p(\theta|\mathbf{x}) \; d\theta$$

  ist Bayes-optimal bei quadratischer Verlustfunktion $L(d, \theta) = (d - \theta)^2$.

- $$\widehat{\theta} = \text{med}(\theta|\mathbf{x})$$

  ist Bayes-optimal bei absoluter Verlustfunktion $L(d, \theta) = |d - \theta|$.

Optimalität von Bayes-Schätzern fortgeführt:

- Der Posteriori-Modus

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(\theta|\mathbf{x})$$

ist Bayes-optimal bei 0-1 Verlustfunktion

$$L_\varepsilon(d, \theta) = \begin{cases} 1, & \text{falls } |d - \theta| \geq \varepsilon, \\ 0, & \text{falls } |d - \theta| < \varepsilon \end{cases}$$

und Grenzübergang $\varepsilon \longrightarrow 0$.

- Anmerkung: Die ML-Schätzung ist optimal bei flacher Priori $p(\theta) \propto 1$ und bei Wahl obiger 0-1-Verlustfunktion.